Πιθανότητες Στατιστική

Πολυχρόνης Οικονόμου
Σόνια Μαλεφάκη
Απόστολος Μπατσίδης

ΚΑΛΛΙΠΟΣ ανοικτές ακαδημαϊκές εκδόσεις

Εθνικό Πρόγραμμα Ανάπτυξης 2021-2025

ΠΙΘΑΝΟΤΗΤΕΣ - ΣΤΑΤΙΣΤΙΚΗ

Πιθανότητες - Στατιστική

Εφαρμογές με χρήση της R

Πολυχρόνης Οικονόμου
Αναπληρωτής Καθηγητής
Πανεπιστήμιο Πατρών

Σόνια Μαλεφάκη
Επίκουρη Καθηγήτρια
Πανεπιστήμιο Πατρών

Απόστολος Μπατσίδης
Αναπληρωτής Καθηγητής
Πανεπιστήμιο Ιωαννίνων

ΚΑΛΛΙΠΟΣ
ΑΝΟΙΚΤΕΣ ΑΚΑΔΗΜΑΪΚΕΣ ΕΚΔΟΣΕΙΣ
www.kallipos.gr

Τίτλος πρωτοτύπου: «*Πιθανότητες - Στατιστική*»

Στην κόρη μου, Σοφιάννα
Π.Ο.

Στον γιο μου, Αλέξανδρο
Σ.Μ.

Στις κόρες μου, Ελένη και Όλγα
Α.Μ.

Συντελεστές έκδοσης
Γλωσσική επιμέλεια: Βασιλική Τυραϊδή

Κεντρική Ομάδα Υποστήριξης
Γλωσσικός Έλεγχος: Γεωργία Τριανταφυλλίδου

Γραφιστικός Έλεγχος: Χρήστος Κεντρωτής

Βιβλιοθηκονομική Επεξεργασία: Έλενα Αδαμοπούλου

ΚΑΛΛΙΠΟΣ

Εθνικό Μετσόβιο Πολυτεχνείο

Ηρώων Πολυτεχνείου 9

15780 Ζωγράφου

www.kallipos.gr

Βιβλιογραφική αναφορά: Οικονόμου, Π., Μαλεφάκη, Σ., & Μπατσίδης, Α. (2023). *Πιθανότητες - Στατιστική* [Προπτυχιακό εγχειρίδιο]. Κάλλιπος, Ανοικτές Ακαδημαϊκές Εκδόσεις.

Διαθέσιμο στο: http://dx.doi.org/10.57713/kallipos-101

ISBN: **978-618-5667-85-6**

ΠΕΡΙΕΧΟΜΕΝΑ

Οικονόμου, Π., Μαλεφάκη, Σ., & Μπατσίδης, Α. (2023). *Πιθανότητες - Στατιστική.*
[Προπτυχιακό εγχειρίδιο].

ΚΑΤΑΛΟΓΟΣ ΣΧΗΜΑΤΩΝ

Οικονόμου, Π., Μαλεφάκη, Σ., & Μπατσίδης, Α. (2023). *Πιθανότητες - Στατιστική.*

ΚΑΤΑΛΟΓΟΣ ΠΙΝΑΚΩΝ ΚΑΤΑΝΟΜΩΝ

ΠΡΟΛΟΓΟΣ

Τις τελευταίες δεκαετίες αυξάνεται συνεχώς η αναγνώριση του κυρίαρχου ρόλου της πιθανοθεωρίας και των στατιστικών μεθόδων στην επίλυση προβλημάτων σε διάφορα επιστημονικά πεδία. Για τον λόγο αυτόν, οι Πιθανότητες και η Στατιστική διδάσκονται σε πληθώρα τμημάτων της τριτοβάθμιας εκπαίδευσης πέραν των Τμημάτων Μαθηματικών και Στατιστικής. Βασικός στόχος της συγγραφικής ομάδας ήταν η δημιουργία ενός εύληπτου συγγράμματος για όλους τους φοιτητές τριτοβάθμιων ιδρυμάτων οι οποίοι διδάσκονται για πρώτη φορά το προαναφερθέν αντικείμενο, εφοδιάζοντάς τους με όλο το απαραίτητο θεωρητικό υπόβαθρο που θα τους επιτρέπει να επιλέγουν και να εφαρμόζουν την κατάλληλη μεθοδολογία για την επίλυση προβλημάτων της επιστημονικής περιοχής τους.

Ως εκ τούτου, η ύλη του παρόντος βιβλίου ανά κεφάλαιο αναλύεται παρακάτω. Στο **Κεφάλαιο 1** παρουσιάζεται η έννοια του πειράματος τύχης, του δειγματικού χώρου, των ενδεχομένων και των πράξεων μεταξύ τους. Εν συνεχεία, γίνεται αναφορά στους διαφορετικούς ορισμούς της πιθανότητας και παρουσιάζονται τα βασικά εργαλεία υπολογισμού της πιθανότητας ενός ενδεχομένου σε ένα πείραμα τύχης, ενώ αναπτύσσονται οι βασικές μέθοδοι για την επίλυση προβλημάτων υπολογισμού πιθανοτήτων. Στο **Κεφάλαιο 2** παρουσιάζονται έννοιες, όπως η δεσμευμένη πιθανότητα και η ανεξαρτησία ενδεχομένων, και αναπτύσσονται μέθοδοι με τη βοήθεια θεμελιωδών θεωρημάτων για την επίλυση σύνθετων προβλημάτων υπολογισμού πιθανοτήτων. Στο **Κεφάλαιο 3** ορίζεται η έννοια της τυχαίας μεταβλητής. Μετά τον ορισμό των τυχαίων μεταβλητών και, αφού αυτές διαχωριστούν σε διακριτές και συνεχείς, εισάγονται και μελετώνται οι έννοιες της αθροιστικής συνάρτησης κατανομής, της συνάρτησης πιθανότητας (για διακριτές τυχαίες μεταβλητές) και της συνάρτησης πυκνότητας πιθανότητας (για συνεχείς τυχαίες μεταβλητές). Επιπροσθέτως, ορίζονται και μελετώνται αριθμητικά χαρακτηριστικά της τυχαίας μεταβλητής και της αντίστοιχης κατανομής της, η γνώση των οποίων μας δίνει χρήσιμες πληροφορίες για την τυχαία μεταβλητή. Τέλος, παρουσιάζονται μεθοδολογίες για την εύρεση της κατανομής μιας συνάρτησης τυχαίας μεταβλητής. Αντικείμενο μελέτης του **Κεφαλαίου 4** αποτελούν οι βασικές διακριτές τυχαίες μεταβλητές που μοντελοποιούν πιθανοθεωρητικά ευρύ φάσμα προβλημάτων και τυχαίων φαινομένων σε διάφορα επιστημονικά πεδία. Στο πλαίσιο αυτό, παρουσιάζονται οι ακόλουθες διακριτές κατανομές: η διακριτή ομοιόμορφη, η διωνυμική, η υπεργεωμετρική, η γεωμετρική, η αρνητική διωνυμική και η Poisson, ενώ στο **Κεφάλαιο 5** οι έννοιες που παρουσιάστηκαν στο Κεφάλαιο 3 αξιοποιούνται για τη μελέτη βασικών συνεχών

Οικονόμου, Π., Μαλεφάκη, Σ., & Μπατσίδης, Α. (2023). *Πιθανότητες - Στατιστική.*
[Προπτυχιακό εγχειρίδιο].

τυχαίων μεταβλητών και των κατανομών τους, που περιγράφουν ένα ευρύ φάσμα προβλημάτων και τυχαίων φαινομένων σε διάφορα επιστημονικά πεδία. Στο πλαίσιο αυτό, παρουσιάζονται οι ακόλουθες συνεχείς κατανομές: η ομοιόμορφη, η εκθετική, η Weibull, η λογαριθμοκανονική, η γάμμα, η βήτα και, τέλος, η σπουδαιότερη όλων, η κανονική κατανομή.

Στο **Κεφάλαιο 6** παρουσιάζονται, εν συντομία, οι πολυδιάστατες τυχαίες μεταβλητές και οι βασικότερες ιδιότητές τους. Επιπλέον, παρουσιάζονται οι έννοιες της δεσμευμένης τυχαίας μεταβλητής, της στοχαστικής ανεξαρτησίας και κάποιες ειδικές πολυδιάστατες τυχαίες μεταβλητές.

Στο **Κεφάλαιο 7**, χρησιμοποιώντας τις μεθόδους της αθροιστικής συνάρτησης κατανομής, του μετασχηματισμού και της ροπογεννήτριας συνάρτησης προσδιορίζονται οι κατανομές συναρτήσεων τυχαίων μεταβλητών. Στο πλαίσιο αυτό, εισάγονται και παρουσιάζονται τρεις πολύ σημαντικές κατανομές, η χι-τετράγωνο, η t και η F κατανομή. Έπειτα, το ενδιαφέρον επικεντρώνεται στον προσδιορισμό της κατανομής του αθροίσματος ανεξάρτητων τυχαίων μεταβλητών. Το κεφάλαιο ολοκληρώνεται με την παρουσίαση ενός θεμελιώδους θεωρήματος, σύμφωνα με το οποίο το άθροισμα ή η μέση τιμή n το πλήθος ανεξάρτητων και ισόνομων τυχαίων μεταβλητών από μια κατανομή με πεπερασμένη μέση τιμή και διακύμανση, ακολουθεί για μεγάλες τιμές του n, προσεγγιστικά την κανονική κατανομή. Το θεώρημα αυτό, που είναι ένα από τα σπουδαιότερα θεωρήματα της Θεωρίας Πιθανοτήτων και της Στατιστικής, είναι γνωστό ως Κεντρικό Οριακό Θεώρημα.

Στο **Κεφάλαιο 8** το ενδιαφέρον μας επικεντρώνεται στην έννοια του τυχαίου δείγματος και των μεθόδων συνοπτικής παρουσίασης ποιοτικών και ποσοτικών δεδομένων μέσω πινάκων, γραφημάτων και περιγραφικών μέτρων. Τέλος, αξίζει να σημειωθεί ότι το κεφάλαιο αυτό λειτουργεί ως συνδετικός κρίκος μεταξύ της Θεωρίας Πιθανοτήτων και των τυχαίων μεταβλητών, που παρουσιάστηκαν στα προηγούμενα κεφάλαια, και της Στατιστικής Συμπερασματολογίας που θα αποτελέσει το κεντρικό αντικείμενο μελέτης των επόμενων κεφαλαίων. Ειδικότερα, στο **Κεφάλαιο 9** εστιάζουμε στη δειγματοληψία από πληθυσμούς με συγκεκριμένα χαρακτηριστικά και θα μελετήσουμε σημαντικές ποσότητες, όπως είναι η δειγματική μέση τιμή, η δειγματική διασπορά και συναρτήσεις αυτών, οι οποίες παίζουν καθοριστικό ρόλο στη στατιστική συμπερασματολογία η οποία παρουσιάζεται εκτενώς στα επόμενα κεφάλαια. Στο **Κεφάλαιο 10** παρουσιάζονται, εν συντομία, βασικές αρχές και μέθοδοι για την εκτίμηση των άγνωστων παραμέτρων ενός πιθανοθεωρητικού μοντέλου σε σημείο και σε διάστημα. Στο πλαίσιο αυτό παρατίθενται αποτελέσματα που αφορούν την εκτίμηση της μέσης τιμής κανονικού πληθυσμού, της διαφοράς των μέσων τιμών κανονικών πληθυσμών με ανεξάρτητα ή εξαρτημένα δείγματα, τη διασπορά κανονικού πληθυσμού, το πηλίκο των διασπορών δύο κανονικών πληθυσμών, τη διωνυμική πιθανότητα και τη διαφορά δύο ποσοστών. Το **Κεφάλαιο 11** έχει ως κύριο στόχο του να παρουσιάσει τους βασικότερους παραμετρικούς στατιστικούς ελέγχους υποθέσεων. Πιο συγκεκριμένα, παρουσιάζονται έλεγχοι για τη μέση τιμή, την αναλογία, τη διασπορά ενός πληθυσμού, αλλά και για τη διαφορά των μέσων και των αναλογιών ή τον λόγο των διασπορών δύο πληθυσμών. Στο **Κεφάλαιο 12** παρουσιάζεται το απλό γραμμικό μοντέλο. Ιδιαίτερη έμφαση δίνεται στη διαδικασία υπολογισμού διάφορων απαραίτητων ποσοτήτων για την εξαγωγή συμπερασμάτων σχετικά με τη γραμμική σχέση δύο μεταβλητών. Τέλος, γίνεται αναλυτική παρουσίαση της χρήσης της R στην προσαρμογή και αξιολόγηση του απλού γραμμικού μοντέλου. Στο **Κεφάλαιο 13** γενικεύεται ο απλός έλεγχος t για τη διαφορά των μέσων τιμών δύο ανεξάρτητων πληθυσμών σε έναν έλεγχο για την ισότητα των μέσων τιμών k το πλήθος ανεξάρτητων πληθυσμών, υπό τις υποθέσεις της κανονικότητας και της ισότητας των διασπορών των πληθυσμών. Επίσης, προτείνονται τεχνικές για επακόλουθες πολλαπλές ζευγαρωτές συγκρίσεις (post-hoc ανάλυση) στην περίπτωση απόρριψης της αρχικής μηδενικής υπόθεσης και στην προσπάθεια διερεύνησης περαιτέρω διαφορών μεταξύ των μέσων τιμών των πληθυσμών. Το παρόν σύγγραμμα ολοκληρώνεται με δύο παραρτήματα και το ευρετήριο όρων. Στο πρώτο παράρτημα παρατίθενται χρήσιμοι πίνακες που αφορούν πολύ γνωστές κατανομές, ενώ στο δεύτερο παράρτημα δίνονται χρήσιμες μαθηματικές γνώσεις για την καλύτερη κατανόηση κάποιων αποδείξεων των υπόλοιπων κεφαλαίων του παρόντος συγγράμματος.

Αξίζει να επισημανθεί ότι σε κάθε κεφάλαιο του παρόντος συγγράμματος περιέχονται βιβλιογραφικές αναφορές για περαιτέρω μελέτη από τον/την ενδιαφερόμενο/όμενη αναγνώστη/στρια, ενώ υπάρχουν ασκήσεις αυτοαξιολόγησης. Η συγγραφική ομάδα προτείνει να λύνονται αυτές οι ασκήσεις παράλληλα με τη μελέτη του κάθε κεφαλαίου και να γίνεται αυτοαξιολόγηση για την κατανόηση του υλικού που έχει προηγηθεί, με βάση τις αναλυτικές λύσεις που δίνονται στο τέλος του κάθε κεφαλαίου. Εν αντιθέσει, η συλλογή των άλυτων ασκήσεων που παρατίθενται στο τέλος του κάθε κεφαλαίου προτείνεται να αποτελεί αντικείμενο μελέτης μετά την ολοκλήρωση κάθε κεφαλαίου και, για αυτόν τον λόγο, οι ασκήσεις αυτές δεν ακολουθούν τη σειρά εμφάνισης των εννοιών στο κάθε κεφάλαιο. Τέλος, αξίζει να αναφερθεί ότι, κατά την επίλυση των παραδειγμάτων ή και των ασκήσεων αυτοαξιολόγησης, πολλές φορές χρησιμοποιείται η R στατιστική γλώσσα προγραμματισμού.

Το παρόν βιβλίο μπορεί να χρησιμοποιηθεί ως ένα εισαγωγικό σύγγραμμα σε μαθήματα Πιθανοτήτων ή Στατιστικής ή Πιθανοτήτων - Στατιστικής. Ειδικότερα, για τη διδασκαλία ενός εισαγωγικού μαθήματος Πιθανοτήτων κρίνεται επαρκής η ύλη των Κεφαλαίων 1-7 με πλήρη έμφαση στις αποδείξεις και στο θεωρητικό υπόβαθρο, ενώ για τη διδασκαλία ενός εισαγωγικού μαθήματος στη Στατιστική θεωρείται επαρκής η ύλη των Κεφαλαίων 8-13 με πλήρη έμφαση στα θεωρητικά αποτελέσματα. Τέλος, για τη διδασκαλία ενός εισαγωγικού μαθήματος Πιθανοτήτων-Στατιστικής ή για τη διδασκαλία του μαθήματος σε Τμήματα άλλα πέραν των Μαθηματικών και Στατιστικής προτείνεται η κάλυψη της ύλης των Κεφαλαίων 1-6 και 8-13, χωρίς να δίνεται έμφαση στις μαθηματικές αποδείξεις. Στο σημείο αυτό να αναφέρουμε ότι η παράθεση προαπαιτούμενων γνώσεων, η καταγραφή των προσδοκώμενων μαθησιακών αποτελεσμάτων και η αναφορά επιστημονικών όρων στην αρχή κάθε κεφαλαίου, σε συνδυασμό με την παράθεση άλυτων ασκήσεων αλλά και ασκήσεων αυτοαξιολόγησης, επιτρέπει τη χρήση του συγγράμματος αυτού και στο πλαίσιο της εξ αποστάσεως εκπαίδευσης ενός εισαγωγικού μαθήματος Πιθανοτήτων και Στατιστικής.

Οι συγγραφείς επιθυμούν να ευχαριστήσουν, από τη θέση αυτή, τους πανεπιστημιακούς τους δασκάλους που τους μετέδωσαν την αγάπη για τις Πιθανότητες και τη Στατιστική, αλλά και το πάθος τους για τη διδασκαλία. Ιδιαίτερες όμως ευχαριστίες αποδίδονται στις οικογένειές τους για την υπομονή που έδειξαν καθόλη τη διάρκεια της συγγραφής του βιβλίου.

Ευχόμαστε το παρόν σύγγραμμα να αποτελέσει έναν πλήρη οδηγό για όλους όσοι διδάσκουν ή διδάσκονται ή απλώς επιθυμούν να πρωτογνωρίσουν τον κόσμο των Πιθανοτήτων και της Στατιστικής. Παρότι έγινε προσπάθεια ελαχιστοποίησης λαθών, αστοχιών και παραλείψεων, με μεγάλη πιθανότητα τέτοια έχουν παραμείνει στο παρόν κείμενο. Για κάθε παράλειψη, λάθος, υπεύθυνα είναι όλα τα μέλη της συγγραφικής ομάδας και κάθε παρατήρηση και σχόλιο για βελτίωση του παρόντος συγγράμματος είναι ευπρόσδεκτα.

Μέρος Ι

ΠΙΘΑΝΟΤΗΤΕΣ

ΚΕΦΑΛΑΙΟ 1

Η ΕΝΝΟΙΑ ΤΗΣ ΠΙΘΑΝΟΤΗΤΑΣ

Σύνοψη

Σε αυτό το κεφάλαιο παρουσιάζονται η έννοια της πιθανότητας και οι διάφοροι ορισμοί της, τα βασικά εργαλεία υπολογισμού της πιθανότητας ενός ενδεχομένου σε ένα πείραμα τύχης, ενώ αναπτύσσονται οι βασικές μέθοδοι για την επίλυση προβλημάτων υπολογισμού πιθανοτήτων.

Προαπαιτούμενη γνώση: Βασικές γνώσεις μαθηματικών.

Προσδοκώμενα μαθησιακά αποτελέσματα: Όταν θα έχετε ολοκληρώσει τη μελέτη του κεφαλαίου αυτού, θα γνωρίζετε

- την έννοια και τους διάφορους ορισμούς της πιθανότητας,
- τις βασικές ιδιότητες των πιθανοτήτων και
- τις βασικές αρχές απαρίθμησης.

Γλωσσάριο επιστημονικών όρων:

- Ασυμβίβαστα/ξένα ενδεχόμενα
- Βασική αρχή απαρίθμησης
- Δειγματοχώρος
- Διατάξεις
- Δυναμοσύνολο
- Ενδεχόμενο
- Ένωση συνόλων/ενδεχομένων
- Μεταθέσεις
- Πείραμα τύχης
- Συμπλήρωμα συνόλου/ενδεχομένου
- Συνδυασμοί
- Τομή συνόλων/ενδεχομένων

Οικονόμου, Π., Μαλεφάκη, Σ., & Μπατσίδης, Α. (2023). *Πιθανότητες - Στατιστική.*

1.1 Εισαγωγή

Οι περισσότερες ανθρώπινες δραστηριότητες αλλά και τα περισσότερα φυσικά φαινόμενα διέπονται από έναν βαθμό αβεβαιότητας ως προς την έκβασή τους. Ακόμα και αν αυτά επαναλαμβάνονται υπό τις ίδιες ακριβώς συνθήκες, δεν είναι πάντα εφικτό να προβλέψουμε με βεβαιότητα ή να προσδιορίσουμε με ακρίβεια την εξέλιξή τους. Παραδείγματα τέτοιων δραστηριοτήτων ή φαινομένων περιλαμβάνουν, μεταξύ άλλων:

- την κατανάλωση καυσίμου από ένα αυτοκίνητο σε μια διαδρομή,
- την αντοχή και τη συμπεριφορά μιας νεόδμητης οικοδομής σε έναν σεισμό,
- το ύψος σε εκατοστά της ετήσιας βροχόπτωσης,
- τον χρόνο ολοκλήρωσης μιας υπολογιστικής διαδικασίας σε έναν ηλεκτρονικό υπολογιστή,
- την επίδοση ενός φοιτητή σε μια εξέταση,
- τον μηνιαίο αριθμό των τροχαίων ατυχημάτων σε ένα συγκεκριμένο οδικό δίκτυο,
- τον αριθμό των ελαττωματικών ανταλλακτικών σε μια παρτίδα 1000 κομματιών,
- την εμφάνιση κυκλοφοριακής συμφόρησης ή όχι σε μια λεωφόρο,
- το αποτέλεσμα μιας εγχείρησης κ.ά.

Στο κεφάλαιο αυτό παρουσιάζονται με τη βοήθεια της θεωρίας Πιθανοτήτων βασικά εργαλεία και κατάλληλες μέθοδοι για την περιγραφή και κατανόηση φαινομένων αντίστοιχων με τα άνωθεν.

1.2 Βασικές έννοιες

Προτού εξηγήσουμε αναλυτικά την έννοια της πιθανότητας, είναι χρήσιμο να παρουσιαστούν κάποιες βασικές έννοιες, οι οποίες θα διευκολύνουν την παρουσίασή της, καθώς και τον σαφή καθορισμό του πεδίου εφαρμογών της. Πρωταρχικό ρόλο για την επίτευξη των προαναφερθέντων στόχων διαδραματίζει η έννοια του **τυχαίου πειράματος** ή **πειράματος τύχης**. Σύμφωνα με τον Siegel (2011): «Ένα τυχαίο πείραμα είναι οποιαδήποτε καλά καθορισμένη διαδικασία που παράγει ένα παρατηρήσιμο αποτέλεσμα που δεν θα μπορούσε να προβλεφθεί επακριβώς εκ των προτέρων». Ενδεικτικά παραδείγματα τυχαίων πειραμάτων είναι οι δραστηριότητες και τα φυσικά φαινόμενα που αναφέρθηκαν στην εισαγωγή του παρόντος κεφαλαίου. Τα τυχαία πειράματα αποτελούν το πεδίο εφαρμογής της θεωρίας Πιθανοτήτων που θα αναπτυχθεί στη συνέχεια.

Μια έννοια στενά συνδεδεμένη με το τυχαίο πείραμα είναι η έννοια του **δειγματοχώρου**.

Ορισμός 1.1

Δειγματοχώρος ή **δειγματικός χώρος** ενός πειράματος τύχης είναι το σύνολο που περιλαμβάνει τα δυνατά αποτελέσματα του πειράματος τύχης και συμβολίζεται, συνήθως, με Ω.

Είναι σημαντικό να αναφερθεί ότι το πλήθος $\nu(\Omega)$ των στοιχείων ενός δειγματοχώρου, δηλαδή το πλήθος των δυνατών αποτελεσμάτων ενός πειράματος τύχης, μπορεί να είναι πεπερασμένο ή άπειρο (αριθμήσιμο ή υπεραριθμήσιμο).

Ορισμός 1.2

Κάθε υποσύνολο A ενός δειγματοχώρου Ω ονομάζεται **ενδεχόμενο**.
Απλό ή **στοιχειώδες ενδεχόμενο** θα ονομάζεται κάθε μονοσύνολο ενδεχόμενο, δηλαδή κάθε υποσύνολο ενός δειγματοχώρου Ω που αποτελείται από ένα και μόνο στοιχείο του.

Όπως θα δούμε στη συνέχεια, η πιθανότητα ορίζεται μόνο για τα ενδεχόμενα, δηλαδή τα υποσύνολα, του δειγματοχώρου ενός πειράματος τύχης. Αξίζει να σημειωθεί ότι κάθε δειγματοχώρος περιλαμβάνει σίγουρα δύο ενδεχόμενα: το βέβαιο και το αδύνατο ενδεχόμενο, που ορίζονται στη συνέχεια.

Ορισμός 1.3

Το **βέβαιο ενδεχόμενο** είναι το ενδεχόμενο που πραγματοποιείται πάντοτε, και επομένως, ταυτίζεται με τον δειγματοχώρο Ω, ενώ το **αδύνατο ενδεχόμενο** είναι το ενδεχόμενο που δεν μπορεί να πραγματοποιηθεί και επομένως ταυτίζεται με το κενό σύνολο, $\varnothing$.

Κλείνοντας την ενότητα των βασικών εννοιών, θα πρέπει να αναφερθεί η έννοια του δυναμοσυνόλου. Το **δυναμοσύνολο** ενός συνόλου Ω, το οποίο συμβολίζεται με 2^{Ω}, είναι το σύνολο όλων των υποσυνόλων του, δηλαδή

$$2^{\Omega} = \{A : A \subseteq \Omega\}.$$

Παράδειγμα 1.1

Προσδιορίστε το δυναμοσύνολο του δειγματοχώρου $\Omega = \{x, y\}$.

Λύση Παραδείγματος 1.1

Το δυναμοσύνολο του Ω είναι το:

$$2^{\Omega} = \{\varnothing, \{x\}, \{y\}, \Omega\}.$$

Παρατήρηση 1.1

Το πλήθος των στοιχείων του δυναμοσυνόλου ενός συνόλου Ω, με πεπερασμένο πλήθος $\nu(\Omega)$ στοιχείων, ισούται με $2^{\nu(\Omega)}$.

Άσκηση Αυτοαξιολόγησης 1.1

Προσδιορίστε τον δειγματοχώρο Ω και το δυναμοσύνολο 2^{Ω} των ακόλουθων πειραμάτων τύχης:

1. της ρίψης ενός ζαριού και της καταγραφής της ένδειξής του,
2. της επιλογής μιας παρτίδας 10 τυχαίων προϊόντων και της καταγραφής του αριθμού των ελαττωματικών τεμαχίων σε αυτήν.

1.2.1 Βασικά στοιχεία θεωρίας συνόλων

Όπως αναφέρθηκε νωρίτερα και θα δούμε στη συνέχεια, η πιθανότητα ορίζεται για τα ενδεχόμενα, δηλαδή τα υποσύνολα, του δειγματοχώρου ενός πειράματος τύχης. Ως εκ τούτου, κρίνεται απαραίτητη μια σύντομη παρουσίαση των βασικών στοιχείων της θεωρίας συνόλων, δηλαδή της θεωρίας που μελετάει τα σύνολα και τις ιδιότητές τους.

Ως σύνολο ορίζεται οποιαδήποτε συλλογή αντικειμένων ή εννοιών. Η θεωρία συνόλων βασίζεται σε μια βασική δυαδική σχέση μεταξύ ενός αντικειμένου x και ενός συνόλου, δηλαδή μιας συλλογής αντικειμένων A. Αν το x είναι μέλος ή αλλιώς ανήκει στο A, τότε γράφουμε $x \in A$ ενώ, αν δεν ανήκει, γράφουμε $x \notin A$. Παρατηρώντας ότι τα σύνολα μπορούν να θεωρηθούν και αυτά αντικείμενα, η παραπάνω δυαδική σχέση μπορεί να επεκταθεί και μεταξύ συνόλων. Έτσι, εάν όλα τα στοιχεία ενός συνόλου A ανήκουν και σε ένα άλλο σύνολο B, τότε λέμε ότι το A είναι **υποσύνολο** του B και γράφουμε $A \subseteq B$. Αν επιπλέον υπάρχει τουλάχιστον ένα στοιχείο y του B που δεν ανήκει στο A, δηλαδή $y \in B$ και $y \notin A$, τότε λέμε ότι το A είναι **γνήσιο υποσύνολο** του B και γράφουμε ότι $A \subset B$.

1.2.1.1 Πράξεις συνόλων

Στη συνέχεια, ορίζονται οι τρεις βασικές πράξεις μεταξύ συνόλων. Για την καλύτερη κατανόησή τους θα χρησιμοποιηθούν τα διαγράμματα Venn, τα οποία αποτελούν γεωμετρικές αναπαραστάσεις συνόλων. Ειδικότερα, η κατασκευή ενός τέτοιου διαγράμματος συνίσταται σε ένα ορθογώνιο που αντιπροσωπεύει τον δειγματικό χώρο και μέσα σε αυτό το ορθογώνιο τα σύνολα αντιπροσωπεύονται συνήθως με κύκλους ή ελλείψεις.

Ορισμός 1.4

Η **ένωση** δύο συνόλων A και B ενός δειγματικού χώρου Ω συμβολίζεται με $A \cup B$ και είναι το σύνολο που αποτελείται από τα αντικείμενα ή στοιχεία που είναι μέλη του A ή του B ή και των δύο. Συμβολικά, μπορούμε να γράψουμε ότι

$$x \in A \cup B, \text{ αν και μόνο αν } x \in A \text{ ή } x \in B.$$

Το γραμμοσκιασμένο μέρος του αριστερού διαγράμματος Venn του Σχήματος 1.1 παριστάνει την ένωση των συνόλων A και B.

Ορισμός 1.5

Η **τομή** δύο συνόλων A και B ενός δειγματικού χώρου Ω συμβολίζεται με $A \cap B$ και είναι το σύνολο που αποτελείται από τα αντικείμενα που είναι μέλη του A και του B. Συμβολικά, μπορούμε να γράψουμε ότι

$$x \in A \cap B, \text{ αν και μόνο αν } x \in A \text{ και } x \in B.$$

Το γραμμοσκιασμένο μέρος του μεσαίου διαγράμματος Venn του Σχήματος 1.1 παριστάνει την τομή των συνόλων A και B.

Ορισμός 1.6

Το **συμπλήρωμα** ενός συνόλου A ενός δειγματικού χώρου Ω συμβολίζεται με A' και είναι το σύνολο που αποτελείται από τα αντικείμενα που είναι μέλη του Ω και δεν ανήκουν στο A. Συμβολικά, μπορούμε να γράψουμε ότι:

$$x \in A', \text{ αν και μόνο αν } x \in \Omega \text{ και } x \notin A.$$

Το γραμμοσκιασμένο μέρος του δεξιού διαγράμματος Venn του Σχήματος 1.1 παριστάνει το συμπλήρωμα του συνόλου A.

Πέρα από τις τρεις παραπάνω πράξεις μεταξύ συνόλων μπορούν να οριστούν και άλλες πιο σύνθετες πράξεις, όπως η διαφορά και η συμμετρική διαφορά δύο συνόλων, οι οποίες ορίζονται στη συνέχεια.

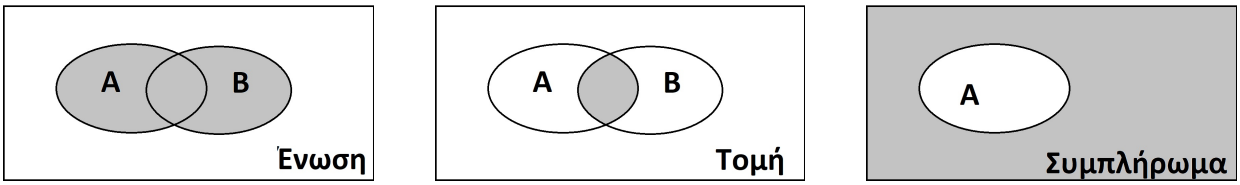
Ορισμός 1.7

Η **διαφορά** δύο συνόλων A και B ενός δειγματικού χώρου Ω συμβολίζεται με $A \setminus B$ και είναι το σύνολο που αποτελείται από αντικείμενα που είναι μέλη του A και δεν ανήκουν στο B, δηλαδή

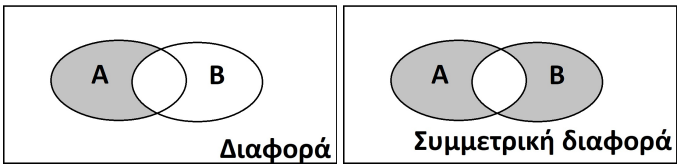
$$x \in A \setminus B, \text{ αν και μόνο αν } x \in A \text{ και } x \notin B.$$

Η **συμμετρική διαφορά** δύο συνόλων A και B ενός δειγματικού χώρου Ω συμβολίζεται με $A \oplus B$ και είναι το σύνολο που αποτελείται από αντικείμενα που είναι μέλη είτε του A είτε του B, δηλαδή

$$x \in A \oplus B, \text{ αν και μόνο αν } (x \in A \text{ και } x \notin B) \text{ ή } (x \notin A \text{ και } x \in B).$$



Σχήμα 1.1: Διαγράμματα Venn για τις βασικές πράξεις μεταξύ συνόλων.



Σχήμα 1.2: Διαγράμματα Venn για τη διαφορά και τη συμμετρική διαφορά δύο συνόλων.

Η διαφορά και η συμμετρική διαφορά δύο συνόλων δεν είναι βασικές πράξεις, αφού μπορεί να εκφραστούν ισοδύναμα ως $A \cap B'$ και $(A \cap B') \cup (A' \cap B)$, αντίστοιχα. Στο Σχήμα 1.2 παρουσιάζονται τα διαγράμματα Venn για τη διαφορά (αριστερά) και τη συμμετρική διαφορά (δεξιά) των συνόλων A και B.

Παράδειγμα 1.2

Ένα νόμισμα με όψεις κορώνα και γράμματα ρίχνεται 6 φορές και καταγράφεται ο αριθμός των φορών εμφάνισης της όψης «γράμματα».

1. Ποιος είναι ο δειγματοχώρος του συγκεκριμένου πειράματος τύχης;
2. Αν $A = \{1,2,3\}$, $B = \{2,4,6\}$ και $\Gamma = \{1,2,4,5,6\}$ τρία ενδεχόμενα του παραπάνω δειγματοχώρου, προσδιορίστε τα $A \cup B$, $A \cap B$, A', $A \cup B'$, $A \cap \Gamma$, $A \setminus \Gamma$, $\Gamma \oplus B$, $A \cap B \cap \Gamma$, $A' \cap B' \cap \Gamma'$ και $(A \cup B \cup \Gamma)'$.

Λύση Παραδείγματος 1.2

1. Ο δειγματοχώρος του συγκεκριμένου πειράματος τύχης είναι ο $\Omega = \{0,1,2,3,4,5,6\}$, αφού μπορεί η όψη γράμματα να εμφανιστεί μηδέν, μία, δύο, τρεις, τέσσερις, πέντε ή έξι φορές.
2. Το $A \cup B$ αποτελείται από τα στοιχεία που ανήκουν ή στο A ή στο B ή και στα δύο. Επομένως, είναι $A \cup B = \{1,2,3,4,6\}$. Από την άλλη πλευρά, το $A \cap B$ αποτελείται από τα κοινά στοιχεία των δύο ενδεχομένων, άρα $A \cap B = \{2\}$. Συνεχίζοντας κατά τον ίδιο τρόπο και αξιοποιώντας τους ορισμούς που δόθηκαν πρωτύτερα, προκύπτουν τα ακόλουθα:
   - $A' = \{0,4,5,6\}$.
   - $A \cup B' = \{1,2,3\} \cup \{0,1,3,5\} = \{0,1,2,3,5\}$.
   - $A \cap \Gamma = \{1,2\}$.
   - $A \setminus \Gamma = \{3\}$.
   - $\Gamma \oplus B = \{1,5\}$.
   - $A \cap B \cap \Gamma = \{2\}$.
   - $A' \cap B' \cap \Gamma' = \{0,4,5,6\} \cap \{0,1,3,5\} \cap \{0.3\} = \{0\}$.
   - $(A \cup B \cup \Gamma)' = \{1,2,3,4,5,6\}' = \{0\}$.

Μία σημαντική έννοια στη θεωρία συνόλων είναι η έννοια των ασυμβίβαστων ή, αλλιώς, ξένων μεταξύ τους συνόλων.

Ορισμός 1.8

Δύο σύνολα A και B ονομάζονται **ασυμβίβαστα** ή **ξένα**, αν η τομή τους είναι το κενό σύνολο, δηλαδή αν $A \cap B = \varnothing$.

Η έννοια των ασυμβίβαστων συνόλων μπορεί να επεκταθεί και σε περισσότερα από δύο σύνολα. Παραδείγματος χάριν, τα σύνολα A, B και Γ λέγονται ασυμβίβαστα, αν $A \cap B \cap \Gamma = \varnothing$.

Ορισμός 1.9

Μια συλλογή συνόλων $A_1, A_2, \dots, A_k$ λέγεται ότι αποτελείται από **ασυμβίβαστα ή ξένα ανά δύο μεταξύ τους σύνολα**, αν $A_i \cap A_j = \varnothing,\ \forall\, i,j = 1,2,\dots,k$, με $i \neq j$.

Άσκηση Αυτοαξιολόγησης 1.2

Αν $A_1, A_2, \dots, A_k$ είναι μια συλλογή συνόλων, τότε ποια από τις έννοιες των ξένων μεταξύ τους και των ξένων ανά δύο ενδεχομένων είναι πιο ισχυρή και συνεπάγεται την άλλη;

1.2.1.2 Πράξεις ενδεχομένων: Ιδιότητες

Αν A είναι ένα ενδεχόμενο ενός δειγματοχώρου Ω, δηλαδή αν $A \subseteq \Omega$, τότε ισχύουν οι παρακάτω χρήσιμες ιδιότητες:

- $A \cup A = A$
- $A \cup \varnothing = A$
- $A \cup \Omega = \Omega$
- $(A')' = A$
- $A \cup A' = \Omega$
- $A \cap A = A$
- $A \cap \varnothing = \varnothing$
- $A \cap \Omega = A$
- $A \cap A' = \varnothing$

Επιπλέον, για το βέβαιο και το αδύνατο ενδεχόμενο ισχύουν οι παρακάτω σχέσεις:

- $\Omega' = \varnothing$
- $\varnothing' = \Omega$

Αν A, B και Γ είναι ενδεχόμενα του ίδιου δειγματοχώρου Ω, δηλαδή $A, B, \Gamma \subseteq \Omega$, τότε ισχύουν και οι ακόλουθες ιδιότητες:

- Αντιμεταθετική: $A \cup B = B \cup A$ και $B \cap A = A \cap B$.
- Προσεταιριστική: $A \cup (B \cup \Gamma) = (A \cup B) \cup \Gamma = A \cup B \cup \Gamma$.
  $A \cap (B \cap \Gamma) = (A \cap B) \cap \Gamma = A \cap B \cap \Gamma$.
- Επιμεριστική: $A \cup (B \cap \Gamma) = (A \cup B) \cap (A \cup \Gamma)$.
  $A \cap (B \cup \Gamma) = (A \cap B) \cup (A \cap \Gamma)$.

Κλείνοντας την ενότητα των ιδιοτήτων των πράξεων μεταξύ ενδεχομένων, θα αναφερθούμε στους αποκαλούμενους **τύπους De Morgan**, οι οποίοι οφείλουν την ονομασία τους στον Βρετανό μαθηματικό Augustus De Morgan (1806-1871). Οι τύποι De Morgan εμπλέκουν τις τομές και τις ενώσεις δύο ή περισσότερων ενδεχομένων με το συμπλήρωμά τους και δίνονται στην πρόταση που ακολουθεί.

Πρόταση 1.1

Αν A, B είναι δύο ενδεχόμενα ενός δειγματικού χώρου Ω, με $A, B \subseteq \Omega$, τότε:

1. $(A \cap B)' = A' \cup B'$ και
2. $(A \cup B)' = A' \cap B'$.

Απόδειξη Πρότασης 1.1

1. Αρχικά θα δείξουμε ότι $(A \cap B)'$ είναι υποσύνολο του $A' \cup B'$. Αρκεί να δείξουμε ότι για οποιοδήποτε $x \in (A \cap B)'$ συνεπάγεται ότι $x \in A' \cup B'$. Καθώς $x \in (A \cap B)'$ συμπεραίνουμε ότι $x \notin (A \cap B)$, που με τη σειρά του συνεπάγεται ότι $x \notin A$ ή $x \notin B$. Επομένως, έχουμε ότι $x \in A'$ ή $x \in B'$ ή, ισοδύναμα, $x \in A' \cup B'$.

   Στη συνέχεια, θα δείξουμε ότι $A' \cup B'$ είναι υποσύνολο του $(A \cap B)'$. Αρκεί να δείξουμε ότι για οποιοδήποτε $y \in A' \cup B'$ συνεπάγεται ότι $y \in (A \cap B)'$. Καθώς $y \in A' \cup B'$, έχουμε ότι $y \in A'$ ή $y \in B'$ ή, ισοδύναμα, $y \notin A$ ή $y \notin B$. Από την τελευταία σχέση προκύπτει ότι $y \notin (A \cap B)$ ή, ισοδύναμα, $y \in (A \cap B)'$. Συνδυάζοντας τα παραπάνω προκύπτει το ζητούμενο.
2. Η απόδειξη είναι παρόμοια και αφήνεται ως άσκηση στον/στην αναγνώστη/στρια.

Οι παραπάνω σχέσεις μπορούν να επεκταθούν και σε μια συλλογή n το πλήθος ενδεχομένων ενός δειγματικού χώρου Ω ως ακολούθως:

1. $(A_1 \cup A_2 \cup \cdots \cup A_n)' = A_1' \cap A_2' \cap \cdots \cap A_n'$ και
2. $(A_1 \cap A_2 \cap \cdots \cap A_n)' = A_1' \cup A_2' \cup \cdots \cup A_n'$.

Παράδειγμα 1.3

Σε συνέχεια του Παραδείγματος 1.2 υπολογίστε τα ενδεχόμενα: $A' \cap \Omega$, $A' \cup \varnothing$, $A' \cup B'$, $(A \cup B)'$, $A' \cap (B' \cup \Gamma)$,

Λύση Παραδείγματος 1.3

Με βάση τους ορισμούς και την πρόταση που προηγήθηκε, εύκολα προκύπτουν τα ακόλουθα:
$A' \cap \Omega = A' = \{0,4,5,6\}$,
$A' \cup \varnothing = A' = \{0,4,5,6\}$,
$A' \cup B' = (A \cap B)' = \{0,1,3,4,5,6\}$,
$(A \cup B)' = A' \cap B' = \{0,4,5,6\} \cap \{0,1,3,5\} = \{0,5\}$, και
$A' \cap (B' \cup \Gamma) = (A' \cap B') \cup (A' \cap \Gamma) = \{0,5\} \cup \{4,5,6\} = \{0,4,5,6\}$.
Παρατηρήστε ότι ισοδύναμα θα μπορούσαμε να γράψουμε ότι: $(A \cup B)' = (\{1,2,3,4,6\})' = \{0,5\}$.

Οι αναγνώστες που ενδιαφέρονται να εμβαθύνουν στη θεωρία συνόλων παραπέμπονται στα συγγράμματα των Halmos (1974), Μητακίδης (1988), Κάλφα (1990), Enderton (2013) και Γεωργίου και Ηλιάδης (2017).

1.2.2 Βασικές αρχές απαρίθμησης

Πέρα από τις προαναφερθείσες πράξεις μεταξύ των συνόλων, όπως θα δούμε στην επόμενη ενότητα, για τον εύκολο, γρήγορο και, κυρίως, σωστό υπολογισμό των πιθανοτήτων πολλές φορές απαιτείται επιπρόσθετα και κάποια στοιχειώδης γνώση τεχνικών απαρίθμησης. Στο πλαίσιο αυτού του συγγράμματος, θα περιοριστούμε στην παρουσίαση της αποκαλούμενης **βασικής αρχής απαρίθμησης** και κάποιων βασικών στοιχείων συνδυαστικής (συνδυασμοί, διατάξεις και μεταθέσεις).

Ορισμός 1.10: Βασική αρχή απαρίθμησης

Αν μια διαδικασία μπορεί να χωριστεί σε n το πλήθος ανεξάρτητες επιμέρους διαδικασίες (στάδια) και η καθεμία από αυτές τις επιμέρους διαδικασίες έχει $\nu_1, \nu_2, \cdots, \nu_n$, αντίστοιχα, το πλήθος δυνατά διαφορετικά αποτελέσματα, τότε η αρχική διαδικασία έχει συνολικά

$$\nu = \nu_1 \cdot \nu_2 \ldots \nu_n$$

το πλήθος δυνατά διαφορετικά αποτελέσματα.

Παράδειγμα 1.4

Το δυαδικό ψηφίο ή μπιτ (bit) είναι η στοιχειώδης μονάδα πληροφορίας στην Επιστήμη Υπολογιστών, το οποίο μπορεί να έχει μόνο δύο πιθανές τιμές. Αυτές οι δύο τιμές συνήθως ερμηνεύονται ως δυαδικά ψηφία και αναπαρίστανται με τους αριθμούς 0 και 1.
Οι $32 - bit$ υπολογιστές χρησιμοποιούν ακολουθίες 32 το πλήθος ψηφίων 0 και 1, για να ορίσουν τις «λέξεις» που χρησιμοποιούν, ενώ οι $64 - bit$ υπολογιστές ορίζουν τις αντίστοιχες «λέξεις» μέσω ακολουθιών 64 το πλήθος ψηφίων. Πόσες παραπάνω «λέξεις» μπορούμε να ορίσουμε σε έναν $64 - bit$ υπολογιστή από ότι σε έναν $32 - bit$ υπολογιστή;

Λύση Παραδείγματος 1.4

Κάθε ψηφίο έχει δύο δυνατές επιλογές, το 0 και 1. Επομένως, με βάση τη βασική αρχή απαρίθμησης μπορούν να δημιουργηθούν $2 \times 2 \times \ldots \times 2 = 2^{32}$ το πλήθος διαφορετικές «λέξεις» σε έναν $32 - bit$ υπολογιστή και 2^{64} το πλήθος διαφορετικές «λέξεις» σε έναν $64 - bit$ υπολογιστή. Οπότε σε έναν $64 - bit$ υπολογιστή μπορούμε να ορίσουμε $2^{64} - 2^{32} = 18446744069414584320$ το πλήθος περισσότερες «λέξεις».

Εκτός από τη βασική αρχή απαρίθμησης, χρήσιμα στον υπολογισμό πιθανοτήτων είναι και κάποια στοιχεία συνδυαστικής. Η συνδυαστική είναι ο κλάδος των μαθηματικών που περιλαμβάνει τεχνικές απαρίθμησης μετρήσιμων διακριτών δομών. Βασικές τεχνικές απαρίθμησης είναι οι συνδυασμοί, οι διατάξεις και οι μεταθέσεις, οι οποίες παρουσιάζονται στη συνέχεια.

Ορισμός 1.11: Συνδυασμοί

Συνδυασμός n το πλήθος διακεκριμένων στοιχείων ανά r $(1 \leq r \leq n)$ ονομάζεται κάθε δυνατή συλλογή r διαφορετικών στοιχείων από τα n, χωρίς να μας ενδιαφέρει η σειρά τους.

Το πλήθος των δυνατών συνδυασμών n διακεκριμένων στοιχείων ανά r συμβολίζεται με $\binom{n}{r}$ και ισούται με

$$\binom{n}{r} = \frac{n!}{r! \cdot (n-r)!}.$$

Επισημαίνεται ότι $n! = 1 \times 2 \times \ldots \times n$, με $0! = 1$. Επιπλέον, ο αριθμός $\binom{n}{r}$ ονομάζεται διωνυμικός συντελεστής, επειδή αριθμοί αυτής της μορφής εμφανίζονται ως συντελεστές στο διωνυμικό ανάπτυγμα

$$(\alpha + \beta)^n = \sum_{j=0}^{n} \binom{n}{j} \cdot \alpha^{n-j} \cdot \beta^j.$$

Για κάθε ακέραιο n ισχύει ότι

$$\binom{n}{0} = \binom{n}{n} = 1.$$

Μία από τις σημαντικότερες ταυτότητες που ικανοποιεί ο διωνυμικός συντελεστής, και η οποία αποδεικνύεται εύκολα με βάση τον ορισμό του, είναι η ταυτότητα του Pascal[1], η οποία διατυπώνεται στη συνέχεια:

$$\binom{n+1}{k} = \binom{n}{k-1} + \binom{n}{k}.$$

όπου n και k θετικοί ακέραιοι αριθμοί με $k \leq n$.

Ορισμός 1.12: Διατάξεις

Διάταξη n διακεκριμένων στοιχείων ανά r $(1 \leq r \leq n)$ ονομάζεται κάθε συλλογή r διαφορετικών στοιχείων από τα n, για την οποία μας ενδιαφέρει η σειρά εμφάνισης των στοιχείων.

Το πλήθος των δυνατών διατάξεων n διακεκριμένων στοιχείων ανά r συμβολίζεται με ${}_nP_r$ και ισούται με

$${}_nP_r = \frac{n!}{(n-r)!}.$$

Ορισμός 1.13: Μεταθέσεις

Η ειδική περίπτωση των διατάξεων n στοιχείων ανά n ονομάζεται **μετάθεση** των n στοιχείων.

Το πλήθος των δυνατών μεταθέσεων n στοιχείων συμβολίζεται με P_n και ισούται με

$$P_n = n!$$

Παράδειγμα 1.5

Ένα κτήριο έχει 10 ορόφους. Με πόσους τρόπους μπορούμε να επιλέξουμε 3 από τους ορόφους, όταν δεν μας ενδιαφέρει η σειρά επιλογής;

Λύση Παραδείγματος 1.5

Στο παράδειγμα αυτό δεν μας ενδιαφέρει η σειρά επιλογής, οπότε θέλουμε το πλήθος των συνδυασμών των 10 ανά 3. Δηλαδή το πλήθος των διαφορετικών τρόπων με τους οποίους μπορούμε να επιλέξουμε 3 από τους 10 ορόφους ισούται με:

$$\binom{10}{3} = \frac{10!}{3! \cdot (10-3)!} = \frac{10!}{3! \cdot 7!} = \frac{10 \cdot 9 \cdot 8 \cdot 7!}{3! \cdot 7!} = \frac{10 \cdot 9 \cdot 8}{1 \cdot 2 \cdot 3} = 120.$$

1.3 Ορισμοί Πιθανότητας

Η έννοια της πιθανότητας, βαθιά ριζωμένη στην ανθρώπινη ιστορία, χρησιμοποιείται καθημερινά και κατανοείται από τους περισσότερους ανθρώπους. Παρόλα αυτά, ο ακριβής και αυστηρός μαθηματικός

[1] Ο Blaise Pascal (1623-1662) ήταν Γάλλος μαθηματικός, φυσικός, συγγραφέας και φιλόσοφος. Ο Pascal ήταν ένα παιδί-θαύμα, αφού σε ηλικία μόλις 12 ετών ανακάλυψε μόνος του ότι το άθροισμα των γωνιών ενός τριγώνου ισούται με δύο ορθές γωνίες. Στα 16 του διατύπωσε ένα σημαντικό συμπέρασμα για τις κωνικές τομές, ενώ στα 18 του σχεδίασε, κατασκεύασε και πούλησε μία αριθμομηχανή, η οποία έκανε προσθέσεις και αφαιρέσεις. Η αριθμομηχανή του Pascal δεν έκανε μεγάλες πωλήσεις, οπότε πολύ γρήγορα σταμάτησε η παραγωγή της. Μαζί με τον Γάλλο νομικό και μαθηματικό Pierre de Fermat (1607-1665) θεωρούνται οι θεμελιωτές της σύγχρονης θεωρίας των Πιθανοτήτων. Το 1645 ο Pascal εγκατέλειψε τα μαθηματικά και αφιερώθηκε στη θεολογία.

ορισμός της δεν είναι εύκολος. Μάλιστα, η διαδικασία ανάπτυξης ενός ορισμού για την πιθανότητα έχει προσεγγιστεί κατά τη διάρκεια των τελευταίων 5-6 αιώνων με περισσότερους από έναν τρόπους, δίνοντάς μας διάφορους ορισμούς. Ειδικότερα, έχουν εισαχθεί στη βιβλιογραφία:

1. ο κλασικός ορισμός της πιθανότητας, ο οποίος αναφέρεται και ως ορισμός πιθανότητας κατά Laplace,
2. ο ορισμός της πιθανότητας κατά Von Mises ή, αλλιώς, εμπειρικός ορισμός, όπου η πιθανότητα προκύπτει ως όριο της σχετικής συχνότητας. Ο ορισμός αυτός αναφέρεται και πιο απλά ως «στατιστικός»,
3. ο υποκειμενικός ορισμός, και
4. ο «αξιωματικός ορισμός της πιθανότητας», ο οποίος διατυπώθηκε πολύ πρόσφατα, μόλις το 1933, από τον Ρώσο μαθηματικό Andrey Kolmogorov (1903-1987). Το έργο του Andrey Kolmogorov δημοσιεύτηκε στα αγγλικά το 1956 (Kolmogorov, 1956), ενώ υπάρχουν πολλές πρόσφατες επανεκδόσεις του, όπως αυτή του 2013 (Kolmogorov, 2013).

Στη συνέχεια της ενότητας αυτής, θα παρουσιαστούν αναλυτικά όλοι οι παραπάνω ορισμοί.

1.3.1 Κλασικός ορισμός

Ο πρώτος ορισμός που θα αναφέρουμε είναι ο αποκαλούμενος κλασικός ορισμός, ο οποίος αναφέρεται και ως ορισμός πιθανότητας κατά Laplace, προς τιμήν του Γάλλου αστρονόμου και μαθηματικού Pierre-Simon Laplace (1749-1827), καθώς ήταν αυτός που διατύπωσε σαφώς τον ορισμό που ακολουθεί (βλ. Laplace, 1814).

Ορισμός 1.14

Έστω Ω ένας δειγματικός χώρος με πεπερασμένο πλήθος στοιχείων, του οποίου όλα τα στοιχειώδη (απλά) ενδεχόμενα είναι εξίσου πιθανά (ισοπίθανα). Τότε η πιθανότητα ενός ενδεχομένου A συμβολίζεται με $P(A)$, και ορίζεται από τη σχέση:

$$P(A) = \frac{\text{πλήθος ευνοϊκών περιπτώσεων}}{\text{πλήθος δυνατών περιπτώσεων}} = \frac{\nu(A)}{\nu(\Omega)}.$$

Ο κλασικός ορισμός είναι πάρα πολύ απλός και μπορεί να εφαρμοστεί εύκολα σε πολλές περιπτώσεις. Ένα τέτοιο παράδειγμα είναι η πιθανότητα εμφάνισης ενός άρτιου αριθμού κατά τη ρίψη ενός «τίμιου» ζαριού, δηλαδή ενός ζαριού που κάθε ένδειξή του είναι το ίδιο πιθανό να εμφανιστεί. Αφού τα δυνατά αποτελέσματα είναι 6 και τα ευνοϊκά 3, είναι εύκολο να υπολογίσουμε ότι η πιθανότητα εμφάνισης ενός άρτιου αριθμού κατά τη ρίψη ενός «τίμιου» ζαριού ισούται με 3/6, δηλαδή 1/2. Αυτό σημαίνει ότι έχουμε 50% πιθανότητα να έχουμε ως αποτέλεσμα έναν άρτιο αριθμό.

Παρά την ευκολία του, ο κλασικός ορισμός παρουσιάζει μια σειρά από σημαντικά μειονεκτήματα. Το πρώτο σημαντικό μειονέκτημα του κλασικού ορισμού είναι ότι μπορεί να εφαρμοστεί μόνο σε πειράματα τύχης που έχουν πεπερασμένο πλήθος δυνατών αποτελεσμάτων. Αυτό έχει ως συνέπεια πολλά ενδιαφέροντα προβλήματα να μην μπορούν να αντιμετωπιστούν με τη χρήση του.

Το δεύτερο σημαντικό μειονέκτημα είναι ότι, ακόμα και για πειράματα τύχης με πεπερασμένο πλήθος δυνατών αποτελεσμάτων, θα πρέπει όλα τα στοιχειώδη (απλά) ενδεχόμενα να είναι ισοπίθανα. Έτσι, στο παραπάνω παράδειγμα, αν είχαμε ένα «μη τίμιο» ζάρι, το οποίο έφερνε πιο συχνά το 1, δεν θα μπορούσαμε να εφαρμόσουμε τον κλασικό ορισμό.

Το τρίτο και τελευταίο μειονέκτημα του κλασικού ορισμού έχει να κάνει περισσότερο με την αναγκαιότητα μιας αυστηρότερης μαθηματικής προσέγγισής του. Η αναγκαιότητα αυτή προκύπτει από το γεγονός ότι ο Ορισμός 1.14 δεν είναι με την αυστηρή έννοια ορισμός, καθώς χρησιμοποιεί την έννοια της πιθανότητας

(του ισοπίθανου) για τον ορισμό της πιθανότητας! Ωστόσο, παρά την προφανή αυτήν αντίφαση, ο κλασικός ορισμός έχει καθιερωθεί ως ένας αποδεκτός ορισμός, καταδεικνύοντας με αυτόν τον τρόπο τη βαθιά σχέση της έννοιας της πιθανότητας με την καθημερινότητά μας.

Παράδειγμα 1.6

Υπολογίστε τις πιθανότητες των ενδεχομένων

$$A \cap B \qquad A' \cap B' \qquad (A \cap B)' \qquad A' \cap (B \cup \Gamma) \qquad A' \cup (B \cap \Gamma),$$

για τη ρίψη ενός «τίμιου» ζαριού, όπου τα ενδεχόμενα A, B και Γ είναι τα:

$$A = \{1,2,3\} \quad B = \{2,4,6\} \quad \Gamma = \{1,2,4,5,6\}$$

Λύση Παραδείγματος 1.6

Επειδή όλα τα στοιχειώδη ενδεχόμενα είναι ισοπίθανα και το πλήθος τους είναι πεπερασμένο (έξι δυνατά στοιχειώδη ενδεχόμενα, $\Omega = \{1,2,3,4,5,6\}$), μπορούμε να εφαρμόσουμε τον κλασικό ορισμό της πιθανότητας.
Άρα, επειδή $A \cap B = \{1,2,3\} \cap \{2,4,6\} = \{2\}$ έχουμε ότι το πλήθος των στοιχείων του ισούται με $\nu(A \cap B) = 1$, οπότε

$$P(A \cap B) = \frac{\nu(A \cap B)}{\nu(\Omega)} = \frac{1}{6} = 0.167.$$

Για τα υπόλοιπα τέσσερα ενδεχόμενα έχουμε ότι:

$$\begin{aligned} A' \cap B' &= (A \cup B)' = \{1,2,3,4,6\}' = \{5\}, \\ (A \cap B)' &= \{2\}' = \{1,3,4,5,6\}, \\ A' \cap (B \cup \Gamma) &= \{4,5,6\} \cap (\{2,4,6\} \cup \{1,2,3,4,5\}) = \{4,5,6\} \cap \{1,2,3,4,5,6\} = \{4,5,6\}, \\ A' \cup (B \cap \Gamma) &= \{4,5,6\} \cup (\{2,4,6\} \cap \{1,2,3,4,5\}) = \{4,5,6\} \cup \{2,4\} = \{2,4,5,6\}, \end{aligned}$$

και, επομένως, έχουμε ότι: $P(A' \cap B') = 1/6$, $P((A \cap B)') = 5/6$, $P(A' \cap (B \cup \Gamma)) = 3/6$ και $P(A' \cup (B \cap \Gamma)) = 4/6$.

Παράδειγμα 1.7

Σε ένα οδικό δίκτυο με 15 διακεκριμένα τούνελ θέλουμε να επιλέξουμε τυχαία

1. ένα τούνελ,
2. δύο τούνελ

για την πραγματοποίηση κάποιων διαγνωστικών ελέγχων. Προφανώς, στη δεύτερη περίπτωση, δεν μπορεί να επιλεχθεί ένα τούνελ δύο φορές. Σε καθεμία από τις περιπτώσεις να βρεθεί ο δειγματικός χώρος, καθώς και η πιθανότητα να ελεγχθεί το υπ' αριθμόν 2 τούνελ.

Λύση Παραδείγματος 1.7

1. Ο δειγματοχώρος είναι $\Omega_1 = \{1,2,3,\ldots,15\}$. Για να απαντήσουμε στο ερώτημα που έχει τεθεί, ορίζουμε το ενδεχόμενο $A = \{$να ελεγχθεί το υπ' αριθμόν 2 τούνελ$\}$. Επειδή η επιλογή του τούνελ είναι τυχαία, όλα τα στοιχειώδη ενδεχόμενα είναι ισοπίθανα. Με τη χρήση του κλασικού ορισμού έχουμε ότι: $P(A) = 1/15$.
2. Ο δειγματοχώρος της επιλογής δύο τούνελ, χωρίς τη δυνατότητα επιλογής του ίδιου τούνελ δύο φορές, μπορεί να περιγραφθεί από το

$$\Omega_2 = \{(i,j), i,j = 1,2,\ldots,15, \text{ με } i \neq j\}.$$

Επειδή η επιλογή γίνεται πάλι τυχαία, όλα τα στοιχειώδη ενδεχόμενα είναι ισοπίθανα και, προφανώς, το πλήθος τους είναι πεπερασμένο, οπότε και μπορούμε να βασιστούμε στον κλασικό ορισμό.

Για τον υπολογισμό της ζητούμενης πιθανότητας ορίζουμε το ενδεχόμενο

$$B = \{\text{ένα από τα δύο τούνελ που ελέγχεται είναι το υπ' αριθμόν 2}\}$$

και η πιθανότητά του δίνεται από τη σχέση

$$P(B) = \frac{\nu(B)}{\nu(\Omega)}.$$

Επομένως, ο υπολογισμός αυτής της πιθανότητας απαιτεί τον προσδιορισμό των ευνοϊκών και των δυνατών αποτελεσμάτων του συγκεκριμένου πειράματος τύχης.

Εδώ, θα μπορούσαμε να καταγράψουμε πλήρως τον δειγματοχώρο και να προσδιορίσουμε τις ευνοϊκές περιπτώσεις και να δείξουμε ότι η ζητούμενη πιθανότητα ισούται με $P(B) = 2/15$.

Καθώς αυτή η διαδικασία, ακόμα και στην περίπτωση αυτού του απλού παραδείγματος, είναι χρονοβόρα, στη συνέχεια, θα δούμε πώς μπορούμε να μετρήσουμε τα ευνοϊκά και τα δυνατά αποτελέσματα του πειράματος τύχης χωρίς να χρειαστεί να τα καταγράψουμε πλήρως. Το πλήθος των διαφορετικών τρόπων με τους οποίους μπορούμε να επιλέξουμε τα 2 τούνελ από τα 15 διαθέσιμα ισούται με $\binom{15}{2}$.

Για να υπολογίσουμε την πιθανότητα ανάμεσα στα δύο τούνελ που θα επιλέξουμε να είναι το υπ' αριθμόν 2 τούνελ, χωρίς βλάβη της γενικότητας, αφού δεν μας ενδιαφέρει η σειρά επιλογής, μπορούμε να θεωρήσουμε ότι είναι το τούνελ που επιλέγεται πρώτο, δηλαδή ότι το i ισούται με 2.

Στη συνέχεια, θα χρησιμοποιήσουμε τη βασική αρχή απαρίθμησης και θα θεωρήσουμε ότι, αντί να επιλέγουμε ταυτόχρονα τα 2 τούνελ, επιλέγουμε πρώτα το υπ' αριθμόν 2 τούνελ και, στη συνέχεια, ένα τούνελ από τα 14 υπόλοιπα. Η πρώτη επιλογή μας, επειδή πρέπει να είναι το υπ' αριθμόν 2 τούνελ, μπορεί να γίνει μόνο με έναν τρόπο, ενώ η δεύτερη επιλογή μας μπορεί να γίνει με τόσους διαφορετικούς τρόπους όσοι και οι συνδυασμοί των 14 ανά 1. Έτσι συνολικά το πλήθος των στοιχείων του B, δηλαδή των ευνοϊκών περιπτώσεων, ισούται με $1 \cdot \binom{14}{1}$. Επομένως, με βάση τον κλασικό ορισμό της πιθανότητας έχουμε, ότι:

$$P(B) = \frac{1 \cdot \binom{14}{1}}{\binom{15}{2}} = \frac{2}{15}.$$

Άσκηση Αυτοαξιολόγησης 1.3

Μπορεί να εφαρμοστεί ο κλασικός ορισμός της πιθανότητας, για τον υπολογισμό των πιθανοτήτων διάφορων ενδεχομένων του Παραδείγματος 1.2; Θα ήταν αρκετό να υποτεθεί ότι το νόμισμά μας είναι «τίμιο» ή μήπως δεν αρκεί ούτε αυτό;

Άσκηση Αυτοαξιολόγησης 1.4

Σε συνέχεια του Παραδείγματος 1.5, να υπολογίσετε την πιθανότητα ανάμεσα στους τρεις ορόφους που θα επιλεχθούν, να είναι ο 10ος όροφος.

1.3.2 Ορισμός πιθανότητας ως το όριο της σχετικής συχνότητας

Ο ορισμός της πιθανότητας ενός ενδεχομένου ως το όριο της σχετικής συχνότητας, ο οποίος αναφέρεται και ως ορισμός πιθανότητας κατά Von Mises ή πιο απλά ως «στατιστικός», προτάθηκε για την αντιμετώπιση των προβλημάτων του κλασικού ορισμού[2]. Πιο συγκεκριμένα, με τον ορισμό της πιθανότητας ως όριο της σχετικής συχνότητας γίνεται μια προσπάθεια ποσοτικοποίησης της αβεβαιότητας ενός ενδεχομένου μέσα από την επανάληψη του πειράματος τύχης, υπό τις ίδιες συνθήκες, πολλές φορές.

Ορισμός 1.15

Έστω A ένα ενδεχόμενο του δειγματικού χώρου Ω ενός πειράματος τύχης, $A \subseteq \Omega$. Υποθέτουμε ότι το πείραμα τύχης επαναλαμβάνεται κάτω από τις ίδιες συνθήκες n το πλήθος φορές. Τότε η πιθανότητα $P(A)$ του ενδεχομένου A ορίζεται ως

$$P(A) = \lim_{n\to\infty} \frac{n_A}{n},$$

όπου n_A είναι ο αριθμός των φορών που πραγματοποιήθηκε το ενδεχόμενο A κατά την επανάληψη n φορών του πειράματος τύχης.

Ο παραπάνω ορισμός, αν και ξεπερνάει όλα τα μειονεκτήματα του κλασικού, εντούτοις απαιτεί την επανάληψη του πειράματος τύχης άπειρες φορές υπό τις ίδιες συνθήκες. Επομένως, ο ακριβής προσδιορισμός της πιθανότητας ενός ενδεχομένου είναι πρακτικά αδύνατος. Παρόλα αυτά, αν επαναλάβουμε το πείραμα έναν αρκετά μεγάλο αριθμό, τότε ο λόγος της σχετικής συχνότητας πραγματοποίησης του A, n_A/n, προσεγγίζει ικανοποιητικά την πραγματική τιμή δίνοντάς μας μία καλή εκτίμηση της τιμής της.

Στη γραφική παράσταση του Σχήματος 1.3 απεικονίζεται η σχετική συχνότητα της εμφάνισης της ένδειξης 3 ενός ζαριού σε σχέση με το πλήθος n ($n = 1, 2, \ldots, 3000$) των ρίψεων, όταν χρησιμοποιείται ένα «τίμιο» (μαύρη γραμμή) και ένα «μη τίμιο» (κόκκινη γραμμή) ζάρι. Η γραφική παράσταση του Σχήματος 1.3 έχει κατασκευαστεί με τη βοήθεια κατάλληλου κώδικα της R[3], ο οποίος παρατίθεται ως Κώδικας 1.1.

Στις γραμμές 8 και 9 του κώδικα ζητάμε από το πρόγραμμα να επιλέξει τυχαία έναν από τους αριθμούς που εμφανίζονται στην κάθε γραμμή. Παρατηρούμε ότι ο αριθμός 1 εμφανίζεται περισσότερες από μία φορά στη γραμμή 9, υποδηλώνοντας ότι το ζάρι είναι «μη τίμιο».

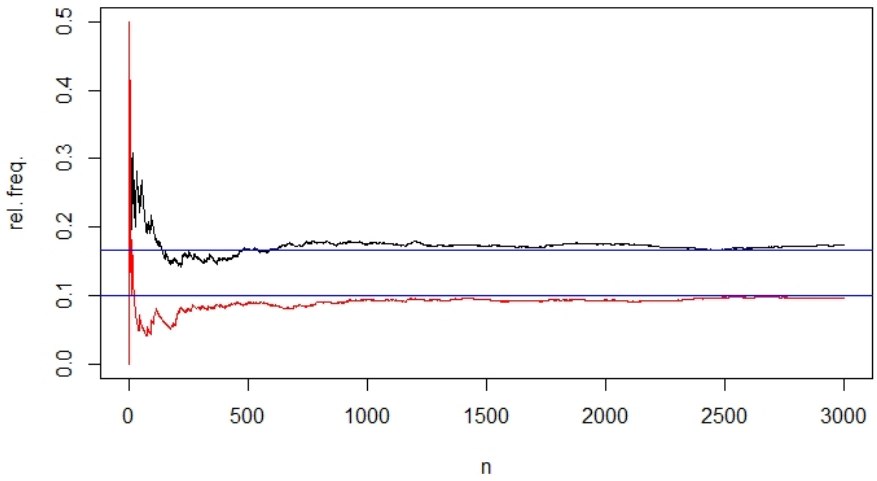
Πιο συγκεκριμένα, ο αριθμός 1 εμφανίζεται 5 φορές και, επομένως, ο αριθμός 3 έχει κάθε φορά πιθανότητα 1/10 να επιλεχθεί. Αντίθετα, στη γραμμή 8 κάθε ακέραιος αριθμός μεταξύ του 1 και του 6 εμφανίζεται μόνο μία φορά και έτσι κάθε αριθμός έχει πιθανότητα 1/6 να επιλεχθεί και η διαδικασία αυτή προσομοιάζει τη ρίψη ενός «τίμιου» ζαριού. Οι τιμές 1/10 και 1/6 έχουν σημειωθεί στο Σχήμα 1.3 με τις οριζόντιες μπλε γραμμές.

Από το Σχήμα 1.3 είναι φανερό ότι η σχετική συχνότητα της εμφάνισης του αριθμού 3 συγκλίνει και για τα δύο ζάρια στην πραγματική της τιμή. Είναι σημαντικό, επίσης, να παρατηρήσουμε ότι, αν ο σκοπός μας ήταν να αποδείξουμε ότι το δεύτερο ζάρι δεν είναι «τίμιο», τότε δεν θα ήταν απαραίτητο να κάνουμε πολλές επαναλήψεις του πειράματος, αφού η εκτίμηση της τιμής της πιθανότητας εμφάνισης του αριθμού 1 είναι σχεδόν για όλους τους αριθμούς επαναλήψεων του πειράματος μακριά από το 1/6, τιμή που θα είχε η πιθανότητα εμφάνισης κάθε ένδειξης, αν το ζάρι ήταν «τίμιο».

---

[2] Η ονομασία «ορισμός κατά Von Mises» έχει δοθεί προς τιμήν του Αυστριακού μαθηματικού Richard Von Mises (1883-1953) και της συνεισφοράς του στην πιθανοθεωρία (βλ. Von Mises, 1919).

[3] Η R είναι μια γλώσσα προγραμματισμού ελεύθερου λογισμικού, διαθέσιμη στον ιστότοπο https://www.r-project.org/, η οποία είναι κατάλληλη για εφαρμογές Πιθανοτήτων και Στατιστικής.



Σχήμα 1.3: Εφαρμογή του ορισμού της πιθανότητας ως το όριο της σχετικής συχνότητας σε ένα «τίμιο» και σε ένα «μη τίμιο» ζάρι.

Παρατήρηση 1.2

Αξίζει να σημειωθεί ότι η σύγκλιση προς την πραγματική τιμή δεν είναι μονότονη. Αυτό σημαίνει ότι η εκτίμηση για ένα μεγαλύτερο n μπορεί να απέχει περισσότερο από την πραγματική τιμή από ότι για ένα μικρότερο n. Αυτό που ισχύει, όμως, είναι ότι η πιθανότητα η σχετική συχνότητα να απέχει κατά απόλυτη τιμή από την πραγματική τιμή περισσότερο από κάποια αυθαίρετη θετική ποσότητα ϵ φθίνει μονότονα, καθώς το n αυξάνεται.

1.3.3 Υποκειμενικός ορισμός

Εν αντιθέσει με τους δύο προηγούμενους ορισμούς, ο υποκειμενικός ορισμός της πιθανότητας βασίζεται όχι σε αντικειμενικές προσεγγίσεις προσδιορισμού της πιθανότητας ενός ενδεχομένου, αλλά σε υποκειμενικές κρίσεις, απόψεις, εκτιμήσεις. Η προσέγγιση αυτή είναι ίσως μία από τις συνηθέστερες προσεγγίσεις εκτίμησης της πιθανότητας ενός ενδεχομένου στην καθημερινή ζωή.

Σε κάθε περίπτωση, όμως, ο υποκειμενικός ορισμός, δηλαδή η ανάθεση μιας τιμής στην πιθανότητα παρατήρησης ενός ενδεχομένου δεν μπορεί να αντιβαίνει τις απαιτήσεις των ιδιοτήτων των πιθανοτήτων, όπως αυτές προκύπτουν από τους αντικειμενικούς ορισμούς της πιθανότητας. Έτσι, παραδείγματος χάριν, δεν είναι σωστή η πρόταση ενός μετεωρολόγου ότι είναι σίγουρος 110% ότι την επόμενη μέρα θα βρέξει σε μια συγκεκριμένη περιοχή. Η πιθανότητα, όπως ορίστηκε στον κλασικό ορισμό αλλά και στον στατιστικό ορισμό, είναι ένας αριθμός πάντα μικρότερος ή ίσος του 1 και μεγαλύτερος ή ίσος του 0.

Λάθη όπως το παραπάνω αλλά και οι αδυναμίες των αντικειμενικών ορισμών οδήγησαν στην ανάγκη ενός πιο αυστηρού, μαθηματικοποιημένου ορισμού της πιθανότητας. Ένας τέτοιος ορισμός είναι ο αποκαλούμενος αξιωματικός ορισμός της πιθανότητας, που παρουσιάζεται στη συνέχεια.

1.3.4 Αξιωματικός ορισμός

Όλοι οι ορισμοί της πιθανότητας που παρουσιάστηκαν παραπάνω έχουν βασικά μειονεκτήματα και περιορισμένο πεδίο εφαρμογής. Το 1933 ο Andrey Kolmogorov διατύπωσε έναν αυστηρό ορισμό της

```r
set.seed(0)
n=3000;
f1=matrix(NA,nrow=n,ncol=1)
f2=matrix(NA,nrow=n,ncol=1)
j1=0;
j2=0;
for(i in 1:n){
  zari1=sample(c(1,2,3,4,5,6), 1)
  zari2=sample(c(1,1,1,1,1,2,3,4,5,6), 1)
  if(zari1==3){
        j1=j1+1;
    }

  if(zari2==3){
        j2=j2+1;
  }
  f1[i]=j1/i;
  f2[i]=j2/i;
}

plot(f1,type="l", ylab={"rel. freq."}, xlab="n")
lines(f2, col = "red")
abline(1/6,0,col="blue")
abline(1/10,0,col="blue")
```

Κώδικας 1.1: Κώδικας R προσομοίωσης ρίψης ενός «τίμιου» και ενός «μη τίμιου» ζαριού 3000 φορές.

πιθανότητας βασιζόμενος σε τρία αξιώματα. Για τον λόγο αυτόν, ο συγκεκριμένος ορισμός αναφέρεται ως «αξιωματικός ορισμός της πιθανότητας». Ο αξιωματικός ορισμός της πιθανότητας ξεπερνάει τις αδυναμίες των προηγούμενων ορισμών και δεν έρχεται σε αντίθεση με τους αρχικούς ορισμούς της πιθανότητας, στις περιπτώσεις που αυτοί μπορούν να εφαρμοστούν.

Ορισμός 1.16

Έστω Ω ο δειγματικός χώρος ενός πειράματος και 2^{Ω} το δυναμοσύνολό του. Ως πιθανότητα ενός ενδεχομένου A, $A \subseteq \Omega$, ορίζεται η τιμή $P(A)$ μιας συνάρτησης $P : 2^{\Omega} \to [0,1]$, που ικανοποιεί τα ακόλουθα τρία αξιώματα:

1. για κάθε ενδεχόμενο A (δηλαδή για κάθε $A \subseteq \Omega$), $P(A) \geq 0$,
2. $P(\Omega) = 1$,
3. για κάθε αριθμήσιμη συλλογή ξένων ανά δύο ενδεχομένων $A_1, A_2, \ldots$, ισχύει

$$P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i).$$

Ο παραπάνω ορισμός ορίζει την πιθανότητα ενός ενδεχομένου ως την τιμή μιας συνάρτησης, της συνάρτησης πιθανότητας ή, αλλιώς, μέτρο πιθανότητας. Η συνάρτηση πιθανότητας ορίζεται πάνω σε σύνολα και είναι φανερό ότι δεν υπάρχει μοναδικός τρόπος να ορίσουμε τη συνάρτηση P για ένα πείραμα τύχης. Το γεγονός αυτό δίνει μια ευελιξία, αφού μπορούμε να ορίσουμε διαφορετικές συναρτήσεις για ένα πείραμα τύχης και να εξετάσουμε, στη συνέχεια, ποια από αυτές περιγράφει καλύτερα το φαινόμενο.

Αξίζει να σημειωθεί ότι ο αξιωματικός ορισμός της πιθανότητας, στην πραγματικότητα, δεν απαιτεί η συνάρτηση πιθανότητας να έχει πεδίο ορισμού το δυναμοσύνολο του Ω, αλλά να ορίζεται σε μια συλλογή υποσυνόλων του που αποτελούν, όπως ονομάζεται, μια $\sigma-$άλγεβρα. Σημειώνεται ότι το δυναμοσύνολο του

Ω αποτελεί τη μεγαλύτερη $\sigma-$άλγεβρα που μπορεί να δημιουργηθεί από υποσύνολα του Ω. Για λόγους πληρότητας παρατίθεται στη συνέχεια ο ορισμός της $\sigma-$άλγεβρας.

Ορισμός 1.17

Ως σ-άλγεβρα $\mathcal{F}$ ορίζεται κάθε συλλογή από υποσύνολα του Ω που ικανοποιεί τις ακόλουθες ιδιότητες:

1. $\Omega \in \mathcal{F}$,
2. αν $A \in \mathcal{F}$, τότε $A' \in \mathcal{F}$ και
3. αν έχουμε μια ακολουθία συνόλων A_n, $n = 1,2,\ldots$ στο $\mathcal{F}$, τότε η ένωση των A_n ανήκει, επίσης, στο $\mathcal{F}$.

Με βάση τον αξιωματικό ορισμό της πιθανότητας μπορούν να αποδειχθούν μια σειρά από ιδιότητες, οι οποίες είναι ιδιαίτερα χρήσιμες για την επίλυση σύνθετων προβλημάτων. Οι ιδιότητες αυτές μπορούν να προκύψουν και από τον κλασικό ορισμό πιθανότητας, αλλά αρκετές φορές με περισσότερο κόπο.

Πρόταση 1.2

Η πιθανότητα του αδύνατου ενδεχομένου ισούται με μηδέν, δηλαδή

$$P(\varnothing) = 0.$$

Απόδειξη Πρότασης 1.2

Αν υποθέσουμε ότι $P(\varnothing) = \alpha$ και θέσουμε $A_1 = A_2 = \ldots = \varnothing$, τότε το αριστερό μέρος του τρίτου αξιώματος εκφράζεται ως

$$P(\cup_{i=1}^{\infty} A_i) = P(\cup_{i=1}^{\infty} \varnothing) = P(\varnothing) = \alpha,$$

ενώ το δεξί μέρος ισούται με $\sum_{i=1}^{\infty} P(A_i) = \sum_{i=1}^{\infty} \alpha$.
Επομένως, η σταθερά α πρέπει να ικανοποιεί την ταυτότητα:

$$\alpha = \sum_{i=1}^{\infty} \alpha.$$

Ο μοναδικός, όμως, αριθμός που ικανοποιεί την παραπάνω ταυτότητα είναι το μηδέν και, επομένως, $P(\varnothing) = 0$.

Πρόταση 1.3

Για κάθε πεπερασμένη συλλογή ξένων ανά δύο ενδεχομένων $A_1, A_2, \ldots, A_n$ ενός δειγματικού χώρου ισχύει ότι:

$$P(\cup_{i=1}^{n} A_i) = \sum_{i=1}^{n} P(A_i).$$

Απόδειξη Πρότασης 1.3

Συμπληρώνοντας την πεπερασμένη συλλογή ξένων ανά δύο ενδεχομένων $A_1, A_2, \ldots, A_n$ με άπειρες επαναλήψεις του αδύνατου ενδεχομένου και λαμβάνοντας υπόψη το τρίτο αξίωμα του αξιωματικού ορισμού της πιθανότητας και την Πρόταση 1.2, προκύπτει ότι:

$$P(\cup_{i=1}^{n} A_i) = P(\cup_{i=1}^{n} A_i \cup \varnothing \cup \varnothing \ldots) = \sum_{i=1}^{n} P(A_i) + \sum_{i=n+1}^{\infty} P(\varnothing) = \sum_{i=1}^{n} P(A_i) + 0 = \sum_{i=1}^{n} P(A_i)$$

το οποίο αποδεικνύει το ζητούμενο.

Πρόταση 1.4

Για κάθε ενδεχόμενο A ενός δειγματικού χώρου ισχύει $P(A') = 1 - P(A)$.

Απόδειξη Πρότασης 1.4

Γνωρίζουμε ότι $A \cup A' = \Omega$ και $A \cap A' = \varnothing$. Αυτό σημαίνει ότι τα A και A' είναι ξένα μεταξύ τους, χρησιμοποιώντας την Πρόταση 1.3 και το δεύτερο αξίωμα του αξιωματικού ορισμού της πιθανότητας, έχουμε ότι:

$$1 = P(\Omega) = P(A \cup A') = P(A) + P(A')$$

από την οποία προκύπτει πράγματι η ζητούμενη σχέση $P(A') = 1 - P(A)$.

Πρόταση 1.5

Για οποιαδήποτε δύο ενδεχόμενα A και B ενός δειγματικού χώρου ισχύει

$$P(A \setminus B) = P(A \cap B') = P(A) - P(A \cap B)$$

Απόδειξη Πρότασης 1.5

Παρατηρήστε ότι

$$(A \setminus B) \cup (A \cap B) = A,$$

δηλαδή ότι το ενδεχόμενο A εκφράζεται ως ένωση δύο ξένων ενδεχομένων. Χρησιμοποιώντας την Πρόταση 1.3 είναι

$$P(A) = P(A \setminus B) + P(A \cap B)$$

από όπου προκύπτει άμεσα το ζητούμενο.

Πρόταση 1.6

Αν $A \subseteq B$, τότε

1. $P(A) \leq P(B)$, και
2. $P(A \cap B) = P(A)$.

Απόδειξη Πρότασης 1.6

1. Αφού το $A \subseteq B$, το B μπορεί να εκφραστεί ως

   $$B = A \cup (B \setminus A),$$

   το οποίο μας οδηγεί στη σχέση:

   $$P(B) = P(A) + P(B \setminus A).$$

   Από το πρώτο αξίωμα του αξιωματικού ορισμού της πιθανότητας προκύπτει ότι η πιθανότητα είναι μη αρνητικός αριθμός και επομένως, καθώς, $P(B \setminus A) \geq 0$, άμεσα προκύπτει ότι: $P(B) - P(A) \geq 0$, που αποδεικνύει τη ζητούμενη ανισότητα.
2. Αφού το $A \subseteq B$, έχουμε ότι $A \cap B = A$. Από την τελευταία σχέση προκύπτει ότι $P(A \cap B) = P(A)$, καθώς η P είναι συνάρτηση.

Πρόταση 1.7

Για κάθε ενδεχόμενο A ενός δειγματικού χώρου ισχύει $P(A) \leq 1$.

Απόδειξη Πρότασης 1.7

Αφού το A είναι ενδεχόμενο, έχουμε ότι $A \subseteq \Omega$. Αυτό με τη σειρά του συνεπάγεται, σύμφωνα με την Πρόταση 1.6, ότι $P(A) \leq P(\Omega) = 1$ (δεύτερο αξίωμα).

Πρόταση 1.8

Για οποιαδήποτε δύο ενδεχόμενα A και B ενός δειγματικού χώρου ισχύει ότι:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

Απόδειξη Πρότασης 1.8

Παρατηρήστε αρχικά ότι το $A \cup B$ εκφράζεται ως ένωση ξένων ανά δύο ενδεχομένων ως ακολούθως:

$$A \cup B = (A \setminus B) \cup (A \cap B) \cup (B \setminus A).$$

Έπειτα, χρησιμοποιώντας τις Προτάσεις 1.3 και 1.5, έχουμε ότι:

$$\begin{aligned} P(A \cup B) &= P(A \setminus B) + P(A \cap B) + P(B \setminus A) \\ &= P(A) - P(A \cap B) + P(A \cap B) + P(B) - P(A \cap B) \\ &= P(A) + P(B) - P(A \cap B). \end{aligned}$$

Η παραπάνω πρόταση μπορεί να γενικευτεί για περισσότερα από δύο ενδεχόμενα και αναφέρεται ως γενικευμένος κανόνας της πρόσθεσης.

Πρόταση 1.9: Γενικευμένος κανόνας της πρόσθεσης

Για οποιαδήποτε n το πλήθος ενδεχόμενα $A_1, A_2, \ldots, A_n$ ενός δειγματοχώρου Ω ισχύει ότι:

$$P(\cup_{i=1}^{n} A_i) = \sum_{i=1}^{n} P(A_i) - \underbrace{\sum_{i} \sum_{j} P(A_i \cap A_j)}_{1 \leq i < j \leq n} + \underbrace{\sum_{i} \sum_{j} \sum_{k} P(A_i \cap A_j \cap A_k)}_{1 \leq i < j < k \leq n} + \ldots + (-1)^{n-1} P(\cap_{i=1}^{n} A_i)$$

Απόδειξη Πρότασης 1.9

Η απόδειξη της πρότασης προκύπτει με εφαρμογή της μεθόδου της μαθηματικής επαγωγής και αφήνεται ως άσκηση στον/στην αναγνώστη/στρια.

Με βάση τις παραπάνω προτάσεις μπορούν να υπολογιστούν οι πιθανότητες διάφορων ενδεχομένων σε σύνθετα προβλήματα, όπως αυτό που ακολουθεί.

Παράδειγμα 1.8

Τα θεμέλια ενός ψηλού κτηρίου μπορούν να αστοχήσουν από θραύση του εδάφους ή από υπερβολικές καθιζήσεις. Η πιθανότητα να αστοχήσουν από την πρώτη αιτία είναι 0.01, ενώ η πιθανότητα να αστοχήσουν από τη δεύτερη αιτία είναι 0.02. Η πιθανότητα να έχουμε αστοχία των θεμελίων λόγω ύπαρξης και των δύο αιτιών είναι ίση με 0.006.

Να βρεθεί η πιθανότητα τα θεμέλια ενός τυχαία επιλεγμένου ψηλού κτηρίου:

1. να αστοχήσουν εξαιτίας μίας τουλάχιστον εκ των δύο αιτιών,
2. να μην αστοχήσουν από καμία από τις δύο αυτές αιτίες,
3. να αστοχήσουν είτε μόνο εξαιτίας της θραύσης του εδάφους είτε μόνο εξαιτίας υπερβολικών καθιζήσεων.

Λύση Παραδείγματος 1.8

Για την επίλυση των ερωτημάτων της άσκησης ορίζονται αρχικά τα ακόλουθα ενδεχόμενα:

$$A = \{\text{τα θεμέλια να αστοχήσουν από θραύση του εδάφους}\}$$
$$B = \{\text{τα θεμέλια να αστοχήσουν από υπερβολικές καθιζήσεις}\}.$$

Με βάση τα δεδομένα της άσκησης έχουμε ότι:

$$P(A) = 0.01,\ \ P(B) = 0.02,\ \ P(A \cap B) = 0.006.$$

1. Το ενδεχόμενο τα θεμέλια ενός ψηλού κτηρίου να αστοχήσουν εξαιτίας μίας τουλάχιστον από τις δύο αυτές αιτίες εκφράζεται ως $A \cup B$ και επομένως:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0.01 + 0.02 - 0.006 = 0.024.$$

2. Η πιθανότητα τα θεμέλια ενός ψηλού κτηρίου να μην αστοχήσουν από καμία από τις δύο αυτές αιτίες εκφράζεται ως $A' \cap B'$ και υπολογίζεται σύμφωνα με τη σχέση:

$$P(A' \cap B') = P((A \cup B)') = 1 - P(A \cup B) = 1 - 0.024 = 0.976.$$

3. Το ζητούμενο ενδεχόμενο είναι η συμμετρική διαφορά των A και B, η οποία συμβολίζεται $A \oplus B$, και όπως έχουμε αναφέρει, εκφράζεται ως:

$$(A \cap B') \cup (A' \cap B).$$

Επομένως, η πιθανότητα τα θεμέλια ενός ψηλού κτηρίου να αστοχήσουν είτε μόνο εξαιτίας της θραύσης του εδάφους είτε μόνο εξαιτίας υπερβολικών καθιζήσεων υπολογίζεται από τη σχέση:

$$\begin{aligned} P(A \oplus B) &= P((A \cap B') \cup (A' \cap B)) \\ &= P(A \cap B') + P(A' \cap B) \\ &= P(A) - P(A \cap B) + P(B) - P(A \cap B) \\ &= P(A) + P(B) - 2 \cdot P(A \cap B) \\ &= 0.01 + 0.02 - 2 \cdot 0.006 = 0.018. \end{aligned}$$

Άσκηση Αυτοαξιολόγησης 1.5

Μια εταιρεία θέλει να συμμετάσχει σε τρεις διαγωνισμούς, τους διαγωνισμούς Α, Β και Γ. Με βάση τις προσφορές που έχει καταθέσει υπολογίζει ότι έχει 20% πιθανότητα να επιλεχθεί στον Α διαγωνισμό, 50% στον Β και 40% στον Γ. Υπολογίζει ακόμα ότι η πιθανότητα να επιλεχθεί σε τουλάχιστον έναν από τους διαγωνισμούς Α και Β ισούται με 60%. Σημειώνεται ότι, λόγω των κανονισμών, η εταιρεία δεν μπορεί να επιλεχθεί ταυτόχρονα και στον Β και στον Γ διαγωνισμό.
Σύμφωνα με τις παραπάνω εκτιμήσεις της εταιρείας και τους κανονισμούς των διαγωνισμών, να βρείτε την πιθανότητα η εταιρεία:

1. να μην επιλεχθεί στον Γ διαγωνισμό,
2. να επιλεχθεί στον Α και στον Β διαγωνισμό,
3. να επιλεχθεί στον Α, αλλά όχι και στον Β,
4. να επιλεχθεί ή μόνο στον Α ή μόνο στον Β,
5. να επιλεχθεί στον Β ή στον Γ,
6. να μην επιλεχθεί ούτε στον Α ούτε στον Β.

Παράδειγμα 1.9

Έστω A και B δύο ενδεχόμενα ενός δειγματικού χώρου Ω. Να δείξετε ότι:

$$P(A) + P(B) \geq 2 \cdot P(A) \cdot P(B).$$

Λύση Παραδείγματος 1.9

Θέλουμε να δείξουμε ότι

$$P(A) + P(B) \geq 2 \cdot P(A) \cdot P(B)$$

ή, ισοδύναμα, ότι

$$P(A) + P(B) - 2 \cdot P(A) \cdot P(B) \geq 0.$$

Είναι

$$\begin{aligned} P(A) + P(B) - 2 \cdot P(A) \cdot P(B) &= P(A) - P(A) \cdot P(B) + P(B) - P(A) \cdot P(B) \\ &= P(A) \cdot (1 - P(B)) + P(B) \cdot (1 - P(A)) \\ &= P(A)P(B') + P(B)P(A') \geq 0, \end{aligned}$$

καθώς για κάθε ενδεχόμενο, έστω E ενός δειγματοχώρου Ω, ισχύει ότι $P(E) \geq 0$.

1.4 Ασκήσεις

Άσκηση 1.1 *Δώστε την κατάλληλη απάντηση (ΣΩΣΤΟ ή ΛΑΘΟΣ) στις κάτωθι προτάσεις. Αιτιολογήστε σύντομα τις απαντήσεις σας.*

1. *Έστω A και B δύο ενδεχόμενα ορισμένα στον ίδιο δειγματοχώρο Ω. Το ενδεχόμενο να πραγματοποιηθεί το πολύ ένα από τα A και B είναι το $(A' \cap B) \cup (A \cap B')$.*
2. *Έστω δύο ασυμβίβαστα ενδεχόμενα A και B, με $P(A) = 0.3$ και $P(A \cup B) = 0.65$. Τότε η πιθανότητα $P(A \cap B)$ ισούται με 0.15.*

Άσκηση 1.2 *Έστω A και B δύο ξένα μεταξύ τους ενδεχόμενα, με $P(A) = 0.3$ και $P(B) = 0.2$. Υπολογίστε τις πιθανότητες: $P(A \cup B)$, $P(A \cap B)$, $P(A' \cup B')$.*

Άσκηση 1.3 *Ένας μηχανικός δηλώνει 1000% σίγουρος ότι ο τρόπος κατασκευής και τα υλικά που χρησιμοποιεί για την πεζογέφυρα που κατασκευάζει εξασφαλίζουν ότι αυτή θα αντέξει ακόμα και τη διέλευση των ποδηλατών ενός αγώνα. Ποια από τις παρακάτω προτάσεις είναι σωστή; Δικαιολογήστε την απάντησή σας.*

1. *Ο μηχανικός είναι υπερβολικός. Θα μπορούσε να κατασκευάσει τη γέφυρα με τέτοιον τρόπο ώστε να είναι 200% σίγουρος.*
2. *Ο μηχανικός έχει κάνει την κατασκευή της πεζογέφυρας με τέτοιον τρόπο που του δίνει τόσο μεγάλη σιγουριά.*
3. *Ο μηχανικός κάνει λάθος. Καμιά κατασκευή δεν μπορεί να έχει τέτοιον βαθμό σιγουριάς.*

Άσκηση 1.4 *Εξετάστε αν ισχύει $P(A \cap B) \leq P(A) + P(B) - 1$. Αναπτύξτε την απάντησή σας.*

Άσκηση 1.5 *Αν A και B είναι δύο ενδεχόμενα ενός δειγματικού χώρου Ω για τα οποία ισχύει $P(A \cap B') = P(A' \cap B) = 0$, τότε ποια από τις παρακάτω προτάσεις ισχύει;*

1. $P(A) = P(B)$.
2. $P(A \cup B) = P(A \cap B)$.
3. $P((A \cap B') \cup (A' \cap B)) = 0$.
4. *Ισχύουν όλα τα παραπάνω.*

Δικαιολογήστε την απάντησή σας.

Άσκηση 1.6 *Αν A και B είναι δύο ενδεχόμενα του ίδιου δειγματικού χώρου Ω και γνωρίζουμε ότι $P(A) = \frac{1}{3}$ και $P(B) = \frac{3}{8}$, να αποδειχθεί ότι:*

$$\frac{3}{8} \leq P(A \cup B) \leq \frac{17}{24}.$$

Άσκηση 1.7 *Χρησιμοποιώντας οποιαδήποτε γλώσσα προγραμματισμού κατασκευάστε ένα πρόγραμμα που επιλέγει τυχαία έναν ακέραιο από το 1 μέχρι το n, $n > 2$, έτσι ώστε η πιθανότητα επιλογής κάθε ακέραιου να είναι ανάλογη της τιμής του.*

1. *Υπολογίστε την πιθανότητα επιλογής του αριθμού k, όπου $k = 1, \ldots, n$.*
2. *Υπολογίστε την πιθανότητα να επιλεχθεί ο αριθμός 1 ή ο αριθμός n.*

Άσκηση 1.8 *Υπολογίστε πόσες διαφορετικές «λέξεις» με 8 χαρακτήρες μπορούμε να φτιάξουμε χρησιμοποιώντας τα σύμβολα 0 και 1.*

Άσκηση 1.9 *Ένας ψυχολόγος θέλει να σχηματίσει λέξεις που αποτελούνται από τρία γράμματα ώστε να τις χρησιμοποιήσει σε ένα τεστ μνήμης. Το πρώτο γράμμα κάθε λέξης μπορεί να είναι ένα από τα σύμφωνα Σ, Κ, Π, Λ και Ν, το δεύτερο ένα από τα φωνήεντα Α, Ε, Η, Ι, Ο και Υ και το τρίτο κάποιο από τα Σ, Ν και Μ. Αν κάθε λέξη έχει την ίδια πιθανότητα να σχηματιστεί, να προσδιοριστεί η πιθανότητα:*

1. *να προκύψει μια λέξη που να μην τελειώνει με το γράμμα Μ και να μην περιέχει το γράμμα Ε,*
2. *να προκύψει μια λέξη που να περιέχει το γράμμα Ν,*
3. *να προκύψει μια λέξη που να έχει ως πρώτο γράμμα της το Ν, γνωρίζοντας ότι η λέξη περιέχει το γράμμα Ν.*

Άσκηση 1.10 *Από τους 20 υπολογιστές ενός εργαστηρίου οι 15 είναι συνδεδεμένοι στο τοπικό δίκτυο. Επιλέγονται τυχαία 7 διαφορετικοί υπολογιστές. Ποια η πιθανότητα 2 από τους 7 που επιλέχθηκαν να μην είναι συνδεδεμένοι στο τοπικό δίκτυο;*

Άσκηση 1.11 *Στο τμήμα επισκευών μιας εταιρείας ηλεκτρικών συσκευών υπάρχουν προς επισκευή 10 τηλεοράσεις και 12 υπολογιστές. Το προσωπικό που διαθέτει η εταιρεία επισκευάζει σε μια μέρα μόνο 6 συσκευές. Αν οι συσκευές που θα επισκευάσει το προσωπικό επιλέγονται τυχαία, να προσδιοριστεί η πιθανότητα σε μια μέρα:*

1. *να επισκευαστούν 3 τηλεοράσεις και 3 υπολογιστές,*
2. *να επισκευαστούν το πολύ 5 τηλεοράσεις.*

1.5 Απαντήσεις στις ασκήσεις αυτοαξιολόγησης

Λύση Άσκησης Αυτοαξιολόγησης 1.1

1. Ο δειγματοχώρος του τυχαίου πειράματος της ρίψης ενός ζαριού και καταγραφής της ένδειξής του είναι ο $\Omega = \{1,2,3,4,5,6\}$. Επιπλέον, το δυναμοσύνολό του είναι το

$$\begin{aligned} 2^{\Omega} = &\{\varnothing,\{1\},\{2\},\dots,\{6\},\\ &\{1,2\},\{1,3\},\dots,\{5,6\},\{1,2,3\},\dots,\\ &\{1,2,3,4\},\dots,\{1,2,3,4,5\},\dots,\Omega\}. \end{aligned}$$

   Το πλήθος των στοιχείων του 2^{Ω} είναι $2^6 = 64$.

2. Κατά την καταγραφή των ελαττωματικών προϊόντων σε μια παρτίδα 10 τεμαχίων μπορούν να βρεθούν $0,1,2,\dots,10$ ελαττωματικά προϊόντα. Επομένως, ο δειγματοχώρος είναι ο $\Omega = \{0,1,2,\dots,10\}$. Η πλήρης καταγραφή των στοιχείων του δυναμοσυνόλου του Ω είναι μια επίπονη διαδικασία, καθώς το πλήθος τους ισούται με $2^{11} = 2048$. Το δυναμοσύνολο όμως περιέχει:
   - το αδύνατο ενδεχόμενο $\varnothing$,
   - όλα τα στοιχειώδη ενδεχόμενα $\{i\},\ i = 0,1\dots,10$,
   - όλα τα ενδεχόμενα με δύο στοιχεία $\{i,j\},\ 0 \le i < j \le 10$,
   - όλα τα ενδεχόμενα με τρία στοιχεία $\{i,j,k\},\ 0 \le i < j < k \le 10$,
   - ...
   - το βέβαιο ενδεχόμενο Ω.

Λύση Άσκησης Αυτοαξιολόγησης 1.2

Αν $A_1, A_2, \dots, A_k$ είναι μια συλλογή ξένων ανά δύο συνόλων, τότε $A_i \cap A_j = \varnothing,\ \forall\, i,j = 1,2,\dots,k$, με $i \neq j$. Αυτό σημαίνει ότι το $A_1 \cap A_2 \cap A_3 \cap \dots \cap A_k$ μπορεί να εκφραστεί ως

$$(A_1 \cap A_2) \cap A_3 \cap \dots \cap A_k = \varnothing \cap A_3 \cap \dots \cap A_k.$$

Η τομή, όμως, του κενού συνόλου με το A_3 είναι πάλι το κενό σύνολο και, ακολουθώντας παρόμοιο συλλογισμό και για τις επόμενες τομές, καταλήγουμε ότι $A_1 \cap A_2 \cap A_3 \cap \dots \cap A_k = \varnothing$, δηλαδή ότι τα $A_1, A_2, \dots, A_k$ είναι ασυμβίβαστα. Επομένως, αν μια συλλογή συνόλων αποτελείται από ξένα ανά δύο μεταξύ τους σύνολα, τότε τα σύνολα που αποτελούν αυτήν τη συλλογή είναι και ασυμβίβαστα.
Το αντίθετο δεν ισχύει, όπως μπορούμε να δούμε από το ακόλουθο αντιπαράδειγμα. Έστω $\Omega = \{1,2,3,4,5,6\}$ είναι ένας δειγματοχώρος και $A = \{1,2,3\}$, $B = \{2,3,4\}$ και $\Gamma = \{4.5.6\}$ τρία υποσύνολά του. Τότε, εύκολα μπορούμε να δούμε ότι η τομή των τριών αυτών συνόλων είναι το κενό σύνολο, αλλά η ανά δύο τομή τους είναι μη κενή. Αυτό σημαίνει ότι τα σύνολα A, B και Γ είναι ασυμβίβαστα, αλλά όχι ξένα ανά δύο μεταξύ τους.

Λύση Άσκησης Αυτοαξιολόγησης 1.3

Όχι, δεν μπορούμε να εφαρμόσουμε τον κλασικό ορισμό της πιθανότητας, για να υπολογίσουμε πιθανότητες διάφορων ενδεχομένων του Παραδείγματος 1.2, γιατί τα στοιχειώδη ενδεχόμενα δεν είναι ισοπίθανα. Ακόμα και αν υποθέσουμε ότι το νόμισμα είναι «τίμιο», δεν μπορούμε να εφαρμόσουμε τον κλασικό ορισμό της πιθανότητας, αφού το να φέρουμε 0 φορές γράμματα στις 6 ρίψεις του νομίσματος δεν μπορεί να έχει την ίδια πιθανότητα με το να φέρουμε 3 φορές γράμματα στις 6 ρίψεις.

Λύση Άσκησης Αυτοαξιολόγησης 1.4

Για να υπολογίσουμε την πιθανότητα ανάμεσα στους τρεις ορόρφους που θα επιλέξουμε να είναι ο 10ος όροφος, θα ακολουθήσουμε ίδια συλλογιστική με αυτήν του Παραδείγματος 1.7. Στο πλαίσιο αυτό, ορίζουμε το ενδεχόμενο

$$A = \{\text{ανάμεσα στους τρεις ορόφους που θα επιλέξουμε να είναι ο 10ος όροφος}\}.$$

Το A αποτελείται από όλες τις δυνατές τριάδες (i,j,k) που περιέχουν το 10. Χωρίς βλάβη της γενικότητας, αφού δεν μας ενδιαφέρει η σειρά επιλογής, μπορούμε να θεωρήσουμε ότι το πρώτο, δηλαδή το i, ισούται με 10. Στη συνέχεια, θα χρησιμοποιήσουμε τη βασική αρχή απαρίθμησης και θα θεωρήσουμε ότι, αντί να επιλέγουμε ταυτόχρονα τους τρεις ορόφους, επιλέγουμε πρώτα τον 10ο όροφο και στη συνέχεια τους άλλους δύο ορόφους από τους υπόλοιπους εννιά το πλήθος ορόφους. Η πρώτη επιλογή μας, επειδή πρέπει να είναι ο 10ος όροφος, μπορεί να γίνει μόνο με έναν τρόπο, ενώ η δεύτερη επιλογή μας (των δύο υπόλοιπων ορόφων) μπορεί να γίνει με τόσους διαφορετικούς τρόπους όσοι και οι συνδυασμοί των 9 ανά 2. Έτσι, συνολικά, το πλήθος των στοιχείων του A, δηλαδή των ευνοϊκών περιπτώσεων, ισούται με $1 \cdot \binom{9}{2}$. Επομένως, με βάση τον κλασικό ορισμό της πιθανότητας έχουμε ότι:

$$P(A) = \frac{1 \cdot \binom{9}{2}}{\binom{10}{3}} = \frac{3}{10} = 0.3.$$

Λύση Άσκησης Αυτοαξιολόγησης 1.5

Αρχικά, ορίζουμε τα ακόλουθα ενδεχόμενα:

$$A = \{\text{η εταιρεία να επιλεχθεί στον Α διαγωνισμό}\},$$
$$B = \{\text{η εταιρεία να επιλεχθεί στον Β διαγωνισμό}\},$$
$$\Gamma = \{\text{η εταιρεία να επιλεχθεί στον Γ διαγωνισμό.}\}$$

Με βάση τα δεδομένα της άσκησης έχουμε ότι:

$$P(A) = 0.2, \;\; P(B) = 0.5, \;\; P(\Gamma) = 0.4, \;\; P(A \cup B) = 0.6$$

και ότι $B \cap \Gamma = \varnothing$, δηλαδή $P(B \cap \Gamma) = 0$.
Οι ζητούμενες πιθανότητες υπολογίζονται χρησιμοποιώντας τον Αξιωματικό Ορισμό της Πιθανότητας και τις προτάσεις που παρουσιάστηκαν στην Ενότητα 1.3.4 ως ακολούθως:

1. $P(\Gamma') = 1 - P(\Gamma) = 1 - 0.4 = 0.6.$
2. $P(A \cap B) = P(A) + P(B) - P(A \cup B) = 0.2 + 0.5 - 0.6 = 0.1.$
3. $P(A \cap B') = P(A) - P(A \cap B) = 0.2 - 0.1 = 0.1.$
4. $P(A \oplus B) = P(A) + P(B) - 2P(A \cap B) = 0.2 + 0.5 - 2 \cdot 0.1 = 0.5.$
5. $P(B \cup \Gamma) = P(B) + P(\Gamma) - P(B \cap \Gamma) = 0.5 + 0.4 - 0 = 0.9.$
6. $P(A' \cap B') = P((A \cup B)') = 1 - P(A \cup B) = 1 - 0.6 = 0.4.$

ΒΙΒΛΙΟΓΡΑΦΙΑ

Ελληνόγλωσση

Enderton, H. B. (2013). *Μια μαθηματική εισαγωγή στη λογική.* Μετάφραση: Παπαδόγγονας, Ιωάννης. Ηράκλειο: Πανεπιστημιακές Εκδόσεις Κρήτης.

Γεωργίου, Δ. και Ηλιάδης, Σ. (2017). *Θεωρία Συνόλων* (2η έκδ.). Θεσσαλονίκη: Τζιόλας.

Κάλφα, Κ. (1990). *Αξιωματική θεωρία συνόλων*. Θεσσαλονίκη: Εκδόσεις Ζήτη.

Μητακίδης, Γ. (1988). *Θεωρία Συνόλων*. Πάτρα: Εκδόσεις Πανεπιστημίου Πατρών.

Ξενόγλωσση

Halmos, P. R. (1974). *Naive set theory*. Undergraduate texts in mathematics. New York: Springer Verlag.

Kolmogorov, A. (1956). *Foundations of the Theory of Probability*. AMS Chelsea Publishing Series. Chelsea Publishing Company.

Kolmogorov, A. (2013). *Foundations of the Theory of Probability*. Martino Fine Books.

Laplace, P. (1814). *A Philosophical Essay on Probabilities.* Translated by Trruscott, Frederick Wilson; Emory, Frederick Lincoln. John Wiley & Son and Chapman & Hall.

Siegel, A. (2011). *Practical Business Statistics*. Elsevier Science.

Von Mises, R. (1919). Grundlagen der Wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift*, *5*, pp. 52–99.

ΚΕΦΑΛΑΙΟ 2

ΔΕΣΜΕΥΜΕΝΗ ΠΙΘΑΝΟΤΗΤΑ, ΘΕΩΡΗΜΑ BAYES ΚΑΙ ΑΝΕΞΑΡΤΗΣΙΑ

Σύνοψη

Σε αυτό το κεφάλαιο παρουσιάζονται έννοιες, όπως η δεσμευμένη πιθανότητα και η ανεξαρτησία ενδεχομένων, και αναπτύσσονται μέθοδοι με τη βοήθεια θεμελιωδών θεωρημάτων για την επίλυση σύνθετων προβλημάτων υπολογισμού πιθανοτήτων.

Προαπαιτούμενη γνώση: Να γνωρίζετε και να έχετε κατανοήσει την έννοια της πιθανότητας.

Προσδοκώμενα μαθησιακά αποτελέσματα: Όταν θα έχετε ολοκληρώσει τη μελέτη του κεφαλαίου αυτού, θα γνωρίζετε

- την έννοια της δεσμευμένης πιθανότητας,
- την έννοια της ανεξαρτησίας ενδεχομένων,
- το Θεώρημα Ολικής Πιθανότητας και το Θεώρημα του Bayes.

Γλωσσάριο επιστημονικών όρων:

- Ανεξάρτητα ενδεχόμενα
- Δεσμευμένη πιθανότητα
- Διαμέριση
- Θεώρημα Ολικής Πιθανότητας
- Θεώρημα του Bayes
- Πολλαπλασιαστικός κανόνας

Οικονόμου, Π., Μαλεφάκη, Σ., & Μπατσίδης, Α. (2023). *Πιθανότητες - Στατιστική.*

2.1 Εισαγωγή

Κάποιες φορές μας ενδιαφέρει η πιθανότητα εμφάνισης ενός ενδεχομένου, όταν το αποτέλεσμα του πειράματος τύχης είναι (ή θεωρούμε ότι είναι) μερικώς γνωστό, ενώ άλλες φορές μας ενδιαφέρει να μελετήσουμε έναν υποπληθυσμό λαμβάνοντας υπόψη πληροφορίες από το σύνολο του πληθυσμού. Ενδεικτικά παραδείγματα τέτοιων περιπτώσεων είναι ο υπολογισμός:

- της πιθανότητας χιονόπτωσης σε μια περιοχή, όταν προβλέπεται ότι η θερμοκρασία θα είναι μικρότερη ή ίση με $-2^{\circ}C$, και
- της πιθανότητας βλάβης ενός κινητού σε μια περίοδο ενός έτους από την αγορά του, όταν γνωρίζουμε την κατασκευάστρια εταιρεία.

Η μερική γνώση του αποτελέσματος ενός πειράματος τύχης συχνά αλλάζει την πιθανότητα εμφάνισης του ζητούμενου ενδεχομένου. Η παραπάνω διατύπωση θα γίνει καλύτερα κατανοητή μέσω του επόμενου παραδείγματος. Υποθέστε ότι το 0.1% του πληθυσμού έχει υψηλό πυρετό μια συγκεκριμένη μέρα. Επιπροσθέτως, γνωρίζουμε ότι τη συγκεκριμένη εποχή κυκλοφορεί ένας ιός που προκαλεί υψηλό πυρετό ανάμεσα σε αυτούς που είναι φορείς του με πιθανότητα 97%. Επομένως, η πιθανότητα κάποιος να έχει υψηλό πυρετό μια συγκεκριμένη μέρα ισούται με μόλις 0.1%, ενώ όταν γνωρίζουμε ότι είναι φορέας του συγκεκριμένου ιού, η πιθανότητα αυξάνεται και ισούται με 97%.

Για να μπορέσουμε να μελετήσουμε τέτοιες περιπτώσεις, χρήσιμο και απαραίτητο εργαλείο είναι η έννοια της **δεσμευμένης πιθανότητας**, η οποία παρουσιάζεται στην επόμενη ενότητα.

2.2 Δεσμευμένη πιθανότητα

Η έννοια της δεσμευμένης πιθανότητας μας επιτρέπει τον υπολογισμό της πιθανότητας πραγματοποίησης ενός ενδεχομένου όταν γνωρίζουμε ότι έχει πραγματοποιηθεί ένα άλλο ενδεχόμενο του ίδιου δειγματικού χώρου.

Ορισμός 2.1

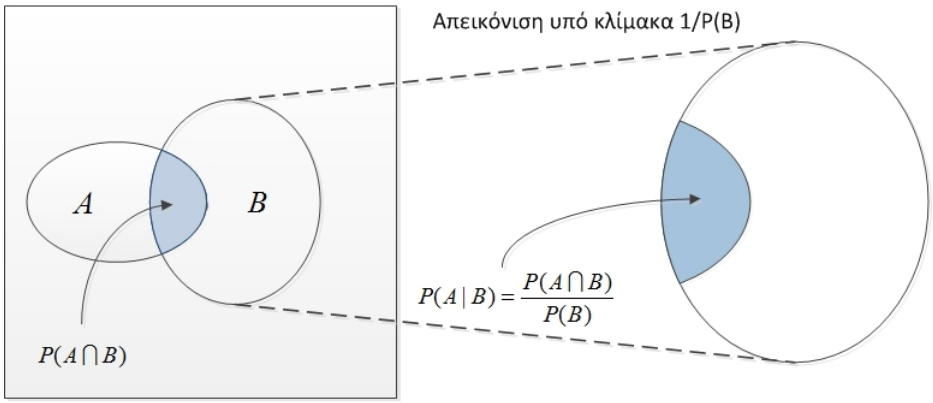
Έστω A και B δύο ενδεχόμενα ενός δειγματοχώρου Ω, με $P(B) > 0$. Η πιθανότητα να πραγματοποιηθεί το A, αν είναι γνωστό ότι έχει πραγματοποιηθεί το B, συμβολίζεται με $P(A|B)$ και ονομάζεται δεσμευμένη πιθανότητα (ή πιθανότητα υπό συνθήκη), ενώ ορίζεται από τη σχέση:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

Η γνώση της πραγματοποίησης του B έχει ως συνέπεια:

- ο δειγματικός χώρος να περιορίζεται στο ενδεχόμενο B, και
- η πραγματοποίηση του A να μπορεί να συμβεί μόνο στην περίπτωση που πραγματοποιηθεί το $A \cap B$.

Τα παραπάνω απεικονίζονται γραφικά στο Σχήμα 2.1. Το δεξί μέρος του σχήματος είναι υπό κλίμακα έτσι ώστε η πιθανότητα του B δοθέντος του ίδιου του B να ισούται με 1.



Σχήμα 2.1: Γραφική απεικόνιση ορισμού δεσμευμένης πιθανότητας.

Παράδειγμα 2.1

Για τα δεδομένα του Παραδείγματος 1.8 να υπολογιστεί η πιθανότητα

1. αστοχίας των θεμελίων ενός τυχαία επιλεγμένου ψηλού κτηρίου λόγω και θραύσης του εδάφους, όταν είναι γνωστό ότι υπερβολικές καθιζήσεις ευθύνονται για την αστοχία των θεμελίων του,
2. αστοχίας των θεμέλιων ενός τυχαία επιλεγμένου ψηλού κτηρίου λόγω υπερβολικών καθιζήσεων, δοθέντος ότι η αστοχία των θεμελίων του οφείλεται στη θραύση του εδάφους.

Λύση Παραδείγματος 2.1

Από τα δεδομένα του Παραδείγματος 1.8 έχουμε ότι:

$$P(A) = 0.01, \;\; P(B) = 0.02, \;\; P(A \cap B) = 0.006,$$

όπου

$$A = \{\text{τα θεμέλια αστοχούν από θραύση του εδάφους}\},$$
$$B = \{\text{τα θεμέλια αστοχούν από υπερβολικές καθιζήσεις}\}.$$

Με βάση τα παραπάνω στοιχεία οι ζητούμενες πιθανότητες υπολογίζονται ως ακολούθως.

1. Αφού είναι γνωστό ότι υπερβολικές καθιζήσεις ευθύνονται για την αστοχία των θεμελίων του κτηρίου, η πιθανότητα να οφείλεται και στη θραύση του εδάφους εκφράζεται από τη δεσμευμένη πιθανότητα $P(A|B)$, η οποία, με βάση τον ορισμό της δεσμευμένης πιθανότητας, ισούται με

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{0.006}{0.02} = 0.3.$$

   Επομένως, η πιθανότητα η αστοχία των θεμελίων του κτηρίου να οφείλεται και στη θραύση του εδάφους, όταν είναι γνωστό ότι υπερβολικές καθιζήσεις ευθύνονται για την αστοχία τους, ισούται με 0.3.

2. Στο ερώτημα αυτό ζητείται ο υπολογισμός της πιθανότητας $P(B|A)$. Είναι εξ ορισμού:

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{0.006}{0.01} = 0.6.$$

   Καταλήγουμε λοιπόν στο συμπέρασμα ότι η πιθανότητα η αστοχία των θεμέλιων του κτηρίου να οφείλεται και στις υπερβολικές καθιζήσεις, δοθέντος ότι η αστοχία τους οφείλεται στη θραύση του εδάφους, ισούται με 0.6.

Παρατήρηση 2.1

Από το παραπάνω παράδειγμα είναι φανερό ότι $P(A|B) \neq P(B|A)$. Η $P(A|B) = P(B|A)$ ισχύει μόνο όταν είτε $P(A) = P(B) > 0$ είτε όταν $P(A \cap B) = 0$, δηλαδή όταν $A \cap B = \varnothing$.
Επομένως, οι δεσμευμένες πιθανότητες $P(A|B)$ και $P(B|A)$ είναι ίσες είτε όταν τα A, B είναι ισοπίθανα ενδεχόμενα είτε όταν είναι ξένα. Στην τελευταία περίπτωση η γνώση πραγματοποίησης ενός εκ των ενδεχομένων A, B καθιστά αδύνατη την πραγματοποίηση του άλλου, αφού είναι ξένα μεταξύ τους.

Άσκηση Αυτοαξιολόγησης 2.1

Για τα δεδομένα της Άσκησης Αυτοαξιολόγησης 1.5 να υπολογίσετε την πιθανότητα η εταιρεία να επιλεγεί στον Β διαγωνισμό, αν είναι γνωστό ότι έχει επιλεγεί στον Α διαγωνισμό.

Είναι σημαντικό να αναφέρουμε ότι οι ιδιότητες των πιθανοτήτων ισχύουν και για τις δεσμευμένες πιθανότητες, υπό την προϋπόθεση ότι η δέσμευση παραμένει σταθερή. Έτσι, για παράδειγμα, αν A, B και Γ είναι τρία ενδεχόμενα ενός δειγματοχώρου Ω, τότε, μεταξύ άλλων, ισχύουν οι ακόλουθες σχέσεις:

- $P(A'|B) = 1 - P(A|B)$, και
- $P(A \cup B|\Gamma) = P(A|\Gamma) + P(B|\Gamma) - P(A \cap B|\Gamma)$.

2.2.1 Πολλαπλασιαστικός κανόνας

Μια ενδιαφέρουσα εφαρμογή του ορισμού της δεσμευμένης πιθανότητας είναι ο υπολογισμός της πιθανότητας της τομής δύο ενδεχομένων με τον αποκαλούμενο **πολλαπλασιαστικό κανόνα** (ή διαφορετικά **κανόνα γινομένου**).

Πρόταση 2.1: Πολλαπλασιαστικός κανόνας

Έστω A και B δύο ενδεχόμενα ενός δειγματοχώρου Ω, με $P(B) > 0$. Τότε ισχύει ότι:

$$P(A \cap B) = P(A|B)P(B).$$

Απόδειξη Πρότασης 2.1

Προκύπτει άμεσα από τον ορισμό της δεσμευμένης πιθανότητας $P(A|B) = \frac{P(A \cap B)}{P(B)}$.

Παρατήρηση 2.2

Από την αντιμεταθετική ιδιότητα της τομής δύο ενδεχομένων προκύπτει ότι αν $P(A) > 0$ και $P(B) > 0$, τότε:

$$P(A \cap B) = P(A|B)P(B) = P(B|A)P(A).$$

Μια ενδιαφέρουσα χρήση του πολλαπλασιαστικού κανόνα αφορά την περίπτωση που η δέσμευση αναφέρεται στο συμπλήρωμα ενός ενδεχομένου. Για παράδειγμα, στην περίπτωση που η δέσμευση αφορά το συμπλήρωμα του B, αυτή μπορεί να εκφραστεί και ως:

$$P(A|B') = \frac{(1 - P(B|A))P(A)}{1 - P(B)},$$

καθώς

$$P(A|B') = \frac{P(A \cap B')}{P(B')} = \frac{P(B'|A)P(A)}{1 - P(B)} = \frac{(1 - P(B|A))P(A)}{1 - P(B)}.$$

Άσκηση Αυτοαξιολόγησης 2.2

Χρησιμοποιώντας τα δεδομένα της Άσκησης Αυτοαξιολόγησης 1.5 να υπολογίσετε την πιθανότητα η εταιρεία να επιλεγεί στον B διαγωνισμό αν είναι γνωστό ότι δεν έχει επιλεγεί στον A διαγωνισμό.

Ο πολλαπλασιαστικός κανόνας, που δόθηκε στην Πρόταση 2.1 για δύο ενδεχόμενα ενός δειγματοχώρου, γενικεύεται στην επόμενη πρόταση σε περισσότερα από δύο ενδεχόμενα.

Πρόταση 2.2: Γενικευμένος πολλαπλασιαστικός κανόνας

Αν $A_1, A_2, \dots, A_n$ είναι μη κενά ενδεχόμενα ενός δειγματικού χώρου Ω, με $A_i \subseteq \Omega$, τέτοια ώστε $A_1 \cap A_2 \cap \cdots A_{n-1} \neq \varnothing$, $n > 2$, τότε:

$$P(A_1 \cap A_2, \cap \dots \cap A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2) \cdots P(A_n|A_1 \cap A_2 \cap \dots \cap A_{n-1}).$$

Απόδειξη Πρότασης 2.2

Ξεκινώντας από το δεξί μέλος της ισότητας, η απόδειξη προκύπτει άμεσα εφαρμόζοντας τον ορισμό της δεσμευμένης πιθανότητας.

2.2.2 Ανεξαρτησία

Σε πολλές περιπτώσεις η γνώση της πραγματοποίησης ενός ενδεχομένου, όπως είδαμε, αλλάζει την πιθανότητα πραγματοποίησης ενός άλλου ενδεχομένου. Ωστόσο, σε κάποιες περιπτώσεις η πληροφορία ότι έχει πραγματοποιηθεί το ένα ενδεχόμενο δεν αλλάζει την πιθανότητα πραγματοποίησης του άλλου. Ως ένα τέτοιο πρώτο παράδειγμα θεωρήστε το πείραμα τύχης της ρίψης ενός «τιμίου» ζαριού δύο φορές. Έστω $A = \{$ ένδειξη πρώτης ρίψης 6$\}$ και $B = \{$ένδειξη δεύτερης ρίψης 5$\}$. Τότε η πιθανότητα πραγματοποίησης του ενδεχομένου B (ίση με 1/6) δεν αλλάζει, αν έχει εμφανιστεί 6 ή όχι στην πρώτη ρίψη. Ως ένα δεύτερο παράδειγμα θεωρήστε το πείραμα τύχης της ρίψης ενός «τίμιου» ζαριού μία φορά. Στο πλαίσιο αυτό, έστω $A = \{$ένδειξη ρίψης άρτιος αριθμός$\}$ και $\mathrm{B} = \{$ένδειξη ρίψης αριθμός μεγαλύτερος ή ίσος του 3$\}$. Σε αυτήν την περίπτωση, έχουμε ότι η πιθανότητα εμφάνισης ενός άρτιου αριθμού είναι $P(A) = 3/6 = 1/2$ και το ερώτημα που τίθεται είναι αν αυτή αλλάζει από τη γνώση της πληροφορίας ότι έχει πραγματοποιηθεί το ενδεχόμενο B. Η γνώση αυτή έχει ως αποτέλεσμα να περιορίζεται ο δειγματικός χώρος και να έχουμε δυνατά αποτελέσματα τα $\{3, 4, 5, 6\}$. Επομένως, υπό αυτή τη γνώση, η πιθανότητα εμφάνισης ενός άρτιου αριθμού είναι ίση με $2/4 = 1/2$. Επομένως, η γνώση της πραγματοποίησης του ενδεχομένου B δεν αλλάζει την πιθανότητα πραγματοποίησης του ενδεχομένου A. Για τις περιπτώσεις αυτές, έχουμε τον ακόλουθο ορισμό.

Ορισμός 2.2

Δύο ενδεχόμενα A και B ενός δειγματικού χώρου Ω, με $P(B) > 0$, ονομάζονται **ανεξάρτητα**, αν και μόνο αν $P(A|B) = P(A)$.

Από τον ορισμό της δεσμευμένης πιθανότητας, τον πολλαπλασιαστικό κανόνα και τον ορισμό των ανεξάρτητων ενδεχομένων προκύπτει το ακόλουθο πόρισμα.

Πόρισμα 2.1

Δύο ενδεχόμενα A και B ενός δειγματικού χώρου Ω είναι **ανεξάρτητα**, αν και μόνο αν

$$P(A \cap B) = P(A)P(B).$$

Απόδειξη Πορίσματος 2.1

Από τον πολλαπλασιαστικό κανόνα έχουμε ότι $P(A \cap B) = P(A|B)P(B)$. Επειδή τα A και B είναι ανεξάρτητα, έχουμε ότι $P(A|B) = P(A)$ και, επομένως,

$$P(A \cap B) = P(A)P(B).$$

Αντίστροφα, αν $P(A \cap B) = P(A)P(B)$, τότε

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)P(B)}{P(B)} = P(A).$$

Παρατήρηση 2.3

- Αν $P(A|B) = P(A)$, τότε και $P(B|A) = P(B)$, για αυτό αρκεί να λέμε ότι τα A και B είναι ανεξάρτητα ενδεχόμενα και όχι ότι το A είναι ανεξάρτητο από το B ή το αντίστροφο.
- Αν τα A και B είναι ανεξάρτητα, τότε και το A με το B' είναι ανεξάρτητα, όπως ανεξάρτητα είναι και το B με το A'. Το ίδιο ισχύει και για το A' με το B'.

Άσκηση Αυτοαξιολόγησης 2.3

Έστω A και B ενδεχόμενα ενός δειγματικού χώρου Ω. Αν τα ενδεχόμενα A και B είναι ανεξάρτητα, τότε αποδείξτε ότι και τα ενδεχόμενα A' και B' είναι ανεξάρτητα με $P(B) > 0$.

Στην περίπτωση που έχουμε τρία ή περισσότερα ενδεχόμενα, τότε μπορούμε να ορίσουμε τις παρακάτω περιπτώσεις ανεξαρτησίας.

Ορισμός 2.3

Τα ενδεχόμενα $A_1, A_2, \dots, A_n$ ενός δειγματικού χώρου Ω, με $A_i \subseteq \Omega$, $i = 1, \dots, n$, ονομάζονται:

- **κατά ζεύγη ανεξάρτητα**, αν και μόνο αν

$$P(A_i \cap A_j) = P(A_i)P(A_j) \quad \forall\, i,j = 1,\dots,n, \text{ με } i \neq j$$

- **(αμοιβαία) ανεξάρτητα** , αν και μόνο αν για κάθε υποσύνολο δεικτών $\{i_1, i_2, \dots, i_k\}$ του $\{1,2,\dots,n\}$, με $k = 2,\dots,n$, ισχύει:

$$P(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k}) = P(A_{i_1}) \cdot P(A_{i_2}) \cdots P(A_{i_k})$$

Με βάση τον παραπάνω ορισμό τρία ενδεχόμενα A, B και C είναι (αμοιβαία) ανεξάρτητα, αν είναι κατά ζεύγη ανεξάρτητα και, επιπλέον, ισχύει ότι $P(A \cap B \cap C) = P(A)P(B)P(C)$. Επίσης, από τον παραπάνω ορισμό είναι φανερό ότι η έννοια των (αμοιβαία) ανεξάρτητων ενδεχομένων είναι πιο ισχυρή από την έννοια των κατά ζεύγη ανεξάρτητων ενδεχομένων. Αυτό μπορεί να δειχθεί και από το παράδειγμα που ακολουθεί.

Παράδειγμα 2.2

Κατασκευάστε έναν δειγματοχώρο Ω και προσδιορίστε τρία ενδεχόμενα αυτού, έστω A, B, C, τέτοια ώστε να ικανοποιείται η σχέση:

$$P(A \cap B \cap C) = P(A)P(B)P(C),$$

ενώ τα ενδεχόμενα A, B και C δεν είναι (αμοιβαία) ανεξάρτητα.

Λύση Παραδείγματος 2.2

Έστω ο δειγματικός χώρος $\Omega = \{1,2,3,4,5,6,7,8\}$ με όλα τα στοιχειώδη ενδεχόμενά του να είναι ισοπίθανα. Ας ορίσουμε τα ενδεχόμενα:

$$A = \{1,2,3,4\} \text{ και } B = C = \{1,5,6,7\}.$$

Από τον ορισμό των ενδεχομένων έχουμε ότι:

$$P(A \cap B \cap C) = P(\{1\}) = \frac{1}{8} = \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = P(A)P(B)P(C).$$

Ωστόσο, προφανώς A, B και C δεν είναι ανεξάρτητα, καθώς $C = B$ (παρατηρήστε ότι $1 = P(C|B) \neq P(C) = \frac{4}{8}$ ή, ισοδύναμα, ότι $P(C \cap B) = \frac{4}{8} \neq P(C) \cdot P(B) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$).
Επομένως, καταλήγουμε στο συμπέρασμα ότι, πράγματι, δεν αρκεί να ικανοποιείται η σχέση

$$P(A \cap B \cap C) = P(A)P(B)P(C),$$

για να είναι τα A, B και C (αμοιβαία) ανεξάρτητα ενδεχόμενα.
Σημειώνεται ότι μπορούμε να καταλήξουμε στο ίδιο συμπέρασμα θεωρώντας τον ίδιο δειγματοχώρο και τα ενδεχόμενα

$$A = \{1,2,3,4\}, B = \{1,3,5,7\} \text{ και } C = \{1,5,6,8\}.$$

για τα οποία δεν ισχύει $C = B$ όπως πριν. Αφήνεται ως άσκηση να το επιβεβαιώσετε.

Παράδειγμα 2.3

Ένα ηλεκτρικό σύστημα αποτελείται από δύο υποσυστήματα που λειτουργούν το ένα ανεξάρτητα από το άλλο. Το σύστημα θεωρείται ότι είναι σε κανονική λειτουργία, αν τουλάχιστον ένα από τα υποσυστήματα δεν έχει παρουσιάσει βλάβη και, επομένως, είναι σε λειτουργία. Θεωρήστε ότι η πιθανότητα βλάβης κάθε υποσυστήματος μέσα σε ένα προκαθορισμένο χρονικό διάστημα λειτουργίας είναι ίση με 0.3.

1. Να υπολογιστεί η πιθανότητα το σύστημα να είναι σε κανονική λειτουργία με το πέρας της προκαθορισμένης περιόδου, όταν τα δύο υποσυστήματα είναι σε παράλληλη σύνδεση.
2. Να απαντήσετε στο προηγούμενο ερώτημα, στην περίπτωση που τα υποσυστήματα είναι συνδεδεμένα σε σειρά.
3. Η προσθήκη επιπλέον ανεξάρτητων υποσυστημάτων σε σειρά αυξάνει ή μειώνει την αξιοπιστία του συστήματος; Τι συμβαίνει αν τα επιπλέον υποσυστήματα είναι συνδεδεμένα παράλληλα;

Λύση Παραδείγματος 2.3

Αρχικά, ορίζουμε τα ενδεχόμενα

$$B_i = \{\text{βλάβη του } i\text{-οστού υποσυστήματος}\},\ i = 1,2.$$

Τα B_i είναι ανεξάρτητα ενδεχόμενα, με $P(B_i) = 0.3$ για κάθε $i = 1,2$.

1. Καθώς τα δύο υποσυστήματα είναι συνδεδεμένα παράλληλα, το σύστημα, για να είναι σε κανονική λειτουργία με το πέρας της προκαθορισμένης περιόδου, αρκεί τουλάχιστον ένα από τα υποσυστήματα να μην έχει παρουσιάσει βλάβη. Αυτό σημαίνει ότι η ζητούμενη πιθανότητα είναι η $P(B_1' \cup B_2')$. Εφαρμόζοντας τον τύπο De Morgan (βλ. Πρόταση 1.1) έχουμε ότι η παραπάνω πιθανότητα ισούται με $P((B_1 \cap B_2)')$. Λαμβάνοντας υπόψη το γεγονός ότι τα υποσυστήματα λειτουργούν ανεξάρτητα το ένα από το άλλο και άρα παρουσιάζουν ανεξάρτητα βλάβη, έχουμε ότι:

$$P(B_1' \cup B_2') = P((B_1 \cap B_2)') = 1 - P(B_1 \cap B_2) = 1 - P(B_1)P(B_2) = 1 - 0.3^2 = 0.91.$$

2. Καθώς τα δύο υποσυστήματα είναι συνδεδεμένα σε σειρά, τότε, για να θεωρείται το σύστημα ότι είναι σε κανονική λειτουργία, θα πρέπει και τα δύο υποσυστήματα να μην έχουν παρουσιάσει βλάβη και, επομένως, να είναι σε λειτουργία. Η πιθανότητα για να συμβεί αυτό, ισούται με

$$P(B_1' \cap B_2') = P(B_1')P(B_2') = (1 - P(B_1))(1 - P(B_2)) = (1 - 0.3)^2 = 0.49,$$

όπου χρησιμοποιήθηκε το γεγονός ότι αν B_1 και B_2 είναι ανεξάρτητα ενδεχόμενα ενός δειγματοχώρου Ω, τότε και τα ενδεχόμενα B_1' και B_2' είναι ανεξάρτητα (βλ. Άσκηση Αυτοαξιολόγησης 2.3).

3. Η προσθήκη επιπλέον ανεξάρτητων υποσυστημάτων σε σειρά, τόσων ώστε συνολικά να είναι n το πλήθος υποσυστήματα, μειώνει την αξιοπιστία του συστήματος. Το παραπάνω αιτιολογείται πλήρως, καθώς η πιθανότητα να μην έχει παρουσιάσει βλάβη κανένα υποσύστημα, ισούται με $(1 - 0.3)^n = 0.7^n$, η οποία είναι φθίνουσα συνάρτηση ως προς n. Το αντίθετο συμβαίνει, αν τα συστήματα είναι συνδεδεμένα παράλληλα, καθώς η αξιοπιστία του ηλεκτρικού συστήματος σε αυτήν την περίπτωση ισούται με $1 - 0.3^n$, η οποία συγκλίνει μονότονα στο 1, καθώς το n αυξάνεται.

Άσκηση Αυτοαξιολόγησης 2.4

Μια κατασκευαστική εταιρεία αναλαμβάνει την κατασκευή τριών διαφορετικών έργων. Η πιθανότητα να ολοκληρώσει το έργο Α μέσα στα χρονικά περιθώρια της σύμβασης είναι 0.7, ενώ οι αντίστοιχες πιθανότητες για τα έργα Β και Γ είναι 0.5 και 0.9, αντίστοιχα.
Αν η κατασκευή ενός έργου είναι ανεξάρτητη της κατασκευής των άλλων, να υπολογίσετε την πιθανότητα:

- να ολοκληρωθούν και τα τρία έργα μέσα στα χρονικά περιθώρια των συμβάσεών τους,
- να ολοκληρωθεί ένα τουλάχιστον από τα έργα μέσα στα χρονικά περιθώρια της σύμβασής του,
- να ολοκληρωθεί μόνο το έργο Β μέσα στα χρονικά περιθώρια της σύμβασής του.

2.3 Θεώρημα Ολικής Πιθανότητας και Θεώρημα Bayes

Στην ενότητα αυτή παρουσιάζονται δύο θεμελιώδη θεωρήματα της θεωρίας Πιθανοτήτων. Προτού, όμως, προβούμε στη διατύπωσή τους, είναι χρήσιμο να παρουσιάσουμε την έννοια της διαμέρισης ενός συνόλου.

Ορισμός 2.4

Η συλλογή συνόλων $A_1, A_2, \dots, A_n$ ονομάζεται **διαμέριση** του Ω, αν ικανοποιούνται οι παρακάτω συνθήκες:

- $\cup_{i=1}^{n} A_i = \Omega$ και
- $A_i \cap A_j = \varnothing$ για κάθε $i, j = 1, \dots, n$ με $i \neq j$.

Το Θεώρημα Ολικής Πιθανότητας είναι το πρώτο από τα δύο θεωρήματα που παρουσιάζονται στη συνέχεια και εκφράζει την ολική πιθανότητα ενός ενδεχομένου μέσω μίας διαμέρισης του Ω.

Θεώρημα 2.1: Θεώρημα Ολικής Πιθανότητας

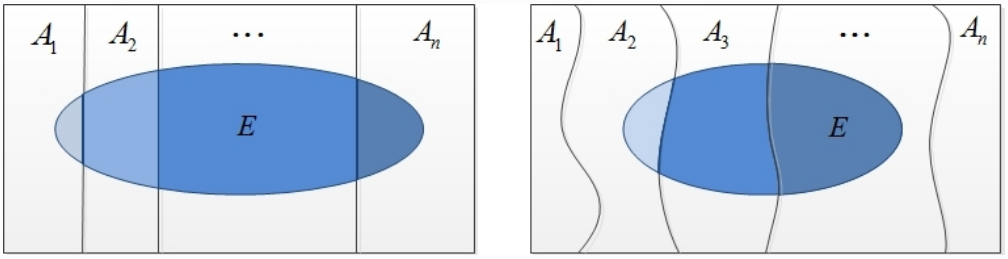
Έστω $A_1, A_2, \dots, A_n$ μια διαμέριση του δειγματικού χώρου Ω με $P(A_i) > 0$, $i = 1, \dots, n$. Τότε για οποιοδήποτε ενδεχόμενο E του δειγματικού χώρου Ω ισχύει ότι:

$$P(E) = P(E|A_1)P(A_1) + P(E|A_2)P(A_2) + \dots + P(E|A_n)P(A_n) = \sum_{i=1}^{n} P(E|A_i)P(A_i).$$

Απόδειξη Θεωρήματος 2.1

Το ενδεχόμενο E μπορεί να διατρέχει ή να μην διατρέχει όλα τα σύνολα της διαμέρισης $A_1, A_2, \dots, A_n$ (βλ. παρακάτω σχήμα). Ωστόσο, σε κάθε περίπτωση το ενδεχόμενο E μπορεί να γραφτεί ως ένωση των ξένων ανά δύο ενδεχομένων $E \cap A_i$, δηλαδή της ένωσης των τομών του E με καθένα μέλος της διαμέρισης του Ω. Αυτό έχει ως συνέπεια (ανατρέξτε στην Πρόταση 1.3) η πιθανότητα του $E = \cup_{i=1}^{n}(E \cap A_i)$ να μπορεί να εκφραστεί ως

$$P(E) = P(\cup_{i=1}^{n}(E \cap A_i)) = P(E \cap A_1) + P(E \cap A_2) + \dots + P(E \cap A_n).$$



Το αποτέλεσμα του θεωρήματος λαμβάνεται εφαρμόζοντας τον πολλαπλασιαστικό κανόνα σε καθεμία τομή $E \cap A_i$, δηλαδή γράφοντας ότι $P(E \cap A_i) = P(E|A_i)P(A_i)$, για $i = 1, 2, \dots, n$.

Στη συνέχεια, διατυπώνεται και αποδεικνύεται ένα από τα σημαντικότερα θεωρήματα της Θεωρίας Πιθανοτήτων, το Θεώρημα Bayes. Η αρχική μορφή του Θεωρήματος Bayes (ή τύπος Bayes) οφείλεται στον Βρετανό κληρικό Thomas Bayes (1701–1761), ο οποίος πρώτος κατέδειξε τη διαδικασία ενσωμάτωσης νέων στοιχείων για την ανανέωση των εκάστοτε πεποιθήσεων/πιθανοτήτων. Η σημερινή μορφή του θεωρήματος, η οποία παρουσιάζεται στη συνέχεια, οφείλεται στον Γάλλο μαθηματικό Pierre-Simon Laplace (1749-1827), ο οποίος, ανεξάρτητα από τον Thomas Bayes, παρουσίασε το αντίστοιχο αποτέλεσμα αρχικά το 1774 και στη συνέχεια το 1812 (Laplace, 1812, στο κλασικό βιβλίο του *Théorie analytique des probabilités*).

Θεώρημα 2.2: Θεώρημα Bayes

Έστω $A_1, A_2, \dots, A_n$ μια διαμέριση του Ω, με $P(A_i) > 0$, $i = 1, \dots, n$. Τότε για οποιοδήποτε ενδεχόμενο $E \subseteq \Omega$, με $P(E) > 0$ ισχύει:

$$P(A_i|E) = \frac{P(E|A_i)P(A_i)}{P(E|A_1)P(A_1) + P(E|A_2)P(A_2) + \dots + P(E|A_n)P(A_n)},$$

ή, ισοδύναμα,

$$P(A_i|E) = \frac{P(E|A_i)P(A_i)}{\sum_{i=1}^{n} P(E|A_i)P(A_i)}.$$

Απόδειξη Θεωρήματος 2.2

Με βάση τον ορισμό της δεσμευμένης πιθανότητας έχουμε ότι:

$$P(A_i|E) = \frac{P(E \cap A_i)}{P(E)}.$$

Ο αριθμητής του κλάσματος γράφεται ως $P(E \cap A_i) = P(E|A_i)P(A_i)$, ενώ από το Θεώρημα Ολικής Πιθανότητας ο παρανομαστής μπορεί να εκφραστεί ως

$$P(E|A_1)P(A_1) + P(E|A_2)P(A_2) + \dots + P(E|A_n)P(A_n).$$

Συνδυάζοντας τα παραπάνω, εύκολα προκύπτει το προς απόδειξη αποτέλεσμα.

Το Θεώρημα Bayes αναπτύχθηκε, όπως αναφέρθηκε και νωρίτερα, για να περιγραφθεί η διαδικασία ενσωμάτωσης νέων στοιχείων για την ανανέωση των εκάστοτε πεποιθήσεων/πιθανοτήτων. Αυτό γίνεται εύκολα αντιληπτό, αν αναλογιστούμε ότι μπορεί να γνωρίζουμε την πιθανότητα κάθε μέλους της διαμέρισης (εκ των προτέρων πιθανότητα) και να θέλουμε να ανανεώσουμε αυτήν την πληροφορία (εκ των υστέρων πιθανότητα) λαμβάνοντας υπόψη την πραγματοποίηση του E (νέα στοιχεία). Είναι φανερό ότι η πραγματοποίηση του E μπορεί να περιέχει πληροφορία για τα A_i, την οποία θα θέλαμε να εκμεταλλευτούμε.

Παραδείγματος χάριν, στο δεξί γράφημα που παρατέθηκε στην απόδειξη του Θεωρήματος Ολικής Πιθανότητας, η πραγματοποίηση του E αποκλείει την ταυτόχρονη πραγματοποίηση του A_1. Η περίπτωση αυτή όμως είναι μια ειδική περίπτωση, όπου ο υπολογισμός του $P(A_1|E) = 0$ είναι άμεσος και εύκολος. Για τον υπολογισμό των εκ των υστέρων πιθανοτήτων των υπόλοιπων μελών της διαμέρισης του Ω χρησιμοποιείται το Θεώρημα Bayes.

Παράδειγμα 2.4

Το 1% των δοκών που φτιάχνει μια εταιρεία για την κατασκευή κτηρίων παρουσιάζει μειωμένη δυνατότητα κάμψης. Κατά τη διαδικασία ελέγχου και τελικής επιλογής εντοπίζεται σωστά το 80% των μη κατάλληλων δοκών, ενώ κρίνεται εσφαλμένα μη κατάλληλο το 5% των κατάλληλων δοκών.

1. Να υπολογίσετε την πιθανότητα μια τυχαία επιλεγμένη δοκός, να κριθεί μη κατάλληλη κατά τη διαδικασία ελέγχου.
2. Να υπολογίσετε την πιθανότητα μια τυχαία επιλεγμένη δοκός που κρίνεται μη κατάλληλη, να μην έχει στην πραγματικότητα μειωμένη δυνατότητα κάμψης.
3. Να υπολογίσετε την πιθανότητα μια τυχαία επιλεγμένη δοκός που κρίνεται κατάλληλη, να έχει στην πραγματικότητα μειωμένη δυνατότητα κάμψης.

Λύση Παραδείγματος 2.4

Αρχικά, ορίζουμε τα ακόλουθα ενδεχόμενα:

$$K = \{\text{πραγματικά κατάλληλη δοκός, δηλαδή δοκός χωρίς μειωμένη δυνατότητα κάμψης}\}$$
$$M = \{\text{χαρακτηρισμός δοκού ως μη κατάλληλης}\}.$$

Με βάση τα δεδομένα της άσκησης έχουμε ότι:

$$P(K') = 0.01,\ \ P(M|K') = 0.8,\ \ P(M|K) = 0.05.$$

Τα ενδεχόμενα K και K' αποτελούν μια διαμέριση του δειγματοχώρου Ω και, επομένως, μπορούμε να εφαρμόσουμε το Θεώρημα Ολικής Πιθανότητας και το Θεώρημα Bayes.

1. Η πιθανότητα μια τυχαία επιλεγμένη δοκός να κριθεί μη κατάλληλη κατά τη διαδικασία ελέγχου υπολογίζεται με βάση το Θεώρημα Ολικής Πιθανότητας ως εξής:

$$\begin{aligned} P(M) &= P(M|K)P(K) + P(M|K')P(K') \\ &= 0.05 \cdot (1 - 0.01) + 0.8 \cdot 0.01 \\ &= 0.0495 + 0.008 = 0.0575. \end{aligned}$$

   Παρατηρήστε ότι παρόλο που το ποσοστό των δοκών που φτιάχνει η εταιρεία και οι οποίες παρουσιάζουν μειωμένη δυνατότητα κάμψης είναι μόλις 1%, η διαδικασία ελέγχου χαρακτηρίζει ως τέτοιες το 5.75% των δοκών.

2. Η πιθανότητα μια δοκός που κρίνεται μη κατάλληλη να μην έχει στην πραγματικότητα μειωμένη δυνατότητα κάμψης, δηλαδή η πιθανότητα $P(K|M)$, μπορεί να υπολογιστεί με τη βοήθεια του

Θεωρήματος Bayes ως εξής:

$$P(K|M) = \frac{P(M|K)P(K)}{P(M|K)P(K) + P(M|K')P(K')} = \frac{0.0495}{0.0575} = 0.8608696.$$

Παρατηρήστε ότι η συντριπτική πλειονότητα των δοκών που κρίνονται ως μη κατάλληλες δεν έχουν στην πραγματικότητα μειωμένη δυνατότητα κάμψης και είναι καλές. Αυτό οφείλεται στο μεγάλο σχετικά ποσοστό (5%) εσφαλμένης κρίσης των κατάλληλων δοκών.

3. Η πιθανότητα μια δοκός που κρίνεται κατάλληλη να έχει στην πραγματικότητα μειωμένη δυνατότητα κάμψης υπολογίζεται με τη βοήθεια του Θεωρήματος Bayes ως εξής:

$$P(K'|M') = \frac{P(M'|K')P(K')}{P(M')} = \frac{(1 - P(M|K'))P(K')}{1 - P(M)} = \frac{(1 - 0.8) \cdot 0.01}{1 - 0.05} = 0.0021.$$

Η παραπάνω τιμή υποδηλώνει πρακτικά ότι μόλις 2 στις 1000 μη κατάλληλες δοκοί διαφεύγουν από τον έλεγχο και δίνονται εν τέλει για την κατασκευή κτηρίων.

Άσκηση Αυτοαξιολόγησης 2.5

Η πιθανότητα να διαγνωστεί επιτυχώς χρησιμοποιώντας ένα κιτ μοριακής ανίχνευσης ένα άτομο που πάσχει από τον ιό της γρίπης είναι 95%. Υπολογίζεται ότι κατά την περίοδο έξαρσης της γρίπης το 15% του πληθυσμού μιας περιοχής προσβάλλεται από τον ιό. Κατά την περίοδο έξαρσης της γρίπης ένα άτομο επιλέγεται τυχαία. Να υπολογίσετε την πιθανότητα:

1. να διαγνωστεί ότι είναι φορέας του ιού της γρίπης,
2. να μην είναι φορέας της γρίπης όταν το διαγνωστικό κιτ δείχνει ότι είναι,
3. να είναι φορέας της γρίπης, αν το διαγνωστικό κιτ δείξει ότι δεν είναι.

Επιπλέον ερώτημα: Για όσους έχουν θετικό τεστ η διαδικασία ελέγχου επαναλαμβάνεται ανεξάρτητα από την προηγούμενη (με τα ίδια χαρακτηριστικά ως προς τη σωστή διάγνωση). Αν το διαγνωστικό κιτ δείξει και τη δεύτερη φορά ότι ένα άτομο είναι φορέας της γρίπης, ποια είναι η πιθανότητα αυτό το άτομο να μην είναι φορέας; Συγκρίνετε το αποτέλεσμα με την τιμή που βρήκατε στο δεύτερο ερώτημα.

2.4 Ασκήσεις

Άσκηση 2.1 *Δώστε την κατάλληλη απάντηση (ΣΩΣΤΟ ή ΛΑΘΟΣ) στις κάτωθι προτάσεις. Αιτιολογήστε σύντομα τις απαντήσεις σας.*

1. *Ο τύπος του Θεωρήματος Bayes δίνεται από τη σχέση:*

$$P(A_i|E) = \frac{P(E|A_i)P(A_i)}{\sum_{i=1}^{n} P(E|A_i)P(A_i)}$$

   όπου τα A_i είναι υποσύνολα του δειγματοχώρου Ω, τέτοια ώστε

$$\cup_{i=1}^{n} A_i = \Omega \text{ και } A_i \cap A_j = \emptyset, \text{ για κάθε } i, j = 1, \ldots, n, \text{ με } i \neq j$$

   και E είναι ένα οποιοδήποτε ενδεχόμενο του Ω.

2. *Αν για δύο ενδεχόμενα A και B, που είναι τέτοια ώστε $A \cap B \neq \emptyset$, ισχύει ότι $P(A) > P(B) > 0$, τότε $P(B|A) > P(A|B)$.*

Άσκηση 2.2 *Έστω Α και Β δύο ενδεχόμενα ενός δειγματικού χώρου Ω για τα οποία ισχύει $P(A) = 0.3$ και $P(A'|B) = 0.7$. Επιλέξτε ποια από τις παρακάτω διατυπώσεις είναι σωστή και αιτιολογήστε σύντομα την απάντησή σας.*

1. *Τα Α, Β είναι ασυμβίβαστα ενδεχόμενα.*
2. *Τα Α, Β είναι ανεξάρτητα ενδεχόμενα.*
3. *Τα Α, Β είναι συμπληρωματικά ενδεχόμενα.*
4. *Δεν ισχύει τίποτα από τα παραπάνω για τα ενδεχόμενα Α και Β.*

Άσκηση 2.3 *Ένα εξάρτημα παρουσιάζει βλάβες τύπου Α και Β, οι οποίες βλάβες εμφανίζονται ανεξάρτητα η μία από την άλλη. Η πιθανότητα να εμφανιστεί βλάβη τύπου Α είναι 12%, ενώ η πιθανότητα να εμφανιστεί βλάβη τύπου Β είναι 15%. Ποιο από τα παρακάτω ενδεχόμενα έχει τη μεγαλύτερη πιθανότητα; Δικαιολογήστε την απάντησή σας.*

1. *Το ενδεχόμενο να εμφανιστούν και οι δύο βλάβες συγχρόνως.*
2. *Το ενδεχόμενο να μην εμφανιστεί καμιά από τις δύο βλάβες.*
3. *Το ενδεχόμενο να εμφανιστεί μία τουλάχιστον από τις βλάβες.*
4. *Το ενδεχόμενο να εμφανιστεί βλάβη τύπου Β, αν είναι γνωστό ότι έχει εμφανιστεί βλάβη τύπου Α.*

Άσκηση 2.4 *Η Δημόσια Επιχείρηση Πετρελαίου (Δ.Ε.Π.) πρέπει να αποφασίσει, αν θα κάνει ή όχι γεωτρήσεις σε διάφορες περιοχές της Ηπείρου. Σύμφωνα με την κρίση των υπευθύνων της εταιρείας, η οποία βασίζεται τόσο σε προηγούμενη εμπειρία αλλά και σε αρχική ανάλυση των χαρακτηριστικών των περιοχών, μια περιοχή μπορεί να χαρακτηριστεί ως: ξερή τρύπα, περιοχή με μέτρια κοιτάσματα πετρελαίου και περιοχή με μεγάλα κοιτάσματα πετρελαίου με πιθανότητες 0.65, 0.30 και 0.05 αντίστοιχα. Οι εταιρείες εξόρυξης πετρελαίου χρησιμοποιούν ένα σεισμικό τεστ για την ανίχνευση ύπαρξης πετρελαίου. Το τεστ αυτό έχει θετικά αποτελέσματα στο 80% των περιπτώσεων ύπαρξης μεγάλου κοιτάσματος, στο 60% των περιπτώσεων ύπαρξης μετρίου κοιτάσματος και στο 20% στις περιοχές ξερής τρύπας. Αν ένα σεισμικό τεστ που κάνει η Δ.Ε.Π. έχει θετικό αποτέλεσμα, η πιθανότητα να έχει γίνει σε περιοχή που χαρακτηρίζεται ξερή τρύπα είναι:*

1. *0.35.*
2. *0.18.*
3. *0.3714.*

4. *Τίποτα από τα παραπάνω δεν ισχύει.*

Δικαιολογήστε πλήρως την απάντησή σας.

Άσκηση 2.5 *Ποια είναι η πιθανότητα το άθροισμα της ρίψης δύο ζαριών να είναι ίσο με 7, δεδομένου ότι κάθε ζάρι είχε ένδειξη μεγαλύτερη ή ίση του 3;*

Άσκηση 2.6 *Οι σημαντικότερες αιτίες κατάρρευσης μίας γέφυρας είναι η θραύση του εδάφους (A) και οι υπερβολικές καθιζήσεις (B). Υποθέτουμε ότι οι αιτίες αυτές θεωρούνται οι μόνες δυνατές αιτίες κατάρρευσης μίας γέφυρας. Δίνονται:*

$$P(A) = 0.02,\ P(B) = 0.07 \text{ και } P(B|A) = 0.60.$$

1. *Υπολογίστε την πιθανότητα να έχουμε κατάρρευση της γέφυρας λόγω ταυτόχρονης ύπαρξης και των δύο αιτιών.*
2. *Υπολογίστε την πιθανότητα να μην καταρρεύσει η γέφυρα.*
3. *Υπολογίστε την πιθανότητα να συμβεί θραύση του εδάφους όταν είναι γνωστό ότι δεν παρουσιάστηκε καθίζηση.*
4. *Εξετάστε αν τα ενδεχόμενα Α και Β είναι ανεξάρτητα.*

Άσκηση 2.7 *Μετά από ισχυρούς σεισμούς, κλιμάκια μηχανικών επιθεωρούν κτήρια ως προς την καταλληλότητά τους ταξινομώντας κάθε κτήριο ως κατοικήσιμο ή μη κατοικήσιμο. Από προηγούμενες μελέτες είναι γνωστό ότι τα κλιμάκια μηχανικών ταξινομούν κάθε κτήριο στη σωστή κατηγορία στο 95% των περιπτώσεων. Αν οι εκτιμήσεις (από το μέγεθος του σεισμού, την παλαιότητα των κτηρίων της περιοχής, τη μορφολογία του εδάφους κ.ά.) είναι ότι το 20% του συνολικού αριθμού των κτηρίων θα πρέπει να έχει υποστεί τέτοιες ζημιές ώστε να μην πρέπει να κατοικείται, να υπολογίσετε την πιθανότητα ένα τυχαία επιλεγμένο κτήριο:*

1. *να είναι κατοικήσιμο, αν έχει κριθεί από το κλιμάκιο ως μη ασφαλές,*
2. *να είναι στην πραγματικότητα μη κατοικήσιμο, αν έχει αξιολογηθεί από το κλιμάκιο ως ασφαλές.*

Άσκηση 2.8 *Το νερό σε μια κατοικημένη περιοχή αντλείται αρχικά από μια γεώτρηση, στη συνέχεια περνά από ένα σύστημα χλωρίωσης και, τέλος, μέσα από ένα φίλτρο νερού. Σύμφωνα με τον μηχανικό επίβλεψης η πιθανότητα να παρουσιάσει, κατά τη διάρκεια ενός έτους, σημαντική βλάβη το σύστημα της γεώτρησης ισούται με 0.1. Οι αντίστοιχες πιθανότητες για τα συστήματα χλωρίωσης και φιλτραρίσματος είναι 0.2 και 0.1.*

Οποιαδήποτε σημαντική βλάβη στην αντλία έχει ως επίπτωση τη μη παροχή νερού στην περιοχή, ενώ βλάβη σε τουλάχιστον ένα από τα άλλα δύο συστήματα καθιστά μη πόσιμο το νερό. Τα ενδεχόμενα παρουσίασης βλάβης στα συστήματα θεωρούνται μεταξύ τους ανεξάρτητα.

1. *Ποια είναι η πιθανότητα κατά τη διάρκεια ενός έτους να μην παρουσιαστεί καμιά βλάβη και, επομένως, η περιοχή να έχει ικανοποιητική παροχή πόσιμου νερού όλο τον χρόνο;*
2. *Κατά τη διάρκεια του έτους παρουσιάστηκε κάποιου είδους βλάβη, η οποία είχε ως συνέπεια η περιοχή να τροφοδοτείται με νερό, αλλά αυτό να μην είναι πόσιμο.*
    1. *Υπολογίστε την πιθανότητα το γεγονός αυτό να οφείλεται στην αποτυχία του συστήματος χλωρίωσης.*
    2. *Υπολογίστε την πιθανότητα το γεγονός αυτό να οφείλεται στην αποτυχία του συστήματος χλωρίωσης και του συστήματος φιλτραρίσματος.*

Άσκηση 2.9 *Η ποιότητα των θαλασσών και των παραλιών αποτελεί βασικό αντικείμενο μελέτης στην Ελλάδα εδώ και πολλά χρόνια. Ας ορίσουμε τα παρακάτω ενδεχόμενα:*

A*={η παραλία είναι μη κατάλληλη για κολύμπι}*

B*={το δείγμα νερού από την παραλία έδειξε ότι αυτή είναι μη κατάλληλη για κολύμπι}*

C*={το κολύμπι επιτρέπεται}.*

Δίνεται ότι $P(A) = 0.3$, $P(B|A) = 0.75$, $P(B|A') = 0.20$, $P(C|A \cap B) = 0.20$, $P(C|A' \cap B) = 0.15$, $P(C|A \cap B') = 0.80$ *και* $P(C|A' \cap B') = 0.90$.

1. *Υπολογίστε την πιθανότητα* $P(A \cap B \cap C)$*. Τι εκφράζει η πιθανότητα αυτή;*
2. *Υπολογίστε την πιθανότητα* $P(B \cap C)$*. Τι εκφράζει η πιθανότητα αυτή;*
3. *Υπολογίστε την πιθανότητα* $P(C)$*. Τι εκφράζει η πιθανότητα αυτή;*
4. *Υπολογίστε την πιθανότητα μια παραλία να είναι μη κατάλληλη για κολύμπι, δοθέντος ότι το κολύμπι επιτρέπεται και το δείγμα νερού από την παραλία έδειξε ότι αυτή είναι κατάλληλη για κολύμπι.*

Άσκηση 2.10 *(Κουτρουβέλης, 2011) Οι διακοπές ηλεκτρικού ρεύματος σε μια περιοχή οφείλονται στο 10% των περιπτώσεων σε βλάβη κάποιου μετασχηματιστή, στο 75% των περιπτώσεων σε κάποια βλάβη στη γραμμή μεταφοράς και στο 2% των περιπτώσεων στην ταυτόχρονη ύπαρξη και των δύο προαναφερθέντων βλαβών.*

1. *Εξετάστε αν τα ενδεχόμενα «βλάβη κάποιου μετασχηματιστή» και «όχι βλάβη στη γραμμή μεταφοράς» είναι ανεξάρτητα.*
2. *Υπολογίστε την πιθανότητα σε μια συγκεκριμένη διακοπή ρεύματος να υπάρχει:*
   (α΄) Το πολύ μία από τις προαναφερθείσες βλάβες.
   (β΄) Καμία από τις προαναφερθείσες βλάβες.

Άσκηση 2.11 *Μια βιομηχανία χρησιμοποιεί ως πρώτη ύλη βωξίτη για να παράγει δύο είδη κουφωμάτων αλουμινίου (συρόμενα, ανοιγόμενα). Τα δύο είδη κουφωμάτων παράγονται εναλλάξ στην ίδια γραμμή παραγωγής. Το Τμήμα Ποιοτικού Ελέγχου της βιομηχανίας επιλέγει κατά τακτά χρονικά διαστήματα ζευγάρια κουφωμάτων και τα εξετάζει ως προς την ποιότητά τους. Από τη μέχρι τώρα εμπειρία γνωρίζουμε ότι:*

$$P(A) = 0.01 \quad P(A' \cap B') = 0.99 \quad P(A'|B') = 0.999,$$

όπου για καθένα τυχαία επιλεγμένο ζευγάρι κουφωμάτων έχουμε συμβολίσει με A *το ενδεχόμενο το συρόμενο κούφωμα να είναι ελαττωματικό και με* B *το ενδεχόμενο το ανοιγόμενο κούφωμα να είναι ελαττωματικό.*

1. *Υπολογίστε την πιθανότητα σε ένα τυχαία επιλεγμένο ζευγάρι κουφωμάτων το ανοιγόμενο κούφωμα να μην είναι ελαττωματικό.*
2. *Υπολογίστε την πιθανότητα σε ένα τυχαία επιλεγμένο ζευγάρι κουφωμάτων τουλάχιστον ένα από τα δύο κουφώματα να είναι ελαττωματικό.*

Άσκηση 2.12 *(Montgomery and Runger, 2018) Η καρδιακή ανεπάρκεια οφείλεται είτε σε φυσικά αίτια (87%) είτε σε εξωτερικούς παράγοντες (13%). Οι εξωτερικοί παράγοντες σχετίζονται με τη χορήγηση ουσιών (73%) ή με ξένα σώματα (27%). Τα φυσικά αίτια προκαλούνται από φράξιμο αρτηριών σε ποσοστό 56%, από ασθένειες σε ποσοστό 27% και από μόλυνση, όπως για παράδειγμα ο σταφυλόκοκκος, σε ποσοστό 17%.*

1. *Υπολογίστε την πιθανότητα η καρδιακή ανεπάρκεια να οφείλεται σε χορήγηση ουσίας.*
2. *Υπολογίστε την πιθανότητα η καρδιακή ανεπάρκεια να οφείλεται σε ασθένεια ή μόλυνση.*

Άσκηση 2.13 *Υπάρχουν τρία είδη γονιδίων, τα Α, Β και Ο, από τα οποία καθορίζεται η ομάδα αίματος ενός ανθρώπου. Κάθε άνθρωπος έχει δύο από τα τρία γονίδια τα οποία κληρονομεί από τους γονείς του, ένα γονίδιο από τον καθένα, ενώ στο παιδί του δίνει ένα από τα δύο γονίδια, το καθένα με πιθανότητα* $1/2$*. Ανάλογα με το ζευγάρι των γονιδίων που έχει ένας άνθρωπος, η ομάδα αίματος είναι αυτή που φαίνεται στον ακόλουθο πίνακα (η σειρά δεν παίζει ρόλο):*

Ζεύγος γονιδίων	*AA*	*AO*	*BB*	*BO*	*AB*	*OO*
Ομάδα αίματος	*A*	*A*	*B*	*B*	*AB*	*O*

Η Δήμητρα έχει μητέρα τη Μαρία και πατέρα τον Κώστα. Στο πλαίσιο κάποιας δικαστικής διαμάχης η πατρότητα του Κώστα αμφισβητήθηκε. Από τα στοιχεία που συνέλεξε το δικαστήριο, αποφασίστηκε ότι η πιθανότητα ο Κώστας να είναι πατέρας της Δήμητρας είναι p*. Επιπλέον, το δικαστήριο αποφάσισε να γίνει και στους τρεις εξέταση αίματος. Η Δήμητρα είχε ομάδα αίματος Β, η Μαρία Ο και ο Κώστας ΑΒ. Υποθέτουμε ότι η πιθανότητα η ομάδα αίματος της Δήμητρας να είναι Β, αν ο Κώστας δεν είναι πατέρας της είναι* 9%*, ίση με την πιθανότητα εμφάνισης της ομάδας αίματος Β στη λευκή φυλή.*

1. *Αν* $p = 0.9$*, ποια είναι η εκ των υστέρων πιθανότητα (μετά δηλαδή από την εξέταση αίματος) ο Κώστας να είναι πατέρας της Δήμητρας;*
2. *Να αποδείξετε ότι οποιαδήποτε τιμή στο* $(0,1)$ *και αν είχε δώσει το δικαστήριο στην εκ των προτέρων πιθανότητα* p*, η πιθανότητα ο Κώστας να είναι ο πατέρας της Δήμητρας θα αυξανόταν με τα συγκεκριμένα αποτελέσματα της εξέτασης αίματος.*

Άσκηση 2.14 *Θέλουμε να μάθουμε το ποσοστό των αθλητών που χρησιμοποιούν αναβολικά και επειδή θεωρείται πολύ πιθανό ότι σε μία τέτοια ερώτηση δεν θα δοθεί ειλικρινής απάντηση από πολλούς αθλητές, ακολουθείται η εξής τεχνική. Σε κάθε αθλητή δίνεται ένα κουτί με 6 κόκκινα χαρτιά, 3 πράσινα και 3 κίτρινα και ο αθλητής καλείται να τραβήξει ένα κρυφά. Αν τραβήξει κόκκινο (Κ) θα απαντήσει με ειλικρίνεια στην ερώτηση «έχεις πάρει ποτέ αναβολικά;», αν τραβήξει πράσινο θα απαντήσει ΝΑΙ (ανεξάρτητα με το αν έχει κάνει ή όχι χρήση αναβολικών) και αν τραβήξει κίτρινο θα απαντήσει ΟΧΙ (ανεξάρτητα με το αν έχει κάνει ή όχι χρήση αναβολικών). Αν το* 28% *των αθλητών απάντησαν ΝΑΙ, ποιο είναι το ποσοστό των αθλητών που πραγματικά έχει χρησιμοποιήσει αναβολικά κάποια στιγμή;*

Άσκηση 2.15 *Ένας φοιτητής απαντάει σε ερωτήσεις ενός διαγωνίσματος πολλαπλής επιλογής με* n *το πλήθος δυνατές απαντήσεις σε κάθε ερώτηση, εκ των οποίων η μία μόνο είναι σωστή. Το ποσοστό των ερωτήσεων για τις οποίες ο φοιτητής γνωρίζει τη σωστή απάντηση είναι* p $(0 < p < 1)$*. Αν δεν γνωρίζει την απάντηση σε μία ερώτηση, τότε επιλέγει μια απάντηση από τις* n *στην τύχη.*

1. *Ποιο είναι το ποσοστό των σωστών απαντήσεων που δίνει ο φοιτητής;*
2. *Αν ο φοιτητής απαντήσει σωστά σε μία ερώτηση, ποια είναι η πιθανότητα να γνωρίζει πραγματικά την απάντηση;*
3. *Αν θέλουμε για έναν διαβασμένο φοιτητή, που γνωρίζει το* 80% *των σωστών απαντήσεων του διαγωνίσματος να υπάρχει πιθανότητα τουλάχιστον* 95% *να γνωρίζει πράγματι τις απαντήσεις των ερωτήσεων στις οποίες απαντάει σωστά, πόσες δυνατές απαντήσεις πρέπει να βάλουμε σε κάθε ερώτηση;*

2.5 Απαντήσεις στις ασκήσεις αυτοαξιολόγησης

Λύση Άσκησης Αυτοαξιολόγησης 2.1

Θέλουμε να υπολογίσουμε την πιθανότητα να επιλεχθεί η εταιρεία στον Β διαγωνισμό, όταν γνωρίζουμε ότι έχει επιλεχθεί στον Α. Επομένως, η ζητούμενη πιθανότητα είναι η $P(B|A)$, η οποία ισούται με:

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{0.1}{0.2} = 0.5$$

Λύση Άσκησης Αυτοαξιολόγησης 2.2

Θέλουμε να υπολογίσουμε την πιθανότητα να επιλεχθεί η εταιρεία στον Β διαγωνισμό, αν γνωρίζουμε ότι δεν έχει επιλεχθεί στον Α. Επομένως, η ζητούμενη πιθανότητα είναι η $P(B|A')$, η οποία υπολογίζεται ως εξής:

$$\begin{aligned} P(B|A') &= \frac{P(B \cap A')}{P(A')} = \frac{P(A'|B)P(B)}{1 - P(A)} \\ &= \frac{(1 - P(A|B))P(B)}{1 - P(A)} = \frac{\left(1 - \frac{P(A \cap B)}{P(B)}\right)P(B)}{1 - P(A)} \\ &= \frac{\left(1 - \frac{0.1}{0.5}\right)0.5}{1 - 0.2} \\ &= 0.5. \end{aligned}$$

Σημείωση: Όπως έχει αναφερθεί στην Ενότητα 2.2 οι ιδιότητες των πιθανοτήτων ισχύουν και για τη δεσμευμένη πιθανότητα υπό την προϋπόθεση ότι η δέσμευση παραμένει σταθερή. Επομένως, εν γένει, $P(B|A') \neq 1 - P(B|A)$, ενώ $P(B|A') = 1 - P(B'|A')$.

Λύση Άσκησης Αυτοαξιολόγησης 2.3

Για να αποδείξουμε ότι τα A' και B' είναι ανεξάρτητα, αρκεί να δείξουμε ότι

$$P(A'|B') = P(A').$$

Από τον ορισμό της δεσμευμένης πιθανότητας έχουμε ότι

$$\begin{aligned} P(A'|B') &= 1 - P(A|B') = 1 - \frac{P(A \cap B')}{P(B')} \\ &= 1 - \frac{P(B'|A)P(A)}{1 - P(B)} = 1 - \frac{(1 - P(B|A))P(A)}{1 - P(B)}. \end{aligned}$$

Όμως τα A και B είναι ανεξάρτητα και, επομένως, $P(B|A) = P(B)$ και η παραπάνω σχέση γράφεται ως

$$P(A'|B') = 1 - \frac{(1 - P(B))P(A)}{1 - P(B)} = 1 - P(A) = P(A')$$

το οποίο αποδεικνύει ότι τα A' και B' είναι πράγματι ανεξάρτητα.

Λύση Άσκησης Αυτοαξιολόγησης 2.4

Αρχικά, ορίζουμε τα ακόλουθα ενδεχόμενα:

$$A = \{\text{να ολοκληρωθεί το έργο Α}\},$$
$$B = \{\text{να ολοκληρωθεί το έργο Β}\},$$

και

$$\Gamma = \{\text{να ολοκληρωθεί το έργο Γ}\}.$$

Με βάση τα δεδομένα της άσκησης έχουμε ότι:

$$P(A) = 0.7, \quad P(B) = 0.5, \quad P(\Gamma) = 0.9$$

και ότι τα ενδεχόμενα A, B και Γ είναι ανεξάρτητα.

- Η πιθανότητα να ολοκληρωθούν και τα τρία έργα μέσα στα χρονικά περιθώρια των συμβάσεών τους, δηλαδή η πιθανότητα του ενδεχομένου $A \cap B \cap \Gamma$, λαμβάνοντας υπόψη την ανεξαρτησία των ενδεχομένων, ισούται με:

$$P(A \cap B \cap \Gamma) = P(A)P(B)P(\Gamma) = 0.7 \cdot 0.5 \cdot 0.9 = 0.315.$$

- Η πιθανότητα να ολοκληρωθεί ένα τουλάχιστον από τα έργα μέσα στα χρονικά περιθώρια της σύμβασής του, δηλαδή του ενδεχομένου $A \cup B \cup \Gamma$, μπορεί να υπολογιστεί με τη βοήθεια του συμπληρωματικού ενδεχομένου του, δηλαδή του ενδεχομένου να μην ολοκληρωθεί κανένα έργο. Είναι

$$\begin{aligned} P(A \cup B \cup \Gamma) &= 1 - P((A \cup B \cup \Gamma)') = 1 - P(A' \cap B' \cap \Gamma') \\ &= 1 - P(A')P(B')P(\Gamma') = 1 - (1 - 0.7)(1 - 0.5)(1 - 0.9) \\ &= 0.985, \end{aligned}$$

  όπου χρησιμοποιήθηκε το γεγονός ότι η ανεξαρτησία των ενδεχομένων A, B, Γ συνεπάγεται την ανεξαρτησία των ενδεχομένων A', B' και Γ'.

- Η πιθανότητα από τα τρία έργα να ολοκληρωθεί μέσα στα χρονικά περιθώρια της σύμβασής του μόνο το έργο Β ισούται με την πιθανότητα του ενδεχομένου $A' \cap B \cap \Gamma'$ και υπολογίζεται ως εξής:

$$P(A' \cap B \cap \Gamma') = P(A')P(B)P(\Gamma') = (1 - 0.7) \cdot 0.5 \cdot (1 - 0.9) = 0.015,$$

  όπου χρησιμοποιήθηκε το γεγονός ότι η ανεξαρτησία των ενδεχομένων A, B, Γ συνεπάγεται την ανεξαρτησία των ενδεχομένων A', B και Γ'.

Λύση Άσκησης Αυτοαξιολόγησης 2.5

Επιτυχής διάγνωση θεωρείται κάθε διάγνωση που είναι θετική, αν κάποιος είναι φορέας του ιού, και αρνητική, αν δεν είναι. Ορίζουμε τα ενδεχόμενα:

$$\Phi = \{\text{φορέας του ιού}\},$$
$$\Theta = \{\text{θετικό τεστ}\}.$$

και εκφράζουμε τα δεδομένα της άσκησης με τη βοήθεια των Φ και Θ. Από την εκφώνηση προκύπτει ότι:

$$P(\Theta|\Phi) = 0.95, \quad P(\Theta'|\Phi') = 0.95, \quad P(\Phi) = 0.15.$$

1. Η πιθανότητα κατά την περίοδο έξαρσης της γρίπης ένα τυχαία επιλεγμένο άτομο να διαγνωστεί ότι είναι φορέας του ιού της γρίπης ισούται με:

$$\begin{aligned} P(\Theta) &= P(\Theta|\Phi)P(\Phi) + P(\Theta|\Phi')P(\Phi') \\ &= P(\Theta|\Phi)P(\Phi) + (1 - P(\Theta'|\Phi'))P(\Phi') \\ &= 0.95 \cdot 0.15 + (1 - 0.95) \cdot (1 - 0.15) = 0.1425 + 0.0425 \\ &= 0.185. \end{aligned}$$

2. Η πιθανότητα κατά την περίοδο έξαρσης της γρίπης ένα τυχαία επιλεγμένο άτομο να μην είναι φορέας της γρίπης, αν το διαγνωστικό κιτ δείξει ότι είναι, ισούται με:

$$\begin{aligned} P(\Phi'|\Theta) &= \frac{P(\Theta|\Phi')P(\Phi')}{P(\Theta)} = \frac{(1 - P(\Theta'|\Phi'))P(\Phi')}{P(\Theta)} \\ &= \frac{(1 - 0.95) \cdot (1 - 0.15)}{0.185} = \frac{0.0425}{0.185} \\ &= 0.22973. \end{aligned}$$

3. Η πιθανότητα κατά την περίοδο έξαρσης της γρίπης ένα τυχαία επιλεγμένο άτομο να είναι φορέας της γρίπης, αν το διαγνωστικό κιτ δείξει ότι δεν είναι, ισούται με:

$$\begin{aligned} P(\Phi|\Theta') &= \frac{P(\Theta'|\Phi)P(\Phi)}{P(\Theta')} = \frac{(1 - P(\Theta|\Phi))P(\Phi)}{1 - P(\Theta)} \\ &= \frac{(1 - 0.95) \cdot 0.15}{1 - 0.185} = \frac{0.0075}{0.815} \\ &= 0.0092. \end{aligned}$$

Επιπλέον ερώτημα:
Για να υπολογίσουμε την πιθανότητα ένα τυχαία επιλεγμένο άτομο να είναι φορέας της γρίπης, αν έχει δύο ανεξάρτητες θετικές διαγνώσεις (με τα ίδια χαρακτηριστικά), πρέπει να ορίσουμε τα ενδεχόμενα

$$\Theta_1 = \{\text{Θετικό το 1ο τεστ}\} \quad \Theta_2 = \{\text{Θετικό το 2ο τεστ}\}.$$

Με βάση τα ενδεχόμενα αυτά η ζητούμενη πιθανότητα εκφράζεται ως $P(\Phi|\Theta_1 \cap \Theta_2)$ και υπολογίζεται ακολουθώντας την εξής διαδικασία:

$$\begin{aligned} P(\Phi|\Theta_1 \cap \Theta_2) &= \frac{P(\Phi \cap (\Theta_1 \cap \Theta_2)}{P(\Theta_1 \cap \Theta_2)} \\ &= \frac{P(\Theta_1 \cap \Theta_2|\Phi)P(\Phi)}{P(\Theta_1 \cap \Theta_2|\Phi)P(\Phi) + P(\Theta_1 \cap \Theta_2|\Phi')P(\Phi')} \\ &= \frac{P(\Theta_1|\Phi)P(\Theta_2|\Phi)P(\Phi)}{P(\Theta_1|\Phi)P(\Theta_2|\Phi)P(\Phi) + P(\Theta_1|\Phi')P(\Theta_2|\Phi')P(\Phi')} \\ &= \frac{P(\Theta_1|\Phi)P(\Theta_2|\Phi)P(\Phi)}{P(\Theta_1|\Phi)P(\Theta_2|\Phi)P(\Phi) + (1 - P(\Theta_1'|\Phi'))(1 - P(\Theta_2'|\Phi'))P(\Phi')} \\ &= \frac{0.95 \cdot 0.95 \cdot 0.15}{0.95 \cdot 0.95 \cdot 0.15 + (1 - 0.95) \cdot (1 - 0.95) \cdot (1 - 0.15)} \\ &= \frac{0.135375}{0.1375} \\ &= 0.984545. \end{aligned}$$

Επιπλέον, από το ερώτημα 2 της άσκησης έχουμε ότι η πιθανότητα να είναι κάποιος φορέας του ιού, αν έχει ένα θετικό τεστ ισούται με $1 - P(\Phi'|\Theta) = 1 - 0.22973 = 0.77027$. Από το επιπλέον ερώτημα της άσκησης προκύπτει ότι αν έχει δύο ανεξάρτητα θετικά τεστ, τότε η πιθανότητα να είναι φορέας αυξάνεται σημαντικά και ισούται με 0.984545. Αυτός είναι και ο λόγος που όταν ένα άτομο διαγνωστεί με μια ασθένεια ζητείται συνήθως και μια δεύτερη ανεξάρτητη διάγνωση.

ΒΙΒΛΙΟΓΡΑΦΙΑ

Ελληνόγλωσση

Montgomery, D. C. και Runger, G. C. (2018). *Εφαρμοσμένη στατιστική και πιθανότητες για μηχανικούς*. 6η αμερικανική έκδοση; Επιστημονική επιμέλεια: Πολυχρόνης Οικονόμου, Μετάφραση Ανδρέας Μπικουβαράκης. Αθήνα: Τζιόλας.

Κουτρουβέλης, Ι. Α. (2011). *Εφαρμοσμένες πιθανότητες και στατιστική*. Συμμετρία.

Ξενόγλωσση

Laplace, P. (1812). *Théorie analytique des probabilités*. Courcier.

ΚΕΦΑΛΑΙΟ 3

ΤΥΧΑΙΕΣ ΜΕΤΑΒΛΗΤΕΣ ΚΑΙ ΚΑΤΑΝΟΜΕΣ

Σύνοψη

Σε αυτό το κεφάλαιο ορίζεται, αρχικά, η έννοια της τυχαίας μεταβλητής (τ.μ.). Στη συνέχεια, αφού οι τυχαίες μεταβλητές διαχωριστούν σε διακριτές και συνεχείς, εισάγονται και μελετώνται οι έννοιες της αθροιστικής συνάρτησης κατανομής, της συνάρτησης πιθανότητας (για διακριτές τυχαίες μεταβλητές) και της συνάρτησης πυκνότητας πιθανότητας (για συνεχείς τυχαίες μεταβλητές). Τέλος, παρουσιάζονται αριθμητικά χαρακτηριστικά της τυχαίας μεταβλητής και της αντίστοιχης κατανομής της, η γνώση των οποίων μας δίνει χρήσιμες πληροφορίες για την τυχαία μεταβλητή, καθώς και μεθοδολογίες για την εύρεση της κατανομής μιας συνάρτησης της τυχαίας μεταβλητής.

Προαπαιτούμενη γνώση: Κεφάλαια 1-2 του παρόντος συγγράμματος και βασικές γνώσεις μαθηματικών.

Προσδοκώμενα μαθησιακά αποτελέσματα: Όταν θα έχετε ολοκληρώσει τη μελέτη του κεφαλαίου αυτού, θα γνωρίζετε

- την έννοια της τυχαίας μεταβλητής και τη διάκριση των τυχαίων μεταβλητών σε διακριτές και συνεχείς,
- την έννοια της αθροιστικής συνάρτησης κατανομής και των ιδιοτήτων της,
- να υπολογίζετε πιθανότητες χρησιμοποιώντας την αθροιστική συνάρτηση κατανομής,
- τη διαφορά μεταξύ συνάρτησης πιθανότητας και συνάρτησης πυκνότητας πιθανότητας και τις ιδιότητες που πρέπει να πληροί μια συνάρτηση για να είναι τέτοια,
- να υπολογίζετε χαρακτηριστικά μέτρα μιας τυχαίας μεταβλητής και της αντίστοιχης κατανομής της,
- να προσδιορίζετε την κατανομή μιας συνάρτησης μιας τυχαίας μεταβλητής και
- να υπολογίζετε πιθανότητες ενδεχομένων που συνδέονται με διακριτές και συνεχείς κατανομές.

Οικονόμου, Π., Μαλεφάκη, Σ., & Μπατσίδης, Α. (2023). *Πιθανότητες - Στατιστική.*

Γλωσσάριο επιστημονικών όρων:

- Αθροιστική συνάρτηση κατανομής
- Διακριτή τυχαία μεταβλητή
- Διακύμανση ή διασπορά
- Διάμεσος
- Κατανομή συνάρτησης τυχαίας μεταβλητής
- Μέση τιμή
- Ποσοστιαία σημεία
- Ροπές
- Ροπογεννήτρια
- Συνάρτηση πιθανότητας
- Συνάρτηση πυκνότητας πιθανότητας
- Συνεχής τυχαία μεταβλητή
- Τυπική απόκλιση
- Τυχαία μεταβλητή

3.1 Εισαγωγή

Στα προηγούμενα δύο κεφάλαια το ενδιαφέρον μας επικεντρώθηκε στον υπολογισμό της πιθανότητας εμφάνισης κάποιων ενδεχομένων ενός δειγματικού χώρου. Στο κεφάλαιο αυτό, αρχικά θα αναδείξουμε τόσο την κοινή πιθανοθεωρητική δομή που έχουν πολλά πειράματα τύχης από διαφορετικά επιστημονικά πεδία και τυχαία φαινόμενα όσο και το γεγονός ότι συχνά μας ενδιαφέρει μόνο η τιμή κάποιας ποσότητας που συνδέεται με το πείραμα τύχης και τον δειγματικό χώρο. Η αναγκαιότητα για την πιθανοθεωρητική μοντελοποίηση και μελέτη αυτών θα μας οδηγήσει στην παρουσίαση δύο θεμελιωδών εννοιών της Θεωρίας Πιθανοτήτων και της Στατιστικής, που είναι η τυχαία μεταβλητή και η κατανομή πιθανότητας. Έπειτα θα οριστούν ποσότητες που συνδέονται με την τυχαία μεταβλητή και την κατανομή πιθανότητας και έχουν στόχο τη συνοπτική περιγραφή μίας τυχαίας μεταβλητής.

3.2 Η έννοια της τυχαίας μεταβλητής

Στην ενότητα αυτή θα εισαχθεί η έννοια της τυχαίας μεταβλητής, αρχικά, μέσω παραδειγμάτων.

Παράδειγμα 3.1

Ένας γιατρός επιλέγει τυχαία από το μητρώο δημοτών μιας περιοχής έναν ενήλικα άνδρα και εξετάζει αν πάσχει από διαβήτη ή όχι. Ποιος είναι ο δειγματικός χώρος Ω_1 του τυχαίου πειράματος; Ο ίδιος γιατρός επιλέγει τυχαία από το αρχείο του έναν ασθενή που έχει παρακολουθήσει δύο ή και περισσότερα έτη και εξετάζει αν έχει τεθεί ο διαβήτης υπό έλεγχο ή όχι. Ποιος είναι ο δειγματικός χώρος Ω_2 του τυχαίου πειράματος; Από την άλλη πλευρά, ένας πολιτικός μηχανικός επιλέγει τυχαία μια δοκό και την υποβάλλει σε μια δοκιμασία αντοχής σε ένα συγκεκριμένο βάρος και εξετάζει αν αντέχει ή όχι. Ποιος είναι ο δειγματικός χώρος Ω_3 του τυχαίου πειράματος; Πώς θα μπορούσαν να αντιμετωπιστούν ενιαία τα παραπάνω τυχαία πειράματα;

Λύση Παραδείγματος 3.1

Ο δειγματικός χώρος του πρώτου τυχαίου πειράματος είναι ο $\Omega_1 = \{\text{Διαβητικός, Φυσιολογικός}\}$, δηλαδή έχει δύο δυνατά αποτελέσματα. Επιπρόσθετα, ο δειγματικός χώρος του δεύτερου τυχαίου πειράματος είναι ο $\Omega_2 = \{\text{Υπό έλεγχο, εκτός ελέγχου}\}$, δηλαδή πάλι έχει δύο δυνατά αποτελέσματα. Τέλος, ο δειγματικός χώρος του τρίτου τυχαίου πειράματος είναι ο $\Omega_3 = \{\text{Αντέχει, Δεν αντέχει}\}$, δηλαδή πάλι έχει δύο δυνατά αποτελέσματα. Ο κατάλογος τέτοιων πειραμάτων τύχης με αυτήν τη δομή είναι, προφανώς, ανεξάντλητος και όλοι μπορούμε να σκεφτούμε πολλά ακόμα τέτοια πειράματα από την καθημερινή μας ζωή. Για την ενιαία αντιμετώπισή τους σκεφτόμαστε ως ακολούθως. Αντιστοιχίζουμε στα δύο δυνατά αποτελέσματα τους αριθμούς 0 και 1. Δηλαδή έχουμε μία απεικόνιση, συνάρτηση που αντιστοιχεί κάθε σημείο του δειγματικού χώρου σε μια αριθμητική τιμή (π.χ. 1=διαβητικός και 0=φυσιολογικός ή 1= υπό έλεγχο και 0= εκτός ελέγχου ή 1= αντέχει και 0 = δεν αντέχει). Κατά αυτόν τον τρόπο, παρότι ίσως δίνεται η εντύπωση της απώλειας άμεσης επαφής με το υπό μελέτη πρόβλημα, έχουμε τη δυνατότητα πιθανοθεωρητικής μοντελοποίησης παρόμοιων τυχαίων φαινομένων και μετατροπής του αρχικού δειγματικού χώρου σε υποσύνολο του συνόλου των πραγματικών αριθμών. Η μετατροπή αυτή επιτρέπει την αξιοποίηση των μαθηματικών ιδιοτήτων που είναι διαθέσιμες για το σύνολο των πραγματικών αριθμών.

Παράδειγμα 3.2

Ένας γιατρός επιλέγει τυχαία από το μητρώο δημοτών μιας περιοχής πέντε ενήλικες άνδρες και εξετάζει αν πάσχουν από διαβήτη ή όχι. Ποιος είναι ο δειγματικός χώρος του τυχαίου πειράματος; Αν ο γιατρός ενδιαφέρεται να μελετήσει τον αριθμό των ατόμων που πάσχουν από διαβήτη στα πέντε άτομα που επέλεξε τυχαία, ποια είναι η απεικόνιση που προκύπτει;

Λύση Παραδείγματος 3.2

Ο δειγματικός χώρος του τυχαίου πειράματος (συμβολίζοντας με Δ το διαβητικό άτομο και με Φ το φυσιολογικό άτομο) είναι ο Ω ={ΔΔΔΔΔ, ΔΔΔΔΦ ,…, ΦΦΦΦΦ} και αποτελείται από $32 = 2^5$ απλά ενδεχόμενα. Ο γιατρός ενδιαφέρεται να μελετήσει το πλήθος των ατόμων που πάσχουν από διαβήτη στα πέντε άτομα που επέλεξε τυχαία. Κάθε απλό ενδεχόμενο του δειγματικού χώρου απεικονίζεται, αντιστοιχίζεται σε μία τιμή. Για παράδειγμα, τα απλά ενδεχόμενα ΔΔΔΔΦ, ΔΔΔΦΔ, ΔΔΦΔΔ, ΔΦΔΔΔ, ΦΔΔΔΔ απεικονίζονται στην τιμή 1, αφού και στα πέντε αυτά απλά ενδεχόμενα ένα άτομο πάσχει από διαβήτη και ούτω καθεξής. Επομένως, αν X είναι το πλήθος των διαβητικών ατόμων στα πέντε που επιλέχθηκαν τυχαία, έχουμε μια συνάρτηση με πεδίο ορισμού τον δειγματικό χώρο Ω και πεδίο τιμών το σύνολο $\{0,1,2,3,4,5\}$. Στο σημείο αυτό, επισημαίνουμε ότι στην ίδια δομή θα καταλήγαμε αν το πείραμα τύχης ήταν, παραδείγματος χάριν, οι πέντε προσπάθειες επίτευξης τρίποντου από ένα άτομο και μας ενδιέφερε το πλήθος των εύστοχων τρίποντων ή ο έλεγχος πέντε προϊόντων και μας ενδιέφερε το πλήθος των ελαττωματικών σε αυτά.

Παράδειγμα 3.3

Ένας γιατρός επιλέγει τυχαία από το μητρώο δημοτών μιας περιοχής ενήλικες άνδρες μέχρις ότου βρει τον πρώτο που έχει διαβήτη. Ποιος είναι ο δειγματικός χώρος του τυχαίου πειράματος; Αν ο γιατρός ενδιαφέρεται για τον αριθμό των ατόμων που θα χρειαστεί να επιλέξει μέχρι να βρει τον πρώτο διαβητικό, ποια είναι η απεικόνιση που προκύπτει;

Λύση Παραδείγματος 3.3

Ο δειγματικός χώρος του πειράματος (συμβολίζοντας με Δ το διαβητικό άτομο και με Φ το φυσιολογικό άτομο) είναι ο Ω = {Δ,ΦΔ,ΦΦΔ,ΦΦΦΔ,ΦΦΦΔ,...}. Αυτό που ενδιαφέρει τον γιατρό είναι ο αριθμός των ατόμων που θα χρειαστεί να επιλέξει μέχρις ότου να βρει τον πρώτο διαβητικό. Κάθε απλό ενδεχόμενο του δειγματικού χώρου απεικονίζεται σε μια τιμή. Για παράδειγμα, το απλό ενδεχόμενο ΦΦΦΔ απεικονίζεται στην τιμή 4 και ούτω καθεξής. Επομένως, αν X είναι ο αριθμός των ατόμων που επιλέγονται μέχρι να βρεθεί ο πρώτος διαβητικός, πάλι έχουμε μια συνάρτηση με πεδίο ορισμού τον δειγματικό χώρο Ω και πεδίο τιμών το σύνολο $\{1,2,3,...\}$. Στο σημείο αυτό, επισημαίνουμε ότι στην ίδια δομή θα καταλήγαμε, αν το πείραμα τύχης ήταν για παράδειγμα η εύρεση του πρώτου ελαττωματικού προϊόντος που κατασκευάζεται από μια εταιρεία και το ενδιαφέρον μας επικεντρωνόταν στο πλήθος των προϊόντων που ελέγχονται μέχρις ότου να βρεθεί το πρώτο ελαττωματικό ή οι προσπάθειες για τρίποντο μέχρι το πρώτο εύστοχο.

Παράδειγμα 3.4

Θεωρήστε το απλό παράδειγμα της ρίψης δύο ζαριών. Ποιος είναι ο δειγματικός χώρος του τυχαίου πειράματος; Ποια απεικόνιση προκύπτει, αν μας ενδιαφέρει το άθροισμα των ενδείξεων των δύο ζαριών;

Λύση Παραδείγματος 3.4

Ο δειγματικός χώρος $\Omega = \{(1,1),(1,2),..,(5,6),(6,6)\}$ αποτελείται από $36 = 6 \cdot 6$ το πλήθος απλά ενδεχόμενα. Αυτό που μας ενδιαφέρει, για παράδειγμα όταν παίζουμε τάβλι, είναι το άθροισμα των ενδείξεων των δύο ζαριών. Τότε κάθε απλό ενδεχόμενο του δειγματικού χώρου απεικονίζεται σε μία τιμή. Για παράδειγμα, το απλό ενδεχόμενο $(1,1)$ απεικονίζεται στην τιμή 2, το $(4,3)$ και το $(5,2)$ απεικονίζονται στην τιμή 7 κ.ο.κ. Επομένως, αν συμβολίσουμε με X το άθροισμα των δύο ενδείξεων των ζαριών μετά τη ρίψη τους, έχουμε ουσιαστικά μια συνάρτηση με πεδίο ορισμού τον δειγματικό χώρο του πειράματος τύχης και πεδίο τιμών το σύνολο $\{2,3,4,5,6,7,8,9,10,11,12\}$.

Παράδειγμα 3.5

Ένα ασθενοφόρο του ΕΚΑΒ εξυπηρετεί τις ανάγκες για επείγουσα παροχή ιατρικής βοήθειας σε ασθενείς ή τη μεταφορά τους στο νοσοκομείο σε ακτίνα 30 χιλιομέτρων από τη βάση του. Το ασθενοφόρο δέχεται κλήση για παροχή βοήθειας σε άτομο που βρίσκεται σε ένα σημείο εντός της περιοχής ευθύνης του. Ποιος είναι ο δειγματικός χώρος του τυχαίου πειράματος; Ποια απεικόνιση προκύπτει, αν το ενδιαφέρον μας επικεντρώνεται στην απόσταση του σημείου από τη βάση του ΕΚΑΒ;

Λύση Παραδείγματος 3.5

Ο δειγματικός χώρος του τυχαίου πειράματος είναι κάθε σημείο ευθύνης του ασθενοφόρου, ο οποίος δειγματικός χώρος μπορεί να παρασταθεί με γεωγραφικές συντεταγμένες ή με τον ακόλουθο τρόπο που εν συνεχεία θα περιγραφθεί. Θεωρούμε, λοιπόν, ένα σύστημα ορθογώνιων αξόνων με το σημείο (0,0) να είναι το σημείο στο οποίο βρίσκεται η βάση του ασθενοφόρου. Κατά αυτόν τον τρόπο ο δειγματικός χώρος μπορεί να γραφτεί $\Omega = \{(x,y) \in \mathbb{R}^2 : 0 \leq x^2 + y^2 \leq 30^2\}$, δηλαδή ουσιαστικά είναι ένας κυκλικός δίσκος ακτίνας 30 χιλιομέτρων. Το ενδιαφέρον μας επικεντρώνεται στην απόσταση του σημείου από το κέντρο του κυκλικού δίσκου, που ισούται με $\sqrt{x^2 + y^2}$, δηλαδή την τετραγωνική ρίζα του αθροίσματος των τετραγώνων των συντεταγμένων x, y. Τότε, κάθε απλό ενδεχόμενο (x,y) του δειγματικού χώρου απεικονίζεται στην τιμή $\sqrt{x^2 + y^2}$. Για παράδειγμα το $(0,10)$ ανατίθεται στην τιμή 10, τα $(2,3)$, $(3,2)$ στην τιμή $\sqrt{13}$ κ.ο.κ. Επομένως, αν συμβολίσουμε με D την απόσταση που πρέπει να διανύσει το ασθενοφόρο, έχουμε ουσιαστικά μια συνάρτηση με πεδίο ορισμού τον δειγματικό χώρο του πειράματος τύχης και πεδίο τιμών το διάστημα $(0,30]$.

Σε όλα τα προηγούμενα παραδείγματα αντιστοιχίστηκε ο αρχικός δειγματικός χώρος ενός τυχαίου πειράματος με ένα υποσύνολο του συνόλου των πραγματικών αριθμών, θέλοντας είτε να κωδικοποιήσουμε τα δυνατά αποτελέσματα του πειράματος τύχης είτε το τυχαίο φαινόμενο που θέλουμε να μελετήσουμε. Η αντιστοίχιση αυτή αποτελεί την ουσία της έννοιας της τυχαίας μεταβλητής. Ο χαρακτηρισμός τυχαία δικαιολογείται πλήρως καθώς μπορεί να πάρει διάφορες τιμές τυχαία ανάλογα με το αποτέλεσμα του πειράματος τύχης. Στη συνέχεια, ακολουθεί ο κλασικός ορισμός της τυχαίας μεταβλητής, καθώς και ο μετροθεωρητικός ορισμός της (για λόγους μαθηματικής ακρίβειας).

Ορισμός 3.1

Τυχαία μεταβλητή (τ.μ.), έστω X, ορίζεται να είναι μια μονοσήμαντη συνάρτηση με πεδίο ορισμού έναν δειγματικό χώρο Ω και τιμές ένα υποσύνολο των πραγματικών αριθμών, δηλαδή $X : \Omega \to S_X \subseteq \mathbb{R}$.

Ορισμός 3.2

Έστω $(\Omega, \mathscr{A}, P)$ είναι χώρος πιθανότητας, δηλαδή Ω είναι ο δειγματικός χώρος του πειράματος τύχης, $\mathscr{A}$ η σ-άλγεβρα υποσυνόλων του Ω και P ένα μέτρο πιθανότητας. Μια συνάρτηση $X : \Omega \to S_X \subseteq \mathbb{R}$ θα λέγεται **τυχαία μεταβλητή**, αν για κάθε $B \subseteq S_X$, όπου το B μπορεί να έχει τη μορφή αριθμήσιμων ενώσεων ή τομών ή συμπληρωμάτων διαστημάτων της μορφής $(-\infty, x]$, $x \in \mathbb{R}$ [α΄] ισχύει ότι:

$$\{X(\omega) \in B\} = \{\omega \in \Omega : X(\omega) \in B\} \subseteq \mathscr{A},$$

δηλαδή $\{X(\omega) \in B\}$ είναι στοιχείο της $\sigma-$άλγεβρας $\mathscr{A}$.

[α΄] Ένα τέτοιο σύνολο B λέμε ότι είναι ένα Borel σύνολο.

Είναι φανερό ότι αν ως σ-άλγεβρα θεωρήσουμε το δυναμοσύνολο του Ω, τότε κάθε πραγματική συνάρτηση του δειγματοχώρου θεωρείται τυχαία μεταβλητή. Για το εν λόγω σύγγραμμα θεωρούμε ότι όλες οι πραγματικές συναρτήσεις που ορίζονται σε έναν δειγματοχώρο είναι τυχαίες μεταβλητές.

Οι τυχαίες μεταβλητές συμβολίζονται συνήθως με κεφαλαία γράμματα, για παράδειγμα $X, Y, \ldots$, ενώ με μικρά γράμματα $x, y, \ldots$ συμβολίζονται οι τιμές των τυχαίων μεταβλητών.

Οι τυχαίες μεταβλητές ταξινομούνται σε διακριτές ή συνεχείς, ανάλογα με το σύνολο των δυνατών τιμών τους.

Ορισμός 3.3

Η τ.μ. X λέμε ότι είναι **διακριτή τυχαία μεταβλητή**, αν το σύνολο τιμών της S_X είναι πεπερασμένο ή το πολύ απείρως αριθμήσιμο.

Το S_X είναι πεπερασμένο, αν υπάρχει μια αντιστοιχία κάθε στοιχείου του με κάποιο τμήμα των φυσικών αριθμών, διαφορετικά στοιχεία του S_X αντιστοιχούν σε διαφορετικά στοιχεία του τμήματος των φυσικών αριθμών και κάθε στοιχείο του τμήματος των φυσικών αριθμών είναι εικόνα ενός στοιχείου του S_X. Από την άλλη πλευρά, το S_X είναι απείρως αριθμήσιμο εάν δεν είναι πεπερασμένο, αλλά είναι ισοπληθικό με το σύνολο των φυσικών αριθμών, δηλαδή εάν υπάρχει μια αμφιμονοσήμαντη αντιστοιχία (ένα-προς-ένα και επί) μεταξύ του S_X και του συνόλου των φυσικών αριθμών. Επομένως, τα σύνολα $\{0,1,2,3,4,5\}$ και $\{0,1\}$ είναι πεπερασμένα, ενώ τα σύνολα $\{0,1,2,3,4,\ldots\}$ και $\{1,2,3,\ldots\}$ είναι απείρως αριθμήσιμα και αυτό συνεπάγεται ότι οι τ.μ. που αντιστοιχούν σε αυτά τα σύνολα τιμών είναι διακριτές τυχαίες μεταβλητές.

Από την άλλη μεριά, ένας πρώτος, μη αυστηρός ορισμός, της συνεχούς τ.μ. είναι ο ακόλουθος: η τυχαία μεταβλητή X λέμε ότι είναι συνεχής αν το σύνολο τιμών της είναι μη αριθμήσιμο. Με άλλα λόγια, μία συνεχής τ.μ. παίρνει οποιαδήποτε τιμή σε ένα διάστημα ή σε ένωση διαστημάτων της ευθείας των πραγματικών αριθμών ή στο σύνολο των πραγματικών αριθμών.

Ο ορισμός της τ.μ. δεν εμπλέκει καθόλου την έννοια της πιθανότητας, παρότι ο μετροθεωρητικός ορισμός είναι τέτοιος ώστε να είναι εφικτός ο υπολογισμός πιθανοτήτων της μορφής $P(X \in (-\infty, x])$. Λαμβάνοντας υπόψη ότι ο απώτερος στόχος του ορισμού μιας τ.μ. είναι η διευκόλυνση στον υπολογισμό των πιθανοτήτων, εισήχθη στη βιβλιογραφία η έννοια της συνάρτησης κατανομής ή αθροιστικής συνάρτησης κατανομής. Η γνώση αυτής της συνάρτησης μας δίνει όλες τις πληροφορίες που απαιτούνται για την τ.μ. Ο ορισμός και η μελέτη των ιδιοτήτων της αθροιστικής συνάρτησης κατανομής αποτελεί αντικείμενο της επόμενης ενότητας και είναι κοινός και για τα δύο είδη μεταβλητών που προαναφέρθηκαν.

3.3 Συνάρτηση κατανομής

Η συνάρτηση κατανομής $F_X(\cdot)$ μιας τ.μ. X επιτρέπει τον υπολογισμό πιθανοτήτων της μορφής $P(X \in \mathrm{B})$ για κάθε B με $B \subseteq \mathbb{R}$, και ορίζεται ως εξής:

Ορισμός 3.4

Έστω $(\Omega, \mathscr{A}, P)$ είναι χώρος πιθανότητας και $X : \Omega \to S_X \subseteq \mathbb{R}$ μια τυχαία μεταβλητή. Η **συνάρτηση κατανομής** ή **αθροιστική συνάρτηση κατανομής** (ασκ) της τυχαίας μεταβλητής X συμβολίζεται με $F_X(\cdot)$ και είναι $F_X : R \to [0,1]$ μια πραγματική συνάρτηση που ορίζεται από τη σχέση,

$$F_X(x) = P(X \leq x) = P(\{\omega \in \Omega : X(\omega) \leq x\}), \quad x \in \mathbb{R}. \tag{3.1}$$

Παρατηρήστε ότι η αθροιστική συνάρτηση κατανομής ορίζεται για κάθε πραγματικό αριθμό ακόμη κι αν αυτός δεν ανήκει στο πεδίο τιμών της τ.μ. Επιπλέον, η $F_X(x)$ εκφράζει την πιθανότητα η τ.μ. να παίρνει μια τιμή που είναι μικρότερη ή ίση με την τιμή x.

Στο επόμενο παράδειγμα παρουσιάζεται αναλυτικά ο τρόπος εύρεσης της ασκ μιας διακριτής τ.μ..

Παράδειγμα 3.6

Θεωρήστε την τυχαία μεταβλητή X με τιμές -2, 2, 0, -1, 3 και αντίστοιχες πιθανότητες $P(X = -2) = P(X = 2) = P(X = 0) = 0.3$ και $P(X = -1) = P(X = 3) = 0.05$. Προσδιορίστε την αθροιστική συνάρτηση κατανομής. Τι παρατηρείτε;

Λύση Παραδείγματος 3.6

Από τον ορισμό της αθροιστικής συνάρτησης κατανομής προκύπτει ότι πρέπει να προσδιοριστεί η πιθανότητα $P(X \leq x)$ για κάθε $x \in \mathbb{R}$. Θα μας διευκόλυνε να διατάξουμε τις δυνατές τιμές της τ.μ. X σε αύξουσα τάξη μεγέθους. Έτσι, το σύνολο των δυνατών τιμών είναι $S_X = \{-2, -1, 0, 2, 3\}$. Είναι τότε: $P(X \leq x) = 0$ για $x < -2$, καθώς σε αυτήν την περίπτωση το $\{X \leq x\} = \varnothing$. Επίσης, για $-2 \leq x < -1$ είναι $P(X \leq x) = P(X = -2) = 0.3$, ενώ για $-1 \leq x < 0$ είναι $P(X \leq x) = P(X = -2) + P(X = -1) = 0.3 + 0.05 = 0.35$. Συνεχίζοντας με παρόμοιο τρόπο έχουμε:

$$F_X(x) = \begin{cases} 0, & x < -2, \\ 0.3, & -2 \leq x < -1, \\ 0.35, & -1 \leq x < 0, \\ 0.65, & 0 \leq x < 2, \\ 0.95, & 2 \leq x < 3, \\ 1, & 3 \leq x. \end{cases} \tag{3.2}$$

Παρατηρούμε ότι η ασκ είναι σε αυτήν την περίπτωση μια βηματική συνάρτηση.

Στη συνέχεια, δίνονται οι ιδιότητες της αθροιστικής συνάρτησης κατανομής. Η απόδειξη κάποιων από αυτές απευθύνεται μόνο σε άτομα με καλό μαθηματικό υπόβαθρο και στηρίζεται στην ακόλουθη πρόταση που συχνά αναφέρεται ως «θεώρημα συνέχειας» (continuity theorem) για αύξουσα ή φθίνουσα ακολουθία ενδεχομένων.

Πρόταση 3.1

Έστω $(A_1, A_2, \ldots)$ είναι μια ακολουθία ενδεχομένων.

1. Αν η ακολουθία ενδεχομένων είναι αύξουσα, δηλαδή $A_n \subseteq A_{n+1}$ για κάθε $n \in N_+$, τότε

$$\lim_{n\to\infty} P(A_n) = P\left(\lim_{n\to\infty} A_n\right) = P\left(\bigcup_{n=1}^{\infty} A_n\right),$$

   όπου εξ ορισμού $\lim_{n\to\infty} A_n = \bigcup_{n=1}^{\infty} A_n$.

2. Αν η ακολουθία ενδεχομένων είναι φθίνουσα, δηλαδή $A_{n+1} \subseteq A_n$ για κάθε $n \in N_+$, τότε

$$\lim_{n\to\infty} P(A_n) = P\left(\lim_{n\to\infty} A_n\right) = P\left(\bigcap_{n=1}^{\infty} A_n\right).$$

Απόδειξη Πρότασης 3.1

1. Ισχύει ότι

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = P\left(\bigcup_{i=1}^{\infty} B_i\right) = \sum_{i=1}^{\infty} P(B_i) = \lim_{n\to\infty} \sum_{i=1}^{n} P(B_i),$$

   όπου $B_1 = A_1$ και $B_i = A_i \cap A_{i-1}'$, $i = 2, 3, \ldots$ είναι ξένα ανά δύο σύνολα με την ίδια ένωση με τα $(A_1, A_2, \ldots)$.

Επομένως, καθώς $P(B_1) = P(A_1)$ και $P(B_i) = P(A_i) - P(A'_{i-1})$ είναι

$$\sum_{i=1}^{n} P(B_i) = P(A_n)$$

και η απόδειξη ολοκληρώνεται συνδυάζοντας τα παραπάνω.

2. Καθώς η ακολουθία $(A_1, A_2, \ldots)$ είναι φθίνουσα, τότε η ακολουθία των συμπληρωματικών συνόλων $(A'_1, A'_2, \ldots)$ είναι αύξουσα και ισχύει ότι:

$$P\left(\bigcap_{i=1}^{\infty} A_i\right) = 1 - P\left(\bigcup_{i=1}^{\infty} A_i^c\right) = 1 - \lim_{n\to\infty} P(A_n^c) = \lim_{n\to\infty} [1 - P(A_n^c)] = \lim_{n\to\infty} P(A_n),$$

όπου χρησιμοποιήθηκε το αποτέλεσμα του πρώτου μέρους της πρότασης.

Χρησιμοποιώντας την παραπάνω πρόταση θα αποδειχθούν κάποιες από τις ιδιότητες της αθροιστικής συνάρτησης κατανομής.

Πρόταση 3.2

Έστω $(\Omega, \mathscr{A}, P)$ είναι χώρος πιθανότητας και $X : \Omega \to S_X \subseteq \mathbb{R}$ μια τυχαία μεταβλητή με αθροιστική συνάρτηση κατανομής $F_X(\cdot)$. Τότε:

1. $0 \leq F_X(x) \leq 1$ για κάθε $x \in \mathbb{R}$.
2. Η $F_X(\cdot)$ είναι αύξουσα συνάρτηση.
3. Η $F_X(\cdot)$ είναι συνεχής από δεξιά, δηλαδή $\lim_{x\to x_0+} F_X(x) = F_X(x_0)$.
4. $\lim_{x\to-\infty} F_X(x) = 0$.
5. $\lim_{x\to+\infty} F_X(x) = 1$.
6. $\lim_{x\to x_0-} F_X(x) = P(X < x_0)$.
7. $P(X = x_0) = \lim_{x\to x_0+} F_X(x) - \lim_{x\to x_0-} F_X(x)$.

Απόδειξη Πρότασης 3.2

1. Η ιδιότητα αυτή προκύπτει άμεσα από τον ορισμό της ασκ και τις ιδιότητες της πιθανότητας.
2. Αν $x < y$, τότε προφανώς ισχύει ότι $\{X \leq x\} \subset \{X \leq y\}$ και επομένως, $F_X(x) = P(X \leq x) \leq P\left(X \leq y\right) = F_X(y)$.
3. Για σταθερό $x \in \mathbb{R}$ ας είναι $x_1, x_2, \ldots, x_n$ φθίνουσα ακολουθία, τέτοια ώστε $x_n \to x$, καθώς το $n \to +\infty$. Τότε η ακολουθία των ενδεχομένων $\{X \leq x_n\}$ είναι φθίνουσα ως προς $n \in N_+$ και τέτοια ώστε η τομή των ενδεχομένων να είναι το σύνολο $\{X \leq x\}$. Το αποτέλεσμα προκύπτει άμεσα εφαρμόζοντας το θεώρημα συνέχειας για φθίνουσα ακολουθία ενδεχομένων (βλ. Πρόταση 3.1).
4. Έστω $x_1, x_2, \ldots, x_n$ μια φθίνουσα ακολουθία, τέτοια ώστε $x_n \to -\infty$, καθώς το $n \to +\infty$. Τότε η ακολουθία των ενδεχομένων $\{X \leq x_n\}$ είναι φθίνουσα ως προς $n \in N_+$ και τέτοια ώστε η τομή των ενδεχομένων να είναι το $\varnothing$. Το αποτέλεσμα προκύπτει άμεσα εφαρμόζοντας το θεώρημα συνέχειας για φθίνουσα ακολουθία ενδεχομένων (βλ. Πρόταση 3.1).
5. Έστω $x_1, x_2, \ldots, x_n$ μια αύξουσα ακολουθία, τέτοια ώστε $x_n \to \infty$, καθώς το $n \to +\infty$. Τότε η

ακολουθία των ενδεχομένων $\{X \leq x_n\}$ είναι αύξουσα ως προς $n \in N_+$ και τέτοια ώστε η ένωση των ενδεχομένων να είναι το σύνολο $\{X \in \mathbb{R}\}$. Το αποτέλεσμα προκύπτει άμεσα εφαρμόζοντας το θεώρημα συνέχειας για αύξουσα ακολουθία ενδεχομένων (βλ. Πρόταση 3.1).

6. Για σταθερό $x \in \mathbb{R}$ έστω $x_1, x_2, \ldots, x_n$ μια αύξουσα ακολουθία, τέτοια ώστε $x_n \to x$, καθώς το $n \to +\infty$. Τότε η ακολουθία των ενδεχομένων $\{X \leq x_n\}$ είναι αύξουσα ως προς $n \in N_+$ και τέτοια ώστε η ένωση των ενδεχομένων να είναι το σύνολο $\{X \leq x\}$. Το αποτέλεσμα προκύπτει άμεσα εφαρμόζοντας το θεώρημα συνέχειας για αύξουσα ακολουθία ενδεχομένων (βλ. Πρόταση 3.1).
7. Προκύπτει με συνδυασμό των ιδιοτήτων 3 και 6.

Οι ιδιότητες 1-5 που αποδείχθηκαν στην προηγούμενη πρόταση αποτελούν επίσης τις συνθήκες τις οποίες μια πραγματική συνάρτηση πρέπει να ικανοποιεί ώστε να είναι συνάρτηση κατανομής. Από την ιδιότητα 7 προκύπτει ότι στα σημεία όπου η αθροιστική συνάρτηση κατανομής είναι συνεχής, η πιθανότητα $P(X = x_0) = 0$.

Παράδειγμα 3.7

Να αποδείξετε ότι η πραγματική συνάρτηση που ορίζεται από τη σχέση:

$$F(x) = \begin{cases} 1 - x^{-2}, & x \geq 1, \\ 0, & x < 1, \end{cases} \tag{3.3}$$

είναι αθροιστική συνάρτηση κατανομής.

Λύση Παραδείγματος 3.7

Αρκεί να αποδείξουμε ότι ισχύουν οι Ιδιότητες 1-5 της προηγούμενης πρότασης.

1. Είναι προφανές ότι $0 \leq F(x) \leq 1$ για κάθε $x \in \mathbb{R}$.
2. Αν $x < y$, με $x, y \in [1, +\infty)$, τότε εύκολα προκύπτει ότι $x^{-2} > y^{-2}$ και $1 - x^{-2} < 1 - y^{-2}$. Επομένως, αν $x < y$, με $x, y \in \mathbb{R}$, τότε $F(x) \leq F(y)$, που αποδεικνύει την Ιδιότητα 2.
3. Για να ικανοποιείται η Ιδιότητα 3, αρκεί να δείξουμε ότι είναι συνεχής από δεξιά στο σημείο 1. Όντως ισχύει ότι: $\lim_{x \to 1^+} F_X(x) = 0 = F_X(1)$.
4. $\lim_{x \to -\infty} F_X(x) = \lim_{x \to -\infty} 0 = 0$.
5. $\lim_{x \to +\infty} F_X(x) = \lim_{x \to +\infty} (1 - x^{-2}) = 1$.

Επομένως, είναι όντως αθροιστική συνάρτηση κατανομής.

Είχαμε αναφέρει στην τελευταία παράγραφο της προηγούμενης ενότητας ότι η ασκ μας βοηθά στον υπολογισμό πιθανοτήτων που σχετίζονται με την τυχαία μεταβλητή. Όντως κάτι τέτοιο επιτυγχάνεται, όπως φαίνεται και στην πρόταση που ακολουθεί[1].

Πρόταση 3.3

Έστω η τυχαία μεταβλητή X και $F_X(\cdot)$ η αθροιστική συνάρτηση κατανομής της. Τότε για $a, b \in \mathbb{R}$, με $a < b$, ισχύουν οι παρακάτω σχέσεις:

1. $P(a < X \leq b) = F_X(b) - F_X(a)$.
2. $P(a < X < b) = F_X(b-) - F_X(a)$.
3. $P(a \leq X < b) = F_X(b-) - F_X(a-)$.

[1] Αν κάποιος/α αναρωτιέται αν όλες αυτές τις σχέσεις πρέπει να τις θυμάται απέξω, η ειλικρινής απάντηση της συγγραφικής ομάδας είναι αρνητική. Εύκολα προκύπτουν όλες οι παραπάνω σχέσεις, όπως θα δείτε, από τον αξιωματικό ορισμό της πιθανότητας λαμβάνοντας υπόψη τον ορισμό της ασκ και τις ιδιότητές της.

4. $P(a \le X \le b) = F_X(b) - F_X(a-)$.
5. $P(X > a) = 1 - F_X(a)$.
6. $P(X \ge a) = 1 - F_X(a-)$.

Σε όλες τις παραπάνω σχέσεις $F_X(x_0-) = \lim_{x \to x_0-} F_X(x)$.

Απόδειξη Πρότασης 3.3

1. Είναι $P(X \le b) = P(\{a < X \le b\} \cup \{X \le a\})$ και, καθώς έχουμε ένωση ξένων συνόλων προκύπτει, χρησιμοποιώντας τον ορισμό της ασκ, ότι $F_X(b) = P(\{a < X \le b\}) + F_X(a)$, που αποδεικνύει το ζητούμενο.
2. Είναι $P(X < b) = P(\{a < X < b\} \cup \{X \le a\})$ και καθώς έχουμε ένωση ξένων συνόλων προκύπτει, χρησιμοποιώντας τον ορισμό της ασκ και τις ιδιότητές της που αποδείχθηκαν πρωτύτερα, ότι
$$\lim_{x \to b-} F_X(x) = P(\{a < X < b\}) + F_X(a),$$
που αποδεικνύει το ζητούμενο.
3. Είναι $P(X < b) = P(\{a \le X < b\} \cup \{X < a\})$ και καθώς έχουμε ένωση ξένων συνόλων προκύπτει, χρησιμοποιώντας τον ορισμό της ασκ και τις ιδιότητές της που αποδείχθηκαν πρωτύτερα, ότι
$$\lim_{x \to b-} F_X(x) = P(\{a \le X < b\}) + \lim_{x \to a-} F_X(x),$$
που αποδεικνύει το ζητούμενο.
4. Είναι $P(X \le b) = P(\{a \le X \le b\} \cup \{X < a\})$ και, καθώς έχουμε ένωση ξένων συνόλων προκύπτει, χρησιμοποιώντας τον ορισμό της ασκ και τις ιδιότητές της που αποδείχθηκαν πρωτύτερα, ότι
$$F_X(b) = P(\{a \le X \le b\}) + \lim_{x \to a-} F_X(x),$$
που αποδεικνύει το ζητούμενο.
5. Είναι $P(X > a) = 1 - P(X \le a) = 1 - F_X(a)$.
6. Είναι $P(X \ge a) = 1 - P(X < a) = 1 - F_X(a-)$.

Παρατήρηση 3.1

Αν δύο τ.μ. X και Y είναι τέτοιες ώστε οι αθροιστικές συναρτήσεις κατανομής τους να ταυτίζονται, τότε λέμε ότι είναι ισόνομες, δηλαδή έχουν την ίδια πιθανοθεωρητική συμπεριφορά.

Παράδειγμα 3.8

Χρησιμοποιώντας την ασκ της τ.μ. X του Παραδείγματος 3.6 να υπολογιστούν οι πιθανότητες: $P(-1 < X \le 1.8)$, $P(X = 2)$, $P(1 \le X \le 3)$, $P(X < 2)$ και $P(X > -1)$.

Λύση Παραδείγματος 3.8

Χρησιμοποιώντας τα αποτελέσματα της Πρότασης 3.3 έχουμε ότι

$$\begin{aligned} P(-1 < X \leq 1.8) &= F_X(1.8) - F_X(-1) = 0.65 - 0.35 = 0.3, \\ P(X = 2) &= F_X(2) - F_X(2-) = 0.95 - 0.65 = 0.3, \\ P(1 \leq X \leq 3) &= F_X(3) - F_X(1-) = 1 - 0.65 = 0.35, \\ P(X < 2) &= F_X(2-) = 0.65, \end{aligned}$$

και τέλος

$$P(X > -1) = 1 - P(X \leq -1) = 1 - F_X(-1) = 1 - 0.35 = 0.65.$$

Παράδειγμα 3.9

Χρησιμοποιώντας την ασκ της τ.μ. X του Παραδείγματος 3.7 να υπολογιστούν οι πιθανότητες: $P(2 < X \leq 3)$, $P(X = 3)$, $P(1 \leq X \leq 5)$, $P(X < 3)$ και $P(X > 2)$.

Λύση Παραδείγματος 3.9

Χρησιμοποιώντας τα αποτελέσματα της Πρότασης 3.3 έχουμε ότι

$$\begin{aligned} P(2 < X \leq 3) &= F_X(3) - F_X(2) = \frac{1}{4} - \frac{1}{9} = \frac{5}{36}, \\ P(X = 3) &= F_X(3) - F_X(3-) = 0, \\ P(1 \leq X \leq 5) &= F_X(5) - F_X(1-) = 1 - \frac{1}{25} - 0 = \frac{24}{25}, \\ P(X < 3) &= F_X(3-) = 1 - \frac{1}{9} = \frac{8}{9} \end{aligned}$$

και τέλος

$$P(X > 2) = 1 - P(X \leq 2) = 1 - F_X(2) = \frac{1}{4}.$$

Παρατηρώντας τις ασκ που προσδιορίστηκαν στα Παραδείγματα 3.6 και 3.7, μπορούμε να διακρίνουμε δύο βασικές διαφορές. Στο πρώτο παράδειγμα η ασκ είναι μια βηματική συνάρτηση, ενώ στο δεύτερο είναι μια συνεχής συνάρτηση. Επιπλέον, στο πρώτο παράδειγμα η πιθανότητα η τ.μ. να πάρει μια εκ των τιμών $\{-2, -1, 0, 2, 3\}$ είναι μη μηδενική, ενώ στο δεύτερο παράδειγμα η πιθανότητα η τ.μ. να πάρει οποιαδήποτε συγκεκριμένη τιμή είναι ίση με μηδέν και είναι μη μηδενική μόνο σε διαστήματα. Οι διαφορές αυτές οφείλονται στο γεγονός ότι η πρώτη είναι ασκ μιας διακριτής τ.μ., ενώ στη δεύτερη περίπτωση είναι ασκ μιας συνεχούς τ.μ. Για αυτούς τους λόγους, στη συνέχεια του κεφαλαίου, οι διακριτές και οι συνεχείς τ.μ. διαχωρίζονται και μελετώνται ξεχωριστά.

3.4 Διακριτή τυχαία μεταβλητή

Σύμφωνα με τον ορισμό που δόθηκε στην προηγούμενη ενότητα, μια τυχαία μεταβλητή X είναι διακριτή αν το σύνολο των τιμών της είναι πεπερασμένο ή το πολύ απείρως αριθμήσιμο. Αυτό επιτρέπει σε κάθε τιμή μιας διακριτής τυχαίας μεταβλητής να μπορεί να αντιστοιχηθεί μια πιθανότητα. Η αντιστοίχιση αυτή οδηγεί στον ορισμό της συνάρτησης πιθανότητας που ακολουθεί.

Ορισμός 3.5

Έστω X μια διακριτή τυχαία μεταβλητή με σύνολο τιμών $S_X = \{x_1, x_2, \dots, x_n, \dots\}$. Η συνάρτηση $p_X : \mathbb{R} \to [0,1]$, που ορίζεται από τη σχέση:

$$p_X(x) = \begin{cases} P(X = x), & x \in S_X, \\ 0, & x \notin S_X, \end{cases}$$

ονομάζεται **συνάρτηση πιθανότητας** (σπ) της τυχαίας μεταβλητής X.

Ας διευκρινίσουμε τον προηγούμενο ορισμό ανατρέχοντας στο Παράδειγμα 3.4.

Παράδειγμα 3.10

Θεωρήστε το πείραμα τύχης της ρίψης δύο ζαριών και την τ.μ. X που παριστάνει το άθροισμα των ενδείξεων των δύο ζαριών. Προσδιορίστε τη σπ της τ.μ. X.

Λύση Παραδείγματος 3.10

Το σύνολο των δυνατών τιμών της τ.μ. X είναι $S_X = \{2,3,4,5,6,7,8,9,10,11,12\}$. Για τον προσδιορισμό της σπ της τ.μ. X αρκεί να προσδιορίσουμε την πιθανότητα εμφάνισης κάθε δυνατής τιμής της. Είναι

$$P(X = 2) = P(\text{ενδείξεις } (1,1)) = \frac{1}{36},$$
$$P(X = 3) = P(\text{ενδείξεις } (1,2),(2,1)) = \frac{2}{36}$$

και συνεχίζοντας με παρόμοιο τρόπο

$$P(X = 12) = P(\text{ενδείξεις } (6,6)) = \frac{1}{36}.$$

Συνοψίζοντας η σπ είναι

$$p_X(x) = \begin{cases} \frac{1}{36}, & x = 2,12, \\ \frac{2}{36}, & x = 3,11, \\ \frac{3}{36}, & x = 4,10, \\ \frac{4}{36}, & x = 5,9, \\ \frac{5}{36}, & x = 6,8, \\ \frac{6}{36}, & x = 7, \\ 0, & x \text{ αλλού.} \end{cases} \tag{3.4}$$

Από τον ορισμό της σπ άμεσα προκύπτει ότι:

$$0 \le p_X(x) \le 1, \text{ για όλα τα } x \in \mathbb{R},$$

και

$$\sum_{x \in S_X} p_X(x) = 1,$$

που είναι και οι ικανές συνθήκες που πρέπει να πληροί μια συνάρτηση για να είναι συνάρτηση πιθανότητας

μιας διακριτής τ.μ. Επιπρόσθετα ισχύει ότι:

$$P(X \in A) = \sum_{x_i \in A} p_X(x_i). \tag{3.5}$$

Στην Πρόταση 3.2 είδαμε ότι

$$P(X = x_i) = \lim_{x \to x_i +} F_X(x) - \lim_{x \to x_i -} F_X(x) = F_X(x_i) - F_X(x_i^-)). \tag{3.6}$$

Επομένως, αν γνωρίζουμε την ασκ μπορούμε να προσδιορίσουμε τη σπ. Ειδικότερα, η σπ δίνεται ως το άλμα που υπάρχει στην ασκ. Ένα εύλογο ερώτημα που προκύπτει είναι αν ισχύει και το αντίστροφο. Σιωπηρά έχουμε απαντήσει στο ερώτημα αυτό στο Παράδειγμα 3.6, το οποίο είπαμε μπορεί να αποτελέσει γνώμονα για κάθε διακριτή τ.μ. Στην ενότητα αυτή θα δώσουμε μια πιο γενική προσέγγιση.

Έστω X μια διακριτή τυχαία μεταβλητή με σύνολο τιμών $S_X = \{x_1, x_2, \dots, x_n, \dots\}$, όπου $x_1 < x_2 < \dots < x_n$ και σπ $p_X(x)$. Θέλουμε να προσδιορίσουμε την ασκ, δηλαδή την

$$F_X(x) = P(X \leq x) = \sum_{x_i \leq x} p_X(x_i), \text{ για κάθε } x \in \mathbb{R}.$$

Τότε, άμεσα, εξ ορισμού προκύπτει ότι $F_X(x) = 0$ για κάθε $x < x_1$. Από εκεί και πέρα η ασκ είναι μια βηματική συνάρτηση σε αριστερά κλειστά, δεξιά ανοικτά διαστήματα με βήματα που πραγματοποιούνται στις τιμές με θετική πιθανότητα, δηλαδή στις τιμές που η σπ είναι μη μηδενική. Ειδικότερα, είναι $F_X(x) = P(X = x_1)$ για κάθε $x_1 \leq x < x_2$, $F_X(x) = P(X = x_1) + P(X = x_2)$ για κάθε $x_2 \leq x < x_3$ και συνεχίζουμε με τον ίδιο τρόπο.

Παράδειγμα 3.11

Προσδιορίστε τη σταθερά k έτσι ώστε η $p_X(x) = k \cdot (2/3)^x$, $x = 0, 1, 2, 3 \dots$, να είναι συνάρτηση πιθανότητας. Για αυτήν την τιμή της σταθεράς k προσδιορίστε την ασκ.

Λύση Παραδείγματος 3.11

Για να είναι σπ, θα πρέπει να ισχύει ότι $0 \leq k \cdot (2/3)^x \leq 1$ για $x = 0, 1, 2, 3 \dots$. Από αυτήν την ιδιότητα προκύπτει ότι $0 \leq k \leq 1$.
Επιπλέον, χρησιμοποιώντας τη σχέση (Β΄.3) του Παραρτήματος Β΄ που δίνει το άθροισμα άπειρων διαδοχικών όρων φθίνουσας γεωμετρικής προόδου, είναι:

$$\sum_{x=0}^{+\infty} k(2/3)^x = k \cdot \sum_{x=0}^{+\infty} (2/3)^x = k \cdot \frac{1}{1 - 2/3} = 3k.$$

Επομένως, πρέπει $k = 1/3$, η οποία ικανοποιεί την ανίσωση $0 \leq k \leq 1$.
Για την εύρεση της ασκ έχουμε από τον τρόπο ορισμού της ότι:

- $F_X(x) = 0$ για $x < 0$,
- $F_X(x) = P(X = 0) = \frac{1}{3}$ για $0 \leq x < 1$,
- $F_X(x) = P(X = 0) + P(X = 1) = \frac{1}{3} + \frac{1}{3}\frac{2}{3} = \frac{5}{9}$, για $1 \leq x < 2$.

Συνεχίζοντας, με παρόμοιο τρόπο, για κάθε $x \geq 0$ έχουμε:

$$F_X(x) = \sum_{y=0}^{d} \frac{1}{3}\left(\frac{2}{3}\right)^y = \frac{1}{3}\frac{1 - \left(\frac{2}{3}\right)^{d+1}}{1 - \frac{2}{3}} = 1 - \left(\frac{2}{3}\right)^{d+1},$$

όπου d είναι ο μεγαλύτερος ακέραιος που είναι μικρότερος ή ίσος του x, δηλαδή το ακέραιο μέρος του x.

Παράδειγμα 3.12

Θεωρήστε το πείραμα τύχης της ρίψης δύο ζαριών και την τ.μ. X που παριστάνει το άθροισμα των ενδείξεων των δύο ζαριών. Προσδιορίστε την ασκ της τ.μ. X χρησιμοποιώντας τη σπ της, που δόθηκε στη σχέση (3.4).

Λύση Παραδείγματος 3.12

Το σύνολο των δυνατών τιμών της τ.μ. X είναι $S_X = \{2,3,4,5,6,7,8,9,10,11,12\}$. Η ασκ είναι

$$F_X(x) = \begin{cases} 0, & x < 2, \\ 1/36, & 2 \le x < 3, \\ 3/36, & 3 \le x < 4, \\ 6/36, & 4 \le x < 5, \\ 10/36, & 5 \le x < 6, \\ 15/36, & 6 \le x < 7, \\ 21/36, & 7 \le x < 8, \\ 26/36, & 8 \le x < 9, \\ 30/36, & 9 \le x < 10, \\ 33/36, & 10 \le x < 11, \\ 35/36, & 11 \le x < 12, \\ 1, & x \ge 12. \end{cases}$$

Παράδειγμα 3.13

Έστω X η τ.μ. που παριστάνει τον αριθμό των λαθών που κάνει στη δακτυλογράφηση μια γραμματέας κατά τη διάρκεια μιας ημέρας. Η ασκ της τ.μ. X δίνεται από τη σχέση:

$$F_X(x) = \begin{cases} 0, & x < 0, \\ 0.16, & 0 \le x < 1, \\ 0.43, & 1 \le x < 2, \\ 0.82, & 2 \le x < 3, \\ 0.95, & 3 \le x < 4, \\ 1, & x \ge 4. \end{cases}$$

Προσδιορίστε τη σπ της τ.μ. X. Υπολογίστε την πιθανότητα ότι η γραμματέας κάνει από 1 έως και 3 λάθη.

Λύση Παραδείγματος 3.13

Η ασκ παρουσιάζει σημεία ασυνέχειας στα σημεία 0, 1, 2, 3 και 4. Στα σημεία αυτά λαμβάνει θετικές τιμές η σπ και χρησιμοποιώντας τη σχέση (3.6) έχουμε ότι: $P(X = x_i) = F_X(x_i) - F_X(x_i^-)$, με $x_i \in \{0,1,2,3,4\}$. Έτσι είναι:

$$p_X(x) = \begin{cases} 0.16, & x = 0, \\ 0.43 - 0.16 = 0.27 & x = 1, \\ 0.82 - 0.43 = 0.39, & x = 2, \\ 0.95 - 0.82 = 0.13, & x = 3, \\ 1 - 0.95 = 0.05, & x = 4, \\ 0, & \text{αλλού.} \end{cases}$$

Η πιθανότητα η γραμματέας να κάνει από 1 έως και 3 λάθη είναι:

$$P(1 \le X \le 3) = \sum_{x=1,2,3} p_X(x) = 0.27 + 0.39 + 0.13 = 0.79,$$

ή, εναλλακτικά, $P(1 \le X \le 3) = F_X(3) - F_X(1^-) = 0.95 - 0.16 = 0.79$.

Παράδειγμα 3.14

Τρία άτομα που ανήκουν σε τρεις διαφορετικές ευπαθείς ομάδες και επιλέγονται τυχαία από αυτές εμβολιάζονται με ένα νέο εμβόλιο. Αν είναι γνωστό ότι η πιθανότητα εμφάνισης παρενέργειας από τον εμβολιασμό για καθεμία από αυτές τις τρεις διαφορετικές ευπαθείς ομάδες είναι 0.05, 0.04 και 0.03, αντίστοιχα, προσδιορίστε τη σπ της τ.μ. που παριστάνει τον αριθμό των ατόμων που εμφάνισαν παρενέργειες στα τρία άτομα που εμβολιάστηκαν. Υπολογίστε την πιθανότητα να εμφάνισε παρενέργειες το πολύ ένα άτομο. Δίνεται ότι οι ομάδες είναι ανεξάρτητες μεταξύ τους.

Λύση Παραδείγματος 3.14

Έστω X η τ.μ. που παριστάνει τον αριθμό των ατόμων που εμφάνισαν παρενέργειες στα τρία άτομα που εμβολιάστηκαν. Οι δυνατές τιμές της X είναι 0, 1, 2 και 3. Θα προσδιορίσουμε τη σπ της τ.μ. X. Είναι $P(X=0) = 0.95 \cdot 0.96 \cdot 0.97 = 0.8846$, καθώς ισούται με την πιθανότητα να μην εμφανίσει κάποιο άτομο παρενέργειες. Η πιθανότητα $P(X=1)$ είναι ισοδύναμη με το να εμφανίσει παρενέργεια ένα από τα τρία άτομα, που συμβαίνει αν εμφανίσει το πρώτο άτομο και κανένα άλλο ή το δεύτερο άτομο και κανένα άλλο ή το τρίτο άτομο και κανένα άλλο. Επομένως,

$$P(X=1) = 0.05 \cdot 0.96 \cdot 0.97 + 0.95 \cdot 0.04 \cdot 0.97 + 0.95 \cdot 0.96 \cdot 0.03 = 0.1108.$$

Η πιθανότητα $P(X=2)$ είναι ισοδύναμη με το να εμφανίσουν παρενέργεια δύο από τα τρία άτομα, που συμβαίνει αν εμφανίσει το πρώτο και το δεύτερο άτομο και όχι το τρίτο ή το πρώτο και το τρίτο άτομο και όχι το δεύτερο ή το δεύτερο και το τρίτο άτομο και όχι το πρώτο. Επομένως,

$$P(X=2) = 0.05 \cdot 0.04 \cdot 0.97 + 0.05 \cdot 0.96 \cdot 0.03 + 0.95 \cdot 0.04 \cdot 0.03 = 0.0045.$$

Η πιθανότητα $P(X=3)$ είναι ισοδύναμη με το να εμφανίσουν παρενέργεια και τα τρία άτομα και υπολογίζεται ως εξής:

$$P(X=3) = 0.05 \cdot 0.04 \cdot 0.03 = 6 \cdot 10^{-5}.$$

Παρατηρήστε ότι $\sum_{x=0,1,2,3} p_X(x) = 1$.

Τέλος, η πιθανότητα να εμφάνισε παρενέργειες το πολύ ένα άτομο υπολογίζεται ως εξής:

$$P(X \leq 1) = P(X=0) + P(X+1) = 0.8846 + 0.1108 = 0.9954.$$

Άσκηση Αυτοαξιολόγησης 3.1

Να προσδιοριστεί η σταθερά k, έτσι ώστε η $p_X(x) = \frac{kx}{n(n+1)}$, $x = 1,2,\ldots,n$ να είναι σπ.

Άσκηση Αυτοαξιολόγησης 3.2

Να προσδιοριστεί η σταθερά k, έτσι ώστε η $p_X(x) = kx^2$, $x = 1,2,3,4$ να είναι σπ.

Άσκηση Αυτοαξιολόγησης 3.3

Μια ηλεκτρική εγκατάσταση αποτελείται από τρία υποσυστήματα σε παράλληλη σύνδεση. Κάθε υποσύστημα έχει πιθανότητα να παρουσιάσει βλάβη κατά τη διάρκεια της λειτουργίας του συστήματος ίση με 0.1, ενώ η λειτουργία κάθε υποσυστήματος είναι ανεξάρτητη των υπολοίπων. Προσδιορίστε τη σπ και την ασκ της τ.μ. που παριστάνει τον αριθμό των υποσυστημάτων που παρουσιάζουν βλάβη κατά τη διάρκεια της λειτουργίας του συστήματος.

3.5 Συνεχής τυχαία μεταβλητή

Στην καθημερινή μας ζωή και στα τυχαία πραγματικά φαινόμενα που θέλει κάποιος ερευνητής να μοντελοποιήσει πιθανοθεωρητικά δεν εμφανίζονται μόνο διακριτές τυχαίες μεταβλητές. Πλήθος παραδειγμάτων μπορούν να αναφερθούν όπου το σύνολο των δυνατών τιμών της υπό μελέτη τυχαίας μεταβλητής είναι ένα διάστημα της ευθείας των πραγματικών αριθμών είτε ένωση τέτοιων διαστημάτων είτε και όλο το σύνολο των πραγματικών αριθμών. Για παράδειγμα, ο χρόνος επιβίωσης ενός καρκινοπαθούς από τη στιγμή της διάγνωσης της κακοήθειας, η αντοχή τάσης ενός υλικού, ο χρόνος λειτουργίας μιας συσκευής, η μέτρηση της χοληστερόλης ενός ατόμου, η θέση ενός σημείου σε μια γραμμή, το βάρος αντοχής μιας δοκού είναι μερικά μόνο παραδείγματα τυχαίων μεταβλητών με σύνολο τιμών ένα υποσύνολο των πραγματικών αριθμών.

Στην Ενότητα 3.3 η αθροιστική συνάρτηση κατανομής ορίστηκε ανεξάρτητα από το αν το σύνολο τιμών της τυχαίας μεταβλητής είναι το πολύ αριθμήσιμο ή υπεραριθμήσιμο. Έτσι, η αθροιστική συνάρτηση κατανομής εξακολουθεί να υπάρχει και να ορίζεται και για συνεχείς τυχαίες μεταβλητές. Το ερώτημα που προκύπτει είναι αν έχει νόημα η συνάρτηση πιθανότητας για μια συνεχή τυχαία μεταβλητή. Δηλαδή τίθεται το ερώτημα αν έχει νόημα στις συνεχείς τ.μ. να μιλάμε για πιθανότητα αυτές να πάρουν συγκεκριμένες τιμές.

Για να απαντήσουμε σε αυτό το ερώτημα και να αναδείξουμε τη διαφοροποίηση των συνεχών τυχαίων μεταβλητών σε σχέση με τις διακριτές τυχαίες μεταβλητές θα παρουσιάσουμε ένα απλό παράδειγμα. Κάποιος επιλέγει τυχαία έναν φυσικό αριθμό από το 1-9, δηλαδή επιλέγει τυχαία έναν αριθμό από το σύνολο $\{1, 2, 3, 4, \dots, 9\}$. Αν κάθε επιλογή είναι το ίδιο πιθανή, τότε η πιθανότητα να επιλέξει κάποια συγκεκριμένη τιμή αυτού του συνόλου είναι $P(X = x) = \frac{1}{9}$, ως το πηλίκο των ευνοϊκών προς τις δυνατές περιπτώσεις. Ας κάνουμε τώρα την εξής τροποποίηση επιλέγοντας τυχαία έναν πραγματικό αριθμό από το 1-9, δηλαδή επιλέγοντας τυχαία έναν αριθμό στο διάστημα [1, 9]. Ποια είναι τώρα η πιθανότητα επιλογής ενός συγκεκριμένου αριθμού; Υποθέτοντας ότι κάθε αριθμός έχει ίδια πιθανότητα επιλογής, αν η πιθανότητα ήταν ίση με κάποιον αριθμό, τότε το άθροισμα δεν θα μπορούσε να κάνει 1, σύμφωνα με τον αξιωματικό ορισμό της πιθανότητας, και δεν θα μπορούσαμε να αθροίσουμε καθώς το σύνολο είναι υπεραριθμήσιμο. Όλα αυτά μας οδηγούν στο να συμπεράνουμε ότι στις περιπτώσεις συνεχών τυχαίων μεταβλητών (δηλαδή μεταβλητών με υπεραριθμήσιμο σύνολο τιμών) ισχύει:

$$P(X = x) = 0, \text{για κάθε } x \in \mathbb{R}.$$

Θα αναρωτηθεί εύλογα κάποιος αν δηλαδή σε μια συνεχή τ.μ. η πιθανότητα παρατήρησης μιας συγκεκριμένης τιμής είναι λογικό να είναι ίση με μηδέν. Η απάντηση είναι ότι στις συνεχείς μεταβλητές αυτό που πραγματικά θέλουμε να υπολογίσουμε είναι η πιθανότητα η τ.μ. X να πάρει τιμές σε οποιοδήποτε διάστημα $(x, x + \Delta x)$ ή $(x - \Delta x, x + \Delta x)$, με το $\Delta x > 0$ να είναι οσοδήποτε μικρό. Επομένως, η έννοια της συνάρτησης πιθανότητας, που είδαμε στην προηγούμενη ενότητα, δεν μπορεί να χρησιμοποιηθεί στις συνεχείς τυχαίες μεταβλητές.

Λαμβάνοντας υπόψη τη σχέση:

$$P(X = x_0) = \lim_{x \to x_0+} F_X(x) - \lim_{x \to x_0-} F_X(x),$$

προκύπτει ότι στις συνεχείς τ.μ.

$$\lim_{x \to x_0+} F_X(x) = \lim_{x \to x_0-} F_X(x).$$

Επομένως, σε αυτήν την περίπτωση η ασκ δεν είναι μόνο δεξιά συνεχής, αλλά είναι συνεχής. Σε πολλά συγγράμματα για αυτόν τον λόγο ορίζεται ως συνεχής τ.μ. αυτή που έχει συνεχή ασκ, σε αντίθεση με τις διακριτές που έχουν βηματική συνάρτηση. Σε κάθε περίπτωση ο ορισμός που ακολουθεί είναι ο πιο μαθηματικά αυστηρός.

Ορισμός 3.6

Έστω X μια τυχαία μεταβλητή με τιμές στο σύνολο των πραγματικών αριθμών R. Η X λέγεται **συνεχής** αν υπάρχει μια μη αρνητική ολοκληρώσιμη πραγματική συνάρτηση $f_X(\cdot)$ ορισμένη στο σύνολο των πραγματικών αριθμών R τέτοια, ώστε:

$$P(X \in B) = \int_B f_X(x)dx, \quad B \subseteq \mathbb{R}.$$

Η συνάρτηση $f_X(\cdot)$ ονομάζεται **συνάρτηση πυκνότητας πιθανότητας** (σππ) της τυχαίας μεταβλητής X.

Άμεσες συνέπειες του ορισμού είναι ότι η σππ της τ.μ. X με σύνολο δυνατών τιμών S_X ικανοποιεί τις ιδιότητες:

$$f_X(x) \geq 0, \text{ για όλα τα } x \in \mathbb{R},$$

και

$$\int_{x \in S_X} f_X(x)dx = 1.$$

Οι παραπάνω ιδιότητες είναι και οι ικανές συνθήκες που πρέπει να πληροί μια πραγματική συνάρτηση για να είναι συνάρτηση πυκνότητας πιθανότητας μιας συνεχούς τυχαίας μεταβλητής.

Ο όρος συνάρτηση πυκνότητας πιθανότητας μπορεί να δικαιολογηθεί ως εξής:

$$P(x < X \leq x + \Delta x) = \int_x^{x+\Delta x} f_X(u)du$$

και υποθέτοντας ότι είναι συνεχής προκύπτει ότι

$$f_X(x) = \lim_{\Delta x \to 0+} \frac{P(x < X \leq x + \Delta x)}{\Delta x}.$$

Από την τελευταία σχέση, κατ' ουσίαν, προκύπτει ο τρόπος προσδιορισμού της σππ της τυχαίας μεταβλητής X από την ασκ, καθώς

$$f_X(x) = \lim_{\Delta x \to 0+} \frac{P(X \leq x + \Delta x) - P(X \leq x)}{\Delta x} = \frac{d}{dx}F_X(x). \tag{3.7}$$

Επιπρόσθετα, άμεσες συνέπειες του ορισμού των συνεχών τ.μ. είναι οι ακόλουθες: $P(X = x) = 0$ (αναμενόμενη σχέση),

$$F_X(x) = P(X \leq x) = \int_{-\infty}^{x} f_X(y)dy, \tag{3.8}$$

και

$$P(a \leq X \leq b) = P(a \leq X < b) = P(a < X \leq b) = P(a < X < b) = \int_a^b f_X(x)dx.$$

Οι σχέσεις (3.7) και (3.8) δίνουν τον τρόπο προσδιορισμού της σππ από την ασκ και αντίστροφα.

Παρατήρηση 3.2

Τονίζουμε ότι στον υπολογισμό πιθανοτήτων συνεχών τ.μ. τα σύμβολα $\leq$ και $<$, καθώς και τα σύμβολα $\geq$ και $>$, μπορούν να εναλλάσσονται χωρίς καμιά διαφοροποίηση καθώς $P(X = x) = 0$. Επιπλέον, καθώς ο υπολογισμός της πιθανότητας η τ.μ. X να πάρει τιμές σε ένα διάστημα ισοδυναμεί με τον υπολογισμό του ολοκληρώματος της συνάρτησης $f_X(\cdot)$ στο διάστημα αυτό, κατ' ουσίαν αυτό σημαίνει ότι η πιθανότητα ισούται με το εμβαδόν του χωρίου που σχηματίζεται κάτω από τη σππ, τον οριζόντιο άξονα των x και τις κάθετες που άγονται στον άξονα των x στα σημεία που ορίζουν το διάστημα.

Παράδειγμα 3.15

Για ποια τιμή της σταθεράς k η συνάρτηση

$$f(x) = k \cdot (1 - x^2),\ 0 \le x \le 1,$$

είναι σππ της τ.μ. X; Στη συνέχεια, για αυτήν την τιμή της σταθεράς k, να προσδιοριστεί η ασκ και να υπολογιστούν οι πιθανότητες $P(X < 0.8)$ και $P(0.2 < X < 1.3)$.

Λύση Παραδείγματος 3.15

Για να είναι η συνάρτηση $f(x)$ σππ θα πρέπει να είναι μη αρνητική πραγματική συνάρτηση, δηλαδή θα πρέπει $k(1 - x^2) \ge 0$ για $0 \le x \le 1$. Από αυτήν την ανισότητα, εύκολα προκύπτει ότι θα πρέπει $k \ge 0$. Επιπλέον, θα πρέπει η συνάρτηση $f(x)$ να είναι τέτοια, ώστε $\int_{-\infty}^{+\infty} f(x)dx = 1$. Συνεπώς, πρέπει

$$k \int_0^1 (1 - x^2)dx = 1$$

ή, ισοδύναμα,

$$k\left(1 - \frac{1}{3}\right) = 1,$$

δηλαδή $k = \frac{3}{2}$, η οποία ικανοποιεί τη συνθήκη $k \ge 0$.
Για την εύρεση της ασκ θα χρησιμοποιήσουμε τον ορισμό $F(x) = P(X \le x)$ για $x \in \mathbb{R}$. Εξ ορισμού είναι $F(x) = 0$ για $x < 0$, ενώ είναι $F(x) = \int_0^1 f(u)du = 1$ για $x \ge 1$. Μένει να προσδιοριστεί η ασκ για $0 \le x < 1$. Είναι τότε

$$F(x) = \int_0^x \frac{3}{2}(1 - u^2)du = \frac{3}{2}\left(x - \frac{x^3}{3}\right),\ 0 \le x < 1.$$

Η $P(X < 0.8)$ μπορεί να υπολογιστεί με τη βοήθεια είτε της σπ είτε της ασκ. Ειδικότερα, είναι

$$P(X < 0.8) = \int_0^{0.8} \frac{3}{2}(1 - x^2)dx = 0.944,$$

ή, ισοδύναμα,

$$P(X < 0.8) = F(0.8) = \frac{3}{2}\left(0.8 - \frac{0.8^3}{3}\right) = 0.944.$$

Τέλος, για τον υπολογισμό της πιθανότητας $P(0.2 < X < 1.3)$ θα πρέπει να είμαστε ιδιαίτερα προσεκτικοί, καθώς το 1.3 είναι εκτός των ορίων που είναι θετική η σππ. Ακολουθώντας παρόμοιο σκεπτικό με προηγουμένως έχουμε:

$$P(0.2 < X < 1.3) = \int_{0.2}^1 \frac{3}{2}(1 - x^2)dx = 0.704,$$

ή, ισοδύναμα,

$$P(0.2 < X < 1.3) = F(1.3) - F(0.2) = 1 - \frac{3}{2}\left(0.2 - \frac{0.2^3}{3}\right) = 0.704.$$

Παράδειγμα 3.16

Για ποια τιμή της σταθεράς k η συνάρτηση

$$f(x) = \begin{cases} x, & 0 \le x < 1, \\ k - x & 1 \le x \le 2, \\ 0, & \text{αλλού}, \end{cases}$$

είναι σππ της τ.μ. X; Στη συνέχεια, για αυτήν την τιμή της σταθεράς k, να προσδιοριστεί η ασκ.

Λύση Παραδείγματος 3.16

Για να είναι η συνάρτηση $f(x)$ σππ, θα πρέπει να είναι μη αρνητική πραγματική συνάρτηση, δηλαδή θα πρέπει η σταθερά k να είναι τέτοια, ώστε $k - x \ge 0$ για $1 \le x \le 2$. Από αυτήν την ανισότητα, εύκολα προκύπτει, ότι θα πρέπει $k \ge 2$. Επιπλέον, θα πρέπει η συνάρτηση $f(x)$ να είναι τέτοια, ώστε $\int_{-\infty}^{+\infty} f(x)dx = 1$. Επομένως, πρέπει:

$$\int_0^1 x dx + \int_1^2 (k - x)dx = 1$$

ή, ισοδύναμα,

$$0.5 + (2k - 2 - (k - 0.5)) = 1,$$

δηλαδή πρέπει $k = 2$. Παρατηρήστε ότι η τιμή αυτή ικανοποιεί και τη σχέση $k - x \ge 0$ για $1 \le x \le 2$. Για την εύρεση της ασκ θα χρησιμοποιήσουμε τον ορισμό $F(x) = P(X \le x)$ για $x \in \mathbb{R}$. Εξ ορισμού είναι $F(x) = 0$, για $x < 0$, ενώ είναι $F(x) = \int_0^2 f(u)du = 1$ για $x \ge 2$. Απομένει να προσδιοριστεί η ασκ για $0 \le x < 1$ και $1 \le x < 2$. Για $0 \le x < 1$ έχουμε:

$$F(x) = \int_0^x u du = \frac{x^2}{2}, \text{ για } 0 \le x < 1,$$

ενώ για $1 \le x < 2$ έχουμε:

$$F(x) = \int_0^1 u du + \int_1^x (2 - u)du = 0.5 + \left(2x - \frac{x^2}{2} - (2 - 0.5)\right) = 2x - \frac{x^2}{2} - 1.$$

Άσκηση Αυτοαξιολόγησης 3.4

Για ποια τιμή της σταθεράς k η συνάρτηση

$$f(x) = k + \frac{3}{8}x, \ 0 \le x \le 2,$$

είναι σππ της τ.μ. X; Στη συνέχεια, για αυτήν την τιμή της σταθεράς k, να προσδιοριστεί η ασκ και να υπολογιστούν οι πιθανότητες $P(X > 1)$ και $P(0.5 < X < 1.5)$.

Άσκηση Αυτοαξιολόγησης 3.5

Έστω X τ.μ. με ασκ

$$F(x) = \begin{cases} 0, & x < 0, \\ x^4 & 0 \le x \le 1, \\ 1, & x > 1. \end{cases}$$

Να προσδιοριστεί η σππ της τ.μ. X.

Άσκηση Αυτοαξιολόγησης 3.6

Για ποια τιμή της σταθεράς k η συνάρτηση

$$f(x) = \begin{cases} k \cdot (x+3) & -3 \le x \le 0 \\ -k \cdot (x-3) & 0 \le x \le 3 \end{cases}$$

είναι σππ της τ.μ. X; Στη συνέχεια, για αυτήν την τιμή της σταθεράς k, να προσδιοριστεί η ασκ.

3.6 Χαρακτηριστικά μέτρα τυχαίων μεταβλητών

Στην ενότητα αυτή θα παρουσιαστούν διάφορα αριθμητικά χαρακτηριστικά μέτρα των τυχαίων μεταβλητών και των κατανομών τους που έχουν ως στόχο τη συνοπτική περιγραφή της συμπεριφοράς τους.

3.6.1 Μαθηματική ελπίδα ή αναμενόμενη ή μέση τιμή

Η μαθηματική ελπίδα ή αναμενόμενη ή μέση τιμή μιας τυχαίας μεταβλητής είναι μία από τις πιο σημαντικές έννοιες στις πιθανότητες και τη στατιστική. Ακολούθως, δίνεται ο ορισμός της και ακολουθεί η ερμηνεία της.

Ορισμός 3.7

Η **μαθηματική ελπίδα ή αναμενόμενη τιμή ή μέση τιμή** της τυχαίας μεταβλητής X συμβολίζεται με $E(X)$ ή με μ. Αν η τ.μ. X είναι διακριτή με σύνολο τιμών $S_X = \{x_1, \dots, x_n, \dots\}$ και συνάρτηση πιθανότητας $p_X(\cdot)$, τότε η μέση τιμή της ορίζεται από τη σχέση:

$$\mu = E(X) = \sum_{x \in S_X} x\, p_X(x), \tag{3.9}$$

με την προϋπόθεση ότι η σειρά που εμφανίζεται συγκλίνει απόλυτα, δηλαδή ότι $\sum_{x \in S_X} |x| p_X(x) < +\infty$, ενώ, αν η τ.μ. X είναι συνεχής με σύνολο τιμών S_X και σππ $f_X(\cdot)$, τότε η μέση τιμή της ορίζεται από τη σχέση:

$$\mu = E(X) = \int_{x \in S_X} x \cdot f_X(x) dx, \tag{3.10}$$

με την προϋπόθεση ότι το ολοκλήρωμα συγκλίνει απόλυτα, δηλαδή ότι $\int_{x \in S_X} |x| f_X(x) dx < +\infty$.

Από τον ορισμό της μέσης τιμής μιας διακριτής τ.μ. μπορούμε εύκολα να αντιληφθούμε τον ρόλο της, παρατηρώντας ότι για παράδειγμα στη διακριτή περίπτωση πρόκειται για ένα σταθμισμένο άθροισμα των πιθανών τιμών που μπορεί να λάβει η τ.μ. με βάρη ή συντελεστές στάθμισης τις αντίστοιχες πιθανότητες. Ειδικότερα, κάθε τιμή x_i συμμετέχει στον υπολογισμό της $E(X)$ πολλαπλασιασμένη με έναν συντελεστή βαρύτητας που είναι η πιθανότητα πραγματοποίησής της, $p_X(x_i)$. Αυτό έχει ως αποτέλεσμα τιμές της τ.μ. με μεγαλύτερη πιθανότητα να έχουν μεγαλύτερη βαρύτητα στον υπολογισμό της μέσης τιμής, η οποία ουσιαστικά εκφράζει το σημείο γύρω από το οποίο κατανέμονται οι τιμές της τ.μ. με τη μεγαλύτερη πιθανότητα πραγματοποίησης. Γεωμετρικά, θα λέγαμε ότι η μέση τιμή εκφράζει το κέντρο βάρους της κατανομής.

Παράδειγμα 3.17

Ο Αποστόλης και ο Πολυχρόνης παίζουν το εξής παιχνίδι. Ο Αποστόλης επιλέγει τυχαία από την τράπουλα ένα φύλλο και, αν είναι φιγούρα (υπάρχουν συνολικά 12 στα 52 φύλλα της τράπουλας), τότε ο Πολυχρόνης του δίνει 3 Ευρώ, αν όχι τότε ο Αποστόλης δίνει a Ευρώ στον Πολυχρόνη. Ποια πρέπει να είναι η τιμή a έτσι ώστε το παιχνίδι να είναι δίκαιο;
Υπόδειξη: Ένα παιχνίδι θεωρείται δίκαιο αν το αναμενόμενο κέρδος του κάθε παίκτη είναι μηδέν.

Λύση Παραδείγματος 3.17

Έστω X η τ.μ. που παριστάνει το κέρδος του Αποστόλη. Οι δυνατές τιμές της X είναι 3 και $-a$, με πιθανότητες 12/52 (καθώς 12 είναι οι φιγούρες της τράπουλας, άρα οι ευνοϊκές περιπτώσεις) και 40/52, αντίστοιχα. Επομένως, είναι:

$$E(X) = 3 \cdot \frac{12}{52} + (-a) \cdot \frac{40}{52} = \frac{36 - 40a}{52}.$$

Συνεπώς, για να είναι το παιχνίδι δίκαιο, δηλαδή για να είναι $E(X) = 0$, πρέπει $a = 36/40 = 0.9$ Ευρώ.

Παράδειγμα 3.18

Έστω η συνεχής τ.μ. X με σππ που δίνεται από τη σχέση:

$$f_X(x) = \frac{3x^2}{64}, 0 < x < 4.$$

Να προσδιοριστεί η μέση τιμή της τ.μ. X.

Λύση Παραδείγματος 3.18

Από τον ορισμό της μέσης τιμής για συνεχή τ.μ. είναι:

$$E(X) = \int_0^4 x \frac{3x^2}{64} dx = \frac{3}{64} \frac{4^4}{4} = 3.$$

Στη συνέχεια, δίνεται ένα παράδειγμα τ.μ. που δεν έχει μέση τιμή (βλ. Παπαϊωάννου, 1993).

Παράδειγμα 3.19

Έστω η συνεχής τ.μ. X με σππ που δίνεται από τη σχέση:

$$f_X(x) = \frac{1}{x^2}, 1 < x < +\infty.$$

Υπάρχει η μέση τιμή της τ.μ. X;

Λύση Παραδείγματος 3.19

Από τον ορισμό της μέσης τιμής για συνεχή τ.μ. είναι:

$$E(X) = \int_1^{+\infty} x \frac{1}{x^2} dx = \int_1^{+\infty} \frac{1}{x} dx = \lim_{t \to +\infty} (\log(t) - \log(1)) = +\infty.$$

Άρα δεν υπάρχει η μέση τιμή, καθώς το ολοκλήρωμα δεν συγκλίνει.

Πολλές φορές αντί να μας ενδιαφέρει η μέση τιμή της τ.μ. X μας ενδιαφέρει η μέση τιμή μιας συνάρτησής της, $g(X)$. Τότε ο ορισμός της μέσης τιμής γενικεύεται ανάλογα, με την υπόθεση της απόλυτης σύγκλισης του αθροίσματος ή του ολοκληρώματος να απαιτείται. Για λόγους πληρότητας ο ορισμός αυτός δίνεται στη συνέχεια.

Ορισμός 3.8

Η μέση τιμή μιας συνάρτησης $g(X)$ της τυχαίας μεταβλητής X, συμβολίζεται με $E(g(X))$ και, αν η τ.μ. X είναι διακριτή με σύνολο τιμών $S_X = \{x_1, \dots x_n, \dots\}$ και συνάρτηση πιθανότητας p_X, ορίζεται από τη σχέση:

$$E[g(X)] = \sum_{x \in S_X} g(x) \cdot p_X(x), \tag{3.11}$$

με την προϋπόθεση ότι η σειρά που εμφανίζεται συγκλίνει απόλυτα, ενώ, αν η τ.μ. X είναι συνεχής με σύνολο τιμών S_X και συνάρτηση πυκνότητας πιθανότητας f_X, ορίζεται από τη σχέση:

$$E[g(X)] = \int_{x \in S_X} g(x) \cdot f_X(x) dx, \tag{3.12}$$

με την προϋπόθεση ότι το ολοκλήρωμα που εμφανίζεται συγκλίνει απόλυτα.

Στην επόμενη πρόταση δίνονται βασικές ιδιότητες της μέσης τιμής. Η απόδειξή τους αφήνεται ως άσκηση καθώς προκύπτουν άμεσα από τους ορισμούς που μόλις δόθηκαν.

Πρόταση 3.4

Έστω X μια τ.μ., $g_1(\cdot),\dots,g_n(\cdot)$ πραγματικές συναρτήσεις, ενώ $a_1,\dots,a_n$ και $b_1,\dots,b_n$ πραγματικοί αριθμοί (σταθερές). Με την προϋπόθεση ότι όλες οι αναμενόμενες τιμές που εμφανίζονται υπάρχουν, ισχύουν τα ακόλουθα:

1. $E(a_i) = a_i$.
2. $E\left(a_1 g_1(X) + b_1)\right) = a_1 E\left[g_1(X)\right] + b_1$.
3. $E\left(\sum_{i=1}^{n}(a_i g_i(X) + b_i)\right) = \sum_{i=1}^{n} a_i E\left(g_i(X)\right) + \sum_{i=1}^{n} b_i$.

Παρατήρηση 3.3

Επισημαίνουμε ότι, αν η συνάρτηση $g(\cdot)$ είναι οποιαδήποτε μη γραμμική συνάρτηση, τότε γενικά $E\left(g(X)\right) \neq g(E(X))$.

Άσκηση Αυτοαξιολόγησης 3.7

Η συνάρτηση πιθανότητας της τ.μ. X δίνεται ως εξής: $p_X(-2) = p_X(2) = p_X(0) = 0.3$, και $p_X(-1) = p_X(3) = 0.05$. Να προσδιοριστούν οι $E(X)$, $E(3X-1)$ και $E(7X^2+8)$.

Άσκηση Αυτοαξιολόγησης 3.8

Η συνάρτηση πυκνότητας πιθανότητας της τ.μ. X δίνεται από τη σχέση:

$$f_X(x) = \begin{cases} x^2, & 0 < x < 1, \\ 1, & 1 < x < \frac{5}{3}. \end{cases}$$

Να προσδιοριστούν οι $E(X)$, $E(3X-1)$ και $E(7X^2+8)$.

3.6.2 Διακύμανση ή διασπορά

Η μέση τιμή της τ.μ., όπως αναφέρθηκε στην προηγούμενη υποενότητα, είναι ένα χαρακτηριστικό γνώρισμά της που μας δίνει τη θέση ή το κέντρο βάρους της κατανομής. Αυτό αιτιολογεί ότι είναι γνωστή και ως μέτρο ή παράμετρος θέσης. Ο τρόπος με τον οποίο μεταβάλλονται οι τιμές της τ.μ. γύρω από αυτό το μέτρο θέσης ποσοτικοποιείται από τη διακύμανση της τ.μ. που παρουσιάζεται σε αυτήν την υποενότητα. Προτού, όμως προχωρήσουμε στον ορισμό της διακύμανσης, ας αναδείξουμε την αναγκαιότητα της εισαγωγής της έννοιάς της μέσω του ακόλουθου (εικονικού) παραδείγματος.

Παράδειγμα 3.20

Υπάρχουν δύο τρόποι μέτρησης της πίεσης, με ένα σύγχρονο ηλεκτρονικό πιεσόμετρο (έστω A) και με ένα πιεσόμετρο χειρός (έστω B). Χρησιμοποιώντας το όργανο μέτρησης A από προηγούμενες μελέτες γνωρίζουμε ότι οι δυνατές τιμές διαφοράς της μέτρησης από την αληθινή πίεση είναι -0.5, 0 και 0.5 με ίσες πιθανότητες εμφάνισης, ενώ χρησιμοποιώντας το όργανο μέτρησης B είναι -1, -0.5, 0, 0.5 και 1 με ίσες πιθανότητες εμφάνισης. Υπολογίστε τη μέση τιμή των τ.μ. που παριστάνουν το σφάλμα μέτρησης κάθε οργάνου. Τι συμπέρασμα εξάγετε;

Λύση Παραδείγματος 3.20

Συμβολίζουμε με X_A και X_B τις τ.μ. που παριστάνουν το σφάλμα μέτρησης όταν χρησιμοποιείται το A ή το B όργανο μέτρησης, αντίστοιχα. Τότε:

$$E(X_A) = -0.5 \cdot \frac{1}{3} + 0 \cdot \frac{1}{3} + 0.5 \cdot \frac{1}{3} = 0,$$

και

$$E(X_B) = -1 \cdot \frac{1}{5} - 0.5 \cdot \frac{1}{5} + 0 \cdot \frac{1}{5} + 0.5 \cdot \frac{1}{5} + 1 \cdot \frac{1}{5} = 0.$$

Επομένως, οι μέσες τιμές ή το μέσο σφάλμα μέτρησης είναι μηδέν και στις δύο περιπτώσεις, με τις δύο τ.μ. όμως να έχουν τελείως διαφορετική συμπεριφορά. Παρατηρώντας τις δυνατές τιμές της τ.μ. X_A βλέπουμε ότι αυτές είναι πιο συγκεντρωμένες γύρω από τη μέση τιμή και αυτό θα μας έκανε να πούμε ότι το ηλεκτρονικό πιεσόμετρο (αν δεν τίθεται θέμα κόστους) είναι προτιμότερο. Παρ' όλα αυτά, αυτή είναι μόνο μια διαισθητική αντίληψη, που θα θέλαμε με τη βοήθεια των μαθηματικών και της στατιστικής να ποσοτικοποιηθεί.

Η ποσοτικοποίηση του πόσο μακριά ή κοντά βρίσκονται οι τιμές της τ.μ. γύρω από τη μέση τιμή επιτυγχάνεται μέσω της διακύμανσης (ή διασποράς), ο ορισμός της οποίας ακολουθεί.

Ορισμός 3.9

Η **διακύμανση** ή **διασπορά** της τυχαίας μεταβλητής X με πεπερασμένη μέση τιμή $\mu = E(X) < +\infty$ συμβολίζεται με $Var(X)$ ή σ^2. Αν η τ.μ. X είναι διακριτή με σύνολο τιμών $S_X = \{x_1, \dots, x_n, \dots\}$ και συνάρτηση πιθανότητας $p_X(\cdot)$, τότε η διακύμανσή της ορίζεται από τη σχέση:

$$\sigma^2 = E(X-\mu)^2 = \sum_{x \in S_X} (x-\mu)^2 \cdot p_X(x), \tag{3.13}$$

με την προϋπόθεση ότι η σειρά που εμφανίζεται συγκλίνει απόλυτα, ενώ, αν η τ.μ. X είναι συνεχής με σύνολο τιμών S_X και συνάρτηση πυκνότητας πιθανότητας $f_X(\cdot)$, τότε η διακύμανσή της ορίζεται από τη σχέση:

$$\sigma^2 = E(X-\mu)^2 = \int_{x \in S_X} (x-\mu)^2 \cdot f_X(x)dx, \tag{3.14}$$

με την προϋπόθεση ότι το ολοκλήρωμα συγκλίνει απόλυτα.

Ένας εναλλακτικός τρόπος προσδιορισμού της διακύμανσης της τυχαίας μεταβλητής X μπορεί να προκύψει χρησιμοποιώντας τις ιδιότητες της μέσης τιμής. Ειδικότερα,

$$\begin{aligned} Var(X) &= E\left((X-\mu)^2\right) = E(X^2 - 2X\mu - \mu^2) = E(X^2) - 2\mu E(X) + \mu^2 \\ &= E(X^2) - 2E(X)E(X) + E(X)^2 \end{aligned}$$

ή, ισοδύναμα,

$$Var(X) = E(X^2) - E(X)^2. \tag{3.15}$$

Παράδειγμα 3.21

Χρησιμοποιώντας τα δεδομένα του Παραδείγματος 3.20 υπολογίστε τη διακύμανση των τ.μ. X_A και X_B που ορίστηκαν εκεί. Τι συμπεράσματα εξάγετε;

Λύση Παραδείγματος 3.21

Από τον ορισμό της διακύμανσης διακριτής τ.μ. και λαμβάνοντας υπόψη τις δυνατές τιμές κάθε μεταβλητής, έχουμε ότι:

$$Var(X_A) = (-0.5-0)^2 \cdot \frac{1}{3} + (0-0)^2 \cdot \frac{1}{3} + (0.5-0)^2 \cdot \frac{1}{3} = \frac{1}{6},$$

ενώ

$$\begin{aligned} Var(X_B) &= (-1-0)^2 \cdot \frac{1}{5} + (-0.5-0)^2 \cdot \frac{1}{5} + (0-0)^2 \cdot \frac{1}{5} + (0.5-0)^2 \cdot \frac{1}{5} + (1-0)^2 \cdot \frac{1}{5} \\ &= \frac{3}{5}. \end{aligned}$$

Επομένως, οι μετρήσεις του πρώτου οργάνου μέτρησης είναι πιο συγκεντρωμένες γύρω από τη μέση τιμή 0 από ότι του δεύτερου οργάνου μέτρησης.

Από το προηγούμενο παράδειγμα αλλά και από τον ορισμό της διακύμανσης, θεωρούμε ότι αναδεικνύεται και η ερμηνεία της. Κατ΄ ουσίαν, η διακύμανση στην περίπτωση διακριτής τ.μ. αποτελεί ένα σταθμισμένο άθροισμα των μέσων τετραγωνικών αποκλίσεων, πιο συγκεκριμένα των αποστάσεων των τιμών της τ.μ. X από τη μέση τιμή της, με τις πιθανότητες να αποτελούν τους αντίστοιχους συντελεστές στάθμισης. Ένα εύλογο ερώτημα είναι αν θα μπορούσε να χρησιμοποιηθεί αντί για το τετράγωνο της απόστασης κάποια άλλης μορφής απόσταση, όπως για παράδειγμα η απόλυτη τιμή. Η απάντηση είναι προφανώς θετική. Ο κύριος λόγος που επιλέχθηκε το τετράγωνο των αποστάσεων αντί της απόλυτης τιμής είναι ότι το τετράγωνο των αποστάσεων είναι πιο εύχρηστο ως συνάρτηση στη στατιστική από την απόλυτη τιμή.

Παρατήρηση 3.4

Από τον ορισμό της διακύμανσης προκύπτει άμεσα ότι αυτή είναι πάντοτε μη αρνητική και ότι έχει μονάδα μέτρησης το τετράγωνο της μονάδας μέτρησης της τ.μ. Θέλοντας να οδηγηθούμε σε ένα μέτρο της μεταβλητότητας που θα έχει ίδιες μονάδες μέτρησης με την τ.μ. εισήχθη στη βιβλιογραφία, υπό την υπόθεση ότι η διακύμανση είναι πεπερασμένη, η τυπική απόκλιση της τυχαίας μεταβλητής. Η **τυπική απόκλιση** συμβολίζεται με σ και ορίζεται ως $\sigma = \sqrt{Var(X)}$.

Ακολούθως δίνουμε τον ορισμό της διακύμανσης μιας συνάρτησης $g(X)$ της τ.μ. X.

Ορισμός 3.10

Έστω $g(X)$ μια συνάρτηση της τυχαίας μεταβλητής X με πεπερασμένη μέση τιμή. Η διακύμανση της $g(X)$ συμβολίζεται με $Var(g(X))$ και, αν η τ.μ. X είναι διακριτή με σύνολο τιμών $S_X = \{x_1, \dots, x_n, \dots\}$ και συνάρτηση πιθανότητας $p_X(\cdot)$, ορίζεται από τη σχέση:

$$\begin{aligned} Var(g(X)) &= \sum_{x \in S_X} (g(x) - E[g(X)])^2 p_X(x) \\ &= E\left(g(X)^2\right) - (E[g(X)])^2, \end{aligned} \tag{3.16}$$

ενώ, αν η τ.μ. X είναι συνεχής με σύνολο τιμών S_X και συνάρτηση πυκνότητας πιθανότητας $f_X(\cdot)$, ορίζεται από τη σχέση:

$$\begin{aligned} Var(g(X)) &= \int_{x \in S_X} (g(x) - E[g(X)])^2 f_X(x) dx \\ &= E\left(g(X)^2\right) - (E[g(X)])^2, \end{aligned} \tag{3.17}$$

με την προϋπόθεση ότι η $E(g(X)^2)$ είναι πεπερασμένη.

Στην επόμενη πρόταση δίνονται βασικές ιδιότητες της διακύμανσης. Η απόδειξή τους αφήνεται ως άσκηση για τον/την αναγνώστη/στρια, καθώς προκύπτουν άμεσα από τους ορισμούς που μόλις δόθηκαν.

Πρόταση 3.5

Έστω X μια τ.μ., $g(\cdot)$ πραγματική συνάρτηση, ενώ a και b πραγματικοί αριθμοί (σταθερές). Με την προϋπόθεση ότι όλες οι διακυμάνσεις που εμφανίζονται υπάρχουν, ισχύουν τα ακόλουθα:

1. $Var(a) = 0$.
2. $Var(ag(X) + b) = a^2 Var\left(g(X)\right)$.

Παράδειγμα 3.22

Να βρεθεί η διακύμανση του κέρδους του Αποστόλη στο παιχνίδι που περιγράφθηκε στο Παράδειγμα 3.17 για την ειδική περίπτωση που $a = 1$.

Λύση Παραδείγματος 3.22

Έστω X η τ.μ. που παριστάνει το κέρδος του Αποστόλη. Με βάση όσα υπολογίστηκαν στη λύση του Παραδείγματος 3.17 για $a = 1$, είναι $E(X) = -\frac{1}{13}$. Η διακύμανση είναι

$$Var(X) = E(X^2) - E(X)^2,$$

και συνεπώς, αρκεί να υπολογιστεί η $E(X^2)$. Είναι

$$E(X^2) = 3^2 \cdot \frac{12}{52} + (-1)^2 \cdot \frac{40}{52} = \frac{148}{52}.$$

Επομένως, $Var(X) = \frac{148}{52} - (\frac{1}{13})^2 = 2.840237$.

Παράδειγμα 3.23

Να υπολογιστεί η διακύμανση της τ.μ. X του Παραδείγματος 3.18.

Λύση Παραδείγματος 3.23

Η διακύμανση δίνεται από τη σχέση $Var(X) = E(X^2) - (E(X))^2$ και, επομένως, αρκεί να υπολογιστεί η $E(X^2)$, καθώς η $E(X) = 3$ από τη λύση του Παραδείγματος 3.18. Επομένως,

$$E(X^2) = \int_0^4 x^2 \frac{3x^2}{64} dx = \frac{3}{64}\frac{4^5}{5} = \frac{48}{5}$$

και άρα, $Var(X) = \frac{48}{5} - 9 = 0.6$.

Άσκηση Αυτοαξιολόγησης 3.9

Για την τ.μ. X της Άσκησης Αυτοαξιολόγησης 3.7 να υπολογιστούν οι διακυμάνσεις $Var(X)$, $Var(3X - 1)$ και $Var(7X^2 + 8)$.

Άσκηση Αυτοαξιολόγησης 3.10

Για την τ.μ. X της Άσκησης Αυτοαξιολόγησης 3.8 να υπολογιστούν οι διακυμάνσεις $Var(X)$, $Var(3X - 1)$ και $Var(7X^2 + 8)$.

Άσκηση Αυτοαξιολόγησης 3.11

Έστω X τ.μ. με ασκ

$$F(x) = \begin{cases} 0, & x < 0, \\ x^4 & 0 \le x \le 1, \\ 1, & x > 1. \end{cases}$$

Να υπολογιστούν η μέση τιμή και η διακύμανση της τ.μ. X.

Άσκηση Αυτοαξιολόγησης 3.12

Έστω X η τ.μ. με σππ

$$f(x) = \begin{cases} \frac{1}{9} \cdot (x + 3) & -3 \le x \le 0, \\ -\frac{1}{9} \cdot (x - 3), & 0 \le x \le 3. \end{cases}$$

Να υπολογιστούν η μέση τιμή και η διακύμανση της τ.μ. X.

Άσκηση Αυτοαξιολόγησης 3.13

Έστω X η τ.μ. με σππ

$$f(x) = \frac{1}{8} + \frac{3}{8}x, \ 0 \le x \le 2.$$

Να υπολογιστούν η μέση τιμή και η διακύμανση της τ.μ. X.

Άσκηση Αυτοαξιολόγησης 3.14

Έστω X η τ.μ. με σππ

$$f(x) = \begin{cases} x, & 0 \le x < 1, \\ 2 - x & 1 \le x \le 2, \\ 0, & \text{αλλού.} \end{cases}$$

Να υπολογιστούν η μέση τιμή και η διακύμανση της τ.μ. X.

3.6.3 Άλλα χαρακτηριστικά μέτρα

Η αναμενόμενη τιμή και η διακύμανση μιας τυχαίας μεταβλητής δεν είναι τα μόνα αριθμητικά μέτρα που έχουν στόχο να μας δώσουν πληροφορίες για τη θέση, το σχήμα και τη μορφή της κατανομής της τ.μ. Στην ενότητα αυτή θα παρουσιαστούν οι ορισμοί κάποιων επιπλέον μέτρων θέσης, διασποράς και σχήματος της κατανομής της τ.μ.

Ορισμός 3.11

Κάθε σημείο k μιας τ.μ. X για το οποίο ισχύει:

$$p_X(k) = \max_x p_X(x), \text{ όταν } X \text{ διακριτή με σπ } p_X(x), \tag{3.18}$$

ή

$$f_X(k) = \max_x f_X(x), \text{ όταν } X \text{ συνεχής με σππ } f_X(x), \tag{3.19}$$

λέγεται **κορυφή ή επικρατούσα τιμή** της τυχαίας μεταβλητής X.

Επομένως, ο προσδιορισμός της κορυφής μιας τ.μ. ανάγεται σε ένα πρόβλημα μεγιστοποίησης και ως τέτοιο μπορεί ή να έχει μοναδική λύση ή περισσότερες από μία λύσεις ή ακόμη και να μην υπάρχει μέγιστο. Η κορυφή ή οι κορυφές αντιστοιχούν στην περίπτωση των διακριτών τ.μ. στις τιμές με τη μεγαλύτερη πιθανότητα πραγματοποίησης, ενώ στην περίπτωση συνεχών τ.μ. στις τιμές που μεγιστοποιούν τη συνάρτηση πυκνότητας πιθανότητας.

Στη συνέχεια, ορίζουμε μια άλλη κατηγορία χαρακτηριστικών σημείων μιας τ.μ. που είναι γνωστά ως ποσοστιαία ή εκατοστιαία σημεία.

Ορισμός 3.12

Κάθε σημείο x_p, $0 < p < 1$, μιας τ.μ. X με ασκ $F_X(\cdot)$ για το οποίο ισχύει:

$$P(X \leq x_p) = F_X(x_p) = p, \text{ όταν } X \text{ συνεχής} \tag{3.20}$$

ή

$$P(X \leq x_p) = F_X(x_p) \geq p, \text{ και } P(X \geq x_p) \geq (1-p), \text{ όταν } X \text{ διακριτή,} \tag{3.21}$$

λέγεται p**-ποσοστιαίο σημείο** της τυχαίας μεταβλητής X. Στην ειδική περίπτωση που $p = 0.5$ το σημείο $x_{0.5}$ λέγεται **διάμεσος**, ενώ τα σημεία $x_{0.25}$, $x_{0.5}$, $x_{0.75}$ ορίζονται να είναι το πρώτο, δεύτερο και τρίτο τεταρτημόριο. Τέλος, η διαφορά $x_{0.75} - x_{0.25}$ ονομάζεται **ενδοτεταρτημοριακό εύρος**.

Γενικεύσεις της αναμενόμενης τιμής και της διακύμανσης αποτελούν οι λεγόμενες ροπές k τάξης, οι οποίες διακρίνονται σε απλές ροπές ή ροπές περί το μηδέν, σε κεντρικές ροπές ή ροπές περί τη μέση τιμή και σε τυπικές ή κανονικοποιημένες ροπές. Οι ορισμοί τους παρατίθενται στη συνέχεια.

Ορισμός 3.13

Η k**−τάξης απλή ροπή** ή k**−τάξης ροπή περί το μηδέν** ή απλά k**−τάξης ροπή** της τυχαίας μεταβλητής X με k θετικό ακέραιο συμβολίζεται με μ_k και ορίζεται ως $\mu_k = E(X^k)$. Αν η τυχαία μεταβλητή X είναι διακριτή με σύνολο τιμών $S_X = \{x_1, \dots x_n, \dots\}$ και συνάρτηση πιθανότητας $p_X(\cdot)$, ορίζεται από τη σχέση:

$$\mu_k = \sum_{x \in S_X} x^k \cdot p_X(x), \tag{3.22}$$

ενώ, αν η τυχαία μεταβλητή X είναι συνεχής με σύνολο τιμών S_X και συνάρτηση πυκνότητας πιθανότητας

$f_X(\cdot)$, ορίζεται από τη σχέση:

$$\mu_k = \int_{S_X} x^k \cdot f_X(x)dx. \tag{3.23}$$

Τα παραπάνω ορίζονται με την προϋπόθεση ότι οι σειρές και τα ολοκληρώματα που εμφανίζονται συγκλίνουν απόλυτα.

Παρατηρήστε ότι άμεσα από τον ορισμό προκύπτει ότι $\mu_1 = E(X) = \mu$ και, έτσι, αιτιολογείται ότι οι k-τάξης ροπές αποτελούν γενίκευση της αναμενόμενης τιμής. Στη συνέχεια, ορίζεται η $k-$τάξης ροπή περί τη μέση τιμή ή αλλιώς η $k-$τάξης κεντρική ροπή. Τα παρακάτω ορίζονται με την προϋπόθεση ότι οι σειρές και τα ολοκληρώματα που εμφανίζονται συγκλίνουν απόλυτα.

Ορισμός 3.14

Έστω X μια τυχαία μεταβλητή με πεπερασμένη αναμενόμενη τιμή $\mu = E(X)$. Η $k-$**τάξης κεντρική ροπή** ή $k-$**τάξης ροπή περί τη μέση τιμή** της τυχαίας μεταβλητής X, με k θετικό ακέραιο, συμβολίζεται με ν_k και ορίζεται ως $\nu_k = E\left((X-\mu)^k\right)$. Ειδικότερα, αν η τυχαία μεταβλητή X είναι διακριτή με σύνολο τιμών $S_X = \{x_1, \dots x_n, \dots\}$ και συνάρτηση πιθανότητας $p_X(\cdot)$, ορίζεται από τη σχέση:

$$\nu_k = \sum_{x \in S_X} (x-\mu)^k \cdot p_X(x), \tag{3.24}$$

ενώ, αν η τυχαία μεταβλητή X είναι συνεχής με σύνολο τιμών S_X και συνάρτηση πυκνότητας πιθανότητας $f_X(\cdot)$, ορίζεται από τη σχέση:

$$\nu_k = \int_{S_X} (x-\mu)^k \cdot f_X(x)dx. \tag{3.25}$$

Παρατηρήστε ότι $\nu_1 = 0$ για οποιαδήποτε τυχαία μεταβλητή και επομένως δεν έχει ως μέτρο καμία συνεισφορά. Επίσης, $\nu_2 = Var(X) = \sigma^2$.

Ορισμός 3.15

Έστω X μια τυχαία μεταβλητή με πεπερασμένη αναμενόμενη τιμή και διακύμανση, $\mu = E(X)$ και $\sigma^2 = Var(X)$, αντίστοιχα. Η $k-$**τάξης τυπική ή κανονικοποιημένη ροπή** της τυχαίας μεταβλητής X, με k θετικό ακέραιο, συμβολίζεται με β_k και ισούται με:

$$\beta_k = E\left[\left(\frac{X-\mu}{\sigma}\right)^k\right]. \tag{3.26}$$

Ιδιαίτερο ενδιαφέρον παρουσιάζουν οι ειδικές περιπτώσεις των κανονικοποιημένων ροπών τρίτης και τέταρτης τάξης. Η ροπή β_3 ονομάζεται **μέτρο ή συντελεστής λοξότητας**. Αν η κατανομή είναι συμμετρική ως προς μ, τότε προφανώς $\beta_3 = 0$, ενώ λέμε ότι είναι λοξή προς τα δεξιά (αριστερά), αν $\beta_3 > (<)0$. Η ροπή β_4 ονομάζεται **μέτρο ή συντελεστής κύρτωσης**. Όταν $\beta_4 = 3$, λέμε ότι η κατανομή είναι μεσόκυρτη, για τιμές $\beta_4 > 3$ λεπτόκυρτη και για τιμές $\beta_4 < 3$ πλατύκυρτη.

Οι ροπές μιας τυχαίας μεταβλητής διαδραματίζουν σημαντικό ρόλο αφού σε μερικές περιπτώσεις η γνώση τους οδηγεί σε προσδιορισμό της αντίστοιχης κατανομής. Μια συνάρτηση που μας επιτρέπει τον υπολογισμό των ροπών κάθε τάξης είναι η αποκαλούμενη ροπογεννήτρια συνάρτηση, ο ορισμός της οποίας έπεται.

Ορισμός 3.16

Η **ροπογεννήτρια συνάρτηση** της τ.μ. X συμβολίζεται με $M_X(t)$ και ορίζεται να είναι η πραγματική συνάρτηση $M_X(t) = E(e^{tX})$ για κάθε t που ανήκει σε ένα διάστημα τέτοιο, ώστε να υπάρχει η παραπάνω μέση τιμή. Ειδικότερα, αν η τυχαία μεταβλητή X είναι διακριτή με σύνολο τιμών $S_X = \{x_1, \dots x_n, \dots\}$ και

συνάρτηση πιθανότητας $p_X(\cdot)$ ορίζεται από τη σχέση:

$$M_X(t) = E(e^{tX}) = \sum_{x \in S_X} e^{tx} \cdot p_X(x), \tag{3.27}$$

για τις τιμές του t για τις οποίες το άθροισμα συγκλίνει, ενώ, αν η τυχαία μεταβλητή X είναι συνεχής με σύνολο τιμών S_X και συνάρτηση πυκνότητας πιθανότητας $f_X(\cdot)$, ορίζεται από τη σχέση:

$$M_X(t) = E(e^{tX}) = \int_{S_X} e^{tx} \cdot f_X(x)dx, \tag{3.28}$$

για τις τιμές του t για τις οποίες συγκλίνει το ολοκλήρωμα.

Η ονομασία ροπογεννήτρια αιτιολογείται πλήρως παρατηρώντας τα ακόλουθα. Υποθέτοντας ότι υπάρχει η ροπογεννήτρια της X σε μια περιοχή γύρω από το 0, τότε θα υπάρχει και η k-τάξης παράγωγός της ως προς t,

$$M_X^{(k)}(t) = E\left(X^k e^{tX}\right),$$

και θέτοντας $t = 0$ έχουμε:

$$M_X^{(k)}(0) = E\left(X^k\right).$$

Επομένως, αν είναι γνωστή η ροπογεννήτρια της X, με την παραπάνω προϋπόθεση, μπορούν να προσδιοριστούν οι k-τάξης ροπές της τ.μ.. Αντίστροφα, χρησιμοποιώντας το ανάπτυγμα της εκθετικής συνάρτησης σύμφωνα με το οποίο:

$$e^{tX} = \sum_{n=0}^{+\infty} \frac{(tX)^n}{n!}$$

έχουμε ότι

$$M_X(t) = E(e^{tX}) = E\left(\sum_{n=0}^{\infty} \frac{(tX)^n}{n!}\right),$$

και, επομένως,

$$M_X(t) = \sum_{n=0}^{\infty} \frac{t^n}{n!} E(X^n).$$

Δηλαδή, αν γνωρίζουμε όλες τις n τάξης ροπές, μπορούμε να προσδιορίσουμε τη ροπογεννήτρια στη μορφή της πιο πάνω σειράς και από εκεί και πέρα το ερώτημα είναι αν μπορεί αυτή να βρεθεί σε κλειστή αναλυτική μορφή ή/και αν συγκλίνει.

Άλλες ιδιότητες της ροπογεννήτριας, που προκύπτουν άμεσα από τον ορισμό της, είναι ότι

$$M_X(0) = E\left(e^0\right) = 1,$$

ενώ, αν $Y = aX + b$, με a, b πραγματικές σταθερές, τότε:

$$M_Y(t) \quad = \quad E(e^{tY}) = E\left(e^{atX+bt}\right) = E\left(e^{atX}e^{bt}\right) = e^{bt}M_X(at).$$

Ωστόσο, η πιο σημαντική ιδιότητα της ροπογεννήτριας συνάρτησης είναι αυτή που διατυπώνεται στο ακόλουθο θεώρημα το οποίο είναι γνωστό ως θεώρημα του μονοσήμαντου των ροπογεννητριών (για την απόδειξή του, η οποία ξεφεύγει από τους σκοπούς του παρόντος συγγράμματος, παραπέμπουμε στο σύγγραμμα Feller, 1968).

Θεώρημα 3.1

Η ροπογεννήτρια μιας τυχαίας μεταβλητής χαρακτηρίζει μονοσήμαντα την κατανομή της, δηλαδή αν X και Y είναι δύο τυχαίες μεταβλητές με ροπογεννήτριες $M_X(t)$ και $M_Y(t)$, αντίστοιχα, που υπάρχουν για $t \in (-c, c)$, με $c > 0$ σταθερά και $M_X(t) = M_Y(t)$ για κάθε $t \in (-c, c)$, τότε οι τυχαίες μεταβλητές X και Y έχουν την ίδια κατανομή.

3.7 Κατανομή συνάρτησης τυχαίας μεταβλητής

Στην ενότητα αυτή, το ενδιαφέρον μας επικεντρώνεται στον προσδιορισμό της κατανομής μιας συνάρτησης μιας τυχαίας μεταβλητής. Ειδικότερα, υποθέτουμε ότι για την τ.μ. X, που είναι είτε διακριτή είτε συνεχής, γνωρίζουμε την κατανομή της, δηλαδή γνωρίζουμε είτε τη σπ ή τη σππ είτε την ασκ της είτε ακόμη, λόγω του μονοσήμαντου των ροπογεννητριών, τη ροπογεννήτριά της. Σε αυτό το πλαίσιο, για λόγους που επιβάλλει η θεωρητική μελέτη ενός προβλήματος, ενδιαφερόμαστε για τον προσδιορισμό της κατανομής της τ.μ. $Y = g(X)$, όπου $g(\cdot)$ μια πραγματική συνάρτηση.

Στην περίπτωση που η τ.μ. είναι διακριτή, τότε ο προσδιορισμός αυτός είναι τις περισσότερες φορές εύκολος, καθώς για την εύρεση της σπ σε κάθε σημείο y_i αρκεί να χρησιμοποιηθεί η σχέση:

$$p_Y(y_i) = \sum_{x_i : g(x_i) = y_i} p_X(x_i).$$

Ο τρόπος εφαρμογής της παραπάνω σχέσης διευκρινίζεται μέσω του παραδείγματος που ακολουθεί.

Παράδειγμα 3.24

Έστω η τ.μ. X με σπ

$$p_X(x) = \begin{cases} 0.16, & x = -2, \\ 0.27 & x = -1, \\ 0.39, & x = 0, \\ 0.13, & x = 1, \\ 0.05, & x = 2, \\ 0, & \text{αλλού.} \end{cases}$$

Να προσδιοριστεί η σπ της τ.μ. $Y = X^2 + 3$.

Λύση Παραδείγματος 3.24

Καθώς το σύνολο των δυνατών τιμών της τ.μ. X είναι $S_X = \{-2, -1, 0, 1, 2\}$ και ο μετασχηματισμός που μας δίνεται είναι $Y = X^2 + 3$, προκύπτει άμεσα ότι το σύνολο των δυνατών τιμών της Y είναι $S_Y = \{(-2)^2 + 3, (-1)^2 + 3, 0^2 + 3, 1^2 + 3, 2^2 + 3\} = \{3, 4, 7\}$ και ισχύει:

$$P(Y = 3) = P(X = 0) = 0.39,\ P(Y = 4) = P(X = 1) + P(X = -1) = 0.13 + 0.27 = 0.4,$$

και

$$P(Y = 7) = P(X = 2) + P(X = -2) = 0.05 + 0.16 = 0.21.$$

Στην περίπτωση, τώρα, που η τ.μ. είναι συνεχής, για τον προσδιορισμό της κατανομής της τ.μ. Y, μπορεί να ακολουθηθεί μία εκ των μεθόδων που ακολουθούν.

Η πρώτη μέθοδος είναι γνωστή ως μέθοδος της αθροιστικής συνάρτησης κατανομής. Η μέθοδος ουσιαστικά έγκειται στο να προσδιοριστεί η αθροιστική συνάρτηση κατανομής της $Y = g(X)$, χρησιμοποιώντας την ασκ της τ.μ. X. Εφόσον προσδιοριστεί η ασκ, η σπ προκύπτει με παραγώγιση. Η μέθοδος διευκρινίζεται μέσω των παραδειγμάτων που ακολουθούν.

Παράδειγμα 3.25

Έστω X τ.μ. με ασκ

$$F_X(x) = \begin{cases} 0, & x < 0, \\ x^4 & 0 \le x \le 1, \\ 1, & x > 1. \end{cases}$$

Να προσδιοριστούν η ασκ και η σπ της τ.μ. $Y = X^2$.

Λύση Παραδείγματος 3.25

Το σύνολο των δυνατών τιμών της Y είναι το διάστημα $[0,1]$. Είναι από τον ορισμό της ασκ

$$F_Y(y) = P(Y \le y) = P(X^2 \le y) = P(-\sqrt{y} \le X \le \sqrt{y}) = F_X(\sqrt{y}) - F_X(-\sqrt{y}),$$

και, επομένως,

$$F_y(y) = \begin{cases} 0, & y < 0, \\ y^2 & 0 \le y \le 1, \\ 1, & y > 1, \end{cases}$$

από όπου προκύπτει ότι:

$$f_y(y) = \begin{cases} 2y, & 0 \le y \le 1, \\ 0, & \text{αλλού.} \end{cases}$$

Παράδειγμα 3.26

Έστω X η τ.μ. με σππ $f_X(x) = \frac{1}{2}$, $x \in [-1,1]$. Να προσδιοριστεί η κατανομή των $Y = e^X$ και $Z = X^2$.

Λύση Παραδείγματος 3.26

Το σύνολο των δυνατών τιμών της Y είναι το διάστημα $[e^{-1}, e]$. Είναι:

$$F_Y(y) = P(Y \le y) = P(e^X \le y) = P(X \le \log(y)) = F_X(\log(y)),$$

και, επομένως,

$$f_Y(y) = \frac{d}{dy} F_y(y) = f_X(\log(y)) \frac{d}{dy}(\log(y)),$$

ή, ισοδύναμα,

$$f_Y(y) = \frac{1}{2y}, y \in [e^{-1}, e].$$

Το σύνολο των δυνατών τιμών της Z είναι το διάστημα $[0,1]$. Είναι:

$$F_Z(z) = P(Z \le z) = P(X^2 \le z) = P(-\sqrt{z} \le X \le \sqrt{z}) = F_X(\sqrt{z}) - F_X(-\sqrt{z}),$$

και, επομένως,

$$f_Z(z) = \frac{d}{dz} F_Z(z) = f_X(\sqrt{z}) \frac{d}{dz}(\sqrt{z}) - f_X(-\sqrt{z}) \frac{d}{dz}(-\sqrt{z}),$$

ή, ισοδύναμα,

$$f_Z(z) = \frac{1}{4}z^{-\frac{1}{2}} + \frac{1}{4}z^{-\frac{1}{2}} = \frac{z^{-\frac{1}{2}}}{2}, z \in [0,1].$$

Η δεύτερη μέθοδος, που είναι γνωστή ως μέθοδος του μετασχηματισμού, στηρίζεται στην εφαρμογή της παρακάτω πρότασης, για την απόδειξη της οποίας παραπέμπουμε, μεταξύ άλλων, στο σύγγραμμα των Hogg and Craig (1978).

Πρόταση 3.6

Έστω συνεχής τυχαία μεταβλητή X με συνάρτηση πυκνότητας πιθανότητας $f_X(x)$ και σύνολο δυνατών τιμών S_X. Θεωρούμε τον μετασχηματισμό $Y = g(X)$ και έστω ότι υπάρχει μια διαμέριση $\{S_1, S_2, \dots, S_n\}$ του S τέτοια, ώστε:

- ο μετασχηματισμός $y = g(x)$ να είναι $1-1$ μετασχηματισμός του S_i στο $g(S_i) = \{y : y = g(x), x \in S_i\}$, $i = 1, \dots, n$, και
- ο αντίστροφος μετασχηματισμός της $g(x)$ για $x \in S_i$, έστω $g_i^{-1}(y)$, να έχει συνεχή πρώτη παράγωγο, η οποία είναι τέτοια, ώστε $\frac{d}{dy}g_i^{-1}(y) \neq 0$ για $y \in g(S_i)$, $i = 1, \dots, n$.

Τότε η τυχαία μεταβλητή $Y = g(X)$ είναι συνεχής με σύνολο τιμών το $g(S_X)$ και συνάρτηση πυκνότητας πιθανότητας:

$$f_Y(y) = \sum_{i=1}^{n} f_X\left(g_i^{-1}(y)\right)\left|\frac{d}{dy}g_i^{-1}(y)\right|, \quad y \in g(S).$$

Ο τρόπος εφαρμογής της μεθόδου διευκρινίζεται μέσω των ακόλουθων παραδειγμάτων.

Παράδειγμα 3.27

Έστω X τ.μ. με σππ

$$f_X(x) = \begin{cases} 4x^3 & 0 \leq x \leq 1, \\ 0, & \text{αλλού.} \end{cases}$$

Να προσδιοριστεί η σππ της τ.μ. $Y = X^2$.

Λύση Παραδείγματος 3.27

Το σύνολο των δυνατών τιμών της X είναι $S_X = [0,1]$. Μας ζητείται η εύρεση της σππ του μετασχηματισμού της τ.μ. X, που ορίζεται από τη σχέση $Y = g(X) = X^2$. Παρατηρούμε ότι ο μετασχηματισμός αυτός είναι $1-1$ μετασχηματισμός του S_X στο $g(S_X) = \{y : y \in [0,1]\}$. Επομένως, δεν χρειάζεται να αναζητήσουμε διαμέριση του S_X, καθώς η συνθήκη της πρότασης για $1-1$ μετασχηματισμό ισχύει σε όλο το σύνολο S_X. Στη συνέχεια, επιλύουμε τη σχέση $y = x^2$ ως προς x για να βρούμε τον αντίστροφο μετασχηματισμό. Είναι τότε $g^{-1}(y) = \sqrt{y}$, $y \in [0,1]$. Η πρώτη παράγωγος του αντίστροφου μετασχηματισμού υπάρχει, είναι συνεχής και ίση με:

$$\frac{d}{dy}g^{-1}(y) = \frac{1}{2y^{\frac{1}{2}}} \neq 0, \text{ για } y \in [0,1].$$

Επομένως, σύμφωνα με την Πρόταση 3.6, έχουμε ότι:

$$f_Y(y) = f_X\left(g^{-1}(y)\right)\left|\frac{d}{dy}g^{-1}(y)\right| = 4\left(\sqrt{y}\right)^3 \frac{1}{2y^{\frac{1}{2}}} = 2y, \; y \in [0,1].$$

Παράδειγμα 3.28

Έστω X η τ.μ. με σππ $f_X(x) = \frac{1}{2}$, $x \in [-1,1]$. Να προσδιοριστεί η κατανομή της τ.μ. $Z = X^2$.

Λύση Παραδείγματος 3.28

Το σύνολο των δυνατών τιμών της X είναι $S_X = [-1,1]$. Μας ζητείται η εύρεση της σππ του μετασχηματισμού της τ.μ. X, που ορίζεται από τη σχέση $Z = g(X) = X^2$ με $g(S_X) = [0,1]$. Παρατηρούμε ότι ο μετασχηματισμός αυτός δεν είναι $1-1$ μετασχηματισμός του S_X στο $g(S_X) = \{z : z \in [0,1]\}$. Το ερώτημα είναι αν υπάρχει διαμέριση του S_X που να είναι τέτοια, ώστε σε κάθε σύνολό της ο μετασχηματισμός που δίνεται να είναι $1-1$. Θεωρούμε τη διαμέριση του $S_X = (-\infty,0) \cup [0,+\infty) = S_1 \cup S_2$, που είναι τέτοια, ώστε ο μετασχηματισμός $Z = g(X) = X^2$ να είναι $1-1$ στα S_1 και S_2 με

$$g_1^{-1}(z) = -\sqrt{z} \text{ και } g_2^{-1}(z) = \sqrt{z}, \text{ αντίστοιχα.}$$

Οι πρώτες παράγωγοι αυτών των αντίστροφων μετασχηματισμών υπάρχουν, είναι συνεχείς και ίσες με:

$$\frac{d}{dz}g_1^{-1}(z) = -\frac{1}{2z^{\frac{1}{2}}} \neq 0 \text{ και } \frac{d}{dz}g_2^{-1}(z) = \frac{1}{2z^{\frac{1}{2}}} \neq 0.$$

Επομένως, σύμφωνα με την Πρόταση 3.6, έχουμε ότι:

$$\begin{aligned} f_Z(z) &= \sum_{i=1}^{2} f_X\left(g_i^{-1}(z)\right)\left|\frac{d}{dy}g_i^{-1}(z)\right| \\ &= \frac{1}{2}\left|-\frac{1}{2z^{\frac{1}{2}}}\right| + \frac{1}{2}\left|\frac{1}{2z^{\frac{1}{2}}}\right| = \frac{1}{2z^{\frac{1}{2}}},\ z \in [0,1]. \end{aligned}$$

Άσκηση Αυτοαξιολόγησης 3.15

Έστω X η τ.μ. με σππ $f_X(x) = \frac{1}{2}$, $x \in [-1,1]$. Να προσδιοριστεί η κατανομή της $Y = e^X$.

Άσκηση Αυτοαξιολόγησης 3.16

Έστω X η τ.μ. με σππ $f_X(x)$, $x \in \mathbb{R}$. Να προσδιοριστεί η κατανομή της $Y = |X|$ με τη μέθοδο της ασκ και του μετασχηματισμού.

Η τρίτη μέθοδος που μπορεί να χρησιμοποιηθεί τόσο για συνεχείς όσο και διακριτές τ.μ. είναι η γνωστή ως μέθοδος της ροπογεννήτριας. Η μέθοδος αυτή ουσιαστικά χρησιμοποιεί το θεώρημα του μονοσήμαντου των ροπογεννητριών και η μεθοδολογία της είναι η ακόλουθη. Προσδιορίζουμε τη ροπογεννήτρια της τ.μ. $Y = g(X)$, δηλαδή τη ροπογεννήτρια του μετασχηματισμού της τ.μ. X. Αν αυτή η ροπογεννήτρια ταυτίζεται με τη ροπογεννήτρια κάποιας γνωστής μας κατανομής, όπως αυτές που θα μελετηθούν στα δύο επόμενα κεφάλαια του παρόντος συγγράμματος, τότε, λόγω του μονοσήμαντου των ροπογεννητριών, η κατανομή της τ.μ. Y είναι αυτή η γνωστή κατανομή. Επομένως, η μέθοδος προϋποθέτει δύο βασικά στοιχεία: να μπορούμε να προσδιορίζουμε τη ροπογεννήτρια του μετασχηματισμού και να είμαστε, ας μας επιτραπεί ο όρος, τυχεροί και να συμπίπτει με τη ροπογεννήτρια συνάρτηση κάποιας γνωστής στη βιβλιογραφία κατανομής.

Παράδειγμα 3.29

Έστω X η τ.μ. με σππ $f_X(x) = \lambda e^{-\lambda x}$, $x \geq 0$. Να προσδιοριστεί η κατανομή της τ.μ. $Y = kX$, $k > 0$.

Λύση Παραδείγματος 3.29

Σύμφωνα με τη μέθοδο της ροπογεννήτριας αρκεί να προσδιορίσουμε τη ροπογεννήτρια της τ.μ. Y. Είναι:

$$M_Y(t) = E(e^{Yt}) = E(e^{kXt}) = M_X(kt).$$

Επομένως, αρκεί να προσδιορίσουμε τη ροπογεννήτρια της τ.μ. X με σππ $f_X(x) = \lambda e^{-\lambda x}$, $x \geq 0$. Επιστρέφουμε στον υπολογισμό της ροπογεννήτριας της X. Είναι

$$M_X(t) = E\left(e^{tX}\right) = \int_0^{+\infty} e^{tx}\lambda e^{-\lambda x}dx = \lambda \int_0^{+\infty} e^{-(\lambda - t)x}dx.$$

Το τελευταίο ολοκλήρωμα είναι πεπερασμένο, αν $\lambda - t > 0$, δηλαδή αν $t < \lambda$ και τότε προκύπτει ότι:

$$M_X(t) = \frac{\lambda}{(\lambda - t)}, t < \lambda. \tag{3.29}$$

Αφού προσδιορίστηκε η ροπογεννήτρια της X είναι:

$$M_Y(t) = M_X(kt) = \frac{\lambda}{(\lambda - kt)}, kt < \lambda,$$

ή, ισοδύναμα,

$$M_Y(t) = \frac{\frac{\lambda}{k}}{(\frac{\lambda}{k} - t)}, t < \frac{\lambda}{k}.$$

Παρατηρούμε τώρα μια ομοιότητα μεταξύ της ροπογεννήτριας της Y και της X, με τη διαφοροποίηση ότι στη θέση του λ έχουμε $\frac{\lambda}{k}$. Επομένως, η ροπογεννήτρια της Y ταυτίζεται με τη ροπογεννήτρια της κατανομής που έχει σππ ίδια με αυτήν της X με την προαναφερθείσα τροποποίηση. Επομένως, η τ.μ. Y έχει σππ:

$$f_Y(y) = \frac{\lambda}{k} e^{-\frac{\lambda y}{k}}, y \geq 0.$$

Παράδειγμα 3.30

Έστω X η τ.μ. με σππ:

$$f_X(x) = \frac{b}{a^b} x^{b-1} e^{-\left(\frac{x}{a}\right)^b},\ x \geq 0, a > 0, b > 0.$$

Να προσδιοριστεί η κατανομή της τ.μ. $Y = \left(\frac{X}{a}\right)^b$.

Λύση Παραδείγματος 3.30

Σύμφωνα με τη μέθοδο της ροπογεννήτριας αρκεί να προσδιορίσουμε τη ροπογεννήτρια της τ.μ. Y. Είναι:

$$M_Y(t) = E(e^{Yt}) = E\left(e^{\left(\frac{X}{a}\right)^b t}\right).$$

Στη συνέχεια, θα χρησιμοποιήσουμε τον ορισμό της μέσης τιμής συνάρτησης τ.μ. και θα έχουμε:

$$\begin{aligned} M_Y(t) &= \int_0^{+\infty} e^{\left(\frac{x}{a}\right)^b t} \frac{b}{a^b} x^{b-1} e^{-\left(\frac{x}{a}\right)^b} dx \\ &= \frac{b}{a^b} \int_0^{+\infty} x^{b-1} e^{-(t-1)\left(\frac{x}{a}\right)^b} dx \\ &= \frac{1}{t-1}, t < 1, \end{aligned}$$

καθώς παρατηρήσαμε ότι:

$$\frac{b}{a^b} x^{b-1} e^{-(t-1)\left(\frac{x}{a}\right)^b} = -\frac{1}{t-1} \frac{d}{dx} \left(e^{-(t-1)\left(\frac{x}{a}\right)^b} \right).$$

Επομένως, προσδιορίστηκε η ροπογεννήτρια της τ.μ. Y. Ανατρέξτε στο προηγούμενο παράδειγμα και διαπιστώστε ότι ταυτίζεται με τη ροπογεννήτρια της σχέσης (3.29) για $\lambda = 1$. Επομένως, η τ.μ. Y ακολουθεί την κατανομή με σππ $f_Y(y) = e^{-y}$, $y \geq 0$.

3.8 Ασκήσεις

Άσκηση 3.1 *Αν $f_1(x)$, $f_2(x)$, $f_3(x)$ συναρτήσεις πυκνότητας πιθανότητας, τότε η συνάρτηση $g(x) = 0.3 \cdot f_1(x) + 0.5 \cdot f_2(x) + 0.2 \cdot f_3(x)$ είναι και αυτή συνάρτηση πιθανότητας.*

Είναι σωστή η παραπάνω πρόταση; Δικαιολογήστε την απάντησή σας.

Άσκηση 3.2 *Για την τ.μ. X δίνεται ότι $E(X) = 3$ και $Var(X) = 15$. Ποια από τις παρακάτω προτάσεις είναι σωστή;*

1. $E(X^2) = 18$.
2. *Αν* $Y = 2X^2 + 3$, *τότε* $E(Y) = 39$.
3. $Var(2X + 3) = 63$.
4. *Δεν ισχύει τίποτε από τα παραπάνω.*

Άσκηση 3.3 *Για ποια τιμή του a η συνάρτηση $P(x) = a\, 4^x$, $x = 1,2,\ldots,n$, με $n \in \{1,2,3,4\ldots\}$, είναι συνάρτηση πιθανότητας;*

Άσκηση 3.4 *Έστω το πείραμα τύχης που αφορά την επιλογή 4 φοιτητών. Η πιθανότητα κάθε φοιτητής να έχει μυωπία είναι ίση με 0.4 και κάθε φοιτητής έχει μυωπία ή όχι ανεξάρτητα από κάθε άλλο. Να προσδιορίσετε την κατανομή της τ.μ. που παριστάνει τον αριθμό των φοιτητών που έχουν μυωπία στους 4 που επιλέχθηκαν.*

Άσκηση 3.5 *Δίνεται η ασκ*

$$F(x) = \begin{cases} 0, & x < -2, \\ 1/2, & -2 \le x < -1, \\ 2/3, & -1 \le x < 0, \\ 11/12, & 0 \le x < 1, \\ 1, & 1 \le x. \end{cases}$$

Να υπολογιστούν οι πιθανότητες

$P(-1 < X \le 1.8)$, $P(X = 2)$, $P(-0.1 \le X \le 0.8)$, $P(X < 1.2)$ *και* $P(X > -1)$.

Άσκηση 3.6 *Δύο ζάρια ρίχνονται και X είναι η τ.μ. που παριστάνει το γινόμενο των ενδείξεών τους. Προσδιορίστε την κατανομή της τ.μ. X.*

Άσκηση 3.7 *Ένα δίκαιο νόμισμα έχει δύο όψεις, τις Α, Β. Το ρίχνουμε τρεις φορές και καταγράφουμε τον αριθμό των εμφανίσεων κάθε όψης, έστω a και b, αντίστοιχα. Έστω X η τ.μ. που παριστάνει τη διαφορά $a - b$. Προσδιορίστε την κατανομή της τ.μ. X.*

Άσκηση 3.8 *Ο Αποστόλης και η Σόνια παίζουν ένα παιχνίδι στο οποίο η Σόνια είναι πολύ καλύτερη και έχει πιθανότητα νίκης 0.7. Συμφωνούν να τελειώσει το παιχνίδι τους όταν κάποιος νικήσει 4 φορές συνολικά (όχι συνεχόμενα). Η Σόνια κέρδισε στο συνολικό παιχνίδι. Να προσδιορίσετε την κατανομή της τ.μ. που παριστάνει τον αριθμό των παιχνιδιών που χρειάστηκε η Σόνια για να κερδίσει.*

Άσκηση 3.9 *Η πιθανότητα κάποιος να αναρρώσει από κοινή γρίπη εντός 3 ημερών είναι 0.8. Έστω X η τ.μ. που παριστάνει τον αριθμό των ατόμων που ανάρρωσαν εντός 3 ημερών σε 5 τυχαία επιλεγμένα άτομα. Προσδιορίστε την κατανομή της τ.μ. X.*

Άσκηση 3.10 *Δίνεται η συνάρτηση* $f(x) = ax^2(2-x)$, $0 < x < 2$. *Για ποια τιμή της παραμέτρου* a *είναι σππ; Να βρεθούν η ασκ, η μέση τιμή και η διακύμανσή της.*

Άσκηση 3.11 *Δίνεται η συνάρτηση* $f(x) = 2 - |x|$, $-2 < x < 2$. *Να δείξετε ότι είναι σππ μιας τ.μ.* X *και να προσδιορίσετε την ασκ της.*

Άσκηση 3.12 *Δίνεται η συνάρτηση* $f(x) = ax(2-x)$, $0 < x < 2$. *Για ποια τιμή της παραμέτρου* a *είναι σππ; Να βρεθούν η ασκ, η μέση τιμή, η διακύμανση και η τυπική απόκλισή της.*

Άσκηση 3.13 *Ένα νόμισμα με όψεις Α και Β ρίχνεται 4 φορές. Έστω* X *η τ.μ. που παριστάνει τον αριθμό των φορών εμφάνισης δύο διαδοχικών όψεων Α. Ποιο είναι το σύνολο των δυνατών τιμών της* X*; Να βρεθούν η κατανομή, η μέση τιμή και η διακύμανσή της.*

Άσκηση 3.14 *Ο Αποστόλης και ο Πολυχρόνης παίζουν το εξής παιχνίδι: ρίχνει ο Αποστόλης δύο ζάρια και αν το άθροισμα των ενδείξεων είναι μικρότερο από 7 κερδίζει 2 Ευρώ, αν το άθροισμα των ενδείξεων είναι μεγαλύτερο από 7 χάνει* a *Ευρώ, ενώ αν το άθροισμα των ενδείξεων είναι 7 τότε δεν κερδίζει ούτε χάνει. Για ποια τιμή του* a *το παιχνίδι είναι δίκαιο, δηλαδή έχει ο Αποστόλης αναμενόμενο κέρδος 0;*

Άσκηση 3.15 *Δίνεται η συνάρτηση* $f(x) = ke^{-\frac{|x-5|}{5}}$, $x \in \mathbb{R}$. *Να προσδιοριστεί η σταθερά* k *έτσι ώστε να είναι σππ μιας τ.μ.* X *και να βρεθούν η μέση τιμή και η διακύμανσή της.*

Άσκηση 3.16 *Σε ένα δοχείο υπάρχουν 3 άσπρες, 2 μαύρες και 3 πράσινες σφαίρες. Επιλέγουμε τυχαία, χωρίς επανατοποθέτηση, δύο σφαίρες από αυτές. Υποθέστε ότι κερδίζουμε 1 Ευρώ για κάθε άσπρη σφαίρα που επιλέγουμε, χάνουμε 1 Ευρώ για κάθε μαύρη σφαίρα που επιλέγουμε, ενώ δεν χάνουμε αλλά ούτε κερδίζουμε κάτι για κάθε πράσινη σφαίρα. Προσδιορίστε το αναμενόμενο κέρδος.*

Άσκηση 3.17 *Σε ένα δοχείο υπάρχουν 3 άσπρες, 2 μαύρες και 3 πράσινες σφαίρες. Επιλέγουμε τυχαία, με επανατοποθέτηση, δύο σφαίρες. Υποθέστε ότι κερδίζουμε 1 Ευρώ για κάθε άσπρη σφαίρα που επιλέγουμε, χάνουμε 1 Ευρώ για κάθε μαύρη σφαίρα που επιλέγουμε, ενώ δεν χάνουμε αλλά ούτε κερδίζουμε κάτι για κάθε πράσινη σφαίρα. Προσδιορίστε το αναμενόμενο κέρδος.*

Άσκηση 3.18 *Δίνεται η συνάρτηση* $f(x) = k(2x - x^2)$, $0 < x < 2$. *Για ποια τιμή του* k *η* $f(x)$ *είναι σππ; Να βρείτε την ασκ, τη μέση τιμή, τη διακύμανσή της.*

Άσκηση 3.19 *Έστω* X *τ.μ. με* $f_X(x) = 4x^3$, $0 \le x \le 1$. *Να βρεθούν οι* $E(X)$ *και* $E\left((X-2)^2\right)$.

Άσκηση 3.20 *Έστω* X *τ.μ. με* $f_X(x) = ke^{-4x}$, $x \ge 0$. *Να βρεθούν η σταθερά* k, *οι* $E(X)$ *και* $E\left((X-2)^2\right)$.

Άσκηση 3.21 *Να βρεθεί η ροπογεννήτρια της τ.μ.* X *με σπ* $P(X=1) = p$ *και* $P(X=-1) = (1-p)$, $0 < p < 1$.

Άσκηση 3.22 *Υπολογίστε τη μέση τιμή, τη διακύμανση και τα τεταρτημόρια της τ.μ.* X *με σππ* $f_X(x) = k(2-x)$, $0 < x < 2$, *όπου* k *κατάλληλη σταθερά.*

Άσκηση 3.23 *Έστω* $M_X(t)$ *η ροπογεννήτρια της τ.μ.* X *και* $K(t) = \log(M_X(t))$. *Δείξτε ότι* $\mu = E(X) = K'(0)$ *και* $\sigma^2 = Var(X) = K''(0)$.

Άσκηση 3.24 *Έστω X διακριτή τ.μ. με σπ που δίνεται στον ακόλουθο πίνακα:*

x	4	5	8	10	12
$p(x)$	0.15	0.25	0.20	0.15	0.25

Να προσδιοριστούν η ασκ, η μέση τιμή, η τυπική απόκλιση της τυχαίας μεταβλητής X και οι πιθανότητες $P(X < 8)$ και $P(X \geq 6)$.

Άσκηση 3.25 *Έστω X η τ.μ. με σππ $f_X(x)$, $x \in [0,+\infty)$. Να προσδιοριστεί η κατανομή της $Y = X^n$ με τη μέθοδο της ασκ και του μετασχηματισμού.*

Άσκηση 3.26 *Έστω X η τ.μ. με σππ $f_X(x)$, $x \in \mathbb{R}$. Να προσδιοριστεί η κατανομή της $Y = X^2$ με τη μέθοδο της ασκ και του μετασχηματισμού.*

Άσκηση 3.27 *Έστω X η τ.μ. με σππ $f_X(x) = \lambda e^{-\lambda x}$, $x \geq 0$. Να προσδιοριστεί με τη μέθοδο της ασκ και του μετασχηματισμού η κατανομή της τ.μ. $Y = e^{-X}$.*

Άσκηση 3.28 *Έστω X η τ.μ. με σππ $f_X(x) = \lambda e^{-\lambda x}$, $x \geq 0$. Να προσδιοριστεί η κατανομή της τ.μ. $Y = ke^X$, $k > 0$.*

Άσκηση 3.29 *Έστω X η τ.μ. με σππ $f_X(x) = \lambda e^{-\lambda x}$, $x \geq 0$. Να προσδιοριστεί η κατανομή της τ.μ. $Y = \sqrt{X}$.*

Άσκηση 3.30 *Έστω X η τ.μ. με σππ $f_X(x) = 60x^2(1-x)^3$, $x \in [0,1]$. Να προσδιοριστεί η κατανομή της τ.μ. $Y = 1 - X$.*

Άσκηση 3.31 *Έστω X τ.μ. με σππ $f_X(x) = 1$, $0 < x < 1$. Προσδιορίστε την κατανομή της τ.μ. $Y = \log(X^2)$.*

3.9 Απαντήσεις στις ασκήσεις αυτοαξιολόγησης

Λύση Άσκησης Αυτοαξιολόγησης 3.1

Για να είναι η $p_X(x)$ σπ θα πρέπει να ισχύει $0 \leq \frac{kx}{n(n+1)} \leq 1$ για $x = 1, 2, \ldots, n$. Από αυτήν την ιδιότητα προκύπτει ότι θα πρέπει: $0 \leq k \leq (n+1)$. Επιπλέον, θα πρέπει να ισχύει:

$$\sum_{x=1}^{n} \frac{kx}{n(n+1)} = 1.$$

Χρησιμοποιώντας τη σχέση (Β΄.1) που δίνει το άθροισμα διαδοχικών όρων αριθμητικής προόδου έχουμε:

$$\sum_{x=1}^{n} \frac{kx}{n(n+1)} = \frac{k}{n(n+1)} \sum_{x=1}^{n} x = \frac{k}{n(n+1)} \frac{n(1+n)}{2} = \frac{k}{2}.$$

Επομένως, $k = 2$, που ικανοποιεί και τη σχέση $0 \leq \frac{kx}{n(n+1)} \leq 1$ για $x = 1, 2, \ldots, n$.

Λύση Άσκησης Αυτοαξιολόγησης 3.2

Για να είναι η $p_X(x)$ σπ θα πρέπει να ισχύει $0 \leq k \cdot x^2 \leq 1$ για $x = 1, 2, 3, 4$. Από αυτήν την ιδιότητα προκύπτει ότι θα πρέπει: $0 \leq k \leq \frac{1}{16}$. Επιπλέον, θα πρέπει να ισχύει: $\sum_{x=1}^{4} kx^2 = 1$, ήτοι

$$k\left(1^2 + 2^2 + 3^2 + 4^2\right) = 1$$

και, επομένως, $k = \frac{1}{30}$, που πληροί και τη σχέση $0 \leq k \leq \frac{1}{16}$.

Λύση Άσκησης Αυτοαξιολόγησης 3.3

Έστω X η τ.μ. που εκφράζει τον αριθμό των υποσυστημάτων που παρουσιάζουν βλάβη κατά τη διάρκεια της λειτουργίας του συστήματος. Έστω, επίσης, A_i το ενδεχόμενο να παρουσιάσει βλάβη το i-οστό υποσύστημα, με $P(A_i) = 0.1$ για $i = 1, 2, 3$. Τότε

$$P(X = 0) = P(A_1' \cap A_2' \cap A_3') = P(A_1')P(A_2')P(A_3') = 0.9^3 = 0.729,$$

$$\begin{aligned} P(X = 1) &= P\left((A_1' \cap A_2' \cap A_3) \cup (A_1' \cap A_2 \cap A_3') \cup (A_1 \cap A_2' \cap A_3')\right) \\ &= P(A_1' \cap A_2' \cap A_3) + P(A_1' \cap A_2 \cap A_3') + P(A_1 \cap A_2' \cap A_3') \\ &= 3 \cdot (1 - 0.9) \cdot 0.9^2 = 3 \cdot 0.1 \cdot (1 - 0.1)^2 = 0.243, \end{aligned}$$

$$P(X = 3) = P(A_1 \cap A_2 \cap A_3) = P(A_1)P(A_3)P(A_3) = 0.1^3 = 0.001,$$

ενώ, εκμεταλλευόμενοι τις ιδιότητες της σπ, έχουμε ότι:

$$P(X = 2) = 1 - P(X = 0) - P(X = 1) - P(X = 3) = 0.027.$$

Η ασκ της τ.μ. X είναι

$$F_X(x) = \begin{cases} 0, & x < 0, \\ 0.729, & 0 \leq x < 1, \\ 0.729 + 0.243 = 0.972 & 1 \leq x < 2, \\ 0.972 + 0.001 = 0.973, & 2 \leq x < 3, \\ 0.973 + 0.027 = 1, & x \geq 3. \end{cases}$$

Λύση Άσκησης Αυτοαξιολόγησης 3.4

Για να είναι η συνάρτηση $f(x)$ σππ θα πρέπει να είναι μη αρνητική πραγματική συνάρτηση, δηλαδή θα πρέπει $k + \frac{3}{8}x \geq 0$ για $0 \leq x \leq 2$. Από αυτήν τη συνθήκη προκύπτει ότι $k \geq 0$. Επιπλέον, θα πρέπει η $f(x)$ να είναι τέτοια, ώστε $\int_{-\infty}^{+\infty} f(x)dx = 1$. Συνεπώς, πρέπει

$$\int_0^2 \left(k + \frac{3}{8}x\right) dx = 1$$

ή, ισοδύναμα,

$$2k + \frac{3}{8}\frac{2^2}{2} = 1,$$

από όπου έχουμε ότι $k = \frac{1}{8}$, τιμή που ικανοποιεί και τη σχέση $k + \frac{3}{8}x \geq 0$ για $0 \leq x \leq 2$.
Για την εύρεση της ασκ θα χρησιμοποιήσουμε τον ορισμό $F(x) = P(X \leq x)$ για $x \in \mathbb{R}$. Εξ ορισμού είναι $F(x) = 0$ για $x < 0$, ενώ είναι $F(x) = \int_0^2 f(u)du = 1$ για $x \geq 2$. Απομένει να προσδιοριστεί η ασκ για $0 \leq x < 2$. Είναι τότε

$$F(x) = \int_0^x \left(\frac{1}{8} + \frac{3}{8}u\right)du = \frac{x}{8} + \frac{3}{16}x^2,\ 0 \leq x < 2.$$

Είναι

$$P(X > 1) = 1 - P(X \leq 1) = 1 - F(1) = 1 - \left(\frac{1}{8} + \frac{3}{16}1^2\right) = 0.6875,$$

και

$$P(0.5 < X < 1.5) = F(1.5) - F(0.5) = \left(\frac{1.5}{8} + \frac{3}{16}1.5^2\right) - \left(\frac{0.5}{8} + \frac{3}{16}0.5^2\right) = 0.5.$$

Λύση Άσκησης Αυτοαξιολόγησης 3.5

Παρατηρούμε ότι η ασκ είναι συνεχής. Αυτό σημαίνει ότι είναι ασκ συνεχούς τ.μ. Θα χρησιμοποιήσουμε τη σχέση που συνδέει την ασκ με τη σππ, δηλαδή τη σχέση (3.7) και θα έχουμε:

$$f(x) = \begin{cases} 4x^3 & 0 \leq x \leq 1, \\ 0, & \text{αλλού.} \end{cases}$$

Λύση Άσκησης Αυτοαξιολόγησης 3.6

Για να είναι η συνάρτηση $f(x)$ σππ θα πρέπει να είναι μη αρνητική πραγματική συνάρτηση, δηλαδή θα πρέπει η σταθερά k να είναι τέτοια, ώστε $k(x+3) \geq 0$ για $-3 \leq x \leq 0$ και, επιπλέον, $-k(x+3) \geq 0$ για $0 \leq x \leq 3$. Αυτές οι δύο σχέσεις μας οδηγούν στη συνθήκη $k \geq 0$. Επιπροσθέτως, θα πρέπει να είναι τέτοια, ώστε $\int_{-\infty}^{+\infty} f(x)dx = 1$. Επομένως, πρέπει

$$\int_{-3}^{0} k(x+3)dx + \int_0^3 -k(x-3)dx = 1$$

ή, ισοδύναμα,

$$-k\left(\frac{9}{2} - 9\right) - k\left(\frac{9}{2} - 9\right) = 1,$$

δηλαδή $k = \frac{1}{9}$ (που επαληθεύει και τη συνθήκη $k \geq 0$, η οποία αναφέρθηκε πρωτύτερα). Για την εύρεση της ασκ θα χρησιμοποιήσουμε τον ορισμό $F(x) = P(X \leq x)$ για $x \in \mathbb{R}$. Εξ ορισμού είναι $F(x) = 0$ για $x < -3$, ενώ είναι $F(x) = 1$ για $x \geq 3$. Απομένει να προσδιοριστεί η ασκ για $-3 \leq x < 0$ και $0 \leq x < 3$.

Είναι τότε:

$$F(x) = \int_{-3}^{x} \frac{1}{9}(u+3)du = \frac{1}{9}\left(\frac{x^2}{2} + 3x - (\frac{9}{2} - 9)\right), \text{ για } -3 \le x < 0,$$

ενώ

$$F(x) = \int_{-3}^{0} \frac{1}{9}(u+3)du + \int_{0}^{x} -\frac{1}{9}(u-3)du = 0.5 + \frac{3}{9}x - \frac{x^2}{18}, \text{ για } 0 \le x < 3.$$

Λύση Άσκησης Αυτοαξιολόγησης 3.7

Η τ.μ. X είναι διακριτή τ.μ. με $S_X = \{-2, 2, 0, -1, 3\}$ και, από τον ορισμό της μέσης τιμής για διακριτή τ.μ., έχουμε ότι:

$$E(X) = -2 \cdot P(X=-2) + 2 \cdot P(X=2) + 0 \cdot P(X=0) - 1 \cdot P(X=-1) + 3 \cdot P(X=3)$$

ή με αντικατάσταση

$$E(X) = -2 \cdot 0.3 + 2 \cdot 0.3 + 0 \cdot 0.3 - 1 \cdot 0.05 + 3 \cdot 0.05 = 0.1.$$

Επίσης, από τις ιδιότητες της μέσης τιμής

$$E(3X-1) = 3E(X) - 1 = -0.7,$$

και

$$E(7X^2 + 8) = 7E(X^2) + 8.$$

Επομένως, αρκεί να υπολογιστεί η $E(X^2)$. Είναι (βλ. τη σχέση (3.11))

$$E(X^2) = (-2)^2 \cdot 0.3 + 2^2 \cdot 0.3 + 0 \cdot 0.3 + (-1)^2 \cdot 0.05 + 3^2 \cdot 0.05 = 2.9$$

ενώ από τις ιδιότητες της μέσης τιμής έχουμε ότι $E(7X^2 + 8) = 7E(X^2) + 8 = 7 \cdot 2.9 + 8 = 28.3$.

Λύση Άσκησης Αυτοαξιολόγησης 3.8

Η τ.μ. X είναι συνεχής τ.μ. Από τον ορισμό της μέσης τιμής για συνεχή τ.μ. έχουμε ότι:

$$E(X) = \int_{0}^{1} x\, x^2 dx + \int_{1}^{5/3} x\, dx = \frac{1}{4} + \frac{8}{9} = \frac{41}{36}.$$

Επίσης, από τις ιδιότητες της μέσης τιμής

$$E(3X-1) = 3E(X) - 1 = 2.4167,$$

και

$$E(7X^2 + 8) = 7E(X^2) + 8.$$

Επομένως, αρκεί να υπολογιστεί η $E(X^2)$. Είναι

$$E(X^2) = \int_{0}^{1} x^2\, x^2 dx + \int_{1}^{5/3} x^2\, dx = \frac{1}{5} + \frac{98}{81} = \frac{571}{405}.$$

Επομένως, από τις ιδιότητες της μέσης τιμής, έχουμε ότι

$$E(7X^2 + 8) = 7E(X^2) + 8 = 7 \cdot \frac{571}{405} + 8 = \frac{7237}{405}.$$

Λύση Άσκησης Αυτοαξιολόγησης 3.9

Η τ.μ. X είναι διακριτή τ.μ. και στη λύση της Άσκησης Αυτοαξιολόγησης 3.7 έχουμε προσδιορίσει ότι $E(X) = 0.1$ και $E(X^2) = 2.9$. Επομένως,

$$Var(X) = E(X^2) - (E(X))^2 = 2.89.$$

Από τις ιδιότητες της διακύμανσης έχουμε ότι

$$Var(3X - 1) = 9 \cdot Var(X) = 9 \cdot 2.89 = 26.01.$$

Τέλος, ισχύει ότι $Var\left(7X^2 + 8\right) = 49 \cdot Var(X^2)$ και, επομένως, η εύρεσή της ανάγεται στην εύρεση της $Var(X^2)$. Είναι

$$Var(X^2) = E(X^4) - \left(E\left(X^2\right)\right)^2,$$

με

$$E(X^4) = (-2)^4 \cdot\ 0.3 + 2^4 \cdot\ 0.3 + 0 \cdot\ 0.3 + (-1)^4 \cdot\ 0.05 + 3^4 \cdot\ 0.05 = 13.7,$$

οπότε $Var(X^2) = 13.7 - (2.9)^2 = 5.29$.

Λύση Άσκησης Αυτοαξιολόγησης 3.10

Η τ.μ. X είναι συνεχής τ.μ. και από τη λύση της Άσκησης Αυτοαξιολόγησης 3.8 έχουμε ότι $E(X) = \frac{41}{36}$ και $E(X^2) = \frac{571}{405}$. Επομένως, $Var(X) = E(X^2) - (E(X))^2 = 16.57207$. Από τις ιδιότητες της διακύμανσης έχουμε ότι:

$$Var(3X - 1) = 9 \cdot Var(X) = 149.1486.$$

Τέλος, ισχύει ότι $Var\left(7X^2 + 8\right) = 49 \cdot Var(X^2)$ και, επομένως, η εύρεσή της ανάγεται στην εύρεση της $Var(X^2)$. Είναι

$$Var\left(X^2\right) = E\left(X^4\right) - \left(E\left(X^2\right)\right)^2,$$

με

$$E(X^4) = \int_0^1 x^4 x^2 dx + \int_1^{5/3} x^4 dx = \frac{1}{7} + \frac{2882}{1215} = 2.514874.$$

Επομένως, $Var(7X^2 + 8) = 49\left\{2.514874 - \left(\frac{571}{405}\right)^2\right\} = 25.82898.$

Λύση Άσκησης Αυτοαξιολόγησης 3.11

Η σπ της τ.μ. X είναι:

$$f(x) = \begin{cases} 4x^3 & 0 \le x \le 1, \\ 0, & \text{αλλού.} \end{cases}$$

Από τον ορισμό της μέσης τιμής είναι:

$$E(X) = \int_0^1 x\ 4x^3 dx = \frac{4}{5},$$

ενώ για τον υπολογισμό της διακύμανσης θα χρησιμοποιηθεί η σχέση

$$Var(X) = E(X^2) - (E(X))^2.$$

Είναι επομένως:

$$Var(X) = \int_0^1 x^2\ 4x^3 dx - \left(\frac{4}{5}\right)^2 = \frac{2}{75}.$$

Λύση Άσκησης Αυτοαξιολόγησης 3.12

Από τον ορισμό της μέσης τιμής είναι:

$$E(X) = \int_{-3}^{0} \frac{1}{9}x(x+3)dx + \int_{0}^{3} -\frac{1}{9}x(x-3)dx = -0.5 + 0.5 = 0.$$

Τότε έχουμε ότι: $Var(X) = E(X^2) - (E(X))^2 = E(X^2)$ και

$$E(X^2) = \int_{-3}^{0} \frac{1}{9}x^2(x+3)dx + \int_{0}^{3} -\frac{1}{9}x^2(x-3)dx = 0.75 + 0.75 = 1.5.$$

Λύση Άσκησης Αυτοαξιολόγησης 3.13

Η μέση τιμή της τ.μ. X ισούται με

$$E(X) = \int_{0}^{} 2x\left(\frac{1}{8} + \frac{3}{8}x\right)dx = \frac{5}{4}.$$

Για τη διακύμανση θα βασιστούμε στη σχέση $Var(X) = E(X^2) - E(X)^2$, όπου

$$E(X^2) = \int_{0}^{} 2x^2\left(\frac{1}{8} + \frac{3}{8}x\right)dx = \frac{11}{6},$$

και άρα $Var(X) = \frac{11}{6} - \left(\frac{5}{4}\right)^2 = \frac{13}{48}$.

Λύση Άσκησης Αυτοαξιολόγησης 3.14

Η μέση τιμή της τ.μ. X ισούται με

$$E(X) = \int_{0}^{1} xxdx + \int_{1}^{2} x(2-x)dx = \frac{1}{3} + \frac{2}{3} = 1.$$

Για τη διακύμανσή της θα βασιστούμε στη σχέση $Var(X) = E(X^2) - E(X)^2$. Για το $E(X^2)$ έχουμε ότι:

$$E(X^2) = \int_{0}^{1} x^2xdx + \int_{1}^{2} x^2(2-x)dx = \frac{1}{4} + \frac{11}{12} = \frac{7}{6},$$

και άρα $Var(X) = \frac{7}{6} - (1)^2 = \frac{1}{6}$.

Λύση Άσκησης Αυτοαξιολόγησης 3.15

Το σύνολο των δυνατών τιμών της X είναι $S_X = [-1, 1]$. Μας ζητείται η εύρεση της σππ του μετασχηματισμού της τ.μ. X, που ορίζεται από τη σχέση $Y = g(X) = e^X$. Παρατηρούμε ότι ο μετασχηματισμός αυτός είναι $1-1$ μετασχηματισμός του S_X στο $g(S_X) = \{y : y \in [e^{-1}, e]\}$. Επομένως, δεν χρειάζεται να αναζητήσουμε διαμέριση του S_X, καθώς η συνθήκη της πρότασης για $1-1$ μετασχηματισμό ισχύει σε όλο το σύνολο S_X. Στη συνέχεια, επιλύουμε τη σχέση $y = e^x$ ως προς x για να βρούμε τον αντίστροφο μετασχηματισμό. Είναι τότε $g^{-1}(y) = \log y$, $y \in [e^{-1}, e]$. Η πρώτη παράγωγος του αντίστροφου μετασχηματισμού υπάρχει, είναι συνεχής και ίση με

$$\frac{d}{dy}g^{-1}(y) = \frac{1}{y} \neq 0, \text{ για } y \in [e^{-1}, e].$$

Επομένως, σύμφωνα με την Πρόταση 3.6, έχουμε ότι:

$$f_Y(y) = f_X\left(g^{-1}(y)\right)\left|\frac{d}{dy}g^{-1}(y)\right| = \frac{1}{2}\cdot\frac{1}{y} = \frac{1}{2y},\ y \in [e^{-1}, e].$$

Λύση Άσκησης Αυτοαξιολόγησης 3.16

Το σύνολο των δυνατών τιμών της X είναι $S_X = [-1,1]$. Μας ζητείται η εύρεση της σππ του μετασχηματισμού της τ.μ. X, που ορίζεται από τη σχέση $Y = |X|$, με $g(S_X) = \{y : y \in [0,+\infty)\}$. Με τη μέθοδο της ασκ είναι:

$$\begin{aligned} F_Y(y) &= P(Y \le y) = P\left(|X| \le y\right) = P\left(-y \le X \le y\right) \\ &= F_X(y) - F_X(-y), y \in [0,+\infty) \end{aligned}$$

και, επομένως, είναι

$$f_Y(y) = \frac{d}{dy}F_Y(y) = f_X(y)\frac{d}{dy}(y) - f_X(-y)\frac{d}{dy}(-y) = f_X(y) + f_X(-y), y \ge 0.$$

Με τη μέθοδο του μετασχηματισμού παρατηρούμε ότι αυτός δεν είναι $1-1$ μετασχηματισμός του S_X στο $g(S_X) = \{y : y \in [0,+\infty)\}$. Το ερώτημα είναι αν υπάρχει διαμέριση του S_X που να είναι τέτοια, ώστε σε κάθε σύνολό της ο μετασχηματισμός που δίνεται να είναι $1-1$. Θεωρούμε τη διαμέριση του $S_X = (-\infty,0) \cup [0,+\infty) = S_1 \cup S_2$ που είναι τέτοια, ώστε ο μετασχηματισμός $Y = g(X) = |X|$ να είναι $1-1$ στα S_1 και S_2, με $g_1^{-1}(y) = -y$ και $g_2^{-1}(y) = y$, αντίστοιχα. Οι πρώτες παράγωγοι αυτών των αντίστροφων μετασχηματισμών υπάρχουν, είναι συνεχείς και ίσες με:

$$\frac{d}{dz}g_1^{-1}(z) = -1 \ne 0 \text{ και } \frac{d}{dz}g_2^{-1}(z) = 1 \ne 0.$$

Επομένως, σύμφωνα με την Πρόταση 3.6, έχουμε ότι:

$$\begin{aligned} f_Y(y) &= \sum_{i=1}^{2} f_X\left(g_i^{-1}(y)\right)\left|\frac{d}{dy}g_i^{-1}(y)\right| \\ &= f_X\left(-y\right)|-1| + f_X\left(y\right)|1| \\ &= f_X(y) + f_X(-y), y \ge 0. \end{aligned}$$

ΒΙΒΛΙΟΓΡΑΦΙΑ

Ελληνόγλωσση

Παπαϊωάννου, Τ. (1993). *Εισαγωγή στις Πιθανότητες και τη Στατιστική, Μέρος Ι: Πιθανότητες*. Ιωάννινα.

Ξενόγλωσση

Feller, W. (1968). *An Introduction to Probability Theory and its Applications.* John Wiley and Sons.
Hogg, R. V. and Craig, A. T. (1978). *Introduction to Mathematical Statistics.* Macmillan.

ΚΕΦΑΛΑΙΟ 4

ΕΙΔΙΚΕΣ ΔΙΑΚΡΙΤΕΣ ΚΑΤΑΝΟΜΕΣ

Σύνοψη

Σε αυτό το κεφάλαιο θα μελετηθούν βασικές διακριτές τυχαίες μεταβλητές που περιγράφουν ένα ευρύ φάσμα προβλημάτων και τυχαίων φαινομένων σε διάφορα επιστημονικά πεδία. Στο πλαίσιο αυτό, θα παρουσιαστούν οι ακόλουθες διακριτές κατανομές: η διακριτή ομοιόμορφη, η διωνυμική, η υπεργεωμετρική, η γεωμετρική, η αρνητική διωνυμική και η Poisson.

Προαπαιτούμενη γνώση: Κεφάλαια 1-3 του παρόντος συγγράμματος.

Προσδοκώμενα μαθησιακά αποτελέσματα: Όταν θα έχετε ολοκληρώσει τη μελέτη του κεφαλαίου αυτού, θα μπορείτε

- να αναγνωρίζετε καθεμία από τις διακριτές κατανομές που παρουσιάζονται σε αυτό το κεφάλαιο και πότε πρέπει να τις εφαρμόζετε και
- να υπολογίζετε πιθανότητες ενδεχομένων που συνδέονται με τις διακριτές κατανομές που παρουσιάζονται.

Γλωσσάριο επιστημονικών όρων

- Αρνητική διωνυμική κατανομή
- Γεωμετρική κατανομή
- Διακριτή ομοιόμορφη κατανομή
- Διωνυμική κατανομή
- Διωνυμικό τυχαίο πείραμα
- Δοκιμή Bernoulli
- Κατανομή Poisson
- Υπεργεωμετρική κατανομή

Οικονόμου, Π., Μαλεφάκη, Σ., & Μπατσίδης, Α. (2023). *Πιθανότητες - Στατιστική.*

4.1 Εισαγωγή

Στο παρόν κεφάλαιο, το ενδιαφέρον μας επικεντρώνεται στην παρουσίαση των σημαντικότερων και πιο ευρέως διαδεδομένων μονοδιάστατων διακριτών κατανομών, που αποτελούν πολύτιμο εργαλείο για τη μοντελοποίηση πλήθους πραγματικών τυχαίων φαινομένων, επιτρέπουν τον υπολογισμό πιθανοτήτων που συνδέονται με αυτά, καθώς και τη μελέτη διάφορων χαρακτηριστικών τους. Στο πλαίσιο αυτό, θα παρουσιαστούν οι ακόλουθες διακριτές κατανομές: η διακριτή ομοιόμορφη, η Bernoulli, η διωνυμική, η υπεργεωμετρική, η γεωμετρική, η αρνητική διωνυμική και η ίσως πιο δημοφιλής από αυτές, η Poisson.

4.2 Διακριτή ομοιόμορφη κατανομή

Στη ενότητα αυτή, θα παρουσιαστεί η απλούστερη διακριτή κατανομή που είναι γνωστή ως διακριτή ομοιόμορφη κατανομή. Η κατανομή αυτή βρίσκει εφαρμογή όταν η υπό μελέτη διακριτή τ.μ. X μπορεί να λάβει ισοπίθανα μια εκ των διακεκριμένων τιμών $x_1,\ldots,x_n$ με το n να είναι ένας πεπερασμένος αριθμός. Αν $S = \{x_1, \ldots, x_n\}$, τότε θα πρέπει $P(S) = 1$, όπου $S = \{x_1\} \cup \{x_2\} \cup \ldots \cup \{x_n\}$ με τα $\{x_1\}$, $\{x_2\},\ldots,\{x_n\}$ να αποτελούν μία διαμέριση του S. Επομένως, προκύπτει ότι $P(S) = \sum_{i=1}^{n} P(X = x_i)$. Λαμβάνοντας υπόψη ότι τα στοιχειώδη ενδεχόμενα είναι ισοπίθανα, προκύπτει άμεσα ότι $P(S) = nP(X = x_i)$ ή, ισοδύναμα, ότι $p_X(x) = P(X = x) = \frac{1}{n}$ για κάθε $x \in S$, ενώ $P(X = x) = 0$ για κάθε $x \notin S$. Τα παραπάνω μας οδηγούν στον ακόλουθο ορισμό.

Ορισμός 4.1

Η τυχαία μεταβλητή X λέγεται ότι ακολουθεί τη **διακριτή ομοιόμορφη κατανομή**, αν $S = \{x_1,\ldots,x_n\}$ είναι το σύνολο των διακεκριμένων δυνατών τιμών της, με n πεπερασμένο και η συνάρτηση πιθανότητάς της δίνεται από τη σχέση:

$$p_X(x) = \begin{cases} \frac{1}{n}, & x \in S, \\ 0, & \text{αλλού.} \end{cases} \tag{4.1}$$

Στην ειδική περίπτωση που οι τιμές $x_1,x_2,\ldots,x_n$ είναι οι διαδοχικοί ακέραιοι από την τιμή a έως την τιμή b, $a < b$, γράφουμε $X \sim DUnif(a,b)$ με $b = a + n - 1$.

Παραδείγματα πραγματικών τυχαίων φαινομένων που μοντελοποιούνται από τη διακριτή ομοιόμορφη κατανομή είναι η ρίψη ενός αμερόληπτου, δηλαδή τίμιου ζαριού, όπου καθεμία από τις 6 όψεις έχει ίδια πιθανότητα εμφάνισης, η τυχαία επιλογή ενός ακέραιου αριθμού από το 1 μέχρι το 45, η τυχαία επιλογή ενός μαθητή για να εκπροσωπήσει μια μαθητική κοινότητα 35 ατόμων σε μια εκδήλωση και άλλα παρόμοια παραδείγματα.

Στη συνέχεια, μελετούμε κάποιες ιδιότητες της κατανομής αυτής.

Πρόταση 4.1

Έστω ότι η τ.μ. X ακολουθεί τη διακριτή ομοιόμορφη κατανομή με τιμές τους n το πλήθος διαδοχικούς ακέραιους a, $a+1,\ldots$, $b-1$, b, δηλαδή $X \sim DUnif(a,b)$, $a < b$ με $b = a + n - 1$. Τότε η ασκ της δίνεται από τη σχέση:

$$F_X(x) = \begin{cases} 0, & x < a, \\ \frac{\lfloor x \rfloor - a + 1}{n}, & x \in [a,b), \\ 1, & x \geq b, \end{cases} \tag{4.2}$$

όπου με $\lfloor x \rfloor$ συμβολίζεται το ακέραιο μέρος του x. Επιπρόσθετα,

$$E(X) = \frac{a+b}{2} = \frac{2a + (n-1)}{2} \text{ και } Var(X) = \frac{n^2 - 1}{12}. \tag{4.3}$$

Απόδειξη Πρότασης 4.1

Χρησιμοποιώντας τη σχέση (4.1) και από τον ορισμό της ασκ έχουμε ότι, αν $X \sim DUnif(a,b)$, $a < b$, τότε:

$$F_X(x) = \begin{cases} 0, & x < a \\ \sum_{y=a}^{\lfloor x \rfloor} \frac{1}{n} = \frac{\lfloor x \rfloor - a + 1}{n}, & a \le x < b \\ 1, & x \ge b, \end{cases}$$

όπου το $\lfloor x \rfloor$ συμβολίζει το ακέραιο μέρος του x.

Από τον ορισμό της μέσης τιμής έχουμε $E(X) = \sum_{x=a}^{b} x\frac{1}{n} = \frac{1}{n}\sum_{x=a}^{b} x$, όπου $\sum_{x=a}^{b} x$ είναι το άθροισμα n διαδοχικών όρων αριθμητικής προόδου με πρώτο όρο το a και n-οστό όρο το b. Επομένως, χρησιμοποιώντας τη σχέση (Β΄.1) του Παραρτήματος Β΄, προκύπτει ότι $\sum_{x=a}^{b} x = \frac{n(a+b)}{2}$. Συνδυάζοντας τα παραπάνω, προκύπτει το ζητούμενο.

Για τον υπολογισμό της διακύμανσης έχουμε ότι: $Var(X) = E(X^2) - (E(X))^2$, όπου $E(X^2) = \frac{1}{n}\sum_{x=a}^{b} x^2$.

Ισοδύναμα, ισχύει ότι:

$$\begin{aligned} nE(X^2) &= a^2 + (a+1)^2 + \ldots + (a+(n-1))^2 \\ &= na^2 + 2a(1 + \ldots + (n-1)) + \left(1^2 + \ldots + (n-1)^2\right) \\ &= na^2 + a(n-1)n + \frac{n(n-1)(2n-1)}{6}. \end{aligned}$$

Σημειώστε ότι στη δεύτερη γραμμή το άθροισμα εντός της πρώτης παρένθεσης είναι το άθροισμα $n-1$ διαδοχικών όρων αριθμητικής προόδου με πρώτο όρο ίσο με 1 και τελευταίο όρο ίσο με $n-1$ και χρησιμοποιήθηκε η σχέση (Β΄.1) του Παραρτήματος Β΄ για τον υπολογισμό του, ενώ επιπρόσθετα λάβαμε υπόψη ότι:

$$\sum_{i=1}^{n} i^2 = \frac{n(n+1)(2n+1)}{6}.$$

Επομένως,

$$\begin{aligned} Var(X) &= a^2 + a(n-1) + \frac{(n-1)(2n-1)}{6} - \frac{(2a+(n-1))^2}{4} \\ &= a^2 + a(n-1) + \frac{(n-1)(2n-1)}{6} - a^2 - a(n-1) - \frac{(n-1)^2}{4} \\ &= \frac{(n-1)(2n-1)}{6} - \frac{(n-1)^2}{4} \\ &= \frac{(n-1)(n+1)}{12}, \end{aligned}$$

και η απόδειξη ολοκληρώθηκε.

Παρατήρηση 4.1

Έστω $X \sim DUnif(a,b)$, $a < b$ με $x \in \{a, a+1, \ldots, b\}$. Στη γλώσσα προγραμματισμού R μπορούμε:

- με τη συνάρτηση `ddunif(x,a,b)` να υπολογίσουμε τη σπ στο σημείο ή στο διάνυσμα σημείων x,
- με τη συνάρτηση `pdunif(x,a,b,lower.tail=TRUE)` να υπολογίσουμε την ασκ στο σημείο ή στο διάνυσμα σημείων x,
- με τη συνάρτηση `pdunif(x,a,b,lower.tail=FALSE)` να υπολογίσουμε την πιθανότητα $P(X > x)$ στο σημείο ή στο διάνυσμα σημείων x,
- με τη συνάρτηση `qdunif(q,a,b,lower.tail=TRUE)` να προσδιορίσουμε το σημείο ή το διάνυσμα σημείων x για τα οποία ισχύει ότι $P(X \leq x) \geq q$, όπου q είναι πιθανότητα ή διάνυσμα με συνιστώσες πιθανότητες, αντίστοιχα,
- με τη συνάρτηση `qdunif(q,a,b,lower.tail=FALSE)` να προσδιορίσουμε το σημείο ή το διάνυσμα σημείων x για τα οποία ισχύει ότι $P(X > x) \geq q$, όπου q είναι πιθανότητα ή διάνυσμα με συνιστώσες πιθανότητες, αντίστοιχα,
- με τη συνάρτηση `rdunif(n,a,b)` να παράγουμε ένα δείγμα μεγέθους n από τη διακριτή ομοιόμορφη κατανομή.

Άσκηση Αυτοαξιολόγησης 4.1

Ένα άτομο επιλέγει τυχαία έναν αριθμό από τους $\{1,2,3,\ldots,20\}$. Υπολογίστε την πιθανότητα ο αριθμός που επέλεξε να είναι μεγαλύτερος από 12. Διαφοροποιείται αυτή η πιθανότητα όταν γνωρίζετε ότι έχει επιλέξει αριθμό μικρότερο από 18 και, μάλιστα, περιττό;

Άσκηση Αυτοαξιολόγησης 4.2

Έστω ότι η τ.μ. X ακολουθεί τη διακριτή ομοιόμορφη κατανομή με τιμές τους n το πλήθος διαδοχικούς ακέραιους a, $a+1$,..., $b-1$, b, δηλαδή $X \sim DUnif(a,b)$, $a < b$ με $b = a + n - 1$. Να δείξετε ότι

$$M_X(t) = \frac{e^{at} - e^{(a+n)t}}{n(1-e^t)} = \frac{e^{at} - e^{(b+1)t}}{n(1-e^t)}. \tag{4.4}$$

4.3 Διωνυμική κατανομή

Η διωνυμική κατανομή είναι μία από τις παλαιότερες κατανομές πιθανότητας. Παρότι ίσως είχε παρουσιαστεί στη βιβλιογραφία νωρίτερα για συγκεκριμένες, ειδικές τιμές των παραμέτρων της (μεταξύ άλλων και από τον Γάλλο Blaise Pascal, 1623 – 1662) αποδίδεται ιστορικά ότι παρουσιάστηκε από τον Ελβετό θεολόγο, μαθηματικό και αστρονόμο Jacob Bernoulli (γνωστός επίσης και ως James ή Jackques, 1655-1705) το 1713 (Bernoulli, 1713). Αυτό το έργο του, που ήταν ατελές όταν πέθανε και δημοσιεύτηκε 8 χρόνια μετά τον θάνατό του, συνδέει ίσως για πρώτη φορά την εφαρμογή των πιθανοτήτων όχι μόνο με τα τυχερά παιχνίδια αλλά και με άλλα επιστημονικά πεδία.

Πριν προχωρήσουμε στη μελέτη της διωνυμικής κατανομής είναι σημαντικό να κατανοήσουμε τις λεγόμενες δοκιμές Bernoulli, οι οποίες εμφανίζονται σε πλήθος τυχαίων φαινομένων από διάφορα επιστημονικά πεδία και επιτρέπουν τη μελέτη τους μέσα από μια κοινή πιθανοθεωρητική δομή. Ο όρος **δοκιμή Bernoulli** χρησιμοποιείται για να δηλώσει ένα πείραμα τύχης που έχει δύο δυνατά αποτελέσματα. Ενδεικτικά παραδείγματα τέτοιων πειραμάτων τύχης είναι τα ακόλουθα:

- η ρίψη ενός νομίσματος (εμφάνιση κορώνας ή γραμμάτων),
- το αποτέλεσμα του μοριακού ελέγχου ενός ατόμου για Covid-19 (νοσεί ή δεν νοσεί),
- η αντοχή μίας δοκού σε ένα φορτίο (αντέχει ή δεν αντέχει),
- το φύλο κατά τη γέννηση ενός παιδιού (αγόρι ή κορίτσι),
- η επιλογή ενός ατόμου από τον γενικό πληθυσμό και ο έλεγχος αν αυτό έχει χοληστερόλη πάνω από μια συγκεκριμένη τιμή ή όχι (ναι ή όχι) και
- αν θα έχει κυκλοφοριακή συμφόρηση ή όχι στην περιφερειακή οδό (έχει ή δεν έχει).

Ο κατάλογος των παραδειγμάτων που έχουν δύο δυνατά αποτελέσματα είναι ανεξάντλητος, καθώς τέτοια εμφανίζονται τόσο στην καθημερινή μας ζωή όσο και σε όλα τα επιστημονικά πεδία. Θέλοντας λοιπόν να μελετήσουμε τέτοια τυχαία φαινόμενα υπό μία ενιαία πιθανοθεωρητική δομή, μπορούμε τα δύο δυνατά αποτελέσματα να τα ονομάσουμε συμβατικά Επιτυχία και Αποτυχία. Ο λόγος που χρησιμοποιείται η λέξη Επιτυχία συμβατικά γίνεται άμεσα αντιληπτός, καθώς η αντιστοίχιση του ενός εκ των δύο δυνατών αποτελεσμάτων στην Επιτυχία είναι αυθαίρετος. Για παράδειγμα, στο φύλο του παιδιού κατά τη γέννηση δεν υπάρχει κανένας λόγος να θεωρηθεί επιτυχία η γέννηση αγοριού ή η γέννηση κοριτσιού και για αυτόν τον λόγο επιλέγεται αυθαίρετα ένα εκ των δύο αποτελεσμάτων για να οριστεί ως Επιτυχία. Σε πολλές περιπτώσεις, ως επιτυχία δηλώνεται η εμφάνιση του αποτελέσματος που θέλουμε να μελετήσουμε ακόμα και αν αυτή δεν έχει θετικό αντίκτυπο. Χαρακτηριστικά τέτοια παραδείγματα, είναι να θεωρείται Επιτυχία ένα θετικό αποτέλεσμα μοριακού ελέγχου για Covid-19, ένα ελαττωματικό εξάρτημα ή ακόμα και ο θάνατος από μία συγκεκριμένη ασθένεια.

Στο παραπάνω πιθανοθεωρητικό πλαίσιο, πολλές φορές ενδιαφερόμαστε για τον αριθμό των επιτυχιών (θα συμβολίζεται με E από εδώ και στο εξής) σε n το πλήθος προκαθορισμένες επαναλήψεις μιας δοκιμής Bernoulli (ή διαφορετικά σε μια ακολουθία δοκιμών Bernoulli). Για αυτές τις επαναλήψεις ισχύουν τα ακόλουθα:

- Το αποτέλεσμα σε οποιαδήποτε από τις n επαναλήψεις δεν επηρεάζει το αποτέλεσμα οποιασδήποτε άλλης.
- Οι συνθήκες εκτέλεσης του τυχαίου πειράματος παραμένουν αμετάβλητες σε κάθε επανάληψη και για αυτόν τον λόγο η πιθανότητα επιτυχίας παραμένει ίδια από επανάληψη σε επανάληψη. Η πιθανότητα επιτυχίας συμβολίζεται με p, δηλαδή $p = P(E)$, $0 < p < 1$, οπότε η πιθανότητα αποτυχίας (θα συμβολίζεται με A από εδώ και στο εξής) είναι $P(A) = 1 - p = q$.

Ένα τυχαίο πείραμα που αποτελείται από n το πλήθος προκαθορισμένες επαναλήψεις μιας δοκιμής Bernoulli για τις οποίες ικανοποιούνται οι παραπάνω προϋποθέσεις ονομάζεται **διωνυμικό τυχαίο πείραμα**. Η διωνυμική κατανομή είναι αυτή που μοντελοποιεί την τ.μ., έστω X, που καταγράφει το πλήθος των επιτυχιών στις n ανεξάρτητες επαναλήψεις Bernoulli, δηλαδή σε ένα διωνυμικό τυχαίο πείραμα. Ο ορισμός της διωνυμικής κατανομής ακολουθεί.

Ορισμός 4.2

Η τυχαία μεταβλητή X λέγεται ότι ακολουθεί τη **διωνυμική κατανομή** με παραμέτρους n και p με $p \in (0,1)$, αν οι δυνατές της τιμές x είναι $x \in \{0,1,\dots,n\}$ και η συνάρτηση πιθανότητάς της δίνεται από τη σχέση:

$$p_X(x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x}, & x = 0,1,\dots,n \\ 0, & \text{αλλού.} \end{cases} \tag{4.5}$$

Στην περίπτωση αυτή, θα γράφουμε $X \sim B(n,p)$.

Παρατηρήστε ότι στη σπ εμφανίζεται ο διωνυμικός συντελεστής, η παρουσία του οποίου αιτιολογεί και την ονομασία της κατανομής.

Δύο εύλογα ερωτήματα που προκύπτουν μετά τον ορισμό της διωνυμικής κατανομής είναι αν όντως η σχέση (4.5) αποτελεί συνάρτηση πιθανότητας και αν μπορεί να χρησιμοποιηθεί για το υπό μελέτη τυχαίο φαινόμενο. Όσον αφορά το πρώτο ερώτημα, άμεσα προκύπτει ότι η συνάρτηση που εμφανίζεται στη σχέση (4.5) είναι μη αρνητική. Επιπλέον, σύμφωνα με το **διωνυμικό θεώρημα**[1] ισχύει ότι:

$$(a+b)^n = \sum_{x=0}^{n} \binom{n}{x} a^x \, b^{n-x}. \tag{4.6}$$

Εφαρμόζοντας τη σχέση (4.6) για $a = p$ και $b = 1 - p$ έχουμε ότι $\sum_{x=0}^{n} \binom{n}{x} p^x \, (1-p)^{n-x} = 1$, που επιβεβαιώνει ότι όντως η συνάρτηση που εμφανίζεται στη σχέση (4.5) είναι σπ.

Σχετικά με το δεύτερο ερώτημα, δηλαδή αν η παραπάνω συνάρτηση πιθανότητας μπορεί να χρησιμοποιηθεί για το υπό μελέτη τυχαίο φαινόμενο, έχουμε, ενθυμούμενοι τον ορισμό της συνάρτησης πιθανότητας, ότι η $p_X(x) = P(X = x)$ ισοδυναμεί ουσιαστικά με τον υπολογισμό της πιθανότητας

$$P(\text{στις } n \text{ δοκιμές έχουν εμφανιστεί ακριβώς } x \text{ το πλήθος } E \text{ και } n-x \text{ το πλήθος } A).$$

Είναι προφανές ότι δεν υπάρχει μοναδικός τρόπος επίτευξης x επιτυχιών στις n το πλήθος επαναλήψεις δοκιμών Bernoulli. Για παράδειγμα, αν το $n = 4$ και το $x = 2$, τότε αυτό μπορεί να επιτευχθεί με τους ακόλουθους έξι διαφορετικούς τρόπους: $EEAA$, $EAAE$, $EAEA$, $AEEA$ $AEAE$ $AAEE$. Εύκολα προκύπτει, ότι ο αριθμός των διαφορετικών n–άδων, στις οποίες εμφανίζονται ακριβώς x το πλήθος επιτυχίες (άρα και $n - x$ αποτυχίες), είναι ίσος με τον αριθμό των συνδυασμών n αντικειμένων ανά x. Επομένως, είναι ίσος με $\binom{n}{x}$, $x = 0, 1, \ldots, n$. Έστω τώρα B_i, $i = 1, 2, \ldots, \binom{n}{x}$, είναι καθένα από αυτά τα $\binom{n}{x}$ το πλήθος ενδεχόμενα στα οποία εμφανίζονται ακριβώς x επιτυχίες στις n επαναλήψεις του διωνυμικού τυχαίου πειράματος. Τότε, λαμβάνοντας υπόψη ότι $B_i \cap B_j = \varnothing, \forall\, i \neq j$, $i, j = 1, \ldots, \binom{n}{x}$, έχουμε ότι:

$$p_X(x) = P\left(\bigcup_{i=1}^{\binom{n}{x}} B_i\right) = \sum_{i=1}^{\binom{n}{x}} P(B_i).$$

Καθώς τώρα οι επαναλήψεις είναι ανεξάρτητες και η πιθανότητα επιτυχίας αμετάβλητη, είναι $P(B_i) = p^x \, (1-p)^{n-x}$ και προκύπτει άμεσα ότι:

$$p_X(x) = \sum_{i=1}^{\binom{n}{x}} \left(p^x \, (1-p)^{n-x}\right) = \binom{n}{x} p^x \, (1-p)^{n-x},$$

που αποδεικνύει το ζητούμενο.

[1] Για περισσότερες πληροφορίες και τρόπους απόδειξης του διωνυμικού θεωρήματος παραπέμπουμε στον ιστότοπο https://en.wikipedia.org/wiki/Binomial_theorem (ημερομηνία προσπέλασης: 1/3/2022).

Παρατήρηση 4.2

Η ειδική περίπτωση που προκύπτει για $n = 1$ είναι γνωστή ως **κατανομή Bernoulli** και έχει σπ $P(X = x) = p^x\ (1-p)^{1-x}$, $x = 0,1$. Ουσιαστικά, η κατανομή Bernoulli μοντελοποιεί πιθανοθεωρητικά την έκβαση μιας δοκιμής Bernoulli. Παρατηρήστε ότι $X = X_1 + ... + X_n$, όπου X_i η i-οστή Bernoulli τ.μ. που περιγράφει το αποτέλεσμα της i-οστής επανάληψης.

Αφού προσδιορίστηκε η σπ της τ.μ. X (ή μετά τον ορισμό της, από όποια οπτική και αν το δούμε) θα προσδιοριστεί η αθροιστική συνάρτηση κατανομής της και θα μελετηθούν κάποιες βασικές ιδιότητές της. Σε αυτό το πλαίσιο, χρησιμοποιώντας τη σχέση (4.5) και τον ορισμό της ασκ, έχουμε ότι, αν $X \sim B(n,p)$, $0 < p < 1$, τότε:

$$F_X(x) = \begin{cases} 0, & x < 0 \\ \sum_{y=0}^{\lfloor x \rfloor} \binom{n}{y} p^y\ (1-p)^{n-y}, & 0 \le x < n \\ 1, & x \ge n \end{cases} \tag{4.7}$$

όπου το $\lfloor x \rfloor$ συμβολίζει το ακέραιο μέρος του x.

Στη συνέχεια, παρουσιάζονται κάποιες χρήσιμες ιδιότητες και χαρακτηριστικά της τ.μ. $X \sim B(n,p)$. Προτού όμως προχωρήσουμε θα παραθέσουμε ένα χρήσιμο λήμμα.

Λήμμα 4.1

Για οποιουσδήποτε ακεραίους n, m με $0 < m \le n$ ισχύει ότι:

$$\binom{n}{m} = \frac{n}{m}\binom{n-1}{m-1}, \text{ για } m \neq 0. \tag{4.8}$$

Απόδειξη Λήμματος 4.1

Προκύπτει άμεσα από τον ορισμό του διωνυμικού συντελεστή ως εξής:

$$\binom{n}{m} = \frac{n!}{m!(n-m)!} = \frac{n}{m} \cdot \frac{(n-1)!}{(m-1)!((n-1)-(m-1))!} = \frac{n}{m}\binom{n-1}{m-1}.$$

Πρόταση 4.2

Έστω X η τυχαία μεταβλητή που ακολουθεί διωνυμική κατανομή με παραμέτρους n και $p \in (0,1)$ με σπ που προσδιορίζεται στη σχέση (4.5). Τότε ισχύει ότι:

$$E(X) = np, \tag{4.9}$$

και

$$Var(X) = np(1-p). \tag{4.10}$$

Απόδειξη Πρότασης 4.2

Υπάρχουν διάφοροι τρόποι προσδιορισμού των $E(X)$ και $Var(X)$ και ακολούθως θα παρουσιαστεί αυτός που στηρίζεται στον ορισμό τους. Από τον ορισμό της μέσης τιμής και με αλγεβρικές πράξεις έχουμε

$$\begin{aligned} E(X) &= \sum_{x=0}^{n} x\binom{n}{x} p^x\,(1-p)^{n-x} = \sum_{x=1}^{n} x\binom{n}{x} p^x\,(1-p)^{n-x} \\ &\overset{(4.8)}{=} \sum_{x=1}^{n} x\frac{n}{x}\binom{n-1}{x-1} p^x\,(1-p)^{n-x} = \sum_{x=1}^{n} n\binom{n-1}{x-1} p^x\,(1-p)^{n-x} \\ &\overset{y=x-1}{=} np\sum_{y=0}^{n-1}\binom{n-1}{y} p^y\,(1-p)^{n-1-y} \overset{(4.6)}{=} n\,p\,[p+(1-p)]^{n-1} = np, \end{aligned}$$

όπου το πρώτο άθροισμα το γράψαμε να ξεκινάει από τη μονάδα καθώς το μηδέν δεν συνεισφέρει τίποτα. Καθώς $Var(X) = E(X^2) - (E(X))^2$ θα υπολογιστεί αρχικά η $E(X^2)$ ως ακολούθως

$$\begin{aligned} E(X^2) &= \sum_{x=0}^{n} x^2\binom{n}{x} p^x\,(1-p)^{n-x} = \sum_{x=1}^{n} x^2\binom{n}{x} p^x\,(1-p)^{n-x} \\ &\overset{(4.8)}{=} \sum_{x=1}^{n} x^2\frac{n}{x}\binom{n-1}{x-1} p^x\,(1-p)^{n-x} = n\sum_{x=1}^{n} x\binom{n-1}{x-1} p^x\,(1-p)^{n-x} \\ &\overset{y=x-1}{=} np\sum_{y=0}^{n-1}(y+1)\binom{n-1}{y} p^y\,(1-p)^{n-1-y} \\ &= np\sum_{y=0}^{n-1} y\binom{n-1}{y} p^y\,(1-p)^{n-1-y} + np\sum_{y=0}^{n-1}\binom{n-1}{y} p^y\,(1-p)^{n-1-y} \\ &= np(n-1)p + np, \end{aligned}$$

όπου

$$\sum_{y=0}^{n-1} y\binom{n-1}{y} p^y\,(1-p)^{n-1-y} = E(Y), \text{ με } Y \sim B(n-1,p).$$

Άρα ισχύει ότι

$$\sum_{y=0}^{n-1} y\binom{n-1}{y} p^y\,(1-p)^{n-1-y} = (n-1)p$$

ενώ από την ιδιότητα της σπ της Y έχουμε ότι $\sum_{y=0}^{n-1}\binom{n-1}{y} p^y\,(1-p)^{n-1-y} = 1$. Το επιθυμητό αποτέλεσμα προκύπτει άμεσα, καθώς $Var(X) = E(X^2) - (E(X))^2$.

Πρόταση 4.3

Έστω X η τυχαία μεταβλητή που ακολουθεί διωνυμική κατανομή με παραμέτρους n και $p \in (0,1)$ με σπ που προσδιορίζεται στη σχέση (4.5). Τότε ισχύει ότι:

$$M_X(t) = \left(q + pe^t\right)^n. \tag{4.11}$$

Απόδειξη Πρότασης 4.3

Λαμβάνοντας υπόψη τη σχέση (3.27) έχουμε ότι:

$$\begin{aligned} M_X(t) = E\left(e^{tX}\right) &= \sum_{x=0}^{n} e^{tx} \binom{n}{x} p^x \, (1-p)^{n-x} \\ &= \sum_{x=0}^{n} \binom{n}{x} \left(pe^t\right)^x \, (1-p)^{n-x} \\ &= \left((1-p) + pe^t\right)^n \end{aligned}$$

όπου χρησιμοποιήθηκε η σχέση (4.6) του διωνυμικού θεωρήματος για $a = pe^t$ και $b = 1-p$, και έτσι αποδείχθηκε το ζητούμενο της πρότασης.

Παρατήρηση 4.3

Από την προηγούμενη πρόταση προκύπτει ένας ακόμη, ίσως πιο εύκολος, τρόπος προσδιορισμού της μέσης τιμής και της διακύμανσης της διωνυμικής κατανομής. Ειδικότερα, ενθυμούμενοι ότι $E(X^k) = \frac{d}{dt^k} M_X^{(k)}(t)|_{t=0}$, προκύπτει ότι:

$$\frac{d}{dt} M_X(t) = n\left((1-p) + pe^t\right)^{n-1} pe^t,$$

και για $t = 0$ έχουμε ότι $E(X) = np$, ενώ

$$\frac{d}{dt^2} M_X^{(2)}(t) = n(n-1)\left((1-p) + pe^t\right)^{n-1} (pe^t)^2 + n\left((1-p) + pe^t\right)^{n-1} p$$

και για $t = 0$ έχουμε ότι $E(X^2) = n(n-1)p^2 + np$. Επομένως, το επιθυμητό αποτέλεσμα προκύπτει άμεσα, καθώς $Var(X) = E(X^2) - (E(X))^2$.

Κάποιες φορές ο υπολογισμός πιθανοτήτων χρησιμοποιώντας τη σπ ή την ασκ της διωνυμικής κατανομής είναι χρονοβόρος. Για τον λόγο αυτό στη βιβλιογραφία υπάρχουν οι λεγόμενοι **διωνυμικοί πίνακες** (βλ. για τη δημιουργία τους Dodge, 2008, και τις εκεί αναφορές). Οι διωνυμικοί πίνακες μας δίνουν απευθείας την τιμή είτε της σπ είτε της ασκ για διάφορες τιμές του n (συνήθως $n \leq 30$) και για τιμές του $p = 0.05, 0.10, \ldots, 0.5$. Απόσπασμα τέτοιων πινάκων δίνεται στο Παράρτημα Α΄ (βλ. Πίνακα Α΄.1). Στους διωνυμικούς πίνακες, ο αριθμός των επαναλήψεων n δίνεται στην πρώτη στήλη, ο αριθμός των επιτυχιών x δίνεται στη δεύτερη στήλη, ενώ ακολούθως επιλέγεται η στήλη στην οποία βρίσκεται η πιθανότητα επιτυχίας p. Αφού εντοπιστεί η γραμμή για το ζητούμενο n και x, τότε η τιμή της πιθανότητας ή η τιμή της ασκ (ανάλογως με το ποιας μορφής πίνακας χρησιμοποιείται) εντοπίζεται στη διασταύρωση της γραμμής αυτής και της στήλης του p. Ένα πρώτο εύλογο ερώτημα που δημιουργείται είναι το τι κάνουμε στην περίπτωση που $p > 0.5$. Η απάντηση δίνεται στην παρατήρηση που ακολουθεί.

Παρατήρηση 4.4

Έστω X και Y οι τ.μ. που παριστάνουν τον αριθμό των επιτυχιών και αποτυχιών, αντίστοιχα, στις n επαναλήψεις ενός διωνυμικού τυχαίου πειράματος. Τότε $X \sim B(n,p)$ και $Y \sim B(n,1-p)$ με $p_X(x) = p_Y(n-x)$. Επομένως, για την εύρεση της $p_X(x)$, όταν $X \sim B(n,p)$ με $p > 0.5$, μπορούν να χρησιμοποιηθούν οι διωνυμικοί πίνακες που είναι διαθέσιμοι για πιθανότητες επιτυχίας μικρότερες ή ίσες από 0.5, υπολογίζοντας τις αντίστοιχες πιθανότητες σε όρους της $Y = n - X$ με την εξής τροποποίηση: επιλέγουμε το n στην πρώτη στήλη, το $n-x$ στη δεύτερη και το $q = 1-p < 0.5$ στη συνέχεια.

Ένα δεύτερο ερώτημα είναι πώς χειριζόμαστε περιπτώσεις πιθανοτήτων επιτυχίας ή/και αριθμού επαναλήψεων, οι οποίες δεν περιέχονται στον διαθέσιμο πίνακα ή ακόμη και περιπτώσεις που απαιτούν πολλούς υπολογισμούς με τον συνήθη τρόπο. Ένας τρόπος αντιμετώπισης του παραπάνω προβλήματος είναι είτε η προσέγγιση Poisson είτε η κανονική προσέγγιση της διωνυμικής κατανομής στις οποίες θα αναφερθούμε στις Ενότητες 4.7 και 7.3.1, αντίστοιχα. Προφανώς, ένας εναλλακτικός και συχνά προτιμότερος τρόπος είναι η χρήση κάποιου προγράμματος του υπολογιστή για τον υπολογισμό των πιθανοτήτων. Σε αυτήν την κατεύθυνση είναι πολύ χρήσιμη η παρατήρηση που ακολουθεί.

Παρατήρηση 4.5

Έστω $X \sim B(n,p)$, $0 < p < 1$, $x \in \{0,1,\dots,n\}$ με σπ που δίνεται από τη σχέση (4.5). Στη γλώσσα προγραμματισμού R μπορούμε:

- με τη συνάρτηση `dbinom(x,n,p)` να υπολογίσουμε τη σπ στο σημείο ή στο διάνυσμα σημείων x,
- με τη συνάρτηση `pbinom(x,n,p,lower.tail=TRUE)` να υπολογίσουμε την ασκ στο σημείο ή στο διάνυσμα σημείων x,
- με τη συνάρτηση `pbinom(x,n,p,lower.tail=FALSE)` να υπολογίσουμε την πιθανότητα $P(X > x)$ στο σημείο ή στο διάνυσμα σημείων x,
- με τη συνάρτηση `qbinom(c,n,p,lower.tail=TRUE)` να προσδιορίσουμε τη μικρότερη τιμή ή το διάνυσμα των μικρότερων τιμών x για τις οποίες ισχύει ότι $P(X \le x) \ge c$, όπου c είναι πιθανότητα ή διάνυσμα με συνιστώσες πιθανότητες, αντίστοιχα,
- με τη συνάρτηση `qbinom(c,n,p,lower.tail=FALSE)` να προσδιορίσουμε τη μικρότερη τιμή ή το διάνυσμα των μικρότερων τιμών x για τις οποίες ισχύει ότι $P(X > x) \ge c$, όπου c είναι πιθανότητα ή διάνυσμα με συνιστώσες πιθανότητες,
- με τη συνάρτηση `rbinom(k,n,p)` να παράγουμε ένα δείγμα μεγέθους k από την κατανομή με σπ (4.5).

Η χρήση κάποιων εκ των παραπάνω εντολών παρουσιάζεται στα παραδείγματα που ακολουθούν.

Παράδειγμα 4.1

Η πιθανότητα ένα είδος δοκού να μην αντέξει ένα φορτίο συγκεκριμένου βάρους είναι 0.25. Υπολογίστε τις πιθανότητες μεταξύ 15 τέτοιων δοκών στις οποίες ασκείται υπό τις ίδιες συνθήκες το συγκεκριμένο βάρος να αντέξουν (α) ακριβώς 6 δοκοί και (β) τουλάχιστον 12 δοκοί.

Λύση Παραδείγματος 4.1

Έστω X η τ.μ. που παριστάνει τον αριθμό των δοκών που δεν αντέχουν φορτίο συγκεκριμένου βάρους στις 15 που επιλέγονται τυχαία και στις οποίες ασκείται υπό τις ίδιες συνθήκες το συγκεκριμένο βάρος. Η τ.μ. X εκφράζει τον αριθμό των επιτυχιών (δοκών που δεν αντέχουν το φορτίο) σε 15 τυχαία επιλεγμένες, δηλαδή ανεξάρτητες, δοκούς που δεν αντέχουν, με την ίδια πιθανότητα, το φορτίο. Με βάση τα παραπάνω, εύκολα συμπεραίνουμε ότι $X \sim B(n=15, p=0.25)$, αφού έχουμε ένα διωνυμικό πείραμα.

Σημειώνουμε ότι ο λόγος που ορίστηκε ως επιτυχία η μη αντοχή στο φορτίο είναι για να έχουμε πιθανότητα επιτυχίας $p < 0.5$ και να μπορεί να χρησιμοποιηθεί ο πίνακας που δίνεται στο Παράρτημα Α΄. Κατά αυτόν τον τρόπο, η πιθανότητα μεταξύ 15 τέτοιων δοκών να αντέξουν ακριβώς 6 ισοδυναμεί με την πιθανότητα να μην αντέξουν ακριβώς 9 και η πιθανότητα να αντέξουν τουλάχιστον 12 ισοδυναμεί με την πιθανότητα να μην αντέξουν το πολύ 3.
(α) Με βάση τα παραπάνω, η ζητούμενη πιθανότητα ισούται με

$$P(X = 9) = \binom{15}{9} \cdot 0.25^9 \cdot 0.75^6 = \frac{15!}{9! \cdot 6!} 0.25^9 \cdot 0.75^6 = 5005 \cdot 0.25^9 \cdot 0.75^6 = 0.0034.$$

Στο ίδιο αποτέλεσμα μπορούμε να καταλήξουμε χρησιμοποιώντας και τον Πίνακα Α΄.1 του Παραρτήματος Α΄, αφού

$$P(X = 9) = P(X \leq 9) - P(X \leq 8) = 0.9992 - 0.9958 = 0.0034$$

(β) Η πιθανότητα $P(X \leq 3)$ υπολογίζεται προσδιορίζοντας αρχικά τα $P(X = i), i = 0, 1, 2, 3$, ως ακολούθως

$$\begin{aligned}
P(X = 0) &= \binom{15}{0} \cdot 0.25^0 \cdot 0.75^{15} = \frac{15!}{0! \cdot 15!} 0.25^0 \cdot 0.75^{15} = 0.0134, \\
P(X = 1) &= \binom{15}{1} \cdot 0.25^1 \cdot 0.75^{14} = \frac{15!}{1! \cdot 14!} 0.25^1 \cdot 0.75^{14} \\
&= 15 \cdot 0.25^1 \cdot 0.75^{14} = 0.0668, \\
P(X = 2) &= \binom{15}{2} \cdot 0.25^2 \cdot 0.75^{13} = \frac{15!}{2! \cdot 13!} 0.25^2 \cdot 0.75^{13} \\
&= \frac{14 \cdot 15}{2} 0.25^2 \cdot 0.75^{13} = 105 \cdot 0.25^2 \cdot 0.75^{13} = 0.1559,
\end{aligned}$$

και

$$\begin{aligned}
P(X = 3) &= \binom{15}{3} \cdot 0.25^3 \cdot 0.75^{12} = \frac{15!}{3! \cdot 12!} 0.25^3 \cdot 0.75^{12} \\
&= \frac{13 \cdot 14 \cdot 15}{6} 0.25^3 \cdot 0.75^{12} = 455 \cdot 0.25^3 \cdot 0.75^{12} = 0.2252.
\end{aligned}$$

Άρα $P(X \leq 3) = 0.0134 + 0.0668 + 0.1559 + 0.2252 = 0.4613$.
Εναλλακτικά, ο υπολογισμός της $P(X \leq 3)$ μπορεί να γίνει απευθείας από τον Πίνακα Α΄.1 του Παραρτήματος Α΄, από τον οποίο προκύπτει άμεσα ότι $P(X \leq 3) = 0.4613$.
Οι παραπάνω πιθανότητες μπορούν να υπολογιστούν στην R εκτελώντας τις εντολές `dbinom(9,15,0.25)`, και `pbinom(3,15,0.25,lower.tail=TRUE)`, αντίστοιχα. Οι εντολές αυτές επιστρέφουν τις τιμές 0.003398065 και 0.4612869, οι οποίες είναι ίσες πρακτικά (εκτός από τα σφάλματα στρογγυλοποίησης στους υπολογισμούς) με τα αποτελέσματα που λάβαμε νωρίτερα.

Παράδειγμα 4.2

Μια φαρμακοβιομηχανία υποστηρίζει ότι η πιθανότητα να έχει παρενέργειες ένα φάρμακο καταπολέμησης της υπέρτασης που παρασκευάζει είναι 0.1. Χορηγείται το φάρμακο σε 20 τυχαία επιλεγμένα άτομα. Υπολογίστε την πιθανότητα το πολύ 2 άτομα να έχουν παρενέργειες. Ποιος είναι ο αναμενόμενος αριθμός των ατόμων που θα έχει παρενέργειες στα 20 τυχαία επιλεγμένα άτομα; Αναφέρετε όλες τις απαραίτητες υποθέσεις για την επίλυση του προβλήματος.

Λύση Παραδείγματος 4.2

Έστω X η τ.μ. που παριστάνει τον αριθμό των ατόμων στα οποία εμφανίστηκαν παρενέργειες μετά τη χορήγηση του φαρμάκου στα 20 τυχαία επιλεγμένα άτομα. Θεωρούμε ότι η πιθανότητα ανάπτυξης παρενεργειών είναι σταθερή σε κάθε άτομο και ίση με $p = 0.1$. Επιπλέον, η εμφάνιση παρενεργειών σε κάποιο άτομο είναι ανεξάρτητη από την εμφάνιση παρενεργειών σε κάποιο άλλο άτομο. Υπό αυτές τις υποθέσεις έχουμε ότι $X \sim B(20, 0.1)$. Ζητείται η $P(X \leq 2)$. Είναι:

$$
\begin{aligned}
P(X=0) &= \binom{20}{0} \cdot 0.1^0 \cdot 0.9^{20} = \frac{20!}{0! \cdot 20!} 0.1^0 \cdot 0.9^{20} = 0.1216, \\
P(X=1) &= \binom{20}{1} \cdot 0.1^1 \cdot 0.9^{19} = \frac{20!}{1! \cdot 19!} 0.1^1 \cdot 0.9^{19} \\
&= 20 \cdot 0.1^1 \cdot 0.9^{19} = 0.2702, \\
P(X=2) &= \binom{20}{2} \cdot 0.1^2 \cdot 0.9^{18} = \frac{20!}{2! \cdot 18!} 0.1^2 \cdot 0.9^{18} \\
&= \frac{19 \cdot 20}{2} 0.1^2 \cdot 0.9^{18} = 0.2852.
\end{aligned}
$$

Άρα $P(X \leq 2) = 0.1216 + 0.2702 + 0.2852 = 0.6770$.

Εναλλακτικά, η παραπάνω πιθανότητα υπολογίζεται με την εντολή `pbinom(2,20,0.1,lower.tail=TRUE)`, η οποία επιστρέφει την τιμή 0.6769268, η οποία ισούται πρακτικά με το αποτέλεσμα που λάβαμε παραπάνω (αγνοώντας τα σφάλματα στογγυλοποίησης).

Τέλος, ο αναμενόμενος αριθμός των ατόμων που θα έχει παρενέργειες στα 20 τυχαία επιλεγμένα άτομα είναι ίσος με $E(X) = np = 20 \cdot 0.1 = 2$.

Άσκηση Αυτοαξιολόγησης 4.3

Ένα ζάρι ρίχνεται 15 φορές. Υπολογίστε την πιθανότητα να φέρουμε:

1. το πολύ 4 φορές την ένδειξη 3,
2. το πολύ 5 φορές ένδειξη μικρότερη ή ίση με 4.

Άσκηση Αυτοαξιολόγησης 4.4

Από προηγούμενες έρευνες είναι γνωστό ότι το ποσοστό των τροχαίων που οφείλονται σε κατανάλωση αλκοόλ είναι 35%. Προσδιορίστε, κάνοντας τις κατάλληλες υποθέσεις, την πιθανότητα ότι σε 5 ατυχήματα που έγιναν, τουλάχιστον 3 να οφείλονταν σε κατανάλωση αλκοόλ.

Άσκηση Αυτοαξιολόγησης 4.5

Είναι γνωστό ότι ο Αποστόλης είναι πολύ καλός στο παιχνίδι της καλοθοσφαίρισης. Αν υποθέσουμε ότι κατά κανόνα ευστοχεί στο 85% των ελεύθερων βολών που επιχειρεί και ότι το αποτέλεσμα σε μία βολή δεν επηρεάζει το αποτέλεσμα σε οποιαδήποτε άλλη, υπολογίστε την πιθανότητα ο Αποστόλης σε 10 προσπάθειες να ευστοχήσει σε 6 έως και 8 βολές. Ποιος είναι ο μικρότερος αριθμός που με πιθανότητα μεγαλύτερη ή ίση με 80% ο Αποστόλης θα ευστοχήσει σε περισσότερες βολές από αυτόν;

4.4 Γεωμετρική κατανομή

Στην Ενότητα 4.3 παρουσιάστηκε η διωνυμική κατανομή για την πιθανοθεωρητική μοντελοποίηση του αριθμού των επιτυχιών σε έναν προκαθορισμένο αριθμό n το πλήθος επαναλήψεων ενός διωνυμικού

τυχαίου πειράματος. Δηλαδή, το ενδιαφέρον επικεντρώθηκε στον αριθμό των επιτυχιών σε n ανεξάρτητες (με την έννοια ότι το αποτέλεσμα της μιας δεν επηρεάζει το αποτέλεσμα κάποιας άλλης) επαναλήψεις ενός τυχαίου φαινομένου με δύο δυνατά αποτελέσματα, που συμβατικά ονομάστηκαν επιτυχία-αποτυχία, με το χαρακτηριστικό ότι η πιθανότητα επιτυχίας παραμένει αμετάβλητη από επανάληψη σε επανάληψη και ίση με p, όπου $0 < p < 1$. Ωστόσο, πολλές φορές, είναι απαραίτητη η πιθανοθεωρητική μοντελοποίηση του αριθμού των ανεξάρτητων επαναλήψεων μιας δοκιμής Bernoulli μέχρι την εμφάνιση της πρώτης επιτυχίας. Επομένως, σε μια τέτοια περίπτωση ο αριθμός των επαναλήψεων δεν είναι πλέον προκαθορισμένος και το ενδιαφέρον επικεντρώνεται στη μελέτη του αριθμού των επαναλήψεων μέχρι την πρώτη επιτυχία ή, ισοδύναμα, του αριθμού των αποτυχιών μέχρι την πρώτη επιτυχία. Για παράδειγμα, υποθέστε ότι γνωρίζουμε ότι η πιθανότητα ένα τυχαίο άτομο του γενικού πληθυσμού να έχει μυωπία είναι ίση με p (ίδια για κάθε άτομο και η ύπαρξη ή όχι μυωπίας σε κάποιο άτομο δεν επηρεάζει την εμφάνιση ή όχι σε κάποιο άλλο) και μας ενδιαφέρει ο αριθμός των ατόμων που θα επιλεχθούν τυχαία από τον γενικό πληθυσμό μέχρις ότου εμφανιστεί το πρώτο άτομο που έχει μυωπία.

Η πιθανοθεωρητική μοντελοποίηση του παραπάνω και παρόμοιων τυχαίων φαινομένων επιτυγχάνεται με τη μελέτη είτε της τ.μ. X που παριστάνει τον αριθμό των ανεξάρτητων δοκιμών Bernoulli που χρειάζονται μέχρι την πρώτη επιτυχία, με σύνολο δυνατών τιμών $\{1,2,3,\ldots\}$, είτε της τ.μ. Y που παριστάνει τον αριθμό των αποτυχιών σε μια διαδικασία ανεξάρτητων επαναλήψεων δοκιμών Bernoulli μέχρις ότου εμφανιστεί η πρώτη επιτυχία, με σύνολο δυνατών τιμών $\{0,1,2,\ldots\}$. Παρατηρήστε ότι οι δύο παραπάνω τ.μ. συνδέονται με την προφανή σχέση $Y = X - 1$, και επομένως, αρκεί να μελετήσουμε τη μία από τις δύο. Παρότι τα πιθανοθεωρητικά μοντέλα που συνδέονται με τις δύο προαναφερθείσες τ.μ. είναι διαφορετικά, εμφανίζονται στη βιβλιογραφία με το ίδιο όνομα, αυτό της γεωμετρικής κατανομής[2]. Για τον λόγο αυτό, συστήνεται κάθε φορά που αναφερόμαστε σε γεωμετρική κατανομή να ορίζεται σαφώς το σύνολο των δυνατών τιμών της, έτσι ώστε να αποφεύγεται η δημιουργία σύγχυσης.

Ορισμός 4.3

Η τυχαία μεταβλητή X λέγεται ότι ακολουθεί τη **γεωμετρική κατανομή** με παράμετρο $p \in (0,1)$, αν οι δυνατές της τιμές x είναι $x \in \{1,2,3,\ldots\}$ και η συνάρτηση πιθανότητάς της δίνεται από τη σχέση:

$$p_X(x) = p(1-p)^{x-1}, \ x = 1,2,\ldots \ . \tag{4.12}$$

Στην περίπτωση αυτή, θα συμβολίζεται $X \sim Geo(p)$.

Παρατήρηση 4.6

Παρατηρώντας ότι $p_X(1) = p$, $p_X(2) = p(1-p)$, $p_X(3) = p(1-p)^2$,...συμπεραίνουμε ότι ουσιαστικά πρόκειται για μια φθίνουσα γεωμετρική πρόοδο με λόγο $1-p < 1$ και πρώτο όρο ίσο με p. Το γεγονός αυτό αιτιολογεί την ονομασία γεωμετρική κατανομή που της έχει δοθεί.

Δύο εύλογα ερωτήματα που προκύπτουν μετά τον ορισμό της γεωμετρικής κατανομής είναι αν η σχέση (4.12) αποτελεί όντως συνάρτηση πιθανότητας και αν πράγματι μπορεί να χρησιμοποιηθεί για το υπό μελέτη τυχαίο φαινόμενο το οποίο αναφέρθηκε πρωτύτερα. Όσον αφορά το πρώτο ερώτημα, όντως η συνάρτηση που εμφανίζεται στη σχέση (4.12) αποτελεί συνάρτηση πιθανότητας, καθώς είναι μη αρνητική και, επιπλέον,

$$\sum_{x=1}^{+\infty} p(1-p)^{x-1} = p\sum_{x=0}^{+\infty}(1-p)^x = p\frac{1}{1-(1-p)} = 1,$$

όπου για τον υπολογισμό του $\sum_{x=0}^{+\infty}(1-p)^x$ παρατηρήσαμε ότι πρόκειται για το άθροισμα των διαδοχικών άπειρων όρων μιας φθίνουσας γεωμετρικής προόδου με πρώτο όρο ίσο με 1 και λόγο $0 < (1-p) < 1$ και

[2] Κάποιοι συγγραφείς αναφέρονται στην κατανομή της τ.μ. X ως μετατοπισμένη γεωμετρική κατανομή (shifted geometric distribution).

χρησιμοποιήσαμε τη σχέση (Β΄.3) του Παραρτήματος Β΄.

Σχετικά με το δεύτερο ερώτημα, ενθυμούμενοι τον ορισμό της συνάρτησης πιθανότητας, έχουμε ότι $p_X(x) = P(X = x)$, που ισοδυναμεί με τον υπολογισμό της πιθανότητας

$$P(X = x) = P(\overbrace{\text{A A A}\cdots\text{A}}^{x-1}\text{E}),\ x = 1,2,\ldots$$

όπου με Α συμβολίζουμε την εμφάνιση αποτυχίας και με Ε την εμφάνιση επιτυχίας. Υπό τις υποθέσεις ότι οι δοκιμές είναι ανεξάρτητες και η πιθανότητα επιτυχίας αμετάβλητη και ίση με p $(0 < p < 1)$, προκύπτει ότι είναι ισοδύναμη με τον υπολογισμό της

$$P(X = x) = \overbrace{P(A)\cdot P(A)\cdots P(A)}^{x-1}\cdot P(E) = (1-p)^{x-1}p,\ x = 1,2,\ldots$$

Παρατήρηση 4.7

Όπως αναφέρθηκε, η γεωμετρική κατανομή μοντελοποιεί τον αριθμό των επαναλήψεων που απαιτούνται σε μια διαδικασία ανεξάρτητων δοκιμών Bernoulli μέχρι την εμφάνιση της πρώτης επιτυχίας. Επομένως, μπορούμε να πούμε ότι μοντελοποιεί τον χρόνο αναμονής μέχρι την πρώτη επιτυχία σε μια διαδικασία επαναλήψεων ανεξάρτητων δοκιμών Bernoulli. Για τον λόγο αυτό, αναφέρεται συχνά ως μια διακριτή κατανομή του χρόνου αναμονής.

Αφού προσδιορίστηκε η σπ της τ.μ. X θα προσδιοριστεί η ασκ της και θα μελετηθούν κάποιες ιδιότητές της. Σε αυτό το πλαίσιο, χρησιμοποιώντας τη σχέση (4.12) και από τον ορισμό της ασκ έχουμε ότι, αν $X \sim Geo(p)$, $0 < p < 1$, τότε:

$$F_X(x) = \begin{cases} 0, & x < 1, \\ \sum_{y=1}^{\lfloor x \rfloor} p\,(1-p)^{y-1}, & x \geq 1, \end{cases} \tag{4.13}$$

όπου το $\lfloor x \rfloor$ συμβολίζει το ακέραιο μέρος του x. Ισοδύναμα, αν $d = \lfloor x \rfloor$, τότε:

$$F_X(x) = \begin{cases} 0, & x < 1 \\ 1-(1-p)^d, & x \geq 1, d = \lfloor x \rfloor \end{cases} \tag{4.14}$$

Άσκηση Αυτοαξιολόγησης 4.6

Έστω Y η τ.μ. που παριστάνει τον αριθμό των αποτυχιών σε μια διαδικασία ανεξάρτητων επαναλήψεων δοκιμών Bernoulli μέχρις ότου να εμφανιστεί η πρώτη επιτυχία, με σύνολο δυνατών τιμών $y \in \{0,1,2,3,\ldots\}$. Να δείξετε ότι:

$$p_Y(y) = p(1-p)^y,\ y = 0,1,2,\ldots \tag{4.15}$$

με ασκ

$$F_Y(y) = \begin{cases} 0, & y < 0, \\ \sum_{l=0}^{\lfloor y \rfloor} p\,(1-p)^l, & y \geq 0, \end{cases} \tag{4.16}$$

όπου το $\lfloor y \rfloor$ συμβολίζει το ακέραιο μέρος του y.

Στη συνέχεια, παρουσιάζονται κάποιες χρήσιμες ιδιότητες και χαρακτηριστικά της τ.μ. $X \sim Geo(p)$.

Πρόταση 4.4

Έστω X η τυχαία μεταβλητή που ακολουθεί γεωμετρική κατανομή με παράμετρο $p \in (0,1)$ με σπ που προσδιορίζεται στη σχέση (4.12). Τότε ισχύει ότι:

$$E(X) = \frac{1}{p}, \tag{4.17}$$

και

$$Var(X) = \frac{1-p}{p^2}. \tag{4.18}$$

Απόδειξη Πρότασης 4.4

Υπάρχουν διάφοροι τρόποι προσδιορισμού των $E(X)$ και $Var(X)$ της γεωμετρικής κατανομής. Στη συνέχεια, θα παρουσιαστεί αυτός που στηρίζεται στον ορισμό. Από τον ορισμό της μέσης τιμής διακριτών τυχαίων μεταβλητών είναι:

$$E(X) = \sum_{x \in S_X} x p_X(x) = \sum_{x=1}^{+\infty} xp(1-p)^{x-1} = p\sum_{x=1}^{+\infty} x(1-p)^{x-1} = p\frac{d}{dp}\left(-\sum_{x=0}^{+\infty}(1-p)^x\right)$$
$$= p\frac{d}{dp}\left(-\frac{1}{1-(1-p)}\right) = p\frac{1}{p^2} = \frac{1}{p}.$$

Καθώς $Var(X) = E(X^2) - (E(X))^2$ θα πρέπει, αρχικά, να υπολογιστεί η $E(X^2)$, η οποία ισούται με

$$E(X^2) = \sum_{x=1}^{+\infty} x^2 p(1-p)^{x-1} = p\sum_{x=1}^{+\infty} x^2(1-p)^{x-1} = -p\frac{d}{dp}\left(\sum_{x=1}^{+\infty} x(1-p)^x\right).$$

Όμως (βλ. και παραπάνω προσδιορισμό της $E(X)$)

$$\sum_{x=1}^{+\infty} x(1-p)^x = (1-p)\sum_{x=1}^{+\infty} x(1-p)^{x-1} = \frac{1-p}{p^2}$$

και, επομένως,

$$E(X^2) = -p\frac{d}{dp}\left(\frac{1-p}{p^2}\right) = -p\frac{p^2-2p}{p^4} = \frac{2p-p^2}{p^3} = \frac{2-p}{p^2}.$$

Το αποτέλεσμα για τη διακύμανση προκύπτει εύκολα συνδυάζοντας τα προηγούμενα.

Πρόταση 4.5

Έστω Y η τ.μ. που παριστάνει τον αριθμό των αποτυχιών σε μια διαδικασία ανεξάρτητων επαναλήψεων δοκιμών Bernoulli μέχρις ότου να εμφανιστεί η πρώτη επιτυχία, με σύνολο δυνατών τιμών $y \in \{0,1,2,3,...\}$, δηλαδή η Y ακολουθεί γεωμετρική κατανομή με παράμετρο $p \in (0,1)$ με σπ που προσδιορίζεται στη σχέση (4.15). Τότε ισχύει ότι:

$$E(Y) = \frac{1-p}{p}, \tag{4.19}$$

και

$$Var(Y) = \frac{1-p}{p^2}. \tag{4.20}$$

Απόδειξη Πρότασης 4.5

Η τ.μ. Y συνδέεται με τη X μέσω της σχέσης $Y = X - 1$ με X να είναι η τ.μ. που ακολουθεί γεωμετρική κατανομή με παράμετρο $p \in (0,1)$ και με σπ που προσδιορίζεται στη σχέση (4.12). Επομένως, $E(Y) = E(X) - 1$ και $Var(Y) = Var(X)$. Το αποτέλεσμα προκύπτει άμεσα από τις σχέσεις (4.17) και (4.18), αντίστοιχα.

Πρόταση 4.6

Έστω X η τυχαία μεταβλητή που ακολουθεί γεωμετρική κατανομή με παράμετρο $p \in (0,1)$ με σπ που προσδιορίζεται στη σχέση (4.12). Τότε:

$$M_X(t) = \frac{pe^t}{1-(1-p)e^t}, \quad \text{για } t < -\log(1-p). \tag{4.21}$$

Απόδειξη Πρότασης 4.6

Λαμβάνοντας υπόψη τη σχέση (3.27) έχουμε ότι:

$$M_X(t) = E\left(e^{tX}\right) = \sum_{x=1}^{+\infty} e^{tx} p(1-p)^{x-1} = pe^t \sum_{x=1}^{+\infty} \left\{(1-p)e^t\right\}^{x-1} = pe^t \sum_{x=0}^{+\infty} \left\{(1-p)e^t\right\}^{x}.$$

Η τελευταία σειρά είναι, ουσιαστικά, το άθροισμα άπειρων διαδοχικών όρων γεωμετρικής προόδου με πρώτο όρο ίσο με 1 και λόγο $(1-p)e^t$. Αυτό είναι πεπερασμένο αν η γεωμετρική πρόοδος είναι φθίνουσα, το οποίο συμβαίνει αν $(1-p)e^t < 1$, δηλαδή αν $t < -\log(1-p)$ και τότε, σύμφωνα με τη σχέση (Β΄.3) του Παραρτήματος Β΄, προκύπτει ότι:

$$M_X(t) = pe^t \frac{1}{1-(1-p)e^t}, t < -\log(1-p),$$

που αποδεικνύει το ζητούμενο.

Παρατήρηση 4.8

Από την προηγούμενη πρόταση προκύπτει ένας ακόμη, ίσως πιο εύκολος, τρόπος προσδιορισμού της μέσης τιμής και της διακύμανσης της γεωμετρικής κατανομής. Ειδικότερα, ενθυμούμενοι ότι $E(X^k) = \frac{d}{dt^k} M_X^{(k)}(t)|_{t=0}$ προκύπτει ότι:

$$\frac{d}{dt} M_X(t) = \frac{pe^t}{\left(1-(1-p)e^t\right)^2}$$

και για $t = 0$ έχουμε $E(X) = p^{-1}$. Επιπρόσθετα, μετά από λίγες αλγεβρικές πράξεις, έχουμε ότι:

$$\frac{d}{dt^2} M_X^{(2)}(t) = \frac{pe^t\left(1+(1-p)e^t\right)}{\left(1-(1-p)e^t\right)^3}$$

και για $t = 0$ λαμβάνουμε ότι $E(X^2) = \frac{2-p}{p^2}$. Το επιθυμητό αποτέλεσμα προκύπτει άμεσα, καθώς $Var(X) = E(X^2) - (E(X))^2$.

Στην πρόταση που ακολουθεί, δίνεται μια χαρακτηριστική ιδιότητα της γεωμετρικής κατανομής, η οποία είναι γνωστή ως ιδιότητα της αμνησίας ή ιδιότητα της έλλειψης μνήμης ή ιδιότητα της μη γήρανσης και η γεωμετρική κατανομή είναι η μοναδική διακριτή κατανομή που την έχει.

Πρόταση 4.7

Η γεωμετρική κατανομή είναι η μοναδική κατανομή μεταξύ των κατανομών με δυνατό σύνολο τιμών $\{1,2,\ldots\}$ που έχει την **ιδιότητα της αμνησίας ή έλλειψης μνήμης ή μη γήρανσης**, η οποία αποδίδεται από τη σχέση:

$$P(X > t+s|X > t) = P(X > s), \text{ για κάθε } s,t \in \{0,1,2,\ldots\}. \tag{4.22}$$

Απόδειξη Πρότασης 4.7

Αρχικά, υποθέτουμε ότι η τ.μ. $X \sim Geo(p)$ με $p > 0$, και θα δείξουμε ότι ικανοποιεί τη σχέση (4.22). Από τον ορισμό της δεσμευμένης ή υπό συνθήκη πιθανότητας και λαμβάνοντας υπόψη τη σχέση (4.14) έχουμε:

$$\begin{aligned} P(X > t+s|X > t) &= \frac{P(X > t+s,\ X > t)}{P(X > t)} = \frac{P(X > t+s)}{P(X > t)} \\ &= \frac{1 - F_X(t+s)}{1 - F_X(t)} = \frac{1 - \left(1 - (1-p)^{t+s}\right)}{1 - \left(1 - (1-p)^{t}\right)} \\ &= (1-p)^s = 1 - F_X(s) = P(X > s), \end{aligned}$$

που αποδεικνύει ότι η γεωμετρική κατανομή πληροί αυτήν την ιδιότητα.

Στη συνέχεια, θα δείξουμε ότι η γεωμετρική κατανομή είναι η μοναδική κατανομή μεταξύ των κατανομών με δυνατό σύνολο τιμών $x \in \{0,1,2,\ldots\}$ που ικανοποιεί την ιδιότητα της αμνησίας. Έστω X μια τ.μ. με σπ και ασκ $p_X(\cdot)$ και $F_X(\cdot)$, αντίστοιχα, για την οποία υποθέτουμε ότι ισχύει η ιδιότητα της αμνησίας, δηλαδή η σχέση (4.22) και ότι $x \in \{1,2,\ldots\}$. Η σχέση αυτή πρωτύτερα μας οδήγησε στην ισοδύναμη έκφραση:

$$\frac{1 - F_X(t+s)}{1 - F_X(t)} = 1 - F_X(s), \text{ για κάθε } s,t \in \{1,2,\ldots\},$$

ή, ισοδύναμα, στη σχέση:

$$S_X(s+t) = S_X(t)S_X(s), \text{ για κάθε } s,t \in \{1,2,\ldots\}, \tag{4.23}$$

όπου $S_X(x) = P(X > x) = 1 - F_X(x)$ με $0 \le S_X(x) \le 1$. Από τον παραπάνω ορισμό και καθώς το σύνολο δυνατών τιμών είναι $x \in \{0,1,2,\ldots\}$, έχουμε ότι $S_X(0) = 1 - F_X(0) = 1 - 0 = 1$. Επιπρόσθετα, αξιοποιώντας τη σχέση (4.23) για $s = t = 1$, έχουμε ότι $S_X(2) = (S_X(1))^2$, ενώ από την ίδια σχέση για $s = 2$ και $t = 1$, έχουμε ότι $S_X(3) = S_X(2) \cdot S_X(1) = (S_X(1))^2 \cdot S_X(1)$. Συνεχίζοντας, κατά τον ίδιο τρόπο, προκύπτει ότι $S_X(n) = (S_X(1))^n$ για κάθε $n \in \{1,2,3,\ldots\}$ ή, ισοδύναμα, $F_X(n) = 1 - S_X(1)^n$ για κάθε $n \in \{1,2,3,\ldots\}$, καθώς εξ ορισμού $S_X(n) = 1 - F_X(n)$. Γράφοντας ότι $S_X(1) = 1 - (1 - S_X(1))$, προκύπτει τελικά ότι

$$F_X(n) = 1 - \{1 - [1 - S_X(1)]\}^n. \tag{4.24}$$

Λαμβάνοντας υπόψη τη σχέση (4.14), έχουμε ότι η (4.24) είναι η ασκ της γεωμετρικής κατανομής με παράμετρο $p = 1 - S_X(1)$ και η απόδειξη ολοκληρώθηκε.

Το ερώτημα που τίθεται τώρα είναι: ποια είναι η πραγματική ερμηνεία της ιδιότητας της αμνησίας; Χωρίς βλάβη της γενικότητας, ας θεωρήσουμε ότι η τ.μ. $X \sim Geo(p)$ περιγράφει τον αριθμό των ατόμων που επιλέγονται μέχρι να εμφανιστεί το πρώτο άτομο που πάσχει από μυωπία στον γενικό πληθυσμό, όπου κάθε άτομο έχει σταθερή πιθανότητα p να έχει μυωπία και το ένα άτομο δεν επηρεάζει το άλλο αναφορικά με τη μυωπία. Με άλλα λόγια, η ιδιότητα της αμνησίας λέει ότι η γνώση ότι δεν έχει εμφανιστεί άτομο με μυωπία στις t το πλήθος επαναλήψεις, δεν επηρεάζει καθόλου την πιθανότητα να χρειαστούν τουλάχιστον άλλες s το πλήθος επαναλήψεις για να εμφανιστεί το πρώτο άτομο με μυωπία. Το παραπάνω είναι απόλυτα λογικό μιας και έχουμε θεωρήσει ότι η πιθανότητα επιτυχίας p παραμένει αμετάβλητη και οι δοκιμές ανεξάρτητες.

Παρατήρηση 4.9

Έστω $X \sim Geo(p)$, $0 < p < 1$, $x \in \{1,2,...\}$ με σπ που δίνεται από τη σχέση (4.12), και Y η τ.μ., με σπ που δίνεται από τη σχέση (4.15), με $Y = X - 1$. Στη γλώσσα προγραμματισμού R χρησιμοποιείται η τ.μ. Y. Αν εμείς θέλουμε αποτελέσματα σχετικά με τη X κάνουμε, μέσω της σχέσης $p_X(x) = p_Y(x-1)$, κάποιες τροποποιήσεις, όπως φαίνεται παρακάτω, και μπορούμε:

- με τη συνάρτηση `dgeom(x-1,p)` να υπολογίσουμε τη σπ στο σημείο ή στο διάνυσμα σημείων x,
- με τη συνάρτηση `pgeom(x-1,p,lower.tail=TRUE)` να υπολογίσουμε την ασκ στο σημείο ή στο διάνυσμα σημείων x,
- με τη συνάρτηση `pgeom(x-1, p,lower.tail=FALSE)` να υπολογίσουμε την πιθανότητα $P(X > x)$ στο σημείο ή στο διάνυσμα σημείων x,
- με τη συνάρτηση `qgeom(q,p,lower.tail=TRUE)` να προσδιορίσουμε τη μικρότερη τιμή ή το διάνυσμα των μικρότερων τιμών $x-1$ για τις οποίες ισχύει ότι $P(X \le x) \ge q$, όπου q είναι πιθανότητα ή διάνυσμα με συνιστώσες πιθανότητες, αντίστοιχα. Επομένως, το διάνυσμα των σημείων x είναι `qgeom(q,p,lower.tail=TRUE)+1`,
- με τη συνάρτηση `qgeom(q,p,lower.tail=FALSE)` να προσδιορίσουμε τη μικρότερη τιμή ή το διάνυσμα των μικρότερων τιμών $x-1$ για τις οποίες ισχύει ότι $P(X > x) \ge q$, όπου q είναι πιθανότητα ή διάνυσμα με συνιστώσες πιθανότητες, αντίστοιχα. Επομένως, το διάνυσμα των σημείων x είναι `qgeom(q,p,lower.tail=FALSE)+1`,
- με τη συνάρτηση `rgeom(n,p)` να παράγουμε ένα δείγμα μεγέθους n από την κατανομή με σπ (4.15) και, επομένως, αρκεί να προσθέσουμε τη μονάδα σε κάθε παρατήρηση για να έχουμε δείγμα μεγέθους n από τη γεωμετρική με σπ (4.12) ή απλώς να γράψουμε `rgeom(n,p)+1`.

Στην περίπτωση που κάποιος θέλει να χρησιμοποιήσει την παραμετροποίηση που δίνεται στη σχέση (4.15), τότε στα παραπάνω αντικαθιστά το $x-1$ με το x και δεν απαιτείται να προσθέτει τη μονάδα.
Η χρήση κάποιων εκ των παραπάνω εντολών παρουσιάζεται στα παραδείγματα που ακολουθούν.

Παράδειγμα 4.3

Ρίχνουμε ένα ζάρι μέχρι να φέρουμε πρώτη φορά 6.

1. Προσδιορίστε τη συνάρτηση πιθανότητας της τ.μ. X που παριστάνει τον αριθμό των ρίψεων μέχρι την εμφάνιση της ένδειξης 6 για πρώτη φορά.
2. Πόσες φορές αναμένεται να ρίξουμε το ζάρι μέχρι να φέρουμε για πρώτη φορά 6;
3. Υπολογίστε την πιθανότητα ότι θα χρειαστούμε τουλάχιστον 5 ρίψεις μέχρι να φέρουμε πρώτη φορά 6.
4. Ποιος είναι ο μικρότερος ακέραιος αριθμός που είναι τέτοιος, ώστε η πιθανότητα να χρειαστούν το πολύ τόσες επαναλήψεις, για να εμφανιστεί για πρώτη φορά 6, να είναι μεγαλύτερη ή ίση με 0.35;

Λύση Παραδείγματος 4.3

Η ρίψη του ζαριού μπορεί να θεωρηθεί ως ένα πείραμα τύχης με δύο δυνατά αποτελέσματα: ένδειξη 6 και ένδειξη διαφορετική από 6. Θεωρούμε επιτυχία την εμφάνιση ένδειξης 6. Τότε πρόκειται για μια δοκιμή Bernoulli, με την πιθανότητα επιτυχίας να παραμένει αμετάβλητη σε κάθε επανάληψη αυτής της δοκιμής Bernoulli και ίση με 1/6. Επιπλέον, το αποτέλεσμα μιας οποιασδήποτε ρίψης του ζαριού δεν επηρεάζει το αποτέλεσμα οποιασδήποτε άλλης, δηλαδή οι επαναλήψεις των δοκιμών Bernoulli είναι ανεξάρτητες.

1. Έστω X η τ.μ. που παριστάνει τον αριθμό των ρίψεων μέχρι την εμφάνιση της ένδειξης 6. Τότε $X \sim Geo(p = 1/6)$ με σπ που δίνεται από τη σχέση (4.12) για $p = 1/6$, δηλαδή $p_X(x) = \frac{1}{6}\left(\frac{5}{6}\right)^{x-1}$, $x = 1,2,...$

2. Ζητείται η $E(X) = \frac{1}{p} = 6$.
3. Θέλουμε να υπολογίσουμε την πιθανότητα $P(X \geq 5)$ ή, ισοδύναμα, την $P(X > 4)$ ή την $1 - P(X \leq 4)$. Είναι

$$P(X \leq 4) = \sum_{x=1}^{4} \frac{1}{6}\left(\frac{5}{6}\right)^{x-1} = \frac{1}{6}\sum_{x=1}^{4}\left(\frac{5}{6}\right)^{x-1} = \frac{1}{6}\frac{1-\left(\frac{5}{6}\right)^4}{1-\frac{5}{6}} = 1-\left(\frac{5}{6}\right)^4 = 0.5177$$

όπου χρησιμοποιήσαμε τη σχέση (Β΄.2) του Παραρτήματος Β΄, καθώς το άθροισμα ισούται με το άθροισμα των τεσσάρων πρώτων διαδοχικών όρων γεωμετρικής προόδου με πρώτο όρο τη μονάδα και λόγο ίσο με 5/6. Επομένως, $P(X \geq 5) = 1 - 0.5177 = 0.4823$.

Χρησιμοποιώντας την R και την εντολή `pgeom(3,1/6,lower.tail=FALSE)` έχουμε ότι $P(X > 4) = 0.4822531$, η οποία ισούται πρακτικά (εκτός από τα σφάλματα στρογγυλοποίησης στους υπολογισμούς) με το αποτέλεσμα που λάβαμε παραπάνω.

4. Θέλουμε να προσδιορίσουμε τον μικρότερο ακέραιο x, για τον οποίο ισχύει ότι $P(X \leq x) \geq 0.35$. Επομένως, χρησιμοποιώντας τη σχέση (4.14), θέλουμε να ισχύει ότι:

$$1 - (1-p)^x \geq 0.35 \text{ ή } 1 - \left(\frac{5}{6}\right)^x \geq 0.35.$$

Έχουμε ισοδύναμα ότι $\left(\frac{5}{6}\right)^x \leq 0.65$ ή $x \log\left(\frac{5}{6}\right) \leq \log(0.65)$, δηλαδή $x \geq 2.3628$. Επομένως, ο μικρότερος ακέραιος για τον οποίο ισχύει η παραπάνω σχέση είναι ο $x = 3$. Εναλλακτικά, χρησιμοποιώντας την R και την εντολή `qgeom(0.35,1/6,lower.tail=TRUE)+1` έχουμε αποτέλεσμα $x = 3$. Πράγματι, η $P(X \leq 3) = 0.4212963$.

Άσκηση Αυτοαξιολόγησης 4.7

Το παιχνίδι της αμερικάνικης ρουλέτας έχει 38 δυνατά αποτελέσματα: 18 κόκκινα, 18 μαύρα και 2 πράσινα. Υποθέστε ότι σε 10 διαδοχικά παιχνίδια έχει έρθει μαύρο ή πράσινο. Υπολογίστε την πιθανότητα να χρειαστούν τουλάχιστον k παιχνίδια ακόμη, μέχρι να εμφανιστεί για πρώτη φορά κόκκινο.

Άσκηση Αυτοαξιολόγησης 4.8

Από προηγούμενες έρευνες είναι γνωστό ότι σε μια περιοχή της Ελλάδας η πιθανότητα κάποιος ενήλικας δημότης της να έχει μυωπία είναι 0.2 και κάθε άτομο έχει ή δεν έχει την πάθηση αυτή ανεξάρτητα από κάθε άλλο άτομο. Υπολογίστε την πιθανότητα σε μια τυχαία δειγματοληψία που διενεργεί ένας οφθαλμίατρος μεταξύ των ενήλικων δημοτών να χρειαστεί ακριβώς 4 επιλογές ατόμων μέχρι να βρει το πρώτο με αυτήν την πάθηση.

4.5 Αρνητική διωνυμική κατανομή

Σε πολλά τυχαία πραγματικά φαινόμενα μας ενδιαφέρει να μελετήσουμε τον αριθμό των ανεξάρτητων επαναλήψεων μιας δοκιμής Bernoulli μέχρις ότου εμφανιστεί η r-οστή επιτυχία με $r \geq 1$. Η κατανομή που μοντελοποιεί τέτοια τυχαία πραγματικά φαινόμενα είναι γνωστή στη βιβλιογραφία ως αρνητική διωνυμική κατανομή ή κατανομή Pascal και, προφανώς, αποτελεί γενίκευση της γεωμετρικής, καθώς για $r = 1$ έχουμε τη γεωμετρική. Η ονομασία κατανομή Pascal δικαιολογείται πλήρως καθώς έχει δοθεί προς τιμήν του Γάλλου μαθηματικού, φυσικού και φιλοσόφου Blaise Pascal (1623-1662), ο οποίος ήταν ο πρώτος που εισήγαγε ειδικές μορφές της (βλ. Pascal, 1679). Η αρνητική διωνυμική εφαρμόζεται σε πλήθος εφαρμογών, όπως για παράδειγμα στη μελέτη ατυχημάτων, στη διάρκεια παραμονής σε ένα νοσοκομείο, στις βιολογικές επιστήμες, στην ψυχολογία και στην επιχειρησιακή έρευνα. Για περισσότερες εφαρμογές και λεπτομέρειες

για τα παραπάνω παραπέμπουμε στους Johnson *et al.* (2005).

Ορισμός 4.4

Η τυχαία μεταβλητή X λέγεται ότι ακολουθεί την **αρνητική διωνυμική κατανομή** με παραμέτρους $r \in \{1,2,3,\ldots\}$ και $p \in (0,1)$, αν οι δυνατές της τιμές x είναι $\{r, r+1, r+2, \ldots\}$ και η συνάρτηση πιθανότητάς της δίνεται από τη σχέση:

$$p_X(x) = \binom{x-1}{r-1} p^r (1-p)^{x-r}, \quad x = r, r+1, r+2, \ldots . \tag{4.25}$$

Στην περίπτωση αυτή, θα συμβολίζεται $X \sim NB(r,p)$.

Παρατήρηση 4.10

Το όνομα της αρνητικής διωνυμικής μπορεί να αιτιολογηθεί, αν παρατηρήσουμε ότι ισχύει:

$$\binom{k+r-1}{k} = \frac{r(r+1)\cdots(r+k-1)}{k!} = \frac{(-1)^k(-r)(-r-1)\cdots(-r-k+1)}{k!} = (-1)^k \binom{-r}{k}.$$

Δηλαδή προκύπτει από το γεγονός ότι εμφανίζεται αρνητικός αριθμός στον διωνυμικό συντελεστή.

Όπως και στην προηγούμενη ενότητα, δύο ερωτήματα προκύπτουν, αν η σχέση (4.25) αποτελεί συνάρτηση πιθανότητας και αν όντως μοντελοποιεί το υπό μελέτη τυχαίο φαινόμενο που αναφέρθηκε πρωτύτερα. Όσον αφορά το πρώτο ερώτημα, όντως η συνάρτηση που εμφανίζεται στη σχέση (4.25) αποτελεί συνάρτηση πιθανότητας, καθώς είναι μη αρνητική και, επιπλέον,

$$\sum_{x=r}^{+\infty} p_X(x) = \sum_{x=r}^{+\infty} \binom{x-1}{r-1} p^r (1-p)^{x-r} = p^r \sum_{y=0}^{+\infty} \binom{y+r-1}{y} (1-p)^y$$

όπως προκύπτει μετά από λίγες αλγεβρικές πράξεις, θέτοντας $y = x - r$. Επομένως, αρκεί να αποδείξουμε ότι

$$\sum_{y=0}^{+\infty} \binom{y+r-1}{y} (1-p)^y = p^{-r} = \big(1-(1-p)\big)^{-r}.$$

Παίρνοντας το ανάπτυγμα MacLaurin της συνάρτησης $f(z) = (1-z)^{-r}$, έχουμε

$$\begin{aligned}(1-z)^{-r} &= 1 + rz + r(r+1)\frac{z^2}{2} + r(r+1)(r+2)\frac{z^3}{3!} + \ldots \\ &= \sum_{y=0}^{+\infty} \binom{y+r-1}{y} z^y, \text{για } |z| < 1\end{aligned} \tag{4.26}$$

και το ζητούμενο προκύπτει θέτοντας $z = 1 - p$.

Σχετικά με το δεύτερο ερώτημα ενθυμούμενοι τον ορισμό της συνάρτησης πιθανότητας έχουμε ότι η $p_X(x) = P(X = x)$ ισοδυναμεί με τον υπολογισμό της πιθανότητας

$$P(r-1 \text{ Επιτυχίες στις } x-1 \text{ πρώτες δοκιμές και Επιτυχία στην } r\text{-οστή δοκιμή}).$$

Επομένως, υπό τις υποθέσεις ότι οι δοκιμές είναι ανεξάρτητες και η πιθανότητα επιτυχίας αμετάβλητη και ίση με p $(0 < p < 1)$, έχουμε ότι είναι ισοδύναμη με τον υπολογισμό της

$$P(r-1 \text{ Επιτυχίες στις } x-1 \text{ πρώτες δοκιμές}) \cdot P(\text{Επιτυχία στην } r\text{-οστή δοκιμή}).$$

Η δεύτερη πιθανότητα είναι, προφανώς, ίση με p, ενώ η πρώτη πιθανότητα ισούται με την πιθανότητα $r-1$ επιτυχιών στις $x-1$ το πλήθος δοκιμές ενός διωνυμικού τυχαίου πειράματος και προσδιορίζεται από τη διωνυμική κατανομή $B(x-1,p)$, χρησιμοποιώντας τη σχέση (4.5) για $n=x-1$ και $x=r-1$, να είναι ίση με:

$$\binom{x-1}{r-1} p^{r-1} (1-p)^{x-1-(r-1)}.$$

Συνδυάζοντας τα παραπάνω, προκύπτει το ζητούμενο.

Παρατήρηση 4.11

Όπως αναφέρθηκε, η αρνητική διωνυμική κατανομή μοντελοποιεί τον αριθμό των ανεξάρτητων επαναλήψεων μιας δοκιμής Bernoulli μέχρι την r-οστή επιτυχία. Για τον λόγο αυτό, μπορούμε να πούμε ότι μοντελοποιεί τον χρόνο αναμονής μέχρι την r-οστή επιτυχία σε μια διαδικασία ανεξάρτητων επαναλήψεων μιας δοκιμής Bernoulli. Αυτό εξηγεί γιατί θεωρείται συχνά ως μια διακριτή κατανομή του χρόνου αναμονής.

Αφού προσδιορίστηκε η σπ της τ.μ. X, θα προσδιοριστεί η ασκ της και θα μελετηθούν κάποιες βασικές ιδιότητές της. Σε αυτό το πλαίσιο, χρησιμοποιώντας τη σχέση (4.25) και από τον ορισμό της ασκ, έχουμε ότι, αν $X \sim NB(r,p)$, $0<p<1$, $r \in \{1,2,\ldots\}$, τότε:

$$F_X(x) = \begin{cases} 0, & x < r, \\ \sum_{y=r}^{\lfloor x \rfloor} \binom{y-1}{r-1} p^r (1-p)^{y-r}, & x \geq r, \end{cases} \tag{4.27}$$

όπου το $\lfloor x \rfloor$ συμβολίζει το ακέραιο μέρος του x.

Στη συνέχεια, παρουσιάζονται κάποιες χρήσιμες ιδιότητες και χαρακτηριστικά της τ.μ. $X \sim NB(r,p)$.

Πρόταση 4.8

Έστω X η τυχαία μεταβλητή που ακολουθεί $NB(r,p)$ με σπ που προσδιορίζεται στη σχέση (4.25). Τότε ισχύει ότι:

$$E(X) = \frac{r}{p}, \tag{4.28}$$

και

$$Var(X) = \frac{r(1-p)}{p^2}. \tag{4.29}$$

Απόδειξη Πρότασης 4.8

Υπάρχουν διάφοροι τρόποι προσδιορισμού των $E(X)$ και $Var(X)$. Στη συνέχεια, θα παρουσιαστεί αυτός που στηρίζεται στον ορισμό της μέσης τιμής, οπότε έχουμε:

$$E(X) = \sum_{x=r}^{+\infty} x \binom{x-1}{r-1} p^r (1-p)^{x-r} = \sum_{x=r}^{+\infty} x \frac{(x-1)!}{(r-1)! \cdot (x-r)!} p^r (1-p)^{x-r} = \sum_{x=r}^{+\infty} r \frac{x!}{r!(x-r)!} p^r (1-p)^{x-r}$$

$$= r \sum_{x=r}^{+\infty} \binom{x}{r} p^r (1-p)^{x-r} \overset{x=y-1}{=} \frac{r}{p} \sum_{y=r+1}^{+\infty} \binom{y-1}{r+1-1} p^{r+1} (1-p)^{y-r-1} = \frac{r}{p} \cdot 1 = \frac{r}{p}$$

όπου $\sum_{y=r+1}^{+\infty} \binom{y-1}{r+1-1} p^{r+1} (1-p)^{y-r-1} = 1$ από τις ιδιότητες της σπ της $NB(r+1,p)$.

Επειδή $Var(X) = E(X^2) - (E(X))^2$, θα υπολογιστεί αρχικά η $E(X^2)$ μέσω της ακόλουθης διαδικασίας:

$$\begin{aligned}
E(X^2) &= \sum_{x=r}^{+\infty} x^2 \binom{x-1}{r-1} p^r\,(1-p)^{x-r} = \sum_{x=r}^{+\infty} x^2 \frac{(x-1)!}{(r-1)!(x-r)!}\, p^r\,(1-p)^{x-r} \\
&= \sum_{x=r}^{+\infty} r\,x\,\frac{x!}{r!(x-r)!}\, p^r\,(1-p)^{x-r} = r\sum_{x=r}^{+\infty} x \binom{x}{r} p^r\,(1-p)^{x-r} \\
&\overset{x=y-1}{=} \frac{r}{p} \sum_{y=r+1}^{+\infty} (y-1) \binom{y-1}{r+1-1} p^{r+1}\,(1-p)^{y-r-1} \\
&= \frac{r}{p} \sum_{y=r+1}^{+\infty} y \binom{y-1}{r+1-1} p^{r+1}\,(1-p)^{y-r-1} - \frac{r}{p} \sum_{y=r+1}^{+\infty} \binom{y-1}{r+1-1} p^{r+1}\,(1-p)^{y-r-1} \\
&= \frac{r}{p}\,\frac{r+1}{p} - \frac{r}{p} = \frac{r^2+r-rp}{p^2},
\end{aligned}$$

όπου $\sum_{y=r+1}^{+\infty} y \binom{y-1}{r+1-1} p^{r+1}\,(1-p)^{y-r-1} = \frac{r+1}{p}$, και $\sum_{y=r+1}^{+\infty} \binom{y-1}{r+1-1} p^{r+1}\,(1-p)^{y-r-1} = 1$, τα οποία προκύπτουν παρατηρώντας ότι αποτελούν τη μέση τιμή της $NB(r+1,p)$ και το άθροισμα των τιμών της σπ της $NB(r+1,p)$ αντίστοιχα. Το επιθυμητό αποτέλεσμα προκύπτει άμεσα, καθώς $Var(X) = E(X^2) - (E(X))^2$.

Πρόταση 4.9

Έστω X η τυχαία μεταβλητή που ακολουθεί $NB(r,p)$ με σπ που προσδιορίζεται στη σχέση (4.25). Τότε:

$$M_X(t) = \left(\frac{pe^t}{1-(1-p)e^t}\right)^r, \quad \text{για } t < -\log(1-p). \tag{4.30}$$

Απόδειξη Πρότασης 4.9

Λαμβάνοντας υπόψη τη σχέση (3.27) έχουμε ότι:

$$\begin{aligned}
M_X(t) = E\left(e^{tX}\right) &= \sum_{x=r}^{+\infty} e^{tx} \binom{x-1}{r-1} p^r\,(1-p)^{x-r} = (pe^t)^r \sum_{x=r}^{+\infty} \binom{x-1}{r-1} p^r \left\{(1-p)e^t\right\}^{x-r} \\
&= (pe^t)^r \sum_{y=0}^{+\infty} \binom{y+r-1}{r-1} \left\{(1-p)e^t\right\}^{y}.
\end{aligned}$$

Όμως από την (4.26), αν $(1-p)e^t < 1$, είναι

$$\sum_{y=0}^{+\infty} \binom{y+r-1}{r-1} \left\{(1-p)e^t\right\}^{y} = \left(1-(1-p)e^t\right)^{-r},$$

που αποδεικνύει το ζητούμενο.

Παρατήρηση 4.12

Από την προηγούμενη πρόταση προκύπτει ένας ακόμη, ίσως πιο εύκολος, τρόπος προσδιορισμού της μέσης τιμής και της διακύμανσης της αρνητικής διωνυμικής κατανομής. Ειδικότερα, ενθυμούμενοι ότι $E(X^k) = \frac{d}{dt^k} M_X^{(k)}(t)|_{t=0}$, προκύπτει ότι:

$$\frac{d}{dt} M_X(t) = \frac{r\,(pe^t)^r}{\left(1-(1-p)e^t\right)^{r+1}}$$

και για $t = 0$ έχουμε $E(X) = r\,p^{-1}$. Επιπρόσθετα, μετά από αρκετές απλές αλγεβρικές πράξεις και για $t = 0$ έχουμε ότι $\frac{d}{dt^2} M_X^{(2)}(t)|_{t=0} = E(X^2) = \frac{r(r+1-p)}{p^2}$. Το επιθυμητό αποτέλεσμα προκύπτει άμεσα, καθώς $Var(X) = E(X^2) - (E(X))^2$.

Στη βιβλιογραφία έχει παρουσιαστεί η αρνητική διωνυμική κατανομή και με άλλες μορφές ή τρόπους ορισμού. Μεταξύ αυτών εξέχουσα θέση έχει εκείνη που ορίζεται για να μοντελοποιήσει την τ.μ. Y, που παριστάνει τον αριθμό των αποτυχιών μέχρι να εμφανιστεί η r-οστή επιτυχία. Τότε το σύνολο των δυνατών τιμών της είναι $\{0,1,2,\ldots\}$ και ισχύει ότι $Y = X - r$ με X να είναι η τυχαία μεταβλητή της σχέσης (4.25). Επομένως, άμεσα προκύπτει ότι η $P(Y = y) = P(X - r = y)$ ή, ισοδύναμα, ότι $P(Y = y) = P(X = y + r)$. Επομένως, είναι

$$P_Y(y) = P_X(X = y + r) = P_X(y + r) = \binom{y+r-1}{r-1} p^r\,(1-p)^y, y = 0,1,2,\ldots, \qquad (4.31)$$

όπου X η τ.μ. με σπ που δόθηκε στη σχέση (4.25).

Παρατήρηση 4.13

Έστω $X \sim NB(r,p)$ με σπ που δίνεται από τη σχέση (4.25) και Y η τ.μ. με σπ που δίνεται από τη σχέση (4.31). Προφανώς, ισχύει η σχέση $Y = X - r$. Στη γλώσσα προγραμματισμού R χρησιμοποιείται η τ.μ. Y. Αν εμείς θέλουμε αποτελέσματα σχετικά με τη X, κάνουμε κάποιες τροποποιήσεις, όπως φαίνεται παρακάτω, και μπορούμε:

- με τη συνάρτηση `dnbinom(x-r, r,p)` να υπολογίσουμε τη σπ στο σημείο ή στο διάνυσμα σημείων x,
- με τη συνάρτηση `pnbinom(x-r,r,p,lower.tail=TRUE)` να υπολογίσουμε την ασκ στο σημείο ή στο διάνυσμα σημείων x,
- με τη συνάρτηση `pnbinom(x-r,r,p lower.tail=FALSE)` να υπολογίσουμε την πιθανότητα $P(X > x)$ στο σημείο ή στο διάνυσμα σημείων x,
- με τη συνάρτηση `qnbinom(q,r, p,lower.tail=TRUE)+r` να προσδιορίσουμε τη μικρότερη τιμή ή το διάνυσμα των μικρότερων τιμών x, για τις οποίες ισχύει ότι $P(X \le x) \ge q$, όπου q είναι πιθανότητα ή διάνυσμα με συνιστώσες πιθανότητες, αντίστοιχα,
- με τη συνάρτηση `qnbinom(q,r,p,lower.tail=FALSE)+r` μπορούμε να προσδιορίσουμε τη μικρότερη τιμή ή το διάνυσμα των μικρότερων τιμών x, για τις οποίες ισχύει ότι $P(X > x) \ge q$, όπου q είναι πιθανότητα ή διάνυσμα με συνιστώσες πιθανότητες, αντίστοιχα,
- με τη συνάρτηση `rnbinom(n,r,p)+r` να παράγουμε ένα δείγμα μεγέθους n από την αρνητική διωνυμική με σπ που δόθηκε στη σχέση (4.25).

Η χρήση κάποιων εκ των παραπάνω εντολών παρουσιάζεται στα παραδείγματα που ακολουθούν.

Παράδειγμα 4.4

Ο Αποστόλης και ο Πολυχρόνης παίζουν το εξής παιχνίδι: επιλέγουν από μία τράπουλα ο καθένας τυχαία ένα φύλλο. Αν και οι δύο επιλέξουν καρό τότε κερδίζει ο Αποστόλης, διαφορετικά κερδίζει ο Πολυχρόνης. Οι δύο παίχτες συμφωνούν να τελειώσει το παιχνίδι όταν θα κερδίσει ο Αποστόλης για τρίτη φορά. Υπολογίστε την πιθανότητα να χρειαστούν 12 ακριβώς επαναλήψεις αυτού του παιχνιδιού. Ποιος είναι ο αναμενόμενος αριθμός των επαναλήψεων του παιχνιδιού μέχρι να τελειώσει;

Λύση Παραδείγματος 4.4

Κάθε επανάληψη του παιχνιδιού έχει δύο δυνατά αποτελέσματα: νίκη Αποστόλη ή νίκη Πολυχρόνη. Θεωρούμε επιτυχία τη νίκη του Αποστόλη. Είναι προφανές ότι η πιθανότητα επιτυχίας παραμένει αμετάβλητη σε κάθε επανάληψη και ίση με $p = \frac{1}{4}\frac{1}{4} = \frac{1}{16}$. Επιπρόσθετα, κάθε επανάληψη είναι ανεξάρτητη από οποιαδήποτε άλλη. Έστω X η τ.μ. που παριστάνει τον αριθμό των επαναλήψεων του παιχνιδιού μέχρι να κερδίσει για τρίτη φορά ο Αποστόλης (και έτσι να τελειώσει το παιχνίδι). Είναι $X \sim NB(r = 3, p = 1/16)$. Αρχικά, ζητείται να υπολογιστεί η πιθανότητα $P(X = 12)$, η οποία υπολογίζεται από τη σχέση (4.25) για $x = 12$, $r = 3$ και $p = 1/16$ ως ακολούθως

$$\begin{aligned} P(X = 12) &= \binom{11}{2}\left(\frac{1}{16}\right)^3\left(\frac{15}{16}\right)^{12-3} = \frac{11!}{2! \cdot 9!}\frac{15^9}{16^{12}} \\ &= \frac{10 \cdot 11}{2}\frac{15^9}{16^{12}} = 55 \cdot \frac{15^9}{16^{12}} \\ &= 0.0075. \end{aligned}$$

Εναλλακτικά, η ζητούμενη πιθανότητα μπορεί να υπολογιστεί χρησιμοποιώντας την R με την εντολή: `dnbinom(12-3,3,1/16)`, η οποία επιστρέφει την τιμή 0.007511804 η οποία ισούται πρακτικά (εκτός από τα σφάλματα στρογγυλοποίησης στους υπολογισμούς) με τα αποτελέσματα που λάβαμε παραπάνω.
Τέλος, ο αναμενόμενος αριθμός επαναλήψεων μέχρι να τελειώσει το παιχνίδι είναι $E(X) = \frac{r}{p} = 48$.

Άσκηση Αυτοαξιολόγησης 4.9

Ρίχνουμε δύο ζάρια ταυτόχρονα.

1. Προσδιορίστε τη συνάρτηση πιθανότητας της τ.μ. X που παριστάνει τον αριθμό των φορών που φέρνουμε άθροισμα μεγαλύτερο από 7 σε 10 ρίψεις των ζαριών.
2. Πόσες φορές αναμένεται να φέρουμε άθροισμα μεγαλύτερο από 7 στις 10 ρίψεις των ζαριών;

Άσκηση Αυτοαξιολόγησης 4.10

Σε συνέχεια της Άσκησης Αυτοαξιολόγησης 4.9

1. υπολογίστε την πιθανότητα να χρειαστούμε τουλάχιστον 11 ρίψεις μέχρι να φέρουμε για τρίτη φορά άθροισμα μεγαλύτερο από 7,
2. προσδιορίστε ποιος είναι ο μικρότερος αριθμός που είναι τέτοιος ώστε η πιθανότητα να χρειαστούν το πολύ τόσες επαναλήψεις μέχρι την τρίτη εμφάνιση ενδείξεων με άθροισμα μεγαλύτερο από επτά να είναι μεγαλύτερη ή ίση με 0.15.

4.6 Υπεργεωμετρική κατανομή

Υποθέτουμε ότι ένας πεπερασμένος πληθυσμός έχει συνολικά N μέλη και κάθε μέλος του ανήκει σε μία και μόνο από δύο ομάδες (έστω ομάδα Α και Β, αντίστοιχα). Υποθέτουμε ότι $N = N_1 + N_2$, όπου N_i, $i = 1,2$, είναι ο αριθμός των μελών της i-οστής ομάδας, $i = 1,2$. Η επιλογή μελών της μιας ομάδας θεωρείται επιτυχία και της άλλης αποτυχία. Σε αυτό το πλαίσιο, η διωνυμική κατανομή που περιγράφθηκε στην Ενότητα 4.3 προσδιορίζει την πιθανότητα k επιτυχιών (k ατόμων της Α ομάδας) σε n επιλογές με επανατοποθέτηση από αυτόν τον πεπερασμένο πληθυσμό. Παρατηρήστε ότι, καθώς οι επιλογές γίνονται με επανατοποθέτηση, η πιθανότητα επιτυχίας σε κάθε επιλογή παραμένει σταθερή και ίση με τη σχετική συχνότητα εμφάνισης της επιτυχίας στον πεπερασμένο πληθυσμό, δηλαδή εδώ N_1/N. Επομένως, η διωνυμική κατανομή στο πλαίσιο της θεωρίας δειγματοληψίας μπορεί να θεωρηθεί ως η κατανομή που μοντελοποιεί τη δειγματοληψία, με επανατοποθέτηση, από δύο ομάδες (τις Α και Β) ενός πεπερασμένου πληθυσμού.

Στην ενότητα αυτή, θα παρουσιαστεί η υπεργεωμετρική κατανομή, η οποία διαφοροποιείται από τη διωνυμική σε δύο βασικά σημεία. Αρχικά, θα χρησιμοποιήσουμε δειγματοληψία χωρίς επανατοποθέτηση από τον πεπερασμένο πληθυσμό και στο πλαίσιο αυτό θα προσδιορίζεται η πιθανότητα k επιτυχιών (k μελών της ομάδας Α) σε n το πλήθος επιλογές. Είναι προφανές ότι, καθώς οι επιλογές γίνονται χωρίς επανατοποθέτηση, η πιθανότητα επιτυχίας δεν παραμένει σταθερή σε κάθε επιλογή, καθώς κάθε επιλογή μειώνει το μέγεθος του πληθυσμού από όπου επιλέγουμε και οι δοκιμές παύουν να είναι ανεξάρτητες. Αυτή είναι και η δεύτερη βασική διαφοροποίηση από τη διωνυμική κατανομή, που προκύπτει ως συνέπεια της πρώτης ότι η δειγματοληψία είναι πλέον χωρίς επανατοποθέτηση.

Η υπεργεωμετρική κατανομή παρουσιάστηκε για πρώτη φορά από τον γνωστό Γάλλο μαθηματικό Abraham de Moivre (1667-1753) στο έργο του με τίτλο de Mensura Sortis (βλ. Moivre, 1712), στο οποίο ασχολήθηκε με 26 προβλήματα τύχης που έχουν οδηγήσει σε πολύ σημαντικά αποτελέσματα της θεωρίας πιθανοτήτων. Ειδικότερα, η υπεργεωμετρική κατανομή παρουσιάστηκε στο πλαίσιο του 14ου προβλήματος. Είναι άξιο αναφοράς ότι το έργο αυτό αποτέλεσε τη βάση για το βιβλίο του Moivre (1718), το πιο σημαντικό βιβλίο στη θεωρία πιθανοτήτων μέχρι τη δημοσίευση από τον Laplace του δικού του συγγράμματος (Laplace, 1812). Για περισσότερες ιστορικές πληροφορίες παραπέμπουμε τον ενδιαφερόμενο αναγνώστη, μεταξύ άλλων, στους Hald *et al.* (1984). Από τότε η υπεργεωμετρική κατανομή μελετήθηκε από διάφορους ερευνητές (βλ. Johnson *et al.*, 2005, και τις εκεί αναφορές), ενώ βρίσκει εφαρμογή σε διάφορα επιστημονικά πεδία.

Χαρακτηριστικό παράδειγμα εφαρμογής της υπεργεωμετρικής κατανομής αποτελεί η εφαρμογή της στον στατιστικό έλεγχο ποιότητας. Παραδείγματος χάριν, αν σε παρτίδες N προϊόντων ξέρουμε ότι υπάρχει ένας αριθμός N_1 ελαττωματικών και αποστέλλουμε σε έναν πελάτη n το πλήθος προϊόντα, τα οποία επιλέγονται στην τύχη, τότε το πλήθος των ελαττωματικών ανάμεσα στα n προϊόντα περιγράφεται από την υπεργεωμετρική κατανομή. Σε αυτές τις περιπτώσεις, ο πελάτης αποδέχεται την παραγγελία, αν τα ελαττωματικά προϊόντα είναι λιγότερα από έναν προκαθορισμένο αριθμό c και η πιθανότητα αποδοχής της παραγγελίας υπολογίζεται με χρήση της υπεργεωμετρικής κατανομής.

Δεύτερο χαρακτηριστικό παράδειγμα εφαρμογής της υπεργεωμετρικής κατανομής είναι η εφαρμογή της στην εκτίμηση του άγνωστου μεγέθους του πληθυσμού ζωντανών οργανισμών, μέσω των επονομαζόμενων capture−recapture ή mark and recapture δεδομένων. Για περισσότερες λεπτομέρειες και εφαρμογές της υπεργεωμετρικής κατανομής παραπέμπουμε, μεταξύ άλλων, στους Joarder (2011) και Johnson *et al.* (2005).

Στη συνέχεια, δίνεται ο ορισμός της υπεργεωμετρικής κατανομής.

Ορισμός 4.5

Η τυχαία μεταβλητή X λέγεται ότι ακολουθεί την **υπεργεωμετρική κατανομή** με παραμέτρους N_1, N_2 και n φυσικούς αριθμούς, με $n \leq N = N_1 + N_2$, αν οι δυνατές της τιμές x είναι φυσικοί αριθμοί τέτοιοι, ώστε $\max\{0, n - N_2\} \leq x \leq \min\{N_1, n\}$ και η συνάρτηση πιθανότητάς της δίνεται από τη σχέση:

$$p_X(x) = \frac{\binom{N_1}{x}\binom{N_2}{n-x}}{\binom{N_1+N_2}{n}}, \quad \max\{0, n - N_2\} \leq x \leq \min\{N_1, n\}. \tag{4.32}$$

Στην περίπτωση αυτή, θα συμβολίζεται $X \sim Hg(N_1, N_2, n)$.

Όπως και στις προηγούμενες ενότητες, δύο ερωτήματα προκύπτουν: αν η σχέση (4.32) αποτελεί συνάρτηση πιθανότητας και αν όντως μοντελοποιεί το υπό μελέτη τυχαίο φαινόμενο που αναφέρθηκε πρωτύτερα. Όσον αφορά το πρώτο ερώτημα, όντως η συνάρτηση που εμφανίζεται στη σχέση (4.32) αποτελεί συνάρτηση πιθανότητας, καθώς είναι μη αρνητική και επιπλέον:

$$\sum_{x=0}^{n} \binom{N_1}{x}\binom{N_2}{n-x} = \binom{N_1+N_2}{n}.$$

Υπάρχουν διάφοροι τρόποι απόδειξης της τελευταίας σχέσης που είναι γνωστή ως ταυτότητα του Vandermonde, καθώς οφείλεται στον Γάλλο μαθηματικό, μουσικό και χημικό Alexandre Theophile Vandermonde (1735-1796). Ο αλγεβρικός τρόπος βασίζεται στις ακόλουθες σχέσεις, οι οποίες προέρχονται από το διωνυμικό θεώρημα που δόθηκε στη σχέση (4.6), αλλά μπορούν να προκύψουν και ως ειδική περίπτωση αναπτυγμάτων Taylor:

$$(1+y)^{N_1+N_2} = \sum_{n=0}^{N_1+N_2} \binom{N_1+N_2}{n} y^n,$$

ενώ

$$(1+y)^{N_1+N2} = (1+y)^{N_1}(1+y)^{N_2} = \left(\sum_{i=0}^{N_1}\binom{N_1}{i} y^i\right)\left(\sum_{j=0}^{N_2}\binom{N_2}{j} y^j\right)$$

$$= \sum_{n=0}^{N_1+N_2}\left(\sum_{x=0}^{n}\binom{N_1}{x}\binom{N_2}{n-x}\right) y^n.$$

Η τελευταία σχέση προήλθε εφαρμόζοντας τη γενική σχέση πολλαπλασιασμού δύο πολυωνύμων του y, σύμφωνα με την οποία

$$\sum_{i=0}^{p} a_i\, y^i \sum_{j=0}^{q} b_j\, y^j = \sum_{n=0}^{p+q}\left(\sum_{x=0}^{n} a_x b_{n-x}\right) y^n$$

για την ειδική περίπτωση $p = N_1$, $q = N_2$, $a_i = \binom{N_1}{i}$ και $b_j = \binom{N_2}{j}$.

Παρατηρήστε ότι τα αριστερά μέλη των παραπάνω σχέσεων είναι ίσα, γεγονός που συνεπάγεται την ισότητα των δεξιών μελών. Γνωρίζουμε τότε ότι, αν $\sum_{n=0}^{N_1+N_2} a_n\, x^n = \sum_{n=0}^{N_1+N_2} b_n\, x^n$, τότε $a_n = b_n$ για κάθε n. Εφαρμόζοντας το παραπάνω, προκύπτει το ζητούμενο. Η παραπάνω ήταν μια αλγεβρική απόδειξη. Διαφορετικά σκεφτείτε

ότι θέλουμε να επιλέξουμε n το πλήθος υποκείμενα από μια συλλογή $N_1 + N_2$ το πλήθος υποκειμένων, με N_1 να ανήκουν σε μια ομάδα και N_2 σε άλλη. Στην τελική επιλογή μπορώ να έχω x άτομα από την πρώτη και $n - x$ από τη δεύτερη ομάδα για όλες τις δυνατές τιμές του x. Οι τρόποι επιλογής των x και $n - x$ είναι $\binom{N_1}{x}, \binom{N_2}{n-x}$, αντίστοιχα. Με τον πολλαπλασιαστικό νόμο και, καθώς αυτό συμβαίνει για κάθε x, προκύπτει το ζητούμενο.

Σχετικά με το δεύτερο ερώτημα, ενθυμούμενοι τον ορισμό της συνάρτησης πιθανότητας, έχουμε ότι η $p_X(x) = P(X = x)$ ισοδυναμεί με τον υπολογισμό της πιθανότητας

$$P(x \text{ υποκείμενα από την ομάδα Α εκλέγονται χωρίς επανατοποθέτηση, ενώ επιλέγω } n).$$

Η σπ προκύπτει λαμβάνοντας υπόψη ότι οι δυνατοί τρόποι επιλογής n το πλήθος στοιχείων από τα $N_1 + N_2$ είναι $\binom{N_1+N_2}{n}$, ενώ οι ευνοϊκές περιπτώσεις επιλογής x και $n - x$ το πλήθος υποκειμένων από καθεμία ομάδα είναι $\binom{N_1}{x}\binom{N_2}{n-x}$ και, χρησιμοποιώντας τον κλασικό ορισμό της πιθανότητας, προκύπτει η σπ.

Στην επόμενη πρόταση θα προσδιορίσουμε τη μέση τιμή και τη διακύμανση της υπεργεωμετρικής κατανομής.

Πρόταση 4.10

Έστω X η τυχαία μεταβλητή που ακολουθεί $Hg(N_1, N_2, n)$ με σπ που προσδιορίζεται στη σχέση (4.32) και $N = N_1 + N_2$. Τότε:

$$E(X) = \frac{nN_1}{N}, \tag{4.33}$$

και

$$Var(X) = \frac{nN_1}{N_1+N_2}\left(\frac{(n-1)(N_1-1)}{N_1+N_2-1} + 1 - \frac{nN_1}{N_1+N_2}\right), \tag{4.34}$$

ή, ισοδύναμα,

$$Var(X) = \frac{nN_1}{N^2(N-1)}\ (N - N_1)\ (N - n). \tag{4.35}$$

Απόδειξη Πρότασης 4.10

Σύμφωνα με τη σχέση (4.8) είναι:

$$x\binom{N_1}{x} = N_1\binom{N_1-1}{x-1} \text{ και } \binom{N}{n} = \frac{N}{n}\binom{N-1}{n-1}. \tag{4.36}$$

Επιπλέον, εξ ορισμού έχουμε ότι:

$$E(X) = \sum_{x=0}^{n} x\frac{\binom{N_1}{x}\binom{N-N_1}{n-x}}{\binom{N}{n}} = \frac{nN_1}{N}\sum_{x=1}^{n}\frac{\binom{N_1-1}{x-1}\binom{N-N_1}{n-x}}{\binom{N-1}{n-1}}$$

$$= \frac{nN_1}{N}\sum_{y=0}^{n-1}\frac{\binom{N_1-1}{y}\binom{N-N_1}{n-y-1}}{\binom{N-1}{n-1}} = \frac{nN_1}{N},$$

όπου το δεύτερο άθροισμα ξεκινά από την τιμή $x = 1$, καθώς δεν λάβαμε υπόψη την τιμή $x = 0$, η οποία δεν συνεισφέρει τίποτα, ενώ το τελευταίο άθροισμα ισούται με 1, καθώς είναι το άθροισμα της σπ της υπεργεωμετρικής $Hg(N_1 - 1, N - N_1, n - 1)$ σε όλες τις δυνατές τιμές της, και η ζητούμενη μέση τιμή προσδιορίστηκε.

Για την εύρεση της διακύμανσης αρκεί να προσδιορίσουμε την $E(X^2)$. Είναι τότε, χρησιμοποιώντας την (4.36):

$$E(X^2) = \sum_{x=0}^{n} x^2 \frac{\binom{N_1}{x}\binom{N-N_1}{n-x}}{\binom{N}{n}} = \frac{nN_1}{N} \sum_{x=1}^{n} x \frac{\binom{N_1-1}{x-1}\binom{N-N_1}{n-x}}{\binom{N-1}{n-1}}$$

$$= \frac{nN_1}{N} \sum_{y=0}^{n-1} (y+1) \frac{\binom{N_1-1}{y}\binom{N-N_1}{n-y-1}}{\binom{N-1}{n-1}}$$

$$= \frac{nN_1}{N} \sum_{y=0}^{n-1} y \frac{\binom{N_1-1}{y}\binom{N-N_1}{n-y-1}}{\binom{N-1}{n-1}} + \frac{nN_1}{N} \sum_{y=0}^{n-1} \frac{\binom{N_1-1}{y}\binom{N-N_1}{n-y-1}}{\binom{N-1}{n-1}},$$

όπου το δεύτερο άθροισμα ξεκινά από την τιμή $x = 1$, καθώς δεν λάβαμε υπόψη την τιμή $x = 0$, η οποία δεν συνεισφέρει τίποτα, ενώ στην τελευταία σχέση έχουμε στον πρώτο όρο τη μέση τιμή της $Hg(N_1 - 1, N - N_1, n - 1)$ και το άθροισμα στον δεύτερο όρο ισούται με μονάδα, καθώς είναι το άθροισμα της σπ της υπεργεωμετρικής $Hg(N_1 - 1, N - N_1, n - 1)$ σε όλες τις δυνατές τιμές της. Η μέση τιμή της $Hg(N_1 - 1, N - N_1, n - 1)$ είναι $\frac{(n-1)(N_1-1)}{N-1}$ και επομένως:

$$E(X^2) = \frac{nN_1}{N}\left(\frac{(n-1)(N_1-1)}{N-1} + 1\right).$$

Η σχέση (4.34) προκύπτει, καθώς $Var(X) = E(X^2) - (E(X))^2$.
Τέλος, από τη σχέση (4.34), ύστερα από λίγη άλγεβρα, προκύπτει η σχέση (4.35).

Όπως αναφέρθηκε στην αρχή αυτής της ενότητας, η διωνυμική κατανομή συνδέεται με τη δειγματοληψία με επανάθεση από έναν πεπερασμένο πληθυσμό, ενώ η υπεργεωμετρική κατανομή συνδέεται με τη δειγματοληψία χωρίς επανάθεση από έναν πεπερασμένο πληθυσμό με $N = N_1 + N_2$ μέλη. Στην επόμενη πρόταση αποδεικνύεται ότι όταν, $N \to +\infty$, έτσι ώστε $\frac{N_1}{N} \to p$, $0 < p < 1$, τότε προκύπτει η διωνυμική κατανομή με παραμέτρους n και p.

Πρόταση 4.11

Έστω η τυχαία μεταβλητή $X \sim Hg(N_1, N_2, n)$ με $N_1 + N_2 = N$. Αν $N \to +\infty$, έτσι ώστε $\frac{N_1}{N} \to p$, $0 < p < 1$, και το n είναι σταθερό, τότε

$$\lim_{N\to+\infty} p_X(x) = \binom{n}{x} p^x (1-p)^{n-x}, x = 0, 1, 2, \ldots, n.$$

Απόδειξη Πρότασης 4.11

Η σπ της υπεργεωμετρικής κατανομής γράφεται:

$$\begin{aligned}\frac{\binom{N_1}{x}\binom{N-N_1}{n-x}}{\binom{N}{n}} &= \frac{N_1!}{x!(N_1-x)!}\frac{(N-N_1)!}{(n-x)!(N-n-(N_1-x))!}\frac{n!(N-n)!}{N!}\\ &= \binom{n}{x}\frac{N_1!/(N_1-x)!}{N!/(N-x)!}\frac{(N-N_1)!(N-n)!}{(N-x)!(N-N_1-(n-x))!}\\ &= \binom{n}{x}\frac{N_1!/(N_1-x)!}{N!/(N-x)!}\frac{(N-N_1)!/(N-N_1-(n-x))!}{(N-n+(n-x))!/(N-n)!}\\ &= \binom{n}{x}\prod_{k=1}^{x}\frac{(N_1-x+k)}{(N-x+k)}\prod_{m=1}^{n-x}\frac{(N-N_1-(n-x)+m)}{(N-n+m)}.\end{aligned}$$

Το ζητούμενο προκύπτει, καθώς

$$\lim_{N\to+\infty}\frac{(N_1-x+k)}{(N-x+k)} = \lim_{N\to\infty}\frac{N_1}{N} = p$$

και

$$\lim_{N\to+\infty}\frac{(N-N_1-(n-x)+m)}{(N-n+m)} = \lim_{N\to+\infty}\frac{N-N_1}{N} = 1-p$$

Η προηγούμενη πρόταση μας λέει ότι, αν το μέγεθος του πληθυσμού είναι πάρα πολύ μεγάλο, με τις δύο ομάδες να αποτελούνται από πολλά μέλη, τότε μπορούμε να προσεγγίσουμε την υπεργεωμετρική κατανομή από τη διωνυμική.

Παρατήρηση 4.14

Στην παρατήρηση αυτή θα αιτιολογηθεί, χωρίς να υπεισέλθουμε σε λεπτομέρειες, η ονομασία της υπεργεωμετρικής κατανομής. Έστω $X \sim Hg(N_1, N_2, n)$ με $N_1 + N_2 = N$. Τότε η σχέση (4.32) μπορεί ισοδύναμα να γραφτεί ως:

$$p_X(x) := h_N(x,n) = \frac{\binom{Np}{x}\binom{N(1-p)}{n-x}}{\binom{N}{n}}, \tag{4.37}$$

όπου $p = N_1/N$ και αποδεικνύεται ότι (η απόδειξη ξεφεύγει από τους στόχους του παρόντος συγγράμματος και παραλείπεται):

$$E(u^X) = \sum_{x=0}^{n} h_N(x,n)u^x = \frac{\binom{N_2}{n}}{\binom{N}{n}}{}_2F_1(-n,-Np;N(1-p)-n+1;u), \tag{4.38}$$

όπου ${}_2F_1(a_1,a_2;b;x)$ είναι η γενικευμένη υπεργεωμετρική συνάρτηση (ονομάζεται έτσι γιατί μπορεί να εκφραστεί σε όρους υπεργεωμετρικών σειρών). Η σύνδεση αυτή της $E(u^X)$, δηλαδή της πιθανογεννήτριας συνάρτησης της X με την υπεργεωμετρική συνάρτηση εξηγεί και την ονομασία της κατανομής.

Παρατήρηση 4.15

Έστω $X \sim Hg(N_1, N_2, n)$ με σπ που δίνεται από τη σχέση (4.32). Στη γλώσσα προγραμματισμού R μπορούμε:

- με τη συνάρτηση `dhyper(x,N_1,N_2,n)` να υπολογίσουμε τη σπ στο σημείο ή στο διάνυσμα σημείων x,
- με τη συνάρτηση `phyper(x,N_1,N_2,n,lower.tail=TRUE)` να υπολογίσουμε την ασκ στο σημείο ή στο διάνυσμα σημείων x,
- με τη συνάρτηση `phyper(x,N_1,N_2,n,lower.tail=FALSE)` να υπολογίσουμε την πιθανότητα $P(X > x)$ στο σημείο ή στο διάνυσμα σημείων x,
- με τη συνάρτηση `qhyper(q,N_1,N_2,n,lower.tail=TRUE)` να προσδιορίσουμε τη μικρότερη τιμή ή το διάνυσμα των μικρότερων τιμών x για τις οποίες ισχύει ότι $P(X \leq x) \geq q$, όπου q είναι πιθανότητα ή διάνυσμα με συνιστώσες πιθανότητες, αντίστοιχα,
- με τη συνάρτηση `qhyper(q,N_1,N_2,n,lower.tail=FALSE)` να προσδιορίσουμε τη μικρότερη τιμή ή το διάνυσμα των μικρότερων τιμών x για τις οποίες ισχύει ότι $P(X > x) \geq q$, όπου q είναι πιθανότητα ή διάνυσμα με συνιστώσες πιθανότητες, αντίστοιχα,
- με τη συνάρτηση `rhyper(k,N_1,N_2,n)` να παράγουμε ένα δείγμα μεγέθους k από την κατανομή με σπ, που δίνεται στη σχέση (4.32).

Η χρήση κάποιων εκ των παραπάνω εντολών παρουσιάζεται στα παραδείγματα που ακολουθούν.

Παράδειγμα 4.5

Σε μία παρτίδα 70 προϊόντων ο υπεύθυνος παραγωγής της εταιρείας γνωρίζει ότι υπάρχουν 7 ελαττωματικά προϊόντα. Ένας πελάτης αγοράζει 25 προϊόντα από αυτήν την παρτίδα, χωρίς να γνωρίζει αυτήν την πληροφορία. Παρ' όλα αυτά κάνει τη συμφωνία ότι θα επιστρέψει την παρτίδα αν βρει περισσότερα από ένα ελαττωματικά προϊόντα. Υπολογίστε την πιθανότητα επιστροφής της παραγγελίας.

Λύση Παραδείγματος 4.5

Τα προϊόντα διακρίνονται σε δύο ομάδες: τα ελαττωματικά (ομάδα Α) και τα μη ελαττωματικά (ομάδα Β). Το σύνολο των προϊόντων είναι $N = 70$ με $N_1 = 7$ και $N_2 = 63$. Επιλέγονται 25 από αυτά τα προϊόντα χωρίς επανατοποθέτηση. Έστω X η τ.μ. που παριστάνει τον αριθμό των ελαττωματικών προϊόντων στα 25 που επιλέγονται από τα 70 συνολικά. Είναι $X \sim Hg(7, 63, 25)$ με σπ που δίνεται από τη σχέση (4.32) για $N_1 = 7, N_2 = 63$ και $n = 25$. Η P(επιστροφής της παραγγελίας) είναι ίση με την $P(X > 1)$. Η τελευταία πιθανότητα είναι $P(X > 1) = 1 - P(X \leq 1) = 1 - (P(X = 0) + P(X = 1))$.

Επομένως, αφού

$$P(X = 0) = \frac{\binom{7}{0}\binom{63}{25}}{\binom{70}{25}} = \frac{63! \cdot 25! \cdot (70-25)!}{25! \cdot (63-25)! \cdot 70!} = \frac{63! \cdot 45!}{70! \cdot 38!} = 0.0379,$$

και

$$P(X = 1) = \frac{\binom{7}{1}\binom{63}{24}}{\binom{70}{25}} = \frac{7! \cdot 63! \cdot 25! \cdot (70-25)!}{1! \cdot 6! \cdot 24!(63-24)! \cdot 70!} = \frac{7 \cdot 25 \cdot 63! \cdot 45!}{70! \cdot 39!} = 0.1699.$$

έχουμε ότι $P(X > 1) = 1 - (0.0379 + 0.1699) = 0.7922$.

Εναλλακτικά, μπορεί να υπολογιστεί με την εντολή `phyper(1,7,63,25,lower.tail=FALSE)` στην R, η οποία επιστρέφει την τιμή 0.7922828, η οποία αποτελεί την τιμή της ζητούμενης πιθανότητας με μεγαλύτερη ακρίβεια.

Το επόμενο παράδειγμα περιγράφει τη μέθοδο capture-recapture ή mark and recapture, το οποίο αναφέρθηκε στην αρχή αυτής της ενότητας ότι χρησιμοποιεί την υπεργεωμετρική κατανομή για την εκτίμηση, για παράδειγμα, του άγνωστου πλήθους ενός είδους ψαριού σε μια λίμνη.

Παράδειγμα 4.6

Προκειμένου να εκτιμηθεί το μέγεθος του πληθυσμού των γαλαζολιόψαρων (ένα είδος ψαριού του γλυκού νερού) σε μια μικρή λίμνη στο Missouri, συλλέγονται και επισημαίνονται συνολικά 250 ψάρια αυτού του είδους και στη συνέχεια, απελευθερώνονται. Αφού δόθηκε αρκετός χρόνος για τη διασπορά των ψαριών με ετικέτα, πιάστηκε ένα δείγμα 150 ψαριών αυτού του είδους και διαπιστώθηκε ότι 15 από αυτά ήταν σημαδεμένα. Υπολογίστε το μέγεθος του πληθυσμού των ψαριών αυτού του είδους σε αυτήν τη λίμνη.
Υπόδειξη: Στο διάστημα μεταξύ των δύο επιλογών υποθέτουμε ότι δεν υπάρχουν γεννήσεις ή θάνατοι ψαριών αυτού του είδους στη λίμνη.

Λύση Παραδείγματος 4.6

Έστω ότι N είναι το άγνωστο μέγεθος του πληθυσμού αυτού του είδους ψαριού στη λίμνη στο Missouri. Από αυτά τα ψάρια κάποια επισημαίνονται και κάποια όχι. Επομένως, χωρίζονται σε δύο κατηγορίες με πλήθος μελών $N_1 = 250$ και $N - N_1$, αντίστοιχα. Επιλέγουμε χωρίς επανατοποθέτηση ένα δείγμα $n = 150$ τέτοιων ψαριών στα οποία βρίσκουμε $x = 16$ σημαδεμένα. Συνεπώς, το ποσοστό των σημαδεμένων ψαριών στο δείγμα των 150 είναι $x/n = 16/150$. Η ιδέα, στη συνέχεια, είναι να υπολογιστεί η τιμή N για την οποία λαμβάνει μέγιστο η πιθανότητα του ενδεχομένου που παρατηρήσαμε, δηλαδή η

$$P(X = x) = \frac{\binom{N_1}{x}\binom{N - N_1}{n - x}}{\binom{N}{n}}$$

ή, ισοδύναμα, κάνοντας τις πράξεις στους διωνυμικούς συντελεστές

$$P(X = x) = \frac{n!N_1!}{x!(N_1 - x)!}\frac{(N - N_1)!(N - n)!}{(N - N_1 - n + x)!(N!)}.$$

Παρατηρήστε ότι το πρώτο κλάσμα δεν εξαρτάται από το N. Επομένως, η μεγιστοποίηση της $P(X = x)$ ανάγεται στη μεγιστοποίηση ως προς N της συνάρτησης

$$f(N) = \frac{(N - N_1)!(N - n)!}{(N - N_1 - n + x)!N!}$$

πάνω από τους φυσικούς αριθμούς.
Παρατηρούμε ότι:

$$\frac{f(N)}{f(N-1)} = \frac{(N - n)}{N}\frac{(N - N_1)}{(N - N_1 - n + x)} = \frac{1 - n/N}{1 - (n - x)/(N - N_1)}.$$

Προκύπτει εύκολα από αυτήν τη σχέση ότι $f(N) > f(N - 1)$, αν και μόνο αν

$$1 - n/N > 1 - (n - x)/(N - N_1)$$

ή, ισοδύναμα, αν $\frac{n}{N} < \frac{n-x}{N-N_1}$, δηλαδή αν $n(N - N_1) < N(n - x)$. Από την τελευταία σχέση έχουμε $N < \frac{nN_1}{x}$. Παρόμοια $f(N) < f(N - 1)$, αν και μόνο αν $N > \frac{nN_1}{x}$. Επομένως, μεγιστοποιείται για $N = \lfloor \frac{nN_1}{x} \rfloor$, δηλαδή για τον μεγαλύτερο ακέραιο που είναι μικρότερος ή ίσος από $N_1 n/x$. Για το συγκεκριμένο παράδειγμα, η τιμή αυτή ισούται με $N = 2343$.

Στον ιστότοπο https://probabilityandstats.wordpress.com/2010/02/16/the-capture-recapture-method/ (ημερομηνία προσπέλασης: 1/3/2022) ο/η ενδιαφερομένος/η μπορεί να αναζητήσει περισσότερες λεπτομέρειες για το προηγούμενο παράδειγμα.

Άσκηση Αυτοαξιολόγησης 4.11

Είναι γνωστό ότι μια παρτίδα 500 τσιπ υπολογιστών περιέχει 10 ελαττωματικά προϊόντα. Επιλέγονται 50 τσιπ στην τύχη χωρίς επανάθεση. Προσδιορίστε τη σπ του αριθμού των ελαττωματικών προϊόντων στα 50 που επιλέχθηκαν. Υπολογίστε τον αριθμό των τσιπ που αναμένονται να είναι ελαττωματικά.

4.7 Κατανομή Poisson

Η κατανομή Poisson είναι ίσως η σημαντικότερη συνήθης διακριτή κατανομή με πλήθος εφαρμογών, όπως αναλυτικά θα αναφέρουμε στη συνέχεια. Ονομάζεται έτσι προς τιμήν του Γάλλου μαθηματικού, μηχανικού και φυσικού Siméon Denis Poisson (1781 –1840), ο οποίος παρουσίασε την κατανομή αυτή σε εργασία του το 1837 (Poisson, 1837). Δείτε, επίσης, τις εργασίες των Stigler (1982) και Hald *et al.* (1984). Ειδικότερα, ο Poisson οδηγήθηκε στη γνωστή μας σήμερα κατανομή Poisson ως το όριο της διωνυμικής κατανομής. Ο Poisson αρχικά διαπίστωσε ότι η χρήση της σπ της διωνυμικής κατανομής (βλ. Ενότητα 4.3) παρουσιάζει αρκετές δυσκολίες για τον υπολογισμό πιθανοτήτων, όταν το n παίρνει μεγάλες τιμές και η τιμή για την οποία θέλουμε να υπολογιστεί η πιθανότητα δεν είναι κοντά είτε στο 0 είτε στο n. Αναγνωρίζοντας, λοιπόν, ότι σε τέτοιες περιπτώσεις η διωνυμική κατανομή είναι δύσχρηστη, θέλησε να προσδιορίσει έναν εναλλακτικό τρόπο υπολογισμού των πιθανοτήτων και οδηγήθηκε, υπό τις υποθέσεις της πρότασης που ακολουθεί, στην κατανομή που σήμερα μας είναι γνωστή ως κατανομή Poisson.

Πρόταση 4.12

Έστω ότι η τ.μ. X ακολουθεί τη διωνυμική κατανομή με παραμέτρους $n, p \in (0,1)$, δηλαδή $X \sim B(n,p)$ με σπ που δόθηκε στη σχέση (4.5). Αν για $n \to +\infty$, η πιθανότητα επιτυχίας p συγκλίνει στο 0, $p \to 0$, έτσι ώστε η αναμενόμενη τιμή της τ.μ. να συγκλίνει σε έναν σταθερό αριθμό $\lambda > 0$, δηλαδή αν $E(X) = np \to \lambda$ με $\lambda > 0$, τότε:

$$\lim_{n\to+\infty} p_X(x) = \lim_{n\to+\infty} \left(\binom{n}{x} p^x (1-p)^{n-x} \right) = \frac{e^{-\lambda}\lambda^x}{x!}, x = 0,1,2,\ldots$$

Απόδειξη Πρότασης 4.12

Η σπ της διωνυμικής μπορεί να εκφραστεί ως

$$\begin{aligned} p_X(x) = \binom{n}{x} p^x (1-p)^{n-x} &= \frac{n!}{x!(n-x)!} \frac{(np)^x}{n^x} \left(1 - \frac{np}{n}\right)^{n-x} \\ &= \frac{n(n-1)\cdots(n-x+1)}{nn\cdots n} \frac{1}{x!} (np)^x \left(1 - \frac{np}{n}\right)^{n-x} \\ &= \frac{n(n-1)\cdots(n-x+1)}{nn\cdots n} \left(1-p\right)^{-x} \frac{1}{x!} (np)^x \left(1 - \frac{np}{n}\right)^{n}. \end{aligned}$$

Όμως, όταν $n \to \infty$,

$$\lim_{n\to+\infty} \frac{n(n-1)\cdots(n-x+1)}{n\cdots n} = 1,$$

ενώ, λαμβάνοντας υπόψη τα δεδομένα της πρότασης, έχουμε ότι

$$\lim_{n\to+\infty} \frac{(np)^x}{x!} = \frac{\lambda^x}{x!}, \lim_{n\to+\infty} (1-p)^{-x} = 1 \text{ και } \lim_{n\to+\infty} \left(1 - \frac{np}{n}\right)^{n} = e^{-\lambda}.$$

Συνδυάζοντας τα παραπάνω, προκύπτει το ζητούμενο.

Η πρακτική αξία του αποτελέσματος είναι ότι σε εκείνες τις περιπτώσεις που το n της διωνυμικής κατανομής είναι πολύ μεγάλο, η πιθανότητα επιτυχίας p σε κάθε επανάληψη της δοκιμής Bernoulli είναι πάρα πολύ μικρή και ο αναμενόμενος αριθμός των επιτυχιών $np \to \lambda > 0$, τότε

$$P(X = x) \approx \frac{e^{-\lambda}\lambda^x}{x!}, x = 0,1,2,\dots, \lambda > 0,$$

δηλαδή μπορούμε να υπολογίσουμε τις πιθανότητες προσεγγιστικά μέσω της παραπάνω σχέσης. Παρατηρήστε ότι είναι πάντοτε

$$\frac{e^{-\lambda}\lambda^x}{x!} \geq 0 \text{ και } \sum_{x=0}^{+\infty} \frac{e^{-\lambda}\lambda^x}{x!} = e^{-\lambda}\sum_{x=0}^{+\infty} \frac{\lambda^x}{x!} = e^{-\lambda}e^{\lambda} = 1.$$

Οπότε, το όριο της διωνυμικής κατανομής, υπό τις υποθέσεις που διατυπώθηκαν, οδήγησε στην πραγματικότητα τον Poisson σε μια νέα κατανομή, ο ορισμός της οποίας ακολουθεί.

Ορισμός 4.6

Η τυχαία μεταβλητή X λέγεται ότι ακολουθεί την **κατανομή Poisson** με παράμετρο λ, $\lambda > 0$, αν οι δυνατές της τιμές x είναι $x \in \{0,1,2,\dots,\}$ και η συνάρτηση πιθανότητάς της δίνεται από τη σχέση:

$$p_X(x) = \frac{e^{-\lambda}\lambda^x}{x!}, \quad x = 0,1,2,\dots \tag{4.39}$$

Στην περίπτωση αυτή, θα συμβολίζεται $X \sim \mathscr{P}(\lambda)$.

Παρατήρηση 4.16

Ένα εύλογο ερώτημα που μπορεί να έχει προκύψει είναι αν υπάρχει κάποιος πρακτικός κανόνας για το πότε προσεγγίζονται καλά οι διωνυμικές πιθανότητες από την κατανομή Poisson. Η απάντηση είναι ότι οι διωνυμικές πιθανότητες προσεγγίζονται ικανοποιητικά όταν $n > 20$ και $p < 0.05$, ενώ στην περίπτωση που το $n > 100$ η προσέγγιση είναι εξαιρετική αν $np < 10$ (παραπέμπουμε, μεταξύ άλλων, στον ιστότοπο https://www.solon-karapanagiotis.com/post/approx_binomial/approximating-binomial-with-poisson/ (ημερομηνία προσπέλασης: 1/3/2022).

Αφού προσδιορίστηκε η σπ της τ.μ. X θα προσδιοριστεί η ασκ της και θα μελετηθούν κάποιες ιδιότητές της. Σε αυτό το πλαίσιο, χρησιμοποιώντας τη σχέση (4.39) και από τον ορισμό της ασκ έχουμε ότι, αν $X \sim \mathscr{P}(\lambda)$, $\lambda > 0$, τότε:

$$F_X(x) = \begin{cases} 0, & x < 0, \\ \sum_{y=0}^{\lfloor x \rfloor} \frac{e^{-\lambda}\lambda^y}{y!}, & x \geq 0, \end{cases} \tag{4.40}$$

όπου το $\lfloor x \rfloor$ συμβολίζει το ακέραιο μέρος του x.

Στη συνέχεια, παρουσιάζονται κάποιες χρήσιμες ιδιότητες και χαρακτηριστικά της τ.μ. $X \sim \mathscr{P}(\lambda)$, $\lambda > 0$.

Πρόταση 4.13

Έστω X η τυχαία μεταβλητή που ακολουθεί $\mathscr{P}(\lambda)$ με σπ που προσδιορίζεται στη σχέση (4.39). Τότε ισχύει ότι:

$$E(X) = \lambda, \tag{4.41}$$

και

$$Var(X) = \lambda. \tag{4.42}$$

Απόδειξη Πρότασης 4.13

Υπάρχουν διάφοροι τρόποι προσδιορισμού των $E(X)$ και $Var(X)$. Ακολούθως, θα παρουσιαστεί αυτός που στηρίζεται στον ορισμό των $E(X)$ και $Var(X)$. Από τον ορισμό της μέσης τιμής έχουμε:

$$E(X) = \sum_{x=0}^{+\infty} x\frac{e^{-\lambda}\lambda^x}{x!} = \lambda\sum_{x=1}^{+\infty}\frac{e^{-\lambda}\lambda^{x-1}}{(x-1)!} = \lambda\sum_{y=0}^{+\infty}\frac{e^{-\lambda}\lambda^y}{y!} = \lambda$$

καθώς εντός του τελευταίου αθροίσματος είναι η σπ της κατανομής Poisson με παράμετρο λ. Καθώς $Var(X) = E(X^2) - (E(X))^2$ θα υπολογιστεί στη συνέχεια, η $E(X^2)$. Είναι:

$$E(X^2) = \sum_{x=0}^{+\infty} x^2\frac{e^{-\lambda}\lambda^x}{x!} = \lambda\sum_{x=1}^{+\infty} x\frac{e^{-\lambda}\lambda^{x-1}}{(x-1)!} = \lambda\sum_{y=0}^{+\infty}(y+1)\frac{e^{-\lambda}\lambda^y}{y!}$$

$$= \lambda\sum_{y=0}^{+\infty} y\frac{e^{-\lambda}\lambda^y}{y!} + \lambda\sum_{y=0}^{+\infty}\frac{e^{-\lambda}\lambda^y}{y!} = \lambda^2 + \lambda,$$

καθώς $\sum_{y=0}^{+\infty} y\frac{e^{-\lambda}\lambda^y}{y!} = \lambda$ ως μέση τιμή της Poisson με παράμετρο λ και $\sum_{y=0}^{+\infty}\frac{e^{-\lambda}\lambda^y}{y!} = 1$ από τη σπ Poisson με παράμετρο λ. Το επιθυμητό αποτέλεσμα προκύπτει άμεσα, καθώς $Var(X) = E(X^2) - (E(X))^2$.

Πρόταση 4.14

Έστω X η τυχαία μεταβλητή που ακολουθεί $\mathscr{P}(\lambda)$ με σπ που προσδιορίζεται στη σχέση (4.39), τότε

$$M_X(t) = \exp[\lambda(e^t - 1)]. \tag{4.43}$$

Απόδειξη Πρότασης 4.14

Λαμβάνοντας υπόψη τη σχέση (3.27) έχουμε ότι

$$M_X(t) = E\left(e^{tX}\right) = \sum_{x=0}^{+\infty} e^{tx}\frac{e^{-\lambda}\lambda^x}{x!} = e^{-\lambda}\sum_{x=0}^{+\infty}\frac{(\lambda e^t)^x}{x!} = e^{-\lambda}e^{\lambda e^t} = e^{\lambda e^t - \lambda},$$

που αποδεικνύει το ζητούμενο.

Παρατήρηση 4.17

Από την προηγούμενη πρόταση προκύπτει ένας ακόμη, ίσως πιο εύκολος, τρόπος προσδιορισμού της μέσης τιμής και της διακύμανσης της $X \sim \mathscr{P}(\lambda)$. Ειδικότερα, ενθυμούμενοι ότι $E(X^k) = \frac{d}{dt^k}M_X^{(k)}(t)|_{t=0}$, προκύπτει ότι:

$$\frac{d}{dt}M_X(t) = \exp[\lambda(e^t - 1)]e^t\lambda$$

και για $t = 0$ έχουμε $E(X) = \lambda$. Επιπρόσθετα, μετά από απλές αλγεβρικές πράξεις και για $t = 0$ έχουμε ότι $\frac{d}{dt^2}M_X^{(2)}(t)|_{t=0} = E(X^2) = \lambda^2 + \lambda$. Το επιθυμητό αποτέλεσμα προκύπτει άμεσα, καθώς $Var(X) = E(X^2) - (E(X))^2$.

Για βοήθεια στον υπολογισμό πιθανοτήτων χρησιμοποιώντας τη σπ ή την ασκ της κατανομής Poisson στη βιβλιογραφία υπάρχουν οι λεγόμενοι **πίνακες Poisson**. Οι πίνακες Poisson μας δίνουν απευθείας την τιμή είτε της σπ είτε της ασκ για διάφορες τιμές του n και του λ. Απόσπασμα τέτοιων πινάκων δίνεται στο Παράρτημα Α΄. Στους πίνακες Poisson η τιμή του x, για το οποίο θέλουμε να υπολογίσουμε τη σπ ή την

ασκ, δίνεται στην πρώτη στήλη του πίνακα, ενώ στη συνέχεια, επιλέγεται η τιμή της παραμέτρου λ. Στη διασταύρωση της γραμμής που εντοπίστηκε η τιμή του x και της στήλης που εντοπίστηκε η τιμή του λ λαμβάνεται, ανάλογα με τη μορφή του πίνακα, είτε η τιμή της σπ είτε της ασκ. Σύμφωνα με όσα θα αναφέρουμε στην Ενότητα 7.3, υπό κάποιες προϋποθέσεις, η Poisson πιθανότητα μπορεί να υπολογιστεί προσεγγιστικά από την τυπική κανονική κατανομή, με τον υπολογισμό των πιθανοτήτων της τυπικής κανονικής να γίνεται μέσω ενός και μόνο πίνακα. Προφανώς, ένας εναλλακτικός τρόπος για τον υπολογισμό Poisson πιθανοτήτων είναι η χρήση κάποιας στατιστικής γλώσσας προγραμματισμού. Σε αυτήν την κατεύθυνση, είναι πολύ χρήσιμη η παρατήρηση που ακολουθεί.

Παρατήρηση 4.18

Έστω $X \sim \mathscr{P}(\lambda)$, $\lambda > 0$, $x \in \{0,1,2,...\}$ με σπ που δίνεται από τη σχέση (4.39). Στη γλώσσα προγραμματισμού R μπορούμε:

- με τη συνάρτηση `dpois(x,λ)` να υπολογίσουμε τη σπ στο σημείο ή στο διάνυσμα σημείων x,
- με τη συνάρτηση `ppois(x,λ,lower.tail=TRUE)` να υπολογίσουμε την ασκ στο σημείο ή στο διάνυσμα σημείων x,
- με τη συνάρτηση `ppois(x,λ,lower.tail=FALSE)` να υπολογίσουμε την πιθανότητα $P(X > x)$ στο σημείο ή στο διάνυσμα σημείων x,
- με τη συνάρτηση `qpois(x,λ,lower.tail=TRUE)` να προσδιορίσουμε τη μικρότερη τιμή ή το διάνυσμα των μικρότερων τιμών x για τις οποίες ισχύει ότι $P(X \leq x) \geq c$, όπου c είναι πιθανότητα ή διάνυσμα με συνιστώσες πιθανότητες, αντίστοιχα,
- με τη συνάρτηση `qpois(x,λ,lower.tail=FALSE)` να προσδιορίσουμε τη μικρότερη τιμή ή το διάνυσμα των μικρότερων τιμών x για τις οποίες ισχύει ότι $P(X > x) \geq c$, όπου c είναι πιθανότητα ή διάνυσμα με συνιστώσες πιθανότητες,
- με τη συνάρτηση `rpois(n,λ)` να παράγουμε ένα δείγμα μεγέθους n από την κατανομή με σπ, η οποία δίνεται στη σχέση (4.39).

Η χρήση κάποιων εκ των παραπάνω εντολών παρουσιάζεται στα παραδείγματα που ακολουθούν.

Από όσα προηγήθηκαν, προκύπτει ότι η κατανομή Poisson αποτελώντας, υπό συγκεκριμένες συνθήκες (n μεγάλο, p μικρό), όριο της διωνυμικής κατανομής χρησιμοποιείται για τη μοντελοποίηση του αριθμού των επιτυχιών σε ένα πάρα πολύ μεγάλο πλήθος από ανεξάρτητες επαναλήψεις μιας δοκιμής Bernoulli με σταθερή και πάρα πολύ μικρή πιθανότητα επιτυχίας σε κάθε δοκιμή. Για τον λόγο αυτό, αναφέρεται και ως **κατανομή των σπάνιων ενδεχομένων**. Για παράδειγμα, θα μπορούσε να χρησιμοποιηθεί για την πιθανοθεωρητική μοντελοποίηση του αριθμού των ατόμων που πάσχουν από σαρκοείδωση[3] στην Ελλάδα. Ενδεικτικό σε αυτήν την κατεύθυνση είναι το παράδειγμα που ακολουθεί (βλ. Feller, 1968).

Παράδειγμα 4.7

Υπολογίστε την πιθανότητα σε μια εταιρεία με 500 άτομα, ακριβώς 3 να έχουν γενέθλια την Πρωτοχρονιά. Προσδιορίστε την πιθανότητα με δύο τρόπους (ακριβώς και προσεγγιστικά).

Λύση Παραδείγματος 4.7

Το ενδιαφέρον μας επικεντρώνεται στην τ.μ. X που παριστάνει τον αριθμό των ατόμων που έχουν γενέθλια μια συγκεκριμένη μέρα του χρόνου στα 500 άτομα μιας εταιρείας. Επομένως, πρόκειται για ένα διωνυμικό τυχαίο πείραμα με επιτυχία να θεωρείται κάποιος εργαζόμενος να έχει γενέθλια την Πρωτοχρονιά, ενώ επιπρόσθετα υποθέτουμε ότι το πότε έχει γενέθλια κάποιος εργαζόμενος είναι ανεξάρτητο από τα γενέθλια οποιουδήποτε άλλου στην εταιρεία. Επομένως, $X \sim B(n,p)$, όπου $n = 500$ μεγάλο και $p = 1/365$ πολύ μικρό, ενώ ταυτόχρονα $np = 500/365 < 10$. Άρα η ζητούμενη πιθανότητα

[3] Μια σπάνια αυτοάνοση φλεγμονώδης νόσος που επηρεάζει πολλαπλά όργανα του σώματος αλλά ως επί το πλείστον τους πνεύμονες και τους λεμφαδένες.

$P(X=3)$ μπορεί να βρεθεί είτε από τη σπ της σχέσης (4.5) για $x=3$, $p=1/365$, $n=500$, δηλαδή

$$P(X=3)=\binom{500}{3}\left(\frac{1}{365}\right)^3\left(\frac{364}{365}\right)^{497}=\frac{500!}{3!497!}\frac{364^{497}}{365^{500}}$$
$$=\frac{498\cdot 499\cdot 500}{6}\frac{364^{497}}{365^{500}}=20708500\frac{364^{497}}{365^{500}}=0.1089191$$

ή με την εντολή `dbinom(3,500,1/365)` στην R.
Στη συνέχεια, θα υπολογιστεί η ζητούμενη πιθανότητα προσεγγιστικά από την κατανομή Poisson χρησιμοποιώντας τη σπ της σχέσης (4.39) για $x=3$ και $\lambda=np=500/365$. Είναι τότε προσεγγιστικά ίση με

$$\frac{e^{-\frac{500}{365}}\left(\frac{500}{365}\right)^3}{3!}=0.108882$$

ή, ακόμα πιο εύκολα, με την εντολή `dpois(3,500/365)`. Η προσέγγιση λοιπόν, όπως αναμενόταν, καθώς $n>100$ και $np<10$, είναι εξαιρετική (το σφάλμα είναι μικρότερο από 0.0001).

Με όσα έχουν αναφερθεί, η κατανομή Poisson περιορίζεται στη μοντελοποίηση του αριθμού των επιτυχιών σε πολύ μεγάλο πλήθος ανεξάρτητων επαναλήψεων μιας δοκιμής Bernoulli, υπό την πρόσθετη υπόθεση ότι η πιθανότητα επιτυχίας είναι σταθερή και πάρα πολύ μικρή. Όμως, αποδεικνύεται ότι η κατανομή αυτή, υπό κάποιους περιορισμούς-υποθέσεις, μπορεί να χρησιμοποιηθεί στη μοντελοποίηση τυχαίων πειραμάτων που εξελίσσονται στον χρόνο και το ενδιαφέρον μας επικεντρώνεται στον αριθμό των εμφανίσεων ενός ενδεχομένου σε αυτό. Ακολούθως παρουσιάζουμε πώς επιτυγχάνεται αυτό (βλ. Feller, 1968; Παπαϊωάννου, 1993; Ζωγράφος, 2008).

Η κεντρική ιδέα είναι να συνδεθεί ο αριθμός των συμβάντων στο διάστημα $(0,t]$, όπως π.χ. ο αριθμός σωματιδίων ενός ορισμένου τύπου, ο αριθμός των κλήσεων πελατών σε ένα τηλεφωνικό κέντρο σε αυτό το διάστημα κ.ά., με μια διωνυμική κατανομή που πληροί όλες εκείνες τις υποθέσεις για να προσεγγίζεται από την κατανομή Poisson. Στο πλαίσιο αυτό, αρχικά υποδιαιρούμε το διάστημα $(0,t]$ σε n υποδιαστήματα ίσου μήκους $\frac{t}{n}$, τα οποία υποδιαστήματα αποτελούν διαμέριση του $(0,t]$. Το πλήθος των υποδιαστημάτων n είναι πάρα πολύ μεγάλο. Μαθηματικά θα λέγαμε $n\to+\infty$, που είναι μία εκ των υποθέσεων για να μπορεί να προσεγγιστεί η διωνυμική κατανομή από την Poisson, παρότι ακόμη δεν έχουμε ορίσει ποιο είναι το διωνυμικό τυχαίο πείραμα. Το γεγονός ότι το n είναι πολύ μεγάλο έχει ως συνέπεια το μήκος κάθε υποδιαστήματος να είναι πολύ μικρό. Το ότι το κάθε υποδιάστημα είναι πολύ μικρό μας κάνει να πιστεύουμε ότι είναι ρεαλιστικό να υποθέσουμε ότι σε καθένα από τα n το πλήθος υποδιαστήματα το πολύ ένα συμβάν μπορεί να εμφανιστεί. Αυτό ουσιαστικά σημαίνει ότι σε κάθε υποδιάστημα είτε θα συμβεί ένα συμβάν (επιτυχία) είτε όχι και ότι είναι αμελητέα η πιθανότητα να συμβούν δύο ή περισσότερα συμβάντα. Από το παραπάνω, ίσως γίνεται αντιληπτό ότι η πιθανότητα να έχουμε x συμβάντα στο διάστημα $(0,t]$ ισοδυναμεί με το να συμβεί ένα συμβάν σε x από τα n υποδιαστήματα (θυμηθείτε x επιτυχίες σε n δοκιμές).

Για να μπορούμε, όμως, να χρησιμοποιήσουμε τη διωνυμική κατανομή θα πρέπει η πιθανότητα επιτυχίας, δηλαδή εδώ πραγματοποίησης ενός συμβάντος στο υποδιάστημα πολύ μικρού μήκους, να είναι σταθερή σε όλα τα υποδιαστήματα και, μάλιστα, για να μπορεί να προσεγγιστεί η διωνυμική κατανομή από την Poisson, να είναι τέτοια, ώστε η πιθανότητα p για πολύ μεγάλο n να πηγαίνει στο 0 και το γινόμενο $n\cdot p$ σε σταθερό αριθμό. Υποθέτουμε, λοιπόν, ότι η πιθανότητα πραγματοποίησης ακριβώς ενός συμβάντος είναι ανάλογη του μήκους του υποδιαστήματος, δηλαδή $p=\frac{\lambda t}{n}$ με $\lambda>0$. Παρατηρήστε ότι για μεγάλο n η πιθανότητα αυτή τείνει στο μηδέν, ενώ $np=\lambda t>0$. Τέλος, για να μπορούμε να μιλάμε για διωνυμική κατανομή θα πρέπει οι επαναλήψεις να είναι ανεξάρτητες, επομένως, η πραγματοποίηση ή όχι ενός συμβάντος σε ένα υποδιάστημα να είναι ανεξάρτητη από την πραγματοποίηση σε οποιοδήποτε άλλο.

Υπό τις παραπάνω υποθέσεις προκύπτει ότι ο αριθμός των συμβάντων στο $(0,t]$ ακολουθεί διωνυμική κατανομή $B\left(n,\frac{\lambda t}{n}\right)$, που είναι τέτοια ώστε για $n \to +\infty$, η πιθανότητα επιτυχίας $p = \frac{\lambda t}{n} \to 0$ με $np \to \lambda t$ και, επομένως, προσεγγίζεται από την κατανομή Poisson με παράμετρο λt, δηλαδή από την κατανομή με σπ:

$$\frac{e^{-\lambda t}(\lambda t)^x}{x!},\ x = 0,1,2,\ldots$$

Συνεπως, η κατανομή Poisson βρίσκει εφαρμογή και σε περιπτώσεις όπου σε ένα τυχαίο πείραμα μας ενδιαφέρει πόσες φορές εμφανίζεται ένα ενδεχόμενο σε χρονικό διάστημα t ή σε μήκος t ή σε επιφάνεια t ή σε όγκο t. Σε τέτοιες περιπτώσεις δεν έχουμε πλέον μια τ.μ., αλλ,ά σύμφωνα με τον επόμενο ορισμό, μια στοχαστική διαδικασία.

Ορισμός 4.7

Μία **στοχαστική διαδικασία** είναι μία συλλογή, μία οικογένεια τυχαίων μεταβλητών $\{X(t) : t \in T\}$, όπου t είναι μία παράμετρος που παίρνει τιμές σε ένα κατάλληλα ορισμένο σύνολο T. Δηλαδή για κάθε t η $X(t)$ είναι τ.μ. Στην περίπτωση που αντιπροσωπεύει τον συνολικό αριθμό των συμβάντων που έχουν πραγματοποιηθεί στο χρόνο t με $X(t) \geq 0$ και σύνολο δυνατών τιμών $\{0,1,2,\ldots\}$, καλείται **διαδικασία καταμέτρησης**.

Μια διαδικασία καταμέτρησης λέμε ότι είναι **διαδικασία Poisson** με μέσο ρυθμό λ στη μονάδα του χρόνου (όγκου, μήκους, ανάλογα), αν $X(0) = 0$ και, επιπλέον, πληροί τις ακόλουθες υποθέσεις:

1. Υπόθεση 1. (*Ιδιότητα Στατικότητας*). Η πιθανότητα πραγματοποίησης ακριβώς ενός γεγονότος σε ένα μικρό χρονικό διάστημα μήκους dt είναι κατά προσέγγιση ανάλογη του μήκους του διαστήματος, δηλαδή
$$P(X(dt) = 1) = \lambda \cdot dt + o(dt),$$
όπου $o(dt)$ χρησιμοποιείται για να δηλώσει μια συνάρτηση τέτοια, ώστε $\lim\limits_{dt\to 0} \frac{o(dt)}{dt} = 0$.
2. Υπόθεση 2. Η πιθανότητα να εμφανιστεί το ενδεχόμενο δύο ή περισσότερες φορές σε ένα μικρό χρονικό διάστημα dt είναι αμελητέα
$$P(X(dt) \geq 2) = o(dt).$$
3. Υπόθεση 3. (*Ιδιότητα Ανεξαρτησίας*). Ο αριθμός των γεγονότων σε ένα χρονικό διάστημα είναι ανεξάρτητος από τον αριθμό των γεγονότων σε ένα οποιοδήποτε άλλο μη επικαλυπτόμενο χρονικό διάστημα ή διαφορετικά οι αριθμοί των συμβάντων που λαμβάνουν χώρα σε μη επικαλυπτόμενα χρονικά διαστήματα είναι ανεξάρτητοι μεταξύ τους.

Οι τρεις παραπάνω υποθέσεις, κατ' ουσίαν, οδηγούν στην ακόλουθη ιδιότητα: για οποιοδήποτε χρονικό διάστημα $(s,s+t]$ με $s \geq 0$ και $t > 0$, ο αριθμός των γεγονότων σε αυτό $X(s+t) - X(s)$ ακολουθεί την κατανομή Poisson με παράμετρο λt, δηλαδή

$$P(X(s+t) - X(s) = x) = \frac{e^{\lambda t}(\lambda t)^x}{x!}, \quad x = 0,1,2,\ldots$$

Η απόδειξη αυτή ξεφεύγει από τους σκοπούς του παρόντος συγγράμματος, καθώς οδηγεί σε ένα σύστημα διαφορικών εξισώσεων και συνήθως γίνεται σε συγγράμματα με αντικείμενο μελέτης τη θεωρία των στοχαστικών διαδικασιών.

Πρόταση 4.15

Αν $X(t)$ είναι διαδικασία Poisson που περιγράφει τον αριθμό των συμβάντων σε ένα χρονικό διάστημα t (ή αντίστοιχα σε μια επιφάνεια εμβαδού ή όγκου t ή σε μια απόσταση t), τότε

$$P(X(t) = x) = \frac{e^{-\lambda t}(\lambda t)^x}{x!}, \quad x = 0,1,2,\ldots \tag{4.44}$$

όπου η παράμετρος λ εκφράζει τον μέσο αριθμό των συμβάντων που πραγματοποιούνται στη μονάδα του χρόνου (ή μήκους ή επιφάνειας ή όγκου) ή αλλιώς τον ρυθμό εμφάνισης.

Οι παραπάνω ιδιότητες έχουν κάνει την κατανομή και τη διαδικασία Poisson να εφαρμόζονται σε πλήθος επιστημονικών πεδίων και εφαρμογές. Ενδεικτικά παραδείγματα αφορούν τη μοντελοποίηση του αριθμού:

- των εναέριων βομβών που έπληξαν το Λονδίνο κατά τη διάρκεια του Β' Παγκοσμίου Πολέμου,
- των ασθενών που καταφθάνουν σε ένα νοσοκομείο κατά τη διάρκεια μιας ώρας,
- των φωτονίων λέιζερ που χτυπούν έναν ανιχνευτή σε ένα χρονικό διάστημα,
- των τηλεφωνικών κλήσεων, των αφίξεων πελατών σε ένα σύστημα εξυπηρέτησης, σε ένα χρονικό διάστημα,
- των μεταλλάξεων σε ένα τμήμα DNA ανά μονάδα μήκους,
- των βακτηρίων σε πλάκα άγαρ συγκεκριμένης επιφάνειας και
- των τερμάτων σε έναν ποδοσφαιρικό αγώνα

και ο κατάλογος είναι ανεξάντλητος, ενώ δεν πρέπει να ξεχνάμε και τις εφαρμογές της ως όριο της διωνυμικής κατανομής. Για περισσότερα παραδείγματα παραπέμπουμε στον ιστότοπο https://en.wikipedia.org/wiki/Poisson_distributionLaw_of_rare_events (ημερομηνία προσπέλασης: 1/3/2022) και στις εκεί αναφορές.

Στο σημείο αυτό, όμως, θα πρέπει να επισημάνουμε ότι παρότι η κατανομή Poisson μοντελοποιεί τον αριθμό των συμβάντων στη μονάδα του χρόνου (όγκου, μήκους, εμβαδού) θα πρέπει να είμαστε ιδιαίτερα προσεκτικοί γιατί θα πρέπει να ικανοποιούνται οι Υποθέσεις 1-3 (στατικότητα, ανεξαρτησία, αμελητέα πιθανότητα να συμβούν περισσότερα του ενός γεγονότα σε ένα μικρό χρονικό διάστημα μήκους dt). Επιπλέον, με βάση τις ιδιότητες της κατανομής Poisson θα πρέπει να χρησιμοποιείται στη μοντελοποίηση τυχαίων φαινομένων όπου η μέση τιμή ταυτίζεται πρακτικά με τη διακύμανση, καθώς στην κατανομή Poisson με παράμετρο λ, είναι $E(X) = Var(X) = \lambda$. Για παράδειγμα, ο αριθμός των πελατών σε ένα εστιατόριο δεν είναι λογικό να μοντελοποιείται από την κατανομή Poisson, καθώς ο ρυθμός άφιξης των πελατών δεν μπορεί να είναι σταθερός (Υπόθεση 1), αλλά και οι αφίξεις τους δεν μπορούν να θεωρηθούν ανεξάρτητες (Υπόθεση 3) μιας και είναι σύνηθες οι πελάτες να φτάνουν σε παρέες.

Από την άλλη, καθώς από τη σπ της κατανομής Poisson προκύπτει ότι η πιθανότητα να μην πραγματοποιηθεί ένα γεγονός στη μονάδα του χρόνου είναι ίση με $P(X = 0) = e^{-\lambda}$, και επομένως, $P(X \geq 1) = 1 - e^{-\lambda}$, η κατανομή Poisson δεν μπορεί να χρησιμοποιηθεί για τη μοντελοποίηση τυχαίων φαινομένων στα οποία είναι βέβαιη η πραγματοποίηση, εμφάνιση ενός γεγονότος[4]. Τέλος, με την ίδια αιτιολόγηση δεν μπορεί να χρησιμοποιηθεί για τη μοντελοποίηση τυχαίων φαινομένων με μεγάλη πιθανότητα μη εμφάνισης γεγονότων σε συγκεκριμένο χρονικό διάστημα[5].

[4] Για τη μοντελοποίηση τέτοιων τυχαίων φαινομένων έχει εισαχθεί στη βιβλιογραφία η λεγόμενη αποκομμένη στο μηδέν κατανομή Poisson (zero truncated Poisson).

[5] Στην περίπτωση αυτή, προτείνεται η χρήση των λεγόμενων zero-inflated models.

Παράδειγμα 4.8

Ο αριθμός των επιληπτικών επεισοδίων που έχει ένα παιδί με βεβαιωμένη επιληψία περιγράφεται ικανοποιητικά από μια κατανομή Poisson με μέσο ρυθμό 1.7 επεισόδια ανά έτος. Υπολογίστε την πιθανότητα να έχει κάποιο παιδί 3 επεισόδια σε 2 έτη. Υπολογίστε την πιθανότητα να μην έχει κάποιο επεισόδιο σε έναν χρόνο. Προσδιορίστε τον αναμενόμενο αριθμό επεισοδίων σε δέκα έτη.

Λύση Παραδείγματος 4.8

Ο αριθμός των επιληπτικών επεισοδίων ενός παιδιού με βεβαιωμένη επιληψία ακολουθεί την κατανομή Poisson με ρυθμό $\lambda = 1.7$ ανά έτος. Συμβολίζουμε με $X(2)$, $X(1)$ και $X(10)$ τις τυχαίες μεταβλητές που παριστάνουν τον αριθμό των επεισοδίων σε δύο, ένα και δέκα έτη αντίστοιχα. Τότε $X(2) \sim \mathscr{P}(\lambda = 2 \cdot 1.7 = 3.4)$, $X(1) \sim \mathscr{P}(\lambda = 1.7)$ και $X(10) \sim \mathscr{P}(\lambda = 10 \cdot 1.7 = 17)$. Ζητείται ο υπολογισμός των: $P(X(2) = 3)$, $P(X(1) = 0)$ και $E(X(10))$. Είναι

$$P(X(2) = 3) = \frac{e^{-3.4} \cdot 3.4^2}{3!} = 0.2186172,$$

και

$$P(X(1) = 0) = e^{-1.7} = 0.1826835,$$

ενώ από τη σχέση (4.41) έχουμε ότι: $E(X(10)) = 17$.
Οι τιμές των πιθανοτήτων $P(X(2) = 3)$ και $P(X(1) = 0)$ μπορούν επίσης να υπολογιστούν και με τις εντολές `dpois(3,3.4)` και `dpois(0,1.7)` της R, αντίστοιχα, οι οποίες επιστρέφουν τις προαναφερθείσες τιμές.

Παράδειγμα 4.9

Ένας κακοήθης όγκος σπάνιας μορφής εμφανίζεται σε 10 περιπτώσεις ανά 1 εκατομμύριο πληθυσμού. Σε μια περιοχή πληθυσμού 10000 ατόμων υπολογίστε την πιθανότητα να εμφανιστούν περισσότεροι από ένας ασθενής. Υπολογίστε την πιθανότητα ακριβώς και προσεγγιστικά. Αν σας έλεγαν ότι στην περιοχή αυτή εμφανίστηκαν 4 ασθενείς, τι θα λέγατε;

Λύση Παραδείγματος 4.9

Έστω X η τ.μ. που παριστάνει τον αριθμό των ασθενών που πάσχουν από τη σπάνιας μορφής κακοήθεια σε 10000 άτομα μιας περιοχής. Τότε, υποθέτοντας ότι η εμφάνιση κακοήθειας σε ένα άτομο είναι ανεξάρτητη από την εμφάνιση κακοήθειας σε οποιοδήποτε άλλο άτομο και η πιθανότητα εμφάνισης είναι ίδια σε όλα τα άτομα και ίση με $p = 10/10^6 = 10^{-5}$, έχουμε ότι $X \sim B(n = 10^4, p = 10^{-5})$. Παρατηρήστε ότι $n > 100$ και $np = 10^{-1}$ και, επομένως, η διωνυμική κατανομή μπορεί να προσεγγιστεί από την Poisson με παράμετρο $np = 10^{-1}$.
Ζητείται η εύρεση της $P(X > 1)$ ή, ισοδύναμα, της $1 - P(X \leq 1) = 1 - P(X = 0) - P(X = 1)$. Είναι τότε με τη διωνυμική κατανομή και χρησιμοποιώντας τη σχέση (4.5):

$$P(X = 0) = \binom{10^4}{0} (1 - 10^{-5})^{10^4} = 0.904837$$

και

$$P(X = 1) = \binom{10^4}{1} 10^{-5} (1 - 10^{-5})^{10^4 - 1} = 0.0904846$$

και, επομένως, $P(X > 1) = 1 - 0.904837 - 0.0904846 = 0.0046784$.
Εναλλακτικά, με την εντολή `pbinom(1,10000,10^(-5),lower.tail=FALSE)` της R παίρνουμε την τιμή 0.004678433, η οποία μας δίνει με μεγαλύτερη ακρίβεια την τιμή της ζητούμενης πιθανότητας.

Στη συνέχεια, θα υπολογιστεί η ζητούμενη πιθανότητα χρησιμοποιώντας την προσέγγιση της κατανομής Poisson με παράμετρο $\lambda = np = 10^{-1} = 0.1$. Τότε η πιθανότητα $P(X = 0)$ είναι κατά προσέγγιση ίση με $e^{-0.1} = 0.9048374$ και η $P(X = 1)$ είναι κατά προσέγγιση ίση με $e^{-0.1}10^{0.1} = 0.09048374$. Επομένως, η $P(X > 1)$ είναι κατά προσέγγιση ίση με $1 - 0.9048374 - 0.09048374 = 0.00467886$.
Εναλλακτικά, θα μπορούσε να χρησιμοποιηθεί και η εντολή `ppois(1,10^(-1),lower.tail=FALSE)` της R από όπου έχουμε ότι $P(X > 1) \simeq 0.00467884$, η οποία αποτελεί την τιμή της ζητούμενης πιθανότητας με μεγαλύτερη ακρίβεια σε σχέση με πριν.
Παρατηρούμε, λοιπόν, ότι η πιθανότητα εμφάνισης περισσότερων από 1 ασθενών είναι πάρα πολύ μικρή και, επομένως, η εμφάνιση 4 ασθενών στην περιοχή είναι ένα μη σύνηθες ενδεχόμενο, που θα πρέπει να προβληματίσει τις τοπικές αρχές.

Άσκηση Αυτοαξιολόγησης 4.12

Ο αριθμός των βακτηρίδιων που βρίσκονται σε ένα συγκεκριμένο υγρό περιγράφεται ικανοποιητικά από την κατανομή Poisson με μέσο αριθμό 10 βακτηρίδια ανά λίτρο. Να υπολογίσετε την πιθανότητα σε 500 ml του υγρού να υπάρχουν τουλάχιστον 4 βακτηρίδια.

Άσκηση Αυτοαξιολόγησης 4.13

Η πιθανότητα παρενεργειών σε ένα εμβόλιο είναι ίση με 0.004. Αν 1000 άτομα εμβολιάζονται, υπολογίστε την πιθανότητα 1 άτομο να έχει παρενέργειες. Υπολογίστε την πιθανότητα 6 άτομα να έχουν παρενέργειες. Ζητείται να βρεθούν οι πιθανότητες τόσο προσεγγιστικά όσο και ακριβώς.

Άσκηση Αυτοαξιολόγησης 4.14

Από προηγούμενες μελέτες είναι γνωστό ότι ο αριθμός των ατόμων που καταφθάνουν στα επείγοντα περιστατικά ενός νοσοκομείου της Δυτικής Ελλάδας περιγράφεται ικανοποιητικά από την κατανομή Poisson με μέσο ρυθμό 2 αφίξεις ανά 30 λεπτά. Υπολογίστε την πιθανότητα να έχουμε 5 αφίξεις ασθενών σε μία ώρα. Υπολογίστε την πιθανότητα σε 15 λεπτά να μην αφιχθεί κανένας ασθενής.

4.8 Ασκήσεις

Άσκηση 4.1 *Από προηγούμενες έρευνες είναι γνωστό ότι σε μια περιοχή της Ελλάδας η πιθανότητα κάποιος ενήλικας δημότης της να έχει μυωπία είναι 0.2 και κάθε άτομο έχει ή δεν έχει την πάθηση αυτή ανεξάρτητα από κάθε άλλο. Υπολογίστε την πιθανότητα σε μια τυχαία δειγματοληψία που διενεργεί ένας οφθαλμίατρος και επιλέγει 20 άτομα να βρεθούν 7 με αυτήν την πάθηση. Ποιος είναι ο αναμενόμενος αριθμός των ατόμων με μυωπία στο δείγμα που επιλέχθηκε;*

Άσκηση 4.2 *Ο Αποστόλης και ο Πολυχρόνης παίζουν το εξής παιχνίδι: επιλέγουν από μία τράπουλα, ο καθένας τυχαία, ένα φύλλο. Αν και οι δύο επιλέξουν καρό τότε σταματούν το παιχνίδι, ενώ διαφορετικά, αφού τοποθετήσουν το φύλλο που επέλεξαν στην τράπουλα και την ανακατέψουν, συνεχίζουν να παίζουν στον επόμενο γύρο με τον ίδιο τρόπο. Υπολογίστε την πιθανότητα να διαρκέσει το παιχνίδι τους ακριβώς 4 γύρους, καθώς και την πιθανότητα να διαρκέσει λιγότερο από 6.*

Άσκηση 4.3 *Η πιθανότητα ένα άτομο με συγκεκριμένο γονίδιο να νοσεί από μια ασθένεια είναι 0.7. Επιλέγονται τυχαία 12 άτομα που έχουν αυτό το γονίδιο. Υπολογίστε την πιθανότητα να νοσούν το πολύ 4, τουλάχιστον 7 και από 3 έως 6. Ένας γιατρός διεξάγει μια έρευνα και θέλει να δημιουργήσει μια δεκαμελή ομάδα ατόμων που έχουν το συγκεκριμένο γονίδιο και οι οποίοι νοσούν. Υπολογίστε την πιθανότητα να επιλέξει τουλάχιστον 16 άτομα από το μητρώο των ατόμων με το συγκεκριμένο γονίδιο για αυτό τον σκοπό.*

Άσκηση 4.4 *Από προηγούμενη εμπειρία είναι γνωστό ότι το 80% των ατόμων που έχουν μεικτή ασφάλιση στο αυτοκίνητό τους είναι νέοι οδηγοί. Επιλέγονται τυχαία 9 κάτοχοι μεικτής ασφάλισης αυτοκινήτου. Υπολογίστε την πιθανότητα ακριβώς 6 να είναι νέοι οδηγοί. Κάποιος επιλέγει τυχαία κατόχους μεικτής ασφάλισης μέχρι να βρει τον τρίτο νέο οδηγό. Υπολογίστε την πιθανότητα να χρειαστεί να επιλέξει τουλάχιστον 6 άτομα.*

Άσκηση 4.5 *Ένα κουτί περιέχει 100 βίδες εγχώριες και άλλες 200 ίδιου τύπου αλλά παραγόμενες σε χώρα του εξωτερικού. Αν τέσσερις βίδες επιλέγονται τυχαία από το κιβώτιο χωρίς επανατοποθέτηση, υπολογίστε την πιθανότητα όλες οι βίδες να είναι εγχώριες και την πιθανότητα τουλάχιστον δύο βίδες να είναι εγχώριες.*

Άσκηση 4.6 *Σε ένα κουτί υπάρχουν σφαίρες αριθμημένες από το 1 έως το 500. Επιλέγουμε στην τύχη μία σφαίρα. Υπολογίστε την πιθανότητα ο αριθμός της σφαίρας που επιλέχθηκε να διαιρείται με το 3. Επανατοποθετούμε τη σφαίρα μέσα και επαναλαμβάνουμε την επιλογή 10 φορές, επανατοποθετώντας κάθε φορά τη σφαίρα μέσα. Υπολογίστε την πιθανότητα να επιλεχθούν 4 σφαίρες ακριβώς με μονό αριθμό.*

Άσκηση 4.7 *Ένας έμπορος παραλαμβάνει 200 τρανζίστορ. Αποφασίζει να κρατήσει το φορτίο, αν βρει τουλάχιστον εννέα τρανζίστορ να λειτουργούν μεταξύ δέκα που θα επιλέξει στην τύχη. Αν το φορτίο περιέχει 50 ελαττωματικά, υπολογίστε την πιθανότητα ο έμπορος να κρατήσει το φορτίο.*

Άσκηση 4.8 *Από προηγούμενη μελέτη είναι γνωστό ότι ένας στους 5 επισκέπτες ενός ηλεκτρονικού βιβλιοπωλείου αγοράζει κάποιο από τα βιβλία του καταστήματος. Επιλέγονται τυχαία επισκέπτες της ιστοσελίδας και ζητείται να βρεθεί η πιθανότητα:*

1. *ο πρώτος πελάτης που θα αγοράσει κάποιο βιβλίο να είναι ο τρίτος,*
2. *να χρειαστεί να επιλέξουμε περισσότερους από δύο επισκέπτες μέχρι να βρούμε τον πρώτο που θα αγοράσει κάποιο βιβλίο,*
3. *ο δεύτερος πελάτης που θα αγοράσει κάποιο βιβλίο να είναι ο έβδομος επισκέπτης,*

4. *να χρειαστεί να επιλέξουμε τουλάχιστον τρεις επισκέπτες μέχρι να βρούμε τον δεύτερο που θα αγοράσει κάποιο βιβλίο.*

Άσκηση 4.9 *Η πιθανότητα κάποιος φοιτητής να καπνίζει είναι* 55%*. Επιλέγονται τυχαία 10 φοιτητές. Υπολογίστε την πιθανότητα να καπνίζουν τουλάχιστον 7. Ένας γιατρός επιθυμεί να διεξάγει ένα πρόγραμμα διακοπής του καπνίσματος. Πόσους φοιτητές πρέπει να επιλέξει μέχρι να βρει 10 καπνιστές; Αναφέρετε πλήρως όλες τις απαραίτητες υποθέσεις για την επίλυση της άσκησης.*

Άσκηση 4.10 *Έστω* $X \sim Geo(p)$*. Αποδείξτε ότι:*

$$E\left[X(X-1)\cdots(X-k+1)\right] = k!\frac{(1-p)^{k-1}}{p^k}, \quad k \in N_+.$$

Θα μπορούσε η παραπάνω σχέση να χρησιμοποιηθεί για την εύρεση της διακύμανσης ή και των ροπών;

Άσκηση 4.11 *Ένας Ιταλός τουρίστας που μιλάει μόνο τη μητρική του γλώσσα επισκέπτεται την Κέρκυρα. Από έρευνες είναι γνωστό ότι το* 30% *των κατοίκων της Κέρκυρας μιλούν ιταλικά. Ο τουρίστας θέλει να πάρει πληροφορίες σχετικά με ένα αξιοθέατο. Υπολογίστε την πιθανότητα να είναι το τρίτο άτομο που θα ρωτήσει το πρώτο που θα μιλάει ιταλικά. Το απόγευμα της ίδιας μέρας συναντά σε μία στάση λεωφορείου 4 άγνωστα άτομα μεταξύ τους. Υπολογίστε την πιθανότητα τουλάχιστον 1 από αυτά τα άτομα να μιλάει ιταλικά.*

Άσκηση 4.12 *Είναι γνωστό ότι ο Αποστόλης είναι πολύ καλός στο παιχνίδι της καλοθοσφαίρισης. Συμμετέχει λοιπόν σε ένα παιχνίδι όπου σουτάρει από τη γραμμή των ελεύθερων βολών συνεχώς μέχρι να αστοχήσει. Αν υποθέσουμε ότι κατά κανόνα χάνει το* 15% *των ελεύθερων βολών που επιχειρεί και ότι το αποτέλεσμα μιας προσπάθειάς του είναι ανεξάρτητο από το αποτέλεσμα οποιασδήποτε άλλης προσπάθειας, υπολογίστε την πιθανότητα ο Αποστόλης να τελειώσει το παιχνίδι σε 5 ελεύθερες βολές και την πιθανότητα να χρειαστούν περισσότερες από 7 ελεύθερες βολές για να ολοκληρωθεί το παιχνίδι.*

Άσκηση 4.13 *Σε μια αίθουσα κινηματογράφου υπάρχουν 15 γυναίκες και 25 άντρες που παρακολουθούν μια κινηματογραφική προβολή. Λόγω ενός διαγωνισμού θα επιλεχθούν τυχαία 7 άτομα τα οποία θα κερδίσουν ένα εισιτήριο για επόμενη προβολή. Υπολογίστε την πιθανότητα να επιλεχθούν 4 γυναίκες ακριβώς.*

Άσκηση 4.14 *Μια εταιρεία κατασκευάζει τετράγωνες πλάκες πεζοδρομίου. Η πιθανότητα μια τυχαία επιλεγμένη πλάκα να έχει εμβαδόν μικρότερο από το ονομαστικό της είναι 0.01. Υπολογίστε:*

1. *την πιθανότητα ανάμεσα σε 100 τυχαία επιλεγμένες πλάκες ο αριθμός των πλακών με εμβαδόν μικρότερο από το ονομαστικό τους να είναι το πολύ ίσος με ένα και*
2. *τον αναμενόμενο αριθμό πλακών με εμβαδόν μικρότερο από το ονομαστικό ανάμεσα σε 100 τυχαία επιλεγμένες πλάκες.*

Άσκηση 4.15 *Ένας εισαγωγέας παραλαμβάνει μια μεγάλη παρτίδα προϊόντων, από τα οποία επιλέγει τυχαία 500 για να επιθεωρήσει. Αν σε αυτά βρεθούν 2 ή λιγότερα ελαττωματικά, τότε ολόκληρη η παρτίδα γίνεται αποδεκτή. Αν υποθέσουμε ότι το πραγματικό ποσοστό των ελαττωματικών προϊόντων είναι 0.001, υπολογίστε την πιθανότητα να γίνει αποδεκτή η παρτίδα χωρίς κάποιον επιπλέον έλεγχο. Προσδιορίστε την ακριβή και προσεγγιστική τιμή της πιθανότητας.*

Άσκηση 4.16 *Μια εταιρεία κατασκευάζει ένα προϊόν, με πιθανότητα να είναι κάποιο από αυτά ελαττωματικό 0.01. Σε ένα δείγμα 400 προϊόντων υπολογίστε την πιθανότητα να εμφανιστούν ακριβώς 5 ελαττωματικά προϊόντα. Υπόδειξη: υπολογίστε την ακριβή και προσεγγιστική τιμή της πιθανότητας.*

Άσκηση 4.17 *Κατά τη διάρκεια μιας εργάσιμης μέρας διέρχονται από ένα συγκεκριμένο σημείο της περιφερειακής οδού των Ιωαννίνων 250 οχήματα την ώρα. Υπολογίστε την πιθανότητα ότι από το συγκεκριμένο σημείο δεν θα περάσει κάποιο όχημα σε χρονικό διάστημα 3 λεπτών. Ποιος είναι ο αναμενόμενος αριθμός των διερχόμενων οχημάτων σε χρονικό διάστημα 2 λεπτών και ποια η πιθανότητα εμφάνισής του;*

Άσκηση 4.18 *Τα τελευταία 200 έτη έχουν γίνει στον ελλαδικό χώρο 83 σεισμικές δονήσεις με ένταση 6 ή και μεγαλύτερη στην κλίμακα Richter. Αν ο αριθμός των σεισμών σε ένα χρονικό διάστημα μοντελοποιείται από την κατανομή Poisson, υπολογίστε την πιθανότητα σε έναν χρόνο να συμβούν ακριβώς 2 σεισμοί με ένταση 6 ή και μεγαλύτερη της κλίμακας Richter.*

Άσκηση 4.19 *Ένας βοτανολόγος μελετά ένα συγκεκριμένο σπάνιο είδος λουλουδιού, που ξέρει ότι ανθεί στην περιοχή των Ζαγοροχωρίων και ότι κατά μέσο όρο εμφανίζονται 3 λουλούδια σε έκταση 8 στρεμμάτων. Κάνοντας κατάλληλες υποθέσεις, υπολογίστε την πιθανότητα ότι σε μια έκταση 10 στρεμμάτων θα βρει 4 λουλούδια και την πιθανότητα ότι σε μια έκταση 5 στρεμμάτων θα βρει τουλάχιστον ένα λουλούδι.*

Άσκηση 4.20 *Μια εταιρεία έχει διαπιστώσει ότι περιστασιακά πρέπει να κάνει επαναφορά (reset) στη μηχανή παραγωγής ενός προϊόντος και ότι κατά μέσο όρο απαιτούνται 6 επαναφορές το δίμηνο. Κάνοντας κατάλληλες υποθέσεις, υπολογίστε την πιθανότητα να χρειαστούν 10 επαναφορές σε ένα δίμηνο, λιγότερες από 2 επαναφορές σε έναν μήνα, περισσότερες από 4 επαναφορές σε ένα δίμηνο.*

Άσκηση 4.21 *Είναι γνωστό ότι μόλις το 0.2% των ατόμων που υποβάλλονται σε μια λαπαρασκοπική επέμβαση έχουν μετεγχειρητικές επιπλοκές. Αν 1500 άτομα υποβάλλονται σε αυτήν την επέμβαση υπολογίστε την πιθανότητα το πολύ 2 άτομα να εμφανίσουν επιπλοκές. Αν σας έλεγαν ότι παρατηρήθηκαν 4 άτομα με επιπλοκές, τι θα λέγατε;*

Άσκηση 4.22 *Ένας εκδότης ξέρει ότι ένα βιβλίο 250 σελίδων έχει κατά μέσο όρο 450 τυπογραφικά λάθη. Υπολογίστε την πιθανότητα σε μια σελίδα που θα επιλεχθεί στην τύχη να μην υπάρχουν λάθη. Υπολογίστε την πιθανότητα να υπάρχουν περισσότερα από 2.*

Άσκηση 4.23 *Ένας ερασιτέχνης δρομέας αποστάσεων έχει κατά μέσο όρο 3 μυϊκούς τραυματισμούς ανά έτος. Κάνοντας κατάλληλες υποθέσεις, υπολογίστε την πιθανότητα να έχει 4 μυϊκούς τραυματισμούς στα δύο χρόνια, λιγότερους από 2 μυϊκούς τραυματισμούς σε έναν χρόνο, έναν ακριβώς μυϊκό τραυματισμό σε ένα εξάμηνο.*

Άσκηση 4.24 *Διάφορες ατέλειες εμφανίζονται συχνά κατά το μήκος συρμάτων. Πιο συγκεκριμένα, σε καθένα μέτρο εμφανίζονται κατά μέσο όρο 1.2 ατέλειες.*

1. *Να υπολογιστεί η πιθανότητα σε ένα τυχαία επιλεγμένο κομμάτι σύρματος μήκους ενός μέτρου να βρεθεί τουλάχιστον μία ατέλεια.*
2. *Να υπολογιστεί η πιθανότητα σε ένα τυχαία επιλεγμένο κομμάτι σύρματος μήκους δύο μέτρων να μην βρεθεί καμία ατέλεια.*
3. *Να υπολογιστεί η πιθανότητα, αν επιλέξουμε τυχαία 5 κομμάτια σύρματος μήκους δύο μέτρων, να μην βρεθεί καμία ατέλεια σε κανένα από αυτά.*

Άσκηση 4.25 *Η εμφάνιση τροχαίων ατυχημάτων σε ένα τμήμα αυτοκινητόδρομου στις πέντε εργάσιμες μέρες της εβδομάδας κατά τη διάρκεια των ωρών αυξημένου κυκλοφοριακού φόρτου, 1:30 μ.μ. με 4:30 μ.μ., ακολουθεί κατανομή Poisson με μέσο ρυθμό 2.4 ατυχήματα την ώρα. Να υπολογιστούν οι πιθανότητες:*

1. *Να συμβούν τουλάχιστον 4 ατυχήματα μεταξύ 2:00 μ.μ. και 3:00 μ.μ. μιας εργάσιμης μέρας.*
2. *Να συμβεί το πολύ ένα ατύχημα μεταξύ 2:00 μ.μ. και 4:00 μ.μ. μιας εργάσιμης μέρας.*
3. *Να μην συμβεί ατύχημα αύριο από τις 1:30 μ.μ. μέχρι τις 3:00 μ.μ.*

Άσκηση 4.26 *Ο αριθμός των καταιγίδων που προκαλούν υπερχείλιση ενός ποταμού περιγράφεται από μια Poisson διαδικασία με μέσο αριθμό τέτοιων καταιγίδων 1 κάθε 10 χρόνια.*

1. *Υπολογίστε την πιθανότητα να υπάρξουν το πολύ δύο τέτοιες καταιγίδες σε μια περίοδο 10 ετών.*
2. *Υπολογίστε την πιθανότητα να υπάρξουν ακριβώς δύο τέτοιες καταιγίδες σε μια περίοδο 20 ετών.*

Άσκηση 4.27 *Στην πόλη των Ιωαννίνων ένα άτομο στα ογδόντα είναι χορτοφάγο. Αν 200 άτομα επιλέγονται στην τύχη υπολογίστε κατά προσέγγιση την πιθανότητα ότι θα υπάρχουν τουλάχιστον 5 χορτοφάγοι. Πόσα άτομα πρέπει να επιλεχθούν τυχαία, έτσι ώστε η πιθανότητα να υπάρχει τουλάχιστον ένας χορτοφάγος να είναι μεγαλύτερη ή ίση με 0.9; Υπόδειξη: Χρησιμοποιήστε την R για την επίλυση της άσκησης.*

4.9 Απαντήσεις στις ασκήσεις αυτοαξιολόγησης

Λύση Άσκησης Αυτοαξιολόγησης 4.1

Έστω X η τ.μ. που παριστάνει τον αριθμό που επιλέχθηκε από το σύνολο $S = \{1,2,3,\ldots,20\}$. Είναι $X \sim DUnif(1,20)$ με $P(X = x) = \frac{1}{20}$ για $x \in S$. Είναι τότε

$$P(X > 12) = P(X \geq 13) = \sum_{x=13}^{20} p_X(x) = \sum_{x=13}^{20} \frac{1}{20} = \frac{8}{20}.$$

Έπειτα μας ζητείται η δεσμευμένη πιθανότητα $P(X > 12|X < 18 \cap X \text{ περιττός})$. Από τον ορισμό της δεσμευμένης πιθανότητας προκύπτει:

$$\begin{aligned} P(X > 12|X < 18 \cap X \text{ περιττός}) &= \frac{P(\{X > 12\} \cap \{X < 18\} \cap \{X \text{ περιττός}\})}{P(\{X < 18\} \cap \{X \text{ περιττός}\})} \\ &= \frac{P(\{X = 13\} \cup \{X = 15\} \cup \{X = 17\})}{P(\{X = 1\} \cup \{X = 3\} \cup \ldots \cup \{X = 17\})} \\ &= \frac{3/20}{9/20} = \frac{1}{3}. \end{aligned}$$

Εναλλακτικά, θα μπορούσαμε να σκεφτούμε ότι ο δειγματικός χώρος έχει περιοριστεί στο σύνολο $\{1,3,5,7,9,11,13,15,17\}$ και, καθώς θέλουμε αριθμό μεγαλύτερο του 12, είναι μόνο τρεις οι ευνοϊκές περιπτώσεις στις 9 δυνατές και προκύπτει το ζητούμενο αποτέλεσμα.

Λύση Άσκησης Αυτοαξιολόγησης 4.2

Από τον ορισμό της ροπογεννήτριας συνάρτησης και λαμβάνοντας υπόψη τη σπ της $X \sim DUnif(a,b)$ έχουμε (θυμηθείτε ότι $b = a + n - 1$)

$$\begin{aligned} M_X(t) &= \sum_{x=a}^{b} e^{tx} \frac{1}{n} = \frac{1}{n}\left(e^{ta} + e^{ta+t} + \ldots + e^{ta+(n-1)t}\right) \\ &= \frac{1}{n} e^{ta}\left(1 + e^{t} + \ldots + e^{(n-1)t}\right) = \frac{1}{n} e^{ta} \sum_{i=0}^{n-1} e^{ti}. \end{aligned}$$

Το τελευταίο άθροισμα είναι ουσιαστικά το άθροισμα των n πρώτων όρων αριθμητικής προόδου με πρώτο όρο τη μονάδα, λόγο e^t και τελευταίο όρο $e^{t(n-1)}$. Λαμβάνοντας υπόψη τη σχέση (Β΄.2) του Παραρτήματος Β΄, είναι:

$$M_X(t) = \frac{1}{n} e^{ta} \frac{1 - e^{tn}}{1 - e^{t}} = \frac{e^{ta} - e^{t(a+n)}}{n(1 - e^{t})} = \frac{e^{at} - e^{(b+1)t}}{n(1 - e^{t})}.$$

Λύση Άσκησης Αυτοαξιολόγησης 4.3

1. Έστω X η τ.μ. που παριστάνει τον αριθμό των επιτυχιών στις 15 ρίψεις του ζαριού, με επιτυχία να θεωρείται η εμφάνιση της ένδειξης 3. Είναι τότε $X \sim B\left(15, p = \frac{1}{6}\right)$. Ζητείται η $P(X \leq 4)$, η οποία, στη συνέχεια, θα υπολογιστεί, αναλυτικά αλλά και με τη βοήθεια της R.

Η $P(X \leq 4)$ υπολογίζεται αναλυτικά ως ακολούθως

$$\begin{aligned}
P(X \leq 4) &= \sum_{x=0}^{4} \binom{15}{x} \left(\frac{1}{6}\right)^x \left(\frac{5}{6}\right)^{15-x} \\
&= \binom{15}{0} \left(\frac{1}{6}\right)^0 \left(\frac{5}{6}\right)^{15} + \binom{15}{1} \left(\frac{1}{6}\right)^1 \left(\frac{5}{6}\right)^{14} \\
&\quad + \binom{15}{2} \left(\frac{1}{6}\right)^2 \left(\frac{5}{6}\right)^{13} + \binom{15}{3} \left(\frac{1}{6}\right)^3 \left(\frac{5}{6}\right)^{12} + \binom{15}{4} \left(\frac{1}{6}\right)^4 \left(\frac{5}{6}\right)^{11} \\
&= \frac{5^{15}}{6^{15}} + \frac{15!}{1!14!} \frac{5^{14}}{6^{15}} + \frac{15!}{2!13!} \frac{5^{13}}{6^{15}} + \frac{15!}{3!12!} \frac{5^{12}}{6^{15}} + \frac{15!}{4!11!} \frac{5^{11}}{6^{15}} \\
&= \frac{5^{15}}{6^{15}} + 15 \cdot \frac{5^{14}}{6^{15}} + \frac{14 \cdot 15}{2} \frac{5^{13}}{6^{15}} + \frac{13 \cdot 14 \cdot 15}{6} \frac{5^{12}}{6^{15}} + \frac{12 \cdot 13 \cdot 14 \cdot 15}{24} \frac{5^{11}}{6^{15}} \\
&= 0.06490547 + 0.1947164 + 0.272603 + 0.2362559 + 0.1417535 \\
&= 0.9102343
\end{aligned}$$

Στο ίδιο αποτέλεσμα οδηγούμαστε και με την R, χρησιμοποιώντας την εντολή `pbinom(4,15,1/6,lower.tail=TRUE)`.

2. Έστω Y η τ.μ. που παριστάνει τον αριθμό των επιτυχιών στις 15 ρίψεις του ζαριού, με επιτυχία να θεωρείται η εμφάνιση ένδειξης μικρότερης ή ίσης με 4. Είναι τότε $Y \sim B\left(15, p = \frac{4}{6}\right)$. Ζητείται η $P(Y \leq 5)$, η οποία υπολογίζεται με τον ακόλουθο τρόπο:

$$\begin{aligned}
P(Y \leq 5) &= \sum_{y=0}^{5} \binom{15}{y} \left(\frac{4}{6}\right)^y \left(\frac{2}{6}\right)^{15-y} \\
&= \binom{15}{0} \left(\frac{4}{6}\right)^0 \left(\frac{2}{6}\right)^{15} + \binom{15}{1} \left(\frac{4}{6}\right)^1 \left(\frac{2}{6}\right)^{14} + \binom{15}{2} \left(\frac{4}{6}\right)^2 \left(\frac{2}{6}\right)^{13} \\
&\quad + \binom{15}{3} \left(\frac{4}{6}\right)^3 \left(\frac{2}{6}\right)^{12} + \binom{15}{4} \left(\frac{4}{6}\right)^4 \left(\frac{2}{6}\right)^{11} + \binom{15}{5} \left(\frac{4}{6}\right)^5 \left(\frac{2}{6}\right)^{10} \\
&= \frac{2^{15}}{6^{15}} + \frac{15!}{1!14!} \frac{4 \cdot 2^{14}}{6^{15}} + \frac{15!}{2!13!} \frac{4^2 \cdot 2^{13}}{6^{15}} \\
&\quad + \frac{15!}{3!12!} \frac{4^3 \cdot 2^{12}}{6^{15}} + \frac{15!}{4!11!} \frac{4^4 \cdot 2^{11}}{6^{15}} \frac{15!}{5!10!} \frac{4^5 \cdot 2^{10}}{6^{15}} \\
&= \frac{2^{15}}{6^{15}} + 60 \cdot \frac{2^{14}}{6^{15}} + \frac{14 \cdot 15}{2} \frac{4^2 \cdot 2^{13}}{6^{15}} + \frac{13 \cdot 14 \cdot 15}{6} \frac{4^3 \cdot 2^{12}}{6^{15}} \\
&\quad + \frac{12 \cdot 13 \cdot 14 \cdot 15}{24} \frac{4^4 \cdot 2^{11}}{6^{15}} + \frac{11 \cdot 12 \cdot 13 \cdot 14 \cdot 15}{120} \frac{4^5 \cdot 2^{10}}{6^{15}} \\
&= 6.969172 \cdot 10^{-8} + 2.090752 \cdot 10^{-6} + 2.927052 \cdot 10^{-6} \\
&\quad + 0.0002536779 + 0.001522067 + 0.006697095 \\
&= 0.008504271.
\end{aligned}$$

Το ίδιο αποτέλεσμα λαμβάνεται και με τη βοήθεια της R εκτελώντας την εντολή `pbinom(5,15,4/6,lower.tail=TRUE)`.

Στα ίδια αποτελέσματα (με κάποια σφάλματα ακριβείας) θα καταλήγαμε, αν χρησιμοποιούσαμε τους Πίνακες της διωνυμικής κατανομής (επιβεβαιώστε το!).

Λύση Άσκησης Αυτοαξιολόγησης 4.4

Έστω X η τ.μ. που παριστάνει τον αριθμό των τροχαίων ατυχημάτων που αποδίδονται στην κατανάλωση αλκοόλ στα 5 ατυχήματα που έγιναν. Στο σημείο αυτό, υποθέτουμε ότι η πιθανότητα εμφάνισης ατυχήματος που αποδίδεται στην κατανάλωση αλκοόλ παραμένει σταθερή (δεν διαφοροποιείται π.χ. ανάλογα με την ημέρα ή ώρα) και ότι κάθε ατύχημα είναι ανεξάρτητο από οποιοδήποτε άλλο. Έχουμε ότι $X \sim B(5, 0.35)$. Ζητείται να προσδιοριστεί η πιθανότητα $P(X \geq 3) = 1 - P(X < 3)$. Αφού η τ.μ. X μπορεί να λάβει τις τιμές $0, 1, 2, 3, 4, 5$, η πιθανότητα $P(X < 3)$ εκφράζεται ως $P(X \leq 2)$. Επομένως, έχουμε ότι $P(X \geq 3) = 1 - P(X \leq 2)$, δηλαδή $P(X \geq 3) = 1 - (P(X = 0) + P(X = 1) + P(X = 2))$. Καθώς

$$P(X = 0) = \binom{5}{0} \cdot 0.35^0 \cdot 0.65^5 = 0.1160291,$$

$$P(X = 1) = \binom{5}{1} \cdot 0.35^1 \cdot 0.65^4 = 5 \cdot 0.35 \cdot 0.65^4 = 0.3123859,$$

και

$$P(X = 2) = \binom{5}{2} \cdot 0.35^2 \cdot 0.65^3 = \frac{5!}{2!3!} 5 \cdot 0.35^2 \cdot 0.65^3$$
$$= \frac{4 \cdot 5}{2} \cdot 0.35^2 \cdot 0.65^3 = 0.3364156$$

έχουμε ότι $P(X \geq 3) = 1 - (0.1160291 + 0.3123859 + 0.3364156) = 0.2351694$.
Το ίδιο αποτέλεσμα λαμβάνεται και με τη βοήθεια της R εκτελώντας την εντολή `pbinom(2,5,0.35, lower.tail=FALSE)` αλλά και χρησιμοποιώντας τους Πίνακες της διωνυμικής κατανομής (με κάποια σφάλματα ακρίβειας - επιβεβαιώστε το!).

Λύση Άσκησης Αυτοαξιολόγησης 4.5

Έστω X η τ.μ. που παριστάνει τον αριθμό των εύστοχων βολών στις 10 που επιχείρησε ο Αποστόλης. Από την εκφώνηση του προβλήματος έχουμε ότι η πιθανότητα μια βολή να είναι εύστοχη παραμένει αμετάβλητη, με πιθανότητα επιτυχίας $p = 0.85$. Επιπλέον, έχουμε ότι οι επαναλήψεις αυτού του πειράματος είναι ανεξάρτητες μεταξύ τους, καθώς το αποτέλεσμα σε μία βολή δεν επηρεάζει το αποτέλεσμα σε οποιαδήποτε άλλη. Επομένως, $X \sim B(n = 10, p = 0.85)$. Ζητείται η $P(6 \leq X \leq 8) = P(X = 6) + P(X = 7) + P(X = 8)$, όπου

$$P(X = 6) = \binom{10}{6} \cdot 0.85^6 \cdot 0.15^4 = \frac{10!}{6!4!} 0.85^6 \cdot 0.15^4$$
$$= \frac{7 \cdot 8 \cdot 9 \cdot 10}{1 \cdot 2 \cdot 3 \cdot 4} 0.85^6 \cdot 0.15^4 = 0.04009571,$$

$$P(X = 7) = \binom{10}{7} \cdot 0.85^7 \cdot 0.15^3 = \frac{10!}{7!3!} 0.85^7 \cdot 0.15^3$$
$$= \frac{8 \cdot 9 \cdot 10}{1 \cdot 2 \cdot 3} 0.85^7 \cdot 0.15^3 = 0.1298337,$$

και

$$P(X = 8) = \binom{10}{8} \cdot 0.85^8 \cdot 0.15^2 = \frac{10!}{8!2!} 0.85^8 \cdot 0.15^2$$
$$= \frac{9 \cdot 10}{1 \cdot 2} 0.85^8 \cdot 0.15^2 = 0.2758967.$$

Άρα $P(6 \leq X \leq 8) = 0.04009571 + 0.1298337 + 0.2758967 = 0.4458261$.
Στην R, παρατηρώντας ότι $P(6 \leq X \leq 8) = P(X \leq 8) - P(X \leq 5)$, μπορούμε να χρησιμοποιήσουμε την εντολή `pbinom(8,10,0.85,lower.tail=TRUE)-pbinom(5,10,0.85, lower.tail=TRUE)`, με τη βοήθεια της οποίας λαμβάνουμε την ίδια ακριβώς τιμή με το παραπάνω αποτέλεσμα.
Τέλος, ζητείται να προσδιορίσουμε τον μικρότερο ακέραιο αριθμό c που είναι τέτοιος, ώστε $P(X \geq c) \geq 0.8$. Με παρόμοιο τρόπο, όπως προηγουμένως, έχουμε ότι $P(X = 10) = 0.1968744$ και $P(X = 9) = 0.3474254$. Επομένως, $P(X \geq 8) = 0.8201965 > 0.8$ και, επομένως, η ζητούμενη τιμή είναι το 8. Στην ίδια τιμή μπορούμε να καταλήξουμε άμεσα χρησιμοποιώντας την εντολή `qbinom(0.8,10,0.85,lower.tail=FALSE)` στην R.

Λύση Άσκησης Αυτοαξιολόγησης 4.6

Έστω Y η τ.μ. που παριστάνει τον αριθμό των αποτυχιών σε μια διαδικασία ανεξάρτητων επαναλήψεων δοκιμών Bernoulli μέχρις ότου εμφανιστεί η πρώτη επιτυχία με σύνολο δυνατών τιμών $\{0,1,2,3,...\}$. Τότε

$$p_Y(y) = P(Y = y) = P(\overbrace{\text{A A A}\cdots\text{A}}^{y}\,\text{E}),\ y = 0,1,2,...$$

$$= \overbrace{P(A)\cdot P(A)...P(A)}^{y}\cdot P(E) = (1-p)^y p,\ y = 0,1,2,...$$

Εναλλακτικά, παρατηρήστε ότι $Y = X - 1$, επομένως,

$$p_Y(y) = P(Y = y) = P(X - 1 = y) = P(X = y + 1) = p_X(y + 1),\ y = 0,1,2,...$$

και το ζητούμενο προκύπτει από τη σχέση (4.12). Τέλος, $\sum_{y=0}^{\infty}(1-p)^y p = p\frac{1}{1-(1-p)} = 1$, χρησιμοποιώντας τη σχέση που μας δίνει το άθροισμα άπειρων διαδοχικών όρων φθίνουσας γεωμετρικής προόδου με λόγο $1 - p$. Σχετικά με την ασκ ένας άμεσος τρόπος είναι να χρησιμοποιηθεί η σχέση $F_Y(y) = F_X(y + 1)$ με X την τ.μ. με ασκ που δίνεται στη σχέση (4.13).

Λύση Άσκησης Αυτοαξιολόγησης 4.7

Έστω X η τ.μ. που παριστάνει τον αριθμό των επαναλήψεων του παιχνιδιού της αμερικάνικης ρουλέτας μέχρι να εμφανιστεί για πρώτη φορά κόκκινο. Επομένως, μπορεί να θεωρηθεί δοκιμή Bernoulli με τα δυνατά αποτελέσματα να είναι κόκκινο και όχι κόκκινο. Κάθε επανάληψη του τυχαίου πειράματος είναι ανεξάρτητη από οποιαδήποτε άλλη και η πιθανότητα επιτυχίας είναι αμετάβλητη και ίση με $p = 18/38$. Επομένως, $X \sim Geo(p = 18/38)$. Ζητείται να υπολογιστεί η $P(X > 10 + k|X > 10)$. Όμως, από την ιδιότητα της αμνησίας της γεωμετρικής κατανομής ισχύει ότι: $P(X > 10 + k|X > 10) = P(X > k)$. Επομένως

$$P(X > 10 + k|X > 10) = 1 - P(X \leq k) = 1 - F_X(k) = (1-p)^k,$$

όπου χρησιμοποιήθηκε η ασκ της γεωμετρικής κατανομής που έχει προσδιοριστεί στη σχέση (4.14).

Λύση Άσκησης Αυτοαξιολόγησης 4.8

Έστω X η τ.μ. που παριστάνει τον αριθμό των ατόμων που επιλέγει ο οφθαλμίατρος μέχρι να βρεθεί ο πρώτος ενήλικας δημότης με μυωπία. Με βάση τις υποθέσεις της εκφώνησης είναι $X \sim Geo(p = 0.2)$ και ζητείται να υπολογιστεί η πιθανότητα $P(X = 4)$. Είναι $P(X = 4) = (1 - 0.2)^3 \cdot 0.2 = 0.1024$ ή, εναλλακτικά, με την εντολή `dgeom(3,0.2)` στην R, η οποία επιστρέφει την προαναφερθείσα τιμή.

Λύση Άσκησης Αυτοαξιολόγησης 4.9

Η ρίψη των δύο ζαριών μπορεί να θεωρηθεί ότι έχει δύο δυνατά αποτελέσματα: εμφάνιση ενδείξεων με άθροισμα μεγαλύτερο από 7 ή όχι, με πιθανότητα επιτυχίας p ίση με την πιθανότητα εμφάνισης αθροίσματος από 8-12. Είναι τότε, από τον κλασικό ορισμό της πιθανότητας, $p = 15/36$. Επιπρόσθετα, κάθε επανάληψη είναι ανεξάρτητη από οποιαδήποτε άλλη.

1. Έστω X η τ.μ. που παριστάνει τον αριθμό των φορών που φέρνουμε άθροισμα μεγαλύτερο από 7 στις 10 ρίψεις. Οι δυνατές τιμές της τ.μ. X είναι $\{0,1,2,\ldots,10\}$ και είναι $X \sim B(n = 10, p = 15/36)$. Ζητείται η πιθανότητα $P(X = 7)$ που μπορεί να υπολογιστεί από τη σχέση (4.5) για $x = 7$, $n = 10$, $p = 15/36$. Είναι, λοιπόν,

$$\begin{aligned} P(X = 7) &= \binom{10}{7} \cdot (15/36)^7 \cdot (1 - 15/36)^3 \\ &= \frac{10!}{7!3!}(15/36)^7 \cdot (1 - 15/36)^3 = \frac{8 \cdot 9 \cdot 10}{6}(15/36)^7 \cdot (1 - 15/36)^3 \\ &= 0.05193414 \end{aligned}$$

   Η παραπάνω τιμή μπορεί να υπολογιστεί και με την εντολή `dbinom(7,10,15/36)` της R.

2. Ζητείται η $E(X) = np = 4.166667$, άρα αναμένεται περίπου 4 φορές να φέρουμε άθροισμα μεγαλύτερο από 7 στις 10 ρίψεις των δύο ζαριών.

Λύση Άσκησης Αυτοαξιολόγησης 4.10

1. Το ενδιαφέρον εδώ επικεντρώνεται στην εμφάνιση για τρίτη φορά ενδείξεων με άθροισμα μεγαλύτερο από 7. Έστω Y η τ.μ. που παριστάνει το πλήθος αυτών των επαναλήψεων, με δυνατές τιμές $\{3,4,5,\ldots\}$. Επομένως, η $Y \sim NB(r = 3, p = 15/36)$ και θέλουμε να υπολογίσουμε την $P(Y = 11)$. Η πιθανότητα αυτή μπορεί να υπολογιστεί από τη σχέση (4.25) για $x = 11$, $r = 3$ και $p = 15/36$ και είναι:

$$\begin{aligned} P(Y = 11) &= \binom{10}{2} \cdot (15/36)^3 \cdot (21/36)^8 \\ &= \frac{10!}{8!2!}(15/36)^3 \cdot (21/36)^8 = 45 \cdot (15/36)^3 \cdot (21/36)^8 \\ &= 0.04364285. \end{aligned}$$

   Η παραπάνω τιμή μπορεί να υπολογιστεί και με την εντολή `dnbinom(11-3,3,15/36)` της R.

2. Ζητείται ο προσδιορισμός του μικρότερου ακέραιου αριθμού, έστω x, που είναι τέτοιος, ώστε $P(Y \le x) \ge 0.15$ ή, ισοδύναμα, του μικρότερου ακέραιου αριθμού για τον οποίο είναι $F_Y(x) \ge 0.15$, όπου $Y \sim NB(r = 3, p = 15/36)$ με ασκ $F_Y(\cdot)$ που προσδιορίστηκε στη σχέση (4.27). Προφανώς, θα πρέπει $x \ge 3$, καθώς

$$P(Y = 3) = (15/36)^3 = 0.07233796.$$

   Από την άλλη, έχουμε ότι:

$$P(Y = 4) = \binom{3}{2} \cdot (15/36)^3 \cdot (21/36)^1 = 3 \cdot (15/36)^3 \cdot (21/36) = 0.1265914,$$

   και άρα $P(Y \le 4) = 0.07233796 + 0.1265914 = 0.1989294 \ge 0.15$, αφού $P(Y \le 4) = P(Y \le 3) + P(Y = 4)$. Επομένως, η ζητούμενη τιμή είναι το 4.

   Στο ίδιο αποτέλεσμα μπορούμε να καταλήξουμε εκτελώντας στην R την εντολή `qnbinom(0.15,3,15/36,lower.tail=TRUE)+3`. Χρησιμοποιώντας την εντολή `pnbinom(4-3,3,15/36,lower.tail=TRUE)`, παρατηρούμε ότι $P(Y \le 4) = 0.1989294$.

Λύση Άσκησης Αυτοαξιολόγησης 4.11

Τα τσιπ των υπολογιστών διακρίνονται σε δύο ομάδες: τα ελαττωματικά και τα μη ελαττωματικά. Το σύνολο των τσιπ είναι $N = 500$ με $N_1 = 10$ και $N_2 = 490$. Επιλέγονται τυχαία $n = 50$ το πλήθος από αυτά τα προϊόντα χωρίς επανατοποθέτηση. Έστω X η τ.μ. που παριστάνει τον αριθμό των ελαττωματικών προϊόντων στα 50 που επιλέγονται τυχαία και χωρίς επανατοποθέτηση από τα 500 συνολικά. Είναι $X \sim Hg(10,490,50)$ με σπ που δίνεται από τη σχέση (4.32) για $N_1 = 10$, $N_2 = 490$ και $n = 50$. Ο αναμενόμενος αριθμός των ελαττωματικών τσιπ είναι $E(X) = nN_1/N = 50 \cdot \frac{10}{500} = 1$.

Λύση Άσκησης Αυτοαξιολόγησης 4.12

Ο αριθμός των βακτηριδίων σε ένα συγκεκριμένο υγρό περιγράφεται από την κατανομή Poisson με ρυθμό $\lambda = 10$ ανά λίτρο. Συμβολίζουμε με $X := X(0.5)$ την τυχαία μεταβλητή που παριστάνει τον αριθμό των βακτηριδίων σε 500 ml=0.5 lt του υγρού. Τότε $X(0.5) \sim \mathscr{P}(\lambda = 0.5 \cdot 10 = 5)$. Ζητείται ο υπολογισμός της $P(X \geq 4)$ ή, ισοδύναμα, της $1 - P(X < 4) = 1 - P(X \leq 3) = 1 - \sum_{x=0}^{3} P(X = x)$. Είναι τότε από τη σχέση (4.39):

$$P(X = 0) = e^{-5} = 0.006737947,\ P(X = 1) = e^{-5}5 = 0.03368973$$

και

$$P(X = 2) = \frac{e^{-5}5^2}{2!} = 0.08422434,\ P(X = 3) = \frac{e^{-5}5^3}{3!} = 0.1403739.$$

Επομένως,

$$P(X \geq 4) = 1 - (0.006737947 + 0.03368973 + 0.08422434 + 0.1403739) = 0.7349741.$$

Παρατηρήστε ότι $P(X \geq 4) = P(X > 3)$ και για αυτό στην R πρέπει να χρησιμοποιούμε την εντολή `ppois(3,5, lower.tail=FALSE)`, για να λάβουμε το παραπάνω αποτέλεσμα.
Στα ίδια αποτελέσματα θα καταλήγαμε, αν χρησιμοποιούσαμε τους πίνακες της κατανομής Poisson (με κάποια σφάλματα ακριβείας - επιβεβαιώστε το!).

Λύση Άσκησης Αυτοαξιολόγησης 4.13

Έστω X η τ.μ. που παριστάνει τον αριθμό των ατόμων με παρενέργειες μετά τον εμβολιασμό στα 1000 άτομα που εμβολιάστηκαν. Τότε, υποθέτοντας ότι η εμφάνιση παρενέργειας σε ένα άτομο είναι ανεξάρτητη από την εμφάνιση παρενέργειας σε οποιοδήποτε άλλο άτομο και η πιθανότητα εμφάνισης είναι ίδια σε όλα τα άτομα και ίση με $p = 0.004$, έχουμε ότι $X \sim B(n = 1000, p = 0.004)$. Παρατηρήστε ότι $n > 100$ και $np = 4 < 10$ και, επομένως, η διωνυμική κατανομή μπορεί να προσεγγιστεί εξαιρετικά (όπως και θα επαληθεύσουμε) από την Poisson με παράμετρο $np = 4$. Ζητούνται οι $P(X = 1)$ και $P(X = 6)$, οι οποίες υπολογίζονται, στη συνέχεια, τόσο με τη διωνυμική κατανομή όσο και προσεγγιστικά από την κατανομή Poisson.
Με τη διωνυμική οι ζητούμενες πιθανότητες υπολογίζονται ως ακολούθως

$$P(X = 1) = \binom{10^3}{1} \cdot 0.0004 \cdot (1 - 0.004)^{10^3 - 1} = 10^3 \cdot 0.0004 \cdot (1 - 0.004)^{10^3 - 1} = 0.07296911$$

και

$$P(X = 6) = \binom{10^3}{6} \cdot 0.0004^6 \cdot (1 - 0.004)^{10^3 - 6} = \frac{10^3}{6!(10^3 - 6)!} \cdot 0.0004^6 \cdot (1 - 0.004)^{10^3 - 6}$$
$$= \frac{995 \cdot 996 \cdots 1000}{1 \cdot 2 \cdots 6} \cdot 0.0004^6 \cdot (1 - 0.004)^{10^3 - 6} = 0.1042998.$$

Στα ίδια αποτελέσματα μπορούμε να καταλήξουμε με την εκτέλεση στην R των εντολών `dbinom(1,1000,0.004)` και `dbinom(6,1000,0.004)`, αντίστοιχα.
Εναλλακτικά, οι παραπάνω πιθανότητες μπορούν να υπολογιστούν προσεγγιστικά με τη βοήθεια της Poisson με παράμετρο $\lambda = 4$. Πιο συγκεκριμένα, έχουμε ότι

$$P(X = 1) \simeq 4 \cdot e^{-4} = 0.07326256 \text{ και } P(X = 6) \simeq \frac{4^6 \cdot e^{-4}}{6!} = 0.1041956$$

Στα παραπάνω αποτελέσματα καταλήγουμε εκτελώντας τις εντολές `dpois(1,4)` και `dpois(6,4)`, αντίστοιχα, στην R.
Στα ίδια αποτελέσματα θα καταλήγαμε, αν χρησιμοποιούσαμε τους πίνακες της κατανομής Poisson (επιβεβαιώστε το!).

Λύση Άσκησης Αυτοαξιολόγησης 4.14

Ο αριθμός των ατόμων που καταφθάνουν στα επείγοντα περιστατικά ενός νοσοκομείου της Δυτικής Ελλάδας περιγράφεται ικανοποιητικά από την κατανομή Poisson με μέσο ρυθμό 2 αφίξεις στα 30 λεπτά. Έστω $X(2)$ και $X(0.5)$ οι τ.μ. που περιγράφουν τον αριθμό των αφίξεων σε 1 ώρα=$2 \cdot 30$ λεπτά και σε 15 λεπτά=$0.5 \cdot 30$ λεπτά. Τότε $X(2) \sim \mathscr{P}(2 \cdot 2 = 4)$ και $X(0.5) \sim \mathscr{P}(0.5 \cdot 2 = 1)$. Ζητείται η εύρεση των $P(X(2) = 5)$ και $P(X(0.5) = 0)$, οι οποίες είναι ίσες με

$$P(X(2) = 5) = \frac{e^{-4}4^5}{5!} = 0.1562935 \text{ και } P(X(0.5) = 0) = e^{-1} = 0.3678794,$$

αντιστοίχως. Τα παραπάνω αποτελέσματα μπορούν να ληφθούν άμεσα από την R με τη βοήθεια των εντολών `dpois(5,4)` και `dpois(0,1)`, αντίστοιχα.
Στα ίδια αποτελέσματα θα καταλήγαμε, αν χρησιμοποιούσαμε τους πίνακες της κατανομής Poisson (με κάποια σφάλματα ακριβείας - επιβεβαιώστε το!).

ΒΙΒΛΙΟΓΡΑΦΙΑ

Ελληνόγλωσση

Ζωγράφος, Κ. (2008). *Πιθανότητες*. Ιωάννινα: Πανεπιστήμιο Ιωαννίνων.

Παπαϊωάννου, Τ. (1993). *Εισαγωγή στις Πιθανότητες και τη Στατιστική, Μέρος Ι: Πιθανότητες*. Ιωάννινα.

Ξενόγλωσση

Bernoulli, J. (1713). *Ars Conjectandi, Opus Posthumum. Accedit Tractatus de Seriebus infinitis, et Epistola Gallice scripta de ludo Pilae recticularis. Impensis Thurnisiorum*. Fratrum, Basel.

Dodge, Y. (2008). Binomial Distribution. In: *The Concise Encyclopedia of Statistics*. New York, NY: Springer New York, pp. 44–45.

Feller, W. (1968). *An Introduction to Probability Theory and its Applications.* John Wiley and Sons.

Hald, A., De Moivre, A. and McClintock, B. (1984). 'De Mensura Sortis' or 'On the Measurement of Chance'. *International Statistical Review*, *3*, pp. 229–262.

Joarder, A. H. (2011). Hypergeometric Distribution and Its Application in Statistics. In: *International Encyclopedia of Statistical Science*. Ed. by M. Lovric. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 641–643.

Johnson, N., Kemp, A. W. and Kotz, S. (2005). *Univariate Discrete Distributions* (3rd ed.). New York: Wiley and Sons, Inc.

Laplace, P. (1812). *Théorie analytique des probabilités*. Courcier.

Moivre, A. de (1712). De Mensura Sortis. *Philosophical Transactions*, *27*, pp. 213–264.

Moivre, A. de (1718). *De Mensura Sortis*. The Doctrine of Chances. 2nd edition 1738, 3nd 1756.

Pascal, B. (1679). *Varia Opera Mathematica. D. Petri de Fermat.* Tolosae.

Poisson, S. D. (1837). *Probabilité des jugements en matière criminelle et en matière civile, précédées des règles générales du calcul des probabilités [Research on the Probability of Judgments in Criminal and Civil Matters] (in French)*. Paris, France: Bachelier.

Stigler, S. M. (1982). Poisson on the poisson distribution. *Statistics & Probability Letters*, *1*, pp. 33–35.

ΚΕΦΑΛΑΙΟ 5

ΕΙΔΙΚΕΣ ΣΥΝΕΧΕΙΣ ΚΑΤΑΝΟΜΕΣ

Σύνοψη

Σε αυτό το κεφάλαιο οι έννοιες που παρουσιάστηκαν στο Κεφάλαιο 3 θα αξιοποιηθούν για τη μελέτη βασικών συνεχών τυχαίων μεταβλητών και των κατανομών τους, που περιγράφουν ευρύ φάσμα προβλημάτων και τυχαίων φαινομένων σε διάφορα επιστημονικά πεδία. Στο πλαίσιο αυτό, θα παρουσιαστούν οι ακόλουθες συνεχείς κατανομές: η ομοιόμορφη, η εκθετική, η Weibull, η λογαριθμοκανονική, η γάμμα, η βήτα και, τέλος, η σπουδαιότερη όλων, η κανονική κατανομή.

Προαπαιτούμενη γνώση: Κεφάλαιο 3 του παρόντος συγγράμματος.

Προσδοκώμενα μαθησιακά αποτελέσματα: Όταν θα έχετε ολοκληρώσει τη μελέτη του κεφαλαίου αυτού,

- θα αναγνωρίζετε καθεμία από τις συνεχείς κατανομές που θα παρουσιαστούν και πότε πρέπει να τις εφαρμόζετε και
- θα μπορείτε να υπολογίζετε πιθανότητες ενδεχομένων που συνδέονται με αυτές.

Γλωσσάριο επιστημονικών όρων

- Βήτα κατανομή
- Γάμμα κατανομή
- Εκθετική κατανομή
- Κανονική κατανομή
- Λογαριθμοκανονική κατανομή
- Συνεχής ομοιόμορφη κατανομή
- Τυπική κανονική κατανομή
- Weibull κατανομή

Οικονόμου, Π., Μαλεφάκη, Σ., & Μπατσίδης, Α. (2023). *Πιθανότητες - Στατιστική.*
[Προπτυχιακό εγχειρίδιο].

5.1 Εισαγωγή

Στο παρόν κεφάλαιο θα παρουσιαστούν οι βασικότερες μονοδιάστατες συνεχείς κατανομές, που αποτελούν τόσο βασικές επιλογές για τη μοντελοποίηση τυχαίων φαινομένων όσο και έχουν αποτελέσει τη βάση για την παρουσίαση στη στατιστική βιβλιογραφία άλλων γενικευμένων κατανομών, που τις περιέχουν ως ειδικές περιπτώσεις.

5.2 Ομοιόμορφη κατανομή

Η απλούστερη συνεχής κατανομή είναι η συνεχής ομοιόμορφη κατανομή (continuous uniform distribution), η οποία συναντάται και ως ορθογώνια κατανομή (rectangular distribution), λόγω της μορφής του γραφήματος της συνάρτησης πυκνότητας πιθανότητάς της (βλ. Σχήμα 5.1). Πρόκειται για μια οικογένεια συμμετρικών κατανομών πιθανότητας με δύο παραμέτρους a και b με $a, b \in \mathbb{R}$ και $a < b$. Οι τιμές a, b είναι η ελάχιστη και η μέγιστη δυνατή τιμή της υπό μελέτη συνεχούς τυχαίας μεταβλητής, αντίστοιχα. Η συνεχής ομοιόμορφη κατανομή προέκυψε ως γενίκευση της διακριτής ομοιόμορφης κατανομής θέλοντας να μοντελοποιήσει τυχαία φαινόμενα με ισοπίθανη έκβαση σε υποδιαστήματα, ίσου μήκους, του διαστήματος (a,b). Για τον λόγο αυτό έχει σταθερή συνάρτηση πυκνότητας πιθανότητας στο (a,b) και, χρησιμοποιώντας τις ιδιότητες που πρέπει να πληροί η συνάρτηση πυκνότητας πιθανότητας, προκύπτει ο ακόλουθος ορισμός.

Ορισμός 5.1

Η τυχαία μεταβλητή X λέγεται ότι ακολουθεί τη **συνεχή ομοιόμορφη κατανομή** ή απλώς **ομοιόμορφη κατανομή** στο διάστημα (a,b) με $a, b \in \mathbb{R}$ και $a < b$, αν οι δυνατές της τιμές ανήκουν στο διάστημα (a,b) και η συνάρτηση πυκνότητας πιθανότητάς της δίνεται από τη σχέση:

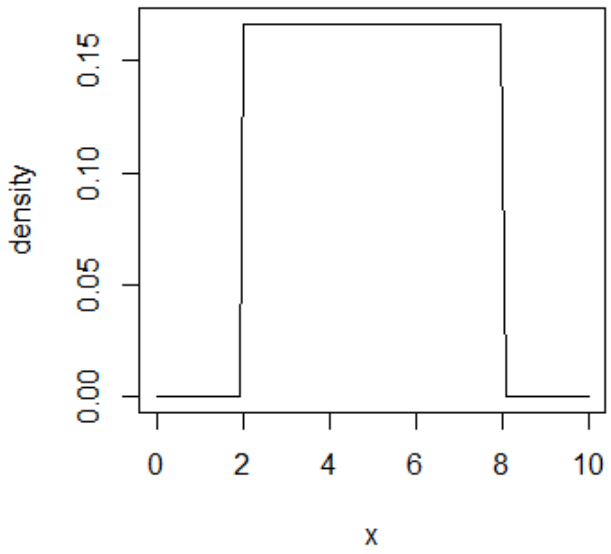
$$f_X(x) = \begin{cases} \frac{1}{b-a}, & x \in (a,b), \\ 0, & \text{αλλού.} \end{cases} \tag{5.1}$$

Στην περίπτωση αυτή, γράφουμε ότι $X \sim U(a,b)$.

Παρατήρηση 5.1

Παρατηρήστε ότι οι παράμετροι της ομοιόμορφης κατανομής αποτελούν τα άκρα του διαστήματος του πεδίου ορισμού της κατανομής. Επίσης, ενθυμούμενοι ότι στις συνεχείς τυχαίες μεταβλητές η πιθανότητα πραγματοποίησης συγκεκριμένης τιμής ισούται με μηδέν, προκύπτει ότι η ομοιόμορφη κατανομή θα μπορούσε ισοδύναμα να οριστεί στο κλειστό διάστημα $[a,b]$.

Γραφικά, η συνάρτηση πυκνότητας πιθανότητας απεικονίζεται ως ένα ευθύγραμμο τμήμα μεταξύ του a και του b για y ίσο με $\frac{1}{b-a}$. Λαμβάνοντας υπόψη ότι το εμβαδόν κάτω από την καμπύλη της σππ είναι πάντοτε ίσο με 1, προκύπτει ότι, καθώς αυξάνεται το μήκος της βάσης, μειώνεται το ύψος του ορθογωνίου που σχηματίζεται κάτω από το ευθύγραμμο τμήμα της σππ. Παρατηρήστε ότι η πυκνότητα της ομοιόμορφης κατανομής παραμένει 0 μέχρι το σημείο a, δηλαδή μέχρι τη μικρότερη τιμή της θεωρηθείσας ομοιόμορφης κατανομής. Στη συνέχεια, ανεβαίνει αμέσως σε πιθανότητα $\frac{1}{b-a}$ και παραμένει σε αυτό το επίπεδο μέχρι να φτάσουμε στην τιμή b (δηλαδή το άνω άκρο της ομοιόμορφης κατανομής μας). Τέλος, για μεγαλύτερες τιμές από το άνω άκρο της ισούται πάλι με 0. Στο Σχήμα 5.1 απεικονίζεται η σππ της $U(2,8)$.



Σχήμα 5.1: Γραφική παράσταση της σππ της $U(2,8)$.

Οι εντολές που χρησιμοποιήθηκαν στην R, για να γίνει το Σχήμα 5.1, ήταν οι ακόλουθες:

```
x = seq(0,10, length=100)
plot(x, dunif(x, 2, 8), ylab="density", type ="l", col=4)
```

Άμεση συνέπεια της σχέσης (5.1) και του ορισμού της ασκ μιας τ.μ. , που δόθηκε στη σχέση (3.1), είναι ότι η ασκ της $X \sim U(a,b)$ δίνεται από τη σχέση:

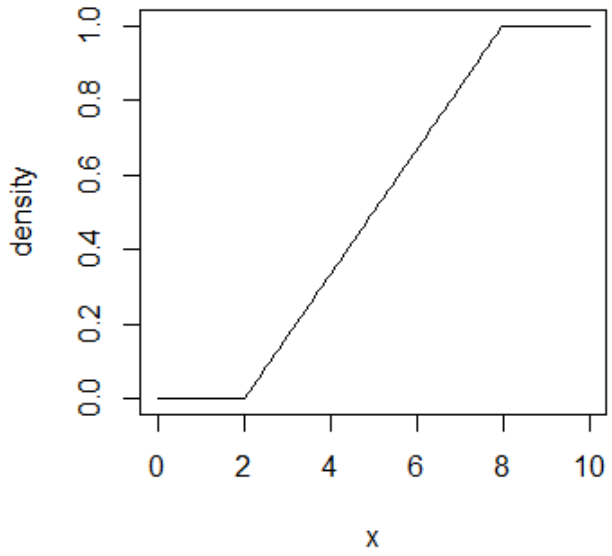
$$F_X(x) = \begin{cases} 0, & x \leq a, \\ \frac{x-a}{b-a}, & x \in [a,b], \\ 1, & x \geq b. \end{cases} \tag{5.2}$$

Παρατηρήστε ότι, όπως αναμενόταν από τις ιδιότητες της ασκ, η ασκ της $U(a,b)$ είναι $F_X(x) = 0$ ($F_X(x) = 1$) για τιμές του x μικρότερες (μεγαλύτερες) από το κάτω (άνω, αντίστοιχα) άκρο του διαστήματος των δυνατών τιμών της τ.μ. X, ενώ η κλίση της ευθείας μεταξύ των άκρων (a,b) είναι ίση με $\frac{1}{b-a}$. Άμεσα προκύπτει ότι όσο το μήκος του διαστήματος (a,b) αυξάνεται, η κλίση της ασκ μειώνεται. Στο Σχήμα 5.2 απεικονίζεται η ασκ της $U(2,8)$.

Παρατήρηση 5.2

Έστω $X \sim U(a,b)$, τότε με τη γλώσσα προγραμματισμού R μπορούμε χρησιμοποιώντας:

- τη συνάρτηση `dunif(x,a,b)` να υπολογίσουμε τη σππ στο σημείο ή στο διάνυσμα σημείων x,
- τη συνάρτηση `punif(x,a,b,lower.tail=TRUE)` να υπολογίσουμε την ασκ στο σημείο ή στο διάνυσμα σημείων x,
- τη συνάρτηση `punif(x,a,b,lower.tail=FALSE)` να υπολογίσουμε την πιθανότητα $P(X > x)$ στο σημείο ή στο διάνυσμα σημείων x,
- τη συνάρτηση `qunif(q,a,b,lower.tail=TRUE)` να προσδιορίσουμε το σημείο ή το διάνυσμα σημείων x για τα οποία ισχύει ότι $P(X \leq x) = q$, όπου q είναι πιθανότητα ή διάνυσμα με συνιστώσες πιθανότητες, αντίστοιχα,



Σχήμα 5.2: Γραφική παράσταση της ασκ της $U(2,8)$.

- τη συνάρτηση `qunif(q,a,b,lower.tail=FALSE` να προσδιορίσουμε το σημείο ή το διάνυσμα σημείων x για τα οποία ισχύει ότι $P(X > x) = q$, όπου q είναι πιθανότητα ή διάνυσμα με συνιστώσες πιθανότητες, αντίστοιχα,
- τη συνάρτηση `runif(n,a,b)` να παράγουμε ένα δείγμα μεγέθους n από αυτήν την κατανομή.

Άσκηση Αυτοαξιολόγησης 5.1

Μπορείτε να γράψετε τις εντολές που πρέπει να χρησιμοποιηθούν στην R για να γίνει η γραφική παράσταση της ασκ της $U(2,8)$;

Στην πρόταση που ακολουθεί, διατυπώνεται η πιο χαρακτηριστική ιδιότητα της ομοιόμορφης κατανομής.

Πρόταση 5.1

Έστω X η τυχαία μεταβλητή που ακολουθεί ομοιόμορφη κατανομή στο διάστημα (a,b), τότε οποιαδήποτε υποδιαστήματα του (a,b) ίσου μήκους έχουν την ίδια πιθανότητα πραγματοποίησης.

Απόδειξη Πρότασης 5.1

Έστω (c,d), όπου c και d δύο αυθαίρετοι πραγματικοί αριθμοί τέτοιοι, ώστε $a < c < d < b$ με $d - c = \ell$. Αρκεί να δείξουμε ότι η $P(c \leq X \leq d)$ δεν εξαρτάται από την επιλογή των (c,d) αλλά μόνο από το μήκος l. Πράγματι, ισχύει ότι:

$$P(c \leq X \leq d) = F_X(d) - F_X(c) = \frac{d-c}{b-a} = \frac{\ell}{b-a}.$$

Αποδεικνύεται ότι ισχύει και το αντίστροφο της παραπάνω πρότασης, ότι δηλαδή, αν για μια συνεχή τ.μ. με τιμές στο (a,b), οι πιθανότητες αυτή να παίρνει τιμές σε οποιοδήποτε υποδιάστημα του (a,b) ίσου μήκους είναι ίσες, τότε η $X \sim U(a,b)$. Για την απόδειξη του αντιστρόφου βλ. μεταξύ άλλων, Ζωγράφος (2008).

Συνδυάζοντας τα παραπάνω, ίσως να μην προκαλεί έκπληξη το γεγονός ότι η ομοιόμορφη κατανομή δεν βρίσκει εφαρμογή στην περιγραφή πολλών τυχαίων πραγματικών φαινομένων, καθώς αυτά θα έπρεπε να διέπονται από το ισοπίθανο των διαστημάτων ίσου μήκους σε όλο το σύνολο τιμών της τυχαίας μεταβλητής. Παρ΄ όλα αυτά, η ομοιόμορφη κατανομή χρησιμοποιείται αρκετά συχνά για τη μοντελοποίηση των σφαλμάτων που προκύπτουν κατά τη μετατροπή αναλογικών σημάτων σε ψηφιακά σήματα. Πιο συγκεκριμένα, η ειδική περίπτωση της ομοιόμορφης $U(-h,h)$ με $h = 0.51^k$, χρησιμοποιείται συχνά για την περιγραφή σφαλμάτων που προκύπτουν κατά τη στρογγυλοποίηση αριθμητικών τιμών σε τιμές με k το πλήθος δεκαδικά ψηφία, ενώ η ομοιόμορφη κατανομή έχει χρησιμοποιηθεί και για τη μοντελοποίηση της κίνησης σε μια ευθεία οδό. Για περισσότερες εφαρμογές σε πραγματικά φαινόμενα παραπέμπουμε τον αναγνώστη στο σύγγραμμα Johnson *et al.* (1994a) και στις εκεί αναφορές.

Από την άλλη πλευρά, η ομοιόμορφη κατανομή αποτελεί τον θεμέλιο λίθο δημιουργίας τυχαίων αριθμών από πλήθος άλλες κατανομές με την ευρέως χρησιμοποιούμενη μέθοδο της αντιστροφής (inverse transform sampling). Η χρησιμότητά της οφείλεται στην ακόλουθη ιδιότητα της ομοιόμορφης κατανομής.

Πρόταση 5.2

Έστω F μια συνεχής ασκ και F^{-1} η αντίστροφή της που ορίζεται από τη σχέση:

$$F^{-1}(u) = inf\{x : F(x) > u\}.$$

Αν η τ.μ. $X \sim U(0,1)$, τότε η τ.μ. $Y = F^{-1}(X)$ έχει την F ως ασκ.

Απόδειξη Πρότασης 5.2

Η αθροιστική συνάρτηση κατανομής της Y είναι:

$$\begin{aligned} F_Y(y) = P(Y \le y) &= P\left(F^{-1}(X) \le y\right) \\ &= P\left(X \le F(y)\right) = F_X\left(F(y)\right) \\ &= \int_0^{F(y)} 1dx = F(y). \end{aligned}$$

Επισημαίνεται ότι αρχικά εφαρμόσαμε την αθροιστική συνάρτηση F στην $P\left(F^{-1}(X) \le y\right)$, ενώ έπειτα χρησιμοποιήθηκε ο ορισμός της αθροιστικής συνάρτησης της τυχαίας μεταβλητής X.

Η ιδιότητα που δόθηκε στην Πρόταση 5.2 επιτρέπει μέσω της δημιουργίας τυχαίων - ή για την ακρίβεια ψευδο-τυχαίων αριθμών[1] - από την $U(0,1)$ και υπό την προϋπόθεση ότι μπορούμε να προσδιορίσουμε την αντίστροφη της F να μπορούμε να δημιουργούμε τυχαίους αριθμούς από την κατανομή με ασκ F.

Στις προτάσεις που ακολουθούν, προσδιορίζονται οι απλές ροπές k τάξης και η ροπογεννήτρια της ομοιόμορφης κατανομής, με βάσει τους ορισμούς αυτών των εννοιών που δίνονται στις σχέσεις (3.23) και (3.28), αντίστοιχα.

Πρόταση 5.3

Έστω X η τυχαία μεταβλητή που ακολουθεί ομοιόμορφη κατανομή στο διάστημα (a,b). Τότε:

$$E(X^k) = \frac{b^{k+1} - a^{k+1}}{(k+1)(b-a)}. \tag{5.3}$$

[1] Ονομάζονται έτσι γιατί παράγονται ξεκινώντας από κάποια αρχική τιμή, η οποία καλείται σπόρος (seed). Αν η ίδια αρχική τιμή χρησιμοποιείται ξανά και ξανά, τότε προκύπτει η ίδια ακολουθία τυχαίων αριθμών. Παρ΄ όλα αυτά, οι αριθμοί που παράγονται συμπεριφέρονται σαν να είναι αληθινά τυχαίοι.

Απόδειξη Πρότασης 5.3

Συνδυάζοντας τις σχέσεις (3.23) και (5.1) έχουμε:

$$\mathrm{E}(X^k) = \int_a^b \frac{x^k}{b-a}dx = \int_a^b \frac{1}{b-a}\frac{d}{dx}\left(\frac{x^{k+1}}{k+1}\right)dx = \frac{b^{k+1}-a^{k+1}}{(k+1)(b-a)}.$$

Πρόταση 5.4

Έστω X η τυχαία μεταβλητή που ακολουθεί ομοιόμορφη κατανομή στο διάστημα (a,b). Τότε:

$$M_X(t) = \frac{e^{tb}-e^{ta}}{t(b-a)}, \text{ για } t \neq 0, \tag{5.4}$$

ενώ $M_X(0) = 1$.

Απόδειξη Πρότασης 5.4

Εξ ορισμού $M_X(0) = 1$, ενώ για $t \neq 0$, συνδυάζοντας τις (3.28) και (5.1), προκύπτει:

$$M_X(t) = E_X\left(e^{tX}\right) = \int_a^b \frac{e^{tx}}{b-a}dx = \int_a^b \frac{1}{b-a}\frac{d}{dx}\left(\frac{e^{tx}}{t}\right) = \frac{e^{tb}-e^{ta}}{t(b-a)}.$$

Συνέπεια των παραπάνω προτάσεων είναι το ακόλουθο Πόρισμα.

Πόρισμα 5.1

Έστω X η τυχαία μεταβλητή που ακολουθεί ομοιόμορφη κατανομή στο διάστημα (a,b). Τότε:

$$\mu = E(X) = \frac{a+b}{2}, \tag{5.5}$$

και

$$\sigma^2 = Var(X) = \frac{(b-a)^2}{12}. \tag{5.6}$$

Απόδειξη Πορίσματος 5.1

Η απόδειξη προκύπτει άμεσα από τη σχέση (5.3) και λαμβάνοντας υπόψη ότι $Var(X) = E(X^2) - (E(X))^2$.

Παράδειγμα 5.1

Ένα συνεργείο οδικής βοήθειας βρίσκεται στην αρχή μιας οδού που συνδέει δύο πόλεις, η οποία οδός έχει συνολικό μήκος 100 χιλιομέτρων. Αν υποθέσουμε ότι το σημείο όπου θα κληθεί το συνεργείο να παρέχει οδική βοήθεια βρίσκεται ισοπίθανα σε οποιοδήποτε σημείο της οδού, υπολογίστε την πιθανότητα να υπάρξει ανάγκη για παροχή βοήθειας μετά το ογδοηκοστό χιλιόμετρο και μέχρι το τέλος της οδού.

Λύση Παραδείγματος 5.1

Έστω X η τ.μ. που παριστάνει την απόσταση (σε χιλιόμετρα) από την αρχή της οδού μέχρι το σημείο της οδού όπου θα κληθεί το συνεργείο να προσφέρει βοήθεια. Τότε, από τα δεδομένα του προβλήματος, προκύπτει ότι $X \sim U(0,100)$ με σππ και ασκ που δίνονται από τις σχέσεις (5.1) και (5.2), αντίστοιχα με $a = 0$ και $b = 100$. Η πιθανότητα να υπάρξει ανάγκη για παροχή βοήθειας μετά το ογδοηκοστό

χιλιόμετρο και μέχρι το τέλος της οδού προσδιορίζεται ως εξής:

$$P(80 < X < 100) = F_X(100) - F_X(80) = 1 - \frac{80-0}{100-0} = 0.2.$$

Σύμφωνα με την Παρατήρηση 5.2, το προηγούμενο αποτέλεσμα θα μπορούσε να υπολογιστεί και μέσω της παρακάτω εντολής της R: `punif(100,0,100,lower.tail=TRUE)-punif(80,0,100,lower.tail=TRUE)`

Άσκηση Αυτοαξιολόγησης 5.2

Θεωρήστε ότι φτάνετε στο κτήριο μιας δημόσιας υπηρεσίας και θέλετε να κάνετε χρήση του ανελκυστήρα. Από τη στιγμή που πιέζετε το κουμπί κλήσης του ανελκυστήρα χρειάζονται από 0 έως 40 δευτερόλεπτα για να έρθει στο σημείο που βρίσκεστε. Υιοθετώντας το μοντέλο της ομοιόμορφης κατανομής, προσδιορίστε:

1. την πιθανότητα να περιμένετε περισσότερο από 12 δευτερόλεπτα,
2. την πιθανότητα να περιμένετε το πολύ 17 δευτερόλεπτα,
3. την πιθανότητα να περιμένετε λιγότερο από 25 δευτερόλεπτα, όταν σας είναι εκ των προτέρων γνωστό ότι θα περιμένετε περισσότερο από 12 δευτερόλεπτα,
4. τον μέσο χρόνο αναμονής,
5. την τιμή του χρόνου, έστω x, για την οποία το 90 τοις εκατό των επισκεπτών της δημόσιας υπηρεσίας θα περιμένει λιγότερο από x δευτερόλεπτα.

5.3 Βήτα κατανομή

Η βήτα κατανομή αποτελεί γενίκευση της ομοιόμορφης $U(0,1)$ και είναι κατ' ουσίαν μια οικογένεια συνεχών κατανομών πιθανότητας που ορίζονται στο διάστημα $(0,1)$ με δύο θετικές παραμέτρους a και b, οι οποίες, όπως θα δούμε, καθορίζουν το σχήμα της κατανομής. Σύμφωνα με τον Sheynin (1971) η βήτα κατανομή οφείλεται στον Thomas Bayes (1701-1761), έναν Άγγλο κληρικό, στατιστικό και φιλόσοφο, πασίγνωστο για το Θεώρημα του Bayes, που παρουσιάστηκε στην Ενότητα 2.3. Ειδικότερα, ο Bayes οδηγήθηκε στη βήτα κατανομή ως τη σππ της πιθανότητας επιτυχίας σε δοκιμές Bernoulli (παραπέμπουμε τον/την ενδιαφερόμενο/η αναγνώστη/στρια στην ενότητα της εργασίας του με τίτλο Applications, Bayesian inference που αναδημοσιεύτηκε στο άρθρο Bayes, 1958), χωρίς παρ΄ όλα αυτά να αναλύσει ιδιότητες της κατανομής. Ωστόσο, η κατανομή είχε εξεταστεί επανειλημμένα από πολλούς μελετητές πριν από τον Bayes (1958). Για παράδειγμα, όπως αναφέρει ο Arjun K. Gupta, η προέλευσή της πηγαίνει πίσω στο 1676 και σε μια επιστολή του Sir Isaac Newton προς τον Henry Oldenberg (Gupta, 2011).

Σε κάθε περίπτωση, η βήτα κατανομή έχει χρησιμοποιηθεί για τη μοντελοποίηση πλήθους τυχαίων μεταβλητών με τιμές σε πεπερασμένου μήκους διάστημα, καθώς τις περισσότερες φορές με κατάλληλο μετασχηματισμό μπορούν να οδηγηθούν στο διάστημα $[0,1]$. Στο πλαίσιο αυτό, έχει χρησιμοποιηθεί στη μοντελοποίηση κλιματικών δεδομένων, όπως είναι η ηλιοφάνεια στη Μαλαισία (Sulaiman *et al.*, 1999), η διάρκεια συννεφιάς στην Αυστραλία (Chia and Hutchinson, 1991), αλλά και στη μοντελοποίηση γενωμικών δεδομένων (genomic data) (Tataru *et al.*, 2015). Για γενικεύσεις αυτής της κατανομής, περισσότερες πληροφορίες για εφαρμογές της και ιστορικά στοιχεία για τη δημιουργία της παραπέμπουμε μεταξύ άλλων, στους Johnson *et al.* (1994b) και Nadarajah and Kotz (2007), καθώς και στις εκεί αναφορές.

Έπειτα από αυτήν την εισαγωγή και τη μικρή παράθεση ιστορικών στοιχείων, ακολουθεί ο ορισμός της βήτα κατανομής.

Ορισμός 5.2

Η τυχαία μεταβλητή X λέγεται ότι ακολουθεί τη **βήτα κατανομή** με παραμέτρους $a > 0$ και $b > 0$, αν οι δυνατές της τιμές x είναι $x \in (0,1)$ και η συνάρτηση πυκνότητας πιθανότητάς της δίνεται από τη σχέση:

$$f_X(x) = \begin{cases} \frac{x^{a-1}(1-x)^{b-1}}{B(a,b)}, & x \in (0,1), \\ 0, & \text{αλλού}, \end{cases} \tag{5.7}$$

όπου $B(a,b)$ είναι η βήτα συνάρτηση που ορίζεται στη σχέση (Β΄.10) του Παραρτήματος Β΄. Στην περίπτωση αυτή, γράφουμε ότι $X \sim Be(a,b)$.

Άσκηση Αυτοαξιολόγησης 5.3

Επιβεβαιώστε ότι η συνάρτηση της σχέσης (5.7) είναι όντως σππ.

Η σππ της βήτα κατανομής μπορεί να λάβει διάφορες μορφές ανάλογα με τις τιμές των παραμέτρων της a και b (Nadarajah and Kotz, 2007). Ειδικότερα, προκύπτει ότι:

(i) είναι συμμετρική γύρω από το $x = 0.5$, όταν $a = b$,

(ii) είναι μονοκόρυφη και το σχήμα της μοιάζει με βουνό (mount shape), όταν $a > 1$ και $b > 1$,

(iii) έχει σχήμα παρόμοιο με το λατινικό γράμμα J (J-shape), για $a > 1$ και $b < 1$,

(iv) έχει σχήμα παρόμοιο με ανάποδο λατινικό γράμμα J, για $a < 1$ και $b > 1$, ενώ, τέλος,

(v) έχει σχήμα παρόμοιο με το λατινικό γράμμα U (U-shape), για $a < 1$ και $b < 1$.

Τα παραπάνω απεικονίζονται στο Σχήμα 5.3 και στο Σχήμα 5.4, όπου παριστάνεται η σππ της βήτα κατανομής για διάφορους συνδυασμούς των παραμέτρων της.

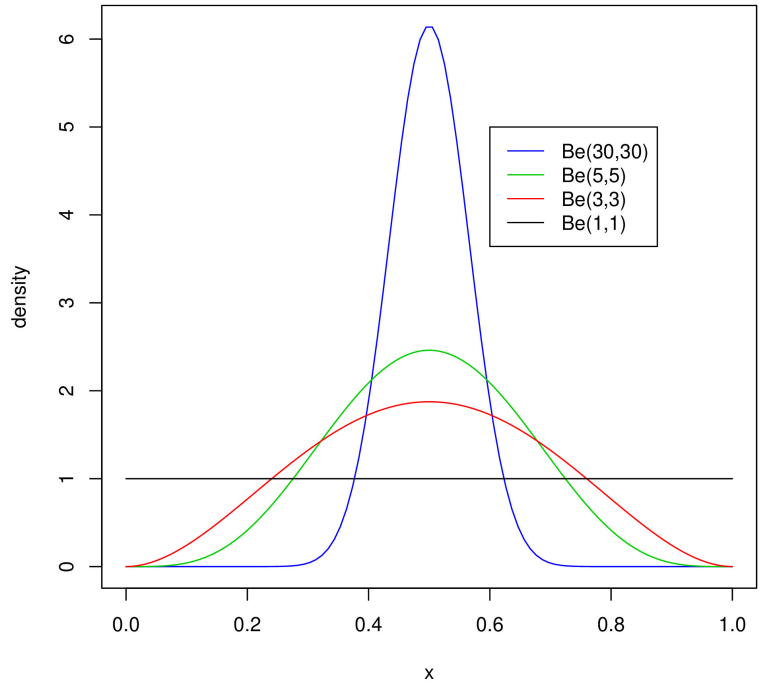
Ενδεικτικά, οι εντολές που χρησιμοποιήθηκαν στην R, για να κατασκευαστεί το Σχήμα 5.3, είναι οι ακόλουθες:

```
x = seq(0,1, length=100)
plot(x, dbeta(x, 30, 30), ylab="density", type ="l", col=4)
lines(x, dbeta(x, 5, 5), type ="l", col=3)
lines(x, dbeta(x, 3, 3), col=2)
lines(x, dbeta(x, 1, 1), col=1)
legend(0.6,5, c("Be(30,30)","Be(5,5)","Be(3,3)","Be(1,1)"),lty=c(1,1,1,1),col=
    c(4,3,2,1))
```

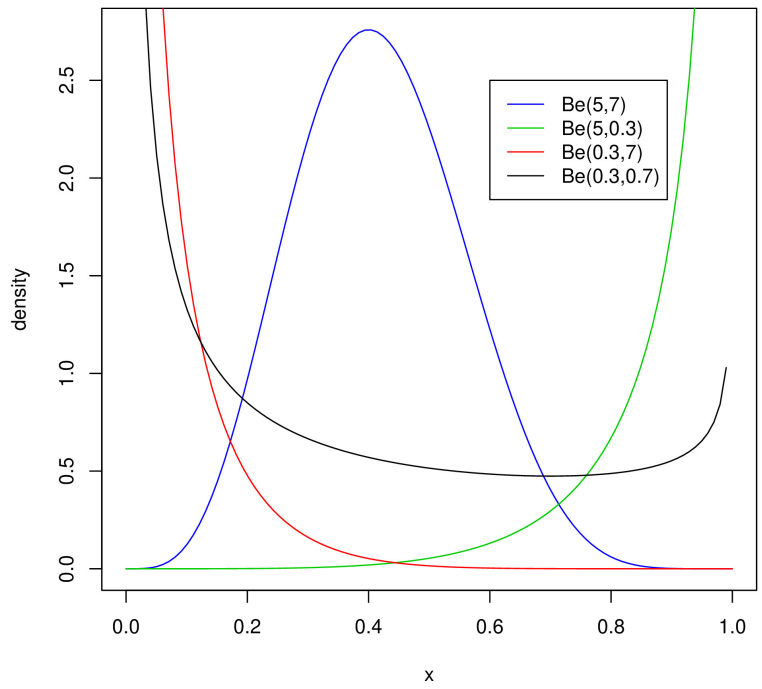
Στη συνέχεια, προσδιορίζεται η ασκ της βήτα κατανομής. Έστω $X \sim Be(a,b)$ με σππ που δίνεται στη σχέση (5.7). Τότε, προφανώς για $x < 0$ ισχύει ότι $F_X(x) = 0$, ενώ για $0 \le x < 1$ εξ ορισμού έχουμε ότι:

$$\begin{aligned} F_X(x) = P(X \le x) &= \int_0^x \frac{t^{a-1}(1-t)^{b-1}}{B(a,b)} dt \\ &= \frac{1}{B(a,b)} \int_0^x t^{a-1}(1-t)^{b-1} dt \\ &= \frac{B(x;a,b)}{B(a,b)} = I_x(a,b), \end{aligned}$$

όπου χρησιμοποιήθηκαν οι σχέσεις (Β΄.18) και (Β΄.19) του Παραρτήματος Β΄.



Σχήμα 5.3: Γραφική παράσταση της σππ της $Be(a,b)$ για $(a,b) = (30,30),(5,5),(3,3),(1,1)$.



Σχήμα 5.4: Γραφική παράσταση της σππ της $Be(a,b)$ για $(a,b) = (5,7),(5,0.3),(0.3,7),(0.3,0.7)$.

Συγκεντρωτικά, η ασκ της τ.μ. $X \sim Be(a,b)$ με σππ που δίνεται από τη σχέση (5.7), προσδιορίζεται από τη σχέση:

$$F_X(x) = \begin{cases} 0, & x < 0, \\ \frac{\mathrm{B}(x;a,b)}{B(a,b)}, & 0 \le x < 1, \\ 1, & x \ge 1, \end{cases} \tag{5.8}$$

και μπορεί να υπολογιστεί με τη βοήθεια κατάλληλου λογισμικού ή πινάκων.

Παρατήρηση 5.3

Έστω $X \sim Be(a,b)$ με σππ που δίνεται από τη σχέση (5.7). Τότε με τη γλώσσα προγραμματισμού R μπορούμε:

- με τη συνάρτηση `dbeta(x,a,b)` να υπολογίσουμε τη σππ στο σημείο ή στο διάνυσμα σημείων x,
- με τη συνάρτηση `pbeta(x,a,b,lower.tail=TRUE)` να υπολογίσουμε την ασκ στο σημείο ή στο διάνυσμα σημείων x,
- με τη συνάρτηση `pbeta(x,a,b,lower.tail=FALSE)` να υπολογίσουμε την πιθανότητα $P(X > x)$ στο σημείο ή στο διάνυσμα σημείων x,
- με τη συνάρτηση `qbeta(q,a,b,lower.tail=TRUE)` να προσδιορίσουμε το σημείο ή το διάνυσμα σημείων x για τα οποία ισχύει ότι $P(X \le x) = q$, όπου q είναι πιθανότητα ή διάνυσμα με συνιστώσες πιθανότητες, αντίστοιχα,
- με τη συνάρτηση `qbeta(q,a,b,lower.tail=FALSE` να προσδιορίσουμε το σημείο ή το διάνυσμα σημείων x για τα οποία ισχύει ότι $P(X > x) = q$, όπου q είναι πιθανότητα ή διάνυσμα με συνιστώσες πιθανότητες, αντίστοιχα,
- με τη συνάρτηση `rbeta(n,a,b)` να δημιουργήσουμε ένα δείγμα μεγέθους n από αυτήν την κατανομή.

Η χρήση κάποιων εκ των παραπάνω εντολών παρουσιάζεται στα παραδείγματα που ακολουθούν.

Άσκηση Αυτοαξιολόγησης 5.4

Να γράψετε τις εντολές που πρέπει να χρησιμοποιηθούν στην R για να κατασκευαστούν οι γραφικές παραστάσεις των ασκ των βήτα κατανομών που εμφανίζονται στο Σχήμα 5.3 και στο Σχήμα 5.4.

Στις προτάσεις που ακολουθούν, προσδιορίζονται οι ροπές k τάξης και η ροπογεννήτρια όταν η τ.μ. $X \sim Be(a,b)$ με σππ που δόθηκε στη σχέση (5.7).

Πρόταση 5.5

Έστω X η τυχαία μεταβλητή που ακολουθεί βήτα κατανομή με παραμέτρους a και b με σππ που προσδιορίζεται στη σχέση (5.7). Τότε η ροπή $k-$τάξης δίνεται από τη σχέση:

$$E(X^k) = \frac{B(a+k,b)}{B(a,b)} = \prod_{r=0}^{k-1} \frac{a+r}{a+b+r}. \tag{5.9}$$

Απόδειξη Πρότασης 5.5

Υπάρχουν πολλοί διαφορετικοί τρόποι απόδειξης της παραπάνω σχέσης, αλλά εδώ θα χρησιμοποιηθούν ο ορισμός και αριθμητικές πράξεις[α΄]. Είναι, εξ ορισμού,

$$\begin{aligned}\mathrm{E}(X^k) &= \int_0^1 x^k \frac{x^{a-1}(1-x)^{b-1}}{B(a,b)}dx \\ &= \frac{1}{B(a,b)}\int_0^1 x^{k+a-1}(1-x)^{b-1}dx, \\ &= \frac{B(a+k,b)}{B(a,b)}\end{aligned}$$

όπου το τελευταίο ολοκλήρωμα προέκυψε με τη βοήθεια του ορισμού της βήτα συνάρτησης που δόθηκε στη σχέση (Β΄.10) του Παραρτήματος Β΄. Επιπρόσθετα, χρησιμοποιώντας τη σχέση (Β΄.11), είναι:

$$\frac{B(a+k,b)}{B(a,b)} = \frac{\Gamma(a+k)\Gamma(b)}{\Gamma(a+b+k)}\frac{\Gamma(a+b)}{\Gamma(b)\Gamma(a)} = \frac{\Gamma(a+k)}{\Gamma(a)}\frac{\Gamma(a+b)}{\Gamma(a+b+k)}.$$

Χρησιμοποιώντας επαναληπτικά την ιδιότητα (Β΄.8) του ίδιου παραρτήματος, έχουμε ότι:

$$\Gamma(a+k) = \Gamma(a)\prod_{r=0}^{k-1}(a+r)$$

και

$$\Gamma(a+b+k) = \Gamma(a+b)\prod_{r=0}^{k-1}(a+b+r),$$

οπότε προκύπτει:

$$\frac{B(a+k,b)}{B(a,b)} = \prod_{r=0}^{k-1}\frac{a+r}{a+b+r}.$$

[α΄] Για παράδειγμα θα μπορούσε να χρησιμοποιηθεί η ροπογεννήτρια συνάρτηση που προσδιορίζεται στην πρόταση που ακολουθεί.

Πρόταση 5.6

Έστω X η τυχαία μεταβλητή που ακολουθεί βήτα κατανομή με παραμέτρους a και b με σππ που προσδιορίζεται στη σχέση (5.7). Τότε:

$$M_X(t) = 1 + \sum_{k=1}^{+\infty}\left(\prod_{r=0}^{k-1}\frac{a+r}{a+b+r}\right)\frac{t^k}{k!}. \tag{5.10}$$

Απόδειξη Πρότασης 5.6

Από τη σχέση (3.28) προκύπτει ότι:

$$\begin{aligned}M_X(t) = E\left(e^{tX}\right) &= \int_0^1 e^{tx}\frac{x^{a-1}(1-x)^{b-1}}{B(a,b)}dx \\ &= \frac{1}{B(a,b)}\int_0^1 e^{tx}x^{a-1}(1-x)^{b-1}dx.\end{aligned}$$

Στη συνέχεια, καθώς $e^{tx} = \sum_{k=0}^{+\infty} \frac{(tx)^k}{k!}$, έχουμε ότι:

$$\begin{aligned} M_X(t) &= \frac{1}{B(a,b)} \int_0^1 \left(\sum_{k=0}^{+\infty} \frac{(tx)^k}{k!} \right) x^{a-1}(1-x)^{b-1} dx \\ &= \frac{1}{B(a,b)} \sum_{k=0}^{+\infty} \frac{t^k}{k!} \int_0^1 x^{a+k-1}(1-x)^{b-1} dx \\ &= \sum_{k=0}^{+\infty} \frac{t^k}{k!} \frac{B(a+k,b)}{B(a,b)} \\ &= 1 + \sum_{k=1}^{+\infty} \frac{t^k}{k!} \frac{B(a+k,b)}{B(a,b)}, \end{aligned}$$

όπου χρησιμοποιήθηκε ο ορισμός της βήτα συνάρτησης που δίνεται στη σχέση (Β΄.10). Επίσης, προηγουμένως δείξαμε ότι $\frac{B(a+k,b)}{B(a,b)} = \prod_{r=0}^{k-1} \frac{a+r}{a+b+r}$ και η απόδειξη ολοκληρώνεται με συνδυασμό των παραπάνω.

Παρατήρηση 5.4

Η παραπάνω σχέση για τον προσδιορισμό της ροπογεννήτριας της βήτα κατανομής μπορεί να φαίνεται ότι στην πράξη δεν μπορεί να χρησιμοποιηθεί, καθώς περιέχει τόσο ένα μη πεπερασμένο άθροισμα όσο και ένα γινόμενο με πλήθος όρων που αυξάνεται. Όμως η συνάρτηση ${}_1F_1(a, a+b, t)$, η οποία ορίζεται ως:

$${}_1F_1(a, a+b, t) = 1 + \sum_{k=1}^{+\infty} \left(\prod_{r=0}^{k-1} \frac{a+r}{a+b+r} \right) \frac{t^k}{k!}$$

και ονομάζεται «confluent hypergeometric function of the first kind», είναι ευρέως γνωστή, καθώς έχει μελετηθεί σε διάφορους κλάδους των μαθηματικών και αλγόριθμοι υπολογισμού της είναι διαθέσιμοι σε πλήθος υπολογιστικών προγραμμάτων.

Από τη σχέση (5.9) για $k = 1, 2$ και λαμβάνοντας υπόψη ότι $Var(X) = E(X^2) - (E(X))^2$, έπειτα από λίγη άλγεβρα, εύκολα προκύπτει το ακόλουθο πόρισμα.

Πόρισμα 5.2

Έστω X η τυχαία μεταβλητή που ακολουθεί βήτα κατανομή με παραμέτρους a και b με σππ που προσδιορίζεται στη σχέση (5.7). Τότε:

$$\mu = E(X) = \frac{a}{a+b}, \tag{5.11}$$

και

$$\sigma^2 = Var(X) = \frac{ab}{(a+b)^2(a+b+1)}. \tag{5.12}$$

Άσκηση Αυτοαξιολόγησης 5.5

Το ποσοστό των ατόμων που πάσχει από την ασθένεια Α σε μια συγκεκριμένη περιοχή μοντελοποιείται από μία τυχαία μεταβλητή που περιγράφεται ικανοποιητικά από τη βήτα κατανομή με παραμέτρους $a = 2$ και $b = 4$.

1. Υπολογίστε την πιθανότητα το ποσοστό των ατόμων που πάσχει από την ασθένεια Α να είναι μικρότερο από 20%.
2. Υπολογίστε την πιθανότητα το ποσοστό των ατόμων που πάσχει από την ασθένεια Α να είναι μεγαλύτερο από 25%, όταν γνωρίζετε ότι είναι μικρότερο από 40%.

Υπόδειξη: Για την εύρεση αριθμητικής τιμής χρησιμοποιήστε την R.

5.4 Εκθετική κατανομή

Η εκθετική κατανομή είναι μια από τις πιο σημαντικές συνεχείς κατανομές με ευρύτατο πλαίσιο εφαρμογών, που έχει ως βάση της μια διαδικασία Poisson. Ειδικότερα, ας θεωρήσουμε ότι $Y(t)$ είναι η τυχαία μεταβλητή που παριστάνει το πλήθος των αφίξεων σε μια διαδικασία Poisson με ρυθμό λ, $\lambda > 0$, σε χρονικό διάστημα μήκους t. Με βάση όσα διατυπώθηκαν στην Ενότητα 4.7, αυτό σημαίνει ότι ο αριθμός των αφίξεων σε ένα χρονικό διάστημα δεν έχει καμία επίδραση στον αριθμό των αφίξεων σε οποιοδήποτε άλλο ξένο διάστημα, η πιθανότητα πραγματοποίησης δύο ή περισσότερων αφίξεων σε ένα πολύ μικρό χρονικό διάστημα είναι αμελητέα και η πιθανότητα πραγματοποίησης ακριβώς μιας άφιξης στο επόμενο μικρό χρονικό διάστημα είναι σταθερή για κάθε χρονικό διάστημα ίδιου μήκους, ενώ

$$P\big(Y(t) = y\big) = \frac{e^{-\lambda t}(\lambda t)^y}{y!}, y = 0,1,2.... \tag{5.13}$$

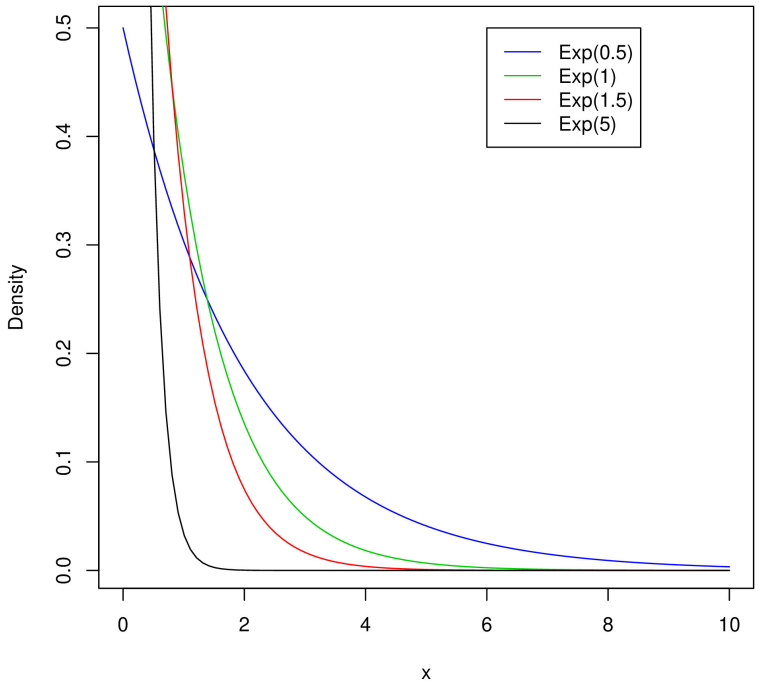
Έστω τώρα X η τ.μ. που παριστάνει τον χρόνο μεταξύ δύο διαδοχικών αφίξεων ή συμβάντων σε μια διαδικασία Poisson με ρυθμό λ, $\lambda > 0$. Προφανώς, οι δυνατές τιμές της X είναι $x \geq 0$. Στόχος μας είναι ο προσδιορισμός της κατανομής της τ.μ. X. Όπως είδαμε στην Ενότητα 3.7, ο προσδιορισμός της κατανομής μιας νέας τυχαίας μεταβλητής που ορίζεται με τη βοήθεια μιας υπάρχουσας τυχαίας μεταβλητής μπορεί να επιτευχθεί μεταξύ άλλων, μέσω του προσδιορισμού της ασκ της νέας τ.μ. Σε αυτό το πλαίσιο έχουμε, λαμβάνοντας υπόψη το σύνολο των δυνατών τιμών της X, ότι $F_X(x) = P(X \leq x) = 0$ για $x < 0$. Από την άλλη πλευρά για $x \geq 0$ έχουμε ότι:

$$F_X(x) = P(X \leq x) = 1 - P(X > x) = 1 - P(Y(x) = 0) = 1 - e^{-\lambda x}, x \geq 0, \tag{5.14}$$

καθώς, αν ο χρόνος μεταξύ δύο διαδοχικών συμβάντων είναι μεγαλύτερος από x, αυτό ουσιαστικά σημαίνει ότι στο χρονικό διάστημα που μεσολαβεί από το πρώτο γεγονός έως τη χρονική στιγμή x, δηλαδή στο χρονικό διάστημα μήκους x, δεν έχει πραγματοποιηθεί καμία άφιξη. Επομένως, $P(X > x) = P(\text{καμία άφιξη σε χρόνο } x) = P(Y(x) = 0)$ και μέσω της διαδικασίας Poisson με ρυθμό λ, $\lambda > 0$, προσδιορίστηκε η ασκ της X. Συγκεντρωτικά, η ασκ της τ.μ. X δίνεται από τη σχέση:

$$F_X(x) = \begin{cases} 0, & x \leq 0, \\ 1 - e^{-\lambda x}, & x \geq 0. \end{cases} \tag{5.15}$$

Έχοντας προσδιορίσει την F_X στο R, παραγωγίζοντάς την ως προς x, προκύπτει η σππ της εκθετικής κατανομής, η οποία για πληρότητα δίνεται στον ακόλουθο ορισμό.



Σχήμα 5.5: Γραφική παράσταση της σππ της $Exp(\lambda)$ για $\lambda = 0.5,1,1.5,5$.

Ορισμός 5.3

Η τυχαία μεταβλητή X λέγεται ότι ακολουθεί την **εκθετική κατανομή** με παράμετρο λ, $\lambda > 0$, αν οι δυνατές της τιμές x είναι $x \geq 0$ και η συνάρτηση πυκνότητας πιθανότητάς της δίνεται από τη σχέση:

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0, \\ 0, & \text{αλλού.} \end{cases} \tag{5.16}$$

Στην περίπτωση αυτή, γράφουμε ότι $X \sim Exp(\lambda)$.

Στο Σχήμα 5.5 απεικονίζεται η σππ της εκθετικής κατανομής για διάφορες τιμές του λ. Οι εντολές που χρησιμοποιήθηκαν στην R, για να γίνει το Σχήμα 5.5, ήταν οι ακόλουθες:

```
x = seq(0,10, length=100)
plot(x, dexp(x, 0.5), ylab="Density", type ="l", col=4)
lines(x, dexp(x, 1), type ="l", col=3)
lines(x, dexp(x, 1.5), col=2)
lines(x, dexp(x, 5), col=1)

legend(6,0.5, c("Exp(0.5)","Exp(1)","Exp(1.5)","Exp(5)"),lty=c(1,1,1,1),col=c
    (4,3,2,1))
```

Άσκηση Αυτοαξιολόγησης 5.6

Αν η τ.μ. $X \sim U(0,1)$, τότε η τ.μ. $Y = -\frac{1}{\lambda}\log(1-X)$ με $\lambda > 0$, ακολουθεί εκθετική κατανομή με παράμετρο $\lambda > 0$. Σχολιάστε πώς μπορεί να αξιοποιηθεί αυτό το αποτέλεσμα για τη δημιουργία μιας παρατήρησης από την εκθετική κατανομή με παράμετρο λ.

Παρατήρηση 5.5

Έστω $X \sim Exp(\lambda)$, τότε με τη γλώσσα προγραμματισμού R μπορούμε:

- με τη συνάρτηση `dexp(x,λ)` να υπολογίσουμε τη σππ στο σημείο ή στο διάνυσμα σημείων x,
- με τη συνάρτηση `pexp(x,λ,lower.tail=TRUE)` να υπολογίσουμε την ασκ στο σημείο ή στο διάνυσμα σημείων x,
- με τη συνάρτηση `pexp(x,λ,lower.tail=FALSE)` να υπολογίσουμε την πιθανότητα $P(X > x)$ στο σημείο ή στο διάνυσμα σημείων x,
- με τη συνάρτηση `qexp(q,λ,lower.tail=TRUE)` να προσδιορίσουμε το σημείο ή το διάνυσμα σημείων x για τα οποία ισχύει ότι $P(X \leq x) = q$, όπου q είναι πιθανότητα ή διάνυσμα με συνιστώσες πιθανότητες, αντίστοιχα,
- με τη συνάρτηση `qexp(q,λ,lower.tail=FALSE` να προσδιορίσουμε το σημείο ή το διάνυσμα σημείων x για τα οποία ισχύει ότι $P(X > x) = q$, όπου q είναι πιθανότητα ή διάνυσμα με συνιστώσες πιθανότητες, αντίστοιχα,
- με τη συνάρτηση `rexp(n,λ)` να δημιουργήσουμε ένα δείγμα μεγέθους n από αυτήν την κατανομή.

Η χρήση κάποιων εκ των παραπάνω εντολών παρουσιάζεται στα παραδείγματα που ακολουθούν.

Άσκηση Αυτοαξιολόγησης 5.7

Γράψτε τις εντολές που πρέπει να χρησιμοποιηθούν στην R, για να κατασκευαστεί η γραφική παράσταση της ασκ της $Exp(\lambda)$, για $\lambda = 0.5,1,1.5,5$.

Από όσα προαναφέρθηκαν, είναι προφανές ότι η εκθετική κατανομή με παράμετρο λ προέκυψε ως ο χρόνος μεταξύ δύο διαδοχικών αφίξεων ή συμβάντων ή γεγονότων σε μια διαδικασία Poisson με ρυθμό λ, $\lambda > 0$. Αυτή η ιδιότητα έχει ως άμεση συνέπεια η εκθετική κατανομή να χρησιμοποιείται στη μοντελοποίηση του χρόνου ζωής (χρόνος μεταξύ δύο συμβάντων: γέννηση-θάνατος), του χρόνου μέχρι την ίαση ενός ασθενούς (χρόνος μεταξύ δύο συμβάντων: πάθηση-ίαση) και κατά ανάλογο τρόπο για τη μοντελοποίηση του χρόνου διάσπασης ενός ραδιενεργού ατόμου, της χρονικής διάρκειας μεταξύ δύο βλαβών μηχανικών εξαρτημάτων (με σταθερό ρυθμό βλαβών), της αναμονής σε ουρά αναμονής (χρόνος εξυπηρέτησης πελατών, χρόνος ανταπόκρισης πυροσβεστικής) και ούτω καθεξής, υπό την προϋπόθεση ότι η πιθανότητα να συμβεί το γεγονός στο επόμενο μικρό χρονικό διάστημα είναι σταθερή για κάθε χρονική στιγμή.

Ωστόσο, μια ακόμη χαρακτηριστική ιδιότητα, πλην της παραπάνω, την οποία η εκθετική κατανομή ικανοποιεί και, όπως αποδεικνύεται, είναι η μόνη συνεχής κατανομή που την πληροί, περιορίζει τη χρήση της στη μοντελοποίηση πλήθους πραγματικών τυχαίων φαινομένων[2]. Η ιδιότητα αυτή, που είναι γνωστή ως ιδιότητα της αμνησίας ή ιδιότητα της έλλειψης μνήμης ή ιδιότητα της μη γήρανσης, αποτελεί αντικείμενο μελέτης της επόμενης πρότασης.

Πρόταση 5.7

Η εκθετική κατανομή είναι η μοναδική κατανομή μεταξύ των κατανομών με συνεχή ασκ και δυνατό σύνολο $S = \{x : x > 0\}$ που έχει την **ιδιότητα της αμνησίας ή έλλειψης μνήμης ή μη γήρανσης**, η οποία αποδίδεται από τη σχέση

$$P(X > t + s|X > t) = P(X > s), \text{ για κάθε } s,t > 0. \tag{5.17}$$

[2]Οι παραπάνω περιορισμοί έχουν οδηγήσει να εμφανιστεί στη βιβλιογραφία πλήθος επεκτάσεων της εκθετικής κατανομής. Στο πλαίσιο του παρόντος συγγράμματος, θα παρουσιαστούν η γάμμα και η Weibull κατανομή, οι οποίες είναι πολύ γνωστές γενικεύσεις της.

Απόδειξη Πρότασης 5.7

Αρχικά, υποθέτουμε ότι η τ.μ. $X \sim Exp(\lambda)$ με $\lambda > 0$, αυθαίρετη παράμετρο, και θα δείξουμε ότι ικανοποιεί τη σχέση (5.17). Από τον ορισμό της δεσμευμένης ή υπό συνθήκη πιθανότητας έχουμε:

$$\begin{aligned} P(X > t+s|X > t) &= \frac{P(X > t+s \text{ και } X > t)}{P(X > t)} = \frac{P(X > t+s)}{P(X > t)} \\ &= \frac{1 - F_X(t+s)}{1 - F_X(t)} = \frac{e^{-\lambda(t+s)}}{e^{-\lambda t}} = e^{-\lambda s} \\ &= 1 - F_X(s) = P(X > s). \end{aligned}$$

Στη συνέχεια, θέλουμε να δείξουμε ότι η εκθετική κατανομή είναι η μοναδική κατανομή μεταξύ των κατανομών με συνεχή ασκ και δυνατό σύνολο τιμών $S = \{x : x > 0\}$ που πληροί την ιδιότητα της αμνησίας. Έστω X μια τ.μ. με σππ και ασκ $f_X(\cdot)$ και $F_X(\cdot)$, αντίστοιχα, για την οποία υποθέτουμε ότι ισχύει η ιδιότητα της αμνησίας, ήτοι η σχέση (5.17). Η σχέση αυτή πρωτύτερα μας οδήγησε στην ισοδύναμη έκφραση:

$$\frac{1 - F_X(t+s)}{1 - F_X(t)} = 1 - F_X(s) \text{ για κάθε } s, t > 0,$$

ή στη μορφή:

$$S_X(s+t) = S_X(t) S_X(s) \text{ για κάθε } s, t > 0, \tag{5.18}$$

όπου $S_X(x) = 1 - F_X(x)$ με $0 \leq S_X(x) \leq 1$. Η σχέση (5.18) ανήκει στις λεγόμενες συναρτησιακές εξισώσεις του Cauchy και μάλιστα είναι γνωστή στη βιβλιογραφία ως **multiplicative Cauchy functional equation** και έχει διαφορετικές μορφές λύσεων ανάλογα με τις ιδιότητες, συνθήκες που ικανοποιεί. Για περισσότερες λεπτομέρειες παραπέμπουμε, μεταξύ άλλων στο σύγγραμμα του Kuczma (2009) και τις εκεί αναφορές. Ειδικότερα, καθώς η συνάρτηση $S_X(t)$ είναι συνεχής και μονότονα φθίνουσα έχουμε ότι η μοναδική λύση της είναι της μορφής $S_X(t) = (S_X(1))^t$. Από αυτήν την τελευταία σχέση, έχουμε ότι:

$$S_X(t) = e^{\log(S_X(1))t} = e^{-\lambda t},$$

ή, ισοδύναμα,

$$F_X(t) = 1 - e^{-\lambda t}, \text{ για } t > 0,$$

όπου θέσαμε $\lambda = -\log(S_X(1)) > 0$, και η απόδειξη ολοκληρώθηκε.

Το ερώτημα που τίθεται τώρα είναι ποια είναι η πραγματική ερμηνεία της ιδιότητας της αμνησίας και αν αυτή περιορίζει τη χρήση της εκθετικής κατανομής στη μοντελοποίηση πραγματικών τυχαίων φαινομένων. Χωρίς βλάβη της γενικότητας, ας θεωρήσουμε ότι η τ.μ. $X \sim Exp(\lambda)$ περιγράφει τον χρόνο ζωής ενός οργανισμού ή τον χρόνο αναμονής μέχρι να πραγματοποιηθεί ένα γεγονός. Η ιδιότητα της αμνησίας λέει: δεδομένου ότι μας είναι γνωστό ότι ο χρόνος ζωής ή αναμονής είναι περισσότερος από t μονάδες του χρόνου, η πιθανότητα ο χρόνος ζωής ή αναμονής να είναι περισσότερος από $t+s$ μονάδες του χρόνου, δηλαδή να ζήσει ή να αναμένει άλλες επιπλέον s μονάδες του χρόνου, ισούται με τη μη δεσμευμένη πιθανότητα ο χρόνος ζωής ή αναμονής να είναι μεγαλύτερος από μια αρχική περίοδο s χρονικών μονάδων. Με άλλα λόγια, άσχετα με πόσο διάστημα έχει ζήσει κάποιος ή τον χρόνο αναμονής του, η κατανομή του εναπομείνοντος χρόνου ζωής ή αναμονής είναι ίδια με τη μη δεσμευμένη αρχική. Δηλαδή, όσο ζει ένας οργανισμός ή περιμένει σε μια ουρά συμπεριφέρεται σαν να γεννήθηκε μόλις ή μόλις να πήγε στην ουρά. Προφανώς, τα περισσότερα πραγματικά τυχαία φαινόμενα δεν ικανοποιούν αυτήν την ιδιότητα.

Παρατήρηση 5.6

Στη βιβλιογραφία έχει παρουσιαστεί και η λεγόμενη Εκθετική Οικογένεια Κατανομών. Η παρουσίαση και η μελέτη της ξεφεύγουν από τους σκοπούς του παρόντος συγγράμματος. Στο σημείο αυτό απλώς θα αναφερθεί ότι η Εκθετική Οικογένεια Κατανομών είναι μια μεγάλη οικογένεια κατανομών, η οποία περιέχει την εκθετική ως ειδική περίπτωση, αλλά και άλλες κατανομές τόσο συνεχείς όσο και διακριτές, όπως τη διωνυμική, την Poisson, την κανονική και πολλές άλλες. Επομένως, δεν θα πρέπει να δημιουργείται σύγχυση μεταξύ εκθετικής κατανομής και Εκθετικής Οικογένειας Κατανομών.

Στις προτάσεις που ακολουθούν, προσδιορίζονται οι ροπές k τάξης και η ροπογεννήτρια της εκθετικής κατανομής με παράμετρο λ.

Πρόταση 5.8

Έστω X τυχαία μεταβλητή που ακολουθεί την εκθετική κατανομή με παράμετρο λ. Τότε η $k-$ τάξης ροπή δίνεται από τη σχέση:

$$E(X^k) = \frac{k!}{\lambda^k}. \tag{5.19}$$

Απόδειξη Πρότασης 5.8

Εξ ορισμού

$$\begin{aligned} \mathrm{E}(X^k) &= \int_0^{+\infty} x^k \lambda e^{-\lambda x} dx = \lambda \int_0^{+\infty} x^k e^{-\lambda x} dx \\ &= \frac{1}{\lambda^k} \int_0^{+\infty} z^k e^{-z} dz \\ &= \frac{\Gamma(k+1)}{\lambda^k} = \frac{k!}{\lambda^k}, \end{aligned}$$

όπου το τελευταίο ολοκλήρωμα προέκυψε από την αλλαγή μεταβλητών $z = \lambda x$ και υπολογίζεται με τη βοήθεια της γάμμα συνάρτησης, λαμβάνοντας υπόψη τις σχέσεις (Β΄.6) και (Β΄.9) του Παραρτήματος Β΄.

Πρόταση 5.9

Έστω X τυχαία μεταβλητή που ακολουθεί την εκθετική κατανομή με παράμετρο λ. Τότε:

$$M_X(t) = \frac{\lambda}{\lambda - t}, \quad \text{για} \;\; t < \lambda. \tag{5.20}$$

Απόδειξη Πρότασης 5.9

Λαμβάνοντας υπόψη τη σχέση (3.28) έχουμε ότι:

$$M_X(t) = E\left(e^{tX}\right) = \int_0^{+\infty} e^{tx} \lambda e^{-\lambda x} dx = \lambda \int_0^{+\infty} e^{(t-\lambda)x} dx.$$

Το τελευταίο ολοκλήρωμα είναι πεπερασμένο, αν $t - \lambda < 0$, δηλαδή αν $t < \lambda$. Τότε για $t < \lambda$

$$M_X(t) = \lambda \int_0^{+\infty} \frac{d}{dx}\left(\frac{e^{(t-\lambda)x}}{t-\lambda}\right)$$

και το επιθυμητό αποτέλεσμα προκύπτει ύστερα από λίγη άλγεβρα.

Παρατηρήστε ότι οι ροπές k-τάξης θα μπορούσαν να προκύψουν (ανατρέξτε στην Ενότητα 3.6.3) χρησιμοποιώντας τη ροπογεννήτρια συνάρτηση και τη σχέση $E(X^k) = \frac{d}{dx} M_X^k(t)|_{t=0}$. Έτσι, λαμβάνοντας

υπόψη τη σχέση (5.20) και μετά από λίγη άλγεβρα, προκύπτει το επιθυμητό αποτέλεσμα. Επιπλέον, από τη σχέση (5.19) και λαμβάνοντας υπόψη ότι $Var(X) = E(X^2) - (E(X))^2$, έπειτα από λίγη άλγεβρα, εύκολα προκύπτει το ακόλουθο πόρισμα.

Πόρισμα 5.3

Έστω X τυχαία μεταβλητή που ακολουθεί την εκθετική κατανομή με παράμετρο λ. Τότε:

$$\mu = E(X) = \frac{1}{\lambda}, \tag{5.21}$$

και

$$\sigma^2 = Var(X) = \frac{1}{\lambda^2}. \tag{5.22}$$

Η παρουσίαση της εκθετικής κατανομής έγινε πρωτύτερα συνδέοντάς την με τη διαδικασία Poisson. Στη συνέχεια, για λόγους πληρότητας, θα παρουσιαστεί η σύνδεση της εκθετικής κατανομής με τη γεωμετρική κατανομή και θα διαπιστωθεί ότι η εκθετική κατανομή μπορεί να θεωρηθεί ως το συνεχές ανάλογο της γεωμετρικής κατανομής (της μόνης απαριθμητής κατανομής που πληροί την ιδιότητα της αμνησίας). Πιο συγκεκριμένα, η γεωμετρική κατανομή περιγράφει τον αριθμό των ανεξάρτητων επαναλήψεων μιας δοκιμής Bernoulli μέχρι την εμφάνιση της πρώτης επιτυχίας, δηλαδή να αλλάξει η κατάσταση, ενώ η εκθετική τον χρόνο που χρειάζεται μέχρι την πραγματοποίηση ενός ενδεχομένου, τουτέστιν την αλλαγή κατάστασης.

Για να γίνει κατανοητή η σύνδεση της γεωμετρικής κατανομής με την εκθετική κατανομή υποθέτουμε αρχικά ότι μεταξύ κάθε ανεξάρτητης επανάληψης δοκιμής Bernoulli μεσολαβεί χρόνος d και σε κάθε δοκιμή η πιθανότητα επιτυχίας είναι $p = \lambda d$ (βλ. μεταξύ άλλων, Pishro-Nik, 2014). Επιπλέον, θεωρούμε ότι το d είναι πολύ μικρό, που έχει ως αποτέλεσμα μεταξύ των δοκιμών να μεσολαβεί πολύ μικρός χρόνος, δηλαδή οι δοκιμές είναι πολύ κοντινές η μία στην άλλη, και η πιθανότητα επιτυχίας κάθε δοκιμής είναι πολύ μικρή. Σε αυτό το πλαίσιο, η σύνδεση της γεωμετρικής με την εκθετική κατανομή δίνεται στην πρόταση που ακολουθεί.

Πρόταση 5.10

Έστω $Y \sim Geo(p = \lambda d)$ με $\lambda > 0, d > 0$ και έστω $X = Yd$. Τότε για οποιοδήποτε $x \in (0, +\infty)$ έχουμε ότι:

$$\lim_{d \to 0} F_X(x) = 1 - e^{-\lambda x}.$$

Απόδειξη Πρότασης 5.10

Για οποιοδήποτε $x \in (0, +\infty)$ ισχύει ότι:

$$F_X(x) = P(X \leq x) = P(Yd \leq x) = P\left(Y \leq \frac{x}{d}\right) = F_Y\left(\frac{x}{d}\right).$$

Καθώς η τ.μ. Y ακολουθεί γεωμετρική κατανομή με παράμετρο $p = \lambda d$, εφαρμόζοντας τη σχέση για την ασκ της γεωμετρικής κατανομής από την Ενότητα 4.4, έχουμε ότι:

$$F_X(x) = 1 - (1 - \lambda d)^{\lfloor \frac{x}{d} \rfloor},$$

όπου το $\lfloor \frac{x}{d} \rfloor$ συμβολίζει το ακέραιο μέρος του $\frac{x}{d}$. Επομένως,

$$\lim_{d \to 0} F_X(x) = \lim_{d \to 0} \left(1 - (1 - \lambda d)^{\lfloor \frac{x}{d} \rfloor}\right) = 1 - e^{-\lambda x}, \text{ για } x > 0,$$

που αποδεικνύει το ζητούμενο.

Για να γίνει κατανοητή η ερμηνεία της παραπάνω ιδιότητας υποθέστε ότι περιμένετε να γίνει ένα συμβάν. Σε κάθε πολύ μικρό χρονικό διάστημα, π.χ. κλάσματα του δευτερολέπτου, ρίχνετε ένα κέρμα με πιθανότητα εμφάνισης κεφαλής πολύ μικρή και κάνετε την υπόθεση ότι, αν προσγειωθεί κεφαλή, τότε συμβαίνει το γεγονός που αναμένεται. Σύμφωνα με την παραπάνω πρόταση, ο χρόνος μέχρι την πραγματοποίηση του επιθυμητού συμβάντος ακολουθεί προσεγγιστικά εκθετική κατανομή. Η παραπάνω ερμηνεία εξηγεί κατά έναν τρόπο και την ιδιότητα της αμνησίας της εκθετικής κατανομής, καθώς, αν η εμφάνιση της πρώτης επιτυχίας συνδεθεί με τη ρίψη ενός νομίσματος, είναι λογικό οι προηγούμενες αποτυχημένες ρίψεις να μην επηρεάζουν τις επόμενες, άρα και τον χρόνο αναμονής μέχρι την πραγματοποίηση του επιθυμητού συμβάντος.

Παράδειγμα 5.2

Υποθέστε ότι ο χρόνος εξυπηρέτησης ενός πελάτη σε ένα κατάστημα λιανικής περιγράφεται ικανοποιητικά από μια εκθετική κατανομή με μέσο χρόνο εξυπηρέτησης ίσο με 3 λεπτά.

1. Υπολογίστε την πιθανότητα η διάρκεια εξυπηρέτησης ενός τυχαίου πελάτη να είναι μεταξύ τεσσάρων έως πέντε λεπτών.
2. Δεδομένου ότι ένας πελάτης έχει εξυπηρετηθεί ήδη τρία λεπτά, υπολογίστε την πιθανότητα ότι θα χρειαστεί τουλάχιστον άλλα δύο λεπτά για να εξυπηρετηθεί.
3. Υπολογίστε την πιθανότητα ένας τυχαίος πελάτης να εξυπηρετηθεί σε λιγότερο από τέσσερα λεπτά.
4. Υπολογίστε την πιθανότητα ένας τυχαίος πελάτης να χρειαστεί για την εξυπηρέτησή του περισσότερο από πέντε λεπτά.
5. Ποιος είναι ο ελάχιστος χρόνος που χρειάζεται για να εξυπηρετηθεί το 70% των πελατών;

Λύση Παραδείγματος 5.2

Έστω X η τ.μ. που παριστάνει σε λεπτά τον χρόνο εξυπηρέτησης ενός πελάτη σε ένα κατάστημα λιανικής. Μας δίνεται ότι έχει μέσο χρόνο εξυπηρέτησης ίσο με 3 λεπτά, επομένως $E(X) = 3$ και, γνωρίζοντας ότι $E(X) = \frac{1}{\lambda}$, άμεσα προκύπτει ότι $\lambda = \frac{1}{3}$.

1. Στο ερώτημα αυτό ζητείται να υπολογιστεί η πιθανότητα $P(4 < X < 5)$. Είναι

$$P(4 < X < 5) = F_X(5) - F_X(4) = \left(1 - \exp\left(-\frac{5}{3}\right)\right) - \left(1 - \exp\left(-\frac{4}{3}\right)\right) = 0.07472154.$$

2. Η ζητούμενη πιθανότητα είναι η $P(X > 3 + 2|X > 3)$. Από την ιδιότητα της αμνησίας της εκθετικής κατανομής έχουμε:

$$P(X > 3 + 2|X > 3) = P(X > 2) = 1 - F_X(2) = \exp\left(-\frac{2}{3}\right) = 0.5134171.$$

3. Στο ερώτημα αυτό ζητείται να υπολογιστεί η πιθανότητα $P(X < 4)$. Από τον ορισμό της ασκ έχουμε:

$$P(X < 4) = F_X(4) = 1 - \exp\left(-\frac{4}{3}\right) = 0.7364029.$$

4. Θέλουμε να υπολογίσουμε την πιθανότητα $P(X > 5)$. Είναι

$$P(X > 5) = 1 - P(X \leq 5) = 1 - F_X(5) = \exp\left(-\frac{5}{3}\right) = 0.1888756.$$

5. Συμβολίζουμε με c τον ζητούμενο ελάχιστο χρόνο που χρειάζεται στην εξυπηρέτηση το 70% των πελατών. Δηλαδή c είναι η τιμή εκείνη για την οποία ισχύει ότι $P(X < c) = 0.7$. Επομένως, από τον ορισμό της ασκ της εκθετικής, έχουμε ότι:

$$1 - \exp\left(-\frac{c}{3}\right) = 0.7$$

ή, μετά από λίγη άλγεβρα, ότι:

$$c = -3 \cdot \log(0.3) = 3.611918.$$

Σύμφωνα με την Παρατήρηση 5.4 τα αποτελέσματα μπορούν να εξαχθούν χρησιμοποιώντας τις ακόλουθες εντολές της R:
1. `pexp(5,1/3,lower.tail=TRUE)-pexp(4,1/3,lower.tail=TRUE)`,
2. `pexp(2,1/3,lower.tail=FALSE)`,
3. `pexp(4,1/3,lower.tail=TRUE)`,
4. `pexp(5,1/3,lower.tail=FALSE)` και
5. `qexp(0.7,1/3,lower.tail=TRUE)`, αντίστοιχα.

Παράδειγμα 5.3

Υποθέστε ότι ο χρόνος ζωής ενός μηχανικού εξαρτήματος περιγράφεται ικανοποιητικά από μια εκθετική κατανομή. Επιπλέον, είναι γνωστό ότι ο μέσος χρόνος ζωής του συγκεκριμένου μηχανικού εξαρτήματος είναι ίσος με 8 χρόνια.

1. Υπολογίστε την πιθανότητα ότι ένα τυχαία επιλεγμένο τέτοιο μηχανικό εξάρτημα θα χρειαστεί αντικατάσταση σε λιγότερο από 6 χρόνια.
2. Υπολογίστε την πιθανότητα ότι ένα τυχαία επιλεγμένο τέτοιο μηχανικό εξάρτημα θα έχει διάρκεια ζωής μεταξύ 6-9 έτη.
3. Ποιος είναι ο ελάχιστος χρόνος ζωής που έχει το 75% των μηχανικών εξαρτημάτων;

Λύση Παραδείγματος 5.3

Έστω X η τ.μ. που παριστάνει σε χρόνια τον χρόνο ζωής του μηχανικού εξαρτήματος. Μας δίνεται ότι έχει μέσο χρόνο ζωής ίσο με 8 χρόνια, επομένως $E(X) = 8$ και, γνωρίζοντας ότι $E(X) = \frac{1}{\lambda}$, άμεσα προκύπτει ότι $\lambda = \frac{1}{8}$.

1. Στο ερώτημα αυτό ζητείται να υπολογιστεί η πιθανότητα $P(X < 6)$. Από τον ορισμό της ασκ έχουμε:

$$P(X < 6) = F_X(6) = 1 - \exp\left(-\frac{6}{8}\right) = 0.5276334.$$

2. Η ζητούμενη πιθανότητα είναι η $P(6 < X < 9)$. Είναι

$$P(6 < X < 9) = F_X(9) - F_X(6) = \left(1 - \exp\left(-\frac{9}{8}\right)\right) - \left(1 - \exp\left(-\frac{6}{8}\right)\right) = 0.1477141.$$

3. Συμβολίζουμε με c τον ζητούμενο ελάχιστο χρόνο ζωής που έχει το 75 τοις εκατό των μηχανικών εξαρτημάτων. Δηλαδή c είναι η τιμή εκείνη για την οποία ισχύει ότι $P(X < c) = 0.75$. Επομένως, από τον ορισμό της ασκ της εκθετικής έχουμε ότι:

$$1 - \exp\left(-\frac{c}{8}\right) = 0.75$$

ή, μετά από λίγη άλγεβρα, ότι:

$$c = -8 \cdot \log(0.25) = 11.09035.$$

Σύμφωνα με την Παρατήρηση 5.4, τα αποτελέσματα θα μπορούσαν να εξαχθούν χρησιμοποιώντας τις ακόλουθες εντολές της R:
1. `pexp(6,1/8,lower.tail=TRUE)`,
2. `pexp(9,1/8,lower.tail=TRUE)-pexp(6,1/8,lower.tail=TRUE)` και
3. `qexp(0.75,1/8,lower.tail=TRUE)`, αντίστοιχα.

Παράδειγμα 5.4

Ο χρόνος αναμονής X ενός πελάτη σε μία τράπεζα είναι μηδέν, εάν αυτός βρει το σύστημα άδειο, και ακολουθεί την εκθετική κατανομή με παράμετρο λ, εάν βρει το σύστημα απασχολημένο. Η πιθανότητα να βρει το σύστημα κενό είναι p. Να βρεθεί η ασκ της X.

Λύση Παραδείγματος 5.4

Έστω K το ενδεχόμενο να βρει ο πελάτης την τράπεζα άδεια. Γνωρίζουμε ότι $P(K) = p$, οπότε η πιθανότητα να βρει ο πελάτης την τράπεζα με κόσμο (όχι άδεια) ισούται με $P(K') = 1 - p$. Οπότε, χρησιμοποιώντας το Θεώρημα Ολικής Πιθανότητας (Κεφάλαιο 2), θα έχουμε $\forall\, x \geq 0$:

$$F(x) = P(X \leq x) = P(X \leq x|K)P(K) + P(X \leq x|K')P(K').$$

Η πιθανότητα $P(X \leq x|K) = 1$, αφού, όταν ο πελάτης βρίσκει την τράπεζα άδεια, ο χρόνος αναμονής του είναι μηδέν, άρα μικρότερος ή ίσος από κάθε $x \geq 0$ με πιθανότητα 1. Επιπλέον, από την εκφώνηση έχουμε ότι, όταν η τράπεζα δεν είναι άδεια και άρα ικανοποιείται το ενδεχόμενο K', ο χρόνος αναμονής ακολουθεί εκθετική κατανομή και επομένως $P(X \leq x|K') = 1 - e^{-\lambda x}$.
Συνοψίζοντας, $\forall\, x \geq 0$ έχουμε

$$F(x) = P(X \leq x) = 1 \cdot p + (1 - e^{-\lambda x}) \cdot (1 - p) = 1 - (1 - p)(1 - e^{-\lambda x}).$$

Τέλος, για $x < 0$ έχουμε, προφανώς, ότι $F(x) = 0$.

Άσκηση Αυτοαξιολόγησης 5.8

Σε μια μικρή επαρχιακή πόλη ο αριθμός των αυτοκινητικών ατυχημάτων περιγράφεται ικανοποιητικά από μια διαδικασία Poisson με μέσο αριθμό ατυχημάτων ίσο με 3 ανά εβδομάδα.

1. Υπολογίστε την πιθανότητα να συμβούν το πολύ 3 ατυχήματα σε μια τυχαία επιλεγμένη εβδομάδα.
2. Υπολογίστε την πιθανότητα να περάσουν τουλάχιστον 2 εβδομάδες μεταξύ δύο διαδοχικών αυτοκινητικών ατυχημάτων.

Άσκηση Αυτοαξιολόγησης 5.9

Υποθέστε ότι σε ένα κατάστημα λιανικής οι πελάτες καταφθάνουν σύμφωνα με μια διαδικασία Poisson με ρυθμό 30 πελάτες την ώρα.

1. Κατά μέσο όρο πόσος χρόνος χρειάζεται για δύο διαδοχικές αφίξεις πελατών;
2. Μετά την άφιξη ενός πελάτη, υπολογίστε την πιθανότητα να χρειαστεί λιγότερο από 1 λεπτό μέχρι την άφιξη του επόμενου πελάτη.
3. Μετά την άφιξη ενός πελάτη, υπολογίστε την πιθανότητα να χρειαστούν περισσότερα από 4 λεπτά

μέχρι την άφιξη του επόμενου πελάτη.

4. Πόσα λεπτά κατά μέγιστο μετά την άφιξη του προηγούμενου πελάτη έρχεται το 80% των πελατών;
5. Είναι η υιοθέτηση των κατανομών που χρησιμοποιήθηκαν στην άσκηση αυτή ρεαλιστική; Δικαιολογήστε την απάντησή σας.

5.5 Γάμμα κατανομή

Στην Ενότητα 5.4 θεωρώντας $Y(t)$ να είναι η τυχαία μεταβλητή που παριστάνει το πλήθος των αφίξεων σε χρονικό διάστημα μήκους t σε μια διαδικασία Poisson με ρυθμό λ, $\lambda > 0$, προσδιορίστηκε η κατανομή του χρόνου μεταξύ δύο διαδοχικών αφίξεων ή συμβάντων σε αυτήν τη διαδικασία και ορίστηκε να είναι η εκθετική κατανομή με παραμέτρο λ. Στο πλαίσιο της ίδιας διαδικασίας Poisson ορίζεται τώρα ως X η τ.μ. που παριστάνει τον χρόνο που μεσολαβεί μεταξύ a διαδοχικών αφίξεων με $a > 0$ οποιονδήποτε φυσικό αριθμό. Στη συνέχεια, προσδιορίζουμε την ασκ της νέας τυχαίας μεταβλητής, της οποίας οι δυνατές τιμές είναι οι $x \in [0, +\infty)$. Επομένως, εύκολα προκύπτει ότι

$F_X(x) = P(X \leq x) = 0$, για $x < 0$. Από την άλλη πλευρά για $x \geq 0$, έχουμε ότι:

$$F_X(x) = P(X \leq x) = 1 - P(X > x) = 1 - P(Y(x) \leq a-1) = 1 - \sum_{k=0}^{a-1} \frac{e^{-\lambda x}(\lambda x)^k}{k!}, x \geq 0, \tag{5.23}$$

όπου χρησιμοποιήθηκε το σκεπτικό ότι το ενδεχόμενο $\{X > x\}$ σημαίνει ότι μέχρι τη χρονική στιγμή x έχουν γίνει το πολύ $a-1$ αφίξεις ή συμβάντα, δηλαδή

$$P(X > x) = P(\text{λιγότερες από } a \text{ αφίξεις σε χρόνο } x) = P(Y(x) \leq a-1).$$

Συγκεντρωτικά, η ασκ της τ.μ. X δίνεται από τη σχέση:

$$F_X(x) = \begin{cases} 0, & x \leq 0, \\ 1 - \sum_{k=0}^{a-1} \frac{e^{-\lambda x}(\lambda x)^k}{k!}, & x \geq 0. \end{cases} \tag{5.24}$$

Έχοντας προσδιορίσει την F_X στο R, παραγωγίζοντάς την ως προς x, προκύπτει η σππ της τ.μ. X. Είναι

$$\begin{aligned} f_X(x) = \frac{d}{dx}F_X(x) &= \frac{d}{dx}\left(1 - e^{-\lambda x} - \sum_{k=1}^{a-1} \frac{e^{-\lambda x}(\lambda x)^k}{k!}\right) \\ &= \lambda e^{-\lambda x} - \frac{d}{dx}\left(\sum_{k=1}^{a-1} \frac{e^{-\lambda x}(\lambda x)^k}{k!}\right) \\ &= \lambda e^{-\lambda x} - \sum_{k=1}^{a-1} \frac{-\lambda e^{-\lambda x}(\lambda x)^k + e^{-\lambda x} k \lambda^k x^{k-1}}{k!} \\ &= \lambda e^{-\lambda x} + \lambda e^{-\lambda x} \sum_{k=1}^{a-1} \frac{(\lambda x)^k}{k!} - \lambda e^{-\lambda x} \sum_{k=1}^{a-1} \frac{k(\lambda x)^{k-1}}{k!}, \end{aligned}$$

από όπου προκύπτει ότι:

$$f_X(x) = \lambda e^{-\lambda x} + \lambda e^{-\lambda x}\left\{\sum_{k=1}^{a-1}\frac{(\lambda x)^k}{k!} - \sum_{k=1}^{a-1}\frac{(\lambda x)^{k-1}}{(k-1)!}\right\}$$
$$= \lambda e^{-\lambda x} + \lambda e^{-\lambda x}\left\{\sum_{k=1}^{a-1}\frac{(\lambda x)^k}{k!} - \sum_{k=0}^{a-2}\frac{(\lambda x)^k}{k!}\right\}$$
$$= \lambda e^{-\lambda x} + \lambda e^{-\lambda x}\left\{\frac{(\lambda x)^{a-1}}{(a-1)!} - 1\right\}.$$

Επομένως,

$$f_X(x) = \begin{cases} \frac{\lambda^a x^{a-1} e^{-\lambda x}}{(a-1)!}, & x \geq 0, \\ 0, & \text{αλλού.} \end{cases} \tag{5.25}$$

Παρατηρήστε ότι από τη σχέση (5.25) για $a = 1$, προκύπτει η εκθετική κατανομή με παράμετρο λ. Η παραπάνω συνεχής κατανομή με σύνολο δυνατών τιμών $x \in (0, \infty)$ είναι μια κατανομή με δύο παραμέτρους, ήτοι την παράμετρο $a \in \{1, 2, 3, ...\}$ (ονομάζεται παράμετρος σχήματος) και την παράμετρο $\lambda > 0$ (ονομάζεται παράμετρος ρυθμού, μιας και συνδέεται με τον ρυθμό της διαδικασίας Poisson). Η κατανομή αυτή είναι γνωστή ως κατανομή **Erlang** ή $\lambda-$ **τάξης κατανομή Erlang**. Η κατανομή Erlang οφείλει το όνομά της στον Agner Krarup Erlang (1878-1929), έναν Δανό μαθηματικό, στατιστικό και μηχανικό, ο οποίος ασχολήθηκε ερευνητικά κυρίως με θέματα της θεωρίας ουρών. Η κατανομή Erlang προέκυψε στο πλαίσιο της μελέτης του σχετικά με τον αριθμό των τηλεφωνικών κλήσεων που θα μπορούσαν να γίνουν ταυτόχρονα στους χειριστές των σταθμών μεταγωγής (στα χρόνια του έτσι λειτουργούσαν οι τηλεφωνικές κλήσεις). Από τότε η κατανομή αυτή έχει χρησιμοποιηθεί σε διάφορα επιστημονικά πεδία, ενώ, όπως θα αποδειχθεί στο Κεφάλαιο 7, είναι η κατανομή του αθροίσματος a το πλήθος ανεξάρτητων και ισόνομων εκθετικών κατανομών με παράμετρο λ.

Παρατήρηση 5.7

Μια εναλλακτική, αλλά ισοδύναμη, παραμετροποίηση για την κατανομή Erlang προκύπτει με χρήση της παραμέτρου $\mu = \frac{1}{\lambda}$. Πιο συγκεκριμένα, από τη σππ της σχέσης (5.25) έχουμε:

$$f_X(x) = \begin{cases} \frac{x^{a-1} e^{-\frac{x}{\mu}}}{\mu^a (a-1)!}, & x \geq 0, \\ 0, & \text{αλλού.} \end{cases} \tag{5.26}$$

Παρατήρηση 5.8

Στην προηγούμενη ενότητα αποδείξαμε ότι η εκθετική κατανομή αποτελεί το συνεχές ανάλογο της γεωμετρικής κατανομής. Με παρόμοιο σκεπτικό αποδεικνύεται ότι η κατανομή Erlang αποτελεί το συνεχές ανάλογο της αρνητικής διωνυμικής κατανομής.

Η γάμμα κατανομή, που αποτελεί αντικείμενο μελέτης σε αυτήν την ενότητα, γενικεύει την κατανομή Erlang επιτρέποντας η παράμετρος a να είναι οποιοσδήποτε θετικός πραγματικός αριθμός. Η γενίκευση αυτή επιτυγχάνεται χρησιμοποιώντας τη γάμμα συνάρτηση αντί για το παραγοντικό (βλ. στο Παράρτημα Β΄, τον ορισμό και τις ιδιότητες της γάμμα συνάρτησης). Έτσι, προκύπτει ο ακόλουθος ορισμός:

Ορισμός 5.4

Η τυχαία μεταβλητή X λέγεται ότι ακολουθεί τη **γάμμα κατανομή** με παραμέτρους $a > 0$ και $\lambda > 0$ αν οι δυνατές της τιμές x είναι $x \geq 0$ και η συνάρτηση πυκνότητας πιθανότητάς της δίνεται από τη σχέση:

$$f_X(x) = \begin{cases} \frac{\lambda^a x^{a-1} e^{-\lambda x}}{\Gamma(a)}, & x \geq 0, \\ 0, & \text{αλλού.} \end{cases} \tag{5.27}$$

Στην περίπτωση αυτή, γράφουμε ότι $X \sim G(a, \lambda)$.

Εναλλακτικά, με παρόμοιο σκεπτικό με την Παρατήρηση 5.5, έχουμε και την ακόλουθη παραμετροποίηση της γάμμα κατανομής:

$$f_X(x) = \begin{cases} \frac{x^{a-1} e^{-\frac{x}{\mu}}}{\mu^a \Gamma(a)}, & x \geq 0, \\ 0, & \text{αλλού.} \end{cases} \tag{5.28}$$

Η παραμετροποίηση με τις παραμέτρους (a, μ) είναι πιο συνηθισμένη σε εφαρμογές στην οικονομετρία και σε διάφορα άλλα επιστημονικά πεδία, ενώ η παραμετροποίηση με τις παραμέτρους (a, λ) χρησιμοποιείται στο πλαίσιο της Μπεϋζιανής Στατιστικής ως συζυγής εκ των προτέρων κατανομή διάφορων παραμέτρων. Στη συνέχεια αυτού του συγγράμματος θα αναφέρουμε ξεκάθαρα σε ποια από τις δύο παραμετροποιήσεις αναφερόμαστε είτε γράφοντας $X \sim G(a, \lambda)$ ή $X \sim G(a, \mu)$ είτε απλώς κάνοντας αναφορά στη σππ.

Παρατήρηση 5.9

Η εκθετική κατανομή με παράμετρο λ προκύπτει από τη σχέση (5.27) για $a = 1$ ή από τη σχέση (5.28) για $\mu = \frac{1}{\lambda}$ και $a = 1$, ενώ στην περίπτωση που η παράμετρος $a \in \{1, 2, ...\}$ οι σχέσεις (5.25) και (5.26) της κατανομής Erlang προκύπτουν άμεσα καθώς τότε $\Gamma(a) = (a-1)!$. Τέλος, στην ειδική περίπτωση που $X \sim G(a = n/2, \lambda = 0.5)$ (ή $X \sim G(a = n/2, \mu = 2)$), δηλαδή όταν έχουμε τη σππ

$$f_X(x) = \begin{cases} \frac{x^{\frac{n-2}{2}} e^{-\frac{x}{2}}}{2^{\frac{n}{2}} \Gamma\left(\frac{n}{2}\right)}, & x \geq 0, \\ 0, & \text{αλλού,} \end{cases} \tag{5.29}$$

λέμε ότι η τ.μ. X ακολουθεί **χι-τετράγωνο κατανομή** με n βαθμούς ελευθερίας και συμβολίζουμε $X \sim \chi^2_n$. Η κατανομή αυτή θα παρουσιαστεί αναλυτικότερα στο Κεφάλαιο 7.

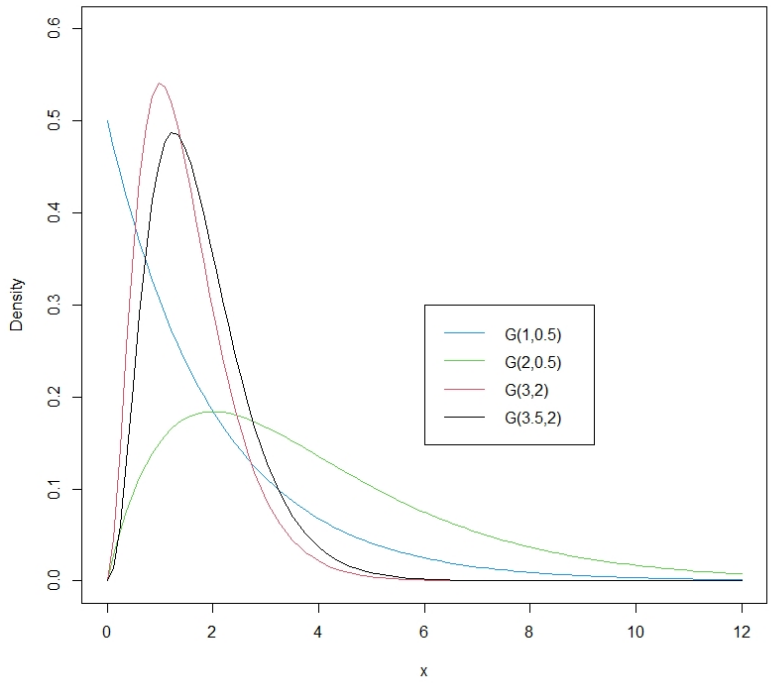
Άσκηση Αυτοαξιολόγησης 5.10

Επιβεβαιώστε ότι η συνάρτηση της σχέσης (5.27) είναι όντως σππ.

Στο Σχήμα 5.6 απεικονίζεται η σππ της γάμμα κατανομής για διάφορους συνδυασμούς των παραμέτρων της, a, λ. Οι εντολές που χρησιμοποιήθηκαν στην R, για να γίνει το Σχήμα 5.6, ήταν οι ακόλουθες:

```
x = seq(0,12, length=100)
plot(x, dgamma(x,1,0.5), ylab="Density", type ="l", col=4, ylim=c(0,0.6))
lines(x, dgamma(x,2,0.5), type ="l", col=3)
lines(x, dgamma(x,3,2), col=2)
lines(x, dgamma(x,3.5,2), col=1)

legend(6,0.3, c("G(1,0.5)","G(2,0.5)","G(3,2)","G(3.5,2)"),lty=c(1,1,1,1),col=
    c(4,3,2,1))
```



Σχήμα 5.6: Γραφική παράσταση της σππ της $G(a,\lambda)$ για $(a,\lambda) = (1,0.5),(2,0.5),(3,2),(3.5,2)$.

Εν συνέχεια, προσδιορίζεται η ασκ της γάμμα κατανομής. Έστω $X \sim G(a,\lambda)$ με σππ που δίνεται στη σχέση (5.27). Τότε προφανώς, για $x \leq 0$ ισχύει ότι $F_X(x) = 0$, ενώ για $x > 0$ εξ ορισμού έχουμε ότι:

$$F_X(x) = P(X \leq x) = \int_0^x \frac{\lambda^a t^{a-1} e^{-\lambda t}}{\Gamma(a)} dt = \frac{\lambda^a}{\Gamma(a)} \int_0^x t^{a-1} e^{-\lambda t} dt = \frac{1}{\Gamma(a)} \int_0^{x\lambda} w^{a-1} e^{-w} dw,$$

όπου το τελευταίο ολοκλήρωμα προέκυψε με την αλλαγή μεταβλητής $\lambda t = w$. Όμως, χρησιμοποιώντας τη σχέση (Β΄.16) του Παραρτήματος Β΄, έχουμε ότι $F_X(x) = \frac{\gamma(a,x\lambda)}{\Gamma(a)}$ για $x > 0$, όπου $\gamma(\cdot,\cdot)$ η κάτω ελλιπής γάμμα συνάρτηση. Συγκεντρωτικά, η ασκ της τ.μ. $X \sim G(a,\lambda)$, με σππ που δίνεται από τη σχέση (5.27), προσδιορίζεται από τη σχέση:

$$F_X(x) = \begin{cases} 0, & x \leq 0, \\ \frac{\gamma(a,x\lambda)}{\Gamma(a)}, & x \geq 0. \end{cases} \tag{5.30}$$

Προφανώς, η ασκ της γάμμα κατανομής δεν δίνεται γενικά σε κλειστή μορφή, αλλά υπολογίζεται με τη βοήθεια λογισμικού ή πινάκων.

Παρατήρηση 5.10

Έστω $X \sim G(a,\lambda)$ με σππ που δίνεται από τη σχέση (5.27). Τότε με τη γλώσσα προγραμματισμού R μπορούμε:

- με τη συνάρτηση `dgamma(x,shape=a,rate=λ)` να υπολογίσουμε τη σππ στο σημείο ή στο διάνυσμα σημείων x,
- με τη συνάρτηση `pgamma(x,shape=a,rate=λ,lower.tail=TRUE)` να υπολογίσουμε την ασκ στο σημείο ή στο διάνυσμα σημείων x,
- με τη συνάρτηση `pgamma(x,shape=a,rate=λ,lower.tail=FALSE)` να υπολογίσουμε την πιθανότητα $P(X > x)$ στο σημείο ή στο διάνυσμα σημείων x,
- με τη συνάρτηση `qgamma(q,shape=a,rate=λ,lower.tail=TRUE)` να προσδιορίσουμε το σημείο ή το διάνυσμα σημείων x για τα οποία ισχύει ότι $P(X \leq x) = q$, όπου

q είναι πιθανότητα ή διάνυσμα με συνιστώσες πιθανότητες, αντίστοιχα,

- με τη συνάρτηση `qgamma(q,shape=a,rate=λ,lower.tail=FALSE` να προσδιορίσουμε το σημείο ή το διάνυσμα σημείων x για τα οποία ισχύει ότι $P(X > x) = q$, όπου q είναι πιθανότητα ή διάνυσμα με συνιστώσες πιθανότητες, αντίστοιχα,
- με τη συνάρτηση `rgamma(n,shape=a,rate=λ)` να δημιουργήσουμε ένα δείγμα μεγέθους n από αυτήν την κατανομή.

Στην περίπτωση που κάποιος θέλει να χρησιμοποιήσει την παραμετροποίηση της γάμμα με σππ που δίνεται στη σχέση (5.28), τότε στα παραπάνω αντικαθιστά το rate=λ με το scale=μ.
Η χρήση κάποιων εκ των παραπάνω εντολών παρουσιάζεται στα παραδείγματα που ακολουθούν.

Άσκηση Αυτοαξιολόγησης 5.11

Να γραφτούν οι εντολές της R που πρέπει να χρησιμοποιηθούν στην R, για να γίνει η γραφική παράσταση της ασκ της $G(a,\lambda)$ για $(a,\lambda) = (1,0.5),(2,0.5),(3,2),(3.5,2)$.

Άσκηση Αυτοαξιολόγησης 5.12

Πώς προκύπτουν από τη σχέση (5.30) ως ειδικές περιπτώσεις οι ασκ της κατανομής Erlang και της εκθετικής κατανομής;

Παρατήρηση 5.11

Σύμφωνα με τους Johnson *et al.* (1994a) η γάμμα κατανομή πρωτοεμφανίστηκε στη βιβλιογραφία από τον Laplace (1836). Από τότε έχει χρησιμοποιηθεί σε διάφορα επιστημονικά πεδία και εφαρμογές. Για παράδειγμα, για τη μοντελοποίηση ασφαλιστικών απαιτήσεων (Boland, 2007), στην υδρολογία (Aksoy, 2000), στην έκφραση βακτηριδιακού γονιδίου (Friedman *et al.*, 2006). Για περισσότερα παραδείγματα εφαρμογών της γάμμα κατανομής παραπέμπουμε μεταξύ άλλων, στο σύγγραμμα των Johnson *et al.* (1994a) και στις εκεί αναφορές.

Στις προτάσεις που ακολουθούν προσδιορίζονται οι ροπές k τάξης και η ροπογεννήτρια συνάρτηση όταν η τ.μ. $X \sim \Gamma(a,\lambda)$ με σππ που δόθηκε στη σχέση (5.27).

Πρόταση 5.11

Έστω X η τυχαία μεταβλητή που ακολουθεί γάμμα κατανομή με παραμέτρους $a > 0$ και $\lambda > 0$, με σππ που προσδιορίζεται στη σχέση (5.27). Τότε η k–τάξης ροπή δίνεται από τη σχέση:

$$E(X^k) = \frac{\Gamma(a+k)}{\lambda^k \Gamma(a)}. \tag{5.31}$$

Απόδειξη Πρότασης 5.11

Εξ ορισμού

$$\begin{aligned} \mathrm{E}(X^k) &= \int_0^{+\infty} x^k \frac{\lambda^a x^{a-1} e^{-\lambda x}}{\Gamma(a)} dx = \frac{\lambda^a}{\Gamma(a)} \int_0^{+\infty} x^{k+a-1} e^{-\lambda x} dx \\ &= \frac{\lambda^a}{\Gamma(a)} \int_0^{+\infty} \frac{z^{k+a-1} e^{-z}}{\lambda^{k+a}} dz = \frac{\Gamma(a+k)}{\lambda^k \Gamma(a)}, \end{aligned}$$

όπου το τελευταίο ολοκλήρωμα προέκυψε από την αλλαγή μεταβλητής $z = \lambda x$ και υπολογίζεται με τη βοήθεια του ορισμού της γάμμα συνάρτησης.

Πρόταση 5.12

Έστω X η τυχαία μεταβλητή που ακολουθεί γάμμα κατανομή με παραμέτρους $a > 0$ και $\lambda > 0$ με σππ που προσδιορίζεται στη σχέση (5.27). Τότε:

$$M_X(t) = \left(1 - \frac{t}{\lambda}\right)^{-a}, \quad \text{για } t < \lambda. \tag{5.32}$$

Απόδειξη Πρότασης 5.12

Λαμβάνοντας υπόψη τη σχέση (3.28), έχουμε ότι:

$$M_X(t) = E\left(e^{tX}\right) = \int_0^{+\infty} e^{tx} \frac{\lambda^a x^{a-1} e^{-\lambda x}}{\Gamma(a)} dx$$
$$= \frac{\lambda^a}{\Gamma(a)} \int_0^{+\infty} x^{a-1} e^{-(\lambda - t)x} dx.$$

Το τελευταίο ολοκλήρωμα είναι πεπερασμένο αν $\lambda - t > 0$, δηλαδή αν $t < \lambda$ και τότε, κάνοντας την αλλαγή μεταβλητής $w = (\lambda - t)x$ και σύμφωνα με τον ορισμό της γάμμα συνάρτησης που δίνεται στη σχέση (Β΄.6), προκύπτει ότι:

$$M_X(t) = \frac{\lambda^a}{\Gamma(a)} \int_0^{+\infty} \frac{w^{a-1}}{(\lambda - t)^a} e^{-w} dw = \frac{\lambda^a}{(\lambda - t)^a}, t < \lambda.$$

Από τις παραπάνω σχέσεις, προφανώς, μπορεί να προκύψουν ως ειδικές περιπτώσεις τα αποτελέσματα που αφορούν τη ροπογεννήτρια και τις ροπές k τάξης της κατανομής Erlang και της εκθετικής με παράμετρο λ. Επιπρόσθετα, από τη σχέση (5.31) για $k = 1,2$ και λαμβάνοντας υπόψη ότι με εφαρμογή της σχέσης (Β΄.9) έχουμε ότι $\Gamma(a+1) = a\Gamma(a)$, $\Gamma(a+2) = (a+1)\Gamma(a+1) = a(a+1)\Gamma(a)$, προκύπτει το ακόλουθο πόρισμα.

Πόρισμα 5.4

Έστω X η τυχαία μεταβλητή που ακολουθεί γάμμα κατανομή με παραμέτρους $a > 0$ και $\lambda > 0$ με σππ που προσδιορίζεται στη σχέση (5.27). Τότε:

$$\mu = E(X) = \frac{a}{\lambda}, \tag{5.33}$$

και

$$\sigma^2 = Var(X) = \frac{a}{\lambda^2}. \tag{5.34}$$

Παρατήρηση 5.12

Στη σχέση (5.28) δόθηκε η σππ της γάμμα κατανομής χρησιμοποιώντας μια εναλλακτική παραμετροποίηση με την παράμετρο $\mu = \frac{1}{\lambda}$. Όλα τα παραπάνω αποτελέσματα (ασκ, ροπογεννήτρια, ροπές k τάξης, μέση τιμή, διακύμανση) μας οδηγούν άμεσα στα αντίστοιχα υπό αυτήν την παραμετροποίηση αντικαθιστώντας την παράμετρο λ με την $\frac{1}{\mu}$.

Παράδειγμα 5.5

Υποθέστε ότι ο χρόνος επισκευής σε ώρες ενός εργοστασιακού μηχανικού εξαρτήματος ακολουθεί την $G(a, \mu)$ με σππ που δίνεται στη σχέση (5.28) με μέση τιμή 1.5 και διακύμανση 0.75.

1. Υπολογίστε την πιθανότητα η επισκευή ενός τέτοιου εργοστασιακού μηχανικού εξαρτήματος να ξεπεράσει σε διάρκεια τις 2 ώρες.

2. Υπολογίστε την πιθανότητα η επισκευή ενός τέτοιου εργοστασιακού μηχανικού εξαρτήματος να διαρκέσει τουλάχιστον 5 ώρες δεδομένου ότι η επισκευή του έχει ήδη ξεπεράσει τις 2 ώρες.
3. Προσδιορίστε την τιμή του χρόνου x που είναι τέτοια, ώστε $P(X \leq x) = 0.75$.

Υπόδειξη: χρησιμοποιήστε την R για τους υπολογισμούς.

Λύση Παραδείγματος 5.5

Έστω X η τ.μ. που παριστάνει τον χρόνο επισκευής (σε ώρες) του εργοστασιακού εξαρτήματος. Σύμφωνα με την εκφώνηση του παραδείγματος η $X \sim G(a, \mu)$ και, λαμβάνοντας υπόψη τις σχέσεις (5.33) και (5.34) υπό αυτήν την παραμετροποίηση, είναι

$$E(X) = a\mu \text{ και } Var(X) = a\mu^2.$$

Επομένως, $a\mu = 1.5$ και $a\mu^2 = 0.75$, από όπου έχουμε ότι $1.5\mu = 0.75$ και έτσι $\mu = 0.5$, άρα και $\lambda = 2$, και $a = 3$. Καταλήξαμε, λοιπόν, ότι $X \sim G(a = 3, \mu = 0.5)$ ή $X \sim G(a = 3, \lambda = 2)$.

1. Είναι
$$P(X > 2) = 1 - \int_0^2 \frac{2^3 x^{3-1} e^{-2x}}{\Gamma(3)} dx.$$
Χρησιμοποιώντας την `pgamma(2,shape=3,rate=2,lower.tail=FALSE)` έχουμε ότι $P(X > 2) = 0.2381033$.
2. Ζητείται η $P(X > 5|X > 2)$. Από τον ορισμό της δεσμευμένης πιθανότητας είναι
$$\begin{aligned} P(X > 5|X > 2) &= \frac{P(X > 5 \text{ και } X > 2)}{P(X > 2)} \\ &= \frac{P(X > 5)}{P(X > 2)} = \frac{0.002769396}{0.2381033} \\ &= 0.01163107, \end{aligned}$$
όπου με την εντολή `pgamma(5,shape=3,rate=2,lower.tail=FALSE)` υπολογίστηκε ο αριθμητής, ενώ ο παρονομαστής είχε υπολογιστεί στο προηγούμενο ερώτημα.
3. Ζητείται να προσδιοριστεί το χρονικό σημείο, έστω x, για το οποίο ισχύει ότι $P(X \leq x) = 0.75$. Η απάντηση προκύπτει χρησιμοποιώντας την εντολή `qgamma(0.75,shape=3,rate=2,lower.tail=TRUE)` και είναι $x = 1.960201$ ώρες.

Άσκηση Αυτοαξιολόγησης 5.13

Έστω ότι ο αριθμός πελατών που καταφθάνουν σε ένα συνοικιακό φαρμακείο ακολουθεί την κατανομή Poisson με μέση τιμή $\lambda = 5$ πελάτες ανά ώρα. Υπολογίστε την πιθανότητα ο χρόνος που μεσολαβεί μεταξύ τριών διαδοχικών πελατών να είναι μικρότερος από 1/3 της ώρας.
Υπόδειξη: για την εύρεση αριθμητικής τιμής χρησιμοποιήστε την R.

5.6 Κανονική κατανομή

Στην ενότητα αυτή θα παρουσιαστεί η πιο ευρέως, για ποικίλους λόγους όπως θα αναφέρουμε, συνεχής κατανομή. Η κατανομή αυτή είναι γνωστή ως κατανομή σφαλμάτων (the law of error, the law of facility of errors and the law of frequency of errors) ή Gaussian κατανομή ή κατανομή Gauss ή κατανομή Laplace-Gauss ή κανονική κατανομή, με την τελευταία να αποτελεί και την κύρια ονομασία της. Ένα πρώτο

εύλογο ερώτημα που μπορεί να τεθεί είναι για ποιο λόγο παρουσιάζεται ή παρουσιάστηκε με όλα αυτά τα διαφορετικά ονόματα. Αρχικά, θα πρέπει να επισημανθεί ότι ο Γερμανός μαθηματικός Johann Carl Friedrich Gauss (1777-1855) στη μονογραφία του με τίτλο “Theoria motus corporum coelestium in sectionibus conicis solem ambientium” οδηγήθηκε στην εισαγωγή αυτής της κατανομής - όπως επίσης και στην παρουσίαση σημαντικών στατιστικών εννοιών, όπως των μεθόδων εκτίμησης ελαχίστων τετραγώνων και μέγιστης πιθανοφάνειας - στο πλαίσιο προσπάθειας μοντελοποίησης σφαλμάτων αστρονομικών μετρήσεων. Τα παραπάνω δικαιολογούν πλήρως την ονομασία Gaussian κατανομή ή κατανομή Gauss ή κατανομή σφαλμάτων. Παρότι ο Gauss ήταν ο πρώτος[3] που παρουσίασε την κατανομή αυτή, ο Γάλλος Pierre-Simon, marquis de Laplace (1749-1827) ήταν αυτός που τη μελέτησε διεξοδικά σε μια σειρά επιστημονικών εργασιών και παρουσίασε μια ιδιότητά της θεμελιώδους σημασίας, διατυπώνοντας το γνωστό στις μέρες μας ως Κεντρικό Οριακό Θεώρημα. Το θεώρημα αυτό θα παρουσιαστεί διεξοδικά στο Κεφάλαιο 7. Η σημαντική συνεισφορά του Laplace είχε ως αποτέλεσμα να αποκαλείται η κατανομή αυτή ως Laplace-Gauss, ειδικά σε γαλλόφωνες περιοχές. Στα τέλη του 19ου αιώνα αρκετοί συγγραφείς[4] άρχισαν να χρησιμοποιούν τον όρο normal όπου η λέξη «κανονική» χρησιμοποιείτο ως επίθετο, θέλοντας να δηλώσουν τη συνήθη κατανομή. Για περισσότερες λεπτομέρειες σχετικά με τις διάφορες ονομασίες παραπέμπουμε μεταξύ άλλων, στον ιστότοπο https://condor.depaul.edu/ntiourir/NormalOrigin.htm (ημερομηνία προσπέλασης: 1/3/2022).

Πριν προχωρήσουμε στον ορισμό της κανονικής κατανομής και σε κάποιες χρήσιμες ιδιότητές της, ας εξηγήσουμε εν συντομία τη σπουδαιότητά της. Αρχικά, η κατανομή αυτή μοντελοποιεί ικανοποιητικά πλήθος τυχαίων φαινομένων σε διάφορα επιστημονικά πεδία. Ενδεικτικά αναφέρουμε το ύψος ενήλικων ατόμων, το βάρος ενός βρέφους κατά τη γέννηση, τη διαστολική και συστολική αρτηριακή πίεση, τα τυχαία σφάλματα που εμφανίζονται σε διάφορες μετρήσεις και ούτω καθεξής. Όμως αυτό που κάνει πραγματικά ξεχωριστή την κανονική κατανομή είναι το Κεντρικό Οριακό Θεώρημα, όπως θα δούμε στο Κεφάλαιο 7, σύμφωνα με το οποίο το άθροισμα και η μέση τιμή n το πλήθος ανεξάρτητων και ισόνομων τυχαίων μεταβλητών από έναν πληθυσμό που περιγράφεται από οποιαδήποτε κατανομή με πεπερασμένη μέση τιμή και διακύμανση, ακολουθούν, για μεγάλες τιμές του n, προσεγγιστικά κανονική κατανομή. Για περισσότερες λεπτομέρειες παραπέμπουμε μεταξύ άλλων, στους Johnson *et al.* (1994a).

Αφού αναφέρθηκαν εν συντομία οι λόγοι της σπουδαιότητας της κανονικής κατανομής, στη συνέχεια δίνεται ο ορισμός της και κάποιες χρήσιμες ιδιότητές της.

Ορισμός 5.5

Η τυχαία μεταβλητή X λέγεται ότι ακολουθεί την **κανονική κατανομή** (normal distribution) με παραμέτρους μ και σ^2 με $\mu \in R$ και $\sigma > 0$, αν η συνάρτηση πυκνότητας πιθανότητάς της δίνεται από τη σχέση:

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}, x \in \mathbb{R}. \quad (5.35)$$

Στην περίπτωση αυτή, γράφουμε ότι $X \sim N(\mu, \sigma^2)$.

[3] Αξίζει να αναφερθεί ότι ταυτόχρονα με τον Gauss και ανεξάρτητα από αυτόν είχε οδηγηθεί στην κανονική κατανομή και ο Ιρλανδός μαθηματικός Robert Adrain (1775-1843) στο πλαίσιο της μοντελοποίησης σφαλμάτων μετρήσεων και της συμβολής του στους εκτιμητές ελαχίστων τετραγώνων. Κάποιοι συγγραφείς θεωρούν ότι ο Γάλλος μαθηματικός Abraham de Moivre (1667-1754) πρωτοπαρουσίασε την κατανομή το 1773, ως προσέγγιση της διωνυμικής κατανομής $B(n,p)$, όταν η παράμετρος n της κατανομής είναι πολύ μεγάλη, αλλά κάποιοι άλλοι ισχυρίζονται ότι απλώς έδωσε έναν προσεγγιστικό κανόνα για τους διωνυμικούς συντελεστές και δεν ανέφερε ότι πρόκειται για σππ.

[4] Μεταξύ αυτών σημαντικό ρόλο διαδραμάτισε ο Άγγλος μαθηματικός και βιοστατιστικός Karl Pearson (1857-1936) στο πλαίσιο της ερευνητικής του ενασχόλησης με δεδομένα που δεν ακολουθούν «κανονικά» αυτήν την κατανομή.

Μια ειδική περίπτωση κανονικής κατανομής, η οποία διαδραματίζει κυρίαρχο ρόλο στην Πιθανοθεωρία και τη Στατιστική, για λόγους που θα δούμε στη συνέχεια, είναι αυτή που προκύπτει από τη $N(\mu,\sigma^2)$ για $\mu=0$ και $\sigma^2=1$, η οποία ονομάζεται **τυπική κανονική κατανομή** (standard normal distribution) και έχει σππ που δίνεται από τη σχέση:

$$f_X(x)=\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{x^2}{2}\right\}, x\in\mathbb{R}. \tag{5.36}$$

Στο σημείο αυτό αξίζει να αναφερθεί ότι, αν η τ.μ. Z ακολουθεί τυπική κανονική κατανομή, δηλαδή αν $Z\sim N(0,1)$, τότε έχει καθιερωθεί στη στατιστική βιβλιογραφία να συμβολίζονται με $\phi(z)$ και $\Phi(z)$ η σππ και η ασκ της, αντίστοιχα.

Ένα πρώτο εύλογο ερώτημα που προκύπτει είναι αν όντως η συνάρτηση της σχέσης (5.35), άρα και της σχέσης (5.36), είναι σππ. Για να επιβεβαιώσουμε ότι η $f_X(x)$ της σχέσης (5.35) είναι όντως σππ, αρκεί να δείξουμε ότι είναι μη αρνητική και το ολοκλήρωμα στο πεδίο ορισμού της είναι ίσο με τη μονάδα (ανατρέξτε στο Κεφάλαιο 3 για τις ιδιότητες που πρέπει να πληροί μια συνάρτηση για να είναι σππ). Από τον τρόπο ορισμού της $f_X(x)$ και καθώς $\sigma>0$, είναι προφανές ότι όντως $f_X(x)\geq 0$. Επιπρόσθετα, ισχύει ότι:

$$\begin{aligned}\int_{-\infty}^{+\infty} f_X(x)dx &= \int_{-\infty}^{+\infty}\frac{1}{\sigma\sqrt{2\pi}}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}dx\\ &= \frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{+\infty}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}dx\\ &= \frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{+\infty}\exp\left(-z^2\right)\sqrt{2}\sigma dz,\end{aligned}$$

όπου το τελευταίο ολοκλήρωμα προέκυψε κάνοντας την αλλαγή μεταβλητής $\frac{x-\mu}{\sqrt{2}\sigma}=z$. Το ολοκλήρωμα που προέκυψε είναι γνωστό στη βιβλιογραφία ως Gaussian integral ή Euler–Poisson integral και υπάρχουν διάφοροι τρόποι υπολογισμού του (βλ., για παράδειγμα, τον ιστότοπο https://en.wikipedia.org/wiki/Gaussian_integral (ημερομηνία προσπέλασης: 1/3/2022)).

Ένας από αυτούς τους τρόπους έχει δοθεί στο Παράρτημα Β΄, σχέση (Β΄.13), από όπου έχουμε ότι:

$$\int_{-\infty}^{+\infty}\exp\left(-z^2\right)=2\int_0^{+\infty}\exp\left(-z^2\right)dz=\sqrt{\pi}.$$

Άρα όντως η συνάρτηση της σχέσης (5.35) είναι σππ.

Πρόταση 5.13

Η σππ της κανονικής κατανομής $N(\mu,\sigma^2)$, όπου $\mu\in\mathbb{R}$ και $\sigma>0$, που δίνεται στη σχέση (5.35), ικανοποιεί τις ακόλουθες ιδιότητες:

1. είναι συμμετρική γύρω από την παράμετρο μ,
2. είναι μονοκόρυφη με κορυφή στο $x=\mu$ με μέγιστη τιμή $f_X(\mu)=\frac{1}{\sigma\sqrt{2\pi}}$,
3. έχει δύο σημεία καμπής, τα σημεία $x=\mu-\sigma$ και $x=\mu+\sigma$, εντός των οποίων στρέφει τα κοίλα προς τα κάτω, ενώ εκατέρωθεν αυτών στρέφει τα κοίλα προς τα άνω,
4. η δεύτερη παράγωγος της σππ πληροί τη σχέση:

$$f_X(x)f_X''(x)\leq\left(f_X'(x)\right)^2,$$

και

5. $f_X(x)\to 0$, όταν $x\to\infty$ ή $x\to-\infty$, δηλαδή η σππ της κανονικής κατανομής προσεγγίζει ασυμπτωτικά τον οριζόντιο άξονα για μεγάλες τιμές.

Απόδειξη Πρότασης 5.13

1. Από τη σχέση (5.35) εύκολα προκύπτει ότι αυτή είναι συμμετρική γύρω από την παράμετρο μ, καθώς ισχύει ότι $f_X(\mu - x) = f_X(\mu + x)$ για κάθε x.
2. Η πρώτη παράγωγος της σππ ικανοποιεί τη σχέση:

$$f'_X(x) = -\frac{x-\mu}{\sigma^2} f_X(x).$$

Επομένως, η πρώτη παράγωγός της μηδενίζεται για $x = \mu$, είναι θετική για $x < \mu$ και είναι αρνητική για $x > \mu$. Άρα η σππ είναι αύξουσα για $x \in (-\infty, \mu)$ και φθίνουσα για $x \in (\mu, +\infty)$. Η μέγιστη τιμή της σππ της κανονικής κατανομής προκύπτει για $x = \mu$ και είναι $f_X(\mu) = \frac{1}{\sigma\sqrt{2\pi}}$.

3. Εύκολα προκύπτει ότι η δεύτερη παράγωγος της σππ ικανοποιεί τη σχέση:

$$f''_X(x) = \frac{(x-\mu)^2 - \sigma^2}{\sigma^4} f_X(x) = \frac{(x-\mu-\sigma)(x-\mu+\sigma)}{\sigma^4} f_X(x).$$

Επομένως, η σππ έχει δύο σημεία καμπής, τα $x = \mu + \sigma$ και $x = \mu - \sigma$, εκατέρωθεν των οποίων αλλάζει πρόσημο η δεύτερη παράγωγος. Ειδικότερα, η $f'_X(x)$ είναι θετική για $x \in (-\infty, \mu - \sigma)$ ή $x \in (\mu + \sigma, +\infty)$ και αρνητική για $x \in (\mu - \sigma, \mu + \sigma)$.

4. Η ζητούμενη σχέση ικανοποιείται καθώς:

$$f_X(x) f''_X(x) = \frac{(x-\mu)^2 - \sigma^2}{\sigma^4} f_X^2(x) \leq \frac{(x-\mu)^2}{\sigma^4} f_X^2(x).$$

Σημειώνουμε ότι για κάθε δύο φορές παραγωγίσιμη, μη αρνητική συνάρτηση $f(\cdot)$, η σχέση:

$$f_X(x) f''_X(x) \leq \left(f'_X(x)\right)^2,$$

αποτελεί ικανή και αναγκαία συνθήκη για να είναι log-concave[α΄].

5. Προκύπτει άμεσα από τις ιδιότητες της εκθετικής συνάρτησης.

[α΄] Η τυχαία μεταβλητή X με σππ $f(\cdot)$ λέμε ότι είναι log-concave, αν για οποιοδήποτε x_1 και x_2 και για οποιοδήποτε $\lambda \in (0,1)$ ισχύει ότι $f(\lambda x_1 + (1-\lambda)x_2) \geq [f(x_1)]^{\lambda}[f(x_2)]^{1-\lambda}$.

Τα αποτελέσματα της παραπάνω πρότασης ήδη μας δίνουν χρήσιμες πληροφορίες για το γράφημα της σππ της κανονικής κατανομής αλλά και τον ρόλο των παραμέτρων μ και σ. Για παράδειγμα, η παράμετρος μ είναι η κορυφή και η διάμεσος της κανονικής κατανομής, ενώ η παράμετρος σ καθορίζει το ύψος της καμπύλης. Ο επιπρόσθετος ρόλος αυτών των παραμέτρων αναδεικνύεται στην επόμενη πρόταση[5].

Πρόταση 5.14

Έστω ότι η τυχαία μεταβλητή $X \sim N(\mu, \sigma^2)$, όπου $\mu \in \mathbb{R}$ και $\sigma > 0$. Τότε:

$$M_X(t) = \exp\left\{\mu t + \frac{1}{2}\sigma^2 t^2\right\}, t \in \mathbb{R}. \tag{5.37}$$

Επιπρόσθετα, $E(X) = \mu$ και $Var(X) = \sigma^2$.

[5] Η απόδειξη ότι $E(X) = \mu$ και $Var(X) = \sigma^2$ θα μπορούσε να γίνει χρησιμοποιώντας και τον ορισμό τους, αλλά θέλοντας να αποφύγουμε τις πολλές πράξεις και παραγοντικές ολοκληρώσεις, προτιμούμε να χρησιμοποιήσουμε τη ροπογεννήτρια συνάρτηση.

Απόδειξη Πρότασης 5.14

Χρησιμοποιώντας τον ορισμό της ροπογεννήτριας συνάρτησης και μετά από λίγες πράξεις έχουμε:

$$\begin{aligned} M_X(t) &= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left\{tx - \frac{(x-\mu)^2}{2\sigma^2}\right\} dx \\ &= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left\{-\frac{x^2 - 2x(\mu + t\sigma^2) + \mu^2}{2\sigma^2}\right\} dx \\ &= \frac{\exp\left\{-\frac{\mu^2}{2\sigma^2}\right\}}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left\{-\frac{x^2 - 2x(\mu + t\sigma^2)}{2\sigma^2}\right\} dx. \end{aligned}$$

Στο τελευταίο ολοκλήρωμα κάνουμε συμπλήρωση της ταυτότητας του τετραγώνου και έτσι προκύπτει:

$$\begin{aligned} M_X(t) &= \frac{\exp\left\{-\frac{\mu^2}{2\sigma^2}\right\}}{\sigma\sqrt{2\pi}} \exp\left\{\frac{(\mu + t\sigma^2)^2}{2}\right\} \int_{-\infty}^{\infty} \exp\left\{-\frac{x^2 - 2x(\mu + t\sigma^2) + (\mu + t\sigma^2)^2}{2\sigma^2}\right\} dx \\ &= \exp\left\{\mu t + \frac{1}{2}\sigma^2 t^2\right\} \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{\left(x - (\mu + t\sigma^2)\right)^2}{2\sigma^2}\right\} dx \\ &= \exp\left\{\mu t + \frac{1}{2}\sigma^2 t^2\right\}, \end{aligned}$$

καθώς το τελευταίο ολοκλήρωμα είναι ίσο με 1 (παρατηρήστε ότι εντός του ολοκληρώματος έχουμε τη σππ της κανονικής με μέση τιμή $(\mu + t\sigma^2)$ και διακύμανση σ^2).
Επιπρόσθετα,

$$\begin{aligned} E(X) &= \frac{d}{dt} M_X(t)|_{t=0} = \frac{d}{dt}\left(\exp\left\{\mu t + \frac{1}{2}\sigma^2 t^2\right\}\right)_{t=0} \\ &= \left(\left(\mu + \sigma^2 t\right) \exp\left\{\mu t + \frac{1}{2}\sigma^2 t^2\right\}\right)|_{t=0} = \mu, \end{aligned}$$

ενώ

$$\begin{aligned} E(X^2) &= \frac{d^2}{dt^2} M_X(t)|_{t=0} = \frac{d}{dt}\left(\mu + \sigma^2 t\right) \exp\left\{\mu t + \frac{1}{2}\sigma^2 t^2\right\}|_{t=0} \\ &= \left(\sigma^2 \exp\left\{\mu t + \frac{1}{2}\sigma^2 t^2\right\} + (\mu + \sigma^2 t)^2 \exp\left\{\mu t + \frac{1}{2}\sigma^2 t^2\right\}\right)|_{t=0} \\ &= \sigma^2 + \mu^2. \end{aligned}$$

Άρα, καθώς $Var(X) = E(X^2) - (E(X))^2$, προκύπτει το ζητούμενο.

Άσκηση Αυτοαξιολόγησης 5.14

Έστω ότι η τυχαία μεταβλητή X ακολουθεί κανονική κατανομή $N(\mu, \sigma^2)$, όπου $\mu \in \mathbb{R}$ και $\sigma > 0$. Τότε, χρησιμοποιώντας τον ορισμό της μέσης τιμής και της διακύμανσης, να αποδείξετε ότι $E(X) = \mu$ και $Var(X) = \sigma^2$.

Πρόταση 5.15

Έστω ότι η τυχαία μεταβλητή $X \sim N(\mu, \sigma^2)$, όπου $\mu \in \mathbb{R}$ και $\sigma > 0$, τότε

$$E\left[\left(X-\mu\right)^k\right] = 0 \text{ για } k \text{ περιττό,} \tag{5.38}$$

ενώ

$$E\left[\left(X-\mu\right)^k\right] = \sigma^k \frac{2^{k/2}}{\sqrt{\pi}}\Gamma\left(\frac{k+1}{2}\right) \text{ για } k \text{ άρτιο.} \tag{5.39}$$

Απόδειξη Πρότασης 5.15

Εξ ορισμού έχουμε ότι:

$$\begin{aligned} E\left[\left(X-\mu\right)^k\right] &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma}(x-\mu)^k e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx \\ &\overset{y=x-\mu}{=} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} y^k e^{-\frac{y^2}{2\sigma^2}} dy \overset{y=\sigma u}{=} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma}\sigma^k u^k e^{-\frac{u^2}{2}}\sigma du \\ &= \sigma^k \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} u^k e^{-\frac{u^2}{2}} du. \end{aligned}$$

Καθώς το ολοκλήρωμα περιττής συνάρτησης σε όλη την ευθεία των πραγματικών αριθμών είναι ίσο με μηδέν, προκύπτει άμεσα ότι το τελευταίο ολοκλήρωμα είναι μηδέν όταν το k είναι περιττός αριθμός. Από την άλλη πλευρά, στην περίπτωση που k είναι άρτιος, δηλαδή όταν $k = 2 \cdot \ell$, έχουμε ότι:

$$\begin{aligned} E\left[\left(X-\mu\right)^k\right] &\overset{k=2\ell}{=} \sigma^{2\ell} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} u^{2l} e^{-\frac{u^2}{2}} du = 2\sigma^{2l} \int_0^{\infty} \frac{1}{\sqrt{2\pi}} u^{2l} e^{-\frac{u^2}{2}} du \\ &\overset{u=\sqrt{2w}}{=} \frac{2\sigma^{2\ell}}{\sqrt{2\pi}} \int_0^{\infty} (2w)^{\ell} e^{-w} \frac{1}{\sqrt{2w}} dw = \frac{2^{\ell}\sigma^{2\ell}}{\sqrt{\pi}} \int_0^{\infty} w^{\ell-1/2} e^{-w} dw = \frac{2^{\ell}\sigma^{2l}}{\sqrt{\pi}}\Gamma\left(\ell+\frac{1}{2}\right) \\ &\overset{\ell=k/2}{=} \sigma^k \frac{2^{k/2}}{\sqrt{\pi}}\Gamma\left(\frac{k+1}{2}\right) \end{aligned}$$

που αποδεικνύει το ζητούμενο.

Σημειώνουμε ότι θα μπορούσαμε να αποδείξουμε τα παραπάνω αποτελέσματα προσδιορίζοντας τις k ροπές της τυχαίας μεταβλητής $Y = X - \mu$ με τη βοήθεια της ροπογεννήτριας συνάρτησής της

$$M_Y(t) = \exp\left\{0.5\sigma^2 t^2\right\}, t \in \mathbb{R}$$

και ανακαλώντας ότι $E(Y^k)$ προσδιορίζεται από την k παράγωγό της ως προς t για $t = 0$. Ο τρόπος αυτός απόδειξης αφήνεται ως άσκηση στον/στην αναγνώστη/στρια.

Παρατήρηση 5.13

Συνοψίζοντας, αν $X \sim N(\mu, \sigma^2)$, τότε η παράμετρος μ αποτελεί το κέντρο συμμετρίας της σππ της κανονικής κατανομής και ταυτόχρονα είναι και η μέση τιμή της τ.μ. Για τον λόγο αυτό ονομάζεται και παράμετρος θέσης. Από την άλλη πλευρά η παράμετρος σ^2 καθορίζει τη μέγιστη τιμή της σππ. Μάλιστα, ισούται με τη διακύμανση της τ.μ. και πολλές φορές καλείται παράμετρος σχήματος.

α. Έστω $X_1 \sim N(\mu, \sigma_1^2)$ και $X_2 \sim N(\mu, \sigma_2^2)$ με $\sigma_1^2 < \sigma_2^2$. Έχουμε, λοιπόν, ότι οι υπό θεώρηση σππ είναι συμμετρικές γύρω από το σημείο μ με τη μέγιστη τιμή της πρώτης ($f_{X_1}(\mu) = \frac{1}{\sigma_1\sqrt{2\pi}}$) να είναι

μεγαλύτερη από την αντίστοιχη της δεύτερης ($f_{X_2}(\mu) = \frac{1}{\sigma_2\sqrt{2\pi}}$). Καθώς το εμβαδόν κάτω από την καμπύλη της σππ είναι ίσο με τη μονάδα συμπεραίνουμε ότι όσο μικρότερη είναι η διακύμανση τόσο λιγότερο, ας μας επιτραπεί η έκφραση, απλωμένη είναι η σππ.

β. Από την άλλη πλευρά, αν $X_1 \sim N(\mu_1,\sigma^2)$ και $X_2 \sim N(\mu_2,\sigma^2)$ έχουμε ότι παρότι έχουν διαφορετικό κέντρο συμμετρίας, η μέγιστη τιμή που λαμβάνει η σππ τους είναι ίδια και ίση με $f_{X_1}(\mu_1) = f_{X_2}(\mu_2) = \frac{1}{\sigma\sqrt{2\pi}}$. Επίσης, παρατηρούμε ότι $f_{X_1}(x) = f_{X_2}(x + \mu_2 - \mu_1)$, για κάθε $x \in R$. Άρα σε αυτήν την περίπτωση το γράφημά τους είναι ίδιο ως προς το σχήμα με μια μετατόπιση κατά $\mu_2 - \mu_1$.

Οι παραπάνω παρατηρήσεις επιβεβαιώνονται και από τα Σχήματα 5.7 και 5.8 στα οποία απεικονίζεται η σππ της κανονικής κατανομής για διάφορους συνδυασμούς των παραμέτρων της.

Οι εντολές που χρησιμοποιήθηκαν στην R, για να γίνει το Σχήμα 5.7, ήταν οι ακόλουθες:

```
x = seq(-6,6, length=100)
plot(x, dnorm(x, 0, sqrt(0.3)), ylab="Density", type ="l", col=4)
lines(x, dnorm(x, 0, 1), type ="l", col=3)
lines(x, dnorm(x, 0, 2), type ="l", col=2)
lines(x, dnorm(x, 0, sqrt(6)), type ="l", col=1)

legend(2,0.4, c("N(0,0.3)","N(0,1)","N(0,4)","N(0,6)"),lty=c(1,1,1,1),col=c
    (4,3,2,1))
```

Για την ασκ της $X \sim N(\mu,\sigma^2)$ έχουμε:

$$F_X(x) = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(t-\mu)^2}{2\sigma^2}\right\} dt$$
$$\overset{\frac{t-\mu}{\sigma}=u}{=} \int_{-\infty}^{\frac{x-\mu}{\sigma}} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{u^2}{2}\right\} du.$$

Τι, όμως, αντιπροσωπεύει το τελευταίο ολοκλήρωμα; Παρατηρήστε ότι εντός του ολοκληρώματος είναι η σππ της τυπικής κανονικής που δόθηκε στη σχέση (5.36) και τα άκρα του ολοκληρώματος και θα προκύψει η προφανής σχέση:

$$F_X(x) = F_Z\left(\frac{x-\mu}{\sigma}\right), \forall\, x \in \mathbb{R}, \text{ όπου } X \sim N(\mu,\sigma^2) \text{ και } Z \sim N(0,1). \tag{5.40}$$

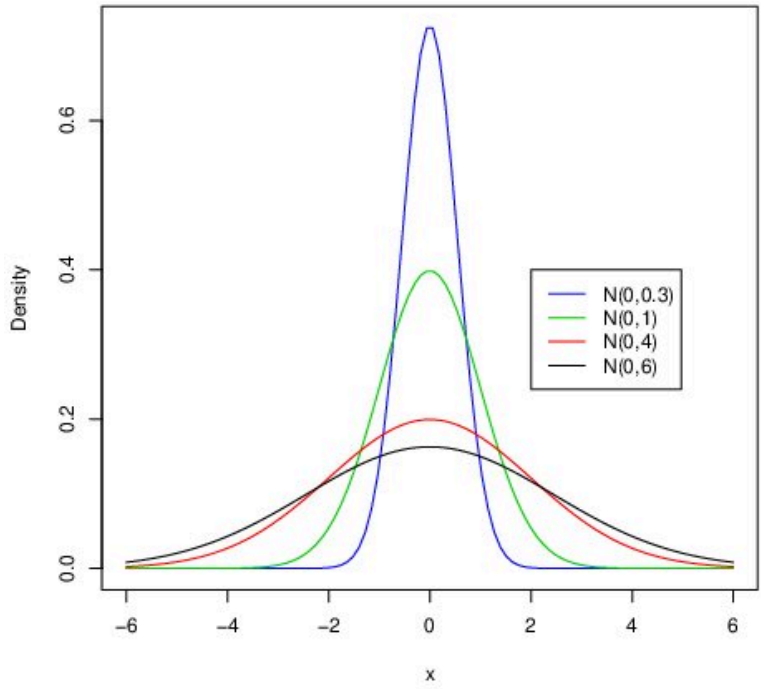
Με παρόμοιο τρόπο μπορούμε να αποδείξουμε ότι:

$$P(a \le X \le b) = P\left(\frac{a-\mu}{\sigma} \le Z \le \frac{b-\mu}{\sigma}\right), \forall\, a,b \in \mathbb{R},\ X \sim N(\mu,\sigma^2), Z \sim N(0,1). \tag{5.41}$$

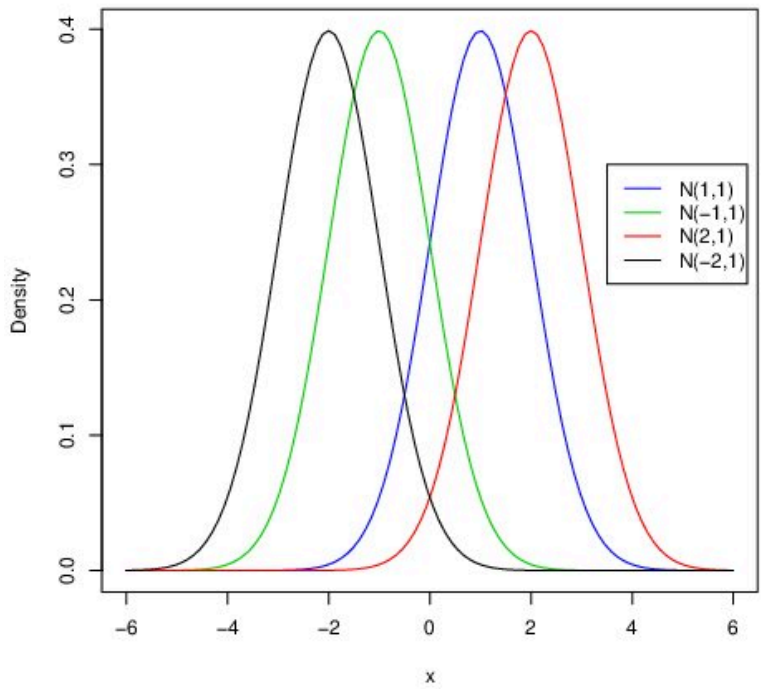
Επομένως, από τα παραπάνω εξάγεται άμεσα το συμπέρασμα ότι τόσο η εύρεση της ασκ της $N(\mu,\sigma^2)$ όσο και ο υπολογισμός πιθανοτήτων με αυτήν ανάγεται στην εύρεση της ασκ της $N(0,1)$ και στον υπολογισμό πιθανοτήτων χρησιμοποιώντας τη $N(0,1)$.

Το εύλογο ερώτημα που προκύπτει τώρα είναι αν η ασκ της τυπικής κανονικής κατανομής $\Phi(z)$ υπολογίζεται σε κλειστή μορφή, αν δηλαδή υπολογίζεται σε κλειστή μορφή το ολοκλήρωμα:

$$\int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{u^2}{2}\right\} du,$$



Σχήμα 5.7: Γραφική παράσταση της σππ της $N(\mu,\sigma^2)$ για $(\mu,\sigma^2)=(0,0.3),(0,1),(0,4),(0,6)$.



Σχήμα 5.8: Γραφική παράσταση της σππ της $N(\mu,\sigma^2)$ για $(\mu,\sigma^2)=(1,1),(-1,1),(2,1),(-2,1)$.

ή αν υπολογίζονται σε κλειστή μορφή ολοκληρώματα της μορφής:

$$\int_{z_1}^{z_2} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{u^2}{2}\right\} du.$$

Οι απαντήσεις, δυστυχώς, στα παραπάνω ερωτήματα είναι αρνητικές. Παρ΄ όλα αυτά ο υπολογισμός αυτών των ολοκληρωμάτων - άρα και των αντίστοιχων πιθανοτήτων - καθίσταται εφικτός με αριθμητικές μεθόδους. Στη βιβλιογραφία υπάρχουν πίνακες που προέκυψαν με αριθμητικές μεθόδους και προσδιορίζουν πιθανότητες μίας εκ των ακόλουθων μορφών: $P(Z \leq z)$, $P(0 \leq Z \leq z)$, $P(Z \geq z)$, $P(-z \leq Z \leq z)$ για $z \geq 0$. Χρησιμοποιώντας τις ιδιότητες της σπ της τυπικής κανονικής (συμμετρία γύρω από το μηδέν, εμβαδόν μεταξύ της καμπύλης της σπ και του οριζόντιου άξονα ίσο με 1) εύκολα γίνεται αντιληπτό ότι οποιασδήποτε μορφής πιθανότητα μας δοθεί μπορεί να υπολογιστεί με τη βοήθεια ενός και μόνο πίνακα από τους παραπάνω. Στο Παράρτημα Α΄ δίνεται πίνακας που υπολογίζει την ασκ της τυπικής κανονικής κατανομής (βλ. τον Πίνακα Α΄.3), δηλαδή τις πιθανότητες $\Phi(z) = P(Z \leq z)$, για $z \geq 0$ με z γνωστό αριθμό. Προφανώς, χρησιμοποιώντας τις σχέσεις (5.40) και (5.41), ο πίνακας αυτός προσδιορίζει την τιμή της ασκ κάθε κανονικής κατανομής, καθώς και οποιασδήποτε μορφής πιθανότητας για αυτήν.

Όλα αυτά θα επεξηγηθούν αναλυτικά μέσω των παραδειγμάτων που ακολουθούν. Προηγουμένως, όμως, θα δοθεί μια πρόταση, πόρισμα της οποίας έχει κατά έναν τρόπο ήδη χρησιμοποιηθεί.

Πρόταση 5.16

Έστω ότι η τυχαία μεταβλητή X ακολουθεί κανονική κατανομή $N(\mu, \sigma^2)$. Τότε η τυχαία μεταβλητή $Y = aX + b$ με $a, b \in \mathbb{R}$ ακολουθεί κανονική κατανομή $N(a\mu + b, (a\sigma)^2)$, δηλαδή κανονική κατανομή με μέση τιμή $a\mu + b$ και τυπική απόκλιση ίση με $|a|\sigma$.

Απόδειξη Πρότασης 5.16

Καθώς η τ.μ. $Y = aX + b$ με $a, b \in R$ είναι συνάρτηση της τ.μ. X, θα χρησιμοποιήσουμε για την απόδειξη της πρότασης τη μέθοδο της ροπογεννήτριας (ανατρέξτε στην Ενότητα 3.7). Είναι

$$M_Y(t) = E(e^{tY}) = E\left(e^{t(aX+b)}\right) = e^{tb}E\left(e^{taX}\right) = e^{tb}M_X(at). \tag{5.42}$$

Όμως η ροπογεννήτρια της $N(\mu, \sigma^2)$ έχει προσδιοριστεί στη σχέση (5.37), από όπου αντικαθιστώντας το t από το at, είναι:

$$M_Y(t) = e^{tb}e^{\mu at + 0.5\sigma^2(at)^2} = \exp\{(a\mu + b)t + 0.5(\sigma a)^2 t^2\},$$

η οποία ταυτίζεται με τη ροπογεννήτρια της $N(a\mu + b, (a\sigma)^2)$ και λόγω του Θεωρήματος του Μονοσήμαντου των Ροπογεννητριών (ανατρέξτε στην Ενότητα 3.6.3) έχουμε το προς απόδειξη αποτέλεσμα.

Σύμφωνα με την παραπάνω πρόταση, αν μια τυχαία μεταβλητή ακολουθεί κανονική κατανομή, τότε και κάθε γραμμικός μετασχηματισμός της τυχαίας μεταβλητής εξακολουθεί να ακολουθεί κανονική κατανομή, με διαφορετικές όμως παραμέτρους. Δηλαδή, μία από τις χρησιμότερες ιδιότητες της κανονικής κατανομής είναι η ιδιότητα του αναλλοίωτου της κατανομής αυτής υπό γραμμικούς μετασχηματισμούς. Στην ειδική περίπτωση που $a = \frac{1}{\sigma}$ και $b = -\frac{\mu}{\sigma}$, προκύπτει το ακόλουθο πόρισμα.

Πόρισμα 5.5

Έστω ότι η τυχαία μεταβλητή X ακολουθεί κανονική κατανομή $N(\mu, \sigma^2)$, $\mu \in \mathbb{R}$, $\sigma > 0$. Τότε η τυχαία μεταβλητή $Z = \frac{X-\mu}{\sigma}$ (**τυπικός μετασχηματισμός**) ακολουθεί τυπική κανονική κατανομή, δηλαδή $Z \sim N(0,1)$.

Επομένως, από τα παραπάνω καταλαβαίνουμε ότι κάθε κανονική κατανομή μπορεί να μετατραπεί σε τυπική κανονική κατανομή και ο υπολογισμός πιθανοτήτων για οποιαδήποτε τ.μ. που ακολουθεί κανονική κατανομή ανάγεται στον υπολογισμό πιθανοτήτων από την τυπική κανονική κατανομή (ή προφανώς οποιαδήποτε άλλης κανονικής για την οποία έχουμε διαθέσιμους πίνακες).

Παρατήρηση 5.14

Έστω $X \sim N(\mu,\sigma^2)$ με σππ που δίνεται από τη σχέση (5.35). Τότε με τη γλώσσα προγραμματισμού R μπορούμε:

- με τη συνάρτηση `dnorm(x,μ,σ)` να υπολογίσουμε τη σππ στο σημείο ή στο διάνυσμα σημείων x,
- με τη συνάρτηση `pnorm(x,μ,σ,lower.tail=TRUE)` να υπολογίσουμε την ασκ στο σημείο ή στο διάνυσμα σημείων x,
- με τη συνάρτηση `pnorm(x,μ,σ,lower.tail=FALSE)` να υπολογίσουμε την πιθανότητα $P(X > x)$ στο σημείο ή στο διάνυσμα σημείων x,
- με τη συνάρτηση `qnorm(q,μ,σ,lower.tail=TRUE)` να προσδιορίσουμε το σημείο ή το διάνυσμα σημείων x για τα οποία ισχύει ότι $P(X \le x) = q$, όπου q είναι πιθανότητα ή διάνυσμα με συνιστώσες πιθανότητες, αντίστοιχα,
- με τη συνάρτηση `qnorm(q,μ,σ,lower.tail=FALSE` να προσδιορίσουμε το σημείο ή το διάνυσμα σημείων x για τα οποία ισχύει ότι $P(X > x) = q$, όπου q είναι πιθανότητα ή διάνυσμα με συνιστώσες πιθανότητες, αντίστοιχα,
- με τη συνάρτηση `rnorm(n,μ,σ)` να δημιουργήσουμε ένα δείγμα μεγέθους n από αυτήν την κατανομή.

Η χρήση κάποιων εκ των παραπάνω εντολών παρουσιάζεται στα παραδείγματα που ακολουθούν.

Παράδειγμα 5.6

Αν η τ.μ. Z ακολουθεί τυπική κανονική κατανομή να υπολογίσετε με τη βοήθεια του πίνακα της τυπικής κανονικής του Παραρτήματος Α΄, τις ακόλουθες πιθανότητες: $P(Z < 1.27)$, $P(Z > 1.06)$, $P(Z < -2.15)$, $P(Z > -3.15)$, $P(1.06 < Z < 3.78)$ και $P(-1.26 < Z < 2.37)$.

Λύση Παραδείγματος 5.6

Ο πίνακας δίνει άμεσα πιθανότητες της μορφής $P(Z \le z)$ για $z \ge 0$ με z γνωστό αριθμό. Επομένως, χρησιμοποιώντας τον μπορούμε να υπολογίσουμε άμεσα την πιθανότητα $P(Z < 1.27)$. Αυτό επιτυγχάνεται με τον ακόλουθο τρόπο: εντοπίζουμε στην πρώτη στήλη του πίνακα την τιμή 1.2 (δηλαδή τον αριθμό μέχρι το πρώτο δεκαδικό ψηφίο) και στην πρώτη γραμμή του πίνακα την τιμή 0.07 (δηλαδή το δεύτερο δεκαδικό ψηφίο του αριθμού). Η ζητούμενη πιθανότητα βρίσκεται στη διασταύρωση της γραμμής και της στήλης που έχουμε εντοπίσει και είναι: $P(Z < 1.27) = 0.89796$.

Για τον υπολογισμό όλων των υπόλοιπων πιθανοτήτων θα πρέπει να χρησιμοποιηθεί η ιδιότητα της συμμετρίας της σππ της τυπικής κανονικής κατανομής γύρω από το 0, καθώς και το γεγονός ότι το εμβαδόν μεταξύ της καμπύλης της και του οριζόντιου άξονα είναι ίσο με 1. Χρησιμοποιώντας τη σχέση $P(A) = 1 - P(A')$ είναι:

$$P(Z > 1.06) = 1 - P(Z \le 1.06) = 1 - 0.85543 = 0.14457.$$

Χρησιμοποιώντας τη συμμετρία της σππ γύρω από το μηδέν προκύπτει ότι:

$$P(Z < -2.15) = P(Z \ge 2.15) = 1 - P(Z < 2.15) = 1 - 0.98422 = 0.01578.$$

Με παρόμοιο τρόπο είναι:

$$\begin{aligned} P(Z > -3.15) &= 1 - P(Z \le -3.15) = 1 - P(Z \ge 3.15) \\ &= 1 - (1 - P(Z < 3.15)) = P(Z < 3.15) = 0.99918, \end{aligned}$$

ή και απευθείας, λόγω συμμετρίας, $P(Z > -3.15) = P(Z < 3.15) = 0.99918$.
Για την απάντηση στο επόμενο ερώτημα χρησιμοποιώντας τη σχέση

$$P(a < Z < b) = \Phi(b) - \Phi(a)$$

προκύπτει ότι:

$$P(1.06 < Z < 3.78) = P(Z < 3.78) - P(Z < 1.06) = 0.99992 - 0.85543 = 0.14449.$$

Με παρόμοιο σκεπτικό και χρησιμοποιώντας τη συμμετρία της σππ, έχουμε ότι:

$$\begin{aligned} P(-1.26 < Z < 2.37) &= P(Z < 2.37) - P(Z < -1.26) = P(Z < 2.37) - P(Z > 1.26) \\ &= P(Z < 2.37) - (1 - P(Z \leq 1.26)) \\ &= P(Z < 2.37) + P(Z < 1.26) - 1 \\ &= 0.99111 + 0.89617 - 1 = 0.88728. \end{aligned}$$

Χρησιμοποιώντας την R τα παραπάνω αποτελέσματα μπορούν να ληφθούν μέσω των παρακάτω εντολών

- `pnorm(1.27,0,1,lower.tail=TRUE)`,
- `pnorm(1.06,0,1,lower.tail=FALSE)`,
- `pnorm(-2.15,0,1,lower.tail=TRUE)`,
- `pnorm(-3.15,0,1,lower.tail=FALSE)`,
- `pnorm(3.78,0,1,lower.tail=TRUE)-pnorm(1.06,0,1,lower.tail=TRUE)` και
- `pnorm(2.37,0,1,lower.tail=TRUE)-pnorm(-1.26,0,1,lower.tail=TRUE)`,

οι οποίες επιστρέφουν τις τιμές 0.8979577, 0.1445723, 0.01577761, 0.9991836, 0.1444939 και 0.8872713, αντίστοιχα. Οι τιμές αυτές είναι πρακτικά ίσες με τις τιμές που λάβαμε χρησιμοποιώντας τον πίνακα της τυπικής κατανομής. Οι όποιες διαφορές εντοπίζονται στο πέμπτο δεκαδικό ψηφίο λόγω της στρογγυλοποίησης που χρησιμοποιείται στην κατασκευή του πίνακα.

Άσκηση Αυτοαξιολόγησης 5.15

Αν η τ.μ. Z ακολουθεί τυπική κανονική κατανομή να υπολογίσετε με τη βοήθεια του πίνακα της $N(0,1)$ του Παραρτήματος Α΄, τις ακόλουθες πιθανότητες: $P(Z < -1.27)$, $P(Z > -1.06)$, $P(Z < 2.15)$, $P(Z > 3.15)$, $P(-3.78 < Z < -1.06)$, $P(-2.37 < Z < 1.26)$ και $P(-2.37 < Z < 1.26)$.

Πρωτύτερα παρουσιάστηκε ο τρόπος υπολογισμού πιθανοτήτων οποιασδήποτε μορφής, όταν η τ.μ. Z ακολουθεί τυπική κανονική κατανομή χρησιμοποιώντας τον πίνακα της $N(0,1)$ του Παραρτήματος Α΄. Ο ίδιος πίνακας, μέσω του τυπικού μετασχηματισμού, χρησιμεύει και για τον υπολογισμό οποιασδήποτε μορφής πιθανότητας, όταν η τ.μ. $X \sim N(\mu, \sigma^2)$, μέσω της σχέσης (5.41). Για την κατανόηση της διαδικασίας ακολουθεί ένα παράδειγμα.

Παράδειγμα 5.7

Το βάρος που αντέχει ένα συρματόσχοινο συγκεκριμένου τύπου περιγράφεται ικανοποιητικά από την κανονική κατανομή με μέση τιμή 3 τόνους και τυπική απόκλιση 0.1 τόνους. Υπολογίστε την πιθανότητα ένα τυχαία επιλεγμένο συρματόσχοινο αυτού του τύπου να αντέξει βάρος μικρότερο από 3.15 τόνους, μεταξύ 3.15 και 3.3 τόνων, μεγαλύτερο από 2.9 τόνους και, τέλος, μεταξύ 2.9 και 3.15 τόνους.

Λύση Παραδείγματος 5.7

Έστω X η τ.μ. που περιγράφει το βάρος που αντέχει ένα συρματόσχοινο του υπό θεώρηση συγκεκριμένου τύπου. Τότε, σύμφωνα με τα δεδομένα του παραδείγματος, είναι $X \sim N(\mu = 3, \sigma^2 = 0.1^2)$. Μας ζητείται να υπολογιστούν οι πιθανότητες $P(X < 3.15)$, $P(3.15 < X < 3.3)$, $P(X > 2.9)$ και, τέλος, $P(2.9 < X < 3.15)$. Οι πιθανότητες αυτές θα αναχθούν σε πιθανότητες τυπικής κανονικής κατανομής μέσω της σχέσης (5.41), ήτοι του τυπικού μετασχηματισμού. Σε αυτό το πλαίσιο, έχουμε:

$$P(X < 3.15) = P\left(\frac{X-3}{0.1} < \frac{3.15-3}{0.1}\right) = P(Z < 1.5) = 0.93319,$$

$$\begin{aligned} P(3.15 < X < 3.3) &= P\left(\frac{3.15-3}{0.1} < \frac{X-3}{0.1} < \frac{3.3-3}{0.1}\right) \\ &= P(1.5 < Z < 3) = 0.99865 - 0.93319 = 0.06546, \end{aligned}$$

$$P(X > 2.9) = P\left(\frac{X-3}{0.1} > \frac{2.9-3}{0.1}\right) = P(Z > -1) = P(Z < 1) = 0.84134,$$

και

$$\begin{aligned} P(2.9 < X < 3.15) &= P\left(\frac{2.9-3}{0.1} < \frac{X-3}{0.1} < \frac{3.15-3}{0.1}\right) \\ &= P(-1 < Z < 1.5) = P(Z < 1.5) - P(Z < -1) \\ &= P(Z < 1.5) - P(Z > 1) \\ &= P(Z < 1.5) + P(Z < 1) - 1 = 0.93319 + 0.84134 - 1 \\ &= 0.77453. \end{aligned}$$

Χρησιμοποιώντας την R, τα παραπάνω αποτελέσματα μπορούν να ληφθούν μέσω των παρακάτω εντολών:

- `pnorm(3.15,3,0.1,lower.tail=TRUE)`,
- `pnorm(3.3,3,0.1,lower.tail=TRUE)-pnorm(3.15,3,0.1,lower.tail=TRUE)`
- `pnorm(2.9,3,0.1,lower.tail=FALSE)` και
- `pnorm(3.15,3,0.1,lower.tail=TRUE)-pnorm(2.9,3,0.1,lower.tail=TRUE)`

οι οποίες επιστρέφουν τις τιμές 0.9331928, 0.0654573, 0.8413447 και 0.7745375, αντίστοιχα, οι οποίες είναι πάλι πρακτικά ίσες με τις τιμές που λάβαμε παραπάνω χρησιμοποιώντας τον πίνακα της τυπικής κατανομής.

Μέχρι τώρα ο πίνακας της $N(0,1)$ του Παραρτήματος Α΄ χρησιμοποιήθηκε για τον υπολογισμό κάθε μορφής πιθανότητας. Κάποιες φορές όμως χρειάζεται να υπολογίσουμε την τιμή z η οποία είναι τέτοια, ώστε $\Phi(z) = p$ με $0 < p < 1$ γνωστή (δοθείσα) πιθανότητα. Πρόκειται, επομένως, για το αντίστροφο πρόβλημα της εύρεσης του σημείου που είναι τέτοιο, ώστε να προκύπτει κάποια δοθείσα πιθανότητα. Ο τρόπος που μπορεί να

επιτευχθεί ο προσδιορισμός του σημείου με τη βοήθεια του ίδιου πίνακα θα εξηγηθεί μέσω παραδειγμάτων, αφού προηγηθεί ο ακόλουθος ορισμός.

Ορισμός 5.6

Έστω $Z \sim N(0,1)$. Ορίζουμε ως z_α την τιμή της τυπικής κανονικής κατανομής που είναι τέτοια, ώστε

$$P(Z > z_\alpha) = \alpha \text{ για } \alpha \in (0,1) \text{ γνωστό αριθμό.} \quad (5.43)$$

Παράδειγμα 5.8

Να υπολογίσετε τα: $z_{0.025}$, $z_{0.05}$, $z_{0.01}$ και $z_{0.005}$.

Λύση Παραδείγματος 5.8

Σύμφωνα με τη σχέση (5.43) έχουμε ότι: $P(Z > z_{0.025}) = 0.025$. Καθώς ο πίνακας της $N(0,1)$ του Παραρτήματος Α΄ δίνει πιθανότητες της μορφής $P(-\infty \leq Z \leq z)$ για $z \geq 0$, για να μπορεί να χρησιμοποιηθεί, θα πρέπει να μετατρέψουμε την παραπάνω πιθανότητα σε αυτή τη μορφή. Χρησιμοποιώντας τη σχέση $P(A) = 1 - P(A')$ έχουμε ότι: $1 - P(Z \leq z_{0.025}) = 0.025$ ή, ισοδύναμα, ότι $P(Z \leq z_{0.025}) = 0.975$. Για τον σκοπό αυτόν αναζητούμε, επομένως, τον αριθμό εκείνο που δίνει πιθανότητα 0.975, ψάχνοντας να εντοπίσουμε την τιμή 0.975 στο κύριο σώμα του πίνακα (εντός του πίνακα). Τότε ο ζητούμενος αριθμός προκύπτει από τη διασταύρωση της γραμμής με τη στήλη στην οποία βρίσκεται (η αντίστροφη διαδικασία από αυτήν για τον υπολογισμό πιθανοτήτων). Έτσι, άμεσα προκύπτει ότι $z_{0.025} = 1.96$.

Με παρόμοιο τρόπο για τον προσδιορισμό του $z_{0.05}$ προσπαθούμε να εντοπίσουμε το 0.95. Καθώς ο αριθμός αυτός δεν υπάρχει (παρατηρήστε ότι υπάρχουν οι αριθμοί 0.94950 και 0.95053) οδηγούμαστε χρησιμοποιώντας τον πίνακα της $N(0,1)$ του Παραρτήματος Α΄ ότι ο ζητούμενος αριθμός είναι μεταξύ του 1.64 και του 1.65. Η ακριβής τιμή μπορεί να προσδιοριστεί με χρήση στατιστικών προγραμμάτων (βλ. συνέχεια) και ισούται με 1.64485. Για αυτόν τον λόγο, όταν γίνεται στρογγυλοποίηση στα δύο δεκαδικά ψηφία, η τιμή αυτή αναγράφεται ως 1.64, ενώ, όταν γίνεται στρογγυλοποίηση στα τρία δεκαδικά ψηφία, ως 1.645. Με το ίδιο σκεπτικό προκύπτει ότι $z_{0.01} = 2.32$, $z_{0.005} = 2.57$.

Χρησιμοποιώντας την R και εκτελώντας την εντολή `qnorm(c(0.025,0.05,0.01,0.005),0,1, lower.tail=FALSE)`, λαμβάνουμε παρόμοια αποτελέσματα με προηγουμένως, αλλά με μεγαλύτερη ακρίβεια. Πιο συγκεκριμένα, λαμβάνουμε τις τιμές 1.959964, 1.644854, 2.326348 και 2.575829, οι οποίες αντιστοιχούν στα σημεία $z_{0.025}$, $z_{0.05}$. $z_{0.01}$ και $z_{0.005}$.

Άσκηση Αυτοαξιολόγησης 5.16

Να προσδιορίσετε την τιμή των a, b και c έτσι ώστε: $P(Z < a) = 0.31918$, $P(Z > b) = 0.77337$ και $P(c < Z < 0.27) = 0.09047$.

Παράδειγμα 5.9

Η ολική ετήσια βροχόπτωση μετρημένη σε εκατοστά σε μία περιοχή είναι γνωστό ότι περιγράφεται ικανοποιητικά από την κανονική κατανομή με μέση τιμή $\mu = 70$ εκατοστά και τυπική απόκλιση $\sigma = 12$ εκατοστά.

1. Υπολογίστε την πιθανότητα η ετήσια ολική βροχόπτωση να υπερβεί τα 82 εκατοστά.
2. Υπολογίστε την πιθανότητα την επόμενη χρονιά η ολική βροχόπτωση να είναι μεταξύ 58 και 94 εκατοστών.
3. Προσδιορίστε τα εκατοστά για τα οποία η πιθανότητα η ολική ετήσια βροχόπτωση να είναι μεγαλύτερη από αυτά, να ισούται με 30.153%.

Λύση Παραδείγματος 5.9

Έστω X η τ.μ. που περιγράφει την ολική ετήσια βροχόπτωση μετρημένη σε εκατοστά στη συγκεκριμένη περιοχή. Είναι τότε $X \sim N(\mu = 70, \sigma^2 = 12^2)$.

1. Ζητείται η $P(X > 82)$. Χρησιμοποιώντας τον τυπικό μετασχηματισμό, έχουμε ότι

$$P(X > 82) = P\left(\frac{X-70}{12} > \frac{82-70}{12}\right) = P(Z > 1),$$

όπου $Z \sim N(0,1)$. Είναι από τον Πίνακα Α΄.3 $P(Z > 1) = 1 - P(Z \leq 1) = 1 - 0.84134$. Άρα είναι $P(X > 82) = 0.15866$.

2. Ζητείται η $P(58 < X < 94)$. Χρησιμοποιώντας τον τυπικό μετασχηματισμό έχουμε ότι

$$P(58 < X < 94) = P\left(\frac{58-70}{12} < \frac{X-70}{12} < \frac{94-70}{12}\right),$$

ή, ισοδύναμα,

$$P(58 < X < 94) = P(-1 < Z < 2) = P(Z < 2) - P(Z < -1).$$

Είναι όμως $P(Z < -1) = P(Z > 1) = 1 - P(Z < 1) = 0.15866$ και, επομένως,

$$P(58 < X < 94) = 0.97725 - 0.15866 = 0.81859.$$

3. Θέλουμε να προσδιορίσουμε τα εκατοστά της ολικής ετήσιας βροχόπτωσης, έστω a για τα οποία ισχύει ότι $P(X > a) = 0.30153$. Για να μπορεί να χρησιμοποιηθεί ο πίνακας της $N(0,1)$ (έστω και αντίστροφα) θα πρέπει να μετατρέψουμε την κανονική κατανομή σε τυπική κανονική κατανομή με τον τυπικό μετασχηματισμό. Είναι

$$P(X > a) = P\left(\frac{X-70}{12} > \frac{a-70}{12}\right) = 0.30153.$$

Επομένως, πρέπει το σημείο $\frac{a-70}{12}$ να είναι θετικό, αφού διαφορετικά η πιθανότητα θα ήταν μεγαλύτερη από 0.5, και επιπλέον

$$P\left(Z < \frac{a-70}{12}\right) = 1 - 0.30153 = 0.69847,$$

όπου $Z = \frac{X-70}{12} \sim N(0,1)$. Από τον πίνακα της $N(0,1)$ προκύπτει ότι $\frac{a-70}{12} = 0.52$ ή, ισοδύναμα, $a = 12 \cdot 0.52 + 70 = 76.24$ εκατοστά.

Στην R οι παραπάνω ποσότητες μπορούν να υπολογιστούν με τη βοήθεια των εντολών:

- `pnorm(82,70,12,lower.tail=FALSE)`,
- `pnorm(94,70,12,lower.tail=TRUE)-pnorm(58,70,12,lower.tail=TRUE)` και
- `qnorm(0.30153,70,12,lower.tail=FALSE)`,

οι οποίες επιστρέφουν τα παραπάνω αποτελέσματα, αλλά με μεγαλύτερη ακρίβεια.

5.7 Άλλες συνήθεις συνεχείς κατανομές

Στην ενότητα αυτή θα παρουσιαστούν, χωρίς την παράθεση λεπτομερειών, κάποιες ακόμη συνεχείς κατανομές, πλην των κατανομών t ή Student, χι-τετράγωνο χ^2 και F, οι οποίες θα παρουσιαστούν σε ξεχωριστή ενότητα στο Κεφάλαιο 7.

5.7.1 Λογαριθμοκανονική κατανομή

Σε πολλές πρακτικές εφαρμογές έχει παρατηρηθεί ότι η τυχαία μεταβλητή που μελετάμε δεν μπορεί να περιγραφθεί ικανοποιητικά από την κανονική κατανομή, αλλά ένας κατάλληλος και απλός μετασχηματισμός αυτής μπορεί να μας οδηγεί σε μία νέα τ.μ. η οποία περιγράφεται ικανοποιητικά από την κανονική κατανομή. Ένας τέτοιος συνήθης μετασχηματισμός είναι ο μετασχηματισμός του λογαρίθμου. Για παράδειγμα, έχει διαπιστωθεί ότι ενώ η τ.μ. , έστω X, που περιγράφει την αντοχή ενός υλικού σε συγκεκριμένη καταπόνηση ή τη συγκέντρωση μιας ουσίας στον ορό αίματος δεν ακολουθεί κανονική κατανομή, η τ.μ. $Y = \log(X)$ ακολουθεί κανονική κατανομή. Λόγω της ιδιαίτερης χρησιμότητας που παρουσιάζουν τέτοιες περιπτώσεις παρουσιάστηκε στη βιβλιογραφία η λεγόμενη λογαριθμοκανονική κατανομή, ο ορισμός της οποίας παρατίθεται στη συνέχεια.

Ορισμός 5.7

Η τυχαία μεταβλητή X λέγεται ότι ακολουθεί τη **λογαριθμοκανονική κατανομή** με παραμέτρους $\mu \in R$ και $\sigma > 0$, αν οι δυνατές της τιμές x είναι $x \in (0, +\infty)$ και η συνάρτηση πυκνότητας πιθανότητάς της δίνεται από τη σχέση:

$$f_X(x) = \begin{cases} \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{(\log x - \mu)^2}{2\sigma^2}\right), & x \in (0, +\infty), \\ 0, & \text{αλλού.} \end{cases} \tag{5.44}$$

Στην περίπτωση αυτή, γράφουμε ότι $X \sim LN(\mu, \sigma^2)$.

Η λογαριθμοκανονική κατανομή βρίσκει πλήθος εφαρμογών, όπως αναφέρθηκε και πρωτύτερα, και έχει χρησιμοποιηθεί μεταξύ άλλων, για την περιγραφή της κατανομής του μεγέθους πετρωμάτων και κοιτασμάτων, της ποσότητας των βροχοπτώσεων, της συγκέντρωσης ρύπων, αλλά ακόμα και της διάρκειας ζωής ηλεκτρονικών εξαρτημάτων. Για λεπτομέρειες σχετικά με τις ιδιότητες και τις εφαρμογές αυτής της κατανομής παραπέμπουμε μεταξύ άλλων, στο σύγγραμμα των Johnson *et al.* (1994a).

Στο σημείο αυτό θα παρουσιαστεί η ιδιότητα που περιφραστικά αναφέρθηκε πριν τον ορισμό 5.7.1 και η οποία είναι ιδιαίτερα χρήσιμη στην απόδειξη θεωρητικών αποτελεσμάτων, όπως αυτά των ασκήσεων αυτοαξιολόγησης που ακολουθούν.

Πρόταση 5.17

Αν η τ.μ. $X \sim LN(\mu, \sigma^2)$, τότε η τυχαία μεταβλητή $Y = \log X$ ακολουθεί κανονική κατανομή με παραμέτρους μ και σ^2, ήτοι $Y = \log X \sim N(\mu, \sigma^2)$.

Απόδειξη Πρότασης 5.17

Με τη μέθοδο του μετασχηματισμού που παρουσιάστηκε στην Ενότητα 3.7 έχουμε ότι:

$$f_Y(y) = f_X(e^y)|e^y| = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right), y \in \mathbb{R}.$$

Άσκηση Αυτοαξιολόγησης 5.17

Έστω $X \sim LN(\mu, \sigma^2)$, να αποδείξετε ότι η ασκ της δίνεται από τη σχέση:

$$F_X(x) = \Phi\left(\frac{\log(x) - \mu}{\sigma}\right), x \in \mathbb{R}.$$

Άσκηση Αυτοαξιολόγησης 5.18

Έστω $X \sim LN(\mu, \sigma^2)$, να αποδείξετε ότι η ροπή k τάξης δίνεται από τη σχέση:

$$E(X^k) = \exp\left(k\mu + 0.5k^2\sigma^2\right).$$

Παρατήρηση 5.15

Έστω $X \sim LN(\mu, \sigma^2)$ με σππ που δίνεται από τη σχέση (5.44). Τότε με τη γλώσσα προγραμματισμού R μπορούμε:

- με τη συνάρτηση `dlnorm(x,μ,σ)` να υπολογίσουμε τη σππ στο σημείο ή στο διάνυσμα σημείων x,
- με τη συνάρτηση `plnorm(x,μ,σ,lower.tail=TRUE)` να υπολογίσουμε την ασκ στο σημείο ή στο διάνυσμα σημείων x,
- με τη συνάρτηση `plnorm(x,μ,σ,lower.tail=FALSE)` να υπολογίσουμε την πιθανότητα $P(X > x)$ στο σημείο ή στο διάνυσμα σημείων x,
- με τη συνάρτηση `qlnorm(q,μ,σ,lower.tail=TRUE)` να προσδιορίσουμε το σημείο ή το διάνυσμα σημείων x για τα οποία ισχύει ότι $P(X \leq x) = q$, όπου q είναι πιθανότητα ή διάνυσμα με συνιστώσες πιθανότητες, αντίστοιχα,
- με τη συνάρτηση `qlnorm(q,μ,σ,lower.tail=FALSE)` να προσδιορίσουμε το σημείο ή το διάνυσμα σημείων x για τα οποία ισχύει ότι $P(X > x) = q$, όπου q είναι πιθανότητα ή διάνυσμα με συνιστώσες πιθανότητες, αντίστοιχα,
- με τη συνάρτηση `rlnorm(n,μ,σ)` να δημιουργήσουμε ένα δείγμα μεγέθους n από αυτήν την κατανομή.

Άσκηση Αυτοαξιολόγησης 5.19

Έστω ότι η τ.μ. X ακολουθεί λογαριθμοκανονική κατανομή με παραμέτρους $\mu = 2.4$ και $\sigma = 1.4$.

1. Υπολογίστε την πιθανότητα $P(1.8 < X < 21.7)$.
2. Προσδιορίστε το 95ο εκατοστιαίο σημείο.

5.7.2 Κατανομή Weibull

Η κατανομή Weibull ονομάστηκε έτσι προς τιμήν του Σουηδού μαθηματικού Waloddi Weibull (1887-1979), ο οποίος την περιέγραψε λεπτομερώς το 1951 (Weibull, 1951), παρότι είχε πρωτοπαρουσιαστεί στη βιβλιογραφία από τον Fréchet (1927). Αποτελεί την πιο συχνά χρησιμοποιούμενη κατανομή στις μελέτες ανάλυσης επιβίωσης και αξιοπιστίας, αλλά βρίσκει πλήθος εφαρμογών σε διάφορα επιστημονικά πεδία. Παραπέμπουμε τον/την ενδιαφερόμενο/η αναγνώστη/στρια ενδεικτικά στους Johnson *et al.* (1994a).

Ορισμός 5.8

Η τυχαία μεταβλητή X λέγεται ότι ακολουθεί την **κατανομή Weibull** με παραμέτρους $a > 0$ και $b > 0$, αν οι δυνατές της τιμές x είναι $x \in (0, +\infty)$ και η συνάρτηση πυκνότητας πιθανότητάς της δίνεται από τη σχέση:

$$f_X(x) = \begin{cases} \frac{a}{b}\left(\frac{x}{b}\right)^{a-1} e^{-(x/b)^a}, & x \in (0, +\infty), \\ 0, & \text{αλλού} \end{cases} \tag{5.45}$$

Στην περίπτωση αυτή, γράφουμε ότι $X \sim \text{Weibull}(a, b)$.

Η παράμετρος a καλείται παράμετρος σχήματος ή μορφής (shape parameter), ενώ η παράμετρος b παράμετρος κλίμακας (scale parameter). Παρατηρήστε ότι για $a = 1$ προκύπτει η εκθετική κατανομή με μέση τιμή b. Η κατανομή Weibull μπορεί να είναι λοξή προς τα αριστερά, προς τα δεξιά ή συμμετρική (αν $\alpha = 3.6$, τότε είναι συμμετρική).

Άσκηση Αυτοαξιολόγησης 5.20

Έστω $X \sim \text{Weibull}(a, b)$, να αποδείξετε ότι η ασκ της δίνεται από τη σχέση:

$$F_X(x) = \begin{cases} 1 - e^{-(x/b)^a}, & x \in (0, +\infty), \\ 0, & \text{αλλού.} \end{cases} \tag{5.46}$$

Άσκηση Αυτοαξιολόγησης 5.21

Έστω $X \sim \text{Weibull}(a, b)$, να αποδείξετε ότι οι ροπές k-τάξης δίνονται από τη σχέση:

$$E(X^k) = b^k \Gamma\left(1 + \frac{k}{a}\right).$$

Παρατήρηση 5.16

Έστω $X \sim \text{Weibull}(a, b)$ με σππ που δίνεται από τη σχέση (5.45). Τότε με τη γλώσσα προγραμματισμού R μπορούμε:

- με τη συνάρτηση `dweibull(x,a,b)` να υπολογίσουμε τη σππ στο σημείο ή στο διάνυσμα σημείων x,
- με τη συνάρτηση `pweibull(x,a,b,lower.tail=TRUE)` να υπολογίσουμε την ασκ στο σημείο ή στο διάνυσμα σημείων x,
- με τη συνάρτηση `pweibull(x,a,b,lower.tail=FALSE)` να υπολογίσουμε την πιθανότητα $P(X > x)$ στο σημείο ή στο διάνυσμα σημείων x,
- με τη συνάρτηση `qweibull(q,a,b,lower.tail=TRUE)` να προσδιορίσουμε το σημείο ή το διάνυσμα σημείων x για τα οποία ισχύει ότι $P(X \leq x) = q$, όπου q είναι πιθανότητα ή διάνυσμα με συνιστώσες πιθανότητες, αντίστοιχα,
- με τη συνάρτηση `qweibull(q,a,b,lower.tail=FALSE)` να προσδιορίσουμε το σημείο ή το διάνυσμα σημείων x για τα οποία ισχύει ότι $P(X > x) = q$, όπου q είναι πιθανότητα ή διάνυσμα με συνιστώσες πιθανότητες, αντίστοιχα,
- με τη συνάρτηση `rweibull(n,a,b)` να δημιουργήσουμε ένα δείγμα μεγέθους n από αυτήν την κατανομή.

5.8 Ασκήσεις

Άσκηση 5.1 *Αν η τυχαία μεταβλητή X ακολουθεί κανονική κατανομή με μέση τιμή μ και διακύμανση σ^2, να υπολογίσετε την $E\left(\left|X-\mu\right|\right)$.*

Άσκηση 5.2 *Κατά τη διάρκεια της ημέρας, τα τρένα μιας συγκεκριμένης γραμμής σε έναν υπόγειο σιδηρόδρομο περνούν κάθε μισή ώρα.*

1. *Ποια είναι η πιθανότητα κάποιος που εισέρχεται στον σταθμό τυχαία στη διάρκεια της ημέρας να περιμένει περισσότερο από 20 λεπτά;*
2. *Πόσος είναι ο μέσος χρόνος αναμονής;*

Άσκηση 5.3 *Έστω ότι επιλέγεται τυχαία ένα σημείο σε μία ράβδο μήκους $1m$ και η ράβδος σπάζεται στο σημείο αυτό. Να βρεθούν*

1. *η κατανομή του μήκους του μεγαλύτερου κομματιού, και*
2. *η μέση τιμή και η διασπορά αυτής της κατανομής.*

Άσκηση 5.4 *Υποθέτουμε ότι η διάρκεια ζωής μιας ηλεκτρονικής συσκευής ακολουθεί την εκθετική κατανομή με μέση τιμή 1.5 χρόνια. Να βρεθούν οι πιθανότητες: μια συσκευή να διαρκέσει τουλάχιστον τρία χρόνια, μια συσκευή ηλικίας 5 ετών, που λειτουργεί ακόμη, να διαρκέσει επιπλέον 3 χρόνια.*

Άσκηση 5.5 *Υποθέστε ότι ο χρόνος ζωής ενός μηχανικού εξαρτήματος περιγράφεται ικανοποιητικά από μια εκθετική κατανομή. Αν είναι γνωστό ότι ο μέσος χρόνος ζωής του συγκεκριμένου μηχανικού εξαρτήματος είναι ίσος με 4 χρόνια:*

1. *υπολογίστε την πιθανότητα ότι ένα τυχαία επιλεγμένο τέτοιο μηχανικό εξάρτημα θα χρειαστεί αντικατάσταση σε λιγότερο από 3 χρόνια,*
2. *υπολογίστε την πιθανότητα ότι ένα τυχαία επιλεγμένο τέτοιο μηχανικό εξάρτημα θα έχει διάρκεια ζωής μεταξύ 3-5 ετών,*
3. *υπολογίστε την τιμή εκείνη που είναι τέτοια ώστε το 65% των μηχανικών εξαρτημάτων να έχει χρόνο ζωής μικρότερο από αυτήν την τιμή.*

Άσκηση 5.6 *Το τηλεφωνικό κέντρο της Πυροσβεστικής Υπηρεσίας μιας μεγάλης αστικής πόλης, δέχεται κατά μέσο όρο 4 κλήσεις ανά δεκάλεπτο. Υποθέστε ότι ο αριθμός των αφίξεων περιγράφεται από μια διαδικασία Poisson. Επιπρόσθετα, θεωρήστε ότι αγνοούμε τον χρόνο εξυπηρέτησης κάθε κλήσης και ότι οι χρόνοι μεταξύ κλήσεων είναι ανεξάρτητοι.*

1. *Πόσος χρόνος σε λεπτά μεσολαβεί κατά μέσο όρο μεταξύ δύο διαδοχικών κλήσεων;*
2. *Υπολογίστε την πιθανότητα ο χρόνος μεταξύ δύο διαδοχικών κλήσεων να είναι μικρότερος από 2 λεπτά.*
3. *Υπολογίστε την πιθανότητα το τηλεφωνικό κέντρο να δεχτεί λιγότερες από 5 κλήσεις σε ένα εικοσάλεπτο.*

Άσκηση 5.7 *Έστω ότι η διαστολική αρτηριακή πίεση ανδρών μιας συγκεκριμένης ηλικιακής ομάδας ακολουθεί κανονική κατανομή με παραμέτρους $\mu = 80$ mmHg και $\sigma = 10$ mmHg.*

1. *Υπολογίστε την πιθανότητα κάποιο άτομο του πληθυσμού να έχει διαστολική αρτηριακή πίεση μεγαλύτερη από 90 mmHg.*
2. *Τι ποσοστό του πληθυσμού έχει αρτηριακή πίεση μικρότερη από 70 mmHg;*

Άσκηση 5.8 *Έστω* $X \sim N(\mu, \sigma^2)$. *Να δειχθεί ότι*

1. $P(\mu - \sigma \leq X \leq \mu + \sigma) = 0.68268$.
2. $P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) = 0.9545$.
3. $P(\mu - 3\sigma \leq X \leq \mu + 3\sigma) = 0.9973$.

Υπόδειξη: Χρησιμοποιήστε τον τυπικό μετασχηματισμό και τον πίνακα της $N(0,1)$ *του Παραρτήματος Α΄.*

Άσκηση 5.9 *Γνωρίζουμε ότι η πιθανότητα αναμονής ενός πελάτη μιας τράπεζας για περισσότερο από 20 λεπτά είναι 0.0239. Αν ο χρόνος αναμονής ακολουθεί κανονική κατανομή με τυπική απόκλιση 3.75 λεπτά, να υπολογιστούν ο μέσος χρόνος αναμονής στην τράπεζα και η πιθανότητα ένας πελάτης να περιμένει στην τράπεζα από 10 μέχρι 15 λεπτά.*

Άσκηση 5.10 *Υποθέστε ότι η ποσότητα υδρογονανθράκων που εκπέμπουν αυτοκίνητα ορισμένου τύπου ακολουθεί την κανονική κατανομή με μέση τιμή* $1gr/km$ *και τυπική απόκλιση* $0.25gr/km$.

1. *Υπολογίστε την πιθανότητα κάποιο αυτοκίνητο αυτού του τύπου να εκπέμπει ποσότητα υδρογονανθράκων μεταξύ* $0.9gr/km$ *και* $1.54gr/km$.
2. *Αν σε κάποιο έλεγχο καυσαερίων κόβονται όλα τα αυτοκίνητα που εκπέμπουν υδρογονάνθρακες πάνω από* $1.4gr/km$, *τι ποσοστό αυτοκινήτων του συγκεκριμένου τύπου περνούν τον έλεγχο;*
3. *Βρείτε το* 15° *εκατοστιαίο σημείο των εκπομπών υδρογονανθράκων των αυτοκινήτων αυτού του τύπου.*

Άσκηση 5.11 *Η ημερήσια ποσότητα λιγνίτη που εξάγεται από μια εταιρεία ακολουθεί την κανονική κατανομή με μέση τιμή 70 τόνους και τυπική απόκλιση 6.5 τόνους. Σε μία δεδομένη μέρα υπολογίστε την πιθανότητα να εξαχθούν από το λιγνιτωρυχείο από 50 έως 80 τόνοι λιγνίτη.*

Άσκηση 5.12 *Ο όγκος των δεδομένων ανά λεπτό που εξυπηρετεί ένας παροχέας υπηρεσιών Internet ακολουθεί την κανονική κατανομή με μέση τιμή τα 2.5 TerraBytes και τυπική απόκλιση τα 0.2 TerraBytes. Να υπολογιστεί η πιθανότητα ένα τυχαία επιλεγμένο λεπτό ο παροχέας να εξυπηρετήσει λιγότερα από 3 TerraBytes, μεταξύ 2.7 και 3 TerraBytes, λιγότερα από 2 TerraBytes, μεταξύ 2.4 και 3 TerraBytes.*

Άσκηση 5.13 *Η διαφορά δυναμικού μεταξύ δύο σημείων ενός αγωγού, όταν διέρχεται από αυτόν σταθερό ρεύμα, περιγράφεται από μια κανονική κατανομή με μέση τιμή 12* $Volts$ *και τυπική απόκλιση 0.3* $Volts$.

1. *Υπολογίστε την πιθανότητα μια τυχαία μέτρηση της διαφοράς δυναμικού να βρεθεί μεταξύ των 11.6 και των 12.4* $Volts$.
2. *Υπολογίστε την τιμή της διαφοράς δυναμικού κάτω από την οποία βρίσκεται το 12% των μικρότερων τιμών που μπορεί να πάρει η διαφορά δυναμικού στον συγκεκριμένο αγωγό.*

Άσκηση 5.14 *Είναι γνωστό ότι το βάρος του πληθυσμού των κοριτσιών ηλικίας 12 ετών ακολουθεί κανονική κατανομή με μέση τιμή 35 κιλά και τυπική απόκλιση 4 κιλά. Να βρεθεί το ποσοστό των κοριτσιών βάρους 32-37 κιλά. Υπολογίστε την πιθανότητα από 12 κορίτσια που επιλέχτηκαν τυχαία, 2 ακριβώς να έχουν βάρος από 32-37 κιλά.*

Άσκηση 5.15 *Σύμφωνα με τη μετεωρολογική υπηρεσία μιας περιοχής το μέσο μηνιαίο ύψος βροχόπτωσης είναι 2.09 εκατοστά. Υποθέτοντας ότι το μηνιαίο ύψος βροχόπτωσης ακολουθεί κανονική κατανομή με τυπική απόκλιση 0.48 εκατοστά, να υπολογίσετε την πιθανότητα κάποιον μήνα το ύψος της βροχόπτωσης στη συγκεκριμένη περιοχή να είναι από 1.5 έως 2.5 εκατοστά. Υπολογίστε την πιθανότητα να περάσουν τρεις μήνες μέχρι για πρώτη φορά να υπάρξει μηνιαίο ύψος βροχόπτωσης μεγαλύτερο από 3 εκατοστά. Σημείωση: η βροχόπτωση κάθε μήνα θεωρείται ανεξάρτητη από τη βροχόπτωση σε οποιονδήποτε άλλο μήνα και κάθε μήνας περιγράφεται από την ίδια κατανομή.*

Άσκηση 5.16 *Ο χρόνος που απαιτείται από έναν φοιτητή για να συμπληρώσει το ερωτηματολόγιο αξιολόγησης του μαθήματος περιγράφεται από την κανονική κατανομή με μέση τιμή 20 λεπτά και τυπική απόκλιση 2 λεπτά. Σε 7 τυχαία επιλεγμένους φοιτητές υπολογίστε την πιθανότητα τουλάχιστον 3 φοιτητές να έχουν χρειαστεί περισσότερο από 22 λεπτά για να συμπληρώσουν το ερωτηματολόγιο.*

Άσκηση 5.17 *Ο αριθμός των αφίξεων πελατών στην ιστοσελίδα του Πανεπιστημίου Ιωαννίνων περιγράφεται από την κατανομή Poisson με ρυθμό αφίξεων 10 ανά λεπτό. Αν ο διακομιστής αντιμετωπίζει μια αστοχία διάρκειας 18 δευτερολέπτων κατά τη διάρκεια της οποίας οι επισκέπτες του ιστότοπου δεν θα έχουν πρόσβαση, υπολογίστε την πιθανότητα η διακοπή αυτή να μην είχε επίπτωση σε κάποιον χρήστη. Υπόδειξη: να δοθεί απάντηση με χρήση τόσο της κατανομής Poisson όσο και της εκθετικής κατανομής.*

Άσκηση 5.18 *Είναι γνωστό ότι το βάρος του πληθυσμού των αγοριών ηλικίας 12 ετών ακολουθεί κανονική κατανομή με μέση τιμή 40 κιλά και τυπική απόκλιση 2 κιλά. Ένα παιδί θεωρείται σε αυτήν την ηλικία παχύσαρκο αν έχει βάρος μεγαλύτερο από 44 κιλά. Υπολογίστε την πιθανότητα να βρεθεί το πρώτο παχύσαρκο παιδί μετά από 10 επιλογές.*

Άσκηση 5.19 *Ο χρόνος ζωής ενός ηχοσυστήματος ακολουθεί εκθετική κατανομή με μέσο χρόνος ζωής τα 8 έτη. Αν ο Αποστόλης αγόρασε ένα μεταχειρισμένο ηχοσύστημα, υπολογίστε την πιθανότητα αυτό να λειτουργεί μετά από 8 έτη από την αγορά του.*

Άσκηση 5.20 *Η διάμετρος ενός εξαρτήματος μιας μηχανής περιγράφεται από την κανονική κατανομή με μέση τιμή 1.1 χιλιοστά και τυπική απόκλιση 0.004 χιλιοστά. Τα εργοστασιακά όρια, για να μην είναι ελαττωματικό ένα εξάρτημα, είναι 1.08-1.14 χιλιοστά. Υπολογίστε την πιθανότητα ένα εξάρτημα να είναι ελαττωματικό. Επιλέγονται τυχαία 100 τέτοια εξαρτήματα. Πόσα εξαρτήματα αναμένεται να βρεθούν ελαττωματικά; Υπολογίστε την πιθανότητα να χρειαστεί να ελεγχθούν 18 εξαρτήματα μέχρι να βρεθεί το τρίτο ελαττωματικό.*

Άσκηση 5.21 *Ο αριθμός των πολιτιστικών εκδηλώσεων που διεξάγονται σε μια επαρχιακή πόλη ακολουθεί κατανομή Poisson. Αν ο ρυθμός των εκδηλώσεων είναι 1.5 εκδηλώσεις στους έξι μήνες, υπολογίστε την πιθανότητα να μην πραγματοποιηθεί κάποια πολιτιστική εκδήλωση σε διάστημα 3 μηνών. Δεδομένου ότι έχουν περάσει 4 μήνες χωρίς καμία εκδήλωση, υπολογίστε την πιθανότητα να περάσουν άλλοι 4 μήνες χωρίς καμία εκδήλωση.*

Άσκηση 5.22 *Το ποσό των χρημάτων που ξοδεύει μηνιαία κατά τη διαμονή του στα Ιωάννινα ένας φοιτητής του Πανεπιστημίου Ιωαννίνων περιγράφεται από την κανονική κατανομή με μέση τιμή 800 Ευρώ και τυπική απόκλιση 100 Ευρώ. Αν επιλέξουμε τυχαία τρεις φοιτητές, υπολογίστε την πιθανότητα τουλάχιστον δύο από αυτούς να ξοδεύουν περισσότερα από 900 Ευρώ. Υπολογίστε την πιθανότητα να χρειαστεί να επιλέξουμε τυχαία περισσότερους από δεκαπέντε φοιτητές μέχρι να επιλεχθεί ο τρίτος που ξοδεύει λιγότερα από 600 Ευρώ.*

Άσκηση 5.23 *Παίρνουμε τυχαία έναν αριθμό στο διάστημα (0,1). Υπολογίστε την πιθανότητα το πρώτο δεκαδικό ψηφίο του να είναι 0. Υπολογίστε την πιθανότητα το τετράγωνο του αριθμού που επιλέχτηκε να ανήκει στο διάστημα (0.3, 0.9).*

Άσκηση 5.24 *Η διάρκεια των τηλεφωνικών κλήσεων στη γραμματεία ενός διαγνωστικού κέντρου περιγράφεται από την εκθετική κατανομή με μέση τιμή 5 λεπτά. Υπολογίστε την πιθανότητα μια τηλεφωνική κλήση σε αυτήν τη γραμματεία να ξεπεράσει τα 4 λεπτά. Δεδομένου ότι μια κλήση έχει διαρκέσει 3 λεπτά, υπολογίστε την πιθανότητα να ολοκληρωθεί στα επόμενα 30 δευτερόλεπτα. Υπολογίστε την πιθανότητα σε τέσσερις κλήσεις μόνο μία να ξεπεράσει τα τέσσερα λεπτά.*

Άσκηση 5.25 *Ο δείκτης ευφυίας IQ ενός πληθυσμού ακολουθεί κανονική κατανομή με μέση τιμή 100 και τυπική απόκλιση 15.*

1. *Ποια είναι η πιθανότητα ένα τυχαία επιλεγμένο άτομο από τον πληθυσμό αυτό να έχει IQ λιγότερο από 90;*
2. *Αν κάποιος έχει IQ περισσότερο από 90, ποια είναι η πιθανότητα να έχει IQ περισσότερο από 110;*
3. *Αν θέλουμε να χωρίσουμε τον πληθυσμό σε 3 ομάδες με χαμηλό, μέσο και υψηλό IQ, έτσι ώστε το* 20% *του πληθυσμού να ανήκει στην πρώτη ομάδα, το* 65% *στη δεύτερη και το υπόλοιπο* 15% *στην τρίτη, ποιες τιμές IQ θα πρέπει να χρησιμοποιηθούν για να διαχωρίσουν την κάθε ομάδα;*
4. *Επιλέγονται 9 άτομα στην τύχη από τον πληθυσμό. Ποια είναι η πιθανότητα δύο από αυτά τα άτομα να έχουν IQ 2 μεγαλύτερο από 115;*
5. *Επιλέγονται 9 άτομα στην τύχη από τον πληθυσμό. Ποια είναι η πιθανότητα το IQ μόνο του δεύτερου και του ένατου να ξεπερνάει το 115;*
6. *Επιλέγονται διαδοχικά άτομα από τον πληθυσμό. Ποια είναι η πιθανότητα το ένατο άτομο να είναι το δεύτερο άτομο που το IQ του ξεπερνάει το 115;*

5.9 Απαντήσεις στις ασκήσεις αυτοαξιολόγησης

Λύση Άσκησης Αυτοαξιολόγησης 5.1

Οι εντολές που πρέπει να χρησιμοποιηθούν στην R, για να γίνει η γραφική παράσταση της $U(2,8)$, είναι οι ακόλουθες:

```
x = seq(0,10, length=100)
plot(x, punif(x, 2, 8), ylab="cdf", type ="l", col=4)
```

Λύση Άσκησης Αυτοαξιολόγησης 5.2

Έστω X η τ.μ. που παριστάνει τον χρόνο σε δευτερόλεπτα που χρειάζεται για να έρθει ο ανελκυστήρας μετά το πάτημα του κουμπιού. Από τα δεδομένα του προβλήματος προκύπτει ότι $X \sim U(0,40)$ με σππ και ασκ που δίνονται από τις σχέσεις (5.1) και (5.2), αντίστοιχα, με $a = 0$ και $b = 40$.

1. Ζητείται η πιθανότητα να περιμένει κάποιος περισσότερο από 12 δευτερόλεπτα, ήτοι η $P(X > 12)$. Είναι τότε:
$$P(X > 12) = 1 - F_X(12) = 1 - \frac{12-0}{40-0} = \frac{28}{40}.$$
2. Ζητείται η πιθανότητα να περιμένει κάποιος το πολύ 17 δευτερόλεπτα, ήτοι η $P(X < 17)$. Είναι τότε:
$$P(X < 17) = F_X(17) = \frac{17-0}{40-0} = \frac{17}{40}.$$
3. Ζητείται ο προσδιορισμός του μέσου χρόνου αναμονής, ήτοι της $E(X)$. Εφαρμόζοντας τη σχέση (5.5) προκύπτει ότι:
$$E(X) = \frac{0+40}{2} = 20.$$
4. Ζητείται ο προσδιορισμός σε δευτερόλεπτα της τιμής του χρόνου, η οποία είναι τέτοια ώστε $P(X < x) = 0.9$. Αυτό ισοδύναμα σημαίνει ότι $F_X(x) = 0.9$, από όπου προκύπτει ότι:
$$\frac{x-0}{40-0} = 0.9 \text{ ή } x = 36.$$

Σύμφωνα με την Παρατήρηση 5.2, τα αποτελέσματα στα πρώτα δύο ερωτήματα θα μπορούσαν να υπολογιστούν χρησιμοποιώντας τις παρακάτω εντολές τις R:

- `punif(12,0,40,lower.tail=FALSE)`,
- `punif(17,0,40,lower.tail=TRUE)` και
- `qunif(0.9,0,40,lower.tail=TRUE)`, αντίστοιχα.

Λύση Άσκησης Αυτοαξιολόγησης 5.3

Για να επιβεβαιώσουμε ότι η $f_X(x)$ της σχέσης (5.7) είναι όντως σππ αρκεί να δείξουμε ότι είναι μη αρνητική και το ολοκλήρωμα στο πεδίο ορισμού της είναι ίσο με τη μονάδα. Καθώς $x \in (0,1)$ και η συνάρτηση βήτα εξ ορισμού είναι θετική (βλ. Παράρτημα Β΄), εύκολα προκύπτει ότι όντως $f_X(x) \geq 0$.

Επιπρόσθετα, ισχύει ότι:

$$\begin{aligned}\int_{-\infty}^{+\infty} f_X(x)dx &= \int_0^1 \frac{x^{a-1}(1-x)^{b-1}}{B(a,b)}dx \\ &= \frac{1}{B(a,b)}\int_0^1 x^{a-1}(1-x)^{b-1}dx \\ &= 1\end{aligned}$$

όπου χρησιμοποιήθηκε η σχέση (Β΄.10) του Παραρτήματος Β΄.

Λύση Άσκησης Αυτοαξιολόγησης 5.4

Οι εντολές που πρέπει να χρησιμοποιηθούν στην R για να γίνει σχήμα ανάλογο με το Σχήμα 5.3 αλλά για την ασκ, είναι οι ακόλουθες:

```
x = seq(0,1, length=100)
plot(x, pbeta(x, 30, 30), ylab="cdf", type ="l", col=4)
lines(x, pbeta(x, 5, 5), type ="l", col=3)
lines(x, pbeta(x, 3, 3), col=2)
lines(x, pbeta(x, 1, 1), col=1)

legend(0.6,0.2, c("Be(30,30)","Be(5,5)","Be(3,3)","Be(1,1)"),lty=c(1,1,1,1)
    ,col=c(4,3,2,1))
```

Με προφανείς τροποποιήσεις προκύπτουν και οι γραφικές παραστάσεις για την περίπτωση των κατανομών που απεικονίζονται στο Σχήμα 5.4.
Παρατηρήστε ότι η διαφοροποίηση έγκειται στη χρησιμοποίηση της συνάρτησης `pbeta` αντί για τη συνάρτηση `dbeta`.

Λύση Άσκησης Αυτοαξιολόγησης 5.5

Έστω X η τ.μ. που παριστάνει το ποσοστό των ασθενών που πάσχει από την ασθένεια Α στη συγκεκριμένη περιοχή. Τότε, σύμφωνα με την εκφώνηση, έχουμε ότι $X \sim Be(a=2, b=4)$.

1. Στο ερώτημα αυτό ζητείται να προσδιοριστεί η $P(X < 0.2)$ και, χρησιμοποιώντας την εντολή `pbeta(0.2,2,4,lower.tail=TRUE)`, έχουμε $P(X < 0.2) = 0.26272$.
2. Ζητείται να υπολογιστεί η δεσμευμένη πιθανότητα $P(X > 0.25|X < 0.4)$. Είναι από τον ορισμό της δεσμευμένης πιθανότητας:

$$\begin{aligned}P(X > 0.25|X < 0.4) &= \frac{P(0.25 < X < 0.4)}{P(X < 0.4)} = \frac{F_X(0.4) - F_X(0.25)}{F_X(0.4)} \\ &= \frac{0.66304 - 0.3671875}{0.66304} = 0.4462061,\end{aligned}$$

όπου οι αριθμητικές τιμές των πιθανοτήτων $F_X(0.4)$ και $F_X(0.25)$ προέκυψαν με χρήση των εντολών

`pbeta(0.4,2,4,lower.tail=TRUE)` και

`pbeta(0.25,2,4,lower.tail=TRUE)`, αντίστοιχα.

Λύση Άσκησης Αυτοαξιολόγησης 5.6

Το αποτέλεσμα μπορεί να προκύψει ως ειδική περίπτωση της Πρότασης 5.2, λαμβάνοντας υπόψη ότι η ασκ της εκθετικής κατανομής δίνεται από τη σχέση (5.15). Εναλλακτικά, με τη μέθοδο μετασχηματισμού της ασκ (ανατρέξτε στην Ενότητα 3.7) είναι για y≥0,

$$\begin{aligned} F_Y(y) &= P\left(-\frac{1}{\lambda}\log(1-X) \le y\right) \\ &= P(\log(1-X) \ge -\lambda y) \\ &= P\left(X \le 1-\exp(-\lambda y)\right) \\ &= 1-\exp(-\lambda y), y \ge 0, \end{aligned}$$

ενώ για $y < 0$ είναι $F_Y(y) = 0$, καθώς η $-\frac{1}{\lambda}\log(1-X)$ είναι πάντοτε θετική.
Άρα προέκυψε η ασκ της εκθετικής κατανομής. Το αποτέλεσμα αυτό μπορεί να αξιοποιηθεί για τη δημιουργία μιας τυχαίας παρατήρησης από την εκθετική κατανομή με παράμετρο λ βασιζόμενοι σε μια τυχαία παρατήρηση από την ομοιόμορφη στο $(0,1)$.

Λύση Άσκησης Αυτοαξιολόγησης 5.7

Οι εντολές που πρέπει να χρησιμοποιηθούν στην R είναι οι ακόλουθες:

```
x = seq(0,10, length=100)
plot(x, pexp(x, 0.5), ylab="Cdf", type ="l", col=4)
lines(x, pexp(x, 1), type ="l", col=3)
lines(x, pexp(x, 1.5), col=2)
lines(x, pexp(x, 5), col=1)

legend(6,0.5, c("Exp(0.5)","Exp(1)","Exp(1.5)","Exp(5)"),lty=c(1,1,1,1),col
   =c(4,3,2,1))
```

Λύση Άσκησης Αυτοαξιολόγησης 5.8

1. Έστω X η τ.μ. που παριστάνει τον αριθμό των αυτοκινητικών ατυχημάτων που συμβαίνουν στη μικρή επαρχιακή πόλη σε μία εβδομάδα. Η τυχαία μεταβλητή X ακολουθεί, εξ υποθέσεως, κατανομή Poisson. Μας δίνεται ότι ο μέσος αριθμός αφίξεων πελατών είναι ίσος με 3 πελάτες ανά εβδομάδα, επομένως $E(X) = 3$ και, γνωρίζοντας ότι $E(X) = \lambda$, άμεσα προκύπτει ότι $\lambda = 3$. Ζητείται να προσδιοριστεί η $P(X \le 3) = F_X(3) = 0.6472319$ χρησιμοποιώντας τον Πίνακα της Poisson στο Παράρτημα Α΄.
2. Έστω Y η τ.μ. που παριστάνει σε εβδομάδες τον χρόνο που μεσολαβεί μεταξύ δύο διαδοχικών αυτοκινητικών ατυχημάτων. Τότε, καθώς ο αριθμός των αφίξεων ατυχημάτων ανά εβδομάδα ακολουθεί κατανομή Poisson με $\lambda = 3$, εξάγουμε το συμπέρασμα ότι ο χρόνος (σε εβδομάδες) που μεσολαβεί μεταξύ δύο διαδοχικών αυτοκινητικών ατυχημάτων ακολουθεί εκθετική κατανομή με παράμετρο $\lambda = 3$. Ζητείται να υπολογιστεί η πιθανότητα $P(Y > 2)$. Είναι

$$P(Y > 2) = 1 - F_Y(2) = 1 - (1 - e^{-6}) = e^{-6} = 0.002478752.$$

Με την R τα παραπάνω αποτελέσματα μπορούν να ληφθούν με την εκτέλεση των εντολών: `ppois(3, 3,lower.tail=TRUE)` και `pexp(2,3,lower.tail=FALSE)`, αντίστοιχα, οι οποίες επιστρέφουν ακριβώς τις παραπάνω τιμές.

Λύση Άσκησης Αυτοαξιολόγησης 5.9

Έστω X η τ.μ. που περιγράφει τον χρόνο με μονάδα μέτρησης την ώρα μεταξύ δύο διαδοχικών αφίξεων πελατών στο κατάστημα λιανικής. Καθώς οι αφίξεις πραγματοποιούνται σύμφωνα με την κατανομή Poisson με μέσο ρυθμό 30 πελάτες ανά ώρα, δηλαδή με $\lambda = 30$ πελάτες ανά ώρα, εξάγουμε το συμπέρασμα ότι $X \sim Exp(\lambda = 30)$.

1. Ζητείται να προσδιοριστεί ο μέσος χρόνος μεταξύ δύο διαδοχικών αφίξεων πελατών. Επομένως, αυτό που ζητείται είναι η εύρεση της $E(X)$. Από τη σχέση (5.21) έχουμε ότι $E(X) = \frac{1}{30}$.
2. Ζητείται να υπολογιστεί η πιθανότητα να χρειαστεί λιγότερο από 1 λεπτό μέχρι την άφιξη του επόμενου πελάτη, δηλαδή λιγότερο από το 1/60 της ώρας μεταξύ δύο διαδοχικών αφίξεων - η ώρα είναι η μονάδα μέτρησης στον ορισμό της τ.μ. X. Επομένως, το πρόβλημα ανάγεται στον υπολογισμό της πιθανότητας $P\left(X < \frac{1}{60}\right)$. Είναι τότε:

$$P\left(X < \frac{1}{60}\right) = F_X(1/60) = 1 - \exp\left(-\frac{30}{60}\right) = 0.3934693.$$

3. Ζητείται να υπολογιστεί η πιθανότητα να χρειαστούν περισσότερα από 4 λεπτά μέχρι την άφιξη του επόμενου πελάτη, δηλαδή περισσότερο από το 4/60=1/15 της ώρας μεταξύ δύο διαδοχικών αφίξεων - η ώρα είναι η μονάδα μέτρησης στον ορισμό της τ.μ. X. Επομένως, το πρόβλημα ανάγεται στον υπολογισμό της πιθανότητας $P\left(X > \frac{1}{15}\right)$. Είναι τότε:

$$P\left(X > \frac{1}{15}\right) = 1 - F_X(1/15) = \exp\left(-\frac{30}{15}\right) = 0.1353353.$$

4. Συμβολίζουμε με c τον ζητούμενο χρόνο με μονάδα μέτρησης την ώρα, που είναι τέτοιος ώστε $P(X < c) = 0.8$. Επομένως, από τον ορισμό της ασκ της εκθετικής, έχουμε ότι

   $$1 - \exp(-30c) = 0.8$$

   από την οποία προκύπτει ότι

   $$c = -\frac{\log(0.2)}{30} = 0.05364793.$$

   Επομένως, χρειάζεται 0.0536 της ώρας ή 3.218876 λεπτά.
5. Υιοθετώντας την κατανομή Poisson ως την κατανομή που μοντελοποιεί τις αφίξεις των πελατών στο κατάστημα λιανικής δεχόμαστε ότι ένας πελάτης φτάνει κάθε φορά και, επιπλέον, ότι ο ρυθμός αφίξεων παραμένει σταθερός κατά τη διάρκεια της ημέρας. Οι δύο αυτές υποθέσεις δεν είναι πάντοτε ρεαλιστικές.

Σύμφωνα με την Παρατήρηση 5.4, τα αποτελέσματα θα μπορούσαν να εξαχθούν εκτελώντας στην R τις εντολές: `pexp(1/60,30,lower.tail=TRUE)`, `pexp(1/15,30,lower.tail=FALSE)`, και `qexp(0.8,30,lower.tail=TRUE)`, αντίστοιχα.

Λύση Άσκησης Αυτοαξιολόγησης 5.10

Για να επιβεβαιώσουμε ότι η $f_X(x)$ της σχέσης (5.27) είναι όντως σππ αρκεί να δείξουμε ότι είναι μη αρνητική και το ολοκλήρωμά της στο πεδίο ορισμού της είναι ίσο με τη μονάδα. Καθώς $x \in [0,+\infty)$ και η γάμμα συνάρτηση εξ ορισμού είναι θετική (βλ. Παράρτημα Β΄), εύκολα προκύπτει ότι όντως $f_X(x) \geq 0$. Επιπρόσθετα, ισχύει ότι:

$$\begin{aligned}\int_{-\infty}^{+\infty} f_X(x)dx &= \int_0^{+\infty} \frac{\lambda^a x^{a-1} e^{-\lambda x}}{\Gamma(a)} dx \\ &= \frac{\lambda^a}{\Gamma(a)} \int_0^{+\infty} x^{a-1} e^{-\lambda x} dx \\ &= \frac{1}{\Gamma(a)} \int_0^{+\infty} w^{a-1} e^{-w} dw = 1,\end{aligned}$$

όπου το τελευταίο ολοκλήρωμα προέκυψε με αλλαγή μεταβλητών $\lambda x = w$ και, χρησιμοποιώντας τη σχέση (Β΄.6) του Παραρτήματος Β΄, ισούται με $\Gamma(a)$.

Λύση Άσκησης Αυτοαξιολόγησης 5.11

Οι εντολές που πρέπει να χρησιμοποιηθούν στην R είναι οι ακόλουθες:

```
x = seq(0,10, length=100)
plot(x, pgamma(x,1,0.5), ylab="Cdf", type ="l", col=4)
lines(x, pgamma(x,2,0.5), type ="l", col=3)
lines(x, pgamma(x,3,2), col=2)
lines(x, pgamma(x,3.5,2), col=1)

legend(6,0.3, c("G(1,0.5)","G(2,0.5)","G(3,2)","G(3.5,2)"),lty=c(1,1,1,1),
    col=c(4,3,2,1))
```

Λύση Άσκησης Αυτοαξιολόγησης 5.12

Η κατανομή Erlang αποτελεί ειδική περίπτωση της γάμμα κατανομής όταν $a \in \{1,2,3,...\}$. Τότε, χρησιμοποιώντας τις σχέσεις (Β΄.9) και (Β΄.17) του Παραρτήματος Β΄, προκύπτει ότι η σχέση (5.30) ανάγεται στην:

$$F_X(x) = \begin{cases} 1 - e^{-\lambda x} \sum_{k=0}^{a-1} \frac{(\lambda x)^k}{k!}, & x \geq 0, \\ 0, & \text{αλλού,} \end{cases}$$

η οποία ταυτίζεται με τη σχέση (5.24). Τέλος, η ασκ της εκθετικής κατανομής με παράμετρο λ προκύπτει άμεσα από την παραπάνω σχέση για $a = 1$.

Λύση Άσκησης Αυτοαξιολόγησης 5.13

Έστω X η τ.μ. που παριστάνει σε ώρες τον χρόνο που μεσολαβεί μεταξύ τριών διαδοχικών αφίξεων πελατών. Τότε, καθώς ο αριθμός των αφίξεων περιγράφεται από μια διαδικασία Poisson με ρυθμό $\lambda = 5$ πελάτες ανά ώρα, από τη θεωρία έχουμε ότι $X \sim G(a = 3, \lambda = 5)$. Στο πλαίσιο αυτό, μας ζητείται η πιθανότητα $P(X < 1/3)$, η οποία μπορεί να υπολογιστεί με τη βοήθεια της R εκτελώντας την εντολή `pgamma(1/3,shape=3,rate=5,lower.tail=TRUE)`, από την οποία προκύπτει ότι $P(X < 1/3) = 0.2340045$.

Λύση Άσκησης Αυτοαξιολόγησης 5.14

Χρησιμοποιώντας τον ορισμό της μέσης τιμής και μετά από λίγες πράξεις, έχουμε:

$$\begin{aligned} E(X) &= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} x \exp\left(-\frac{\left(x-\mu\right)^2}{2\sigma^2}\right) dx \\ &\overset{\frac{x-\mu}{\sqrt{2}\sigma}=t}{=} \frac{\sqrt{2}\sigma}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} \left(\sqrt{2}\sigma t + \mu\right) \exp\left(-t^2\right) dt \\ &= \frac{1}{\sqrt{\pi}} \left(\sqrt{2}\sigma \int_{-\infty}^{\infty} t \exp\left(-t^2\right) dt + \mu \int_{-\infty}^{\infty} \exp\left(-t^2\right) dt\right) \\ &= \frac{1}{\sqrt{\pi}} \left(0 + \mu\sqrt{\pi}\right) = \mu, \end{aligned}$$

όπου χρησιμοποιήθηκε ότι το ολοκλήρωμα περιττής συνάρτησης σε συμμετρικό διάστημα είναι ίσο με μηδέν. Για τον προσδιορισμό της διακύμανσης, καθώς $Var(X) = E(X^2) - (E(X))^2$, αρκεί να προσδιοριστεί η $E(X^2)$. Με παρόμοιο τρόπο όπως προηγουμένως έχουμε:

$$\begin{aligned} E(X^2) &= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} x^2 \exp\left(-\frac{\left(x-\mu\right)^2}{2\sigma^2}\right) dx \\ &\overset{\frac{x-\mu}{\sqrt{2}\sigma}=t}{=} \frac{\sqrt{2}\sigma}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} \left(\sqrt{2}\sigma t + \mu\right)^2 \exp\left(-t^2\right) dt \\ &= \frac{2\sigma^2}{\sqrt{\pi}} \int_{-\infty}^{\infty} t^2 \exp\left(-t^2\right) dt + \frac{2\sqrt{2}\sigma\mu}{\sqrt{\pi}} \int_{-\infty}^{\infty} t \exp\left(-t^2\right) dt + \frac{\mu^2}{\sqrt{\pi}} \int_{-\infty}^{\infty} \exp\left(-t^2\right) dt. \end{aligned}$$

Όμως το δεύτερο ολοκλήρωμα ως ολοκλήρωμα περιττής συνάρτησης σε συμμετρικό διάστημα ισούται με μηδέν, ενώ το τρίτο ολοκλήρωμα από τη σχέση (Β΄.13) είναι ίσο με $\sqrt{\pi}$. Επομένως, απομένει να υπολογιστεί το πρώτο ολοκλήρωμα. Είναι με τη μέθοδο της παραγοντικής ολοκλήρωσης,

$$\begin{aligned} \int_{-\infty}^{\infty} t^2 \exp\left(-t^2\right) dt &= \int_{-\infty}^{\infty} -0.5t \frac{d}{dt} \left\{\exp\left(-t^2\right)\right\} \\ &= \left[-\frac{t}{2} \exp\left(-t^2\right)\right]_{-\infty}^{\infty} + \frac{1}{2} \int_{-\infty}^{\infty} \exp\left(-t^2\right) dt \\ &= \frac{\sqrt{\pi}}{2}. \end{aligned}$$

Συνδυάζοντας τα παραπάνω, προκύπτει το ζητούμενο.

Λύση Άσκησης Αυτοαξιολόγησης 5.15

Χρησιμοποιώντας την ιδιότητα της συμμετρίας της σππ της τυπικής κανονικής κατανομής:

$$P(Z < -1.27) = P(Z > 1.27) = 1 - P(Z \leq 1.27) = 1 - 0.91466 = 0.08534.$$

Πάλι από την ιδιότητα της συμμετρίας της σππ της τυπικής κανονικής κατανομής:

$$P(Z > -1.06) = 1 - P(Z < 1.06) = 1 - 0.85543 = 0.14457.$$

Άμεσα από τον πίνακα έχουμε $P(Z < 2.15) = 0.98422$, ενώ, χρησιμοποιώντας τη σχέση $P(A) = 1 -$

$P(A')$, έχουμε:

$$P(Z > 3.15) = 1 - P(Z \leq 3.15) = 1 - 0.99918 = 0.00082.$$

Αξιοποιώντας τη συμμετρία της σππ γύρω από το μηδέν έχουμε ότι:

$$P(-3.78 < Z < -1.06) = P(1.06 < Z < 3.78) = 0.14449.$$

Η πιθανότητα $P(1.06 < Z < 3.78)$ έχει αναλυτικά υπολογιστεί στο Παράδειγμα 5.6. Τέλος,

$$\begin{aligned} P(-2.37 < Z < 1.26) &= P(Z < 1.26) - P(Z < -2.37) = P(Z < 1.26) - P(Z > 2.37) \\ &= P(Z < 1.26) - (1 - P(Z \leq 2.37) \\ &= 0.99324 + 0.89617 - 1 = 0.88941. \end{aligned}$$

Εναλλακτικά, κάποιος θα μπορούσε να παρατηρήσει ότι λόγω συμμετρίας η πιθανότητα αυτή είναι ίση με την $P(-1.26 < Z < 2.37)$, που έχει υπολογιστεί στο Παράδειγμα 5.6.

Λύση Άσκησης Αυτοαξιολόγησης 5.16

Καθώς μας δίνεται ότι $P(Z < a) = 0.31918$, δηλαδή πιθανότητα μικρότερη από 0.5, καταλαβαίνουμε αμέσως ότι ο αριθμός a είναι αρνητικός, δηλαδή $a < 0$. Λόγω συμμετρίας προκύπτει ότι $P(Z > -a) = 0.31918$ με $-a > 0$. Επομένως, είναι $1 - P(Z \leq -a) = 0.31918$ ή, ισοδύναμα, $P(Z \leq -a) = 1 - 0.31918 = 0.68082$. Από τον πίνακα της $N(0,1)$ προκύπτει ότι $-a = 0.47$ και, επομένως, $a = -0.47$.
Επίσης, επειδή $P(Z > b) = 0.77337$, καταλαβαίνουμε ότι ο αριθμός $b < 0$, γιατί διαφορετικά δεν θα μπορούσε να είναι η πιθανότητα μεγαλύτερη από 0.5. Λόγω συμμετρίας προκύπτει ότι $P(Z > b) = P(Z < -b) = 0.77337$, από όπου, χρησιμοποιώντας τον πίνακα της $N(0,1)$, έχουμε ότι $-b = 0.75$ και $b = -0.75$.
Για την τελευταία πιθανότητα έχουμε ότι $P(c < Z < 0.27) = P(Z < 0.27) - P(Z < c) = 0.60642 - P(Z < c)$ και, επομένως, $P(Z < c) = 0.60642 - 0.09047 = 0.51595$. Με χρήση του πίνακα της $N(0,1)$ έχουμε $c = 0.04$.

Λύση Άσκησης Αυτοαξιολόγησης 5.17

Χρησιμοποιώντας τη μονοτονία της συνάρτησης του λογαρίθμου έχουμε:

$$F_X(X) = P(X \leq x) = P(\log(X) \leq \log(x)).$$

Όμως αν $X \sim LN(\mu, \sigma^2)$, τότε $\log(X) \sim N(\mu, \sigma^2)$, οπότε:

$$P(\log(X) \leq \log(x)) = P\left(\frac{\log(X) - \mu}{\sigma} \leq \frac{\log(x) - \mu}{\sigma}\right).$$

Επομένως, καθώς $Z = \frac{\log(X) - \mu}{\sigma} \sim N(0,1)$, προκύπτει ότι:

$$F_X(x) = \Phi\left(\frac{\log(x) - \mu}{\sigma}\right), x \in \mathbb{R}.$$

Λύση Άσκησης Αυτοαξιολόγησης 5.18

Χρησιμοποιώντας τις ιδιότητες της εκθετικής συνάρτησης και του λογαρίθμου έχουμε ότι: $E(X^k) = E(\exp(\log(X))^k)$. Όμως όταν $X \sim LN(\mu,\sigma^2)$, τότε $Y = \log(X) \sim N(\mu,\sigma^2)$ και, επομένως,

$$E(X^k) = E[(\exp(Y))^k] = M_Y(k),$$

όπου M_Y η ροπογεννήτρια της κανονικής κατανομής $N(\mu,\sigma^2)$. Από τη σχέση (5.37) προκύπτει ότι

$$E(X^k) = M_Y(k) = \exp(k\mu + 0.5k\sigma^2).$$

Λύση Άσκησης Αυτοαξιολόγησης 5.19

Είναι $X \sim LN(2.4,1.4^2)$ και, επομένως, $Y = \log(X) \sim N(2.4,1.4^2)$.

1. Ζητείται η $P(1.8 < X < 21.7)$, η οποία ανάγεται στην $P(\log(1.8) < Y < \log(21.7))$, την οποία θα υπολογίσουμε ως εξής:

$$\begin{aligned} P(\log(1.8) < Y < \log(21.7)) &= P\left(\frac{\log(1.8) - 2.4}{1.4} < \frac{Y - 2.4}{1.4} < \frac{\log(21.7) - 2.4}{1.4}\right) \\ &= P(-1.29 < Z < 0.48) \\ &= P(Z < 0.48) - P(Z < -1.29) \\ &= P(Z < 0.48) - P(Z > 1.29) \\ &= P(Z < 0.48) + P(Z < 1.29) - 1 \\ &= 0.68439 + 0.90147 - 1 = 0.58586, \end{aligned}$$

   όπου χρησιμοποιήσαμε ότι $Z = \frac{Y-2.4}{1.4} \sim N(0,1)$.

2. Θέλουμε να προσδιορίσουμε το σημείο x που είναι τέτοιο, ώστε $P(X < x) = 0.95$. Με παρόμοιο σκεπτικό προκύπτει ότι η εύρεση ανάγεται στην εύρεση του x που πληροί την ακόλουθη σχέση:

$$P\left(\frac{\log(X) - 2.4}{1.4} \leq \frac{\log(x) - 2.4}{1.4}\right) = 0.95.$$

   Είναι τότε $\frac{\log(x)-2.4}{1.4} = 1.64$ ή $\log(x) = 4.696$, άρα $x = 109.5083$.

   Σημειώνεται ότι, εκτελώντας την εντολή `qlnorm(0.95,2.4,1.4)`, μπορούμε να προσδιορίσουμε με τη βοήθεια της R το σημείο x που είναι τέτοιο, ώστε $P(X < x) = 0.95$. Η εντολή αυτή επιστρέφει την τιμή 110.2549, η οποία φαινομενικά διαφέρει αρκετά από την τιμή που βρήκαμε νωρίτερα. Ο λόγος είναι ότι η R χρησιμοποιεί την τιμή 1.644854 για το 0.95 ποσοστιαίο σημείο της τυπικής κανονικής κατανομής και όχι τη στρογγυλοποιημένη τιμή, 1.64, όπως εμείς με βάση τους πίνακες.

Λύση Άσκησης Αυτοαξιολόγησης 5.20

Χρησιμοποιώντας τον ορισμό της ασκ προφανώς για $x < 0$ είναι $F_X(x) = P(X \leq x) = 0$, ενώ για $x > 0$ έχουμε ότι:

$$F_X(x) = P(X \leq x) = \int_0^x \frac{a}{b}\left(\frac{t}{b}\right)^{a-1} e^{-(t/b)^a} dt.$$

Το επιθυμητό αποτέλεσμα προκύπτει άμεσα παρατηρώντας ότι:

$$\frac{a}{b}\left(\frac{t}{b}\right)^{a-1} e^{-(t/b)^a} = -\frac{d}{dt}e^{-(t/b)^a}$$

Λύση Άσκησης Αυτοαξιολόγησης 5.21

Εξ ορισμού είναι

$$E(X^k) = \int_0^{+\infty} x^k \frac{a}{b}\left(\frac{x}{b}\right)^{a-1} e^{-(x/b)^a} dx.$$

Το ολοκλήρωμα αυτό μπορεί να υπολογιστεί κάνοντας την αλλαγή μεταβλητής: $(x/b)^a = y$, οπότε προκύπτει:

$$E(X^k) = b^k \int_0^{+\infty} y^{k/a} e^{-y} dy,$$

και το επιθυμητό αποτέλεσμα προκύπτει χρησιμοποιώντας τη σχέση (Β΄.6) του Παραρτήματος Β΄.

ΒΙΒΛΙΟΓΡΑΦΙΑ

Ελληνόγλωσση

Ζωγράφος, Κ. (2008). *Πιθανότητες*. Ιωάννινα: Πανεπιστήμιο Ιωαννίνων.

Ξενόγλωσση

Aksoy, H. (2000). Use of gamma distribution in hydrological analysis. *Turkish Journal of Engineering and Environmental Sciences*, *24*, pp. 419–428.

Bayes, T. (1958). An essay towards solving a problem in the doctrine of chances. *Reprinted in Biometrika*, *45*, pp. 296–315.

Boland, P. (2007). *Statistical and Probabilistic Methods in Actuarial Science*. Chapman and Hall/CRC.

Chia, E. and Hutchinson, M. (1991). The beta distribution as a probability model for daily cloud duration. *Agricultural and Forest Meteorology*, *56*, pp. 195–208.

Fréchet, M. (1927). Sur la loi de probabilité de l'écart maximum. *Annales de la Société Polonaise de Mathématique*, *6*, pp. 93–116.

Friedman, N., Cai, L. and Xie, X. (2006). Linking stochastic dynamics to population distribution: An analytical framework of gene expression. *Physical Review Letters*, *97*, p. 168302.

Gupta, A. K. (2011). *Beta Distribution*. In: *International Encyclopedia of Statistical Science, Lovric, Miodrag* (Ed.).

Johnson, N., Kotz, S. and Balakrishnan, N. (1994a). *Continuous Univariate Distributions, Vol. 1* (2nd ed.). New York: Wiley and Sons, Inc.

Johnson, N., Kotz, S. and Balakrishnan, N. (1994b). *Continuous Univariate Distributions, Vol. 2* (2nd ed.). New York: Wiley and Sons, Inc.

Kuczma, M. (2009). *An Introduction to the Theory of Functional Equations and Inequalities: Cauchy's Equation and Jensen's Inequality* (2nd ed.). Basel: Birkhauser.

Laplace, P. (1836). *Théorie analytique des probabilités, Supplement to 3rd edition*.

Nadarajah, S. and Kotz, S. (2007). Multitude of beta distributions with applications. *Statistics*, *41*, pp. 153–179.

Pishro-Nik, H. (2014). *Introduction to Probability, Statistics, and Random Processes*. Kappa Research, LLC.

Sheynin, O. B. (1971). Studies in the History of Probability and Statistics. XXV. On the history of some statistical laws of distribution. *Biometrika*, *58*, pp. 234–236.

Sulaiman, M., Hlaing Oo, W., Abd Wahab, M. and Zakaria, A. (1999). Application of beta distribution model to Malaysian sunshine data. *Renewable Energy*, *18*(4), pp. 573–579.

Tataru, P., Bataillon, T. and Hobolth, A. (2015). Inference Under a Wright-Fisher Model Using an Accurate Beta Approximation. *Genetics*, *201*, pp. 1133–1141.

Weibull, W. (1951). A Statistical Distribution Function Of Wide Applicability. *Journal of Applied Mechanics*, *18*, pp. 293–297.

ΚΕΦΑΛΑΙΟ 6

ΠΟΛΥΔΙΑΣΤΑΤΕΣ ΤΥΧΑΙΕΣ ΜΕΤΑΒΛΗΤΕΣ-ΣΤΟΧΑΣΤΙΚΗ ΑΝΕΞΑΡΤΗΣΙΑ

Σύνοψη

Σε αυτό το κεφάλαιο παρουσιάζονται εν συντομία οι πολυδιάστατες τυχαίες μεταβλητές και οι βασικότερες ιδιότητές τους. Έπειτα παρουσιάζονται οι έννοιες της δεσμευμένης τυχαίας μεταβλητής και της στοχαστικής ανεξαρτησίας. Το κεφάλαιο ολοκληρώνεται με την παρουσίαση δύο ειδικών περιπτώσεων πολυδιάστατων κατανομών, της πολυωνυμικής κατανομής και της διδιάστατης κανονικής κατανομής.

Προαπαιτούμενη γνώση: Κεφάλαια 1, 2 και 3 του παρόντος συγγράμματος.

Προσδοκώμενα μαθησιακά αποτελέσματα: Όταν θα έχετε ολοκληρώσει τη μελέτη του κεφαλαίου αυτού, θα γνωρίζετε

- την έννοια της πολυδιάστατης τυχαίας μεταβλητής ή τυχαίου διανύσματος,
- την έννοια της από κοινού αθροιστικής συνάρτησης κατανομής,
- την έννοια της από κοινού συνάρτησης πυκνότητας πιθανότητας,
- την έννοια της από κοινού συνάρτησης πιθανότητας,
- την έννοια της περιθώριας κατανομής,
- την αναμενόμενη τιμή συναρτήσεων πολυδιάστατων τυχαίων διανυσμάτων,
- την έννοια της υπό συνθήκη ή δεσμευμένης κατανομής,
- την έννοια της ανεξαρτησίας τυχαίων μεταβλητών,
- να προσδιορίζετε την κατανομή συναρτήσεων τυχαίων διανυσμάτων και
- κάποιες ειδικές πολυδιάστατες κατανομές.

Οικονόμου, Π., Μαλεφάκη, Σ., & Μπατσίδης, Α. (2023). *Πιθανότητες - Στατιστική.* [Προπτυχιακό εγχειρίδιο].

Γλωσσάριο επιστημονικών όρων

- Ανεξάρτητες τυχαίες μεταβλητές
- Από κοινού συνάρτηση κατανομής
- Από κοινού συνάρτηση πιθανότητας
- Από κοινού συνάρτηση πυκνότητας πιθανότητας
- Δεσμευμένη διασπορά
- Δεσμευμένη κατανομή (ή υπό συνθήκη κατανομή)
- Δεσμευμένη μέση τιμή
- Διδιάστατη κανονική κατανομή
- Περιθώρια κατανομή
- Πολυδιάστατη κανονική κατανομή
- Πολυωνυμική κατανομή
- Συνδιασπορά – Συνδιακύμανση
- Συνέλιξη
- Συντελεστής συσχέτισης

6.1 Εισαγωγή

Στα προηγούμενα κεφάλαια ασχοληθήκαμε με τυχαία φαινόμενα που διέπονται από τη συμπεριφορά μιας μεμονωμένης τυχαίας μεταβλητής και για τα οποία αρκεί η παρατήρηση αυτής και μόνο αυτής της τυχαίας μεταβλητής για την περιγραφή τους. Ωστόσο, σε πολλές περιπτώσεις, διάφορα υπό μελέτη φαινόμενα καθορίζονται και επηρεάζονται από περισσότερες από μία τυχαίες μεταβλητές. Στη συνέχεια, ενδεικτικά, αναφέρονται κάποια παραδείγματα τέτοιων τυχαίων φαινομένων ή πειραμάτων:

- σε μια αγροτική περιοχή κατά τη διάρκεια μιας χρονικής περιόδου μας ενδιαφέρει να μελετήσουμε τόσο το ύψος της βροχόπτωσης (σε εκατοστά) όσο και την κατανάλωση νερού για άρδευση (σε κυβικά μέτρα),
- σε μια κλινική μελέτη θα μπορούσε κάποιος/κάποια να μελετήσει τον δείκτη βάρους σώματος, τον εβδομαδιαίο χρόνο άσκησης και τη χοληστερόλη του ατόμου,
- σε μια έρευνα θα μπορούσε κάποιος/κάποια να μελετήσει τον αριθμό των παιδιών μιας οικογένειας, την οικονομική της κατάσταση και το μορφωτικό επίπεδο του πατέρα.

Στα παραπάνω παραδείγματα εμφανίζονται περισσότερες από μία τυχαίες μεταβλητές και έχει ιδιαίτερο ενδιαφέρον να μελετηθούν η συμπεριφορά της καθεμίας ξεχωριστά, η από κοινού συμπεριφορά τους, αλλά και η συμπεριφορά της μίας σε σχέση με τη συμπεριφορά των υπολοίπων ή κάποιων εκ των υπολοίπων. Στο πλαίσιο αυτό, στο παρόν κεφάλαιο γενικεύεται η έννοια της μονοδιάστατης τυχαίας μεταβλητής και κατανομής στον χώρο των k διαστάσεων ($k \geq 2$). Οι έννοιες της αθροιστικής συνάρτησης κατανομής, της συνάρτησης (πυκνότητας) πιθανότητας, της μέσης τιμής, της διακύμανσης και της ροπογεννήτριας, που εισήχθησαν στο Κεφάλαιο 3, γενικεύονται με ανάλογο τρόπο στην πολυδιάστατη περίπτωση, ενώ παρουσιάζονται νέες έννοιες, όπως είναι η δεσμευμένη κατανομή, η περιθώρια κατανομή, η στοχαστική ανεξαρτησία, η συνδιακύμανση και άλλες. Τέλος, παρουσιάζονται εν συντομία δύο ειδικές περιπτώσεις πολυδιάστατων κατανομών, η πολυωνυμική κατανομή και η διδιάστατη κανονική κατανομή.

6.2 Τυχαίο διάνυσμα και από κοινού αθροιστική συνάρτηση κατανομής

Στο Κεφάλαιο 3 η τυχαία μεταβλητή ορίστηκε να είναι μια μονοσήμαντη συνάρτηση με πεδίο ορισμού έναν δειγματικό χώρο και τιμές ένα υποσύνολο των πραγματικών αριθμών. Ο ορισμός της πολυδιάστατης τυχαίας μεταβλητής ή του τυχαίου διανύσματος αποτελεί γενίκευση αυτού του ορισμού και παρατίθεται στη συνέχεια.

Ορισμός 6.1

Μια **k-διάστατη τυχαία μεταβλητή** ή ένα **k-διάστατο τυχαίο διάνυσμα**, έστω $X = (X_1, \dots, X_k)^t$, ορίζεται να είναι μια μονοσήμαντη συνάρτηση με πεδίο ορισμού έναν δειγματικό χώρο Ω και τιμές ένα υποσύνολο, έστω S_X, του $\mathbb{R}^k$, δηλαδή $X : \Omega \to S_X \subseteq \mathbb{R}^k$.

Από τον παραπάνω ορισμό γίνεται άμεσα αντιληπτό ότι ένα k-διάστατο τυχαίο διάνυσμα είναι ένα k-διάστατο διάνυσμα στήλη με καθεμία από τις k το πλήθος συνιστώσες του να αποτελεί μια τυχαία μεταβλητή. Αυτές οι k το πλήθος τυχαίες μεταβλητές μπορεί να είναι είτε όλες διακριτές είτε όλες συνεχείς είτε κάποιες συνεχείς και κάποιες διακριτές. Στην πρώτη (δεύτερη) περίπτωση λέμε ότι έχουμε ένα διακριτό (συνεχές, αντίστοιχα) τυχαίο διάνυσμα, ενώ στην τρίτη περίπτωση λέμε ότι έχουμε ένα μεικτό τυχαίο διάνυσμα. Στο παρόν σύγγραμμα, θα ασχοληθούμε μόνο με τη μελέτη τυχαίων διανυσμάτων που ανήκουν στις δύο πρώτες περιπτώσεις.

Ο ορισμός του τυχαίου διανύσματος, κατά πλήρη αντιστοιχία με τον ορισμό της τυχαίας μεταβλητής, δεν εμπλέκει καθόλου την έννοια της πιθανότητας. Ο απώτερος όμως στόχος του ορισμού του τυχαίου διανύσματος είναι ο υπολογισμός πιθανοτήτων για την από κοινού συμπεριφορά των τυχαίων μεταβλητών που αποτελούν το τυχαίο διάνυσμα. Όπως αναφέρθηκε στην Ενότητα 3.3, η αθροιστική συνάρτηση κατανομής μας βοηθά στον υπολογισμό αυτών των πιθανοτήτων. Ο ορισμός της γενικεύεται στην περίπτωση k-διάστατου τυχαίου διανύσματος ως ακολούθως.

Ορισμός 6.2

Έστω $X : \Omega \to S_X \subseteq \mathbb{R}^k$ ένα k-διάστατο τυχαίο διάνυσμα με $X = (X_1, \dots, X_k)^t$. Η **συνάρτηση κατανομής** ή **η αθροιστική συνάρτηση κατανομής** (ασκ) του τυχαίου διανύσματος X συμβολίζεται με $F_X(\cdot)$ και είναι $F_X : \mathbb{R}^k \to [0,1]$ μια πραγματική συνάρτηση που ορίζεται από τη σχέση

$$F_X(x) = F_{X_1,X_2,\dots,X_k}(x_1, x_2, \dots, x_k) = P(X_1 \le x_1, X_2 \le x_2, \dots, X_k \le x_k), \tag{6.1}$$

για $x_i \in \mathbb{R}$, $i = 1,2,\dots,k$.

Παρατηρήστε ότι η αθροιστική συνάρτηση κατανομής ορίζεται για κάθε πραγματικό αριθμό ακόμη κι αν αυτός δεν ανήκει στο πεδίο τιμών του τυχαίου διανύσματος. Επιπλέον, η $F_{X_1,X_2,\dots,X_k}(x_1, x_2, \dots, x_k)$ εκφράζει την πιθανότητα οι τυχαίες μεταβλητές $X_1, X_2, \dots, X_k$ να πάρουν τιμές μικρότερες ή ίσες από τις τιμές $x_1, x_2, \dots, x_k$, αντίστοιχα.

Οι ιδιότητες της αθροιστικής συνάρτησης κατανομής τυχαίου διανύσματος αποτελούν ουσιαστικά γενίκευση των ιδιοτήτων της αθροιστικής συνάρτησης κατανομής τυχαίας μεταβλητής. Στην πρόταση που ακολουθεί διατυπώνονται κάποιες εξ αυτών.

Πρόταση 6.1

Έστω $X : \Omega \to S_X \subseteq \mathbb{R}^k$ ένα k-διάστατο τυχαίο διάνυσμα με $X = (X_1, \dots, X_k)^t$ και αθροιστική συνάρτηση κατανομής $F_X(\cdot)$. Τότε:

1. $0 \le F_X(x) \le 1$ για κάθε $x \in \mathbb{R}^k$.
2. Η $F_X(\cdot)$ είναι μη φθίνουσα συνάρτηση ως προς καθεμία συνιστώσα της.
3. Η $F_X(\cdot)$ είναι δεξιά συνεχής συνάρτηση ως προς καθεμία συνιστώσα της, δηλαδή

$$\lim_{x_i \to x_0+} F_X(x) = F_X(x_1, \dots, x_{i-1}, x_0, x_{i+1}, \dots, x_k).$$

4. $F(-\infty, -\infty, \dots, -\infty) = \lim_{\substack{x_1 \to -\infty \\ \dots \\ x_k \to -\infty}} F_{X_1,X_2,\dots,X_k}(x_1, x_2, \dots, x_k) = 0.$
5. $F(x_1, \dots, -\infty, \dots, x_k) = \lim_{x_j \to -\infty} F_{X_1,X_2,\dots,X_k}(x_1, \dots, x_j, \dots, x_k) = 0,\ j = 1, \dots, k.$
6. $F(+\infty, +\infty, \dots, +\infty) = \lim_{\substack{x_1 \to +\infty \\ \dots \\ x_k \to +\infty}} F_{X_1,X_2,\dots,X_k}(x_1, x_2, \dots, x_k) = 1.$

Απόδειξη Πρότασης 6.1

Η απόδειξη των ιδιοτήτων της από κοινού αθροιστικής συνάρτησης κατανομής ενός τυχαίου διανύσματος αφήνεται ως άσκηση για τον/την αναγνώστη/στρια.

Οι παραπάνω ιδιότητες αποτελούν ουσιαστικά τις συνθήκες τις οποίες μια πραγματική συνάρτηση πρέπει να ικανοποιεί ώστε να είναι αθροιστική συνάρτηση κατανομής.

Από την αθροιστική συνάρτηση κατανομής του τυχαίου διανύσματος $X = (X_1, \dots, X_k)^t$, που είναι γνωστή και ως από κοινού συνάρτηση κατανομής των τυχαίων μεταβλητών $X_1, \dots, X_k$, μπορεί να προκύψει η (από κοινού) αθροιστική συνάρτηση κατανομής οποιουδήποτε υποσυνόλου αυτών των τυχαίων μεταβλητών ή η ασκ μιας μεμονωμένης τυχαίας μεταβλητής. Οι κατανομές αυτές ονομάζονται **περιθώριες κατανομές**. Παραδείγματος χάριν, η αθροιστική συνάρτηση κατανομής της τυχαίας μεταβλητής X_1 προκύπτει από την από κοινού αθροιστική συνάρτηση κατανομής των $X_1, X_2, \dots, X_k$ ως εξής:

$$\begin{aligned} F_{X_1}(x_1) &= P(X_1 \le x_1) = P(X_1 \le x_1, X_2 \le +\infty, \dots, X_k \le +\infty) \\ &= P\left(\lim_{\substack{x_2 \to +\infty \\ \dots \\ x_k \to +\infty}} \{X_1 \le x_1, X_2 \le x_2, \dots, X_k \le x_k\} \right) \\ &= \lim_{\substack{x_2 \to +\infty \\ \dots \\ x_k \to +\infty}} P(X_1 \le x_1, X_2 \le x_2, \dots, X_k \le x_k) \\ &= \lim_{\substack{x_2 \to +\infty \\ \dots \\ x_k \to +\infty}} F_{X_1, X_2, \dots, X_k}(x_1, x_2, \dots, x_k) \\ &= F_{X_1, X_2, \dots, X_k}(x_1, +\infty, \dots, +\infty). \end{aligned}$$

Με παρόμοιο τρόπο η από κοινού αθροιστική συνάρτηση κατανομής των X_1 και X_3 είναι:

$$\begin{aligned} F_{X_1, X_3}(x_1, x_3) &= P(X_1 \le x_1, X_3 \le x_3) = P(X_1 \le x_1, X_2 \le +\infty, X_3 \le x_3, X_4 \le +\infty \dots, X_k \le +\infty) \\ &= \lim_{\substack{x_2 \to +\infty \\ x_4 \to +\infty \\ \dots \\ x_k \to +\infty}} F_{X_1, X_2, \dots, X_k}(x_1, x_2, \dots, x_k) \\ &= F_{X_1, X_2, \dots, X_k}(x_1, +\infty, x_3, +\infty \dots, +\infty). \end{aligned}$$

Τέλος, είναι σημαντικό να αναφερθεί ότι με βάση την από κοινού κατανομή μπορεί να υπολογιστεί η πιθανότητα διάφορων ενδεχομένων πέρα από το ενδεχόμενο $\{X_1 \le x_1, X_2 \le x_2, \dots, X_k \le x_k\}$ που εμφανίζεται στη σχέση ορισμού της.

Παράδειγμα 6.1

Έστω $F_{X,Y,Z}(x, y, z)$ η από κοινού αθροιστική συνάρτηση κατανομής των τυχαίων μεταβλητών X, Y και Z. Προσδιορίστε με τη βοήθεια της $F_{X,Y,Z}(\cdot, \cdot, \cdot)$ την πιθανότητα του ενδεχομένου $\{X > x \cup Z > z\}$.

Λύση Παραδείγματος 6.1

Για τον υπολογισμό της ζητούμενης πιθανότητας μπορούμε να εργαστούμε ως εξής:

$$\begin{aligned} P(\{X > x \cup Z > z\}) &= 1 - P(\{X > x \cup Z > z\}') = 1 - P(\{X > x\}' \cap \{Z > z\}') \\ &= 1 - P(\{X \le x\} \cap \{Z \le z\}) = 1 - P(X \le x, Z \le z) \\ &= 1 - F_{X,Z}(x, z) = 1 - F_{X,Y,Z}(x, +\infty, z) = 1 - \lim_{y \to +\infty} F_{X,Y,Z}(x, y, z), \end{aligned}$$

όπου με $F_{X,Z}(\cdot, \cdot)$ συμβολίσαμε την από κοινού περιθώρια αθροιστική συνάρτηση κατανομής των X και Z.

Άσκηση Αυτοαξιολόγησης 6.1

Έστω $F_{XY}(x, y)$ η από κοινού αθροιστική συνάρτηση κατανομής των τυχαίων μεταβλητών X και Y. Προσδιορίστε με τη βοήθεια της $F_{XY}(\cdot, \cdot)$ την πιθανότητα του ενδεχομένου $\{X > x \cap Y > y\}$.

Με παρόμοιο σκεπτικό με αυτό που αναφέρθηκε στο Κεφάλαιο 3 και θέλοντας να γενικεύσουμε τις έννοιες της συνάρτησης πιθανότητας και της συνάρτησης πυκνότητας πιθανότητας στην περίπτωση k-διάστατων διακριτών και συνεχών τυχαίων διανυσμάτων, αντίστοιχα, στη συνέχεια μελετώνται ξεχωριστά οι παραπάνω δύο κατηγορίες τυχαίων διανυσμάτων.

6.3 Διακριτό τυχαίο διάνυσμα - Από κοινού συνάρτηση πιθανότητας

Στην περίπτωση που όλες οι μεταβλητές που αποτελούν τις συνιστώσες ενός τυχαίου διανύσματος είναι διακριτές μπορούμε να ορίσουμε και την από κοινού συνάρτηση πιθανότητας.

Ορισμός 6.3

Έστω $X = (X_1, \dots, X_k)^t$ ένα k-διάστατο διακριτό τυχαίο διάνυσμα με σύνολο τιμών S_X με S_X το πολύ αριθμήσιμο σύνολο. Η συνάρτηση $p_X : \mathbb{R}^k \to [0,1]$, που ορίζεται από τη σχέση:

$$p_X(x) = \begin{cases} P(X = x) = P(X_1 = x_1, \dots, X_k = x_k), & x \in S_X, \\ 0, & x \notin S_X. \end{cases}$$

ονομάζεται **από κοινού συνάρτηση πιθανότητας** των $X_1,\dots,X_k$.

Η από κοινού συνάρτηση πιθανότητας των τ.μ. $X_1,\dots,X_k$ στο διάνυσμα $x = (x_1, .., x_k)^t$ εκφράζει την πιθανότητα οι τ.μ. $X_1,\dots,X_k$ να πάρουν ταυτόχρονα τις τιμές $x_1,..,x_k$, αντίστοιχα.

Από τον ορισμό της από κοινού σπ άμεσα προκύπτει ότι:

$$0 \le p_X(x) \le 1, \text{ για όλα τα } x \in \mathbb{R}^k$$

και

$$\sum_{x \in S_X} p_X(x) = 1,$$

που είναι και οι ικανές συνθήκες που πρέπει να πληροί μια συνάρτηση για να είναι η από κοινού συνάρτηση πιθανότητας ενός διακριτού τυχαίου διανύσματος. Επιπροσθέτως ισχύει ότι:

$$P(X \in A) = \sum_{x \in A} p_X(x). \tag{6.2}$$

Η από κοινού συνάρτηση πιθανότητας συνδέεται με την από κοινού συνάρτηση κατανομής μέσω της σχέσης:

$$F_{X_1,\dots,X_k}(x_1, \dots, x_k) = P(X_1 \le x_1, \dots, X_k \le x_k) = \sum_{u_1 \le x_1} \dots \sum_{u_k \le x_k} p_X(u_1, \dots, u_k). \tag{6.3}$$

Τέλος, από την από κοινού συνάρτηση πιθανότητας μπορούμε να προσδιορίσουμε τις περιθώριες συναρτήσεις πιθανότητας. Για παράδειγμα, η περιθώρια κατανομή της X_i υπολογίζεται αν αθροίσουμε την από κοινού συνάρτηση πιθανότητας των $X_1,\dots,X_k$ πάνω από όλες τις δυνατές τιμές των $X_1,\dots,X_{i-1}$, $X_{i+1},\dots,X_k$. Ειδικότερα, αν X και Y είναι δύο διακριτές τυχαίες μεταβλητές με από κοινού συνάρτηση πιθανότητας $p_{X,Y}(x,y) = P(X = x, Y = y)$, τότε η περιθώρια κατανομή της τ.μ. Y δίνεται από τη σχέση:

$$p_Y(y) = \sum_{x : p_{X,Y}(x,y) > 0} p_{X,Y}(x,y), \tag{6.4}$$

ενώ η περιθώρια κατανομή της τ.μ. X από τη σχέση:

$$p_X(x) = \sum_{y : p_{X,Y}(x,y) > 0} p_{X,Y}(x,y). \tag{6.5}$$

Παράδειγμα 6.2

Για την κάλυψη του χώρου μιας βιβλιοθήκης χρησιμοποιούνται δύο δίκτυα ασύρματου internet. Από τη μελέτη που έχει κάνει ο τεχνικός εγκατάστασης είναι γνωστό ότι παρουσιάζονται προβλήματα σύνδεσης (αδύνατο σήμα - όχι καλή κάλυψη) σε ένα μικρό τμήμα της βιβλιοθήκης. Μάλιστα, ο τεχνικός έχει υπολογίσει ότι ένας υπολογιστής με μια συγκεκριμένη κάρτα δικτύου επιτυγχάνει να συνδεθεί σε ένα από τα δύο δίκτυα μία φορά στις τέσσερις προσπάθειες. Έστω X και Y οι τυχαίες μεταβλητές που περιγράφουν τον αριθμό των φορών που συνδέεται ένας υπολογιστής στο δίκτυο Α και Β, αντίστοιχα. Η από κοινού συνάρτηση πιθανότητας των τυχαίων αυτών μεταβλητών δίνεται από τη σχέση:

$$p_{X,Y}(x,y) = \begin{cases} \frac{1}{1+x+y}\binom{5}{x+y}0.25^{x+y}(1-0.25)^{5-x-y}, & x,y \in \{0,1,\ldots,5\}, \text{ με } 0 \le x+y \le 5, \\ 0, & \text{αλλού.} \end{cases}$$

- Να υπολογιστούν οι πιθανότητες για όλα τα δυνατά ενδεχόμενα.
- Να βρεθούν οι περιθώριες συναρτήσεις πιθανότητας των τυχαίων μεταβλητών X και Y.

Λύση Παραδείγματος 6.2

Το σύνολο των δυνατών τιμών του τυχαίου διανύσματος (X,Y) είναι

$$S_{X,Y} = \{(x,y) : x,y \in \{0,1,\ldots,5\}, \text{ με } 0 \le x+y \le 5\}$$

ή, ισοδύναμα, $S_{X,Y} = \{(0,0),(0,1),\ldots,(0,5),(1,0),\ldots(1,4),\ldots(4,0),(4,1),(5,0)\}$.
Για να υπολογίσουμε τις πιθανότητες όλων των δυνατών αποτελεσμάτων, αρκεί να χρησιμοποιήσουμε την από κοινού συνάρτηση πιθανότητας και να υπολογίσουμε την τιμή της πάνω από όλες τις δυνατές τιμές. Παραδείγματος χάριν, αν $X = 0$ και $Y = 1$, τότε

$$\begin{aligned} p_{X,Y}(0,1) &= \frac{1}{1+1+0}\binom{5}{1+0}0.25^{1+0}(1-0.25)^{5-1-0} \\ &= \frac{1}{2}\cdot 5 \cdot 0.25^1 \cdot 0.75^4 = 0.197754 \approx 0.198. \end{aligned}$$

Ακολουθώντας παρόμοια διαδικασία και για τα υπόλοιπα ενδεχόμενα μπορούμε να κατασκευάσουμε τον επόμενο πίνακα στον οποίο παρουσιάζονται (με κάποιες απαραίτητες στρογγυλοποιήσεις κάθε φορά) οι πιθανότητες όλων των ενδεχομένων.

	$X=0$	$X=1$	$X=2$	$X=3$	$X=4$	$X=5$	$p_Y(y)$
$Y=0$	0.237	0.198	0.088	0.022	0.003	0.000	0.548
$Y=1$	0.198	0.088	0.022	0.003	0.000	0.000	0.311
$Y=2$	0.088	0.022	0.003	0.000	0.000	0.000	0.113
$Y=3$	0.022	0.003	0.000	0.000	0.000	0.000	0.025
$Y=4$	0.003	0.000	0.000	0.000	0.000	0.000	0.003
$Y=5$	0.000	0.000	0.000	0.000	0.000	0.000	0.000
$p_X(x)$	0.548	0.311	0.113	0.025	0.003	0.000	1.000

Στα περιθώρια του παραπάνω πίνακα εμφανίζονται και οι περιθώριες συναρτήσεις πιθανότητας των X και Y, οι οποίες έχουν υπολογιστεί αθροίζοντας τις τιμές στα κελιά κατά στήλη και γραμμή αντίστοιχα. Για παράδειγμα, για την εύρεση της $P(X=1)$ έχουμε ότι

$$P(X=1) = \sum_{y:p_{X,Y}(1,y)>0} p_{X,Y}(1,y)$$

ή, ισοδύναμα,

$$P(X=1) = p_{X,Y}(1,0) + p_{X,Y}(1,1) + p_{X,Y}(1,2) + p_{X,Y}(1,3) + p_{X,Y}(1,4) + p_{X,Y}(1,5) = 0.31071.$$

Άσκηση Αυτοαξιολόγησης 6.2

Η από κοινού συνάρτηση πιθανότητας των τ.μ. X και Y δίνεται από τη σχέση:

$$p_{X,Y}(x,y) = \begin{cases} \frac{(x+y)^2}{48}, & x,y \in \{0,1,2\}, \\ 0, & \text{αλλού.} \end{cases}$$

- Να υπολογιστούν οι πιθανότητες κάθε δυνατού ενδεχομένου.
- Να υπολογιστούν οι περιθώριες συναρτήσεις πιθανότητας των τ.μ. X και Y.

6.4 Συνεχές τυχαίο διάνυσμα - Από κοινού συνάρτηση πυκνότητας πιθανότητας

Στην περίπτωση που όλες οι μεταβλητές που αποτελούν τις συνιστώσες ενός τυχαίου διανύσματος είναι συνεχείς μπορούμε να ορίσουμε, για τους ίδιους λόγους που αναφέρθηκαν στην περίπτωση μιας συνεχούς τυχαίας μεταβλητής, την από κοινού συνάρτηση πυκνότητας πιθανότητας.

Ορισμός 6.4

Έστω $X = (X_1, \dots, X_k)^t$ ένα $k-$διάστατο τυχαίο διάνυσμα με τιμές στο σύνολο $\mathbb{R}^k$. Το τυχαίο διάνυσμα X λέγεται **συνεχές τυχαίο διάνυσμα** αν υπάρχει μια μη αρνητική ολοκληρώσιμη πραγματική συνάρτηση $f_X(\cdot)$ ορισμένη στο σύνολο $\mathbb{R}^k$, τέτοια ώστε:

$$\begin{aligned} P(X \in C) &= \int_C f_X(x)dx, \quad C \subseteq \mathbb{R}^k \\ &= \int \int \dots \int_C f_{X_1,\dots,X_k}(x_1, \dots, x_k)dx_1 \dots dx_k. \end{aligned}$$

Η συνάρτηση $f_X(\cdot)$ ονομάζεται **από κοινού συνάρτηση πυκνότητας πιθανότητας** των $X_1,\dots,X_k$ ή συνάρτηση πυκνότητας πιθανότητας του τυχαίου διανύσματος X.

Άμεσες συνέπειες του ορισμού είναι ότι η από κοινού σππ του τυχαίου διανύσματος X με σύνολο δυνατών τιμών S_X ικανοποιεί τις ιδιότητες:

$$f_X(x) \geq 0 \text{ για όλα τα } x \in \mathbb{R}^k$$

και

$$\int_{x \in S_X} f_X(x)dx = \int \int \dots \int_{S_X} f_{X_1,\dots,X_k}(x_1, \dots, x_k)dx_1 \dots dx_k = 1,$$

που είναι και οι ικανές συνθήκες που πρέπει να πληροί μια πραγματική συνάρτηση για να είναι η από κοινού συνάρτηση πυκνότητας πιθανότητας ενός συνεχούς τυχαίου διανύσματος.

Παρατήρηση 6.1

Στην περίπτωση που το C μπορεί να γραφτεί ως το καρτεσιανό γινόμενο των συνόλων $A_1,\dots A_k$, δηλαδή αν το C είναι τέτοιο ώστε $X_1 \in A_1$, $X_2 \in A_2,\dots$, $X_k \in A_k$, τότε από τον ορισμό της από κοινού σππ προκύπτει ότι:

$$P(\{X_1, \dots, X_k\} \in C) = \int_{A_1} \dots \int_{A_k} f_{X_1,\dots,X_k}(x_1, \dots, x_k)dx_k \dots dx_1. \tag{6.6}$$

Από τη σχέση (6.6) ουσιαστικά προκύπτει ο τρόπος προσδιορισμού της από κοινού αθροιστικής συνάρτησης κατανομής του τυχαίου διανύσματος X από την από κοινού συνάρτηση πυκνότητας πιθανότητάς του, καθώς

$$F_X(x) = P(X_1 \leq x_1, \dots, X_k \leq x_k) = \int_{-\infty}^{x_k} \dots \int_{-\infty}^{x_1} f_{X_1,\dots,X_k}(u_1, \dots, u_k) du_1 \dots du_k. \tag{6.7}$$

Τέλος, υπό την προϋπόθεση ότι οι μερικές παράγωγοι ορίζονται, ισχύει ότι

$$f_{X_1,\dots,X_k}(x_1, \dots, x_k) = \frac{\partial^k F_{X_1,\dots,X_k}(x_1, \dots, x_k)}{\partial x_1 \dots \partial x_k}. \tag{6.8}$$

Παρατήρηση 6.2

Στην ειδική περίπτωση του διδιάστατου τυχαίου διανύσματος (X, Y) οι σχέσεις (6.6) και (6.7) μπορούν να συνδυαστούν και να προκύψει ότι

$$\begin{aligned} P(x_1 < X < x_2, y_1 < Y < y_2) &= \int_{x_1}^{x_2} \int_{y_1}^{y_2} f_{X,Y}(x,y) dy dx \\ &= F_{X,Y}(x_2, y_2) - F_{X,Y}(x_2, y_1) - F_{X,Y}(x_1, y_2) + F_{X,Y}(x_1, y_1), \end{aligned}$$

μια σχέση ιδιαίτερα χρήσιμη για τον υπολογισμό της πιθανότητας οποιουδήποτε ορθογωνίου στο $\mathbb{R}^2$ με χρήση της από κοινού αθροιστικής συνάρτησης κατανομής.

Έστω $X = (X_1, \dots, X_k)^t$ ένα $k-$διάστατο συνεχές τυχαίο διάνυσμα. Μπορούμε να ορίσουμε, εκτός από τις περιθώριες αθροιστικές συναρτήσεις κατανομής καθεμίας εκ των $X_1,\dots,X_k$ ή ενός υποσυνόλου αυτών, και τις περιθώριες συναρτήσεις πυκνότητας πιθανότητας αυτών. Η περιθώρια συνάρτηση πυκνότητας πιθανότητας της X_i, $i = 1, 2, \dots, k$ ισούται με

$$f_{X_i}(x_i) = \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} f_{X_1,\dots,X_k}(x_1, \dots, x_k) dx_1 \dots dx_{i-1} dx_{i+1} \dots dx_k. \tag{6.9}$$

Ειδικότερα, αν X και Y είναι δύο συνεχείς τυχαίες μεταβλητές με από κοινού συνάρτηση πυκνότητας πιθανότητας $f_{X,Y}(x,y)$, τότε η περιθώρια κατανομή της τυχαίας μεταβλητής Y δίνεται από τη σχέση:

$$f_Y(y) = \int_{x: f_{X,Y}(x,y)>0} f_{X,Y}(x,y) dx, \tag{6.10}$$

ενώ η περιθώρια κατανομή της τυχαίας μεταβλητής X δίνεται από τη σχέση:

$$f_X(x) = \int_{y: f_{X,Y}(x,y)>0} f_{X,Y}(x,y) dy. \tag{6.11}$$

Παράδειγμα 6.3

Έστω X και Y οι τυχαίες μεταβλητές που παριστάνουν την αξιολόγηση της απόδοσης ενός ηλεκτρονικού υπολογιστή, στην κλίμακα 0-10, από δύο άτομα. Η από κοινού συνάρτηση πυκνότητας πιθανότητας του τυχαίου διανύσματος (X, Y) δίνεται από τη σχέση

$$f_{X,Y}(x,y) = cxy, \quad 0 < x < 10, 0 < y < 10.$$

1. Να προσδιοριστεί η σταθερά c.

2. Να υπολογιστεί η πιθανότητα η βαθμολογία του πρώτου ατόμου να είναι μικρότερη από την τιμή 5 και του δεύτερου ατόμου μεγαλύτερη από την τιμή 5.
3. Να υπολογιστεί η πιθανότητα η βαθμολογία του πρώτου ατόμου να είναι μικρότερη από τη βαθμολογία του δεύτερου ατόμου.
4. Να βρεθεί η περιθώρια συνάρτηση πυκνότητας πιθανότητας της βαθμολογίας του δεύτερου ατόμου.

Λύση Παραδείγματος 6.3

1. Η σταθερά c θα προσδιοριστεί από τις δύο συνθήκες που πρέπει να ικανοποιεί η από κοινού συνάρτηση πυκνότητας πιθανότητας των X και Y, δηλαδή από τις

   (α΄) $f_{X,Y}(x,y) \geq 0, \quad \forall (x,y) \in \mathbb{R}^2,$

   (β΄) $\int\int f_{X,Y}(x,y)dydx = 1.$

   Από την πρώτη συνθήκη είναι φανερό ότι η σταθερά c πρέπει να είναι μη αρνητική, ενώ από τη δεύτερη έχουμε ότι:

$$\begin{aligned}\int_x\int_y f_{X,Y}(x,y)dydx &= \int_0^{10}\int_0^{10} cxy\ dydx = c\int_0^{10}\left(\frac{y^2}{2}\Big|_0^{10}\right)x\ dx\\ &= c\int_0^{10} 50\ x\ dx = 50c\left(\frac{x^2}{2}\Big|_0^{10}\right)\\ &= 50^2 c.\end{aligned}$$

   Επομένως πρέπει $50^2 c = 1$, δηλαδή $c = 1/2500$. Η τιμή αυτή είναι μη αρνητική, οπότε ικανοποιεί και την πρώτη συνθήκη. Άρα η από κοινού συνάρτηση πυκνότητας πιθανότητας των X και Y δίνεται από τη σχέση:

$$f_{X,Y}(x,y) = \frac{xy}{2500}, \qquad 0 < x < 10, 0 < y < 10.$$

2. Η πιθανότητα η βαθμολογία του πρώτου ατόμου να είναι μικρότερη από την τιμή 5 και του δεύτερου ατόμου μεγαλύτερη από την τιμή 5 δίνεται από τη σχέση:

$$\begin{aligned}P(X < 5, Y > 5) &= \int_0^5\int_5^{10}\frac{1}{2500}xy\ dydx = \frac{1}{2500}\int_0^5\left(\frac{y^2}{2}\Big|_5^{10}\right)x\ dx\\ &= \frac{1}{2500}\int_0^5\frac{100-25}{2}\ x\ dx = \frac{1}{2500}\frac{75}{2}\left(\frac{x^2}{2}\Big|_0^5\right)\\ &= \frac{1}{2500}\frac{75}{2}\frac{25}{2}\\ &= 0.1875.\end{aligned}$$

3. Η πιθανότητα η βαθμολογία του πρώτου ατόμου να είναι μικρότερη από τη βαθμολογία του δεύτερου ατόμου υπολογίζεται ως εξής:

$$P(X<Y)=\int_0^{10}\int_x^{10}\frac{1}{2500}xy\,dydx=\frac{1}{2500}\int_0^{10}\left(\frac{y^2}{2}\Big|_x^{10}\right)x\,dx$$
$$=\frac{1}{2500}\int_0^{10}\frac{100-x^2}{2}\,x\,dx=\frac{1}{2500}\int_0^{10}50x-\frac{x^3}{2}\,dx$$
$$=\frac{1}{2500}\left(\frac{50x^2}{2}-\frac{x^4}{2\cdot 4}\Big|_0^{10}\right)=\frac{1}{2500}\left(25\cdot 100^2-\frac{100^4}{8}\right)$$
$$=\frac{1250}{2500}=1/2.$$

Σημειώνεται ότι η παραπάνω πιθανότητα υπολογίστηκε αφήνοντας την τυχαία μεταβλητή X να κινείται ελεύθερη στο πεδίο τιμών της, δηλαδή το (0, 10), και την Y να περιορίζεται από κάτω από την τιμή της X και από πάνω από το 10 (την ανώτερη δυνατή τιμή της). Εναλλακτικά, θα μπορούσαμε να είχαμε υπολογίσει την παραπάνω πιθανότητα αφήνοντας την τυχαία μεταβλητή Y να κινείται ελεύθερη στο πεδίο τιμών της, δηλαδή το (0, 10), και τη X να περιορίζεται κάτω από την τιμή 0 και πάνω από την τιμή της τ.μ. Y, δηλαδή

$$P(X<Y)=\int_0^{10}\int_0^{y}\frac{1}{2500}xy\,dxdy=\ldots=1/2.$$

4. Η περιθώρια συνάρτηση πυκνότητας πιθανότητας της βαθμολογίας του Y υπολογίζεται ολοκληρώνοντας την από κοινού συνάρτηση πιθανότητας ως προς όλες τις τιμές της X για τις οποίες $f_{X,Y}(x,y)>0$. Δηλαδή, έχουμε ότι:

$$f_Y(y)=\int_0^{10}\frac{1}{2500}xy\,dx=\frac{1}{2500}y\left(\frac{x^2}{2}\Big|_0^{10}\right)$$
$$=\frac{50}{2500}y=\frac{1}{50}y,\qquad 0<y<10.$$

Άσκηση Αυτοαξιολόγησης 6.3

Μια μηχανή αποτελείται από δύο εξαρτήματα. Έστω X, Y οι τυχαίες μεταβλητές που παριστάνουν τη διάρκεια ζωής του πρώτου και δεύτερου εξαρτήματος, αντίστοιχα. Η από κοινού αθροιστική συνάρτηση κατανομής της διάρκειας ζωής τους (σε έτη) δίνεται από τη σχέση:

$$F_{X,Y}(x,y)=(1-e^{-2x})(1-e^{-2y}),\qquad x,y>0,$$

ενώ $F_{X,Y}(x,y)=0$ αλλού.

1. Να προσδιοριστεί η από κοινού συνάρτηση πυκνότητας πιθανότητας του τυχαίου διανύσματος (X,Y). Υπόδειξη: βλ. τη σχέση (6.8).
2. Να βρεθεί η πιθανότητα η διάρκεια ζωής του πρώτου εξαρτήματος να είναι από ένα έως δύο έτη και του δεύτερου μικρότερη των δύο ετών.
3. Να προσδιοριστούν οι περιθώριες συναρτήσεις πυκνότητας πιθανότητας των τυχαίων μεταβλητών X και Y.

6.5 Υπό συνθήκη ή δεσμευμένες κατανομές

Έστω $X = (X_1, \dots, X_k)^t$ ένα $k-$διάστατο (συνεχές ή διακριτό) τυχαίο διάνυσμα. Σε πολλές περιπτώσεις μας ενδιαφέρει η κατανομή ενός υποσυνόλου των τυχαίων μεταβλητών όταν γνωρίζουμε την τιμή των υπόλοιπων τυχαίων μεταβλητών. Οι κατανομές αυτές ονομάζονται **υπό συνθήκη ή δεσμευμένες κατανομές**.

Στη συνέχεια, για την καλύτερη κατανόησή τους, οι δεσμευμένες κατανομές θα παρουσιαστούν για την ειδική περίπτωση του διδιάστατου διακριτού ή συνεχούς τυχαίου διανύσματος. Στο πλαίσιο αυτό, έστω $X = (X_1, X_2)^t$ ένα διδιάστατο διακριτό τυχαίο διάνυσμα. Η δεσμευμένη ή υπό συνθήκη κατανομή της τ.μ. X_2, δοθέντος ότι $X_1 = x_1$, με βάση τον ορισμό της δεσμευμένης πιθανότητας, ισούται με

$$p_{X_2|X_1=x1}(x_2) = P(X_2 = x_2|X_1 = x_1) = \frac{P(X_1 = x_1, X_2 = x_2)}{P(X_1 = x_1)} = \frac{p_{X_1,X_2}(x_1,x_2)}{p_{X_1}(x_1)},$$

υπό την προϋπόθεση ότι το x_1 είναι τέτοιο ώστε $p_{X_1}(x_1) > 0$.

Με παρόμοιο τρόπο ορίζεται και η συνάρτηση πυκνότητας πιθανότητας της υπό συνθήκη κατανομής X_2, δοθέντος ότι $X_1 = x_1$ για την περίπτωση όπου το $X = (X_1, X_2)^t$ είναι ένα διδιάστατο συνεχές τυχαίο διάνυσμα. Είναι τότε:

$$f_{X_2|X_1=x_1}(x_2) = \frac{f_{X_1,X_2}(x_1,x_2)}{f_{X_1}(x_1)}, \quad \text{με } x_1 : f_{X_1}(x_1) > 0.$$

Σημειώνεται ότι με ανάλογο τρόπο μπορεί να οριστεί η υπό συνθήκη κατανομή της X_1, δοθέντος ότι $X_2 = x_2$.

Παρατήρηση 6.3

Όπως προαναφέρθηκε, οι υπό συνθήκη κατανομές για λόγους απλούστευσης παρουσιάστηκαν για τη διδιάστατη περίπτωση, ενώ με παρόμοιο σκεπτικό προκύπτουν αντίστοιχες σχέσεις στην περίπτωση περισσότερων από δύο τυχαίων μεταβλητών. Παραδείγματος χάριν, έστω $X = (X_1, X_2, X_3, X_4)^t$ ένα συνεχές τυχαίο διάνυσμα με από κοινού σππ $f_{X_1,X_2,X_3,X_4}(x_1, x_2, x_3, x_4)$, τότε η υπό συνθήκη κατανομή των X_1 και X_2, δοθέντος ότι $X_3 = x_3$ και $X_4 = x_4$, μπορεί να περιγραφθεί από τη σχέση

$$f_{X_1,X_2|X_3=x_3,X_4=x_4}(x_1,x_2) = \frac{f_{X_1,X_2,X_3,X_4}(x_1,x_2,x_3,x_4)}{f_{X_3,X_4}(x_3,x_4)}, \quad \text{με } (x_3,x_4) : f_{X_3,X_4}(x_3,x_4) > 0.$$

Παράδειγμα 6.4

Έστω X και Y οι τυχαίες μεταβλητές που παριστάνουν το κόστος (σε εκατοντάδες χιλιάδες ευρώ) προμήθειας πρώτων υλών και απασχόλησης εργαζομένων (εργατικό κόστος), αντίστοιχα, για την κατασκευή μιας πεζογέφυρας. Η από κοινού συνάρτηση πυκνότητας πιθανότητας των τ.μ. X και Y δίνεται από τη σχέση:

$$f_{X,Y}(x,y) = \begin{cases} xe^{-x(y+1)}, & x > 0,\ y > 0, \\ 0, & \text{αλλού.} \end{cases}$$

Να υπολογιστεί η υπό συνθήκη συνάρτηση πυκνότητας πιθανότητας του εργατικού κόστους Y, δοθέντος ότι το κόστος X της προμήθειας των πρώτων υλών ισούται με $X = x$.

Λύση Παραδείγματος 6.4

Η υπό συνθήκη συνάρτηση πυκνότητας πιθανότητας του εργατικού κόστους Y, δοθέντος ότι το κόστος X της προμήθειας των πρώτων υλών ισούται με $X = x$, δίνεται από τη σχέση:

$$f_{Y|X=x}(y) = \frac{f_{X,Y}(x,y)}{f_X(x)} = \frac{xe^{-x(y+1)}}{\int_0^{+\infty} xe^{-x(y+1)}dy}$$
$$= \frac{xe^{-x(y+1)}}{e^{-x}} = xe^{-xy},\ y > 0.$$

Επομένως, η $Y|X = x$ είναι μια εκθετική κατανομή με παράμετρο x.

Άσκηση Αυτοαξιολόγησης 6.4

Για την από κοινού συνάρτηση πυκνότητας πιθανότητας του Παραδείγματος 6.4 να υπολογιστεί η υπό συνθήκη συνάρτηση πυκνότητας πιθανότητας του κόστους X της προμήθειας των πρώτων υλών, αν γνωρίζουμε ότι το εργατικό κόστος ισούται με $Y = y$.

6.6 Χαρακτηριστικά μέτρα πολυδιάστατων τυχαίων κατανομών

Όπως και στην περίπτωση των μονοδιάστατων τυχαίων μεταβλητών, έτσι και στις πολυδιάστατες τυχαίες μεταβλητές είναι χρήσιμες κάποιες αριθμητικές τιμές που έχουν ως στόχο τη συνοπτική παρουσίαση της συμπεριφοράς των μεταβλητών που αποτελούν τις συνιστώσες ενός τυχαίου διανύσματος. Για τον σκοπό αυτό, στην ενότητα αυτή, θα παρουσιαστούν επεκτάσεις/ γενικεύσεις κάποιων εννοιών που παρουσιάστηκαν στη μονοδιάστατη περίπτωση (αναμενόμενη τιμή, ροπογεννήτρια) αλλά και κάποιες νέες έννοιες, όπως η συνδιασπορά και η συσχέτιση.

6.6.1 Αναμενόμενη τιμή και ροπογεννήτρια

Στην ενότητα αυτή γενικεύονται οι έννοιες της αναμενόμενης τιμής και της ροπογεννήτριας τυχαίου διανύσματος.

Ορισμός 6.5

Έστω $X = (X_1, \dots, X_k)^t$ ένα k-διάστατο τυχαίο διάνυσμα (διακριτό ή συνεχές). Η αναμενόμενη τιμή του τυχαίου διανύσματος ή αλλιώς το μέσο διάνυσμα ή απλώς η μέση τιμή του τυχαίου διανύσματος είναι το διάνυσμα των αναμενόμενων τιμών των αντίστοιχων μονοδιάστατων τυχαίων μεταβλητών, δηλαδή $E(X) = (E(X_1), \dots, E(X_k))^t$.

Ορισμός 6.6

Έστω $X = (X_1, \dots, X_k)^t$ ένα k-διάστατο τυχαίο διάνυσμα (διακριτό ή συνεχές) με από κοινού συνάρτηση πιθανότητας $p_{X_1,\dots,X_k}(x_1, \dots, x_k)$ ή με από κοινού συνάρτηση πυκνότητας πιθανότητας $p_{X_1,\dots,X_k}(x_1, \dots, x_k)$, ανάλογα αν είναι διακριτό ή συνεχές, αντίστοιχα. Η αναμενόμενη τιμή μιας συνάρτησης $g(X_1, \dots, X_k)$ των τυχαίων μεταβλητών $X_1, \dots, X_k$ συμβολίζεται με $E\left(g(X_1, X_2, \dots, X_k)\right)$ και ορίζεται από τη σχέση

$$E(g(X_1, \dots, X_k)) = \sum_{x_1} \cdots \sum_{x_k} g(x_1, \dots, x_k) p_{X_1,\dots,X_k}(x_1, \dots, x_k)$$

αν οι $X_1, X_2, \dots, X_k$ είναι διακριτές τυχαίες μεταβλητές, και από τη σχέση:

$$E(g(X_1, \dots, X_k)) = \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} g(x_1, \dots, x_k) f_{X_1,\dots,X_k}(x_1, \dots, x_k) dx_1 \dots dx_k$$

αν είναι συνεχείς. Τα παραπάνω ισχύουν υπό την προϋπόθεση ότι το άθροισμα ή το ολοκλήρωμα, αντίστοιχα, συγκλίνει απόλυτα.

Με βάση τον παραπάνω ορισμό, είναι φανερό ότι οι ιδιότητες της αναμενόμενης τιμής, όπως αυτές παρουσιάστηκαν στη μονοδιάστατη περίπτωση, μπορούν να επεκταθούν και στην περίπτωση αυτή. Παραδείγματος χάριν, αν $g_1(\dots),\dots, g_n(\dots)$ είναι πραγματικές συναρτήσεις και $a_1,\dots,a_n$ και $b_1,\dots,b_n$ είναι πραγματικοί αριθμοί (σταθερές), τότε έχουμε ότι

$$E\left[\sum_{j=1}^{n}\left(a_j g_j(X_1, \dots, X_k) + b_j\right)\right] = \sum_{j=1}^{n}\left(a_j E\left[g_j(X_1, \dots, X_k)\right] + b_j\right).$$

Παρατήρηση 6.4

Με βάση τον παραπάνω ορισμό μπορούν να οριστούν και οι αναμενόμενες τιμές $\mu_i = E(X_i)$, οι ροπές $E(X_i^r)$ και οι διασπορές $Var(X_i) = E[(X_i - \mu_i)^2]$ καθεμίας μεμονωμένης τυχαίας μεταβλητής X_i, δεδομένου ότι οι X_i, X_i^r και $(X_i - \mu_i)^2$ αποτελούν ειδικές περιπτώσεις συναρτήσεων των τυχαίων μεταβλητών $X_1, \dots, X_k$. Μάλιστα, οι ποσότητες αυτές ταυτίζονται με τις αντίστοιχες ποσότητες των μονοδιάστατων τυχαίων μεταβλητών. Παραδείγματος χάριν, η διασπορά της συνεχούς τυχαίας μεταβλητής X_1 υπολογίζεται μέσω της σχέσης

$$\begin{aligned} Var(X_1) = E[(X_1 - \mu_1)^2] &= \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} (x_1 - \mu_1)^2 f_{X_1,\dots,X_k}(x_1, \dots, x_k) dx_1 \dots dx_k \\ &= \int_{-\infty}^{+\infty} (x_1 - \mu_1)^2 \left(\int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} f_{X_1,\dots,X_k}(x_1, \dots, x_k) dx_2 \dots dx_k \right) dx_1. \end{aligned}$$

Παρατηρώντας ότι τα εσωτερικά ολοκληρώματα ορίζουν την περιθώρια συνάρτηση πυκνότητας πιθανότητας της X_1, έχουμε ότι

$$Var(X_1) = \int_{-\infty}^{+\infty} (x_1 - \mu_1)^2 f_{X_1}(x_1) dx_1.$$

Είναι πλέον προφανές ότι η παραπάνω σχέση ταυτίζεται με τη σχέση ορισμού της διασποράς στη μονοδιάστατη περίπτωση.

Παράδειγμα 6.5

Σε συνέχεια του Παραδείγματος 6.3 να βρεθούν οι $E(XY)$, $E(Y)$ και $E(X - 2Y)$.

Λύση Παραδείγματος 6.5

Η αναμενόμενη τιμή της XY δίνεται από τη σχέση:

$$\begin{aligned} E(XY) &= \int_0^{10}\int_0^{10} xy \frac{1}{2500} xy \, dxdy = \int_0^{10}\int_0^{10} \frac{1}{2500} x^2 y^2 dxdy \\ &= \frac{1}{2500}\left(\frac{x^3}{3}\Big|_0^{10}\right)\left(\frac{y^3}{3}\Big|_0^{10}\right) = \frac{1}{2500}\frac{1000}{3}\frac{1000}{3} = \frac{400}{9}. \end{aligned}$$

Η αναμενόμενη τιμή της Y δίνεται από τη σχέση:

$$E(Y) = \int_0^{10}\int_0^{10} y\frac{1}{2500}xy\,dxdy = \int_0^{10}\frac{1}{2500}y^2\left(\frac{x^2}{2}\Big|_0^{10}\right)dy$$
$$= \int_0^{10} y^2\frac{50}{2500}ydy = \frac{50}{2500}\left(\frac{y^3}{3}\Big|_0^{10}\right)$$
$$= \frac{50}{2500}\cdot\frac{1000}{3} = \frac{20}{3}.$$

Η αναμενόμενη τιμή της $X-2Y$ δίνεται από τη σχέση:

$$E(X-2Y) = E(X) - E(2Y) = E(X) - 2E(Y)$$
$$= \int_0^{10}\int_0^{10} x\frac{1}{2500}xy\,dxdy - 2\int_0^{10}\int_0^{10} y\frac{1}{2500}xy\,dxdy$$
$$= \frac{20}{3} - 2\cdot\frac{20}{3} = -\frac{20}{3}.$$

Εναλλακτικά, για τον προσδιορισμό της αναμενόμενης τιμής των X και Y θα μπορούσαν να χρησιμοποιηθούν οι περιθώριες κατανομές των X και Y. Τότε για παράδειγμα

$$\mathrm{E}(Y) = \int_0^{10} y f_Y(y)dy = \int_0^{10} y\frac{y}{50}dy = \frac{1}{50}\left(\frac{y^3}{3}\Big|_0^{10}\right) = \frac{20}{3},$$

και με παρόμοιο τρόπο $E(X) = \frac{20}{3}$.

Άσκηση Αυτοαξιολόγησης 6.5: (Walpole *et al.*, 2017)

Η από κοινού συνάρτηση πυκνότητας πιθανότητας των τυχαίων μεταβλητών X και Y δίνεται από τη σχέση:

$$f_{X,Y}(x,y) = \begin{cases} 24xy, & 0<x<1,\ 0<y<1,\ 0<x+y<1, \\ 0, & \text{αλλού.} \end{cases}$$

1. Να υπολογιστεί η αναμενόμενη τιμή της X.
2. Να υπολογιστεί η αναμενόμενη τιμή της Y^2.
3. Να υπολογιστεί η αναμενόμενη τιμή της $X+Y^2$.

Άσκηση Αυτοαξιολόγησης 6.6

Σε μια αβαρή δοκό τοποθετούνται σε απόσταση α και 2α από το σημείο στήριξής της δύο τυχαία φορτία βάρους W_1 και W_2 με από κοινού συνάρτηση πυκνότητας πιθανότητας

$$f(w_1,w_2) = \begin{cases} w_1 e^{-(w_1+w_2)}, & w_1>0,\ w_2>0 \\ 0, & \text{αλλού.} \end{cases}$$

Να υπολογιστεί η αναμενόμενη τιμή της ροπής κάμψης στο σημείο στήριξής της.
Σημειώνεται ότι στη συγκεκριμένη περίπτωση η ροπή κάμψης στο σημείο στήριξης της δοκού για συγκεκριμένα βάρη w_1 και w_2 δίνεται από τη σχέση $M = \alpha w_1 + 2\alpha w_2$.

Η έννοια της ροπογεννήτριας μιας πολυδιάστατης κατανομής αποτελεί άμεση επέκταση της αντίστοιχης έννοιας της μονοδιάστατης περίπτωσης και οι ιδιότητές τους ταυτίζονται. Στη συνέχεια, παρατίθεται ο ορισμός.

Ορισμός 6.7

Έστω $X = (X_1, \dots, X_k)^t$ ένα k-διάστατο τυχαίο διάνυσμα (διακριτό ή συνεχές). Η από κοινού ροπογεννήτρια των $X_1,\dots,X_k$ ορίζεται από τη σχέση:

$$M_X(t) = M_{X_1,\dots,X_k}(t_1,\dots,t_k) = E\left(e^{t_1X_1+\dots+t_kX_k}\right)$$

υπό την προϋπόθεση ότι η αναμενόμενη τιμή υπάρχει για κάθε $(t_1,\dots,t_k) \in (-h_1,h_1) \times \dots \times (-h_k,h_k)$ με $h_1 > 0,\dots,h_k > 0$.

6.6.2 Δεσμευμένη αναμενόμενη τιμή και διασπορά

Η μέση τιμή και η διασπορά της υπό συνθήκη κατανομής μιας τυχαίας μεταβλητής ονομάζονται δεσμευμένη αναμενόμενη τιμή και δεσμευμένη διασπορά, αντίστοιχα. Έτσι στην περίπτωση δύο τυχαίων μεταβλητών X_1 και X_2 έχουμε τους ακόλουθους ορισμούς.

Ορισμός 6.8

Έστω X_1 και X_2 δύο τυχαίες μεταβλητές και $g(X_1,X_2)$ μια συνάρτηση αυτών. Η δεσμευμένη αναμενόμενη τιμή της $g(X_1,X_2)$, δοθέντος $X_2 = x_2$, ορίζεται από τη σχέση:

$$E(g(X_1,X_2)|X_2 = x_2) = \sum_{x_1} g(x_1,x_2)p_{X_1|X_2=x_2}(x_1) \tag{6.12}$$

αν το τυχαίο διάνυσμα (X_1,X_2) είναι διακριτό. Από την άλλη πλευρά, η δεσμευμένη αναμενόμενη τιμή της $g(X_1,X_2)$, δοθέντος $X_2 = x_2$, ορίζεται από τη σχέση:

$$E(g(X_1,X_2)|X_2 = x_2) = \int_{-\infty}^{\infty} g(x_1,x_2)f_{X_1|X_2=x_2}(x_1)dx_1 \tag{6.13}$$

αν το τυχαίο διάνυσμα (X_1,X_2) είναι συνεχές με $p_{X_1|X_2=x_2}(x_1)$ να είναι η υπό συνθήκη συνάρτηση πιθανότητας και $f_{X_1|X_2=x_2}(x_1)$ η υπό συνθήκη συνάρτηση πυκνότητας πιθανότητας, αντίστοιχα.

Εφαρμόζοντας τον παραπάνω ορισμό, άμεσα προκύπτει ότι η μέση τιμή και η διασπορά της υπό συνθήκη κατανομής προσδιορίζονται από τις σχέσεις:

$$\mu_{X_1|X_2=x2} = E(X_1|X_2 = x_2) = \sum_{x_1} x_1 p_{X_1|X_2=x_2}(x_1)$$

$$Var(X_1|X_2 = x_2)) = \sum_{x_1} (x_1 - \mu_{X_1|X_2=x2})^2 p_{X_1|X_2=x_2}(x_1)$$

στην περίπτωση διδιάστατων διακριτών τυχαίων διανυσμάτων ή από τις σχέσεις

$$\mu_{X_1|X_2=x2} = E(X_1|X_2 = x_2) = \int_{-\infty}^{+\infty} x_1 f_{X_1|X_2=x_2}(x_1)dx_1$$

$$Var(X_1|X_2 = x_2) = \int_{-\infty}^{+\infty} (x_1 - \mu_{X_1|X_2=x2})^2 f_{X_1|X_2=x_2}(x_1)dx_1$$

στην περίπτωση των διδιάστατων συνεχών τυχαίων διανυσμάτων.

Παρατήρηση 6.5

Σημειώνεται ότι οι παραπάνω σχέσεις γενικεύονται εύκολα στην περίπτωση των πολυδιάστατων κατανομών.

Άσκηση Αυτοαξιολόγησης 6.7

Για την από κοινού συνάρτηση πυκνότητας πιθανότητας του Παραδείγματος 6.4 να υπολογιστεί η αναμενόμενη τιμή του κόστους X της προμήθειας των πρώτων υλών, αν γνωρίζουμε ότι το εργατικό κόστος ισούται με $Y = y$.

Η δεσμευμένη μέση τιμή και η δεσμευμένη διασπορά έχουν τις ίδιες ιδιότητες με την αναμενόμενη τιμή και διασπορά, αντίστοιχα. Παραδείγματος χάριν, ακολουθούν τους ίδιους κανόνες για τους γραμμικούς μετασχηματισμούς των τυχαίων μεταβλητών. Ωστόσο παρατηρήστε ότι η ποσότητα $E(g(X_1, X_2)|X_2 = x_2)$ είναι μια σταθερά ποσότητα, ενώ η $E(g(X_1, X_2)|X_2)$ είναι τυχαία μεταβλητή συνάρτηση της τυχαίας μεταβλητής X_2. Αυτή η παρατήρηση είναι πολύ χρήσιμη για την απόδειξη των σχέσεων που ακολουθούν (βλ. και Παπαϊωάννου, 1997).

Θεώρημα 6.1

Αν (X, Y) μια διδιάστατη τυχαία μεταβλητή, τότε υπό την προϋπόθεση ότι οι μέσες τιμές που εμφανίζονται υπάρχουν, ισχύει ότι:

$$\mathrm{E}(g(X,Y)) = \mathrm{E}[\mathrm{E}(g(X,Y)|Y)].$$

Απόδειξη Θεωρήματος 6.1

Η απόδειξη θα γίνει χωρίς βλάβη της γενικότητας για τη συνεχή περίπτωση, ενώ η διακριτή περίπτωση αποδεικνύεται ανάλογα. Είναι

$$\mathrm{E}[\mathrm{E}(g(X,Y)|Y)] = \int_{-\infty}^{+\infty} \mathrm{E}(g(X,Y)|Y = y) f_Y(y) dy.$$

Όμως η $\mathrm{E}(g(X,Y)|Y = y)$ ισούται με

$$\mathrm{E}(g(X,Y)|Y = y) = \int_{-\infty}^{+\infty} g(x,y) f_{X|Y=y}(x|y) dx$$

και, επομένως, η $\mathrm{E}[\mathrm{E}(g(X,Y)|Y)]$ εκφράζεται ως

$$\mathrm{E}[\mathrm{E}(g(X,Y)|Y)] = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} g(x,y) f_{X|Y=y}(x|y) f_Y(y) dx dy.$$

Όμως η $f_{X|Y=y}(x|y) f_Y(y)$ ισούται με την από κοινού συνάρτηση πυκνότητας πιθανότητας των (X, Y) και έτσι έχουμε ότι

$$\mathrm{E}[\mathrm{E}(g(X,Y)|Y)] = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} g(x,y) f_{X,Y}(x,y) dx dy.$$

Το δεξί μέλος της παραπάνω ισότητας είναι η $E(g(X, Y))$, γεγονός που αποδεικνύει τη ζητούμενη σχέση.

Άμεσα προκύπτει το επόμενο πόρισμα.

Πόρισμα 6.1

Αν (X, Y) μια διδιάστατη τυχαία μεταβλητή, τότε υπό την προϋπόθεση ότι οι μέσες τιμές που εμφανίζονται υπάρχουν, ισχύει ότι:

$$\mathrm{E}(X) = \mathrm{E}[\mathrm{E}(X|Y)].$$

Απόδειξη Πορίσματος 6.1

Η απόδειξη προκύπτει άμεσα από την προηγούμενη πρόταση για $g(X, Y) = X$.

Παρατήρηση 6.6

Πολύ συχνά οι δεσμευμένες μέσες τιμές $\mathrm{E}[\mathrm{E}(g(X, Y)|Y)]$ και $\mathrm{E}[\mathrm{E}(X|Y)]$ σημειώνονται ως $\mathrm{E}_Y[\mathrm{E}_X(g(X, Y)|Y)]$ και $\mathrm{E}_Y[\mathrm{E}_X(X|Y)]$, αντίστοιχα, έτσι ώστε να δηλώνεται ξεκάθαρα ως προς ποια μεταβλητή ορίζεται η κάθε μέση τιμή.

Σημαντική συνέπεια του Θεωρήματος 6.1 είναι η σχέση:

$$Var(X) = \mathrm{E}_Y[Var_X(X|Y)] + Var_Y[E_Y(X|Y)]$$

για τη διασπορά της τυχαίας μεταβλητής X εκφρασμένη με τη βοήθεια των υπό συνθήκη κατανομών. Η απόδειξη του παραπάνω αποτελέσματος αφήνεται ως άσκηση για τον/την αναγνώστη/στρια, ενώ παραπέμπουμε μεταξύ άλλων, στο σύγγραμμα Παπαϊωάννου (1997).

6.6.3 Συνδιασπορά και συσχέτιση

Αντικείμενο μελέτης αυτής της ενότητας είναι οι έννοιες της συνδιακύμανσης - ή αλλιώς συνδιασποράς - και της συσχέτισης δύο τυχαίων μεταβλητών X και Y.

Έστω δύο τυχαίες μεταβλητές X και Y, τότε είναι γνωστό ότι η διασπορά ή διακύμανση κάθε τυχαίας μεταβλητής αποτελεί ένα μέτρο της μεταβλητότητάς της γύρω από τη μέση τιμή της, χωρίς να μας δίνει καμία πληροφορία για το πώς μεταβάλλεται η μία σε σχέση με την άλλη. Κάτι τέτοιο επιτυγχάνεται με την έννοια της συνδιασποράς ή συνδιακύμανσης που ορίζεται στη συνέχεια.

Ορισμός 6.9

Έστω X και Y δύο τυχαίες μεταβλητές με πεπερασμένες διακυμάνσεις. Τότε η ποσότητα

$$\begin{aligned} Cov(X, Y) &= E\left[(X - E(X))(Y - E(Y))\right] \\ &= E(XY) - E(X)E(Y). \end{aligned}$$

ονομάζεται **συνδιασπορά** ή αλλιώς **συνδιακύμανση** των X και Y.

Στη συνέχεια, παρατίθεται μια σειρά από ιδιότητες της συνδιασποράς, οι οποίες προκύπτουν άμεσα από από τον ορισμό της. Η απόδειξή τους αφήνεται ως άσκηση στον/στην αναγνώστη/στρια.

- $Cov(X, Y) = Cov(Y, X)$.
- $Var(X) = Cov(X, X)$.
- $Cov(\alpha X + \gamma, \beta Y + \delta) = \alpha\beta Cov(X, Y)$.
- $Var(\alpha X + \beta Y) = \alpha^2 Var(X) + \beta^2 Var(Y) + 2\alpha\beta Cov(X, Y)$.

Παράδειγμα 6.6

Έστω X και Y οι τυχαίες μεταβλητές που παριστάνουν το ποσοστό των εντολών που εξυπηρετεί καθένας από τους δύο πυρήνες ενός διπύρηνου επεξεργαστή. Η από κοινού συνάρτηση πυκνότητας πιθανότητας των X και Y δίνεται από τη σχέση:

$$f_{X,Y}(x,y) = \begin{cases} 4(x+y^2), & x,y \geq 0, x+y \leq 1, \\ 0, & \text{αλλού.} \end{cases}$$

Να υπολογιστεί η συνδιασπορά $Cov(X,Y)$.

Λύση Παραδείγματος 6.6

Για να υπολογίσουμε τη συνδιασπορά $Cov(X,Y)$, αρκεί να υπολογίσουμε τις $E(XY)$, $E(X)$ και $E(Y)$, οι οποίες δίνονται από τις σχέσεις

$$E(XY) = \int_0^1 \int_0^{1-x} xy f_{X,Y}(x,y) dy dx,$$
$$E(X) = \int_0^1 \int_0^{1-x} x f_{X,Y}(x,y) dy dx,$$

και

$$E(Y) = \int_0^1 \int_0^{1-x} y f_{X,Y}(x,y) dy dx,$$

αντίστοιχα. Από τα παραπάνω ολοκληρώματα έχουμε μετά από αλγεβρικές πράξεις (αφήνονται για εξάσκηση στον/στην αναγνώστη/στρια), ότι $E(XY) = 0.1$, $E(X) = 0.4$ και $E(Y) = 11/30$. Επομένως, είναι:

$$Cov(X,Y) = E(XY) - E(X)E(Y) = 0.1 - 0.4 \cdot \frac{11}{30} = -\frac{7}{150}.$$

Παρατήρηση 6.7

Στην περίπτωση που $X = (X_1, \ldots, X_k)^t$ είναι ένα k-διάστατο τυχαίο διάνυσμα με $\Sigma = Cov(X)$ συμβολίζουμε τον **πίνακα διακυμάνσεων-συνδιακυμάνσεών** του, ο οποίος είναι ένας συμμετρικός $k \times k$ πίνακας με το (i,ℓ) στοιχείο του να ισούται με $Cov(X_i, X_\ell)$, $i,\ell = 1, \ldots, k$, $i \neq \ell$, ενώ τα διαγώνια στοιχεία του είναι ίσα με $Cov(X_i, X_i) = Var(X_i)$, $i = 1, \ldots, k$.

Η συνδιακύμανση είναι ένα μέτρο με το οποίο μετράμε την αλληλεπίδραση δύο μεταβλητών ποσοτικοποιώντας τον τρόπο που αυτές συμμεταβάλλονται. Σημειώνεται ότι:

- θετικές τιμές της συνδιασποράς προκύπτουν όταν έχουν μεγαλύτερη πιθανότητα να παρατηρηθούν ζεύγη παρατηρήσεων που διαφέρουν από τον μέσο όρο προς την ίδια κατεύθυνση, ενώ
- αρνητικές τιμές της συνδιασποράς προκύπτουν όταν είναι πιο πιθανό να παρατηρηθούν ζεύγη τα οποία διαφέρουν από τον μέσο όρο τους σε αντίθετες κατευθύνσεις.

Ωστόσο παρατηρήστε ότι η συνδιασπορά έχει ως μονάδα μέτρησης το γινόμενο των μονάδων μέτρησης κάθε μεταβλητής. Επομένως, για τα ίδια δεδομένα που αφορούν για παράδειγμα το ύψος (X) και το βάρος (Y), αν δύο άτομα χρησιμοποιήσουν ως μονάδες μέτρησης το μέτρο και τα κιλά και τα εκατοστά και τα γραμμάρια, αντίστοιχα, θα οδηγηθούν σε διαφορετικά αποτελέσματα. Για να ξεπεραστεί το παραπάνω μειονέκτημα του μέτρου της συνδιασποράς, έχει παρουσιαστεί στη βιβλιογραφία ένα άλλο μέτρο της σχέσης μεταξύ δύο μεταβλητών. Το μέτρο αυτό είναι ο συντελεστής συσχέτισης που ορίζεται ως εξής.

Ορισμός 6.10

Έστω X και Y δύο τυχαίες μεταβλητές με πεπερασμένες διακυμάνσεις. Τότε η ποσότητα

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}.$$

ονομάζεται **συντελεστής συσχέτισης** των X και Y.

Στο επόμενο θεώρημα θα αποδείξουμε ότι ο συντελεστής συσχέτισης είναι καθαρός αριθμός (αριθμός απαλλαγμένος από μονάδες μέτρησης) με συγκεκριμένο εύρος τιμών, ενώ θα προσδιοριστεί πότε λαμβάνει τη μέγιστη και πότε την ελάχιστη τιμή του.

Θεώρημα 6.2

Έστω X και Y δύο τυχαίες μεταβλητές με συντελεστή συσχέτισης $\rho(X,Y)$. Τότε ισχύουν τα ακόλουθα:

1. Ο συντελεστής συσχέτισης είναι καθαρός αριθμός και ικανοποιεί την ανίσωση
$$-1 \leq \rho(X,Y) \leq 1.$$
2. Αν $\rho(X,Y) = 1$, τότε υπάρχουν πραγματικές σταθερές $a > 0$ και $b \in \mathbb{R}$ τέτοιες ώστε $Y = aX + b$ και αντίστροφα.
3. Αν $\rho(X,Y) = -1$, τότε υπάρχουν πραγματικές σταθερές $a < 0$ και $b \in \mathbb{R}$ τέτοιες ώστε $Y = aX + b$ και αντίστροφα.

Απόδειξη Θεωρήματος 6.2

1. Αρχικά παρατηρούμε ότι οι μονάδες μέτρησης της συνδιασποράς είναι το γινόμενο των μονάδων μέτρησης των X και Y. Όμως και οι τετραγωνικές ρίζες των διασπορών, δηλαδή οι τυπικές αποκλίσεις, έχουν τις ίδιες μονάδες μέτρησης με τις αντίστοιχες τυχαίες μεταβλητές. Επομένως, ο παρανομαστής του συντελεστή συσχέτισης έχει τις ίδιες μονάδες με τον αριθμητή, δηλαδή με τη συνδιασπορά. Τα παραπάνω αιτιολογούν πλήρως το γεγονός ότι ο συντελεστής συσχέτισης είναι ένας καθαρός αριθμός.

   Για την απόδειξη της ανίσωσης $-1 \leq \rho(X,Y) \leq 1$ χρειάζεται να παρατηρήσουμε ότι η συνάρτηση $G(t) = Var(tX - Y)$ είναι αυστηρά μη αρνητική και μπορεί να εκφραστεί ως

$$\begin{aligned} G(t) &= Var(tX - Y) = Var(tX) + Var(Y) - 2tCov(X,Y) \\ &= t^2 Var(X) + Var(Y) - 2tCov(X,Y) \\ &= Var(X)\left(t - \frac{Cov(X,Y)}{Var(X)}\right)^2 + Var(Y) - \frac{(Cov(X,Y))^2}{Var(X)}. \end{aligned}$$

   Επειδή όμως η $G(t)$ είναι μη αρνητική για κάθε t, άρα και για $t = \frac{Cov(X,Y)}{Var(X)}$ έχουμε ότι πρέπει να ισχύει

$$\frac{(Cov(X,Y))^2}{Var(X)} \leq Var(Y),$$

   δηλαδή ότι

$$\frac{(Cov(X,Y))^2}{Var(X)Var(Y)} = \left(\rho(X,Y)\right)^2 \leq 1.$$

   Από την παραπάνω σχέση προκύπτει ότι:

$$-1 \leq \rho(X,Y) \leq 1.$$

2. Αν έχουμε ότι $\rho(X,Y)=1$, τότε $Cov(X,Y)=\sqrt{Var(X)}\sqrt{Var(Y)}=\sigma_X\sigma_Y$. Από το πρώτο μέρος της απόδειξης έχουμε ότι για $t=\frac{Cov(X,Y)}{Var(X)}$ είναι $G(t)=Var(tX-Y)=0$. Επομένως, καθώς $Var\left(\frac{Cov(X,Y)}{Var(X)}X-Y\right)=0$, προκύπτει ότι

$$\frac{Cov(X,Y)}{Var(X)}X-Y=c,$$

όπου $c\in\mathbb{R}$ σταθερά. Ύστερα από λίγη άλγεβρα (λαμβάνοντας υπόψη ότι, όταν $\rho(X,Y)=1$, είναι $Cov(X,Y)=\sigma_X\sigma_Y$) προκύπτει ότι:

$$X=\frac{Var(X)}{Cov(X,Y)}Y+c\frac{Var(X)}{Cov(X,Y)}=\frac{\sigma_x}{\sigma_y}X+c\frac{\sigma_x}{\sigma_y}=aX+b,$$

με $a=\frac{\sigma_x}{\sigma_y}>0$ και $b=c\frac{\sigma_x}{\sigma_y}\in\mathbb{R}$.

Αντίστροφα, αν $X=aY+b$ με $a>0$ είναι $Var(X)=a^2Var(Y)$ με $\sqrt{Var(X)}=a\sqrt{Var(Y)}$, και:

$$Cov(X,Y)=Cov(aY+b,Y)=aCov(Y,Y)=aVar(Y).$$

Επομένως, προκύπτει άμεσα ότι $\rho(X,Y)=1$.

3. Αν έχουμε ότι $\rho(X,Y)=-1$, τότε $Cov(X,Y)=-\sqrt{Var(X)}\sqrt{Var(Y)}=-\sigma_X\sigma_Y$. Από το πρώτο μέρος της απόδειξης έχουμε ότι για $t=\frac{Cov(X,Y)}{Var(X)}$ είναι $G(t)=Var(tX-Y)=0$. Επομένως, καθώς $Var\left(\frac{Cov(X,Y)}{Var(X)}X-Y\right)=0$ προκύπτει ότι

$$\frac{Cov(X,Y)}{Var(X)}X-Y=c,$$

όπου $c\in\mathbb{R}$ σταθερά. Ύστερα από λίγη άλγεβρα (λαμβάνοντας υπόψη ότι, όταν $\rho(X,Y)=-1$ είναι $Cov(X,Y)=-\sigma_X\sigma_Y$) προκύπτει ότι:

$$X=\frac{Var(X)}{Cov(X,Y)}Y+c\frac{Var(X)}{Cov(X,Y)}=-\frac{\sigma_x}{\sigma_y}X-c\frac{\sigma_x}{\sigma_y}=aX+b,$$

με $a=-\frac{\sigma_x}{\sigma_y}<0$ και $b=-c\frac{\sigma_x}{\sigma_y}\in\mathbb{R}$.

Αντίστροφα, αν $X=aY+b$ με $a<0$, είναι $Var(X)=a^2Var(Y)$ με $\sqrt{Var(X)}=-a\sqrt{Var(Y)}$ και:

$$Cov(X,Y)=Cov(aY+b,Y)=aCov(Y,Y)=aVar(Y).$$

Επομένως, άμεσα προκύπτει ότι $\rho(X,Y)=-1$.

Παρατήρηση 6.8

Από το παραπάνω θεώρημα θα πρέπει να συνειδητοποιήσουμε ότι ο συντελεστής συσχέτισης είναι στην πραγματικότητα ένα μέτρο της γραμμικής εξάρτησης δύο τυχαίων μεταβλητών και δεν μπορεί να αποκαλύψει πιθανές μη γραμμικές εξαρτήσεις τους. Σημειώνεται ότι τιμές κοντά στο 1 σημαίνουν ισχυρή θετική γραμμική συσχέτιση, τιμές κοντά στο -1 σημαίνουν ισχυρή αρνητική γραμμική συσχέτιση, ενώ τιμές κοντά στο 0 σημαίνουν ότι οι μεταβλητές είναι γραμμικά ασυσχέτιστες. Από την άλλη, αν δύο μεταβλητές είναι ασυσχέτιστες μεταξύ τους ($\rho(X,Y)=0$), δεν συνεπάγεται ότι δεν υπάρχει κάποιου άλλου είδους σχέση μεταξύ τους (βλ. το επόμενο παράδειγμα).

Παράδειγμα 6.7

Έστω $X \sim N(0,1)$ και $Y = X^2$. Να υπολογιστεί η συνδιασπορά των X και Y.

Λύση Παραδείγματος 6.7

Η συνδιασπορά των X και Y δίνεται από τη σχέση

$$Cov(X,Y) = E(XY) - E(X)E(Y).$$

Επειδή $Y = X^2$, η παραπάνω σχέση εκφράζεται ως

$$\begin{aligned} Cov(X,Y) &= E(X\,X^2) - E(X)E(X^2) \\ &= E(X^3) - E(X)E(X^2). \end{aligned}$$

Επειδή όμως $X \sim N(0,1)$, έχουμε, από τη σχέση (5.38) της Πρότασης 5.15, ότι οι ροπές περιττής τάξης είναι ίσες με μηδέν. Είναι δηλαδή $E(X^3) = E(X) = 0$ και, επομένως, προκύπτει άμεσα ότι $Cov(X,Y) = 0$. Παρατηρούμε ότι, ενώ οι $X \sim N(0,1)$ και $Y = X^2$ συνδέονται μεταξύ τους, έχουμε ότι είναι ασυσχέτιστες. Αυτό οφείλεται στο γεγονός ότι η Y δεν είναι γραμμικά συσχετισμένη με τη X.

Παράδειγμα 6.8

Έστω X και Y τ.μ που ακολουθούν την τυπική κανονική κατανομή. Αν $Cov(X,Y) = -0.81$, τότε ποια από τις επόμενες απαντήσεις είναι σωστή;

1. $\rho(X,Y) = 0.81$.
2. $\rho(X,Y) = -0.81$.
3. $\rho(X,Y) = 0.9$.
4. Δεν έχουμε αρκετά στοιχεία για να υπολογίσουμε το $\rho(X,Y)$.

Λύση Παραδείγματος 6.8

Καθώς οι τ.μ. X και Y ακολουθούν την τυπική κανονική κατανομή έχουμε ότι $\sigma_X = \sigma_Y = 1$. Επομένως, καθώς $Cov(X,Y) = -0.81$ είναι

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sigma_X \cdot \sigma_Y} = \frac{-0.81}{1 \cdot 1} = -0.81.$$

Επομένως, σωστή απάντηση είναι η δεύτερη.

Παράδειγμα 6.9

Για τις από κοινού κατανεμημένες τυχαίες μεταβλητές X, Y ισχύει $E(X \cdot Y) = E(X) \cdot E(Y)$. Τότε, ποια από τις παρακάτω προτάσεις ΔΕΝ ΙΣΧΥΕΙ;

1. Η συνδιασπορά των τ.μ. X, Y ισούται $Cov(X,Y) = 0$.
2. $Var(3X + 6Y) = 9 \cdot Var(X) + 36 \cdot Var(Y)$.
3. Οι τ.μ. X και Y συνδέονται με μια γραμμική σχέση.
4. Ο συντελεστής συσχέτισης $\rho(X,Y) = 0$.

Λύση Παραδείγματος 6.9

Δεν ισχύει η τρίτη από τις παραπάνω προτάσεις, καθώς στην περίπτωση αυτή ο συντελεστής συσχέτισης είναι

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} = \frac{E(X\cdot Y) - E(X)\cdot E(Y)}{\sqrt{Var(X)Var(Y)}} = 0.$$

Επομένως, ισχύει η τέταρτη πρόταση, ενώ το γεγονός ότι ο συντελεστής συσχέτισης είναι ίσος με μηδέν σημαίνει ότι οι τ.μ. ΔΕΝ συνδέονται με μια γραμμική σχέση. Ακόμα, από τον ορισμό της συνδιακύμανσης έχουμε ότι $Cov(X,Y) = E(X\cdot Y) - E(X)\cdot E(Y) = 0$, οπότε η πρόταση 1 ισχύει. Τέλος, $Var(3X + 6Y) = 9\cdot Var(X) + 36\cdot Var(Y) + 2\cdot 3\cdot 5\cdot Cov(X,Y) = 9\cdot Var(X) + 36\cdot Var(Y)$, άρα ισχύει και η δεύτερη πρόταση.

Καθώς ο συντελεστής συσχέτισης μπορεί να ανακαλύψει την ύπαρξη ή μη μόνο γραμμικών σχέσεων μεταξύ των μεταβλητών, αντικείμενο μελέτης της επόμενης ενότητας αποτελεί η γενικότερη έννοια της στοχαστικής ανεξαρτησίας των τυχαίων μεταβλητών.

6.7 Ανεξαρτησία τυχαίων μεταβλητών

Η έννοια της στοχαστικής ανεξαρτησίας τυχαίων μεταβλητών είναι μια από τις σημαντικότερες της Θεωρίας Πιθανοτήτων και αποτελεί φυσική επέκταση της έννοιας της ανεξαρτησίας ενδεχομένων. Στη συνέχεια, δίνεται ο ορισμός της στοχαστικής ανεξαρτησίας k το πλήθος με $k \geq 2$, τυχαίων μεταβλητών.

Ορισμός 6.11

Οι τυχαίες μεταβλητές $X_1,\ldots,X_k$, $k \geq 2$, λέμε ότι είναι στοχαστικά ανεξάρτητες αν για οποιαδήποτε συλλογή από (Borel) υποσύνολα των πραγματικών αριθμών $B_1,\ldots,B_k$ ισχύει ότι:

$$P(X_1 \in B_1,\ldots,X_k \in B_k) = \prod_{i=1}^{k} P(X_i \in B_i). \tag{6.14}$$

Από τον παραπάνω ορισμό άμεσα προκύπτουν οι ακόλουθες προτάσεις.

Πρόταση 6.2

Αν οι τυχαίες μεταβλητές $X_1,\ldots,X_k$, $k \geq 2$, είναι στοχαστικά ανεξάρτητες, τότε οποιεσδήποτε m από αυτές ($2 \leq m \leq k$) είναι στοχαστικά ανεξάρτητες.

Απόδειξη Πρότασης 6.2

Χωρίς βλάβη της γενικότητας θεωρούμε τις τ.μ. $X_1,\ldots,X_m$, $2 \leq m \leq k$. Τότε για οποιαδήποτε συλλογή από (Borel) υποσύνολα των πραγματικών αριθμών $B_1,\ldots,B_m$ ισχύει ότι

$$\begin{aligned} P(X_1 \in B_1,\ldots,X_m \in B_m) &= P(X_1 \in B_1,\ldots,X_m \in B_m, X_{m+1} \in \mathbb{R},\ldots X_k \in \mathbb{R}) \\ &= P(X_1 \in B_1)\cdots P(X_m \in B_m)\cdot P(X_{m+1} \in \mathbb{R})\cdots P(X_k \in \mathbb{R}) \\ &= P(X_1 \in B_1)\cdots P(X_m \in B_m)\cdot 1\cdots 1 \\ &= \prod_{i=1}^{m} P(X_i \in B_i), \end{aligned}$$

όπου στην τελευταία σχέση χρησιμοποιήθηκε η ανεξαρτησία των $X_1,\ldots,X_k$.

Πρόταση 6.3

Έστω $X_1,\ldots,X_k$, $k \geq 2$, ανεξάρτητες τυχαίες μεταβλητές είτε διακριτές είτε συνεχείς. Επιπλέον, έστω k το πλήθος πραγματικές συναρτήσεις $g_i : \mathbb{R} \to \mathbb{R}$, $i = 1, \ldots, k$. Τότε οι τυχαίες μεταβλητές $Y_1,\ldots,Y_k$, $k \geq 2$ με $Y_i = g_i(X_i)$, $i = 1, \ldots, k$, είναι ανεξάρτητες.

Απόδειξη Πρότασης 6.3

Σύμφωνα με τον ορισμό της ανεξαρτησίας αρκεί να αποδείξουμε ότι για οποιαδήποτε συλλογή από (Borel) υποσύνολα των πραγματικών αριθμών $B_1,\ldots,B_k$:

$$P(Y_1 \in B_1, \ldots, Y_k \in B_k) = \prod_{i=1}^{k} P(Y_i \in B_i). \tag{6.15}$$

Θεωρούμε τις αντίστροφες εικόνες των συνόλων $B_1,\ldots,B_k$ με τις συναρτήσεις $g_1,\ldots,g_k$. Τότε τα ενδεχόμενα $\{Y_i \in B_i\}$ και $\{X_i \in g_i^{-1}(B_i)\}$ είναι ισοδύναμα. Επιπλέον, λόγω της ανεξαρτησίας των X_1, …, X_k, έχουμε ότι:

$$\begin{aligned} P(Y_1 \in B_1, \ldots, Y_k \in B_k) &= P(X_1 \in g_1^{-1}(B_1), \ldots, X_k \in g_k^{-1}(B_k) \\ &= \prod_{i=1}^{k} P(X_i \in g_i^{-1}(B_i)) = \prod_{i=1}^{k} P(Y_i \in B_i). \end{aligned}$$

Στην πράξη κάποιες φορές αντιλαμβανόμαστε άμεσα την ανεξαρτησία των τυχαίων μεταβλητών που μας ενδιαφέρουν όταν αυτές αναφέρονται σε ανεξάρτητα πειράματα τύχης. Για παράδειγμα, οι τυχαίες μεταβλητές $X_1,\ldots,X_k$ που παριστάνουν τη μέτρηση χοληστερίνης k ατόμων ενός πληθυσμού, θεωρούνται ανεξάρτητες. Ωστόσο, στην πλειονότητα των περιπτώσεων, η ανεξαρτησία των τυχαίων μεταβλητών θα πρέπει να διαπιστωθεί. Η χρήση του ορισμού για την εξέταση της στοχαστικής ανεξαρτησίας k το πλήθος τ.μ. εμπεριέχει την εξέταση της ικανοποίησης ή όχι της σχέσης (6.14) για οποιαδήποτε συλλογή υποσυνόλων. Αυτό καθιστά αδύνατη τη χρήση του στην πράξη. Για να ξεπεραστεί αυτό το πρόβλημα έχουν εμφανιστεί στη βιβλιογραφία τα λεγόμενα κριτήρια ανεξαρτησίας.

Πρόταση 6.4: 1ο κριτήριο ανεξαρτησίας.

Έστω $X_1,\ldots,X_k$, $k \geq 2$, τυχαίες μεταβλητές με από κοινού συνάρτηση κατανομής $F_{X_1,\ldots,X_k}(x_1, \ldots, x_k)$ και περιθώριες συναρτήσεις κατανομής $F_{X_1}(x_1),\ldots,F_{X_k}(x_k)$, αντίστοιχα. Τότε οι $X_1,\ldots,X_k$ είναι ανεξάρτητες αν και μόνο αν

$$F_{X_1,\ldots,X_k}(x_1, \ldots, x_k) = \prod_{i=1}^{k} F_{X_i}(x_i), \tag{6.16}$$

για κάθε $(x_1, \ldots, x_k) \in \mathbb{R}^k$.

Απόδειξη Πρότασης 6.4

Η απόδειξη του αντίστροφου της παραπάνω πρότασης απαιτεί γνώσεις Θεωρίας Μέτρου και για τον λόγο αυτό παραλείπεται, ενώ το ευθύ προκύπτει άμεσα από τον ορισμό της στοχαστικής ανεξαρτησίας.

Σύμφωνα με την προηγούμενη πρόταση, η απόδειξη της ανεξαρτησίας δύο ή περισσότερων τ.μ. γίνεται μέσω της εξέτασης αν ικανοποιείται ή όχι η παραπάνω συνθήκη που αφορά την από κοινού αθροιστική συνάρτηση κατανομής και τις αντίστοιχες περιθώριες ασκ. Εναλλακτικά, έχουμε το κριτήριο ανεξαρτησίας που δίνεται στην επόμενη πρόταση μέσω της από κοινού συνάρτησης (πυκνότητας) πιθανότητας και των αντίστοιχων περιθωρίων της.

Πρόταση 6.5: 2ο κριτήριο ανεξαρτησίας.

Έστω $X_1,\ldots,X_k$, $k \geq 2$, διακριτές (συνεχείς, αντίστοιχα) τυχαίες μεταβλητές με από κοινού συνάρτηση πιθανότητας (συνάρτηση πυκνότητας πιθανότητας, αντίστοιχα) $f_{X_1,\ldots,X_k}(x_1,\ldots,x_k)$ και περιθώριες συναρτήσεις πιθανότητας (συναρτήσεις πυκνότητας πιθανότητας) $f_{X_1}(x_1),\ldots,f_{X_k}(x_k)$, αντίστοιχα. Τότε οι $X_1,\ldots,X_k$ είναι ανεξάρτητες αν και μόνο αν

$$f_{X_1,\ldots,X_k}(x_1,\ldots,x_k) = \prod_{i=1}^{k} f_{X_i}(x_i), \tag{6.17}$$

για κάθε $(x_1,\ldots,x_k) \in S_{X_1,\ldots,X_k}$ στη διακριτή περίπτωση και για κάθε $(x_1,\ldots,x_k) \in \mathbb{R}^k$ στη συνεχή περίπτωση.

Απόδειξη Πρότασης 6.5

Συνεχής περίπτωση. Έστω ότι οι τ.μ. $X_1,\ldots,X_k$, $k \geq 2$ είναι ανεξάρτητες. Τότε, σύμφωνα με το 1ο κριτήριο ανεξαρτησίας, ισχύει ότι:

$$F_{X_1,\ldots,X_k}(x_1,\ldots,x_k) = \prod_{i=1}^{k} F_{X_i}(x_i),$$

για κάθε $(x_1,\ldots,x_k) \in \mathbb{R}^k$. Παραγωγίζοντας την ισότητα αυτή διαδοχικά ως προς x_1, $x_2,\ldots,x_k$ προκύπτει ότι:

$$f_{X_1,\ldots,X_k}(x_1,\ldots,x_k) = \prod_{i=1}^{k} f_{X_i}(x_i).$$

Για την απόδειξη του αντιστρόφου υποθέτουμε ότι ισχύει η σχέση $f_{X_1,\ldots,X_k}(x_1,\ldots,x_k) = \prod_{i=1}^{k} f_{X_i}(x_i)$ και θα αποδείξουμε ότι ισχύει η σχέση του 1ου κριτηρίου ανεξαρτησίας. Είναι

$$\begin{aligned} F_{X_1,\ldots,X_k}(x_1,\ldots,x_k) &= \int_{-\infty}^{x_k} \cdots \int_{-\infty}^{x_1} f_{X_1,\ldots,X_k}(u_1,\ldots,u_k)du_1\cdots du_k \\ &= \int_{-\infty}^{x_k} \cdots \int_{-\infty}^{x_1} f_{X_1}(u_1)\cdots f_{X_k}(u_k)du_1\cdots du_k \\ &= F_{X_1}(x_1)\cdots F_{X_k}(x_k) \end{aligned}$$

που αποδεικνύει το ζητούμενο.

Διακριτή περίπτωση. Έστω ότι οι τ.μ. $X_1,\ldots,X_k$, $k \geq 2$ είναι ανεξάρτητες. Τότε, σύμφωνα με τον ορισμό της ανεξαρτησίας των τ.μ., ισχύει ότι για οποιαδήποτε συλλογή από (Borel) υποσύνολα των πραγματικών αριθμών $B_1,\ldots,B_k$:

$$P(X_1 \in B_1,\ldots,X_k \in B_k) = \prod_{i=1}^{k} P(X_i \in B_i). \tag{6.18}$$

Εφαρμόζουμε την παραπάνω σχέση για $B_i = \{x_i\}$, $i = 1,\ldots,k$ και, επομένως, προκύπτει ότι:

$$f_{X_1,\ldots,X_k}(x_1,\ldots,x_k) = \prod_{i=1}^{k} f_{X_i}(x_i).$$

Για την απόδειξη του αντιστρόφου υποθέτουμε ότι ισχύει η σχέση $f_{X_1,\dots,X_k}(x_1,\dots,x_k)=\prod_{i=1}^{k} f_{X_i}(x_i)$. Τότε για οποιαδήποτε συλλογή από (Borel) υποσύνολα των πραγματικών αριθμών $B_1,\dots,B_k$ είναι:

$$\begin{aligned} P(X_1\in B_1,\dots,X_k\in B_k) &= \sum_{x_1\in B_1}\cdots\sum_{x_k\in B_k} f_{X_1,\dots,X_k}(x_1,\dots,x_k)\\ &= \sum_{x_1\in B_1}\cdots\sum_{x_k\in B_k} f_{X_1}(x_1)\cdots f_{x_k}(x_k)\\ &= \left(\sum_{x_1\in B_1} f_{X_1}(x_1)\right)\cdots\left(\sum_{x_k\in B_k} f_{X_k}(x_k)\right)\\ &= \prod_{i=1}^{k} P(X_i\in B_i), \end{aligned}$$

που αποδεικνύει το ζητούμενο.

Άσκηση Αυτοαξιολόγησης 6.8

Η από κοινού συνάρτηση πυκνότητας πιθανότητας του χρόνου X που απαιτείται για την παραλαβή των ανταλλακτικών ενός μηχανήματος και του χρόνου Y επισκευής του μηχανήματος δίνεται από τη σχέση:

$$f_{X,Y}(x,y)=\begin{cases} \frac{1}{4}xye^{-\frac{x}{2}-y}, & x,y>0,\\ 0, & \text{αλλού.}\end{cases}$$

Να εξετάσετε αν ο χρόνος επισκευής του μηχανήματος είναι ανεξάρτητος από τον χρόνο παραλαβής των ανταλλακτικών.

Με βάση τα παραπάνω κριτήρια ανεξαρτησίας προκύπτουν ορισμένες χρήσιμες ιδιότητες που ισχύουν για ανεξάρτητες τυχαίες μεταβλητές.

Πρόταση 6.6

Έστω $X_1,\dots,X_k$, $k\geq 2$, ανεξάρτητες τυχαίες μεταβλητές είτε διακριτές είτε συνεχείς. Τότε, οι δεσμευμένες κατανομές m το πλήθος τυχαίων μεταβλητών από τις k συνολικά τυχαίες μεταβλητές ($2\leq m\leq k$), δοθέντος όλων των υπολοίπων ($k-m$ τυχαίων μεταβλητών) ή ενός υποσυνόλου τους, συμπίπτουν με τις αντίστοιχες μη δεσμευμένες κατανομές τους, δηλαδή τις περιθώριές τους.

Απόδειξη Πρότασης 6.6

Θεωρούμε χωρίς βλάβη της γενικότητας τις μεταβλητές $X_1,\dots,X_m$ με $2\leq m\leq k$. Θέλουμε να δείξουμε ότι:

$$f_{X_1,\dots,X_m|X_{m+1},\dots,X_k}(x_1,\dots,x_m|x_{m+1},\dots,x_k)=f_{X_1,\dots,X_m}(x_1,\dots,x_m).$$

Από τον ορισμό της δεσμευμένης κατανομής έχουμε:

$$f_{X_1,\dots X_m|X_{m+1},\dots,X_k}(x_1,\dots,x_m|x_{m+1},\dots,x_k)=\frac{f_{X_1,\dots,X_k}(x_1,\dots,x_k)}{f_{X_{m+1},\dots,X_k}(x_{m+1},\dots,x_k)}.$$

Λόγω της ανεξαρτησίας των τ.μ. $X_1,\dots,X_k$, σύμφωνα με το 2ο κριτήριο ανεξαρτησίας, ισχύει ότι:

$$f_{X_1,\dots,X_k}(x_1,\dots,x_k)=\prod_{i=1}^{k} f_{X_i}(x_i),$$

ενώ και κάθε υποσύνολο αυτών είναι ανεξάρτητες τ.μ. και, επομένως,

$$f_{X_{m+1},\ldots,X_k}(x_1,\ldots,x_k) = \prod_{i=m+1}^{k} f_{X_i}(x_i).$$

Συνδυάζοντας τα παραπάνω:

$$f_{X_1,\ldots,X_m|X_{m+1},\ldots,X_k}(x_1,\ldots,x_m|x_{m+1},\ldots,x_k) = \frac{\prod_{i=1}^{k} f_{X_i}(x_i)}{\prod_{i=m+1}^{k} f_{X_i}(x_i)},$$

που αποδεικνύει το ζητούμενο.

Άμεση συνέπεια της προηγούμενης πρότασης είναι το ακόλουθο πόρισμα.

Πόρισμα 6.2

Έστω $X_1,\ldots,X_k$, $k \geq 2$, ανεξάρτητες τυχαίες μεταβλητές είτε διακριτές είτε συνεχείς. Τότε η δεσμευμένη μέση τιμή της X_i, δοθέντος των τιμών κάποιων από τις X_j, $i \neq j$, $i,j = 1,\ldots,k$, είναι ίση με τη μέση τιμή της X_i, δηλαδή ισχύουν οι σχέσεις:

$$E(X_i|X_1 = x_1,\ldots,X_{i-1} = x_{i-1},X_{i+1} = x_{i+1},\ldots,X_k = x_k) = E(X_i),$$

και

$$E\left(X_i|X_j = x_j\right) = E(X_i).$$

Επιπρόσθετα, για οποιαδήποτε πραγματική συνάρτηση g, ισχύει ότι:

$$E\left(g(X_i)|X_1 = x_1,\ldots,X_{i-1} = x_{i-1},X_{i+1} = x_{i+1},\ldots,X_k = x_k\right) = E\left(g(X_i)\right),$$

και

$$E\left(g(X_i)|X_j = x_j\right) = E\left(g(X_i)\right).$$

Απόδειξη Πορίσματος 6.2

Δίνεται ότι οι τυχαίες μεταβλητές $X_1,\ldots,X_k$, $k \geq 2$, είναι ανεξάρτητες τυχαίες μεταβλητές. Από την Πρόταση 6.6 που προηγήθηκε έχουμε ότι η δεσμευμένη κατανομή της X_i, δοθέντος των τιμών κάποιων από τις υπόλοιπες, είναι ίση με τη μη δεσμευμένη κατανομή. Επομένως αφού η δεσμευμένη κατανομή της X_i, δοθέντος των υπολοίπων, ταυτίζεται με τη μη δεσμευμένη κατανομή, συνεπάγεται ότι ταυτίζονται οι μέσες τιμές τους. Τέλος, γνωρίζουμε ότι όταν οι τ.μ. είναι ανεξάρτητες το ίδιο ισχύει και για τις συναρτήσεις αυτών και με ανάλογο τρόπο αποδεικνύεται το δεύτερο σκέλος του πορίσματος.

Πρόταση 6.7

Έστω $X_1,\ldots,X_k$, $k \geq 2$, ανεξάρτητες τυχαίες μεταβλητές είτε διακριτές είτε συνεχείς. Τότε ισχύει ότι:

$$E(X_1X_2\cdots X_k) = \prod_{i=1}^{k} E(X_i) \tag{6.19}$$

με την προϋπόθεση ότι οι αναμενόμενες τιμές που εμφανίζονται στο δεξί μέλος υπάρχουν. Επιπρόσθετα,

έστω k το πλήθος πραγματικές συναρτήσεις $g_i : \mathbb{R} \to \mathbb{R}$, $i = 1, \dots, k$. Τότε

$$E\left(g_1(X_1)g_2(X_2)\cdots g_k(X_k)\right) = \prod_{i=1}^{k} E\left(g_i(X_i)\right).$$

Απόδειξη Πρότασης 6.7

Χωρίς βλάβη της γενικότητας θα αποδείξουμε την πρόταση στην περίπτωση που οι k το πλήθος τ.μ. είναι συνεχείς. Η απόδειξη για τη διακριτή περίπτωση είναι παρόμοια με την αντικατάσταση των ολοκληρωμάτων με τα αντίστοιχα αθροίσματα.
Έστω $f_{X_1,\dots,X_k}(x_1, \dots, x_k)$ η από κοινού συνάρτηση πυκνότητας πιθανότητας των $X_1,\dots,X_k$. Τότε ισχύει ότι:

$$E(X_1X_2\cdots X_k) = \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} x_1 \dots x_k \, f_{X_1,\dots,X_k}(x_1, \dots, x_k)dx_1 \cdots dx_k.$$

Σύμφωνα με το 2ο κριτήριο ανεξαρτησίας για κάθε $(x_1, \dots, x_k) \in S_{X_1,\dots,X_k}$, ισχύει ότι $f_{X_1,\dots,X_k}(x_1, \dots, x_k) = \prod_{i=1}^{k} f_{X_i}(x_i)$, όπου $f_{X_i}(x_i)$ οι περιθώριες συναρτήσεις πυκνότητας πιθανότητας των X_i, $i = 1, \dots, k$. Επομένως, είναι:

$$E(X_1X_2\cdots X_k) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} x_1 \dots x_k \, f_{X_1}(x_1)\cdots f_{X_k}(x_k)dx_1 \cdots dx_k = \prod_{i=1}^{k} E(X_i),$$

που αποδεικνύει το πρώτο μέρος της πρότασης. Το δεύτερο μέρος της πρότασης προκύπτει άμεσα συνδυάζοντας το αποτέλεσμα που μόλις αποδείχθηκε με την Πρόταση 6.3.

Από την παραπάνω πρόταση προκύπτει ότι, αν X και Y είναι δύο ανεξάρτητες μεταβλητές, τότε $Cov(X, Y) = 0$. Το αντίστροφο δεν ισχύει όπως φαίνεται και στα (αντι)παραδείγματα που ακολουθούν, εκτός και αν οι (X, Y) ακολουθούν τη διδιάστατη κανονική κατανομή, όπως θα δούμε σε επόμενη ενότητα.

Παράδειγμα 6.10

Έστω X μια εκφυλισμένη τυχαία μεταβλητή, δηλαδή μια τυχαία μεταβλητή που λαμβάνει την τιμή α ($\alpha \neq 0$) με πιθανότητα 1. Ορίζουμε την τυχαία μεταβλητή Y ως $Y = 1/X$. Να δείξετε ότι $E(XY) = E(X)E(Y)$. Τι παρατηρείτε;

Λύση Παραδείγματος 6.10

Για τις τυχαίες μεταβλητές X και Y έχουμε ότι

$$E(X) = \alpha,\ E(Y) = 1/\alpha \text{ και } E(XY) = E\left(X\frac{1}{X}\right) = E(1) = 1.$$

Άρα ισχύει η σχέση $E(XY) = E(X)E(Y)$, χωρίς προφανώς οι τ.μ. X και Y να είναι ανεξάρτητες.

Παράδειγμα 6.11

Δίνεται η τυχαία μεταβλητή X με συνάρτηση πιθανότητας

$$p_X(x) = \begin{cases} 1/4, & \text{για } x = -2, \\ 1/4, & \text{για } x = -1, \\ 1/4, & \text{για } x = 1, \\ 1/4, & \text{για } x = 2 \\ 0 & \text{αλλού .} \end{cases}$$

Δείξτε ότι για τις τ.μ. X και $Y = |X|$ ισχύει η $E(XY) = E(X)E(Y)$, χωρίς αυτές να είναι ανεξάρτητες.

Λύση Παραδείγματος 6.11

Είναι φανερό ότι η γνώση της τιμής της X καθορίζει πλήρως την τιμή της Y και, επομένως, δεν είναι ανεξάρτητες.
Η τυχαία μεταβλητή Y μπορεί να λάβει τις τιμές 1 και 2 και η συνάρτηση πιθανότητάς της δίνεται από τη σχέση:

$$p_Y(y) = \begin{cases} 1/2, & \text{για } y = 1, \\ 1/2, & \text{για } y = 2, \\ 0, & \text{αλλού,} \end{cases}$$

αφού η $p_Y(y) = P(Y = y) = P(|X| = y) = P(X = -y \cup X = y) = P(X = -y) + P(X = y) = 1/4 + 1/4 = 2/4$, για $y = 1,2$.
Είναι

$$E(Y) = 0.5 \cdot 1 + 0.5 \cdot 2 = 1.5 \text{ και } E(X) = 0.25 \cdot (-2) + 0.25 \cdot (-1) + 0.25 \cdot (2) + 0.25 \cdot (1) = 0.$$

Από την άλλη, η τ.μ. $Z = XY = X|X|$ έχει δυνατό σύνολο τιμών -4, -1, 1 και 4 με συνάρτηση πιθανότητας $P(Z = z) = 0.25$, για $z = -4,1,1,4$. Επομένως, είναι:

$$E(Z) = 0.25 \cdot (-4) + 0.25 \cdot (-1) + 0.25 \cdot (1) + 0.25 \cdot (4) = 0.$$

Έχουμε λοιπόν ότι ισχύει η

$$\mathrm{E}(XY) = \mathrm{E}(X)\mathrm{E}(Y),$$

αφού και τα δύο μέλη της ισότητας είναι ίσα με μηδέν.

Τέλος, στη βιβλιογραφία έχει εμφανιστεί και ένα ακόμη κριτήριο ανεξαρτησίας που αξιοποιεί το μονοσήμαντο της ροπογεννήτριας συνάρτησης. Το κριτήριο αυτό δίνεται στην πρόταση που ακολουθεί.

Πρόταση 6.8: 3ο κριτήριο ανεξαρτησίας

Οι τυχαίες μεταβλητές $X_1,\ldots,X_k$, $k \geq 2$, είναι ανεξάρτητες αν και μόνο αν για κάθε $(t_1,\ldots,t_k) \in (-h_1,h_1) \times \ldots \times (-h_k,h_k)$ με $h_1 > 0,\ldots,h_k > 0$, ισχύει ότι:

$$M_{X_1,\ldots,X_k}(t_1,\ldots,t_k) = \prod_{i=1}^{k} M_{X_i}(t_i), \tag{6.20}$$

Απόδειξη Πρότασης 6.8

Έστω $X_1,\ldots,X_k$, $k \geq 2$, είναι ανεξάρτητες τυχαίες μεταβλητές. Από τον ορισμό της ροπογεννήτριας συνάρτησης και από τις ιδιότητες της εκθετικής συνάρτησης έχουμε ότι

$$M_{X_1,\ldots,X_k}(t_1,\ldots,t_k) = E\left(e^{t_1X_1+\cdots+t_kX_k}\right) = E\left(e^{t_1X_1}\cdots e^{t_kX_k}\right).$$

Επομένως, είναι:

$$M_{X_1,\ldots,X_k}(t_1,\ldots,t_k) = E\left(g_1(X_1)\cdots g_k(X_k)\right),$$

με $g_i(X_i) = e^{t_iX_i}$, $i = 1,\ldots,k$. Λόγω της ανεξαρτησίας των $X_1,\ldots,X_k$ είναι

$$E\left(g_1(X_1)\cdots g_k(X_k)\right) = \prod_{i=1}^{k} E\left(g_i(X_i)\right) = \prod_{i=1}^{k} M_{X_i}(t_i),$$

και το ευθύ του κριτηρίου αποδείχθηκε.

Για το αντίστροφο, χωρίς βλάβη της γενικότητας, θεωρούμε τη συνεχή περίπτωση. Έστω ότι για κάθε $(t_1,\ldots,t_k) \in (-h_1,h_1)\times\ldots\times(-h_k,h_k)$ με $h_1 > 0,\ldots,h_k > 0$, ισχύει ότι:

$$M_{X_1,\ldots,X_k}(t_1,\ldots,t_k) = \prod_{i=1}^{k} M_{X_i}(t_i).$$

Ωστόσο

$$\begin{aligned} M_{X_1,\ldots,X_k}(t_1,\ldots,t_k) &= \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} e^{t_1x_1+\cdots+t_kx_k} f_{X_1,\ldots,X_k}(x_1,\ldots,x_k)dx_1\cdots dx_k \\ &= \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} e^{t_1x_1}\cdots e^{t_kx_k} f_{X_1,\ldots,X_k}(x_1,\ldots,x_k)dx_1\cdots dx_k \end{aligned}$$

και

$$\begin{aligned} \prod_{i=1}^{k} M_{X_i}(t_i) &= \prod_{i=1}^{k} e^{t_ix_i} f_{X_i}(x_i)dx_i \\ &= \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} e^{t_1x_1}\cdots e^{t_kx_k}\left(\prod_{i=1}^{k} f_{X_i}(x_i)\right)dx_1\cdots dx_k. \end{aligned}$$

Συγκρίνοντας τις δύο παραπάνω σχέσεις και από το θεώρημα του μονοσήμαντου των ροπογεννητριών προκύπτει ότι $f_{X_1,\ldots,X_k}(x_1,\ldots,x_k) = \prod_{i=1}^{k} f_{X_i}(x_i)$. Επομένως, χρησιμοποιώντας το 2ο κριτήριο ανεξαρτησίας, έχουμε ότι $X_1,\ldots,X_k$, $k \geq 2$, είναι ανεξάρτητες τυχαίες μεταβλητές.

Άσκηση Αυτοαξιολόγησης 6.9

Έστω X και Y δύο συνεχείς τυχαίες μεταβλητές με από κοινού συνάρτηση πυκνότητας πιθανότητας:

$$f_{X,Y}(x,y) = \begin{cases} \frac{x}{5} + cy, & 0 < x < 1,\ 1 < y < 5, \\ 0, & \text{αλλού}, \end{cases}$$

όπου c κατάλληλη σταθερά. Εξετάστε αν οι τ.μ. X και Y είναι ανεξάρτητες.

6.8 Ειδικές πολυδιάστατες κατανομές

Κλείνοντας το κεφάλαιο των πολυδιάστατων τυχαίων μεταβλητών παρουσιάζονται δύο ειδικές πολυδιάστατες κατανομές: η πολυωνυμική και η διδιάστατη κανονική κατανομή, ενώ απλά δίνεται ο ορισμός της πολυδιάστατης κανονικής κατανομής. Για λεπτομέρειες για αυτές τις ειδικές περιπτώσεις πολυδιάστατων κατανομών και ιδιότητες αυτών παραπέμπουμε μεταξύ άλλων, στα συγγράμματα των Παπαϊωάννου (1997) και Κούτρας (2005).

6.8.1 Πολυωνυμική κατανομή

Η πολυωνυμική κατανομή αποτελεί γενίκευση της διωνυμικής κατανομής και χρησιμοποιείται για την περιγραφή των αποτελεσμάτων ενός πειράματος τύχης με k το πλήθος δυνατά αποτελέσματα. Ειδικότερα, ας θεωρήσουμε ένα τυχαίο πείραμα σε κάθε επανάληψη του οποίου μπορεί να εμφανιστεί ένα μόνο από τα k το πλήθος δυνατά αποτελέσματα, έστω $\mathrm{A}_1,\ldots,A_k$. Επιπρόσθετα, υποθέτουμε ότι η πιθανότητα εμφάνισης κάθε αποτελέσματος είναι ίση με $p_i = P(A_i)$, $i = 1,\ldots,k$ με $\sum_{i=1}^{k} p_i = 1$ και οι πιθανότητες αυτές παραμένουν σταθερές κατά τη διάρκεια των n το πλήθος επαναλήψεων του πειράματος. Τέλος, θεωρούμε ότι οι επαναλήψεις είναι ανεξάρτητες μεταξύ τους με την έννοια ότι το αποτέλεσμα οποιασδήποτε επανάληψης δεν επηρεάζει το αποτέλεσμα κάποιας άλλης. Το παραπάνω τυχαίο πείραμα θα το λέμε στη συνέχεια πολυωνυμικό τυχαίο πείραμα. Στο πλαίσιο αυτό η πολυωνυμική κατανομή ορίζεται ως ακολούθως.

Ορισμός 6.12

Έστω $X_1, X_2, \ldots, X_{k-1}$ οι τυχαίες μεταβλητές που παριστάνουν το πλήθος των φορών που εμφανίζεται στις n ανεξάρτητες επαναλήψεις ενός πολυωνυμικού τυχαίου πειράματος το A_1, A_2,…, A_{k-1}, αποτέλεσμα, αντίστοιχα. Το τυχαίο διάνυσμα $(X_1,\ldots,X_{k-1})^t$ ή, ισοδύναμα, το τυχαίο διάνυσμα $(X_1,\ldots,X_{k-1},X_k)^t$ θα λέμε ότι ακολουθεί την **πολυωνυμική κατανομή** με παραμέτρους n, $p_1,\ldots,p_{k-1}$ και από κοινού συνάρτηση πιθανότητας που δίνεται από τη σχέση:

$$p_{X_1,\ldots,X_{k-1}}(x_1,\ldots,x_{k-1}) = P(X_1 = x_1,\ldots,X_k = x_k) = \frac{n!}{x_1!\cdots x_k!} p_1^{x_1}\cdots p_k^{x_k},$$

όπου $\sum_{i=1}^{k} p_i = 1$, $\sum_{i=1}^{k} x_i = n$, $x_i \in \{0,1,\ldots,n\}$ και $0 \le p_i \le 1$, $i = 1,\ldots,k$. Η από κοινού συνάρτηση πιθανότητας εκφράζεται ισοδύναμα και ως

$$p_{X_1,\ldots,X_{k-1}}(x_1,\ldots,x_{k-1}) = \frac{n!}{x_1!\cdots x_{k-1}!(n-\sum_{i=1}^{k-1} x_i)!} p_1^{x_1}\cdots p_{k-1}^{x_{k-1}}\left(1-\sum_{i=1}^{k-1} p_i\right)^{n-\sum_{i=1}^{k-1} x_i}.$$

Η πολυωνυμική κατανομή συμβολίζεται με $(X_1, X_2, \ldots, X_{k-1}) \sim M(n, p_1, \ldots, p_{k-1})$.

Παρατηρήστε ότι καθώς ισχύει ότι $X_k = n - X_1 - \ldots - X_{k-1}$, στο τυχαίο διάνυσμα $(X_1, \ldots, X_{k-1})^t$ δεν συμπεριλαμβάνεται η X_k.

Η πολυωνυμική κατανομή ικανοποιεί τις ακόλουθες προτάσεις, η απόδειξη των οποίων αφήνεται ως άσκηση στον/στην αναγνώστη/στρια.

Πρόταση 6.9

Αν $(X_1, X_2, \dots, X_{k-1}) \sim M(n, p_1, \dots, p_{k-1})$, τότε

- $X_i \sim B(n, p_i)$.
- $(X_i, X_j) \sim M(n, p_i, p_j)$.
- $(X_{i_1}, X_{i_2}, \dots, X_{i_m}) \sim M(n, p_{i_1}, p_{i_2}, \dots, p_{i_m})$ για κάθε συλλογή διαφορετικών $i_j \in \{1, \dots, k-1\}$.
- $X_1 | X_2 = x_2, \dots, X_{k-1} = x_{k-1} \sim B\left(n - x_1 - \dots - x_{k-1}, \frac{p_1}{1-p_1-\dots-p_{k-1}}\right)$.

Παράδειγμα 6.12

Κατά τη διάρκεια μιας κλινικής μελέτης τα άτομα χωρίζονται ανεξάρτητα το ένα από το άλλο σε τρεις διαφορετικές ομάδες στις οποίες χορηγούνται διαφορετικές φαρμακευτικές αγωγές. Η πιθανότητα ένα άτομο να ενταχθεί στην πρώτη ομάδα είναι 50%, ενώ η αντίστοιχη πιθανότητα για καθεμία από τις άλλες δύο ομάδες είναι 25%. Επιλέγονται τυχαία δέκα άτομα. Να υπολογίσετε την πιθανότητα να ενταχθούν 5 άτομα στην πρώτη, 3 στη δεύτερη και 2 στην τρίτη ομάδα.

Λύση Παραδείγματος 6.12

Θα συμβολίσουμε με X_1 και X_2 τις τ.μ. που παριστάνουν τον αριθμό των ατόμων που εντάσσονται στην πρώτη και στη δεύτερη ομάδα από το σύνολο των 10 ατόμων. Προφανώς, για την κατανομή των τυχαίων μεταβλητών X_1 και X_2 έχουμε ότι:

$$(X_1, X_2) \sim M(10, 0.5, 0.25).$$

Η ζητούμενη πιθανότητα, επομένως, ισούται με:

$$\begin{aligned} p(5,3) &= P(X_1 = 5, X_2 = 3, X_3 = 10 - 5 - 3) \\ &= \frac{10!}{5!3!2!} 0.5^5 \cdot 0.25^3 \cdot 0.25^2 \\ &= \frac{6 \cdot 7 \cdot 8 \cdot 9 \cdot 10}{2 \cdot 3 \cdot 2} 0.5^5 \cdot 0.25^5 \\ &= 4 \cdot 7 \cdot 9 \cdot 10 \cdot 0.5^5 \cdot 0.25^5 \\ &= 0.077, \end{aligned}$$

όπου με X_3 έχουμε συμβολίσει την τ.μ. που παριστάνει τον αριθμό των ατόμων που εντάσσονται στην τρίτη ομάδα στο σύνολο των 10 ατόμων που επιλέγονται τυχαία.

Άσκηση Αυτοαξιολόγησης 6.10

Σε ένα τουρνουά γρήγορου σκακιού ο νικητής προκύπτει από την έκβαση 5 (ανεξάρτητων) παιχνιδιών ανάμεσα στους δύο διαγωνιζόμενους του τελικού. Από τις προηγούμενες αναμετρήσεις των δύο διαγωνιζόμενων του τελικού γνωρίζουμε ότι η πιθανότητα ένα παιχνίδι τους να λήξει

- με ισοπαλία είναι 0.05,
- με νικητή τον Α είναι 0.45 και
- με νικητή τον Β είναι 0.50.

Να υπολογίσετε την πιθανότητα στη σειρά των 5 παιχνιδιών να έχουμε 3 νίκες για τον Α και 2 για τον Β.

Παρατήρηση 6.9

Έστω $(X_1, X_2, \dots, X_{k-1}) \sim M(n, p_1, \dots, p_{k-1})$ με $\sum_{i=1}^{k-1} p_i = 1 - p_k$. Τότε με τη γλώσσα προγραμματισμού R μπορούμε:

- με τη συνάρτηση `dmultinom(x, size =n, prob)` να υπολογίσουμε τη σπ στο k–διάστατο διάνυσμα x, δηλώνοντας στο όρισμα `prob` το k-διάστατο διάνυσμα των πιθανοτήτων,
- με τη συνάρτηση `rmultinom(m, size, prob)` να δημιουργήσουμε ένα δείγμα μεγέθους m από αυτήν την κατανομή.

6.8.2 Διδιάστατη κανονική κατανομή

Η πολυδιάστατη κανονική κατανομή αποτελεί γενίκευση της κανονικής κατανομής σε περισσότερες των μία διαστάσεων και, όπως και η μονοδιάστατη κανονική κατανομή, παρουσιάζει ιδιαίτερα μεγάλη αξία λόγω των εφαρμογών στις οποίες εμφανίζεται. Στην ενότητα αυτή, αρχικά θα παρουσιαστεί η διδιάστατη κανονική κατανομή και, στη συνέχεια, θα γίνει μια σύντομη αναφορά στην πολυδιάστατη περίπτωση.

Ορισμός 6.13

Το διδιάστατο τυχαίο διάνυσμα $X = (X_1, X_2)^t$ με σύνολο δυνατών τιμών $S_X = \mathbb{R}^2$ λέμε ότι ακολουθεί τη διδιάστατη κανονική κατανομή με παραμέτρους μ_1, μ_2, $\sigma_1 > 0$, $\sigma_2 > 0$ και $|\rho| \leq 1$ αν η από κοινού συνάρτηση πυκνότητας πιθανότητας δίνεται από τη σχέση

$$f_{X_1,X_2}(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \cdot \exp\left(-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right) + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right]\right).$$

Η παραπάνω από κοινού συνάρτηση πυκνότητας πιθανότητας ισοδύναμα εκφράζεται ως

$$f_{X_1,X_2}(x_1, x_2) = \frac{1}{2\pi|\Sigma|^{1/2}} \cdot e^{-\frac{1}{2}(x-\mu)^t\Sigma^{-1}(x-\mu)}$$

όπου $x = (x_1, x_2)^t$, $\mu = (\mu_1, \mu_2)^t$ και

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \rho\,\sigma_1\sigma_2 \\ \rho\,\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$$

ο αποκαλούμενος πίνακας διασποράς-συνδιασποράς των X_1 και X_2. Συμβολικά γράφουμε ότι $X = (X_1, X_2)^t \sim N_2(\mu, \Sigma)$.

Στη συνέχεια παρατίθεται, χωρίς απόδειξη, μια σειρά από ιδιότητες της διδιάστατης κανονικής κατανομής. Για την απόδειξη αυτών των ιδιοτήτων ενδεικτικά παραπέμπουμε μεταξύ άλλων, στους Κούτρας (2010) και Muirhead (1982).

Πρόταση 6.10

Για τη διδιάστατη κανονική κατανομή $X = (X_1, X_2)^t \sim N_2(\mu, \Sigma)$, με $\mu = (\mu_1, \mu_2)^t$ και πίνακα διασποράς-συνδιασποράς που δόθηκε στον Ορισμό 6.13, ισχύουν οι παρακάτω προτάσεις:

- $X_1 \sim N(\mu_1, \sigma_1^2)$ και $X_2 \sim N(\mu_1, \sigma_1^2)$.
- $\rho(X_1, X_2) = \rho$.
- $X_1 | X_2 = x_2 \sim N(\mu_1 + \rho\frac{\sigma_1}{\sigma_2}(x_2 - \mu_2), \sigma_1^2(1-\rho^2))$.
- $X_2 | X_1 = x_1 \sim N(\mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x_1 - \mu_1), \sigma_2^2(1-\rho^2))$.
- Αν οι X_1 και X_2 είναι ασυσχέτιστες, τότε είναι και ανεξάρτητες.
- Αν A είναι 2×2 πίνακας και B είναι 2×1 διάνυσμα, τότε $AX + B \sim N_2(A\mu + B, A\Sigma A^t)$.

Παρατήρηση 6.10

Έστω $(X_1, X_2)^t \sim N_2(\mu, \Sigma)$. Τότε με τη βιβλιοθήκη `mnormt` της γλώσσας προγραμματισμού R μπορούμε:

- με τη συνάρτηση `pmvnorm(lower, upper, mean, sigma)` να υπολογίσουμε την πιθανότητα το τυχαίο διάνυσμα να ανήκει στην περιοχή που δηλώνεται στα ορίσματα `lower, upper`,
- με τη συνάρτηση `dmvnorm(x, mean, sigma)` να υπολογίσουμε τη σππ στο διάνυσμα x, δηλώνοντας στα ορίσματα `mean` και `sigma` τις παραμέτρους της,
- με τη συνάρτηση `rmvnorm(m, mean, sigma)` να δημιουργήσουμε ένα δείγμα μεγέθους m από αυτήν την κατανομή.

Παράδειγμα 6.13

Μια βιομηχανία κατασκευάζει έμβολα για μηχανές. Έστω X_1 και X_2 οι τυχαίες μεταβλητές που παριστάνουν τη διάμετρο και το μήκος των εμβόλων, αντίστοιχα. Γνωρίζουμε ότι η από κοινού κατανομή των X_1 και X_2 είναι διδιάστατη κανονική κατανομή με $\mu_1 = 60mm$, $\mu_2 = 110mm$, $\sigma_1 = \sigma_2 = 0.2mm$, και ρ. Είναι γνωστό ότι ένα έμβολο μπορεί να χρησιμοποιηθεί μόνο όταν η διάμετρος και το μήκος του βρίσκονται μέσα στο διάστημα μιας τυπικής απόκλισης από τη μέση τιμή τους.

1. Να βρεθεί το ποσοστό των εμβόλων που μπορεί να χρησιμοποιηθεί αν $\rho = 0,\ 0.5,\ 0.9$.
2. Για τις παραπάνω τιμές του ρ να υπολογίσετε την πιθανότητα ένα έμβολο να έχει διάμετρο εντός των ορίων όταν είναι γνωστό ότι έχει μήκος $110.01mm$.

Υπόδειξη: χρησιμοποιήστε την R για τους υπολογισμούς.

Λύση Παραδείγματος 6.13

Από όσα δίνονται στην εκφώνηση της άσκησης έχουμε ότι:

$$X = (X_1, X_2)^t \sim N_2(\mu, \Sigma)$$

όπου $\mu = (60, 110)^t$ και

$$\Sigma = \begin{bmatrix} 0.2^2 & \rho\, 0.2^2 \\ \rho\, 0.2^2 & 0.2^2 \end{bmatrix}$$

ο πίνακας διασποράς-συνδιασποράς.

1. Αφού ένα έμβολο μπορεί να χρησιμοποιηθεί μόνο όταν η διάμετρος και το μήκος του βρίσκονται μέσα στο διάστημα μιας τυπικής απόκλισης από τη μέση τιμή τους, έχουμε ότι το ποσοστό των

εμβόλων που μπορεί να χρησιμοποιηθεί μπορεί να υπολογιστεί από τη σχέση:

$$P(59.8 < X_1 < 60.2, 109.8 < X_2 < 110.2) = \int_{59.8}^{60.2} \int_{109.8}^{110.2} f_{X_1,X_2}(x_1,x_2)dx_2dx_1,$$

όπου $f_{X_1,X_2}(x_1,x_2)$ η από κοινού συνάρτηση πυκνότητας πιθανότητας της διδιάστατης κανονικής κατανομής. Η παραπάνω πιθανότητα θα υπολογιστεί ενθυμούμενοι ότι

$$\begin{aligned} P(x_1 < X_1 < x_2, y_1 < X_2 < y_2) &= \int_{x_1}^{x_2} \int_{y_1}^{y_2} f_{X_1,X_2}(x_1,x_2)dx_2dx_1 \\ &= F(x_2,y_2) - F(x_2,y_1) - F(x_1,y_2) + F(x_1,y_1). \end{aligned}$$

όπου $F := F_{X_1,X_2}$ η ασκ της διδιάστατης κανονικής κατανομής. Πιο συγκεκριμένα, έχουμε με τη βοήθεια της R, κάνοντας χρήση της αθροιστικής συνάρτησης κατανομής της διδιάστατης κανονικής κατανομής (`pmnorm`), μέσω των εντολών:

```
library(mnormt)
mu<-c(60,110)
rhoAll<-c(0,0.5,0.9)
x1<-c(59.8,60.2)
x2<-c(109.8,110.2)
for (i in 1:3){
  rho<-rhoAll[i]
  sigma <- matrix(c(0.2^2,rho*0.2^2, rho*0.2^2, 0.2^2), nrow = 2)
  print(pmnorm(cbind(x1[2], x2[2]), mu, sigma) - pmnorm(cbind(x1[2],
     x2[1]), mu, sigma) -
    pmnorm(cbind(x1[1], x2[2]), mu, sigma) +   pmnorm(cbind(x1[1], x2
       [1]), mu, sigma))
}
```

έχουμε ότι οι πιθανότητες αυτές είναι ίσες με

$$0.4660649,\ 0.4979718,\ 0.5963599,$$

για $\rho = 0, 0.5, 0.9$, αντίστοιχα.

2. Από τις ιδιότητες της διδιάστατης κανονικής κατανομής έχουμε ότι

$$X_1|X_2 = x_2 \sim N\left(\mu_1 + \rho\frac{\sigma_1}{\sigma_2}(x_2 - \mu_2), \sigma_1^2(1-\rho^2)\right).$$

Επομένως, όταν $X_2 = 110.01$ έχουμε

$$X_1|X_2 = 110.01 \sim N\left(60 + \rho\frac{0.2}{0.2}(110.01 - 110), 0.2^2(1-\rho^2)\right),$$

δηλαδή

$$X_1|X_2 = 110.01 \sim N(60 + \rho \cdot 0.01, 0.2^2(1-\rho^2)).$$

Επομένως, μπορούμε να υπολογίσουμε την πιθανότητα του ενδεχομένου $\{59.8 < X_1 < 60.2|X_2 = 110.01\}$, για τις διάφορες τιμές του ρ, χρησιμοποιώντας τις ιδιότητες της μονοδιάστατης κανονικής κατανομής. Οι τιμές των πιθανοτήτων αυτών είναι ίσες με

$$0.6826895,\ 0.7515899,\ 0.9775151$$

για $\rho = 0$, 0.5 και 0.9, αντίστοιχα. Ο αναλυτικός υπολογισμός των πιθανοτήτων αυτών αφήνεται ως άσκηση στον/στην αναγνώστη/στρια.

Στο Σχήμα 6.1 εμφανίζονται τα διαγράμματα της από κοινού συνάρτησης πυκνότητας πιθανότητας $f_{X_1,X_2}(x_1,x_2)$ του προηγούμενου παραδείγματος και τα αντίστοιχα διαγράμματα ισοϋψών για $\rho = 0, 0.5$ και 0.9. Από τα γραφήματα είναι φανερό ότι η ζητούμενη πιθανότητα εξαρτάται σημαντικά από τον συντελεστή συσχέτισης ρ, αφού το σχήμα της $f_{X_1,X_2}(x_1, x_2)$ επηρεάζεται σημαντικά από την τιμή του συντελεστή συσχέτισης.

Σημειώνεται ότι οι γραφικές παραστάσεις του Σχήματος 6.1 έχουν δημιουργηθεί με τη βοήθεια της R με τη χρήση των παρακάτω εντολών.

```
library(mnormt)

f <- function(x, y,mu, sigma)     dmnorm(cbind(x, y), mu, sigma)

x <- seq(59.5, 60.5, 0.025)
y <- seq(109.5, 110.5, 0.025)

mu <- c(60,110)
rhoAll <- c(0,0.5,0.9)

for (i in 1:3){
  rho<-rhoAll[i]
  sigma <- matrix(c(0.2^2,rho*0.2^2, rho*0.2^2 , 0.2^2), nrow = 2)

  z <- outer(x, y, f, mu, sigma)
  persp(x, y, z, theta = -30, phi = 35,
        shade = 0.275, col = "steelblue", expand = 0.75, r = 40, d=.3,
        ltheta = 25, ticktype = "detailed")
  contour(x, y, z)
}
```

Ο ορισμός της διδιάστατης κανονικής κατανομής έχει γενικευτεί σε περισσότερες διαστάσεις και έχει οδηγήσει στον ορισμό της πολυδιάστατης κανονικής κατανομής.

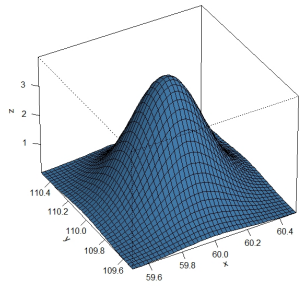
Ορισμός 6.14

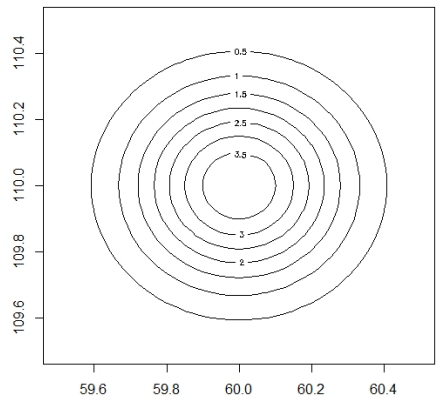
Το k-διάστατο τυχαίο διάνυσμα $X = (X_1, X_2, \dots, X_k)^t$ λέμε ότι ακολουθεί την k-διάστατη κανονική κατανομή με παραμέτρους μ, Σ αν η από κοινού συνάρτηση πυκνότητας πιθανότητάς τους δίνεται από τη σχέση

$$f_{X_1,\dots,X_k}(x_1,\dots,x_k) = \frac{1}{(2\pi)^{k/2}|\Sigma|^{1/2}} \cdot e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^t\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})},$$

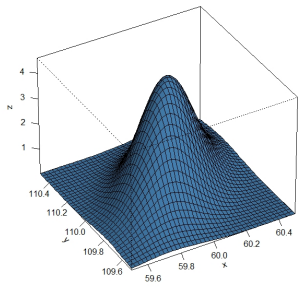
όπου $x = (x_1, \dots, x_k)^t \in \mathbb{R}^k$, $\boldsymbol{\mu} = (\mu_1 \dots, \mu_k)^t \in \mathbb{R}^k$ και Σ ο συμμετρικός, θετικά ορισμένος πίνακας διασποράς συνδιασποράς των X_i με τα στοιχεία της κύριας διαγωνίου να είναι ίσα με τις διασπορές των X_i και το (i,j) με $i \neq j$ στοιχείο να ισούται με τη συνδιασπορά των X_i και X_j. Συμβολικά γράφουμε ότι $X \sim N_k(\mu,\Sigma)$.

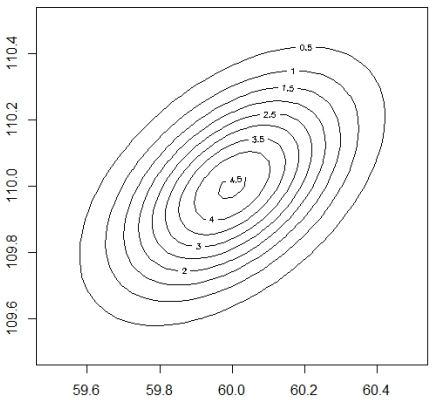
$\rho = 0$



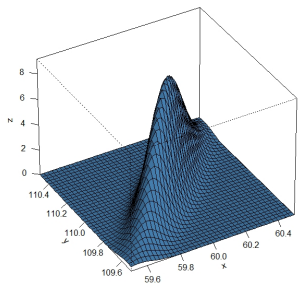


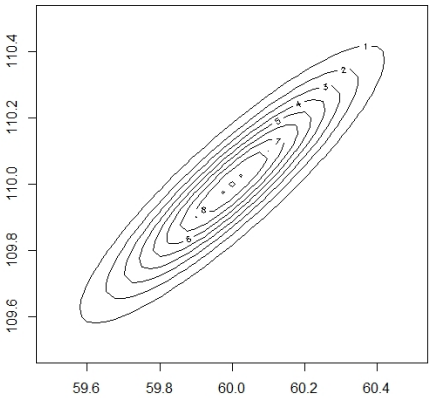
$\rho = 0.5$





$\rho = 0.9$





Σχήμα 6.1: Διαγράμματα της από κοινού συνάρτησης πυκνότητας πιθανότητας του Παραδείγματος 6.13 και τα αντίστοιχα διαγράμματα ισοϋψών για $\rho = 0, 0.5$ και 0.9.

6.9 Ασκήσεις

Άσκηση 6.1 *Η από κοινού σππ των τυχαίων μεταβλητών X και Y δίνεται από τη σχέση $f_{X,Y}(x,y) = c \cdot x \cdot y$, $0 < x < 1, 0 < y < 4, c > 0$. Να υπολογιστούν η σταθερά c και στη συνέχεια οι πιθανότητες $P(X < 0.2, Y < 3)$ και $P(X < 1.1, Y > 2)$.*

Άσκηση 6.2 *Η από κοινού συνάρτηση των μεταβλητών X και Y δίνεται από τη σχέση $f_{X,Y}(x,y) = c(x+y)$, $0 < x < 1, 0 < y < 1, c > 0$. Να υπολογιστούν η σταθερά c και οι περιθώριες κατανομές των τ.μ. X και Y. Να υπολογιστούν η $E(X|Y = y)$ και η $Var(X|Y = y)$.*

Άσκηση 6.3 *Η από κοινού σππ των τυχαίων μεταβλητών X και Y δίνεται από τη σχέση:*

$$f_{X,Y}(x,y) = \frac{1}{2\pi}\exp\left(-\frac{x^2+y^2}{2}\right) + \frac{1}{4\pi\exp(1)}x^3 \cdot y^3, \text{ για } (x,y) \in \mathbb{R}^2.$$

Είναι οι τυχαίες μεταβλητές X και Y ανεξάρτητες;

Άσκηση 6.4 *Η από κοινού συνάρτηση πιθανότητας των X, Y, Z δίνεται από τη σχέση:*

$$p_{X,Y,Z}(x,y,z) = 0.25, \text{ για } (x,y,z) = (1,0,0),(0,1,0),(0,0,1),(1,1,1).$$

Είναι οι τυχαίες μεταβλητές X, Y, Z ανά δύο ανεξάρτητες; Είναι όλες μαζί ανεξάρτητες;

Άσκηση 6.5 *Τα δύο κυριότερα είδη λαθών που κάνουν οι προγραμματιστές είναι λάθη λογικής και λάθη σύνταξης. Σε μία απλή γλώσσα, όπως η R, ο αριθμός τέτοιων λαθών είναι συνήθως πολύ μικρός. Έστω X και Y οι τ.μ. που παριστάνουν τον αριθμό λαθών λογικής και σύνταξης, αντίστοιχα, που γίνονται στο πρώτο τρέξιμο ενός προγράμματος σε γλώσσα R. Υποθέστε ότι η κοινή συνάρτηση πιθανότητας των τ.μ. X και Y είναι αυτή που δίνεται στον παρακάτω πίνακα*

X, Y	0	1	2	3	4
0	0.405	0.310	0.030	0.012	0.005
1	0.112	0.030	0.015	0.008	0.004
2	0.020	0.012	0.009	0.007	0.003
3	0.006	0.004	0.004	0.003	0.001

1. *Υπολογίστε την πιθανότητα ένα τυχαίο πρόγραμμα R να έχει το πολύ ένα λάθος λογικής και το πολύ δύο λάθη σύνταξης.*
2. *Να βρείτε τις περιθώριες συναρτήσεις πιθανότητας των τ.μ. X και Y.*
3. *Υπολογίστε την πιθανότητα ένα τυχαίο πρόγραμμα να έχει το πολύ τρία λάθη σύνταξης.*
4. *Υπολογίστε την πιθανότητα ένα τυχαίο πρόγραμμα να έχει τουλάχιστον δύο λάθη λογικής.*
5. *Υπολογίστε την πιθανότητα ένα τυχαίο πρόγραμμα να έχει τουλάχιστον τρία λάθη σύνταξης όταν είναι γνωστό ότι έχει δύο λάθη λογικής.*

Άσκηση 6.6 *Έστω X και Y οι τ.μ. που παριστάνουν τους βαθμούς συγκέντρωσης δύο ρύπων και έστω*

$$f_{X,Y}(x,y) = c(x+y), 0 < x < 10, 0 < y < 10$$

η από κοινού συνάρτηση πυκνότητας πιθανότητάς τους.

1. *Προσδιορίστε την τιμή της σταθεράς c.*

2. *Βρείτε την περιθώρια σππ της τ.μ. X και της τ.μ. Y.*
3. *Εξετάστε αν οι τυχαίες μεταβλητές X και Y είναι ανεξάρτητες.*
4. *Βρείτε την πιθανότητα $P(Y > 5|X < 4)$.*

Άσκηση 6.7 *Έστω X η τ.μ. που παριστάνει την τιμή πώλησης ενός προϊόντος από τον έμπορο στον παραγωγό (σε ευρώ ανά κιλό) και Y η τ.μ. που παριστάνει την τιμή αγοράς του προϊόντος από τον καταναλωτή. Υποθέστε ότι η από κοινού σππ των τ.μ. X και Y δίνεται από τη σχέση:*

$$f(x,y) = \begin{cases} c \cdot y, & 0 < x < y < 2, \\ 0, & \text{αλλού.} \end{cases}$$

1. *Βρείτε την τιμή της σταθεράς c.*
2. *Να βρεθεί η πιθανότητα ο καταναλωτής να αγοράζει το προϊόν τουλάχιστον μισό ευρώ ακριβότερα από την τιμή που πουλάει ο παραγωγός.*
3. *Να βρεθούν οι περιθώριες συναρτήσεις των τ.μ. X και Y.*
4. *Υπολογίστε την πιθανότητα ο παραγωγός να πουλάει το προϊόν μεταξύ 25 και 60 λεπτών το κιλό.*
5. *Υπολογίστε την πιθανότητα ο καταναλωτής να αγοράζει το προϊόν το πολύ 1.5 ευρώ, όταν ο παραγωγός το έχει πουλήσει μεταξύ 25 και 50 λεπτών.*
6. *Είναι ανεξάρτητες οι τυχαίες μεταβλητές X και Y;*

Άσκηση 6.8 *Έστω X και Y οι τυχαίες μεταβλητές που περιγράφουν τον χρόνο ζωής σε ημέρες των μικροοργανισμών Α και Β, αντίστοιχα. Είναι γνωστό ότι η από κοινού συνάρτηση πυκνότητας πιθανότητάς τους δίνεται από τη σχέση:*

$$f_{X,Y}(x,y) = \frac{1}{2}, 0 < x < y, 0 < y < 2.$$

Να προσδιορίσετε την πιθανότητα ο χρόνος ζωής του μικροοργανισμού Α να είναι μικρότερος από μισή μέρα όταν είναι γνωστό ότι ο χρόνος ζωής του μικροοργανισμού Β είναι ίσος με 1.5 μέρα.

Άσκηση 6.9 *Έστω X και Y οι τυχαίες μεταβλητές που περιγράφουν τον χρόνο ζωής σε ημέρες των μικροοργανισμών Α και Β, αντίστοιχα. Είναι γνωστό ότι η από κοινού συνάρτηση πυκνότητας πιθανότητάς τους δίνεται από τη σχέση*

$$f_{X,Y}(x,y) = \frac{2}{9}, 0 < x < y, 0 < y < 3.$$

Να προσδιορίσετε τον αναμενόμενο χρόνο ζωής του μικροοργανισμού Β όταν γνωρίζετε ότι ο χρόνος ζωής του μικροοργανισμού Α είναι ίσος με 1.5 μέρα.

Άσκηση 6.10 *Αν T, W είναι δύο ανεξάρτητες κανονικές τυχαίες κατανομές με μέση τιμή και διασπορά μ_T, σ_T^2 και μ_W, σ_W^2 αντίστοιχα, να προσδιορίσετε την από κοινού κατανομή των τυχαίων μεταβλητών X και Y που ορίζονται*

$$X = aT + bW, \quad Y = cT + dW$$

με a, b, c, d σταθερές.

Άσκηση 6.11 *Έστω (X, Y) τυχαίες μεταβλητές με από κοινού συνάρτηση πυκνότητας πιθανότητας*

$$f(x,y) = \begin{cases} cxy^2, & \text{αν } 0 < x < y < 1, \\ 0, & \text{αλλού.} \end{cases}$$

Να υπολογιστεί η σταθερά c και να υπολογιστεί η μέση τιμή του X. Ελέγξτε αν οι τυχαίες μεταβλητές X και Y είναι ανεξάρτητες. Τέλος, υπολογίστε την πιθανότητα $P(X > 0.4|Y < 0.6)$.

Άσκηση 6.12 *Μια ηλεκτρική εγκατάσταση αποτελείται από 2 ανεξάρτητα υποσυστήματα. Η διάρκεια ζωής X_i (σε χρόνια) του i-οστού υποσυστήματος έχει συνάρτηση πυκνότητας πιθανότητας:*

$$f_{X_i}(x) = cx\exp(-x),\ x > 0.$$

1. *Υπολογίστε την τιμή του c.*
2. *Υπολογίστε την πιθανότητα ένα τυχαίο υποσύστημα να λειτουργεί το πολύ 2 χρόνια.*
3. *Βρείτε την κατανομή της διάρκειας ζωής της ηλεκτρικής εγκατάστασης αν το σύστημα είναι συνδεδεμένο παράλληλα.*
4. *Βρείτε τον μέσο χρόνο ζωής της ηλεκτρικής εγκατάστασης.*
5. *Υπολογίστε την πιθανότητα η ηλεκτρική εγκατάσταση να λειτουργεί περισσότερο από 3 χρόνια.*

Άσκηση 6.13 *Σε μια αποθήκη ηλεκτρολογικού υλικού υπάρχουν καλώδια των οποίων η διάμετρος X (σε χιλιοστά) δεν είναι σταθερή, αλλά ακολουθεί την ομοιόμορφη κατανομή στο διάστημα $(5,\ 5.8)$. Ο μηχανικός ενδιαφέρεται να γνωρίζει αν το καλώδιο ανήκει σε μία από τις παρακάτω κατηγορίες: $\{X < 5.2$ χιλ.$\}$, $\{5.2 \leq X \leq 5.5$ χιλ.$\}$ και $\{X > 5.5$ χιλ.$\}$. Επιλέγει τυχαία 20 καλώδια. Υπολογίστε την πιθανότητα τρία καλώδια να έχουν διάμετρο μικρότερη από 5.2, 10 να έχουν διάμετρο από 5.2 μέχρι 5.5 και τα υπόλοιπα να έχουν διάμετρο μεγαλύτερη από 5.5.*

Άσκηση 6.14 *Ο βωξίτης είναι το μετάλλευμα που χρησιμοποιείται για την παραγωγή αλουμίνας (Al_2O_3). Οικονομικά εκμεταλλεύσιμος για παραγωγή αλουμίνας (η οποία χρησιμοποιείται για την παραγωγή μεταλλικού αλουμινίου) θεωρείται ο βωξίτης που περιέχει τουλάχιστον 50% Al_2O_3 και λιγότερο από 5% πυρίτιο (Si). Από προηγούμενες γεωλογικές μελέτες είναι γνωστό ότι η περιεκτικότητα Al_2O_3 στα κοιτάσματα βωξίτη μιας περιοχής περιγράφεται ικανοποιητικά από μια κανονική κατανομή με μέση τιμή 55% και τυπική απόκλιση 3.5%, η περιεκτικότητα σε Si περιγράφεται ικανοποιητικά από μια ομοιόμορφη κατανομή στο διάστημα 1% έως 10%, ενώ οι συγκεντρώσεις των δύο υλικών σε ένα κοίτασμα βωξίτη θεωρούνται ανεξάρτητες. Από την περιοχή αυτή επιλέγονται τυχαία διάφορα κοιτάσματα και εξετάζεται η δυνατότητα εμπορικής εκμετάλλευσής τους.*

1. *Να υπολογιστεί η πιθανότητα ένα κοίτασμα βωξίτη να είναι εμπορικά εκμεταλλεύσιμο.*
2. *Να υπολογιστεί η πιθανότητα σε ένα δείγμα 10 ανεξάρτητων κοιτασμάτων βωξίτη από τη συγκεκριμένη περιοχή να είναι εμπορικά εκμεταλλεύσιμα λιγότερα από 2.*

6.10 Απαντήσεις στις ασκήσεις αυτοαξιολόγησης

Λύση Άσκησης Αυτοαξιολόγησης 6.1

Για τον υπολογισμό της ζητούμενης πιθανότητας μπορούμε να εργαστούμε ως ακολούθως:

$$\begin{aligned}
P(\{X > x \cap Y > y\}) &= 1 - P(\{X > x \cap Y > y\}') \\
&= 1 - P(\{X > x\}' \cup \{Y > y\}') \\
&= 1 - P(\{X \leq x\} \cup \{Y \leq y\}) \\
&= 1 - \Big(P(X \leq x) + P(Y \leq y) - P(X \leq x, Y \leq y)\Big) \\
&= 1 - \Big(F_X(x) + F_Y(y) - F_{XY}(x,y)\Big) \\
&= 1 - F_X(x) - F_Y(y) + F_{XY}(x,y),
\end{aligned}$$

όπου $F_X(\cdot)$ και $F_Y(\cdot)$ οι περιθώριες κατανομές των X και Y, οι οποίες προσδιορίζονται από τις σχέσεις $F_X(x) = F_{X,Y}(x, +\infty)$ και $F_Y(y) = F_{X,Y}(+\infty, y)$, αντίστοιχα.

Λύση Άσκησης Αυτοαξιολόγησης 6.2

Στον επόμενο πίνακα παρουσιάζονται οι πιθανότητες (έχουν γίνει οι απαραίτητες στρογγυλοποιήσεις) όλων των ενδεχομένων για την από κοινού συνάρτηση πιθανότητας του προβλήματος, καθώς και οι περιθώριες συναρτήσεις πιθανότητας των δύο τυχαίων μεταβλητών.

	$X=0$	$X=1$	$X=2$	$p_Y(y)$
$Y=0$	0.000	0.021	0.083	0.104
$Y=1$	0.021	0.083	0.188	0.292
$Y=2$	0.083	0.188	0.333	0.604
$p_X(x)$	0.104	0.292	0.604	1.000

Κάθε τιμή στον πίνακα έχει υπολογιστεί από την αντικατάσταση των δυνατών τιμών των τυχαίων μεταβλητών X και Y στην από κοινού συνάρτηση πιθανότητας, ενώ οι περιθώριες συναρτήσεις πιθανότητας των X και Y έχουν υπολογιστεί αθροίζοντας τις τιμές στα κελιά κατά στήλη και γραμμή, αντίστοιχα.

Λύση Άσκησης Αυτοαξιολόγησης 6.3

1. Η από κοινού συνάρτηση πυκνότητας πιθανότητας των X και Y υπολογίζεται από τη σχέση:

$$\begin{aligned}
f_{X,Y}(x,y) &= \frac{\partial^2 F_{X,Y}(x,y)}{\partial x \partial y} \\
&= \frac{\partial^2 (1 - e^{-2x})(1 - e^{-2y})}{\partial x \partial y} \\
&= 2e^{-2x} \cdot 2e^{-2y} = 4e^{-2x-2y},\ x, y > 0,
\end{aligned}$$

   ενώ $f_{X,Y}(x,y) = 0$ για όλες τις υπόλοιπες τιμές των (x,y).

2. Η πιθανότητα η τυχαία μεταβλητή X να είναι μεταξύ του 1 και του 2 και η Y μικρότερη του 2

υπολογίζεται από τη σχέση

$$P(1<X<2, Y<2) = \int_1^2 \int_0^2 f_{X,Y}(x,y)dydx$$
$$= \int_1^2 \int_0^2 4e^{-2x-2y}dydx.$$

Το παραπάνω ολοκλήρωμα εύκολα προκύπτει ότι ισούται με $(e^{-2}-e^{-4})\cdot(1-e^{-4}) = 0.114876$. Στο ίδιο αποτέλεσμα μπορούμε να καταλήξουμε παρατηρώντας ότι

$$P(1<X<2, Y<2) = P(X<2, Y<2) - P(X<1, Y<2)$$
$$= F_{X,Y}(2,2) - F_{X,Y}(1,2)$$
$$= (1-e^{-4})(1-e^{-4}) - (1-e^{-2})(1-e^{-4}).$$

3. Οι περιθώριες συναρτήσεις πυκνότητας πιθανότητας των X και Y υπολογίζονται από τις σχέσεις:

$$f_X(x) = \int_{y:f_{X,Y}(x,y)>0} f_{X,Y}(x,y)dy = \int_0^{+\infty} f_{X,Y}(x,y)dy,$$

και

$$f_Y(y) = \int_{x:f_{X,Y}(x,y)} f_{X,Y}(x,y)dx = \int_0^{+\infty} f_{X,Y}(x,y)dx,$$

αντίστοιχα. Υπολογίζοντας τα παραπάνω ολοκληρώματα καταλήγουμε ότι

$$f_X(x) = 2e^{-2x}, \text{ και } f_Y(y) = 2e^{-2y}$$

για $x, y > 0$. Δηλαδή και οι δύο τυχαίες μεταβλητές ακολουθούν την εκθετική κατανομή με παράμετρο 2.

Εναλλακτικά, μπορούμε να προσδιορίσουμε την $F_X(x) = F_{X,Y}(x,+\infty) = 1-e^{-2x}$ και παραγωγίζοντας να προκύψει η ζητούμενη συνάρτηση πυκνότητας πιθανότητας. Παρόμοια $F_Y(y) = F_{X,Y}(+\infty, y) = 1-e^{-2y}$.

Λύση Άσκησης Αυτοαξιολόγησης 6.4

Η υπό συνθήκη συνάρτηση πυκνότητας πιθανότητας του κόστους X της προμήθειας των πρώτων υλών, αν γνωρίζουμε ότι το εργατικό κόστος ισούται με $Y = y$ δίνεται από τη σχέση:

$$f_{X|Y=y}(x) = \frac{f_{X,Y}(x,y)}{f_Y(y)} = \frac{xe^{-x(y+1)}}{\int_0^{+\infty} xe^{-x(y+1)}dx}$$
$$= \frac{xe^{-x(y+1)}}{\frac{1}{(1+y)^2}} = x(y+1)^2 e^{-x(y+1)},\ x>0.$$

Η $X|Y=y$ είναι, όπως λέγεται, μια σταθμισμένη ως προς το μήκος εκθετική κατανομή με παράμετρο $(y+1)$.

Οι σταθμισμένες ως προς το μήκος κατανομές αποτελούν ειδική περίπτωση των σταθμισμένων ως προς το μέγεθος κατανομών που ορίζονται από τη σχέση:

$$f_X^r(x) = \frac{x^r f_X(x)}{E(X^r)}, x>0$$

για $r = 1$. Οι σταθμισμένες κατανομές εμφανίζονται στη μεροληπτική δειγματοληψία. Ο αναγνώστης που ενδιαφέρεται να διαβάσει περισσότερα για αυτές τις κατανομές παραπέμπεται, ενδεικτικά, στα άρθρα Patil and Rao (1978), Economou and Tzavelas (2013) και Economou *et al.* (2021).

Λύση Άσκησης Αυτοαξιολόγησης 6.5

1. Η αναμενόμενη τιμή της X ισούται με

$$E(X) = \int_0^1 \int_0^{1-x} x f_{X,Y}(x,y) dy dx,$$

αφού τα x και y πρέπει να ικανοποιούν την ανίσωση $x + y < 1$. Επομένως, όταν η X ισούται με x, η Y μπορεί να πάρει τιμές στο διάστημα $(0, 1-x)$. Υπολογίζοντας το παραπάνω ολοκλήρωμα βρίσκουμε ότι $E(X) = 0.4$.

Εναλλακτικά, αρχικά προσδιορίζουμε την περιθώρια κατανομή της τυχαίας μεταβλητής X. Είναι τότε:

$$f_X(x) = \int_0^{1-x} 24xy dy = 12x(1-x)^2, 0 < x < 1.$$

Επομένως,

$$E(X) = \int_0^1 12x^2(1-x)^2 dx = 12 \int_0^1 x^2(1-x)^2 dx.$$

Το τελευταίο ολοκλήρωμα (βλ. το Παράρτημα Β΄) είναι ίσο με $B(3,3)$ και ισχύει ότι:

$$B(3,3) = \frac{\Gamma(3)\Gamma(3)}{\Gamma(6)} = \frac{2!2!}{5!} = 1/30.$$

Επομένως, είναι $E(X) = 12/30 = 0.4$.

Σημειώνεται ότι το ολοκλήρωμα $\int_0^1 x^2(1-x)^2 dx$ μπορεί να υπολογιστεί και αναλυτικά χωρίς τη χρήση της Βήτα συνάρτησης.

2. Ακολουθώντας παρόμοιο σκεπτικό με το προηγούμενο ερώτημα έχουμε ότι η αναμενόμενη τιμή της Y^2 δίνεται από τη σχέση:

$$E(Y^2) = \int_0^1 \int_0^{1-x} y^2 f_{X,Y}(x,y) dy dx$$

και, μετά από λίγη άλγεβρα, προκύπτει ότι ισούται με 0.2. Σημειώνεται ότι θα καταλήγαμε στο ίδιο αποτέλεσμα, αν είχαμε φράξει το x και όχι το y, δηλαδή αν εκφράζαμε την αναμενόμενη τιμή της Y^2 ως

$$E(Y^2) = \int_0^1 \int_0^{1-y} y^2 f_{X,Y}(x,y) dx dy$$

Εναλλακτικά, αρχικά προσδιορίζουμε την περιθώρια κατανομή της τυχαίας μεταβλητής Y. Είναι τότε:

$$f_Y(y) = \int_0^{1-y} 24xy dx = 12y(1-y)^2, 0 < y < 1.$$

Επομένως,

$$E(Y^2) = \int_0^1 12y^3(1-y)^2 dy = 12 \int_0^1 y^3(1-y)^2 dy.$$

Το τελευταίο ολοκλήρωμα όμως είναι ίσο με $B(4,3)$ (βλ. σχέση (Β΄.10) του Παραρτήματος Β΄) και είναι:

$$B(4,3) = \frac{\Gamma(4)\Gamma(3)}{\Gamma(7)} = \frac{3!2!}{6!} = 1/60.$$

Επομένως, είναι $E(Y^2) = 12/60 = 0.2$.

3. Η αναμενόμενη τιμή της $X + Y^2$ προκύπτει από το άθροισμα των αναμενόμενων τιμών των X και Y^2. Επομένως, έχουμε ότι

$$E(X + Y^2) = E(X) + E(Y^2) = 0.4 + 0.2 = 0.6.$$

Σημειώνεται ότι το ολοκλήρωμα $\int_0^1 y^3(1-y)^2 dy$ μπορεί να υπολογιστεί και αναλυτικά χωρίς τη χρήση της Βήτα συνάρτησης.

Λύση Άσκησης Αυτοαξιολόγησης 6.6

Η αναμενόμενη τιμή της ροπής κάμψης στο σημείο στήριξης της δοκού ισούται με

$$E(M) = E(\alpha W_1 + 2\alpha W_2)$$

η οποία, χρησιμοποιώντας τις ιδιότητες της μέσης τιμής, μπορεί να εκφραστεί ως

$$E(M) = \alpha \mathrm{E}(W_1) + 2\alpha \mathrm{E}(W_2).$$

Επομένως, για να υπολογιστεί η αναμενόμενη τιμή της ροπής κάμψης αρκεί να υπολογιστούν οι αναμενόμενες τιμές των W_1 και W_2. Οι αναμενόμενες αυτές τιμές υπολογίζονται από τις σχέσεις:

$$E(W_1) = \int_0^{+\infty} \int_0^{+\infty} w_1 w_1 e^{-(w_1+w_2)} dw_2 dw_1,$$

και

$$E(W_2) = \int_0^{+\infty} \int_0^{+\infty} w_2 w_1 e^{-(w_1+w_2)} dw_2 dw_1,$$

αντίστοιχα. Για το πρώτο διπλό ολοκλήρωμα έχουμε ότι:

$$E(W_1) = \left(\int_0^{+\infty} w_1^2 e^{-w_1} dw_1\right) \cdot \left(\int_0^{+\infty} e^{-w_2} dw_2\right)$$

Το δεύτερο από αυτά είναι ίσο με ένα, καθώς πρόκειται για το ολοκλήρωμα της συνάρτησης πυκνότητας πιθανότητας της Εκθετικής κατανομής με παράμετρο 1, ενώ το πρώτο από τη σχέση (Β΄.6) είναι ίσο με $\Gamma(3) = 2! = 2$. Επομένως, το διπλό ολοκλήρωμα είναι ίσο με 2.
Για το δεύτερο διπλό ολοκλήρωμα έχουμε ότι:

$$E(W_2) = \left(\int_0^{+\infty} w_1 e^{-w_1} dw_1\right) \cdot \left(\int_0^{+\infty} w_2 e^{-w_2} dw_2\right).$$

Από τη σχέση (Β΄.6) προκύπτει ότι καθένα εκ των δύο είναι ίσο με $\Gamma(2) = 1! = 1$. Επομένως, τα παραπάνω διπλά ολοκληρώματα είναι ίσα με 2 και 1, αντίστοιχα, και

$$E(M) = \alpha \mathrm{E}(W_1) + 2\alpha \mathrm{E}(W_2) = 2\alpha + 2\alpha = 4\alpha.$$

Εναλλακτικά, θα μπορούσε κάποιος να προσδιορίσει τις περιθώριες κατανομές των W_1 και W_2. Εύκολα προκύπτει, χρησιμοποιώντας τη σχέση (Β΄.6) του Παραρτήματος Β΄, ότι η τυχαία μεταβλητή W_1 ακολουθεί Γάμμα κατανομή με παραμέτρους 2 και 1 (άρα μέση τιμή 2), ενώ η τυχαία μεταβλητή W_2 ακολουθεί Εκθετική κατανομή με παράμετρο 1 (άρα μέση τιμή 1).

Λύση Άσκησης Αυτοαξιολόγησης 6.7

Ζητείται η αναμενόμενη τιμή του κόστους X της προμήθειας των πρώτων υλών, όταν γνωρίζουμε ότι το εργατικό κόστος ισούται με $Y = y$, δηλαδή η

$$E(X|Y=y) = \int_0^{\infty} x f_{X|Y=y}(x)dx = \int_0^{\infty} x\, x(y+1)^2 e^{-x(y+1)}dx$$
$$= \frac{2}{(y+1)^3}, \text{ με } y > 0.$$

Για τον υπολογισμό του ολοκληρώματος παρατηρήστε ότι

$$\int_0^{+\infty} x^2 e^{-x(y+1)}dx = \int_0^{+\infty} x^{3-1} e^{-x(y+1)}dx$$
$$= \frac{\Gamma(3)}{(y+1)^3}\int_0^{+\infty} \frac{(y+1)^3 x^{3-1} e^{-x(y+1)x}}{\Gamma(3)}dx$$
$$= \frac{\Gamma(3)}{(y+1)^3} = \frac{2}{(y+1)^3},$$

όπου το ολοκλήρωμα της δεύτερης ισότητας είναι ίσο με 1, καθώς εμφανίζεται η σππ της Γάμμα κατανομής που δίνεται στη σχέση (5.27) για $\lambda = (y+1)$ και $a = 3$.

Λύση Άσκησης Αυτοαξιολόγησης 6.8

Για να εξετάσουμε αν δύο τυχαίες μεταβλητές είναι ανεξάρτητες, αρκεί να εξετάσουμε αν ισχύει $f_{X,Y}(x,y) = f_X(x)f_Y(y)$. Οι περιθώριες συναρτήσεις πυκνότητας πιθανότητας του χρόνου X που απαιτείται για την παραλαβή των ανταλλακτικών ενός μηχανήματος και του χρόνου Y επισκευής του μηχανήματος δίνονται από τις σχέσεις:

$$f_X(x) = \int_0^{\infty} \frac{1}{4} x y e^{-\frac{x}{2}-y}dy = \frac{1}{4}xe^{-x/2},\ x > 0,$$

και

$$f_Y(y) = \int_0^{\infty} \frac{1}{4} x y e^{-\frac{x}{2}-y}dx = ye^{-y},\ y > 0,$$

αντίστοιχα. Εύκολα μπορούμε να διαπιστώσουμε ότι $f_{X,Y}(x,y) = f_X(x)f_Y(y)$ για κάθε $(x,y) \in \mathbb{R}^2$. Επομένως, ο χρόνος συναρμολόγησης είναι ανεξάρτητος από τον χρόνο παραλαβής των ανταλλακτικών του μηχανήματος αυτού.

Λύση Άσκησης Αυτοαξιολόγησης 6.9

Για να είναι ανεξάρτητες οι τ.μ. Χ και Υ πρέπει $f_{X,Y}(x,y) = f_X(x)f_Y(x)$ για κάθε $(x,y) \in \mathbb{R}^2$. Οι περιθώριες των X και Y προκύπτουν από την από κοινού συνάρτηση πυκνότητας πιθανότητας ολοκληρώνοντας ως προς x και y, αντίστοιχα, και δίνονται από τις σχέσεις:

$$f_X(x) = \int_1^5 f_{X,Y}(x,y)dy = \int_1^5 \left(\frac{x}{5} + cy\right)dy = \left(\frac{x}{5}y + c\frac{y^2}{2}\right)\Big|_{y=1}^{5} = \frac{4x}{5} + 12c,\ 0 < x < 1,$$

$$f_Y(y) = \int_0^1 f_{X,Y}(x,y)dx = \int_0^1 \left(\frac{x}{5} + cy\right)dx = \left(\frac{x^2}{10} + cyx\right)\Big|_{x=0}^{1} = \frac{1}{10} + cy,\ 1 < y < 5.$$

Σημειώνεται ότι η σταθερά c μπορεί να προσδιοριστεί είτε από την από κοινού συνάρτηση πυκνότητας

πιθανότητας είτε από τις περιθώριες και μπορούμε να δείξουμε ότι ισούται με 1/20. Επιπλέον, είναι

$$f_{X,Y}(0.25,2) = \frac{1}{20} + \frac{1}{20} \cdot 2 = \frac{3}{20},$$

ενώ

$$f_X(0.25) = \frac{4}{20} + 12 \cdot \frac{1}{20} = \frac{16}{20}$$

και

$$f_Y(2) = \frac{1}{10} + 2 \cdot \frac{1}{20} = \frac{4}{20}.$$

Επομένως, βρήκαμε ένα $(x,y) \in \mathbb{R}^2$ για το οποίο $f_X(x)f_Y(y) \neq f_{X,Y}(x,y)$, καθώς $\frac{3}{20} \cdot \frac{16}{20} \neq \frac{4}{20}$. Επομένως, οι τ.μ. X και Y δεν είναι ανεξάρτητες.

Λύση Άσκησης Αυτοαξιολόγησης 6.10

Θα συμβολίσουμε με X_A και X_B τις τ.μ. που παριστάνουν τον αριθμό των παιχνιδιών που κερδίζει ο A και ο B, αντίστοιχα, στη σειρά των 5 παιχνιδιών. Οι τυχαίες μεταβλητές X_A και X_B κατανέμονται σύμφωνα με την $(X_A, X_B) \sim M(5, 0.45, 0.5)$. Επομένως, η ζητούμενη πιθανότητα ισούται με

$$\begin{aligned} p_{X_A,X_B}(3,2) &= P(X_A = 3, X_B = 2, X_\Gamma = 0) \\ &= \frac{5!}{3!2!0!} 0.45^3 \cdot 0.5^2 \cdot 0.05^0 = \frac{2 \cdot 3 \cdot 4 \cdot 5}{2 \cdot 2 \cdot 3} 0.45^3 \cdot 0.5^2 \\ &= 2 \cdot 5 \cdot 0.45^3 \cdot 0.5^2 = 0.2278, \end{aligned}$$

όπου με X_Γ έχουμε συμβολίσει την τ.μ. που παριστάνει τον αριθμό των παιχνιδιών που λήγουν ισόπαλα στη σειρά των 5 παιχνιδιών.

ΒΙΒΛΙΟΓΡΑΦΙΑ

Ελληνόγλωσση

Κούτρας, Μ. (2005). *Εισαγωγή στις Πιθανότητες. Μέρος ΙΙ (Θεωρία και Εφαρμογές)*. Εκδόσεις Σταμούλη.

Κούτρας Μ. και Ευαγγελάρας, Χ. (2010). *Ανάλυση Παλινδρόμησης. Θεωρία και Εφαρμογές*. Εκδόσεις Σταμούλη.

Παπαϊωάννου, Τ. (1997). *Θεωρία πιθανοτήτων και στατιστικής*. Σταμούλη Α.Ε.

Ξενόγλωσση

Economou, P., Batsidis, A., Tzavelas, G. and Malefaki, S. (2021). Understanding the Sampling Bias: A Case Study on NBA Drafts. *Journal of Statistical Theory and Practice*, *15*, pp. 1–20.

Economou, P. and Tzavelas, G. (2013). Sample Tests for Detection of Size-Biased Sampling Mechanism. *Communications in Statistics - Theory and Methods*, *42*, pp. 3280–3295.

Muirhead, R. J. (1982). *Aspects of Multivariate Statistical Theory*. John Wiley and Sons, Inc.

Patil, G. P. and Rao, G. P. (1978). Weighted Distributions and Size-Biased Sampling with Applications to Wildlife Populations and Human Families. *Biometrics*, *34*, pp. 179–189.

Walpole, R., Myers, R., Myers, S. and Ye, K. (2017). *Probability & Statistics for Engineers & Scientists*. Pearson.

ΚΕΦΑΛΑΙΟ 7

ΜΕΤΑΣΧΗΜΑΤΙΣΜΟΙ ΤΥΧΑΙΩΝ ΜΕΤΑΒΛΗΤΩΝ - ΚΕΝΤΡΙΚΟ ΟΡΙΑΚΟ ΘΕΩΡΗΜΑ

Σύνοψη

Σε αυτό το κεφάλαιο γενικεύονται οι μέθοδοι της αθροιστικής συνάρτησης κατανομής, του μετασχηματισμού και της ροπογεννήτριας για τον προσδιορισμό της κατανομής συναρτήσεων τυχαίων μεταβλητών. Στο πλαίσιο αυτό, εισάγονται και παρουσιάζονται τρεις πολύ σημαντικές κατανομές, η χι-τετράγωνο, η t και η F κατανομή. Έπειτα, το ενδιαφέρον μας επικεντρώνεται στον προσδιορισμό της κατανομής του αθροίσματος ανεξάρτητων τυχαίων μεταβλητών. Το κεφάλαιο ολοκληρώνεται με την παρουσίαση ενός θεμελιώδους θεωρήματος, σύμφωνα με το οποίο το άθροισμα ή η μέση τιμή n το πλήθος ανεξάρτητων και ισόνομων τυχαίων μεταβλητών από μια κατανομή με πεπερασμένη μέση τιμή και διακύμανση ακολουθεί, για μεγάλες τιμές του n, προσεγγιστικά κανονική κατανομή. Το θεώρημα αυτό, που είναι ένα από τα σπουδαιότερα της Θεωρίας Πιθανοτήτων και της Στατιστικής, είναι γνωστό ως Κεντρικό Οριακό Θεώρημα.

Προαπαιτούμενη γνώση: Οι έννοιες που παρουσιάζονται στα προηγούμενα κεφάλαια αυτού του συγγράμματος.

Προσδοκώμενα μαθησιακά αποτελέσματα: Όταν θα έχετε ολοκληρώσει τη μελέτη του κεφαλαίου αυτού, θα μπορείτε

- να προσδιορίζετε την κατανομή συναρτήσεων τυχαίων μεταβλητών,
- να αναγνωρίζετε καθεμία από τις κατανομές χι-τετράγωνο, t και F και πώς/πότε αυτές προκύπτουν,
- να διαβάζετε τους πίνακες για την εύρεση των α-συμπληρωματικών ποσοστιαίων σημείων των κατανομών χι-τετράγωνο, t και F και
- να εφαρμόζετε το Κεντρικό Οριακό Θεώρημα.

Οικονόμου, Π., Μαλεφάκη, Σ., & Μπατσίδης, Α. (2023). *Πιθανότητες - Στατιστική.*

Γλωσσάριο επιστημονικών όρων

- Κατανομή Αθροίσματος ανεξάρτητων τυχαίων μεταβλητών
- Κατανομή συναρτήσεων τυχαίων μεταβλητών
- Κατανομή χ^2
- Κατανομή F
- Κατανομή t
- Κεντρικό Οριακό Θεώρημα (ΚΟΘ)
- Μέθοδος αθροιστικής συνάρτησης κατανομής
- Μέθοδος μετασχηματισμού
- Μέθοδος ροπογεννήτριας

7.1 Εισαγωγή

Στο Κεφάλαιο 3 ασχοληθήκαμε, μεταξύ άλλων με την εύρεση της κατανομής μιας συνάρτησης μιας τ.μ. Αρχικά, στο κεφάλαιο αυτό το ενδιαφέρον μας θα επικεντρωθεί στην εύρεση της κατανομής συναρτήσεων τυχαίων μεταβλητών. Ως ειδική περίπτωση θα οριστούν οι κατανομές χι-τετράγωνο, t και F, ενώ το ενδιαφέρον μας επικεντρώνεται και στην ειδική περίπτωση του προσδιορισμού της κατανομής του αθροίσματος n το πλήθος ανεξάρτητων και ισόνομων τυχαίων μεταβλητών. Το κεφάλαιο ολοκληρώνεται με την παρουσίαση του Κεντρικού Οριακού Θεωρήματος, ενός θεμελιώδους θεωρήματος, σύμφωνα με το οποίο, υπό μη αυστηρές προϋποθέσεις, μπορεί η κατανομή του αθροίσματος ή του μέσου όρου n το πλήθος ανεξάρτητων και ισόνομων τυχαίων μεταβλητών να προσεγγιστεί από την κανονική κατανομή.

7.2 Κατανομές συναρτήσεων τυχαίων μεταβλητών

Υποθέτουμε ότι γνωρίζουμε την από κοινού κατανομή των k το πλήθος τυχαίων μεταβλητών $X_1,\ldots,X_k$ και θέλουμε να προσδιορίσουμε την κατανομή m ($1 \leq m \leq k$) το πλήθος τυχαίων μεταβλητών $Y_1,\ldots,Y_m$ που ορίζονται ως συνάρτηση των αρχικών, δηλαδή $Y_i = g_i(X_1,\ldots,X_k)$, $i = 1,\ldots,m$. Με παρόμοιο σκεπτικό με αυτό που αναπτύχθηκε στην Ενότητα 3.7, χρησιμοποιούνται και σε αυτήν την ενότητα η μέθοδος της ασκ, η μέθοδος του μετασχηματισμού και η μέθοδος της ροπογεννήτριας.

Στη συνέχεια, παρουσιάζονται οι παραπάνω μέθοδοι και αναλύονται μέσω παραδειγμάτων.

7.2.1 Μέθοδος της αθροιστικής συνάρτησης κατανομής

Η μέθοδος έγκειται στον προσδιορισμό της από κοινού ασκ των τυχαίων μεταβλητών $Y_i = g_i(X_1,\ldots,X_k)$, $i = 1,\ldots,m$. Ο προσδιορισμός αυτός επιτυγχάνεται μέσω της σχέσης:

$$\begin{aligned} F_{Y_1,Y_2,\ldots,Y_m}(y_1,y_2,\ldots,y_m) &= P(Y_1 \leq y_1, Y_2 \leq y_2,\ldots,Y_m \leq y_m) \\ &= P(g_1(X) \leq y_1, g_2(X) \leq y_2, g_m(X) \leq y_m), \end{aligned}$$

όπου $X = (X_1,\ldots,X_k)^t$. Η τελευταία πιθανότητα υπολογίζεται αθροίζοντας (ή ολοκληρώνοντας) την από κοινού συνάρτηση (πυκνότητας) πιθανότητας των $X_1,X_2,\ldots,X_k$ πάνω στην περιοχή που καθορίζεται από τα αντίστοιχα ενδεχόμενα. Κάποιες φορές αυτό μπορεί να επιτευχθεί σχετικά εύκολα, αλλά αρκετά συχνά εμφανίζονται δυσκολίες καθώς ο προσδιορισμός των περιοχών, πάνω από τις οποίες υπολογίζονται οι πιθανότητες των ενδεχομένων αυτών, είναι αρκετά δύσκολος.

Τα παραπάνω θα γίνουν αντιληπτά μέσω των προτάσεων και του παραδείγματος που ακολουθεί. Αρχικά, θα προσδιοριστούν οι κατανομές δύο εκ των σημαντικότερων συναρτήσεων k του πλήθους ανεξάρτητων και ισόνομων τυχαίων μεταβλητών, του μεγίστου και του ελαχίστου τους, που βρίσκουν εφαρμογή σε πληθώρα επιστημονικών πεδίων.

Πρόταση 7.1

Έστω $X_1,X_2,\ldots,X_k$, k το πλήθος ανεξάρτητες και ισόνομες τυχαίες μεταβλητές με αθροιστική συνάρτηση κατανομής $F_X(\cdot)$ και συνάρτηση πυκνότητας πιθανότητας ή συνάρτηση πιθανότητας $f_X(\cdot)$. Τότε η τυχαία μεταβλητή $W = max\{X_1,X_2,\ldots,X_k\}$ έχει αθροιστική συνάρτηση κατανομής που δίνεται από τη σχέση:

$$F_W(w) = [F_X(w)]^k,$$

και συνάρτηση πιθανότητας που δίνεται από τη σχέση:

$$f_W(w) = k[F_X(w)]^{k-1} f_X(w).$$

Απόδειξη Πρότασης 7.1

Αρκεί να προσδιορίσουμε την αθροιστική συνάρτηση κατανομής της W που προφανώς είναι μια συνάρτηση των $X_1, X_2, \dots, X_k$. Είναι $F_W(w) = P(W \leq w)$, και καθώς $W \leq w$ σημαίνει ότι $max\{X_1, X_2, \dots, X_k\} \leq w$, έχουμε ότι όλα τα X_i είναι μικρότερα ή ίσα από w. Επομένως, λόγω ανεξαρτησίας και ισονομίας, προκύπτει ότι:

$$\begin{aligned} F_W(w) &= P(W \leq w) = P(X_1 \leq w \cap X_2 \leq w \cap \cdots \cap X_k \leq w) \\ &= P(X_1 \leq w) \cdot P(X_2 \leq w) \cdots P(X_k \leq w) = [P(X \leq w)]^k \\ &= [F_X(w)]^k . \end{aligned}$$

Η συνάρτηση πυκνότητας πιθανότητας προκύπτει άμεσα με παραγώγιση της $F_W(w)$.

Πρόταση 7.2

Έστω $X_1, X_2, \dots, X_k$, k το πλήθος ανεξάρτητες και ισόνομες τυχαίες μεταβλητές με αθροιστική συνάρτηση κατανομής $F_X(\cdot)$ και συνάρτηση πυκνότητας πιθανότητας ή συνάρτηση πιθανότητας $f_X(\cdot)$. Τότε η τυχαία μεταβλητή $Z = min\{X_1, X_2, \dots, X_k\}$ έχει αθροιστική συνάρτηση κατανομής που δίνεται από τη σχέση:

$$F_Z(z) = 1 - (1 - F_X(z))^k ,$$

και συνάρτηση πιθανότητας που δίνεται από τη σχέση:

$$f_Z(z) = k[1 - F_X(z)]^{k-1} f_X(z).$$

Απόδειξη Πρότασης 7.2

Αρκεί να προσδιορίσουμε την αθροιστική συνάρτηση κατανομής της Z που προφανώς είναι μια συνάρτηση των $X_1, X_2, \dots, X_k$. Είναι $F_Z(z) = P(Z \leq z)$, όμως τώρα το γεγονός ότι $min\{X_1, X_2, \dots, X_k\} \leq z$ δεν μας οδηγεί σε κάποιο συμπέρασμα για όλα τα X_i. Ωστόσο, $F_Z(z) = 1 - P(Z > z)$ και το ενδεχόμενο $min\{X_1, X_2, \dots, X_k\} > z$ σημαίνει ότι όλα τα X_i είναι μεγαλύτερα από z, δηλαδή προκύπτει το ενδεχόμενο

$$X_1 > z \cap X_2 > z \cap \cdots \cap X_k > z.$$

Επομένως, λόγω ανεξαρτησίας και ισονομίας, προκύπτει ότι:

$$\begin{aligned} F_Z(z) &= 1 - P(Z > z) = 1 - P(X_1 > z \cap X_2 > z \cap \cdots \cap X_k > z) \\ &= 1 - P(X_1 > z)P(X_2 > z) \cdots P(X_k > z) = 1 - (P(X > z))^k \\ &= 1 - (1 - F_X(z))^k . \end{aligned}$$

Η συνάρτηση πυκνότητας πιθανότητας προκύπτει άμεσα με παραγώγιση της $f_Z(z)$.

Παράδειγμα 7.1

Μια ηλεκτρική εγκατάσταση αποτελείται από k το πλήθος ανεξάρτητα υποσυστήματα. Η διάρκεια ζωής των υποσυστημάτων ακολουθεί την εκθετική κατανομή με μέση τιμή τις 100 ώρες λειτουργίας. Να βρεθεί η κατανομή της διάρκειας ζωής του μηχανήματος,

1. αν τα υποσυστήματα είναι σε παράλληλη σύνδεση και
2. αν τα υποσυστήματα είναι συνδεδεμένα σε σειρά.

Λύση Παραδείγματος 7.1

Αρχικά, θα προσδιορίσουμε την κατανομή κάθε υποσυστήματος. Αφού η διάρκεια ζωής κάθε υποσυστήματος, έστω X_i, $i = 1, \dots, k$, περιγράφεται από μια εκθετική κατανομή με μέση τιμή τις 100 ώρες λειτουργίας, έχουμε ότι η συνάρτηση κατανομής της διάρκειας ζωής τους δίνεται από τη σχέση

$$F_{X_i}(x) = \begin{cases} 0 & x < 0 \\ 1 - e^{-\frac{1}{100}x} & \text{αν } x \geq 0. \end{cases}$$

1. Αν τα υποσυστήματα είναι σε παράλληλη σύνδεση, τότε η διάρκεια ζωής του μηχανήματος ορίζεται από τον χρόνο ζωής του τελευταίου υποσυστήματος που θα χαλάσει, δηλαδή από την τυχαία μεταβλητή $W = max\{X_1, X_2, \dots, X_k\}$. Αφού τα υποσυστήματα είναι ανεξάρτητα, η συνάρτηση κατανομής της W δίνεται από τη σχέση (βλ. την Πρόταση 7.1):

$$F_W(w) = [F_X(w)]^k = \left(1 - e^{-\frac{1}{100}w}\right)^k,$$

για $w > 0$ και μηδέν διαφορετικά. Η συνάρτηση πυκνότητας πιθανότητας της W για $w > 0$ δίνεται από τη σχέση

$$f_W(w) = \frac{1}{100} k e^{-w/100} \left(1 - e^{-w/100}\right)^{k-1}.$$

2. Αν τα υποσυστήματα είναι συνδεδεμένα σε σειρά, τότε η διάρκεια ζωής του μηχανήματος ορίζεται από τον χρόνο ζωής του πρώτου υποσυστήματος που θα χαλάσει, δηλαδή από την τυχαία μεταβλητή $Z = min\{X_1, X_2, \dots, X_k\}$. Αφού τα υποσυστήματα είναι ανεξάρτητα, η συνάρτηση κατανομής της Z δίνεται από τη σχέση:

$$F_Z(z) = 1 - (1 - F_X(z))^k = \left(1 - (1 - 1 + e^{-\frac{1}{100}z})\right)^k$$
$$= 1 - \left(e^{-\frac{1}{100}z}\right)^k = 1 - \left(e^{-\frac{k}{100}z}\right)$$

για $z > 0$ και μηδέν διαφορετικά.
Παρατηρήστε ότι η κατανομή του ελαχίστου ανεξάρτητων και ισόνομων τυχαίων μεταβλητών εκθετικά κατανεμημένων είναι ξανά εκθετική, αλλά με διαφορετική παράμετρο, καθώς

$$f_Z(z) = \frac{k}{100} e^{-\frac{k}{100}z}, z > 0.$$

Άσκηση Αυτοαξιολόγησης 7.1

Αν X_1, X_2,…,X_k είναι ανεξάρτητες και ισόνομες ομοιόμορφες κατανομές στο $(0, \alpha)$, να υπολογίσετε τη συνάρτηση πυκνότητας πιθανότητας της $W = max\{X_1, X_2, \dots, X_k\}$ και να υπολογίσετε την αναμενόμενη τιμή της.

Στη συνέχεια, μέσω της μεθόδου της αθροιστικής συνάρτησης κατανομής θα προσδιορίσουμε την κατανομή του γινομένου δύο από κοινά κατανεμημένων τυχαίων μεταβλητών X και Y.

Παράδειγμα 7.2: Παπαϊωάννου (1997)

Έστω $(X,Y)^t$ τυχαίο διάνυσμα με από κοινού συνάρτηση πυκνότητας πιθανότητας:

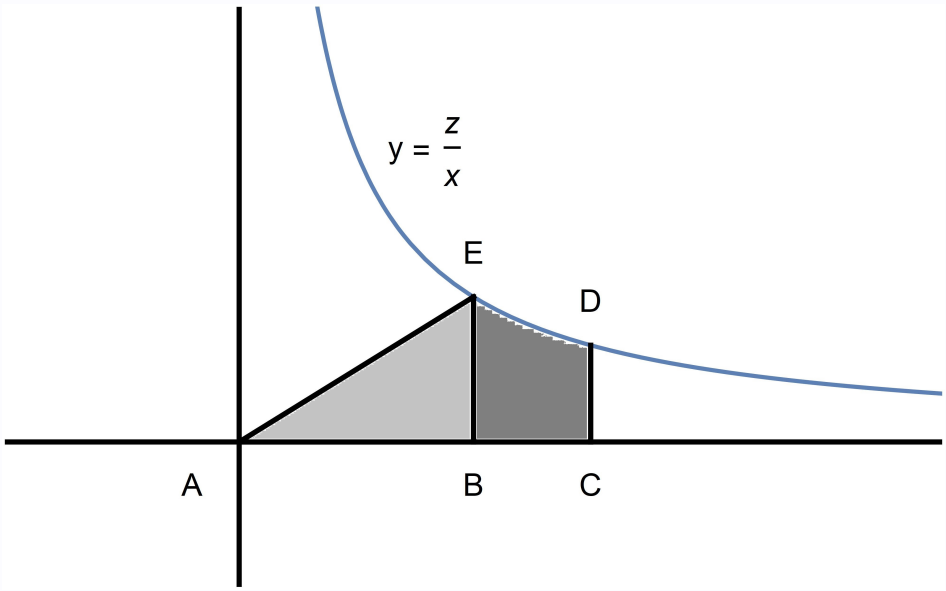
$$f_{X,Y}(x,y) = \begin{cases} 3x, & 0 \le y \le x \le 1, \\ 0, & \text{αλλού.} \end{cases}$$

Να υπολογιστεί η κατανομή της $Z = XY$.

Λύση Παραδείγματος 7.2

Για να βρούμε τη συνάρτηση κατανομής της $Z = XY$, παρατηρούμε ότι το $Z \le z$ σημαίνει $XY \le z$, το οποίο σημαίνει ότι για δοθείσα τιμή της X, έστω x, η Y πρέπει να ικανοποιεί την $Y \le z/x$. Επιπροσθέτως, από τον ορισμό της από κοινού κατανομής των X και Y, έχουμε ότι η Y πρέπει να ικανοποιεί και την $Y \le x$.

Συνοψίζοντας, για δοθείσα τιμή της X, έστω x, η Y πρέπει να ικανοποιεί την $Y \le min\{x, z/x\}$. Τότε για $0 \le x \le \sqrt{z}$ είναι $min\{x, z/x\} = x$, ενώ για $\sqrt{z} \le x \le 1$ είναι $min\{x, z/x\} = z/x$. Εναλλακτικά, τα ζευγάρια (x,y) που ικανοποιούν το $Z \le z$ ορίζονται από το χωρίο $ACDE$ του παρακάτω σχήματος.



Σε κάθε περίπτωση, η συνάρτηση κατανομής της Z δίνεται από τη σχέση:

$$F_Z(z) = P(Z \le z) = P(XY \le z) = \iint_{ABE} 3x\, dydx + \iint_{BCDE} 3x\, dydx$$

$$= \int_0^{\sqrt{z}} \int_0^x 3x\, dydx + \int_{\sqrt{z}}^1 \int_0^{z/x} 3x\, dydx$$

$$= 3z - 2z\sqrt{z},\ 0 \le z \le 1.$$

Από το προηγούμενο παράδειγμα, ίσως έγινε σαφής η δυσκολία χρήσης της μεθόδου. Για τον λόγο αυτό στις επόμενες δύο υποενότητες περιγράφονται δύο εναλλακτικές μεθοδολογίες.

7.2.2 Μέθοδος μετασχηματισμού

Υποθέτουμε ότι γνωρίζουμε την από κοινού κατανομή των k το πλήθος τυχαίων μεταβλητών $X_1,\ldots,X_k$ και θέλουμε να προσδιορίσουμε την από κοινού συνάρτηση (πυκνότητας) πιθανότητας m $(1 \leq m \leq k)$ το πλήθος τυχαίων μεταβλητών $Y_1,\ldots,Y_m$ που ορίζονται ως συνάρτηση των αρχικών, δηλαδή $Y_i = g_i(X_1,\ldots,X_k)$, $i = 1,\ldots,m$. Για την παρουσίαση της μεθόδου διακρίνουμε δύο περιπτώσεις: το τυχαίο διάνυσμα $X = (X_1,\ldots,X_k)^t$ να είναι διακριτό και το τυχαίο διάνυσμα $X = (X_1,\ldots,X_k)^t$ να είναι συνεχές.

Διακριτές τυχαίες μεταβλητές.

Αν $X = (X_1,\ldots,X_k)^t$ ένα διακριτό τυχαίο διάνυσμα, τότε, προφανώς, οι τυχαίες μεταβλητές $Y_i = g_i(X_1,\ldots,X_k)$, $i = 1,\ldots,m$ είναι διακριτές και ισχύει:

$$\begin{aligned} p_{Y_1,\ldots,Y_m}(y_1,\ldots,y_m) &= P(Y_1 = y_1,\ldots,Y_m = y_m) \\ &= P\left(g_1(X_1,\ldots,X_k) = y_1,\ldots,g_m(X_1,\ldots,X_k)\right) = y_m \\ &= \sum_{(x_1,\ldots,x_k):g_i(x_1,\ldots,x_k)=y_i,i=1,\ldots,m} p_{X_1,\ldots,X_m}(x_1,\ldots,x_m), \end{aligned}$$

όπου $p_{Y_1,\ldots,Y_m}$ και $p_{X_1,\ldots,X_m}$ η από κοινού συνάρτηση πιθανότητας των $(Y_1,\ldots,Y_m)^t$ και $(X_1,\ldots,X_m)^t$, αντίστοιχα.

Η μέθοδος διευκρινίζεται μέσω του επόμενου παραδείγματος.

Παράδειγμα 7.3

Έστω X και Y οι τ.μ. που παριστάνουν τις ενδείξεις δύο τίμιων ζαριών. Να υπολογιστεί η από κοινού συνάρτηση πιθανότητας των Z και U, όπου Z και U οι τυχαίες μεταβλητές που παριστάνουν το άθροισμα και την απόλυτη διαφορά αντίστοιχα των ενδείξεων των δύο ζαριών, δηλαδή $Z = X + Y$ και $U = |X - Y|$. Να προσδιορίσετε έπειτα και την από κοινού αθροιστική συνάρτηση κατανομής των Z και U.

Λύση Παραδείγματος 7.3

Οι τυχαίες μεταβλητές X και Y λαμβάνουν τις τιμές $\{1,2,3,4,5,6\}$ καθεμία με πιθανότητα 1/6. Οι τυχαίες μεταβλητές Z και U λαμβάνουν τις τιμές $\{2,3,4,\ldots,11,12\}$ και $\{0,1,2,3,4,5\}$ αντίστοιχα, όπως φαίνεται και στον επόμενο πίνακα που παρουσιάζονται οι τιμές (z,u) των Z και U με βάση τις τιμές των X και Y.

	Y					
X	1	2	3	4	5	6
1	(2,0)	(3,1)	(4,2)	(5,3)	(6,4)	(7,5)
2	(3,1)	(4,0)	(5,1)	(6,2)	(7,3)	(8,4)
3	(4,2)	(5,1)	(6,0)	(7,1)	(8,2)	(9,3)
4	(5,3)	(6,2)	(7,1)	(8,0)	(9,1)	(10,2)
5	(6,4)	(7,3)	(8,2)	(9,1)	(10,0)	(11,1)
6	(7,5)	(8,4)	(9,3)	(10,2)	(11,1)	(12,0)

Λόγω της υπόθεσης ότι έχουμε δύο «τίμια» ζάρια, κάθε ζευγάρι (x,y) έχει πιθανότητα 1/36 και, επομένως, η πιθανότητα κάθε ζευγαριού (z,u) ισούται με το πλήθος των κελιών, στα οποία αυτό εμφανίζεται στον παραπάνω πίνακα, πολλαπλασιασμένο με το 1/36. Επομένως, η από κοινού συνάρτηση πιθανότητας των Z και U μπορεί να δοθεί στον παρακάτω πίνακα:

$p_{Z,U}(z,u)$	U					
Z	0	1	2	3	4	5
2	1/36	0	0	0	0	0
3	0	2/36	0	0	0	0
4	1/36	0	2/36	0	0	0
5	0	2/36	0	2/36	0	0
6	1/36	0	2/36	0	2/36	0
7	0	2/36	0	2/36	0	2/36
8	1/36	0	2/36	0	2/36	0
9	0	2/36	0	2/36	0	0
10	1/36	0	2/36	0	0	0
11	0	2/36	0	0	0	0
12	1/36	0	0	0	0	0

Από τον πίνακα αυτόν μπορούμε να υπολογίσουμε και την από κοινού συνάρτηση κατανομής των Z και U, όπως αποτυπώνεται στη συνέχεια:

$F_{Z,U}(z,u)$	U					
Z	0	1	2	3	4	5
2	1/36	1/36	1/36	1/36	1/36	1/36
3	1/36	3/36	3/36	3/36	3/36	3/36
4	2/36	4/36	6/36	6/36	6/36	6/36
5	2/36	6/36	8/36	10/36	10/36	10/36
6	3/36	7/36	11/36	13/36	15/36	15/36
7	3/36	9/36	13/36	17/36	19/36	21/36
8	4/36	10/36	16/36	20/36	24/36	26/36
9	4/36	12/36	18/36	24/36	28/36	30/36
10	5/36	13/36	21/36	27/36	31/36	33/36
11	5/36	15/36	23/36	29/36	33/36	35/36
12	6/36	16/36	24/36	30/36	34/36	36/36

Συνεχείς τυχαίες μεταβλητές.

Η δεύτερη μέθοδος προσδιορισμού της κατανομής συναρτήσεων τ.μ. έχει ιδιαίτερο ενδιαφέρον στην περίπτωση που οι τυχαίες μεταβλητές είναι συνεχείς. Όπως είχαμε δει και στην Ενότητα 3.7, έχουμε περιπτώσεις μετασχηματισμών που είναι είτε εξ ολοκλήρου αμφιμονοσήμαντοι είτε κομματιαστά αμφιμονοσήμαντοι. Στο παρόν σύγγραμμα και στο πλαίσιο της πολυδιάστατης περίπτωσης θα περιοριστούμε στην περίπτωση των εξ ολοκλήρου αμφιμονοσήμαντων μετασχηματισμών. Για την περίπτωση των κομματιαστά αμφιμονοσήμαντα μετασχηματισμών ενδεικτικά παραπέμπουμε τον/την αναγνώστη/στρια στο σύγγραμμα του Παπαϊωάννου (1997). Για την περίπτωση των συνεχών τυχαίων μεταβλητών και εξ ολοκλήρου αμφιμονοσήμαντων μεταβλητών έχουμε το ακόλουθο θεώρημα, η απόδειξη του οποίου παραλείπεται και παραπέμπουμε, μεταξύ άλλων, στο σύγγραμμα Παπαϊωάννου (1997).

Θεώρημα 7.1

Έστω συνεχές τυχαίο διάνυσμα $X = (X_1, \dots, X_k)^t$ με από κοινού συνάρτηση πυκνότητας πιθανότητας $f_X(x) = f_{X_1,\dots,X_k}(x_1, \dots, x_k)$ και σύνολο δυνατών τιμών S_X. Θεωρούμε τους μετασχηματισμούς $Y_i = g_i(X)$, όπου $g_i(\cdot)$ πραγματικές συναρτήσεις και έστω $T = \{y = (y_1, \dots, y_k) : y_i = g_i(x), x \in S_X, i = 1, \dots, k\}$. Υποθέτουμε ότι

1. για κάθε $(y_1, y_2, \dots, y_k)$ υπάρχει μοναδικό $(x_1^*, x_2^*, \dots, x_k^*)$, τέτοιο ώστε

$$y_1 = g_1(x_1^*, x_2^*, \dots, x_k^*), y_2 = g_2(x_1^*, x_2^*, \dots, , x_k^*), \dots, y_k = g_k(x_1^*, x_2^*, \dots, , x_k^*)$$

2. και οι αντίστροφες συναρτήσεις $x_1^* = g_1^{-1}(y_1, y_2, \dots, y_k)$, $x_2^* = g_2^{-1}(y_1, y_2, \dots, y_k)$,…, $x_k^* = g_k^{-1}(y_1, y_2, \dots, y_k)$
   - έχουν συνεχείς μερικές παραγώγους πρώτης τάξης και
   - ικανοποιούν τη συνθήκη $|J(y_1, y_2, \dots, y_k)| \neq 0, \forall (y_1, y_2, \dots, y_k) \in T$ για την ορίζουσα του Ιακωβιανού πίνακα

$$J(y_1, y_2, \dots, y_k) = \begin{bmatrix} \frac{\partial g_1^{-1}(y_1,y_2,\dots,y_k)}{\partial y_1} & \frac{\partial g_1^{-1}(y_1,y_2,\dots,y_k)}{\partial y_2} & \dots & \frac{\partial g_1^{-1}(y_1,y_2,\dots,y_k)}{\partial y_k} \\ \frac{\partial g_2^{-1}(y_1,y_2,\dots,y_k)}{\partial y_1} & \frac{\partial g_2^{-1}(y_1,y_2,\dots,y_k)}{\partial y_2} & \dots & \frac{\partial g_2^{-1}(y_1,y_2,\dots,y_k)}{\partial y_k} \\ \dots & \dots & \dots & \dots \\ \frac{\partial g_k^{-1}(y_1,y_2,\dots,y_k)}{\partial y_1} & \frac{\partial g_k^{-1}(y_1,y_2,\dots,y_k)}{\partial y_2} & \dots & \frac{\partial g_k^{-1}(y_1,y_2,\dots,y_k)}{\partial y_k} \end{bmatrix}.$$

Τότε, η από κοινού συνάρτηση πυκνότητας πιθανότητας των $Y_1, Y_2, \dots, Y_k$ δίνεται από τη σχέση:

$$f_{Y_1,Y_2,\dots,Y_k}(y_1, y_2, \dots, y_k) = f_{X_1,X_2,\dots,X_k}(x_1^*, x_2^*, \dots, x_k)|J|.$$

Παρατήρηση 7.1

Όταν μας ενδιαφέρει η κατανομή m το πλήθος τυχαίων μεταβλητών $Y_i = g_i(X_1, X_2, \dots, X_k)$, $i = 1, \dots, m$ με $m < k$, το παραπάνω θεώρημα εφαρμόζεται συμπληρώνοντας αυθαίρετα τις υπόλοιπες $k - m$ το πλήθος μεταβλητές, όπως, για παράδειγμα, θέτοντας, αυθαίρετα, ότι $Y_i = X_i$ για $i = m + 1, \dots, k$. Με αυτόν τον τρόπο αρχικά, προσδιορίζεται η από κοινού συνάρτηση πυκνότητας πιθανότητας των $Y_1, \dots, Y_k$ και στη συνέχεια, από αυτήν η περιθώρια των $Y_1, \dots, Y_m$.

Στη συνέχεια, δίνονται κάποια παραδείγματα για την κατανόηση της μεθόδου.

Παράδειγμα 7.4

Η από κοινού συνάρτηση πυκνότητας πιθανότητας των X και Y δίνεται από τη σχέση

$$f_{X,Y}(x, y) = \begin{cases} 2e^{-2x-y}, & x, y > 0, \\ 0, & \text{αλλού.} \end{cases}$$

Να βρεθεί η συνάρτηση πυκνότητας πιθανότητας της τυχαίας μεταβλητής X/Y.

Λύση Παραδείγματος 7.4

Αρχικά, ορίζουμε τις μεταβλητές $W = g_1(X, Y) = X/Y$ και $Z = g_2(X, Y) = Y$. Από τον ορισμό των συναρτήσεων $g_1(X, Y)$ και $g_2(X, Y)$ έχουμε ότι:

1. για κάθε (w, z) υπάρχει μοναδικό (x^*, y^*), τέτοιο ώστε

$$w = g_1(x^*, y^*), z = g_2(x^*, y^*)$$

2. οι αντίστροφες συναρτήσεις $x^* = g_1^{-1}(w, z) = wz$, $y^* = g_2^{-1}(w, z) = z$
   - έχουν προφανώς συνεχείς μερικές παραγώγους πρώτης τάξης και

- ικανοποιούν τη συνθήκη $|J(w,z)| \neq 0, \forall (w,z) \in \mathbb{R}^2_+$, αφού

$$J(w,z) = \begin{bmatrix} z & w \\ 0 & 1 \end{bmatrix}$$

και, επομένως, $|J(w,z)| = z \neq 0$.

Άρα σύμφωνα με το Θεώρημα 7.1, η από κοινού συνάρτηση πυκνότητας πιθανότητας των W και Z δίνεται από τη σχέση:

$$\begin{aligned} f_{W,Z}(w,z) &= f_{X,Y}\left(x^*, y^*\right)|J| \\ &= 2e^{-2wz-z}z \\ &= 2ze^{-z(2w+1)}, \text{ με } w,z > 0. \end{aligned}$$

Για να βρούμε την κατανομή της X/Y αρκεί να υπολογίσουμε την περιθώρια της W, η οποία δίνεται από τη σχέση:

$$f_W(w) = \int_0^{+\infty} 2ze^{-z(2w+1)}dz = 2\frac{1}{w+1}\int_0^{+\infty} z(w+1)e^{-z(w+1)}dz.$$

Όμως, η υπό ολοκλήρωση ποσότητα είναι η αναμενόμενη τιμή μιας εκθετικής κατανομής με παράμετρο $w+1$ και, επομένως, έχουμε ότι η συνάρτηση πυκνότητας πιθανότητας της W ισούται με:

$$f_W(w) = 2\frac{1}{w+1} \cdot \frac{1}{w+1} = \frac{2}{(w+1)^2},\ w > 0.$$

Μια ενδιαφέρουσα εφαρμογή του Θεωρήματος 7.1 προκύπτει κατά τον προσδιορισμό της συνάρτησης πυκνότητας πιθανότητας του αθροίσματος δύο συνεχών τυχαίων μεταβλητών X_1 και X_2. Εφαρμόζοντας το θεώρημα για τις συνεχείς τυχαίες μεταβλητές $Y_1 = X_1$ και $Y_2 = X_1 + X_2$, προκύπτει η παρακάτω σχέση για τη συνάρτηση πυκνότητας πιθανότητας της Y_2:

$$f_{Y_2}(y_2) = \int_{-\infty}^{\infty} f_{X_1,X_2}(y_1, y_2 - y_1)dy_1,$$

η οποία σχέση αναφέρεται και ως τύπος της συνέλιξης. Στην περίπτωση διακριτών τυχαίων μεταβλητών, η αντίστοιχη σχέση για τη συνάρτηση πιθανότητας είναι:

$$p_{Y_2}(y_2) = \sum_{y_1} p_{X_1,X_2}(y_1, y_2 - y_1).$$

Παράδειγμα 7.5

Μια συσκευή αποτελείται από δύο ανεξάρτητα υποσυστήματα. Όταν τίθεται σε λειτουργία, ενεργοποιείται το πρώτο υποσύστημα. Μόλις παρουσιάσει βλάβη στο πρώτο υποσύστημα ενεργοποιείται αυτόματα το δεύτερο υποσύστημα έτσι ώστε να συνεχίσει απρόσκοπτα η λειτουργία της συσκευής. Η διάρκεια ζωής σε ώρες κάθε υποσυστήματος ακολουθεί την εκθετική κατανομή με μέση τιμή τις 1500 ώρες. Να βρεθεί η συνάρτηση πυκνότητας πιθανότητας του συνολικού χρόνου λειτουργίας της συσκευής. Ποιος είναι ο μέσος χρόνος λειτουργίας της συσκευής;

Λύση Παραδείγματος 7.5

Έστω X_1 και X_2 οι τυχαίες μεταβλητές που παριστάνουν σε ώρες τη διάρκεια ζωής κάθε υποσυστήματος. Μας ζητείται η συνάρτηση πυκνότητας πιθανότητας της τυχαίας μεταβλητής που παριστάνει τον συνολικό χρόνο λειτουργίας της συσκευής, δηλαδή η συνάρτηση πυκνότητας πιθανότητας της τυχαίας μεταβλητής $Y = X_1 + X_2$, η οποία μπορεί να προσδιοριστεί από τον τύπο της συνέλιξης. Πιο συγκεκριμένα, είναι:

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X_1,X_2}(x_1, y - x_1)dx_1.$$

Από την εκφώνηση της άσκησης έχουμε, λόγω της ανεξαρτησίας των υποσυστημάτων, ότι $f_{X_1,X_2}(x_1,x_2) = f_{X_1}(x_1)f_{X_2}(x_2)$ και, επομένως, ότι

$$f_{X_1,X_2}(x_1, y - x_1) = \frac{1}{1500}e^{-\frac{1}{1500}x_1}\frac{1}{1500}e^{-\frac{1}{1500}(y-x_1)}$$

όταν $x_1 > 0$ και $y - x_1 > 0$ (δηλαδή όταν $0 < x_1 < y$) και 0 αλλού. Επομένως, είναι

$$\begin{aligned} f_Y(y) &= \int_0^y \frac{1}{1500}e^{-\frac{1}{1500}x_1}\frac{1}{1500}e^{-\frac{1}{1500}(y-x_1)}dx_1 \\ &= \frac{y}{2250000}e^{-y/1500},\ y > 0. \end{aligned}$$

Άρα, η τυχαία μεταβλητή Y περιγράφεται από τη γάμμα κατανομή με παραμέτρους 2 και 1500 με $E(Y) = 2 \cdot 1500 = 3000$ ώρες, δηλαδή ο μέσος συνολικός χρόνος λειτουργίας είναι ίσος με 3000 ώρες.

7.2.3 Μέθοδος ροπογεννήτριας

Η τρίτη μέθοδος μετασχηματισμού είναι η μέθοδος της ροπογεννήτριας και βασίζεται στην ιδιότητα του μονοσήμαντου της ροπογεννήτριας (βλ. Θεώρημα 3.1). Ειδικότερα, έστω $X = (X_1, \dots, X_k)^t$ το αρχικό k-διάστατο τυχαίο διάνυσμα του οποίου την κατανομή γνωρίζουμε και $Y_i = g_i(X)$, $i = 1, \dots, m$, m το πλήθος τυχαίες μεταβλητές, την από κοινού κατανομή των οποίων θέλουμε να προσδιορίσουμε. Αρχικά, προσδιορίζουμε τη ροπογεννήτρια συνάρτηση του $Y = (Y_1, \dots, Y_m)^t$ από τη σχέση:

$$M_Y(t_1, \dots, t_m) = E\left(e^{t_1Y_1+\dots+t_mY_m}\right) = E\left(e^{t_1g_1(X)+\dots+t_mg_m(X)}\right).$$

Επομένως, στην περίπτωση διακριτού τυχαίου διανύσματος έχουμε

$$M_Y(t_1, \dots, t_m) = \sum_{x \in S_X} e^{t_1g_1(X)+\dots+t_mg_m(X)}p_X(x_1, \dots, x_k),$$

ενώ στην περίπτωση συνεχούς τυχαίου διανύσματος έχουμε

$$M_Y(t_1, \dots, t_m) = \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} e^{t_1g_1(X)+\dots+t_mg_m(X)}f_X(x_1, \dots, x_k)dx_1 \cdots dx_k.$$

Αν η ροπογεννήτρια που θα προκύψει ταυτίζεται με τη ροπογεννήτρια γνωστής κατανομής, λόγω του μονοσημάντου, έχει κατ' ουσίαν προσδιοριστεί η κατανομή του τυχαίου διανύσματος Y. Η μέθοδος αυτή αποδεικνύεται κάποιες φορές, όπως θα δούμε στην επόμενη ενότητα, ιδιαίτερα χρήσιμη για τον προσδιορισμό της κατανομής του αθροίσματος ανεξάρτητων τυχαίων μεταβλητών.

7.2.4 Κατανομή αθροίσματος ανεξάρτητων τυχαίων μεταβλητών

Στο Παράδειγμα 7.5 παρουσιάστηκε η κατανομή του αθροίσματος δύο ανεξάρτητων τυχαίων μεταβλητών. Συχνά όμως ενδιαφερόμαστε για το άθροισμα περισσότερων των δύο ανεξάρτητων τυχαίων μεταβλητών. Για τον προσδιορισμό της κατανομής του αθροίσματος σε αυτήν την περίπτωση ιδιαίτερα χρήσιμη, κάποιες φορές, είναι η μέθοδος της ροπογεννήτριας.

Έστω X_1, X_2,...,X_k, k το πλήθος ανεξάρτητες τυχαίες μεταβλητές, τότε η ροπογεννήτρια συνάρτηση της τυχαίας μεταβλητής $Y = X_1 + ... + X_k$ δίνεται από τη σχέση

$$M_Y(y) = E(e^{tY}) = E(e^{t(X_1+...+X_k)}) = E(e^{tX_1} \cdots e^{tX_k}) = E(e^{tX_1}) \cdots E(e^{tX_k}).$$

Σύμφωνα με τη μέθοδο της ροπογεννήτριας, αν το γινόμενο στο αριστερό μέρος αναγνωριστεί ως ροπογεννήτρια γνωστής κατανομής, τότε έχει προσδιοριστεί και η κατανομή του αθροίσματος.

Πρόταση 7.3

Έστω X_1, X_2,...,X_k, k το πλήθος ανεξάρτητες τυχαίες μεταβλητές και $Y = X_1 + ... + X_k$. Τότε,

- αν οι τ.μ. X_i, $i = 1, ..., k$, ακολουθούν κατανομή Poisson με παράμετρο λ, το άθροισμα Y ακολουθεί κατανομή Poisson με παράμετρο $k\lambda$,
- αν οι τ.μ. X_i, $i = 1, ..., k$, ακολουθούν $Bernoulli(p)$, το άθροισμα Y ακολουθεί $B(k,p)$ κατανομή,
- αν οι τ.μ. X_i, $i = 1, ..., k$, ακολουθούν $Exp(\lambda)$, το άθροισμα Y ακολουθεί $Erlang(k,\lambda)$,
- αν οι τ.μ. X_i, $i = 1, ..., k$, ακολουθούν κατανομή γάμμα με παραμέτρους (a_i,λ), το άθροισμα Y ακολουθεί κατανομή γάμμα με παραμέτρους $\sum_{i=1}^{k} a_i$ και λ,
- αν οι τ.μ. X_i, $i = 1, ..., k$, ακολουθούν $N(\mu_i,\sigma_i^2)$, το άθροισμα Y ακολουθεί κανονική κατανομή με παραμέτρους $\sum_{i=1}^{k} \mu_i$ και $\sum \sigma_i^2$.

Απόδειξη Πρότασης 7.3

Καθώς οι τ.μ. X_i, $i = 1, ..., k$, είναι ανεξάρτητες, ισχύει ότι:

$$M_Y(t) = E(e^{tY}) = E(e^{t(X_1+...+X_k)}) = E(e^{tX_1} \cdots e^{tX_k})$$

$$= E(e^{tX_1}) \cdots E(e^{tX_k}) = \prod_{i=1}^{k} M_{X_i}(t).$$

- Σε αυτήν την περίπτωση είναι (βλ. τη σχέση (4.43) για τη ροπογεννήτρια της Poisson) $M_{X_i}(t) = \exp[\lambda(e^t - 1)]$. Επομένως, είναι

$$M_Y(t) = \prod_{i=1}^{k} \exp[\lambda(e^t - 1)] = \exp[k\lambda(e^t - 1)],$$

  που ταυτίζεται με τη ροπογεννήτρια της Poisson με παράμετρο $k\lambda$.

- Σε αυτήν την περίπτωση είναι (βλ. τη σχέση (4.11) για $n = 1$) $M_{X_i}(t) = \left(q + pe^t\right)$. Επομένως, προκύπτει άμεσα ότι $m_Y(t) = \left(q + pe^t\right)^k$ που ταυτίζεται με τη ροπογεννήτρια της διωνυμικής με παραμέτρους k και p.

- Σε αυτήν την περίπτωση είναι (βλ. τη σχέση (5.20) για τη ροπογεννήτρια της Εκθετικής)

$$M_{X_i}(t) = \frac{\lambda}{\lambda - t}, \quad \text{για } t < \lambda.$$

Επομένως, προκύπτει ότι:

$$M_Y(t) = \left(\frac{\lambda}{\lambda - t}\right)^k = \left(\frac{\lambda - t}{\lambda}\right)^{-k} = \left(1 - \frac{t}{\lambda}\right)^{-k}, t < \lambda.$$

Παρατηρούμε ότι η ροπογεννήτρια αυτή ταυτίζεται με αυτήν που προκύπτει από τη σχέση (5.32) που δίνει τη ροπογεννήτρια της γάμμα κατανομής με παραμέτρους k και λ με σππ που προσδιορίζεται στη σχέση (5.27). Καθώς k ακέραιος, έχουμε ότι πρόκειται ουσιαστικά για την κατανομή Erlang.

- Σε αυτήν την περίπτωση είναι (βλ. σχέση (5.32) για τη ροπογεννήτρια της Γάμμα)

$$M_{X_i}(t) = \left(1 - \frac{t}{\lambda}\right)^{-a_i}, \quad \text{για } t < \lambda.$$

Επομένως, προκύπτει ότι:

$$M_Y(t) = \left(1 - \frac{t}{\lambda}\right)^{-\sum_{i=1}^k a_i}, \quad \text{για } t < \lambda,$$

η οποία ταυτίζεται με τη ροπογεννήτρια της γάμμα κατανομής με παραμέτρους $\sum_{i=1}^k a_i$ και λ με σππ που προσδιορίζεται στη σχέση (5.27).

- Σε αυτήν την περίπτωση είναι (βλ. σχέση (5.37) για τη ροπογεννήτρια της κανονικής)

$$M_{X_i}(t) = \exp\left\{\mu_i\, t + \frac{1}{2}\sigma_i^2\, t^2\right\}, t \in \mathbb{R}.$$

Επομένως, προκύπτει ότι:

$$M_Y(t) = \exp\left\{t\sum_{i=1}^k \mu_i + \frac{1}{2}t^2\sum_{i=1}^k \sigma_i^2\right\}, t \in \mathbb{R},$$

η οποία ταυτίζεται με τη ροπογεννήτρια της κανονικής κατανομής με παραμέτρους $(\sum_{i=1}^k \mu_i,\ \sum_{i=1}^k \sigma_i^2)$.

Πρόταση 7.4

Έστω $X_1, X_2,\ldots,X_k$, k το πλήθος ανεξάρτητες τυχαίες μεταβλητές με $X_i \sim N(\mu_i, \sigma_i^2)$, $i = 1, \ldots, k$. Τότε η τυχαία μεταβλητή $Y = a_1X_1 + \ldots + a_kX_k$ ακολουθεί κανονική κατανομή με παραμέτρους $\left(\sum_{i=1}^k a_i\mu_i, \sum_{i=1}^k (a_i\sigma_i)^2\right)$.

Απόδειξη Πρότασης 7.4

Καθώς οι τ.μ. X_i, $i = 1, \ldots, k$, είναι ανεξάρτητες ισχύει ότι:

$$M_Y(y) = E(e^{tY}) = E(e^{t(a_1X_1 + \ldots + a_kX_k)}) = E(e^{ta_1X_1} \cdots e^{ta_kX_k})$$

$$= E(e^{ta_1X_1}) \cdots E(e^{ta_kX_k}) = \prod_{i=1}^k M_{X_i}(a_it).$$

Σε αυτήν την περίπτωση είναι (βλ. σχέση (5.37) για τη ροπογεννήτρια της κανονικής)

$$M_{X_i}(a_i t) = \exp\left\{\mu_i a_i t + \frac{1}{2}\sigma_i^2 (a_i t)^2\right\}, t \in \mathbb{R}.$$

Επομένως, προκύπτει ότι:

$$M_Y(t) = \exp\left\{t\sum_{i=1}^{k} a_i\mu_i + \frac{1}{2}t^2\sum_{i=1}^{k}(a_i\sigma_i)^2\right\}, t \in \mathbb{R},$$

η οποία ταυτίζεται με τη ροπογεννήτρια της κανονικής κατανομής με παραμέτρους $\left(\sum_{i=1}^{k} a_i\mu_i, \sum_{i=1}^{k} a_i^2\sigma_i^2\right)$.

Άσκηση Αυτοαξιολόγησης 7.2

Δείξτε ότι το άθροισμα k ανεξάρτητων κατανομών Poisson με παραμέτρους λ_i ακολουθεί κατανομή Poisson με παράμετρο $\lambda_1 + ... + \lambda_k$.

Άσκηση Αυτοαξιολόγησης 7.3

Έστω X_1, X_2,...,X_k, k το πλήθος ανεξάρτητες τυχαίες μεταβλητές. Προσδιορίστε την κατανομή του αθροίσματος $Y = X_1 + ... + X_k$, όταν $X_i \sim B(n_i, p)$, $i = 1, ..., k$.

Άσκηση Αυτοαξιολόγησης 7.4

Έστω X_1, X_2,...,X_k, k το πλήθος ανεξάρτητες τυχαίες μεταβλητές. Προσδιορίστε την κατανομή του αθροίσματος $Y = X_1 + ... + X_k$, όταν $X_i \sim NB(n_i, p)$, $i = 1, ..., k$.

7.2.5 Κατανομές χι-τετράγωνο, t και F

Στην ενότητα αυτή παρουσιάζονται τρεις πολύ σημαντικές, για τη στατιστική συμπερασματολογία, κατανομές, η χι-τετράγωνο (χ^2), t ή Student και F, οι οποίες προκύπτουν ως συναρτήσεις n το πλήθος ανεξάρτητων και ισόνομων τυχαίων μεταβλητών $X_1,...,X_n$ που ακολουθούν κανονική κατανομή με μέση τιμή μ και διακύμανση σ^2.

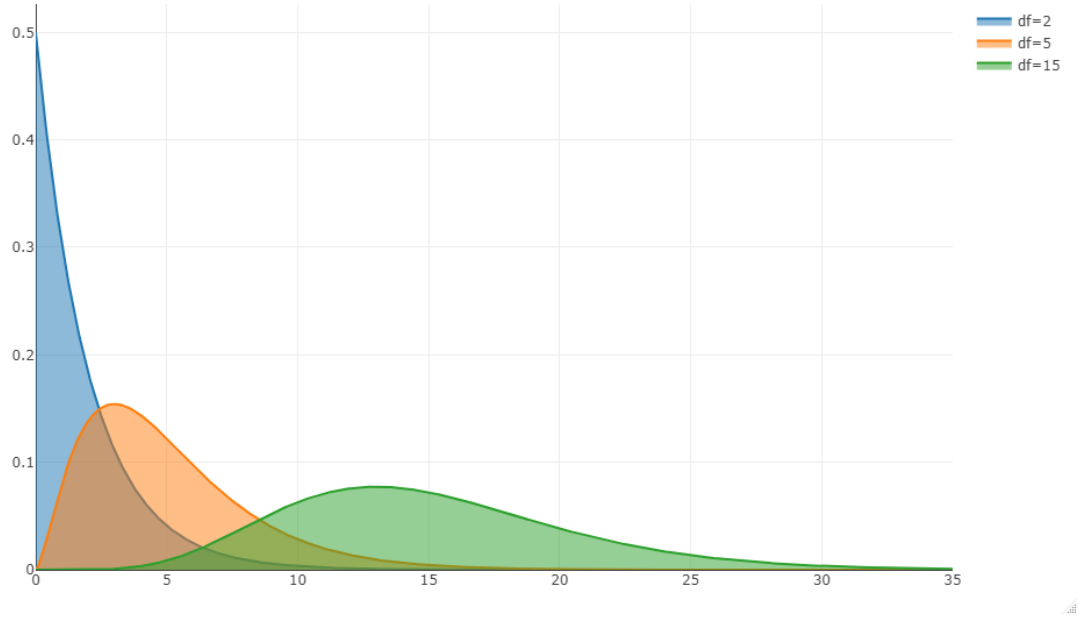
Η πρώτη από τις κατανομές που αναφέρθηκε, η χι-τετράγωνο, έχει ουσιαστικά ήδη οριστεί στην Ενότητα 5.5, ως ειδική περίπτωση της γάμμα κατανομής. Παρόλα αυτά για λόγους πληρότητας παραθέτουμε και στην ενότητα αυτήν τον ορισμό της.

Ορισμός 7.1

Η τυχαία μεταβλητή Y λέγεται ότι ακολουθεί τη **χι-τετράγωνο κατανομή** με n βαθμούς ελευθερίας, αν οι δυνατές της τιμές y είναι $y \geq 0$ και η συνάρτηση πυκνότητας πιθανότητάς της δίνεται από τη σχέση:

$$f_y(y) = \begin{cases} \dfrac{y^{\frac{n-2}{2}} e^{-\frac{y}{2}}}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})}, & y \geq 0, \\ 0, & \text{αλλού}, \end{cases} \tag{7.1}$$

Στην περίπτωση αυτή, γράφουμε ότι $Y \sim \chi^2_n$.



Σχήμα 7.1: Γραφική παράσταση της σππ της χ^2_n για $n = 2, 5$ και 15 βαθμούς ελευθερίας με μπλε, πορτοκαλί και πράσινο χρώμα, αντίστοιχα.

Εύκολα διαπιστώνουμε ότι αποτελεί ειδική περίπτωση της γάμμα κατανομής και ειδικότερα ότι $Y \sim G(a = n/2, \lambda = 0.5)$ ή $Y \sim G(a = n/2, \mu = 2)$, ανάλογα με ποια εκ των παραμετροποιήσεων της γάμμα κατανομής που δίνονται στις σχέσεις (5.27) και (5.28), θα χρησιμοποιηθεί.

Στο Σχήμα 7.1 απεικονίζεται η σππ της χι-τετράγωνο κατανομής για $n = 2, 5$ και 15 βαθμούς ελευθερίας με μπλε, πορτοκαλί και πράσινο χρώμα αντίστοιχα. Οι εντολές που χρησιμοποιήθηκαν στην R, για να γίνει το Σχήμα 7.1, είναι οι ακόλουθες:

```
library(plotly)

x = seq(0, 35, length = 1000)
d1 <- dchisq(x = x,df = 2)
d2 <- dchisq(x = x,df = 5)
d3 <- dchisq(x = x,df = 15)

fig <- plot_ly(x = x, y = d1, type = 'scatter', mode = 'lines',
               name = 'df=2', fill = 'tozeroy')
fig <- fig %>% add_trace(x = x, y = d2, name = 'df=5', fill = 'tozeroy')
fig <- fig %>% add_trace(x = x, y = d3, name = 'df=15', fill = 'tozeroy')
fig
```

Από το Σχήμα 7.1 παρατηρούμε ότι η χ^2_n κατανομή έχει θετική λοξότητα (αφήνεται ως άσκηση στον/στην αναγνώστη/στρια να αποδείξει ότι ο συντελεστής λοξότητάς της είναι $\alpha_3 = 4/\sqrt{2n} > 0$. Υπόδειξη: χρησιμοποιήστε τα αποτελέσματα της πρότασης που ακολουθεί). Επίσης, παρατηρούμε ότι όσο μεγαλώνει το n, δηλαδή όσο αυξάνονται οι βαθμοί ελευθερίας, τόσο το α_3 μικραίνει και για $n \to \infty$ προκύπτει ότι $\alpha_3 \to 0$ και η χ^2_n κατανομή τείνει να γίνει συμμετρική.

Άμεσα από τα αποτελέσματα που έχουν δοθεί για τη γάμμα κατανομή προκύπτουν τα ακόλουθα.

Πρόταση 7.5

Έστω X η τυχαία μεταβλητή που ακολουθεί χι τετράγωνο κατανομή με n βαθμούς ελευθερίας με σππ που προσδιορίζεται στη σχέση (7.1). Τότε:

$$E(X^k) = \frac{2^k \Gamma\left(\frac{n+2k}{2}\right)}{\Gamma\left(\frac{n}{2}\right)}, \tag{7.2}$$

με

$$E(X) = n \text{ και } Var(X) = 2n,$$

ενώ

$$M_X(t) = (1-2t)^{-\frac{n}{2}}, \quad \text{για } t < 0.5. \tag{7.3}$$

Απόδειξη Πρότασης 7.5

Προκύπτουν άμεσα από τις σχέσεις (5.31) και (5.32) για $a = n/2$ και $\lambda = 0.5$.

Κάποιες φορές χρειάζεται να υπολογίσουμε την τιμή του σημείου εκείνου που είναι τέτοιο, ώστε να προκύπτει κάποια δοθείσα τιμή για την πιθανότητα του ενδεχομένου η χ^2_n κατανομή να πάρει τιμές μεγαλύτερες από το σημείο αυτό. Για τα σημεία αυτά έχουμε τον ακόλουθο ορισμό.

Ορισμός 7.2

Έστω $X \sim \chi^2_n$. Ορίζουμε ως $\chi^2_{n,\alpha}$ την τιμή της χ^2_n κατανομής που είναι τέτοια, ώστε

$$P(X > \chi^2_{n,\alpha}) = \alpha, \text{ για } \alpha \in (0,1) \text{ γνωστό αριθμό.} \tag{7.4}$$

Υπάρχουν πίνακες που δίνουν τα σημεία $\chi^2_{n,\alpha}$ (που συχνά αναφέρονται ως συμπληρωματικά α-ποσοστιαία σημεία) της χ^2_n κατανομής για διάφορες τιμές των βαθμών ελευθερίας n και της πιθανότητας α (βλ. τους Πίνακες Α΄.5 και Α΄.6 του Παραρτήματος Α΄).

Εναλλακτικά, μπορεί να χρησιμοποιηθεί η R για τον προσδιορισμό οποιουδήποτε ποσοστιαίου σημείου της χ^2_n κατανομής ή πιθανότητας.

Παρατήρηση 7.2

Έστω $X \sim \chi^2_n$ με σππ που δίνεται από τη σχέση (7.1). Τότε με τη γλώσσα R μπορούμε:

- με τη συνάρτηση `dchisq(x,df=n,ncp=0)` να υπολογίσουμε τη σππ στο σημείο ή στο διάνυσμα σημείων x,
- με τη συνάρτηση `pchisq(x,df=n,ncp=0,lower.tail=TRUE)` να υπολογίσουμε την ασκ στο σημείο ή στο διάνυσμα σημείων x,
- με τη συνάρτηση `pchisq(x,df=n,ncp=0,lower.tail=FALSE)` να υπολογίσουμε την πιθανότητα $P(X > x)$ στο σημείο ή στο διάνυσμα σημείων x,
- με τη συνάρτηση `qchisq(q,df=n,ncp=0,lower.tail=TRUE)` να προσδιορίσουμε το σημείο ή το διάνυσμα σημείων x για τα οποία ισχύει ότι $P(X \leq x) = q$, όπου q είναι πιθανότητα ή διάνυσμα με συνιστώσες πιθανότητες, αντίστοιχα,
- με τη συνάρτηση `qchisq(q,df=n,ncp=0,lower.tail=FALSE)` να προσδιορίσουμε το σημείο ή το διάνυσμα σημείων x για τα οποία ισχύει ότι $P(X > x) = q$, όπου q είναι πιθανότητα ή διάνυσμα με συνιστώσες πιθανότητες, αντίστοιχα,
- με τη συνάρτηση `rchisq(k,df=n,ncp=0)` να δημιουργήσουμε ένα δείγμα μεγέθους k από αυτήν την κατανομή.

Ο τρόπος που συνδέεται η χι-τετράγωνο κατανομή με την κανονική κατανομή προσδιορίζεται στην επόμενη πρόταση.

Πρόταση 7.6

Έστω $X_1, \dots, X_n$ είναι n το πλήθος ανεξάρτητες και ισόνομες τυχαίες μεταβλητές με $X_i \sim N(\mu, \sigma^2)$ $i = 1, \dots, n$. Τότε ισχύουν τα ακόλουθα:

1. $Y_i = \left(\frac{X_i - \mu}{\sigma}\right)^2 \sim \chi^2_1$.
2. $\sum_{i=1}^n Y_i \sim \chi^2_n$.

Απόδειξη Πρότασης 7.6

Όταν $X_1, \dots, X_n$ είναι n το πλήθος ανεξάρτητες και ισόνομες τυχαίες μεταβλητές με $X_i \sim N(\mu, \sigma^2)$ $i = 1, \dots, n$, τότε άμεσα από τον τυπικό μετασχηματισμό προκύπτει ότι

$$Z_i = \frac{X_i - \mu}{\sigma},\ i = 1, \dots, n,$$

είναι (ως συναρτήσεις ανεξάρτητων) ανεξάρτητες και ισόνομες τυχαίες μεταβλητές με $Z_i \sim N(0,1)$, $i = 1, \dots, n$. Επομένως, αρκεί να δείξουμε ότι $Z_i^2 \sim \chi^2_1$ και $\sum_{i=1}^n Z_i^2 \sim \chi^2_n$.

1. Για την απόδειξη θα χρησιμοποιηθεί η μέθοδος μετασχηματισμού της αθροιστικής συνάρτησης κατανομής που παρουσιάστηκε στην Ενότητα 3.7. Είναι

$$F_{Z_i^2}(z) = P(Z_i^2 \le z) = P(-\sqrt{z} \le Z_i \le \sqrt{z}) = \Phi(\sqrt{z}) - \Phi(-\sqrt{z}),\ \text{για } z \in \mathbb{R},$$

όπου $\Phi(\cdot)$ η αθροιστική συνάρτηση κατανομής της τυπικής κανονικής κατανομής. Επομένως, η συνάρτηση πυκνότητας πιθανότητας της Z_i^2 είναι

$$\begin{aligned} f_{Z_i^2}(z) &= \frac{d}{dz} F_{Z_i^2}(z) = \frac{d}{dz}(\Phi(\sqrt{z}) - \Phi(-\sqrt{z})) \\ &= \frac{1}{2\sqrt{z}}\phi(\sqrt{z}) - \left(-\frac{1}{2\sqrt{z}}\right)\phi(\sqrt{z}) \\ &= \frac{1}{2\sqrt{z}}\frac{1}{\sqrt{2\pi}}e^{-z/2} + \frac{1}{2\sqrt{z}}\frac{1}{\sqrt{2\pi}}e^{-z/2} \\ &= \frac{z^{1/2-1}e^{-z/2}}{2^{1/2}\sqrt{\pi}} = \frac{z^{1/2-1}e^{-z/2}}{2^{1/2}\Gamma(1/2)},\ \text{για } z \in \mathbb{R}, \end{aligned}$$

όπου $\phi(\cdot)$ η σπ της τυπικής κανονικής και χρησιμοποιήσαμε ότι $\Gamma(1/2) = \sqrt{\pi}$. Παρατηρήστε ότι η σπ της Z_i^2 ταυτίζεται με τη σπ της γάμμα κατανομής με παραμέτρους $(0.5, 2)$ ή, ισοδύναμα, από τη σχέση (7.1) με τη σπ της χι-τετράγωνο κατανομής με 1 βαθμό ελευθερίας.

2. Δείξαμε ότι $Y_i = Z_i^2 \sim \chi^2_1$. Ακολουθώντας την Πρόταση 7.3 και χρησιμοποιώντας τη σχέση (7.3), έχουμε ότι

$$M_{\Sigma_{i=1}^n Y_i}(t) = \prod_{i=1}^n M_{Y_i}(t) = \prod_{i=1}^n (1-2t)^{-\frac{1}{2}} = (1-2t)^{-\frac{n}{2}},\ \text{για } t < 0.5.$$

Επομένως, ταυτίζεται με τη ροπογεννήτρια της χι-τετράγωνο κατανομής με n βαθμούς ελευθερίας και αποδεικνύεται το ζητούμενο.

Στην επόμενη πρόταση παρατίθεται μια εφαρμογή της μεθόδου του μετασχηματισμού που ουσιαστικά οδηγεί στον ορισμό μιας νέας κατανομής.

Πρόταση 7.7

Έστω Z τυχαία μεταβλητή που ακολουθεί τυπική κανονική κατανομή, $Z \sim N(0,1)$, και Y τυχαία μεταβλητή που ακολουθεί χι-τετράγωνο κατανομή με n βαθμούς ελευθερίας, $Y \sim \chi^2_n$. Υπό την υπόθεση της ανεξαρτησίας των Y και Z, η τυχαία μεταβλητή $T = \frac{Z}{\sqrt{Y/n}}$ έχει συνάρτηση πυκνότητας πιθανότητας:

$$f_T(t) = \frac{\Gamma((n+1)/2)}{\sqrt{\pi n}\cdot\Gamma(n/2)} \cdot \left(1 + \frac{t^2}{n}\right)^{-(n+1)/2}, \; t \in \mathbb{R}. \tag{7.5}$$

Τότε λέμε ότι η τυχαία μεταβλητή T ακολουθεί **κατανομή t ή κατανομή του Student με n βαθμούς ελευθερίας** και συμβολίζουμε $T \sim t_n$.

Απόδειξη Πρότασης 7.7

Η συνάρτηση πυκνότητας πιθανότητας της τυχαίας μεταβλητής $Z \sim N(0,1)$ είναι:

$$f_Z(z) = (2\pi)^{-\frac{1}{2}} e^{-\frac{z^2}{2}}, \; z \in \mathbb{R},$$

ενώ η συνάρτηση πυκνότητας πιθανότητας της τυχαίας μεταβλητής $Y \sim \chi^2_n$ δίνεται από τη σχέση:

$$f_y(y) = \frac{y^{\frac{n-2}{2}} e^{-\frac{y}{2}}}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)}, \; y \geq 0.$$

Καθώς οι τυχαίες μεταβλητές Z και Y είναι ανεξάρτητες, η από κοινού συνάρτηση πυκνότητας πιθανότητάς τους είναι το γινόμενό τους, δηλαδή

$$f_{Z,Y}(z,y) = f_Z(z) f_Y(y), \; (z,y) \in \mathbb{R} \times (0,+\infty).$$

Καθώς θέλουμε να βρούμε την κατανομή της τυχαίας μεταβλητής $T = \frac{Z}{\sqrt{Y/n}}$ σύμφωνα με το Θεώρημα 7.1, συμπληρώνουμε τον μετασχηματισμό με μία ακόμη συνάρτηση. Στο πλαίσιο αυτό θεωρούμε την $U = Y$. Από τη λύση του συστήματος που προκύπτει έχουμε ότι:

$$Y = U \text{ και } Z = T\sqrt{U/n}$$

με Ιακωβιανή ορίζουσα που προκύπτει, ύστερα από λίγη άλγεβρα, να είναι ίση με $J = \sqrt{u/n}$. Επομένως, η από κοινού συνάρτηση πυκνότητας πιθανότητας των T και U είναι:

$$f_{T,U}(t,u) = f_{Z,Y}(t\sqrt{u/n}, u)|J|, \; u > 0, t \in \mathbb{R},$$

και, επομένως,

$$f_T(t) = \int_0^{+\infty} f_{T,U}(t,u)du, t \in \mathbb{R}.$$

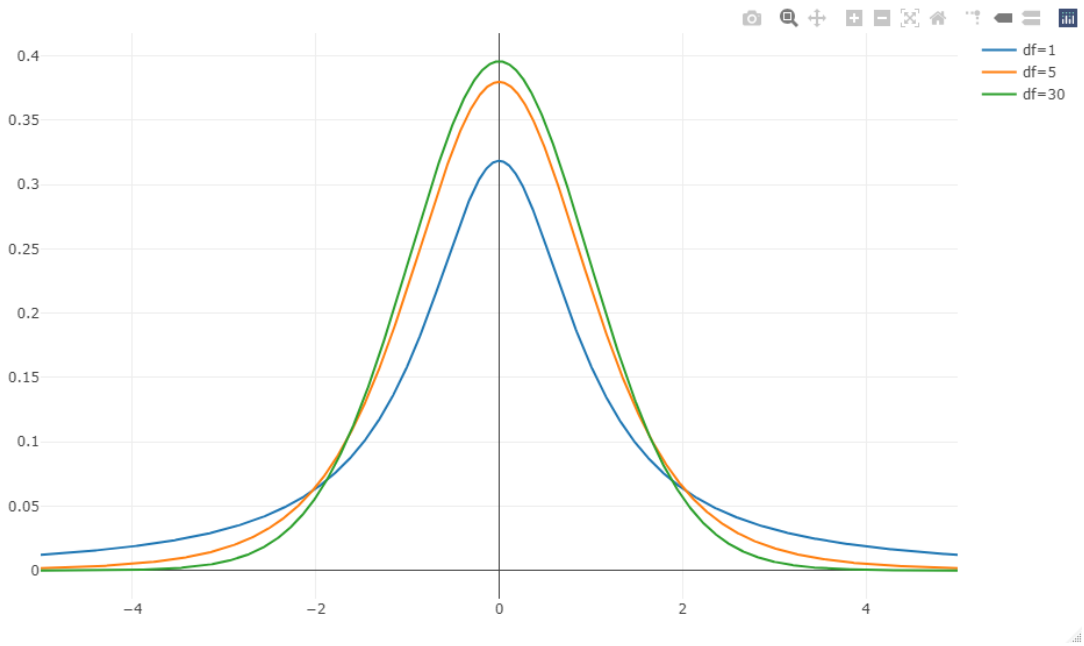
Με αντικατάσταση έχουμε

$$f_T(t) = \frac{2^{-\frac{n+1}{2}}\pi^{-0.5}}{\Gamma(\frac{n}{2})n^{0.5}} \int_0^{+\infty} u^{\frac{n+1}{2}-1} e^{-u\left(0.5\frac{t^2}{n}+0.5\right)}du.$$

Όμως συνδέοντας το τελευταίο ολοκλήρωμα με τη γάμμα κατανομή με παραμέτρους $(n+1)/2$ και $\left(0.5\frac{t^2}{n}+0.5\right)^{-1}$ έχουμε ότι:

$$\int_0^{+\infty} u^{\frac{n+1}{2}-1} e^{-u\left(0.5\frac{t^2}{n}+0.5\right)}du = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\left(0.5\frac{t^2}{n}+0.5\right)^{\frac{n+1}{2}}}.$$

Συνδυάζοντας τα παραπάνω προκύπτει το ζητούμενο.



Σχήμα 7.2: Γραφική παράσταση της σππ της t κατανομής για $n = 1,\ 5$ και 30 βαθμούς ελευθερίας με μπλε, πορτοκαλί και πράσινο χρώμα, αντίστοιχα.

Η κατανομή t παρότι είχε εμφανιστεί στη βιβλιογραφία σε άρθρα των Helmert, Lüroth και Pearson πήρε το όνομά της στην αγγλική βιβλιογραφία από την εργασία του Student (1908). Ο William Sealy Gosset (1876-1937) ήταν στατιστικός που είχε προσληφθεί από ένα ιρλανδικό ζυθοποιείο που απαγόρευε τη δημοσίευση της έρευνας από μέλη του προσωπικού του. Για να παρακάμψει αυτόν τον περιορισμό, δημοσίευσε το έργο του κρυφά με το όνομα Student. Η κατανομή αυτή έγινε πιο δημοφιλής χάρη στην εργασία του Fisher (1925).

Στο Σχήμα 7.2 απεικονίζεται η σππ της κατανομής t για $n = 1,5$ και 30 βαθμούς ελευθερίας με μπλε, πορτοκαλί και πράσινο χρώμα, αντίστοιχα. Οι εντολές που χρησιμοποιήθηκαν στην R, για να δημιουργηθεί το Σχήμα 7.2, ήταν οι εξής:

```
library(plotly)

x = seq(−5, 5, length = 1000)
d1 <- dt(x = x,df = 1)
d2 <- dt(x = x,df = 5)
d3 <- dt(x = x,df = 30)

fig <- plot_ly(x = x, y = d1, type = 'scatter', mode = 'lines', name = 'df=1')
fig <- fig %>% add_trace(x = x, y = d2, name = 'df=5')
fig <- fig %>% add_trace(x = x, y = d3, name = 'df=30')
fig
```

Προκύπτει ότι η κατανομή t είναι συμμετρική γύρω από το μηδέν, καθώς ισχύει ότι $f_T(t) = f_T(-t)$, για κάθε $t \in \mathbb{R}$, ενώ η μορφή της κατανομής εξαρτάται από την παράμετρο n, $n > 0$, που ονομάζεται βαθμός ελευθερίας. Επιπρόσθετα, φαίνεται ότι η κατανομή t μοιάζει πολύ με την τυπική κανονική κατανομή. Όσο αυξάνονται οι βαθμοί ελευθερίας τόσο η κατανομή t τείνει στην τυπική κανονική κατανομή, ενώ οι κατανομές αυτές ταυτίζονται ασυμπτωτικά (για $n \to \infty$). Στην πράξη όμως θεωρούμε ότι για βαθμούς ελευθερίας $n > 30$ η τυπική κανονική κατανομή είναι μία πολύ καλή προσέγγιση της κατανομής t.

Πρόταση 7.8

Έστω X η τυχαία μεταβλητή που ακολουθεί κατανομή t με n βαθμούς ελευθερίας με σππ που προσδιορίζεται στη σχέση (7.5). Τότε:

$$E(X) = 0, \text{ για } n > 1, \text{ ενώ για } n \leq 1 \text{ δεν υπάρχει,} \tag{7.6}$$

ενώ

$$Var(X) = \frac{n}{n-2}, \text{ για } n > 2, \text{ ενώ για } n \leq 2 \text{ δεν υπάρχει.} \tag{7.7}$$

Απόδειξη Πρότασης 7.8

Η απόδειξη προκύπτει χρησιμοποιώντας τον ορισμό της μέσης τιμής και της διακύμανσης και αφήνεται ως άσκηση για τον/την αναγνώστη/στρια.

Κάποιες φορές χρειάζεται να υπολογίσουμε την τιμή του σημείου εκείνου που είναι τέτοιο, ώστε να προκύπτει κάποια δοθείσα τιμή για την πιθανότητα του ενδεχομένου η t_n κατανομή να πάρει τιμές μεγαλύτερες από το σημείο αυτό. Για τα σημεία αυτά έχουμε τον ακόλουθο ορισμό.

Ορισμός 7.3

Έστω $X \sim t_n$. Ορίζουμε ως $t_{n,\alpha}$ την τιμή της t_n κατανομής που είναι τέτοια, ώστε:

$$P(X > t_{n,\alpha}) = \alpha, \text{ για } \alpha \in (0,1) \text{ γνωστό αριθμό.} \tag{7.8}$$

Υπάρχουν πίνακες που δίνουν τα σημεία $t_{n,\alpha}$ (που συχνά αναφέρονται ως συμπληρωματικά α-ποσοστιαία σημεία) της t_n κατανομής για διάφορες τιμές των βαθμών ελευθερίας n και της πιθανότητας α (βλ. τον Πίνακα Α΄.4 του Παραρτήματος Α΄). Για παράδειγμα, αν $T \sim t_{14}$ η τιμή της κατανομής t με 14 βαθμούς ελευθερίας που αφήνει δεξιά της πιθανότητα 0.025, δηλαδή η τιμή t που είναι τέτοια, ώστε $P(T > t) = 0.025$ ισούται με $t_{14,0.025} = 2.145$. Παρατηρήστε ότι αν $T \sim t_n$, τότε $P\left(T \leq t_{n,\alpha}\right) = 1 - \alpha$. Τότε, λόγω συμμετρίας της κατανομής t γύρω από το 0, έχουμε ότι $P\left(T \geq -t_{n,\alpha}\right) = 1 - \alpha$ και, επομένως, ισχύει ότι

$$t_{n,1-\alpha} = -t_{n,\alpha}. \tag{7.9}$$

Εναλλακτικά, μπορεί να χρησιμοποιηθεί η R για τον προσδιορισμό οποιουδήποτε ποσοστιαίου σημείου της t_n κατανομής ή πιθανότητας.

Παρατήρηση 7.3

Έστω $X \sim t_n$ με σππ που δίνεται από τη σχέση (7.5). Τότε με τη γλώσσα R μπορούμε:

- με τη συνάρτηση `dt(x,df=n,ncp=0)` να υπολογίσουμε τη σππ στο σημείο ή στο διάνυσμα σημείων x,
- με τη συνάρτηση `pt(x,df=n,ncp=0,lower.tail=TRUE)` να υπολογίσουμε την ασκ στο σημείο ή στο διάνυσμα σημείων x,
- με τη συνάρτηση `pt(x,df=n,ncp=0,lower.tail=FALSE)` να υπολογίσουμε την πιθανότητα $P(X > x)$ στο σημείο ή στο διάνυσμα σημείων x,
- με τη συνάρτηση `qt(q,df=n,ncp=0,lower.tail=TRUE)` να προσδιορίσουμε το σημείο ή το διάνυσμα σημείων x για τα οποία ισχύει ότι $P(X \leq x) = q$, όπου q είναι πιθανότητα ή διάνυσμα με συνιστώσες πιθανότητες, αντίστοιχα,
- με τη συνάρτηση `qt(q,df=n,ncp=0,lower.tail=FALSE` να προσδιορίσουμε το σημείο ή το διάνυσμα σημείων x για τα οποία ισχύει ότι $P(X > x) = q$, όπου q είναι πιθανότητα ή διάνυσμα με συνιστώσες πιθανότητες, αντίστοιχα,
- με τη συνάρτηση `rt(k,df=n,ncp=0)` να δημιουργήσουμε ένα δείγμα μεγέθους k από αυτήν την κατανομή.

Στη συνέχεια, χρησιμοποιώντας τη μέθοδο του μετασχηματισμού, προκύπτει ο ορισμός μιας νέας κατανομής, η οποία στο Κεφάλαιο 9 θα δούμε ότι ανήκει στις λεγόμενες δειγματικές ή δειγματοληπτικές κατανομές, όπως ανήκουν και η χι-τετράγωνο και t κατανομές.

Πρόταση 7.9

Έστω X_1 και X_2 δύο ανεξάρτητες τ.μ. με κατανομή $\chi^2_{n_1}$ και $\chi^2_{n_2}$, αντίστοιχα. Τότε η τ.μ. $F = \frac{X_1/n_1}{X_2/n_2}$ έχει συνάρτηση πυκνότητας πιθανότητας:

$$f(x) = \frac{\Gamma\left(\frac{n_1+n_2}{2}\right)}{\Gamma\left(\frac{n_1}{2}\right)\Gamma\left(\frac{n_2}{2}\right)}\left(\frac{n_1}{n_2}\right)^{n_1/2}\frac{x^{(n_1-2)/2}}{\left(\frac{n_2+n_1x}{n_2}\right)^{(n_1+n_2)/2}}, \quad x>0. \tag{7.10}$$

Τότε λέμε ότι η τυχαία μεταβλητή F ακολουθεί **κατανομή F με βαθμούς ελευθερίας** n_1, n_2 και συμβολίζουμε $F \sim F_{n_1,n_2}$.

Απόδειξη Πρότασης 7.9

Η απόδειξη είναι παρόμοια με την απόδειξη της Πρότασης 7.7 και αφήνεται ως άσκηση για τον/την αναγνώστη/στρια.

Παρατήρηση 7.4

Άμεσα από τον ορισμό της κατανομής F με βαθμούς ελευθερίας n_1, n_2 προκύπτει ότι, αν $F \sim F_{n_1,n_2}$, τότε $\frac{1}{F} \sim F_{n_2,n_1}$, καθώς $F = \frac{X_1/n_1}{X_2/n_2} \sim F_{n_1,n_2}$ τότε η $1/F = \frac{1}{\frac{X_1/n_1}{X_2/n_2}} = \frac{X_2/n_2}{X_1/n_1} \sim F_{n_2,n_1}$.

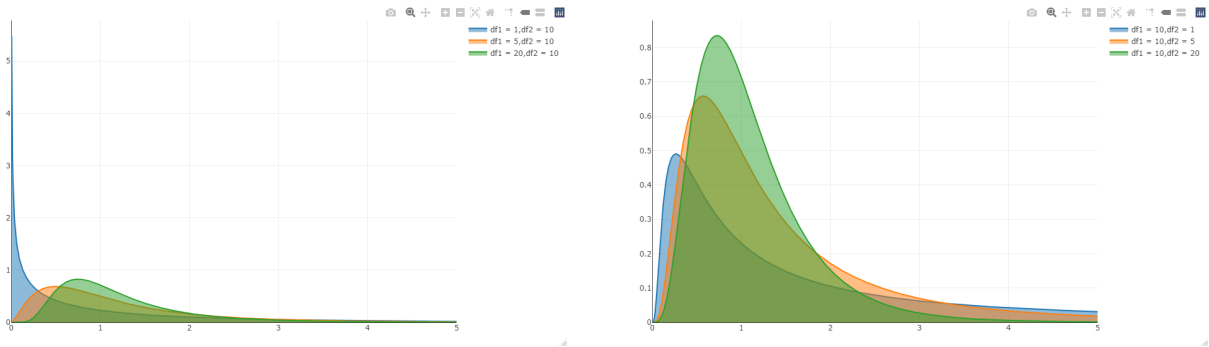
Στο Σχήμα 7.3 (αριστερά) απεικονίζεται η σπ της κατανομής F για $n_1 = 1,\ 5$ και 20 και $n_2 = 10$ βαθμούς ελευθερίας με μπλε, πορτοκαλί και πράσινο χρώμα, αντίστοιχα και (δεξιά) η σπ της κατανομής F για $n_1 = 10$ και $n_2 = 1,5$ και 20 βαθμούς ελευθερίας με μπλε, πορτοκαλί και πράσινο χρώμα, αντίστοιχα. Οι εντολές που χρησιμοποιήθηκαν στην R, για τη δημιουργία του Σχήματος 7.3, είναι οι ακόλουθες:

```
library(plotly)
##### F(df1, df2=10)
x = seq(0, 5, length = 1000)
d1 <- df(x = x,df1 = 1,df2 = 10)
d2 <- df(x = x,df1 = 5,df2 = 10)
d3 <- df(x = x,df1 = 20,df2 = 10)

fig <- plot_ly(x = x, y = d1, type = 'scatter', mode = 'lines',
               name = 'df1 = 1,df2 = 10', fill = 'tozeroy')
fig <- fig %>% add_trace(x = x, y = d2, name = 'df1 = 5,df2 = 10', fill = '
    tozeroy')
fig <- fig %>% add_trace(x = x, y = d3, name = 'df1 = 20,df2 = 10', fill = '
    tozeroy')
fig

##### F(df1=10, df2)
x = seq(0, 5, length = 1000)
d1 <- df(x = x,df1 = 10,df2 = 1)
d2 <- df(x = x,df1 = 10,df2 = 5)
d3 <- df(x = x,df1 = 10,df2 = 20)

fig <- plot_ly(x = x, y = d1, type = 'scatter', mode = 'lines',
               name = 'df1 = 10,df2 = 1', fill = 'tozeroy')
```



Σχήμα 7.3: Γραφική παράσταση της σππ της F κατανομής με διάφορους βαθμούς ελευθερίας $n_1 = 1, 5, 20$ και $n_2 = 10$ βαθμούς ελευθερίας με μπλε, πορτοκαλί και πράσινο χρώμα, αντίστοιχα (αριστερά) και $n_1 = 10$ και $n_2 = 1, 5, 20$ βαθμούς ελευθερίας με μπλε, πορτοκαλί και πράσινο χρώμα, αντίστοιχα (δεξιά).

```
22 fig <- fig %>% add_trace(x = x, y = d2, name = 'df1 = 10,df2 = 5', fill = '
       tozeroy')
23 fig <- fig %>% add_trace(x = x, y = d3, name = 'df1 = 10,df2 = 20', fill = '
       tozeroy')
24 fig
```

Πρόταση 7.10

Έστω X η τυχαία μεταβλητή που ακολουθεί κατανομή F με βαθμούς ελευθερίας n_1, n_2 με σππ που προσδιορίζεται στη σχέση (7.10). Τότε:

$$E(X) = \frac{n_2}{n_2 - 2},\ n_2 > 2, \quad Var(X) = \frac{2n_2^2(n_1 + n_2 - 2)}{n_1(n_2 - 4)(n_2 - 2)^2},\ n_2 > 4.$$

Απόδειξη Πρότασης 7.10

Η απόδειξη προκύπτει χρησιμοποιώντας τον ορισμό της μέσης τιμής και της διακύμανσης και αφήνεται ως άσκηση για τον/την αναγνώστη/στρια.

Πρόταση 7.11

Έστω W τυχαία μεταβλητή που ακολουθεί κατανομή t με n βαθμούς ελευθερίας, τότε $W^2 \sim F_{1,n}$.

Απόδειξη Πρότασης 7.11

Έστω $Z \sim N(0,1)$ και $V \sim \chi^2_n$ ανεξάρτητες τυχαίες μεταβλητές, τότε από τον ορισμό της t_n κατανομής προκύπτει ότι $W = \frac{Z}{\sqrt{V/n}}$. Επίσης $Z^2 \sim \chi^2_1$. Επομένως,

$$W^2 = \frac{Z^2}{V/n} = \frac{Z^2/1}{V/n} \sim F_{1,n}$$

από τον ορισμό της $F_{1,n}$ κατανομής.

Κάποιες φορές χρειάζεται να υπολογίσουμε την τιμή του σημείου εκείνου που είναι τέτοιο, ώστε να προκύπτει κάποια δοθείσα τιμή για την πιθανότητα του ενδεχομένου η F_{n_1,n_2} κατανομή να πάρει τιμές μεγαλύτερες από το σημείο αυτό. Για τα σημεία αυτά έχουμε τον ακόλουθο ορισμό.

Ορισμός 7.4

Έστω $X \sim F_{n_1,n_2}$. Ορίζουμε ως $F_{n_1,n_2,\alpha}$ την τιμή της F_{n_1,n_2} κατανομής που είναι τέτοια, ώστε

$$P(X > F_{n_1,n_2,\alpha}) = \alpha, \text{ για } \alpha \in (0,1) \text{ γνωστό αριθμό .} \quad (7.11)$$

Υπάρχουν πίνακες που δίνουν τα σημεία $F_{n_1,n_2,\alpha}$ (που συχνά αναφέρονται ως συμπληρωματικά α-ποσοστιαία σημεία) της F_{n_1,n_2} κατανομής για διάφορες τιμές των βαθμών ελευθερίας n_1 και n_2, της πιθανότητας α (βλ. τους Πίνακες Α΄.7-Α΄.13 του Παραρτήματος Α΄). Για παράδειγμα, η τιμή της κατανομής F με βαθμούς ελευθερίας (4, 5), η οποία αφήνει δεξιά της πιθανότητα 0.05, ισούται με $F_{4,5,0.05} = 5.19$. Παρατηρήστε ότι οι πίνακες που δίνονται στο παράρτημα Α΄ αντιστοιχούν σε μικρές τιμές του α (συνήθως 0.01, 0.05), οπότε ίσως εύλογα αναρωτιέται κάποιος/α τι γίνεται για μεγάλες τιμές του $\alpha \in (0,1)$ (για παράδειγμα $\alpha = 0.95, 0.99$). Αυτό σημαίνει ότι $P\left(X \le F_{n_1,n_2,\alpha}\right) = 1 - \alpha$ με την τιμή $1 - \alpha$ να υπάρχει στον πίνακα. Η παραπάνω πιθανότητα ισοδύναμα γράφεται $P\left(\frac{1}{X} \ge \frac{1}{F_{n_1,n_2,\alpha}}\right) = 1 - \alpha$, οπότε από την Παρατήρηση 7.4 έχουμε $\frac{1}{X} \sim F_{n_2,n_1}$ και, επομένως, ισχύει ότι:

$$F_{n_2,n_1,1-\alpha} = \frac{1}{F_{n_1,n_2,\alpha}}. \quad (7.12)$$

Εναλλακτικά, μπορεί να χρησιμοποιηθεί η R για τον προσδιορισμό οποιουδήποτε ποσοστιαίου σημείου ή πιθανότητας της κατανομής F_{n_1,n_2}.

Παρατήρηση 7.5

Έστω $X \sim F_{n_1,n_2}$ με σππ που δίνεται από τη σχέση (7.10). Τότε με τη γλώσσα R μπορούμε:

- με τη συνάρτηση `df(x, df1=n1,df2=n2,ncp=0)` να υπολογίσουμε τη σππ στο σημείο ή στο διάνυσμα σημείων x,
- με τη συνάρτηση `pf(x,df1=n1,df2=n2,ncp=0,lower.tail=TRUE)` να υπολογίσουμε την ασκ στο σημείο ή στο διάνυσμα σημείων x,
- με τη συνάρτηση `pf(x,df1=n1,df2=n2,ncp=0,lower.tail=FALSE)` να υπολογίσουμε την πιθανότητα $P(X > x)$ στο σημείο ή στο διάνυσμα σημείων x,
- με τη συνάρτηση `qf(q,df1=n1,df2=n2,ncp=0,lower.tail=TRUE)` να προσδιορίσουμε το σημείο ή το διάνυσμα σημείων x για τα οποία ισχύει ότι $P(X \le x) = q$, όπου q είναι πιθανότητα ή διάνυσμα με συνιστώσες πιθανότητες, αντίστοιχα,
- με τη συνάρτηση `qf(q,df1=n1,df2=n2,ncp=0,lower.tail=FALSE` να προσδιορίσουμε το σημείο ή το διάνυσμα σημείων x για τα οποία ισχύει ότι $P(X > x) = q$, όπου q είναι πιθανότητα ή διάνυσμα με συνιστώσες πιθανότητες, αντίστοιχα,
- με τη συνάρτηση `rf(k,df1=n1,df2=n2,ncp=0)` να δημιουργήσουμε ένα δείγμα μεγέθους k από αυτήν την κατανομή.

7.3 Κεντρικό Οριακό Θεώρημα και εφαρμογές

Έστω $X_1, X_2,\ldots,X_n$, n το πλήθος ανεξάρτητες και ισόνομες τυχαίες μεταβλητές από μια κατανομή με σππ ή σπ $f_X(\cdot)$, δηλαδή αποτελούν αυτό που λέμε ένα **τυχαίο δείγμα** από τη συγκεκριμένη κατανομή μεγέθους n. Στην Ενότητα 7.2.4 το ενδιαφέρον επικεντρώθηκε στον προσδιορισμό της κατανομής του αθροίσματος $\sum_{i=1}^{n} X_i$. Μεταξύ άλλων (βλ. αποτελέσματα της Πρότασης 7.4) αποδείχθηκε το ακόλουθο πολύ σημαντικό αποτέλεσμα.

Πρόταση 7.12

Έστω $X_1, \dots, X_n$ είναι n το πλήθος ανεξάρτητες και ισόνομες τυχαίες μεταβλητές με $X_i \sim N(\mu, \sigma^2)$, $i = 1, \dots, n$, δηλαδή είναι ένα τυχαίο δείγμα μεγέθους n από την προαναφερθείσα κατανομή. Τότε ισχύουν τα ακόλουθα:

$$Z = \frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}} \sim N(0,1), \tag{7.13}$$

ή, ισοδύναμα,

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1), \tag{7.14}$$

όπου $\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$.

Απόδειξη Πρότασης 7.12

Το πρώτο μέρος της πρότασης έχει αποδειχθεί στην Πρόταση 7.4, ενώ το δεύτερο προκύπτει από την ίδια πρόταση για $a_i = \frac{1}{n}$.

Η παραπάνω πρόταση είναι ιδιαίτερα χρήσιμη καθώς σε αυτήν, όπως θα δούμε στα επόμενα κεφάλαια του παρόντος συγγράμματος, στηρίζονται αρκετά αποτελέσματα που αφορούν τη στατιστική συμπερασματολογία για τη μέση τιμή κανονικού πληθυσμού.

Παράδειγμα 7.6

Μια τσιμεντοβιομηχανία συσκευάζει τσιμέντο σε τσουβάλια των 50 κιλών. Υποθέτουμε ότι το βάρος των συσκευαζόμενων τσουβαλιών ακολουθεί κανονική κατανομή με μέση τιμή 50 κιλά και τυπική απόκλιση 0.87 κιλά.

1. Αν ένας πελάτης αγοράσει 100 τσουβάλια από την εν λόγω τσιμεντοβιομηχανία, ποια είναι η πιθανότητα να καταλήξετε με λιγότερο από 4975 κιλά τσιμέντο;
2. Προκειμένου ο πελάτης να παραλάβει την παρτίδα, ελέγχει κάθε φορά 10 τσουβάλια και τη δέχεται μόνο αν το μέσο βάρος τους ξεπερνά τα 49.75 κιλά. Υπολογίστε την πιθανότητα ο πελάτης να παραλάβει την επόμενη παρτίδα.

Λύση Παραδείγματος 7.6

1. Έστω X_i η τυχαία μεταβλητή που παριστάνει το βάρος του i-οστού συσκευασμένου τσουβαλιού, $i = 1, \dots, 100$. Από την εκφώνηση έχουμε ότι $X_i \sim N(50, 0.87^2)$ και, επομένως, άμεσα συμπεραίνουμε ότι $\sum_{i=1}^{100} X_i \sim N(100 \cdot 50,\ 100 \cdot 0.87^2)$. Θέλουμε να προσδιορίσουμε την πιθανότητα $P\left(\sum_{i=1}^{100} X_i < 4975\right)$ με $\sum_{i=1}^{100} X_i \sim N(100 \cdot 50,\ 100 \cdot 0.87^2)$. Χρησιμοποιώντας τον τυπικό μετασχηματισμό

$$Z = \frac{\sum_{i=1}^{100} X_i - 5000}{\sqrt{100 \cdot 0.87^2}} \sim N(0,1),$$

έχουμε ότι:

$$\begin{aligned} P\left(\sum_{i=1}^{100} X_i < 4975\right) &= P\left(Z < \frac{4975 - 5000}{8.7}\right) \\ &= P(Z < -2.87) = P(Z > 2.87) = 1 - P(Z < 2.87) \\ &= 1 - 0.99795 = 0.00205, \end{aligned}$$

όπου χρησιμοποιήσαμε τον πίνακα της τυπικής κανονικής κατανομής του Παραρτήματος Α΄.

2. Έστω X_i η τυχαία μεταβλητή που παριστάνει το βάρος του i-οστού συσκευασμένου τσουβαλιού που ελέγχεται από τον πελάτη, $i = 1, \ldots, 10$. Από την εκφώνηση έχουμε ότι $X_i \sim N(50, 0.87^2)$ και, επομένως, άμεσα συμπεραίνουμε ότι $\bar{X} = \frac{\sum_{i=1}^{10} X_i}{n} \sim N(50,\ 0.87^2/10)$. Θέλουμε να προσδιορίσουμε την πιθανότητα $P(\bar{X} > 49.75)$ με $\bar{X} \sim N(50,\ 0.87^2/10)$. Χρησιμοποιώντας τον τυπικό μετασχηματισμό

$$Z = \frac{\bar{X} - 50}{\sqrt{0.87^2/100}} \sim N(0,1)$$

έχουμε ότι:

$$P(\bar{X} > 49.75) = P\left(Z > \frac{49.75 - 50}{0.87/\sqrt{10}}\right) = P(Z > -0.91) = P(Z < 0.91)$$
$$= 0.81859$$

όπου χρησιμοποιήσαμε τον πίνακα της τυπικής κανονικής κατανομής του Παραρτήματος Α΄ (στις πράξεις διατηρήθηκαν δύο δεκαδικά ψηφία).

Άσκηση Αυτοαξιολόγησης 7.5

Σε έναν πληθυσμό ανδρών το βάρος τους ακολουθεί κανονική κατανομή με μέση τιμή 80 κιλά και τυπική απόκλιση 8 κιλά.

1. Υπολογίστε την πιθανότητα το μέσο βάρος ενός τυχαίου δείγματος 5 ατόμων να ξεπερνάει τα 85 κιλά.
2. Το όριο λειτουργίας ενός ανελκυστήρα είναι 450 κιλά. Υπολογίστε την πιθανότητα να μπουν 5 άτομα από αυτόν τον πληθυσμό στον ανελκυστήρα και να μην λειτουργήσει.

Ένα ερώτημα που ίσως έχει προκύψει είναι αν χρησιμοποιώντας τη μέθοδο της ροπογεννήτριας που παρουσιάστηκε στην Ενότητα 7.2.4 είναι πάντοτε εφικτός ο προσδιορισμός της κατανομής του αθροίσματος, όπως παραδείγματος χάριν στην περίπτωση της κανονικής κατανομής. Η απάντηση στο ερώτημα αυτό, δυστυχώς, είναι αρνητική, όπως μπορεί να γίνει εύκολα αντιληπτό αν θεωρήσουμε ότι οι τυχαίες μεταβλητές $X_i \sim \Gamma(a_i,\ \lambda_i)$, $i = 1, \ldots, k$ με σππ που δόθηκε στη σχέση (5.27). Τότε, σύμφωνα με τη σχέση (5.32), είναι

$$M_{X_i}(t) = \left(1 - \frac{t}{\lambda_i}\right)^{-a_i}, \quad \text{για } t < \lambda_i,$$

και η ροπογεννήτρια της τυχαίας μεταβλητής $Y = \sum_{i=1}^{n} X_i$ είναι

$$M_Y(t) = \prod_{i=1}^{n} \left(1 - \frac{t}{\lambda_i}\right)^{-a_i}, \quad \text{για } t < \lambda_i,$$

που προφανώς δεν ταυτίζεται με τη ροπογεννήτρια κάποιας γνωστής κατανομής.

Σε τέτοιες περιπτώσεις, όπως της γάμμα κατανομής, για να προσδιοριστεί η κατανομή του αθροίσματος θα μπορούσαμε να ακολουθήσουμε μια διαφορετική προσέγγιση και, πιο συγκεκριμένα, τη μέθοδο του μετασχηματισμού. Σύμφωνα με αυτήν θα πρέπει να θεωρήσουμε επιπλέον $n - 1$ συναρτήσεις των $X_1,\ldots,X_n$ (βλ. το Θεώρημα 7.1 και την παρατήρηση που έπεται αυτού) και η κατανομή θα βρεθεί προσδιορίζοντας στη συνέχεια, την περιθώρια κατανομή της Y. Η παραπάνω διαδικασία δεν είναι πάντοτε εύκολη ή ακόμα και εφικτή.

Μια διέξοδο στα παραπάνω προβλήματα δίνει ένα από τα σπουδαιότερα θεωρήματα της Θεωρίας Πιθανοτήτων και της Στατιστικής, το οποίο είναι γνωστό ως **Κεντρικό Οριακό Θεώρημα**. Σύμφωνα με αυτό, το άθροισμα ή η μέση τιμή n το πλήθος ανεξάρτητων και ισόνομων τυχαίων μεταβλητών από μια κατανομή

με πεπερασμένη μέση τιμή και διακύμανση ακολουθεί, για μεγάλες τιμές του n, προσεγγιστικά κανονική κατανομή. Για την ιστορία του Κεντρικού Οριακού Θεωρήματος παραπέμπουμε στο σύγγραμμα Fischer (2011). Ακολουθεί παρακάτω η διατύπωση του θεωρήματος. Όμως, η απόδειξη του θεωρήματος είναι πέρα από τον σκοπό του συγκεκριμένου συγγράμματος, αλλά όποιος επιθυμεί μπορεί να τη βρει - ενδεικτικά - στο σύγγραμμα των Stuart and Ord (1987) και Κούτρας (2005).

Θεώρημα 7.2: Κεντρικό Οριακό Θεώρημα (ΚΟΘ)

Έστω $X_1, X_2, \ldots, X_n$, n το πλήθος ανεξάρτητες και ισόνομες τυχαίες μεταβλητές με πεπερασμένη μέση τιμή μ και διασπορά σ^2. Τότε η αθροιστική συνάρτηση κατανομής της τυχαίας μεταβλητής

$$Z = \frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}}$$

συγκλίνει, για $n \to \infty$, στην αθροιστική συνάρτηση κατανομής της τυπικής κανονικής κατανομής $N(0,1)$. Εναλλακτικά, λέμε ότι η τ.μ. $\sum_{i=1}^{n} X_i$ ακολουθεί προσεγγιστικά (ασυμπτωτικά) την κανονική κατανομή $N(n\mu, n\sigma^2)$.

Καθώς η τυχαία μεταβλητή Z που εμφανίζεται στο Κεντρικό Οριακό Θεώρημα μπορεί να γραφτεί

$$Z = \frac{\frac{\sum_{i=1}^{n} X_i - n\mu}{n}}{\frac{\sigma\sqrt{n}}{n}}$$

ή, ισοδύναμα,

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

έχουμε το ακόλουθο πόρισμα, που πολλές φορές αποτελεί τη διατύπωση του Κεντρικού Οριακού Θεωρήματος.

Πόρισμα 7.1

Έστω $X_1, X_2, \ldots, X_n$, n το πλήθος ανεξάρτητες και ισόνομες τυχαίες μεταβλητές με πεπερασμένη μέση τιμή μ και διασπορά σ^2. Τότε, η αθροιστική συνάρτηση κατανομής της τυχαίας μεταβλητής

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

συγκλίνει, για $n \to \infty$, στην αθροιστική συνάρτηση κατανομής της τυπικής κανονικής κατανομής $N(0,1)$. Εναλλακτικά, λέμε ότι η $\bar{X}$ ακολουθεί ασυμπτωτικά (προσεγγιστικά) την κανονική κατανομή $N(\mu, \sigma^2/n)$.

Παρατήρηση 7.6

Η σύγκλιση που εμφανίζεται στη διατύπωση του Κεντρικού Οριακού Θεωρήματος είναι γνωστή ως σύγκλιση κατά κατανομή, για αυτόν τον λόγο αρκετές φορές αντί της έκφρασης «ακολουθεί ασυμπτωτικά» χρησιμοποιείται η έκφραση «συγκλίνει κατά κατανομή». Για άλλους τύπους σύγκλισης και κριτήρια σύγκλισης ακολουθιών τυχαίων μεταβλητών παραπέμπουμε τον/την ενδιαφερόμενο/η αναγνώστη/στρια στο σύγγραμμα του van der Vaart (1998) και τις εκεί αναφορές. Σε όσα ακολουθούν θα χρησιμοποιούμε τον συμβολισμό $Z \xrightarrow{d} N(0,1)$ για να δηλώσουμε ότι η αθροιστική συνάρτηση κατανομής της τυχαίας μεταβλητής Z συγκλίνει, για $n \to \infty$, στην αθροιστική συνάρτηση κατανομής της τυπικής κανονικής κατανομής $N(0,1)$.

Παρατήρηση 7.7

Έστω $X_1, X_2, \dots, X_n$, n το πλήθος ανεξάρτητες και ισόνομες τυχαίες μεταβλητές με πεπερασμένη μέση τιμή μ και διασπορά σ^2. Τότε έχουμε ότι:

$$E\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} E(X_i) = \sum_{i=1}^{n} \mu = n\mu \text{ και } E(\bar{X}) = \mu, \tag{7.15}$$

ενώ

$$Var\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} Var(X_i) = \sum_{i=1}^{n} \sigma^2 = n\sigma^2 \text{ και } Var\left(\bar{X}\right) = \frac{\sigma^2}{n}. \tag{7.16}$$

Με βάση τα παραπάνω μπορούμε να πούμε ότι η μέση της τυχαίας μεταβλητής $\bar{X}$ είναι ίδια με αυτή της αρχικής κατανομής, ενώ η διασπορά της είναι μικρότερη και μάλιστα η μείωση είναι ανάλογη με το μέγεθος του δείγματος. Επιπρόσθετα, το ΚΟΘ μας λέει ότι οποιαδήποτε και αν είναι η κατανομή των ανεξάρτητων και ισόνομων (διακριτών ή συνεχών) τυχαίων μεταβλητών X_i, $i = 1, \dots, n$, η κατανομή του αθροίσματος και η κατανομή της μέση τιμής τους ακολουθεί προσεγγιστικά κανονική κατανομή με τις αντίστοιχες παραμέτρους. Μόνη προϋπόθεση για να ισχύει το αποτέλεσμα αυτό είναι να είναι πεπερασμένη η μέση τιμή και η διασπορά των τυχαίων μεταβλητών X_i.

Παράδειγμα 7.7

Έστω $X_1, X_2, \dots, X_n$ ανεξάρτητες και ισόνομες τυχαίες μεταβλητές από μια κατανομή με συνάρτηση πιθανότητας

$$P(X_i = 0) = P(X_i = 1) = P(X_i = 2) = 1/3.$$

Να βρεθεί η ασυμπτωτική κατανομή του $\bar{X}$.

Λύση Παραδείγματος 7.7

Έχουμε ότι

$$\mu = E(X_i) = \sum_{j=0}^{2} j \cdot P(X_i = j) = 0 \cdot 1/3 + 1 \cdot 1/3 + 2 \cdot 1/3 = 1$$

και

$$\begin{aligned} \sigma^2 = Var(X_i) &= E(X_i^2) - (E(X_i))^2 \\ &= 0^2 \cdot 1/3 + 1^2 \cdot 1/3 + 2^2 \cdot 1/3 - 1^2 \\ &= 2/3. \end{aligned}$$

Εφαρμόζοντας το ΚΟΘ, προκύπτει, καθώς $n \to \infty$, ότι

$$Z = \frac{\sum_{i=1}^{n} X_i - n}{\sqrt{2/3}\sqrt{n}} = \frac{\sum_{i=1}^{n} X_i - n}{\sqrt{2n/3}} \xrightarrow{d} N(0,1),$$

ή, ισοδύναμα,

$$Z = \frac{\bar{X} - 1}{\sqrt{2/3}/\sqrt{n}} = \frac{\bar{X} - 1}{\sqrt{\frac{2}{3n}}} \xrightarrow{d} N(0,1).$$

δηλαδή ο $\bar{X}$ ακολουθεί ασυμπτωτικά (προσεγγιστικά) την κανονική κατανομή $N(1,\ 2/(3 \cdot n))$.

Ένα εύλογο ερώτημα που προκύπτει είναι πόσο μεγάλο πρέπει να είναι το πλήθος n των τυχαίων μεταβλητών, έτσι ώστε να ισχύει το Κεντρικό Οριακό Θεώρημα, καθώς στη διατύπωσή του αναφέρεται ότι n πρέπει να τείνει στο άπειρο. Αποδεικνύεται ότι η ταχύτητα σύγκλισης εξαρτάται από τη λοξότητα της κοινής κατανομής που ακολουθούν οι τυχαίες μεταβλητές X_i, $i = 1, \dots, n$. Στην πραγματικότητα, όταν η κατανομή είναι συμμετρική, όπως αυτή του Παραδείγματος 7.7, η κατανομή του $\bar{X}$ προσεγγίζεται ικανοποιητικά από την κανονική κατανομή ακόμα και για πολύ μικρά n, επί παραδείγματι $n = 4$ ή 5.

Από την άλλη πλευρά, σε περίπτωση που υπάρχει μεγάλη λοξότητα στην κοινή κατανομή των X_i, $i = 1, \dots, n$, το μέγεθος του δείγματος θα πρέπει να είναι $n \geq 30$ (Κουτρουβέλης, 2011) για να είναι ικανοποιητική η σύγκλιση της κατανομής του $\bar{X}$ στην κανονική κατανομή.

Άσκηση Αυτοαξιολόγησης 7.6

Επιλέξτε τη σωστή απάντηση ανάμεσα στις (1.)-(4.) και δικαιολογήστε σύντομα την απάντησή σας. Έστω $X_1, X_2, \dots, X_{100}$ ανεξάρτητες και ισόνομες τυχαίες μεταβλητές από μια κατανομή με πεπερασμένη μέση τιμή μ και διασπορά $\sigma^2 = 100^2$. Τότε η τυπική απόκλιση του δειγματικού μέσου $\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$ (που καλείται, όπως θα δούμε στο Κεφάλαιο 9, τυπικό σφάλμα του δειγματικού μέσου) ισούται με:

1. 1
2. 10
3. 100
4. 1000

Άσκηση Αυτοαξιολόγησης 7.7

Επιλέξτε τη σωστή απάντηση ανάμεσα στις (α`)-(δ`) και δικαιολογήστε σύντομα την απάντησή σας.

1. Έστω $X_1, X_2, \dots, X_{144}$ ένα τυχαίο δείγμα από την εκθετική κατανομή με παράμετρο $\lambda = 1/2$.
   (α΄) Η κατανομή του $\bar{X} - 2$ είναι προσεγγιστικά η τυπική κανονική κατανομή.
   (β΄) Η μέση τιμή του $\bar{X}$ είναι 2.
   (γ΄) Η τυπική απόκλιση του $\bar{X}$ είναι $1/24$.
   (δ΄) Ο δειγματικός μέσος $\bar{X}$ ακολουθεί προσεγγιστικά κανονική κατανομή με μέση τιμή 2 και τυπική απόκλιση $1/3$.
2. Έστω $X_1, X_2, \dots, X_{144}$ ένα τυχαίο δείγμα από την ομοιόμορφη κατανομή στο (0, 12).
   (α΄) Η κατανομή του $\bar{X} - 6$ είναι προσεγγιστικά τυπική κανονική κατανομή.
   (β΄) Ο δειγματικός μέσος $\bar{X}$ ακολουθεί προσεγγιστικά την κανονική κατανομή με μέση τιμή 6 και διασπορά 12.
   (γ΄) Η τυπική απόκλιση του $\bar{X}$ είναι 12.
   (δ΄) Τίποτα από τα παραπάνω δεν ισχύει.

Μέχρι τώρα υποθέταμε ότι οι τυχαίες μεταβλητές είναι ανεξάρτητες και ισόνομες. Το ΚΟΘ συνεχίζει να ισχύει ακόμα και αν οι τυχαίες μεταβλητές είναι ανεξάρτητες, αλλά όχι ισόνομες. Ειδικότερα, ισχύει η ακόλουθη πρόταση.

Πρόταση 7.13

Έστω $X_1, X_2, \dots, X_n$, n το πλήθος ανεξάρτητες τυχαίες μεταβλητές με πεπερασμένη μέση τιμή μ_i και διασπορά σ_i^2. Τότε, η αθροιστική συνάρτηση κατανομής της τυχαίας μεταβλητής

$$Z = \frac{\sum_{i=1}^{n} X_i - \mu}{\sqrt{\sum_{i=1}^{n} \sigma_i^2}}$$

ή, ισοδύναμα, της τυχαίας μεταβλητής

$$Z = \frac{\bar{X} - \frac{\mu}{n}}{\frac{\sqrt{\sum_{i=1}^{n} \sigma_i^2}}{n}},$$

όπου $\mu = \sum_{i=1}^{n} \mu_i$, και $\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$, συγκλίνει, για $n \to \infty$, στην αθροιστική συνάρτηση κατανομής της τυπικής κανονικής κατανομής $N(0,1)$, Εναλλακτικά, λέμε ότι η τ.μ. $\sum_{i=1}^{n} X_i$ ακολουθεί ασυμπτωτικά (προσεγγιστικά) την κανονική κατανομή $N(\mu, \sum_{i=1}^{n} \sigma_i^2)$.

Απόδειξη Πρότασης 7.13

Ακολουθεί παρόμοια βήματα με αυτήν του κλασικού Κεντρικού Οριακού Θεωρήματος και παραλείπεται.

Παράδειγμα 7.8

Έστω $X_1, X_2, \dots, X_n$ ανεξάρτητες τυχαίες μεταβλητές που ακολουθούν εκθετική κατανομή με παραμέτρους $(\alpha + \beta \cdot y_i)^{-1}$, $i = 1, \dots, n$, δηλαδή $X_i \sim Exp((\alpha + \beta \cdot y_i)^{-1})$, όπου $y_1, \dots, y_n$ γνωστοί θετικοί αριθμοί όχι όλοι ίσοι μεταξύ τους και $\alpha, \beta > 0$. Να βρεθεί η ασυμπτωτική κατανομή του $\bar{X}$.

Λύση Παραδείγματος 7.8

Σύμφωνα με την Πρόταση 7.13, αρχικά προσδιορίζουμε τη μέση τιμή και τη διασπορά των X_i. Είναι τότε (βλ. Ενότητα 5.4)

$$E(X_i) = \alpha + \beta \cdot y_i \text{ και } Var(X_i) = (\alpha + \beta \cdot y_i)^2.$$

Επομένως, η $\bar{X}$ ακολουθεί προσεγγιστικά για μεγάλο n κανονική κατανομή με μέση τιμή

$$\frac{\sum_{i=1}^{n}(\alpha + \beta \cdot y_i)}{n} = \frac{n\alpha + \beta \sum_{i=1}^{n} y_i}{n} = \alpha + \beta \bar{y}$$

και διακύμανση

$$\frac{\sum_{i=1}^{n}(\alpha + \beta \cdot y_i)^2}{n^2} = \frac{n\alpha^2 + 2\alpha\beta \sum_{i=1}^{n} y_i + \beta^2 \sum_{i=1}^{n} y_i^2}{n^2}$$
$$= \frac{\alpha^2}{n} + \frac{2\alpha\beta\bar{y}}{n} + \beta^2 \frac{\sum_{i=1}^{n} y_i^2}{n^2},$$

όπου $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$.

7.3.1 Προσέγγιση της διωνυμικής από την κανονική κατανομή

Ο υπολογισμός των πιθανοτήτων κάποιων ενδεχομένων που συνδέονται με τη διωνυμική κατανομή χρησιμοποιώντας τη συνάρτηση πιθανότητάς της ή την αθροιστική συνάρτηση κατανομής της είναι κάποιες φορές ιδιαίτερα χρονοβόρος ή πολύπλοκος όταν γίνεται χωρίς τη χρήση υπολογιστή, ειδικά σε περιπτώσεις όπου το n είναι πολύ μεγάλο (βλ. Rosner, 2015). Το ερώτημα που προκύπτει είναι αν αυτές οι πιθανότητες μπορούν να προσεγγιστούν χρησιμοποιώντας το Κεντρικό Οριακό Θεώρημα και, αν ναι, πότε η προσέγγιση αυτή είναι ικανοποιητική. Στην ενότητα αυτή θα δοθεί απάντηση στα παραπάνω ερωτήματα.

Έχουμε δει στην Πρόταση 7.3 ότι αν $X_i \sim B(1, p)$ $i = 1, \dots, n$, με X_i να είναι ανεξάρτητες τυχαίες μεταβλητές, τότε η τ.μ. $Y = \sum_{i=1}^{n} X_i$ ακολουθεί $B(n, p)$. Επομένως, για την τ.μ. $Y \sim B(n, p)$ μπορεί να χρησιμοποιηθεί το Κεντρικό Οριακό Θεώρημα, καθώς μπορεί να γραφτεί ως άθροισμα n το πλήθος

ανεξάρτητων και ισόνομων τυχαίων μεταβλητών που η καθεμία ακολουθεί $B(1,p)$, δηλαδή μια κατανομή Bernoulli με πιθανότητα επιτυχίας p. Έτσι οδηγούμαστε στην πρόταση που ακολουθεί.

Πρόταση 7.14

Έστω Y τ.μ. που ακολουθεί διωνυμική κατανομή $B(n,p)$, τότε η τυχαία μεταβλητή $\frac{Y-np}{\sqrt{np(1-p)}}$ ακολουθεί προσεγγιστικά την τυπική κανονική κατανομή $N(0,1)$.

Απόδειξη Πρότασης 7.14

Σύμφωνα με την ανάλυση που προηγήθηκε, η τυχαία μεταβλητή Y γράφεται ως άθροισμα των n το πλήθος ανεξάρτητων και ισόνομων τυχαίων μεταβλητών με κατανομή Bernoulli με παράμετρο p, δηλαδή $Y = \sum_{i=1}^{n} X_i$ με $X_i \sim B(1,p)$, $i = 1,\dots,n$. Τότε (βλ. Ενότητα 4.3) $\mu = E(X_i) = p$ και $\sigma^2 = Var(X_i) = p(1-p)$. Το ζητούμενο αποτέλεσμα προκύπτει εφαρμόζοντας το Κεντρικό Οριακό Θεώρημα.

Από το προηγούμενο θεώρημα προκύπτει η πρόταση που ακολουθεί.

Πρόταση 7.15

Έστω $X_1,\dots,X_n$ ανεξάρτητες και ισόνομες τυχαίες κατανομές με κατανομή Bernoulli με παράμετρο p, δηλαδή $X_i \sim B(1,p)$, $i = 1,\dots,n$. Τότε, καθώς $n \to \infty$, ισχύει ότι:

$$\frac{\sum_{i=1}^{n} X_i}{n} \xrightarrow{d} N\left(p, \frac{(1-p)p}{n}\right).$$

Απόδειξη Πρότασης 7.15

Προκύπτει άμεσα συνδυάζοντας το ΚΟΘ και την Πρόταση 7.14.

Παρατήρηση 7.8

Από την Πρόταση 7.3 γνωρίζουμε ότι αν $X_i \sim B(1,p)$ $i = 1,\dots,n$, ανεξάρτητες τυχαίες μεταβλητές, τότε $Y = \sum_{i=1}^{n} X_i$ ακολουθεί $B(n,p)$ και επομένως:

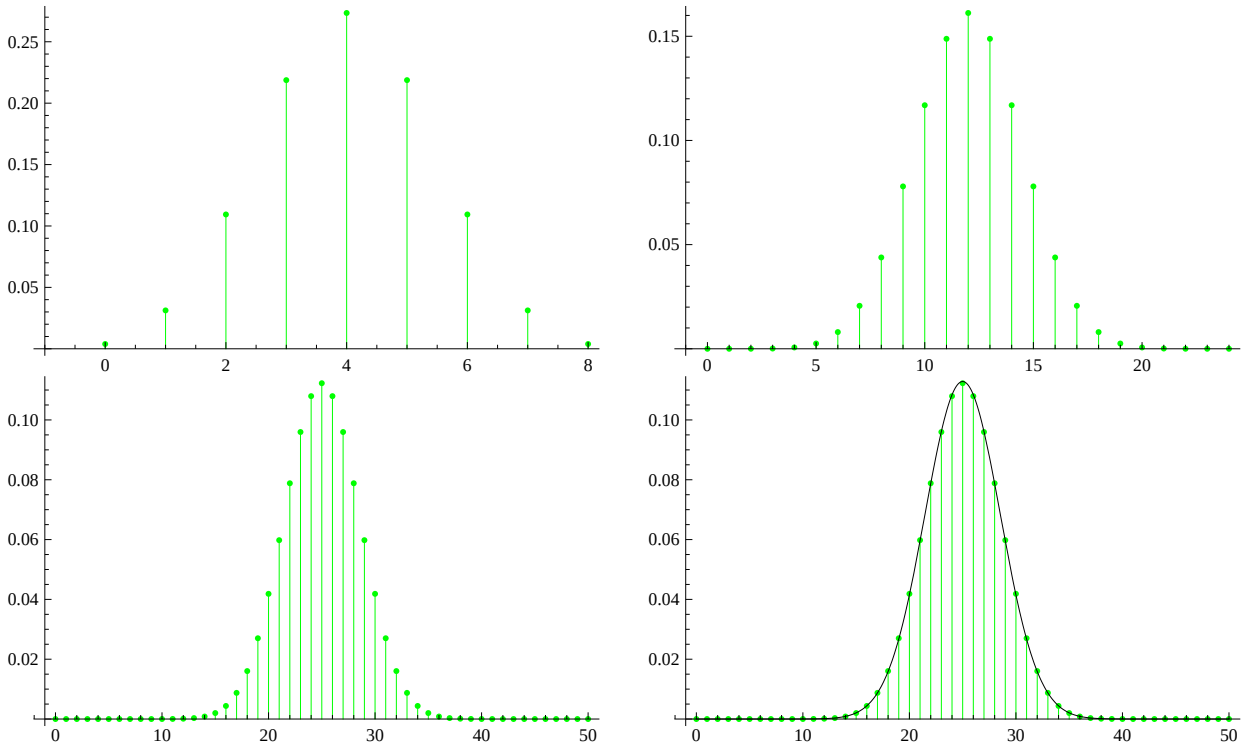
$$P(Y = i) = \begin{pmatrix} n \\ i \end{pmatrix} p^i (1-p)^{n-i},\ i = 0,\dots,n.$$

Επιπρόσθετα με τη μέθοδο του μετασχηματισμού για διακριτές τυχαίες μεταβλητές προκύπτει ότι

$$P\left(\frac{\sum_{i=1}^{n} X_i}{n} = w\right) = \begin{pmatrix} n \\ wn \end{pmatrix} p^{wn} (1-p)^{n-wn},\ w = 0,1/n,2/n\dots,1.$$

Οι παραπάνω σχέσεις μας δίνουν τις ακριβείς τιμές των πιθανοτήτων.

Αφού απαντήθηκε το ερώτημα για το αν μπορεί να χρησιμοποιηθεί το Κεντρικό Οριακό Θεώρημα για την προσέγγιση διωνυμικών πιθανοτήτων, μένει να απαντηθεί πότε η προσέγγιση αυτή είναι ικανοποιητική. Σύμφωνα με τον Rosner (2015), αν η παράμετρος n είναι σχετικά μεγάλη (π.χ. 100) και η πιθανότητα επιτυχίας είναι είτε κοντά στο μηδέν είτε κοντά στο 1, τότε η διωνυμική κατανομή είναι πολύ θετικά ή αρνητικά λοξή, οπότε σε αυτήν την περίπτωση η προσέγγιση της διωνυμικής από την κανονική δεν είναι ικανοποιητική. Το ίδιο ισχύει και όταν το n είναι πολύ μικρό για οποιαδήποτε τιμή της πιθανότητας επιτυχίας p. Από την άλλη πλευρά η προσέγγιση είναι ικανοποιητική, αν n είναι σχετικά μεγάλο και το p είναι ούτε πολύ μικρό ούτε πολύ μεγάλο. Στην πράξη πολύ συχνά τα παραπάνω συνοψίζονται στον λεγόμενο κανόνα του πέντε (rule of five), σύμφωνα με τον οποίο η κανονική προσέγγιση της διωνυμικής είναι ικανοποιητική όταν $np(1-p) \geq 5$, καθώς επίσης και όταν $n > 30$ και το p δεν είναι κοντά στο 0 ή στο 1. Στο Σχήμα 7.4



Σχήμα 7.4: Προσέγγιση της διωνυμικής κατανομής $B(n,\, 1/2)$, $n = 8,\, 25,\, 50$ από την κανονική κατανομή.

απεικονίζονται οι γραφικές παραστάσεις των συναρτήσεων πιθανότητας των διωνυμικών κατανομών $B(8,\, 1/2)$ (πάνω αριστερά), $B(25,\, 1/2)$ (πάνω δεξιά), $B(50,\, 1/2)$ (κάτω αριστερά) και $B(50,\, 1/2)$ μαζί με τη σππ της κανονικής κατανομής με την αντίστοιχη μέση τιμή και διασπορά (κάτω δεξιά).

Επομένως, το ΚΟΘ μας δίνει τη δυνατότητα να υπολογίσουμε προσεγγιστικά πιθανότητες της διωνυμικής κατανομής. Όμως κατά τον προσεγγιστικό υπολογισμό της πιθανότητας της διωνυμικής (και κάθε διακριτής κατανομής) με τη χρήση της κανονικής κατανομής και του ΚΟΘ προκύπτει το πρόβλημα ότι η πιθανότητα $P(Y = y)$, για κάποιο $y = 0, \ldots, n$, όταν προσεγγίζεται από την κανονική ισούται με μηδέν, καθώς η κανονική κατανομή είναι συνεχής. Το πρόβλημα αυτό ξεπερνιέται με τη λεγόμενη διόρθωση συνέχειας, όπου προτείνεται να υπολογίζεται προσεγγιστικά η πιθανότητα $P(y - 0.5 \leq Y \leq y + 0.5)$ για $y = 1, \ldots, n-1$, η $P(Y \leq 0.5)$ για $y = 0$ και η $P(Y \geq n - 0.5)$ για $y = n$. Το ίδιο σκεπτικό εφαρμόζεται και για τις πιθανότητες της μορφής $P(\alpha \leq Y \leq \beta)$, $P(\alpha \leq Y)$ και $P(Y \leq \beta)$, για τις οποίες υπολογίζουμε προσεγγιστικά τις $P(\alpha - 0.5 \leq Y \leq \beta + 0.5)$, $P(\alpha - 0.5 \leq Y)$ και $P(Y \leq \beta + 0.5)$, αντίστοιχα.

Παράδειγμα 7.9

Ρίχνουμε δύο ζάρια 180 φορές και καταγράφουμε το άθροισμα των εδρών του. Να υπολογιστεί προσεγγιστικά η πιθανότητα το άθροισμα 9

1. να εμφανιστεί 25 φορές,
2. να εμφανιστεί τουλάχιστον 25 φορές.

Λύση Παραδείγματος 7.9

Έστω Y η τ.μ που παριστάνει το πλήθος των φορών που το άθροισμα των δύο εδρών είναι 9 στις 180 ρίψεις των δύο ζαριών. Η τυχαία μεταβλητή $Y \sim B(n = 180, p)$, όπου p είναι η πιθανότητα εμφάνισης αθροίσματος 9. Καθώς το άθροισμα 9 εμφανίζεται όταν οι δύο ενδείξεις είναι $\{(5,4),(4,5),(6,3),(3,6)\}$, έχουμε ότι $p = 4/36 = 1/9$. Καθώς n μεγάλο και $np(1-p) > 5$ μπορούμε να χρησιμοποιήσουμε την κανονική προσέγγιση της διωνυμικής κατανομής, δηλαδή να προσεγγίσουμε τη διωνυμική κατανομή από

την κανονική με μέση τιμή $np = 180 \cdot 1/9 = 20$ και διακύμανση $np(1-p) = 180 \cdot 1/9 \cdot 8/9 = 160/9$. Οι ζητούμενες πιθανότητες προσεγγίζονται ως εξής:

1. Ζητείται να υπολογίσουμε την $P(Y = 25)$. Χρησιμοποιώντας ότι η τ.μ. Y είναι διακριτή και τη διόρθωση συνέχειας θα υπολογίσουμε ισοδύναμα την πιθανότητα $P(24.5 < Y < 25.5)$. Καθώς η τ.μ. Y προσεγγίζεται από την κανονική με μέση τιμή 20 και διακύμανση 160/9, έχουμε ότι

$$\begin{aligned} P\left(\frac{24.5-20}{\sqrt{160/9}} < \frac{Y-20}{\sqrt{160/9}} < \frac{25.5-20}{\sqrt{160/9}}\right) &\approx P\left(\frac{24.5-20}{\sqrt{160/9}} < Z < \frac{25.5-20}{\sqrt{160/9}}\right) \\ &= P(1.07 < Z < 1.3) = P(Z < 1.3) - P(Z < 1.07) \\ &= 0.90320 - 0.85769 = 0.04551 \end{aligned}$$

   όπου $Z \sim N(0,1)$. Επισημαίνεται ότι διατηρήσαμε αρχικά, 2 δεκαδικά ψηφία στους υπολογισμούς και έγινε χρήση του πίνακα της τυπικής κανονικής του Παραρτήματος Α΄.
2. Ζητείται να υπολογίσουμε την $P(Y \geq 25)$ ή, ισοδύναμα, την $P(Y > 24.5)$. Με παρόμοιο σκεπτικό με πριν έχουμε:

$$\begin{aligned} P\left(\frac{Y-20}{\sqrt{160/9}} > \frac{24.5-20}{\sqrt{160/9}}\right) &\approx P\left(Z > \frac{24.5-20}{\sqrt{160/9}}\right) \\ &= P(Z > 1.08) = 1 - P(Z < 1.08) \\ &= 1 - 0.85993 = 0.14007, \end{aligned}$$

όπου $Z \sim N(0,1)$. Επισημαίνεται ότι διατηρήσαμε αρχικά, 2 δεκαδικά ψηφία στους υπολογισμούς και έγινε χρήση του πίνακα της τυπικής κανονικής του Παραρτήματος Α΄.
Αφήνεται ως άσκηση στον/στην αναγνώστη/στρια να βρει με χρήση της R τις αντίστοιχες ακριβείς τιμές. Να σχολιάσετε τι παρατηρείτε.

Άσκηση Αυτοαξιολόγησης 7.8

Έστω $X_1, X_2, \ldots X_{144}$ ανεξάρτητες και ισόνομες τυχαίες μεταβλητές τέτοιες, ώστε $P(X_i = 1) = p$ και $P(X_i = 0) = 1 - p$. Ποια από τις παρακάτω δηλώσεις είναι σωστή;

1. Προσεγγιστικά ισχύει ότι $P(\bar{X} = 1) = p$ και $P(\bar{X} = 0) = 1 - p$.
2. $\bar{X} \xrightarrow{d} N(p, p(1-p)/12)$.
3. $\sum_{i=1}^{144} X_i \xrightarrow{d} N(144p, 144p(1-p))$.
4. Τίποτα από τα παραπάνω δεν ισχύει.

Δικαιολογήστε σύντομα την απάντησή σας.

7.3.2 Προσέγγιση της Poisson από την κανονική κατανομή

Η κανονική κατανομή, μέσω του ΚΟΘ, μπορεί επίσης να χρησιμοποιηθεί για να προσεγγίσει την κατανομή Poisson. Η προσέγγιση αυτή είναι ιδιαίτερα χρήσιμη, καθώς για μεγάλες τιμές της παραμέτρου λ της κατανομής Poisson ο υπολογισμός πιθανοτήτων, χρησιμοποιώντας είτε τη συνάρτηση πιθανότητας είτε την αθροιστική συνάρτηση κατανομής της, είναι πολύπλοκος.

Όπως είδαμε, το Κεντρικό Οριακό Θεώρημα χρησιμοποιείται για την προσέγγιση του αθροίσματος n το πλήθος ανεξάρτητων τυχαίων μεταβλητών ή της μέσης τιμής τους για μεγάλες τιμές του n. Επομένως, το ερώτημα ανάγεται στο αν η τυχαία μεταβλητή $Y \sim P(\lambda)$ μπορεί να γραφτεί ως ένα τέτοιο άθροισμα ή ως μια τέτοια μέση τιμή. Η απάντηση στο ερώτημα αυτό είναι θετική. Ειδικότερα, αν λ είναι ακέραιος αριθμός, τότε

από την Πρόταση 7.3 έχουμε ότι, αν $X_i \sim P(1),\ i = 1, \dots, \lambda,$ ανεξάρτητες τυχαίες μεταβλητές, τότε $Y = \sum_{i=1}^{\lambda} X_i$ ακολουθεί $P(\lambda)$. Επομένως, μπορεί να γραφτεί ως άθροισμα λ το πλήθος ανεξάρτητων και ισόνομων τυχαίων μεταβλητών που η καθεμία ακολουθεί $P(1)$, δηλαδή μια κατανομή Poisson με μέση τιμή και διακύμανση ίση με 1 και να χρησιμοποιηθεί το ΚΟΘ για μεγάλες τιμές του λ. Δηλαδή, για μεγάλες τιμές του λ έχουμε ότι

$$\frac{Y - \lambda}{\sqrt{\lambda}} \xrightarrow{d} N(0,1).$$

Θα αναρωτηθεί κανείς αν το παραπάνω αποτέλεσμα αποδεικνύεται και για περιπτώσεις όπου η παράμετρος λ δεν είναι ακέραιος αριθμός. Η απάντηση είναι θετική, όμως η απόδειξη αυτού παραλείπεται καθώς, εκτός από το Κεντρικό Οριακό Θεώρημα, χρησιμοποιεί και ένα άλλο αποτέλεσμα της ασυμπτωτικής θεωρίας, τουτέστιν το Θεώρημα του Slutsky (ο/η ενδιαφερόμενος/η αναγνώστης/στρια παραπέμπεται στο σύγγραμμα του van der Vaart, 1998). Τέλος, η προσέγγιση είναι ικανοποιητική για $\lambda > 10$ (βλ. Rosner, 2015).

Άσκηση Αυτοαξιολόγησης 7.9: Κουτρουβέλης (2000)

Η άφιξη οχημάτων σε μία γέφυρα μπορεί να θεωρηθεί διαδικασία Poisson με μέσο ρυθμό αφίξεων 12 οχήματα το λεπτό. Να υπολογιστούν προσεγγιστικά οι πιθανότητες κατά τη διάρκεια της επόμενης ώρας να φτάσουν στη γέφυρα περισσότερα από 11 οχήματα το λεπτό κατά μέσο όρο.

7.4 Ασκήσεις

Άσκηση 7.1 *Έστω X και Y ανεξάρτητες και ισόνομες τ.μ. που ακολουθούν την Ομοιόμορφη κατανομή στο διάστημα (a,b). Να προσδιοριστεί η κατανομή των τ.μ. $min\{X,Y\}$, $max\{X,Y\}$ και $X+Y$.*

Άσκηση 7.2 *Έστω X και Y ανεξάρτητες και ισόνομες τ.μ. που ακολουθούν την $N(0,1)$. Να προσδιοριστεί η κατανομή της τ.μ. $Z=X/Y$.*

Άσκηση 7.3 *Έστω X και Y ανεξάρτητες και ισόνομες τ.μ. που ακολουθούν τη γάμμα κατανομή με παραμέτρους (6, 2) και (7, 2), αντίστοιχα. Να προσδιοριστεί η κατανομή της τ.μ. $X+Y$.*

Άσκηση 7.4 *Ένα τυχαίο πείραμα με δύο δυνατά αποτελέσματα (Επιτυχία-Αποτυχία) εκτελείται 1600 φορές. Θεωρούμε ότι σε κάθε επανάληψη του τυχαίου πειράματος η πιθανότητα επιτυχίας p παραμένει αμετάβλητη και το αποτέλεσμα κάθε επανάληψης είναι ανεξάρτητο από το αποτέλεσμα οποιασδήποτε άλλης. Στο πλαίσιο αυτό, έστω X η τ.μ. που παριστάνει τον αριθμό των επιτυχιών στις 1600 επαναλήψεις του τυχαίου πειράματος. Για $p=0.25$ να βρεθεί η ακριβής και η προσεγγιστική τιμή της πιθανότητας $P(350\leq X<450)$.*

Άσκηση 7.5 *Από μια τράπουλα που έχει 52 φύλλα εκλέγουμε με επανατοποθέτηση 1600 φύλλα. Έστω X η τ.μ. που παριστάνει τον αριθμό των καρό που επιλέγονται στα 1600 αυτά φύλλα. Να βρεθεί η προσεγγιστική τιμή της $P(165\leq X\leq 225)$.*

Άσκηση 7.6 *Η ποσότητα της νικοτίνης που περιέχεται σε ένα τσιγάρο συγκεκριμένης μάρκας είναι τυχαία μεταβλητή με μέση τιμή 0.8 mgr και τυπική απόκλιση 0.1 mgr. Αν ένα άτομο καπνίζει κάθε βδομάδα 5 πακέτα των 20 τσιγάρων, να βρεθεί κατά προσέγγιση η πιθανότητα η συνολική ποσότητα νικοτίνης, στην οποία θα εκτεθεί σε μία εβδομάδα, να είναι τουλάχιστον 82 mgr.*

Άσκηση 7.7 *Ο χρόνος λειτουργίας 5 αντλιών νερού που είναι συνδεδεμένες σε σειρά περιγράφεται από ανεξάρτητες και ισόνομες κατανομές που ακολουθούν την εκθετική κατανομή με μέση τιμή 0.5 έτη. Το δίκτυο του νερού θεωρείται ότι βρίσκεται σε λειτουργία αν και μόνο αν λειτουργούν και οι 5 αντλίες νερού, δηλαδή παύει να λειτουργεί αν έστω και μία αντλία τεθεί εκτός λειτουργίας. Να βρεθεί η πιθανότητα ο χρόνος λειτουργίας του δικτύου νερού να είναι μεγαλύτερος από 0.75 έτη.*

Άσκηση 7.8 *Ο αριθμός των ελαττωματικών προϊόντων ενός εργοστασίου περιγράφεται από την κατανομή Poisson με μέση τιμή 3 προϊόντα σε μία εβδομάδα. Να υπολογιστεί η πιθανότητα σε 80 ανεξάρτητες εβδομάδες λειτουργίας, ο συνολικός αριθμός των ελαττωματικών προϊόντων να μην υπερβαίνει τα 230.*

Άσκηση 7.9 *Τέσσερα άτομα επιλέγουν τυχαία από έναν αριθμό στο διάστημα (0, 10), ανεξάρτητα ο ένας από τον άλλο. Να υπολογιστούν: i) η πιθανότητα όλοι οι αριθμοί που επιλέχθηκαν να είναι μεγαλύτεροι του 5 και ii) η πιθανότητα ο μικρότερος αριθμός που επιλέχθηκε να βρίσκεται στο διάστημα (2,4).*

Άσκηση 7.10 *Έστω $X_1,\ldots,X_n$, n το πλήθος ανεξάρτητες και ισόνομες τυχαίες κατανομές με συνάρτηση πυκνότητας πιθανότητας*

$$f(x)=(1+\theta)x^{\theta},\ 0<x<1.$$

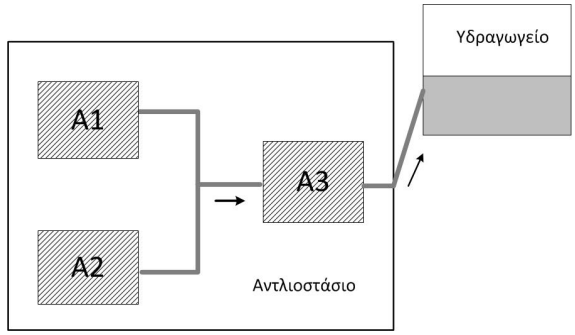
Να προσδιοριστεί η κατανομή του $W=\max\{X_1,\ldots,X_n\}$ και να υπολογιστεί η μέση της $W=\max\{X_1,\ldots,X_n\}$.

Άσκηση 7.11 *Έστω $(X,Y)^t$ τυχαίο διάνυσμα με από κοινού συνάρτηση πυκνότητας πιθανότητας*

$$f_{X,Y}(x,y) = \begin{cases} 3x & 0 \le y \le x \le 1 \\ 0 & \text{αλλού} \end{cases}$$

Να υπολογιστεί η κατανομή της $U = X/Y$.

Άσκηση 7.12 *Ένα αντλιοστάσιο έχει δύο ηλεκτρικές αντλίες Α1 και Α2 συνδεδεμένες παράλληλα και μία τρίτη Α3 συνδεδεμένη σε σειρά με τις προηγούμενες (όπως φαίνεται στο σχήμα).*



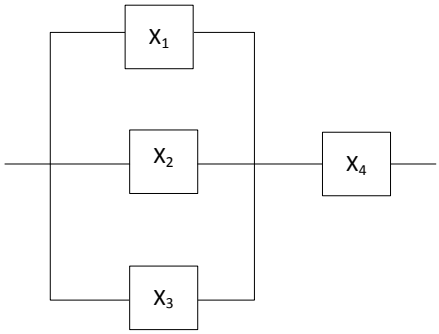
Αν ο χρόνος ζωής της κάθε αντλίας ακολουθεί εκθετική κατανομή με παράμετρο λ και η καθεμία αντλία λειτουργεί ανεξάρτητα από την άλλη

1. *να βρεθεί η κατανομή του χρόνου λειτουργίας (χρόνου ζωής) του αντλιοστασίου και*
2. *να υπολογιστεί ο μέσος χρόνος ζωής του αντλιοστασίου.*

Άσκηση 7.13 *Έστω ένα σύστημα που αποτελείται από τέσσερα υποσυστήματα, συνδεδεμένα με τον τρόπο που φαίνεται στο παρακάτω σχήμα. Έστω X_i, $i = 1, \dots, 4$ η τ.μ. που παριστάνει τον χρόνο ζωής του i-οστού υποσυστήματος, ο οποίος υποθέτουμε ότι ακολουθεί εκθετική κατανομή με παράμετρο $\lambda_i > 0$. Έστω T η τ.μ. που παριστάνει τον χρόνο ζωής του συστήματος. Ως αξιοπιστία του συστήματος $R(t)$ στον χρόνο t ορίζουμε την πιθανότητα να λειτουργεί το σύστημα στον χρόνο t*

$$R(t) = P(\text{το σύστημα λειτουργεί στον χρόνο } t)$$

Υπολογίστε την αξιοπιστία του συγκεκριμένου συστήματος για $t = 2$ χρονικές μονάδες.



Άσκηση 7.14 *Έστω X_1 και X_2 ανεξάρτητες τ.μ. που ακολουθούν την Εκθετική κατανομή με παράμετρο 2. Να προσδιορίσετε την από κοινού κατανομή των $U = X_1 + X_2$ και $V = \frac{X_1}{X_1+X_2}$ και να εξετάσετε αν είναι ανεξάρτητες.*

7.5 Απαντήσεις στις ασκήσεις αυτοαξιολόγησης

Λύση Άσκησης Αυτοαξιολόγησης 7.1

Η συνάρτηση πυκνότητας πιθανότητας της τυχαίας μεταβλητής X_i που ακολουθεί ομοιόμορφη κατανομή στο $(0,\alpha)$ δίνεται από τη σχέση:

$$f_{X_i}(x) = \begin{cases} \frac{1}{\alpha}, & 0 < x < \alpha, \\ 0, & \text{αλλού.} \end{cases}$$

Από την παραπάνω σχέση προκύπτει εύκολα ότι η αθροιστική συνάρτηση κατανομής των X_i δίνεται από τη σχέση

$$F_{X_i}(x_i) = \begin{cases} 0, & x < 0, \\ \frac{x}{\alpha}, & 0 \le x < \alpha, \\ 1, & x \ge \alpha. \end{cases}$$

Επομένως, η αθροιστική συνάρτηση κατανομής της W ισούται με

$$F_W(w) = [F_X(w)]^k = \begin{cases} 0, & w \le 0, \\ \left(\frac{w}{\alpha}\right)^k, & 0 < w < \alpha, \\ 1, & w \ge \alpha. \end{cases}$$

Η συνάρτηση πυκνότητας πιθανότητας της W, για $0 < w < \alpha$, δίνεται από τη σχέση:

$$f_W(w) = k\frac{1}{\alpha}\left(\frac{w}{\alpha}\right)^{k-1}, 0 < w < \alpha.$$

Η αναμενόμενη τιμή της W ισούται με

$$E(W) = \int_0^\alpha w f_W(w) dw = \int_0^\alpha w \frac{k}{\alpha}\left(\frac{w}{\alpha}\right)^{k-1} dw = \frac{\alpha k}{1+k}.$$

Λύση Άσκησης Αυτοαξιολόγησης 7.2

Γνωρίζουμε ότι η ροπογεννήτρια συνάρτηση της Poisson(λ_i) δίνεται από τη σχέση

$$M_X(t) = e^{\lambda_i e^t - \lambda_i}.$$

Από την άλλη, η ροπογεννήτρια συνάρτηση της τυχαίας μεταβλητής $Y = X_1 + \ldots + X_k$ δίνεται από τη σχέση

$$M_Y(y) = E(e^{tX_1}) \cdots E(e^{tX_k}) = e^{\lambda_1 e^t - \lambda_1} \cdots e^{\lambda_k e^t - \lambda_k} = e^{\sum_{i=1}^k \lambda_i e^t - \sum_{i=1}^k \lambda_i},$$

η οποία ταυτίζεται με τη ροπογεννήτρια συνάρτηση της Poisson κατανομής με παράμετρο $\sum_{i=1}^k \lambda_i$, γεγονός που αποδεικνύει το ζητούμενο.

Λύση Άσκησης Αυτοαξιολόγησης 7.3

Καθώς οι τ.μ. X_i, $i = 1, \ldots, k$ είναι ανεξάρτητες, ισχύει ότι:

$$M_Y(y) = E(e^{tY}) = E(e^{t(X_1+\ldots+X_k)}) = E(e^{tX_1} \cdots e^{tX_k})$$

$$= E(e^{tX_1}) \cdots E(e^{tX_k}) = \prod_{i=1}^k M_{X_i}(t).$$

Χρησιμοποιώντας τη σχέση (4.11) για τη ροπογεννήτρια της διωνυμικής κατανομής, σε αυτήν την περίπτωση έχουμε ότι:

$$M_{X_i}(t) = \left(q + pe^t\right)^{n_i}.$$

Επομένως, είναι

$$M_Y(t) = \prod_{i=1}^{k} \left(q + pe^t\right)^{n_i} = \left(q + pe^t\right)^{\sum_{i=1}^{k} n_i},$$

που ταυτίζεται με τη ροπογεννήτρια της διωνυμικής με παραμέτρους $\sum_{i=1}^{k} n_i$ και p.

Λύση Άσκησης Αυτοαξιολόγησης 7.4

Καθώς είναι οι τ.μ. X_i, $i = 1, \ldots, k$ ανεξάρτητες ισχύει ότι

$$M_Y(y) = E(e^{tY}) = E(e^{t(X_1 + \ldots + X_k)}) = E(e^{tX_1} \cdots e^{tX_k})$$

$$= E(e^{tX_1}) \cdots E(e^{tX_k}) = \prod_{i=1}^{k} M_{X_i}(t).$$

Χρησιμοποιώντας τη σχέση (4.30) για τη ροπογεννήτρια της αρνητικής διωνυμικής, σε αυτήν την περίπτωση έχουμε ότι:

$$M_{X_i}(t) = \left(\frac{pe^t}{1 - (1-p)e^t}\right)^{n_i}, \quad \text{για } t < -\log(1-p).$$

Επομένως, είναι:

$$M_Y(t) = \prod_{i=1}^{k} \left(\frac{pe^t}{1 - (1-p)e^t}\right)^{n_i} = \left(\frac{pe^t}{1 - (1-p)e^t}\right)^{\sum_{i=1}^{k} n_i}, \quad \text{για } t < -\log(1-p)$$

που ταυτίζεται με τη ροπογεννήτρια της αρνητικής διωνυμικής με παραμέτρους $\sum_{i=1}^{k} n_i$ και p.

Λύση Άσκησης Αυτοαξιολόγησης 7.5

1. Έστω X_i η τυχαία μεταβλητή που παριστάνει το βάρος του i-οστού ατόμου $i = 1, \ldots, 5$. Από την εκφώνηση έχουμε ότι $X_i \sim N(80, 8^2)$ και, επομένως, άμεσα συμπεραίνουμε ότι $\bar{X} = \frac{\sum_{i=1}^{5} X_i}{n} \sim N(80, 8^2/5)$. Θέλουμε να προσδιορίσουμε την $P(\bar{X} > 85)$ με $\bar{X} \sim N(80, 8^2/5)$. Χρησιμοποιώντας τον τυπικό μετασχηματισμό

$$Z = \frac{\bar{X} - 80}{\sqrt{8^2/5}} \sim N(0,1)$$

έχουμε ότι:

$$P(\bar{X} > 85) = P\left(Z > \frac{85 - 80}{8/\sqrt{5}}\right)$$

$$= P(Z > 1.4) = 1 - P(Z < 1.4)$$

$$= 1 - 0.91924 = 0.08076$$

όπου χρησιμοποιήσαμε τον πίνακα της τυπικής κανονικής του Παραρτήματος Α΄ (στις αρχικές πράξεις διατηρήθηκαν δύο δεκαδικά ψηφία).

2. Έστω X_i η τυχαία μεταβλητή που παριστάνει το βάρος του i-οστού ατόμου $i = 1, \ldots, 5$. Από την εκφώνηση έχουμε ότι $X_i \sim N(80, 8^2)$ και, επομένως, άμεσα συμπεραίνουμε ότι $\sum_{i=1}^{5} X_i \sim N(5 \cdot 80,\ 5 \cdot 8^2)$. Θέλουμε να προσδιορίσουμε την $P\left(\sum_{i=1}^{5} X_i > 450\right)$ με $\sum_{i=1}^{5} X_i \sim N(400,\ 5 \cdot 8^2)$. Χρησιμοποιώντας τον τυπικό μετασχηματισμό

$$Z = \frac{\sum_{i=1}^{5} X_i - 400}{\sqrt{5 \cdot 8^2}} \sim N(0,1)$$

έχουμε ότι:

$$\begin{aligned} P\left(\sum_{i=1}^{5} X_i > 450\right) &= P\left(Z > \frac{450 - 400}{8 \cdot \sqrt{5}}\right) \\ &= P(Z > 2.8) \\ &= 1 - P(Z < 2.8) \\ &= 1 - 0.99744 = 0.00256 \end{aligned}$$

όπου χρησιμοποιήσαμε τον πίνακα της τυπικής κανονικής του Παραρτήματος Α΄.

Λύση Άσκησης Αυτοαξιολόγησης 7.6

Σωστή απάντηση είναι η (β`), καθώς $Var(\bar{X}) = \sigma^2/n$ και, επομένως, η τυπική απόκλιση του δειγματικού μέσου ισούται με $\sigma/\sqrt{n} = 100/\sqrt{100} = 100/10 = 10$.

Λύση Άσκησης Αυτοαξιολόγησης 7.7

1. Έχουμε ότι $X_i \sim Exp(0.5)$, επομένως από τις ιδιότητες της Εκθετικής κατανομής προκύπτει ότι: $E(X_i) = 1/0.5 = 2$ και $Var(X_i) = 1/0.5^2 = 4$. Επομένως, από το Κεντρικό Οριακό Θεώρημα για μεγάλο μέγεθος n ισχύει ότι $\bar{X} \xrightarrow{d} N(2,\ 4/144)$ και $\sqrt{Var\bar{X}} = \sqrt{4/144} = 1/6$. Από τα παραπάνω, εύκολα προκύπτει ότι σωστή απάντηση είναι η (β`).
2. Έχουμε ότι $X_i \sim U(0,12)$, επομένως από τις ιδιότητες της Ομοιόμορφης κατανομής προκύπτει ότι: $E(X_i) = 6$ και $Var(X_i) = 12^2/12 = 12$. Επομένως, από το Κεντρικό Οριακό Θεώρημα για μεγάλο μέγεθος n ισχύει ότι $\bar{X} \xrightarrow{d} N(6,\ 12/144)$, δηλαδή $\bar{X} - 6 \xrightarrow{d} N(0,\ 12/144)$. Από τα παραπάνω, εύκολα προκύπτει ότι σωστή απάντηση είναι η (δ`).

Λύση Άσκησης Αυτοαξιολόγησης 7.8

Ουσιαστικά έχουμε ότι $X_i \sim B(1,p)$, επομένως από τις ιδιότητες της διωνυμικής κατανομής προκύπτει ότι: $E(X_i) = p$ και $Var(X_i) = p(1-p)$. Επομένως, από το Κεντρικό Οριακό Θεώρημα για μεγάλο μέγεθος n (σε αυτήν την περίπτωση $n = 144$) ισχύει ότι $\bar{X} \xrightarrow{d} N\left(p,\ \frac{p(1-p)}{n}\right)$ και $\sum_{i=1}^{n} X_i \xrightarrow{d} N\left(np, np(1-p)\right)$. Από τα παραπάνω, εύκολα προκύπτει ότι σωστή απάντηση είναι η (γ`). Τέλος, η απάντηση (α`) προφανώς είναι λανθασμένη, αφού το σύνολο των δυνατών τιμών της τυχαίας μεταβλητής $\bar{X}$ είναι το $\{0, 1/n, \ldots, 1\}$ (βλ. και την Παρατήρηση 7.8).

Λύση Άσκησης Αυτοαξιολόγησης 7.9

Έστω $X_1, X_2, \ldots, X_{60}$ το πλήθος των οχημάτων που φτάνουν στη γέφυρα σε καθένα από τα επόμενα 60 λεπτά της ώρας. Επειδή οι αφίξεις πραγματοποιούνται σύμφωνα με τη διαδικασία Poisson με $\lambda = 12$ ανά λεπτό, οι τ.μ. $X_1, X_2, \ldots, X_{60}$ είναι ένα τυχαίο δείγμα από την κατανομή Poisson με $\lambda = 12$. Επειδή το n είναι μεγάλο, ισχύει το ΚΟΘ, οπότε θα έχουμε ότι

$$\bar{X} \xrightarrow{d} N\left(12, \frac{12}{60}\right).$$

Θέλουμε να προσδιορίσουμε την $P(\bar{X} > 11)$ με $\bar{X} \xrightarrow{d} N(12, \frac{12}{60})$. Καθώς η τ.μ. $\bar{X}$ προσεγγίζεται από την κανονική με μέση τιμή 12 και διακύμανση 12/60 έχουμε, εφαρμόζοντας διόρθωση συνέχειας και το Κεντρικό Οριακό Θεώρημα, ότι:

$$\begin{aligned} P(\bar{X} > 11.0) = P(\bar{X} \geq 11.5) &= P\left(\frac{\bar{X} - 12}{\sqrt{\frac{12}{60}}} \geq \frac{11.5 - 12}{\sqrt{\frac{12}{60}}}\right) \\ &\approx P\left(Z \geq \frac{11.5 - 12}{\sqrt{\frac{12}{60}}}\right) \\ &= P(Z \geq -1.12) \\ &= P(Z < 1.12) \\ &= 0.86864 \end{aligned}$$

όπου $Z \sim N(0,1)$. Επισημαίνεται ότι διατηρήσαμε αρχικά, 2 δεκαδικά ψηφία στους υπολογισμούς και έγινε χρήση του πίνακα της τυπικής κανονικής κανονικής του Παραρτήματος Α΄.

ΒΙΒΛΙΟΓΡΑΦΙΑ

Ελληνόγλωσση

Κούτρας, Μ. (2005). *Εισαγωγή στις Πιθανότητες. Μέρος ΙΙ (Θεωρία και Εφαρμογές)*. Εκδόσεις Σταμούλη.

Κουτρουβέλης, Ι. Α. (2000). *Βασικά Εργαλεία και Μέθοδοι για τον Έλεγχο Ποιότητας: Πιθανότητες και Στατιστική ΙΙ (Τόμος Β΄)*. Πάτρα: Ελληνικό Ανοικτό Πανεπιστήμιο.

Κουτρουβέλης, Ι. Α. (2011). *Εφαρμοσμένες πιθανότητες και στατιστική*. Συμμετρία.

Παπαϊωάννου, Τ. (1997). *Θεωρία πιθανοτήτων και στατιστικής*. Σταμούλη Α.Ε.

Ξενόγλωσση

Fischer, H. (2011). *A history of the central limit theorem. From Classical to Modern Probability Theory*. Springer, New York.

Fisher, R. (1925). Applications of "Student's" distribution. *Metron*, *5*, pp. 90–104.

Rosner, B. (2015). *Fundamentals of Biostatistics, Eight Edition.* Cengage Learning.

Stuart, A. and Ord, J. K. (1987). *Kendall's Advanced Theory of Statistics. Volume 1. Distribution Theory* (5th ed.). Griffin.

Student (1908). The Probable Error of a Mean. *Biometrika*, *6*, pp. 1–25.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press.

Μέρος ΙΙ

ΣΤΑΤΙΣΤΙΚΗ

ΚΕΦΑΛΑΙΟ 8

ΠΕΡΙΓΡΑΦΙΚΗ ΣΤΑΤΙΣΤΙΚΗ

Σύνοψη

Σε αυτό το κεφάλαιο παρουσιάζονται βασικά στοιχεία της Περιγραφικής Στατιστικής. Η Περιγραφική Στατιστική παρέχει τα βασικά εργαλεία συνοπτικής περιγραφής και παρουσίασης των πληροφοριών που περιέχονται στις παρατηρήσεις ενός τυχαίου δείγματος από έναν πληθυσμό. Στο κεφάλαιο αυτό, το ενδιαφέρον μας επικεντρώνεται στην έννοια του τυχαίου δείγματος και των μεθόδων συνοπτικής παρουσίασης ποιοτικών και ποσοτικών δεδομένων μέσω πινάκων, γραφημάτων και περιγραφικών μέτρων. Τέλος, αξίζει να σημειωθεί ότι το κεφάλαιο αυτό έχει ως επιπρόσθετο στόχο να λειτουργήσει ως συνδετικός κρίκος μεταξύ της θεωρίας πιθανοτήτων και των τυχαίων μεταβλητών, που παρουσιάστηκαν στα προηγούμενα κεφάλαια, και της στατιστικής συμπερασματολογίας, η οποία θα αποτελέσει το κεντρικό αντικείμενο μελέτης των επόμενων κεφαλαίων.

Προαπαιτούμενη γνώση: Βασικές έννοιες τυχαίων μεταβλητών.

Προσδοκώμενα μαθησιακά αποτελέσματα: Όταν θα έχετε ολοκληρώσει τη μελέτη του κεφαλαίου αυτού, θα γνωρίζετε:

- την έννοια του τυχαίου δείγματος,
- να διακρίνετε τις διάφορες κατηγορίες μεταβλητών,
- να επιλέγετε κατάλληλες τεχνικές συνοπτικής παρουσίασης και περιγραφής των δεδομένων και
- να σχολιάζετε και να ερμηνεύετε τα αποτελέσματά τους.

Οικονόμου, Π., Μαλεφάκη, Σ., & Μπατσίδης, Α. (2023). *Πιθανότητες - Στατιστική.*
[Προπτυχιακό εγχειρίδιο].

Γλωσσάριο επιστημονικών όρων:

- Αθροιστική συχνότητα
- Δείγμα
- Δειγματική διάμεσος
- Δειγματική διασπορά ή διακύμανση
- Δειγματικό εύρος
- Δειγματική μέση τιμή
- Δειγματική τυπική απόκλιση
- Διάγραμμα διασκόρπισης-διασποράς
- Διάγραμμα μόσχου-φύλλου
- Θηκόγραμμα
- Ιστόγραμμα
- Κορυφή
- Πίνακας συχνοτήτων
- Πληθυσμός
- Ποιοτικά δεδομένα
- Ποσοτικά δεδομένα
- Ραβδόγραμμα
- Συντελεστής μεταβλητότητας
- Συχνότητα
- Σχετική Συχνότητα

8.1 Εισαγωγή

Στα προηγούμενα κεφάλαια παρουσιάστηκε η έννοια της τυχαίας μεταβλητής. Οι τυχαίες μεταβλητές παρέχουν τα απαραίτητα εργαλεία για τη θεωρητική περιγραφή διάφορων τυχαίων φαινομένων, αλλά και διάφορων χαρακτηριστικών των μελών ενός πληθυσμού. Με τον όρο πληθυσμό αναφερόμαστε σε μια ομάδα οντοτήτων ή καταστάσεων ή γεγονότων, της οποίας τα μέλη μοιράζονται κοινά χαρακτηριστικά, τα οποία χαρακτηριστικά διέπονται από κάποιο βαθμό τυχαιότητας/αβεβαιότητας. Στόχος μας είναι η μελέτη κάποιου/κάποιων χαρακτηριστικού/χαρακτηριστικών της ομάδας ενδιαφέροντος. Παραδείγματα τέτοιων χαρακτηριστικών είναι:

- ο ελάχιστος χρόνος ζωής μιας ηλεκτρικής συσκευής,
- η αντοχή του σκυροδέματος μιας εταιρείας,
- ο αριθμός μηνυμάτων ηλεκτρονικού ταχυδρομείου που λαμβάνουμε στον υπολογιστή μας κατά τη διάρκεια μιας ημέρας,
- ο αριθμός των τερμάτων που επιτυγχάνονται σε έναν ποδοσφαιρικό αγώνα.

Στις περισσότερες περιπτώσεις, η κατανομή που διέπει το υπό μελέτη τυχαίο φαινόμενο ή τον υπό μελέτη πληθυσμό είναι άγνωστη ή, ακόμα και αν είναι γνωστή η συναρτησιακή της μορφή, δηλαδή η οικογένεια στην οποία ανήκει (π.χ. κανονική κ.ά.), δεν είναι πλήρως καθορισμένη, αφού οι παράμετροί της είναι άγνωστες.

Για τις περιπτώσεις αυτές, καθώς η μελέτη ολόκληρου του πληθυσμού τις περισσότερες φορές είναι πολύ δύσκολη, αν όχι αδύνατη, καλούμαστε να λάβουμε ένα υποσύνολο του πληθυσμού, το οποίο καλείται δείγμα του πληθυσμού, και να εξάγουμε συμπεράσματα για τον πληθυσμό που μας ενδιαφέρει μελετώντας το δείγμα και γενικεύοντας τα συμπεράσματα από το δείγμα στον πληθυσμό. Οι μέθοδοι που έχουν ως αντικείμενο τη συλλογή, οργάνωση, παρουσίαση και μελέτη των δεδομένων ενός δείγματος αναφέρονται ως Περιγραφική Στατιστική και αποτελούν το αντικείμενο του κεφαλαίου αυτού. Από την άλλη πλευρά, τα απαραίτητα εργαλεία για τη γενίκευση των συμπερασμάτων από τη μελέτη του δείγματος στον πληθυσμό αναφέρονται ως Στατιστική Συμπερασματολογία ή Επαγωγική Στατιστική και κάποια αποτελέσματα αυτής παρουσιάζονται στα επόμενα κεφάλαια.

Αν και έχουμε ήδη χρησιμοποιήσει τους όρους πληθυσμός και δείγμα, κρίνεται απαραίτητο, προτού προχωρήσουμε στην Περιγραφική Στατιστική, να δοθούν οι αυστηροί ορισμοί των εννοιών αυτών.

Ορισμός 8.1

Με τον όρο **πληθυσμό** ορίζουμε κάθε καλά ορισμένη συλλογή οντοτήτων ή καταστάσεων ή αντικειμένων ή ατόμων που μοιράζονται κοινά χαρακτηριστικά, τα οποία διέπονται από κάποιο βαθμό τυχαιότητας. Ο όρος πληθυσμός περιλαμβάνει και όλες τις δυνατές περιπτώσεις επαναλαμβανόμενων μετρήσεων/παρατηρήσεων/πειραμάτων για τη μελέτη φαινομένων. Το πλήθος των στοιχείων ενός πληθυσμού καλείται **μέγεθος του πληθυσμού**.

Από τον ορισμό της έννοιας του πληθυσμού γίνεται φανερό ότι το μέγεθος ενός πληθυσμού μπορεί να είναι πεπερασμένο, όπως π.χ. οι φοιτητές του Πανεπιστημίου Πατρών, οι κάτοικοι μιας πόλης κ.ά., ή άπειρο (αριθμήσιμο ή υπεραριθμήσιμο), όπως π.χ. οι παρατηρήσεις που παίρνουμε μετρώντας την ατμοσφαιρική ρύπανση, τη θερμοκρασία, την ταχύτητα του αέρα κάθε μέρα κ.ά. Είναι προφανές ότι η μελέτη όλων των μελών ενός πληθυσμού άπειρου μεγέθους είναι αδύνατη. Ακόμα όμως και στην περίπτωση ενός πληθυσμού πεπερασμένου μεγέθους, η συνολική μελέτη των μελών του τις περισσότερες φορές είναι πρακτικά αδύνατη λόγω χρονικών ή/και οικονομικών περιορισμών. Για τον λόγο αυτό, καταφεύγουμε στη λήψη ενός δείγματος.

Ορισμός 8.2

Δείγμα ενός πληθυσμού ονομάζεται κάθε υποσύνολο αυτού του πληθυσμού. Το πλήθος των στοιχείων του δείγματος καλείται **μέγεθος του δείγματος**.

Είναι φανερό ότι οι n το πλήθος παρατηρήσεις ενός δείγματος αποτελούν ένα μέρος του πληθυσμού, το οποίο περιλαμβάνει σημαντικές πληροφορίες για τη συμπεριφορά του πληθυσμού. Το μέγεθος του δείγματος είναι ένα σημαντικό χαρακτηριστικό του, αφού μεγαλύτερα σε μέγεθος δείγματα περιέχουν περισσότερη πληροφορία, αλλά δεν είναι το σημαντικότερο. Ένα δείγμα πρέπει να είναι αντιπροσωπευτικό του πληθυσμού, δηλαδή να είναι τέτοιο ώστε όλες οι υποομάδες του πληθυσμού να αντιπροσωπεύονται στο δείγμα και το δείγμα να αποτελεί κατά αυτόν τον τρόπο μια μικρογραφία του πληθυσμού. Εν τοιαύτη περιπτώσει, είναι σχετικά εύκολη η γενίκευση των συμπερασμάτων από το δείγμα στον πληθυσμό, όπως θα δούμε σε επόμενα κεφάλαια.

Από τα παραπάνω γίνεται εύκολα αντιληπτό ότι ο τρόπος επιλογής ενός δείγματος είναι ιδιαίτερα κρίσιμος για την ποιότητα των συμπερασμάτων που θα εξαχθούν για τον υπό μελέτη πληθυσμό. Υπάρχουν πολλοί τρόποι επιλογής δείγματος, αλλά εδώ θα αναφερθούμε στην απλή δειγματοληψία και στην απλή τυχαία δειγματοληψία. Κατά την απλή τυχαία δειγματοληψία λαμβάνονται n το πλήθος παρατηρήσεις από έναν πληθυσμό χωρίς επανατοποθέτηση, έτσι ώστε κάθε συλλογή n το πλήθος διαφορετικών στοιχείων να έχει την ίδια πιθανότητα να παρατηρηθεί. Κάθε παρατήρηση από τον πληθυσμό είναι μία τιμή από την τυχαία μεταβλητή X, που παριστάνει το υπό μελέτη τυχαίο φαινόμενο. Στην περίπτωση αυτή, οι n το πλήθος παρατηρήσεις $x_1, \dots, x_n$ αποτελούν την πραγματοποίηση n το πλήθος ισόνομων και εξαρτημένων τυχαίων μεταβλητών $X_1, \dots, X_n$. Η τυχαία δειγματοληψία διαφοροποιείται από την απλή τυχαία δειγματοληψία ως προς το γεγονός ότι οι παρατηρήσεις λαμβάνονται με επανατοποθέτηση. Στην περίπτωση αυτή, οι παρατηρήσεις του τυχαίου δείγματος αποτελούν την πραγματοποίηση n το πλήθος ισόνομων και ανεξάρτητων τυχαίων μεταβλητών. Είναι φανερό ότι οι δύο προαναφερθείσες τεχνικές πρακτικά ταυτίζονται στην περίπτωση όπου ο υπό μελέτη πληθυσμός έχει άπειρα μέλη ή το μέγεθός του είναι αρκετά μεγάλο σε σχέση με το μέγεθος του δείγματος n. Για τον λόγο αυτό, στη συνέχεια του συγγράμματος θα αναφερόμαστε σε τυχαία δείγματα, χωρίς να γίνεται ιδιαίτερη διάκριση μεταξύ των τυχαίων και των απλών τυχαίων δειγμάτων.

Το υπόλοιπο κεφάλαιο επικεντρώνεται στην παρουσίαση βασικών εργαλείων της Περιγραφικής Στατιστικής για τη συνοπτική παρουσίαση των πληροφοριών του δείγματος, όπως είναι κατάλληλοι στατιστικοί πίνακες (πίνακας συχνοτήτων, ομαδοποιημένος πίνακας κ.ά.) και γραφικές παραστάσεις (ραβδόγραμμα, κυκλικό διάγραμμα, ιστόγραμμα, θηκόγραμμα κ.ά.), καθώς και διάφορα αριθμητικά μέτρα (δειγματική μέση τιμή, δειγματική διακύμανση, δειγματική τυπική απόκλιση, δειγματική διάμεσος, κορυφή, δειγματικός συντελεστής μεταβλητότητας, λοξότητας και κύρτωσης κ.ά.). Η εφαρμογή των εργαλείων αυτών για τη συνοπτική παρουσίαση των δεδομένων εξαρτάται άμεσα από τη φύση των μεταβλητών που περιλαμβάνονται σε αυτά. Για τον λόγο αυτό, στις επόμενες ενότητες, προτού παρουσιαστούν οι προαναφερθείσες μέθοδοι παρουσίασης και ανάλυσης των δεδομένων, προηγείται μια σύντομη αναφορά στις δύο βασικές κατηγορίες δεδομένων, δηλαδή των ποιοτικών και των ποσοτικών δεδομένων.

Παρατήρηση 8.1

Κλείνοντας την ενότητα αυτή, αξίζει να σημειωθεί ότι οι παρατηρήσεις $x_1, \dots, x_n$ ενός συγκεκριμένου τυχαίου δείγματος, δηλαδή η πραγματοποίηση των n τυχαίων μεταβλητών $X_1, \dots, X_n$ που αντιστοιχούν στην πρώτη, ..., στην n-οστή παρατήρηση αναφέρονται ως **δεδομένα** και αποτελούν το αντικείμενο μελέτης της περιγραφικής στατιστικής. Η μελέτη των δεδομένων αυτών αποτελεί το πρώτο βήμα για την εξαγωγή συμπερασμάτων για τον πληθυσμό υπό την προϋπόθεση ότι το διαθέσιμο δείγμα είναι ένα αντιπροσωπευτικό υποσύνολό του.

Η αξιοπιστία και ο βαθμός βεβαιότητας των συμπερασμάτων αυτών βασίζονται όχι τόσο στις αριθμητικές

τιμές κάποιων χαρακτηριστικών του δείγματος, όπως στη μέση τιμή και στη διασπορά του δείγματος που θα ορίσουμε στη συνέχεια, αλλά στη συμπεριφορά των αντίστοιχων τυχαίων μεταβλητών. Πιο συγκεκριμένα, πρέπει να σημειωθεί ότι οποιαδήποτε ποσότητα υπολογίζεται επάνω στις παρατηρήσεις ενός δείγματος δεν είναι τίποτα άλλο παρά η πραγματοποίηση μιας τυχαίας μεταβλητής, αφού υπολογίζεται με τη βοήθεια των τιμών των τυχαίων μεταβλητών $X_1, \ldots, X_n$. Κατά συνέπεια, όπως θα δούμε στο επόμενο κεφάλαιο, η γενίκευση των συμπερασμάτων από το δείγμα στον πληθυσμό βασίζεται στην κατανομή αυτών των τυχαίων μεταβλητών, δηλαδή των τυχαίων μεταβλητών που ορίζονται ως συναρτήσεις των τυχαίων παρατηρήσεων ενός δείγματος.

8.2 Συνοπτική παρουσίαση ποιοτικών δεδομένων

Αντικείμενο μελέτης αυτής της ενότητας είναι η συνοπτική παρουσίαση ποιοτικών δεδομένων. Με τον όρο ποιοτικά δεδομένα αναφερόμαστε στα δεδομένα που λαμβάνονται από την καταγραφή των παρατηρήσεων των ποιοτικών ή αλλιώς κατηγορικών χαρακτηριστικών, δηλαδή των χαρακτηριστικών που αποτυπώνουν μη αριθμητικές τιμές και ιδιότητες. Παραδείγματα τέτοιων χαρακτηριστικών είναι:

- η εκπαιδευτική βαθμίδα,
- η ψυχολογική κατάσταση,
- το φύλο,
- η χώρα καταγωγής,
- το μοντέλο του κινητού τηλεφώνου,
- η κατάσταση λειτουργίας (on-off) μιας μηχανής,
- το χρώμα των μαλλιών.

Τέτοια χαρακτηριστικά του πληθυσμού που μεταβάλλονται κατά ποιότητα ή είδος, αλλά όχι κατά μέγεθος, είναι φανερό ότι μπορούν να μοντελοποιηθούν με τη βοήθεια διακριτών τυχαίων μεταβλητών. Από τα παραπάνω παραδείγματα, ίσως είναι φανερό ότι οι κατηγορικές μεταβλητές μπορούν να χωριστούν σε δύο μεγάλες υποκατηγορίες:

- τις ονομαστικές, που περιλαμβάνουν ποιοτικές μεταβλητές, οι τιμές των οποίων δεν έχουν μια λογική ιεράρχηση , όπως π.χ. το φύλο, το χρώμα ματιών, ο τόπος γέννησης κ.ά. Δηλαδή περιλαμβάνουν ποιοτικές μεταβλητές των οποίων οι τιμές δεν μπορούν να διαταχθούν σε τάξη μεγέθους, και
- τις διατάξιμες, που περιλαμβάνουν ποιοτικές μεταβλητές, οι τιμές των οποίων έχουν μια λογική ιεράρχηση, όπως είναι η εκπαιδευτική βαθμίδα, η οποία μπορεί να πάρει τις τιμές «πρωτοβάθμια», «δευτεροβάθμια», «τριτοβάθμια». Δηλαδή οι διατάξιμες μεταβλητές περιλαμβάνουν ποιοτικές μεταβλητές των οποίων οι τιμές μπορούν να διαταχθούν σε τάξη μεγέθους.

Βασικά εργαλεία για τη συνοπτική παρουσίαση των δεδομένων μίας ποιοτικής μεταβλητής (Ζωγράφος, 2002) αποτελούν:

α) ο πίνακας συχνοτήτων των δεδομένων και

β) κατάλληλες γραφικές παραστάσεις, που, μεταξύ άλλων, περιλαμβάνουν το ραβδόγραμμα και το κυκλικό διάγραμμα.

Όλες αυτές οι μέθοδοι ανάλυσης και παρουσίασης των ποιοτικών δεδομένων παρουσιάζονται στη συνέχεια.

8.2.1 Πίνακας συχνοτήτων ποιοτικών δεδομένων

Ο πίνακας συχνοτήτων μιας ποιοτικής μεταβλητής προκύπτει από την απαρίθμηση των παρατηρήσεων στην αντίστοιχη κατηγορία. Ένας ολοκληρωμένος πίνακας συχνοτήτων μίας ποιοτικής μεταβλητής περιλαμβάνει:

- τη στήλη των Συχνοτήτων ν_i - η συχνότητα παριστάνει τον αριθμό των φορών που μία κατηγορία της ποιοτικής μεταβλητής εμφανίζεται στο δείγμα μεγέθους n και, προφανώς, ικανοποιεί τη σχέση $0 \leq \nu_i \leq n$ και
- τη στήλη των Σχετικών συχνοτήτων f_i - η σχετική συχνότητα ισούται με τον αριθμό των φορών που μία κατηγορία της ποιοτικής μεταβλητής εμφανίζεται στο δείγμα διαιρεμένο με το μέγεθος του δείγματος n, δηλαδή $f_i = \nu_i/n$.

Σημειώνεται ότι η σχετική συχνότητα συχνά εκφράζεται και επί τοις εκατό και τότε παριστάνει το ποσοστό των φορών που μία κατηγορία της ποιοτικής μεταβλητής εμφανίζεται στο δείγμα.

Αν η ποιοτική μεταβλητή είναι μια διατάξιμη μεταβλητή πέρα των προαναφερθέντων, μπορούν να συμπεριληφθούν στον πίνακα συχνοτήτων και

- η στήλη των αθροιστικών συχνοτήτων N_i, δηλαδή το πλήθος των τιμών του δείγματος που είναι μικρότερες ή το πολύ ίσες από τη συγκεκριμένη τιμή της ποιοτικής διατάξιμης μεταβλητής, καθώς και
- η στήλη των αθροιστικών σχετικών συχνοτήτων F_i που αποτυπώνει το πλήθος των τιμών του δείγματος που είναι μικρότερες ή το πολύ ίσες από τη συγκεκριμένη τιμή της ποιοτικής διατάξιμης μεταβλητής διαιρεμένο με το μέγεθος του δείγματος n, δηλαδή $F_i = N_i/n$.

Οι αθροιστικές σχετικές συχνότητες δύναται να εκφραστούν και αυτές ως επί τοις εκατό και τότε παριστάνουν το ποσοστό επί τοις εκατό των τιμών του δείγματος που είναι μικρότερες ή ίσες από μία τιμή.

Οι ανωτέρω ποσότητες ικανοποιούν τις ιδιότητες που δίνονται στην πρόταση που ακολουθεί.

Πρόταση 8.1

Η συχνότητα ν_i, η σχετική συχνότητα f_i καθώς και οι αντίστοιχες αθροιστικές έννοιες, όποτε αυτές ορίζονται, ικανοποιούν τις ακόλουθες ιδιότητες

- $0 \leq f_i \leq 1$
- $N_i = \sum_{j=1}^{i} \nu_j$
- $F_i = \sum_{j=1}^{i} \frac{\nu_j}{n} = \sum_{j=1}^{i} f_j$
- $N_{min} = \nu_{min}$ και $N_{max} = n$,
- $F_{min} = f_{min}$ και $F_{max} = 1$,

όπου με min και max συμβολίζουμε την πρώτη (μικρότερη) και την τελευταία (μεγαλύτερη) διατεταγμένη κατηγορία.

Απόδειξη Πρότασης 8.1

Η απόδειξή τους προκύπτει εύκολα από τον ορισμό τους και αφήνεται ως άσκηση για τον/την αναγνώστη/αναγνώστρια.

Παράδειγμα 8.1

Σε μια έρευνα για το επίπεδο ικανοποίησης από την παρεχομένη εκπαίδευση σε ένα Πανεπιστημιακό ίδρυμα επιλέχθηκαν τυχαία 100 φοιτητές και ρωτήθηκαν για το επίπεδο σπουδών τους. Από τα 100 άτομα τα 20 δήλωσαν ότι είναι λίγο ικανοποιημένοι, ενώ ο αριθμός των φοιτητών που δήλωσαν μέτρια και πολύ ικανοποιημένοι ήταν 25 και 55, αντίστοιχα. Να κατασκευαστεί ο πίνακας συχνοτήτων των δεδομένων αυτών.

Λύση Παραδείγματος 8.1

Οι αριθμοί που μας δίνονται στην πραγματικότητα αφορούν τις συχνότητες των τριών (διατεταγμένων) κατηγοριών της ποιοτικής διατάξης μεταβλητής X: επίπεδο ικανοποίησης και το μέγεθος n του δείγματος. Με βάση αυτά μπορούμε να κατασκευάσουμε τον ακόλουθο πίνακα συχνοτήτων. Επισημαίνεται ότι, όταν πρόκειται για διατάξιμα δεδομένα, κατασκευάζεται ο πίνακας έτσι ώστε οι τιμές της μεταβλητής να τοποθετούνται σε αύξουσα τάξη μεγέθους. Επομένως, σε αυτό το παράδειγμα η σειρά είναι Λίγο, Μέτρια και Πολύ.

Επίπεδο ικανοποίησης	ν_i	f_i	$f_i\%$	N_i	F_i	$F_i\%$
Λίγο	20	0.2	20%	20	0.2	20%
Μέτρια	25	0.25	25%	45	0.45	45%
Πολύ	55	0.55	55%	100	1	100%
Σύνολο	100	1	100%			

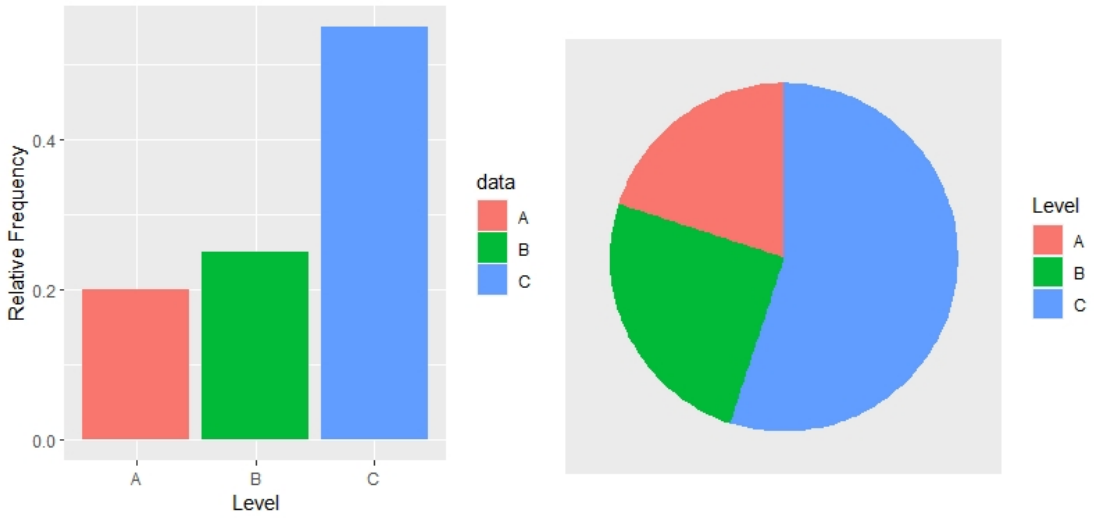
Για τον υπολογισμό των ποσοτήτων στον παραπάνω πίνακα χρησιμοποιήσαμε τους ορισμούς και τις ιδιότητες των συχνοτήτων, των σχετικών συχνοτήτων και των λοιπών συνδεδεμένων με αυτών εννοιών. Για παράδειγμα, η τιμή 0.45 στη στήλη F_i για την απάντηση Μέτρια έχει προκύψει από τη σχέση:

$$F_2 = \sum_{j=1}^{2} \frac{\nu_j}{n} = \frac{20+25}{100} = \frac{45}{100} = 0.45.$$

Επομένως, το 45% των φοιτητών που έλαβαν μέρος στην έρευνα δηλώνουν Μέτρια ή λιγότερο από μέτρια, δηλαδή Λίγο, ικανοποιημένοι από την παρεχομένη εκπαίδευση.

8.2.2 Γραφικές παραστάσεις ποιοτικών δεδομένων

Οι πίνακες συχνοτήτων που παρουσιάστηκαν παραπάνω συνοψίζουν όλη τη διαθέσιμη πληροφορία του δείγματος για την κατανομή της ποιοτικής μεταβλητής. Ωστόσο, ειδικά αν το πλήθος των κατηγοριών είναι μεγάλο, γίνονται εκτενείς και δύσκολοι στην κατανόησή τους. Ένας εναλλακτικός τρόπος κατανόησης και αναγνώρισης των χαρακτηριστικών της κατανομής μπορεί να επιτευχθεί με το ραβδόγραμμα (bar chart), μια από τις πιο δημοφιλείς γραφικές παραστάσεις για ποιοτικά δεδομένα. Το **ραβδόγραμμα** κατασκευάζεται τοποθετώντας στον οριζόντιο άξονα τις κατηγορίες στις οποίες τα μέλη του πληθυσμού κατατάσσονται και στον κατακόρυφο άξονα τις αντίστοιχες συχνότητες (ή εναλλακτικά τις αντίστοιχες σχετικές συχνότητες). Στο ραβδόγραμμα οι ορθογώνιες στήλες είναι μη εφαπτόμενες, με ίσες βάσεις και ύψος ίσο με την αντίστοιχη συχνότητα ή σχετική συχνότητα, ανάλογα. Συνηθέστερα οι μη εφαπτόμενες στήλες τοποθετούνται σε ίση απόσταση η μία από την άλλη (βλ. Ζωγράφος, 2002). Στο αριστερό μέρος του Σχήματος 8.1 παρουσιάζεται το ραβδόγραμμα για τα δεδομένα του Παραδείγματος 8.1. Από το γράφημα είναι φανερό ότι οι φοιτητές που είναι πολύ ικανοποιημένοι από την παρεχόμενη εκπαίδευση (C) αποτελούν την πλειονότητα των ατόμων του δείγματος.



Σχήμα 8.1: Το ραβδόγραμμα (αριστερά) και το κυκλικό διάγραμμα (δεξιά) γα τα δεδομένα του Παραδείγματος 8.1. Με A συμβολίζεται το λίγο επίπεδο ικανοποίησης από την παρεχόμενη εκπαίδευση, ενώ με B και C το μέτριο και πολύ, αντίστοιχα.

Μια δεύτερη γραφική παράσταση για ποιοτικά δεδομένα, εναλλακτική του ραβδογράμματος, είναι το **κυκλικό διάγραμμα** (pie chart), το οποίο κατασκευάζεται χωρίζοντας έναν κυκλικό δίσκο σε τομείς, όσες και οι κατηγορίες στις οποίες τα μέλη του πληθυσμού κατατάσσονται. Η επίκεντρη γωνία (ή, ισοδύναμα, το εμβαδόν) κάθε τομέα καθορίζεται ανάλογα με το ποσοστό των ατόμων που ανήκουν στην αντίστοιχη κατηγορία. Στο δεξί γράφημα του Σχήματος 8.1 παρουσιάζεται το κυκλικό διάγραμμα για τις τρεις κατηγορίες του επιπέδου ικανοποίησης των φοιτητών από την παρεχόμενη εκπαίδευση με βάση τα δεδομένα του Παραδείγματος 8.1.

Παρατήρηση 8.2

Οι γραφικές παραστάσεις του Σχήματος 8.1 έχουν δημιουργηθεί με την R εκτελώντας τις ακόλουθες εντολές.

```
library(ggplot2)
library(ggpubr)

data <- c(rep("A",20),rep("B",25),rep("C",55))
data<-as.data.frame(data)
bar<-ggplot(data=data, aes(x=data, fill=data)) +
  geom_bar(aes(y = (..count..)/sum(..count..)))+labs(x = "Level",y = "
      Relative Frequency")
pie <- ggplot(data, aes(x = factor(1), fill = factor(data))) +
  geom_bar(width = 1) + coord_polar(theta = "y")+
  labs(fill = "Level",x = NULL,y = NULL)+
  theme(axis.text = element_blank(),axis.ticks = element_blank(),
  panel.grid  = element_blank())
ggarrange(bar, pie, ncol = 2, nrow = 1)
```

8.2.3 Αριθμητικά μεγέθη διατάξιμων ποιοτικών δεδομένων

Για τις ποιοτικές μεταβλητές ο υπολογισμός αριθμητικών συγκεντρωτικών μεγεθών δύναται να έχει νόημα μόνο για τις διατάξιμες μεταβλητές. Για τον ορισμό των αριθμητικών αυτών μεγεθών πρέπει, αρχικά, να γίνει μια αυθαίρετη αντιστοίχιση των διατεταγμένων κατηγοριών με κάποιους αριθμούς. Συνηθέστερα, επιλέγονται οι φυσικοί αριθμοί, 1, 2, 3, ... ή οι αριθμοί 0, 1, 2, 3,... Στη συνέχεια με βάση αυτήν την αντιστοίχιση (που μπορεί να θεωρηθεί ανάλογη με αυτήν του ορισμού των τυχαίων μεταβλητών) έχει νόημα να υπολογιστούν επί των τιμών αυτών διάφορα αριθμητικά μεγέθη, όπως η δειγματική μέση τιμή (mean), η δειγματική διάμεσος (median) και η κορυφή (mode), ή αλλιώς η επικρατούσα τιμή. Σημειώνεται ότι τα αριθμητικά αυτά μεγέθη θα οριστούν και θα μελετηθούν αυστηρά στην επόμενη ενότητα, η οποία αφορά τη συνοπτική παρουσίαση ποσοτικών δεδομένων. Στο σημείο αυτό, θα παρουσιαστεί ο τρόπος υπολογισμού τους για διατάξιμα ποιοτικά δεδομένα με τη βοήθεια του παραδείγματος που ακολουθεί.

Παράδειγμα 8.2

Να υπολογίσετε τη μέση τιμή, τη διάμεσο και την επικρατούσα τιμή για τα δεδομένα του Παραδείγματος 8.1.

Λύση Παραδείγματος 8.2

Αρχικά, θα πρέπει να κάνουμε την αντιστοίχιση των διατεταγμένων κατηγοριών με κάποιους αριθμούς. Εδώ, θα υιοθετήσουμε την ακόλουθη αντιστοίχιση:

- Λίγο→1,
- Μέτρια→2,
- Πολύ→3.

Η μέση τιμή, η οποία συμβολίζεται με $\bar{x}$ των παρατηρήσεων $x_1, \ldots, x_{100}$ υπολογίζεται αθροίζοντας όλες τις τιμές και διαιρώντας με το πλήθος τους, δηλαδή

$$\bar{x} = \frac{1 + 1 + \ldots + 3}{100} = \frac{20 \cdot 1 + 25 \cdot 2 + 55 \cdot 3}{100}$$
$$= \frac{235}{100} = 2.35.$$

Η μέση τιμή 2.35 μπορεί να ερμηνευθεί ως το μέσο επίπεδο ικανοποίησης των ατόμων του δείγματος, το οποίο στην προκείμενη περίπτωση είναι μεταξύ του Μέτριου (αντιστοίχιση 2) και του Πολύ (αντιστοίχιση 3). Από το αποτέλεσμα είναι φανερό ότι η μέση τιμή ενός δείγματος όχι μόνο δεν είναι απαραίτητο να είναι ένας αριθμός που έχει παρατηρηθεί στο δείγμα, αλλά δύναται να μην μπορεί να ληφθεί από τη μεταβλητή. Για τον λόγο αυτό, πρέπει να είμαστε ιδιαίτερα προσεκτικοί στην ερμηνεία, αν είναι δυνατή, της τιμής αυτής. Επίσης, αξίζει να αναφερθεί ότι αν είχαμε υιοθετήσει διαφορετική κωδικοποίηση, αλλά είχαμε κρατήσει τη λογική διάταξη, θα προέκυπτε μια διαφορετική τιμή, η οποία όμως πάλι θα αποτύπωνε το μέσο επίπεδο ικανοποίησης των ατόμων στο δείγμα.

Η διάμεσος τιμή m είναι εκείνη η τιμή που ικανοποιεί ταυτόχρονα τις επόμενες δύο απαιτήσεις:

1. τουλάχιστον το 50% των παρατηρήσεων έχει τιμή μικρότερη ή ίση με αυτή και
2. τουλάχιστον το 50% των παρατηρήσεων έχει τιμή μεγαλύτερη ή ίση με αυτή.

Παρατηρώντας τον πίνακα συχνοτήτων των δεδομένων του Παραδείγματος 8.1, παρατηρούμε ότι η τιμή που ικανοποιεί τις παραπάνω απαιτήσεις είναι το 2. Αυτό σημαίνει ότι τουλάχιστον οι μισοί φοιτητές είναι μέτρια ή λιγότερο από μέτρια ικανοποιημένοι και ταυτόχρονα τουλάχιστον το 50% είναι μέτρια ή πολύ ικανοποιημένοι.

Τέλος, η επικρατούσα τιμή m_0 είναι η τιμή 3, δηλαδή το Πολύ, αφού αυτή η κατηγορία έχει τη μεγαλύτερη συχνότητα εμφάνισης στο δείγμα.

Παρατήρηση 8.3

Αξίζει κανείς να παρατηρήσει ότι η σχέση υπολογισμού της μέσης τιμής του δείγματος στο προηγούμενο παράδειγμα μπορεί να εκφραστεί ως

$$\bar{x} = \frac{\nu_1 \cdot 1 + \nu_2 \cdot 2 + \nu_3 \cdot 3}{n} = 1 \cdot \frac{\nu_1}{n} + 2 \cdot \frac{\nu_2}{n} + 3 \cdot \frac{\nu_3}{n}$$

όπου ν_i, $i = 1,2,3,$ οι συχνότητες κάθε κατηγορίας και 1, 2, 3 οι τιμές που αντιστοιχίστηκαν στις τρεις κατηγορίες της μεταβλητής. Αν υποθέσουμε ότι παρατηρούμε όλο και μεγαλύτερο μέρος του πληθυσμού, δηλαδή αν αφήσουμε το n να μεγαλώνει και να απειρίζεται, τότε τα κλάσματα $\frac{\nu_i}{n}$ τείνουν με βάση τον ορισμό της πιθανότητας ως όριο της σχετικής συχνότητας, στην $P(X = i)$. Αυτό έχει ως συνέπεια να έχουμε ότι

$$\lim_{n\to\infty}\left(1 \cdot \frac{\nu_1}{n} + 2 \cdot \frac{\nu_2}{n} + 3 \cdot \frac{\nu_3}{n}\right) = \sum_{x=1}^{3} xP(X = x)$$

που δεν είναι τίποτα άλλο από τη μέση τιμή της τυχαίας μεταβλητής, έστω X, που παριστάνει το επίπεδο ικανοποίησης των ατόμων από την παρεχόμενη εκπαίδευση, με δυνατές τιμές 1, 2 και 3, αντίστοιχα. Αυτό σημαίνει ότι $\lim_{n\to\infty} \bar{x} = E(X)$ με πιθανότητα 1. Στην πραγματικότητα, το παραπάνω αποτέλεσμα αποτυπώνει τη σύνδεση μεταξύ των πιθανοτήτων και των τυχαίων μεταβλητών, που παρουσιάστηκαν στα προηγούμενα κεφάλαια, και της στατιστικής συμπερασματολογίας, που ακολουθεί στα επόμενα κεφάλαια και η οποία με βάση το δείγμα, προσπαθεί να γενικεύσει τα συμπεράσματα στον πληθυσμό. Αυτό επιτυγχάνεται ορίζοντας στο δείγμα ποσότητες, τις λεγόμενες **δειγματικές συναρτήσεις**, οι οποίες εκτιμούν και παρέχουν πληροφορίες για τα αντίστοιχα χαρακτηριστικά του πληθυσμού. Τέτοιο παράδειγμα είναι ο δειγματικός μέσος, που αναφέρθηκε νωρίτερα, ο οποίος παρέχει σημαντικές πληροφορίες για τον πληθυσμιακό μέσο.

Όπως είδαμε, όταν το μέγεθος του δείγματος τείνει στο άπειρο, ο δειγματικός μέσος συγκλίνει στην πραγματική μέση τιμή του πληθυσμού. Όταν όμως το μέγεθος του δείγματος είναι πεπερασμένο, όπως άλλωστε συμβαίνει πάντα, κάποια απόκλιση από την πραγματική τιμή είναι αναμενόμενη λόγω της τυχαιότητας των παρατηρήσεων του δείγματος. Αυτό συμβαίνει γιατί οτιδήποτε υπολογίζεται με τη βοήθεια των παρατηρήσεων ενός τυχαίου δείγματος (δειγματική συνάρτηση) δεν μπορεί παρά να είναι και αυτό τυχαίο και να ακολουθεί μια κατανομή. Η αξιοπιστία της τιμής της δειγματικής συνάρτησης που παρατηρούμε σε ένα δείγμα και η αβεβαιότητά της ως προς την πληροφορία που μας δίνει για το αντίστοιχο χαρακτηριστικό του πληθυσμού αποτυπώνονται στην κατανομή της. Η μελέτη των κατανομών αυτών παίζει κεντρικό ρόλο στη διαδικασία γενίκευσης των συμπερασμάτων από το δείγμα στον πληθυσμό και θα παρουσιαστεί αναλυτικά στο επόμενο κεφάλαιο.

Παρατήρηση 8.4

Η αυθαίρετη αντιστοίχιση των κατηγοριών με κάποιους αριθμούς υιοθετείται συχνά και για τις ονομαστικές μεταβλητές. Στην περίπτωση αυτή, η αντιστοίχιση των κατηγοριών με τους αριθμούς είναι πλήρως αυθαίρετη και δεν έχει νόημα ο υπολογισμός αριθμητικών μεγεθών. Μοναδική εξαίρεση αποτελεί η επικρατούσα τιμή, η οποία δηλώνει την κατηγορία με τη μεγαλύτερη συχνότητα.

8.3 Συνοπτική παρουσίαση ποσοτικών δεδομένων

Αντικείμενο μελέτης αυτής της ενότητας είναι η συνοπτική παρουσίαση **ποσοτικών δεδομένων**. Με τον όρο ποσοτικά δεδομένα αναφερόμαστε στα δεδομένα που λαμβάνονται από την καταγραφή των παρατηρήσεων ποσοτικών μεταβλητών, δηλαδή μεταβλητών που περιγράφουν χαρακτηριστικά γνωρίσματα που μπορούν να μετρηθούν και λαμβάνουν αριθμητικές τιμές. Παραδείγματα τέτοιων μεταβλητών είναι:

- ο αριθμός των φοιτητών που συμμετέχουν σε μια εξέταση,
- ο αριθμός των οχημάτων που διέρχονται από μία διασταύρωση κατά τη διάρκεια που είναι αναμμένος ο πράσινος σηματοδότης,
- ο ημερήσιος αριθμός κρουσμάτων μιας ασθένειας σε μια γεωγραφική περιοχή,
- το βάρος ενός ατόμου,
- η απόσταση που διανύει ένα όχημα μέχρι να χρειαστεί ανεφοδιασμό,
- η αντοχή ενός συρματόσχοινου.

Τα ποσοτικά δεδομένα μπορούν να χωριστούν με τη σειρά τους σε δύο μεγάλες υποκατηγορίες. Τα δεδομένα που προέρχονται

- από διακριτές τυχαίες μεταβλητές, δηλαδή από μεταβλητές που λαμβάνουν τιμές σε ένα πεπερασμένο ή το πολύ απείρως αριθμήσιμο σύνολο και
- από συνεχείς τυχαίες μεταβλητές, δηλαδή από μεταβλητές που λαμβάνουν τιμές σε ένα διάστημα ή σε μία ένωση διαστημάτων.

Από τα παραδείγματα που αναφέρθηκαν παραπάνω, τα τρία πρώτα προέρχονται από διακριτές τυχαίες μεταβλητές, ενώ τα τρία τελευταία από συνεχείς.

Τα εργαλεία συνοπτικής παρουσίασης των δεδομένων μίας ποσοτικής μεταβλητής είναι παρόμοια με αυτά των ποιοτικών δεδομένων και παρουσιάζονται αναλυτικά στη συνέχεια.

8.3.1 Πίνακας συχνοτήτων και ομαδοποιημένος πίνακας συχνοτήτων ποσοτικών δεδομένων

Όταν τα διαθέσιμα δεδομένα είναι ποσοτικά, ο πίνακας συχνοτήτων κατασκευάζεται πρακτικά με τον ίδιο τρόπο με αυτόν για τα ποιοτικά δεδομένα. Ειδικότερα, έστω $x_1,...,x_n$ οι διαθέσιμες παρατηρήσεις. Αρχικά, οι μετρήσεις αυτές διατάσσονται κατά αύξουσα τάξη μεγέθους και πολλές φορές προκύπτει ότι στην πραγματικότητα υπάρχουν ℓ το πλήθος διακεκριμένες τιμές, έστω οι $x_{(1)},\ldots,x_{(\ell)}$. Στη συνέχεια, κατασκευάζεται ο (διατεταγμένος) πίνακας συχνοτήτων:

Διακεκριμένη Μέτρηση $x_{(i)}$	$\nu_{(i)}$	$f_{(i)}$	$f_{(i)}\%$	$\mathrm{N}_{(i)}$	$F_{(i)}$	$F_{(i)}\%$
$x_{(1)}$	$\nu_{(1)}$	$f_{(1)}$	$f_{(1)}\%$	$\mathrm{N}_{(1)}$	$F_{(1)}$	$F_{(1)}\%$
$x_{(2)}$	$\nu_{(2)}$	$f_{(2)}$	$f_{(2)}\%$	$\mathrm{N}_{(2)}$	$F_{(2)}$	$F_{(2)}\%$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$x_{(\ell)}$	$\nu_{(\ell)}$	$f_{(\ell)}$	$f_{(\ell)}\%$	$\mathrm{N}_{(\ell)}=n$	$F_{(\ell)}=1$	$F_{(\ell)}\%=100\%$
Σύνολο	n	1	100%			

Είναι προφανές ότι αν το πλήθος των διακεκριμένων μετρήσεων ℓ είναι μεγάλο, γεγονός πολύ σύνηθες όταν έχουμε να κάνουμε με ποσοτικά δεδομένα, ο παραπάνω πίνακας δεν δίνει συνοπτική πληροφορία[1]. Συνέπεια αυτού είναι, ότι χωρίς ομαδοποίηση των τιμών ή διακριτοποίηση των συνεχών τιμών, οι πίνακες συχνοτήτων τείνουν να είναι εκτενείς και συχνά χωρίς πληροφορία, αφού οι συχνότητες των τιμών είναι όλες μικρές. Για τον λόγο αυτό χρειάζεται η ομαδοποίηση των δεδομένων σε ομάδες ή διαφορετικά σε κλάσεις και έπειτα η κατασκευή του λεγόμενου ομαδοποιημένου πίνακα συχνοτήτων. Η διαδικασία ομαδοποίησης των δεδομένων και ο ορισμός των ομάδων/κλάσεων είναι μια διαδικασία, η οποία εξαρτάται σε μεγάλο βαθμό από το εύρος των τιμών του δείγματος, δηλαδή από τη διαφορά της μέγιστης από την ελάχιστη τιμή, και το μέγεθος του δείγματος.

Ένα πρώτο εύλογο ερώτημα που προκύπτει αφορά το πλήθος των ομάδων. Το πλήθος k των ομάδων καθορίζεται έτσι ώστε η μέση συχνότητα των ομάδων να μην είναι ούτε πολύ μεγάλη ούτε πολύ μικρή. Μια εμπειρική διαδικασία για τον καθορισμό του πλήθους των ομάδων βασίζεται στον τύπο του Sturges, ο οποίος προτείνει ως βέλτιστο αριθμό ομάδων k τον πλησιέστερο ακέραιο στον αριθμό

$$k = 1 + 3.322 \log_{10} n,$$

όπου n το μέγεθος του δείγματος και $\log_{10}$ ο λογάριθμος με βάση το 10.

Το επόμενο εύλογο ερώτημα που προκύπτει είναι πώς κατασκευάζονται αυτές οι k το πλήθος ομάδες. Οι ομάδες αυτές κατασκευάζονται είτε υιοθετώντας κάποιο λογικό, στο πλαίσιο του προβλήματος και της φύσης του προβλήματος, μοτίβο, όπως το $[0, 1)$, $[1, 2)$, ... , $[9, 10]$ ή ακολουθώντας μια πιο αυτοματοποιημένη διαδικασία. Στη δεύτερη περίπτωση, οι ομάδες κατασκευάζονται ως ακολούθως: η πρώτη ομάδα έχει κάτω όριο μια τιμή λίγο μικρότερη από την ελάχιστη τιμή των δεδομένων (αν πρόκειται για ακέραιες τιμές αφαιρούμε, συνήθως, την τιμή 0.5) και η τελευταία έχει άνω όριο μια τιμή λίγο μεγαλύτερη από τη μέγιστη (αν πρόκειται για ακέραιες τιμές προσθέτουμε, συνήθως, την τιμή 0.5). Τα υπόλοιπα όρια των ομάδων καθορίζονται διαιρώντας το διάστημα που ορίζεται από τις προαναφερθείσες τιμές σε k ισομήκη διαστήματα. Αυτό επιτυγχάνεται εύκολα επιλέγοντας το μήκος κάθε ομάδας, έστω d, να δίνεται από τη σχέση:

$$d = \frac{R + 1}{k} = \frac{\max x_i - \min x_i + 1}{k}.$$

Σημειώνεται ότι στην περίπτωση που έχουμε τιμές με ένα δεκαδικό ψηφίο, προσθέτουμε και αφαιρούμε (συνήθως) σε κάθε τιμή την τιμή 0.05. Σε αυτές τις περιπτώσεις, το μήκος κάθε ομάδας δίνεται από έναν παρόμοιο τύπο με τον παραπάνω, στον οποίο όμως το εύρος, R, των τιμών δεν προσαυξάνεται κατά μία μονάδα αλλά κατά 0.1.

Οι προαναφερθείσες πρόσθεση και αφαίρεση των τιμών 0.5 ή 0.05 από τη μέγιστη και ελάχιστη τιμή, αντίστοιχα, του δείγματος, εξασφαλίζουν ότι η πρώτη και η τελευταία ομάδα συμπεριλαμβάνουν, ως εσωτερικά σημεία, την ελάχιστη και τη μέγιστη τιμή, αντίστοιχα του δείγματος. Επιπρόσθετα, τα αριστερά όρια των ομάδων συνηθίζεται να είναι κλειστά και τα δεξιά ανοικτά έτσι ώστε να μην υπάρχει αλληλοκάλυψη των ομάδων.

Μετά την ομαδοποίηση ή διακριτοποίηση των ποσοτικών δεδομένων, μπορεί να χρησιμοποιηθεί ο αποκαλούμενος ομαδοποιημένος πίνακας συχνοτήτων στον οποίο καταγράφεται αρχικά, το πλήθος, δηλαδή η συχνότητα ν_i των τιμών που ανήκουν σε κάθε κλάση i. Στη συνέχεια, ο πίνακας συχνοτήτων συμπληρώνεται με τις στήλες των σχετικών συχνοτήτων f_i, των αθροιστικών συχνοτήτων N_i και των

[1] Στην περίπτωση που τα δεδομένα προέρχονται από μια συνεχή μεταβλητή, τότε η συχνότητα κάθε παρατήρησης ισούται με ένα. Οποιαδήποτε διαφορετική τιμή οφείλεται σε λάθος στρογγυλοποίησης και αδυναμίας καταγραφής της ακριβούς τιμής κάθε παρατήρησης. Σε αυτές τις περιπτώσεις είναι φανερό ότι το πλήθος των διακεκριμένων μετρήσεων ℓ είναι ίσο (ή πρακτικά ίσο) με το μέγεθος του δείγματος και, επομένως, ο (διατεταγμένος) πίνακας συχνοτήτων δεν παρέχει κάποια συνοπτική πληροφορία. Για τον λόγο αυτό, σε τέτοιες περιπτώσεις τα δεδομένα ομαδοποιούνται, όπως αναφέρεται στη συνέχεια.

αθροιστικών σχετικών συχνοτήτων. Συχνά, όπως και στην περίπτωση των ποιοτικών δεδομένων, η σχετική συχνότητα και η αθροιστική σχετική συχνότητα εκφράζονται και επί τοις εκατό.

Παράδειγμα 8.3

Σε μία έρευνα που αφορά το ύψος των φοιτητριών του Τμήματος Μαθηματικών του Πανεπιστημίου Ιωαννίνων καταγράφεται το ύψος σε εκατοστά 40 τυχαία επιλεγμένων φοιτητριών. Οι μετρήσεις είναι:

172 182 179 166 168 168 176 170 172 159
161 170 167 177 180 170 182 178 165 178
159 175 172 173 167 187 171 181 177 171
178 168 187 176 188 173 180 165 171 172.

Να παρουσιαστούν τα παραπάνω δεδομένα συνοπτικά με τη βοήθεια ενός ομαδοποιημένου πίνακα συχνοτήτων.

Λύση Παραδείγματος 8.3

Πριν ξεκινήσουμε την ανάλυση των δεδομένων πρέπει να σημειωθεί ότι τα δεδομένα μας προέρχονται από μια συνεχή μεταβλητή, το ύψος. Ως εκ τούτου, οι οποιεσδήποτε ισοβαθμίες στα δεδομένα προκύπτουν από την αδυναμία καταγραφής της ακριβούς τιμής του ύψους των φοιτητριών. Ως συνεχή που είναι τα δεδομένα, για να κατασκευάσουμε τον πίνακα συχνοτήτων, θα πρέπει αρχικά, να τα ομαδοποιήσουμε. Το πλήθος των ομάδων k με βάση τον τύπο του Sturges είναι:

$$\begin{aligned} k &= 1 + 3.322 \log_{10} n \\ &= 1 + 3.322 \log_{10} 40 \\ &= 1 + 3.322 \cdot 1.6021 \\ &= 6.3220. \end{aligned}$$

Επομένως, θα πρέπει να χρησιμοποιηθούν 6 το πλήθος ομάδες με μήκος κάθε ομάδας

$$\begin{aligned} d &= \frac{R+1}{k} \\ &= \frac{\max x_i - \min x_i + 1}{k} \\ &= \frac{188 - 159 + 1}{6} \\ &= 5. \end{aligned}$$

Επιπρόσθετα, το κάτω όριο της πρώτης ομάδας είναι 159-0.5=158.5 και το άνω όριο της έκτης ομάδας είναι 188+0.5=188.5. Εύκολα μετά προκύπτει ο ακόλουθος ομαδοποιημένος πίνακας συχνοτήτων.

Αύξων αριθμός ομάδας	Όρια ομάδας	$\nu_{(i)}$	$f_{(i)}$	$f_{(i)}\%$	$N_{(i)}$	$F_{(i)}$	$F_{(i)}\%$
1	[158.5-163.5)	3	0.075	7.5%	3	0.075	7.5%
2	[163.5-168.5)	8	0.2	20%	11	0.275	27.5%
3	[168.5-173.5)	12	0.3	30%	23	0.575	57.5%
4	[173.5-178.5)	8	0.2	20%	31	0.775	77.5%
5	[178.5-183.5)	6	0.15	15%	37	0.925	92.5%
6	[183.5-188.5)	3	0.075	7.5%	40	1	100%
	Σύνολο	40	1	100%			

8.3.2 Αριθμητικά μεγέθη ποσοτικών δεδομένων

Τα αριθμητικά μεγέθη αποτελούν ένα πολύ χρήσιμο εργαλείο για τη συνοπτική παρουσίαση των δεδομένων ποσοτικών μεταβλητών. Τα **αριθμητικά μεγέθη** διακρίνονται σε τρεις κατηγορίες: τα μέτρα θέσης, τα μέτρα μεταβλητότητας και τα μέτρα σχήματος ή μορφής (μέτρα λοξότητας και κύρτωσης). Οι οικογένειες αυτών των μέτρων περιλαμβάνουν η καθεμία διάφορα μέτρα, τα οποία συνοψίζουν διαφορετικά χαρακτηριστικά των δεδομένων και παρέχουν μια πλήρη εικόνα για τα δεδομένα ποσοτικών μεταβλητών.

8.3.2.1 Μέτρα θέσης

Ως **μέτρα θέσης** εννοούνται οι τιμές κάποιων συναρτήσεων των τιμών του δείγματος, οι οποίες προσδιορίζουν σημεία, δηλαδή συγκεκριμένες αριθμητικές τιμές, που αποτυπώνουν χαρακτηριστικές ιδιότητες των παρατηρήσεών μας. Τέτοιες χαρακτηριστικές ιδιότητες περιλαμβάνουν το σημείο γύρω από το οποίο συγκεντρώνονται οι περισσότερες παρατηρήσεις ή το κέντρο βάρους των παρατηρήσεων.

Τα βασικότερα μέτρα θέσης ενός δείγματος $x_1, \dots, x_n$ είναι

- η δειγματική μέση τιμή,
- η δειγματική διάμεσος,
- η κορυφή, ή αλλιώς, επικρατούσα τιμή,
- τα δειγματικά ποσοστιαία σημεία (quantile points).

Ορισμός 8.3

Η **δειγματική μέση τιμή**, $\bar{x}$, ενός τυχαίου δείγματος $x_1, \dots, x_n$ δίνεται από τη σχέση:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}. \tag{8.1}$$

Η μέση τιμή αποτελεί ίσως το σημαντικότερο μέτρο θέσης και μπορεί να ερμηνευθεί ως το κέντρο βάρους των παρατηρήσεών μας. Πράγματι, η δειγματική μέση τιμή ικανοποιεί τη σχέση:

$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0$$

και είναι εκείνη η τιμή που η συνολική απόστασή της από τις τιμές που είναι μικρότερες από αυτήν ισούται με τη συνολική απόστασή της από τις τιμές που είναι μεγαλύτερές της. Όπως είναι φανερό, η δειγματική μέση τιμή ικανοποιεί την ανίσωση

$$x_{(1)} \leq \bar{x} \leq x_{(n)}$$

όπου με $x_{(i)}$ συμβολίζουμε την $i-$οστή διατεταγμένη παρατήρηση. Αυτό σημαίνει ότι με $x_{(1)}$ συμβολίζουμε την ελάχιστη και με $x_{(n)}$ τη μέγιστη τιμή του δείγματος. Η ισότητα στην παραπάνω σχέση ισχύει αν και μόνο αν όλες οι τιμές του δείγματος είναι ίσες μεταξύ τους, δηλαδή αν δεν υπάρχει καθόλου μεταβλητότητα στα δεδομένα. Επιπρόσθετα, η δειγματική μέση τιμή είναι η τιμή εκείνη που ελαχιστοποιεί ως προς a τη συνάρτηση

$$g(a) = \sum_{i=1}^{n}(x_i - a)^2.$$

Η σημασία της ιδιότητας αυτής θα αναδειχθεί όταν θα οριστεί η δειγματική διασπορά (βλ. επόμενη υποενότητα).

Αξίζει να σημειωθεί ότι στην περίπτωση που δεν έχουμε πρόσβαση στα πρωτογενή δεδομένα, αλλά τα δεδομένα δίνονται μέσω ενός ομαδοποιημένου πίνακα συχνοτήτων με k το πλήθος ομάδες, τότε η δειγματική μέση τιμή δίνεται από τη σχέση:

$$\bar{x} = \frac{\sum_{l=1}^{k} k_\ell \cdot \nu_\ell}{n}, \tag{8.2}$$

όπου k_ℓ είναι η μέση τιμή ή, αλλιώς, κεντρική τιμή της ℓ-οστής ομάδας/κλάσης, που προκύπτει ως το ημιάθροισμα των ορίων της, ενώ ν_ℓ είναι η συχνότητα της ℓ-οστής ομάδας/κλάσης, $l = 1, \dots, k$.

Η **διάμεσος**, m, εκφράζει εκείνη την τιμή του δείγματος που ικανοποιεί ταυτόχρονα τις ακόλουθες δύο ιδιότητες:

- τουλάχιστον το 50% των παρατηρήσεων έχει τιμή μικρότερη ή ίση με αυτήν και
- τουλάχιστον το 50% των παρατηρήσεων έχει τιμή μεγαλύτερη ή ίση με αυτήν.

Στην περίπτωση που στη διάθεσή μας έχουμε τις ακριβείς τιμές των παρατηρήσεων, η διάμεσος υπολογίζεται από την σχέση:

$$m = \begin{cases} x_{\left(\frac{n+1}{2}\right)}, & \text{για περιττό } n, \\ \frac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2}, & \text{για άρτιο } n. \end{cases}$$

Η παραπάνω σχέση πρακτικά σημαίνει ότι, αν το πλήθος των δεδομένων είναι περιττός αριθμός, τότε η διάμεσος είναι η τιμή της $(n+1)/2$ διατεταγμένης παρατήρησης. Από την άλλη πλευρά, αν το πλήθος των δεδομένων είναι άρτιος αριθμός, τότε η διάμεσος ορίζεται ως το ημιάθροισμα των τιμών της $n/2$ και $n/2+1$ διατεταγμένης παρατήρησης, παρόλο που οποιαδήποτε τιμή στο διάστημα $[x_{\left(\frac{n}{2}\right)}, x_{\left(\frac{n}{2}+1\right)}]$ ικανοποιεί τις δύο προαναφερθείσες ιδιότητες της διαμέσου.

Στην περίπτωση που στη διάθεσή μας έχουμε μόνο τα ομαδοποιημένα δεδομένα ενός δείγματος, η διάμεσος υπολογίζεται από τη σχέση:

$$m = L_i + d \cdot \frac{\frac{n}{2} - N_{i-1}}{\nu_i}, \tag{8.3}$$

όπου

- i είναι η ομάδα/κλάση στην οποία βρίσκεται η διάμεσος, δηλαδή η μικρότερη τιμή i για την οποία ισχύει ότι $N_{i-1} \leq n/2 \leq N_i$,
- L_i το κάτω όριο της ομάδας, όπου εντοπίστηκε ότι ανήκει η διάμεσος,
- N_{i-1} η αθροιστική συχνότητα της $(i-1)$-οστής ομάδας,
- ν_i η συχνότητα της i-οστής ομάδας και
- d το μήκος της ομάδας.

Στην ουσία, η παραπάνω σχέση εκφράζει τα βήματα που πρέπει να κάνει κάποιος μέσα στην ομάδα που περιέχει τη διάμεσο έτσι ώστε να εντοπίσει την κεντρική τιμή των δεδομένων. Τα βήματα αυτά εκφράζονται από τη διαφορά $\frac{n}{2} - N_{i-1}$. Αν η τιμή αυτή είναι μικρή σε σχέση με τη συχνότητα της ομάδας αυτής, τότε η διάμεσος εκτιμάται ότι είναι κοντά στο κάτω όριο της ομάδας. Αντίθετα, αν η διαφορά $\frac{n}{2} - N_{i-1}$ είναι μεγάλη, τότε πρέπει να πάρουμε το μεγαλύτερο μέρος του πλάτους d του διαστήματος που ορίζει την i-οστή ομάδα για να βρούμε τη διάμεσο.

Ορισμός 8.4

Κορυφή ή, αλλιώς, **επικρατούσα τιμή**, m_0, ορίζεται η τιμή εκείνη με τη μεγαλύτερη συχνότητα.

Αν όλες οι μετρήσεις εμφανίζονται με την ίδια συχνότητα δεν υπάρχει κορυφή. Επειδή στα ποσοτικά δεδομένα οι διαφορετικές τιμές που παρατηρούμε στο δείγμα είναι πολυάριθμες, η συχνότητα κάθε τιμής δεν μας δίνει ξεκάθαρα πληροφορία για το ποια τιμή ή ποιες τιμές είναι οι πιο συχνά παρατηρούμενες στο δείγμα. Στην περίπτωση μάλιστα που τα ποσοτικά δεδομένα προέρχονται από μια συνεχή τυχαία μεταβλητή, τότε κάθε παρατήρηση, αν αγνοήσουμε λάθη στρογγυλοποίησης, είναι διαφορετική από τις άλλες και έχει συχνότητα ένα. Παρ΄όλα αυτά, ακόμα και σε αυτές τις περιπτώσεις μπορούμε να ορίσουμε την κορυφή. Η κορυφή σε μια τέτοια περίπτωση αντιστοιχεί στην τιμή εκείνη γύρω από την οποία συγκεντρώνονται οι περισσότερες παρατηρήσεις. Πιο συγκεκριμένα, το διάστημα ή τα διαστήματα με τη μεγαλύτερη συχνότητα περιέχουν την κορυφή. Η παρατήρηση αυτή μας οδηγεί στο να συνδέσουμε τον υπολογισμό της κορυφής με την ομάδα/κλάση με τη μεγαλύτερη συχνότητα, δηλαδή με το ψηλότερο ορθογώνιο σε ένα ιστόγραμμα. Επομένως, η περίπτωση των ποσοτικών δεδομένων από συνεχείς τυχαίες μεταβλητές συνδέεται άμεσα με την ομαδοποίηση των δεδομένων και δίνεται από τη σχέση:

$$m_0 = L_i + \frac{\Delta_1}{\Delta_1 + \Delta_2} \cdot d$$

όπου L_i το κάτω όριο της ομάδας που έχει τη μεγαλύτερη συχνότητα (επικρατούσα ομάδα) και $\Delta_1 = \nu_i - \nu_{i-1}$ και $\Delta_2 = \nu_i - \nu_{i+1}$ οι διαφορές της συχνότητας της επικρατούσας ομάδας από τις γειτονικές της.

Σημειώνεται ότι η έννοια της κορυφής μπορεί να επεκταθεί και στις περιπτώσεις που υπάρχουν περισσότερες από μία ομάδες που παρουσιάζουν μεγαλύτερη συχνότητα από τις γειτονικές τους. Ένα τέτοιο χαρακτηριστικό παράδειγμα αποτελούν τα δείγματα των οποίων τα ιστογράμματα εμφανίζονται στην κάτω γραμμή του Σχήματος 8.2, τα οποία ιστογράμματα παρουσιάζουν δύο κορυφές. Τα δείγματα που αποτυπώνονται στα πάνω γραφήματα του Σχήματος 8.2 εμφανίζουν περισσότερες από 2 κορυφές, κυρίως λόγω του σχετικά μικρού αριθμού παρατηρήσεων. Από τα παραπάνω γίνεται φανερό ότι η κορυφή μπορεί να μην είναι μοναδική.

Ορισμός 8.5

Τα **δειγματικά ποσοστιαία σημεία**, q_p, $0 < p < 1$, αντιπροσωπεύουν τιμές που χωρίζουν το δείγμα σε δύο μέρη ανάλογα με το ποσοστό $p \times 100\%$ των παρατηρήσεων που είναι μικρότερα ή ίσα από αυτές τις τιμές.

Από τον παραπάνω ορισμό άμεσα προκύπτει ότι η δειγματική διάμεσος είναι το 0.5 ποσοστιαίο σημείο. Για τον προσδιορισμό ενός p-ποσοστιαίου σημείου, q_p, ενός δείγματος $x_1, \ldots, x_n$ έχουν προταθεί διάφοροι τρόποι (Hyndman and Fan, 1996). Εδώ θα περιοριστούμε στην παρουσίαση ενός από αυτούς. Το p-ποσοστιαίο σημείο μπορεί να υπολογιστεί με βάση τον ακόλουθο γραμμικό συνδυασμό:

$$q_p = x_{(\lfloor (n-1)p+1 \rfloor)} + \gamma (x_{(\lfloor (n-1)p+1 \rfloor + 1)} - x_{(\lfloor (n-1)p+1 \rfloor)}), \tag{8.4}$$

όπου

$$\gamma = (n-1)p + 1 - \lfloor (n-1)p + 1 \rfloor, \tag{8.5}$$

με $\lfloor a \rfloor$ να συμβολίζει το ακέραιο μέρος του αριθμού a.

Ιδιαίτερη αναφορά πρέπει να γίνει σε κάποιες ιδιαίτερες τιμές του p που ορίζουν κάποια σημαντικά ποσοστιαία σημεία. Αυτές οι τιμές είναι

1. οι 0.25 και 0.75, οι οποίες ορίζουν το 1ο και το 3ο τεταρτημόριο (Q_1 και Q_3, αντίστοιχα) και
2. οι 0.1, 0.2, …, 0.9, οι οποίες ορίζουν το 1ο, 2ο,… , και το 9ο δεκατημόριο.

Παρατήρηση 8.5

Στις περιπτώσεις που στη διάθεσή μας έχουμε μόνο τα ομαδοποιημένα δεδομένα ενός δείγματος, τα δειγματικά ποσοστιαία σημεία, q_p, $0 < p < 1$, υπολογίζονται με μια σχέση παρόμοια με αυτήν της διαμέσου, που παρουσιάστηκε νωρίτερα. Πιο συγκεκριμένα, η τιμή τους προσδιορίζεται με τη βοήθεια της σχέσης:

$$q_p = L_i + d \cdot \frac{p \cdot n - N_{i-1}}{\nu_i}, \tag{8.6}$$

όπου

- i είναι η ομάδα/κλάση στην οποία βρίσκεται το ζητούμενο ποσοστιαίο σημείο, δηλαδή η μικρότερη τιμή i, για την οποία ισχύει ότι $N_{i-1} \leq p \cdot n \leq N_i$,
- L_i το κάτω όριο της ομάδας, όπου εντοπίστηκε ότι ανήκει το ζητούμενο ποσοστιαίο σημείο,
- N_{i-1} η αθροιστική συχνότητα της $(i-1)$-οστής ομάδας,
- ν_i η συχνότητα της i-οστής ομάδας και
- d το μήκος της ομάδας.

Παρατηρήστε ότι για $q = 0.5$ προκύπτει η σχέση (8.3), δηλαδή η σχέση προσδιορισμού της διαμέσου για ομαδοποιημένα δεδομένα.

Παράδειγμα 8.4

Να βρεθούν η μέση τιμή, η διάμεσος, καθώς και το 25ο ποσοστιαίο σημείο, δηλαδή το 1ο τεταρτημόριο, χρησιμοποιώντας τα πρωτογενή δεδομένα του Παραδείγματος 8.3. Στη συνέχεια, χρησιμοποιώντας τον ομαδοποιημένο πίνακα του παραδείγματος, υπολογίστε τη μέση τιμή, τη διάμεσο και το 25ο ποσοστιαίο σημείο.

Λύση Παραδείγματος 8.4

Η δειγματική μέση τιμή, χρησιμοποιώντας τα πρωτογενή δεδομένα, είναι (βλ. τη σχέση (8.1)):

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{172 + 182 + \ldots + 172}{40} = \frac{6932}{40} = 173.3.$$

Καθώς το μέγεθος του δείγματος είναι $n = 40$, δηλαδή άρτιος αριθμός, η διάμεσος είναι το ημιάθροισμα της $20^{\text{ης}}$ $(n/2)$ και της $21^{\text{ης}}$ $(n/2+1)$ διατεταγμένης παρατήρησης. Προκύπτει ότι είναι η τιμή 172 (ελέγξτε το!).

Παρατηρήστε ότι αν είχαμε χρησιμοποιήσει τη σχέση (8.4) για $q = 0.5$, θα είχαμε ότι η διάμεσος ισούται πάλι με 172, καθώς

$$q_{0,5} = x_{(\lfloor (40-1)0.5+1 \rfloor)} + \gamma(x_{(\lfloor (40-1)0.5+1 \rfloor+1)} - x_{(\lfloor (40-1)0.5+1 \rfloor)}) = x_{(20)} + 0.5(x_{(21)} - x_{(20)}) = 172$$

όπου από τη σχέση (8.5) έχουμε για $n = 40$ και $p = 0.5$ ότι:

$$\gamma = (40-1)0.5 + 1 - \lfloor (40-1)0.5 + 1 \rfloor = 20.5 - 20 = 0.5.$$

Με παρόμοιους υπολογισμούς μπορούμε να προσδιορίσουμε και το 1ο τεταρτημόριο ως ακολούθως:

$$q_{0.25} = x_{(\lfloor (40-1)0.25+1 \rfloor)} + \gamma(x_{(\lfloor (40-1)0.25+1 \rfloor+1)} - x_{(\lfloor (40-1)0.25+1 \rfloor)}) = x_{(10)} + 0.75(x_{(11)} - x_{(10)}) = 168$$

όπου τώρα

$$\gamma = (40-1)0.25 + 1 - \lfloor (40-1)0.25 + 1 \rfloor = 0.75.$$

Τα παραπάνω μέτρα μπορούν να υπολογιστούν και για τον ομαδοποιημένο πίνακα συχνοτήτων που παρουσιάστηκε στο Παράδειγμα 8.3, δηλαδή θεωρώντας ότι στη διάθεσή μας έχουμε τον ακόλουθο πίνακα και όχι τα πρωτογενή δεδομένα.

Αύξων αριθμός ομάδας	Όρια ομάδας	$\nu_{(i)}$	$f_{(i)}$	$f_{(i)}\%$	$N_{(i)}$	$F_{(i)}$	$F_{(i)}\%$
1	[158.5-163.5)	3	0.075	7.5%	3	0.075	7.5%
2	[163.5-168.5)	8	0.2	20%	11	0.275	27.5%
3	[168.5-173.5)	12	0.3	30%	23	0.575	57.5%
4	[173.5-178.5)	8	0.2	20%	31	0.775	77.5%
5	[178.5-183.5)	6	0.15	15%	37	0.925	92.5%
6	[183.5-188.5)	3	0.075	7.5%	40	1	100%
	Σύνολο	40	1	100%			

Σε αυτήν την περίπτωση, για να προσδιορίσουμε τη δειγματική μέση τιμή, θα χρησιμοποιήσουμε τη σχέση (8.2), από όπου, παρατηρώντας ότι οι μέσες τιμές των ομάδων είναι ίσες με 161, 166, 171, 176, 181 και 186, έχουμε ότι:

$$\bar{x} = \frac{161 \cdot 3 + 166 \cdot 8 + 171 \cdot 12 + 176 \cdot 8 + 181 \cdot 6 + 186 \cdot 3}{40} = \frac{6915}{40} = 172.875.$$

Η διάμεσος και το 1ο τεταρτημόριο υπολογίζονται από τη σχέση (8.6) για $q = 0.5$ και $q = 0.25$, αντίστοιχα. Ειδικότερα, η διάμεσος και το 1ο τεταρτημόριο υπολογίζονται από τις σχέσεις

$$q_{0.5} = L_i + d \cdot \frac{0.5 \cdot n - N_{i-1}}{\nu_i} = 168.5 + 5 \cdot \frac{0.5 \cdot 40 - 11}{12} = 168.5 + 3.75 = 172.25$$

και

$$q_{0.25} = L_i + d \cdot \frac{0.25 \cdot n - N_{i-1}}{\nu_i} = 163.5 + 5 \cdot \frac{0.25 \cdot 40 - 3}{8} = 163.5 + 4.375 = 167.875$$

αντίστοιχα, αφού η ομάδα που ανήκει η διάμεσος είναι η ομάδα $[168.5 - 173.5)$, ενώ η ομάδα που ανήκει το 1ο τεταρτημόριο είναι η ομάδα $[163.5 - 168.5)$ (γιατί;).

Οι παραπάνω τιμές δεν είναι ίδιες με τις αντίστοιχες τιμές που υπολογίστηκαν από τα πρωτογενή δεδομένα αλλά είναι αρκετά κοντά σε αυτές. Αυτό συμβαίνει καθώς με την ομαδοποίηση των δεδομένων χάνεται κάποια πληροφορία από τις αρχικές τιμές έχοντας ως συνέπεια να επηρεάζονται οι τιμές που υπολογίζουμε για τα μέτρα αυτά.

Άσκηση Αυτοαξιολόγησης 8.1

Να βρεθούν η μέση τιμή, η διάμεσος, καθώς και το 75ο ποσοστιαίο σημείο, δηλαδή το 3ο τεταρτημόριο, για τα παρακάτω δεδομένα

1.10	2.12	2.11	0.98	5.01
3.35	1.89	2.97	4.12	5.12
4.12	4.64	4.21	3.11	1.53
2.10	1.01	2.97	13.2	0.22

Στη συνέχεια, ομαδοποιήστε τα παραπάνω δεδομένα χρησιμοποιώντας τα διαστήματα $[0,1), [1,2), \ldots [13, 14)$ και προσδιορίστε εκ νέου τη μέση τιμή, τη διάμεσο και το 75ο ποσοστιαίο σημείο.

8.3.2.2 Μέτρα μεταβλητότητας

Τα μέτρα μεταβλητότητας αποτυπώνουν χαρακτηριστικά σχετικά με τη διασκόρπιση των δεδομένων, δηλαδή τον τρόπο που αυτά είναι συγκεντρωμένα ή απλωμένα. Τα βασικότερα μέτρα μεταβλητότητας ενός δείγματος $x_1, \dots, x_n$ είναι:

- το δειγματικό εύρος ή αλλιώς δειγματική έκταση (range),
- το δειγματικό ενδοτεταρτημοριακό εύρος (interqunatile range),
- η δειγματική διασπορά ή δειγματική διακύμανση (variance),
- η δειγματική τυπική απόκλιση (standard deviation),
- ο συντελεστής μεταβλητότητας (coefficient of variation).

Ορισμός 8.6

Το **δειγματικό εύρος** R ή, αλλιώς, **δειγματική έκταση** ορίζεται να είναι η διαφορά της ελάχιστης τιμής από τη μέγιστη τιμή του δείγματος, δηλαδή

$$R = x_{(n)} - x_{(1)},$$

όπου $x_{(i)}$ συμβολίζει την i-οστή διατεταγμένη παρατήρηση.

Το δειγματικό εύρος είναι το πιο απλό και εύκολα υπολογίσιμο μέτρο μεταβλητότητας και αποτυπώνει το πλάτος του διαστήματος μέσα στο οποίο παρατηρήθηκαν όλες οι τιμές. Είναι φανερό ότι δείγματα με μεγαλύτερο εύρος παρουσιάζουν μεγαλύτερη μεταβλητότητα σε σχέση με άλλα που έχουν μικρότερο εύρος.

Το δειγματικό εύρος, που παρουσιάστηκε παραπάνω, είναι πράγματι ένα απλό και εύκολο μέτρο μεταβλητότητας. Παρ' όλα αυτά βασίζεται μόνο σε δύο τιμές και μάλιστα στις δύο πιο απομακρυσμένες τιμές. Αυτό το καθιστά ιδιαίτερα ευαίσθητο σε ακραίες τιμές ή σε λάθη καταγραφής των τιμών ενός δείγματος. Ένα παρόμοιας λογικής με το δειγματικό εύρος μέτρο μεταβλητότητας είναι το δειγματικό ενδοτεταρτημοριακό εύρος, IQR (Interquantile range).

Ορισμός 8.7

Το **δειγματικό ενδοτεταρτημοριακό εύρος** ορίζεται να είναι η διαφορά της τιμής του 1ου τεταρτημορίου από την τιμή του 3ου τεταρτημορίου, δηλαδή

$$IQR = q_{0.75} - q_{0.25} = Q_3 - Q_1. \tag{8.7}$$

Το δειγματικό ενδοτεταρτημοριακό εύρος αποτυπώνει το πλάτος του διαστήματος μέσα στο οποίο βρίσκεται το κεντρικό 50% των παρατηρήσεων. Προφανώς, το δειγματικό ενδοτεταρτημοριακό εύρος δεν αποτυπώνει την ολική μεταβλητότητα των παρατηρήσεων του δείγματος, αλλά μόνο του κεντρικού τμήματος. Παρ΄όλα αυτά, μας δίνει σημαντικές πληροφορίες για τη μεταβλητότητα των δεδομένων μας. Επιπρόσθετα, συγκρίνοντάς το με το δειγματικό εύρος, το δειγματικό ενδοτεταρτημοριακό εύρος δεν είναι ευαίσθητο στην ύπαρξη ακραίων τιμών.

Το βασικότερο μέτρο μεταβλητότητας είναι η δειγματική διασπορά, της οποίας ο ορισμός παρουσιάζεται στη συνέχεια.

Ορισμός 8.8

Η **δειγματική διασπορά** εκφράζει τη διασπορά ή τη μεταβλητότητα ενός συνόλου αριθμητικών δεδομένων από τη μέση τιμή τους και δίνεται από τη σχέση:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2. \tag{8.8}$$

Η παραπάνω σχέση μπορεί να δοθεί ισοδύναμα και από τις παρακάτω σχέσεις:

$$s^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right) = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right)$$

Παρατηρήστε ότι η δειγματική διασπορά λαμβάνει υπόψη της όλες τις παρατηρήσεις του δείγματος. Επιπλέον, η λογική πίσω από τη σχέση ορισμού της βασίζεται στη διαπίστωση ότι ένα μέτρο μεταβλητότητας θα μπορούσε να βασιστεί στις αποστάσεις των παρατηρήσεων από ένα σταθερό σημείο. Αν οι αποστάσεις αυτές είναι μεγάλες και πολύ διαφορετικές μεταξύ τους, τότε και η μεταβλητότητα των δεδομένων θα μπορούσε να χαρακτηριστεί ως μεγάλη. Αν αντίθετα ήταν μικρές και όλες παρόμοιες, τότε οι παρατηρήσεις θα ήταν κοντά η μία στην άλλη και η μεταβλητότητα μικρή.

Το πρώτο και βασικότερο βήμα στην υλοποίηση αυτού του σκεπτικού είναι η επιλογή του σταθερού σημείου από το οποίο θα υπολογιστούν οι αποστάσεις των παρατηρήσεων. Το σημείο αυτό θα μπορούσε να είναι ένας οποιοσδήποτε αυθαίρετα επιλεγμένος πραγματικός αριθμός ή ένα χαρακτηριστικό μέγεθος του δείγματος, όπως η μέση τιμή, η οποία είναι και η τιμή που χρησιμοποιήθηκε.

Επιστρέφοντας στον αρχικό μας συλλογισμό, αν υπολογίσουμε τις αποστάσεις των παρατηρήσεων από τη (δειγματική) μέση τιμή τους, θα παρατηρήσουμε ότι αυτές αθροίζουν στο μηδέν, αφού:

$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0.$$

Ο λόγος είναι ότι υπάρχουν παρατηρήσεις που είναι μικρότερες από τη μέση τιμή και άρα έχουν αρνητική απόσταση από αυτήν και άλλες που είναι μεγαλύτερες και άρα η απόστασή τους είναι θετική. Ένας τρόπος να απαλλαγούμε από τις αρνητικές τιμές είναι να υψώσουμε τις αποστάσεις στο τετράγωνο. Αν μια απόσταση (αρνητική ή θετική) είναι κατά απόλυτα τιμή μεγάλη, τότε μεγάλη θα είναι και η τιμή του τετραγώνου της. Αν από την άλλη είναι μικρή, τότε μικρή θα είναι και η τιμή του τετραγώνου. Επομένως, τροποποιώντας λίγο τον αρχικό μας συλλογισμό, μπορούμε να πούμε ότι ένα μέτρο μεταβλητότητας μπορεί να βασιστεί στο τετράγωνο των αποστάσεων των παρατηρήσεων από τη μέση τιμή τους.

Ολοκληρώνοντας το σκεπτικό μας, έτσι ώστε να καταλήξουμε στη σχέση ορισμού της διασποράς, αρκεί να παρατηρήσουμε ότι, για να ποσοτικοποιήσουμε τη μεταβλητότητα των δεδομένων, δεν αρκεί να υπολογίσουμε ένα ή κάποια από τα τετράγωνα των αποστάσεων. Ένα ολικό μέτρο μεταβλητότητας θα πρέπει να βασίζεται πάνω σε όλες τις αποστάσεις ή, καλύτερα, στη μέση τιμή των τετραγώνων των αποστάσεων των παρατηρήσεων από τη μέση τιμή τους. Ο λόγος που η διασπορά ορίζεται με βάση τη σχέση (8.8), δηλαδή ως κάτι που δεν είναι ακριβώς η μέση τιμή των τετραγώνων των αποστάσεων (αφού δεν διαιρείται με το n αλλά με το $n-1$) αφορά τις ιδιότητες της τυχαίας μεταβλητής της διασποράς, S^2, και, πιο συγκεκριμένα, το ότι η αναμενόμενη τιμή της δειγματικής διασποράς S^2 είναι η πραγματική τιμή της πληθυσμιακής διασποράς. Οι ιδιότητες της τυχαίας μεταβλητής της δειγματικής διασποράς S^2 μελετώνται αναλυτικά στο επόμενο κεφάλαιο.

Κλείνοντας την αναφορά μας στη δειγματική διασπορά, αξίζει να σημειώσουμε ότι η επιλογή του δειγματικού μέσου για τον προσδιορισμό των τετραγώνων των αποστάσεων των παρατηρήσεων δικαιολογείται, πέρα

από τις ιδιότητες που έχει η S^2, και από το γεγονός ότι η δειγματική μέση τιμή είναι η τιμή για την οποία ελαχιστοποιείται, όπως έχουμε αναφέρει και νωρίτερα, ως προς a η συνάρτηση:

$$g(a) = \sum_{i=1}^{n}(x_i - a)^2.$$

Η δειγματική διασπορά είναι ένα μέτρο μεταβλητότητας που υπολογίζεται λαμβάνοντας υπόψη όλες τις παρατηρήσεις του δείγματος. Για τον λόγο αυτό, και προτιμάται σε σχέση με το εύρος και το ενδοτεταρτημοριακό εύρος. Ωστόσο η δειγματική διασπορά έχει ένα βασικό μειονέκτημα, καθώς οι μονάδες μέτρησής της είναι το τετράγωνο των μονάδων μέτρησης των παρατηρήσεων του δείγματος. Παραδείγματος χάριν, αν οι μετρήσεις αφορούν θερμοκρασίες σε βαθμούς Celsius, τότε οι μονάδες μέτρησης της διασποράς είναι οι βαθμοί Celsius^2.

Ένας εύκολος τρόπος να επιστρέψουμε στις αρχικές μονάδες μέτρησης είναι να πάρουμε την τετραγωνική ρίζα της διασποράς. Η διαδικασία αυτή ορίζει την τυπική απόκλιση των παρατηρήσεων.

Ορισμός 8.9

Η **δειγματική τυπική απόκλιση** των παρατηρήσεων δίνεται από τη σχέση:

$$s = \sqrt{s^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}. \tag{8.9}$$

Παρατήρηση 8.6

Στην περίπτωση που στη διάθεσή μας έχουμε μόνο τα ομαδοποιημένα δεδομένα ενός δείγματος, η δειγματική διακύμανση υπολογίζεται μέσω της σχέσης:

$$s^2 = \frac{1}{n-1}\sum_{\ell=1}^{k}\nu_\ell\,(k_\ell - \bar{x})^2, \tag{8.10}$$

όπου k_ℓ η κεντρική τιμή της ℓ-οστής ομάδας/κλάσης.

Οι τιμές της διασποράς ή της τυπικής απόκλισης από μόνες τους δεν μπορούν να καταδείξουν αν οι διαφορές μεταξύ των παρατηρήσεων είναι σημαντικά μεγάλες ή όχι και αυτό γιατί δεν υπάρχει κάποιο κριτήριο για το ποιες τιμές θεωρούνται μεγάλες και ποιες μικρές. Ο λόγος είναι ότι τα μέτρα αυτά βασίζονται στις διαφορές που έχουν οι τιμές από τη μέση τιμή τους και όχι στο μέγεθός τους. Για να γίνει αυτό πιο κατανοητό, αρκεί να παρατηρήσουμε ότι οι τιμές της τυπικής απόκλισης των τιμών

$$100,\ 200,\ 300$$

και των

$$999900,\ 1000000,\ 1000100$$

είναι ίσες μεταξύ τους. Όμως οι αποστάσεις από τη μέση τιμή στο δεύτερο σύνολο δεδομένων μπορούν να θεωρηθούν μικρές, εν αντιθέσει με τις αντίστοιχες αποστάσεις των τιμών του πρώτου δείγματος από τη δική τους μέση τιμή. Παρατηρήστε ότι στο δεύτερο σύνολο δεδομένων οι τιμές είναι, πρακτικά, όλες ίσες μεταξύ τους εξαιτίας του μεγάλου μεγέθους τους.

Η παραπάνω παρατήρηση μας οδηγεί στην ανάγκη ορισμού ενός μέτρου σχετικής μεταβλητότητας, του συντελεστή μεταβλητότητας cv.

Ορισμός 8.10

Ο **δειγματικός συντελεστής μεταβλητότητας** είναι ένα μέτρο της μεταβλητότητας απαλλαγμένο από τη μέση τιμή και τις μονάδες μέτρησής της, που ορίζεται ως το πηλίκο της δειγματικής τυπικής απόκλισης προς την απόλυτη τιμή της δειγματικής μέσης τιμής, δηλαδή ισούται με:

$$cv = \frac{s}{|\bar{x}|}.$$

Ο συντελεστής μεταβλητότητας μπορεί να χρησιμοποιηθεί για τη σύγκριση της μεταβλητότητας διαφορετικών συνόλων δεδομένων. Παραδείγματος χάριν, για τα παραπάνω σύνολα δεδομένων, η τιμή του cv για το πρώτο είναι 0.5, ενώ για το δεύτερο 0.0001. Η διαφορά τους αποτυπώνει το γεγονός ότι στο πρώτο σύνολο δεδομένων οι διαφορές μεταξύ των παρατηρήσεων είναι σημαντικές και ουσιαστικές, ενώ στο δεύτερο όχι. Στην πραγματικότητα, υπάρχει ένας εμπειρικός κανόνας ο οποίος χαρακτηρίζει τα δείγματα με $cv < 0.1$ (ή, ισοδύναμα, με $cv < 10\%$) ως ομοιογενή, ενώ τα δείγματα με μεγαλύτερες τιμές του cv ως ετερογενή ή ανομοιογενή.

Παρατήρηση 8.7

Σημειώνεται ότι όλα τα μέτρα μεταβλητότητας είναι αυστηρά μη αρνητικοί αριθμοί.

Παράδειγμα 8.5

Να υπολογιστούν το εύρος και ο συντελεστής μεταβλητότητας για τα δεδομένα του Παραδείγματος 8.3. Στη συνέχεια, χρησιμοποιώντας τον ομαδοποιημένο πίνακα του παραδείγματος, υπολογίστε εκ νέου τον συντελεστή μεταβλητότητας.

Λύση Παραδείγματος 8.5

Το εύρος των δεδομένων του Παραδείγματος 8.3 ισούται με

$$R = x_{(n)} - x_{(1)} = 188 - 159 = 29\ cm.$$

Για τον υπολογισμό του συντελεστή μεταβλητότητας απαραίτητος είναι ο υπολογισμός της μέσης τιμής, την οποία έχουμε προσδιορίσει (βλ. τη λύση του Παραδείγματος 8.3) ότι ισούται με 173.3 και ο υπολογισμός της δειγματικής τυπικής απόκλισης, η οποία ισούται με

$$s = \sqrt{s^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} = \sqrt{\frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right)} = \ldots = 7.157567\ cm.$$

Επομένως, ο συντελεστής μεταβλητότητας ισούται με:

$$cv = \frac{7.157567}{|173.3|} = 0.0413016.$$

Η παραπάνω τιμή είναι μικρότερη του 0.1, οπότε το δείγμα των υψών των φοιτητριών μπορεί να χαρακτηριστεί ως ομοιογενές.

Χρησιμοποιώντας τον ομαδοποιημένο πίνακα που δόθηκε στη λύση του Παραδείγματος 8.3, η

δειγματική τυπική απόκλιση υπολογίζεται ως ακολούθως (βλ. τη σχέση (8.10)):

$$s = \sqrt{s^2} = \sqrt{\frac{1}{n-1}\sum_{\ell=1}^{k} \nu_\ell (k_\ell - \bar{x})^2}$$
$$= \sqrt{\frac{1}{40-1}\left(3 \cdot (161 - 172.875)^2 + \ldots + 3 \cdot (186 - 172.875)^2\right)}$$
$$= \sqrt{\frac{1792.375}{40-1}} = 6.779258\ cm.$$

Επομένως, ο συντελεστής μεταβλητότητας με βάση τα ομαδοποιημένα δεδομένα ισούται με:

$$cv = \frac{6.779258}{|172.875|} = 0.03921479,$$

η τιμή του οποίου είναι πάλι μικρότερη του 0.1, οπότε το δείγμα των υψών των φοιτητριών μπορεί πάλι να χαρακτηριστεί ως ομοιογενές.

Σημειώνεται ότι κατά τον υπολογισμό της τυπικής απόκλισης και του συντελεστή μεταβλητότητας στα ομαδοποιημένα δεδομένα χρησιμοποιήθηκε η δειγματική μέση τιμή με βάση τα ομαδοποιημένα δεδομένα και όχι τα πρωτογενή.

Άσκηση Αυτοαξιολόγησης 8.2

Να υπολογιστούν το εύρος και ο συντελεστής μεταβλητότητας για τα δεδομένα της Άσκησης Αυτοαξιολόγησης 8.1. Στη συνέχεια, χρησιμοποιώντας τον ομαδοποιημένο πίνακα, υπολογίστε εκ νέου τον συντελεστή μεταβλητότητας.

8.3.2.3 Μέτρα σχήματος ή μορφής

Όπως για τις τυχαίες μεταβλητές, έτσι και για το δείγμα μπορούμε να ορίσουμε τον **δειγματικό συντελεστή λοξότητας** (skewness) και τον **δειγματικό συντελεστή κύρτωσης** (kurtosis), δηλαδή εκείνα τα μέτρα που μας δίνουν πληροφορίες για το σχήμα ή, αλλιώς, τη μορφή της κατανομής του δείγματος. Για να ορίσουμε τους συντελεστές αυτούς χρειάζεται πρώτα να ορίσουμε τις δειγματικές κεντρικές ροπές[2].

Η **δειγματική κεντρική ροπή** k **τάξεως** ορίζεται ως

$$m_k = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^k.$$

Με τη βοήθεια των δειγματικών κεντρικών ροπών ορίζουμε τον δειγματικό συντελεστή λοξότητας α_3 από τη σχέση:

$$\alpha_3 = \frac{1}{n}\sum_{i=1}^{n}\frac{(x_i - \bar{x})^3}{m_2^{3/2}} = \frac{m_3}{m_2^{3/2}},$$

ενώ ο δειγματικός συντελεστής κύρτωσης α_4 ορίζεται από τη σχέση:

$$\alpha_4 = \frac{1}{n}\sum_{i=1}^{n}\frac{(x_i - \bar{x})^4}{m_2^{4/2}} = \frac{1}{n}\sum_{i=1}^{n}\frac{(x_i - \bar{x})^4}{m_2^2} = \frac{m_4}{m_2^2}.$$

[2] Σημειώνεται ότι μπορούμε να ορίσουμε και τις δειγματικές (μη κεντρικές) ροπές k τάξης ως $m_k' = \frac{1}{n}\sum_{i=1}^{n} x_i^k$.

Ο δειγματικός συντελεστής λοξότητας εκφράζει τη λοξότητα της κατανομής και, πιο συγκεκριμένα,:

- αν $\alpha_3 > 0$, τότε έχουμε ένα δείγμα από έναν δεξιά λοξευμένο ή, αλλιώς, θετικά ασύμμετρο πληθυσμό,
- αν $\alpha_3 < 0$, τότε έχουμε ένα δείγμα από έναν αριστερά λοξευμένο πληθυσμό, ενώ
- αν $\alpha_3 \approx 0$, τότε έχουμε ένα δείγμα από έναν περίπου συμμετρικό πληθυσμό.

Από την άλλη πλευρά, χρησιμοποιώντας τον δειγματικό συντελεστή κύρτωσης έχουμε ότι:

- αν $\alpha_4 > 3$, τότε το δείγμα προέρχεται από έναν λεπτόκυρτο πληθυσμό,
- αν $\alpha_4 < 3$, τότε το δείγμα προέρχεται από έναν πλατύκυρτο πληθυσμό, ενώ
- αν $\alpha_4 \approx 3$, τότε το δείγμα προέρχεται από έναν μεσόκυρτο πληθυσμό.

Παρατήρηση 8.8

Σημειώνεται ότι για τον χαρακτηρισμό της συμμετρίας ενός δείγματος μπορούν να χρησιμοποιηθούν και τα μέτρα θέσης. Παραδείγματος χάριν,

- όταν $\alpha_3 > 0$, τότε συνήθως έχουμε ότι $m_0 < m < \bar{x}$ και $Q_3 - m > m - Q_1$,
- όταν $\alpha_3 < 0$, τότε συνήθως έχουμε ότι $\bar{x} < m < m_0$ και $Q_3 - m < m - Q_1$,
- όταν $\alpha_3 \approx 0$, τότε συνήθως έχουμε ότι $\bar{x} \approx m \approx m_0$ και $Q_3 - m \approx m - Q_1$.

8.3.3 Γραφικές παραστάσεις ποσοτικών δεδομένων

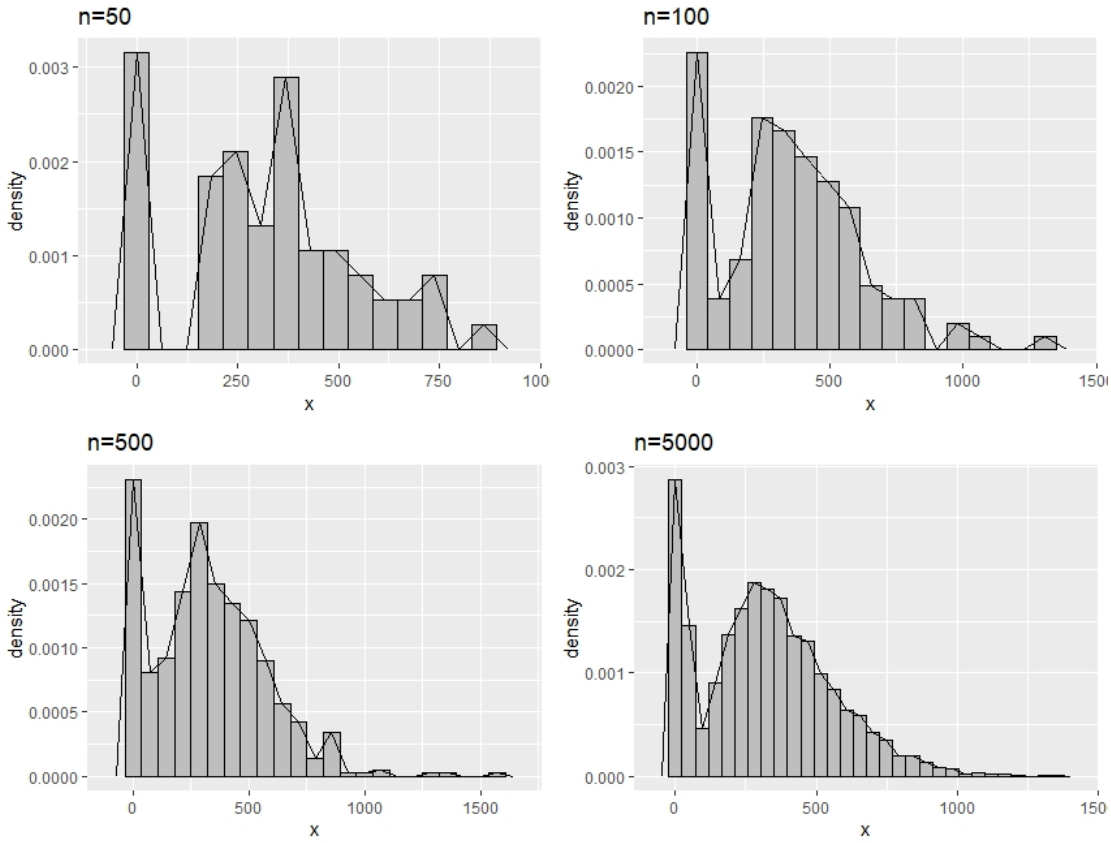
Για τις ποσοτικές μεταβλητές υπάρχει μια σειρά από γραφικές παραστάσεις που οπτικοποιούν πολλά και συνήθως διαφορετικά χαρακτηριστικά των δεδομένων. Δύο από τις πιο συχνά χρησιμοποιούμενες γραφικές παραστάσεις ποσοτικών δεδομένων είναι:

α) το **ραβδόγραμμα** (bar chart), όταν πρόκειται για ποσοτικά δεδομένα διακριτών τυχαίων μεταβλητών και

β) το **ιστόγραμμα** (histogram), όταν πρόκειται για ποσοτικά δεδομένα συνεχών τυχαίων μεταβλητών.

Το ραβδόγραμμα κατασκευάζεται με τον ίδιο ακριβώς τρόπο που παρουσιάστηκε στις ποιοτικές μεταβλητές και για τον λόγο αυτό, η κατασκευή του δεν σχολιάζεται περαιτέρω. Για την κατασκευή του ιστογράμματος οι τιμές της συνεχούς ποσοτικής μεταβλητής αρχικά ομαδοποιούνται σε μορφή κλειστών αριστερά, ανοικτών δεξιά συνεχόμενων διαστημάτων ακολουθώντας την ίδια διαδικασία με την κατασκευή των ομαδοποιημένων πινάκων συχνοτήτων που παρουσιάστηκε νωρίτερα. Στη συνέχεια, το ιστόγραμμα συχνοτήτων κατασκευάζεται χρησιμοποιώντας ένα σύνολο συγγενών (εφαπτόμενων) ορθογώνιων παραλληλόγραμμων, των οποίων το ύψος είναι ανάλογο με τη συχνότητα (ή τη σχετική συχνότητα) κάθε ομάδας και το μήκος της βάσης τους είναι ίσο με το μήκος του διαστήματος που ορίζεται από τα άκρα της κάθε ομάδας.

Ενδιαφέρον παρουσιάζει η περίπτωση όπου το ύψος κάθε ορθογωνίου παραλληλόγραμμου καθορίζεται έτσι ώστε το εμβαδόν του να ισούται με τη σχετική συχνότητα της κάθε ομάδας. Στην περίπτωση αυτή, το συνολικό εμβαδόν του ιστογράμματος ισούται με 1. Μάλιστα, εδώ, έχει ιδιαίτερη αξία και το αποκαλούμενο **πολύγωνο σχετικών συχνοτήτων**. Το πολύγωνο σχετικών συχνοτήτων κατασκευάζεται ενώνοντας με ευθύγραμμα τμήματα τα διαδοχικά μέσα των άνω πλευρών των ορθογωνίων του ιστογράμματος, ξεκινώντας από το μέσο μιας εικονικής πρώτης ομάδας ίδιου μήκους και καταλήγοντας στο μέσο μιας εικονικής ομάδας μετά την τελευταία (Σχήμα 8.2). Το πολύγωνο αυτό, στην περίπτωση που το μέγεθος του δείγματος αυξάνεται και τα πλάτη των ομάδων μικραίνουν, συγκλίνει σε μια καμπύλη, η οποία αντιπροσωπεύει τη συνάρτηση πυκνότητας πιθανότητας του πληθυσμού (κάτω δεξιά γραφική παράσταση του Σχήματος 8.2, για $n = 5000$). Επομένως, το ιστόγραμμα και το πολύγωνο σχετικών συχνοτήτων μπορούν να μας δώσουν πληροφορίες για την κατανομή του υπό μελέτη πληθυσμού και τις ιδιότητές της. Παραδείγματος χάριν, όλα τα ιστογράμματα του Σχήματος 8.2, ακόμα και αυτά με μικρότερο μέγεθος δείγματος, όπως για $n = 50$ (πάνω αριστερή γραφική παράσταση Σχήματος 8.2), φανερώνουν ότι τα δεδομένα προέρχονται από έναν δικόρυφο, δεξιά λοξευμένο πληθυσμό. Η δικόρυφη κατανομή αναφέρεται



Σχήμα 8.2: Ιστόγραμμα και πολύγωνο συχνοτήτων για δείγματα διάφορων μεγεθών n από έναν δικόρυφο, δεξιά λοξευμένο πληθυσμό.

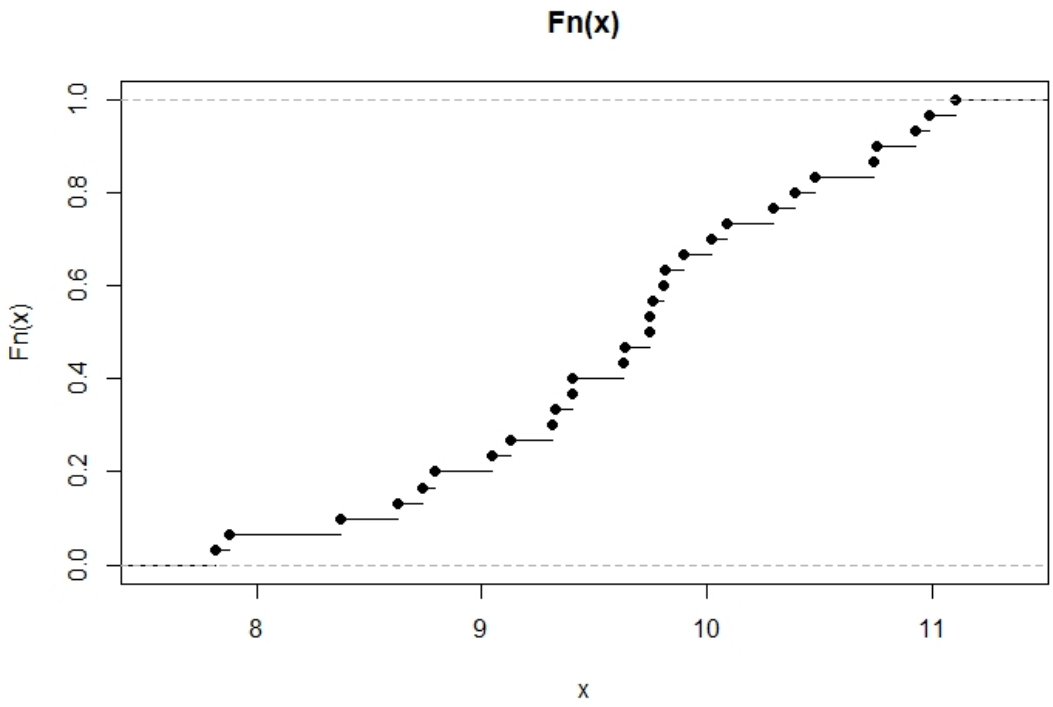
στις περιπτώσεις που το ιστόγραμμα παρουσιάζει δύο (τοπικά) μέγιστα, τις αποκαλούμενες κορυφές, ενώ η έννοια της δεξιάς λοξευμένης κατανομής αναφέρεται στην έλλειψη συμμετρίας των δεδομένων και την παρουσία περισσότερων μεγάλων τιμών από ότι μικρών.

Η γραφική παρουσίαση των δεδομένων ποσοτικών συνεχών μεταβλητών δεν περιορίζεται στο ιστόγραμμα ή στο πολύγωνο (σχετικών) συχνοτήτων. Στην πραγματικότητα, υπάρχει μια πληθώρα γραφικών παραστάσεων, η καθεμία από τις οποίες αποτυπώνει διαφορετικά χαρακτηριστικά των δεδομένων. Ενδεικτικά αναφέρουμε:

1. το διάγραμμα (ή το πολύγωνο) των αθροιστικών (σχετικών) συχνοτήτων, το οποίο παρέχει πληροφορίες για τη μορφή της συνάρτησης κατανομής $F(x)$ του πληθυσμού,
2. το διάγραμμα μόσχου-φύλλου (stem and leaf plot), το οποίο μπορεί να θεωρηθεί ως παραλλαγή του ιστογράμματος, στο οποίο αποτυπώνονται οι αριθμητικές τιμές των δεδομένων, καθώς και οι συχνότητες κάθε ομάδας και
3. το θηκόγραμμα (boxplot), το οποίο παρουσιάζεται αναλυτικά στη συνέχεια.

8.3.3.1 Διάγραμμα αθροιστικών σχετικών συχνοτήτων

Το διάγραμμα αθροιστικών σχετικών συχνοτήτων είναι ένα διάγραμμα που στον οριζόντιο άξονα τοποθετούνται οι τιμές των παρατηρήσεων διατεταγμένες κατά αύξουσα τάξη και στον κατακόρυφο άξονα οι αθροιστικές σχετικές συχνότητές τους. Το διάγραμμα αθροιστικών σχετικών συχνοτήτων έχει τη μορφή του Σχήματος 8.3, δηλαδή πρόκειται για μια δεξιά συνεχή, μη φθίνουσα συνάρτηση, η οποία μπορεί να χρησιμοποιηθεί για την εκτίμηση της συνάρτησης κατανομής, $F(x)$, του πληθυσμού από τον οποίο έχουμε



Σχήμα 8.3: Διάγραμμα αθροιστικών σχετικών συχνοτήτων.

λάβει το δείγμα. Το διάγραμμα αθροιστικών σχετικών συχνοτήτων παρουσιάζει άλματα στα σημεία που υπάρχουν παρατηρήσεις, ενώ το άλμα (αύξηση της τιμής) ισούται με τη σχετική συχνότητα της παρατήρησης.

8.3.3.2 Διάγραμμα μόσχου-φύλλου

Το διάγραμμα μόσχου-φύλλου αποτελεί μια παραλλαγή του ιστογράμματος, στο οποίο αποτυπώνονται οι αριθμητικές τιμές των δεδομένων, καθώς και οι συχνότητες κάθε ομάδας. Στη συνέχεια, παρουσιάζεται το διάγραμμα μόσχου-φύλλου ενός δείγματος μεγέθους 30, στο οποίο οι παρατηρήσεις έχουν χωριστεί με βάση τις ομάδες [7.5, 8), [8, 8.5), ..., [11.0, 11.5) και έχει καταγραφεί το πρώτο δεκαδικό τους ψηφίο (φύλλο).

```
------------------------------------------------------------------------
1 | 2: represents 1.2
 leaf unit: 0.1
            n: 30
   2     7. | 89
   3     8* | 4
   6     8. | 678
  12     9* | 013344
  (8)    9. | 66778889
  10    10* | 0134
   6    10. | 5789
   2    11* | 01
------------------------------------------------------------------------
```

Από την εικόνα του παραπάνω γραφήματος μπορούμε όχι μόνο να διακρίνουμε χαρακτηριστικά όπως τη συχνότητα κάθε ομάδας, π.χ. η συχνότητα της ομάδας [8-8.5) ισούται με 1, αλλά ταυτόχρονα να δούμε τις τιμές που ανήκουν σε καθεμία ομάδα, π.χ. η παρατήρηση που ανήκει στην ομάδα [8-8.5) είναι η 8.4. Επιπρόσθετα, στην πρώτη στήλη παρουσιάζονται οι αθροιστικές συχνότητες της κάθε κλάσης από την πρώτη μέχρι και την κλάση πριν από αυτήν που περιέχει τη διάμεσο. Η κλάση που περιέχει τη διάμεσο διαχωρίζεται από τις υπόλοιπες καθώς στην πρώτη στήλη, μέσα σε παρένθεση, καταγράφεται η συχνότητα εμφάνισής της. Για τις κλάσεις από τη διάμεσο και μετά δίνονται οι αθροιστικές συχνότητες από την τελευταία κλάση προς την κλάση που περιέχει τη διάμεσο. Στο παράδειγμά μας, η κεντρική ομάδα είναι η ομάδα [9.5, 10), η οποία περιέχει 8 παρατηρήσεις, τις 9.6, 9.6, 9.7, 9.7, 9.8, 9.8, 9.8 και 9.9. Καθώς έχουμε 30 παρατηρήσεις και η αθροιστική συχνότητα της ομάδας πριν την κεντρική ισούται με 12, η διάμεσος προκύπτει από το ημιάθροισμα της $15^{ης}$ και της $16^{ης}$ παρατήρησης, δηλαδή από το ημιάθροισμα των τιμών 9.7, 9.7, οπότε ισούται με 9.7. Σημειώνεται ότι τα αστεράκια που εμφανίζονται δίπλα σε κάποιους από τους μόσχους, υπάρχουν για να διακρίνονται πιο εύκολα οι ομάδες που είναι της μορφής $[x.0,\ x.5)$ από τις ομάδες $[x.5,\ (x+1).0)$, όπου x οι τιμές των μόσχων.

8.3.3.3 Θηκόγραμμα

Το θηκόγραμμα έχει τη μορφή του Σχήματος 8.4 και χαρακτηρίζεται από ένα ορθογώνιο κουτί, του οποίου το κάτω και πάνω άκρο αντιστοιχούν στην τιμή του 25ου και 75ου ποσοστιαίου σημείου, αντίστοιχα, δηλαδή, στο 1ο (Q_1) και στο 3ο (Q_3) τεταρτημόριο, αντίστοιχα. Η διάμεσος παριστάνεται από μία οριζόντια γραμμή μέσα στο κουτί. Στην αρχή και στην κορυφή του σχήματος σημειώνονται δύο γραμμές, που αναφέρονται ως φράχτες (whiskers). Η πάνω γραμμή επεκτείνεται μέχρι το σημείο

$$max\{x_i : x_i \leq Q_3 + 1.5(Q_3 - Q_1)\}$$

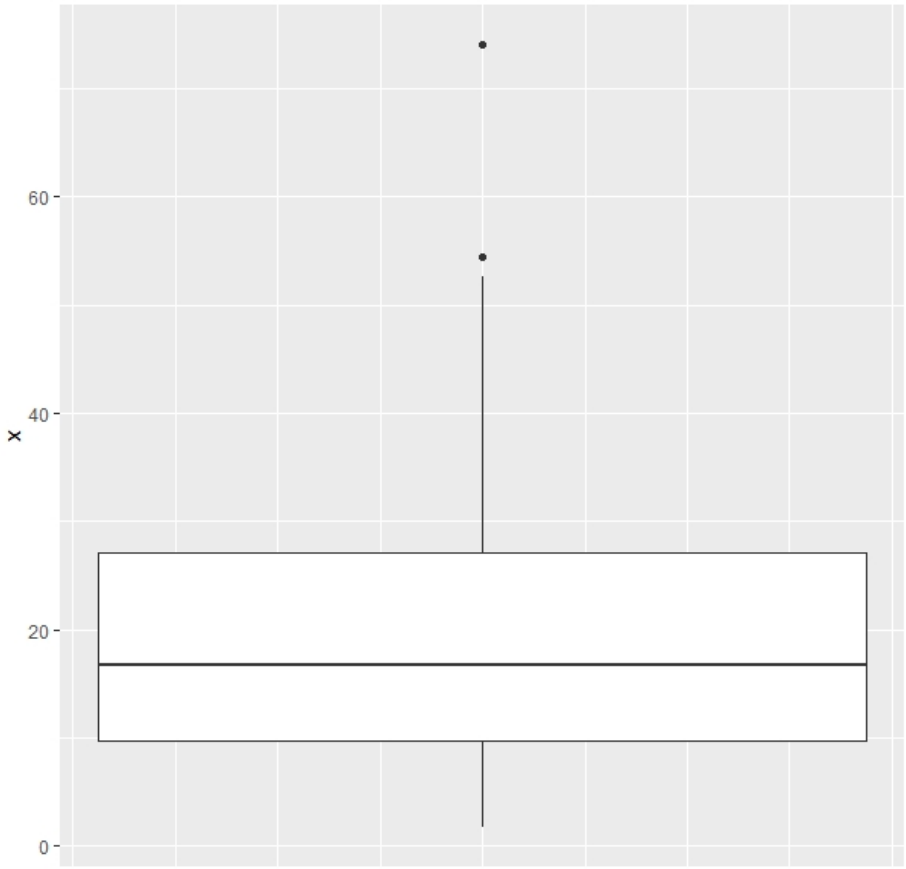
ενώ η κάτω μέχρι το σημείο

$$min\{x_i : x_i \geq Q_1 - 1.5(Q_3 - Q_1)\}.$$

Το εύρος της 1.5 φοράς της διαφοράς $(Q_3 - Q_1)$, που αντιπροσωπεύει το κεντρικό 50% των παρατηρήσεων, δεξιά και αριστερά από το κεντρικό κουτί, αποτυπώνει την περιοχή των τιμών που θεωρούνται μη αντιφατικές, δηλαδή μη παράξενες, σε σχέση με τις υπόλοιπες του δείγματος. Οποιαδήποτε τιμή έξω από τα όρια αποτυπώνεται με μια τελεία (ή με έναν άλλο οποιονδήποτε χαρακτήρα) στο γράφημα και μπορεί να χαρακτηριστεί ως ακραία τιμή (outlier). Με τον όρο ακραία τιμή χαρακτηρίζεται μια παρατηρούμενη τιμή που είναι αντιφατική σε σχέση με τις υπόλοιπες παρατηρούμενες τιμές του συνόλου των δεδομένων.

Από την εικόνα του θηκογράμματος μπορούν να εξαχθούν σημαντικές πληροφορίες για το υπό μελέτη δείγμα. Παραδείγματος χάριν με βάση το Σχήμα 8.4, μπορούμε να πούμε ότι το δείγμα προέρχεται από έναν δεξιά λοξευμένο πληθυσμό, καθώς:

1. το κεντρικό 50% των παρατηρήσεων δεν είναι συμμετρικά κατανεμημένο γύρω από τη διάμεσο (η απόσταση μεταξύ του τρίτου τεταρτημορίου και της διαμέσου είναι μεγαλύτερη από την απόσταση μεταξύ της διαμέσου και του 1ου τεταρτημορίου),
2. η πάνω γραμμή επεκτείνεται σε όλο (ή σχεδόν σε όλο) το μήκος της 1.5 φοράς της διαφοράς $(Q_3 - Q_1)$, γεγονός που δεν ισχύει για την κάτω γραμμή. Αυτό υποδηλώνει ότι το κάτω 25% των παρατηρήσεων είναι συγκεντρωμένο σε ένα μικρότερο διάστημα σε σχέση με το πάνω 25% των παρατηρήσεων και
3. υπάρχουν δύο ακραίες μεγάλες τιμές, ενώ δεν υπάρχουν αντίστοιχα ακραίες μικρές τιμές.



Σχήμα 8.4: Θηκόγραμμα για ένα δείγμα από έναν δεξιά λοξευμένο πληθυσμό.

8.4 Εφαρμογή-Παράδειγμα

Στον παρακάτω πίνακα παρουσιάζεται τμήμα των μετρήσεων που αφορούν τον όγκο της εισροής λυμάτων (σε χιλιάδες κυβικά μέτρα) σε μια μονάδα επεξεργασίας αστικών λυμάτων στην Ισπανία κατά τη διάρκεια μιας βάρδιας εργαζομένων (8 ώρες). Οι μετρήσεις αφορούν 509 τυχαία επιλεγμένες μέρες και, στην πραγματικότητα, αποτελούν τμήμα μιας ευρύτερης μελέτης αξιολόγησης της λειτουργίας της μονάδας επεξεργασίας αστικών λυμάτων.

44101	39024	32229	35023	36924	38572	....

Με τη βοήθεια της R λαμβάνουμε τα περιγραφικά στοιχεία του Πίνακα 8.1.

Από τα αποτελέσματα αυτά, διαπιστώνουμε ότι η μέση εισροή λυμάτων στη συγκεκριμένη μονάδα επεξεργασίας αστικών λυμάτων στην Ισπανία κατά τη διάρκεια των 509 μετρήσεων της μελέτης μας ήταν περίπου 37227 χιλιάδες κυβικά μέτρα ανά οκτάωρο, ενώ τις μισές φορές ο όγκος ήταν μικρότερος ή ίσος με 35990 χιλιάδες κυβικά μέτρα ανά οκτάωρο.

Η εισροή λυμάτων στη μονάδα κατά τη διάρκεια της μελέτης κυμάνθηκε από το ελάχιστο των 10050 μέχρι το μέγιστο των 60081 χιλιάδων κυβικών μέτρων ανά οκτάωρο, υποδεικνύοντας μια έντονη μεταβλητότητα στη λειτουργία της μονάδας. Η μεταβλητότητα αυτή είναι φανερή και από την τιμή του συντελεστή μεταβλητότητας (που προκύπτει ως το πηλίκο $s/|\bar{x}|$). Πιο συγκεκριμένα, ο συντελεστής μεταβλητότητας ισούται με 17.83% υποδηλώνοντας μια έντονη ετερογένεια στο δείγμα μας ($cv > 10\%$).

Παρότι τα δεδομένα παρουσιάζουν έντονη ετερογένεια και μεγάλο εύρος τιμών ($R = x_{(509)} - x_{(1)} = 60081 - 10050 = 50031$) παρατηρούμε ότι το κεντρικό 50% των παρατηρήσεών μας (δηλαδή οι παρατηρήσεις μεταξύ

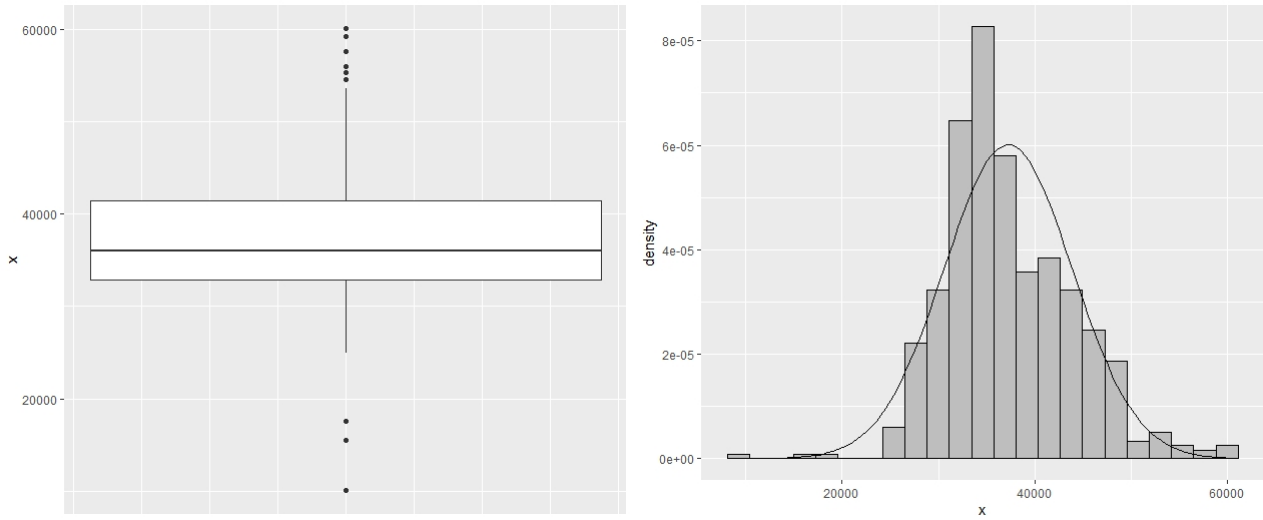
Πίνακας 8.1: Περιγραφικά στατιστικά μέτρα για τα δεδομένα του όγκου της εισροής λυμάτων (σε χιλιάδες κυβικά μέτρα ανά οκτάωρο).

---

n	mean	sd	var	cv
509	37226.5678	6635.9998	44036493.4270	0.1783

min	25%	50%	75%	max
10050	32888	35990	41372	60081

skewness	kurtosis
0.4694	0.9905



Σχήμα 8.5: Το θηκόγραμμα (αριστερό γράφημα) και το ιστόγραμμα (δεξιό γράφημα) για τις μετρήσεις της εισροής λυμάτων (σε χιλιάδες κυβικά μέτρα ανά οκτάωρο) σε μια μονάδα επεξεργασίας αστικών λυμάτων στην Ισπανία.

του 1ου και του 3ου τεταρτημορίου) συγκεντρώνονται σε ένα μικρό εύρος τιμών ($IQR = 41372 - 32888 = 8484$). Το γεγονός αυτό υποδηλώνει ότι τουλάχιστον τις μισές μέρες το εργοστάσιο λειτουργεί με σχετικά σταθερούς ρυθμούς, αν και κάποιες μέρες παρατηρούνται είτε πολύ μεγάλες είτε πολύ μικρές τιμές.

Καθώς ο συντελεστής λοξότητας είναι θετικός (skewness=0.4694), η κατανομή των παρατηρήσεών μας παρουσιάζει μια λοξότητα προς τα δεξιά. Αυτό συνεπάγεται ότι παρατηρούνται πιο πολλές μεγάλες ακραίες τιμές από ότι μικρές. Αυτό επαληθεύεται και από το θηκόγραμμα του Σχήματος 8.5, στο οποίο πέρα από τη θετική ασυμμετρία των παρατηρήσεών μας παρατηρούμε και ότι οι μεγάλες ακραίες τιμές είναι πολύ περισσότερες από τις μικρές ακραίες τιμές.

Τη δεξιά ασυμμετρία των μετρήσεών μας τη διαπιστώνουμε και από το ιστόγραμμα, το οποίο παρουσιάζει μια σχετικά έντονη ασυμμετρία προς τα δεξιά. Από το ιστόγραμμα μπορούμε να υπολογίσουμε προσεγγιστικά και την κορυφή, η οποία είναι μικρότερη από 35000 χιλιάδες κυβικά μέτρα ανά οκτάωρο. Το γεγονός ότι

$$m_0 < m < \bar{x}$$

αποτελεί μια ακόμα ένδειξη για το γεγονός ότι τα δεδομένα μας παρουσιάζουν μια δεξιά ασυμμετρία.

Από το ιστόγραμμα (δεξιό γράφημα Σχήματος 8.5) παρατηρούμε ότι τα δεδομένα μας είναι πιο «αιχμηρά» σε σχέση με την κανονική κατανομή, της οποίας η συνάρτηση πυκνότητας πιθανότητας έχει τοποθετηθεί πάνω στο γράφημα, υποδηλώνοντας ότι η κατανομή της οκτάωρης εισροής λυμάτων είναι λεπτόκυρτη. Το γεγονός αυτό επιβεβαιώνεται και από την τιμή του συντελεστή κύρτωσης (kurtosis), ο οποίος ισούται με 3+0.9905=3.9905[3]. Αυτό σημαίνει ότι γύρω από την κορυφή συγκεντρώνεται μεγάλο πλήθος παρατηρήσεων, το οποίο αντικατοπτρίζει το γεγονός ότι παρόλη την έντονη ετερογένεια των δεδομένων μας το «κέντρο» της κατανομής παρουσιάζει λιγότερο έντονες διαφορές.

Παρατήρηση 8.9

Στη συνέχεια, παρουσιάζονται οι εντολές στην R που δίνουν τα αποτελέσματα και τις γραφικές παραστάσεις αυτής της ενότητας. Το αρχείο των δεδομένων μπορεί να ανακτηθεί από το https://github.com/peconom/DataAnalysis/blob/main/SPAIN.TXT

```
1  library(ggplot2)
2  library(gridExtra)
3  library(e1071)
4  DesFUN = function(x,probs) {
5    m = mean(x,na.rm = TRUE)
6    s = sd(x,na.rm = TRUE)
7    v = var(x,na.rm = TRUE)
8    q = quantile(x,probs=probs,na.rm=TRUE)
9    results<- c(n = length(which(!is.na(x))),m,s,v,cv=s/abs(m),min(x,na.rm =
       TRUE),q,max(x,na.rm = TRUE),skewness(x,na.rm=TRUE),kurtosis(x,na.rm=
       TRUE))
10   names(results)<-c("n","mean","sd","var","cv","min",names(q),"max","
       skewness","kurtosis")
11   return(results)
12 }
13 options("scipen"=100, "digits"=4)
14
15 x<-read.table("SPAIN.TXT", quote="\"", comment.char=""))
16
17 n<-length(x)
18 DesFUN(x,c(0.25,0.5,0.75))
19
20 colnames(x)<-"x"
21 df<-data.frame(x)
22 p <- list()
23 p[[1]]<-ggplot(df, aes(x=x)) + geom_boxplot()+ coord_flip()+
24   theme(axis.title.x=element_blank(),
25         axis.text.x=element_blank(),
26            axis.ticks.x=element_blank())
27
28 d=(max(x)-min(x))/(1+3.322*log(n))
29 p[[2]] <- ggplot(df, aes(x=x)) +
30   geom_histogram(aes(y=..density..),binwidth = d, fill = "grey", color = "
       black")
31
32 grid.arrange(p[[1]],p[[2]], nrow=1, ncol=2)
```

[3] Σημειώνεται ότι η R τυπώνει την τιμή $\alpha_4 - 3$ για τον συντελεστή κύρτωσης.

Άσκηση Αυτοαξιολόγησης 8.3

Στο αρχείο δεδομένων faithful του πακέτου datasets της R εμφανίζονται 272 μετρήσεις του χρόνου σε λεπτά μεταξύ διαδοχικών εκτοξεύσεων θερμού νερού και της διάρκειας σε λεπτά της κάθε έκρηξης του θερμοπίδακα Old Faithful του Εθνικού Πάρκου του Yellowstone στις ΗΠΑ. Να υπολογίσετε με τη βοήθεια της R κατάλληλα περιγραφικά στατιστικά μέτρα, κατάλληλες γραφικές παραστάσεις και να σχολιάσετε τα αποτελέσματα για τη διάρκεια σε λεπτά της κάθε έκρηξης.

Σημειώνεται ότι τα δεδομένα που αφορούν τη διάρκεια σε λεπτά κάθε έκρηξης δίνονται στην πρώτη στήλη του αρχείου δεδομένων faithful.

8.5 Ασκήσεις

Άσκηση 8.1 *Να κατασκευάσετε τον πίνακα συχνοτήτων για τα παρακάτω δεδομένα.*

1	*1*	*0*	*-1*	*0*	*3*	*3*	*2*	*2*	*2*
2	*1*	*0*	*1*	*0*	*-1*	*0*	*3*	*1*	*-1*
-2	*-1*	*1*	*3*	*3*	*-1*	*1*	*3*	*3*	*3*

Στη συνέχεια, να βρεθούν η μέση τιμή, η διάμεσος, τα τεταρτημόρια, η επικρατούσα τιμή, το εύρος και ο συντελεστής μεταβλητότητας.

Άσκηση 8.2 *Να βρεθούν η μέση τιμή, η διάμεσος, τα τεταρτημόρια, το εύρος και ο συντελεστής μεταβλητότητας για τα παρακάτω δεδομένα.*

3.1	*3.1*	*1.1*	*0.3*	*7.3*	*1.3*	*0.8*	*6.0*	*4.0*	*0.1*
8.0	*3.6*	*1.8*	*1.1*	*7.7*	*3.3*	*7.4*	*4.4*	*0.9*	

Στη συνέχεια, ομαδοποιήστε τα παραπάνω δεδομένα χρησιμοποιώντας τον τύπο του Sturges για να προσδιορίσετε τον βέλτιστο αριθμό ομάδων και προσδιορίστε εκ νέου τη μέση τιμή και τη διάμεσο. Χρησιμοποιώντας τα ομαδοποιημένα δεδομένα, υπολογίστε την κορυφή.

Άσκηση 8.3 *Να υπολογίσετε με τη βοήθεια της R κατάλληλα περιγραφικά στατιστικά μέτρα, κατάλληλες γραφικές παραστάσεις και να σχολιάσετε τα αποτελέσματα για καθεμία από τις τέσσερις συνεχείς μεταβλητές που βρίσκονται στο σύνολο δεδομένων iris στο πακέτο datasets της R.*

Άσκηση 8.4 *Στο σύνολο δεδομένων iris στο πακέτο datasets της R εμφανίζεται και μια ποιοτική μεταβλητή όπου καταγράφεται το είδος των λουλουδιών. Κατασκευάστε τον πίνακα συχνοτήτων για τις παρατηρήσεις αυτές. Ποια είναι η επικρατούσα ομάδα;*

Άσκηση 8.5 *Να υπολογίσετε με τη βοήθεια της R κατάλληλα περιγραφικά στατιστικά μέτρα, κατάλληλες γραφικές παραστάσεις και να σχολιάσετε τα αποτελέσματα για τον χρόνο σε λεπτά μεταξύ διαδοχικών εκτοξεύσεων θερμού νερού του θερμοπίδακα Old Faithful του Εθνικού Πάρκου του Yellowstone στις ΗΠΑ.*

Σημειώνεται ότι τα δεδομένα που αφορούν τη διάρκεια σε λεπτά κάθε έκρηξης δίνονται στη δεύτερη στήλη του αρχείου δεδομένων faithful.

8.6 Απαντήσεις στις ασκήσεις αυτοαξιολόγησης

Λύση Άσκησης Αυτοαξιολόγησης 8.1

Η δειγματική μέση τιμή των δεδομένων (βλ. τη σχέση (8.1)) ισούται με:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{1.10 + 2.12 + ... + 0.22}{20} = \frac{65.88}{20} = 3.294.$$

Καθώς το μέγεθος του δείγματος είναι $n = 20$, δηλαδή άρτιος αριθμός, η διάμεσος ισούται με το ημιάθροισμα της $10^{\text{ης}}$ $(n/2)$ και της $11^{\text{ης}}$ $(n/2 + 1)$ διατεταγμένης παρατήρησης. Διατάσσοντας τις παρατηρήσεις κατά αύξουσα σειρά μεγέθους, προκύπτει ότι η διάμεσος ισούται με 2.97.
Το 3ο τεταρτημόριο υπολογίζεται από τη σχέση (8.4) για $p = 0.75$ και $n = 20$, ως ακολούθως :

$$\begin{aligned} q_{0.75} &= x_{(\lfloor (20-1)0.75+1 \rfloor)} + \gamma(x_{(\lfloor (20-1)0.75+1 \rfloor+1)} - x_{(\lfloor (20-1)0.75+1 \rfloor)}) \\ &= x_{(15)} + 0.25(x_{(16)} - x_{(15)}) = 4.12 + 0.25(4.21 - 4.12) = 4.1425, \end{aligned}$$

όπου, από τη σχέση (8.5), είναι $\gamma = (20-1)0.75 + 1 - \lfloor (20-1)0.75 + 1 \rfloor = 15.25 - 15 = 0.25$.
Τα παραπάνω μέτρα μπορούν να υπολογιστούν και από τον ομαδοποιημένο πίνακα συχνοτήτων σύμφωνα με την ομαδοποίηση που μας δίνεται στην εκφώνηση, δηλαδή μη θεωρώντας τα πρωτογενή δεδομένα, αλλά τον πίνακα που ακολουθεί.

Αύξων αριθμός ομάδας	Όρια ομάδας	$\nu_{(i)}$	$f_{(i)}$	$f_{(i)}\%$	$N_{(i)}$	$F_{(i)}$	$F_{(i)}\%$
1	[0-1)	2	0.1	10%	2	0.1	10%
2	[1-2)	4	0.2	20%	6	0.3	40%
3	[2-3)	5	0.25	25%	11	0.55	55%
4	[3-4)	2	0.1	10%	13	0.65	65%
5	[4-5)	4	0.2	20%	17	0.85	85%
6	[5-6)	2	0.1	10%	19	0.95	95%
7	[6-5)	0	0	0%	19	0.95	95%
...							
14	[13-14)	1	0.05	5%	20	1	100%
	Σύνολο	20	1	100%			

Παρατηρώντας ότι οι μέσες τιμές των ομάδων είναι ίσες με 0.5, 1.5, …, η δειγματική μέση τιμή υπολογίζεται μέσω της σχέσης (8.2), δηλαδή

$$\bar{x} = \frac{0.5 \cdot 2 + 1.5 \cdot 4 + 2.5 \cdot 5 + 3.5 \cdot 2 + 4.5 \cdot 4 + 5.5 \cdot 2 + 13.5 \cdot 1}{20} = \frac{69}{20} = 3.45.$$

Η διάμεσος ανήκει στην ομάδα $[2-3)$, ενώ το 3ο τεταρτημόριο ανήκει στην ομάδα $[4-5)$ (γιατί;) και υπολογίζονται από τη σχέση (8.6) για $p = 0.5$ και $p = 0.75$, αντίστοιχα. Επομένως είναι:

$$q_{0.5} = L_i + d \cdot \frac{0.5 \cdot n - N_{i-1}}{\nu_i} = 2 + 1 \cdot \frac{0.5 \cdot 20 - 6}{5} = 2 + 0.8 = 2.8,$$

και

$$q_{0.75} = L_i + d \cdot \frac{0.75 \cdot n - N_{i-1}}{\nu_i} = 4 + 1 \cdot \frac{0.75 \cdot 20 - 13}{4} = 4 + 0.5 = 4.5.$$

Λύση Άσκησης Αυτοαξιολόγησης 8.2

Το εύρος των δεδομένων της Άσκησης Αυτοαξιολόγησης 8.1 ισούται με

$$R = x_{(n)} - x_{(1)} = 13.2 - 0.22 = 29.$$

Για τον υπολογισμό του συντελεστή μεταβλητότητας απαραίτητος είναι ο υπολογισμός της μέσης τιμής, που την έχουμε προσδιορίσει στη λύση της Άσκησης Αυτοαξιολόγησης 8.1 ότι ισούται με 3.294 και ο υπολογισμός της δειγματικής τυπικής απόκλισης, η οποία ισούται με (βλ. τη σχέση (8.8))

$$s = \sqrt{s^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} = \sqrt{\frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right)} = \ldots = 2.741428.$$

Επομένως, ο συντελεστής μεταβλητότητας ισούται με:

$$cv = \frac{2.741428}{|3.294|} = 0.8322488.$$

Η παραπάνω τιμή είναι μεγαλύτερη του 0.1, οπότε το δείγμα μπορεί να χαρακτηριστεί ως ανομοιογενές. Χρησιμοποιώντας τα ομαδοποιημένα δεδομένα της άσκησης, η δειγματική τυπική απόκλιση υπολογίζεται ως ακολούθως (βλ. τη σχέση (8.3)):

$$\begin{aligned} s = \sqrt{s^2} &= \sqrt{\frac{1}{n-1}\sum_{\ell=1}^{k}\nu_\ell (k_\ell - \bar{x})^2} \\ &= \sqrt{\frac{1}{20-1}\left(2\cdot(0.5-3.45)^2 + \ldots + 1\cdot(13.5-3.45)^2\right)} \\ &= \sqrt{\frac{150.95)}{20-1}} = 2.818641. \end{aligned}$$

Επομένως, ο συντελεστής μεταβλητότητας με βάση τα ομαδοποιημένα δεδομένα, ισούται με:

$$cv = \frac{2.818641}{|3.45|} = 0.8169974,$$

δηλαδή η τιμή του συντελεστή μεταβλητότητας είναι πάλι μεγαλύτερη του 0.1, οπότε το δείγμα μπορεί πάλι να χαρακτηριστεί ως ανομοιογενές.

Λύση Άσκησης Αυτοαξιολόγησης 8.3

Εκτελώντας στην R τις ίδιες εντολές με αυτές που εμφανίζονται στην Παρατήρηση 8.9, αλλά έχοντας ως x αυτό που προκύπτει σύμφωνα με την παρακάτω εντολή

```
1 x<-faithful[,1]
```

λαμβάνουμε τα ακόλουθα αποτελέσματα και τις ακόλουθες γραφικές παραστάσεις.

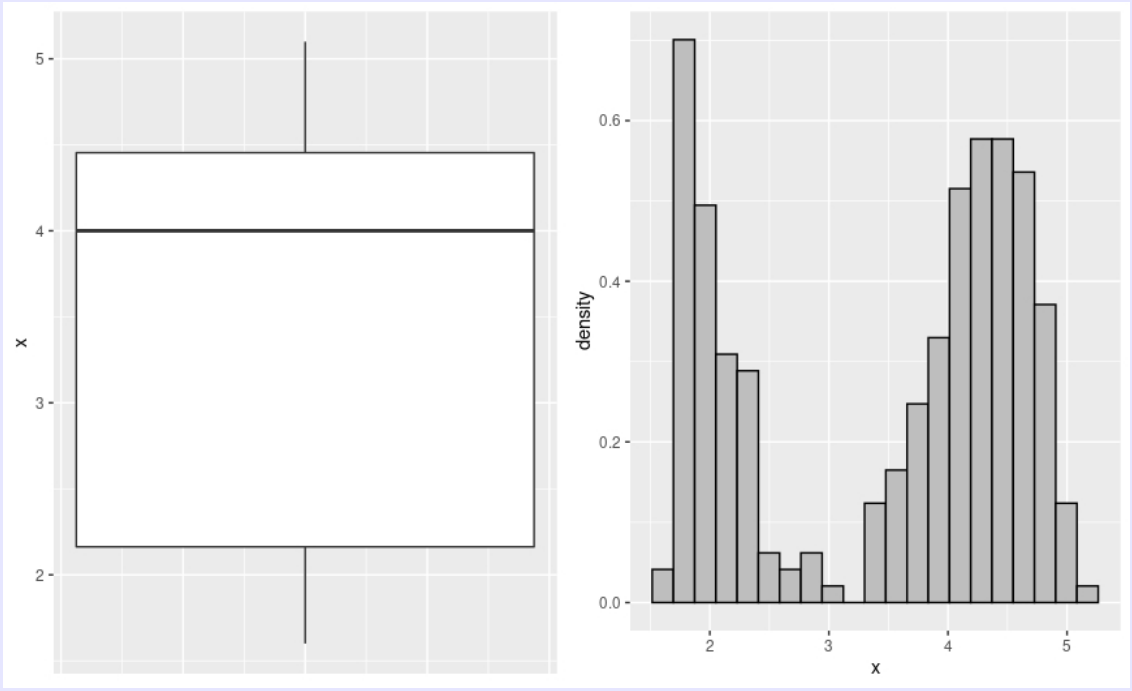
```
------------------------------------------------------------------------------

       n     mean       sd      var
272.0000   3.4878   1.1414   1.3027

      cv      min      25%      50%
  0.3272   1.6000   2.1627   4.0000

     75%      max skewness kurtosis
  4.4543   5.1000  -0.4135  -1.5116

------------------------------------------------------------------------------
```



Από τα παραπάνω είναι φανερό ότι η διάρκεια σε λεπτά κάθε έκρηξης θερμού νερού του θερμοπίδακα Old Faithful του Εθνικού Πάρκου του Yellowstone στις ΗΠΑ.

- έχει μέση διάρκεια (μέση τιμή) ίση με 3.4878 λεπτά,
- έχει τυπική απόκλιση 1.1414 λεπτά,
- παρουσιάζει ανομοιογένεια ($cv = 0.3272$)),
- έχει εύρος τιμών 3.5 λεπτά ($R = 5.1 - 1.6 = 3.5$),
- παρουσιάζει αρνητική ασυμμετρία (-0.4135), ενώ
- είναι και πλατύκυρτη (ο συντελεστής κύρτωσης ισούται με -1.5116+3).

Από το θηκόγραμμα παρατηρούμε ότι δεν υπάρχουν ακραίες τιμές, ενώ είναι φανερή και η αρνητική ασυμμετρία της κατανομής των δεδομένων. Από την άλλη, από το ιστόγραμμα παρατηρούμε ότι τα δεδομένα είναι δικόρυφα. Μάλιστα, φαίνεται να ορίζονται δύο καλά διαχωρισμένες ομάδες παρατηρήσεων.

ΒΙΒΛΙΟΓΡΑΦΙΑ

Ελληνόγλωσση

Ζωγράφος, Κ. (2002). *Μαθήματα Πιθανοτήτων και Στατιστικής*. Τυπογραφείο Πανεπιστημίου Ιωαννίνων.

Ξενόγλωσση

Hyndman, R. J. and Fan, Y. (1996). Sample Quantiles in Statistical Packages. *The American Statistician*, *50*, pp. 361–365.

ΚΕΦΑΛΑΙΟ 9

ΔΕΙΓΜΑΤΟΛΗΠΤΙΚΕΣ ΚΑΤΑΝΟΜΕΣ

Σύνοψη

Σε αυτό το κεφάλαιο εστιάζουμε στη δειγματοληψία από πληθυσμούς με συγκεκριμένα χαρακτηριστικά και μελετούμε σημαντικές ποσότητες, όπως είναι η δειγματική μέση τιμή, η δειγματική διασπορά και συναρτήσεις αυτών, οι οποίες παίζουν καθοριστικό ρόλο στη στατιστική συμπερασματολογία η οποία θα παρουσιαστεί εκτενώς στα επόμενα κεφάλαια.

Προαπαιτούμενη γνώση: Κεφάλαια 4, 5 και 7 του παρόντος συγγράμματος.

Προσδοκώμενα μαθησιακά αποτελέσματα: Όταν θα έχετε ολοκληρώσει τη μελέτη του κεφαλαίου αυτού, θα γνωρίζετε

- την κατανομή του δειγματικού μέσου,
- την κατανομή της δειγματικής διασποράς,
- την κατανομή της δειγματικής αναλογίας,
- την κατανομή της διαφοράς των δειγματικών μέσων τιμών δύο ανεξάρτητων/εξαρτημένων πληθυσμών,
- την κατανομή της διαφοράς των δειγματικών αναλογιών δύο ανεξάρτητων πληθυσμών και
- την κατανομή του λόγου των δειγματικών διασπορών.

Γλωσσάριο επιστημονικών όρων

- Δειγματοληπτικές κατανομές
- Κατανομή δειγματικής αναλογίας
- Κατανομή δειγματικής διασποράς
- Κατανομή δειγματικής μέσης τιμής
- Στατιστικές συναρτήσεις
- Τυχαίο δείγμα

Οικονόμου, Π., Μαλεφάκη, Σ., & Μπατσίδης, Α. (2023). *Πιθανότητες - Στατιστική.*

9.1 Στατιστικές συναρτήσεις

Συχνά χρειάζεται να μελετήσουμε και να εξάγουμε συμπεράσματα για κάποιο χαρακτηριστικό ενός ή περισσότερων πληθυσμών που μας ενδιαφέρουν. Καθώς, όπως αναφέρθηκε στο Κεφάλαιο 8, η μελέτη ολόκληρου του πληθυσμού τις περισσότερες φορές είναι πολύ δύσκολη, αν όχι αδύνατη, καλούμαστε να εξάγουμε συμπεράσματα για τον πληθυσμό που μας ενδιαφέρει στηριζόμενοι σε ένα υποσύνολό του, που καλείται δείγμα. Η εξαγωγή αυτών των συμπερασμάτων, όπως αναλυτικά θα δούμε στα επόμενα δύο κεφάλαια του παρόντος συγγράμματος, βασίζεται σε κατάλληλες συναρτήσεις ενός τυχαίου δείγματος $(X_1, X_2, \dots, X_n)$[1] του πληθυσμού. Μια τέτοια συνάρτηση $h(X_1, \dots, X_n)$ ονομάζεται **στατιστική συνάρτηση (σσ)**. Σύμφωνα με τον παραπάνω ορισμό, παραδείγματα στατιστικών συναρτήσεων είναι η **δειγματική μέση τιμή**

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n},$$

και η **δειγματική διασπορά (διακύμανση)**

$$S^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1},$$

πάνω στις οποίες βασίζεται όλη η στατιστική συμπερασματολογία για τη μέση τιμή μ και τη διασπορά σ^2 ενός πληθυσμού.

Παρατήρηση 9.1

Στο σημείο αυτό θα πρέπει να επισημάνουμε ότι η σσ $h(X_1, X_2, \dots, X_n)$ είναι τυχαία μεταβλητή, αφού είναι συνάρτηση των τ.μ. $(X_1, X_2, \dots, X_n)$. Στην περίπτωση που θα παρατηρηθεί ένα δείγμα $(x_1, x_2, \dots, x_n)$ από τον συγκεκριμένο πληθυσμό, η σσ $h(X_1, X_2, \dots, X_n)$ παίρνει την τιμή $h(x_1, x_2, \dots, x_n)$. Για παράδειγμα, οι σσ $\bar{X}$ και S^2 όταν παρατηρηθεί ένα δείγμα παίρνουν τις τιμές $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ και $s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$, αντίστοιχα.

Οι στατιστικές συναρτήσεις ως τ.μ. έχουν μία κατανομή την οποία ονομάζουμε **δειγματοληπτική κατανομή**. Σκοπός του κεφαλαίου αυτού είναι να παρουσιαστούν αποτελέσματα που αφορούν τις δειγματοληπτικές κατανομές βασικών στατιστικών συναρτήσεων, όπως η δειγματική μέση τιμή, η δειγματική διασπορά και άλλες, οι οποίες αποτελούν τη βάση για την εξαγωγή συμπερασμάτων για τον υπό μελέτη πληθυσμό ή τους υπό μελέτη πληθυσμούς.

Παρατήρηση 9.2

Η δειγματική κατανομή των $\bar{X}$ και S^2 μπορεί να αποτυπωθεί, παραδείγματος χάριν, με τη βοήθεια ενός ιστογράμματος αν πολλά δείγματα ίδιου μεγέθους n από έναν συγκεκριμένο πληθυσμό είναι διαθέσιμα. Ειδικότερα, η δειγματική κατανομή περιγράφει τη μεταβλητότητα του δειγματικού μέσου $\bar{X}$ και της δειγματικής διασποράς S^2 γύρω από το μέσο μ και τη διασπορά σ^2 του πληθυσμού, αντίστοιχα.

9.2 Δειγματική μέση τιμή και διασπορά

Στην ενότητα αυτή θα δοθούν τα κυριότερα αποτελέσματα που αφορούν τη δειγματική μέση τιμή και τη δειγματική διασπορά.

[1] Αν οι τυχαίες μεταβλητές $(X_1, X_2, \dots, X_n)$ είναι ανεξάρτητες και ισόνομες, έχουν δηλαδή όλες την ίδια κατανομή, τότε λέμε ότι αποτελούν ένα τυχαίο δείγμα από τη συγκεκριμένη κατανομή.

Πρόταση 9.1

Έστω $(X_1, X_2, \dots, X_n)$ ένα τυχαίο δείγμα από έναν πληθυσμό με πεπερασμένη μέση τιμή μ και διασπορά σ^2. Είναι τότε

$$E(\bar{X}) = \mu \text{ και } Var(\bar{X}) = \frac{\sigma^2}{n}.$$

Απόδειξη Πρότασης 9.1

Από τον ορισμό της δειγματικής μέσης τιμής και τις ιδιότητες της μέσης τιμής έχουμε:

$$E(\bar{X}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n}\sum_{i=1}^{n} \mu = \mu.$$

Επιπρόσθετα, από τις ιδιότητες της διακύμανσης, έχουμε:

$$Var(\bar{X}) = Var\left(\frac{\sum_{i=1}^{n} X_i}{n}\right) = \frac{\sum_{i=1}^{n} Var(X_i)}{n^2} = \frac{\sum_{i=1}^{n} \sigma^2}{n^2} = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

Με βάση τα παραπάνω, μπορούμε να πούμε ότι η μέση τιμή της δειγματικής μέσης τιμής είναι ίδια με τη μέση τιμή του πληθυσμού, ενώ η διασπορά της είναι μικρότερη από τη διασπορά του πληθυσμού και μάλιστα η μείωση είναι ανάλογη με το μέγεθος του δείγματος.

Από την Πρόταση 9.1 προκύπτει ότι η τυπική απόκλιση του δειγματικού μέσου $\bar{X}$ ισούται με $\frac{\sigma}{\sqrt{n}}$. Η τυπική απόκλιση του δειγματικού μέσου λέγεται **τυπικό σφάλμα** του δειγματικού μέσου.

Όσον αφορά την κατανομή της δειγματικής μέσης τιμής ισχύουν τα ακόλουθα.

Πρόταση 9.2

Έστω $X_1, \dots, X_n$, n το πλήθος ανεξάρτητες και ισόνομες τυχαίες μεταβλητές με πεπερασμένη μέση τιμή μ και διασπορά σ^2.

α) Αν ο πληθυσμός είναι κανονικός, τότε

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

β) Αν ο πληθυσμός δεν είναι κανονικός, τότε για μεγάλο μέγεθος δείγματος

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} N(0,1),$$

ή, ισοδύναμα,

$$Z = \frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}} \xrightarrow{d} N(0,1).$$

Απόδειξη Πρότασης 9.2

Παραπέμπουμε τον αναγνώστη στην Πρόταση 7.12 και στο Θεώρημα 7.2, αντίστοιχα.

Παράδειγμα 9.1

Ρίχνουμε ένα ζάρι 100 φορές και καταγράφουμε το άθροισμα των ρίψεων. Να προσδιοριστεί κατά προσέγγιση η πιθανότητα το άθροισμα των ρίψεων του ζαριού να ξεπερνάει το 400.

Λύση Παραδείγματος 9.1

Έστω X_i, $i = 1, \dots, 100$, το αποτέλεσμα τις $i-$οστής ρίψης του ζαριού και έστω $Y = \sum_{i=1}^{100} X_i$ το άθροισμα των 100 ρίψεων του ζαριού. Θέλουμε να υπολογίσουμε, κατά προσέγγιση, την πιθανότητα $P(Y > 400)$. Για να το κάνουμε αυτό, χρειάζεται να βρούμε την προσεγγιστική κατανομή του Y. Είναι

$$E(X_i) = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} = 3.5$$

και

$$E(X_i^2) = 1^2 \cdot \frac{1}{6} + 2^2 \cdot \frac{1}{6} + 3^2 \cdot \frac{1}{6} + 4^2 \cdot \frac{1}{6} + 5^2 \cdot \frac{1}{6} + 6^2 \cdot \frac{1}{6} = 15.1667$$

οπότε

$$Var(X_i) = E(X_i^2) - (E(X_i))^2 = 15.1667 - 3.5^2 = 2.9187.$$

Οπότε $E(Y) = 100 \cdot 3.5 = 350$ και $Var(Y) = 100 \cdot 2.9187 = 291.87$, άρα

$$Y \xrightarrow{d} N(350,\ 291.87).$$

Η ζητούμενη πιθανότητα εφαρμόζοντας διόρθωση συνέχειας ισούται με

$$\begin{aligned} P(Y > 400) = P(Y \geq 400.5) &= P\left(\frac{Y - 350}{\sqrt{291.87}} \geq \frac{400.5 - 350}{\sqrt{291.87}}\right) \\ &\approx P(Z \geq 2.96) = 1 - P(Z < 2.96) \\ &= 1 - 0.99846 = 0.00154, \end{aligned}$$

όπου $Z \sim N(0,1)$ και χρησιμοποιήθηκε ο Πίνακας Α΄.3 της τυπικής κανονικής κατανομής του Παραρτήματος Α΄.

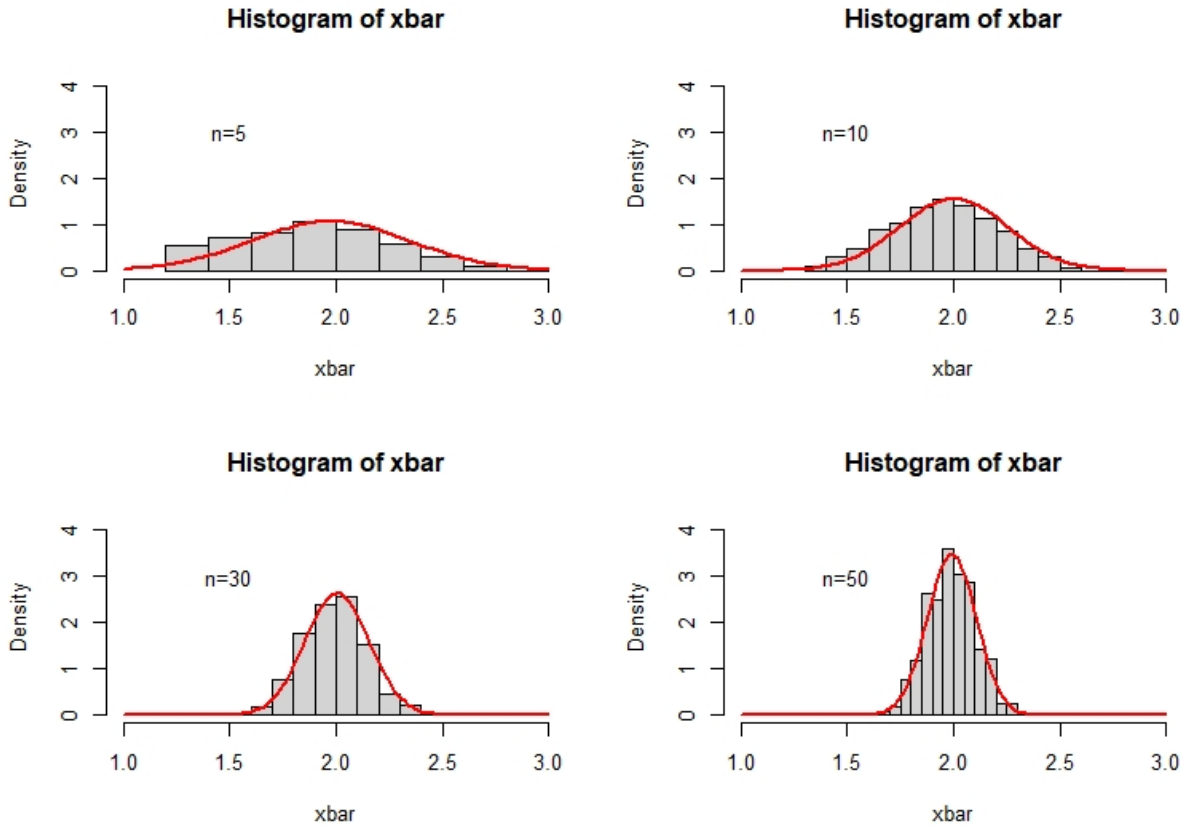
Άσκηση Αυτοαξιολόγησης 9.1

Έστω $X_1, X_2, \dots, X_n$, n το πλήθος ανεξάρτητες και ισόνομες τυχαίες μεταβλητές από μια κατανομή με πεπερασμένη μέση τιμή μ και διασπορά σ^2. Ποια από τα παρακάτω ΔΕΝ είναι ιδιότητα του δειγματικού μέσου $\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$;

1. Η διασπορά του δειγματικού μέσου μικραίνει καθώς αυξάνεται το μέγεθος του δείγματος n.
2. Η μέση τιμή του δειγματικού μέσου ισούται με μ.
3. Ο δειγματικός μέσος $\bar{X}$ ισούται πάντα με μ.
4. Η διασπορά του δειγματικού μέσου είναι πάντα μικρότερη ή ίση με τη διασπορά του πληθυσμού.

Παρατηρούμε, λοιπόν, ότι, όταν, το μέγεθος του δείγματος n είναι μεγάλο, ισχύει το ΚΟΘ και όποια και αν είναι η κατανομή του τυχαίου δείγματος $(X_1, \dots, X_n)$, ακόμα και αν είναι διακριτή, ο δειγματικός μέσος $\bar{X}$ θα ακολουθεί προσεγγιστικά την κανονική κατανομή. Τι σημαίνει όμως στην πράξη μεγάλο μέγεθος δείγματος; Όπως έχει ήδη αναφερθεί στην Ενότητα 7.3 η ταχύτητα σύγκλισης εξαρτάται από τη λοξότητα της κοινής κατανομής που ακολουθούν οι τυχαίες μεταβλητές X_i, $i = 1, \dots, n$. Ειδικότερα, όταν η κατανομή είναι συμμετρική η κατανομή του $\bar{X}$ προσεγγίζεται ικανοποιητικά από την κανονική κατανομή ακόμα και για πολύ μικρά n (π.χ. $n = 4$ ή 5), ενώ σε περίπτωση που υπάρχει μεγάλη λοξότητα στην κοινή κατανομή των X_i, $i = 1, \dots, n$, το μέγεθος του δείγματος θα πρέπει να είναι $n \geq 30$ (Κουτρουβέλης, 2011) για να είναι ικανοποιητική η σύγκλιση της κατανομής του $\bar{X}$ στην κανονική κατανομή.

Προκειμένου, να διαπιστώσουμε τη σύγκλιση της κατανομής της δειγματικής μέσης τιμής στην κανονική κατανομή, με τη βοήθεια της R, προσομοιώνουμε 1000 τυχαία δείγματα μεγέθους $5, 10, 30$ και 50 από την κατανομή του Παραδείγματος 7.7. Στη συνέχεια για καθένα από αυτά τα δείγματα υπολογίζεται η δειγματική



Σχήμα 9.1: Ιστογράμματα της $\bar{X}$ για διάφορες τιμές του n και συμμετρική κατανομή για τον αρχικό πληθυσμό.

μέση τιμή και στο τέλος κατασκευάζονται τα αντίστοιχα ιστογράμματα για κάθε μέγεθος δείγματος, τα οποία παρατίθενται στο Σχήμα 9.1. Ο κώδικας της R που χρησιμοποιήθηκε για την κατασκευή των γραφικών παραστάσεων του Σχήματος 9.1 είναι ο ακόλουθος.

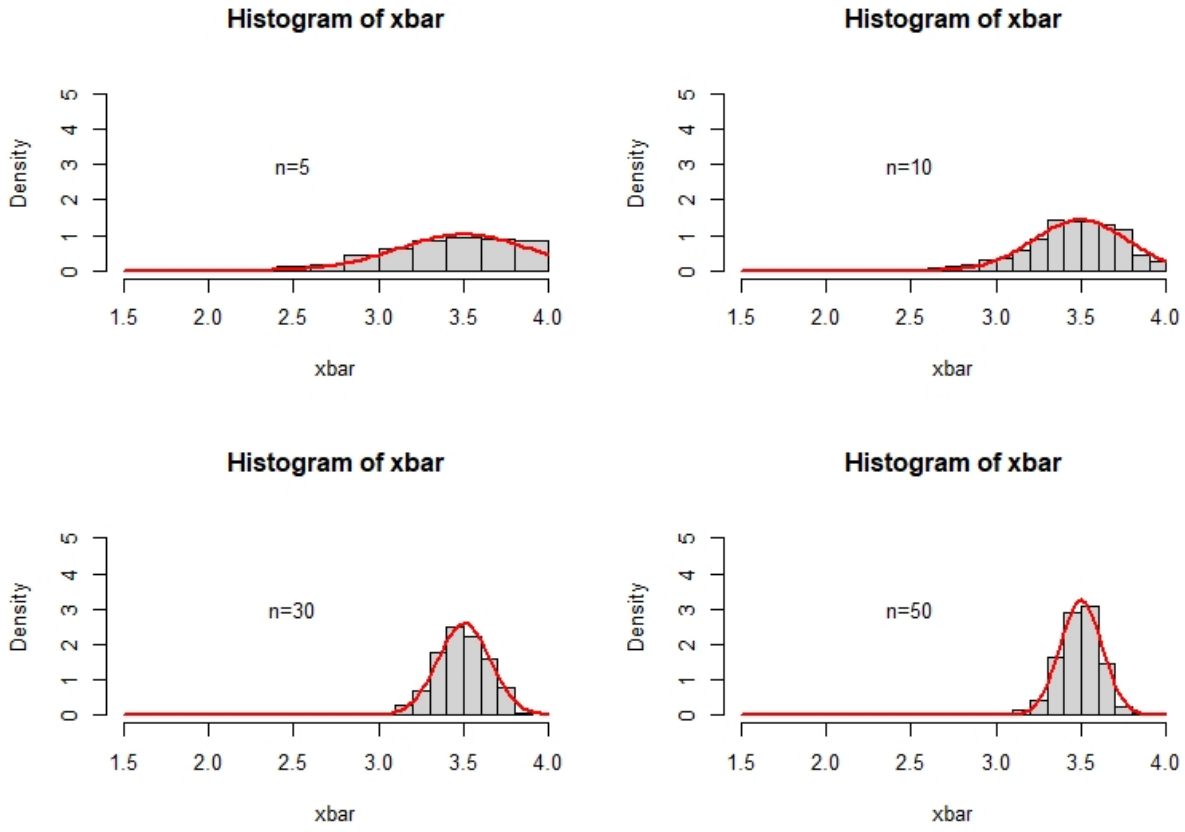
```
N <- 1000
nall <- c(5,10,30,50)

xbar <- numeric(N)

par(mfrow=c(2,2))

for(n in nall){
  for (j in 1:N){
    x <- sample(c(1,2,3),n,replace=TRUE,prob=c(1/3,1/3,1/3))
    xbar[j] <- mean(x)
  }
  hist(xbar,freq = F,ylim = c(0,4.2),xlim=c(1,3))
  curve(dnorm(x,mean = mean(xbar),sd = sd(xbar)),add = T,lwd=2,col="red")
  text(1.5, 3, paste("n=" ,n , sep=""))
}
```

Όπως παρατηρούμε στο Σχήμα 9.1 έχουμε μία αρκετά καλή προσέγγιση της κανονικής κατανομής ακόμα και για $n = 5$ (γράφημα πάνω αριστερά). Επομένως, επιβεβαιώνεται αυτό που προαναφέρθηκε για τη σύγκλιση στην περίπτωση συμμετρικών κατανομών. Επιπρόσθετα, παρατηρήστε ότι καθώς το n αυξάνεται, η διασπορά των παρατηρούμενων δειγματικών μέσων τιμών μικραίνει (οι τιμές ανήκουν σε ένα μικρότερο



Σχήμα 9.2: Ιστογράμματα της $\bar{X}$ για διάφορες τιμές του n και λοξή κατανομή για τον αρχικό πληθυσμό.

εύρος τιμών), το οποίο οφείλεται στο γεγονός ότι η διασπορά του δειγματικού μέσου είναι μια φθίνουσα συνάρτηση ως προς n.

Στη συνέχεια, για να δούμε την επίδραση της λοξότητας της αρχικής κατανομής στην ταχύτητα σύγκλισης της κατανομής της δειγματικής μέσης τιμής στην κανονική κατανομή, με τη βοήθεια της R, προσομοιώσαμε, με παρόμοιο τρόπο όπως προηγουμένως, 1000 τυχαία δείγματα μεγέθους $5, 10, 30$ και 50 από την κατανομή με συνάρτηση πιθανότητας που δίνεται από τη σχέση:

$$P(X = 1) = 0.05,\ P(X = 2) = 0.1,\ P(X = 3) = 0.15, \text{ και } P(X = 4) = 0.7.$$

Τα ιστογράμματα των δειγματικών μέσων τιμών για τις διάφορες τιμές του n των παραπάνω προσομοιώσεων απεικονίζονται στο Σχήμα 9.2. Παρατηρούμε ότι για $n = 5$ και 10 δεν μπορούμε να ισχυριστούμε ότι η κατανομή της δειγματικής μέσης τιμής είναι συμμετρική, ενώ για $n = 30$ και $n = 50$ είμαστε αρκετά κοντά στη συμμετρία.

Στην Πρόταση 9.2 προσδιορίστηκε η κατανομή της δειγματικής μέσης τιμής υπό την υπόθεση ότι μας είναι γνωστή η πληθυσμιακή διασπορά σ^2. Στην πράξη όμως, τις περισσότερες φορές, η διασπορά του πληθυσμού είναι άγνωστη, οπότε την αντικαθιστούμε με το δειγματικό της ανάλογο που είναι η δειγματική διασπορά S^2. Στη συνέχεια θα παρουσιαστούν αποτελέσματα που αφορούν τόσο την κατανομή της δειγματικής διασποράς όσο και την κατανομή της δειγματικής μέσης τιμής όταν η πληθυσμιακή διασπορά είναι άγνωστη.

Πρόταση 9.3

Έστω $(X_1, X_2, \ldots, X_n)$ ένα τυχαίο δείγμα από έναν πληθυσμό με πεπερασμένη μέση τιμή μ και διασπορά σ^2. Είναι τότε

$$E(S^2) = \sigma^2.$$

Απόδειξη Πρότασης 9.3

Από τον ορισμό της δειγματικής διακύμανσης και έπειτα από λίγη άλγεβρα, έχουμε ότι:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right).$$

Επιπλέον, είναι:

$$E(X_i^2) = Var(X_i) + (E(X_i))^2 = \sigma^2 + \mu^2$$

και

$$E(\bar{X}^2) = Var(\bar{X}) + (E(\bar{X}))^2 = Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) + \mu^2 = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) + \mu^2 = \frac{\sigma^2}{n} + \mu^2.$$

Συνδυάζοντας τα παραπάνω έχουμε:

$$\begin{aligned} E(S^2) &= \frac{1}{n-1}\left(\sum_{i=1}^{n}\left(\sigma^2 + \mu^2\right) - n\left(\frac{\sigma^2}{n} + \mu^2\right)\right) \\ &= \frac{1}{n-1}\left(n\sigma^2 + n\mu^2 - \sigma^2 - n\mu^2\right) = \sigma^2, \end{aligned}$$

που ολοκληρώνει την απόδειξη.

Παρατήρηση 9.3

Από την Πρόταση 9.3 προκύπτει ότι η ιδιότητα $E(S^2) = \sigma^2$ ισχύει χωρίς την υπόθεση της κανονικότητας του τυχαίου δείγματος, αρκεί το τυχαίο δείγμα να είναι από μια κατανομή με πεπερασμένη μέση τιμή και διακύμανση, μ και σ^2, αντίστοιχα.

Στην πρόταση που ακολουθεί προσδιορίζεται η κατανομή της δειγματικής διασποράς, υπό την υπόθεση ότι το τυχαίο δείγμα προέρχεται από κανονικό πληθυσμό, αλλά και η ανεξαρτησία των τυχαίων μεταβλητών $\bar{X}$ και S^2.

Πρόταση 9.4

Έστω $X_1, \ldots, X_n$ είναι n το πλήθος ανεξάρτητες και ισόνομες τυχαίες μεταβλητές με $X_i \sim N(\mu, \sigma^2)$, $i = 1, \ldots, n$, δηλαδή $(X_1, X_2, \ldots, X_n)$ είναι ένα τυχαίο δείγμα μεγέθους n από την προαναφερθείσα κατανομή. Τότε, ισχύουν τα ακόλουθα:

1. Η τυχαία μεταβλητή $U = \frac{(n-1)S^2}{\sigma^2}$ ακολουθεί χι-τετράγωνο κατανομή με $n-1$ βαθμούς ελευθερίας, δηλαδή
$$U = \frac{(n-1)S^2}{\sigma^2} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{\sigma^2} \sim \chi^2_{n-1}.$$
2. Οι τυχαίες μεταβλητές $\bar{X}$ και S^2 είναι ανεξάρτητες.

Απόδειξη Πρότασης 9.4

Η απόδειξη αυτών των αποτελεσμάτων ξεφεύγει από τους σκοπούς του παρόντος συγγράμματος και παραπέμπουμε, ενδεικτικά, στο σύγγραμμα Ηλιόπουλος (2006).

Πρόταση 9.5

Έστω $X_1, \ldots, X_n$ είναι n το πλήθος ανεξάρτητες και ισόνομες τυχαίες μεταβλητές με $X_i \sim N(\mu, \sigma^2)$, $i = 1, \ldots, n$, δηλαδή $(X_1, X_2, \ldots, X_n)$ είναι ένα τυχαίο δείγμα μεγέθους n από την προαναφερθείσα κατανομή. Τότε

$$S^2 \sim G\left(\frac{n-1}{2}, \frac{2\sigma^2}{n-1}\right),$$

με

$$E(S^2) = \sigma^2 \text{ και } Var(S^2) = \frac{2\sigma^4}{n-1}.$$

Απόδειξη Πρότασης 9.5

Η απόδειξη προκύπτει άμεσα χρησιμοποιώντας τις ιδιότητες της γάμμα κατανομής. Παρατηρήστε ότι $S^2 = \frac{U \cdot \sigma^2}{n-1}$ και, επομένως, η τ.μ. S^2 είναι συνάρτηση της τ.μ. U. Η κατανομή της μπορεί να προσδιοριστεί είτε με τη μέθοδο της αλλαγής μεταβλητών είτε με τη μέθοδο της ροπογεννήτριας. Ειδικότερα, με τη μέθοδο της ροπογεννήτριας έχουμε ότι:

$$M_{S^2}(t) = E(e^{tS^2}) = E(e^{\frac{\sigma^2 t}{n-1}U}) = M_U\left(\frac{\sigma^2 t}{n-1}\right).$$

Η ροπογεννήτρια της τυχαίας μεταβλητής $U \sim \chi^2_{n-1}$ δίνεται από τη σχέση (7.3) με την τροποποίηση ότι στη θέση του n βάζουμε το $n-1$. Έτσι προκύπτει ότι:

$$M_{S^2}(t) = \left(1 - 2\frac{\sigma^2 t}{n-1}\right)^{-\frac{n-1}{2}}, \quad \text{για } \frac{\sigma^2 t}{n-1} < 0.5,$$

ή, ισοδύναμα,

$$M_{S^2}(t) = \left(1 - \frac{t}{\lambda}\right)^{-\frac{n-1}{2}}, \quad \text{για } t < \frac{n-1}{2\sigma^2} = \lambda.$$

Συνδυάζοντας την παραπάνω σχέση με τη σχέση (5.32) προκύπτει ότι $S^2 \sim G\left(\frac{n-1}{2}, \frac{2\sigma^2}{n-1}\right)$. Τέλος, από τις ιδιότητες της μέσης τιμής και της διακύμανσης της γάμμα κατανομής έχουμε:

$$E(S^2) = \frac{n-1}{2} \cdot \frac{2\sigma^2}{n-1} = \sigma^2,$$

και

$$Var(S^2) = \frac{n-1}{2} \cdot \left(\frac{2\sigma^2}{n-1}\right)^2 = \frac{2\sigma^4}{n-1}.$$

Σχετικά με την προσεγγιστική κατανομή της δειγματικής διακύμανσης όταν έχουμε τυχαίο δείγμα από μη κανονικό πληθυσμό παραπέμπουμε, μεταξύ άλλων, στο σύγγραμμα του van der Vaart (1998).

Άσκηση Αυτοαξιολόγησης 9.2

Ποια από τις παρακάτω ΔΕΝ είναι ιδιότητα της δειγματικής διασποράς;

1. Η μέση τιμή της δειγματικής διασποράς ισούται με τη διασπορά του πληθυσμού.
2. Η δειγματική διασπορά ακολουθεί χι-τετράγωνο κατανομή με $n-1$ βαθμούς ελευθερίας.
3. Η διασπορά της δειγματικής διασποράς μειώνεται καθώς αυξάνεται το μέγεθος του δείγματος.
4. Όλες οι παραπάνω είναι ιδιότητες της δειγματικής διασποράς.

Στην Πρόταση 9.2 προσδιορίζεται η κατανομή της δειγματικής μέσης τιμής υπό την υπόθεση ότι η πληθυσμιακή διασπορά σ^2 είναι γνωστή. Ωστόσο, στην πράξη, τις περισσότερες φορές, η διασπορά του πληθυσμού είναι άγνωστη, οπότε την αντικαθιστούμε με τη δειγματική διασπορά S^2. Στην πραγματικότητα, στην τυποποιημένη έκφραση του δειγματικού μέσου αντικαθιστούμε μία σταθερά, το σ, με μία τυχαία μεταβλητή, το S, οπότε η κατανομή της νέας στατιστικής συνάρτησης $\frac{\bar{X}-\mu}{S/\sqrt{n}}$ αναμένεται να μην είναι η τυπική κανονική κατανομή. Μία απλοϊκή εξήγηση αυτού είναι ότι με αυτήν την αντικατάσταση του θετικού αριθμού σ με την τυχαία μεταβλητή S έχουμε εισάγει επιπλέον μεταβλητότητα. Σε αυτήν την περίπτωση ισχύει η πρόταση που ακολουθεί.

Πρόταση 9.6

Έστω $X_1,\dots,X_n$ είναι n το πλήθος ανεξάρτητες και ισόνομες τυχαίες μεταβλητές με $X_i \sim N(\mu,\sigma^2)$, $i=1,\dots,n$, δηλαδή $(X_1,X_2,\dots,X_n)$ είναι ένα τυχαίο δείγμα μεγέθους n από την προαναφερθείσα κατανομή. Τότε,

$$T=\frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t_{n-1},$$

όπου

$$\bar{X}=\frac{\sum_{i=1}^{n} X_i}{n} \text{ και } S^2=\frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{n-1}.$$

Απόδειξη Πρότασης 9.6

Παρατηρήστε ότι

$$T=\frac{\bar{X}-\mu}{S/\sqrt{n}}=\frac{\left(\bar{X}-\mu\right)/\sigma}{S/(\sigma\sqrt{n})}=\frac{\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}}{\sqrt{S^2/\sigma^2}}=\frac{Z}{\sqrt{V/(n-1)}}$$

όπου $Z=\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ και $V=\frac{(n-1)S^2}{\sigma^2}$. Όμως, έχουμε ότι $Z \sim N(0,1)$ και από την Πρόταση 9.4 ισχύει ότι $V \sim \chi^2_{n-1}$ και, επιπλέον, Z και V είναι ανεξάρτητες τυχαίες μεταβλητές ως συναρτήσεις των ανεξάρτητων τυχαίων μεταβλητών $\bar{X}$ και S^2. Με βάση τα παραπάνω μπορούμε να πούμε ότι η τ.μ. T ορίζεται ως ο λόγος δύο ανεξάρτητων τ.μ. οι οποίες ακολουθούν την κανονική και τη χ^2_{n-1} κατανομή και, εν τέλει, το ζητούμενο προκύπτει από τον ορισμό της t κατανομής.

Στην περίπτωση όπου ο πληθυσμός από τον οποίο προέρχεται το τυχαίο δείγμα δεν είναι κανονικός, αλλά το μέγεθος δείγματος είναι μεγάλο (συνήθως $n \geq 30$), τότε έχουμε το ακόλουθο αποτέλεσμα.

Πρόταση 9.7

Έστω $X_1,\dots,X_n$ είναι n το πλήθος ανεξάρτητες και ισόνομες τυχαίες μεταβλητές από έναν πληθυσμό με πεπερασμένη μέση τιμή και διακύμανση. Τότε, για μεγάλο μέγεθος δείγματος

$$Z=\frac{\bar{X}-\mu}{S/\sqrt{n}} \xrightarrow{d} N(0,1),$$

όπου

$$\bar{X}=\frac{\sum_{i=1}^{n} X_i}{n} \text{ και } S^2=\frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{n-1}.$$

Απόδειξη Πρότασης 9.7

Η απόδειξη ξεφεύγει από τους σκοπούς του παρόντος συγγράμματος, καθώς προκύπτει συνδυάζοντας το Κεντρικό Οριακό Θεώρημα και ένα ακόμη ασυμπτωτικό αποτέλεσμα (βλ. Θεώρημα του Slutsky για το οποίο παραπέμπουμε, μεταξύ άλλων, στο σύγγραμμα van der Vaart, 1998).

9.3 Κατανομή της διαφοράς δύο δειγματικών μέσων τιμών

Στην πράξη πολύ συχνά δεν ενδιαφερόμαστε να μελετήσουμε τη μέση τιμή ενός πληθυσμού, αλλά μας ενδιαφέρει να συγκρίνουμε τις μέσες τιμές δύο πληθυσμών. Για τον λόγο αυτό παίρνουμε ένα τυχαίο δείγμα $(X_1, \dots, X_n)$ από τον πρώτο πληθυσμό και ένα τυχαίο δείγμα $(Y_1, \dots, Y_m)$ από έναν δεύτερο πληθυσμό. Επιπλέον, υποθέτουμε ότι οι δύο πληθυσμοί έχουν πεπερασμένες μέσες τιμές και διακυμάνσεις μ_i και σ_i^2, αντίστοιχα, για $i = 1,2$. Το ενδιαφέρον μας σε αυτήν την ενότητα επικεντρώνεται στον προσδιορισμό της κατανομής της διαφοράς των δειγματικών μέσων $\bar{X} - \bar{Y}$ που αποτελεί το δειγματικό ανάλογο της διαφοράς $\mu_1 - \mu_2$. Ο προσδιορισμός αυτός θα γίνει διακρίνοντας δύο περιπτώσεις: τα δύο τυχαία δείγματα να είναι ανεξάρτητα (π.χ. να μελετήσουμε το επίπεδο χοληστερίνης σε άντρες και γυναίκες ή τον μηνιαίο μισθό των Ελλήνων σε σχέση με τους υπόλοιπους Ευρωπαίους) ή εξαρτημένα (π.χ. η αντοχή ενός σκυροδέματος πριν και μετά την επεξεργασία του, η επίδοση των μαθητών στην αρχή της σχολικής χρονιάς και μετά την ολοκλήρωσή της).

9.3.1 Ανεξάρτητα δείγματα

Έστω ότι έχουμε ένα τυχαίο δείγμα $(X_1, \dots, X_n)$ από τον πρώτο πληθυσμό με πεπερασμένη μέση τιμή και διακύμανση, μ_1 και σ_1^2, αντίστοιχα, και ένα ανεξάρτητο τυχαίο δείγμα $(Y_1, \dots, Y_m)$ από έναν δεύτερο πληθυσμό με πεπερασμένη μέση τιμή και διακύμανση, μ_2 και σ_2^2, αντίστοιχα. Στη συνέχεια, θα διακρίνουμε τις ακόλουθες περιπτώσεις:

Κανονικοί πληθυσμοί με γνωστές διασπορές.

Σε αυτήν την περίπτωση ισχύει η ακόλουθη πρόταση.

Πρόταση 9.8

Έστω $(X_1, \dots, X_n)$ και $(Y_1, \dots, Y_m)$ ανεξάρτητα τυχαία δείγματα από κανονικούς πληθυσμούς με κατανομές $N(\mu_1, \sigma_1^2)$ και $N(\mu_2, \sigma_2^2)$, αντίστοιχα. Τότε

$$\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}} \sim N(0,1).$$

Απόδειξη Πρότασης 9.8

Είναι γνωστό από την Πρόταση 9.2 ότι $\bar{X} \sim N(\mu_1, \sigma_1^2/n)$ και $\bar{Y} \sim N(\mu_2, \sigma_2^2/m)$. Εφόσον $\bar{X}$ και $\bar{Y}$ ανεξάρτητες τ.μ. που ακολουθούν την κανονική κατανομή, τότε και η διαφορά τους (ως γραμμικός συνδυασμός ανεξάρτητων κανονικών κατανομών) θα ακολουθεί την κανονική κατανομή. Πιο συγκεκριμένα, έχουμε ότι:

$$\bar{X} - \bar{Y} \sim N(\mu_{\bar{X}-\bar{Y}}, \sigma^2_{\bar{X}-\bar{Y}}),$$

όπου

$$\mu_{\bar{X}-\bar{Y}} = E(\bar{X} - \bar{Y}) = E(\bar{X}) - E(\bar{Y}) = \mu_1 - \mu_2,$$

και

$$\sigma^2_{\bar{X}-\bar{Y}} = Var(\bar{X} - \bar{Y}) = Var(\bar{X}) + Var(\bar{Y}) = \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}.$$

Επομένως, έχουμε ότι:

$$\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}} \sim N(0,1).$$

Κανονικοί πληθυσμοί με άγνωστες αλλά ίσες διασπορές.

Σε αυτήν την περίπτωση ισχύει η ακόλουθη πρόταση.

Πρόταση 9.9

Έστω $X_1, \dots, X_n$ ένα τυχαίο δείγμα μεγέθους n από την κανονική κατανομή με παραμέτρους (μ_1, σ^2). Επιπρόσθετα, έστω ότι $Y_1, \dots, Y_m$ ένα τυχαίο δείγμα μεγέθους m από την κανονική κατανομή με παραμέτρους (μ_2, σ^2). Υποθέτοντας ότι τα δύο δείγματα είναι ανεξάρτητα και η κοινή διασπορά των πληθυσμών άγνωστη, έχουμε ότι:

$$T = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n} + \frac{1}{m}}} \sim t_{n+m-2},$$

όπου

$$S_p^2 = \frac{(n-1)S_1^2 + (m-1)S_2^2}{n+m-2}.$$

Απόδειξη Πρότασης 9.9

Από την Πρόταση 7.12 έχουμε ότι:

$$\bar{X} \sim N(\mu_1, \frac{\sigma^2}{n}) \text{ και } \bar{Y} \sim N(\mu_2, \frac{\sigma^2}{m}),$$

οπότε, καθώς τα δύο δείγματα είναι ανεξάρτητα, έχουμε ότι

$$\bar{X} - \bar{Y} \sim N\left(\mu_1 - \mu_2, \frac{\sigma^2}{n} + \frac{\sigma^2}{n}\right),$$

ή, ισοδύναμα,

$$Z = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n} + \frac{1}{m}}} \sim N(0,1). \tag{9.1}$$

Επιπρόσθετα, από την Πρόταση 9.4 έχουμε ότι:

$$\frac{(n-1)S_1^2}{\sigma^2} \sim \chi^2_{n-1} \text{ και } \frac{(m-1)S_2^2}{\sigma^2} \sim \chi^2_{m-1}. \tag{9.2}$$

Καθώς έχουμε ανεξάρτητα δείγματα, οι τ.μ. S_1^2 και S_2^2 είναι και αυτές ανεξάρτητες ως συνάρτηση ανεξάρτητων τ.μ. Από τις ιδιότητες της χι-τετράγωνο κατανομής προκύπτει ότι

$$V = \frac{(n-1)S_1^2}{\sigma^2} + \frac{(m-1)S_2^2}{\sigma^2} = \frac{(n-1)S_1^2 + (m-1)S_2^2}{\sigma^2} \sim \chi^2_{n+m-2}. \tag{9.3}$$

Από την Πρόταση 9.4 οι τυχαίες μεταβλητές Z και V είναι ανεξάρτητες και, επομένως, από τον ορισμό της t κατανομής, έχουμε ότι:

$$T = \frac{Z}{\sqrt{V/(n+m-2)}} \sim t_{n+m-2}$$

που αποδεικνύει το ζητούμενο.

Παρατήρηση 9.4

Υψώνοντας τη στατιστική συνάρτηση T, η οποία ορίστηκε στην Πρόταση 9.9, στο τετράγωνο και διαιρώντας αριθμητή και παρονομαστή με το σ^2 έχουμε ότι:

$$\mathrm{T}^2 = \frac{\left(\bar{X} - \bar{Y} - (\mu_1 - \mu_2)\right)^2}{S_p^2\left(\frac{1}{n} + \frac{1}{m}\right)} = \frac{\frac{\left(\bar{X}-\bar{Y}-(\mu_1-\mu_2)\right)^2}{\sigma^2\cdot\left(\frac{1}{n}+\frac{1}{m}\right)}}{\frac{S_p^2}{\sigma^2}} = \frac{\frac{\left(\bar{X}-\bar{Y}-(\mu_1-\mu_2)\right)^2}{\sigma^2\cdot\left(\frac{1}{n}+\frac{1}{m}\right)}}{\frac{(n-1)S_1^2+(m-1)S_2^2}{\sigma^2}}.$$

Λαμβάνοντας υπόψη το αποτέλεσμα της σχέσης (9.1) προκύπτει ότι ο αριθμητής της παραπάνω στατιστικής συνάρτησης ακολουθεί τη χ^2 κατανομή με 1 βαθμό ελευθερίας. Επιπρόσθετα, από τη σχέση (9.3) προκύπτει ότι ο παρονομαστής ακολουθεί τη χ^2 κατανομή με $n + m - 2$ βαθμούς ελευθερίας. Επομένως, από τον ορισμό της F κατανομής, καθώς οι δειγματικές μέσες τιμές είναι ανεξάρτητες από τις δειγματικές διασπορές και τα δύο δείγματα ανεξάρτητα μεταξύ τους, έχουμε ότι:

$$T^2 \sim F_{1,n+m-2}.$$

Κανονικοί πληθυσμοί με άγνωστες αλλά άνισες διασπορές.

Στην παραπάνω πρόταση υποθέσαμε ότι οι δύο κανονικές κατανομές έχουν κοινή διακύμανση. Στην περίπτωση που δεν έχουν κοινή διακύμανση ισχύει η ακόλουθη πρόταση.

Πρόταση 9.10

Έστω $X_1, \ldots, X_n$ ένα τυχαίο δείγμα μεγέθους n από την κανονική κατανομή με παραμέτρους (μ_1, σ^2). Επιπρόσθετα, έστω ότι $Y_1, \ldots, Y_m$ ένα τυχαίο δείγμα μεγέθους m από την κανονική κατανομή με παραμέτρους (μ_2, σ^2). Υποθέτοντας ότι τα δύο δείγματα είναι ανεξάρτητα, έχουμε ότι:

$$T = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n} + \frac{S_2^2}{m}}} \sim t_u.$$

όπου

$$u = \frac{\left(\frac{S_1^2}{n} + \frac{S_2^2}{m}\right)^2}{\frac{\left(S_1^2/n\right)^2}{n-1} + \frac{\left(S_2^2/m\right)^2}{m-1}}.$$

Απόδειξη Πρότασης 9.10

Η απόδειξη ξεφεύγει από τους σκοπούς αυτού του συγγράμματος και για αυτόν τον λόγο παραλείπεται. Ο/Η ενδιαφερόμενος/η παραπέμπεται στην εργασία του Welch (1947).

Η παραπάνω εκτίμηση των βαθμών ελευθερίας καλείται Satterthwaite προσέγγιση (Satterthwaite, 1946). Το u συνήθως δεν είναι ακέραιος, οπότε το στρογγυλοποιούμε στον μικρότερο ακέραιο. Στην πραγματικότητα για τη στατιστική συνάρτηση T της Πρότασης 9.10 ισχύει ότι $P(T > t_{u,\alpha}) \approx \alpha$.

Μη Κανονικοί πληθυσμοί και μεγάλα μεγέθη δειγμάτων

Αν το μέγεθος του δείγματος είναι μεγάλο τότε οι παραπάνω σχέσεις δεν ισχύουν μόνο για κανονικούς πληθυσμούς, αλλά για οποιαδήποτε κατανομή κι αν ακολουθούν τα αρχικά δείγματά μας.

Για παράδειγμα, έστω $X_1, \ldots, X_n$ ένα τυχαίο δείγμα από τον πληθυσμό με μέση τιμή μ_1 και διασπορά σ_1^2, ενώ $Y_1, \ldots, Y_m$ ένα τυχαίο δείγμα από τον πληθυσμό με μέση τιμή μ_2 και διασπορά σ_2^2. Επιπλέον, υποθέτουμε ότι

τα δύο δείγματα είναι ανεξάρτητα και οι άγνωστες διασπορές είναι πεπερασμένες. Στην περίπτωση που ένας από τους δύο πληθυσμούς δεν είναι κανονικός, αλλά το διαθέσιμο δείγμα από αυτόν είναι μεγάλου μεγέθους (συνήθως μεγαλύτερο από 30), τότε

$$\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n} + \frac{S_2^2}{m}}} \xrightarrow{d} N(0,1),$$

με S_1^2 και S_2^2 να είναι οι δειγματικές διακυμάνσεις που προκύπτουν από τα ανεξάρτητα τ.δ. $X_1, \dots, X_n$ και $Y_1, \dots, Y_m$, αντίστοιχα, ενώ υπενθυμίζουμε ότι το σύμβολο $\xrightarrow{d}$ δηλώνει ότι η κατανομή είναι προσεγγιστική.

Άσκηση Αυτοαξιολόγησης 9.3

Έστω $\bar{X}_1$ και $\bar{X}_2$ δειγματικοί μέσοι δύο τυχαίων δειγμάτων μεγέθους $n_1 > 1$ και $n_2 > 1$, αντίστοιχα, που προέρχονται από 2 ανεξάρτητους πληθυσμούς με διασπορές σ_1^2 και σ_2^2. Ποια από τις παρακάτω δηλώσεις είναι σωστή;

1. Η διασπορά της διαφοράς $\bar{X}_1 - \bar{X}_2$ είναι ίση με $\sigma_1^2 + \sigma_2^2$.
2. Η διασπορά της διαφοράς $\bar{X}_1 - \bar{X}_2$ είναι ίση με $\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$.
3. Η διασπορά της διαφοράς $\bar{X}_1 - \bar{X}_2$ είναι ίση με $\sigma_1^2 - \sigma_2^2$.
4. Η διασπορά της διαφοράς $\bar{X}_1 - \bar{X}_2$ είναι ίση με $\frac{(n_1-1)\sigma_1^2 + (n_2-1)\sigma_2^2}{n_1+n_2-2}$.

Άσκηση Αυτοαξιολόγησης 9.4

Έστω $X_1, X_2, \dots, X_n$ τυχαίες μεταβλητές από έναν κανονικά κατανεμημένο πληθυσμό $N(0, \sigma_1^2)$ και $Y_1, Y_2, \dots, Y_m$ τυχαίες μεταβλητές από έναν άλλο κανονικά κατανεμημένο πληθυσμό $N(0, \sigma_2^2)$. Θεωρώντας ότι οι δύο πληθυσμοί είναι ανεξάρτητοι μεταξύ τους και ότι οι διασπορές τους είναι γνωστές, ποια από τις παρακάτω προτάσεις δεν ισχύει;

1. Οι κατανομές των $\bar{X} + \bar{Y}$ και $\bar{X} - \bar{Y}$ έχουν την ίδια μέση τιμή και διασπορά.
2. Αν τετραπλασιαστεί το μέγεθος του δείγματος του κάθε πληθυσμού με $\bar{X}'$ και $\bar{Y}'$ να είναι οι δειγματικές μέσες τιμές στα τετραπλάσια σε μέγεθος δείγματα, τότε η διασπορά της κατανομής του $\bar{X}' + \bar{Y}'$ είναι τετραπλάσια σε σχέση με τη διασπορά της κατανομής του $\bar{X} + \bar{Y}$.
3. Αν τετραπλασιαστεί το μέγεθος του δείγματος του κάθε πληθυσμού με $\bar{X}'$ και $\bar{Y}'$ να είναι οι δειγματικές μέσες τιμές στα τετραπλάσια σε μέγεθος δείγματα, τότε η τυπική απόκλιση της κατανομής του $\bar{X} - \bar{Y}$ είναι διπλάσια σε σχέση με τη διασπορά της κατανομής του $\bar{X}' - \bar{Y}'$.
4. Αν τετραπλασιαστεί το μέγεθος του δείγματος του κάθε πληθυσμού με $\bar{X}'$ και $\bar{Y}'$ να είναι οι δειγματικές μέσες τιμές στα τετραπλάσια σε μέγεθος δείγματα, τότε η μέση τιμή της κατανομής του $\bar{X}' + \bar{Y}'$ και $\bar{X}' - \bar{Y}'$ είναι ίδια με τη μέση τιμή των $\bar{X} + \bar{Y}$ και $\bar{X} - \bar{Y}$, αντίστοιχα.

9.3.2 Εξαρτημένα δείγματα

Σε πολλές πραγματικές εφαρμογές οι πληθυσμοί μας δεν είναι ανεξάρτητοι, όπως για παράδειγμα στην περίπτωση της μέτρησης της αντοχής του σκυροδέματος από δύο διαφορετικά όργανα, της μέτρησης της αντοχής του σκυροδέματος πριν και μετά από ειδική επεξεργασία που υποθέτουμε ότι αυξάνει την αντοχή του, της μέτρησης του βάρους πριν και μετά τη χορήγηση συγκεκριμένης διατροφής σε πειραματόζωα κ.ά. Σε τέτοιες περιπτώσεις λέμε ότι έχουμε **ζευγαρωτές παρατηρήσεις** (paired observations), οι οποίες είναι της μορφής (X_i, Y_i), $i = 1, \dots, n$. Τότε αντί να μελετήσουμε τη διαφορά των μέσων τιμών $(\bar{X} - \bar{Y})$, αγνοώντας την εξάρτηση των παρατηρήσεών μας, μελετάμε τη μέση τιμή των διαφορών. Πιο συγκεκριμένα, για κάθε ζεύγος τιμών κατασκευάζουμε τις διαφορές $d_i = x_i - y_i$, $i = 1, \dots, n$. Αυτές οι τιμές είναι

παρατηρήσεις των τυχαίων μεταβλητών $D_i = X_i - Y_i$, $i = 1, \dots, n$. Τότε ισχύει η πρόταση που ακολουθεί.

Πρόταση 9.11

Έστω $X_1, \dots, X_n$ ένα τυχαίο δείγμα από έναν πληθυσμό με μέση τιμή μ_1 και $Y_1, \dots, Y_n$ ένα τυχαίο δείγμα από έναν πληθυσμό με μέση τιμή μ_2, με τα δύο δείγματα να είναι εξαρτημένα. Έστω D_i, $i = 1, \dots, n$, οι τ.μ. που δίνονται από τη σχέση $D_i = X_i - Y_i$, $i = 1, \dots, n$ με μέση τιμή $\mu_\delta = \mu_1 - \mu_2$ και άγνωστη διασπορά σ_δ^2. Υπό την υπόθεση ότι ο πληθυσμός των διαφορών είναι κανονικός ισχύει ότι:

$$\mathrm{T} = \frac{\bar{D} - \mu_\delta}{S_D/\sqrt{n}} \sim t_{n-1}$$

όπου

$$\bar{D} = \frac{\sum_{i=1}^{n} D_i}{n},$$

και

$$S_D^2 = \frac{\sum_{i=1}^{n} (D_i - \bar{D})^2}{n-1}.$$

Απόδειξη Πρότασης 9.11

Το αποτέλεσμα αυτό αποτελεί ειδική περίπτωση της Πρότασης 9.6.

Όταν ο πληθυσμός των διαφορών δεν ικανοποιεί την υπόθεση της κανονικότητας, τότε για μεγάλο μέγεθος δείγματος n το παραπάνω αποτέλεσμα ισχύει προσεγγιστικά.

Παρατήρηση 9.5

Η άγνωστη διασπορά σ_δ^2 δίνεται από τη σχέση (βλ. την Ενότητα 6.6.3):

$$\sigma_\delta^2 = Var(X - Y) = \sigma_1^2 + \sigma_2^2 - 2 \cdot Cov(X_i, Y_i),$$

όπου $\sigma_1^2 = Var(X)$ και $\sigma_2^2 = Var(Y)$.

9.4 Κατανομή δειγματικής αναλογίας

Πολλές φορές μας ενδιαφέρει να μελετήσουμε την αναλογία κάποιου χαρακτηριστικού σε έναν πληθυσμό, π.χ. το ποσοστό των ελαττωματικών προϊόντων που παράγονται από μία γραμμή παραγωγής ή το ποσοστό των ατόμων που πάσχουν από μία σπάνια ασθένεια ή το ποσοστό των ατόμων που έχουν πρόθεση να ψηφίσουν ένα κόμμα στις επόμενες εκλογές και ούτω καθεξής. Σε μία τέτοια περίπτωση αν $X_1, X_2, \dots X_n$ είναι ένα τυχαίο δείγμα από τον πληθυσμό που μας ενδιαφέρει, τότε κάθε τ.μ. X_i μπορεί να πάρει μόνο δύο τιμές, την τιμή 1 με πιθανότητα p αν το άτομο έχει το χαρακτηριστικό που μας ενδιαφέρει ή την τιμή 0, διαφορετικά. Άρα μπορούμε να ισχυριστούμε ότι η κάθε παρατήρηση προέρχεται από έναν πληθυσμό με κατανομή Bernoulli με πιθανότητα επιτυχίας p, όπου p η πραγματική αναλογία στον πληθυσμό του χαρακτηριστικού που μας ενδιαφέρει.

Τότε είναι απόλυτα λογικό από το τυχαίο δείγμα $X_1, X_2, \dots X_n$ να βρούμε το πλήθος των ατόμων, έστω Y, που έχουν το χαρακτηριστικό που μελετάμε, με $Y = \sum_{i=1}^{n} X_i$, ή, εναλλακτικά, τη δειγματική αναλογία, έστω $\hat{P}$, των ατόμων που έχουν το χαρακτηριστικό που μελετάμε, με $\hat{P} = Y/n = \sum_{i=1}^{n} X_i/n$. Είναι φανερό ότι $Y \sim B(n,p)$, αφού το άθροισμα n ανεξάρτητων τ.μ. Bernoulli ακολουθεί διωνυμική κατανομή με παραμέτρους (n,p), ενώ οι κατανομές των Y και $\hat{P}$ μπορούν να προσεγγιστούν μέσω του Κεντρικού Οριακού Θεωρήματος από την κανονική κατανομή. Τα αποτελέσματα αυτά συνοψίζονται στην πρόταση που ακολουθεί.

Πρόταση 9.12

Έστω $X_1,\ldots,X_n$ ανεξάρτητες και ισόνομες τυχαίες κατανομές με κατανομή Bernoulli με παράμετρο p, δηλαδή $X_i \sim B(1,p)$, $i = 1,\ldots,n$. Τότε ισχύουν τα ακόλουθα:

1. Η τ.μ. $Y = \sum_{i=1}^{n} X_i$ ακολουθεί $B(n,p)$, ενώ η τ.μ. $\hat{P} = Y/n = \sum_{i=1}^{n} X_i/n$ έχει συνάρτηση πιθανότητας:
$$P\left(\hat{P} = w\right) = \left(\begin{array}{c} n \\ wn \end{array} \right) p^{wn}(1-p)^{n-wn},\ w = 0,1/n,2/n \ldots,1.$$
Επιπλέον, είναι $E(Y) = np$ και $Var(Y) = np(1-p)$.
2. Για μεγάλο μέγεθος δείγματος n, η τυχαία μεταβλητή $\frac{\hat{P}-p}{\sqrt{p(1-p)/n}}$ ακολουθεί προσεγγιστικά την τυπική κανονική κατανομή $N(0,1)$, δηλαδή
$$\frac{\hat{P}-p}{\sqrt{p(1-p)/n}} \xrightarrow{d} N(0,1).$$
3. Για μεγάλο μέγεθος δείγματος n, η τυχαία μεταβλητή $\frac{\hat{P}-p}{\sqrt{\hat{P}(1-\hat{P})/n}}$ ακολουθεί προσεγγιστικά την τυπική κανονική κατανομή $N(0,1)$, δηλαδή
$$\frac{\hat{P}-p}{\sqrt{\hat{P}(1-\hat{P})/n}} \xrightarrow{d} N(0,1).$$

Απόδειξη Πρότασης 9.12

1. Για την απόδειξη παραπέμπουμε τον/την αναγνώστη/στρια στην Πρόταση 7.3 και την Παρατήρηση 7.8, καθώς και στις ιδιότητες της μέσης τιμής και της διακύμανσης της διωνυμικής κατανομής.
2. Το αποτέλεσμα προκύπτει άμεσα με εφαρμογή του Κεντρικού Οριακού Θεωρήματος, όπως αναλυτικά είδαμε στην Ενότητα 7.3.1 και στην Πρόταση 7.15.
3. Το αποτέλεσμα προκύπτει συνδυάζοντας το Κεντρικό Οριακό Θεώρημα και ένα επιπρόσθετο ασυμπτωτικό θεώρημα που είναι γνωστό ως Θεώρημα του Slutsky. Η αναλυτική παρουσίαση της απόδειξης ξεφεύγει από τους σκοπούς του συγγράμματος και για αυτό και παραλείπεται (βλ., μεταξύ άλλων van der Vaart, 1998).

Σχετικά με το πότε οι παραπάνω προσεγγίσεις είναι ικανοποιητικές παραπέμπουμε για λεπτομέρειες στην Ενότητα 7.3.1.

9.5 Κατανομή διαφοράς δειγματικών αναλογιών

Πολλές φορές θέλουμε να συγκρίνουμε τα ποσοστά ενός χαρακτηριστικού σε δύο ανεξάρτητους πληθυσμούς. Ενδεικτικά ενδιαφερόμαστε να συγκρίνουμε το ποσοστό των ατόμων με παχυσαρκία σε δύο ηλικιακές ομάδες, στα άτομα ηλικίας κάτω από 30 και πάνω από 30 χρονών ή το ποσοστό που παίρνει το κόμμα Α στους άντρες και στις γυναίκες ψηφοφόρους. Σε αυτές αλλά και σε παρόμοιες περιπτώσεις μας ενδιαφέρει, όπως είναι αναμενόμενο, να μελετήσουμε τη διαφορά $p_1 - p_2$, με p_i να είναι το ποσοστό του χαρακτηριστικού που μας ενδιαφέρει στον i-οστό πληθυσμό. Σε μία τέτοια περίπτωση, θεωρούμε ένα τυχαίο δείγμα $X_1, X_2, \ldots, X_n$ από τον πρώτο πληθυσμό που μας ενδιαφέρει και ένα τυχαίο δείγμα $Y_1, Y_2, \ldots, Y_m$ από τον δεύτερο πληθυσμό που μας ενδιαφέρει. Επιπλέον, καθεμία από τις τ.μ. X_i (Y_j) λαμβάνει μόνο δύο τιμές, την τιμή 1 με πιθανότητα p_1 (p_2, αντίστοιχα) αν το άτομο έχει το χαρακτηριστικό που μας ενδιαφέρει ή την τιμή 0, διαφορετικά. Στο παραπάνω πλαίσιο είναι απόλυτα λογικό η εξαγωγή

συμπερασμάτων για την πραγματική διαφορά των αναλογιών $p_1 - p_2$ να στηρίζεται στην αντίστοιχη διαφορά των αναλογιών στα δύο τυχαία δείγματα, δηλαδή στην τυχαία μεταβλητή $\hat{P}_1 - \hat{P}_2 = \sum_{i=1}^{n} X_i/n - \sum_{j=1}^{m} Y_j/m$. Υποθέτουμε ότι τα δύο δείγματα είναι ανεξάρτητα μεταξύ τους. Τότε από το Κεντρικό Οριακό Θεώρημα έχουμε ότι:

$$\hat{P}_1 \xrightarrow{d} N\left(p_1, \frac{p_1(1-p_1)}{n}\right),$$

και

$$\hat{P}_2 \xrightarrow{d} N\left(p_2, \frac{p_2(1-p_2)}{m}\right).$$

Άρα η διαφορά $\hat{P}_1 - \hat{P}_2$ ως γραμμικός συνδυασμός ανεξάρτητων τυχαίων μεταβλητών θα ακολουθεί προσεγγιστικά κανονική κατανομή, οπότε έχουμε:

$$\frac{\hat{P}_1 - \hat{P}_2 - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n} + \frac{p_2(1-p_2)}{m}}} \xrightarrow{d} N(0,1).$$

Επιπρόσθετα, με χρήση του Θεωρήματος Slutsky προκύπτει ότι (βλ., μεταξύ άλλων, van der Vaart, 1998)

$$\frac{\hat{P}_1 - \hat{P}_2 - (p_1 - p_2)}{\sqrt{\frac{\hat{P}_1(1-\hat{P}_1)}{n} + \frac{\hat{P}_2(1-\hat{P}_2)}{m}}} \xrightarrow{d} N(0,1).$$

9.6 Κατανομή λόγου δειγματικών διασπορών

Εκτός από τη σύγκριση των μέσων τιμών δύο ανεξάρτητων πληθυσμών πολλές φορές χρειάζεται να συγκρίνουμε και τις διασπορές τους. Η σύγκριση αυτή, όπως θα δούμε στα επόμενα δύο κεφάλαια, βασίζεται στον λόγο των δειγματικών διασπορών. Για τον λόγο αυτό, αντικείμενο μελέτης αυτής της ενότητας αποτελεί η κατανομή του λόγου των δειγματικών διασπορών δύο ανεξάρτητων τυχαίων δειγμάτων.

Πρόταση 9.13

Έστω $X_1, \ldots, X_n$ ένα τυχαίο δείγμα μεγέθους n από την κανονική κατανομή με παραμέτρους (μ_1, σ_1^2). Επιπρόσθετα, έστω ότι $Y_1, \ldots, Y_m$ ένα τυχαίο δείγμα μεγέθους m από την κανονική κατανομή με παραμέτρους (μ_2, σ_2^2). Υποθέτοντας ότι τα δύο δείγματα είναι ανεξάρτητα, έχουμε ότι

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} = \frac{S_1^2}{S_2^2} \cdot \frac{\sigma_2^2}{\sigma_2^2} \sim F_{n-1,m-1},$$

όπου

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} \text{ και } S_1^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1},$$

και

$$\bar{Y} = \frac{\sum_{j=1}^{n} Y_j}{m} \text{ και } S_2^2 = \frac{\sum_{j=1}^{m} (Y_j - \bar{Y})^2}{m-1}.$$

Απόδειξη Πρότασης 9.13

Έστω οι τ.μ. $Z = \frac{(n-1)S_1^2}{\sigma_1^2}$ και $V = \frac{(m-1)S_2^2}{\sigma_2^2}$. Οι τ.μ. Z και V είναι ανεξάρτητες ως συναρτήσεις ανεξάρτητων τυχαίων μεταβλητών και, σύμφωνα με την Πρόταση 9.4, είναι: $Z \sim \chi^2_{n-1}$ και $V \sim \chi^2_{m-1}$. Επομένως, το ζητούμενο προκύπτει από τον τρόπο ορισμού της F κατανομής.

Σε περίπτωση μη κανονικών πληθυσμών και μεγάλων δειγμάτων μπορεί να προσδιοριστεί η κατανομή του λόγου των δειγματικών διασπορών χρησιμοποιώντας τη μέθοδο που είναι γνωστή στη στατιστική βιβλιογραφία ως **μέθοδος δέλτα**. Για σχετικές πληροφορίες παραπέμπουμε, μεταξύ άλλων, στο σύγγραμμα του van der Vaart (1998).

9.7 Ασκήσεις

Άσκηση 9.1 *Έστω $(X_1, \dots, X_{16})$ τυχαίο δείγμα από κανονικό πληθυσμό με μέση τιμή μ και τυπική απόκλιση 2. Υπολογίστε την πιθανότητα ο δειγματικός μέσος $\bar{X}$ να βρίσκεται στο διάστημα $\mu_{\bar{X}} - 1.9\sigma_{\bar{X}}$ και $\mu_{\bar{X}} + 1.9\sigma_{\bar{X}}$.*

Άσκηση 9.2 *Έστω $X_1, \dots, X_n$ ένα τυχαίο δείγμα από την κανονική κατανομή $N(0, \sigma^2)$.*

1. *Να βρεθεί η κατανομή του $X_i - \bar{X}$, για i σταθερό.*
2. *Να προσδιοριστεί η κατανομή του $n \cdot \frac{\bar{X}^2}{\sigma^2}$.*
3. *Αν $\bar{X}_k = \frac{\sum_{i=1}^{k} X_i}{k}$ και $\bar{X}_{n-k} = \frac{\sum_{i=k+1}^{n} X_i}{n-k}$, να βρεθεί η κατανομή της στατιστικής συνάρτησης $(\bar{X}_k + \bar{X}_{n-k})/2$.*

Άσκηση 9.3 *Έστω $X_1, \dots, X_{16}$ και $Y_1, \dots, Y_9$ τυχαία δείγματα από ανεξάρτητους κανονικούς πληθυσμούς, όπου $X_i \sim N(0,25)$, $i = 1, \dots, 16$ και $Y_i \sim N(0,16)$, $i = 1, \dots, 9$. Να υπολογιστεί η πιθανότητα $P(\bar{X} - \bar{Y} > 0)$.*

Άσκηση 9.4 *Η τυχαία μεταβλητή X που περιγράφει τον χρόνο εξυπηρέτησης ενός πελάτη από τον ταμία μιας τράπεζας ακολουθεί κανονική κατανομή με μέση τιμή 7 λεπτά και τυπική απόκλιση 2 λεπτά. Αν στον χώρο αναμονής της τράπεζας περιμένουν να εξυπηρετηθούν 65 πελάτες, ποια είναι η πιθανότητα ο ταμίας να τους εξυπηρετήσει όλους μέσα στο οκτάωρό του;*

Υπόδειξη: υποθέτουμε ότι οι πελάτες εξυπηρετούνται διαδοχικά ο ένας μετά τον άλλον, και οι χρόνοι εξυπηρέτησής τους είναι ανεξάρτητοι.

Άσκηση 9.5 *Στα ζώα μιας κτηνοτροφικής μονάδας δίνεται τροφή τρεις φορές την ημέρα. Η ποσότητα θερμίδων που παίρνουν κάθε φορά είναι τυχαία μεταβλητή που ακολουθεί κανονική κατανομή. Το διαιτολόγιο έχει ρυθμιστεί έτσι, ώστε την πρώτη φορά που δίνεται τροφή στα ζώα η μέση ποσότητα θερμίδων που παίρνουν να είναι $500\ cal$ με τυπική απόκλιση $50\ cal$, τη δεύτερη $1700\ cal$ με τυπική απόκλιση $200\ cal$ και την τρίτη να είναι $800\ cal$ με τυπική απόκλιση $100\ cal$.*

1. *Υπολογίστε την πιθανότητα η συνολική ημερήσια ποσότητα θερμίδων που παίρνει ένα τυχαία επιλεγμένο ζώο της μονάδας να ξεπερνάει τις $3100\ cal$.*
2. *Αν μετρήσουμε την ημερήσια κατανάλωση θερμίδων πέντε ζώων της μονάδας αυτής, υπολογίστε την πιθανότητα το πολύ δύο από αυτά τα ζώα να καταναλώνουν περισσότερες από $3100\ cal$ ημερησίως.*
3. *Υπολογίστε την πιθανότητα, η μέση ποσότητα θερμίδων που θα πάρει ένα ζώο σε έναν χρόνο (365 ημέρες) να είναι μεταξύ $2975\ cal$ και $3025\ cal$.*

Άσκηση 9.6 *Η αντοχή κυλινδρικών δοκιμίων από σκυρόδεμα (σε kg/cm^2) στο αρχικό στάδιο παρασκευής τους ακολουθεί κανονική κατανομή $N(80,9)$.*

1. *Μετά από ειδική επεξεργασία η αντοχή του σκυροδέματος αυξάνεται κατά 10%. Υπολογίστε την πιθανότητα ένα τυχαία επιλεγμένο επεξεργασμένο δοκίμιο να έχει αντοχή περισσότερη από $90\ kg/cm^2$.*
2. *Αν επιλεχθούν πέντε επεξεργασμένα δοκίμια στην τύχη, υπολογίστε την πιθανότητα η μέση αντοχή τους να είναι λιγότερη από $80\ kg/cm^2$.*
3. *Αν επιλεχθούν πέντε επεξεργασμένα δοκίμια στην τύχη, υπολογίστε την πιθανότητα δύο από αυτά να έχουν αντοχή περισσότερη από $90\ kg/cm^2$.*
4. *Αν επιλεχθούν πέντε επεξεργασμένα δοκίμια και δύο δοκίμια στην αρχική τους μορφή (πριν την επεξεργασία) στην τύχη, υπολογίστε την πιθανότητα η μέση αντοχή τους να είναι περισσότερη από $90\ kg/cm^2$.*

5. *Αν επιλεχθούν πέντε επεξεργασμένα δοκίμια και δύο δοκίμια στην αρχική τους μορφή (πριν την επεξεργασία) στην τύχη, υπολογίστε την πιθανότητα όλα τα δοκίμια να έχουν αντοχή περισσότερη από* $90\ kg/cm^2$.
6. *Υπολογίστε την πιθανότητα το δεύτερο δοκίμιο με αντοχή περισσότερη από* $90\ kg/cm^2$ *να είναι το πέμπτο στη σειρά από τυχαία επιλεγμένα δοκίμια.*
7. *Κατά τον ποιοτικό έλεγχο, έπειτα από εντολή του υπεύθυνου μηχανικού, απορρίπτονται όλα τα επεξεργασμένα δοκίμια που έχουν αντοχή μικρότερη από* $80\ kg/cm^2$. *Αν μετά τον ποιοτικό έλεγχο ο μηχανικός επιλέξει ένα δοκίμιο στην τύχη, υπολογίστε την πιθανότητα η αντοχή του να είναι μεγαλύτερη από* $90\ kg/cm^2$.

Άσκηση 9.7 *Γνωρίζουμε ότι η πιθανότητα αναμονής ενός πελάτη μιας τράπεζας για περισσότερο από 20 λεπτά είναι 0.0239. Υποθέτουμε ότι ο χρόνος αναμονής ακολουθεί κανονική κατανομή με τυπική απόκλιση 3.75 λεπτά.*

1. *Ποιος είναι ο μέσος χρόνος αναμονής στην τράπεζα;*
2. *Υπολογίστε την πιθανότητα ένας πελάτης να περιμένει στην τράπεζα από 10 μέχρι 15 λεπτά.*
3. *Αν, επιπλέον, ο χρόνος εξυπηρέτησης του πελάτη ακολουθεί κανονική κατανομή με μέση τιμή 7.5 λεπτά και τυπική απόκλιση 2 λεπτά, υπολογίστε την πιθανότητα ο συνολικός χρόνος παραμονής του πελάτη στην τράπεζα να ξεπεράσει τα 20 λεπτά.*
4. *Υπολογίστε την πιθανότητα τις επόμενες τέσσερις φορές που θα πάει ο πελάτης στην τράπεζα, τη μια να περιμένει λιγότερο από 10, τις δύο να περιμένει 10 με 15 λεπτά και τη μια να περιμένει περισσότερο από 15 λεπτά.*

Άσκηση 9.8 *Σε μία πόλη θέλουμε να εκτιμήσουμε το ποσοστό των ψηφοφόρων ενός κόμματος Α. Για τον σκοπό αυτό ρωτήθηκαν 250 τυχαία επιλεγμένοι ψηφοφόροι. Να υπολογιστεί (προσεγγιστικά) η πιθανότητα η πλειονότητα των ψηφοφόρων του δείγματος να είναι υπέρ του κόμματος Α, ενώ στην πραγματικότητα μόλις το* 40% *των ψηφοφόρων της πόλης είναι υπέρ του κόμματος Α.*

9.8 Απαντήσεις στις ασκήσεις αυτοαξιολόγησης

Λύση Άσκησης Αυτοαξιολόγησης 9.1

Η ιδιότητα που δεν ισχύει είναι αυτή που διατυπώνεται στην τρίτη πρόταση.
Η τρίτη πρόταση δεν ισχύει, αφού ο $\bar{X}$ είναι συνεχής τυχαία μεταβλητή, οπότε $P(\bar{X} = \mu) = 0$.
Η πρώτη πρόταση είναι σωστή, αφού $Var(\bar{X}) = \sigma^2/n$, οπότε όσο το n αυξάνεται η $Var(\bar{X})$ μειώνεται.
Η δεύτερη πρόταση είναι σωστή, αφού $E(\bar{X}) = \mu$.
Η τέταρτη πρόταση είναι σωστή, αφού $Var(\bar{X}) = \sigma^2/n$ και $n \geq 1$ άρα $\sigma^2/n \leq \sigma^2$.

Λύση Άσκησης Αυτοαξιολόγησης 9.2

Από την Πρόταση 9.5 έχουμε ότι $E(S^2) = \sigma^2$, άρα η πρώτη πρόταση είναι αληθής. Επιπλέον από την Πρόταση 9.5 έχουμε ότι $Var(S^2) = \frac{2\sigma^4}{n-1}$, άρα όσο το n αυξάνεται τόσο η διασπορά της δειγματικής διασποράς μειώνεται. Επομένως, συμπεραίνουμε ότι η τρίτη πρόταση είναι και αυτή αληθής. Από την άλλη μεριά από την Πρόταση 9.5 έχουμε ότι η δειγματική διασπορά ακολουθεί γάμμα κατανομή με παραμέτρους $\left(\frac{n-1}{2}, \frac{2\sigma^2}{n-1}\right)$ και, επομένως, η δεύτερη πρόταση δεν είναι αληθής. Άρα σωστή απάντηση είναι η δεύτερη.

Λύση Άσκησης Αυτοαξιολόγησης 9.3

Από την Πρόταση 9.2 έχουμε ότι:

$$\bar{X}_1 \sim N\left(\mu_1, \sigma_1^2/n_1\right)$$

και

$$\bar{X}_2 \sim N\left(\mu_2, \sigma_2^2/n_2\right).$$

Λόγω της ανεξαρτησίας των $\bar{X}_1$ και $\bar{X}_2$ άμεσα από τις ιδιότητες της κανονικής κατανομής προκύπτει ότι

$$\bar{X}_1 - \bar{X}_2 \sim N\left(\mu_1 - \mu_2, \sigma_1^2/n_1 + \sigma_2^2/n_2\right).$$

Επομένως, συμπεραίνουμε ότι σωστή απάντηση είναι η δεύτερη.

Λύση Άσκησης Αυτοαξιολόγησης 9.4

Από την Πρόταση 9.2 έχουμε ότι:

$$\bar{X} \sim N\left(0, \sigma_1^2/n\right)$$

και

$$\bar{Y} \sim N\left(0, \sigma_2^2/m\right).$$

Λόγω της ανεξαρτησίας των $\bar{X}$ και $\bar{Y}$ άμεσα από τις ιδιότητες της κανονικής κατανομής προκύπτει ότι

$$\bar{X} - \bar{Y} \sim N\left(0, \sigma_1^2/n + \sigma_2^2/m\right),$$

και

$$\bar{X} + \bar{Y} \sim N\left(0, \sigma_1^2/n + \sigma_2^2/m\right),$$

Επομένως, η πρώτη πρόταση είναι αληθής.
Αν τώρα $\bar{X}'$ και $\bar{Y}'$ οι δειγματικές μέσες τιμές που προκύπτουν από τα τετραπλάσια σε μέγεθος δείγματα, τότε με παρόμοιο σκεπτικό προκύπτει ότι:

$$\bar{X}' - \bar{Y}' \sim N\left(0, \sigma_1^2/(4n) + \sigma_2^2/(4m)\right),$$

και

$$\bar{X}' + \bar{Y}' \sim N\left(0,\, \sigma_1^2/(4n) + \sigma_2^2/(4m)\right).$$

Συνεπώς, συμπεραίνουμε ότι η μέση τιμή της κατανομής των $\bar{X}' + \bar{Y}'$ και $\bar{X}' - \bar{Y}'$ είναι ίδια με αυτή των $\bar{X} + \bar{Y}$ και $\bar{X} - \bar{Y}$, αντίστοιχα (άρα η τέταρτη πρόταση είναι αληθής). Επιπρόσθετα, είναι

$$\begin{aligned} Var(\bar{X}' + \bar{Y}') &= \sigma_1^2/(4n) + \sigma_2^2/(4m) \\ &= (\sigma_1^2/n + \sigma_2^2/m)/4 = Var(\bar{X} + \bar{Y})/4, \end{aligned}$$

άρα η διασπορά της κατανομής του $\bar{X}' + \bar{Y}'$ υποτετραπλασιάζεται (άρα η δεύτερη πρόταση δεν είναι αληθής). Τέλος, είναι

$$\begin{aligned} Var(\bar{X}' - \bar{Y}') &= \sigma_1^2/(4n) + \sigma_2^2/(4m) \\ &= (\sigma_1^2/n + \sigma_2^2/m)/4 = Var(\bar{X} - \bar{Y})/4, \end{aligned}$$

και, επομένως,

$$\sigma_{\bar{X}'-\bar{Y}'} = \sqrt{Var(\bar{X} + \bar{Y})/4} = \sigma_{\bar{X}-\bar{Y}}/2.$$

Άρα η τρίτη πρόταση είναι αληθής.
Συνοψίζοντας, δεν ισχύει η δεύτερη πρόταση (άρα σωστή απάντηση είναι η δεύτερη).

ΒΙΒΛΙΟΓΡΑΦΙΑ

Ελληνόγλωσση

Ηλιόπουλος, Γ. (2006). *Βασικές μέθοδοι εκτίμησης παραμέτρων με σημείο και με διάστημα*. Αθήνα: Αθ. Σταμούλης.

Κουτρουβέλης, Ι. Α. (2011). *Εφαρμοσμένες πιθανότητες και στατιστική*. Συμμετρία.

Ξενόγλωσση

Satterthwaite, F. E. (1946). An Approximate Distribution of Estimates of Variance Components. *Biometrics*, *2*, pp. 110–114.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press.

Welch, B. L. (1947). The Generalization of `Student's' Problem when Several Different Population Variances are Involved. *Biometrika*, *34*, pp. 28–35.

ΚΕΦΑΛΑΙΟ 10

ΕΚΤΙΜΗΤΙΚΗ

Σύνοψη

Σε αυτό το κεφάλαιο παρουσιάζονται αποτελέσματα για την εκτίμηση, σε σημείο και σε διάστημα, των άγνωστων παραμέτρων γνωστών πιθανοθεωρητικών μοντέλων. Ειδικότερα, τα αποτελέσματα αυτά αφορούν την εκτίμηση της μέσης τιμής κανονικού πληθυσμού, της διαφοράς των μέσων τιμών κανονικών πληθυσμών με ανεξάρτητα ή εξαρτημένα δείγματα, της διασποράς κανονικού πληθυσμού, του πηλίκου των διασπορών δύο κανονικών πληθυσμών, της διωνυμικής πιθανότητας και της διαφοράς δύο ποσοστών.

Προαπαιτούμενη γνώση: Κεφάλαια 3-9 του παρόντος συγγράμματος.

Προσδοκώμενα μαθησιακά αποτελέσματα: Όταν θα έχετε ολοκληρώσει τη μελέτη του κεφαλαίου αυτού, θα γνωρίζετε:

- να εκτιμάτε σε σημείο μια παράμετρο ενός πληθυσμού,
- τη διαδικασία εκτίμησης σε διάστημα μιας άγνωστης παραμέτρου ενός πληθυσμού,
- τι είναι διάστημα εμπιστοσύνης, τι πιθανότητα κάλυψης και τι συντελεστής εμπιστοσύνης ενός διαστήματος,
- τα σημαντικότερα διαστήματα εμπιστοσύνης της Στατιστικής και
- να χρησιμοποιείτε την R για να υπολογίζετε διαστήματα εμπιστοσύνης.

Οικονόμου, Π., Μαλεφάκη, Σ., & Μπατσίδης, Α. (2023). *Πιθανότητες - Στατιστική.* [Προπτυχιακό εγχειρίδιο].

Γλωσσάριο επιστημονικών όρων

- Αμερόληπτος εκτιμητής
- Αμεροληψία
- Αντιστρεπτή ποσότητα
- Α.Ο.Ε.Δ εκτιμητής
- Βέλτιστος εκτιμητής
- Διάστημα εμπιστοσύνης
- Διάστημα εμπιστοσύνης ελαχίστου μήκους
- Διάστημα εμπιστοσύνης ίσων ουρών
- Εκτίμηση
- Εκτιμήτρια συνάρτηση
- Επίπεδο εμπιστοσύνης
- Μέσο τετραγωνικό σφάλμα
- Πιθανότητα κάλυψης
- Ποσό μεροληψίας
- Ποσότητα οδηγός
- Συντελεστής εμπιστοσύνης
- Τυπικό σφάλμα

10.1 Εισαγωγή

Στα αρχικά στάδια της ανάπτυξής της η Στατιστική περιοριζόταν στην απλή καταγραφή στοιχείων και τη συνοπτική παρουσίασή τους, δηλαδή περιοριζόταν στις μεθόδους της Περιγραφικής Στατιστικής που παρουσιάστηκαν στο Κεφάλαιο 8. Από τις αρχές του 20ού αιώνα όμως με την ανάπτυξη και τη μαθηματική θεμελίωση της Θεωρίας Πιθανοτήτων, η Στατιστική άρχισε να προσλαμβάνει αυστηρή μαθηματική μορφή έχοντας ως στόχο την ανάπτυξη μεθοδολογιών για την επέκταση των συμπερασμάτων από τη μελέτη του δείγματος στον πληθυσμό (Επαγωγική Στατιστική). Αντικείμενο του κεφαλαίου αυτού είναι ο κλάδος της Επαγωγικής Στατιστικής που ονομάζεται Εκτιμητική και περιλαμβάνει τις μεθοδολογίες προσδιορισμού από το τυχαίο δείγμα μιας στατιστικής συνάρτησης για την προσέγγιση-εκτίμηση της άγνωστης παραμέτρου του πληθυσμού με δύο τρόπους: σε σημείο και σε διάστημα. Στο πλαίσιο του παρόντος συγγράμματος, όσον αφορά τη σημειοεκτιμητική δεν θα δοθεί βαρύτητα στις μεθόδους εύρεσης εκτιμητών, αλλά σε επιθυμητές ιδιότητες που πρέπει να έχει ένας εκτιμητής σε σημείο. Στη συνέχεια και αφού αναδειχθεί ο λόγος της μετάβασης από τη σημειακή εκτίμηση στην εκτίμηση σε διάστημα, θα παρουσιαστεί η έννοια του διαστήματος εμπιστοσύνης. Το κεφάλαιο ολοκληρώνεται με την παράθεση αποτελεσμάτων που αφορούν την εκτίμηση σε διάστημα της μέσης τιμής κανονικού πληθυσμού με γνωστή ή άγνωστη πληθυσμιακή διακύμανση, τη διαφορά των μέσων τιμών κανονικών πληθυσμών με ανεξάρτητα ή εξαρτημένα δείγματα, της διασποράς κανονικού πληθυσμού, του πηλίκου των διασπορών δύο κανονικών πληθυσμών, της διωνυμικής πιθανότητας και της διαφοράς δύο ποσοστών.

10.2 Εκτίμηση σε σημείο

Έστω $X_1,\ldots,X_n$ ένα τυχαίο δείγμα από έναν πληθυσμό με σππ ή σπ $f(x;\theta)$, όπου θ είναι άγνωστη παράμετρος (ή διάνυσμα άγνωστων παραμέτρων). Επιπρόσθετα, σε όσα ακολουθούν το σύνολο των δυνατών τιμών της παραμέτρου θ θα συμβολίζεται με Θ και θα καλείται παραμετρικός χώρος. Τα παραπάνω σημαίνουν ότι μας είναι γνωστή η συναρτησιακή μορφή της σππ ή της σπ των τυχαίων μεταβλητών X_i, $i=1,\ldots,n$, αλλά μας είναι άγνωστη η τιμή της παραμέτρου θ που εμπλέκεται στον ορισμό αυτής. Για την καλύτερη κατανόηση των παραπάνω θα διατυπωθούν στη συνέχεια κάποια παραδείγματα.

Από προηγούμενες μελέτες ξέρουμε ότι ο αριθμός των τηλεφωνημάτων που δέχεται η γραμματεία ενός ιατρικού κέντρου στη μονάδα του χρόνου ακολουθεί κατανομή Poisson, αλλά μας είναι άγνωστη η παράμετρος, έστω θ αυτής, με $\theta \in (0,+\infty)$. Επομένως, σε αυτό το παράδειγμα είναι $f(x;\theta)=\frac{\theta^x e^{-\theta x}}{x!}$, $x=0,1,2\ldots$ με $\theta \in \Theta=(0,\infty)$. Από άλλες μελέτες είναι γνωστό ότι ο δείκτης νοημοσύνης των παιδιών προσχολικής ηλικίας περιγράφεται ικανοποιητικά από την κανονική κατανομή, αλλά μας είναι άγνωστες η πληθυσμιακή μέση τιμή και η διακύμανση. Άρα, σε αυτό το δεύτερο παράδειγμα, είναι $X \sim N(\mu,\sigma^2)$ με $\theta=(\mu,\sigma^2)$, δηλαδή έχουμε διάνυσμα άγνωστων παραμέτρων, ενώ ο παραμετρικός χώρος είναι $\Theta=\mathbb{R}\times(0,\infty)$, δηλαδή το καρτεσιανό γινόμενο του συνόλου των πραγματικών αριθμών με το σύνολο $(0,+\infty)$.

Ο στόχος της Εκτιμητικής είναι η εκτίμηση των άγνωστων παραμέτρων του πληθυσμού ή κάποιας συνάρτησης αυτών, χρησιμοποιώντας την πληροφορία που είναι διαθέσιμη για αυτές τις παραμέτρους στο δείγμα. Επομένως, είναι προφανές ότι η εκτίμηση πρέπει να στηρίζεται στο δείγμα και, πιο συγκεκριμένα, σε στατιστικές συναρτήσεις, δηλαδή συναρτήσεις των $X_1,\ldots,X_n$ και γνωστών ποσοτήτων. Αυτές οι στατιστικές συναρτήσεις ονομάζονται εκτιμήτριες συναρτήσεις ή εκτιμητές, ενώ οι αριθμητικές τιμές που προκύπτουν από την εφαρμογή τους στο δείγμα ονομάζονται εκτίμηση της άγνωστης παραμέτρου θ. Καθώς οι εκτιμήτριες συναρτήσεις είναι στατιστικές συναρτήσεις συμπεραίνουμε ότι ένας εκτιμητής είναι μια τυχαία μεταβλητή που δεν εξαρτάται από την άγνωστη παράμετρο θ, ενώ η εκτίμησή της δεν είναι τίποτε άλλο παρά η παρατηρηθείσα τιμή αυτής της τυχαίας μεταβλητής.

Στο σημείο αυτό θα πρέπει να τονίσουμε ότι η εκτίμηση της παραμέτρου θ μπορεί να γίνει με δύο τρόπους: με την εκτίμηση σε σημείο και την εκτίμηση σε διάστημα. Η εκτίμηση σε σημείο αναφέρεται σε εκείνες τις μεθοδολογίες που οδηγούν στην επιλογή μιας στατιστικής συνάρτησης που μας βοηθά στην προσέγγιση-εκτίμηση της άγνωστης πληθυσμιακής παραμέτρου και στη συνέχεια στον υπολογισμό της τιμής της στατιστικής συνάρτησης για τα συγκεκριμένα δεδομένα. Από την άλλη πλευρά, η εκτίμηση σε διάστημα έγκειται στην κατασκευή ενός διαστήματος και στη συνέχεια στον υπολογισμό των άκρων αυτού του διαστήματος για τα συγκεκριμένα δεδομένα με στόχο να είμαστε π.χ. 95% βέβαιοι ότι η αληθινή παράμετρος είναι εντός αυτού του διαστήματος. Στη συνέχεια αυτής της ενότητας, το ενδιαφέρον μας επικεντρώνεται στην εκτίμηση σε σημείο, ενώ ο τρόπος εκτίμησης σε διάστημα μιας άγνωστης παραμέτρου και η ερμηνεία του αποτελέσματος που προκύπτει αποτελούν αντικείμενο μελέτης της επόμενης ενότητας.

Έστω, λοιπόν, ότι έχουμε στη διάθεσή μας ένα τυχαίο δείγμα $X_1,\ldots,X_n$ από έναν πληθυσμό με σππ ή σπ $f(x;\theta)$, όπου $\theta \in \Theta$ είναι η άγνωστη παράμετρος. Στόχος της εκτίμησης σε σημείο είναι η εύρεση μιας τιμής που υπολογίζεται από το δείγμα, μέσω των στατιστικών συναρτήσεων, η οποία θα εκτιμά την άγνωστη πληθυσμιακή παράμετρο. Ένα πρώτο εύλογο ερώτημα, που ίσως έχει προκύψει, είναι αν για την εκτίμηση μιας άγνωστης παραμέτρου μπορεί να υπάρχουν διάφοροι υποψήφιοι εκτιμητές. Η απάντηση σε αυτό το ερώτημα θα δοθεί μέσω του παραδείγματος που αναφέρθηκε στην αρχή αυτής της ενότητας και αφορά τον δείκτη νοημοσύνης των παιδιών προσχολικής ηλικίας. Ας επικεντρώσουμε αρχικά το ενδιαφέρον μας στην εκτίμηση της πληθυσμιακής μέσης τιμής. Τότε είναι λογικό κάποιος να προτείνει ως εκτιμητή της πληθυσμιακής μέσης τιμής την αντίστοιχη δειγματική ποσότητα, δηλαδή τη δειγματική μέση τιμή ($\bar{X}$). Είναι, όμως, η μόνη εκτιμήτρια που μπορεί να προταθεί; Η απάντηση είναι αρνητική καθώς κάποιος θα μπορούσε να προτείνει τη δειγματική διάμεσο ή ακόμη και τον μέσο όρο της ελάχιστης και μέγιστης δειγματικής τιμής. Το ίδιο φαινόμενο προφανώς ισχύει και για την εκτίμηση της άγνωστης πληθυσμιακής διακύμανσης όπου κάποιος θα μπορούσε να προτείνει τη δειγματική διακύμανση ή το δειγματικό εύρος.

Από τη συζήτηση που προηγήθηκε γίνεται αντιληπτό ότι δεν υπάρχει ένας μοναδικός εκτιμητής για κάθε άγνωστη παράμετρο. Για τον λόγο αυτό, όπως χαρακτηριστικά αναφέρουν οι Lehmann and Casella (1998) με στόχο τον περιορισμό των πιθανών εκτιμητών κάποιος/α μπορεί να εισάγει κριτήρια και ιδιότητες που ένας εκτιμητής πρέπει να πληροί και βάσει των οποίων πρέπει να επιλέγεται, αλλά και μεθόδους εύρεσης εκτιμητριών συναρτήσεων. Είναι φανερό ότι επιθυμούμε να επιλέξουμε έναν εκτιμητή που να έχει τιμή κοντά στην αληθινή άγνωστη προς εκτίμηση παράμετρο. Επομένως, αρχικά θα πρέπει να καθοριστεί ένας τρόπος μέτρησης της απόστασης του εκτιμητή από την εκτιμώμενη ποσότητα και, στη συνέχεια, προσδιορισμού πότε αυτή η απόσταση θα θεωρείται κοντινή και πότε όχι. Αν $\widehat{\theta} := \widehat{\theta}(X_1, \ldots, X_n)$ ένας εκτιμητής της παραμέτρου θ, τότε ως απόσταση μεταξύ της εκτίμησης από την εκτιμώμενη τιμή θα μπορούσε να χρησιμοποιηθεί είτε το απόλυτο σφάλμα $|\widehat{\theta} - \theta|$ είτε το τετραγωνικό σφάλμα $\left(\widehat{\theta} - \theta\right)^2$ είτε οποιαδήποτε άλλη συνάρτηση μπορεί να χρησιμοποιηθεί ως ένα είδος απόστασης. Ωστόσο κάτι τέτοιο δεν μπορεί να γίνει, γιατί η παράμετρος θ είναι άγνωστη, ενώ ο εκτιμητής $\widehat{\theta}$ είναι τυχαία μεταβλητή και δεν λαμβάνει συγκεκριμένη τιμή. Για να ξεπεραστούν τα παραπάνω προβλήματα θεωρούμε την αναμενόμενη τιμή της απόστασης του εκτιμητή από την εκτιμώμενη τιμή και οδηγούμαστε στους ακόλουθους ορισμούς.

Ορισμός 10.1

Έστω $\widehat{\theta} := \widehat{\theta}(X_1, \ldots, X_n)$ ένας εκτιμητής της παραμέτρου θ. Το Μέσο Απόλυτο Σφάλμα (ΜΑΣ) και το Μέσο Τετραγωνικό Σφάλμα (ΜΤΣ) του $\widehat{\theta}$ ως εκτιμητή της παραμέτρου θ ορίζεται από τις σχέσεις:

$$\text{ΜΑΣ}(\widehat{\theta}, \theta) = E\left[\left|\widehat{\theta} - \theta\right|\right], \tag{10.1}$$

και

$$\text{ΜΤΣ}(\widehat{\theta}, \theta) = E\left[\left(\widehat{\theta} - \theta\right)^2\right], \tag{10.2}$$

αντίστοιχα. Τέλος, η θετική τετραγωνική ρίζα του μέσου τετραγωνικού σφάλματος καλείται τυπικό σφάλμα του $\widehat{\theta}$ ως εκτιμητή της παραμέτρου θ.

Στη στατιστική βιβλιογραφία έχει καθιερωθεί να υιοθετείται ως κριτήριο αξιολόγησης των εκτιμητών το μέσο τετραγωνικό σφάλμα. Είναι προφανές τότε ότι μεταξύ δύο εκτιμητών, έστω $\widehat{\theta}_1$ και $\widehat{\theta}_2$, του θ καλύτερος είναι ο $\widehat{\theta}_1$, αν είναι τέτοιος ώστε:

$$\text{ΜΤΣ}(\widehat{\theta}_1, \theta) \leq \text{ΜΤΣ}(\widehat{\theta}_2, \theta), \text{ για κάθε } \theta \in \Theta$$

και

$$\text{ΜΤΣ}(\widehat{\theta}_1, \theta_0) < \text{ΜΤΣ}(\widehat{\theta}_2, \theta_0), \text{ για τουλάχιστον ένα } \theta_0 \in \Theta.$$

Το ερώτημα που εύκολα προκύπτει είναι αν μπορεί να προσδιοριστεί μεταξύ των πιθανών εκτιμητών μιας παραμέτρου ο εκτιμητής εκείνος που είναι καλύτερος από όλους με βάση το κριτήριο του μέσου τετραγωνικού σφάλματος ή, όπως αλλιώς θα λέγαμε, είναι ο βέλτιστος εκτιμητής. Δυστυχώς, η απάντηση στην ερώτηση αυτή είναι αρνητική (βλ. την Πρόταση 1.4.1 στο σύγγραμμα του Ηλιόπουλος, 2006) και αναδεικνύει έτσι την αναγκαιότητα εισαγωγής ενός άλλου κριτηρίου για την αξιολόγηση ενός εκτιμητή.

Ειδικότερα, από τη σχέση (10.2) εύκολα προκύπτει, χρησιμοποιώντας τις ιδιότητες της μέσης τιμής και της διακύμανσης, ότι:

$$\text{ΜΤΣ}(\widehat{\theta}, \theta) = Var(\widehat{\theta}) + \left\{E\left(\widehat{\theta}\right) - \theta\right\}^2. \tag{10.3}$$

Η παραπάνω σχέση είναι αυτή που μας οδηγεί στο να περιοριζόμαστε σε εκτιμητές που μηδενίζουν τον δεύτερο όρο της, δηλαδή τον όρο $E\left(\widehat{\theta}\right) - \theta$, και, στη συνέχεια, να επιλέγουμε τον εκτιμητή εκείνον με τη μικρότερη διακύμανση. Πιο συγκεκριμένα, μια πρώτη ιδιότητα που ένας εκτιμητής πρέπει να ικανοποιεί είναι το κριτήριο της αμεροληψίας, που ορίζεται στη συνέχεια.

Ορισμός 10.2

Ένας εκτιμητής της παραμέτρου θ, έστω $\widehat{\theta} := \widehat{\theta}(X_1, \ldots, X_n)$, λέμε ότι είναι αμερόληπτος εκτιμητής της, αν

$$\mathrm{E}(\widehat{\theta}) = \theta, \ \forall \ \theta \in \Theta,$$

ενώ διαφορετικά λέμε ότι είναι μεροληπτικός και η διαφορά $\mathrm{E}(\widehat{\theta}) - \theta$ ονομάζεται **ποσό μεροληψίας**.

Παρατήρηση 10.1

Η αμεροληψία εκφράζει την ιδιότητα η αναμενόμενη τιμή της εκτιμήτριας να ισούται με την άγνωστη παράμετρο. Αυτό πρακτικά σημαίνει ότι αν επαναλάβουμε τη δειγματοληψία έναν μεγάλο αριθμό φορών, έστω B, και κάθε φορά καταγράφουμε την τιμή της εκτιμήτριας συνάρτησης, τότε η μέση τιμή αυτών των καταγεγραμμένων τιμών θα πλησιάζει την άγνωστη παράμετρο.

Στην επόμενη πρόταση διατυπώνεται ότι η δειγματική μέση τιμή και η δειγματική διακύμανση είναι πάντοτε αμερόληπτες εκτιμήτριες της πληθυσμιακής μέσης τιμής και διακύμανσης, αντίστοιχα.

Πρόταση 10.1

Έστω $X_1, \ldots, X_n$ ένα τυχαίο δείγμα από έναν πληθυσμό με πεπερασμένη μέση τιμή μ και διακύμανση σ^2. Είναι τότε

$$E(\bar{X}) = \mu \text{ και } E(S^2) = \sigma^2.$$

Απόδειξη Πρότασης 10.1

Τα αποτελέσματα αυτά έχουν αποδειχθεί στην Πρόταση 9.1 και στην Πρόταση 9.3, αντίστοιχα.

Παρατήρηση 10.2

Από την προηγούμενη πρόταση προκύπτει άμεσα ότι ένας αμερόληπτος εκτιμητής της άγνωστης πιθανότητας επιτυχίας p ενός διωνυμικού τυχαίου πειράματος είναι η εκτιμήτρια συνάρτηση $\widehat{P} = \frac{1}{n}\sum_{i=1}^{n} X_i$, όπου X_i η τ.μ. που παριστάνει την έκβαση της i-οστής δοκιμής του διωνυμικού τυχαίου πειράματος, δηλαδή $X_i \sim B(1,p)$ με την τιμή 1 να αντιστοιχεί σε επιτυχία και την τιμή 0 σε αποτυχία. Το αποτέλεσμα προκύπτει άμεσα αρκεί να παρατηρήσετε ότι $E(X_i) = p$. Τέλος, ένας αμερόληπτος εκτιμητής της παραμέτρου λ της κατανομής Poisson είναι η δειγματική μέση τιμή, καθώς σε αυτήν την περίπτωση η παράμετρος λ ταυτίζεται με την πληθυσμιακή μέση τιμή.

Ένα λογικό ερώτημα που μπορεί να έχει προκύψει είναι αν το κριτήριο της αμεροληψίας βοηθά στο να προσδιορίζουμε εκτιμήτρια συνάρτηση με την έννοια αν υπάρχει μοναδική εκτιμήτρια που το πληροί. Η απάντηση στο παραπάνω ερώτημα, δυστυχώς, είναι αρνητική και μπορεί να το διαπιστώσουμε εύκολα καθώς εκτός από τη δειγματική μέση τιμή που είναι αμερόληπτη εκτιμήτρια της πληθυσμιακής μέσης τιμής οποιοδήποτε ημιάθροισμα δύο εκ των $X_1,\ldots,X_n$ είναι αμερόληπτος εκτιμητής της μ. Ειδικότερα, ισχύει το ακόλουθο γενικό αποτέλεσμα.

Πρόταση 10.2

Το σύνολο των αμερόληπτων εκτιμητών μιας παραμέτρου θ είναι είτε το κενό είτε μονοσύνολο είτε μη αριθμήσιμο.

Απόδειξη Πρότασης 10.2

Για την απόδειξη παραπέμπουμε στο ηλεκτρονικό σύγγραμμα των Κουρούκλης κ.ά. (2015).

Η παραπάνω πρόταση, αλλά και η συζήτηση που προηγήθηκε αυτής, μας οδηγεί στο συμπέρασμα ότι το κριτήριο της αμεροληψίας από μόνο του δεν αρκεί για την επιλογή ενός εκτιμητή μιας παραμέτρου. Για τον λόγο αυτό επιθυμούμε μια εκτιμήτρια συνάρτηση να πληροί και άλλες ιδιότητες. Για παράδειγμα, λαμβάνοντας υπόψη τη σχέση (10.3) μεταξύ των αμερόληπτων εκτιμητών επιλέγουμε εκείνον που έχει τη μικρότερη διακύμανση. Ένας τέτοιος εκτιμητής, αν υπάρχει, λέμε ότι είναι **Αμερόληπτος Ομοιομόρφως Ελάχιστης Διακύμανσης** (Α.Ο.Ε.Δ.).

Για περισσότερα κριτήρια αξιολόγησης ενός εκτιμητή, όπως αυτά για παράδειγμα της επάρκειας και της συνέπειας, καθώς και για τρόπους εύρεσης εκτιμητών που πληρούν αυτά και όσα προηγούμενα αναφέρθηκαν παραπέμπουμε σε συγγράμματα Μαθηματικής Στατιστικής, όπως των Παπαϊωάννου και Φερεντίνος (2000), Ηλιόπουλος (2006) και Κουρούκλης κ.ά. (2015). Στο πλαίσιο του παρόντος συγγράμματος αναφέρουμε απλώς ότι με την υιοθέτηση αυτών των μεθοδολογιών προκύπτουν τα ακόλουθα συμπεράσματα:

- οι καλύτεροι εκτιμητές των παραμέτρων μ και σ^2 της κανονικής κατανομής είναι η δειγματική μέση τιμή $\bar{X}$ και η δειγματική διακύμανση S^2, αντίστοιχα,
- ο καλύτερος εκτιμητής της άγνωστης διωνυμικής πιθανότητας p είναι η εκτιμήτρια συνάρτηση $\widehat{P} = \frac{1}{n}\sum_{i=1}^{n} X_i$, όπου X_i η τ.μ. που παριστάνει την έκβαση της i-οστής δοκιμής Bernoulli, δηλαδή $X_i \sim B(1,p)$,
- ο καλύτερος εκτιμητής της παραμέτρου λ της κατανομής Poisson είναι η δειγματική μέση τιμή.

Παρατήρηση 10.3

Σε όσα προηγήθηκαν ξεκινήσαμε διαισθητικά και ελέγξαμε αν η δειγματική μέση τιμή και η δειγματική διακύμανση πληρούν ένα από τα επιθυμητά κριτήρια ενός εκτιμητή. Κάποιος/α ίσως διερωτάται αν υπάρχουν εκτός από επιθυμητά κριτήρια που πρέπει ένας εκτιμητής να πληροί και θεμελιωμένες μεθοδολογίες εύρεσης εκτιμητριών συναρτήσεων. Η απάντηση είναι καταφατική και, μάλιστα, στη βιβλιογραφία έχουν εισαχθεί αρκετές μεθοδολογίες εύρεσης εκτιμητριών συναρτήσεων. Οι πιο δημοφιλείς από αυτές είναι η μέθοδος των ροπών και η μέθοδος της μέγιστης πιθανοφάνειας. Η πρώτη βασίζεται στην επίλυση του συστήματος που προκύπτει εξισώνοντας τις πληθυσμιακές ροπές με τις αντίστοιχες δειγματικές ροπές, ενώ η δεύτερη βασίζεται στην εύρεση της τιμής της παραμέτρου θ που μεγιστοποιεί την από κοινού σππ ή σπ των $X_1,\ldots,X_n$. Οι μεθοδολογίες αυτές δεν αποτελούν αντικείμενο μελέτης αυτού του συγγράμματος και παραπέμπουμε ενδεικτικά στα συγγράμματα των Παπαϊωάννου και Φερεντίνος (2000), Ηλιόπουλος (2006) και Κουρούκλης κ.ά. (2015).

10.3 Εκτίμηση σε διάστημα - Διαστήματα εμπιστοσύνης

Στην προηγούμενη ενότητα είδαμε ότι προκειμένου να εκτιμήσουμε σε σημείο μια άγνωστη πληθυσμιακή παράμετρο θ χρησιμοποιούμε μια εκτιμήτρια $\widehat{\theta}$, η οποία είναι στην ουσία τυχαία μεταβλητή και ακολουθεί κάποια κατανομή (άλλοτε μπορούμε να την προσδιορίσουμε με ακρίβεια και άλλοτε ασυμπτωτικά). Τότε η σημειακή εκτίμηση είναι η τιμή που προκύπτει υπολογίζοντας την εκτιμήτρια για τις παρατηρηθείσες τιμές $x_1,x_2,\ldots,x_n$ των τυχαίων μεταβλητών $X_1,X_2,\ldots,X_n$. Αυτό ουσιαστικά σημαίνει ότι κάθε φορά που λαμβάνουμε ένα δείγμα από τον υπό μελέτη πληθυσμό είναι πάρα πολύ πιθανό να παίρνουμε μια διαφορετική τιμή της εκτιμήτριας συνάρτησης.

Για να κατανοήσουμε την παραπάνω παρατήρηση ας θεωρήσουμε το ακόλουθο παράδειγμα. Έστω ότι ο πληθυσμός που μελετάμε είναι τα παιδιά προσχολικής ηλικίας της Ελλάδας. Αν υποθέσουμε ότι ο πληθυσμός αυτός έχει $N = 10000$ μέλη και εμείς έχουμε αποφασίσει να λάβουμε ένα δείγμα μεγέθους $n = 100$ από αυτόν, τότε υπάρχουν $\binom{N}{n} = 6.520847 \times 10^{241}$ το πλήθος τέτοια δυνατά δείγματα. Προφανώς, αν διαφοροποιείται έστω και μια τιμή μεταξύ δύο εξ αυτών των δειγμάτων, που είναι και απόλυτα αναμενόμενο, θα έχουμε διαφορετική εκτίμηση για την πληθυσμιακή μέση τιμή του δείκτη νοημοσύνης των παιδιών προσχολικής ηλικίας. Μπορούμε λοιπόν ουσιαστικά να έχουμε τόσους διαφορετικούς εκτιμητές για το θ όσα και τα δυνατά δείγματα μεγέθους n που μπορούμε να πάρουμε από τον συγκεκριμένο πληθυσμό.

Από την παραπάνω συζήτηση προκύπτει ότι η εκτίμηση μιας άγνωστης παραμέτρου σε σημείο είναι παρακινδυνευμένη καθώς ακόμα και μη μεροληπτικοί εκτιμητές είναι αδύνατο να εκτιμήσουν ακριβώς την παράμετρο του πληθυσμού που μας ενδιαφέρει (βλ. Ζωγράφος, 2002). Από την άλλη πλευρά, είναι αλήθεια ότι η ακρίβεια της εκτίμησης αυξάνεται με την αύξηση του μεγέθους του δείγματος, αλλά όσο μεγάλο κι αν είναι το δείγμα δεν μπορούμε να περιμένουμε ότι η σημειακή εκτίμηση θα είναι ακριβώς ίση με την παράμετρο του πληθυσμού. Τα παραπάνω κατ΄ ουσίαν οδήγησαν στην εκτίμηση των παραμέτρων με διάστημα (interval estimate) και στην εισαγωγή της έννοιας του διαστήματος εμπιστοσύνης που ορίζεται στη συνέχεια (βλ. Mood *et al.*, 1974).

Ορισμός 10.3

Έστω $X_1,\ldots,X_n$ τυχαίο δείγμα από έναν πληθυσμό με σππ ή σπ $f(x;\theta)$. Επιπλέον, έστω $T_1(X_1,\ldots,X_n)$ και $T_2(X_1,\ldots,X_n)$ στατιστικές συναρτήσεις με $T_1(X_1,\ldots,X_n) \leq T_2(X_1,\ldots,X_n)$, τέτοιες ώστε

$$P(T_1(X_1,\ldots,X_n) < g(\theta) < T_2(X_1,\ldots,X_n)) = 1-\alpha, \text{ για κάθε } \theta \in \Theta. \quad (10.4)$$

Τότε λέμε ότι το τυχαίο διάστημα $(T_1(X_1,\ldots,X_n), T_2(X_1,\ldots,X_n))$ είναι ένα $100(1-\alpha)\%$ διάστημα εμπιστοσύνης (δ.ε.) ή ένα τυχαίο διάστημα με συντελεστή εμπιστοσύνης $100(1-\alpha)\%$ για την παραμετρική συνάρτηση $g(\theta)$, όπου $1-\alpha$ είναι προκαθορισμένη μεγάλη πιθανότητα που ονομάζεται **επίπεδο εμπιστοσύνης**. Επιπλέον, η πιθανότητα στη σχέση (10.4) ονομάζεται πιθανότητα κάλυψης. Τέλος, μια τιμή $(T_1(x_1,\ldots,x_n), T_2(x_1,\ldots,x_n))$ του τυχαίου διανύσματος $(T_1(X_1,\ldots,X_n), T_2(X_1,\ldots,X_n))$ λέγεται επίσης ένα $100(1-\alpha)\%$ διάστημα εμπιστοσύνης για την $g(\theta)$.

Σε ορισμένες περιπτώσεις, συνήθως σε διακριτές κατανομές, δεν είναι πάντοτε δυνατόν για κάθε α να βρούμε δ.ε. τέτοιο, ώστε

$$P(T_1(X_1,\ldots,X_n) < \theta < T_2(X_1,\ldots,X_n)) = 1-\alpha, \forall\, \theta \in \Theta.$$

Σε αυτές τις περιπτώσεις βρίσκουμε στατιστικές συναρτήσεις τέτοιες, ώστε:

$$P(T_1(X_1,\ldots,X_n) < \theta < T_2(X_1,\ldots,X_n)) \geq 1-\alpha, \forall\, \theta \in \Theta.$$

Όταν ισχύει η ισότητα μιλάμε για $100(1-\alpha)\%$ ακριβές δ.ε., ενώ διαφορετικά για δ.ε. τουλάχιστον $100(1-\alpha)\%$. Σε όσα ακολουθούν σε αυτό το σύγγραμμα θα παραλείπεται ο προσδιορισμός «ακριβές». Τέλος, συχνά χρησιμοποιούμε στατιστικές συναρτήσεις που είναι τέτοιες, ώστε για μεγάλο μέγεθος δείγματος να ισχύει ότι:

$$P(T_1(X_1,\ldots,X_n) < \theta < T_2(X_1,\ldots,X_n)) \approx 1-\alpha, \forall\, \theta \in \Theta.$$

Σε αυτές τις περιπτώσεις λέμε ότι προκύπτει ένα $100(1-\alpha)\%$ ασυμπτωτικό δ.ε.

Παρατήρηση 10.4

Από τον Ορισμό 10.3 προκύπτει ότι αν το τυχαίο διάνυσμα $(T_1(X_1,\ldots,X_n), T_2(X_1,\ldots,X_n))$ είναι ένα $100(1-\alpha)\%$ διάστημα εμπιστοσύνης (δ.ε.) για μια παράμετρο θ, τότε δεν είναι λάθος να πούμε ότι υπάρχει πιθανότητα $100(1-\alpha)\%$ η παράμετρος θ να ανήκει σε αυτό το διάστημα. Από την άλλη πλευρά, αν $(T_1(x_1,\ldots,x_n), T_2(x_1,\ldots,x_n))$ είναι μια τιμή του παραπάνω τυχαίου διανύσματος, είναι λάθος να πούμε ότι υπάρχει πιθανότητα $100(1-\alpha)\%$ να ισχύει $T_1(x_1,\ldots,x_n) < \theta < T_2(x_1,\ldots,x_n)$. Αυτό ισχύει καθώς όλες οι ποσότητες που εμφανίζονται στην παραπάνω ανισότητα είναι αριθμοί και, επομένως, δεν υπάρχει πουθενά τυχαιότητα!

Η ακριβής ερμηνεία του διαστήματος εμπιστοσύνης είναι ότι αν θα μπορούσαμε να λάβουμε πάρα πολλά (θεωρητικά άπειρα) τυχαία δείγματα μεγέθους n από τον υπό μελέτη πληθυσμό και να κατασκευάσουμε ένα δ.ε. για κάθε δείγμα, τότε το (περίπου) $100(1-\alpha)\%$ αυτών θα περιέχουν την πραγματική τιμή της παραμέτρου. Το διάστημα εμπιστοσύνης $(T_1(x_1,\ldots,x_n), T_2(x_1,\ldots,x_n))$ που υπολογίσαμε είναι απλώς ένα από τα άπειρα διαφορετικά διαστήματα που θα μπορούσε να τύχει να έχουμε υπολογίσει. Επιπρόσθετα, μπορούμε να πούμε ότι είμαστε $100(1-\alpha)\%$ σίγουροι ότι ισχύει $T_1(x_1,\ldots,x_n) < \theta < T_2(x_1,\ldots,x_n)$ (βλ. Κουτρουβέλης, 2000; Ηλιόπουλος, 2006).

Μετά τον ορισμό του δ.ε. και την ερμηνεία του, προκύπτει το ερώτημα για το πώς κατασκευάζεται ένα δ.ε. Στη συνέχεια, θα παρουσιαστεί εν συντομία και για λόγους πληρότητας μια γενική μέθοδος κατασκευής δ.ε., ενώ για λεπτομέρειες παραπέμπουμε τον/την ενδιαφερόμενο/η αναγνώστη/στρια σε συγγράμματα Μαθηματικής Στατιστικής, όπως των Παπαϊωάννου και Φερεντίνος (2000), Ηλιόπουλος (2006) και Κουρούκλης κ.ά. (2015).

Η γενική μέθοδος κατασκευής των δ.ε. είναι η εξής:

1. Βρίσκουμε μία συνάρτηση των δεδομένων που εξαρτάται από την άγνωστη παράμετρο θ, έστω $h(X_1,\dots,X_n;\theta)$, της οποίας η κατανομή είναι γνωστή και δεν εξαρτάται από το θ.
2. Στη συνέχεια, προσδιορίζουμε σταθερές c_1, c_2, με $c_1 < c_2$ (που εξαρτώνται από την κατανομή της $h(X_1,\dots,X_n;\theta)$, αλλά όχι από το θ), για τις οποίες ισχύει η σχέση:

$$P(c_1 < h(X_1,\dots,X_n;\theta) < c_2) = 1-\alpha, \forall\, \theta \in \Theta.$$

3. Αντιστρέφουμε την παραπάνω διπλή ανισότητα ως προς θ (για τον λόγο αυτόν η τυχαία μεταβλητή $h(X_1,\dots,X_n;\theta)$ όταν πληροί και αυτήν την ιδιότητα αναφέρεται στη βιβλιογραφία ως **αντιστρεπτή ποσότητα** ή **ποσότητα οδηγός**) και εφόσον αυτό επιτυγχάνεται οδηγούμαστε στη σχέση:

$$P(T_1(X_1,\dots,X_n) < \theta < T_2(X_1,\dots,X_n)) = 1-\alpha, \forall\, \theta \in \Theta,$$

και χρησιμοποιώντας τα διαθέσιμα δεδομένα υπολογίζουμε τις τιμές των άκρων του δ.ε. που προέκυψε.

Παρατήρηση 10.5

Καθώς υπάρχουν άπειρες σταθερές c_1, c_2 που ικανοποιούν τη σχέση που δόθηκε στο βήμα 2 παραπάνω, μία λογική επιλογή για τα c_1, c_2 θα μπορούσε να είναι αυτή που ελαχιστοποιεί το μήκος του διαστήματος. Σε μια τέτοια περίπτωση προκύπτει το λεγόμενο δ.ε. ελαχίστου μήκους. Η ελαχιστοποίηση αυτή δεν είναι πάντοτε εφικτή σε κλειστή αναλυτική μορφή. Σε αυτές τις περιπτώσεις οι σταθερές c_1 και c_2 επιλέγονται έτσι ώστε να ικανοποιούνται οι ακόλουθες σχέσεις:

$$P(h(X_1,\dots,X_n;\theta) < c_1) = \alpha/2 \text{ και } P(h(X_1,\dots,X_n;\theta) > c_2) = \alpha/2.$$

Τα δ.ε. που προκύπτουν με αυτήν την επιλογή, η οποία χρησιμοποιείται συχνότερα στην πράξη καθώς τα δ.ε. ελαχίστου μήκους δεν υπάρχουν πάντοτε, ονομάζονται **διαστήματα εμπιστοσύνης ίσων ουρών**. Ο/Η ενδιαφερόμενος/η αναγνώστης/στρια παραπέμπεται στην εργασία των Ferentinos and Karakostas (2006) και στις εκεί αναφορές σχετικά με το πότε τα δ.ε. ελαχίστου μήκους και ίσων ουρών ταυτίζονται. Στο πλαίσιο αυτού του συγγράμματος θα αναφέρουμε ότι ταυτίζονται όταν η αντιστρεπτή ποσότητα ακολουθεί κανονική κατανομή ή t κατανομή ή όταν είναι συμμετρική και μονοκόρυφη.

Παρατήρηση 10.6

Αν $(T_1(X_1,\dots,X_n), T_2(X_1,\dots,X_n))$ είναι ένα $100(1-\alpha)\%$ δ.ε. για την παράμετρο θ και $g(\theta)$ είναι μια γνησίως μονότονη συνάρτηση της παραμέτρου θ, τότε μπορούμε άμεσα να εξάγουμε το συμπέρασμα ότι το $(g(T_1(X_1,\dots,X_n)), g(T_2(X_1,\dots,X_n)))$ είναι $100(1-\alpha)\%$ δ.ε. για την $g(\theta)$.

Παρατήρηση 10.7

Υπάρχουν περιπτώσεις που αντί να ενδιαφερόμαστε να προσδιορίσουμε ένα δ.ε. για μια άγνωστη παράμετρο θ, ενδιαφερόμαστε για ένα άνω ή κάτω φράγμα της. Για παράδειγμα, όπως αναφέρουν οι Κουρούκλης κ.ά. (2015), κάτι τέτοιο είναι επιθυμητό για τον μέσο χρόνο ζωής ενός οργανισμού. Σε αυτό το πλαίσιο, ένα κάτω φράγμα $L(X_1,\dots,X_n)$ για την παράμετρο θ με συντελεστή εμπιστοσύνης 100 $(1-\alpha)\%$ ορίζεται από τη σχέση:

$$P(L(X_1,\dots,X_n) < \theta) = 1-\alpha, \forall\, \theta \in \Theta,$$

ενώ ένα άνω φράγμα $U(X_1,\dots,X_n)$ για την παράμετρο θ με συντελεστή εμπιστοσύνης 100 $(1-\alpha)\%$ ορίζεται από τη σχέση:

$$P(U(X_1,\dots,X_n) > \theta) = 1-\alpha, \forall\, \theta \in \Theta.$$

Οι τιμές $L(x_1,\dots,x_n)$ και $U(x_1,\dots,x_n)$ των στατιστικών συναρτήσεων $L(X_1,\dots,X_n)$ και $U(X_1,\dots,X_n)$ είναι οι τιμές του κάτω και άνω φράγματος για την παράμετρο θ με συντελεστή εμπιστοσύνης $100(1-\alpha)\%$, αντίστοιχα.
Τα παραπάνω συχνά αναφέρονται και ως **μονόπλευρα δ.ε.**

Στις επόμενες υποενότητες, θα παρουσιάσουμε τα διαστήματα εμπιστοσύνης για διάφορες παραμέτρους ενός ή δύο πληθυσμών, που έχουν προκύψει με εφαρμογή της παραπάνω γενικής μεθόδου.

10.3.1 Διάστημα εμπιστοσύνης για την πληθυσμιακή μέση τιμή

Στην ενότητα αυτήν, το ενδιαφέρον μας επικεντρώνεται στην κατασκευή δ.ε. για τη μέση τιμή μ ενός πληθυσμού όταν έχουμε διαθέσιμο ένα τυχαίο δείγμα $X_1, X_2, \dots, X_n$ από αυτόν τον πληθυσμό. Στο πλαίσιο αυτό, θα διακρίνουμε τρεις περιπτώσεις.

Παρατήρηση 10.8

Σε όλες τις περιπτώσεις που ακολουθούν σε αυτήν αλλά και στις υπόλοιπες ενότητες που πραγματεύονται δ.ε. για τη μέση τιμή ή τη διασπορά ενός πληθυσμού ή τη διαφορά δύο πληθυσμιακών μέσων ή το πηλίκο δύο πληθυσμιακών διασπορών θα υποθέτουμε ότι τα δεδομένα μας δεν περιέχουν ακραίες τιμές, δηλαδή παρατηρήσεις με πολύ μεγάλες ή πολύ μικρές τιμές σε σχέση με τις υπόλοιπες.

Περίπτωση Ι. Κανονικός πληθυσμός με γνωστή πληθυσμιακή διασπορά σ.

Σε αυτήν την περίπτωση, η αντιστρεπτή ποσότητα είναι η στατιστική συνάρτηση

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1),$$

οπότε, σύμφωνα με τη γενική μεθοδολογία κατασκευής δ.ε. που περιγράψαμε προηγουμένως, θα έχουμε ότι:

$$P\left(c_1 < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < c_2\right) = 1-\alpha,\ \forall\, \mu \in R,$$

ή, ισοδύναμα, μετά από απλές αλγεβρικές πράξεις:

$$P\left(\bar{X} - \frac{c_2\cdot\sigma}{\sqrt{n}} < \mu < \bar{X} - \frac{c_1\cdot\sigma}{\sqrt{n}}\right) = 1-\alpha,\ \forall\, \mu \in R.$$

Οι σταθερές c_1 και c_2 προσδιορίζονται έτσι ώστε να ελαχιστοποιείται το μήκος του διαστήματος $l = (c_2 - c_1)\frac{\sigma}{\sqrt{n}}$ (δ.ε. ελαχίστου μήκους) είτε έτσι ώστε $P\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < c_1\right) = \alpha/2$ και $P\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} > c_2\right) = \alpha/2$ (δ.ε. ίσων ουρών). Καθώς η αντιστρεπτή ποσότητα ακολουθεί τυπική κανονική κατανομή (άρα είναι συμμετρική και μονοκόρυφη) τα δύο διαστήματα ταυτίζονται. Επομένως, προκύπτει ότι $c_1 = z_{1-\alpha/2} = -z_{\alpha/2}$, ενώ $c_2 = z_{\alpha/2}$.

Συνδυάζοντας τα παραπάνω, καταλήγουμε ότι το $100(1-\alpha)\%$ δ.ε. ελαχίστου μήκους και ίσων ουρών για τη μέση τιμή μ ενός κανονικού πληθυσμού με γνωστή διασπορά σ^2 είναι το:

$$\left(\bar{X} - z_{a/2}\frac{\sigma}{\sqrt{n}},\ \bar{X} + z_{a/2}\frac{\sigma}{\sqrt{n}}\right). \tag{10.5}$$

Λαμβάνοντας υπόψη τον Ορισμό 10.3, μια τιμή του παραπάνω τυχαίου διανύσματος που δίνεται από τη σχέση

$$\left(\bar{x} - z_{a/2}\frac{\sigma}{\sqrt{n}},\ \bar{x} + z_{a/2}\frac{\sigma}{\sqrt{n}}\right) \tag{10.6}$$

είναι ένα $100(1-\alpha)\%$ δ.ε. για τη μέση τιμή μ ενός κανονικού πληθυσμού με γνωστή διασπορά σ^2.

Παρατήρηση 10.9

Το πλάτος ενός διαστήματος εμπιστοσύνης καθορίζει την ακρίβειά του, καθώς όσο μικρότερο είναι τόσο ακριβέστερο είναι το διάστημα εμπιστοσύνης. Από τη σχέση (10.5) εύκολα προκύπτει ότι το πλάτος του δ.ε. ελαχίστου μήκους για τη μέση τιμή μ ενός κανονικού πληθυσμού με γνωστή διασπορά ισούται με $d = \frac{2 \cdot z_{\alpha/2}\ \sigma}{\sqrt{n}}$ και, επομένως, εξαρτάται από την τυπική απόκλιση του πληθυσμού, το μέγεθος του δείγματος και το επίπεδο εμπιστοσύνης.

Παρατηρήστε ότι όσο μεγαλύτερο είναι το μέγεθος του δείγματος, τόσο μικρότερο γίνεται το πλάτος του διαστήματος εμπιστοσύνης, όταν οι υπόλοιπες ποσότητες παραμένουν σταθερές. Από την άλλη πλευρά, για σταθερό μέγεθος δείγματος και τυπική απόκλιση όσο αυξάνεται το επίπεδο εμπιστοσύνης, τόσο μεγαλώνει το πλάτος του δ.ε., δηλαδή μειώνεται η ακρίβεια του διαστήματος. Με άλλα λόγια, αυτό που κερδίζουμε σε εμπιστοσύνη το χάνουμε στην ακρίβεια του διαστήματος.

Παρατήρηση 10.10

Πολύ συχνά θέλουμε να υπολογίσουμε το κατάλληλο μέγεθος δείγματος προκειμένου να εξαχθούν συμπεράσματα με προκαθορισμένη ακρίβεια. Πιο συγκεκριμένα, συχνά επιλέγουμε μέγεθος δείγματος, τέτοιο ώστε το πλάτος του δ.ε. για τη μέση τιμή μ ενός κανονικού πληθυσμού με γνωστή διασπορά σ^2 να μην υπερβαίνει μία προκαθορισμένη σταθερά ϵ. Σε μια τέτοια περίπτωση, ισχύουν τα ακόλουθα:

$$\frac{2\, z_{\alpha/2}\, \sigma}{\sqrt{n}} \leq \epsilon \Rightarrow \sqrt{n} \geq \frac{2\, z_{\alpha/2}\, \sigma}{\epsilon} \Rightarrow n \geq 4\left(\frac{z_{\alpha/2}\, \sigma}{\epsilon}\right)^2.$$

Παράδειγμα 10.1

Σε ένα πείραμα μελέτης του βαθμού οξύτητας (pH) δύο τύπων χημικών διαλυμάτων επιλέχθηκαν στην τύχη 5 διαλύματα τύπου Α και 7 διαλύματα τύπου Β. Η ανάλυσή τους έδωσε τα παρακάτω αποτελέσματα

Διάλυμα Α	6.33	6.28	6.50	6.40	6.45		
Διάλυμα Β	6.51	6.55	6.43	6.51	6.62	6.40	6.48

Να κατασκευαστεί και να ερμηνευτεί ένα 95% δ.ε. για τον μέσο βαθμό pH τόσο του χημικού διαλύματος τύπου Α όσο και του Β, υποθέτοντας ότι οι πληθυσμοί είναι κανονικοί με γνωστές πληθυσμιακές διασπορές 0.008 και 0.005, αντίστοιχα.

Λύση Παραδείγματος 10.1

Έστω X και Y οι τ.μ. που παριστάνουν τον βαθμό οξύτητας του χημικού διαλύματος τύπου Α και Β, αντίστοιχα. Από την εκφώνηση έχουμε ότι $X \sim N(\mu_1, 0.008)$ και $Y \sim N(\mu_2, 0.005)$. Θέλουμε να κατασκευάσουμε 95% δ.ε. για τον μέσο βαθμό pH του διαλύματος Α (μ_1) και του διαλύματος Β (μ_2), στηριζόμενοι στα δείγματα $X_1,\ldots,X_n$ και $Y_1,\ldots,Y_m$ από αυτούς του πληθυσμούς με $n = 5$ και $m = 7$. Τα δ.ε. θα προκύψουν χρησιμοποιώντας τη σχέση (10.6) με $\bar{x} = 6.392$, $\bar{y} = 6.5$, $\alpha = 0.05$ και $z_{\alpha/2} = z_{0.025} = 1.96$. Επομένως, ένα 95% δ.ε. για τον μέσο βαθμό οξύτητας του χημικού διαλύματος τύπου Α είναι το:

$$\left(6.392 - 1.96 \cdot \frac{\sqrt{0.008}}{\sqrt{5}}, 6.392 + 1.96 \cdot \frac{\sqrt{0.008}}{\sqrt{5}}\right) = (6.314, 6.47),$$

ενώ ένα 95% δ.ε. για τον μέσο βαθμό οξύτητας του χημικού διαλύματος τύπου Β είναι:

$$\left(6.5 - 1.96 \cdot \frac{\sqrt{0.005}}{7}, 6.5 + 1.96 \cdot \frac{\sqrt{0.005}}{7}\right) = (6.448, 6.552),$$

αντίστοιχα.

Εναλλακτικά, τα παραπάνω αποτελέσματα μπορούν να ληφθούν χρησιμοποιώντας την εντολή `ci.mu.z` του πακέτου `asbio` της R, όπως φαίνεται παρακάτω.

```
library(asbio)

A <- c(6.33, 6.28, 6.5, 6.40, 6.45)
B <- c(6.51, 6.55, 6.43, 6.51, 6.62, 6.40, 6.48)

ci.mu.z(A, conf = 0.95, sigma = sqrt(0.008))
ci.mu.z(B, conf = 0.95, sigma = sqrt(0.005))
```

Στο πρώτο όρισμα της εντολής `ci.mu.z` θέσαμε τα δεδομένα του δείγματος, στο δεύτερο όρισμα δόθηκε ο συντελεστής εμπιστοσύνης, ενώ στο τρίτο όρισμα δόθηκε η γνωστή πληθυσμιακή τυπική απόκλιση. Τα αποτελέσματα που λαμβάνουμε είναι:

```
95% z Confidence interval for population mean
Estimate     2.5%    97.5%
6.392000 6.313601 6.470399

95% z Confidence interval for population mean
Estimate     2.5%    97.5%
6.500000 6.447618 6.552382
```

Ερμηνεία δ.ε.: Ένα 95% δ.ε. για τον μέσο βαθμό οξύτητας του χημικού διαλύματος τύπου Α είναι το $(6.314,\ 6.47)$. Αυτό σημαίνει ότι αν μπορούσαμε να πάρουμε πάρα πολλά (θεωρητικά άπειρα) τυχαία δείγματα μεγέθους $n = 5$ από τον κανονικό πληθυσμό με μέση τιμή μ_1 και διακύμανση 0.008 και για καθένα υπολογίζαμε το αντίστοιχο 95% δ.ε. για το μ_1, τότε το 95% (περίπου) από αυτά θα περιείχε το μ_1, όποια και αν είναι η πραγματική τιμή του (βλ., για παράδειγμα, Ηλιόπουλος, 2006). Το διάστημα εμπιστοσύνης που υπολογίσαμε είναι απλώς ένα από τα άπειρα διαφορετικά διαστήματα που θα μπορούσε να τύχει να έχουμε υπολογίσει. Αυτό που μπορούμε ωστόσο να πούμε (βλ. Κουτρουβέλης, 2000) είναι ότι είμαστε 95% σίγουροι ότι ισχύει $6.314 < \mu_1 < 6.47$, δηλαδή ότι ο μέσος βαθμός οξύτητας του χημικού διαλύματος τύπου Α βρίσκεται μεταξύ 6.314 και 6.47.

Παρόμοια είναι η ερμηνεία του δ.ε. για τον μέσο βαθμό οξύτητας του χημικού διαλύματος τύπου Β, καθώς και κάθε άλλου δ.ε. που θα εμφανιστεί στο υπόλοιπο αυτού του κεφαλαίου και για τον λόγο αυτό παραλείπεται.

Παρατήρηση 10.11

Στο παράδειγμα που προηγήθηκε, μπορεί κάποιος/α να παρατηρήσει ότι υπάρχουν πάρα πολύ μικρές διαφοροποιήσεις μεταξύ των αποτελεσμάτων με την κλασική μέθοδο επίλυσης και των αποτελεσμάτων με χρήση της R. Παρόμοια συμπεριφορά θα παρατηρηθεί και στα υπόλοιπα παραδείγματα και ασκήσεις αυτού του κεφαλαίου, καθώς οι διαφοροποιήσεις αυτές οφείλονται σε σφάλματα στρογγυλοποίησης στις αριθμητικές πράξεις και στη χρήση των πινάκων του Παραρτήματος Α΄ για τον προσδιορισμό των ποσοστιαίων σημείων των κατανομών.

Παρατήρηση 10.12

Η εντολή `ci.mu.z` μπορεί να χρησιμοποιηθεί για την κατασκευή δ.ε., ακόμα και αν δεν έχουμε το σύνολο των αρχικών δεδομένων, αλλά μας δίνονται η δειγματική μέση τιμή, το μέγεθος δείγματος και, προφανώς, η πληθυσμιακή τυπική απόκλιση. Ειδικότερα, αυτό μπορεί να επιτευχθεί με την παρακάτω εντολή:

```
1 ci.mu.z(conf=0.95,sigma=sqrt(0.005),summarized=TRUE,xbar=6.5, n=7)
```

Άσκηση Αυτοαξιολόγησης 10.1

Από προηγούμενες εκτεταμένες μελέτες είναι γνωστό ότι η ροή (σε κυβικά μέτρα/*sec*) ενός ποταμού ακολουθεί την κανονική κατανομή με τυπική απόκλιση τα 5 κυβικά μέτρα/*sec*. Μετρήθηκε η ροή του ποταμού σε ένα δείγμα 25 μετρήσεων και η μέση τιμή υπολογίστηκε να είναι ίση με 81.2 κυβικά μέτρα/*sec*. Να κατασκευαστεί ένα 95% δ.ε. για τη μέση ροή ύδατος.

Άσκηση Αυτοαξιολόγησης 10.2

Να βρεθεί το μέγεθος n ενός τυχαίου δείγματος από κανονικό πληθυσμό με μέση τιμή μ και διακύμανση 25, ώστε το 96% δ.ε. ελαχίστου μήκους της μέσης τιμής του να έχει πλάτος μικρότερο ή ίσο από 1.

Περίπτωση II. Κανονικός πληθυσμός με άγνωστη πληθυσμιακή διασπορά σ^2.

Στην πράξη τις περισσότερες φορές, αν όχι πάντοτε, η διασπορά του κανονικού πληθυσμού είναι άγνωστη, οπότε σε αυτήν την περίπτωση η στατιστική συνάρτηση $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ δεν μπορεί να χρησιμοποιηθεί, καθώς, εκτός από τη μέση τιμή μ, περιέχει και την άγνωστη τυπική απόκλιση σ. Το πρόβλημα αυτό ξεπερνιέται αντικαθιστώντας την άγνωστη τυπική απόκλιση με την εκτίμησή της. Σε αυτές τις περιπτώσεις έχουμε ότι

$$\frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t_{n-1},$$

όπου $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ και $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$. Εφαρμόζοντας τη γενική μεθοδολογία, με παρόμοιο τρόπο, όπως στην περίπτωση κανονικού πληθυσμού με γνωστή διασπορά, καταλήγουμε ότι το $100(1-\alpha)\%$ δ.ε. ελαχίστου μήκους και ίσων ουρών για τη μέση τιμή μ ενός κανονικού πληθυσμού με άγνωστη διασπορά είναι το:

$$\left(\bar{X} - t_{n-1,\alpha/2}\frac{S}{\sqrt{n}},\ \bar{X} + t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}\right). \tag{10.7}$$

Λαμβάνοντας υπόψη τον Ορισμό 10.3, μια τιμή του παραπάνω τυχαίου διανύσματος που δίνεται από τη σχέση

$$\left(\bar{x} - t_{n-1,\alpha/2}\frac{s}{\sqrt{n}},\ \bar{x} + t_{n-1,\alpha/2}\frac{s}{\sqrt{n}}\right) \tag{10.8}$$

λέγεται, επίσης, ένα $100(1-\alpha)\%$ δ.ε. για τη μέση τιμή μ ενός κανονικού πληθυσμού με άγνωστη διασπορά σ^2.

Παράδειγμα 10.2

Χρησιμοποιώντας τα δεδομένα του Παραδείγματος 10.1 να κατασκευαστεί ένα 95% δ.ε. για τον μέσο βαθμό pH του διαλύματος Α και του διαλύματος Β, υποθέτοντας μόνο ότι οι πληθυσμοί είναι κανονικοί.

Λύση Παραδείγματος 10.2

Καθώς οι πληθυσμιακές διασπορές είναι άγνωστες, θα κατασκευάσουμε δ.ε. για τον μέσο βαθμό οξύτητας των δύο διαλυμάτων χρησιμοποιώντας το δ.ε. ελαχίστου μήκους που δόθηκε στη σχέση (10.8). Μετά από αλγεβρικές πράξεις, προκύπτει ότι η δειγματική μέση τιμή και διακύμανση του δείγματος διαλυμάτων τύπου Α είναι $\bar{x} = 6.392$ και $s_1^2 = 0.00787$, αντίστοιχα, ενώ οι αντίστοιχες τιμές για το δείγμα διαλυμάτων τύπου Β είναι $\bar{y} = 6.5$ και $s_2^2 = 0.0054$. Επιπλέον, είναι $t_{n-1,\alpha/2} = t_{4,0.025} = 2.776$ και $t_{m-1,\alpha/2} = t_{6,0.025} = 2.447$. Επομένως, ένα 95% δ.ε. για τον μέσο βαθμό οξύτητας του χημικού διαλύματος τύπου Α είναι το:

$$\left(6.392 - 2.776 \cdot \frac{\sqrt{0.00787}}{\sqrt{5}}, 6.392 + 2.776 \cdot \frac{\sqrt{0.00787}}{\sqrt{5}}\right) = (6.281866, 6.502134)$$

ενώ το 95% δ.ε. για τον μέσο βαθμό οξύτητας του χημικού διαλύματος τύπου Β είναι το:

$$\left(6.5 - 2.447 \cdot \frac{\sqrt{0.0054}}{\sqrt{7}}, 6.5 + 1.96 \cdot \frac{\sqrt{0.005}}{\sqrt{7}}\right) = (6.432036, 6.567964).$$

Ένας εναλλακτικός τρόπος για να εξαχθούν τα παραπάνω αποτελέσματα είναι χρησιμοποιώντας την εντολή `ci.mu.t` του πακέτου `asbio` της R, όπως φαίνεται παρακάτω.

```
library(asbio)
A <- c(6.33, 6.28, 6.5, 6.40, 6.45)
B <- c(6.51, 6.55, 6.43, 6.51, 6.62, 6.40, 6.48)
ci.mu.t(A, conf = 0.95)
ci.mu.t(B, conf = 0.95)
```

Τότε προκύπτουν τα ακόλουθα αποτελέσματα

```
95% t Confidence interval for population mean
Estimate     2.5%    97.5%
6.392000 6.281848 6.502152

95% t Confidence interval for population mean
Estimate     2.5%    97.5%
6.500000 6.432038 6.567962
```

Παρατήρηση 10.13

Η εντολή `ci.mu.t` μπορεί να χρησιμοποιηθεί για την κατασκευή δ.ε., ακόμα και αν δεν έχουμε το αρχικό σύνολο δεδομένων, αλλά μας δίνονται η δειγματική μέση τιμή, η δειγματική τυπική απόκλιση και το μέγεθος δείγματος, όπως φαίνεται παρακάτω:

```
ci.mu.t(conf=0.95,summarized=TRUE,xbar=6.392, sd=sqrt(0.00787), n=5)
ci.mu.t(conf=0.95,summarized=TRUE,xbar=6.5, sd=sqrt(0.0054), n=7)
```

Άσκηση Αυτοαξιολόγησης 10.3

Χρησιμοποιώντας τα δεδομένα της Άσκησης Αυτοαξιολόγησης 10.1 να κατασκευαστεί ένα 95% δ.ε. για τη μέση ροή ύδατος, υποθέτοντας μόνο ότι ο πληθυσμός είναι κανονικός. Δίνεται ότι η δειγματική τυπική απόκλιση ισούται με 5.2 κυβικά μέτρα/*sec*.

Άσκηση Αυτοαξιολόγησης 10.4

Μηχανή παράγει προϊόντα συγκεκριμένου τύπου, των οποίων το βάρος ακολουθεί κανονική κατανομή. Πέντε προϊόντα τυχαία επιλεγμένα έχουν βάρος 6.6, 4.6, 5.4, 5.8, 5.5. Ποιο είναι το πλάτος του δ.ε. ελαχίστου μήκους επιπέδου εμπιστοσύνης 95% για το μέσο βάρος των προϊόντων;

Περίπτωση ΙΙΙ. Μη κανονικός πληθυσμός και μεγάλο δείγμα.

Στην περίπτωση όπου ο πληθυσμός από τον οποίο προέρχεται το τυχαίο δείγμα δεν είναι κανονικός αλλά το μέγεθος δείγματος είναι μεγάλο (συνήθως $n \geq 30$), τότε, υποθέτοντας ότι έχει πεπερασμένη μέση τιμή μ και διακύμανση σ^2, ισχύει ότι:

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \xrightarrow{d} N(0,1)$$

και, ακολουθώντας τη γενική μέθοδο κατασκευής δ.ε. προκύπτει, ότι ένα ασυμπτωτικό $100(1-\alpha)\%$ δ.ε. για τη μέση τιμή μ ενός πληθυσμού με πεπερασμένη διασπορά σ^2 είναι το:

$$\left(\bar{X} - z_{a/2}\frac{S}{\sqrt{n}},\ \bar{X} + z_{a/2}\frac{S}{\sqrt{n}}\right). \tag{10.9}$$

Προφανώς, αν η πληθυσμιακή διακύμανση είναι γνωστή στην παραπάνω σχέση, αντικαθίσταται η δειγματική τυπική απόκλιση S από την πληθυσμιακή τυπική απόκλιση σ.

Λαμβάνοντας υπόψη τον Ορισμό 10.3, μια τιμή του παραπάνω τυχαίου διανύσματος που δίνεται από τη σχέση:

$$\left(\bar{x} - z_{a/2}\frac{s}{\sqrt{n}},\ \bar{x} + z_{a/2}\frac{s}{\sqrt{n}}\right) \tag{10.10}$$

λέγεται, επίσης, ένα $100(1-\alpha)\%$ ασυμπτωτικό δ.ε. για τη μέση τιμή μ ενός πληθυσμού με πεπερασμένη διασπορά σ^2.

Παράδειγμα 10.3

Μια εταιρεία κατασκευής λαμπτήρων μελετά τον μέσο χρόνο ζωής των λαμπτήρων που κατασκευάζει. Για τον λόγο αυτόν εξέτασε 100 λαμπτήρες και βρήκε ότι ο μέσος χρόνος ζωής τους ήταν ίσος με 2850 ώρες, ενώ η τυπική απόκλιση του χρόνου ζωής τους ήταν ίση με 100 ώρες. Κατασκευάστε ένα (ασυμπτωτικό) 95% δ.ε. για τον πραγματικό μέσο χρόνο ζωής των λαμπτήρων.

Λύση Παραδείγματος 10.3

Έστω X η τ.μ. που παριστάνει τον χρόνο ζωής σε ώρες ενός λαμπτήρα που κατασκευάζει η εταιρεία με μέση τιμή μ και πεπερασμένη διακύμανση σ^2. Ένα ασυμπτωτικό 95% δ.ε. για τη μέση τιμή μ δίνεται από τη σχέση (10.10) με $n = 100$, $\alpha = 0.05$, $\bar{x} = 2850$ και $s^2 = 10000$, δηλαδή είναι το:

$$\left(2850 - 1.96\frac{\sqrt{10000}}{\sqrt{100}},\ 2850 + 1.96\frac{\sqrt{10000}}{\sqrt{100}}\right) = (2830.4, 2869.6).$$

Άσκηση Αυτοαξιολόγησης 10.5

Θέλοντας ένας επιστήμονας να μελετήσει τον χρόνο ζωής του ινδικού χοιριδίου καταγράφει τον χρόνο ζωής σε ημέρες 64 τυχαία επιλεγμένων τέτοιων χοιριδίων. Κατασκευάστε ένα 90% δ.ε. για τον μέσο χρόνο ζωής, όταν ο επιστήμονας έχει υπολογίσει ότι ο μέσος χρόνος ζωής των 64 χοιριδίων είναι 345.2 μέρες με τυπική απόκλιση 222.2 ημέρες.

Παρατήρηση 10.14

Ένα ερώτημα που παραμένει αναπάντητο από τα παραπάνω είναι πώς προσδιορίζεται ένα δ.ε. για τη μέση τιμή ενός πληθυσμού, όταν αυτός δεν είναι κανονικός και το μέγεθος του δείγματος είναι μικρότερο από 30. Σε μία τέτοια περίπτωση, μια λύση είναι να προσπαθήσουμε να βρούμε έναν μετασχηματισμό των δεδομένων έτσι ώστε αυτά να ακολουθούν κανονική κατανομή. Αν κάτι τέτοιο είναι εφικτό, για παράδειγμα χρησιμοποιώντας τον μετασχηματισμό του λογαρίθμου, τότε θα κατασκευαστεί δ.ε. για τον πληθυσμιακό μέσο λογάριθμο και με κατάλληλο μετασχηματισμό θα οδηγηθούμε σε δ.ε. για τον αρχικό πληθυσμιακό μέσο. Στην περίπτωση που η εύρεση μετασχηματισμού που να διορθώνει το πρόβλημα της μη κανονικότητας δεν είναι εφικτή, τότε μια λύση που προτείνεται είναι η χρήση μεθόδων υπολογιστικής στατιστικής (για παράδειγμα της μεθόδου bootstrap). Αυτές οι μεθοδολογίες δεν αποτελούν αντικείμενο μελέτης αυτού του συγγράμματος και ο/η ενδιαφερόμενος/η αναγνώστης/στρια παραπέμπεται στο σύγγραμμα των Efron and Tibshirani (1993) και τις εκεί αναφορές, καθώς και στην ευρέως χρησιμοποιούμενη βιβλιοθήκη `boot` της R.

10.3.2 Διάστημα εμπιστοσύνης για τη διαφορά δύο μέσων τιμών με ανεξάρτητα δείγματα

Έστω $X_1, \ldots, X_n$ ένα τυχαίο δείγμα από τον πληθυσμό με μέση τιμή μ_1 και διασπορά σ_1^2, ενώ $Y_1, \ldots, Y_m$ ένα τυχαίο δείγμα από τον πληθυσμό με μέση τιμή μ_2 και διασπορά σ_2^2. Επιπλέον, υποθέτουμε ότι τα δύο δείγματα είναι ανεξάρτητα. Στην ενότητα αυτήν το ενδιαφέρον επικεντρώνεται στην κατασκευή δ.ε. για τη διαφορά των μέσων τιμών $\mu_1 - \mu_2$. Στο παραπάνω πλαίσιο, θα διακρίνουμε τέσσερις περιπτώσεις.

Περίπτωση Ι. Κανονικοί πληθυσμοί με γνωστές πληθυσμιακές διασπορές.

Σε αυτήν την περίπτωση, η αντιστρεπτή ποσότητα είναι η στατιστική συνάρτηση:

$$\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n + \sigma_2^2/m}} \sim N(0,1),$$

οπότε, σύμφωνα με τη γενική μεθοδολογία κατασκευής δ.ε. που περιγράψαμε προηγουμένως, προκύπτει ότι το $100(1-\alpha)\%$ δ.ε. ελαχίστου μήκους και ίσων ουρών για τη διαφορά των μέσων τιμών $\mu_1 - \mu_2$ δύο κανονικών πληθυσμών με γνωστές πληθυσμιακές διασπορές σ_1^2 και σ_2^2, αντίστοιχα, είναι το:

$$\left(\bar{X} - \bar{Y} - z_{a/2}\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}},\ \bar{X} - \bar{Y} + z_{a/2}\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}} \right). \tag{10.11}$$

Λαμβάνοντας υπόψη τον Ορισμό 10.3, μια τιμή του παραπάνω τυχαίου διανύσματος που δίνεται από τη σχέση

$$\left(\bar{x} - \bar{y} - z_{a/2}\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}},\ \bar{x} - \bar{y} + z_{a/2}\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}} \right), \tag{10.12}$$

το οποίο αναφέρεται ως ένα $100(1-\alpha)\%$ δ.ε. για τη διαφορά των μέσων τιμών $\mu_1 - \mu_2$ δύο κανονικών πληθυσμών με γνωστές πληθυσμιακές διασπορές σ_1^2 και σ_2^2, αντίστοιχα.

Παράδειγμα 10.4: (συνέχεια Παραδείγματος 10.1)

Χρησιμοποιώντας τα δεδομένα και τις υποθέσεις του Παραδείγματος 10.1, κατασκευάστε ένα 95% διάστημα εμπιστοσύνης για τη διαφορά των μέσων βαθμών οξύτητας των δύο διαλυμάτων.

Λύση Παραδείγματος 10.4

Καθώς οι πληθυσμιακές διασπορές είναι γνωστές και οι πληθυσμοί υποθέτουμε ότι είναι κανονικοί, θα κατασκευάσουμε δ.ε. για τη μέση διαφορά των βαθμών οξύτητας των δύο διαλυμάτων, χρησιμοποιώντας το δ.ε. που δόθηκε στη σχέση (10.12) με $\bar{x} = 6.392$, $\bar{y} = 6.5$, $n = 5$, $m = 7$, $\alpha = 0.05$ και $z_{\alpha/2} = z_{0.025} = 1.96$. Επομένως, ένα 95% δ.ε. για τη διαφορά $\mu_1 - \mu_2$ είναι το:

$$\left(6.392 - 6.5 - 1.96\sqrt{\frac{0.008}{5} + \frac{0.005}{7}},\ 6.392 + 6.5 - 1.96\sqrt{\frac{0.008}{5} + \frac{0.005}{7}}\right)$$

ή, μετά από λίγη άλγεβρα, το $(-0.2022898, -0.01371023)$.

Εναλλακτικά, τα παραπάνω αποτελέσματα μπορούν να ληφθούν εκτελώντας τις παρακάτω εντολές στην R.

```
A <- c(6.33, 6.28, 6.5, 6.40, 6.45)
B <- c(6.51, 6.55, 6.43, 6.51, 6.62, 6.40, 6.48)
sample.na <- length(A)
sample.nb <- length(B)
population.vara <- 0.008
population.varb <- 0.005
var.meanerrora <- population.vara/sample.na
var.meanerrorb <- population.varb /sample.nb
alpha = 0.05
z.score = qnorm(p=alpha/2,lower.tail=F)
sample.meana <-mean(A)
sample.meanb <- mean(B)
lower.bound <- (sample.meana- sample.meanb)- z.score * sqrt(var.meanerrora+
    var.meanerrorb)
upper.bound <- (sample.meana- sample.meanb)+ z.score * sqrt(var.meanerrora+
    var.meanerrorb)
cat("CI for mean A-mean B = (",lower.bound,",", upper.bound, ")\n")
```

Τα αποτελέσματα των παραπάνω εντολών είναι τα ακόλουθα:

```
CI for mean A-mean B = (-0.202288 , -0.01371197)
```

Άσκηση Αυτοαξιολόγησης 10.6

Ένα τεστ Αγγλικών διεξάγεται στους μαθητές των πόλεων Α και Β. Από προηγούμενες έρευνες γνωρίζουμε ότι η βαθμολογία των μαθητών στο τεστ Αγγλικών στις δύο πόλεις περιγράφεται ικανοποιητικά από μία κανονική κατανομή με τυπική απόκλιση $\sqrt{400}$ και $\sqrt{500}$ μονάδες, αντίστοιχα. Εξετάζονται στα Αγγλικά 18 και 25 άτομα, αντίστοιχα, από τις πόλεις Α και Β και διαπιστώνεται ότι η μέση βαθμολογία τους είναι 1000 και 985, αντίστοιχα. Να κατασκευάσετε ένα 97% δ.ε. για τη μέση διαφορά των βαθμολογιών των μαθητών των δύο πόλεων.

Περίπτωση ΙΙ. Κανονικοί πληθυσμοί με άγνωστες αλλά ίσες πληθυσμιακές διασπορές.

Στην πράξη, αν όχι πάντοτε, τις περισσότερες φορές οι διασπορές των πληθυσμών είναι άγνωστες. Σε αυτήν την περίπτωση, θα πρέπει αρχικά να ελέγξουμε αν μπορούν να θεωρηθούν ίσες ή όχι (μέσω κατάλληλου στατιστικού ελέγχου ο οποίος θα παρουσιαστεί στο Κεφάλαιο 11) και να επιλέξουμε το αντίστοιχο διάστημα εμπιστοσύνης.

Στην περίπτωση όπου οι άγνωστες διασπορές των κανονικών πληθυσμών μπορούν να θεωρηθούν ίσες, παίρνουμε ως αντιστρεπτή ποσότητα την:

$$\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{S_p\sqrt{1/n + 1/m}} \sim t_{n+m-2}$$

όπου

$$S_p^2 = \frac{(n-1)S_1^2 + (m-1)S_2^2}{n+m-2}, \tag{10.13}$$

είναι ο εκτιμητής της κοινής διακύμανσης $\sigma^2 = \sigma_1^2 = \sigma_2^2$ με S_1^2 και S_2^2 να είναι οι δειγματικές διακυμάνσεις, που προκύπτουν από τα ανεξάρτητα τ.δ. $X_1, \ldots, X_n$ και $Y_1, \ldots, Y_m$, αντίστοιχα. Τότε το $100(1-\alpha)\%$ δ.ε. ελαχίστου μήκους και ίσων ουρών για τη διαφορά των μέσων τιμών $\mu_1 - \mu_2$ δύο κανονικών πληθυσμών με άγνωστες, αλλά ίσες πληθυσμιακές διασπορές $\sigma_1^2 = \sigma_2^2$, είναι το:

$$\left(\bar{X} - \bar{Y} - t_{n+m-2,\alpha/2} \cdot S_p \cdot \sqrt{1/n + 1/m},\ \bar{X} - \bar{Y} + t_{n+m-2,\alpha/2} \cdot S_p \cdot \sqrt{1/n + 1/m}\right). \tag{10.14}$$

Λαμβάνοντας υπόψη τον Ορισμό 10.3, μια τιμή του παραπάνω τυχαίου διανύσματος που δίνεται από τη σχέση:

$$\left(\bar{x} - \bar{y} - t_{n+m-2,\alpha/2} \cdot s_p \cdot \sqrt{1/n + 1/m},\ \bar{x} - \bar{y} + t_{n+m-2,\alpha/2} \cdot s_p \cdot \sqrt{1/n + 1/m}\right) \tag{10.15}$$

είναι ένα $100(1-\alpha)\%$ δ.ε. για τη διαφορά των μέσων τιμών $\mu_1 - \mu_2$ δύο κανονικών πληθυσμών με άγνωστες, αλλά ίσες πληθυσμιακές διασπορές $\sigma_1^2 = \sigma_2^2$.

Παράδειγμα 10.5

Χρησιμοποιώντας τα δεδομένα του Παραδείγματος 10.1, να κατασκευαστεί ένα 95% δ.ε. για τη διαφορά των μέσων βαθμών οξύτητας των δύο διαλυμάτων, υποθέτοντας ότι προέρχονται από κανονικούς πληθυσμούς με ίσες διακυμάνσεις.

Λύση Παραδείγματος 10.5

Έστω X και Y οι τ.μ. που παριστάνουν την οξύτητα των δύο τύπων χημικών διαλυμάτων, αντίστοιχα. Από την εκφώνηση έχουμε ότι $X \sim N(\mu_1, \sigma_1^2)$ και $Y \sim N(\mu_2, \sigma_2^2)$ με $\sigma_1^2 = \sigma_2^2 = \sigma^2$. Καθώς οι πληθυσμοί είναι κανονικοί και οι πληθυσμιακές διασπορές είναι άγνωστες, αλλά ίσες εξ υποθέσεως, θα κατασκευάσουμε ένα δ.ε. για τη μέση διαφορά της οξύτητας των δύο τύπων διαλυμάτων χρησιμοποιώντας τη σχέση (10.15) με $\bar{x} = 6.392$, $\bar{y} = 6.5$, $s_1^2 = 0.00787$, $s_2^2 = 0.0054$, $n = 5$, $m = 7$, $\alpha = 0.05$ και $t_{n+m-2,\alpha/2} = t_{10,0.025} = 2.228$, ενώ

$$s_p^2 = \frac{(n-1)s_1^2 + (m-1)s_2^2}{n+m-2} = \frac{4 \cdot 0.00787 + 6 \cdot 0.0054}{10} = 0.006388.$$

Επομένως, ένα 95% δ.ε. για τη διαφορά $\mu_1 - \mu_2$ είναι το:

$$\left(6.392 - 6.5 - 2.228 \cdot \sqrt{0.006388}\sqrt{1/5 + 1/7},\ 6.392 - 6.5 + 2.228 \cdot \sqrt{0.06388}\sqrt{1/5 + 1/7}\right),$$

ή, μετά από λίγη άλγεβρα, το $(-0.2122688,\ -0.003731232)$.
Εναλλακτικά, χρησιμοποιώντας την R και εκτελώντας τις ακόλουθες εντολές:

```
A <- c(6.33, 6.28, 6.5, 6.40, 6.45)
B <- c(6.51, 6.55, 6.43, 6.51, 6.62, 6.40, 6.48)

t.test(A,B, paired = FALSE, var.equal=TRUE,conf.level = 0.95)$conf.int
```

λαμβάνουμε τα παρακάτω αποτελέσματα

```
[1] -0.212275266 -0.003724734
attr(,"conf.level")
[1] 0.95
```

τα οποία είναι ισοδύναμα με τα αποτελέσματα που υπολογίσαμε νωρίτερα.

Άσκηση Αυτοαξιολόγησης 10.7

Ένα τεστ Αγγλικών διεξάγεται στους μαθητές των πόλεων Α και Β. Από προηγούμενες έρευνες γνωρίζουμε ότι η βαθμολογία των μαθητών των δύο πόλεων περιγράφεται ικανοποιητικά από κανονική κατανομή με την ίδια διακύμανση. Εξετάζονται στα Αγγλικά 10 άτομα από καθεμία από τις δύο πόλεις και προκύπτουν τα ακόλουθα αποτελέσματα:

Πόλη Α:	12.9	10.2	7.4	7.0	10.5	11.9	7.1	9.9	14.4	11.3
Πόλη Β:	10.2	6.9	10.9	11.0	10.1	5.3	7.5	10.3	9.2	8.8

Να κατασκευαστεί ένα 95% δ.ε. για τη διαφορά της μέσης βαθμολογίας των μαθητών των δύο πόλεων.

Περίπτωση ΙΙΙ. Κανονικοί πληθυσμοί με άγνωστες και άνισες πληθυσμιακές διασπορές.

Αν θα πρέπει να απορριφθεί η υπόθεση της ισότητας των πληθυσμιακών διασπορών, τότε, σε αυτήν την περίπτωση, δεν μπορεί να βρεθεί ποσότητα οδηγός ή, αλλιώς, αντιστρεπτή ποσότητα, και χρησιμοποιείται το ακόλουθο $100(1-\alpha)\%$ (προσεγγιστικό) δ.ε. για τη διαφορά των μέσων τιμών $\mu_1 - \mu_2$ δύο κανονικών πληθυσμών με άγνωστες και άνισες πληθυσμιακές διασπορές ($\sigma_1^2 \neq \sigma_2^2$):

$$\left(\bar{X} - \bar{Y} - t_{v,\alpha/2}\sqrt{\frac{S_1^2}{n} + \frac{S_2^2}{m}}, \bar{X} - \bar{Y} + t_{v,\alpha/2}\sqrt{\frac{S_1^2}{n} + \frac{S_2^2}{m}}\right), \tag{10.16}$$

όπου

$$v = \frac{\left(S_1^2/n + S_2^2/m\right)^2}{\frac{(S_1^2/n)^2}{n-1} + \frac{(S_2^2/m)^2}{m-1}}, \tag{10.17}$$

με S_1^2 και S_2^2 να είναι οι δειγματικές διακυμάνσεις που προκύπτουν από τα ανεξάρτητα τ.δ. $X_1, \ldots, X_n$ και $Y_1, \ldots, Y_m$, αντίστοιχα. Το παραπάνω δ.ε. είναι γνωστό ως **Welch's t διάστημα εμπιστοσύνης**.

Λαμβάνοντας υπόψη τον Ορισμό 10.3, μια τιμή του παραπάνω τυχαίου διανύσματος που δίνεται από τη σχέση

$$\left(\bar{x} - \bar{y} - t_{v,\alpha/2}\sqrt{\frac{s_1^2}{n} + \frac{s_2^2}{m}}, \bar{x} - \bar{y} + t_{v,\alpha/2}\sqrt{\frac{s_1^2}{n} + \frac{s_2^2}{m}}\right), \tag{10.18}$$

είναι ένα $100(1-\alpha)\%$ (προσεγγιστικό) δ.ε. για τη διαφορά των μέσων τιμών $\mu_1 - \mu_2$ δύο κανονικών πληθυσμών με άγνωστες και άνισες πληθυσμιακές διασπορές $\sigma_1^2 \neq \sigma_2^2$.

Στην R μπορούμε εύκολα να προσδιορίσουμε το π.χ. 95% δ.ε. της σχέσης (10.16) με την εντολή

```
1 t.test(A,B, paired = FALSE, var.equal=FALSE,conf.level = 0.95)$conf.int
```

Επισημαίνεται ότι σε περίπτωση που δεν μας δίνονται τα σύνολα δεδομένων, αλλά οι τιμές των δειγματικών μέσων τιμών και τυπικών αποκλίσεων, τότε για να προσδιορίσουμε το π.χ. 95% δ.ε. της σχέσης (10.16), χρησιμοποιούμε τις ακόλουθες εντολές της R:

```
library(BSDA)
tsum.test(mean.x,s.x,n.x,mean.y,s.y,n.y,alternative = "two.sided",mu = 0,var.
    equal = FALSE,conf.level = 0.95)$conf.int
```

Παράδειγμα 10.6: (Walpole *et al.*, 2017)

Μια μελέτη πραγματοποιήθηκε από το Τμήμα Ζωολογίας του πανεπιστημίου της Virginia για να εκτιμηθεί η διαφορά στις ποσότητες του χημικού ορθοφωσφόρου (orthophosphorus) που μετρήθηκε σε δύο διαφορετικούς σταθμούς στον ποταμό James με μονάδα μέτρησης τα χιλιοστόγραμμα ανά λίτρο. Συλλέχθηκαν δεκαπέντε παρατηρήσεις από τον σταθμό A με μέση περιεκτικότητα σε ορθοφώσφορο 3.84 χιλιοστόγραμμα ανά λίτρο και τυπική απόκλιση 3.07 χιλιοστόγραμμα ανά λίτρο. Επιπρόσθετα, συλλέχθηκαν 12 παρατηρήσεις από τον σταθμό B με μέση περιεκτικότητα 1.49 χιλιοστόγραμμα ανά λίτρο και τυπική απόκλιση 0.8 χιλιοστόγραμμα ανά λίτρο. Υπολογίστε το 95% διάστημα εμπιστοσύνης για τη διαφορά του πραγματικού μέσου περιεχομένου ορθοφωσφόρου σε αυτούς τους δύο σταθμούς, υποθέτοντας ότι οι παρατηρήσεις προήλθαν από κανονικούς πληθυσμούς με διαφορετικές διασπορές.

Λύση Παραδείγματος 10.6

Έστω X και Y οι τ.μ. που παριστάνουν την ποσότητα ορθοφωσφόρου στην περιοχή του σταθμού A και B, αντίστοιχα. Από την εκφώνηση έχουμε ότι $X \sim N(\mu_1, \sigma_1^2)$ και $Y \sim N(\mu_2, \sigma_2^2)$ με $\sigma_1^2 \neq \sigma_2^2$. Καθώς οι πληθυσμοί υποθέτουμε ότι είναι κανονικοί και οι πληθυσμιακές διασπορές είναι άγνωστες, αλλά άνισες, θα κατασκευάσουμε δ.ε. για τη μέση διαφορά της συγκέντρωσης ορθοφωσφόρου χρησιμοποιώντας το δ.ε. που δόθηκε στη σχέση (10.18) με $\bar{x} = 3.84$, $\bar{y} = 1.49$, $s_1^2 = 3.07^2 = 9.4249$, $s_2^2 = 0.8^2 = 0.64$, $n = 15, m = 12, \alpha = 0.05$, ενώ χρησιμοποιώντας τη σχέση (10.17) οι βαθμοί ελευθερίας προσδιορίζονται ως εξής:

$$v = \frac{\left(s_1^2/n + s_2^2/m\right)^2}{\frac{(s_1^2/n)^2}{n-1} + \frac{(s_2^2/m)^2}{m-1}} = \frac{(9.4249/15 + 0.64/12)^2}{\frac{(9.4249/15)^2}{14} + \frac{(0.64/12)^2}{11}} = \frac{0.4646604}{0.0281996 + 0.0002585859} = 16.32783.$$

Για να χρησιμοποιήσουμε τον Πίνακα Α΄.4 της κατανομής t, κρατάμε μόνο το ακέραιο μέρος από την παραπάνω σχέση και συνεχίζουμε την επίλυση με $v = 16$ βαθμούς ελευθερίας, έχοντας ότι $t_{v,\alpha/2} = t_{16,0.025} = 2.120$. Επομένως, ένα 95% δ.ε. για τη διαφορά $\mu_1 - \mu_2$ είναι το:

$$\left(3.84 - 1.49 - 2.12\sqrt{\frac{9.4249}{15} + \frac{0.64}{12}}, 3.84 - 1.49 + 2.12\sqrt{\frac{9.4249}{15} + \frac{0.64}{12}}\right)$$

το οποίο, μετά από λίγη άλγεβρα, καταλήγει να είναι το $(0.5996707, 4.100329)$.
Εναλλακτικά, μπορούμε να χρησιμοποιήσουμε την ακόλουθη εντολή στην R:

```
library(BSDA)
tsum.test(mean.x=3.84,s.x =3.07,n.x = 15,mean.y =1.49,s.y = 0.8,n.y = 12,
    alternative = "two.sided",mu = 0,var.equal = FALSE,conf.level = 0.95)$
    conf.int
```

και έχουμε το αποτέλεσμα

```
[1] 0.6026006 4.0973994
attr(,"conf.level")
[1] 0.95
```

Σημειώνουμε ότι στην R δεν γίνεται στρογγυλοποίηση στους βαθμούς ελευθερίας, για αυτό και υπάρχει αυτή η διαφοροποίηση στα αποτελέσματα.

Άσκηση Αυτοαξιολόγησης 10.8

Έστω X και Y οι τ.μ. που παριστάνουν τη βαθμολογία στο τεστ Αγγλικών στις πόλεις Α και Β, αντίστοιχα. Από προηγούμενες έρευνες γνωρίζουμε ότι η βαθμολογία στις δύο πόλεις περιγράφεται ικανοποιητικά από κανονική κατανομή με άνισες διακυμάνσεις. Εξετάζονται στα Αγγλικά 10 άτομα από καθεμία από τις δύο πόλεις και προκύπτουν τα ακόλουθα αποτελέσματα:

Πόλη Α:	$\bar{x} = 10.26$	$s_1 = 2.51$
Πόλη Β:	$bary = 9.02$	$s_2 = 1.9$

Να κατασκευαστεί ένα 95% δ.ε. διάστημα εμπιστοσύνης για τη διαφορά της μέσης βαθμολογίας στο τεστ Αγγλικών στις δύο πόλεις.

Περίπτωση IV. Μη κανονικοί πληθυσμοί και μεγάλα δείγματα.

Έστω $X_1, \dots, X_n$ ένα τυχαίο δείγμα από τον πληθυσμό με μέση τιμή μ_1 και διασπορά σ_1^2, ενώ $Y_1, \dots, Y_m$ ένα τυχαίο δείγμα από τον πληθυσμό με μέση τιμή μ_2 και διασπορά σ_2^2. Επιπλέον, υποθέτουμε ότι τα δύο δείγματα είναι ανεξάρτητα και οι άγνωστες διασπορές είναι πεπερασμένες. Στην περίπτωση που ένας από τους δύο πληθυσμούς δεν είναι κανονικός, αλλά το διαθέσιμο δείγμα από αυτόν είναι μεγάλου μεγέθους (συνήθως μεγαλύτερο από 30), τότε ένα ασυμπτωτικό $100(1-\alpha)\%$ δ.ε. για τη διαφορά των μέσων τιμών $\mu_1 - \mu_2$ είναι το:

$$\left(\bar{X} - \bar{Y} - z_{a/2}\sqrt{\frac{S_1^2}{n} + \frac{S_2^2}{m}},\ \bar{X} - \bar{Y} + z_{a/2}\sqrt{\frac{S_1^2}{n} + \frac{S_2^2}{m}} \right), \tag{10.19}$$

με S_1^2 και S_2^2 να είναι οι δειγματικές διακυμάνσεις που προκύπτουν από τα ανεξάρτητα τ.δ. $X_1, \dots, X_n$ και $Y_1, \dots, Y_m$, αντίστοιχα.

Λαμβάνοντας υπόψη τον Ορισμό 10.3, μια τιμή του παραπάνω τυχαίου διανύσματος που δίνεται από τη σχέση:

$$\left(\bar{x} - \bar{y} - z_{a/2}\sqrt{\frac{s_1^2}{n} + \frac{s_2^2}{m}},\ \bar{x} - \bar{y} + z_{a/2}\sqrt{\frac{s_1^2}{n} + \frac{s_2^2}{m}} \right), \tag{10.20}$$

είναι ένα ασυμπτωτικό $100(1-\alpha)\%$ δ.ε. για τη διαφορά των μέσων τιμών $\mu_1 - \mu_2$.

Επισημαίνεται ότι το παραπάνω δ.ε. προέκυψε λαμβάνοντας υπόψη ότι:

$$\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n} + \frac{S_2^2}{m}}} \xrightarrow{d} N(0,1)$$

και ακολουθώντας τη γενική μέθοδο κατασκευής διαστημάτων εμπιστοσύνης.

10.3.3 Διάστημα εμπιστοσύνης για τη διαφορά δύο μέσων τιμών με εξαρτημένα δείγματα

Πολύ συχνά καλούμαστε να εκτιμήσουμε την επίδραση μιας θεραπείας/παρέμβασης πάνω σε ένα σύνολο οντοτήτων. Μία πάγια τεχνική για την εκτίμηση αυτήν είναι να πάρουμε μετρήσεις/παρατηρήσεις για το χαρακτηριστικό που μας ενδιαφέρει πριν και μετά την παρέμβαση. Σε αυτήν την περίπτωση, έχουμε εξαρτημένα δείγματα, που πολλές φορές καλούνται και δείγματα κατά ζεύγη ή ζευγαρωτές παρατηρήσεις. Για παράδειγμα, έστω ότι μας ενδιαφέρει να μελετήσουμε την επίδραση μιας δίαιτας στην απώλεια βάρους.

Επιλέγονται n το πλήθος γυναίκες και ζυγίζονται πριν την έναρξη της δίαιτας και έξι μήνες μετά και καταγράφεται σε κάθε περίπτωση το βάρος τους σε κιλά. Εναλλακτικά, θα μπορούσαμε να μελετήσουμε την αποτελεσματικότητα ενός αντι-υπερτασικού φαρμάκου μετρώντας την αρτηριακή πίεση n το πλήθος ατόμων πριν την έναρξη της θεραπείας και έξι μήνες μετά. Σε γενικό πλαίσιο, έστω $X_1,\ldots,X_n$ οι τυχαίες μεταβλητές που παριστάνουν τις μετρήσεις των n πειραματικών μονάδων στο υπό μελέτη χαρακτηριστικό πριν την παρέμβαση και $Y_1,\ldots,Y_n$ οι αντίστοιχες μετρήσεις στο υπό μελέτη χαρακτηριστικό μετά την παρέμβαση στις ίδιες πειραματικές μονάδες. Σε όσα ακολουθούν υποθέτουμε ότι τα ζεύγη $(X_i,\ Y_i),\ i = 1,\ldots,n$ είναι μεταξύ τους ανεξάρτητα. Το ενδιαφέρον σε αυτήν την ενότητα επικεντρώνεται στην κατασκευή δ.ε. για τη διαφορά των πληθυσμιακών μέσων $\mu_\delta = \mu_1 - \mu_2$, όπου μ_1 και μ_2 η μέση τιμή πριν και μετά την παρέμβαση, αντίστοιχα.

Προκειμένου να ληφθεί υπόψη η εξάρτηση, η κατασκευή του διαστήματος εμπιστοσύνης για τη διαφορά $\mu_\delta = \mu_1 - \mu_2$ στηρίζεται στις διαφορές $D_1 = X_1 - Y_1,\ldots,D_n = X_n - Y_n$. Υπό την υπόθεση ότι οι διαφορές αυτές προέρχονται από έναν κανονικό πληθυσμό με μέση τιμή μ_δ, τότε η κατασκευή του δ.ε. για τη διαφορά των μέσων τιμών $\mu_1 - \mu_2$ στη βάση δύο εξαρτημένων δειγμάτων ανάγεται στην κατασκευή δ.ε. για τη μέση τιμή ενός κανονικού πληθυσμού με άγνωστη διασπορά. Επομένως, από τη σχέση (10.7) προκύπτει ότι το $100(1-\alpha)\%$ δ.ε. ελαχίστου μήκους και ίσων ουρών για τη διαφορά των μέσων τιμών $\mu_1 - \mu_2$ προσδιορίζεται από τη σχέση:

$$\left(\bar{D} - t_{n-1,\alpha/2}\frac{S_D}{\sqrt{n}},\ \bar{D} + t_{n-1,\alpha/2}\frac{S_D}{\sqrt{n}}\right) \tag{10.21}$$

όπου $\bar{D}$ και S_D η δειγματική μέση τιμή και η δειγματική τυπική απόκλιση των διαφορών, αντίστοιχα.

Λαμβάνοντας υπόψη τον Ορισμό 10.3, μια τιμή του παραπάνω τυχαίου διανύσματος που δίνεται από τη σχέση:

$$\left(\bar{d} - t_{n-1,\alpha/2}\frac{s_D}{\sqrt{n}},\ \bar{d} + t_{n-1,\alpha/2}\frac{s_D}{\sqrt{n}}\right), \tag{10.22}$$

είναι, επίσης, ένα $100(1-\alpha)\%$ δ.ε. για τη διαφορά των μέσων τιμών $\mu_1 - \mu_2$.

Σε περίπτωση όπου η υπόθεση της κανονικότητας δεν ικανοποιείται, αλλά το πλήθος των ζευγών είναι μεγάλο ($n \geq 30$), τότε ένα $100(1-\alpha)\%$ ασυμπτωτικό δ.ε. για τη διαφορά των μέσων τιμών $\mu_1 - \mu_2$ προσδιορίζεται από τη σχέση:

$$\left(\bar{D} - z_{\alpha/2}\frac{S_D}{\sqrt{n}},\ \bar{D} + z_{\alpha/2}\frac{S_D}{\sqrt{n}}\right). \tag{10.23}$$

Λαμβάνοντας υπόψη τον Ορισμό 10.3, μια τιμή του παραπάνω τυχαίου διανύσματος που δίνεται από τη σχέση:

$$\left(\bar{d} - z_{\alpha/2}\frac{s_D}{\sqrt{n}},\ \bar{d} + z_{\alpha/2}\frac{s_D}{\sqrt{n}}\right), \tag{10.24}$$

είναι ένα $100(1-\alpha)\%$ ασυμπτωτικό δ.ε. για τη διαφορά των μέσων τιμών $\mu_1 - \mu_2$.

Παράδειγμα 10.7

Ένας ερευνητής χρησιμοποιεί τις μεθόδους Α και Β για τη μέτρηση του βάρους (σε γραμμάρια) ενός οργανισμού. Επιλέγει τυχαία 9 τέτοιους οργανισμούς και καταγράφει το βάρος τους σε γραμμάρια με τις δύο μεθόδους. Τα αποτελέσματα του πειράματος δίνονται στον επόμενο πίνακα:

Μέθοδος Α	326.5	326.6	326.6	326.8	326.3	326.6	326.7	326.7	326.3
Μέθοδος Β	326.5	326.6	326.5	326.7	326.3	326.5	326.7	326.6	326.2

Βρείτε, κάνοντας κατάλληλες υποθέσεις, ένα 95% δ.ε. για τη μέση διαφορά των μετρήσεων με τις δύο μεθόδους.

Λύση Παραδείγματος 10.7

Έστω X και Y οι τυχαίες μεταβλητές που παριστάνουν τη μέτρηση του βάρους με τη μέθοδο Α και Β, αντίστοιχα με μέση τιμή μ_1 και μ_2, αντίστοιχα. Έχουμε διαθέσιμα τα δείγματα $X_1,\ldots,X_n$ και $Y_1,\ldots,Y_n$ με $n = 9$, τα οποία είναι εξαρτημένα, καθώς πρόκειται για μετρήσεις στους ίδιους οργανισμούς. Προκειμένου να ληφθεί υπόψη η εξάρτηση, η κατασκευή του διαστήματος εμπιστοσύνης για τη διαφορά $\mu_\delta = \mu_1 - \mu_2$ στηρίζεται στις δειγματικές διαφορές $d_1 = x_1 - y_1, \ldots, d_n = x_n - y_n$. Οι διαφορές αυτές είναι

$d_i = x_i - y_i$	0	0	0.1	0.1	0	0.1	0	0.1	0.1

με δειγματική μέση τιμή και διακύμανση ίση με $\bar{d} = 0.05555556$ και $s_D^2 = 0.002777778$, αντίστοιχα. Υποθέτοντας ότι αυτές οι διαφορές προέρχονται από κανονικό πληθυσμό και χρησιμοποιώντας τη σχέση (10.22), ένα 95% δ.ε. για τη μέση διαφορά των μετρήσεων με τις δύο μεθόδους, λαμβάνοντας υπόψη ότι $t_{n-1,\alpha/2} = t_{8,0.025} = 2.306$, είναι το:

$$\left(0.05555556 - 2.306\frac{\sqrt{0.002777778}}{\sqrt{9}},\ 0.05555556 + 2.306\frac{\sqrt{0.002777778}}{\sqrt{9}}\right),$$

δηλαδή, μετά από λίγη άλγεβρα, το $(0.01504327, 0.09606785)$.
Η υλοποίηση με τη βοήθεια της R γίνεται μέσω των εντολών

```
1 data1<- c(326.5,326.6,326.6,326.8,326.3,326.6,326.7,326.7,326.3)
2 data2 <- c(326.5,326.6,326.5,326.7,326.3,326.5,326.7,326.6,326.2)
3
4 t.test(data1, data2, paired = TRUE)$conf.int
```

Η παραπάνω εντολή δίνει ως αποτέλεσμα το δ.ε. (0.01504319, 0.09606792).

Άσκηση Αυτοαξιολόγησης 10.9

Ένας ερευνητής χρησιμοποιεί τις μεθόδους Α και Β για τη μέτρηση της συγκέντρωσης μιας ποσότητας στο αίμα ενήλικων ανδρών. Επιλέγει τυχαία 15 ενήλικες άνδρες και καταγράφει τη συγκέντρωση της ποσότητας με τις δύο μεθόδους. Τα αποτελέσματα του πειράματος είναι τα ακόλουθα:

Μέθοδος Α	1.94	1.44	1.56	1.58	2.06	1.66	1.75	1.77
Μέθοδος Β	1.27	1.63	1.47	1.39	1.93	1.26	1.71	1.67
Μέθοδος Α	1.78	1.92	1.25	1.93	2.04	1.62	2.08	
Μέθοδος Β	1.28	1.85	1.02	1.34	2.02	1.59	1.97	

Βρείτε, κάνοντας κατάλληλες υποθέσεις, ένα 95% δ.ε. για τη μέση διαφορά των μετρήσεων με τις δύο μεθόδους.

10.3.4 Διάστημα εμπιστοσύνης για τη διακύμανση κανονικού πληθυσμού

Σε πολλές πραγματικές εφαρμογές πολλές φορές δεν μας αρκεί να μελετήσουμε τη μέση τιμή ενός ή περισσότερων πληθυσμών, αλλά χρειάζεται να εξάγουμε συμπεράσματα και για τη μεταβλητότητα αυτού ή αυτών, αντίστοιχα. Επομένως, επιθυμούμε να κατασκευάσουμε διαστήματα εμπιστοσύνης για τη διασπορά ή/και την τυπική απόκλιση του πληθυσμού (ή το πηλίκο των διασπορών, όπως θα δούμε σε επόμενη ενότητα). Στην ενότητα αυτή το ενδιαφέρον επικεντρώνεται στα δ.ε. για τη διασπορά και την τυπική απόκλιση ενός κανονικού πληθυσμού.

Έστω $X_1, \ldots, X_n$ ένα τυχαίο δείγμα από κανονικά κατανεμημένο πληθυσμό με μέση τιμή μ (άγνωστη) και τυπική απόκλιση σ. Τότε χρησιμοποιώντας τη στατιστική συνάρτηση $\frac{(n-1)S^2}{\sigma^2}$ ως αντιστρεπτή ποσότητα και γνωρίζοντας ότι ισχύει ότι:

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$

έχουμε ακολουθώντας τη γενική τεχνική κατασκευής των διαστημάτων εμπιστοσύνης ότι το $100(1-\alpha)\%$ δ.ε. ίσων ουρών για τη διακύμανση σ^2 δίνεται από τη σχέση:

$$\left(\frac{(n-1)S^2}{\chi^2_{n-1,\alpha/2}}, \frac{(n-1)S^2}{\chi^2_{n-1,1-\alpha/2}}\right). \tag{10.25}$$

Λαμβάνοντας υπόψη τον Ορισμό 10.3, μια τιμή του παραπάνω τυχαίου διανύσματος που δίνεται από τη σχέση:

$$\left(\frac{(n-1)s^2}{\chi^2_{n-1,\alpha/2}}, \frac{(n-1)s^2}{\chi^2_{n-1,1-\alpha/2}}\right) \tag{10.26}$$

είναι, επίσης, ένα $100(1-\alpha)\%$ δ.ε. για τη διακύμανση σ^2.

Επιπρόσθετα, ενθυμούμενοι και την Παρατήρηση 10.6, προκύπτει ότι το $100(1-\alpha)\%$ δ.ε. ίσων ουρών για την πληθυσμιακή τυπική απόκλιση δίνεται από τη σχέση:

$$\left(\sqrt{\frac{(n-1)S^2}{\chi^2_{n-1,\alpha/2}}}, \sqrt{\frac{(n-1)S^2}{\chi^2_{n-1,1-\alpha/2}}}\right),$$

ενώ λαμβάνοντας υπόψη τον Ορισμό 10.3, μια τιμή του παραπάνω τυχαίου διανύσματος που δίνεται από τη σχέση:

$$\left(\sqrt{\frac{(n-1)s^2}{\chi^2_{n-1,\alpha/2}}}, \sqrt{\frac{(n-1)s^2}{\chi^2_{n-1,1-\alpha/2}}}\right),$$

είναι, επίσης, ένα $100(1-\alpha)\%$ δ.ε για την πληθυσμιακή τυπική απόκλιση.

Παρατήρηση 10.15

Τα παραπάνω δ.ε. είναι πολύ ευαίσθητα στην υπόθεση της κανονικότητας των πληθυσμών. Μικρές αποκλίσεις από την υπόθεση της κανονικότητας έχουν ως συνέπεια μεγάλες διαφορές μεταξύ του πραγματικού επιπέδου εμπιστοσύνης του διαστήματος και της ονομαστικής του τιμής. Σε τέτοιες περιπτώσεις προτείνεται η χρήση μεθόδων bootstrap, οι οποίες ξεφεύγουν από τους σκοπούς του παρόντος συγγράμματος. Ο/Η ενδιαφερόμενος/η αναγνώστης/στρια παραπέμπεται στο σύγγραμμα των Efron and Tibshirani (1993) και τις εκεί αναφορές, καθώς και στην ευρέως χρησιμοποιούμενη βιβλιοθήκη `boot` της R.

Παράδειγμα 10.8

Από προηγούμενες μελέτες είναι γνωστό ότι το βάρος των ξηρών καρπών που συσκευάζει μια εταιρεία ακολουθεί κανονική κατανομή με μέση τιμή μ και διακύμανση σ^2. Ο υπεύθυνος ποιοτικού ελέγχου της εταιρείας θέλοντας να βρει ένα 95% δ.ε. για την πραγματική διακύμανση επιλέγει 10 τυχαία προϊόντα, καταγράφει το βάρος τους και υπολογίζει την τυπική απόκλισή τους να είναι ίση με $\sqrt{4.2}$ γραμμάρια. Προσδιορίστε το δ.ε. που θέλει ο υπεύθυνος ποιοτικού ελέγχου.

Λύση Παραδείγματος 10.8

Έστω X η τ.μ. που παριστάνει το βάρος της συσκευασίας των ξηρών καρπών. Από την εκφώνηση έχουμε ότι $X \sim N(\mu, \sigma^2)$ και επίσης ότι είναι διαθέσιμο τυχαίο δείγμα μεγέθους 10 με δειγματική διακύμανση $s^2 = 4.2$. Από τη σχέση (10.26) για $\alpha = 0.05$ προκύπτει ότι ένα 95% δ.ε. για την άγνωστη πληθυσμιακή διακύμανση είναι το:

$$\left(\frac{(n-1)s^2}{\chi^2_{n-1,\alpha/2}}, \frac{(n-1)s^2}{\chi^2_{n-1,1-\alpha/2}}\right) = \left(\frac{9\cdot 4.2}{\chi^2_{4,0.025}}, \frac{9\cdot 4.2}{\chi^2_{4,0.975}}\right) = \left(\frac{9\cdot 4.2}{19.02}, \frac{9\cdot 4.2}{2.7}\right) = (1.99, 14.0).$$

Στην R θα μπορούσαμε να κατασκευάσουμε ένα δ.ε. για τη διασπορά και την τυπική απόκλιση κανονικού πληθυσμού χρησιμοποιώντας τη βιβλιοθήκη `Ecfun` και τις εντολές `confint.var` και `confint.sd`, που συντάσσονται όπως φαίνεται παρακάτω:

```
library(Ecfun)
confint.var(4.2,10-1,level=0.95)
confint.sd(sqrt(4.2),10-1,level=0.95)
```

Τότε προκύπτουν τα ακόλουθα αποτελέσματα:

```
        lower    upper
[1,] 1.987093 13.99798
attr(,"level")
[1] 0.95

      lower    upper
[1,] 1.409643 3.741388
attr(,"level")
[1] 0.95
```

Παρατήρηση 10.16

Σε όσα προηγήθηκαν υποθέσαμε ότι η πληθυσμιακή μέση τιμή του κανονικού πληθυσμού είναι άγνωστη. Στη σπάνια περίπτωση που μας είναι γνωστή η πληθυσμιακή μέση τιμή τότε χρησιμοποιείται ως αντιστρεπτή ποσότητα η στατιστική συνάρτηση $\frac{\sum_{i=1}^n (X_i-\mu)^2}{\sigma^2}$ για την οποία ισχύει ότι:

$$\frac{\sum_{i=1}^n \left(X_i - \mu\right)^2}{\sigma^2} \sim \chi^2_n.$$

Έπειτα, ακολουθώντας τη γενική τεχνική κατασκευής των δ.ε., προκύπτει ότι το $100(1-\alpha)\%$ δ.ε. ίσων ουρών για τη διακύμανση σ^2 δίνεται από τη σχέση:

$$\left(\frac{\sum_{i=1}^n \left(X_i - \mu\right)^2}{\chi^2_{n,\alpha/2}}, \frac{\sum_{i=1}^n \left(X_i - \mu\right)^2}{\chi^2_{n,1-\alpha/2}}\right), \tag{10.27}$$

ενώ, λαμβάνοντας υπόψη τον Ορισμό 10.3, μια τιμή του παραπάνω τυχαίου διανύσματος που δίνεται από τη σχέση:

$$\left(\frac{\sum_{i=1}^n \left(x_i - \mu\right)^2}{\chi^2_{n,\alpha/2}}, \frac{\sum_{i=1}^n \left(x_i - \mu\right)^2}{\chi^2_{n,1-\alpha/2}}\right), \tag{10.28}$$

είναι ένα $100(1-\alpha)\%$ δ.ε. για τη διακύμανση σ^2, όταν η πληθυσμιακή μέση τιμή του κανονικού πληθυσμού είναι γνωστή.

Άσκηση Αυτοαξιολόγησης 10.10

Μετρήθηκε ο χρόνος ζωής (σε ημέρες) 18 λαμπτήρων και υπολογίστηκε ότι έχουν τυπική απόκλιση 1.8 ημέρες. Να βρεθεί ένα 95% δ.ε. για την πληθυσμιακή διακύμανση, υπό την υπόθεση ότι ο χρόνος ζωής του λαμπτήρα ακολουθεί κανονική κατανομή.

10.3.5 Διάστημα εμπιστοσύνης για το πηλίκο δύο διακυμάνσεων κανονικών πληθυσμών με ανεξάρτητα δείγματα

Έστω $X_1,\ldots,X_n$ τυχαίο δείγμα από έναν κανονικό πληθυσμό με μέση τιμή μ_1 και διασπορά σ_1^2, ενώ $Y_1,\ldots,Y_m$ είναι ένα άλλο ανεξάρτητο τυχαίο δείγμα από έναν κανονικό πληθυσμό με μέση τιμή μ_2 και διασπορά σ_2^2. Στην ενότητα αυτή, το ενδιαφέρον επικεντρώνεται στην κατασκευή ενός δ.ε. για το πηλίκο των διασπορών, δηλαδή για το πηλίκο $\frac{\sigma_1^2}{\sigma_2^2}$.

Υπό την υπόθεση ότι $X_i \sim N(\mu_1, \sigma_1^2)$ για $i = 1, \ldots, n$ και $Y_j \sim N(\mu_2, \sigma_2^2)$ για $j = 1, \ldots, m$, και υπό την ανεξαρτησία των δύο τυχαίων δειγμάτων έχουμε ότι:

$$\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F_{n-1,m-1}$$

όπου S_1^2 και S_2^2 οι δειγματικές διακυμάνσεις που βασίζονται στα τ.δ. $X_1,\ldots,X_n$ και $Y_1,\ldots,Y_m$, αντίστοιχα. Ακολουθώντας τη γενική τεχνική κατασκευής δ.ε. προκύπτει το ακόλουθο $100(1-\alpha)\%$ δ.ε. ελαχίστων ουρών για το πηλίκο των διασπορών $\frac{\sigma_1^2}{\sigma_2^2}$:

$$\left(\frac{1}{F_{n-1,m-1,a/2}} \cdot \frac{S_1^2}{S_2^2}, \frac{1}{F_{n-1,m-1,1-a/2}} \cdot \frac{S_1^2}{S_2^2} \right). \tag{10.29}$$

Λαμβάνοντας υπόψη τον Ορισμό 10.3, μια τιμή του παραπάνω τυχαίου διανύσματος που δίνεται από τη σχέση:

$$\left(\frac{1}{F_{n-1,m-1,a/2}} \cdot \frac{s_1^2}{s_2^2}, \frac{1}{F_{n-1,m-1,1-a/2}} \cdot \frac{s_1^2}{s_2^2} \right) \tag{10.30}$$

είναι ένα $100(1-\alpha)\%$ δ.ε. για το πηλίκο των διασπορών $\frac{\sigma_1^2}{\sigma_2^2}$.

Στην R μπορούμε να βρούμε ένα 95% δ.ε. για το πηλίκο δύο διακυμάνσεων κανονικών πληθυσμών μέσω της παρακάτω εντολής:

```
1 var.test(A, B, ratio = 1, alternative = c("two.sided"), conf.level = 0.95)
```

Παρατήρηση 10.17

Το διάστημα εμπιστοσύνης για τον λόγο των διασπορών είναι και αυτό πολύ ευαίσθητο στην υπόθεση της κανονικότητας των πληθυσμών και ισχύουν οι ίδιες παρατηρήσεις και υποδείξεις που έγιναν στην Παρατήρηση 10.15.

Παράδειγμα 10.9

Σε μια μελέτη συλλέχθηκαν δεκαέξι παρατηρήσεις ορθοφωσφόρου από τον σταθμό Α με μέση περιεκτικότητα σε ορθοφώσφορο 3.84 χιλιοστόγραμμα ανά λίτρο και τυπική απόκλιση 3.07 χιλιοστόγραμμα ανά λίτρο. Επιπρόσθετα, συλλέχθηκαν 13 παρατηρήσεις από τον σταθμό Β με μέση περιεκτικότητα 1.49 χιλιοστόγραμμα ανά λίτρο και τυπική απόκλιση 0.8 χιλιοστόγραμμα ανά λίτρο. Υπολογίστε ένα 90% διάστημα εμπιστοσύνης για το πηλίκο των διακυμάνσεων των συγκεντρώσεων ορθοφωσφόρου στους δύο σταθμούς, υποθέτοντας ότι οι παρατηρήσεις προήλθαν από κανονικούς πληθυσμούς.

Λύση Παραδείγματος 10.9

Έστω X και Y οι τ.μ. που παριστάνουν την ποσότητα ορθοφωσφόρου στην περιοχή του σταθμού Α και Β, αντίστοιχα. Από την εκφώνηση έχουμε ότι $X \sim N(\mu_1, \sigma_1^2)$ και $Y \sim N(\mu_2, \sigma_2^2)$. Καθώς οι πληθυσμοί υποθέτουμε ότι είναι κανονικοί θα κατασκευάσουμε δ.ε. για το πηλίκο των πληθυσμιακών διασπορών, χρησιμοποιώντας τη σχέση (10.30) με $n = 16$, $m = 13$, $\alpha = 0.1$, $s_1^2 = 3.07^2 = 9.4249$ και $s_2^2 = 0.8^2 = 0.64$. Είναι τότε:

$$\left(\frac{1}{F_{n-1,m-1,\alpha/2}} \cdot \frac{s_1^2}{s_2^2}, \frac{1}{F_{n-1,m-1,1-\alpha/2}} \cdot \frac{s_1^2}{s_2^2} \right) = \left(\frac{1}{F_{15,12,0.05}} \cdot \frac{9.4249}{0.64}, \frac{1}{F_{15,12,0.95}} \cdot \frac{9.4249}{0.64} \right).$$

Από τον Πίνακα Α΄.8 του Παραρτήματος έχουμε $F_{15,12,0.05} = 2.62$, ενώ $F_{15,12,0.95} = \frac{1}{F_{12,15,0.05}} = \frac{1}{2.48} = 0.4032258$. Μετά από λίγη άλγεβρα, προκύπτει ότι ένα 90% δ.ε. για το πηλίκο των διακυμάνσεων είναι το $(5.620766, 36.52149)$.

Στην R θα μπορούσαμε να κατασκευάσουμε ένα δ.ε. για το πηλίκο των διασπορών με τις ακόλουθες εντολές:

```
1  sample.na <- 16
2  sample.nb <- 13
3  sample.vara <- 3.07^2
4  sample.varb <- 0.8^2
5  df1<-sample.na-1
6  df2<-sample.nb-1
7  alpha = 0.1
8  score1 = qf(alpha/2, df1, df2, lower.tail = F)
9  score2=qf(1-alpha/2, df1, df2, lower.tail = F)
10 lower.bound <- (1/score1)* (sample.vara/sample.varb)
11 upper.bound<-(1/score2)* (sample.vara/sample.varb)
12 cat("CI for the ratio of pop varA/varB = (",lower.bound,",", upper.bound, "
       )\n")
```

Τότε προκύπτει το ακόλουθο αποτέλεσμα:

```
CI for the ratio of pop varA/varB = ( 5.627529 , 36.45246 )
```

Άσκηση Αυτοαξιολόγησης 10.11

Χρησιμοποιώντας τα δεδομένα της Άσκησης Αυτοαξιολόγησης 10.7 να βρείτε ένα 90% δ.ε. για το πηλίκο των πληθυσμιακών διακυμάνσεων.

10.3.6 Διάστημα εμπιστοσύνης για το ποσοστό ενός πληθυσμού

Το ενδιαφέρον σε αυτήν την ενότητα επικεντρώνεται στην εκτίμηση μέσω διαστήματος του άγνωστου ποσοστού επιτυχίας p ενός διωνυμικού τυχαίου πειράματος. Στο πλαίσιο αυτό, έστω X_i, $i = 1, \dots, n$ η τ.μ. που παριστάνει την έκβαση μιας δοκιμής Bernoulli στην i-οστή επανάληψη με $X_i = 1$, αν έχουμε επιτυχία στο $i-$οστό πείραμα, και $X_i = 0$, διαφορετικά. Τότε το $X = \sum_{i=1}^{n} X_i$ εκφράζει το πλήθος των επιτυχιών στις n δοκιμές και γνωρίζουμε ότι $X \sim B(n, p)$. Τότε από την Πρόταση 9.12 έχουμε ότι για μεγάλο μέγεθος δείγματος n

$$\frac{\hat{P} - p}{\sqrt{\frac{\hat{P}(1-\hat{P})}{n}}} \xrightarrow{d} N(0,1), \quad \text{με } \hat{P} = X/n.$$

Χρησιμοποιώντας την παραπάνω ως αντιστρεπτή ποσότητα, προκύπτει ότι ένα ασυμπτωτικό $100(1-\alpha)\%$ δ.ε. για την παράμετρο p δίνεται από τη σχέση:

$$\left(\hat{P} - z_{a/2}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}},\ \hat{P} + z_{a/2}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}} \right), \tag{10.31}$$

ενώ, λαμβάνοντας υπόψη τον Ορισμό 10.3, μια τιμή του παραπάνω τυχαίου διανύσματος που δίνεται από τη σχέση:

$$\left(\hat{p} - z_{a/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\ \hat{p} + z_{a/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right) \tag{10.32}$$

είναι επίσης ένα ασυμπτωτικό $100(1-\alpha)\%$ δ.ε. για την παράμετρο p.

Παρατήρηση 10.18

Στη βιβλιογραφία έχουν προταθεί και άλλες μεθοδολογίες κατασκευής δ.ε. για την πιθανότητα p, πλην της προηγούμενης που στηρίζεται στην κανονική προσέγγιση και κάποιες φορές αναφέρεται ως δ.ε. με τη μέθοδο του Wald. Ειδικότερα, μεταξύ αυτών, είναι:

- η μέθοδος των Pearson-Klopper, η οποία βασίζεται στη διωνυμική κατανομή και όχι στην κανονική προσέγγιση,
- η μέθοδος Wilson, η οποία αποτελεί μια βελτίωση της κανονικής προσέγγισης και χρησιμοποιεί μία πιο εξελιγμένη μέθοδο για τον προσδιορισμό των άκρων του διαστήματος εμπιστοσύνης,
- η μέθοδος των Agresti–Coull, η οποία δίνει με διαφορετικό τρόπο ένα προσεγγιστικό διάστημα εμπιστοσύνης.

Οι παραπάνω μέθοδοι ξεφεύγουν από τους σκοπούς του παρόντος συγγράμματος και ο/η ενδιαφερόμενος/η αναγνώστης/στρια παραπέμπεται, μεταξύ άλλων, στην εργασία των Agresti and Coull (1998).

Παράδειγμα 10.10

Επιλέγονται τυχαία 100 άτομα και σε αυτά παρατηρείται ότι 45 άτομα είναι κατά της ψήφισης ενός νομοσχεδίου. Να κατασκευάσετε ένα 95% (ασυμπτωτικό) διάστημα εμπιστοσύνης για το πραγματικό ποσοστό των ατόμων που είναι εναντίον του νομοσχεδίου. Ποιες υποθέσεις στην ουσία χρησιμοποιήσατε;

Λύση Παραδείγματος 10.10

Ένα 95% (ασυμπτωτικό) διάστημα εμπιστοσύνης για το πραγματικό ποσοστό p των ατόμων που είναι εναντίον του νομοσχεδίου θα υπολογιστεί από τη σχέση (10.32) για $\alpha = 0.05$, $n = 100$ και $\hat{p} = 45/100 = 0.45$. Επομένως, είναι το:

$$\left(0.45 - 1.96\sqrt{\frac{0.45 \cdot 0.55}{100}},\ 0.45 + 1.96\sqrt{\frac{0.45 \cdot 0.55}{100}}\right) = (0.3524912,\ 0.5475088).$$

Στην R μπορούμε να χρησιμοποιήσουμε την εντολή `binom.confint` της βιβλιοθήκης `binom`, όπως φαίνεται παρακάτω:

```
1 library(binom)
2 binom.confint(45, 100, alpha=0.05,methods=c("asymptotic"))
```

Παρατηρήστε ότι στο πρώτο όρισμα της εντολής `binom.confint` δηλώνουμε τον αριθμό των επιτυχιών, ενώ στο δεύτερο όρισμα τον αριθμό των επαναλήψεων. Τα αποτελέσματα που προκύπτουν είναι τα ακόλουθα.

```
    method  x   n mean    lower    upper
1 asymptotic 45 100 0.45 0.352493 0.547507
```

Τα παραπάνω ισχύουν υπό την προϋπόθεση ότι η πιθανότητα κάποιο άτομο να είναι εναντίον του νομοσχεδίου παραμένει αμετάβλητη από άτομο σε άτομο και η απάντηση ενός ατόμου είναι ανεξάρτητη από την απάντηση οποιουδήποτε άλλου ατόμου.

Παρατήρηση 10.19

Πολύ συχνά, όπως αναφέρθηκε και στην Παρατήρηση 10.10, θέλουμε να υπολογίσουμε το κατάλληλο μέγεθος δείγματος προκειμένου να εξαχθούν συμπεράσματα με προκαθορισμένη ακρίβεια. Πιο συγκεκριμένα, συχνά επιλέγουμε μέγεθος δείγματος, τέτοιο ώστε το πλάτος του (προσεγγιστικού) δ.ε. για το πραγματικό ποσοστό p να μην υπερβαίνει μία προκαθορισμένη σταθερά ϵ.
Αν από προηγούμενες έρευνες υπάρχει μια αρχική εκτίμηση $\hat{p}$ του πραγματικού ποσοστού, τότε το ζητούμενο μέγεθος δείγματος μπορεί να προσδιοριστεί από την ακόλουθη σχέση

$$\frac{2\, z_{a/2}\sqrt{\hat{p}(1-\hat{p})}}{n} \leq \epsilon \Rightarrow n \geq 4\left(\frac{z_{\alpha/2}\ \sqrt{\hat{p}(1-\hat{p})}}{\epsilon}\right)^2.$$

Συχνά όμως μια τέτοια αρχική εκτίμηση του p δεν είναι διαθέσιμη, αλλά, ακόμα και αν είναι, τίποτα δεν εξασφαλίζει ότι η εκτίμηση στο δείγμα που θα λάβουμε θα είναι πολύ κοντά με την εκτίμηση αυτή έτσι ώστε να ισχύει η παραπάνω ανισότητα. Για τον λόγο αυτόν συνιστάται να χρησιμοποιείται η σχέση

$$n \geq 4\left(\frac{z_{\alpha/2}\ \sqrt{0.5(1-0.5)}}{\epsilon}\right)^2,$$

η οποία προκύπτει από την προηγούμενη, αντικαθιστώντας το $\hat{p}$ με το 0.5. Η τιμή 0.5 είναι αυτή που μεγιστοποιεί τη συνάρτηση $\hat{p}(1-\hat{p})$, γεγονός που εξασφαλίζει ότι οποιαδήποτε και να είναι η τιμή του $\hat{p}$ στο δείγμα, το πλάτος του (προσεγγιστικού) δ.ε. δεν θα υπερβαίνει το ϵ. Ωστόσο, επισημαίνεται ότι η παραπάνω σχέση οδηγεί γενικά σε μεγαλύτερο από το πραγματικά αναγκαίο μέγεθος δείγματος.

Άσκηση Αυτοαξιολόγησης 10.12

Σε δείγμα 2500 ατόμων που παρακολουθούν αθλητικά στην τηλεόραση 920 ήταν γυναίκες. Να βρεθεί ένα (ασυμπτωτικό) 95% διάστημα εμπιστοσύνης για το πραγματικό ποσοστό γυναικών που παρακολουθούν αθλητικές εκπομπές.

10.3.7 Διάστημα εμπιστοσύνης για τη διαφορά δύο πληθυσμιακών ποσοστών με ανεξάρτητα δείγματα

Πολλές φορές μας ενδιαφέρει να κατασκευάσουμε διάστημα εμπιστοσύνης για τη διαφορά των ποσοστών εμφάνισης ενός χαρακτηριστικού σε δύο πληθυσμούς, δηλαδή για τη διαφορά $p_1 - p_2$ με p_i το ποσοστό εμφάνισης στον i-οστό πληθυσμό. Για τον λόγο αυτόν παίρνουμε ένα τυχαίο δείγμα $X_1, \ldots, X_n$ από τον πρώτο πληθυσμό, όπου $X_i = 1$, αν έχουμε επιτυχία στην i-οστή επανάληψη δοκιμής Bernoulli με πιθανότητα επιτυχίας $P(X_i = 1) = p_1$, και 0 διαφορετικά για $i = 1, \ldots, n$ και ένα τυχαίο δείγμα $Y_1, \ldots, Y_m$ από τον δεύτερο πληθυσμό, όπου $Y_j = 1$, αν έχουμε επιτυχία στην j-οστή επανάληψη της δοκιμής Bernoulli με πιθανότητα επιτυχίας $P(Y_j = 1) = p_2$, και 0 διαφορετικά για $j = 1, \ldots, m$. Υποθέτουμε ότι τα δύο δείγματα είναι ανεξάρτητα μεταξύ τους. Σε αυτήν την περίπτωση, $\hat{P}_1 = \frac{X}{n}$ και $\hat{P}_2 = \frac{Y}{m}$, όπου $X = \sum_{i=1}^{n} X_i$ και $Y = \sum_{j=1}^{m} Y_j$. Με παρόμοιο σκεπτικό με αυτό που αναφέρθηκε στην προηγούμενη ενότητα, προκύπτει ότι ένα ασυμπτωτικό $100(1-\alpha)\%$ δ.ε. για το $p_1 - p_2$ δίνεται από τη σχέση:

$$\left(\hat{P}_1 - \hat{P}_2 - z_{\alpha/2} \sqrt{\frac{\hat{P}_1(1-\hat{P}_1)}{n} + \frac{\hat{P}_2(1-\hat{P}_2)}{m}},\ \hat{P}_1 - \hat{P}_2 + z_{\alpha/2} \sqrt{\frac{\hat{P}_1(1-\hat{P}_1)}{n} + \frac{\hat{P}_2(1-\hat{P}_2)}{m}} \right)$$

ενώ, λαμβάνοντας υπόψη τον Ορισμό 10.3, μια τιμή του παραπάνω τυχαίου διανύσματος που δίνεται από τη σχέση:

$$\left(\hat{p}_1 - \hat{p}_2 - z_{\alpha/2} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n} + \frac{\hat{p}_2(1-\hat{p}_2)}{m}},\ \hat{p}_1 - \hat{p}_2 + z_{\alpha/2} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n} + \frac{\hat{p}_2(1-\hat{p}_2)}{m}} \right) \tag{10.33}$$

είναι, επίσης, ένα ασυμπτωτικό $100(1-\alpha)\%$ δ.ε. για το $p_1 - p_2$.

Παράδειγμα 10.11

Θέλοντας ένας πολιτικός να μελετήσει τη διαφορά των ποσοστών αποδοχής ενός νομοσχεδίου μεταξύ ανδρών και γυναικών, συγκέντρωσε ένα τυχαίο δείγμα 62 ανδρών και ένα τυχαίο δείγμα 85 γυναικών και κατέγραψε τις απόψεις τους. Αν 26 άνδρες και 24 γυναίκες ήταν υπέρ του νομοσχεδίου σε αυτά τα δύο ανεξάρτητα τυχαία δείγματα, κατασκευάστε ένα (ασυμπτωτικό) 95% διάστημα εμπιστοσύνης για τη διαφορά των ποσοστών αποδοχής του νομοσχεδίου στους δύο πληθυσμούς.

Λύση Παραδείγματος 10.11

Θέλουμε να κατασκευάσουμε ένα 95% (ασυμπτωτικό) διάστημα εμπιστοσύνης για τη διαφορά των ποσοστών $p_1 - p_2$ με p_1 να είναι το ποσοστό των ανδρών που είναι υπέρ του νομοσχεδίου, ενώ p_2 να είναι το ποσοστό των γυναικών που είναι υπέρ του νομοσχεδίου. Το ζητούμενο δ.ε. θα υπολογιστεί από τη σχέση (10.33) για $\alpha = 0.05$, $n = 62$, $\hat{p}_1 = 26/62 = 0.4193548$, $m = 85$, $\hat{p}_2 = 24/85 = 0.2823529$. Επομένως, το κάτω άκρο του δ.ε. είναι το

$$0.4193548 - 0.2823529 - 1.96 \sqrt{\frac{0.4193548 \cdot (1 - 0.4193548)}{62} + \frac{0.2823529 \cdot (1 - 0.2823529)}{85}}$$

ενώ το πάνω άκρο του είναι το:

$$0.4193548 - 0.2823529 + 1.96\sqrt{\frac{0.4193548 \cdot (1 - 0.4193548)}{62} + \frac{0.2823529 \cdot (1 - 0.2823529)}{85}}.$$

Μετά από λίγη άλγεβρα, προκύπτει ότι ένα 95% (ασυμπτωτικό) διάστημα εμπιστοσύνης για τη διαφορά των αναλογιών $p_1 - p_2$ είναι το $(-0.01870699,\ 0.2927108)$.
Εναλλακτικά, θα μπορούσαμε να χρησιμοποιήσουμε τη βιβλιοθήκη `DescTools` της R και την εντολή `BinomDiffCI`, όπως φαίνεται παρακάτω:

```
1 library(DescTools)
2 BinomDiffCI(26, 62, 24, 85, conf.level = 0.95, method ="wald")
```

Τότε προκύπτουν τα ακόλουθα αποτελέσματα:

```
           est      lwr.ci    upr.ci
[1,] 0.1370019 -0.01870414 0.2927079
```

Άσκηση Αυτοαξιολόγησης 10.13

Σε ένα τυχαίο δείγμα 100 γυναικών ηλικίας 45 και άνω βρέθηκαν 56 να έχουν πρεσβυωπία, ενώ σε ένα άλλο ανεξάρτητο τυχαίο δείγμα 250 ανδρών ηλικίας 45 και άνω βρέθηκαν 125 να έχουν πρεσβυωπία. Κατασκευάστε ένα 95% δ.ε. για τη διαφορά των αναλογιών στους δύο πληθυσμούς.

10.4 Ασκήσεις

Άσκηση 10.1 *Επιλέξτε τη σωστή απάντηση και δικαιολογήστε σύντομα την απάντησή σας.*

1. *Το ελάχιστο μέγεθος δείγματος που πρέπει να ληφθεί από έναν κανονικό πληθυσμό με διασπορά 3, έτσι ώστε το πλάτος του 95% διαστήματος εμπιστοσύνης ελαχίστου μήκους για τη μέση τιμή να είναι μικρότερο από 1/2 είναι:*
   (α΄) 184,
   (β΄) 14,
   (γ΄) 185,
   (δ΄) 13.
2. *Ένα 95% διάστημα εμπιστοσύνης για τη μέση τιμή ενός πληθυσμού είναι το (44.2, 54.2). Ας υποθέσουμε ότι υπολογίζετε ένα διάστημα εμπιστοσύνης 99% χρησιμοποιώντας τις ίδιες πληροφορίες. Ποια από τις ακόλουθες προτάσεις είναι σωστή;*
   (α΄) Τα διαστήματα θα έχουν το ίδιο πλάτος.
   (β΄) Το 99% διάστημα θα έχει μικρότερο πλάτος.
   (γ΄) Το 99% διάστημα θα έχει μεγαλύτερο πλάτος.
   (δ΄) Η απάντηση δεν μπορεί να προσδιοριστεί από τις πληροφορίες που δόθηκαν.
3. *Μία εταιρεία θέλοντας να ελέγξει τη διάρκεια ζωής των ηλεκτρικών λαμπτήρων που παράγονται στις εγκαταστάσεις της, επιλέγει τυχαίο δείγμα από $n = 30$ λαμπτήρες. Το δείγμα έδωσε ένα 90% διάστημα εμπιστοσύνης για τη μέση διάρκεια ζωής ενός λαμπτήρα (σε ώρες) ίσο με (982.40, 997.60). Αν υποθέσουμε ότι η κατανομή της διάρκειας ενός λαμπτήρα είναι κανονική, ποια από τις παρακάτω προτάσεις είναι σωστή;*
   (α΄) Με πιθανότητα 90% η (πληθυσμιακή) μέση διάρκεια ζωής των λαμπτήρων θα βρίσκεται στο διάστημα $(982.40, 997.60)$.
   (β΄) Η δειγματική μέση διάρκεια ζωής ενός λαμπτήρα θα είναι ίση με $\bar{x} = 980$ ώρες.
   (γ΄) Η μέση διάρκεια ζωής ενός λαμπτήρα θα είναι σίγουρα μεγαλύτερη από 982.40 ώρες.
   (δ΄) Τίποτε από τα παραπάνω δεν ισχύει.
4. *Από ένα συγκεκριμένο τυχαίο δείγμα κατασκευάσαμε τα 85%, 90%, 95% και 99% διαστήματα εμπιστοσύνης για την αναλογία επιτυχιών σε έναν πληθυσμό. Τα διαστήματα αυτά εμφανίζονται στις παρακάτω επιλογές. Ποιο από αυτά είναι το 90% διάστημα εμπιστοσύνης;*
   (α΄) (0.325, 0.505),
   (β΄) (0.347, 0.483),
   (γ΄) (0.358, 0.472),
   (δ΄) (0.365, 0.465).
5. *Προκειμένου να ελεγχθεί η ποιότητα χαρτιού που παράγει μία βιομηχανία, επιλέχθηκαν τυχαία από την παραγωγή μίας εβδομάδας 16 ρολά χαρτιού. Από κάθε ρολό αποκόπηκε τυχαία 1 τετράγωνο φύλλο χαρτιού εμβαδού ενός τετραγωνικού μέτρου και καταγράφηκε το βάρος τους σε γραμμάρια. Θεωρώντας ότι η κατανομή του βάρους είναι κανονική, υπολογίστηκε το 95% διάστημα εμπιστοσύνης για το μέσο βάρος των φύλλων χαρτιού και βρέθηκε (79.467, 80.533). Αυτό σημαίνει ότι:*
   (α΄) ο δειγματικός μέσος των 16 φύλλων χαρτιού ισούται με 80.05 γραμμάρια,
   (β΄) είμαστε 95% σίγουροι ότι το μέσο βάρος των φύλλων χαρτιού του δείγματος βρίσκεται στο διάστημα (79.467, 80.0533),
   (γ΄) υπάρχει πιθανότητα 95% να είναι η άγνωστη μέση τιμή μ του βάρους ενός φύλλου χαρτιού εμβαδού ενός τετραγωνικού μέτρου μεταξύ 79.467 και 80.533 γραμμαρίων,
   (δ΄) η δειγματική τυπική απόκλιση είναι 1 γραμμάριο.

Άσκηση 10.2 *Επιλέχθηκαν τυχαία 10 φοιτητές/φοιτήτριες και ζύγιζαν ο καθένας/καθεμία 53, 69, 62, 78, 81, 55, 66, 62, 74, 60. Υποθέτοντας ότι το βάρος των φοιτητών/φοιτητριών ακολουθεί κανονική κατανομή* $\mathrm{N}(\mu,100)$*, να κατασκευάσετε ένα 95% δ.ε. για το μέσο βάρος του πληθυσμού των φοιτητών/φοιτητριών.*

Άσκηση 10.3 *Ο σύλλογος καταναλωτών για να εξακριβώσει αν τα κουτιά των 100 γραμμαρίων καφέ περιέχουν πραγματικά 100 γραμμάρια καφέ, ζύγισε 9 κουτιά τυχαία επιλεγμένα από διάφορα καταστήματα και βρήκε μέσο βάρος 96 γραμμάρια με τυπική απόκλιση 1.8 γραμμάρια. Να κατασκευαστεί ένα 95% δ.ε. για το πραγματικό μέσο βάρος των κουτιών καφέ.*

Άσκηση 10.4 *Δύο τύποι αυτοκινήτων Α και Β δοκιμάζονται για την αξιοπιστία των φρένων τους. Πήραμε 64 αυτοκίνητα από κάθε τύπο και μετρήσαμε την απόσταση που διατρέχουν μέχρι να σταματήσουν όταν με 40 Km/h πατήσουμε φρένο. Από προηγούμενες μελέτες γνωρίζουμε ότι οι τ.μ.* X *και* Y *που παριστάνουν την απόσταση που διατρέχουν τα αυτοκίνητα τύπου Α και Β ακολουθούν κανονική κατανομή με διαφορετικές διασπορές. Αν* $\bar{x} = 118$, $\bar{y} = 112$, $s_1^2 = 102$, $s_2^2 = 87$, *κατασκευάστε ένα 95% δ.ε. για τη διαφορά των πληθυσμιακών μέσων τιμών. Τέλος, να βρεθεί ένα 90% δ.ε. για το πηλίκο των πληθυσμιακών διακυμάνσεων.*

Άσκηση 10.5 *Το μέσο βάρος 50 φοιτητών που γυμνάζονται είναι 68.2 κιλά με τυπική απόκλιση 2.5 κιλά, ενώ το μέσο βάρος 50 φοιτητών που δεν γυμνάζονται βρέθηκε 67.5 κιλά με τυπική απόκλιση 2.8 κιλά. Να προσδιορίσετε ένα 99% δ.ε. για τη μέση πληθυσμιακή διαφορά του βάρους των φοιτητών όταν είναι γνωστό ότι οι πληθυσμοί δεν είναι κανονικοί.*

Άσκηση 10.6 *Δέκα πρωτοετείς φοιτητές έχουν τους ακόλουθους σφυγμούς ανά λεπτό: 59, 72, 58, 65, 77, 83, 72, 77, 62, 62. Αν οι μετρήσεις αυτές προέρχονται από κανονικό πληθυσμό, βρείτε ένα 95% δ.ε. για τους μέσους σφυγμούς του πληθυσμού των πρωτοετών φοιτητών.*

Άσκηση 10.7 *Σε τυχαίο δείγμα* $n = 36$ *ατόμων βρήκαμε 2 άτομα να πάσχουν από την ασθένεια Α. Βρείτε ένα 95% δ.ε. για το αληθινό ποσοστό των ατόμων που πάσχουν από την ασθένεια Α.*

Άσκηση 10.8 *Σε μία διεργασία ανακύκλωσης πλαστικών μπουκαλιών το σημαντικότερο υλικό που παράγεται είναι το πλαστικό PET. Υπάρχει όμως σοβαρό πρόβλημα λόγω της εμφάνισης αλουμινίου στο παραγόμενο υλικό, κάτι που δημιουργεί δυσκολίες στη μετέπειτα χρήση του υλικού. Σε ένα διάστημα μιας εβδομάδας επιλέχθηκαν 49 δοκίμια του πλαστικού και βρέθηκε μέση περιεκτικότητα του πλαστικού σε αλουμίνιο ίση με 172.8 με τυπική απόκλιση 34.6. Να υπολογίσετε ένα* 95% *δ.ε. για τη μέση περιεκτικότητα του πλαστικού σε αλουμίνιο.*

Άσκηση 10.9 *Έστω ότι θέλουμε να εκτιμήσουμε τη μέση διάμετρο κάποιου τύπου ροδέλας που παράγεται από μία μηχανή. Μας είναι γνωστό ότι η μέτρηση της διαμέτρου της ροδέλας ακολουθεί κανονική κατανομή με τυπική απόκλιση* $\sigma = 0.06\ cm$*. Πόσες ροδέλες πρέπει να ελέγξουμε ώστε να είμαστε* 90% *σίγουροι ότι το μέγιστο σφάλμα της εκτίμησης της μέσης τιμής θα είναι το πολύ* $0.02\ cm$*;*

Άσκηση 10.10 *Μελετάται η απόδοση ενός καινούριου αλγόριθμου για τη λύση διαφορικών εξισώσεων. Από ένα δείγμα 52 προγραμμάτων βρήκαμε μέσο χρόνο λύσης 0.81 δευτερόλεπτα και τυπική απόκλιση 1.5 δευτερόλεπτα. Να βρείτε ένα* 99% *δ.ε. για τον μέσο χρόνο λύσης που χρειάζεται το νέο πρόγραμμα για να λύσει τις διαφορικές εξισώσεις και ένα* 95% *δ.ε. για τη διασπορά του χρόνου λύσης.*

Άσκηση 10.11 *Ένας ερευνητής θέλει να συγκρίνει τη μέση αύξηση βάρους πειραματόζωων με δύο διαφορετικές δίαιτες, τις δίαιτες Α και Β. Ο ερευνητής χορηγεί τη δίαιτα Α σε 10 από τα διαθέσιμα πειραματόζωα και σε 6 τη δίαιτα Β και καταγράφει την αύξηση του βάρους τους, όπως φαίνεται στον πίνακα που ακολουθεί.*

A	*113*	*135*	*91*	*104*	*135*	*107*	*152*	*97*	*145*	*129*
B	*126*	*73*	*102*	*110*	*79*	*104*				

Να κατασκευάσετε ένα 95% *διάστημα εμπιστοσύνης για τη διαφορά των μέσων αυξήσεων βάρους με τις δύο δίαιτες υπό την υπόθεση ότι οι δύο πληθυσμοί είναι κανονικοί και έχουν ίσες διασπορές.*

Άσκηση 10.12 *Συγκεντρώθηκε ένα τυχαίο δείγμα από 100 σταγόνες δακρύων. Το δείγμα έδωσε μια μέση περιεκτικότητα σε αλάτι 0.01 και τυπική απόκλιση 0.01. Να βρεθεί ένα* 95% *δ.ε. για τη μέση περιεκτικότητα σε αλάτι των δακρύων ανά σταγόνα.*

Άσκηση 10.13 *Ειδικοί εξέτασαν 144 παιδιά του δημοτικού και διαπίστωσαν ότι το* 30% *των παιδιών έχουν προβλήματα όρασης. Βρείτε ένα* 95% *δ.ε. για το πραγματικό ποσοστό των παιδιών που έχει πρόβλημα όρασης.*

Άσκηση 10.14 *Ένας ερευνητής θέλει να κάνει σύγκριση δύο μεθόδων Α και Β για τη μέτρηση της ανάπτυξης ενζύμων σε μία διαδικασία ζύμωσης. Από μία δεξαμενή, όπου γίνεται η ζύμωση, επιλέγονται 9 δείγματα υλικού. Κάθε δείγμα χωρίζεται σε δύο ίσα μέρη για ανάλυση με τη μέθοδο Α και Β. Τα αποτελέσματα του πειράματος δίνονται στον επόμενο πίνακα.*

Μέθοδος Α	*326.5*	*326.6*	*326.6*	*326.8*	*326.3*	*326.6*	*326.7*	*326.7*	*326.3*
Μέθοδος Β	*326.5*	*326.6*	*326.5*	*326.7*	*326.3*	*326.5*	*326.7*	*326.6*	*326.2*

Κάνοντας κατάλληλες υποθέσεις, βρείτε ένα 95% *δ.ε. για τη διαφορά των μέσων μετρήσεων με τις δύο μεθόδους.*

Άσκηση 10.15 *Η δυνατότητα καταχώρισης χαρακτήρων στον υπολογιστή ενός υπαλλήλου του Πανεπιστημίου ακολουθεί κανονική κατανομή με μέση τιμή 45 χαρακτήρες ανά λεπτό. Πρόσφατα, ο υπάλληλος παρακολούθησε ένα σεμινάριο χρήσης μικροϋπολογιστών με στόχο τη βελτίωσή του. Σε 3 δοκιμές που έγιναν, μετά την ολοκλήρωση του σεμιναρίου, ο υπάλληλος καταχώρισε 43, 42 και 47 χαρακτήρες το λεπτό, αντίστοιχα. Με βάση αυτά τα δεδομένα να κατασκευάσετε ένα* 90% *δ.ε. για τον μέσο αριθμό χαρακτήρων που πληκτρολογεί ο υπάλληλος ανά λεπτό και ένα* 95% *δ.ε. για την τυπική απόκλιση του αριθμού χαρακτήρων που πληκτρολογεί ο υπάλληλος ανά λεπτό.*

Άσκηση 10.16 *Είναι γνωστό ότι η πετρελαϊκή ρύπανση των θαλασσών προκαλεί, μεταξύ άλλων, την ανάπτυξη ενός συγκεκριμένου τύπου βακτηριδίων. Μια ομάδα ερευνητών, προκειμένου να μελετήσει αυτό το φαινόμενο σε μια θαλάσσια περιοχή που έχει πληγεί από πετρελαϊκή ρύπανση, συνέλεξε νερό από 9 διαφορετικά σημεία αυτής της περιοχής και έκανε σχετικές μετρήσεις. Συγκεκριμένα, μέτρησε τον αριθμό των βακτηριδίων ανά 100 milliliters νερού και ο μέσος αριθμός βακτηριδίων ανά 100 milliliters ήταν 59.2 με τυπική απόκλιση 10.4. Κατασκευάστε, κάνοντας κατάλληλες υποθέσεις, ένα* 95% *δ.ε. για τον μέσο αριθμό βακτηριδίων ανά 100 milliliters νερού στην υπό μελέτη θαλάσσια περιοχή.*

Άσκηση 10.17 *Σε μία μεταλλουργία έγινε ένα πείραμα σύγκρισης δύο μεθόδων Α και Β για την κατασκευή ενός κράματος μετάλλων. Η Α είναι η καθιερωμένη μέθοδος κατασκευής, ενώ η Β είναι μία νέα μέθοδος πρόσμειξης των μετάλλων στο κράμα. Αφού επιλέχθηκαν στην τύχη 10 δοκίμια από κράμα που κατασκευάστηκε με τη μέθοδο Α και 10 με τη μέθοδο Β, προσδιορίστηκε η αντοχή θραύσης κάθε δοκιμίου. Στον παρακάτω πίνακα δίνονται τα αποτελέσματα του πειράματος σε* kg/cm^2.

A	*828*	*819*	*858*	*839*	*841*	*856*	*829*	*838*	*845*	*863*
B	*862*	*848*	*835*	*865*	*829*	*872*	*859*	*827*	*868*	*847*

Βρείτε, κάνοντας κατάλληλες υποθέσεις, ένα 90% *δ.ε. για τη διαφορά των μέσων αντοχών θραύσης με τις δύο μεθόδους κατασκευής.*

Άσκηση 10.18 *Σε μία μελέτη ερευνάται η επίδραση της θερμοκρασίας στο όριο θραύσης κάποιου είδους δέρματος. Οκτώ κομμάτια του ίδιου μεγέθους δέρματος κόπηκαν σε δύο ίσα μέρη. Το ένα μέρος από κάθε κομμάτι δέρματος τεντώθηκε μέχρι να σπάσει και το άλλο θερμάνθηκε στους* $50^o\ C$ *και, στη συνέχεια, τεντώθηκε μέχρι να σπάσει. Στον παρακάτω πίνακα δίνονται τα βάρη σε κιλά που χρειάστηκαν για να τεντωθούν μέχρι να σπάσει τα μέρη του δέρματος.*

Μη Θερμανθέντα	*36*	*41*	*25*	*33*	*34*	*41*	*37*	*29*
Θερμανθέντα	*37*	*42*	*21*	*33*	*32*	*37*	*34*	*25*

Βρείτε ένα 95% *δ.ε. για τη διαφορά των μέσων ορίων θραύσης του δέρματος μετά από την επεξεργασία με τις δύο παραπάνω μεθόδους.*

10.5 Απαντήσεις στις ασκήσεις αυτοαξιολόγησης

Λύση Άσκησης Αυτοαξιολόγησης 10.1

Έστω X η τ.μ. που παριστάνει τη ροή (σε κυβικά μέτρα/*sec*) του ποταμού, τότε $X \sim N(\mu, 5^2)$. Είναι διαθέσιμο το τ.δ. $X_1,\ldots,X_n$ με $n = 25$ και $\bar{x} = 81.2$ και θέλουμε να κατασκευάσουμε ένα 95% δ.ε. για τη μέση ροή ύδατος. Η πληθυσμιακή διακύμανση είναι γνωστή και, επομένως, θα χρησιμοποιήσουμε τη σχέση (10.6) με $\alpha = 0.05$. Επομένως, ένα 95% δ.ε. για τη μέση ροή ύδατος είναι το:

$$\left(81.2 - 1.96 \cdot \frac{\sqrt{25}}{5},\ 81.2 + 1.96 \cdot \frac{\sqrt{25}}{5}\right) = (79.24,\ 83.16).$$

Εναλλακτικά, μπορούμε να χρησιμοποιήσουμε την ακόλουθη εντολή στην R:

```
1 ci.mu.z(conf=0.95,sigma=5,summarized=TRUE, xbar=81.2, n=25)
```

και έχουμε το αποτέλεσμα

```
95% z Confidence interval for population mean
Estimate     2.5%    97.5%
81.20000 79.24004 83.15996
```

Λύση Άσκησης Αυτοαξιολόγησης 10.2

Το πλάτος του $100 \cdot (1-\alpha)\%$ διαστήματος εμπιστοσύνης ελαχίστου μήκους για τη μέση τιμή μ ενός κανονικού πληθυσμού με γνωστή διασπορά ισούται με $d = \frac{2 \cdot z_{\alpha/2}\ \sigma}{\sqrt{n}}$. Επομένως, θέλουμε

$$\frac{2 \cdot z_{0.04/2} \cdot \sqrt{25}}{\sqrt{n}} \leq 1$$

ή, ισοδύναμα,

$$\sqrt{n} \geq 2 \cdot z_{0.04/2} \cdot \sqrt{25} = 10 \cdot 2.05.$$

Άρα πρέπει $n \geq 420.25$, που μας οδηγεί στο συμπέρασμα ότι το μικρότερο μέγεθος τυχαίου δείγματος που ικανοποιεί το ζητούμενο είναι $n = 421$.

Λύση Άσκησης Αυτοαξιολόγησης 10.3

Έστω X η τ.μ. που παριστάνει τη ροή (σε κυβικά μέτρα/*sec*) του ποταμού, τότε $X \sim N(\mu, \sigma^2)$. Είναι διαθέσιμο το τ.δ. $X_1,\ldots,X_n$ με $n = 25, \bar{x} = 81.2$ και $s = 5.2$. Θέλουμε να κατασκευάσουμε ένα 95% δ.ε. για τη μέση ροή ύδατος. Η πληθυσμιακή διακύμανση είναι άγνωστη και, επομένως, θα χρησιμοποιήσουμε τη σχέση (10.8) με $\alpha = 0.05$ και $t_{n-1,\alpha/2} = t_{24,0.025} = 2.064$. Επομένως, ένα 95% δ.ε. για τη μέση ροή ύδατος είναι το:

$$\left(81.2 - 2.064 \cdot \frac{5.2}{5},\ 81.2 + 2.04 \cdot \frac{5.2}{5}\right) = (79.05344,\ 83.34656).$$

Εναλλακτικά, μπορούμε να χρησιμοποιήσουμε την ακόλουθη εντολή στην R:

```
1 ci.mu.t(conf=0.95,summarized=TRUE, xbar=81.2, sd=5.2, n=25)
```

και έχουμε το αποτέλεσμα

```
95% t Confidence interval for population mean
Estimate     2.5%    97.5%
81.20000 79.05355 83.34645
```

Λύση Άσκησης Αυτοαξιολόγησης 10.4

Το πλάτος του $100 \cdot (1-\alpha)\%$ διαστήματος εμπιστοσύνης ελαχίστου μήκους για τη μέση τιμή μ ενός κανονικού πληθυσμού με άγνωστη διασπορά ισούται με $d = \frac{2t_{n-1,\alpha/2}\, S}{\sqrt{n}}$. Στη συγκεκριμένη άσκηση είναι $n = 5$ και $\alpha = 0.05$, ενώ από τις τιμές που δίνονται προκύπτει ότι $s^2 = 0.522$. Επομένως, το πλάτος του δ.ε. είναι (βλ. και τη σχέση (10.8))

$$\frac{2 \cdot t_{4,0.025} \cdot \sqrt{0.522}}{\sqrt{5}} = \frac{2 \cdot 2.776 \cdot \sqrt{0.522}}{\sqrt{5}} = 1.793906.$$

Εναλλακτικά, μπορούμε να χρησιμοποιήσουμε τις ακόλουθες εντολές στην R:

```
1 library(asbio)
2 A<-c(6.6,4.6,5.4,5.8,5.5)
3 ci.mu.t(A, conf=0.95)
```

και έχουμε το αποτέλεσμα

```
95% t Confidence interval for population mean
Estimate     2.5%    97.5%
5.580000 4.682903 6.477097
```

Καθώς έχουμε προσδιορίσει το δ.ε., άμεσα βρίσκουμε το πλάτος του.

Λύση Άσκησης Αυτοαξιολόγησης 10.5

Έστω X η τ.μ. που παριστάνει τον χρόνο ζωής του ινδικού χοιριδίου σε μέρες με μέση τιμή μ και πεπερασμένη διακύμανση σ^2. Ένα ασυμπτωτικό 90% δ.ε. για τη μέση τιμή μ δίνεται από τη σχέση (10.10) με $n = 64$, $\alpha = 0.1$, $z_{\alpha/2} = z_{0.05} = 1.645$, $\bar{x} = 345.2$ και $s = 222.2$, δηλαδή είναι το:

$$\left(345.2 - 1.645 \cdot \frac{222.2}{\sqrt{64}},\ 345.2 + 1.645 \cdot \frac{222.2}{\sqrt{64}}\right) = (299.5101,\ 390.8899).$$

Εναλλακτικά, μπορούμε να χρησιμοποιήσουμε την ακόλουθη εντολή στην R:

```
1 ci.mu.z(conf=0.90,sigma=222.2,summarized=TRUE,xbar=345.2, n=64)
```

και έχουμε το αποτέλεσμα

```
90% z Confidence interval for population mean
Estimate       5%      95%
345.2000 299.5142 390.8858
```

Παρατηρήστε ότι χρησιμοποιήθηκε η εντολή της R για την εύρεση δ.ε. για τη μέση τιμή κανονικού πληθυσμού με γνωστή διασπορά, θέτοντάς την ίση με τη δειγματική διακύμανση.

Λύση Άσκησης Αυτοαξιολόγησης 10.6

Έστω X και Y οι τ.μ. που παριστάνουν τη βαθμολογία στο τεστ Αγγλικών στις πόλεις Α και Β, αντίστοιχα. Από την εκφώνηση έχουμε ότι $X \sim N(\mu_1, \sigma_1^2 = 400)$ και $Y \sim N(\mu_2, \sigma_2^2 = 500)$. Καθώς οι πληθυσμιακές διασπορές είναι γνωστές και οι πληθυσμοί υποθέτουμε ότι είναι κανονικοί, θα κατασκευάσουμε ένα δ.ε. για τη μέση διαφορά των βαθμολογιών στα Αγγλικά στις δύο πόλεις χρησιμοποιώντας το δ.ε. που δόθηκε στη σχέση (10.12) με $\bar{x} = 1000$, $\bar{y} = 985$, $n = 18$, $m = 25$, $\alpha = 0.03$ και $z_{\alpha/2} = z_{0.015} = 2.17$. Επομένως, ένα 97% δ.ε. για τη διαφορά $\mu_1 - \mu_2$ είναι το:

$$\left(1000 - 985 - 2.17\sqrt{\frac{400}{18} + \frac{500}{25}},\ 1000 - 985 - 2.17\sqrt{\frac{400}{18} + \frac{500}{25}}\right)$$

ή, μετά από λίγη άλγεβρα, το $(0.8996375,\ 29.10036)$.

```
1  sample.na <- 18
2  sample.nb <- 25
3  population.vara <- 400
4  population.varb <- 500
5  var.meanerrora <- population.vara/sample.na
6  var.meanerrorb <- population.varb /sample.nb
7  alpha = 0.03
8  z.score = qnorm(p=alpha/2,lower.tail=F)
9  sample.meana <-1000
10 sample.meanb <- 985
11 lower.bound <- (sample.meana- sample.meanb)- z.score * sqrt(var.meanerrora+
       var.meanerrorb)
12 upper.bound <- (sample.meana- sample.meanb)+ z.score * sqrt(var.meanerrora+
       var.meanerrorb)
13
14 cat("CI for mean A-mean B = (",lower.bound,",", upper.bound, ")\n")
```

Το αποτέλεσμα των παραπάνω εντολών είναι το ακόλουθο

```
CI for mean A-mean B = ( 0.8990503 , 29.10095 )
```

Λύση Άσκησης Αυτοαξιολόγησης 10.7

Έστω X και Y οι τ.μ. που παριστάνουν τη βαθμολογία στις πόλεις Α και Β, αντίστοιχα. Από την εκφώνηση έχουμε ότι $X \sim N(\mu_1, \sigma_1^2)$ και $Y \sim N(\mu_2, \sigma_2^2)$ με $\sigma_1^2 = \sigma_2^2 = \sigma^2$. Καθώς οι πληθυσμοί υποθέτουμε ότι είναι κανονικοί και οι πληθυσμιακές διασπορές είναι άγνωστες αλλά ίσες, θα κατασκευάσουμε ένα δ.ε. για τη μέση διαφορά των βαθμολογιών στα Αγγλικά στις δύο πόλεις χρησιμοποιώντας το δ.ε. που δόθηκε στη σχέση (10.15) με $\bar{x} = 10.26$, $\bar{y} = 9.02$, $s_1^2 = 6.318222$, $s_2^2 = 3.597333$, $n = 10$, $m = 10$, $\alpha = 0.05$ και $t_{n+m-2,\alpha/2} = t_{18,0.025} = 2.101$, ενώ

$$s_p^2 = \frac{(n-1)s_1^2 + (m-1)s_2^2}{n+m-2} = \frac{9 \cdot 6.318222 + 9 \cdot 3.597333}{18} = 4.957778.$$

Επομένως, το κάτω άκρο του δ.ε. είναι:

$$10.26 - 9.02 - 2.101 \cdot \sqrt{4.957778}\sqrt{1/10 + 1/10} = -0.8521104,$$

ενώ το πάνω άκρο είναι:

$$10.26 - 9.02 - 2.101 \cdot \sqrt{4.957778}\sqrt{1/10 + 1/10} = 3.33211,$$

δηλαδή ένα 95% δ.ε. για τη διαφορά $\mu_1 - \mu_2$ είναι το $(-0.8521104,\ 3.33211)$.
Εναλλακτικά, χρησιμοποιώντας την R εκτελούμε τις εντολές:

```
data1<-c(12.9,10.2, 7.4, 7.0, 10.5, 11.9, 7.1, 9.9, 14.4, 11.3)
data2<-c(10.2, 6.9, 10.9, 11.0, 10.1, 5.3, 7.5, 10.3, 9.2, 8.8)
t.test(data1,data2, paired = FALSE, var.equal=TRUE,conf.level = 0.95)$conf.
    int
```

και έχουμε το αποτέλεσμα

```
[1] -0.8520327  3.3320327
attr(,"conf.level")
[1] 0.95
```

Λύση Άσκησης Αυτοαξιολόγησης 10.8

Έστω X και Y οι τ.μ. που παριστάνουν τη βαθμολογία στα Αγγλικά στην πόλη Α και Β, αντίστοιχα. Από την εκφώνηση έχουμε ότι $X \sim N(\mu_1,\sigma_1^2)$ και $Y \sim N(\mu_2,\sigma_2^2)$ με $\sigma_1^2 \neq \sigma_2^2$. Καθώς οι πληθυσμοί υποθέτουμε ότι είναι κανονικοί και οι πληθυσμιακές διασπορές είναι άγνωστες αλλά άνισες, θα κατασκευάσουμε ένα δ.ε. για τη μέση διαφορά της βαθμολογίας, χρησιμοποιώντας το δ.ε. που δόθηκε στη σχέση (10.18) με $\bar{x} = 10.26$, $\bar{y} = 9.02$, $s_1^2 = 2.51^2 = 6.3001$, $s_2^2 = 1.9^2 = 3.61$, $n = 10$, $m = 10$, $\alpha = 0.05$, ενώ χρησιμοποιώντας τη σχέση (10.17) οι βαθμοί ελευθερίας προσδιορίζονται ως εξής:

$$v = \frac{\left(s_1^2/n + s_2^2/m\right)^2}{\frac{(s_1^2/n)^2}{n-1} + \frac{(s_2^2/m)^2}{m-1}} = \frac{(6.3001/10 + 3.61/10)^2}{\frac{(6.3001/10)^2}{9} + \frac{(3.61/10)^2}{9}} = \frac{0.9821008}{0.0441014 + 0.01448011} = 16.76469.$$

Για να χρησιμοποιήσουμε τον Πίνακα Α΄.4 της κατανομής t κρατάμε μόνο το ακέραιο μέρος από την παραπάνω σχέση και συνεχίζουμε την επίλυση με $v = 16$ βαθμούς ελευθερίας με $t_{v,\alpha/2} = t_{16,0.025} = 2.120$. Επομένως, ένα 95% δ.ε. για τη διαφορά $\mu_1 - \mu_2$ είναι το:

$$\left(10.26 - 9.02 - 2.12\sqrt{\frac{6.3001}{10} + \frac{3.61}{10}}, 10.26 - 9.02 + 2.12\sqrt{\frac{6.3001}{10} + \frac{3.61}{10}}\right)$$

δηλαδή το $(-0.8704491,\ 3.350449)$.
Εναλλακτικά, χρησιμοποιώντας την R, εκτελούμε τις εντολές:

```
library(BSDA)
tsum.test(mean.x=10.26,s.x =2.51,n.x = 10,mean.y =9.02,s.y = 1.9,n.y = 10,
    alternative = "two.sided",mu = 0,var.equal = FALSE,conf.level = 0.95)$
    conf.int
```

και έχουμε το αποτέλεσμα

```
[1] -0.8625585  3.3425585
attr(,"conf.level")
[1] 0.95
```

Σημειώνουμε ότι στην R δεν γίνεται στρογγυλοποίηση στους βαθμούς ελευθερίας, για αυτό και υπάρχει αυτή η διαφοροποίηση στα αποτελέσματα.

Λύση Άσκησης Αυτοαξιολόγησης 10.9

Έστω X και Y οι τυχαίες μεταβλητές που παριστάνουν τη μέτρηση της ποσότητας με τη μέθοδο Α και Β, αντίστοιχα με μέση τιμή μ_1 και μ_2, αντίστοιχα. Έχουμε διαθέσιμα τα δείγματα $X_1,\ldots,X_n$ και $Y_1,\ldots,Y_n$ με $n = 15$, τα οποία είναι εξαρτημένα, καθώς πρόκειται για μετρήσεις στους ίδιους ενήλικες άνδρες. Προκειμένου να ληφθεί υπόψη η εξάρτηση, η κατασκευή ενός διαστήματος εμπιστοσύνης για τη διαφορά $\mu_\delta = \mu_1 - \mu_2$ στηρίζεται στις δειγματικές διαφορές $d_1 = x_1 - y_1, \ldots, d_n = x_n - y_n$. Οι διαφορές αυτές είναι 0.67, -0.19, 0.09, 0.19, 0.13, 0.40, 0.04, 0.10, 0.50, 0.07, 0.23, 0.59, 0.02, 0.03, 0.11 με δειγματική μέση τιμή και διακύμανση ίση με $\bar{d} = 0.1986667$ και $s_D^2 = 0.05678381$. Υποθέτοντας ότι αυτές οι διαφορές προέρχονται από κανονικό πληθυσμό και χρησιμοποιώντας τη σχέση (10.24), ένα 95% δ.ε. για τη μέση διαφορά των μετρήσεων με τις δύο μεθόδους, λαμβάνοντας υπόψη ότι $t_{n-1,\alpha/2} = t_{14,0.025} = 2.145$, είναι το:

$$\left(0.1986667 - 2.145\frac{\sqrt{0.05678381}}{\sqrt{15}},\ 0.1986667 + 2.145\frac{\sqrt{0.05678381}}{\sqrt{15}}\right),$$

δηλαδή το $(0.06669101,\ 0.3306424)$.
Εναλλακτικά, χρησιμοποιώντας την R εκτελούμε τις εντολές:

```
1 data1<-c(1.94, 1.44, 1.56, 1.58, 2.06, 1.66, 1.75, 1.77, 1.78, 1.92, 1.25,
    1.93, 2.04, 1.62, 2.08)
2 data2<-c(1.27, 1.63, 1.47, 1.39, 1.93, 1.26, 1.71, 1.67, 1.28, 1.85, 1.02,
    1.34, 2.02, 1.59, 1.97)
3 t.test(data1,data2,paired=TRUE)$conf.int
```

και έχουμε το αποτέλεσμα

```
[1] 0.0667041 0.3306292
attr(,"conf.level")
[1] 0.95
```

Λύση Άσκησης Αυτοαξιολόγησης 10.10

Έστω X η τυχαία μεταβλητή που παριστάνει τον χρόνο ζωής σε ημέρες ενός λαμπτήρα. Τότε $X \sim N(\mu,\sigma^2)$ με μ και σ^2 άγνωστες παραμέτρους. Σε ένα δείγμα μεγέθους $n = 18$ υπολογίστηκε ότι $s = 1.8$. Θα προσδιορίσουμε ένα 95% δ.ε. για την πληθυσμιακή διασπορά χρησιμοποιώντας τη σχέση (10.26) με $\alpha = 0.05$. Επομένως, ένα 95% δ.ε. για την πληθυσμιακή διακύμανση του χρόνου ζωής είναι το:

$$\left(\frac{(n-1)s^2}{\chi^2_{n-1,\alpha/2}},\ \frac{(n-1)s^2}{\chi^2_{n-1,1-\alpha/2}}\right) = \left(\frac{17\cdot 1.8^2}{\chi^2_{17,0.025}},\ \frac{17\cdot 1.8^2}{\chi^2_{17,0.975}}\right) = \left(\frac{17\cdot 3.24}{30.191},\ \frac{17\cdot 3.24}{7.564}\right),$$

δηλαδή το $(1.824385,\ 7.281861)$.

Στην R θα μπορούσαμε να κατασκευάσουμε ένα δ.ε. χρησιμοποιώντας τις ακόλουθες εντολές:

```
1 library(Ecfun)
2 confint.var(1.8^2,18-1,level=0.95)
```

και θα λαμβάναμε το παρακάτω αποτέλεσμα:

```
         lower    upper
[1,] 1.824384 7.281682
attr(,"level")
```

Λύση Άσκησης Αυτοαξιολόγησης 10.11

Έστω X και Y οι τ.μ. που παριστάνουν τη βαθμολογία στο τεστ Αγγλικών στις πόλεις Α και Β, αντίστοιχα. Υποθέτουμε ότι $X \sim N(\mu_1, \sigma_1^2)$ και $Y \sim N(\mu_2, \sigma_2^2)$ και θέλουμε να κατασκευάσουμε ένα δ.ε. για το πηλίκο των πληθυσμιακών διασπορών χρησιμοποιώντας τη σχέση (10.30) με $n = 10, m = 10, \alpha = 0.1$, $s_1^2 = 6.318222$ και $s_2^2 = 3.597333$. Είναι τότε:

$$\left(\frac{1}{F_{n-1,m-1,a/2}} \cdot \frac{s_1^2}{s_2^2}, \frac{1}{F_{n-1,m-1,1-a/2}} \cdot \frac{s_1^2}{s_2^2}\right) = \left(\frac{1}{F_{9,9,0.05}} \cdot \frac{6.318222}{3.597333}, \frac{1}{F_{15,12,0.95}} \cdot \frac{6.318222}{3.597333}\right).$$

Από τον Πίνακα Α΄.8 του Παραρτήματος είναι $F_{9,9,0.05} = 3.18$, ενώ $F_{9,9,0.95} = \frac{1}{F_{9,9,0.05}} = \frac{1}{3.18} = 0.3144654$. Μετά από λίγη άλγεβρα, προκύπτει ότι ένα 90% δ.ε. για το πηλίκο των διακυμάνσεων είναι το $(0.5523154, 5.585234)$.

Στην R θα μπορούσαμε να κατασκευάσουμε ένα δ.ε. χρησιμοποιώντας τις ακόλουθες εντολές:

```
1 data1<-c(12.9, 10.2, 7.4, 7.0, 10.5, 11.9, 7.1, 9.9, 14.4, 11.3)
2 data2<-c(10.2, 6.9, 10.9, 11.0, 10.1, 5.3, 7.5, 10.3, 9.2, 8.8)
3 var.test(data1,data2,ratio=1,alternative="two.sided", conf.level =0.90)$
    conf.int
```

και θα είχαμε το ακόλουθο αποτέλεσμα:

```
[1] 0.5525076 5.5832894
attr(,"conf.level")
[1] 0.9
```

Λύση Άσκησης Αυτοαξιολόγησης 10.12

Ένα 95% (ασυμπτωτικό) διάστημα εμπιστοσύνης για το πραγματικό ποσοστό p των γυναικών που παρακολουθούν αθλητικά στην τηλεόραση θα υπολογιστεί από τη σχέση (10.32), για $\alpha = 0.05, n = 2500$ και $\hat{p} = 920/2500 = 0.368$. Επομένως, είναι το:

$$\left(0.368 - 1.96\sqrt{\frac{0.368 \cdot 0.632}{2500}},\ 0.368 + 1.96\sqrt{\frac{0.368 \cdot 0.632}{2500}}\right) = (0.3490954,\ 0.3869046).$$

Εναλλακτικά, στην R μπορούμε να χρησιμοποιήσουμε τις παρακάτω εντολές:

```
1 library(binom)
2 binom.confint(920, 2500, alpha=0.05,methods=c("asymptotic"))
```

Τα αποτελέσματα που προκύπτουν είναι τα ακόλουθα:

```
      method   x    n  mean     lower     upper
1 asymptotic 920 2500 0.368 0.3490957 0.3869043
```

Τα παραπάνω ισχύουν υπό την προϋπόθεση ότι η πιθανότητα κάποια γυναίκα να παρακολουθεί αθλητικά στην τηλεόραση παραμένει αμετάβλητη από γυναίκα σε γυναίκα και η παρακολούθηση αθλητικών από μια γυναίκα είναι ανεξάρτητη από την παρακολούθηση αθλητικών από οποιαδήποτε άλλη γυναίκα.

Λύση Άσκησης Αυτοαξιολόγησης 10.13

Θέλουμε να κατασκευάσουμε ένα 95% (ασυμπτωτικό) διάστημα εμπιστοσύνης για τη διαφορά των αναλογιών $p_1 - p_2$ με p_1 να είναι το ποσοστό των γυναικών ηλικίας 45 και άνω που έχουν πρεσβυωπία, ενώ p_2 να είναι το ποσοστό των ανδρών ηλικίας 45 και άνω που έχουν πρεσβυωπία. Το ζητούμενο δ.ε. θα υπολογιστεί από τη σχέση (10.33), για $\alpha = 0.05$, $n = 100$, $\hat{p}_1 = 56/100 = 0.56$, $m = 250$, $\hat{p}_2 = 125/250 = 0.5$. Επομένως, το άνω άκρο του δ.ε. είναι το:

$$0.56 - 0.5 - 1.96\sqrt{\frac{0.56 \cdot (1 - 0.56)}{100} + \frac{0.5 \cdot (1 - 0.5)}{250}},$$

ενώ το κάτω άκρο είναι το

$$0.56 - 0.5 + 1.96\sqrt{\frac{0.56 \cdot (1 - 0.56)}{100} + \frac{0.5 \cdot (1 - 0.5)}{250}}.$$

Μετά από λίγη άλγεβρα, προκύπτει ότι ένα 95% (ασυμπτωτικό) διάστημα εμπιστοσύνης για τη διαφορά των ποσοστών $p_1 - p_2$ είναι το $(-0.05535728,\ 0.1753573)$.
Εναλλακτικά, στην R μπορούμε να χρησιμοποιήσουμε τις παρακάτω εντολές:

```
1 library(DescTools)
2 BinomDiffCI(56, 100, 125, 250, conf.level = 0.95, method ="wald")
```

Τότε προκύπτουν τα ακόλουθα αποτελέσματα.

```
     est       lwr.ci    upr.ci
[1,] 0.06 -0.05535516 0.1753552
```

ΒΙΒΛΙΟΓΡΑΦΙΑ

Ελληνόγλωσση

Ζωγράφος, Κ. (2002). *Μαθήματα Πιθανοτήτων και Στατιστικής*. Τυπογραφείο Πανεπιστημίου Ιωαννίνων.

Ηλιόπουλος, Γ. (2006). *Βασικές μέθοδοι εκτίμησης παραμέτρων με σημείο και με διάστημα*. Αθήνα: Αθ. Σταμούλης.

Κουρούκλης, Σ., Πετρόπουλος, Κ. και Πιπερίγκου, Β. (2015). *Θέματα παραμετρικής στατιστικής συμπερασματολογίας* [Προπτυχιακό εγχειρίδιο]. Αθήνα: Κάλλιπος, Ανοικτές Ακαδημαϊκές Εκδόσεις. URL: http://hdl.handle.net/11419/5687

Κουτρουβέλης, Ι. Α. (2000). *Βασικά Εργαλεία και Μέθοδοι για τον Έλεγχο Ποιότητας: Πιθανότητες και Στατιστική ΙΙ (Τόμος Β΄)*. Πάτρα: Ελληνικό Ανοικτό Πανεπιστήμιο.

Παπαϊωάννου, Τ. και Φερεντίνος, Κ. (2000). *Μαθηματική Στατιστική*. Εκδόσεις Σταμούλης.

Ξενόγλωσση

Agresti, A. and Coull, B. A. (1998). Approximate is better than 'exact' for interval estimation of binomial proportions. *The American Statistician*, *52*, pp. 119–126.

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap.* Boca Raton, Florida, USA: Chapman & Hall/CRC.

Ferentinos, K. and Karakostas, K. X. (2006). More on Shortest and Equal Tails Confidence Intervals. *Communication in Statistics-Theory and Methods*, *35*, pp. 821–829.

Lehmann, E. and Casella, G. (1998). *Theory of Point Estimation* (2nd ed.). Springer-Verlang, New York.

Mood, A. M., Graybill, F. and Boes, D. C. (1974). *Introduction to the Theory of Statistics, 3rd Edition*. New York: McGraww-Hill Book Company.

Walpole, R., Myers, R., Myers, S. and Ye, K. (2017). *Probability & Statistics for Engineers & Scientists*. Pearson.

ΚΕΦΑΛΑΙΟ 11

ΕΛΕΓΧΟΙ ΣΤΑΤΙΣΤΙΚΩΝ ΥΠΟΘΕΣΕΩΝ

Σύνοψη

Σε αυτό το κεφάλαιο παρουσιάζονται οι βασικότεροι παραμετρικοί στατιστικοί έλεγχοι υποθέσεων. Πιο συγκεκριμένα, θα παρουσιαστούν έλεγχοι για τη μέση τιμή, την αναλογία, τη διασπορά ενός πληθυσμού αλλά και για τη διαφορά των μέσων και των αναλογιών ή τον λόγο των διασπορών δύο πληθυσμών.

Προαπαιτούμενη γνώση: Κεφάλαια 5, 9 και 10 του παρόντος συγγράμματος.

Προσδοκώμενα μαθησιακά αποτελέσματα: Όταν θα έχετε ολοκληρώσει τη μελέτη του κεφαλαίου αυτού, θα μπορείτε

- να ελέγχετε υποθέσεις για τη μέση τιμή, την αναλογία και τη διασπορά ενός πληθυσμού,
- να ελέγχετε υποθέσεις για τη διαφορά των μέσων τιμών και των αναλογιών, καθώς και τον λόγο των διασπορών δύο πληθυσμών,
- να υπολογίζετε το παρατηρούμενο επίπεδο σημαντικότητας (p-value) των παραπάνω ελέγχων στατιστικών υποθέσεων,
- να προσδιορίζετε πιθανότητες σφάλματος τύπου II,
- να σχολιάζετε και να ερμηνεύετε τα αποτελέσματα των παραπάνω ελέγχων στατιστικών υποθέσεων και
- να χρησιμοποιείτε την R για να υλοποιείτε τους παραπάνω στατιστικούς ελέγχους.

Οικονόμου, Π., Μαλεφάκη, Σ., & Μπατσίδης, Α. (2023). *Πιθανότητες - Στατιστική.*

Γλωσσάριο επιστημονικών όρων

- Δίπλευρος έλεγχος υπόθεσης
- Έλεγχοι στατιστικών υποθέσεων
- Εναλλακτική υπόθεση
- Επίπεδο σημαντικότητας ελέγχου
- Ισχύς ελέγχου
- Κρίσιμη περιοχή
- Μηδενική υπόθεση
- Μονόπλευρος έλεγχος υπόθεσης
- Παρατηρούμενο επίπεδο σημαντικότητας ελέγχου ή $p-$τιμή (p-value)
- Περιοχή απόρριψης
- Στατιστική συνάρτηση ελέγχου
- Σφάλμα τύπου I
- Σφάλμα τύπου II

11.1 Εισαγωγή

Πέρα από την εκτίμηση παραμέτρων, είτε σε σημείο είτε με διάστημα, πολύ συχνά καλούμαστε να επιλύσουμε προβλήματα αποφάσεων, δηλαδή με βάση τα διαθέσιμα δεδομένα καλούμαστε να επιλέξουμε μία από δύο αντικρουόμενες υποθέσεις. Η διαδικασία της Στατιστικής Συμπερασματολογίας με την οποία καταλήγουμε σε μία απόφαση επιλογής της μίας εκ των δύο αντικρουόμενων υποθέσεων ονομάζεται (στατιστικός) έλεγχος υποθέσεων ή έλεγχος στατιστικών υποθέσεων. Στη συνέχεια αυτού του κεφαλαίου, θα παρουσιαστούν οι σημαντικότεροι παραμετρικοί έλεγχοι υπόθεσης και τα βασικότερα χαρακτηριστικά τους.

11.2 Βασικά χαρακτηριστικά ελέγχων υποθέσεων

Πολλές εφαρμογές της στατιστικής αφορούν προβλήματα αποφάσεων στα οποία είναι απαραίτητο να γίνει επιλογή μίας εκ των δύο αντικρουόμενων υποθέσεων βάσει διαθέσιμων παρατηρήσεων. Παραδείγματα τέτοιων υποθέσεων θα μπορούσαν να είναι τα ακόλουθα:

- Το ποσοστό των ελαττωματικών προϊόντων που παράγονται από μία γραμμή παραγωγής ενός εργοστασίου ξεπερνάει το 5% ή όχι;
- Το ποσοστό παιδικής παχυσαρκίας αυξήθηκε στην Ελλάδα τον τελευταίο χρόνο ή όχι;
- Το κόμμα Α θα πάρει το ίδιο ποσοστό στις επόμενες εκλογές με αυτό που πήρε στις προηγούμενες ή όχι;
- Το μέσο pH του διαλύματος Α είναι ίσο με 5.5 ή όχι;
- Η μέθοδος Α για τη μέτρηση της ανάπτυξης ενζύμων σε μία δεξαμενή ζύμωσης είναι ισοδύναμη με τη μέθοδο Β ή όχι;
- Το καινούριο φάρμακο που βγάζει μία φαρμακευτική εταιρεία είναι αποτελεσματικότερο από το παλαιότερο ή όχι;

Οι παραπάνω καθώς και παρόμοιες υποθέσεις, όταν μπορούν να διατυπώνονται σε όρους των παραμέτρων ενός ή περισσότερων πληθυσμών λέγονται στατιστικές υποθέσεις και η διαδικασία της στατιστικής συμπερασματολογίας με την οποία θα καταλήξουμε σε μία απόφαση λέγεται **(στατιστικός) έλεγχος υποθέσεων**.

Στον στατιστικό έλεγχο υποθέσεων η μία εκ των δύο υποθέσεων ονομάζεται **μηδενική υπόθεση** και συμβολίζεται με H_0, ενώ η άλλη ονομάζεται **εναλλακτική υπόθεση** και συμβολίζεται με H_1 ή H_A. Συνήθως, η μηδενική υπόθεση εκφράζει μία καθορισμένη τιμή της παραμέτρου και η εναλλακτική αποκλίσεις από αυτή. Η μηδενική υπόθεση είναι συνήθως της μορφής $H_0 : \theta = \theta_0$, όπου θ μία (άγνωστη) παράμετρος του πληθυσμού και θ_0 μία γνωστή σταθερά, ενώ η εναλλακτική έχει μία εκ των τριών παρακάτω μορφών:

$$H_1 : \theta \neq \theta_0 \text{ ή } H_1 : \theta > \theta_0 \text{ ή } H_1 : \theta < \theta_0.$$

Η εναλλακτική υπόθεση $H_1 : \theta \neq \theta_0$ ονομάζεται δίπλευρη εναλλακτική υπόθεση, ενώ οι υποθέσεις $H_1 : \theta > \theta_0$ και $H_1 : \theta < \theta_0$ ονομάζονται μονόπλευρες εναλλακτικές υποθέσεις. Ποια από τις τρεις μορφές θα έχει η εναλλακτική μας υπόθεση εξαρτάται από τη φύση του προβλήματος που έχουμε να επιλύσουμε.

Παράδειγμα 11.1

Τα βιομηχανικά απόβλητα που ρίχνονται στα ποτάμια απορροφούν το διαλυμένο στο νερό οξυγόνο με συνέπεια αυτό να μειώνεται και, όταν η μέση τιμή του δεν υπερβαίνει τα 5 ppm, να δημιουργείται σοβαρό πρόβλημα επιβίωσης των υδρόβιων οργανισμών. Το πρόβλημα αυτό είχε διαπιστωθεί, πριν από αρκετά χρόνια, και στον ποταμό Καλαμά$^{\alpha'}$. Για την αντιμετώπιση του προβλήματος εφαρμόστηκε ειδικό πρόγραμμα αποκατάστασης και προστασίας του ποταμού. Για να ελεγχθεί αν απέδωσαν τα μέτρα προστασίας, μεταξύ άλλων δεικτών, μελετήθηκε η ποσότητα διαλυμένου οξυγόνου στα νερά του

ποταμού. Διατυπώστε τη μηδενική και την εναλλακτική υπόθεση του στατιστικού ελέγχου που πρέπει να γίνει, ώστε να διαπιστωθεί αν απέδωσαν τα μέτρα προστασίας και αποκατάστασης του ποταμού.

[α] Ο Θύαμις ή Καλαμάς είναι ο μεγαλύτερος ποταμός της Ηπείρου και ο έβδομος μεγαλύτερος στην Ελλάδα. Θύαμις είναι το αρχαίο όνομά του, ενώ Καλαμάς αποκαλούνταν στο παρελθόν ο μεγαλύτερος παραπόταμός του, αλλά με το πέρασμα του χρόνου οι δύο ονομασίες ταυτίστηκαν. Το μήκος του είναι 115 χιλιόμετρα. Οι πηγές του βρίσκονται στο όρος Δούσκο, κοντά στα σύνορα του νομού Ιωαννίνων με την Αλβανία. Εκβάλλει στο Ιόνιο πέλαγος, βόρεια της Ηγουμενίτσας, σχηματίζοντας Δέλτα.

Λύση Παραδείγματος 11.1

Έστω μ η μέση τιμή του οξυγόνου στο νερό του ποταμού Καλαμά. Τότε αν τα μέτρα έχουν αποδώσει θα πρέπει το μέσο διαλυμένο στο νερό οξυγόνο να είναι περισσότερο από την τιμή των $5\ ppm$. Άρα η μηδενική υπόθεση σε αυτό το πρόβλημα είναι η $H_0 : \mu = 5$, ενώ η εναλλακτική έχει τη μορφή $H_1 : \mu > 5$. Επομένως, αν απορριφθεί η μηδενική υπόθεση θα συμπεράνουμε ότι τα μέτρα προστασίας και αποκατάστασης του ποταμού απέδωσαν.

Σημειώνεται ότι οι υποθέσεις που θέλουμε να ελέγξουμε δεν θα μπορούσαν να ήταν οι $H_0 : \mu = 5$ και $H_1 : \mu < 5$, καθώς σε μια τέτοια περίπτωση

- αν απορρίπταμε τη μηδενική υπόθεση προς όφελος της εναλλακτικής, θα συμπεραίναμε ότι τα μέτρα προστασίας και αποκατάστασης του ποταμού δεν απέδωσαν, ενώ
- αν δεν απορρίπταμε τη μηδενική υπόθεση, δεν θα μπορούσαμε να απαντήσουμε αν το μέσο διαλυμένο στο νερό οξυγόνο υπερβαίνει την τιμή των $5\ ppm$ και, επομένως, δεν θα μπορούσαμε να συμπεράνουμε αν υπάρχει σοβαρό πρόβλημα επιβίωσης των υδρόβιων οργανισμών στον ποταμό.

Στην ίδια αδυναμία εξαγωγής ξεκάθαρου συμπεράσματος θα καταλήγαμε, αν η εναλλακτική υπόθεση ήταν η $H_1 : \mu \neq 5$ και απορρίπταμε τη μηδενική υπόθεση.

Παρατηρούμε, λοιπόν, ότι η διατύπωση της μηδενικής και της εναλλακτικής υπόθεσης γίνεται με τέτοιο τρόπο, έτσι ώστε οποιοδήποτε και να είναι το αποτέλεσμα του ελέγχου (απόρριψη ή μη της μηδενικής υπόθεσης) να μπορούμε να εξάγουμε ένα ξεκάθαρο συμπέρασμα για το ερώτημα που μας απασχολεί.

Παράδειγμα 11.2

Ένας έμπορος έχει παραγγείλει μια μεγάλη παρτίδα κάποιου τυποποιημένου προϊόντος. Από τις προδιαγραφές είναι γνωστό ότι το βάρος κάθε τεμαχίου του προϊόντος ακολουθεί κανονική κατανομή με μέση τιμή 2 κιλά και τυπική απόκλιση 50 γραμμάρια. Ο έμπορος υποψιάζεται ότι το μέσο βάρος των παραγόμενων κομματιών είναι μικρότερο από αυτό που ορίζεται στις προδιαγραφές. Διατυπώστε τη μηδενική και την εναλλακτική υπόθεση του ελέγχου που πρέπει να γίνει ώστε να ελεγχθούν οι υποψίες του εμπόρου.

Λύση Παραδείγματος 11.2

Ο έμπορος θέλει να ελέγξει, αν το πραγματικό μέσο βάρος, έστω μ, του προϊόντος είναι πράγματι ίσο με 2 κιλά ή μικρότερο. Δηλαδή, σε αυτήν την περίπτωση, η μηδενική υπόθεση είναι $H_0 : \mu = 2\ kg$ και η εναλλακτική $H_1 : \mu < 2\ kg$. Επομένως, αν απορριφθεί η μηδενική υπόθεση θα σημαίνει ότι το μέσο βάρος του προϊόντος είναι μικρότερο από τις προδιαγραφές.

Συνεπώς, το πρώτο και ίσως το σημαντικότερο βήμα ενός ελέγχου στατιστικών υποθέσεων είναι ο καθορισμός της μηδενικής και της εναλλακτικής υπόθεσης. Στη συνέχεια, επιλέγεται ένα τυχαίο δείγμα από τον υπό μελέτη ή τους υπό μελέτη πληθυσμούς και η απόφαση, αν θα απορρίψουμε τη μηδενική υπόθεση H_0 ή όχι, βασίζεται στην παρατηρούμενη τιμή μιας στατιστικής συνάρτησης. Η στατιστική συνάρτηση που χρησιμοποιείται λέγεται **στατιστική συνάρτηση ελέγχου (σσε)** ή στατιστικό του ελέγχου (test statistic). Για

να μπορεί να είναι μια συνάρτηση σσε, θα πρέπει να έχει μια πλήρως καθορισμένη κατανομή υπό τη μηδενική υπόθεση, δηλαδή όταν $\theta = \theta_0$. Έπειτα το σύνολο των τιμών της σσε χωρίζεται σε δύο περιοχές, που είναι ξένες μεταξύ τους. Στη μία περιοχή γίνεται αποδεκτή η μηδενική υπόθεση, ενώ στην άλλη απορρίπτεται. Η περιοχή που απορρίπτεται η μηδενική υπόθεση ονομάζεται **περιοχή απόρριψης** ή **κρίσιμη περιοχή** (κπ) (critical region) του ελέγχου. Στη συνέχεια, υπολογίζεται η τιμή της σσε στο συγκεκριμένο δείγμα και ανάλογα με το αν βρίσκεται η τιμή αυτή στην κρίσιμη περιοχή ή όχι, απορρίπτουμε τη μηδενική υπόθεση ή όχι και εξάγουμε τα αντίστοιχα συμπεράσματα. Συνοψίζοντας, θα μπορούσαμε να πούμε ότι η διαδικασία ενός στατιστικού ελέγχου υποθέσεων αποτελείται από τα παρακάτω έξι βήματα:

1. Ορισμός μηδενικής υπόθεσης.
2. Ορισμός εναλλακτικής υπόθεσης.
3. Εύρεση μιας στατιστικής συνάρτησης ελέγχου.
4. Ορισμός περιοχής απόρριψης της H_0.
5. Υπολογισμός της τιμής της στατιστικής συνάρτησης ελέγχου στο δείγμα μας.
6. Εξαγωγή και διατύπωση συμπερασμάτων.

Είναι προφανές ότι η απόφαση που παίρνουμε υπέρ ή κατά της εναλλακτικής υπόθεσης στηρίζεται σε μία παρατήρηση της σσε, πράγμα που σημαίνει ότι δεν μπορούμε να είμαστε απόλυτα βέβαιοι για την ορθότητα των αποφάσεών μας. Ειδικότερα, σε κάθε στατιστικό έλεγχο υποθέσεων κινδυνεύουμε να κάνουμε δύο ειδών λάθη: είτε να απορρίψουμε την H_0, ενώ στην πραγματικότητα αυτή ισχύει (**σφάλμα τύπου Ι**), είτε να δεχτούμε τη μηδενική υπόθεση, ενώ στην πραγματικότητα ισχύει η εναλλακτική (**σφάλμα τύπου ΙΙ**). Προφανώς, σφάλμα τύπου Ι μπορούμε να έχουμε μόνο στην περίπτωση που η τιμή της σσε είναι στην κπ, ενώ σφάλμα τύπου ΙΙ έχουμε μόνο στην περίπτωση που η τιμή της σσε βρίσκεται στην περιοχή που δεν μπορούμε να απορρίψουμε τη μηδενική υπόθεση.

Στην πραγματικότητα, δεν γνωρίζουμε, αλλά ούτε και μπορούμε να μάθουμε αν ισχύει η μηδενική υπόθεση, οπότε δεν μπορούμε να ξέρουμε με σιγουριά αν και τι είδους σφάλμα διαπράξαμε. Ωστόσο, αυτό που μπορούμε να υπολογίσουμε, είναι οι πιθανότητες σφάλματος τύπου Ι και τύπου ΙΙ. Η πιθανότητα σφάλματος τύπου Ι ονομάζεται **επίπεδο σημαντικότητας** του ελέγχου, συμβολίζεται με α και ορίζεται μονοσήμαντα λόγω της μορφής της μηδενικής υπόθεσης. Από την άλλη πλευρά, η πιθανότητα σφάλματος τύπου ΙΙ, λόγω της μορφής της εναλλακτικής υπόθεσης, συνήθως δεν ορίζεται μονοσήμαντα, αφού η τιμή της εξαρτάται από την πραγματική άγνωστη τιμή της παραμέτρου θ υπό την H_1. Για τον λόγο αυτόν η πιθανότητα σφάλματος τύπου ΙΙ συμβολίζεται με $\beta(\theta)$ και έχει πεδίο ορισμού την περιοχή τιμών της θ που καθορίζει η εναλλακτική υπόθεση. Η γραφική παράσταση της συνάρτησης $\beta(\theta)$ ονομάζεται **χαρακτηρίζουσα καμπύλη του ελέγχου**.

Τέλος, η πιθανότητα να απορρίψουμε τη μηδενική υπόθεση, ενώ στην πραγματικότητα ισχύει η εναλλακτική υπόθεση, ονομάζεται **ισχύς** του ελέγχου και ισούται με $\gamma(\theta) = 1 - \beta(\theta)$. Η πιθανότητα σφάλματος τύπου ΙΙ, άρα και η ισχύς, εξαρτάται, εκτός, από την πραγματική τιμή του θ, από το μέγεθος του δείγματος αλλά και από το επίπεδο σημαντικότητας του ελέγχου. Είναι προφανές ότι η πιθανότητα $\beta(\theta)$ μπορεί να είναι πραγματικά μεγάλη όταν το μέγεθος του δείγματος είναι μικρό ή όταν η πραγματική τιμή του θ διαφέρει λίγο από την τιμή θ_0, που καθορίζει η H_0.

Ο τρόπος υπολογισμού των πιθανοτήτων σφάλματος Ι και ΙΙ παρουσιάζεται στο επόμενο παράδειγμα.

Παράδειγμα 11.3

Έστω $X_1, X_2, \ldots, X_9$ τυχαίο δείγμα από κανονικό πληθυσμό με μέση τιμή μ και τυπική απόκλιση 2. Βρείτε τις πιθανότητες σφάλματος τύπου Ι και τύπου ΙΙ στον έλεγχο

$$H_0 : \mu = 10 \quad \text{κατά} \quad H_1 : \mu < 10$$

με κρίσιμη περιοχή $C = \{\bar{X} : \ \bar{X} < 9.1\}$ όταν η πραγματική μέση τιμή είναι $\mu = 8.5$.

Λύση Παραδείγματος 11.3

Η πιθανότητα σφάλματος τύπου Ι ισούται με:

$$\begin{aligned}\alpha &= P(\text{ σφάλμα τύπου Ι}) = P(\text{ απόρριψη } H_0 | H_0 \text{ ισχύει}) = P(\bar{X} < 9.1 | \mu = 10)\\ &= P\left(\frac{\bar{X}-10}{2/3} < \frac{9.1-10}{2/3}\right) = P(Z < -1.35) = P(Z > 1.35) = 1 - P(Z < 1.35)\\ &= 1 - 0.9115 = 0.0885,\end{aligned}$$

όπου χρησιμοποιήθηκε ότι, υπό τη μηδενική υπόθεση, η τυχαία μεταβλητή $\bar{X} \sim N\left(10, \frac{4}{9}\right)$.
Η πιθανότητα σφάλματος τύπου ΙΙ ισούται με:

$$\begin{aligned}\beta(\mu) &= P(\text{ σφάλμα τύπου ΙΙ}) = P(\text{ μη απόρριψη } H_0 | \mu < 10) = P(\bar{X} > 9.1 | \mu < 10)\\ &= P\left(\frac{\bar{X}-\mu}{2/3} > \frac{9.1-\mu}{2/3}\right) = 1 - P\left(Z < \frac{9.1-\mu}{2/3}\right),\end{aligned}$$

όπου τώρα χρησιμοποιήθηκε ότι υπό την εναλλακτική υπόθεση η τυχαία μεταβλητή $\bar{X} \sim N\left(\mu, \frac{4}{9}\right)$.
Όταν $\mu = 8.5$ είναι:

$$\begin{aligned}\beta(8.5) &= 1 - P\left(Z < \frac{9.1-8.5}{2/3}\right) = 1 - P(Z < 0.9)\\ &= 1 - 0.8160 = 0.1840.\end{aligned}$$

Στα ίδια αποτελέσματα μπορούμε να καταλήξουμε χρησιμοποιώντας τις ακόλουθες εντολές της R:

`pnorm(9.1,10,2/3,lower.tail=T)` και `pnorm(9.1,8.5,2/3,lower.tail=F),`

οι οποίες δίνουν ως αποτελέσματα τις τιμές 0.08850799 και 0.1840601, αντίστοιχα.

Είναι προφανές ότι από όλες τις κρίσιμες περιοχές για έναν έλεγχο υποθέσεων προτιμότερη θα ήταν εκείνη για την οποία οι πιθανότητες σφάλματος τύπου Ι και ΙΙ είναι ελάχιστες. Μπορεί όμως να αποδειχθεί (η απόδειξη αφήνεται ως άσκηση στον/στην αναγνώστη/στρια) ότι η ταυτόχρονη ελαχιστοποίηση αυτών των πιθανοτήτων είναι αδύνατη. Για να ξεπεραστεί το πρόβλημα αυτό, προκαθορίζεται το επίπεδο σημαντικότητας α του ελέγχου και χρησιμοποιούνται κατάλληλες σσε και κπ τέτοιες ώστε να δώσουν τις μικρότερες δυνατές τιμές για τη συνάρτηση $\beta(\theta)$ (βλ. Neyman and Pearson, 1933). Για τον λόγο αυτόν η μηδενική υπόθεση διατυπώνεται με την ελπίδα να απορριφθεί, καθώς τότε είναι προκαθορισμένη η πιθανότητα σφάλματος τύπου Ι (είναι ίση με α).

Στην πράξη όμως, όταν δουλεύουμε με κάποιο στατιστικό πακέτο, όπως η R, δεν χρειάζεται να προκαθορίσουμε το α, αλλά προσδιορίζεται το μέγεθος των ενδείξεων που δίνουν τα δεδομένα κατά της μηδενικής υπόθεσης και αυτό επιτυγχάνεται μέσω της p-τιμής (p-value) ή του παρατηρούμενου επιπέδου σημαντικότητας (Κουτρουβέλης, 2000).

Ορισμός 11.1

Παρατηρούμενο επίπεδο σημαντικότητας ($p-$τιμή) είναι η πιθανότητα να παρατηρήσουμε, υπό τη μηδενική υπόθεση, μία τιμή της σσε που να είναι ίση ή ακόμα πιο ακραία από αυτή που παρατηρούμε.

Σύμφωνα με τον παραπάνω ορισμό, αν για κάποιον έλεγχο η μηδενική υπόθεση απορρίπτεται για μεγάλες τιμές της σσε $T(X_1,\dots,X_n)$, δηλαδή αν η κπ είναι της μορφής $T(X_1,\dots,X_n) > C$, όπου C σταθερά, αν $T(x_1,\dots,x_n)$ είναι η παρατηρηθείσα τιμή της $T(X_1,\dots,X_n)$, τότε η $p-$τιμή του ελέγχου ισούται με την πιθανότητα $P(T(X_1,\dots,X_n) > T(x_1,\cdots,x_n)|H_0$ αληθής).

Το παρατηρούμενο επίπεδο σημαντικότητας θα μπορούσε να ερμηνευθεί ως η πιθανότητα να παρατηρήσουμε κάτι πιο ακραίο από αυτό που παρατηρήσαμε, δηλαδή κάτι που να υποστηρίζει περισσότερο την εναλλακτική υπόθεση, δοθέντος ότι η μηδενική υπόθεση αληθεύει.

Έλεγχοι που βασίζονται στο παρατηρούμενο επίπεδο σημαντικότητας αναφέρονται συχνά και ως **έλεγχοι σημαντικότητας**. Ο υπολογισμός του παρατηρούμενου επιπέδου σημαντικότητας γίνεται υπό την H_0, άρα όσο μικρότερη είναι η τιμή της, τόσο μεγαλύτερες είναι οι ενδείξεις κατά της H_0. Το παρατηρούμενο επίπεδο σημαντικότητας δεν εξαρτάται από το επίπεδο σημαντικότητας α του ελέγχου. Όμως, καθορίζοντας ένα επίπεδο σημαντικότητας α για τον έλεγχο, απορρίπτουμε την H_0 όταν $p-value < \alpha$. Στην πραγματικότητα, το παρατηρούμενο επίπεδο σημαντικότητας δίνει το μικρότερο επίπεδο σημαντικότητας α ενός ελέγχου για το οποίο η μηδενική υπόθεση θα απορριφθεί με βάση τα δεδομένα ή, ισοδύναμα, θα μπορούσαμε να πούμε ότι ισούται με το μεγαλύτερο επίπεδο σημαντικότητας α ενός ελέγχου για το οποίο δεν θα απορριπτόταν η μηδενική υπόθεση με βάση τα διαθέσιμα δεδομένα.

Στη συνέχεια, θα παρουσιαστούν οι σημαντικότεροι έλεγχοι στατιστικής υπόθεσης και η υλοποίησή τους μέσω της R. Επισημαίνεται ότι θα δοθούν οι σσε και οι κπ των ελέγχων υποθέσεων που θα μας απασχολήσουν με προκαθορισμένο επίπεδο σημαντικότητας α και με τις μικρότερες δυνατές πιθανότητες σφάλματος τύπου II χωρίς απόδειξη. Όλοι οι έλεγχοι, που θα παρουσιαστούν στις επόμενες υποενότητες, βασίζονται στην υπόθεση της κανονικότητας των πληθυσμών, εκτός από αυτούς που αφορούν μεγάλα δείγματα. Στην περίπτωση, όμως, που η κατανομή του πληθυσμού ή των πληθυσμών δεν είναι κανονική/-ές και τα διαθέσιμα δείγματα είναι μικρά, δεν μπορούμε να χρησιμοποιούμε τους παραμετρικούς ελέγχους υποθέσεων. Σε αυτήν την περίπτωση χρησιμοποιούμε μία ομάδα ελέγχων που ονομάζονται μη παραμετρικοί, γιατί δεν υποθέτουν καμία συγκεκριμένη κατανομή για τον πληθυσμό (Bethea *et al.*, 1995; Sprent and Smeeton, 2016).

11.3 Έλεγχος υπόθεσης για την πληθυσμιακή μέση τιμή

Έστω $X_1,X_2,\dots,X_n$ ένα τυχαίο δείγμα από έναν πληθυσμό με μέση τιμή μ και διασπορά σ^2. Στην ενότητα αυτή το ενδιαφέρον μας επικεντρώνεται στον έλεγχο της υπόθεσης $H_0:\ \mu = \mu_0$ κατά μίας από τις εναλλακτικές $H_1:\ \mu \neq \mu_0$ ή $H_1:\ \mu < \mu_0$ ή $H_1:\ \mu > \mu_0$. Στο πλαίσιο αυτό, θα διακρίνουμε τρεις περιπτώσεις.

Παρατήρηση 11.1

Σε όλες τις περιπτώσεις που ακολουθούν σε αυτήν αλλά και στις υπόλοιπες ενότητες που πραγματεύονται ελέγχους υποθέσεων για τη μέση τιμή ή τη διασπορά ενός πληθυσμού ή τη διαφορά δύο πληθυσμιακών μέσων ή το πηλίκο δύο πληθυσμιακών διασπορών, θα υποθέτουμε ότι τα δεδομένα μας δεν περιέχουν ακραίες τιμές, δηλαδή παρατηρήσεις με πολύ μεγάλες ή πολύ μικρές τιμές σε σχέση με τις υπόλοιπες.

Περίπτωση Ι. Κανονικός πληθυσμός με γνωστή διασπορά.

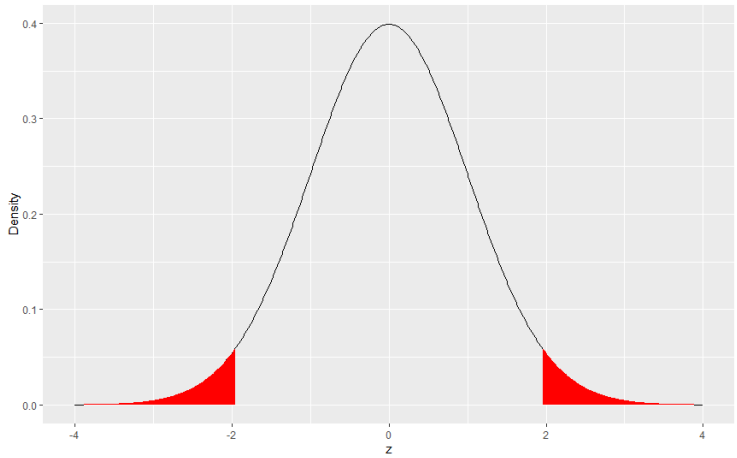
Σε αυτήν την περίπτωση εκτιμούμε τη μέση τιμή μ του πληθυσμού με τον δειγματικό μέσο $\bar{X}$ για τον οποίο γνωρίζουμε ότι: $\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ ή, ισοδύναμα, $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$, όπου μ είναι η αληθινή πληθυσμιακή μέση τιμή.

Επομένως, υπό τη μηδενική υπόθεση ισχύει (βλ. το Κεφάλαιο 9)

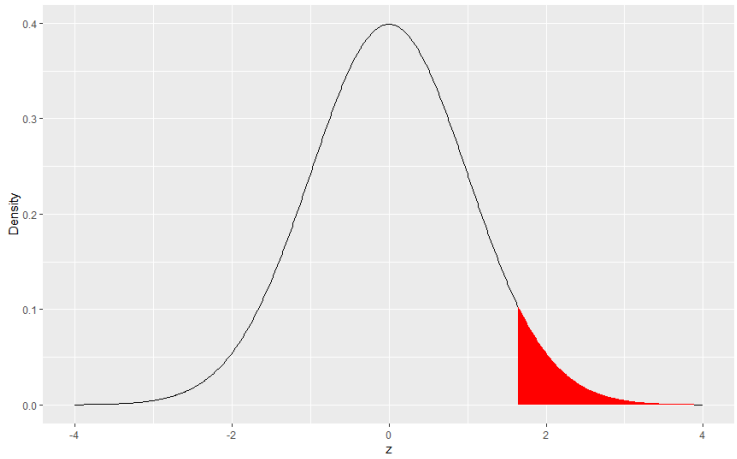
$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \overset{H_0}{\sim} N(0,1) \tag{11.1}$$

όπου το σύμβολο $\overset{H_0}{\sim}$ δηλώνει ότι η τ.μ. Z ακολουθεί την τυπική κανονική κατανομή υπό τη μηδενική υπόθεση. Η παραπάνω στατιστική συνάρτηση αποτελεί και τη σσε για τον συγκεκριμένο έλεγχο. Επιπλέον, έχοντας τις παρατηρήσεις $x_1, \ldots, x_n$, η τιμή της σσε ισούται με $z = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$, όπου $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ και σε επίπεδο σημαντικότητας α η κπ του παραπάνω ελέγχου είναι τέτοια ώστε:

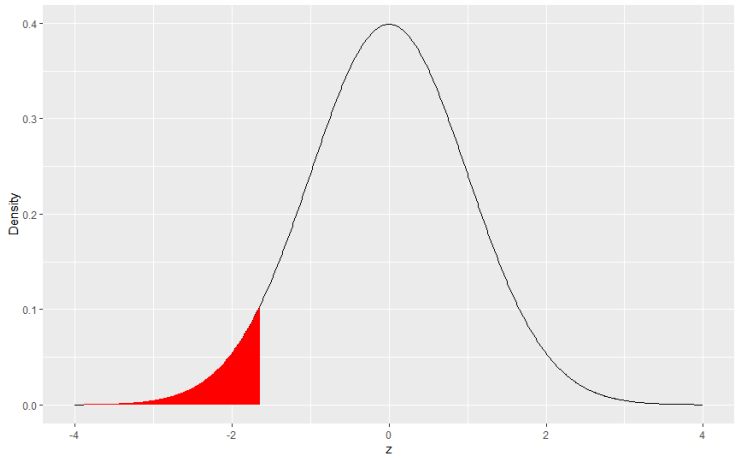
1. αν $H_1 : \mu \neq \mu_0$, τότε η μηδενική υπόθεση απορρίπτεται, αν $z < -z_{\alpha/2}$ ή $z > z_{\alpha/2}$.
   Η κρίσιμη αυτή περιοχή προκύπτει από το γεγονός ότι, για να απορρίψουμε τη μηδενική υπόθεση, θα πρέπει να παρατηρήσουμε μια δειγματική μέση τιμή που να διαφέρει αρκετά από την τιμή μ_0, δηλαδή είτε να είναι πολύ μικρότερή της είτε πολύ μεγαλύτερή της, αφού η εναλλακτική υπόθεση είναι η $H_1 : \mu \neq \mu_0$. Σε μια τέτοια περίπτωση, ο αριθμητής, άρα και η σσε, λαμβάνουν τιμές μακριά από το μηδέν. Συνέπεια αυτού είναι τιμές της σσε που απέχουν κατά απόλυτη τιμή πολύ από το μηδέν να υποδηλώνουν ότι η μηδενική υπόθεση πρέπει να απορριφθεί. Το τι θεωρείται μακριά από το μηδέν καθορίζεται από το επίπεδο σημαντικότητας του ελέγχου και την κατανομή της σσε υπό τη μηδενική υπόθεση.
   Στη συγκεκριμένη περίπτωση, η κατανομή της σσε υπό τη μηδενική υπόθεση είναι η κανονική και το επίπεδο σημαντικότητας, δηλαδή η πιθανότητα σφάλματος Ι, ισούται με α. Αυτό σημαίνει ότι μπορούμε να προσδιορίσουμε δύο σημεία, z_1 και z_2, της τυπικής κανονικής κατανομής, έτσι ώστε η $P(Z < z_1 \cup Z > z_2 | H_0) = \alpha$. Λόγω της συμμετρίας της τυπικής κανονικής κατανομής δύο προφανείς επιλογές για τα z_1 και z_2 είναι τα σημεία $-z_{\alpha/2}$ και $z_{\alpha/2}$, τα οποία αφήνουν εμβαδόν χωρίου $\alpha/2$ αριστερά και δεξιά τους αντίστοιχα, όπου z_α το $(1-\alpha)-$ποσοστιαίο σημείο της τυπικής κανονικής κατανομής (βλ. ακόλουθο σχήμα). Άρα η κρίσιμη περιοχή είναι πράγματι η $z < -z_{\alpha/2}$ ή $z > z_{\alpha/2}$.



2. Αν $H_1 : \mu > \mu_0$, τότε η μηδενική υπόθεση απορρίπτεται, αν $z > z_\alpha$.
   Η αιτιολόγηση της παραπάνω κρίσιμης περιοχής ακολουθεί το ίδιο σκεπτικό με πριν, με τη διαφοροποίηση ότι υποστηρικτικές τιμές για την απόρριψη της μηδενικής υπόθεσης είναι μόνο οι πολύ μεγάλες τιμές, αφού η εναλλακτική υπόθεση σε αυτήν την περίπτωση είναι της μορφής $H_1 : \mu > \mu_0$. Η κπ περιοχή αποτυπώνεται και γραφικά στο παρακάτω σχήμα και ικανοποιεί τη συνθήκη $P(Z > z_\alpha | H_0) = \alpha$.



3. Αν $H_1 : \mu < \mu_0$, τότε η μηδενική υπόθεση απορρίπτεται, αν $z < -z_\alpha$.
   Η αιτιολόγηση της παραπάνω κρίσιμης περιοχής, η οποία αποτυπώνεται και γραφικά στο σχήμα που ακολουθεί, για αυτήν την περίπτωση, προκύπτει ακολουθώντας ανάλογο σκεπτικό με πριν και αφήνεται στον/στην αναγνώστη/στρια για εξάσκηση.



Όπως έχει προαναφερθεί, κάθε έλεγχος περιέχει την πιθανότητα λήψης λανθασμένης απόφασης. Η κπ του ελέγχου κατασκευάστηκε με τέτοιο τρόπο έτσι ώστε να εξασφαλίζεται το επίπεδο σημαντικότητας, δηλαδή η πιθανότητα λανθασμένης απόρριψης της μηδενικής υπόθεσης, ενώ αυτή ισχύει (σφάλμα τύπου I). Πέρα από το σφάλμα τύπου I υπάρχει και το ενδεχόμενο να υποπέσουμε σε σφάλμα τύπου II. Μάλιστα, η πιθανότητα τύπου II καθορίζει την ισχύ του ελέγχου. Στη συνέχεια, υπολογίζεται η ισχύς του ελέγχου $H_0 : \mu = \mu_0$ κατά $H_1 : \mu < \mu_0$.

Για τον υπολογισμό της ισχύος του έλεγχου πρέπει να υιοθετήσουμε μια τιμή μ για τη μέση τιμή του πληθυσμού διαφορετική της μ_0, δηλαδή μια τιμή που δεν αντιπροσωπεύει τη μηδενική υπόθεση και είναι τέτοια ώστε να ικανοποιείται η εναλλακτική υπόθεση. Έχοντας αυτό υπόψη, η ισχύς του ελέγχου υπολογίζεται με την ακόλουθη διαδικασία.

$$\begin{aligned}\gamma(\mu) &= P(\text{ απόρριψη } H_0 | \text{ ισχύει } H_1) = P(Z < -z_\alpha | \text{ ισχύει } H_1) \\ &= P\left(\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} < -z_\alpha | \text{ ισχύει } H_1\right) = P\left(\bar{X} < \mu_0 - \frac{z_\alpha \sigma}{\sqrt{n}} | \text{ ισχύει } H_1\right) \\ &= P\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < -z_\alpha + \frac{\mu_0 - \mu}{\sigma/\sqrt{n}}\right) = P\left(Z' < -z_\alpha - \frac{(\mu - \mu_0)\sqrt{n}}{\sigma}\right),\end{aligned}$$

με $Z' = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$ υπό την εναλλακτική υπόθεση. Επομένως,

$$\gamma(\mu) = \Phi\left(-z_a - \frac{(\mu-\mu_0)\sqrt{n}}{\sigma}\right), \quad \text{με } \mu < \mu_0. \tag{11.2}$$

Με παρόμοιο τρόπο μπορούμε να προσδιορίσουμε την ισχύ του ελέγχου $H_0 : \mu = \mu_0$ κατά $H_1 : \mu > \mu_0$, η οποία δίνεται από τη σχέση

$$\gamma(\mu) = 1 - \Phi\left(z_\alpha - \frac{(\mu-\mu_0)\sqrt{n}}{\sigma}\right), \quad \text{με } \mu > \mu_0, \tag{11.3}$$

ενώ η ισχύς του ελέγχου $H_0 : \mu = \mu_0$ κατά $H_1 : \mu \neq \mu_0$ δίνεται από τη σχέση

$$\gamma(\mu) = 1 - \Phi\left(z_{\alpha/2} - \frac{(\mu-\mu_0)\sqrt{n}}{\sigma}\right) + \Phi\left(-z_{\alpha/2}\frac{(\mu-\mu_0)\sqrt{n}}{\sigma}\right), \quad \text{με } \mu \neq \mu_0. \tag{11.4}$$

Οι αποδείξεις του τύπου της ισχύος για τις δύο τελευταίες περιπτώσεις αφήνονται ως άσκηση στον/στην αναγνώστη/στρια.

Από τις παραπάνω σχέσεις μπορούμε να δούμε ότι για δεδομένο επίπεδο σημαντικότητας α και δεδομένο μέγεθος δείγματος n όσο η πραγματική τιμή μ απομακρύνεται από το μ_0, τόσο μεγαλύτερη είναι η ισχύς του ελέγχου. Με άλλα λόγια, όταν η πραγματική τιμή είναι αρκετά μακριά από το μ_0, ο έλεγχος έχει μεγάλη πιθανότητα απόρριψης της H_0.

Από την άλλη, όταν η πραγματική τιμή είναι πολύ κοντά στην τιμή μ_0, η ισχύς του ελέγχου ισούται περίπου με το επίπεδο σημαντικότητας του ελέγχου α, ανεξάρτητα από το μέγεθος του δείγματος. Συνέπεια αυτού είναι η πιθανότητα σφάλματος τύπου ΙΙ, σε αυτήν την περίπτωση να είναι περίπου ίση με $1-\alpha$. Αυτό προκύπτει επειδή, αν το μ είναι πολύ κοντά στο μ_0, τότε η ποσότητα $\frac{(\mu-\mu_0)\sqrt{n}}{\sigma}$ σχεδόν μηδενίζεται. Επομένως, για τον έλεγχο της $H_0 : \mu = \mu_0$ κατά $H_1 : \mu < \mu_0$ θα έχουμε άμεσα από τη σχέση (11.2) ότι:

$$\gamma(\mu) \approx \Phi(-z_\alpha) = \alpha \quad \text{με } \mu < \mu_0, \text{ αλλά πολύ κοντά στο } \mu_0.$$

Επιπλέον, για τον έλεγχο της $H_0 : \mu = \mu_0$ κατά $H_1 : \mu > \mu_0$ θα έχουμε άμεσα από τη σχέση (11.3) ότι:

$$\gamma(\mu) \approx 1 - \Phi(z_\alpha) = \Phi(-z_\alpha) = a \quad \text{με } \mu > \mu_0, \text{ αλλά πολύ κοντά στο } \mu_0.$$

Τέλος, για τον έλεγχο της $H_0 : \mu = \mu_0$ κατά $H_1 : \mu \neq \mu_0$ θα έχουμε άμεσα από τη σχέση (11.3) ότι:

$$\gamma(\mu) \approx 1 - \Phi\left(z_{\alpha/2}\right) + \Phi\left(-z_{\alpha/2}\right) = 2\Phi\left(-z_{\alpha/2}\right) = 2 \cdot \frac{\alpha}{2} = \alpha, \quad \text{όταν } \mu \neq \mu_0, \text{ αλλά πολύ κοντά στο } \mu_0.$$

Τέλος, για δεδομένο επίπεδο σημαντικότητας α και δεδομένη τιμή μ η ισχύς του ελέγχου αυξάνεται, όταν αυξάνεται το μέγεθος του δείγματος.

Παράδειγμα 11.4: (συνέχεια Παραδείγματος 11.2)

Σε συνέχεια του Παραδείγματος 11.2 και θέλοντας να ελέγξει τις υποψίες του ο έμπορος επιλέγει τυχαία 16 προϊόντα και καταγράφει το βάρος τους.

1. Προσδιορίστε κατάλληλη στατιστική συνάρτηση και κρίσιμη περιοχή για τον έλεγχο υπόθεσης που θέλει να κάνει ο έμπορος, όταν έχει αποφασίσει ότι με πιθανότητα 0.95 θα κάνει δεκτή την παρτίδα όταν ισχύουν οι προδιαγραφές στην πραγματικότητα.
2. Αν στα 16 τυχαία επιλεγμένα προϊόντα προκύψει μέσο βάρος ίσο με 1.985 κιλά, είναι βάσιμες οι υποψίες του εμπόρου;

Λύση Παραδείγματος 11.4

1. Έστω X η τ.μ. που παριστάνει το βάρος ενός προϊόντος. Από την εκφώνηση του Παραδείγματος 11.2 έχουμε ότι το βάρος ακολουθεί κανονική κατανομή με τυπική απόκλιση 50 γραμμάρια και θέλουμε να ελέγξουμε αν η μέση τιμή του είναι ίση με δοθείσα γνωστή τιμή ($H_0 : \mu = 2$) έναντι της εναλλακτικής ότι είναι μικρότερη ($H_1 : \mu < 2$). Καθώς η διασπορά του κανονικού πληθυσμού είναι γνωστή, η σσε δίνεται από τη σχέση (11.1) με $\mu_0 = 2$, $\sigma = 0.05$ και $n = 16$, δηλαδή θα χρησιμοποιήσουμε τη σσε
$$Z = \frac{\bar{X} - 2}{0.05/\sqrt{16}} = \frac{\bar{X} - 2}{0.0125}.$$
Η μηδενική υπόθεση απορρίπτεται, σε επίπεδο σημαντικότητας α, όταν η παρατηρούμενη τιμή z της τυχαίας μεταβλητής Z είναι $z < -z_\alpha$. Από την εκφώνηση έχουμε ότι ο έμπορος επιθυμεί η $P($ μη απόρριψη της $H_0 | H_0$ αληθής$) = 0.95$ ή, ισοδύναμα, η $P($ απόρριψη της $H_0 | H_0$ αληθής$) = 1 - 0.95 = 0.05$. Επομένως, επιθυμεί έλεγχο με επίπεδο σημαντικότητας $\alpha = 0.05$ και η μηδενική υπόθεση απορρίπτεται αν $z < -z_{0.05} = -1.645$ (Πίνακας Α΄.3, Παραρτήματος Α΄).
2. Η τιμή της σσε είναι $z = \frac{1.985-2}{0.0125} = -1.2$, οπότε καθώς $z > -z_{0.05}$ με βάση αυτά τα δεδομένα δεν μπορούμε να απορρίψουμε τη μηδενική υπόθεση. Επομένως, σε επίπεδο σημαντικότητας 0.05, συμπεραίνουμε ότι δεν είναι βάσιμες οι υποψίες του.
Στο ίδιο συμπέρασμα μπορούμε να καταλήξουμε, εκτελώντας στην R τις ακόλουθες εντολές

```
1 library(PASWR)
2 zsum.test(mean.x=1.985, sigma.x =0.05, n.x =16,alternative = "less",
    mu = 2,
3 conf.level = 0.95)
```

οι οποίες επιστρέφουν τα εξής αποτελέσματα:

```
      One-sample z-Test

data:  Summarized x
z = -1.2, p-value = 0.1151
alternative hypothesis: true mean is less than 2
95 percent confidence interval:
     -Inf 2.005561
sample estimates:
mean of x
    1.985
```

Από τα παραπάνω αποτελέσματα παρατηρούμε ότι η $p-$τιμή του ελέγχου ισούται με $0.1151 > 0.05$, άρα δεν απορρίπτεται η μηδενική υπόθεση και, επομένως, πράγματι καταλήγουμε στο ίδιο συμπέρασμα με πριν.

Παράδειγμα 11.5

Σε συνέχεια του προηγούμενου παραδείγματος απαντήστε τα ακόλουθα ερωτήματα:

1. Υπολογίστε την πιθανότητα να απορρίψει ο έμπορος την παρτίδα, αν το πραγματικό μέσο βάρος των παραγόμενων προϊόντων είναι 1.98 κιλά.
2. Διατηρώντας το ίδιο επίπεδο σημαντικότητας στον έλεγχο, ποιος είναι ο ελάχιστος αριθμός προϊόντων που θα πρέπει να ελέγξει ο έμπορος ώστε η πιθανότητα σφάλματος τύπου ΙΙ να είναι το πολύ ίση με 0.10, όταν η αληθινή τιμή του μέσου βάρους προϊόντος είναι $\mu = 1.98$;

Λύση Παραδείγματος 11.5

1. Κατ΄ ουσίαν, ζητείται η ισχύς του ελέγχου όταν $\mu = 1.98$. Από τη σχέση (11.2) για $\sigma = 0.05$, $n = 16$, $\mu = 1.98$ και $\mu_0 = 2$ έχουμε ότι

$$\gamma(1.98) = P(\text{απόρριψη } H_0|\mu = 1.98) = \Phi\left(-z_{0.05} - \frac{(1.98-2)\sqrt{16}}{0.05}\right)$$
$$= \Phi(-1.645 + 1.6) = \Phi(-0.045) = 0.4821.$$

   Χρησιμοποιώντας την R και εκτελώντας την εντολή `pnorm(-qnorm(0.95,0,1)-(1.98-2)*sqrt(16)/(0.05),0,1)`, λαμβάνουμε αποτέλεσμα 0.482112, το οποίο πρακτικά ισούται με την προαναφερθείσα τιμή. Η διαφοροποίηση των τιμών οφείλεται στο γεγονός ότι η R χρησιμοποιεί μια πιο ακριβή τιμή για $z_{0.05}$ από ότι προηγουμένως που βασιστήκαμε στους πίνακες της τυπικής κατανομής που δίνονται στο Παράρτημα Α΄.
   Παρατηρούμε ότι η ισχύς του παραπάνω ελέγχου είναι πολύ μικρή, γεγονός που δικαιολογείται απόλυτα λόγω τόσο της μικρής απόστασης του μ από το μ_0 όσο και του μικρού μεγέθους δείγματος ($n = 16$). Ένας τρόπος για να αυξηθεί η ισχύς του ελέγχου είναι να πάρουμε μεγαλύτερο σε μέγεθος δείγμα.
2. Θέλουμε να προσδιορίσουμε το ελάχιστο μέγεθος του δείγματος n, έτσι ώστε $P(\text{μη απόρριψη της } H_0|\mu = 1.98) \leq 0.1$ ή, ισοδύναμα, έτσι ώστε $\gamma(1.98) \geq 0.9$. Από τη σχέση (11.2) για $\sigma = 0.05$, $\alpha = 0.05$, $\mu = 1.98$ και $\mu_0 = 2$, έχουμε ότι:

$$0.9 \leq \Phi\left(-z_{0.05} - \frac{(1.98-2)\sqrt{n}}{0.05}\right).$$

   Λαμβάνοντας υπόψη ότι z_c είναι το σημείο της τυπικής κανονικής κατανομής για το οποίο ισχύει $P(Z \geq z_c) = c$ ή, ισοδύναμα, $\Phi(z_c) = 1 - c$ προκύπτει ότι:

$$\Phi(z_{0.1}) \leq \Phi\left(-z_{0.05} + 0.4\sqrt{n}\right).$$

   Επίσης, λαμβάνοντας υπόψη τη μονοτονία της ασκ, έχουμε άμεσα ότι:

$$z_{0.1} \leq -z_{0.05} + 0.4\sqrt{n},$$

   και, επομένως, ύστερα από λίγη άλγεβρα, έχουμε

$$n \geq \left(\frac{z_{0.1} + z_{0.05}}{0.4}\right)^2 = \left(\frac{1.645 + 1.282}{0.4}\right)^2 = 53.54.$$

   Άρα για να έχει ο έλεγχός μας πιθανότητα σφάλματος τύπου II το πολύ 0.1, θα πρέπει να πάρουμε μέγεθος δείγματος τουλάχιστον $n = 54$.

Άσκηση Αυτοαξιολόγησης 11.1

Βρείτε το κατάλληλο μέγεθος δείγματος και την κατάλληλη κρίσιμη περιοχή, ώστε ο έλεγχος υπόθεσης

$$H_0 : \mu = 150 \quad \text{κατά} \quad H_1 : \mu > 150$$

να έχει επίπεδο σημαντικότητας $\alpha = 0.01$ και $\gamma(\mu) = 0.92922$, όταν η πραγματική τιμή του $\mu = 156$. Υποθέστε ότι ο πληθυσμός είναι κανονικός με τυπική απόκλιση 12.

Παράδειγμα 11.6

Θεωρήστε έναν κανονικό πληθυσμό με τυπική απόκλιση 2. Ελέγξτε αν η μέση τιμή του πληθυσμού είναι 30 έναντι της εναλλακτικής ότι είναι διαφορετική από 30 σε επίπεδο σημαντικότητας 0.1, αν από ένα τυχαίο δείγμα μεγέθους 25 έχουμε βρει $\bar{x} = 30.94$. Υπολογίστε και το παρατηρούμενο επίπεδο σημαντικότητας του ελέγχου.

Λύση Παραδείγματος 11.6

Θέλουμε να ελέγξουμε αν η μέση τιμή κανονικού πληθυσμού είναι ίση με δοθείσα γνωστή τιμή (H_0 : $\mu = 30$) έναντι της εναλλακτικής ότι είναι διαφορετική ($H_1 : \mu \neq 30$). Καθώς η διασπορά του κανονικού πληθυσμού είναι γνωστή, η σσε δίνεται από τη σχέση (11.1) για $\mu_0 = 30, \sigma = 2$ και $n = 25$. Με αλγεβρικές πράξεις προκύπτει ότι η παρατηρούμενη τιμή z της στατιστικής συνάρτησης ελέγχου είναι:

$$z = \frac{30.94 - 30}{2/\sqrt{25}} = 2.35.$$

Η μηδενική υπόθεση απορρίπτεται σε επίπεδο σημαντικότητας 0.1, αν $|z| > z_{\alpha/2} = z_{0.05} = 1.645$. Επομένως, καθώς $2.35 > 1.645$, η μηδενική υπόθεση απορρίπτεται σε επίπεδο σημαντικότητας 0.1. Με βάση αυτά τα δεδομένα, σε επίπεδο σημαντικότητας 0.1, δεν μπορούμε να ισχυριστούμε ότι η μέση τιμή του πληθυσμού ισούται με 30.
Το παρατηρούμενο επίπεδο σημαντικότητας είναι:

$$\begin{aligned} p - \text{τιμή} &= 2 \cdot P(Z \geq |2.35|) = 2 \cdot (1 - \Phi(2.35)) \\ &= 2 \cdot (1 - 0.99061) = 0.01878. \end{aligned}$$

Στο ίδιο συμπέρασμα μπορούμε να καταλήξουμε εκτελώντας στην R τις ακόλουθες εντολές

```
1 library(PASWR)
2 zsum.test(mean.x=30.94, sigma.x =2, n.x =25,alternative = "two.sided", mu =
      30,
3 conf.level = 0.90)
```

οι οποίες επιστρέφουν τα εξής αποτελέσματα:

```
One-sample z-Test

data:  Summarized x
z = 2.35, p-value = 0.01877
alternative hypothesis: true mean is not equal to 30
90 percent confidence interval:
 30.28206 31.59794
sample estimates:
mean of x
    30.94
```

Από τα παραπάνω αποτελέσματα παρατηρούμε ότι η $p-$τιμή του ελέγχου ισούται με $0.01877 < 0.1$. Άρα δεν απορρίπτεται η μηδενική υπόθεση και, επομένως, πράγματι καταλήγουμε στο ίδιο συμπέρασμα με πριν.

Άσκηση Αυτοαξιολόγησης 11.2

Μία εταιρεία παράγει εδώ και πολλά χρόνια λυχνίες τηλεοράσεων και γνωρίζει ότι η μέση διάρκεια ζωής τους είναι 1200 ώρες με τυπική απόκλιση 300 ώρες. Το τμήμα ερευνών της, όμως, προτείνει μία νέα διαδικασία παραγωγής με τον ισχυρισμό ότι αυτή θα αυξήσει τη διάρκεια ζωής των λυχνιών. Για να ελέγξει τον ισχυρισμό αυτό, ο διευθυντής της εταιρείας ζητά να παραχθούν 100 λυχνίες με τη νέα διαδικασία και να μετρηθεί η διάρκεια ζωής τους. Έστω ότι η μέση διάρκεια ζωής των λυχνιών αυτών ήταν 1265 ώρες:

1. να ελεγχθεί ο ισχυρισμός του τμήματος ερευνών σε επίπεδο σημαντικότητας 0.05,
2. να υπολογίσετε το παρατηρούμενο επίπεδο σημαντικότητας του ελέγχου,
3. να υπολογίσετε την ισχύ του ελέγχου αν στην πραγματικότητα η μέση διάρκεια ζωής των λυχνιών είναι $\mu = 1240$ ώρες.

Δίνεται ότι ο χρόνος ζωής της λυχνίας μιας τηλεόρασης περιγράφεται ικανοποιητικά από την κανονική κατανομή και ότι ο νέος τρόπος παραγωγής δεν επιφέρει αλλαγή στην τυπική απόκλιση του χρόνου ζωής.

Σχέση μεταξύ ελέγχων υποθέσεων και διαστημάτων εμπιστοσύνης

Στην περίπτωση που θέλουμε να υλοποιήσουμε τον δίπλευρο έλεγχο

$$H_0 : \mu = \mu_0 \quad \text{κατά} \quad H_1 : \mu \neq \mu_0$$

είδαμε ότι η μηδενική υπόθεση δεν μπορεί να απορριφθεί όταν

$$-z_{\alpha/2} < z < z_{\alpha/2}$$

ή, ισοδύναμα, όταν

$$-z_{\alpha/2} < \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} < z_{\alpha/2}$$

ή

$$\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu_0 < \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

Παρατηρούμε, λοιπόν, ότι η μηδενική υπόθεση του δίπλευρου ελέγχου δεν απορρίπτεται σε επίπεδο σημαντικότητας α, όταν η τιμή μ_0 ανήκει στο $100(1 - \alpha)\%$ διάστημα εμπιστοσύνης της μέσης τιμής κανονικού πληθυσμού με γνωστή διασπορά. Επομένως, είναι φανερό ότι υπάρχει μία ισοδυναμία μεταξύ των δίπλευρων ελέγχων υποθέσεων και των αντίστοιχων διαστημάτων εμπιστοσύνης.

Παρατήρηση 11.2

Στο σημείο αυτό, θα πρέπει να επισημάνουμε ότι ένα διάστημα εμπιστοσύνης δίνει πολύ περισσότερες πληροφορίες από έναν έλεγχο υπόθεσης, αφού με το ίδιο διάστημα μπορούμε να ελέγξουμε ταυτόχρονα πολλές υποθέσεις για διαφορετικές τιμές μ_0 του μ, διατηρώντας το ίδιο επίπεδο σημαντικότητας.

Παρατήρηση 11.3

Στη περίπτωση που έχουμε μια μονόπλευρη εναλλακτική υπόθεση, τότε το ισοδύναμο δ.ε. του ελέγχου υποθέσεων είναι το αντίστοιχο μονόπλευρο δ.ε.

Περίπτωση ΙΙ. Κανονικός πληθυσμός με άγνωστη διασπορά.

Ο παραπάνω έλεγχος που στηρίζεται στη σχέση (11.1) δεν είναι καθόλου ρεαλιστικός, αφού το να γνωρίζουμε τη διασπορά ενός πληθυσμού, ενώ δεν γνωρίζουμε τη μέση τιμή του, είναι σχεδόν αδύνατον. Στις περισσότερες πραγματικές εφαρμογές η διασπορά του πληθυσμού είναι και αυτή άγνωστη. Σε αυτές τις περιπτώσεις αντικαθιστούμε την άγνωστη διασπορά με την εκτίμησή της, οπότε η σσε παίρνει τη μορφή

$T = \frac{\bar{X}-\mu_0}{S/\sqrt{n}}$ και ακολουθεί, υπό τη μηδενική υπόθεση, την κατανομή t με $n-1$ βαθμούς ελευθερίας (βλ. σχετικά στο Κεφάλαιο 9). Επομένως, έχουμε

$$T = \frac{\bar{X}-\mu_0}{S/\sqrt{n}} \overset{H_0}{\sim} t_{n-1}. \tag{11.5}$$

Κατά αντιστοιχία με τον προηγούμενο έλεγχο, αν t είναι η παρατηρηθείσα τιμή της τ.μ. T, η κπ του ελέγχου σε επίπεδο σημαντικότητας α θα είναι τέτοια, ώστε:

1. αν $H_1 : \mu \neq \mu_0$, να απορρίπτεται η μηδενική υπόθεση όταν $t < -t_{n-1,\alpha/2}$, ή $t > t_{n-1,\alpha/2}$,
2. αν $H_1 : \mu > \mu_0$, να απορρίπτεται η μηδενική υπόθεση όταν $t > t_{n-1,\alpha}$ και
3. αν $H_1 : \mu < \mu_0$, να απορρίπτεται η μηδενική υπόθεση όταν $t < -t_{n-1,\alpha}$.

Παρατήρηση 11.4

Ο παραπάνω δίπλευρος έλεγχος, καθώς και όλοι οι αντίστοιχοι έλεγχοι που στηρίζονται στην κατανομή t, μπορούν να υλοποιηθούν και μέσω της κατανομής F. Πιο συγκεκριμένα, από την Πρόταση 7.11 γνωρίζουμε ότι αν $W \sim t_n$, τότε $W^2 \sim F_{1,n}$, ενώ επίσης ισχύει ότι $t^2_{n-1,\alpha/2} = F_{1,n-1,\alpha}$. Επομένως, προκύπτει ότι απορρίπτεται η μηδενική υπόθεση, αν $t^2 > F_{1,n-1,\alpha}$. Ενδεχομένως μία τέτοια αναπαράσταση τώρα να μοιάζει περιττή, αλλά όπως θα δούμε στο Κεφάλαιο 13 αυτού του συγγράμματος, αν θέλουμε να ελέγξουμε την ισότητα των μέσων σε περισσότερους από δύο πληθυσμούς, οι έλεγχοι που θα συζητηθούν στηρίζονται στην κατανομή F.

Παράδειγμα 11.7: (συνέχεια Παραδείγματος 10.1)

Με βάση τα αριθμητικά δεδομένα του Παραδείγματος 10.1 και υποθέτοντας ότι οι πληθυσμοί είναι κανονικοί, μπορούμε να ισχυριστούμε σε επίπεδο σημαντικότητας 0.05 ότι το μέσο pH του διαλύματος Α ισούται με 6.4;

Λύση Παραδείγματος 11.7

Έστω X η τ.μ. που παριστάνει τον βαθμό οξύτητας του χημικού διαλύματος τύπου Α με $X \sim N(\mu_A, \sigma^2_A)$. Θέλουμε να ελέγξουμε τις υποθέσεις

$$H_0 : \mu_A = 6.4 \text{ κατά } H_1 : \mu_A \neq 6.4.$$

Έχουμε δειγματοληψία από κανονικό πληθυσμό με άγνωστη διασπορά. Επομένως, θα χρησιμοποιηθεί η σσε της σχέσης (11.5), η οποία παίρνει τιμή

$$t = \frac{\bar{x}-6.4}{s/\sqrt{n}} = \frac{6.392-6.4}{0.0887/\sqrt{5}} = -0.20165,$$

καθώς με αλγεβρικές πράξεις η δειγματική μέση τιμή και η δειγματική τυπική απόκλιση είναι $\bar{x} = \frac{6.33+6.28+6.5+6.4+6.45}{5} = 6.392$ και $s = \sqrt{\frac{\sum_{i=1}^{5}(x_i-6.392)^2}{5-1}} = 0.0887$, αντίστοιχα.

Γνωρίζουμε ότι η μηδενική υπόθεση απορρίπτεται σε επίπεδο σημαντικότητας α, αν $|t| \geq t_{n-1,\alpha/2}$. Είναι $t_{n-1,\alpha/2} = t_{4,0.025} = 2.776$ (Πίνακας Α΄.4, Παραρτήματος Α΄), οπότε, καθώς $-2.776 < -0.20165 < 2.776$, δεν μπορούμε να απορρίψουμε τη μηδενική υπόθεση.

Εναλλακτικά, θα μπορούσαμε να υλοποιήσουμε τον παραπάνω έλεγχο χρησιμοποιώντας τις ακόλουθες εντολές της R:

```
1 A <- c(6.33, 6.28, 6.5, 6.40, 6.45)
2 t.test(A, alternative = c("two.sided"), mu = 6.4)
```

από τις οποίες λαμβάνουμε τα παρακάτω αποτελέσματα:

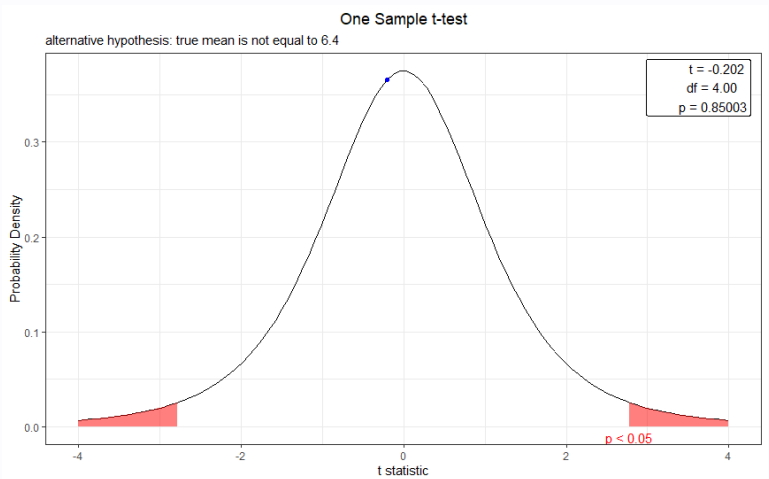
```
data:  A
t = -0.20165, df = 4, p-value = 0.85
alternative hypothesis: true mean is not equal to 6.4
95 percent confidence interval:
 6.281848 6.502152
sample estimates:
mean of x
    6.392
```

Στη δεύτερη γραμμή των αποτελεσμάτων της R δίνεται η τιμή της σσε, δηλαδή $t = -0.20165$, οι βαθμοί ελευθερίας $df = 4 (= n - 1)$ και η p-τιμή του ελέγχου $p - value = 0.85$. Αφού $p - value = 0.85 > 0.05 = \alpha$, δεν μπορούμε να απορρίψουμε τη μηδενική υπόθεση σε επίπεδο σημαντικότητας 0.05. Στην τρίτη γραμμή βλέπουμε ποια είναι η εναλλακτική υπόθεση, ενώ στις επόμενες δύο γραμμές μας δίνεται το 95% διάστημα εμπιστοσύνης για το μέσο pH του διαλύματος Α $(6.281848,\ 6.502152)$. Τέλος, μας δίνεται η δειγματική μέση τιμή $\bar{x}_A = 6.392$ του διαλύματος Α.

Επιπρόσθετα, χρησιμοποιώντας την R έχουμε τη δυνατότητα να απεικονίσουμε σε ένα γράφημα τη σππ της κατανομής της σσε υπό τη μηδενική υπόθεση με κόκκινο χρώμα την κπ του ελέγχου και με μπλε κουκίδα την τιμή της στατιστικής συνάρτησης. Κατά αυτόν τον τρόπο γίνεται αμέσως αντιληπτό ότι η μηδενική υπόθεση δεν μπορεί να απορριφθεί, αφού η τιμή της σσε δεν βρίσκεται στην κπ, ενώ ταυτόχρονα αναγράφονται οι τιμές της σσε, των βαθμών ελευθερίας του ελέγχου και η $p-$τιμή του. Όλα αυτά επιτυγχάνονται με τις ακόλουθες εντολές

```
install.packages("webr")
library(webr)
testa <-t.test(A,mu=6.4)
plot(testa)
```

και προκύπτει το παρακάτω σχήμα.



Παράδειγμα 11.8: (συνέχεια Παραδείγματος 10.1)

Με βάση τα αριθμητικά δεδομένα του Παραδείγματος 10.1 και υποθέτοντας ότι οι πληθυσμοί είναι κανονικοί μπορούμε να ισχυριστούμε σε επίπεδο σημαντικότητας 0.05 ότι το μέσο pH του διαλύματος Β ξεπερνάει το 6.4;

Λύση Παραδείγματος 11.8

Έστω Y η τ.μ. που παριστάνει τον βαθμό οξύτητας του χημικού διαλύματος Β με $Y \sim N(\mu_B, \sigma_B^2)$. Θέλουμε να ελέγξουμε τις υποθέσεις

$$H_0 : \mu_B = 6.4 \text{ κατά } H_1 : \mu_B > 6.4.$$

Έχουμε δειγματοληψία από κανονικό πληθυσμό με άγνωστη διασπορά. Επομένως, θα χρησιμοποιηθεί η σσε της σχέσης (11.5), η οποία παίρνει τιμή

$$t = \frac{6.5 - 6.4}{0.07348469/\sqrt{7}} = 3.6004125,$$

καθώς με αλγεβρικές πράξεις η δειγματική μέση τιμή και η δειγματική τυπική απόκλιση είναι $\bar{y} = 6.5$ και $s = 0.07348469$, αντίστοιχα.

Γνωρίζουμε ότι η μηδενική υπόθεση απορρίπτεται σε επίπεδο σημαντικότητας α, αν $t > t_{n-1,\alpha}$. Είναι $t_{n-1,\alpha} = t_{6,0.05} = 1.943$ (Πίνακας Α΄.4, Παραρτήματος Α΄), οπότε δεν μπορούμε να απορρίψουμε τη μηδενική υπόθεση.

Εναλλακτικά, θα μπορούσαμε να υλοποιήσουμε τον παραπάνω έλεγχο χρησιμοποιώντας τις ακόλουθες εντολές της R:

```
B <- c(6.51, 6.55, 6.43, 6.51, 6.62, 6.40, 6.48)
t.test(B, alternative = c("greater"), mu = 6.4)
```

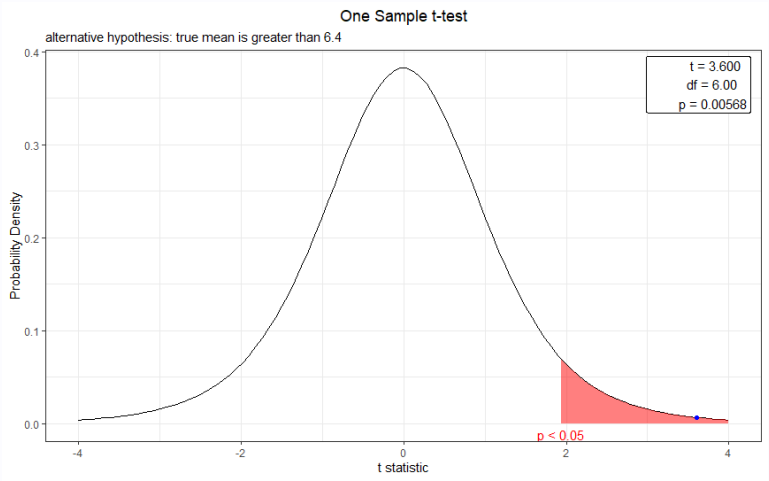
από τις οποίες λαμβάνουμε τα παρακάτω αποτελέσματα:

```
data:  B
t = 3.6004, df = 6, p-value = 0.00568
alternative hypothesis: true mean is greater than 6.4
95 percent confidence interval:
 6.446029      Inf
sample estimates:
mean of x
      6.5
```

Σε αυτήν την περίπτωση η $p - value = 0.00568 < 0.05$, άρα απορρίπτουμε σε επίπεδο σημαντικότητας 0.05 τη μηδενική υπόθεση. Επομένως, με βάση αυτά τα δεδομένα, μπορούμε να ισχυριστούμε ότι το μέσο pH του διαλύματος Β είναι μεγαλύτερο από 6.4.

Στο παρακάτω σχήμα απεικονίζεται η σππ της κατανομής της σσε υπό τη μηδενική υπόθεση, με κόκκινο χρώμα η κπ του ελέγχου και με μπλε κουκκίδα η τιμή της σσε, ενώ ταυτόχρονα αναγράφονται οι τιμές της σσε, των βαθμών ελευθερίας και η $p-$τιμή του ελέγχου. Το σχήμα που εμφανίζεται στη συνέχεια δημιουργήθηκε με τις ακόλουθες εντολές:

```
install.packages("webr")
library(webr)
testb <-t.test(B,mu=6.4)
plot(testb)
```



Παρατήρηση 11.5

Γενικότερα η εντολή `t.test` έχει την εξής μορφή:

```
t.test(x, y = NULL,
       alternative = c("two.sided", "less", "greater"),
       mu = 0, paired = FALSE, var.equal = FALSE,
       conf.level = 0.95)
```

- `x` ένα διάνυσμα παρατηρήσεων.
- `y` ένα δεύτερο διάνυσμα παρατηρήσεων (προαιρετικό). Δηλώνεται όταν έχουμε δύο δείγματα.
- `alternative` ένα όρισμα αλφαριθμητικού τύπου που καθορίζει ποια είναι η μορφή της εναλλακτικής υπόθεσης και παίρνει μία εκ των τριών τιμών, δηλαδή `"two.sided"` `(default)`, `"greater"` `or` `"less"`. Μπορούμε να χρησιμοποιούμε μόνο το πρώτο γράμμα των παραπάνω λέξεων.
- `mu` η τιμή για την οποία γίνεται ο έλεγχος.
- `paired` μια λογική μεταβλητή που παίρνει την τιμή `TRUE` αν θέλουμε να υλοποιήσουμε έλεγχο με ζευγαρωτές παρατηρήσεις.
- `var.equal` μια λογική μεταβλητή που χρησιμοποιείται στην περίπτωση που έχουμε έλεγχο για τη διαφορά των μέσων τιμών δύο ανεξάρτητων πληθυσμών και αν πάρει την τιμή `TRUE` υποθέτουμε ότι οι διασπορές των δύο ανεξάρτητων πληθυσμών είναι άγνωστες αλλά ίσες. Σε αυτήν την περίπτωση χρησιμοποιείται η από κοινού εκτίμηση της διασποράς τους (pooled variance).
- `conf.level` επίπεδο εμπιστοσύνης του διαστήματος εμπιστοσύνης.

Άσκηση Αυτοαξιολόγησης 11.3

Οι προδιαγραφές για την παραγωγή ενός κράματος μετάλλων ορίζουν ότι το 18.5% πρέπει να είναι χαλκός. Δέκα αναλύσεις του κράματος έδωσαν μέσο ποσοστό χαλκού 18.7% με τυπική απόκλιση 0.26%. Μπορούμε να συμπεράνουμε σε επίπεδο σημαντικότητας 0.05 ότι το κράμα πληροί τις προδιαγραφές; Δίνεται ότι το ποσοστό χαλκού σε ένα κράμα μετάλλου ακολουθεί κανονική κατανομή.

Άσκηση Αυτοαξιολόγησης 11.4

Με μία νέα μέθοδο προσδιορισμού του σημείου τήξης μετάλλων προέκυψαν οι παρακάτω μετρήσεις για το μαγγάνιο:

$$1267, 1262, 1267, 1263, 1258, 1263, 1268.$$

Να εξεταστεί αν η νέα μέθοδος σφάλλει σε επίπεδο σημαντικότητας 0.05, δεδομένου ότι το σημείο τήξης του μαγγανίου είναι 1260^oC. Δίνεται ότι το σημείο τήξης ενός μετάλλου περιγράφεται ικανοποιητικά από την κανονική κατανομή.

Περίπτωση III. Μη κανονικός πληθυσμός και μεγάλο δείγμα.

Στην περίπτωση που το διαθέσιμο δείγμα είναι μεγάλο ($n > 30$), η υπόθεση της κανονικότητας του πληθυσμού δεν χρειάζεται. Σε αυτήν την περίπτωση, από το Κεντρικό Οριακό Θεώρημα η σσε $T = \frac{\bar{X}-\mu_0}{S/\sqrt{n}}$, υπό την H_0, προσεγγίζεται από την τυπική κανονική κατανομή. Δηλαδή, υπό τη μηδενική υπόθεση ισχύει ότι:

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \xrightarrow{d} N(0,1)$$

Επιπλέον, αν t είναι η τιμή της T, έχουμε σε (προσεγγιστικό) επίπεδο σημαντικότητας α ότι:

1. αν $H_1 : \mu \neq \mu_0$, τότε απορρίπτεται η μηδενική υπόθεση αν $t < -z_{\alpha/2}$, ή $t > z_{\alpha/2}$,
2. αν $H_1 : \mu > \mu_0$, τότε απορρίπτεται η μηδενική υπόθεση αν $t > z_\alpha$ και
3. αν $H_1 : \mu < \mu_0$, τότε απορρίπτεται η μηδενική υπόθεση αν $t < -z_\alpha$.

Παρατήρηση 11.6

Ένα εύλογο ερώτημα που ίσως έχει δημιουργηθεί είναι πώς διεξάγεται ο έλεγχος υπόθεσης για τη μέση τιμή ενός πληθυσμού όταν αυτός δεν είναι κανονικός και το μέγεθος του δείγματος είναι μικρότερο από 30. Σε μια τέτοια περίπτωση, μία λύση είναι να προσπαθήσουμε να βρούμε έναν μετασχηματισμό των δεδομένων έτσι ώστε τα μετασχηματισμένα δεδομένα να ακολουθούν κανονική κατανομή. Αν κάτι τέτοιο είναι εφικτό, για παράδειγμα χρησιμοποιώντας τον μετασχηματισμό του λογαρίθμου, τότε θα γίνει έλεγχος για τον πληθυσμιακό μετασχηματισμένο μέσο. Στην περίπτωση που η εύρεση μετασχηματισμού που να διορθώνει το πρόβλημα της μη κανονικότητας δεν είναι εφικτή, τότε μπορούμε να χρησιμοποιήσουμε μία ομάδα ελέγχων που ονομάζονται **μη παραμετρικοί**, γιατί δεν υποθέτουν κάποια συγκεκριμένη κατανομή για τον πληθυσμό (Bethea *et al.*, 1995; Sprent and Smeeton, 2016).

11.4 Έλεγχος υπόθεσης για τη διαφορά δύο μέσων τιμών με ανεξάρτητα δείγματα

Έστω $X_1, \ldots, X_n$ είναι ένα τυχαίο δείγμα από έναν πληθυσμό με μέση τιμή μ_1 και διασπορά σ_1^2, ενώ $Y_1, \ldots, Y_m$ είναι ένα τυχαίο δείγμα από έναν πληθυσμό με μέση τιμή μ_2 και διασπορά σ_2^2. Επιπλέον, υποθέτουμε ότι τα δύο δείγματα είναι ανεξάρτητα. Έστω ότι μας ενδιαφέρει να ελέγξουμε τις υποθέσεις:

$$\mathrm{H}_0 : \mu_1 - \mu_2 = \delta \text{ κατά } \mathrm{H}_1 : \mu_1 - \mu_2 \neq \delta \text{ ή } \mathrm{H}_1 : \mu_1 - \mu_2 > \delta \text{ ή } \mathrm{H}_1 : \mu_1 - \mu_2 \neq \delta,$$

όπου δ γνωστός αριθμός. Στο παραπάνω πλαίσιο, θα διακρίνουμε τέσσερις περιπτώσεις.

Περίπτωση Ι. Κανονικοί πληθυσμοί με γνωστές διασπορές.

Σε αυτήν την περίπτωση ισχύει ότι (βλ. Ενότητα 9.3.1):

$$Z = \frac{\bar{X} - \bar{Y} - \delta}{\sqrt{\sigma_1^2/n + \sigma_2^2/m}} \overset{H_0}{\sim} \mathrm{N}(0,1), \tag{11.6}$$

η οποία αποτελεί τη στατιστική συνάρτηση του παραπάνω ελέγχου.

Η κπ του ελέγχου, σε επίπεδο σημαντικότητας α, ανάλογα με την εναλλακτική υπόθεση, είναι τέτοια, ώστε:

1. αν $H_1 : \mu_1 - \mu_2 \neq \delta$, να απορρίπτεται η μηδενική υπόθεση αν $z < -z_{\alpha/2}$ ή $z > z_{\alpha/2}$,
2. αν $H_1 : \mu_1 - \mu_2 > \delta$, να απορρίπτεται η μηδενική υπόθεση αν $z > z_{\alpha}$,
3. αν $H_1 : \mu_1 - \mu_2 < \delta$, να απορρίπτεται η μηδενική υπόθεση αν $z < -z_{\alpha}$,

όπου z η τιμή της σσε, όταν είναι διαθέσιμες οι παρατηρήσεις $x_1, \ldots, x_n$ και $y_1, \ldots, y_m$.

Παράδειγμα 11.9: (συνέχεια Παραδείγματος 10.1)

Χρησιμοποιώντας τα δεδομένα και τις υποθέσεις του Παραδείγματος 10.1, μπορούμε να ισχυριστούμε σε επίπεδο σημαντικότητας 0.05 ότι το μέσο pH του διαλύματος Β είναι μεγαλύτερο από το μέσο pH του διαλύματος Α;

Λύση Παραδείγματος 11.9

Έστω X και Y οι τ.μ. που παριστάνουν τον βαθμό οξύτητας του χημικού διαλύματος Α και Β, αντίστοιχα. Από την εκφώνηση έχουμε ότι $X \sim N(\mu_1, 0.008)$ και $Y \sim N(\mu_2, 0.005)$. Θέλουμε να ελέγξουμε τις υποθέσεις

$$H_0 : \mu_1 - \mu_2 = 0 \text{ κατά της } H_1 : \mu_1 - \mu_2 < 0.$$

Έχουμε ανεξάρτητα δείγματα από κανονικούς πληθυσμούς με γνωστές διασπορές. Επομένως, θα χρησιμοποιηθεί η σσε της σχέσης (11.6), η οποία παίρνει τιμή

$$z = \frac{6.392 - 6.5}{\sqrt{\frac{0.008}{5} + \frac{0.005}{7}}} = -\frac{0.108}{0.04811} = -2.245.$$

Γνωρίζουμε ότι η μηδενική υπόθεση απορρίπτεται σε επίπεδο σημαντικότητας 0.05, αν $z < -z_{\alpha} = -z_{0.05} = -1.645$ (Πίνακας Α΄.3, Παραρτήματος Α΄). Επομένως, απορρίπτεται η μηδενική υπόθεση και με βάση αυτά τα δεδομένα μπορούμε να ισχυριστούμε ότι το μέσο pH του διαλύματος Β είναι μεγαλύτερο από αυτό του διαλύματος Α σε επίπεδο σημαντικότητας 0.05.
Εναλλακτικά, για την υλοποίηση του παραπάνω ελέγχου μπορούμε να χρησιμοποιήσουμε τις ακόλουθες εντολές της R:

```
1 library(BSDA)
2 A <- c(6.33, 6.28, 6.5, 6.40, 6.45)
3 B <- c(6.51, 6.55, 6.43, 6.51, 6.62, 6.40, 6.48)
4 z.test(A, B, alternative='less', mu=0, sigma.x=sqrt(0.008), sigma.y=sqrt
    (0.005),conf.level=.95)
```

οι οποίες επιστρέφουν τα εξής αποτελέσματα:

```
    Two-sample z-Test

data:  A and B
z = -2.245, p-value = 0.01238
alternative hypothesis: true difference in means is less than 0
```

```
95 percent confidence interval:
          NA -0.02887099
sample estimates:
mean of x mean of y
    6.392     6.500
```

Η τιμή της σσε ισούται με -2.245 και το παρατηρούμενο επίπεδο σημαντικότητας είναι $p-value = 0.01238 < 0.05$, άρα απορρίπτουμε τη μηδενική υπόθεση σε επίπεδο σημαντικότητας 0.05. Επομένως, με βάση αυτά τα δεδομένα μπορούμε να ισχυριστούμε, σε επίπεδο σημαντικότητας 0.05, ότι το μέσο pH του διαλύματος Β είναι στατιστικά σημαντικά μεγαλύτερο από το μέσο pH του διαλύματος Α.

Περίπτωση ΙΙ. Κανονικοί πληθυσμοί με άγνωστες, αλλά ίσες διασπορές.

Στην περίπτωση όπου οι διασπορές είναι άγνωστες θα πρέπει να προσδιοριστεί αν οι άγνωστες διασπορές μπορούν να θεωρηθούν ίσες μεταξύ τους ή όχι (ο έλεγχος ισότητας διασπορών θα παρουσιαστεί σε επόμενη ενότητα).

Αν οι άγνωστες διασπορές των δύο πληθυσμών μπορούν να θεωρηθούν ίσες, τότε τις εκτιμούμε με την κοινή δειγματική διασπορά:

$$S_p^2 = \frac{(n-1)S_1^2 + (m-1)S_2^2}{n+m-2},$$

όπου S_1^2 και S_2^2 είναι οι δειγματικές διακυμάνσεις που προκύπτουν από τα ανεξάρτητα τ.δ. $X_1, \dots, X_n$ και $Y_1, \dots, Y_m$, αντίστοιχα.

Τότε, (βλ. Ενότητα 9.3.1):

$$T = \frac{\bar{X} - \bar{Y} - \delta}{S_p\sqrt{1/n + 1/m}} \overset{H_0}{\sim} t_{n+m-2} \tag{11.7}$$

και η κπ του ελέγχου, σε επίπεδο σημαντικότητας α, ανάλογα με την εναλλακτική υπόθεση, είναι τέτοια, ώστε:

1. αν $H_1 : \mu_1 - \mu_2 \neq \delta$, να απορρίπτεται η μηδενική υπόθεση αν $t < -t_{n+m-2,\alpha/2}$ ή $t > t_{n+m-2,\alpha/2}$,
2. αν $H_1 : \mu_1 - \mu_2 > \delta$, να απορρίπτεται η μηδενική υπόθεση αν $t > t_{n+m-2,\alpha}$,
3. αν $H_1 : \mu_1 - \mu_2 < \delta$, να απορρίπτεται η μηδενική υπόθεση αν $t < -t_{n+m-2,\alpha}$,

όπου με t συμβολίζουμε την τιμή της σσε T.

Παράδειγμα 11.10: (συνέχεια Παραδείγματος 10.1)

Χρησιμοποιώντας τα δεδομένα του Παραδείγματος 10.1, να ελέγξετε σε επίπεδο σημαντικότητας 0.05 τον ισχυρισμό ότι το μέσο pH του διαλύματος Β είναι μεγαλύτερο από το μέσο pH του διαλύματος Α, υποθέτοντας ότι προέρχονται από κανονικούς πληθυσμούς με άγνωστες, αλλά ίσες διακυμάνσεις.

Λύση Παραδείγματος 11.10

Έστω X και Y οι τ.μ. που παριστάνουν την οξύτητα των δύο τύπων χημικών διαλυμάτων, αντίστοιχα. Από την εκφώνηση έχουμε ότι $X \sim N(\mu_1, \sigma_1^2)$ και $Y \sim N(\mu_2, \sigma_2^2)$ με $\sigma_1^2 = \sigma_2^2 = \sigma^2$. Θέλουμε να ελέγξουμε τις υποθέσεις

$$H_0 : \mu_1 - \mu_2 = 0 \text{ κατά της } H_1 : \mu_1 - \mu_2 < 0.$$

Καθώς οι πληθυσμοί είναι κανονικοί και οι πληθυσμιακές διασπορές είναι άγνωστες, αλλά ίσες εξ υποθέσεως, θα χρησιμοποιήσουμε τη σσε που δίνεται στη σχέση (11.7) με $\bar{x} = 6.392$, $\bar{y} = 6.5$, $s_1^2 = 0.00787$, $s_2^2 = 0.0054$, $n = 5$, $m = 7$ και

$$s_p^2 = \frac{(n-1)s_1^2 + (m-1)s_2^2}{n+m-2} = \frac{4 \cdot 0.00787 + 6 \cdot 0.0054}{10} = 0.006388.$$

Προκύπτει τότε ότι η τιμή της σσε είναι

$$t = \frac{6.392 - 6.5}{\sqrt{0.006388}\sqrt{1/5 + 1/7}} = -2.307728.$$

Η κπ είναι $t < -t_{n+m-2,\alpha} = -t_{10,0.05} = -1.812$. Επομένως, απορρίπτεται η μηδενική υπόθεση σε επίπεδο σημαντικότητας 0.05.
Εναλλακτικά, για την υλοποίηση του παραπάνω ελέγχου μπορούμε να χρησιμοποιήσουμε τις ακόλουθες εντολές της R:

```
A <- c(6.33, 6.28, 6.5, 6.40, 6.45)
B <- c(6.51, 6.55, 6.43, 6.51, 6.62, 6.40, 6.48)
t.test(A, B, alternative = c("less"), mu = 0,  var.equal = TRUE)
```

οι οποίες επιστρέφουν τα εξής αποτελέσματα:

```
  Two Sample t-test

data:  A and B
t = -2.3077, df = 10, p-value = 0.02184
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
        -Inf -0.02317815
sample estimates:
mean of x mean of y
    6.392     6.500
```

Η τιμή της σσε ισούται με -2.3077 (πρακτικά ισούται με την τιμή που υπολογίσαμε νωρίτερα) και το παρατηρούμενο επίπεδο σημαντικότητας είναι $p - value = 0.02184 < 0.05$, άρα απορρίπτουμε και πάλι τη μηδενική υπόθεση σε επίπεδο σημαντικότητας 0.05. Επομένως, με βάση αυτά τα δεδομένα μπορούμε να ισχυριστούμε, σε επίπεδο σημαντικότητας 0.05, ότι το μέσο pH του διαλύματος Β είναι στατιστικά σημαντικά μεγαλύτερο από το μέσο pH του διαλύματος Α.

Άσκηση Αυτοαξιολόγησης 11.5

Σε ένα πείραμα σύγκρισης της θερμικής ικανότητας του γαιάνθρακα που εξάγεται από δύο ορυχεία έγιναν πέντε μετρήσεις στο καθένα ορυχείο. Οι μετρήσεις σε εκατομμύρια θερμίδες ανά τόνο δίνονται στον παρακάτω πίνακα.

Ορυχείο 1	8240	8110	8330	8050	8320
Ορυχείο 2	7980	7920	7930	8170	7950

Με βάση αυτά τα δεδομένα, μπορούμε να ισχυριστούμε σε επίπεδο σημαντικότητας 0.05 ότι η μέση θερμική ικανότητα του γαιάνθρακα που εξάγεται από τα δύο ορυχεία δεν διαφέρει στατιστικά σημαντικά; Δίνεται ότι οι ποσότητες γαιάνθρακα που εξάγονται από τα δύο ορυχεία ακολουθούν κανονικές κατανομές με ίσες διασπορές.

Άσκηση Αυτοαξιολόγησης 11.6

Τα παρακάτω δεδομένα αφορούν τα φορτία θραύσης (σε tn/cm^2) δύο διαφορετικών τύπων σύνθετων νημάτων.

Τύπος Ι	1.2	0.3	0.8	0.5	0.4	0.9	1.0
Τύπος ΙΙ	1.4	1.5	1.1	1.0	0.8	1.7	0.9

Με βάση αυτά τα δεδομένα, μπορούμε να ισχυριστούμε, σε επίπεδο σημαντικότητας 0.05 ότι οι δύο αυτοί τύποι σύνθετων νημάτων έχουν τα ίδια μέσα φορτία θραύσης; Δίνεται ότι οι πληθυσμοί είναι κανονικοί με ίσες διασπορές.

Περίπτωση ΙΙΙ. Κανονικοί πληθυσμοί με άγνωστες, αλλά άνισες διασπορές.

Αν οι άγνωστες διασπορές των δύο κανονικών πληθυσμών δεν μπορούν να θεωρηθούν ίσες, τότε ισχύει ότι (βλ. Ενότητα 9.3.1):

$$T = \frac{\bar{X} - \bar{Y} - \delta}{\sqrt{S_1^2/n + S_2^2/m}} \overset{H_0}{\sim} t_v, \tag{11.8}$$

με

$$v = \frac{\left(\frac{s_1^2}{n} + \frac{s_2^2}{m}\right)^2}{\frac{\left(\frac{s_1^2}{n}\right)^2}{n-1} + \frac{\left(\frac{s_2^2}{m}\right)^2}{m-1}}. \tag{11.9}$$

Η κπ του ελέγχου, σε επίπεδο σημαντικότητας α, ανάλογα με την εναλλακτική υπόθεση, είναι τέτοια, ώστε

1. αν $H_1 : \mu_1 - \mu_2 \neq \delta$, να απορρίπτεται η μηδενική υπόθεση αν $t < -t_{v,\alpha/2}$ ή $t > t_{v,\alpha/2}$,
2. αν $H_1 : \mu_1 - \mu_2 > \delta$, να απορρίπτεται η μηδενική υπόθεση αν $t > t_{v,a}$,
3. αν $H_1 : \mu_1 - \mu_2 < \delta$, να απορρίπτεται η μηδενική υπόθεση αν $t < -t_{v,a}$,

όπου με t συμβολίζουμε την τιμή της σσε T.

Παράδειγμα 11.11: (συνέχεια Παραδείγματος 10.6)

Με βάση τα δεδομένα και τις υποθέσεις που παρουσιάζονται στο Παράδειγμα 10.6, μπορούμε να ισχυριστούμε σε επίπεδο σημαντικότητας 0.05 ότι δεν υπάρχει διαφορά στο πραγματικό μέσο περιεχόμενο ορθοφωσφόρου σε αυτούς τους δύο σταθμούς;

Λύση Παραδείγματος 11.11

Έστω X και Y οι τ.μ. που παριστάνουν την ποσότητα ορθοφωσφόρου στην περιοχή του σταθμού Α και Β, αντίστοιχα. Από την εκφώνηση έχουμε ότι $X \sim N(\mu_1, \sigma_1^2)$ και $Y \sim N(\mu_2, \sigma_2^2)$ με $\sigma_1^2 \neq \sigma_2^2$. Θέλουμε να ελέγξουμε την

$$H_0 : \mu_1 - \mu_2 = 0 \text{ κατά της } H_1 : \mu_1 - \mu_2 \neq 0.$$

Καθώς οι πληθυσμοί είναι κανονικοί και οι πληθυσμιακές διασπορές είναι άγνωστες και άνισες, θα χρησιμοποιήσουμε τη σσε που δίνεται στη σχέση (11.8). Είναι $\bar{x} = 3.84$, $\bar{y} = 1.49$, $s_1^2 = 3.07^2 = 9.4249$, $s_2^2 = 0.8^2 = 0.64$, $n = 15$, $m = 12$ και

$$v = \frac{\left(s_1^2/n + s_2^2/m\right)^2}{\frac{(s_1^2/n)^2}{n-1} + \frac{(s_2^2/m)^2}{m-1}} = \frac{(9.4249/15 + 0.64/12)^2}{\frac{(9.4249/15)^2}{14} + \frac{(0.64/12)^2}{11}} = \frac{0.4646604}{0.0281996 + 0.0002585859} = 16.32783.$$

Για να χρησιμοποιήσουμε τον Πίνακα Α΄.4 της κατανομής t, κρατάμε μόνο το ακέραιο μέρος του v, οπότε έχουμε $v = 16$ βαθμούς ελευθερίας και $t_{v,\alpha/2} = t_{16,0.025} = 2.120$. Επομένως, η παρατηρούμενη τιμή της

σσε είναι

$$t = \frac{3.84 - 1.49}{\sqrt{\frac{9.4249}{15} + \frac{0.64}{12}}} = 2.846322$$

Γνωρίζουμε ότι η μηδενική υπόθεση απορρίπτεται σε επίπεδο σημαντικότητας 0.05, αν $|t| > t_{v,\alpha/2} = t_{16,0.025} = 2.120$ (Πίνακας Α΄.4, Παραρτήματος Α΄). Επομένως, καθώς $2.846322 > 2.120$, σε επίπεδο σημαντικότητας 0.05 απορρίπτεται η μηδενική υπόθεση, δηλαδή υπάρχει στατιστικά σημαντική διαφοροποίηση στα μέσα επίπεδα συγκέντρωσης ορθοφωσφόρου στους δύο σταθμούς.
Εναλλακτικά, για την υλοποίηση του παραπάνω ελέγχου μπορούμε να χρησιμοποιήσουμε τις ακόλουθες εντολές της R:

```
1 library(BSDA)
2 tsum.test(mean.x=3.84,s.x = 3.07,n.x =15,mean.y = 1.49,s.y = 0.8,n.y = 12,
    alternative = "two.sided",mu = 0,var.equal = FALSE,conf.level = 0.95)
```

οι οποίες επιστρέφουν τα εξής αποτελέσματα:

```
  Welch Modified Two-Sample t-Test

data:  Summarized x and y
t = 2.8463, df = 16.328, p-value = 0.01149
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.6026006 4.0973994
sample estimates:
mean of x mean of y
     3.84      1.49
```

Παρατηρούμε ότι η $p-$τιμή του ελέγχου ισούται με $0.01149 < 0.05$. Άρα απορρίπτεται η μηδενική υπόθεση και καταλήγουμε στο ίδιο συμπέρασμα με πριν.

Περίπτωση IV. Μη κανονικοί πληθυσμοί και μεγάλα δείγματα.

Στην περίπτωση όπου ένας από τους δύο πληθυσμούς δεν είναι κανονικός, αλλά το διαθέσιμο δείγμα από αυτόν είναι μεγάλου μεγέθους (συνήθως μεγαλύτερο από 30), υπό την υπόθεση ότι οι πληθυσμιακές διασπορές είναι πεπερασμένες, τότε η στατιστική συνάρτηση ελέγχου είναι η

$$T = \frac{\bar{X} - \bar{Y} - \delta}{\sqrt{S_1^2/n + S_2^2/m}}.$$

Με εφαρμογή του Κεντρικού Οριακού Θεωρήματος, προκύπτει ότι η T ακολουθεί, υπό τη μηδενική υπόθεση, προσεγγιστικά την τυπική κανονική κατανομή $N(0,1)$. Επιπρόσθετα, η κπ του ελέγχου, σε επίπεδο σημαντικότητας α, ανάλογα με την εναλλακτική υπόθεση, είναι τέτοια, ώστε:

1. αν $H_1 : \mu_1 - \mu_2 \neq \delta$, να απορρίπτεται η μηδενική υπόθεση αν $t < -z_{\alpha/2}$ ή $t > z_{a/2}$,
2. αν $H_1 : \mu_1 - \mu_2 > \delta$, να απορρίπτεται η μηδενική υπόθεση αν $t > z_\alpha$,
3. αν $H_1 : \mu_1 - \mu_2 < \delta$, να απορρίπτεται η μηδενική υπόθεση αν $t < -z_\alpha$,

όπου με t συμβολίζεται η τιμή της σσε T.

11.5 Έλεγχος υπόθεσης για τη διαφορά δύο μέσων τιμών με εξαρτημένα δείγματα

Στην περίπτωση όπου υπάρχει εξάρτηση μεταξύ των παρατηρήσεων δύο πληθυσμών, δηλαδή έχουμε ζευγαρωτές παρατηρήσεις (x_i, y_i), $i = 1, \dots, n$, προκειμένου να λάβουμε υπόψη μας αυτήν την εξάρτηση, αντί να ελέγξουμε αν η διαφορά των μέσων ισούται με δ, ελέγχουμε αν η μέση τιμή της διαφοράς τους ισούται με αυτήν την τιμή χρησιμοποιώντας το δείγμα των διαφορών.

Παράδειγμα τέτοιας εφαρμογής αποτελεί ενδεικτικά η μελέτη της επίδρασης μίας παρέμβασης/θεραπείας σε μια ομάδα ατόμων. Σε αυτές τις περιπτώσεις μελετάμε το χαρακτηριστικό που μας ενδιαφέρει πριν και μετά την παρέμβαση, οπότε για κάθε πειραματική μονάδα έχουμε ένα ζεύγος παρατηρήσεων (x_i, y_i), $i = 1, \dots, n$. Ένα δεύτερο παράδειγμα αποτελούν οι περιπτώσεις όπου μας ενδιαφέρει να ελέγξουμε την επίδραση δύο θεραπειών. Σε αυτές τις περιπτώσεις, προκειμένου να μειώσουμε τη μεταβλητότητα που συνήθως υπάρχει εφαρμόζοντας τις θεραπείες σε ανεξάρτητα δείγματα λόγω άλλων εξωτερικών χαρακτηριστικών τους, ομαδοποιούμε τις παρατηρήσεις σε ζεύγη με βάση κατάλληλο εξωτερικό παράγοντα. Κατά αυτόν τον τρόπο οι συνήθεις έλεγχοι υποθέσεων γίνονται πιο αποτελεσματικοί στον εντοπισμό των διαφορών που οφείλονται στις θεραπείες και όχι σε άλλους εξωτερικούς παράγοντες.

Στα προαναφερθέντα παραδείγματα οι υποθέσεις που θέλουμε να ελέγξουμε είναι οι εξής

$$H_0 : \mu_\delta = \delta_0 \quad \text{κατά} \quad H_1 : \mu_\delta \neq \delta_0 \quad \text{ή} \quad \mu_\delta < \delta_0 \quad \text{ή} \quad \mu_\delta > \delta_0,$$

όπου μ_δ η πληθυσμιακή μέση τιμή της διαφοράς $\Delta = X - Y$ και δ_0 η τιμή της μέσης διαφοράς που θέλουμε να ελέγξουμε. Επομένως, ο έλεγχος για τη διαφορά των μέσων τιμών $\mu_1 - \mu_2$ στη βάση δύο εξαρτημένων δειγμάτων, ανάγεται στην πραγματικότητα σε έλεγχο για τη μέση τιμή ενός πληθυσμού με άγνωστη διασπορά και παραπέμπουμε σε όσα αναφέρθηκαν στην Ενότητα 11.3. Για χάρη πληρότητας και για να μην υπάρχει σύγχυση ως προς τον συμβολισμό σημειώνεται ότι η σσε στην περίπτωση αυτή συμβολίζεται ως

$$T = \frac{\bar{D} - \delta}{S_D/\sqrt{n}}$$

όπου $\bar{D}$ η τ.μ. της δειγματικής μέσης τιμής των διαφορών, S_D η αντίστοιχη τυπική απόκλιση και n το πλήθος των ζευγαρωτών παρατηρήσεων.

Παράδειγμα 11.12: (συνέχεια Παραδείγματος 10.7)

Με βάση τα δεδομένα του Παραδείγματος 10.7, μπορούμε να ισχυριστούμε, σε επίπεδο σημαντικότητας 0.05, ότι οι δύο μέθοδοι μέτρησης είναι αντίστοιχοι; Υπόδειξη: να γίνουν κατάλληλες υποθέσεις.

Λύση Παραδείγματος 11.12

Έστω X και Y οι τυχαίες μεταβλητές που παριστάνουν τη μέτρηση του βάρους με τη μέθοδο Α και Β, αντίστοιχα με μέση τιμή μ_1 και μ_2, αντίστοιχα. Έχουμε διαθέσιμα τα δείγματα $X_1,\dots,X_n$ και $Y_1,\dots,Y_n$ με $n = 9$, τα οποία είναι εξαρτημένα, καθώς πρόκειται για μετρήσεις στους ίδιους οργανισμούς. Προκειμένου να ληφθεί υπόψη η εξάρτηση, οι υποθέσεις που πρέπει να ελέγξουμε είναι οι εξής:

$$H_0 : \mu_\delta = 0 \text{ κατά } H_1 : \mu_\delta \neq 0$$

Για τον λόγο αυτόν αρχικά υπολογίζουμε τις διαφορές $d_1 = x_1 - y_1, \dots, d_9 = x_9 - y_9$. Οι διαφορές αυτές είναι

$$0,\ 0,\ 0.1,\ 0.1,\ 0,\ 0.1,\ 0,\ 0.1,\ 0.1,$$

με δειγματική μέση τιμή και διακύμανση ίση με $\bar{d} = 0.05555556$ και $s_D^2 = 0.002777778$, αντίστοιχα.

Υποθέτοντας ότι αυτές οι διαφορές προέρχονται από κανονικό πληθυσμό, θα χρησιμοποιήσουμε τη σσε

$$T = \frac{\bar{D} - 0}{S_D/\sqrt{n}},$$

με κπ $|T| > t_{n-1,\alpha/2}$. Η τιμή της σσε ισούται με $t = \frac{\bar{d}}{s_d/\sqrt{n}} = \frac{0.05555556}{0.05270463/\sqrt{9}} = 3.162278$. Έχουμε ότι $t_{8,0.025} = 2.306$ (Πίνακας Α΄.4, Παραρτήματος Α΄). Καθώς είναι $3.162278 > 2.306$ συμπεραίνουμε ότι απορρίπτουμε τη μηδενική υπόθεση σε επίπεδο σημαντικότητας 0.05. Δηλαδή με βάση αυτά τα δεδομένα, σε επίπεδο σημαντικότητας 0.05, δεν μπορούμε να ισχυριστούμε ότι οι δύο μέθοδοι μέτρησης είναι ισοδύναμες.
Για να υλοποιήσουμε τον αντίστοιχο έλεγχο μέσω της R χρησιμοποιούμε τις ακόλουθες εντολές:

```
data1<- c(326.5,326.6,326.6,326.8,326.3,326.6,326.7,326.7,326.3)
data2 <- c(326.5,326.6,326.5,326.7,326.3,326.5,326.7,326.6,326.2)
t.test(data1, data2, paired = TRUE)
```

και λαμβάνουμε τα παρακάτω αποτελέσματα:

```
Paired t-test

data:  data1 and data2
t = 3.1623, df = 8, p-value = 0.01335
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.01504319 0.09606792
sample estimates:
mean of the differences
             0.05555556
```

Από τα παραπάνω αποτελέσματα έχουμε ότι $p - value = 0.01335 < 0.05$. Άρα απορρίπτουμε τη μηδενική υπόθεση και σε επίπεδο σημαντικότητας 0.05 μπορούμε να ισχυριστούμε ότι υπάρχει στατιστικά σημαντική διαφοροποίηση στο μέσο επίπεδο μέτρησης μεταξύ των δύο μεθόδων μέτρησης.

Άσκηση Αυτοαξιολόγησης 11.7

Χρησιμοποιώντας τα δεδομένα της Άσκησης Αυτοαξιολόγησης 10.9 να ελέγξετε τον ισχυρισμό ότι δεν διαφέρουν σε επίπεδο σημαντικότητας 0.05 οι δύο μέθοδοι μέτρησης.
Υπόδειξη: να γίνουν κατάλληλες υποθέσεις για την απάντηση του ερωτήματος.

11.6 Έλεγχος για τη διασπορά κανονικού πληθυσμού

Έστω $X_1, \ldots, X_n$ ένα τυχαίο δείγμα από κανονικά κατανεμημένο πληθυσμό με μέση τιμή μ (άγνωστη) και διασπορά σ^2. Μας ενδιαφέρει να ελέγξουμε αν η διασπορά του πληθυσμού μπορεί να θεωρηθεί ότι έχει μια συγκεκριμένη τιμή σ_0^2. Σε αυτήν την περίπτωση, η μηδενική και η εναλλακτική υπόθεση είναι της μορφής:

$$H_0 : \sigma^2 = \sigma_0^2 \quad \text{κατά} \quad H_1 : \sigma^2 \neq \sigma_0^2 \quad \text{ή} \quad \sigma^2 < \sigma_0^2 \quad \text{ή} \quad \sigma^2 > \sigma_0^2.$$

Προφανώς, η υπόθεση αυτή είναι ισοδύναμη με το να ελέγξουμε αν η τυπική απόκλιση ισούται με σ_0.

Στο παραπάνω πλαίσιο έχουμε ότι:

$$X^2 = \frac{(n-1)S^2}{\sigma_0^2} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{\sigma_0^2} \overset{H_0}{\sim} \chi^2_{n-1}, \tag{11.10}$$

δηλαδή ότι η σσε X^2 υπό τη μηδενική υπόθεση, ακολουθεί χι-τετράγωνο κατανομή με $n-1$ βαθμούς ελευθερίας. Η κπ του ελέγχου σε επίπεδο σημαντικότητας α ανάλογα με την εναλλακτική υπόθεση είναι τέτοια, ώστε:

1. αν $H_1 : \sigma^2 \neq \sigma_0^2$, να απορρίπτεται η μηδενική υπόθεση αν $\chi^2 < \chi^2_{n-1,1-\alpha/2}$, ή $\chi^2 > \chi^2_{n-1,\alpha/2}$,
2. αν $H_1 : \sigma^2 > \sigma_0^2$, να απορρίπτεται η μηδενική υπόθεση αν $\chi^2 > \chi^2_{n-1,\alpha}$,
3. αν $H_1 : \sigma^2 < \sigma_0^2$, να απορρίπτεται η μηδενική υπόθεση αν $\chi^2 < \chi^2_{n-1,1-\alpha}$,

όπου χ^2 είναι η τιμή που λαμβάνει η σσε X^2 με βάση τα διαθέσιμα δεδομένα και $\chi^2_{n-1,1-\alpha}$ το $(1-\alpha)$ ποσοστιαίο σημείο της χι-τετράγωνο με $n-1$ βαθμούς ελευθερίας.

Παράδειγμα 11.13: (συνέχεια Παραδείγματος 10.1)

Με βάση τα δεδομένα του Παραδείγματος 10.1 μπορούμε να ισχυριστούμε, σε επίπεδο σημαντικότητας 0.05, ότι η τυπική απόκλιση του pH του διαλύματος Β ξεπερνάει τις 0.05 μονάδες. Θα υπήρχε διαφοροποίηση στο συμπέρασμά σας, αν το επίπεδο σημαντικότητας ήταν 0.01;

Λύση Παραδείγματος 11.13

Έστω X η τ.μ. που παριστάνει τον βαθμό οξύτητας του χημικού διαλύματος Β. Υποθέτουμε ότι $X \sim N(\mu, \sigma^2)$ και θέλουμε να ελέγξουμε σε επίπεδο σημαντικότητας $\alpha = 0.05$ τις υποθέσεις

$$H_0 : \sigma = 0.05 \text{ κατά } H_1 : \sigma > 0.05,$$

ή, ισοδύναμα,

$$H_0 : \sigma^2 = 0.05^2 \text{ κατά } H_1 : \sigma^2 > 0.05^2.$$

Έχουμε $\bar{x} = \frac{\sum_{i=1}^{7} x_i}{7} = 6.5$ και $s^2 = \frac{\sum_{i=1}^{7}(x_i - \bar{x})^2}{7-1} = \frac{\sum_{i=1}^{7} x_i^2 - 7 \cdot 6.5^2}{6} = 0.0054$. Η τιμή της σσε ισούται με

$$\chi^2 = \frac{(7-1)\cdot s^2}{0.05^2} = \frac{6 \cdot 0.0052}{0.0025} = 12.96.$$

Έχουμε ότι $\chi^2_{6,0.05} = 12.592$ (Πίνακας Α΄.6, Παραρτήματος Α΄), οπότε η κπ είναι η $\chi^2 > \chi^2_{6,0.05}$. Επομένως, απορρίπτουμε τη μηδενική υπόθεση σε επίπεδο σημαντικότητας 0.05. Δηλαδή, με βάση αυτά τα δεδομένα και σε επίπεδο σημαντικότητας 0.05, μπορούμε να ισχυριστούμε ότι η τυπική απόκλιση του pH του διαλύματος Β είναι μεγαλύτερη από 0.05 μονάδες.
Από την άλλη πλευρά, αν διεξαχθεί ο έλεγχος σε επίπεδο σημαντικότητας 0.01, δεν μπορεί να απορριφθεί η μηδενική υπόθεση, καθώς $\chi^2 < \chi^2_{6,0.01} = 16.812$ (Πίνακας Α΄.6, Παραρτήματος Α΄).
Εναλλακτικά, για να υλοποιήσουμε τον παραπάνω έλεγχο, θα μπορούσαμε να χρησιμοποιήσουμε τις παρακάτω εντολές της R:

```
library(EnvStats)
B <- c(6.51, 6.55, 6.43, 6.51, 6.62, 6.40, 6.48)
varTest(B, alternative = "greater", conf.level = 0.95,
        sigma.squared = 0.0025, data.name = NULL)
```

από τις οποίες λαμβάνουμε τα παρακάτω αποτελέσματα:

```
Chi-Squared Test on Variance
```

```
data:  B
Chi-Squared = 12.96, df = 6, p-value = 0.04368
alternative hypothesis: true variance is greater than 0.0025
95 percent confidence interval:
 0.002573147         Inf
sample estimates:
variance
  0.0054
```

Παρατηρούμε ότι η $p-value = 0.04368$, άρα σε επίπεδο σημαντικότητας 0.05 απορρίπτεται η μηδενική υπόθεση, ενώ σε επίπεδο σημαντικότητας 0.01 δεν μπορεί να απορριφθεί, δηλαδή καταλήγουμε στα ίδια συμπεράσματα με πριν.

Άσκηση Αυτοαξιολόγησης 11.8

Από προηγούμενες μελέτες είναι γνωστό ότι το βάρος των ξηρών καρπών που συσκευάζει μια εταιρεία ακολουθεί κανονική κατανομή με μέση τιμή μ και διακύμανση σ^2. Ο υπεύθυνος ποιοτικού ελέγχου της εταιρείας θέλοντας να ελέγξει, σε επίπεδο σημαντικότητας 5%, αν η πραγματική τυπική απόκλιση του βάρους είναι 2 γραμμάρια ή μεγαλύτερη, επιλέγει 10 τυχαία προϊόντα, καταγράφει το βάρος του καθενός και υπολογίζει ότι η τυπική απόκλιση του βάρους τους είναι ίση με 2.14 γραμμάρια. Ποιο είναι το συμπέρασμά του;

11.7 Έλεγχος υπόθεσης του λόγου των διασπορών δύο κανονικών πληθυσμών

Σε πολλές στατιστικές μεθόδους, όπως για παράδειγμα στην κατασκευή διαστημάτων εμπιστοσύνης για τη διαφορά των μέσων δύο ανεξάρτητων πληθυσμών με άγνωστες διασπορές, χρειάζεται να γνωρίζουμε αν οι διασπορές των δύο ανεξάρτητων πληθυσμών μπορούν να θεωρηθούν ίσες ή όχι. Για τον λόγο αυτόν, στην ενότητα αυτή, το ενδιαφέρον επικεντρώνεται στον έλεγχο υποθέσεων που αφορούν το πηλίκο διασπορών. Στο πλαίσιο αυτό, έστω $X_1,\ldots,X_n$ τυχαίο δείγμα από έναν κανονικό πληθυσμό με μέση τιμή μ_1 και διασπορά σ_1^2, ενώ $Y_1,\ldots,Y_m$ είναι ένα άλλο ανεξάρτητο τυχαίο δείγμα από έναν κανονικό πληθυσμό με μέση τιμή μ_2 και διασπορά σ_2^2. Θέλουμε να ελέγξουμε τις υποθέσεις

$$H_0 : \sigma_2^2/\sigma_1^2 = 1 \quad \text{κατά} \quad H_1 : \sigma_2^2/\sigma_1^2 \neq 1 \quad \text{ή} \quad \sigma_2^2/\sigma_1^2 > 1 \quad \text{ή} \quad \sigma_2^2/\sigma_1^2 < 1$$

Σε αυτήν την περίπτωση, έχουμε ότι:

$$F = \frac{S_1^2}{S_2^2} \overset{H_0}{\sim} F_{n-1,m-1}, \tag{11.11}$$

με S_1^2 και S_2^2 να είναι οι δειγματικές διακυμάνσεις που βασίζονται στα τ.δ. $X_1,\ldots,X_n$ και $Y_1,\ldots,Y_m$, αντίστοιχα.

Η κπ των παραπάνω ελέγχων σε επίπεδο σημαντικότητας α είναι τέτοια, ώστε:

1. αν $H_1 : \sigma_2^2/\sigma_1^2 \neq 1$, να απορρίπτεται η μηδενική υπόθεση αν $f < F_{n-1,m-1,1-\alpha/2}$ ή $f > F_{n-1,m-1,\alpha/2}$,
2. αν $H_1 : \sigma_2^2/\sigma_1^2 > 1$, να απορρίπτεται η μηδενική υπόθεση αν $f < F_{n-1,m-1,1-\alpha}$,
3. αν $H_1 : \sigma_2^2/\sigma_1^2 < 1$, να απορρίπτεται η μηδενική υπόθεση αν $f > F_{n-1,m-1,\alpha}$,

όπου f η τιμή της σσε F.

Παρατήρηση 11.7

Γενικότερα, θα μπορούσαμε να ελέγξουμε αν ο λόγος των δύο διασπορών ισούται με κάποια γνωστή τιμή λ_0 έναντι μίας εκ των εναλλακτικών να είναι μεγαλύτερη ή μικρότερη ή διαφορετική από λ_0. Σε μια τέτοια περίπτωση έχουμε ότι:

$$F = \frac{S_1^2}{S_2^2} \cdot \lambda_0 \overset{H_0}{\sim} F_{n-1,m-1}, \tag{11.12}$$

Επομένως, όλα τα παραπάνω συνεχίζουν να ισχύουν με την τροποποίηση ότι f είναι η παρατηρηθείσα τιμή της σσε F που ορίστηκε στη σχέση (11.12).

Παρατήρηση 11.8

Οι έλεγχοι αυτής της ενότητας είναι πολύ ευαίσθητοι στην απόκλιση από την υπόθεση της κανονικότητας για τους πληθυσμούς. Αν η υπόθεση αυτή δεν είναι ορθή, τότε το πραγματικό επίπεδο σημαντικότητας α μπορεί να διαφέρει σημαντικά από την ονομαστική τιμή του και η ισχύς του ελέγχου να είναι μικρή. Για αυτό είναι προτιμότερο να χρησιμοποιούνται δείγματα ίσου μεγέθους από τους δύο πληθυσμούς, καθώς σε αυτήν την περίπτωση, ο έλεγχος είναι λιγότερο ευαίσθητος σε απόκλιση από την υπόθεση της κανονικότητας των κατανομών των δύο πληθυσμών (Κουτρουβέλης, 2000).

Παράδειγμα 11.14: (συνέχεια Παραδείγματος 10.1)

Με βάση τα αριθμητικά δεδομένα του Παραδείγματος 10.1, μπορούμε να ισχυριστούμε, σε επίπεδο σημαντικότητας 0.1, ότι οι διασπορές των pH των δύο διαλυμάτων είναι ίσες;

Λύση Παραδείγματος 11.14

Έστω X και Y οι τ.μ. που παριστάνουν τον βαθμό οξύτητας του χημικού διαλύματος τύπου Α και Β, αντίστοιχα. Υποθέτουμε ότι $X \sim N(\mu_1,\sigma_1^2)$ και $Y \sim N(\mu_2,\sigma_2^2)$. Στηριζόμενοι στα δείγματα $X_1,\ldots,X_n$ και $Y_1,\ldots,Y_m$ από αυτούς του πληθυσμούς με $n = 5$ και $m = 7$ θέλουμε να ελέγξουμε την

$$H_0 : \frac{\sigma_2^2}{\sigma_1^2} = 1 \text{ κατά } H_1 : \frac{\sigma_2^2}{\sigma_1^2} \neq 1.$$

Η τιμή f της σσε F της σχέσης (11.11) είναι ίση με $f = \frac{s_1^2}{s_2^2} = \frac{0.00787}{0.0054} = 1.457407$, ενώ σε επίπεδο σημαντικότητας α απορρίπτεται η μηδενική υπόθεση, αν

$$f < F_{n-1,m-1,1-\alpha/2} = F_{4,6,0.95} = \frac{1}{F_{6,4,0.05}} = \frac{1}{6.1631} = 0.162256,$$

ή, αν

$$f > F_{n-1,m-1,a/2} = F_{4,6,0.05} = 4.5337.$$

Επομένως, σε επίπεδο σημαντικότητας 0.1, δεν απορρίπτεται η υπόθεση της ισότητας των διασπορών των pH των δύο διαλυμάτων. Σημειώνεται ότι τα κρίσιμα σημεία της F κατανομής έχουν ληφθεί από τον Πίνακα Α΄.7 του Παραρτήματος Α΄.

Για να υλοποιήσουμε τον παραπάνω έλεγχο μέσω της R χρησιμοποιούμε τις ακόλουθες εντολές

```
A <- c(6.33, 6.28, 6.5, 6.40, 6.45)
B <- c(6.51, 6.55, 6.43, 6.51, 6.62, 6.40, 6.48)
var.test(A, B, ratio = 1, alternative = c("two.sided"),conf.level = 0.90)
```

οι οποίες επιστρέφουν τα παρακάτω αποτελέσματα:

```
  F test to compare two variances

data:  A and B
F = 1.4574, num df = 4, denom df = 6, p-value = 0.6468
alternative hypothesis: true ratio of variances is not equal to 1
90 percent confidence interval:
 0.3214626 8.9821946
sample estimates:
ratio of variances
          1.457407
```

Παρατηρούμε ότι η $p-value = 0.6468 > 0.1$, άρα δεν μπορούμε να απορρίψουμε και με τη χρήση του παρατηρούμενου επιπέδου σημαντικότητας τη μηδενική υπόθεση σε επίπεδο σημαντικότητας 0.1. Επομένως, σε επίπεδο σημαντικότητας 0.1, οι πληθυσμιακές διασπορές δεν διαφέρουν στατιστικά σημαντικά.

Άσκηση Αυτοαξιολόγησης 11.9

Χρησιμοποιώντας τα δεδομένα και τις υποθέσεις του Παραδείγματος 10.9, ελέγξτε σε επίπεδο σημαντικότητας $0,1$ τον ισχυρισμό ότι το πηλίκο των διακυμάνσεων των συγκεντρώσεων ορθοφωσφόρου στους δύο σταθμούς είναι ίσο με ένα.

11.8 Έλεγχος υπόθεσης για το ποσοστό ενός πληθυσμού

Το ενδιαφέρον σε αυτήν την ενότητα επικεντρώνεται σε ελέγχους υποθέσεων για το άγνωστο ποσοστό επιτυχίας p μιας δοκιμής Bernoulli. Για παράδειγμα, δεν είναι λίγες οι φορές που καλούμαστε να αποφασίσουμε αν το ποσοστό των ελαττωματικών προϊόντων μιας γραμμής παραγωγής ξεπερνάει, παραδείγματος χάριν, το 1% ή αν το ποσοστό των ατόμων που πάσχουν από παχυσαρκία είναι μεγαλύτερο από μία συγκεκριμένη τιμή και πολλά άλλα.

Έστω X_i, $i = 1, \dots, n$ η τ.μ. που παριστάνει την έκβαση της δοκιμής Bernoulli στην i-οστή ανεξάρτητη επανάληψη με $X_i = 1$, αν έχουμε επιτυχία στο $i-$οστό πείραμα, και $X_i = 0$, διαφορετικά με $P(X_i = 1) = p$. Στον έλεγχο υπόθεσης για την αναλογία p ενός πληθυσμού η μηδενική και η εναλλακτική υπόθεση έχουν τη μορφή:

$$H_0 : p = p_0 \quad \text{κατά} \quad H_1 : p \neq p_0 \quad \text{ή} \quad p < p_0 \quad \text{ή} \quad p > p_0.$$

Στο πλαίσιο αυτού του συγγράμματος θα ασχοληθούμε με ελέγχους υποθέσεων για την αναλογία ενός πληθυσμού μόνο στις περιπτώσεις όπου το μέγεθος του δείγματος είναι μεγάλο. Τότε (βλ. και Ενότητα 10.3.6), υπό τη μηδενική υπόθεση, για μεγάλο μέγεθος δείγματος n, ισχύει ότι:

$$Z = \frac{\hat{P} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \to^{d}_{H_0} N(0,1), \tag{11.13}$$

με $\hat{P} = \sum_{i=1}^{n} X_i/n$, δηλαδή υπό τη μηδενική υπόθεση η Z ακολουθεί προσεγγιστικά την τυπική κανονική κατανομή και αποτελεί τη σσε. Σε αυτήν την περίπτωση η κπ του ελέγχου σε (ασυμπτωτικό) επίπεδο σημαντικότητας α, ανάλογα με την εναλλακτική υπόθεση, είναι τέτοια, ώστε:

1. αν $H_1 : p \neq p_0$, να απορρίπτεται η μηδενική υπόθεση αν $z > |z_{\alpha/2}|$,

2. αν $H_1 : p > p_0$, να απορρίπτεται η μηδενική υπόθεση αν $z > z_\alpha$,
3. αν $H_1 : p < p_0$, να απορρίπτεται η μηδενική υπόθεση αν $z < -z_\alpha$,

όπου z η τιμή της Z στο συγκεκριμένο σύνολο δεδομένων.

Παράδειγμα 11.15

Ρίχνουμε ένα νόμισμα 100 φορές και έρχεται 45 φορές γράμματα.

1. Με βάση αυτά τα δεδομένα, μπορούμε, σε επίπεδο σημαντικότητας 0.05, να ισχυριστούμε ότι το νόμισμα δεν είναι κάλπικο;
2. Να υπολογίσετε το εύρος του πλήθους των φορών που πρέπει να έρθει γράμματα στις 100 έτσι ώστε να μην μπορεί να απορριφθεί, σε επίπεδο σημαντικότητας 0.05, η υπόθεση ότι το νόμισμα δεν είναι κάλπικο.

Λύση Παραδείγματος 11.15

1. Έστω X_i, $i = 1, \dots, 100$ η τ.μ. που παριστάνει την έκβαση της i-οστής ρίψης του νομίσματος με $X_i = 1$, αν έχουμε κεφαλή στην i–οστή ρίψη και $X_i = 0$, διαφορετικά, με $P(X_i = 1) = p$. Για να μην είναι ένα νόμισμα κάλπικο, θα πρέπει στις μισές φορές κατά μέσο όρο που το ρίχνουμε να έρχεται κεφαλή και στις μισές γράμματα. Άρα ο έλεγχος που πρέπει να κάνουμε είναι ο

$$H_0 : p = 0.5 \text{ κατά } H_1 : p \neq 0.5.$$

Καθώς το μέγεθος δείγματος είναι μεγάλο, θα χρησιμοποιήσουμε τη σσε που δίνεται στη σχέση (11.13). Η τιμή της σσε είναι

$$z = \frac{\hat{p} - 0.5}{\sqrt{\frac{0.5(1-0.5)}{100}}} = \frac{0.45 - 0.5}{\sqrt{\frac{0.5(1-0.5)}{100}}} = -1,$$

καθώς $\hat{p} = \frac{45}{100}$. Η μηδενική υπόθεση απορρίπτεται σε επίπεδο σημαντικότητας 0.05, αν $|z| > z_{\alpha/2} = z_{0.025} = 1.96$ (Πίνακας Α΄.3, Παραρτήματος Α΄). Οπότε με βάση αυτά τα δεδομένα, δεν μπορεί σε επίπεδο σημαντικότητας 0.05 να απορριφθεί η μηδενική υπόθεση, δηλαδή η υπόθεση ότι το νόμισμα δεν είναι κάλπικο.
Ο παραπάνω έλεγχος θα μπορούσε, εναλλακτικά, να εκτελεστεί με τη βοήθεια της ακόλουθης εντολής στην R

```
1  prop.test(45, 100, p =0.5, alternative = "two.sided",
2            correct = FALSE)
```

η οποία επιστρέφει τα ακόλουθα αποτελέσματα

```
 1-sample proportions test without continuity correction

data:  45 out of 100, null probability 0.5
X-squared = 1, df = 1, p-value = 0.3173
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.3561454 0.5475540
sample estimates:
   p
0.45
```

Τα παραπάνω αποτελέσματα μας πληροφορούν ότι η p-τιμή του ελέγχου είναι ίση με 0.3173. Σε αυτό το σημείο αξίζει να επισημανθεί ότι ο έλεγχος που διεξάγει η παραπάνω εντολή χρησιμοποιεί ως στατιστική συνάρτηση ελέγχου το τετράγωνο της σσε που δίνεται στη σχέση (11.13), η οποία προσεγγιστικά ακολουθεί υπό τη μηδενική υπόθεση χι-τετράγωνο κατανομή με 1 βαθμό ελευθερίας.

2. Για να μην μπορούμε να απορρίψουμε την υπόθεση ότι το νόμισμα δεν είναι κάλπικο, δηλαδή τη μηδενική υπόθεση, θα πρέπει η τιμή της στατιστικής συνάρτησης να μην ανήκει στην περιοχή απόρριψης. Επομένως, θα πρέπει $-z_{\alpha/2} < z < z_{\alpha/2}$ ή, ισοδύναμα, ότι

$$-z_{\alpha/2} < \frac{\hat{p} - 0.5}{\sqrt{\frac{0.5(1-0.5)}{100}}} < z_{\alpha/2}$$

από όπου, μετά από λίγη άλγεβρα, προκύπτει ότι

$$0.402 < \frac{x}{100} < 0.598 \text{ ή } 40.2 < x < 59.8 \text{ ή } 41 \leq x \leq 59.$$

Οπότε, αν έρθει γράμματα από 41 μέχρι και 59 φορές, δεν μπορούμε να απορρίψουμε τη μηδενική υπόθεση σε επίπεδο σημαντικότητας 0.05 και μπορούμε να ισχυριστούμε ότι το νόμισμα δεν είναι κάλπικο.

Άσκηση Αυτοαξιολόγησης 11.10

Μια ομάδα ανθρώπων ισχυρίζεται ότι λιγότεροι από τους μισούς κατοίκους που ζουν στην ευρύτερη περιοχή υποστηρίζουν την κατασκευή ενός φράγματος. Σε μία δημοσκόπηση βρέθηκε ότι από τους 600 κατοίκους που πήραν μέρος οι 275 ήταν υπέρ της κατασκευής του φράγματος. Με βάση αυτά τα δεδομένα, θα μπορούσαμε να ισχυριστούμε σε επίπεδο σημαντικότητας 0.05 ότι ευσταθεί ο ισχυρισμός της συγκεκριμένης ομάδας ανθρώπων;

11.9 Έλεγχος υπόθεσης για τη διαφορά δύο πληθυσμιακών ποσοστών με ανεξάρτητα δείγματα

Πολλές φορές μας ενδιαφέρει να ελέγξουμε υποθέσεις που αφορούν τη διαφορά των ποσοστών εμφάνισης ενός χαρακτηριστικού σε δύο πληθυσμούς, δηλαδή τη διαφορά $p_1 - p_2$ με p_i το ποσοστό εμφάνισης του χαρακτηριστικού που μας ενδιαφέρει στον i-οστό πληθυσμό. Για παράδειγμα, μας ενδιαφέρει να ελέγξουμε αν το ποσοστό των καπνιστών είναι το ίδιο στους άντρες και στις γυναίκες ή αν το ποσοστό των ελαττωματικών προϊόντων που παράγονται στο εργοστάσιο Α είναι μεγαλύτερο από το ποσοστό των ελαττωματικών προϊόντων που παράγονται στο Β. Επομένως, το ενδιαφέρον επικεντρώνεται στον έλεγχο

$$H_0 : p_1 - p_2 = \delta \quad \text{κατά} \quad H_1 : p_1 - p_2 \neq \delta \quad \text{ή} \quad p_1 - p_2 < \delta \quad \text{ή} \quad p_1 - p_2 > \delta,$$

με δ μια δοθείσα τιμή στο διάστημα $[-1,1]$.

Για τον λόγο αυτόν παίρνουμε ένα τυχαίο δείγμα $X_1, \ldots, X_n$ από τον πρώτο πληθυσμό και ένα τυχαίο δείγμα $Y_1, \ldots, Y_m$ από τον δεύτερο πληθυσμό. Σημειώνεται ότι $X_i = 1,\ i = 1, \ldots, n$, αν έχουμε επιτυχία στην i-οστή επανάληψη δοκιμής Bernoulli με πιθανότητα επιτυχίας $P(X_i = 1) = p_1$ και 0, διαφορετικά και όπου $Y_j = 1,\ j = 1, \ldots, m$ αν έχουμε επιτυχία στην j-οστή επανάληψη της δοκιμής Bernoulli με πιθανότητα επιτυχίας $P(Y_j = 1) = p_2$ και 0, διαφορετικά. Στο πλαίσιο αυτού του συγγράμματος θα ασχοληθούμε με ελέγχους υποθέσεων για τη διαφορά των ποσοστών $p_1 - p_2$ μόνο στις περιπτώσεις όπου τα δύο δείγματα είναι ανεξάρτητα μεταξύ τους και μεγάλα σε μέγεθος. Σε όσα ακολουθούν είναι $\hat{P}_1 = \frac{X}{n}$ και $\hat{P}_2 = \frac{Y}{m}$, όπου $X = \sum_{i=1}^{n} X_i$ και $Y = \sum_{j=1}^{m} Y_j$.

Σε αυτήν την περίπτωση, η μορφή της στατιστικής συνάρτησης ελέγχου εξαρτάται από το δ. Αν $\delta = 0$, η σσε έχει τη μορφή

$$Z = \frac{\hat{P}_1 - \hat{P}_2}{\sqrt{\hat{P}(1-\hat{P})(1/n + 1/m)}}$$

όπου $\hat{P} = \frac{X+Y}{n+m}$.

Από την άλλη, αν $\delta \neq 0$

$$Z = \frac{\hat{P}_1 - \hat{P}_2 - \delta}{\sqrt{\hat{P}_1(1-\hat{P}_1)/n + \hat{P}_2(1-\hat{P}_2)/m}}.$$

Η διαφορά αυτών των δύο σσε έγκειται στο γεγονός ότι αν $\delta = 0$, τότε υπό τη μηδενική υπόθεση οι δύο πληθυσμοί έχουν την ίδια διασπορά (θυμηθείτε τη σχέση ορισμού της διασποράς της κατανομής Bernoulli, όπως αυτή προκύπτει ως ειδική περίπτωση της διωνυμικής κατανομής) και, επομένως, η κοινή τιμή της διασποράς μπορεί να εκτιμηθεί συνδυάζοντας τα δύο δείγματα. Στην αντίθετη περίπτωση, δηλαδή όταν $\delta \neq 0$, τότε οι δύο πληθυσμοί διαφέρουν και, επομένως, η διασπορά τους εκτιμάται ξεχωριστά από κάθε δείγμα. Στην πραγματικότητα, η διαφοροποίηση αυτή στις δύο σσε είναι ανάλογη με αυτή που παρουσιάστηκε στην Ενότητα 11.4.

Σημειώνεται ότι και στις δύο περιπτώσεις η σσε Z ακολουθεί υπό τη μηδενική υπόθεση προσεγγιστικά τυπική κανονική κατανομή $N(0,1)$ και, επομένως, η κπ του ελέγχου σε (ασυμπτωτικό) επίπεδο σημαντικότητας α, ανάλογα με τη μορφή της εναλλακτικής υπόθεσης, είναι τέτοια, ώστε:

1. αν $H_1 : p_1 - p_2 \neq \delta$, να απορρίπτεται η μηδενική υπόθεση αν $z > |z_{\alpha/2}|$,
2. αν $H_1 : p_1 - p_2 > \delta$, να απορρίπτεται η μηδενική υπόθεση αν $z > z_\alpha$,
3. αν $H_1 : p_1 - p_2 < \delta$, να απορρίπτεται η μηδενική υπόθεση αν $z < -z_\alpha$,

όπου με z συμβολίζεται η τιμή της σσε Z.

Παράδειγμα 11.16

Μία βιομηχανία προμηθεύεται κάποιο εξάρτημα από δύο κατασκευαστές, Α και Β. Επειδή η βιομηχανία είναι δυσαρεστημένη από τον κατασκευαστή Α, θέλει να αλλάξει τη μέχρι τώρα πολιτική της και να αγοράζει αποκλειστικά από τον κατασκευαστή Β, εφόσον η ποιότητα εξαρτημάτων του Β δεν υστερεί αυτής του Α. Σε έναν έλεγχο 110 εξαρτημάτων από τον Α και 150 εξαρτημάτων από τον Β βρέθηκαν 4 και 12 εξαρτήματα ελαττωματικά, αντίστοιχα. Υπάρχει λόγος να μην αγοράσει η βιομηχανία αποκλειστικά από τον κατασκευαστή Β σε επίπεδο σημαντικότητας 0.1; Υπάρχει διαφοροποίηση στο συμπέρασμά σας, αν το επίπεδο σημαντικότητας είναι 0.05; Ποια είναι η $p-$τιμή του ελέγχου;

Λύση Παραδείγματος 11.16

Η μηδενική και η εναλλακτική υπόθεση που πρέπει να ελεγχθούν είναι οι εξής:

$$H_0 : p_1 = p_2 \text{ κατά } H_1 : p_1 < p_2,$$

ή, ισοδύναμα,

$$H_0 : p_1 - p_2 = 0 \text{ κατά } H_1 : p_1 - p_2 < 0,$$

όπου p_1 και p_2 το πραγματικό ποσοστό των ελαττωματικών εξαρτημάτων της βιομηχανίας Α και Β, αντίστοιχα.

Θα χρησιμοποιήσουμε τη σσε

$$Z = \frac{\hat{P}_1 - \hat{P}_2}{\sqrt{\hat{P}(1-\hat{P})(1/n + 1/m)}}.$$

Από τα δεδομένα της εκφώνησης είναι $\hat{p}_1 = 4/110 = 0.03636$, $\hat{p}_2 = 12/150 = 0.08$ και $\hat{p} = \frac{4+12}{110+150} =$

$\frac{16}{260} = 0.06154$. Επομένως, η τιμή της σσε είναι

$$z = \frac{0.03636 - 0.08}{\sqrt{0.06154(1 - 0.06154) \cdot (1/110 + 1/150)}} = -1.44651.$$

Γνωρίζουμε ότι η μηδενική υπόθεση απορρίπτεται, σε επίπεδο σημαντικότητας 0.1, αν $z < -z_\alpha = -z_{0.1} = -1.28$ (Πίνακας Α΄.3, Παραρτήματος Α΄). Επομένως, απορρίπτεται η μηδενική υπόθεση σε επίπεδο σημαντικότητας 0.1. Με βάση αυτά τα δεδομένα μπορούμε να ισχυριστούμε ότι υπάρχει στατιστικά σημαντική διαφορά στο πραγματικό ποσοστό ελαττωματικών εξαρτημάτων των δύο κατασκευαστών, άρα υπάρχει λόγος η βιομηχανία να συνεχίσει να αγοράζει από τον κατασκευαστή Α.
Αν τον παραπάνω έλεγχο τον κάνουμε σε επίπεδο σημαντικότητας 0.05, καθώς $-z_{0.05} = -1.645$ (Πίνακας Α΄.3, Παραρτήματος Α΄) και $z > -z_{0.05}$, δεν θα μπορούσε να απορριφθεί η μηδενική υπόθεση και θα είχαμε οδηγηθεί σε διαφορετικά αποτελέσματα.
Υπολογίζοντας, την $p-$τιμή του ελέγχου μπορούμε να βρούμε ποιο είναι το μικρότερο επίπεδο σημαντικότητας για το οποίο μπορούμε να απορρίψουμε τη μηδενική υπόθεση. Ειδικότερα, για τον παραπάνω έλεγχο έχουμε

$$p - value = P(Z < -1.44651) = P(Z > 1.44651) = 1 - \Phi(1.44651) = 0.07402.$$

Επομένως, για κάθε $\alpha > 0.07402$ απορρίπτεται η μηδενική υπόθεση.
Σημειώνουμε ότι όλα τα παραπάνω είναι προσεγγιστικά, καθώς η κατανομή της σσε Z είναι προσεγγιστικά η τυπική κανονική.

Ο παραπάνω έλεγχος θα μπορούσε να πραγματοποιηθεί με τη βοήθεια της R εκτελώντας τις ακόλουθες εντολές

```
x<-c(4,12)
n<-c(110,150)
prop.test(x, n,  alternative = "less",correct =F)
```

οι οποίες επιστρέφουν τα παρακάτω αποτελέσματα

```
  2-sample test for equality of proportions without continuity
  correction

data:  x out of n
X-squared = 2.0924, df = 1, p-value = 0.07402
alternative hypothesis: less
95 percent confidence interval:
 -1.000000000  0.003154566
sample estimates:
    prop 1     prop 2
0.03636364 0.08000000
```

Τα παραπάνω αποτελέσματα μας πληροφορούν ότι η p-τιμή του ελέγχου είναι ίση με 0.07402. Με βάση αυτήν την τιμή, καταλήγουμε στα ίδια συμπεράσματα με πριν.
Σε αυτό το σημείο αξίζει να επισημανθεί ότι ο έλεγχος που διεξάγει η παραπάνω εντολή χρησιμοποιεί ως στατιστική συνάρτηση ελέγχου το τετράγωνο της σσε Z, η οποία προσεγγιστικά ακολουθεί υπό τη μηδενική υπόθεση χι-τετράγωνο κατανομή με 1 βαθμό ελευθερίας.

Άσκηση Αυτοαξιολόγησης 11.11

Από την παραγωγή δύο μηχανών τυχαία δείγματα μεγέθους 450 και 400 έδωσαν αριθμό ελαττωματικών προϊόντων 17 και 12 αντίστοιχα. Να εξεταστεί αν υπάρχει διαφορά στα ποσοστά παραγωγής ελαττωματικών προϊόντων από τις δύο μηχανές σε επίπεδο σημαντικότητας 0.01.

11.10 Ασκήσεις

Άσκηση 11.1 *Απαντήστε στις παρακάτω ερωτήσεις δικαιολογώντας τις απαντήσεις σας.*

1. *Ένας ερευνητής έκανε έναν μονόπλευρο στατιστικό έλεγχο σε επίπεδο σημαντικότητας $\alpha = 0.01$ και το αποτέλεσμα του ελέγχου ήταν ότι η μηδενική υπόθεση απορρίπτεται. Ένας άλλος ερευνητής, χρησιμοποιώντας τα ίδια δεδομένα, έκανε έναν δίπλευρο έλεγχο σε επίπεδο σημαντικότητας $\alpha = 0.05$ και το αποτέλεσμα του έλεγχου ήταν ότι η μηδενική υπόθεση δεν απορρίπτεται. Είναι δυνατόν να είναι σωστά και τα δύο αποτελέσματα;*
2. *Αν σε επίπεδο σημαντικότητας α, η μηδενική υπόθεση $H_0 : \mu = \mu_0$ απορρίπτεται υπέρ της εναλλακτικής $H_0 : \mu > \mu_0$, τότε στο ίδιο επίπεδο σημαντικότητας θα πρέπει να απορρίπτεται απαραίτητα η $H_0 : \mu = \mu_0$ υπέρ της $H_0 : \mu \neq \mu_0$;*
3. *Ένας ερευνητής θέλει να ελέγξει αν η ασπιρίνη επηρεάζει την τιμή ενός αιματολογικού δείκτη ο οποίος σχετίζεται με την πηκτικότητα του αίματος και τη δημιουργία θρόμβων. Για τον σκοπό αυτόν επιλέγει ένα τυχαίο δείγμα 12 ατόμων και μετράει σε κάθε άτομο την τιμή του δείκτη πριν και τρεις ώρες μετά τη λήψη δύο δισκίων ασπιρίνης. Τι έλεγχο πρέπει να κάνει ο ερευνητής για να ελέγξει, αν τα πειραματικά του δεδομένα υποστηρίζουν ότι η μέση τιμή του δείκτη αλλάζει μετά τη λήψη των δισκίων ασπιρίνης;*
4. *Έστω $X_1, X_2, \ldots, X_n$ ένα τυχαίο δείγμα από κανονική κατανομή με μέση τιμή μ και γνωστή διασπορά σ^2. Μπορούμε να χρησιμοποιήσουμε διάστημα εμπιστοσύνης για να ελέγξουμε την υπόθεση $H_0 : \mu = \mu_0$ κατά $H_1 : \mu \neq \mu_0$; Αν ναι, πότε απορρίπτουμε τη μηδενική υπόθεση;*

Άσκηση 11.2 *Έστω X_1 και X_2 ανεξάρτητες παρατηρήσεις από κανονικό πληθυσμό με μέση τιμή μ και τυπική απόκλιση 1. Θέλουμε να ελέγξουμε την υπόθεση*

$$H_0 : \mu = 10 \text{ κατά } H_1 : \mu > 10.$$

1. *Αν η κρίσιμη περιοχή του ελέγχου είναι $\sum_{i=1}^{n} X_i > 12.5 \cdot n$, να υπολογιστούν οι πιθανότητες σφάλματος τύπου I και σφάλματος τύπου II, αν ο πραγματικός μέσος του πληθυσμού είναι $\mu = 14$.*
2. *Πόσο θα πρέπει να είναι το ελάχιστο μέγεθος του δείγματος έτσι ώστε η πιθανότητα σφάλματος τύπου II να μην ξεπερνάει το 0.001;*
3. *Ποιο είναι το ελάχιστο επίπεδο σημαντικότητας για το οποίο δεν μπορεί να απορριφθεί η μηδενική υπόθεση, αν έχουν παρατηρηθεί οι τιμές 10.37 και 13.43;*

Άσκηση 11.3 *Ένας δημοσιογράφος σε μια τηλεοπτική συζήτηση ισχυρίζεται ότι συμβαίνει κάτι ύποπτο στην εταιρεία μέτρησης τηλεθέασης GBA. Η εταιρεία έχει διαθέσει 1150 μηχανάκια σε ισάριθμους τηλεθεατές, 150 από τα οποία έχουν δοθεί σε νέους συνεργάτες (τηλεθεατές που έχουν ξεκινήσει τη συνεργασία τους με την GBA το τελευταίο εξάμηνο). Ο δημοσιογράφος διαπιστώνει ότι, σύμφωνα με την GBA, το κανάλι Α παρουσιάζει τηλεθέαση 30% στο σύνολο των τηλεθεατών (παλιών και νέων), ενώ έχει μόνο 22% στους νέους τηλεθεατές. Ο δημοσιογράφος το θεωρεί αυτό πολύ ύποπτο και υποψιάζεται ότι οι νέοι τηλεθεατές επιλέχθηκαν με τέτοιο τρόπο, ώστε να μειωθεί σημαντικά το ποσοστό τηλεθέασης του καναλιού Α.*

1. *Με βάση αυτά τα δεδομένα μπορούμε να ισχυριστούμε, σε επίπεδο σημαντικότητας 0.05, ότι ευσταθεί ο ισχυρισμός του δημοσιογράφου;*
2. *Τι σφάλμα κινδυνεύουμε να κάνουμε σε αυτόν τον έλεγχο και γιατί;*

Άσκηση 11.4 *Κατασκευαστής ισχυρίζεται ότι το πολύ 2% των προϊόντων του είναι ελαττωματικά. Σε ένα τυχαίο δείγμα 900 προϊόντων βρέθηκαν 27 ελαττωματικά. Να ελεγχθεί ο ισχυρισμός του κατασκευαστή σε επίπεδο σημαντικότητας 0.05.*

Άσκηση 11.5 *Από έρευνες που έχουν γίνει σε σχολεία, όταν οι μαθητές καλούνται να επιλέξουν τυχαία έναν αριθμό από το ένα (1) μέχρι το είκοσι (20), οι επιλογές τους μοιάζουν να έχουν μια μεροληψία υπέρ του αριθμού δεκαεπτά 17. Σε μία ομάδα 371 μαθητών 25 επέλεξαν το 17.*

1. *Με βάση αυτά τα δεδομένα μπορούμε να ισχυριστούμε σε επίπεδο σημαντικότητας 0.05 ότι όντως υπάρχει μια μεροληψία υπέρ του αριθμού 17; Ο έλεγχος να γίνει με χρήση του παρατηρούμενου επιπέδου σημαντικότητας.*
2. *Να υπολογιστεί η ισχύς του παραπάνω ελέγχου, αν το πραγματικό ποσοστό των μαθητών που επιλέγουν τον αριθμό 17 είναι 7.5%.*
3. *Πάνω από πόσες φορές αν επιλεχθεί ο αριθμός 17 θα μπορούμε να θεωρήσουμε ότι υπάρχει μεροληψία υπέρ του, σε επίπεδο σημαντικότητας 0.05;*

Άσκηση 11.6 *Σε μια χημική διαδικασία το ειδικό διάλυμα που προκύπτει μπορεί να χρησιμοποιηθεί μόνο αν έχει pH το πολύ 8.20. Μία μέθοδος που προσδιορίζει το pH των διαλυμάτων αυτού του τύπου είναι γνωστό ότι δίνει μετρήσεις οι οποίες κατανέμονται κανονικά με μέση τιμή την πραγματική μέση τιμή του pH και τυπική απόκλιση 0.2.*

1. *Αν από εννιά ανεξάρτητες μετρήσεις του pH του διαλύματος που προκύπτει από τη συγκεκριμένη χημική διαδικασία παίρναμε μέση τιμή pH 8.12, θα μπορούσαμε να ισχυριστούμε σε επίπεδο σημαντικότητας 0.05 ότι το συγκεκριμένο διάλυμα μπορεί να χρησιμοποιηθεί;*
2. *Ποιο είναι το παρατηρούμενο επίπεδο σημαντικότητας (p--τιμή) του ελέγχου;*
3. *Επιθυμούμε να σχεδιάσουμε έναν έλεγχο έτσι ώστε, αν το pH είναι πραγματικά 8.20, να απορρίψουμε τη μηδενική υπόθεση με πιθανότητα 5%. Από την άλλη πλευρά, αν το pH είναι 8.03, η πιθανότητα να αποδεχτούμε τη μηδενική υπόθεση να είναι 5%. Ποιο είναι το απαιτούμενο μέγεθος δείγματος για να ικανοποιούνται οι παραπάνω υποθέσεις;*

Άσκηση 11.7 *Μια αεροπορική εταιρεία θέλει να υπολογίσει το ποσοστό των επιβατών που ταξιδεύουν για επαγγελματικούς λόγους σε ένα νέο δρομολόγιο, το οποίο ενσωμάτωσε στις προγραμματισμένες πτήσεις της Αθήνα - Λονδίνο. Από ένα τυχαίο δείγμα 432 επιβατών αυτού του δρομολογίου το 62.5% ταξίδευε για επαγγελματικούς λόγους.*

1. *Με βάση αυτά τα δεδομένα μπορεί η εταιρεία να ισχυριστεί, σε επίπεδο σημαντικότητας 0.05, ότι το 70% των επιβατών του νέου δρομολογίου ταξιδεύουν για επαγγελματικούς λόγους;*
2. *Να υπολογιστεί το παρατηρούμενο επίπεδο σημαντικότητας του ελέγχου.*
3. *Πόσα άτομα θα πρέπει να ταξιδεύουν για επαγγελματικούς λόγους σε αυτό το δρομολόγιο για να μην μπορούμε να απορρίψουμε τη μηδενική υπόθεση;*
4. *Σε ένα αντίστοιχο δρομολόγιο Θεσσαλονίκη - Λονδίνο της εταιρείας, από τους 276 επιβάτες που ρωτήθηκαν ταξιδεύουν για επαγγελματικούς λόγους οι 168. Με βάση αυτά τα δεδομένα, μπορούμε να ισχυριστούμε, σε επίπεδο σημαντικότητας 0.1, ότι τα ποσοστά των επιβατών που ταξιδεύουν για επαγγελματικούς λόγους δεν διαφέρουν στα δύο δρομολόγια;*

Άσκηση 11.8 *Τα βιομηχανικά απόβλητα που ρίχνονται στα ποτάμια απορροφούν το διαλυμένο στο νερό οξυγόνο με συνέπεια αυτό να μειώνεται και, όταν η μέση τιμή του δεν υπερβαίνει τα $5\ ppm$, να δημιουργείται σοβαρό πρόβλημα επιβίωσης των υδρόβιων οργανισμών. Το πρόβλημα αυτό είχε διαπιστωθεί, πριν από αρκετά χρόνια, και στον ποταμό Καλαμά. Για την αντιμετώπισή του εφαρμόστηκε ειδικό πρόγραμμα αποκατάστασης και προστασίας του ποταμού. Για να ελεγχθεί αν απέδωσαν τα μέτρα προστασίας, έπρεπε, μεταξύ άλλων δεικτών, να μελετηθεί η ποσότητα διαλυμένου οξυγόνου στα νερά του ποταμού, η οποία ακολουθεί κανονική κατανομή με τυπική απόκλιση 0.18. Για τον σκοπό αυτόν, συλλέχθηκαν μετρήσεις από 10 σημεία της κοίτης του ποταμού. Οι 10 μετρήσεις έδωσαν μέση τιμή διαλυμένου οξυγόνου (σε ppm) 5.15.*

1. *Με βάση αυτά τα δεδομένα, μπορούμε να ισχυριστούμε, σε επίπεδο σημαντικότητας 0.05, ότι τα μέτρα προστασίας του ποταμού ήταν αποτελεσματικά;*

2. *Να υπολογίσετε την p-τιμή του παραπάνω ελέγχου.*
3. *Να υπολογίσετε το Σφάλμα τύπου II του παραπάνω ελέγχου, αν η πραγματική μέση ποσότητα διαλυμένου οξυγόνου στα νερά του ποταμού είναι 5.1* ppm.
4. *Τι μέγεθος δείγματος πρέπει να επιλέξει ο ερευνητής, έτσι ώστε το Σφάλμα τύπου II να είναι* 10%*;*

Άσκηση 11.9 *Ένα εργοστάσιο κατασκευάζει ηλεκτρικά καλώδια και τα συσκευάζει σε τεμάχια των 10 μέτρων. Τεμάχια που έχουν τουλάχιστον μία ατέλεια χαρακτηρίζονται ως ελαττωματικά και απορρίπτονται από το ίδιο το εργοστάσιο ή επιστρέφονται από τον αγοραστή, προκαλώντας οικονομική ζημιά. Για την αξιολόγηση της παραγωγικής διαδικασίας του εργοστασίου ελήφθη τυχαίο δείγμα 200 τεμαχίων, το* 6.5% *των οποίων βρέθηκαν ελαττωματικά. Το ποσοστό των ελαττωματικών τεμαχίων κρίθηκε ιδιαίτερα μεγάλο και το εργοστάσιο προχώρησε σε εκτεταμένη αναθεώρηση όλων των σταδίων της παραγωγικής διαδικασίας. Μετά την ολοκλήρωση της αναθεώρησης ελήφθη δείγμα 250 τεμαχίων και εντοπίστηκαν 5 ελαττωματικά τεμάχια.*

1. *Υπάρχουν στατιστικά σημαντικές ενδείξεις, σε επίπεδο σημαντικότητας* 0.03*, ότι η εκτεταμένη αναθεώρηση όλων των σταδίων κατασκευής των καλωδίων βελτίωσε την παραγωγική διαδικασία;*
2. *Να υπολογιστεί το παρατηρούμενο επίπεδο σημαντικότητας του παραπάνω ελέγχου.*
3. *Ποιο είναι το μεγαλύτερο πλήθος ελαττωματικών τεμαχίων που μπορούμε να έχουμε στο δείγμα των 250 τεμαχίων μετά την αναθεώρηση, ώστε να μπορούμε να απορρίψουμε τη μηδενική υπόθεση;*

Άσκηση 11.10 *Μια ομάδα γενετιστών ισχυρίζεται ότι τα γονίδια που καθορίζουν αν κάποιος είναι αριστερόχειρας επηρεάζουν επίσης την ανάπτυξη των γλωσσικών κέντρων του εγκεφάλου. Σύμφωνα με τον παραπάνω ισχυρισμό θα περιμέναμε οι αριστερόχειρες να έχουν πιο αναπτυγμένες γλωσσικές ικανότητες. Προκειμένου να ελεγχθεί ο παραπάνω ισχυρισμός επιλέχθηκαν τυχαία 807 φοιτητές οι οποίοι εξετάστηκαν στη γλώσσα. Τα αποτελέσματα της εξέτασης ταξινομήθηκαν σε τρεις κατηγορίες: χαμηλή, μέση και υψηλή επίδοση. Επίσης, σημειώθηκε και αν ο αντίστοιχος φοιτητής είναι αριστερόχειρας ή όχι. Τα αποτελέσματα δίνονται στον παρακάτω πίνακα*

		Επίδοση στο τεστ		
		χαμηλή	*μέση*	*υψηλή*
χέρι	*αριστερό*	*18*	*40*	*22*
	δεξί	*201*	*360*	*166*

1. *Είναι το ποσοστό των αριστερόχειρων με υψηλή επίδοση στο τεστ γλώσσας μεγαλύτερο από το αντίστοιχο ποσοστό των δεξιόχειρων σε επίπεδο σημαντικότητας* 0.05*; Ο έλεγχος να γίνει με χρήση του παρατηρούμενου επιπέδου σημαντικότητας.*
2. *Με βάση αυτά τα δεδομένα, μπορούμε να ισχυριστούμε, σε επίπεδο σημαντικότητας* 0.01*, ότι οι μισοί φοιτητές έχουν μέση επίδοση στο τεστ της γλώσσας;*

11.11 Απαντήσεις στις ασκήσεις αυτοαξιολόγησης

Λύση Άσκησης Αυτοαξιολόγησης 11.1

Θέλουμε να ελέγξουμε αν η μέση τιμή κανονικού πληθυσμού είναι ίση με δοθείσα γνωστή τιμή ($H_0 : \mu = 150$) έναντι της εναλλακτικής ότι είναι μεγαλύτερη ($H_1 : \mu > 150$). Καθώς η διασπορά του κανονικού πληθυσμού είναι γνωστή, η σσε δίνεται από τη σχέση (11.1) για $\mu_0 = 150$, $\sigma = 12$, δηλαδή είναι

$$Z = \frac{\bar{X} - 150}{12/\sqrt{n}}.$$

Η μηδενική υπόθεση απορρίπτεται, όταν η παρατηρούμενη τιμή z της τυχαίας μεταβλητής Z είναι $z > z_\alpha$, για $\alpha = 0.01$, δηλαδή απορρίπτεται η μηδενική υπόθεση όταν $z > 2.32$ (Πίνακας Α΄.3, Παραρτήματος Α΄).

Επιπλέον, επιθυμούμε να ισχύει $\gamma(\mu) = 0.93$, όταν η πραγματική τιμή του $\mu = 156$. Η ισχύς του ελέγχου έχει προσδιοριστεί στη σχέση (11.3), από όπου έχουμε ότι:

$$\gamma(156) = 1 - \Phi\left(z_{0.01} - \frac{(156 - 150)\sqrt{n}}{12}\right) = 1 - \Phi\left(2.32 - 0.5\sqrt{n}\right).$$

Επομένως, θα πρέπει να προσδιοριστεί το μέγεθος του δείγματος, έτσι ώστε:

$$0.07 = \Phi\left(2.32 - 0.5\sqrt{n}\right),$$

ή, ισοδύναμα,

$$\Phi(-1.47) = \Phi\left(2.32 - 0.5\sqrt{n}\right)$$

από όπου έχουμε ότι $2.32 - 0.5\sqrt{n} = -1.47$, δηλαδή $n = 57.4564$. Επομένως, πρέπει να έχουμε δείγμα μεγέθους 58.

Λύση Άσκησης Αυτοαξιολόγησης 11.2

1. Έστω X η τ.μ. που παριστάνει τον χρόνο ζωής μιας λυχνίας τηλεόρασης που παράγεται με τον νέο τρόπο με $X \sim N(\mu, \sigma^2 = 300^2)$. Θέλουμε να ελέγξουμε τις υποθέσεις

   $$H_0 : \mu = 1200 \text{ κατά } H_1 : \mu > 1200.$$

   Έχουμε δειγματοληψία από κανονικό πληθυσμό με γνωστή διασπορά. Επομένως, θα χρησιμοποιηθεί η σσε της σχέσης (11.1), η οποία παίρνει τιμή

   $$z = \frac{1265 - 1200}{300/\sqrt{100}} = 2.17.$$

   Η μηδενική υπόθεση απορρίπτεται σε επίπεδο σημαντικότητας 0.05, όταν η παρατηρούμενη τιμή z της τυχαίας μεταβλητής Z της σχέσης (11.1) είναι $z > z_\alpha$ για $\alpha = 0.05$. Επομένως, η μηδενική υπόθεση απορρίπτεται σε επίπεδο σημαντικότητας 0.05, καθώς $z > 1.645$. Με βάση αυτά τα δεδομένα μπορούμε, σε επίπεδο σημαντικότητας 0.05, να ισχυριστούμε ότι ο μέσος χρόνος ζωής μίας λυχνίας τηλεόρασης ξεπερνάει τις 1200 ώρες.
2. Το παρατηρούμενο επίπεδο σημαντικότητας είναι:

   $$p - \text{τιμή} = P(Z > 2.17) = 1 - \Phi(2.17) = 1 - 0.985 = 0.015.$$

3. Η ισχύς του ελέγχου υπολογίζεται από τη σχέση (11.3), οπότε έχουμε

   $$\gamma(1240) = 1 - \Phi\left(z_{0.05} - \frac{(1240 - 1200)\sqrt{100}}{300}\right) = 1 - \Phi(0.33) = 1 - 0.62930 = 0.3707,$$

   όπου η τιμή $\Phi(0.33)$ έχει προσδιοριστεί με τη βοήθεια του Πίνακα Α΄.3 του Παραρτήματος Α΄.

Εναλλακτικά, εκτελώντας στην R τις εντολές

```
1 library(PASWR)
2 zsum.test(mean.x=1265, sigma.x =300, n.x =100,alternative = "greater", mu =
      1200, conf.level = 0.95)
3 1-pnorm(qnorm(0.95,0,1)-(1240-1200)*sqrt(100)/(300),0,1)
```

λαμβάνουμε τα εξής αποτελέσματα:

```
      One-sample z-Test

data:  Summarized x
z = 2.1667, p-value = 0.01513
alternative hypothesis: true mean is greater than 1200
95 percent confidence interval:
 1215.654      Inf
sample estimates:
mean of x
     1265

[1] 0.3777026
```

από τα οποία, αρχικά, παρατηρούμε ότι η $p-$τιμή του ελέγχου ισούται με $0.01513 < 0.05$, τιμή που πρακτικά ταυτίζεται με την τιμή που βρήκαμε νωρίτερα και, επομένως, μας οδηγεί στο ίδιο συμπέρασμα με πριν, δηλαδή την απόρριψη της μηδενικής υπόθεσης. Η τιμή 0.3777026 που εμφανίζεται στο τέλος των αποτελεσμάτων υπολογίζεται από την τελευταία εντολή και υπολογίζει την ισχύ του ελέγχου για $\mu = 1240$. Η τιμή αυτή έρχεται σε συμφωνία με την τιμή που υπολογίστηκε με τη βοήθεια του πίνακα της τυπικής κανονικής κατανομής.

Λύση Άσκησης Αυτοαξιολόγησης 11.3

Έστω X η τ.μ. που παριστάνει το ποσοστό χαλκού σε ένα κράμα μέταλλου με $X \sim N(\mu, \sigma^2)$. Θέλουμε να ελέγξουμε την

$$H_0 : \mu = 18.5 \text{ κατά } H_1 : \mu \neq 18.5.$$

Έχουμε δειγματοληψία από κανονικό πληθυσμό με άγνωστη διασπορά. Επομένως, θα χρησιμοποιηθεί η σσε της σχέσης (11.5), η οποία παίρνει τιμή

$$t = \frac{\bar{x} - 18.5}{s/\sqrt{n}} = \frac{18.7 - 18.5}{0.26/\sqrt{10}} = 2.432521.$$

Γνωρίζουμε ότι η μηδενική υπόθεση απορρίπτεται, αν $|t| \geq t_{n-1,\alpha/2}$. Είναι $t_{n-1,\alpha/2} = t_{9,0.025} = 2.262$ (Πίνακας Α΄.4, Παραρτήματος Α΄). Επομένως, καθώς η τιμή της σσε ανήκει στην κρίσιμη περιοχή εξάγουμε το συμπέρασμα ότι απορρίπτεται η μηδενική υπόθεση σε επίπεδο σημαντικότητας 0.05. Επομένως, με βάση αυτά τα δεδομένα μπορούμε να ισχυριστούμε, σε επίπεδο σημαντικότητας 0.05, ότι το κράμα μετάλλων δεν πληροί τις προδιαγραφές.
Εναλλακτικά, χρησιμοποιώντας τις παρακάτω εντολές της R

```
1 library(BSDW)
2 tsum.test(mean.x=18.7,s.x =0.26,n.x = 10,alternative = "two.sided",mu =
    18.5,conf.level = 0.95)
```

λαμβάνουμε τα εξής αποτελέσματα:

```
        One-sample t-Test

data:  Summarized x
t = 2.4325, df = 9, p-value = 0.03783
alternative hypothesis: true mean is not equal to 18.5
95 percent confidence interval:
 18.51401 18.88599
sample estimates:
mean of x
     18.7
```

Παρατηρούμε ότι η $p-$τιμή του ελέγχου ισούται με $0.03783 < 0.05$. Άρα απορρίπτεται η μηδενική υπόθεση και καταλήγουμε στο ίδιο συμπέρασμα με πριν.

Λύση Άσκησης Αυτοαξιολόγησης 11.4

Έστω X η τ.μ. που παριστάνει το σημείο τήξης με $X \sim N(\mu, \sigma^2)$. Θέλουμε να ελέγξουμε τις υποθέσεις

$$H_0 : \mu = 1260 \text{ κατά } H_1 : \mu \neq 1260.$$

Έχουμε δειγματοληψία από κανονικό πληθυσμό με άγνωστη διασπορά. Επομένως, θα χρησιμοποιηθεί η σσε της σχέσης (11.5), η οποία παίρνει τιμή

$$t = \frac{\bar{x} - 1260}{s/\sqrt{n}} = \frac{1264 - 1260}{3.559026/\sqrt{7}} = 2.973568.$$

Γνωρίζουμε ότι η μηδενική υπόθεση απορρίπτεται, αν $|t| \geq t_{n-1,\alpha/2}$. Είναι $t_{n-1,\alpha/2} = t_{6,0.025} = 2.447$ (Πίνακας Α΄.4, Παραρτήματος Α΄). Επομένως, καθώς η τιμή της σσε ανήκει στην κρίσιμη περιοχή, εξάγουμε το συμπέρασμα ότι απορρίπτεται η μηδενική υπόθεση σε επίπεδο σημαντικότητας 0.05. Επομένως, με βάση αυτά τα δεδομένα, σε επίπεδο σημαντικότητας 0.05, δεν μπορούμε να ισχυριστούμε ότι το μέσο σημείο τήξης ισούται με 1260.
Εναλλακτικά, χρησιμοποιώντας τις παρακάτω εντολές της R

```
1 data<-c(1267,1262,1267,1263,1258,1263,1268)
2 t.test(data,alternative = "two.sided",mu=1260,conf.level = 0.95)
```

λαμβάνουμε τα εξής αποτελέσματα:

```
        One Sample t-test

data:  data
t = 2.9736, df = 6, p-value = 0.02484
alternative hypothesis: true mean is not equal to 1260
95 percent confidence interval:
 1260.708 1267.292
sample estimates:
mean of x
     1264
```

Παρατηρούμε ότι η $p-$τιμή του ελέγχου ισούται με $0.02484 < 0.05$. Άρα απορρίπτεται η μηδενική υπόθεση και καταλήγουμε στο ίδιο συμπέρασμα με πριν.

Λύση Άσκησης Αυτοαξιολόγησης 11.5

Έστω X και Y οι τ.μ. που παριστάνουν τη θερμική ικανότητα του γαιάνθρακα που εξάγεται από το ορυχείο 1 και ορυχείο 2, αντίστοιχα. Από την εκφώνηση έχουμε ότι $X \sim N(\mu_1, \sigma_1^2)$ και $Y \sim N(\mu_2, \sigma_2^2)$ με $\sigma_1^2 = \sigma_2^2 = \sigma^2$. Θέλουμε να ελέγξουμε τις υποθέσεις

$$H_0 : \mu_1 - \mu_2 = 0 \text{ κατά της } H_1 : \mu_1 - \mu_2 \neq 0.$$

Καθώς οι πληθυσμοί είναι κανονικοί και οι πληθυσμιακές διασπορές είναι άγνωστες, αλλά ίσες, θα χρησιμοποιήσουμε τη σσε που δίνεται στη σχέση (11.7). Προκύπτει τότε ότι η τιμή της σσε είναι

$$t = \frac{8210 - 7990}{\sqrt{13200}\sqrt{1/5 + 1/5}} = 3.02765,$$

καθώς $\bar{x} = 8210$, $\bar{y} = 7990$, $s_1^2 = 15750$, $s_2^2 = 10650$, $n = 5$, $m = 5$ και

$$s_p^2 = \frac{(n-1)s_1^2 + (m-1)s_2^2}{n + m - 2} = \frac{4 \cdot 15750 + 4 \cdot 10650}{8} = 13200.$$

Γνωρίζουμε ότι η μηδενική υπόθεση απορρίπτεται, αν $|t| > -t_{n+m-2,\alpha/2} = t_{10,0.025} = 2.228$ (Πίνακας Α΄.4, Παραρτήματος Α΄). Επομένως, απορρίπτεται η μηδενική υπόθεση σε επίπεδο σημαντικότητας 0.05. Άρα με βάση αυτά τα δεδομένα, δεν μπορούμε να ισχυριστούμε ότι η μέση θερμική ικανότητα του γαιάνθρακα είναι η ίδια στα δύο ορυχεία.
Εναλλακτικά, χρησιμοποιώντας τις παρακάτω εντολές της R

```
1 a<-c(8240,8110,8330,8050,8320)
2 b<-c(7980,7920,7930,8170,7950)
3 t.test(a,b,alternative=c("two.sided"),mu=0,var.equal=TRUE)
```

λαμβάνουμε τα εξής αποτελέσματα:

```
Two Sample t-test

data:  a and b
t = 3.0277, df = 8, p-value = 0.01637
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  52.43742 387.56258
sample estimates:
mean of x mean of y
     8210      7990
```

Παρατηρούμε ότι η p–τιμή του ελέγχου ισούται με $0.01637 < 0.05$. Άρα απορρίπτεται η μηδενική υπόθεση και καταλήγουμε στο ίδιο συμπέρασμα με πριν.

Λύση Άσκησης Αυτοαξιολόγησης 11.6

Έστω X και Y οι τ.μ. που παριστάνουν τα φορτία θραύσης σύνθετων νημάτων τύπου I και II, αντίστοιχα. Από την εκφώνηση έχουμε ότι $X \sim N(\mu_1, \sigma_1^2)$ και $Y \sim N(\mu_2, \sigma_2^2)$ με $\sigma_1^2 = \sigma_2^2 = \sigma^2$. Θέλουμε να ελέγξουμε την

$$H_0 : \mu_1 - \mu_2 = 0 \text{ κατά της } H_1 : \mu_1 - \mu_2 \neq 0.$$

Καθώς οι πληθυσμοί είναι κανονικοί και οι πληθυσμιακές διασπορές είναι άγνωστες αλλά ίσες εξ

υποθέσεως, θα χρησιμοποιήσουμε τη σσε που δίνεται στη σχέση (11.7).
Από τα δεδομένα της εκφώνησης έχουμε ότι: $\bar{x} = 0.7285714$, $\bar{y} = 1.2$, $s_1^2 = 0.112381$, $s_2^2 = 0.1133333$, $n = 7$, $m = 7$ και

$$s_p^2 = \frac{(n-1)s_1^2 + (m-1)s_2^2}{n+m-2} = \frac{6 \cdot 0.112381 + 6 \cdot 0.1133333}{12} = 0.1128572.$$

Επομένως, προκύπτει ότι η τιμή της σσε σχέση (11.7) είναι

$$t = \frac{0.7285714 - 1.2}{\sqrt{0.1128572}\sqrt{1/7 + 1/7}} = -2.625339.$$

Γνωρίζουμε ότι η μηδενική υπόθεση απορρίπτεται, αν $|t| > -t_{n+m-2,\alpha/2} = t_{12,0.025} = 2.179$ (Πίνακας Α΄.4, Παραρτήματος Α΄). Επομένως, καθώς $|t| = 2.625339$, απορρίπτεται η μηδενική υπόθεση σε επίπεδο σημαντικότητας 0.05. Οπότε μπορούμε να ισχυριστούμε ότι το μέσο φορτίο θραύσης των δύο τύπων σύνθετων νημάτων δεν είναι το ίδιο.
Εναλλακτικά, χρησιμοποιώντας την R και τις εντολές

```
1 a1<-c(1.2,0.3,0.8,0.5,0.4,0.9,1.0)
2 b1<-c(1.4,1.5,1.1,1.0,0.8,1.7,0.9)
3 t.test(a1,b1,alternative=c("two.sided"),mu=0,var.equal=TRUE)
```

λαμβάνουμε τα εξής αποτελέσματα:

```
     Two Sample t-test

data:  a1 and b1
t = -2.6253, df = 12, p-value = 0.02217
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.86267504 -0.08018211
sample estimates:
mean of x mean of y
0.7285714 1.2000000
```

Παρατηρούμε ότι η $p-$τιμή του ελέγχου ισούται με $0.02217 < 0.05$, άρα απορρίπτουμε τη μηδενική υπόθεση και καταλήγουμε στο ίδιο συμπέρασμα με πριν.

Λύση Άσκησης Αυτοαξιολόγησης 11.7

Έστω X και Y οι τυχαίες μεταβλητές που παριστάνουν τη μέτρηση της ποσότητας με τη μέθοδο Α και Β με μέση τιμή μ_1 και μ_2, αντίστοιχα. Έχουμε διαθέσιμα τα δείγματα $x_1,\ldots,x_n$ και $y_1,\ldots,y_n$ με $n = 15$, τα οποία είναι εξαρτημένα, καθώς πρόκειται για μετρήσεις στους ίδιους ενήλικες άνδρες. Προκειμένου να ληφθεί υπόψη η εξάρτηση, ο έλεγχος υποθέσεων για τη διαφορά $\mu_\delta = \mu_1 - \mu_2$ στηρίζεται στις δειγματικές διαφορές $d_1 = x_1 - y_1, \ldots, d_n = x_n - y_n$. Οι διαφορές αυτές είναι

$$0.67,\ -0.19,\ 0.09,\ 0.19,\ 0.13,\ 0.40,\ 0.04,\ 0.10,\ 0.50,\ 0.07,\ 0.23,\ 0.59,\ 0.02,\ 0.03,\ 0.11,$$

με δειγματική μέση τιμή και διακύμανση ίση με $\bar{d} = 0.1986667$ και $s_d^2 = 0.05678381$ αντίστοιχα.
Θέλουμε να ελέγξουμε τις υποθέσεις

$$H_0 : \mu_\delta = 0 \text{ κατά } H_1 : \mu_\delta \neq 0.$$

Υποθέτοντας ότι οι διαφορές προέρχονται από κανονικό πληθυσμό θα χρησιμοποιήσουμε τη σσε

$$T = \frac{\bar{D} - 0}{S_D/\sqrt{n}},$$

με κπ $|T| > t_{n-1,\alpha/2}$. Η σσε παίρνει την τιμή

$$t = \frac{\bar{d} - 0}{s_d/\sqrt{n}} = \frac{0.1986667}{\sqrt{0.05678381}/\sqrt{15}} = 3.228929.$$

Έχουμε ότι $t_{14,0.025} = 2.145$ (Πίνακας Α΄.4, Παραρτήματος Α΄) και καθώς είναι $3.228929 > 2.145$ συμπεραίνουμε ότι απορρίπτουμε τη μηδενική υπόθεση σε επίπεδο σημαντικότητας 0.05. Επομένως, με βάση αυτά τα δεδομένα, δεν μπορούμε να ισχυριστούμε ότι οι δύο μέθοδοι μέτρησης είναι ισοδύναμες. Για να υλοποιήσουμε τον αντίστοιχο έλεγχο μέσω της R, χρησιμοποιούμε τις ακόλουθες εντολές:

```
1 data1<- c(1.94, 1.44, 1.56, 1.58, 2.06, 1.66, 1.75, 1.77, 1.78, 1.92, 1.25,
      1.93, 2.04, 1.62, 2.08)
2 data2 <- c(1.27, 1.63, 1.47, 1.39, 1.93, 1.26, 1.71, 1.67, 1.28, 1.85,
     1.02, 1.34, 2.02, 1.59, 1.97)
3 t.test(data1, data2, paired = TRUE)
```

και λαμβάνουμε τα παρακάτω αποτελέσματα:

```
Paired t-test

data:  data1 and data2
t = 3.2289, df = 14, p-value = 0.006062
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.0667041 0.3306292
sample estimates:
mean of the differences
              0.1986667
```

Από τα παραπάνω αποτελέσματα έχουμε ότι η p-τιμή του ελέγχου είναι $p - value = 0.006062 < 0.05$. Άρα συμπεραίνουμε ότι απορρίπτεται η μηδενική υπόθεση και σε επίπεδο σημαντικότητας 0.05 μπορούμε να ισχυριστούμε ότι υπάρχει στατιστικά σημαντική διαφοροποίηση μεταξύ των δύο μεθόδων μέτρησης.

Λύση Άσκησης Αυτοαξιολόγησης 11.8

Έστω X η τ.μ. που παριστάνει το βάρος της συσκευασίας των ξηρών καρπών. Είναι από την εκφώνηση $X \sim N(\mu, \sigma^2)$ και έχουμε διαθέσιμο το τυχαίο δείγμα $X_1, \ldots, X_n$ με $n = 10$ και δειγματική διακύμανση $s^2 = 2.14^2 = 4.5796$. Θέλουμε να ελέγξουμε σε επίπεδο σημαντικότητας $\alpha = 0.05$ τις υποθέσεις

$$H_0 : \sigma = 2 \text{ κατά } H_1 : \sigma > 2$$

ή, ισοδύναμα,

$$H_0 : \sigma^2 = 4 \text{ κατά } H_1 : \sigma^2 > 4.$$

Θα χρησιμοποιήσουμε για αυτόν τον έλεγχο τη στατιστική συνάρτηση της σχέσης (11.10). Η τιμή της

σσε ισούται με

$$\chi^2 = \frac{(10-1)\cdot s^2}{2^2} = \frac{9\cdot 4.5796}{4} = 10.3041.$$

Γνωρίζουμε ότι η μηδενική υπόθεση απορρίπτεται σε επίπεδο σημαντικότητας $\alpha = 0.05$, αν $\chi^2 > \chi^2_{n-1,\alpha} = \chi^2_{9,0.05} = 16.919$ Επομένως, δεν απορρίπτουμε τη μηδενική υπόθεση σε επίπεδο σημαντικότητας 0.05, δηλαδή με βάση αυτά τα δεδομένα δεν μπορούμε να ισχυριστούμε ότι η τυπική απόκλιση του βάρους των ξηρών καρπών είναι μεγαλύτερη από 2 γραμμάρια.
Εναλλακτικά, στην R εκτελούμε τις ακόλουθες εντολές

```
var.sample<-2.14^2
n<-10
test.value<-2^2

test.stat<-(n-1)*var.sample/(test.value)
alpha<-0.05

criticalvalue=qchisq(alpha, df=n-1, lower.tail = FALSE)

if (test.stat>criticalvalue){
  print("The null hypothesis is rejected")
} else {
  print("The null hypothesis is not  rejected")
}

pchisq(test.stat,n-1,lower.tail =F)
```

και λαμβάνουμε τα παρακάτω αποτελέσματα:

```
[1] "The null hypothesis is not  rejected"
```

```
[1] 0.3264322
```

Τα παραπάνω αποτελέσματα μας πληροφορούν ότι η μηδενική υπόθεση δεν απορρίπτεται σε επίπεδο σημαντικότητας 0.05 και ότι η p-τιμή του ελέγχου είναι ίση με 0.3264322.

Λύση Άσκησης Αυτοαξιολόγησης 11.9

Έστω X και Y οι τ.μ. που παριστάνουν την ποσότητα ορθοφωσφόρου στην περιοχή του σταθμού Α και Β, αντίστοιχα. Από την εκφώνηση έχουμε ότι $X \sim N(\mu_1,\sigma_1^2)$ και $Y \sim N(\mu_2,\sigma_2^2)$. Θέλουμε να ελέγξουμε την

$$H_0 : \frac{\sigma_2^2}{\sigma_1^2} = 1 \text{ κατά } H_1 : \frac{\sigma_2^2}{\sigma_1^2} \neq 1.$$

Η τιμή f της σσε F της σχέσης (11.11) είναι ίση με $f = \frac{s_1^2}{s_2^2} = \frac{9.4249}{0.64} = 14.72641$. Σε επίπεδο σημαντικότητας α απορρίπτεται η μηδενική υπόθεση, αν

$$f < F_{n-1,m-1,1-\alpha/2} = F_{15,12,0.95} = \frac{1}{F_{12,15,0.05}} = \frac{1}{2.48} = 0.4032258,$$

ή αν

$$f > F_{n-1,m-1,\alpha/2} = F_{15,12,0.05} = 2.62.$$

Επομένως, σε επίπεδο σημαντικότητας 0.1 απορρίπτεται η υπόθεση της ισότητας των διασπορών των συγκεντρώσεων ορθοφωσφόρου στους δύο σταθμούς.

Εναλλακτικά, στην R εκτελούμε τις ακόλουθες εντολές

```
nA<-16
var.sampleB<-0.8^2
nB<-13
test.value<-1
test.stat<-(var.sampleA/var.sampleB)*test.value
alpha<-0.10
crit1=qf(1-alpha/2, nA-1, nB-1, lower.tail = FALSE)
crit2=qf(alpha/2, nA-1, nB-1, lower.tail = FALSE)
crit2<-qf(0.10/2, 4, 6, lower.tail = FALSE)
if (test.stat<crit1 | test.stat>crit2){
   print("The null hypothesis is rejected")
} else {
   print("The null hypothesis is not  rejected")
}
pf(test.stat,nA-1,nB-1,lower.tail =F)
```

και λαμβάνουμε τα παρακάτω αποτελέσματα:

```
[1] "The null hypothesis is rejected"

[1] 1.772651e-05
```

Τα παραπάνω αποτελέσματα μας πληροφορούν ότι η μηδενική υπόθεση απορρίπτεται σε επίπεδο σημαντικότητας 0.1 και ότι η p-τιμή του ελέγχου είναι ίση με 0.00018.

Λύση Άσκησης Αυτοαξιολόγησης 11.10

Έστω X_i, $i = 1, \ldots, 600$, η τ.μ. που παριστάνει τη γνώμη του i-οστού ατόμου για την κατασκευή του φράγματος με $X_i = 1$, αν το i–οστό άτομο είναι υπέρ της κατασκευής του φράγματος και $X_i = 0$ διαφορετικά με $P(X_i = 1) = p$. Ο έλεγχος που πρέπει να κάνουμε είναι ο

$$H_0 : p = 0.5 \text{ κατά } H_1 : p < 0.5.$$

Καθώς το μέγεθος δείγματος είναι μεγάλο, θα χρησιμοποιήσουμε τη σσε που δίνεται στη σχέση (11.13). Η τιμή της σσε είναι

$$z = \frac{\hat{p} - 0.5}{\sqrt{\frac{0.5(1-0.5)}{600}}} = \frac{0.4583333 - 0.5}{\sqrt{\frac{0.5(1-0.5)}{600}}} = -2.041243$$

καθώς $\hat{p} = \frac{275}{600}$. Η κπ του ελέγχου είναι $z < -z_\alpha = -z_{0.05} = -1.645$. Επομένως, με βάση αυτά τα δεδομένα απορρίπτεται η μηδενική υπόθεση, δηλαδή σε επίπεδο σημαντικότητας 0.05 ευσταθεί ο ισχυρισμός ότι λιγότεροι από τους μισούς κατοίκους υποστηρίζουν την κατασκευή του φράγματος.
Εναλλακτικά, θα μπορούσαμε να χρησιμοποιήσουμε την ακόλουθη εντολή στην R

```
prop.test(275, 600, p =0.5, alternative = "two.sided",correct = FALSE)
```

και λαμβάνουμε τα παρακάτω αποτελέσματα:

```
  1-sample proportions test without continuity correction

data:  275 out of 600, null probability 0.5
X-squared = 4.1667, df = 1, p-value = 0.04123
```

```
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.4188561 0.4983407
sample estimates:
        p
0.4583333
```

Τα παραπάνω αποτελέσματα μας πληροφορούν ότι η p-τιμή του ελέγχου είναι ίση με 0.04123. Σε αυτό το σημείο αξίζει να επισημανθεί ότι ο έλεγχος που διεξάγει η παραπάνω εντολή χρησιμοποιεί ως στατιστική συνάρτηση ελέγχου το τετράγωνο της σσε που δίνεται στη σχέση (11.13), η οποία προσεγγιστικά ακολουθεί, υπό τη μηδενική υπόθεση, χι-τετράγωνο κατανομή με 1 βαθμό ελευθερίας.

Λύση Άσκησης Αυτοαξιολόγησης 11.11

Η μηδενική και η εναλλακτική υπόθεση που πρέπει να ελεγχθούν είναι οι εξής:

$$H_0 : p_1 = p_2 \text{ κατά } H_1 : p_1 \neq p_2$$

ή, ισοδύναμα,

$$H_0 : p_1 - p_2 = 0 \text{ κατά } H_1 : p_1 - p_2 \neq 0$$

όπου p_1 και p_2 το πραγματικό ποσοστό των ελαττωματικών εξαρτημάτων των δύο μηχανών. Επειδή $\delta = 0$, θα χρησιμοποιήσουμε τη στατιστική συνάρτηση

$$Z = \frac{\hat{P}_1 - \hat{P}_2}{\sqrt{\hat{P}(1-\hat{P})(1/n + 1/m)}}.$$

Από τα δεδομένα της εκφώνησης είναι $\hat{p}_1 = 17/450 = 0.03777778$, $\hat{p}_2 = 12/400 = 0.03$ και $\hat{p} = \frac{17+12}{400+450} = \frac{29}{850} = 0.03411765$. Επομένως, η τιμή της σσε είναι

$$z = \frac{0.03777778 - 0.03}{\sqrt{0.03411765(1-0.03411765)} \cdot (1/400 + 1/450)} = 0.6234918.$$

Η μηδενική υπόθεση απορρίπτεται σε επίπεδο σημαντικότητας 0.05, αν $|z| > z_{\alpha/2} = z_{0.025} = 1.96$. Επομένως, δεν απορρίπτεται η μηδενική υπόθεση σε επίπεδο σημαντικότητας 0.05. Με βάση αυτά τα δεδομένα μπορούμε να ισχυριστούμε ότι δεν υπάρχει στατιστικά σημαντική διαφορά στο πραγματικό ποσοστό ελαττωματικών εξαρτημάτων των δύο μηχανών.
Εναλλακτικά, θα μπορούσαμε να χρησιμοποιήσουμε την ακόλουθη εντολή στην R

```
x<-c(17,12)
n<-c(450,400)
prop.test(x, n,  alternative = "two.sided",correct =F)
```

και λαμβάνουμε τα παρακάτω αποτελέσματα:

```
        2-sample test for equality of proportions without continuity
         correction

data:  x out of n
X-squared = 0.38874, df = 1, p-value = 0.533
alternative hypothesis: two.sided
95 percent confidence interval:
```

```
 -0.01650752  0.03206308
sample estimates:
    prop 1     prop 2
0.03777778 0.03000000
```

Τα παραπάνω αποτελέσματα μας πληροφορούν ότι η p-τιμή του ελέγχου είναι ίση με 0.533. Σε αυτό το σημείο αξίζει να επισημανθεί ότι ο έλεγχος που διεξάγει η παραπάνω εντολή χρησιμοποιεί ως στατιστική συνάρτηση ελέγχου το τετράγωνο της σσε Z, η οποία προσεγγιστικά ακολουθεί, υπό τη μηδενική υπόθεση, χι-τετράγωνο κατανομή με 1 βαθμό ελευθερίας.

ΒΙΒΛΙΟΓΡΑΦΙΑ

Ελληνόγλωσση

Κουτρουβέλης, Ι. Α. (2000). *Βασικά Εργαλεία και Μέθοδοι για τον Έλεγχο Ποιότητας: Πιθανότητες και Στατιστική ΙΙ (Τόμος Β΄)*. Πάτρα: Ελληνικό Ανοικτό Πανεπιστήμιο.

Ξενόγλωσση

Bethea, R., Duran, B. and Boullion, T. (1995). *Statistical methods for engineers and scientists*. New York: Marcel Dekker, Inc.

Neyman, J. and Pearson, E. S. (1933). On the Problem of the Most Efficient Tests of Statistical Hypotheses. *Philosophical Transactions of the Royal Society of London. Series A*, *231*, pp. 289–337.

Sprent, P. and Smeeton, N. (2016). *Applied Nonparametric Statistical Methods*. Chapman & Hall/CRC Texts in Statistical Science. CRC Press.

ΚΕΦΑΛΑΙΟ 12

ΑΠΛΗ ΓΡΑΜΜΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ

Σύνοψη

Σε αυτό το κεφάλαιο παρουσιάζεται το απλό γραμμικό μοντέλο. Ιδιαίτερη έμφαση δίνεται στη διαδικασία υπολογισμού διάφορων απαραίτητων ποσοτήτων για την εξαγωγή συμπερασμάτων σχετικά με τη γραμμική σχέση δύο μεταβλητών. Τέλος, το κεφάλαιο ολοκληρώνεται με τη χρήση της R για την ανάλυση ενός πραγματικού συνόλου δεδομένων με τη βοήθεια του απλού γραμμικού μοντέλου.

Προαπαιτούμενη γνώση: Κεφάλαια 5, 10 και 11 του παρόντος συγγράμματος.

Προσδοκώμενα μαθησιακά αποτελέσματα: Όταν θα έχετε ολοκληρώσει τη μελέτη του κεφαλαίου αυτού, θα

- έχετε κατανοήσει την έννοια του απλού γραμμικού μοντέλου,
- μπορείτε να εκτιμήσετε και να ερμηνεύσετε την εκτιμώμενη ευθεία παλινδρόμησης,
- μπορείτε να εφαρμόζετε στατιστικούς ελέγχους και να υπολογίζετε χαρακτηριστικά μεγέθη για την αξιολόγηση της εκτιμώμενης ευθείας παλινδρόμησης και
- μπορείτε να χρησιμοποιείτε την R για την προσαρμογή του απλού γραμμικού μοντέλου.

Οικονόμου, Π., Μαλεφάκη, Σ., & Μπατσίδης, Α. (2023). *Πιθανότητες - Στατιστική.*
[Προπτυχιακό εγχειρίδιο].

Γλωσσάριο επιστημονικών όρων

- Άθροισμα τετραγώνων από την παλινδρόμηση
- Άθροισμα τετραγώνων των υπολοίπων
- Ανάλυση διασποράς
- Διάγραμμα διασκόρπισης
- Διαστήματα εμπιστοσύνης και πρόβλεψης
- Έλεγχος σημαντικότητας του μοντέλου
- Εκτιμήτριες ελαχίστων τετραγώνων
- Εξίσωση παλινδρόμησης
- Θεώρημα Gauss–Markov
- Ολικό άθροισμα τετραγώνων
- Συντελεστής γραμμικής συσχέτισης του Pearson
- Συντελεστής προσδιορισμού
- Τυχαία σφάλματα
- Υπόλοιπα

12.1 Εισαγωγή

Συχνά δύο ή περισσότερες μεταβλητές σχετίζονται μεταξύ τους και η τιμή της μιας επηρεάζεται και καθορίζεται ως έναν βαθμό από τις τιμές της άλλης ή των άλλων, αντίστοιχα. Χαρακτηριστικά παραδείγματα περιλαμβάνουν, μεταξύ άλλων,

- την απόδοση μιας χημικής αντίδρασης σε σχέση με τις αρχικές ποσότητες των αντιδρώντων,
- το βάρος ενός νεογέννητου σε σχέση με το φύλο, το ύψος και τις εβδομάδες που ολοκληρώθηκε η κύηση,
- τα κέρδη μιας επιχείρησης σε σχέση με το ποσό που διαθέτει για διαφήμιση σε διάφορα μέσα κοινωνικής δικτύωσης.

Βασικό εργαλείο για την περιγραφή των παραπάνω αλλά και παρόμοιων φαινομένων αποτελούν τα στατιστικά μοντέλα παλινδρόμησης (regression models). Με τον όρο «μοντέλο» εννοούμε τη μορφή της σχέσης, δηλαδή της εξίσωσης, που συνδέει δύο ή περισσότερες μεταβλητές. Στα στατιστικά μοντέλα η προαναφερθείσα σχέση είναι εν μέρει στοχαστική. Παραδείγματος χάριν, αν υποθέσουμε ότι έχουμε μόνο δύο μεταβλητές, έστω τις X και Y, τότε η σχέση τους μπορεί να είναι της μορφής:

$$Y_{|X=x} = f(x) + \varepsilon,$$

όπου ε είναι ένα τυχαίο σφάλμα (random error) με μέση τιμή μηδέν και (άγνωστη) διασπορά σ^2, $f(x)$ μια γνωστή συνάρτηση των τιμών της X και $Y_{|X=x}$ η δεσμευμένη τυχαία μεταβλητή της Y, όταν $X = x$.

Υπό την παραπάνω υπόθεση για την αναμενόμενη τιμή της τυχαίας μεταβλητής ε, δηλαδή υπό την υπόθεση ότι $E(\varepsilon) = 0$, εύκολα καταλήγουμε στη σχέση:

$$E(Y|X = x) = f(x).$$

Η παραπάνω σχέση αποτυπώνει το γεγονός ότι η συνάρτηση $f(x)$ εκφράζει τη μεταβολή της μέσης τιμής της τυχαίας μεταβλητής Y σε σχέση με τις τιμές της τυχαίας μεταβλητής X και δεν προσδιορίζει επακριβώς τις παρατηρούμενες τιμές της.

Παρατήρηση 12.1

Από τα παραπάνω είναι φανερό ότι τα στατιστικά μοντέλα βασίζονται σε συναρτησιακές σχέσεις, οι οποίες εκφράζουν την αναμενόμενη τιμή και όχι την ακριβή παρατηρούμενη τιμή μιας τυχαίας μεταβλητής. Επομένως, ένα στατιστικό μοντέλο δεν είναι και δεν πρέπει να ερμηνεύεται ως μια συναρτησιακή σχέση μεταξύ δύο ή περισσότερων μεταβλητών.

Παρατήρηση 12.2

Ένα ακόμα χαρακτηριστικό που διαφοροποιεί τα στατιστικά μοντέλα από τα μοντέλα που συναντάμε στη Φυσική και σε άλλες επιστήμες είναι ότι:

- οι συναρτησιακές σχέσεις των άλλων επιστημών προκύπτουν από γνώση του μηχανισμού της επίδρασης της X επάνω στην Y και, επομένως, η Y εξαρτάται και επηρεάζεται άμεσα από τη X, ενώ
- στα στατιστικά μοντέλα μπορεί να μην υπάρξει καν τέτοια εξάρτηση.

Παραδείγματος χάριν, είναι κοινά αποδεκτό ότι στην κατάσβεση μεγάλων πυρκαγιών συμμετέχει συνήθως μεγαλύτερο πλήθος πυροσβεστών από ότι στις μικρότερες. Επομένως, υπάρχει μια ισχυρή θετική συσχέτιση μεταξύ του μεγέθους της φωτιάς και του αριθμού των πυροσβεστών. Άρα, κάποιος θα μπορούσε να εκτιμήσει το μέγεθος της φωτιάς από τον αριθμό των πυροσβεστών που εμπλέκονται στην κατάσβεσή της και να καταλήξει στο συμπέρασμα ότι όσο περισσότεροι πυροσβέστες εμπλέκονται στην

επιχείρηση κατάσβεσης, τόσο μεγαλύτερη είναι η φωτιά. Παρ' όλα αυτά, είναι σαφές ότι δεν μπορούμε να ισχυριστούμε ότι το μεγαλύτερο πλήθος πυροσβεστών προκαλεί μεγαλύτερες φωτιές. Η παραπάνω παρατήρηση αναφέρεται συχνά ως ότι **η συσχέτιση δεν σημαίνει αιτιότητα (correlation does not imply causation)**.

12.1.1 Γραμμικά Μοντέλα Παλινδρόμησης

Ένα από τα βασικά παραδείγματα στατιστικών μοντέλων είναι το αποκαλούμενο απλό γραμμικό μοντέλο, το οποίο εκφράζεται μέσω της σχέσης:

$$Y_{|X=x} = \beta_0 + \beta_1 x + \varepsilon.$$

Είναι φανερό ότι για το απλό γραμμικό μοντέλο η $f(x)$ δίνεται από τη σχέση:

$$f(x) = \beta_0 + \beta_1 x,$$

όπου τα β_0 και β_1 είναι οι άγνωστες παράμετροι του μοντέλου.

Γενίκευση του απλού γραμμικού μοντέλου αποτελεί το πολλαπλό γραμμικό μοντέλο, το οποίο εκφράζεται μέσω της σχέσης:

$$Y_{|\boldsymbol{X}=\boldsymbol{x}} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon,$$

και το οποίο δηλώνει τη σχέση μεταξύ της τυχαίας μεταβλητής Y και των μεταβλητών $\boldsymbol{X} = (X_1, X_2, \ldots, X_k)$.

Στα παραπάνω μοντέλα

- η μεταβλητή Y ονομάζεται **εξαρτημένη** μεταβλητή ή **μεταβλητή απόκρισης**, και
- οι μεταβλητές $X_1, X_2, \ldots, X_k$ (ή απλώς η X για το απλό γραμμικό μοντέλο) ονομάζονται **ανεξάρτητες** ή **επεξηγηματικές** ή **προβλέπουσες** μεταβλητές ή απλώς **συμμεταβλητές**.

Το γραμμικό μοντέλο καλείται γραμμικό όχι επειδή η Y είναι γραμμική συνάρτηση των τιμών $x_1, \ldots, x_k$ των μεταβλητών $X_1, \ldots, X_k$, αλλά επειδή οι προαναφερθείσες εξισώσεις είναι γραμμικές ως προς τις άγνωστες παραμέτρους του μοντέλου, δηλαδή τα $\beta_0, \beta_1, \ldots, \beta_k$. Αυτό σημαίνει ότι η συνάρτηση $f(x)$ θα μπορούσε να είναι μη γραμμική ως προς τα $\boldsymbol{X}$, αλλά το μοντέλο να χαρακτηρίζεται ως γραμμικό, αν όντως είναι γραμμικό ως προς τους συντελεστές β_i. Για παράδειγμα, οι ακόλουθες σχέσεις θεωρούνται γραμμικές

$$Y_{|X=x} = \beta_0 + \beta_1 e^{-x} + \varepsilon,$$

$$Y_{|X=x} = \beta_0 + \beta_1 \frac{1}{x} + \varepsilon,$$

καθώς τα μοντέλα αυτά είναι γραμμικά ως προς τις άγνωστες παραμέτρους β_0, και β_1.

Στη συνέχεια του κεφαλαίου θα επικεντρωθούμε στο απλό γραμμικό μοντέλο. Για το πολλαπλό γραμμικό μοντέλο ο/η αναγνώστης/στρια παραπέμπεται ενδεικτικά στα βιβλία των Καρώνη και Οικονόμου (2017) και Draper and Smith (2014).

12.2 Απλό γραμμικό μοντέλο

Το απλό γραμμικό μοντέλο περιλαμβάνει, όπως προαναφέρθηκε, δύο μεταβλητές, την ανεξάρτητη μεταβλητή X και την εξαρτημένη μεταβλητή Y, οι οποίες συνδέονται μεταξύ τους μέσω της σχέσης:

$$Y_{|X=x} = \beta_0 + \beta_1 x + \varepsilon,$$

όπου ε είναι μια τυχαία μεταβλητή με μέση τιμή μηδέν και (άγνωστη) διασπορά σ^2, η οποία εκφράζει το τυχαίο σφάλμα, ενώ με β_0 και β_1, όπως έχει προαναφερθεί, συμβολίζονται οι άγνωστες παράμετροι του γραμμικού μοντέλου ή, αλλιώς, οι συντελεστές παλινδρόμησης. Η παραπάνω σχέση εκφράζεται ισοδύναμα ως:

$$E(Y|X = x) = E(Y_x) = \beta_0 + \beta_1 x = \mu_x,$$

η οποία αποτελεί και την αποκαλούμενη εξίσωση παλινδρόμησης.

Ο προσδιορισμός ή, καλύτερα, η εκτίμηση των άγνωστων παραμέτρων του μοντέλου γίνεται με τη βοήθεια ενός τυχαίου δείγματος n το πλήθος ανεξάρτητων παρατηρήσεων (x_i, y_i), $i = 1, 2, \dots, n$, οι τιμές των οποίων συνδέονται μεταξύ τους μέσω της σχέσης:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

όπου ε_i το τυχαίο σφάλμα που αντιστοιχεί στην i-οστή παρατήρηση. Τα τυχαία σφάλματα αποτελούν βασικό στοιχείο του μοντέλου και διαδραματίζουν, όπως θα δούμε, κεντρικό ρόλο στη συμπερασματολογία του. Για τον λόγο αυτό στην επόμενη υποενότητα παρουσιάζονται αναλυτικά οι υποθέσεις που υιοθετούνται για τα τυχαία σφάλματα στο απλό γραμμικό μοντέλο αλλά και γενικότερα στα γραμμικά μοντέλα.

12.2.1 Τυχαία σφάλματα

Όπως αναφέρθηκε παραπάνω οι ποσότητες ε_i ονομάζονται τυχαία σφάλματα και παριστάνουν για δοθείσα τιμή x_i την κατακόρυφη απόκλιση της τιμής y_i από την (άγνωστη) ευθεία της συνάρτησης παλινδρόμησης, όπως φαίνεται και στο Σχήμα 12.1[1].

Για τα τυχαία σφάλματα υποθέτουμε ότι ικανοποιούν τις ακόλουθες υποθέσεις:

- $E(\varepsilon_i) = 0$, για κάθε $i = 1, \dots, n$,
- $Var(\varepsilon_i) = \sigma^2$, για κάθε $i = 1, \dots, n$,
- $Cov(\varepsilon_i, \varepsilon_j) = 0$, για $i, j = 1, \dots, n$ με $i \neq j$.

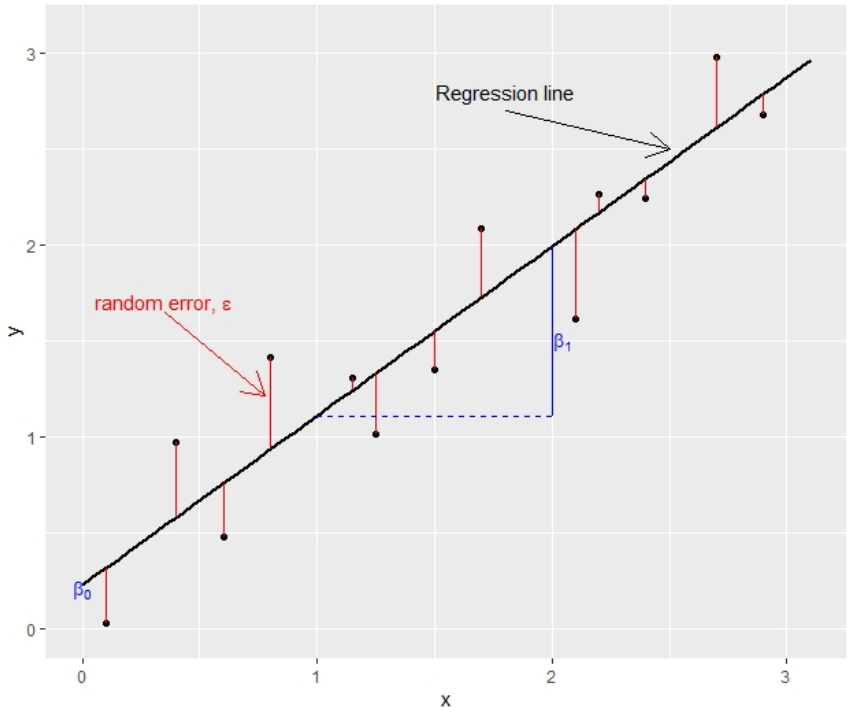
Επίσης, υποθέτουμε ότι η κατανομή των τυχαίων σφαλμάτων ε_i είναι η κανονική. Επομένως, λαμβάνοντας υπόψη και τις παραπάνω υποθέσεις έχουμε ότι $\varepsilon_i \sim N(0, \sigma^2)$, $i = 1, \dots, n$, ανεξάρτητες και ισόνομες τυχαίες μεταβλητές. Εναλλακτικά, όλες οι παραπάνω υποθέσεις μπορούν να εκφραστούν μέσω του τυχαίου διανύσματος $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)^t$ ως:

$$\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma^2 \boldsymbol{I}),$$

όπου με $N_n(\mathbf{0}, \sigma^2 \boldsymbol{I})$ συμβολίζουμε την n-διάστατη κανονική κατανομή με αναμενόμενη τιμή $E(\boldsymbol{\varepsilon}) = \mathbf{0}$ και πίνακα διασποράς-συνδιασποράς

$$Var(\boldsymbol{\varepsilon}) = E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^t) = \sigma^2 \boldsymbol{I},$$

[1] Στο Σχήμα 12.1 αποτυπώνεται και η φυσική ερμηνεία των συντελεστών β_0 και β_1, η οποία όμως θα παρουσιαστεί στη συνέχεια. Για το β_1 ιδιαίτερη έμφαση πρέπει να δοθεί στο γεγονός ότι αντιστοιχεί στη μεταβολή που παρατηρείται στην αναμενόμενη τιμή της εξαρτημένης μεταβλητής όταν η τιμή της ανεξάρτητης μεταβλητής αυξηθεί κατά μία μονάδα.



Σχήμα 12.1: Το διάγραμμα διασκόρπισης ενός συνόλου δεδομένων, η εξίσωση παλινδρόμησης και τα τυχαία σφάλματα.

όπου $\boldsymbol{I}$ ο μοναδιαίος $n \times n$ πίνακας.

Λόγω της σχέσης

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, i = 1 \dots, n,$$

οι παραπάνω υποθέσεις για τα τυχαία σφάλματα μπορούν να εκφραστούν ισοδύναμα για τις τυχαίες μεταβλητές Y_i, τις τιμές y_i των οποίων παρατηρούμε σε ένα τυχαίο δείγμα:

- $E(Y_i) = \beta_0 + \beta_1 x_i$, για κάθε $i = 1, \dots, n$,
- $Var(Y_i) = \sigma^2$, για κάθε $i = 1, \dots, n$,
- $Cov\left(Y_i, Y_j\right) = 0$, για $i, j = 1, \dots, n$ με $i \neq j$.

Επομένως, υπό την πρόσθετη υπόθεση της κανονικότητας, μπορούμε να καταλήξουμε ότι

$$\boldsymbol{Y} \sim N_n(\beta_0 1_n + \beta_1 \boldsymbol{x}, \sigma^2 \boldsymbol{I}),$$

όπου $1_n = (1, \dots, 1)^t$ το n-διάστατο διάνυσμα με όλες τις συνιστώσες μονάδα, $\boldsymbol{Y} = (Y_1, \dots, Y_n)^t$ και $\boldsymbol{x} = (x_1, \dots, x_n)^t$.

12.2.2 Εκτίμηση παραμέτρων - Μέθοδος ελαχίστων τετραγώνων

Η εκτίμηση των παραμέτρων, β_0 και β_1, της εξίσωσης παλινδρόμησης του απλού γραμμικού μοντέλου βασίζεται στη χρήση της μεθόδου ελαχίστων τετραγώνων. Η μέθοδος ελαχίστων τετραγώνων χρησιμοποιείται για τον προσδιορισμό της εξίσωσης εκείνης με την καλύτερη προσαρμογή στις n το πλήθος ανεξάρτητες παρατηρήσεις $\left(x_i, y_i\right)$, $i = 1, \dots, n$ ενός τυχαίου δείγματος.

Η μέθοδος ελαχίστων τετραγώνων έγκειται στην ελαχιστοποίηση της συνάρτησης

$$S\left(\beta_0,\beta_1\right)=\sum_{i=1}^{n}\left(y_i-E(Y_i)\right)^2=\sum_{i=1}^{n}\left(y_i-\beta_0-\beta_1x_i\right)^2$$

ως προς τις άγνωστες παραμέτρους β_0 και β_1. Η ιδέα της μεθόδου ελαχίστων τετραγώνων εδράζεται στην επιθυμία του προσδιορισμού της ευθείας που περνά (κατά μέσο όρο) πιο κοντά από όλα τα παρατηρηθέντα σημεία $\left(x_i,y_i\right)$, $i=1,\ldots,n$. Επομένως, η επιλογή μιας συνάρτησης που εμπλέκει τις διαφορές $\left(y_i-E(y_i)\right)$ μοιάζει φυσιολογική. Οι διαφορές αυτές, οι οποίες εκφράζουν τις κατακόρυφες αποστάσεις των σημείων από την ευθεία, μπορεί να είναι θετικές ή αρνητικές με συνέπεια να αλληλοαναιρούνται. Για τον λόγο αυτό η μέθοδος δεν βασίζεται στην ελαχιστοποίηση των διαφορών, αλλά στην ελαχιστοποίηση των τετραγώνων των διαφορών. Ο βασικός λόγος της επιλογής των τετραγώνων για τον χειρισμό των αρνητικών ποσοτήτων, έναντι κυρίως της έτερης δημοφιλούς επιλογής, δηλαδή της απόλυτης τιμής, είναι ότι οι τετραγωνικές συναρτήσεις είναι παντού παραγωγίσιμες - μια ιδιαίτερα χρήσιμη ιδιότητα όταν θέλουμε να προσδιορίσουμε το ελάχιστο - ενώ οι συναρτήσεις που εμπλέκουν απόλυτες τιμές παρουσιάζουν σημεία όπου η παράγωγός τους δεν υπάρχει.

Για την ελαχιστοποίηση της $S\left(\beta_0,\beta_1\right)$

- παραγωγίζουμε αρχικά τη συνάρτηση $S\left(\beta_0,\beta_1\right)$ ως προς β_0 και ως προς β_1,
- στη συνέχεια εξισώνουμε τις μερικές παραγώγους με το μηδέν και
- επιλύουμε το 2×2 σύστημα εξισώσεων που προκύπτει.

Ειδικότερα, προκύπτει το ακόλουθο σύστημα εξισώσεων (γνωστό και ως σύστημα των κανονικών εξισώσεων):

$$\sum_{i=1}^{n}(y_i-\beta_0-\beta_1\cdot x_i)=0$$

$$\sum_{i=1}^{n}(y_i-\beta_0-\beta_1\cdot x_i)x_i=0.$$

Από την επίλυση του συστήματος λαμβάνουμε τις εκτιμήτριες ελαχίστων τετραγώνων των β_0 και β_1, οι οποίες συμβολίζονται $\hat{\beta}_0$ και $\hat{\beta}_1$, αντίστοιχα, και προσδιορίζονται από τις σχέσεις:

$$\hat{\beta}_1=\frac{\sum_{i=1}^{n}x_iy_i-n\bar{x}\bar{y}}{\sum_{i=1}^{n}x_i^2-n\bar{x}^2}=\frac{\sum_{i=1}^{n}x_iy_i-\frac{\sum_{i=1}^{n}x_i\sum_{i=1}^{n}y_i}{n}}{\sum_{i=1}^{n}x_i^2-\frac{\left(\sum_{i=1}^{n}x_i\right)^2}{n}}=\frac{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)\left(y_i-\bar{y}\right)}{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2},\tag{12.1}$$

και

$$\hat{\beta}_0=\bar{y}-\hat{\beta}_1\bar{x}.\tag{12.2}$$

Η σχέση (12.1) συντομότερα γράφεται και $\hat{\beta}_1=\frac{s_{xy}}{s_{xx}}$, όπου

$$s_{xy}=\sum_{i=1}^{n}\left(x_i-\bar{x}\right)\left(y_i-\bar{y}\right)$$

$$s_{xx}=\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2.$$

Κατά αυτόν τον τρόπο προσδιορίζεται η εκτιμώμενη ή προσαρμοσμένη ευθεία παλινδρόμησης που δίνεται από τη σχέση

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x.$$

Παρατήρηση 12.3

Το ότι η συνάρτηση $S\left(\beta_0, \beta_1\right)$ λαμβάνει την ελάχιστη τιμή της στο $(\hat{\beta}_0, \hat{\beta}_1)$ προκύπτει από το γεγονός ότι ο πίνακας των δεύτερων παραγώγων της S είναι θετικά ορισμένος σε αυτό το σημείο. Σημειώνεται ότι ένας 2×2 πίνακας είναι θετικά ορισμένος, όταν το στοιχείο της πρώτης γραμμής και πρώτης στήλης του είναι θετικό, καθώς και η ορίζουσά του είναι θετική. Τα παραπάνω μπορούν εύκολα να διαπιστωθούν ότι ισχύουν για το απλό γραμμικό μοντέλο.

12.2.3 Η προσαρμοσμένη ευθεία παλινδρόμησης

Από την προσαρμογή του μοντέλου προκύπτει η προσαρμοσμένη ευθεία $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$, η οποία διέρχεται από τα σημεία $\left(x_i, \hat{y}_i\right)$, $i = 1, 2, \ldots, n$, όπου

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i,$$

είναι η εκτίμηση της αναμενόμενης τιμής της μεταβλητής Y για $X = x_i$. Η εκτιμώμενη ευθεία παλινδρόμησης εκφράζει την πίστη μας για την αναμενόμενη τιμή της τ.μ. Y, όταν η η ανεξάρτητη τ.μ. X λάβει μια συγκεκριμένη τιμή. Από την άλλη,

- η εκτίμηση $\hat{\beta}_1$ εκφράζει την αναμενόμενη μεταβολή της εξαρτημένης μεταβλητής Y σε κάθε μοναδιαία αύξηση της ανεξάρτητης μεταβλητής X, ενώ
- η εκτίμηση $\hat{\beta}_0$ εκφράζει την αναμενόμενη τιμή της εξαρτημένης μεταβλητής Y, όταν $X = 0$.

Παρατήρηση 12.4

Ιδιαίτερη προσοχή πρέπει να δίνεται στην ερμηνεία της $\hat{\beta}_0$, όταν το μηδέν δεν ανήκει στο εύρος των τιμών της X, αφού στις περιπτώσεις αυτές η ερμηνεία του $\hat{\beta}_0$ δεν έχει νόημα.

Παράδειγμα 12.1

Η απόδοση των μηχανισμών εκρηκτικών υλών που χρησιμοποιούνται για τη διάνοιξη σηράγγων σε ευσταθή πετρώματα εξαρτάται από τη δύναμη συνοχής Y μεταξύ του προωθητικού εκρηκτικού αερίου και του αερίου συντήρησης που περιέχει. Εικάζεται ότι η δύναμη αυτή εξαρτάται από την ηλικία X (σε εβδομάδες) του αερίου συντήρησης. Για να διαπιστωθεί αυτή η εικασία, μετρήθηκαν η δύναμη αυτή και η αντίστοιχη ηλικία του αερίου συντήρησης σε 20 μηχανισμούς εκρηκτικών υλών. Στη συνέχεια, δίνονται τα συγκεντρωτικά στοιχεία των παρατηρήσεων.

$$\sum_{i=1}^{20} x_i = 267.25 \qquad \sum_{i=1}^{20} y_i = 42627.15$$

$$\sum_{i=1}^{20} x_i^2 = 4677.69 \qquad \sum_{i=1}^{20} x_i y_i = 528492.64 \qquad \sum_{i=1}^{20} y_i^2 = 92547433.46.$$

Να προσαρμοστεί το απλό γραμμικό μοντέλο στα δεδομένα και να ερμηνευτεί η εκτιμώμενη ευθεία παλινδρόμησης με τους όρους του προβλήματος.

Λύση Παραδείγματος 12.1

Από τις σχέσεις (12.1) και (12.2) έχουμε ότι:

$$\begin{aligned}\hat{\beta}_1 &= \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} \\ &= \frac{528492.64 - \frac{267.25 \cdot 42627.15}{20}}{4677.69 - \frac{267.25^2}{20}} \\ &= \frac{-41112.65}{1106.56} \\ &= -37.1536,\end{aligned}$$

και

$$\begin{aligned}\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \\ &= \frac{42627.15}{20} - (-37.1536)\frac{267.25}{20} \\ &= 2131.3575 + 37.1536 \cdot 13.3625 \\ &= 2627.8225.\end{aligned}$$

Επομένως, η εκτιμώμενη ευθεία παλινδρόμησης είναι η

$$\hat{y} = 2627.8225 - 37.1536x,$$

η οποία εκφράζει την εκτίμηση της αναμενόμενης τιμής της δύναμης συνοχής Y μεταξύ του προωθητικού εκρηκτικού αερίου και του αερίου συντήρησης που περιέχει ο εκρηκτικός μηχανισμός σε σχέση με την ηλικία X (σε εβδομάδες) του αερίου συντήρησης. Επιπροσθέτως,

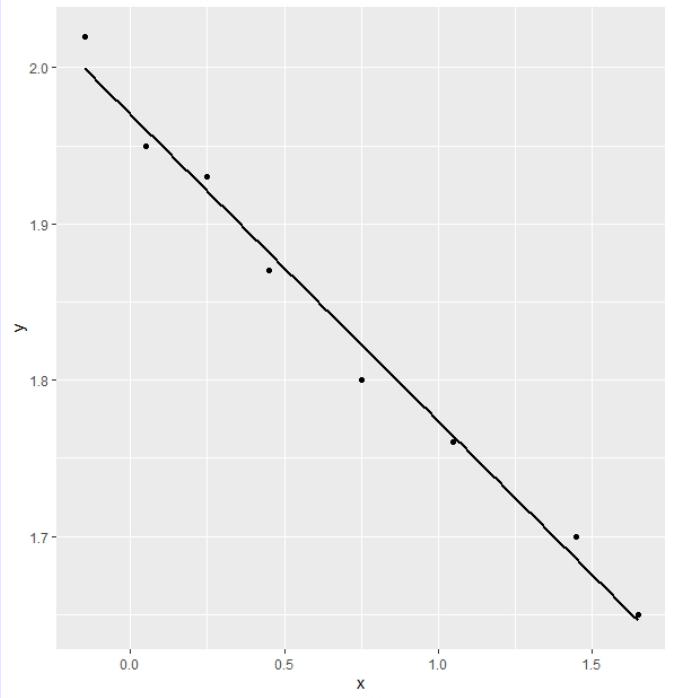
- η τιμή $\hat{\beta}_1 = -37.1536$ εκφράζει το γεγονός ότι η αναμενόμενη τιμή της δύναμης συνοχής Y μεταξύ του προωθητικού εκρηκτικού αερίου και του αερίου συντήρησης μειώνεται, καθώς αυξάνεται η ηλικία του αερίου συντήρησης. Πιο συγκεκριμένα, για καθεμία επιπλέον εβδομάδα (μοναδιαία αύξηση της ανεξάρτητης μεταβλητής X) αναμένουμε η δύναμη συνοχής να μειώνεται κατά μέσο όρο 37.1536 μονάδες.
- Επίσης, η τιμή $\hat{\beta}_0 = 2627.8225$ μπορεί να ερμηνευτεί ως η αναμενόμενη τιμή της δύναμης συνοχής Y μεταξύ του προωθητικού εκρηκτικού αερίου και του αερίου συντήρησης, όταν η ηλικία του αερίου συντήρησης είναι μηδέν, δηλαδή κατά τη στιγμή παραγωγής. Φυσικά, θα πρέπει να είμαστε ιδιαίτερα προσεκτικοί στην υιοθέτηση της παραπάνω ερμηνείας, αφού η τ.μ. X μπορεί να λάβει θεωρητικά την τιμή μηδέν, αλλά δεν είμαστε σίγουροι ότι οι μετρήσεις περιλαμβάνουν την τιμή μηδέν ή έστω τιμές αρκετά κοντά στο μηδέν έτσι ώστε να έχει νόημα η παραπάνω ερμηνεία.

Άσκηση Αυτοαξιολόγησης 12.1

Στο διάγραμμα διασκόρπισης που εμφανίζεται στο τέλος της άσκησης αποτυπώνονται οι παρατηρήσεις από ένα τυχαίο δείγμα και η εκτιμώμενη ευθεία παλινδρόμησης με τη μέθοδο ελαχίστων τετραγώνων. Ποια από τις παρακάτω εξισώσεις είναι η εκτιμώμενη ευθεία;

- $\hat{y} = 1.9699 - 0.1962x$
- $\hat{y} = 1.9699 + 0.1962x$
- $\hat{y} = -1.9699 - 0.1962x$
- $\hat{y} = -1.9699 + 0.1962x$

Δώστε την ερμηνεία της εκτίμησης της ευθείας παλινδρόμησης.



Άσκηση Αυτοαξιολόγησης 12.2

Σε ένα σημείο ενός αυτοκινητόδρομου έχει εγκατασταθεί ένα σύστημα παρακολούθησης της μέσης ταχύτητας Y και του αριθμού X των αυτοκινήτων που διέρχονται από το σημείο κάθε ώρα. Τα συγκεντρωτικά στοιχεία που παρουσιάζονται στη συνέχεια αφορούν ένα τυχαία επιλεγμένο δείγμα 20 ωρών.

$$n = 20 \qquad \sum_{i=1}^{20} x_i = 12090 \qquad \sum_{i=1}^{20} y_i = 1924.87$$

$$\sum_{i=1}^{20} x_i^2 = 7837148 \qquad \sum_{i=1}^{20} y_i^2 = 189187.08 \qquad \sum_{i=1}^{20} x_i y_i = 1118277.29.$$

Από τα παραπάνω δεδομένα προκύπτει ότι η εκτίμηση ελαχίστων τετραγώνων του συντελεστή της επεξηγηματικής μεταβλητής ισούται με -0.085687. Να εκτιμήσετε το απλό γραμμικό μοντέλο παλινδρόμησης και, στη συνέχεια, να ερμηνεύσετε την ευθεία παλινδρόμησης, καθώς και την κλίση της ευθείας με τους όρους του προβλήματος.

12.3 Στατιστική συμπερασματολογία

Τα $\hat{\beta}_0$ και $\hat{\beta}_1$ αποτελούν σημειακές εκτιμήσεις των παραμέτρων β_0 και β_1 αντίστοιχα, οι οποίες λαμβάνονται με τη βοήθεια ενός τυχαίου δείγματος (x_i, y_i), $i = 1, \ldots, n$. Αυτό σημαίνει ότι με ένα άλλο δείγμα θα λαμβάναμε διαφορετικές (πιθανώς παραπλήσιες, αλλά όχι ίδιες) εκτιμήσεις με αυτές από το αρχικό δείγμα. Επομένως, για να μπορέσουμε να εξάγουμε ασφαλέστερα συμπεράσματα για τις παραμέτρους του γραμμικού μοντέλου, θα πρέπει αρχικά να μελετήσουμε τις δειγματοληπτικές κατανομές των $\hat{\beta}_0$ και $\hat{\beta}_1$. Στη συνέχεια και με βάση τις κατανομές αυτές, μπορούμε να κατασκευάσουμε διαστήματα εμπιστοσύνης αλλά και να διεξάγουμε ελέγχους υποθέσεων για τις παραμέτρους β_0 και β_1.

Πέρα από τη συμπερασματολογία για τις παραμέτρους του γραμμικού μοντέλου σημαντική είναι:

- η συνολική αξιολόγηση του εκτιμώμενου μοντέλου παλινδρόμησης, η οποία βασίζεται στην αποκαλούμενη ανάλυση διασποράς και
- η χρήση του εκτιμώμενου μοντέλου παλινδρόμησης για την κατασκευή
  - διαστημάτων εμπιστοσύνης για την αναμενόμενη τιμή της εξαρτημένης μεταβλητής για συγκεκριμένες τιμές της ανεξάρτητης μεταβλητής X και
  - διαστημάτων πρόβλεψης για τις τιμές μελλοντικών παρατηρήσεων για συγκεκριμένες τιμές της ανεξάρτητης μεταβλητής X.

Στην παρούσα ενότητα θα παρουσιάσουμε τα βασικά στοιχεία των παραπάνω εννοιών, δίνοντας παράλληλα έμφαση στην εφαρμογή των αποτελεσμάτων αυτών σε πραγματικά δεδομένα.

12.3.1 Κατανομή των εκτιμητριών ελαχίστων τετραγώνων

Οι εκτιμήτριες ελαχίστων τετραγώνων $\hat{\beta}_0$ και $\hat{\beta}_1$ των β_0 και β_1 μπορούν να εκφραστούν ισοδύναμα από τις σχέσεις:

$$\hat{\beta}_1 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} y_i, \tag{12.3}$$

και

$$\hat{\beta}_0 = \sum_{i=1}^{n} \frac{y_i}{n} - \hat{\beta}_1 \bar{x} = \sum_{i=1}^{n} \left(\frac{1}{n} - \frac{(x_i - \bar{x})\bar{x}}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right) y_i. \tag{12.4}$$

Επισημαίνουμε ότι η σχέση (12.3) προκύπτει εύκολα από τη σχέση (12.1) λαμβάνοντας υπόψη ότι $\sum_{i=1}^{n} (x_i - \bar{x})\bar{y} = 0$.

Παρατήρηση 12.5

Στο σημείο αυτό, θα πρέπει να τονίσουμε ότι τα $\hat{\beta}_i$, $i = 0, 1$, στις παραπάνω σχέσεις είναι τυχαίες μεταβλητές σε αντίθεση με τα $\hat{\beta}_i$ στις σχέσεις (12.3) και (12.4) που είναι οι παρατηρηθείσες τιμές τους (πραγματικοί αριθμοί). Στη συνέχεια, θα χρησιμοποιείται το ίδιο σύμβολο και για τις δύο αυτές περιπτώσεις, καθώς κάτι τέτοιο απλοποιεί τους συμβολισμούς και δεν δημιουργεί ιδιαίτερα προβλήματα στην παρακολούθηση του κειμένου.

Επομένως, από τις (12.3) και (12.4), οι εκτιμήτριες ελαχίστων τετραγώνων $\hat{\beta}_0$ και $\hat{\beta}_1$ μπορούν να ιδωθούν ως γραμμικοί συνδυασμοί των τιμών y_i των τυχαίων μεταβλητών Y_i. Αυτό σημαίνει ότι υπό την υπόθεση της κανονικής κατανομής των τυχαίων σφαλμάτων, και επομένως και των Y_i, οι εκτιμήτριες $\hat{\beta}_0$ και $\hat{\beta}_1$ είναι και αυτές κανονικά κατανεμημένες. Μάλιστα, μπορεί εύκολα να αποδειχθεί (βλ., μεταξύ άλλων, τα συγγράμματα των Rencher and Schaalje, 2008; Κούτρας, 2010) ότι:

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right),$$

και

$$\hat{\beta}_0 \sim N\left(\beta_0, \frac{\sigma^2 \sum_{i=1}^{n} x_i^2}{n \sum_{i=1}^{n} (x_i - \bar{x})^2} \right).$$

Παρατήρηση 12.6

Από τα παραπάνω αποτελέσματα είναι φανερό ότι οι εκτιμήτριες ελαχίστων τετραγώνων $\hat{\beta}_0$ και $\hat{\beta}_1$ είναι αμερόληπτες εκτιμήτριες των β_0 και β_1, αντίστοιχα.

Η διασπορά σ^2 των τυχαίων σφαλμάτων, η οποία εμφανίζεται στην προηγούμενη σχέση, είναι άγνωστη. Για τον λόγο αυτό θα πρέπει, αρχικά, να εκτιμηθεί διασπορά σ^2, προτού μπορέσουμε να χρησιμοποιήσουμε τα παραπάνω αποτελέσματα, για την πραγματοποίηση ελέγχων υποθέσεων και την κατασκευή διαστημάτων εμπιστοσύνης για τις πραγματικές τιμές των β_0 και β_1. Η εκτίμηση της διασποράς σ^2 γίνεται μέσω της ποσότητας:

$$s^2 = \frac{1}{n-2}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2.$$

Η αντίστοιχη τυχαία μεταβλητή, S^2, η οποία ονομάζεται και **μέσο άθροισμα τετραγώνων των υπολοίπων** (MSE) ή μέσο τετραγωνικό σφάλμα, όπως θα δούμε σε επόμενη παράγραφο, μπορεί να αποδειχθεί ότι είναι αμερόληπτη εκτιμήτρια της σ^2, δηλαδή ότι $E(S^2) = \sigma^2$.

Επομένως, η εκτιμώμενη διασπορά του $\hat{\beta}_1$ δίνεται από τη σχέση:

$$\widehat{Var}(\hat{\beta}_1) = S^2\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^{-1},$$

ενώ η εκτιμώμενη τυπική απόκλισή του από τη σχέση:

$$se(\hat{\beta}_1) := \sqrt{\widehat{Var}(\hat{\beta}_1)} = S\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^{-1/2}.$$

Ακόμα σημειώνεται ότι $(n-2)\,S^2/\sigma^2 \sim \chi^2_{n-2}$ και ότι οι εκτιμήτριες των συντελεστών β_0 και β_1 είναι ανεξάρτητες του S^2 (για την απόδειξη ο/η αναγνώστης/στρια παραπέμπεται, ενδεικτικά, στο βιβλίο των Καρώνη και Οικονόμου, 2017). Αυτό έχει ως συνέπεια από τον ορισμό της t κατανομής ότι:

$$\frac{\hat{\beta}_1 - \beta_1}{S\sqrt{\frac{1}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}} \sim t_{n-2}. \tag{12.5}$$

Παρατήρηση 12.7

Παρόμοια αποτελέσματα μπορούμε να λάβουμε και για το $\hat{\beta}_0$. Πιο συγκεκριμένα, μπορεί να αποδειχτεί ότι

$$\frac{\hat{\beta}_0 - \beta_0}{se(\hat{\beta}_0)} \sim t_{n-2}, \tag{12.6}$$

όπου

$$se(\hat{\beta}_0) := \sqrt{\widehat{Var}(\hat{\beta}_1)} = S\sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n\sum_{i=1}^{n}(x_i - \bar{x})^2}}.$$

Παρατήρηση 12.8

Οι εκτιμήτριες ελαχίστων τετραγώνων $\hat{\beta}_0$ και $\hat{\beta}_1$ γενικά δεν είναι ασυσχέτιστες μεταξύ τους, αφού

$$Cov(\hat{\beta}_0, \hat{\beta}_1) = -\frac{\bar{x}}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sigma^2$$

ενώ η εκτίμηση της συνδιασποράς τους μπορεί να γίνει μέσω της σχέσης

$$\widehat{Cov}(\hat{\beta}_0, \hat{\beta}_1) = -\frac{\bar{x}}{\sum_{i=1}^{n}(x_i - \bar{x})^2}S^2.$$

Για μια απόδειξη του παραπάνω αποτελέσματος παραπέμπουμε ενδεικτικά στο βιβλίο Καρακώστας (2002).

Προτού ολοκληρωθεί η παρούσα ενότητα, αξίζει να αναφερθούμε στο θεώρημα των Gauss-Markov, το οποίο δηλώνει ότι οι εκτιμήτριες ελαχίστων τετραγώνων $\hat{\beta}_0$ και $\hat{\beta}_1$ όχι μόνο είναι αμερόληπτες εκτιμήτριες των β_0 και β_1, αντίστοιχα, αλλά έχουν και τη μικρότερη διασπορά ανάμεσα σε όλες τις αμερόληπτες γραμμικές εκτιμήτριες.

Θεώρημα 12.1: Θεώρημα Gauss-Markov

Στο απλό γραμμικό μοντέλο οι εκτιμήτριες ελαχίστων τετραγώνων $\hat{\beta}_0$ και $\hat{\beta}_1$ είναι οι αμερόληπτες γραμμικές εκτιμήτριες των β_0 και β_1 με την ελάχιστη δυνατή διασπορά (βλ., μεταξύ άλλων, Rencher and Schaalje, 2008).

12.3.2 Στατιστικοί έλεγχοι και διαστήματα εμπιστοσύνης

Με βάση τα αποτελέσματα των σχέσεων (12.6) και (12.5) μπορούμε να πραγματοποιήσουμε ελέγχους υποθέσεων και να κατασκευάσουμε διαστήματα εμπιστοσύνης για τις πραγματικές τιμές των β_0 και β_1, αντίστοιχα. Πιο συγκεκριμένα, οι στατιστικές συναρτήσεις

$$T_i = \frac{\hat{\beta}_i - \beta_i}{se(\hat{\beta}_i)} \sim t_{n-2},\ i = 1,2,$$

μπορούν να χρησιμοποιηθούν

- για τον έλεγχο των υποθέσεων

  $$\mathrm{H}_0 : \beta_i = \beta_{i0} \quad \text{έναντι} \quad \mathrm{H}_1 : \beta_i \neq \beta_{i0},\ i = 1,2,$$

  χρησιμοποιώντας ως κρίσιμη περιοχή την $C = \{|T_i| > t_{n-2,\alpha/2}\}$, σε επίπεδο σημαντικότητας α, αλλά και

- για την εύρεση $100(1-\alpha)\%$ διαστημάτων εμπιστοσύνης για τις παραμέτρους β_i, $i = 1, 2$, καταλήγοντας στις σχέσεις:

  $$\left(\hat{\beta}_i - t_{n-2,\alpha/2} \cdot se(\hat{\beta}_i),\ \hat{\beta}_i + t_{n-2,\alpha/2} \cdot se(\hat{\beta}_i)\right).$$

Παράδειγμα 12.2

Για τα δεδομένα του Παραδείγματος 12.1 να εξετάσετε αν υπάρχουν στατιστικά σημαντικές ενδείξεις, σε επίπεδο σημαντικότητας 0.05, ότι ο συντελεστής της επεξηγηματικής μεταβλητής X (ηλικία του αερίου συντήρησης) είναι διάφορος του μηδενός. Στη συνέχεια, κατασκευάστε και ερμηνεύστε ένα 95% διάστημα εμπιστοσύνης για τον συντελεστή αυτόν.
Δίνεται ότι $s = 96.10601$[α΄].

[α΄] Ο τρόπος υπολογισμού του από τα στοιχεία που δίνονται στην εκφώνηση του παραδείγματος παρουσιάζεται στην επόμενη ενότητα.

Λύση Παραδείγματος 12.2

Οι υποθέσεις που ζητείται να εξετάσουμε, σε επίπεδο σημαντικότητας 0.05, είναι οι

$$\mathrm{H}_0 : \beta_1 = 0 \quad \text{έναντι} \quad \mathrm{H}_1 : \beta_1 \neq 0.$$

Η στατιστική συνάρτηση του ελέγχου είναι η

$$\mathrm{T}_1 = \frac{\hat{\beta}_1 - \beta_1}{se(\hat{\beta}_1)}$$

η οποία υπό τη μηδενική υπόθεση εκφράζεται ως

$$\mathrm{T}_1 = \frac{\hat{\beta}_1 - 0}{se(\hat{\beta}_1)}$$

και κατανέμεται σύμφωνα με την $t_{20-2} = t_{18}$ κατανομή. Η κρίσιμη περιοχή του ελέγχου είναι η

$$C = \{|T_1| > t_{18,0.05/2}\} = \{|\mathrm{T}_1| > 2.101\},$$

όπου η τιμή $t_{18,0.05/2}$ προσδιορίστηκε με τη βοήθεια του Πίνακα Α΄.4 του Παραρτήματος Α΄.
Η παρατηρούμενη τιμή, $\mathrm{T}_{1,obs}$, της στατιστικής συνάρτησης ελέγχου T_1 για τα δεδομένα του παραδείγματος ισούται με

$$\mathrm{T}_{1,obs} = \frac{\hat{\beta}_1 - 0}{se(\hat{\beta}_1)} = \frac{-37.1536 - 0}{96.10601\sqrt{\frac{1}{1106.562}}} = -12.86,$$

αφού

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)}{n} = 4677.69 - \frac{267.25^2}{20} = 1106.562.$$

Επομένως, αφού

$$|T_{1,obs}| = 12.86 > 2.101 = t_{18,0.05/2}$$

η μηδενική υπόθεση $\mathrm{H}_0 : \beta_1 = 0$ απορρίπτεται σε επίπεδο σημαντικότητας 0.05 προς όφελος της $H_1 : \beta_1 \neq 0$. Αυτό σημαίνει ότι η ηλικία του αερίου συντήρησης είναι στατιστικά σημαντική για τη δύναμη συνοχής Y μεταξύ του προωθητικού εκρηκτικού αερίου και του αερίου συντήρησης που περιέχει ο εκρηκτικός μηχανισμός.
Ένα 95% διάστημα εμπιστοσύνης για το β_1 δίνεται από τη σχέση:

$$\hat{\beta}_1 \pm t_{20-2,0.05/2} \cdot se(\hat{\beta}_1),$$

δηλαδή ένα 95% διάστημα εμπιστοσύνης προκύπτει από τη σχέση:

$$-37.1536 \pm 2.101 \cdot 96.10601\sqrt{\frac{1}{1106.562}}.$$

Πραγματοποιώντας τις πράξεις στην παραπάνω σχέση καταλήγουμε ότι το

$$(-43.224, -31.084)$$

είναι ένα 95% διάστημα εμπιστοσύνης για το β_1. Επομένως, είμαστε 95% σίγουροι ότι η πραγματική τιμή του συντελεστή β_1 είναι μεγαλύτερη από -43.224 και μικρότερη από -31.084. Με βάση το διάστημα αυτό μπορούμε να εξετάσουμε και τη σημαντικότητα της ανεξάρτητης μεταβλητής, δηλαδή τη σημαντικότητα του μοντέλου. Πιο συγκεκριμένα, αφού το μηδέν δεν ανήκει στο παραπάνω διάστημα, η μηδενική υπόθεση $H_0 : \beta_1 = 0$ απορρίπτεται σε επίπεδο σημαντικότητας 0.05 προς όφελος της $H_1 : \beta_1 \neq 0$.

Παρατήρηση 12.9

Ο έλεγχος των υποθέσεων

$$\mathrm{H}_0 : \beta_1 = 0 \quad \text{έναντι} \quad \mathrm{H}_1 : \beta_1 \neq 0,$$

στην πραγματικότητα εκφράζει, όπως αναφέρθηκε παραπάνω, τη σημαντικότητα της ανεξάρτητης μεταβλητής στο μοντέλο παλινδρόμησης, δηλαδή το αν η ανεξάρτητη μεταβλητή επηρεάζει ή όχι την εξαρτημένη μεταβλητή. Για τον λόγο αυτόν, ο έλεγχος των υποθέσεων αυτών αναφέρεται συχνά και ως έλεγχος σημαντικότητας του μοντέλου. Οι υποθέσεις αυτές αποτελούν αντικείμενο μελέτης στην επόμενη ενότητα.

Άσκηση Αυτοαξιολόγησης 12.3

Για τα δεδομένα του Παραδείγματος 12.1 να εξετάσετε αν υπάρχουν στατιστικά σημαντικές ενδείξεις, σε επίπεδο σημαντικότητας 0.05, ότι ο σταθερός όρος της εξίσωσης παλινδρόμησης είναι διάφορος του 2600.

12.3.3 Ανάλυση διασποράς

Η προσαρμογή και η ερμηνεία της εκτιμώμενης εξίσωσης παλινδρόμησης είναι τα βασικά βήματα στην εφαρμογή του απλού γραμμικού μοντέλου. Για να έχει όμως (πρακτική) αξία η παραπάνω διαδικασία θα πρέπει να εξεταστούν

- αρχικά, η σημαντικότητα του μοντέλου και, στη συνέχεια,
- η ερμηνευτική ικανότητα (βαθμός) του μοντέλου.

Σχετικά με τη σημαντικότητα του μοντέλου είδαμε στην προηγούμενη ενότητα, στο Παράδειγμα 12.2, ότι αυτή αφορά τον έλεγχο των υποθέσεων:

$$\mathrm{H}_0 : \beta_1 = 0 \quad \text{έναντι} \quad \mathrm{H}_1 : \beta_1 \neq 0,$$

και μπορεί να πραγματοποιηθεί είτε με τον έλεγχο t είτε με την κατασκευή διαστημάτων εμπιστοσύνης για το β_1. Πέρα όμως από τις δύο προαναφερθείσες προσεγγίσεις υπάρχει και μια τρίτη, ο έλεγχος F, ο οποίος βασίζεται στη λεγόμενη Ανάλυση Διασποράς. Η ανάλυση διασποράς προσφέρει και επιπλέον πληροφορίες, αφού σε αυτήν βασίζεται και ο υπολογισμός του συντελεστή προσδιορισμού R^2, ο οποίος χρησιμοποιείται για την ποσοτικοποίηση της ερμηνευτικής ικανότητας του εκτιμώμενου μοντέλου. Στη συνέχεια της ενότητας θα παρουσιαστούν αναλυτικά τα δύο αυτά στοιχεία μέσα από την ανάπτυξη της ανάλυσης διασποράς.

Η μεταβλητότητα των τιμών y_i της εξαρτημένης μεταβλητής που παρατηρούμε σε ένα τυχαίο δείγμα οφείλεται

- στη μεταβλητότητα των τιμών της X και
- σε άλλους, τυχαίους παράγοντες, οι οποίοι ενσωματώνουν την επίδρασή τους στα τυχαία σφάλματα.

Η ανάλυση διασποράς αποσκοπεί στο να διακρίνει τη συνεισφορά κάθε πηγής μεταβλητότητας στην παρατηρούμενη (ολική) μεταβλητότητα της εξαρτημένης μεταβλητής. Πιο συγκεκριμένα με την ανάλυση διασποράς αναλύουμε την ολική μεταβλητότητα των y_i στις δύο παραπάνω πηγές. Ένα μέτρο της μεταβλητότητας των y_i είναι η διασπορά τους ή, ισοδύναμα, η ποσότητα $\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2$, δηλαδή η διασπορά χωρίς τον σταθερό όρο $1/(n-1)$. Η ποσότητα $\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2$, η οποία συμβολίζεται με SST και ονομάζεται **ολικό άθροισμα τετραγώνων**, μπορεί να εκφραστεί ως

$$\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2=\sum_{i=1}^{n}\left(y_i-\hat{y}_i+\hat{y}_i-\bar{y}\right)^2=\sum_{i=1}^{n}\left(\hat{y}_i-\bar{y}\right)^2+\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2 \tag{12.7}$$

αφού $2\sum_{i=1}^{n}\left(\hat{y}_i-\bar{y}\right)\left(y_i-\hat{y}_i\right)=0$, καθώς οι $\hat{\beta}_0$ και $\hat{\beta}_1$ επαληθεύουν το σύστημα των εξισώσεων που δόθηκε πριν στις σχέσεις (12.1) και (12.2).

Ο πρώτος όρος στο δεξί μέρος της παραπάνω σχέσης εμπλέκει τα σημεία $\hat{y}_i=\hat{\beta}_0+\hat{\beta}_1x_i$, δηλαδή την επίδραση της επεξηγηματικής μεταβλητής στη διαμόρφωση της τιμής της Y. Ο όρος αυτός συμβολίζεται με SSR και αναφέρεται ως **άθροισμα τετραγώνων από την παλινδρόμηση**. Από την άλλη, ο δεύτερος όρος βασίζεται στην κατακόρυφη απόκλιση της παρατηρούμενης από την εκτιμώμενη και στην ουσία εμπλέκει τα **υπόλοιπα** ή, αλλιώς, **κατάλοιπα** που ορίζονται από τη σχέση:

$$e_i=y_i-\hat{y}_i=y_i-\hat{\beta}_0-\hat{\beta}_1x_i.$$

Τα υπόλοιπα μπορούν να θεωρηθούν ως οι εκτιμήσεις των άγνωστων τυχαίων σφαλμάτων ε_i και συγκεντρώνουν την επίδραση άλλων, τυχαίων παραγόντων, στη διαμόρφωση της παρατηρούμενης τιμής της Y. Ο όρος $\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2$ συμβολίζεται με SSE και αναφέρεται ως **άθροισμα τετραγώνων των υπολοίπων**.

Με βάση τους παραπάνω συμβολισμούς η σχέση (12.7) εκφράζεται ως

$$SST=SSR+SSE. \tag{12.8}$$

Η παραπάνω ανάλυση της ολικής διασποράς της εξαρτημένης μεταβλητής μπορεί να χρησιμοποιηθεί, όπως έχει προαναφερθεί, για την κατασκευή του ελέγχου F για τη σημαντικότητα του μοντέλου παλινδρόμησης και τον υπολογισμό του συντελεστή προσδιορισμού R^2.

12.3.3.1 Έλεγχος F

Ο **έλεγχος** F για τη σημαντικότητα της παλινδρόμησης ελέγχει τις υποθέσεις

$$\mathrm{H}_0:\beta_1=0 \quad \text{έναντι} \quad \mathrm{H}_1:\beta_1\neq 0$$

και βασίζεται στην παρατήρηση ότι υπό τη μηδενική υπόθεση το SSR, δηλαδή το μέρος της μεταβλητότητας των y_i που οφείλεται στην επεξηγηματική μεταβλητή, πρέπει να είναι σχετικά μικρό, ενώ το SSE πρέπει να είναι σχεδόν ίσο με το SST. Αντιθέτως, αν η μηδενική υπόθεση δεν ευσταθεί, τότε αλλαγές στην τιμή της επεξηγηματικής μεταβλητής επηρεάζουν την τιμή της εξαρτημένης μεταβλητής. Αυτό σημαίνει ότι το SSR θα πρέπει να αποτελεί μεγαλύτερο τμήμα του SST. Οι παρατηρήσεις αυτές οδήγησαν στην κατασκευή της στατιστικής συνάρτησης F

$$F=\frac{SSR/1}{SSE/(n-2)} \quad \left(=\frac{MSR}{MSE}\right),$$

η οποία υπό τη μηδενική υπόθεση (και τις υποθέσεις για τα τυχαία σφάλματα) αποδεικνύεται ότι ακολουθεί την $F_{1,(n-2)}$ κατανομή. Με βάση τις παραπάνω παρατηρήσεις είναι φανερό ότι η $\mathrm{H}_0 : \beta_1 = 0$ απορρίπτεται, όταν ο αριθμητής της στατιστικής συνάρτησης ελέγχου F είναι κατά πολύ μεγαλύτερος από τον παρανομαστή, δηλαδή όταν η F λαμβάνει μεγάλες τιμές. Πιο συγκεκριμένα, η $\mathrm{H}_0 : \beta_1 = 0$ απορρίπτεται σε επίπεδο σημαντικότητας α, όταν $F > F_{1,(n-2),\alpha}$, όπου $F_{1,(n-2),\alpha}$ το άνω α ποσοστιαίο σημείο της $F_{1,(n-2)}$ κατανομής.

Παρατήρηση 12.10

Η ιδέα της απόδειξης του παραπάνω αποτελέσματος βασίζεται στον ορισμό της F κατανομής και στην παρατήρηση ότι οι ποσότητες SSR/σ^2 και SSE/σ^2 είναι ανεξάρτητες τυχαίες μεταβλητές που ακολουθούν χ^2 κατανομή με 1 και $n-2$ βαθμούς ελευθερίας, αντίστοιχα. Για μια αναλυτική απόδειξη του παραπάνω παραπέμπουμε ενδεικτικά στο βιβλίο Κουτρουβέλης (2000).
Σημειώνεται ότι κατά την απόδειξη του παραπάνω αποτελέσματος οι ποσότητες SSR και SSE εκλαμβάνονται ως τυχαίες μεταβλητές. Παρ' όλα αυτά για να απλοποιηθεί ο συμβολισμός χρησιμοποιείται το ίδιο σύμβολο και για την περίπτωση όπου οι ποσότητες αυτές εκλαμβάνονται ως τυχαίες μεταβλητές και για την περίπτωση όπου αναφερόμαστε στις τιμές τους.

Παρατήρηση 12.11

Στην περίπτωση του απλού γραμμικού μοντέλου ο έλεγχος t και ο έλεγχος F είναι ισοδύναμοι για την υπόθεση $H_0 : \beta_1 = 0$, αφού ισχύει ότι $F = T^2$ και $F_{1,n-2,\alpha} = t^2_{n-2,\alpha/2}$.

Τα παραπάνω αποτελέσματα μπορούν να συγκεντρωθούν στον αποκαλούμενο **πίνακα ανάλυσης διασποράς** (Πίνακας 12.1). Στον πίνακα αυτόν, εκτός από τα αθροίσματα τετράγωνων και τον έλεγχο F, παρατίθενται ακόμα δύο στήλες:

- οι **βαθμοί ελευθερίας** (degrees of freedom) και
- το **μέσο άθροισμα τετραγώνων**.

Οι τιμές στη στήλη του μέσου αθροίσματος τετραγώνων προκύπτουν από τη διαίρεση των αθροισμάτων τετραγώνων με τους αντίστοιχους βαθμούς ελευθερίας. Από την άλλη, οι βαθμοί ελευθερίας εκφράζουν το πλήθος των τιμών στον τελικό υπολογισμό μιας στατιστικής συνάρτησης, όπου είναι ελεύθερες να μεταβάλλονται, έτσι ώστε να λάβουμε ένα προκαθορισμένο αποτέλεσμα. Για να γίνει αυτό πιο κατανοητό, αρκεί να παρατηρήσουμε, για παράδειγμα, ότι, αν και το SST υπολογίζεται από n το πλήθος όρους $\left(y_i - \bar{y}\right)^2$, δεν πρόκειται για n το πλήθος ξεχωριστές πληροφορίες λόγω του περιορισμού

$$\sum_{i=1}^{n} \left(y_i - \bar{y}\right) = 0.$$

Έτσι, το SST έχει $n-1$ βαθμούς ελευθερίας, διότι, αν γνωρίζουμε $n-1$ το πλήθος αποκλίσεις, έστω τις $\left(y_i - \bar{y}\right)$, $i = 1, \ldots, n-1$, τότε προφανώς και η εναπομείνασα απόκλιση $\left(y_n - \bar{y}\right)$ καθορίζεται από τις υπόλοιπες και δεν προσθέτει καμία νέα πληροφορία, καθώς $\left(y_n - \bar{y}\right) = -\sum_{i=1}^{n-1} \left(y_i - \bar{y}\right)$.

Παρόμοια, το SSE έχει $n-2$ βαθμούς ελευθερίας, γιατί υπάρχουν δύο περιορισμοί

$$\sum_{i=1}^{n} \left(y_i - \hat{y}_i\right) = 0 \ \text{ και } \ \sum_{i=1}^{n} x_i \left(y_i - \hat{y}_i\right) = 0,$$

που πρέπει να ικανοποιούνται. Οι περιορισμοί αυτοί προκύπτουν κατά την εκτίμηση των παραμέτρων β_0 και β_1 με τη μέθοδο ελαχίστων τετραγώνων, καθώς οι εκτιμητές αυτών ικανοποιούν το σύστημα των εξισώσεων που δόθηκε προηγουμένως στις σχέσεις προσδιορισμού τους (12.1) και (12.2).

Πίνακας 12.1: Πίνακας ανάλυσης διασποράς απλού γραμμικού μοντέλου.

Πηγή μεταβλητότητας	Άθροισμα τετραγώνων	Βαθμοί ελ/ρίας	Μέσο άθροισμα τετραγώνων	Έλεγχος F
Παλινδρόμηση	$SSR = \sum_{i=1}^{n} \left(\hat{y}_i - \bar{y}\right)^2$	1	$MSR = \frac{SSR}{1}$	$F = \frac{MSR}{MSE}$
Υπόλοιπα	$SSE = \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2$	$n-2$	$MSE = S^2 = \frac{SSE}{n-2}$	
Σύνολο	$SST = \sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2$	$n-1$		

Το SSR επίσης έχει 1 βαθμό ελευθερίας, γιατί μπορεί να υπολογιστεί από μία μόνο πληροφορία, τη $\hat{\beta}_1$, αφού

$$SSR = \sum_{i=1}^{n} \left(\hat{y}_i - \bar{y}\right)^2 = \hat{\beta}_1^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 .$$

Παρατηρήστε ότι το άθροισμα των βαθμών ελευθερίας του SSR και του SSE ισούται με τους βαθμούς ελευθερίας του SST.

Παρατήρηση 12.12

Οι υπολογισμοί των ποσοτήτων SST, SSR και SSE μπορούν να γίνουν και με τις παρακάτω, ισοδύναμες με τους ορισμούς τους, σχέσεις:

$$SST = \sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}$$

$$SSR = \hat{\beta}_1 \left(\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n} \right)$$

$$SSE = SST - SSR.$$

Η τελευταία σχέση μας δίνει έναν εύκολο και συχνά πιο γρήγορο τρόπο υπολογισμού της εκτίμησης της τυπικής απόκλισης των τυχαίων σφαλμάτων. Πιο συγκεκριμένα, από τον υπολογισμό των SST και SSR υπολογίζουμε στη συνέχεια το SSE. Από τη διαίρεση του SSE με τους βαθμούς ελευθερίας προκύπτει το **μέσο τετραγωνικό σφάλμα**, MSE, το οποίο αποτελεί μια αμερόληπτη εκτιμήτρια της διασποράς σ^2 των τυχαίων σφαλμάτων. Επομένως, παίρνοντας την τετραγωνική ρίζα του MSE λαμβάνουμε την εκτίμηση της τυπικής απόκλισης των τυχαίων σφαλμάτων.

12.3.3.2 Συντελεστής Προσδιορισμού R^2

Από τον πίνακα ανάλυσης διασποράς μπορούμε να υπολογίσουμε τον αποκαλούμενο συντελεστή προσδιορισμού R^2 μέσω της σχέσης:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}.$$

Ο συντελεστής προσδιορισμού R^2 εκφράζει το ποσοστό της μεταβλητότητας της εξαρτημένης μεταβλητής Y που εξηγείται από τη X και λαμβάνει τιμές στο [0,1] ή μεταξύ του μηδέν και του 100, όταν εκφράζεται ως ποσοστό επί τοις εκατό. Όσο πιο κοντά είναι η τιμή του R^2 στη μονάδα τόσο μεγαλύτερο το ποσοστό της

μεταβλητότητας της Y που ερμηνεύεται από τη X ή, ισοδύναμα, από το γραμμικό μοντέλο και, επομένως, τόσο ισχυρότερη είναι η γραμμική σχέση εξάρτησης των τ.μ. Y και X. Επιπλέον, τιμές κοντά στο μηδέν φανερώνουν μια ασθενή γραμμική συσχέτιση μεταξύ των δύο μεταβλητών.

Παρατήρηση 12.13

Σημειώνεται ότι στο απλό γραμμικό μοντέλο ο συντελεστής προσδιορισμού R^2 συνδέεται άμεσα με τον δειγματικό συντελεστή γραμμικής συσχέτισης του Pearson, ο οποίος όχι μόνο εκφράζει τον βαθμό της γραμμικής συσχέτισης μεταξύ δύο μεταβλητών X και Y, αλλά και την κατεύθυνση της σχέσης αυτής. Πιο συγκεκριμένα, ισχύει ότι $R^2 = r_{xy}^2$, όπου

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})\left(y_i - \bar{y}\right)}{\sqrt{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 \sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}} = \frac{s_{xy}}{\sqrt{s_{xx}s_{yy}}}$$

ο δειγματικός συντελεστής συσχέτισης Pearson. Οι τιμές που μπορεί να πάρει ο συντελεστής αυτός είναι μεταξύ -1 και 1. Απόλυτες τιμές κοντά στο ένα υποδηλώνουν ισχυρή γραμμική συσχέτιση, ενώ τιμές κοντά στο μηδέν ασθενή. Θετικές τιμές του r_{xy} φανερώνουν ότι οι δύο μεταβλητές σχετίζονται θετικά, ενώ, αν λαμβάνει αρνητικές τιμές, ότι σχετίζονται αρνητικά. Θετική συσχέτιση σημαίνει ότι αύξηση της μιας συνεπάγεται αύξηση (κατά μέσο όρο) της άλλης, ενώ το αντίθετο σημαίνει η αρνητική συσχέτιση.

Παράδειγμα 12.3

Για τα δεδομένα του Παραδείγματος 12.1 να εξετάσετε με τη βοήθεια του ελέγχου F και σε επίπεδο σημαντικότητας 0.05, αν η παλινδρόμηση είναι στατιστικά σημαντική. Στη συνέχεια υπολογίστε τον συντελεστή προσδιορισμού και δώστε την ερμηνεία της τιμής του.

Λύση Παραδείγματος 12.3

Αρχικά, θα κατασκευάσουμε τον πίνακα ανάλυσης διασποράς και εν συνεχεία με τη βοήθειά του, θα πραγματοποιήσουμε τον έλεγχο των υποθέσεων

$$H_0 : \beta_1 = 0 \quad \text{έναντι} \quad \mathrm{H}_1 : \beta_1 \neq 0.$$

Τέλος, θα υπολογίσουμε την τιμή του συντελεστή προσδιορισμού R^2.

Για να κατασκευάσουμε τον πίνακα ανάλυσης διασποράς χρειάζεται να υπολογίσουμε πρώτα τα SST, SSR και SSE. Ο υπολογισμός των ποσοτήτων αυτών γίνεται ως ακολούθως:

$$SST = \sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n} = 92547433.46 - 42627.15^2/20 = 1693737.60,$$

$$SSR = \hat{\beta}_1\left(\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}\right)$$

$$= -37.1536\,(528492.64 - 267.25 \cdot 42627.15/20) = 1527483.02,$$

$$SSE = SST - SSR = 1693737.60 - 1527483.02 = 166254.58.$$

Επομένως, ο πίνακας ανάλυσης διασποράς είναι ο

Πηγή μεταβλητότητας	Άθροισμα τετραγώνων	Βαθμοί ελ/ρίας	Μέσο άθροισμα τετραγώνων	Έλεγχος F
Παλινδρόμηση	1527483.02	1	1527483.02	165.377
Υπόλοιπα	166254.58	20-2=18	9236.366	
Σύνολο	1693737.60	20-1=19		

Η κρίσιμη περιοχή του ελέγχου F, σε επίπεδο σημαντικότητας 0.05, είναι η $F > F_{0.05,1,18} = 4.41$ (Πίνακας Α΄.9, Παραρτήματος Α΄) και, επομένως, αφού η παρατηρούμενη τιμή F_{obs} του ελέγχου F ικανοποιεί την

$$F_{obs} = 183.377 > 4.41 = F_{0.05,1,18}$$

η μηδενική υπόθεση $\mathrm{H}_0 : \beta_1 = 0$ απορρίπτεται σε επίπεδο σημαντικότητας 0.05 και η παλινδρόμηση είναι στατιστικά σημαντική[α΄].

Ο συντελεστής προσδιορισμού δίνεται από τη σχέση:

$$R^2 = \frac{SSR}{SST} = \frac{1527483.02}{1693737.60} = 0.9018$$

και, επομένως, μπορούμε να πούμε ότι λίγο παραπάνω από το 90% της μεταβλητότητας που παρατηρούμε στη δύναμη συνοχής Y μεταξύ του προωθητικού εκρηκτικού αερίου και του αερίου συντήρησης ενός εκτοξευτήρα ερμηνεύεται από την ηλικία X του αερίου συντήρησης και λιγότερο από το 10% οφείλεται σε άλλους παράγοντες.

[α΄] Παρατηρήστε ότι στο ίδιο συμπέρασμα είχαμε καταλήξει και με τον έλεγχο t και με το διάστημα εμπιστοσύνης για β_1. Επίσης, παρατηρήστε ότι, αν αγνοήσουμε σφάλματα στρογγυλοποίησης, ισχύει πράγματι ότι $\mathrm{T}^2_{1,obs} = F_{obs}$. Επιπλέον, παρατηρήστε ότι $\sqrt{MSE} = \sqrt{9236.366} = 96.10601$, όσο ακριβώς είχε δοθεί στο Παράδειγμα 12.2.

Άσκηση Αυτοαξιολόγησης 12.4

Ποια από τις παρακάτω τιμές μοιάζει πιο αληθοφανής για τον συντελεστή γραμμικής συσχέτισης του Pearson για τα δεδομένα της Άσκησης Αυτοαξιολόγησης 12.1; Δικαιολογήστε την απάντησή σας.

- 0.99352
- -0.37521
- -0.99352
- 0.37521

12.3.4 Διαστήματα εμπιστοσύνης και πρόβλεψης

Η εκτιμώμενη ευθεία παλινδρόμησης στην ουσία μας δίνει μια σημειακή εκτίμηση της αναμενόμενης τιμής της εξαρτημένης μεταβλητής Y για οποιαδήποτε τιμή (στο πειραματικό εύρος τιμών) της επεξηγηματικής μεταβλητής X. Η σημειακή αυτή εκτίμηση μπορεί να χρησιμοποιηθεί και για τη (σημειακή) πρόβλεψη της τιμής μιας νέας παρατήρησης της Y για δοθείσα τιμή $X = x_0$. Η εκτίμηση αυτή δίνεται από τη σχέση

$$\hat{y}_{x_0} = \hat{\beta}_0 + \hat{\beta}_1 x_0 .$$

Η τιμή $\hat{y}_{x_0}$ μπορεί να θεωρηθεί ως μια πραγματοποίηση με βάση το παρατηρηθέν δείγμα της τυχαίας μεταβλητής $\hat{Y}_{x_0}$, η οποία προκύπτει από τον γραμμικό συνδυασμό κανονικά κατανεμημένων τυχαίων μεταβλητών, των $(\hat{\beta}_0, \hat{\beta}_1)$, οπότε μπορούμε να αποδείξουμε ότι

$$\hat{Y}_{x_0} \sim N\left(\mu_{x_0}, \sigma^2\left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]\right),$$

όπου $\mu_{x_0} = \beta_0 + \beta_1 x_0$ (Searle and Gruber, 2016). Η παραπάνω σχέση εκφράζει την κατανομή της εκτιμήτριας της αναμενόμενης τιμής της τυχαίας μεταβλητής $Y|X = x_0$ και μπορεί να χρησιμοποιηθεί για την

κατασκευή διαστημάτων εμπιστοσύνης για την $E(Y|X = x_0)$. Ένα $(1-\alpha)100\%$ διάστημα εμπιστοσύνης για την αναμενόμενη τιμή της Y για $X = x_0$ είναι το

$$\left(\hat{y}_{x_0} - t_{\alpha/2,n-2} \cdot s \cdot \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}},\ \hat{y}_{x_0} + t_{\alpha/2,n-2} \cdot s \cdot \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}\right)$$

όπου $\hat{y}_{x_0}$ η σημειακή εκτίμηση της $E(Y|X = x_0) = \mu_{x_0}$ με βάση την εκτιμώμενη ευθεία παλινδρόμησης, $t_{\alpha/2,n-2}$ το άνω $\alpha/2$ ποσοστιαίο σημείο της t_{n-2} κατανομής και s η εκτίμηση της τυπικής απόκλισης, σ, των τυχαίων σφαλμάτων.

Πέρα όμως από το διάστημα εμπιστοσύνης για την αναμενόμενη τιμή της Y για $X = x_0$, μπορούμε να κατασκευάσουμε και διαστήματα πρόβλεψης για μια νέα παρατήρηση της Y, όταν $X = x_0$. Σε αυτήν την περίπτωση, πέρα από την αβεβαιότητα στην εκτίμηση της μέσης τιμής πρέπει να λάβουμε υπόψη μας, και την αβεβαιότητα κάθε μεμονωμένης παρατήρησης, όπως αυτή εκφράζεται μέσα από το τυχαίο σφάλμα. Για τον λόγο αυτό, η κατασκευή των διαστημάτων πρόβλεψης για μια νέα παρατήρηση της Y, όταν $X = x_0$, βασίζεται στην κατανομή του αποκαλούμενου σφάλματος της πρόβλεψης, $\hat{Y}_{x_0} - y_{x_0}$, η οποία μπορούμε να αποδείξουμε (Searle and Gruber, 2016) ότι είναι η

$$\hat{Y}_{x_0} - y_{x_0} \sim N\left(0,\ \sigma^2\left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]\right).$$

Επομένως, ένα $(1-\alpha)100\%$ διάστημα πρόβλεψης για την τιμή μιας νέας παρατήρησης για την Y για $X = x_0$ δίνεται από τη σχέση

$$\left(\hat{y}_{x_0} - t_{\alpha/2,n-2} \cdot s \cdot \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}},\ \hat{y}_{x_0} + t_{\alpha/2,n-2} \cdot s \cdot \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}\right).$$

Παράδειγμα 12.4

Για τα δεδομένα του Παραδείγματος 12.1 να κατασκευαστεί και να ερμηνευτεί ένα 95% διάστημα πρόβλεψης για τη δύναμη συνοχής Y μεταξύ του προωθητικού εκρηκτικού αερίου και του αερίου συντήρησης ενός εκτοξευτήρα 12.5 εβδομάδων.

Για τα ίδια δεδομένα να κατασκευαστεί και να ερμηνευτεί ένα 95% διάστημα εμπιστοσύνης για την αναμενόμενη τιμή της δύναμης συνοχής Y μεταξύ του προωθητικού εκρηκτικού αερίου και του αερίου συντήρησης ενός εκτοξευτήρα 12.5 εβδομάδων.

Λύση Παραδείγματος 12.4

Για την κατασκευή των ζητούμενων διαστημάτων είναι απαραίτητος πρώτα ο υπολογισμός

- της σημειακής εκτίμησης της $E(Y|X = x_0)$, $\hat{y}_{x_0}$,
- του άνω $\alpha/2$ ποσοστιαίου σημείου της t_{n-2} κατανομής, $t_{\alpha/2,n-2}$, για $n = 20$ και $\alpha = 0.05$,
- της εκτίμησης της τυπικής απόκλισης των τυχαίων σφαλμάτων, s,
- της ποσότητας $(x_0 - \bar{x})^2$ και
- της ποσότητας $\sum_{i=1}^{n}(x_i - \bar{x})^2$.

Αρκετές από αυτές τις ποσότητες τις έχουμε υπολογίσει σε προηγούμενα παραδείγματα. Ειδικότερα, η εκτίμηση της τυπικής απόκλισης των τυχαίων σφαλμάτων είναι $s = 96.10601$ (βλ. τη λύση του Παραδείγματος 12.3), η ποσότητα $\sum_{i=1}^{n}(x_i - \bar{x})^2 = 1106.562$ (βλ. τη λύση του Παραδείγματος 12.2), το άνω $0.05/2$ ποσοστιαίο σημείο της t_{18} κατανομής είναι $t_{0.05/2,18} = 2.101$ και η μέση τιμή $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} =$

13.3625 (κατά την εκτίμηση του β_0). Επομένως, αυτά που επιπλέον χρειάζεται να υπολογίσουμε είναι η σημειακή εκτίμηση της $E(Y|X = x_0)$, για $X = x_0 = 12.5$ και η ποσότητα $(x_0 - \bar{x})^2$. Η ποσότητα $(x_0 - \bar{x})^2$ ισούται με

$$(x_0 - \bar{x})^2 = (12.5 - 13.3625)^2 = 0.7439,$$

ενώ η $\hat{y}_{x_0}$ δίνεται από τη σχέση:

$$\hat{y}_{12.5} = 2627.8225 - 37.1536 \cdot 12.5 = 2163.4025.$$

Συνεπώς, ένα 95% διάστημα πρόβλεψης για τη δύναμη συνοχής Y μεταξύ του προωθητικού εκρηκτικού αερίου και του αερίου συντήρησης ενός εκτοξευτήρα 12.5 βδομάδων υπολογίζεται ως

$$\hat{y}_{x_0} \pm t_{\alpha/2,n-2} \cdot s \cdot \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}$$

$$2163.4025 \pm 2.101 \cdot 96.10601 \sqrt{1 + \frac{1}{20} + \frac{0.7439}{1106.562}}$$

το οποίο μας δίνει το διάστημα $(1956.4311,\ 2370.3739)$.
Ένα 95% διάστημα εμπιστοσύνης για την αναμενόμενη τιμή $E(Y|X = 12.5)$ υπολογίζεται ως

$$\hat{y}_{x_0} \pm t_{\alpha/2,n-2} \cdot s \cdot \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}$$

$$2163.4025 \pm 2.101 \cdot 96.10601 \sqrt{\frac{1}{20} + \frac{0.7439}{1106.562}},$$

το οποίο μας δίνει το διάστημα $(2117.9496,\ 2208.8554)$. Στη συνέχεια, παρατίθεται η ερμηνεία των δύο παραπάνω διαστημάτων εμπιστοσύνης.

Διάστημα πρόβλεψης: Είμαστε 95% σίγουροι ότι μια νέα μέτρηση της δύναμης συνοχής Y μεταξύ του προωθητικού εκρηκτικού αερίου και του αερίου συντήρησης ενός εκτοξευτήρα 12.5 εβδομάδων θα βρεθεί στο διάστημα $(1956.4311, 2370.3739)$.

Διάστημα εμπιστοσύνης: Είμαστε 95% σίγουροι ότι η πραγματική μέση τιμή της δύναμης συνοχής Y μεταξύ του προωθητικού εκρηκτικού αερίου και του αερίου συντήρησης ενός εκτοξευτήρα 12.5 βδομάδων είναι μεγαλύτερη από 2117.9496 και μικρότερη από 2208.8554.

Παρατήρηση 12.14

Το διάστημα πρόβλεψης είναι πάντα πλατύτερο από το αντίστοιχο διάστημα εμπιστοσύνης για το ίδιο επίπεδο εμπιστοσύνης και για το ίδιο x_0 (βλ. τους αντίστοιχους τύπους). Αυτό αποτυπώνει τη μεγαλύτερη αβεβαιότητα που έχουμε στην εκτίμηση/πρόβλεψη μιας μεμονωμένης μελλοντικής παρατήρησης σε σχέση με την εκτίμηση της αναμενόμενης τιμής της Y.

Τόσο το διάστημα πρόβλεψης όσο και το διάστημα εμπιστοσύνης δεν είναι παράλληλα με την ευθεία παλινδρόμησης. Αυτό οφείλεται στον όρο

$$\frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

που εμφανίζεται και στα δύο διαστήματα. Η ποσότητα αυτή μηδενίζεται (άρα ελαχιστοποιείται) όταν $x_0 = \bar{x}$. Επομένως, για $x_0 = \bar{x}$ τα προαναφερθέντα διαστήματα παρουσιάζουν το ελάχιστο πλάτος, το οποίο πλάτος αυξάνεται καθώς απομακρυνόμαστε από το $\bar{x}$.

12.4 Εφαρμογή στην R

Στο αρχείο δεδομένων `faithful` της βιβλιοθήκης `datasets` της R καταγράφονται ο χρόνος σε λεπτά μεταξύ διαδοχικών εκτοξεύσεων θερμού νερού και η διάρκεια σε λεπτά της κάθε έκρηξης του θερμοπίδακα Old Faithful του Εθνικού Πάρκου του Yellowstone στις ΗΠΑ.

Το συγκεκριμένο σύνολο δεδομένων μελετήθηκε αρχικά από τους Azzalini and Bowman (1990) και έχει γίνει αντικείμενο μελέτης από πολλούς ερευνητές. Μπορούμε να δούμε τη βασική του δομή με την εντολή:

```
1 str(faithful)
```

η οποία επιστρέφει το ακόλουθο αποτέλεσμα:

```
------------------------------------------------------------
'data.frame': 272 obs. of  2 variables:
 $ eruptions: num  3.6 1.8 3.33 2.28 4.53 ...
 $ waiting  : num  79 54 74 62 85 55 88 85 51 85 ...
 ------------------------------------------------------------
```

Στον παραπάνω πίνακα αποτελεσμάτων φαίνεται ότι το σύνολο δεδομένων περιλαμβάνει 272 μετρήσεις σε δύο μεταβλητές, οι οποίες είναι συνεχείς ποσοτικές και λαμβάνουν τιμές στο $(0, \infty)$.

Για να εξετάσουμε αν οι δύο μεταβλητές σχετίζονται γραμμικά μεταξύ τους και, επομένως, αν το απλό γραμμικό μοντέλο είναι κατάλληλο για να περιγράφει τη σχέση τους, αρχικά μπορούμε να υπολογίσουμε τον συντελεστή γραμμικής συσχέτισης του Pearson και να κατασκευάσουμε το διάγραμμα διασκόρπισης ή διασποράς, τα οποία μπορούμε να λάβουμε μέσω των ακόλουθων εντολών της R:

```
1 cor(faithful$waiting,faithful$eruptions)
2 plot(faithful$waiting,faithful$eruptions, col='blue', pch=20, cex=1.2,
3      main = "faithful data: Eruptions of Old Faithful",
4      ylab = "Eruption time (min)",
5      xlab = "Waiting time to next eruption (min)")
```

Η τιμή του συντελεστή γραμμικής συσχέτισης του Pearson που επιστρέφει η R για τις δύο μεταβλητές ισούται με 0.9008112, υποδηλώνοντας μια έντονη θετική συσχέτιση μεταξύ των δύο μεταβλητών. Η θετική αυτή συσχέτιση αποτυπώνεται και στο διάγραμμα διασκόρπισης του Σχήματος 12.2.

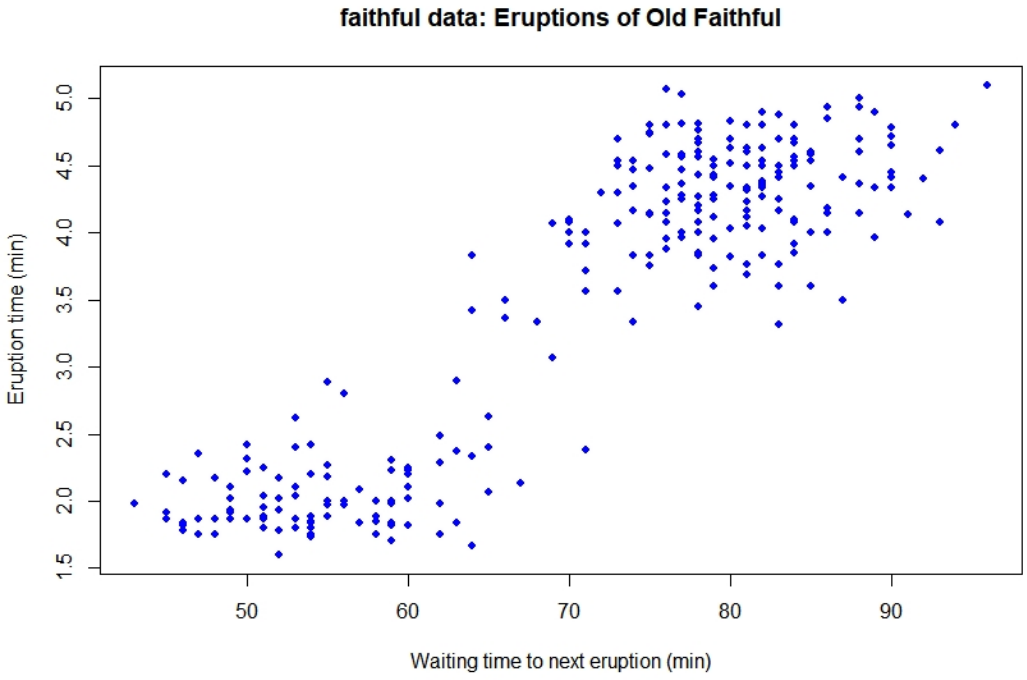
Η προσαρμογή του απλού γραμμικού μοντέλου γίνεται με τις παρακάτω εντολές

```
1 linearMod <- lm(eruptions ~ waiting , data=faithful)
2 summary(linearMod)
```

Η δεύτερη εντολή επιστρέφει τα ακόλουθα συγκεντρωτικά αποτελέσματα της ανάλυσης.

```
------------------------------------------------------------
Call:
lm(formula = eruptions ~ waiting, data = faithful)

Residuals:
     Min       1Q   Median       3Q      Max
```



Σχήμα 12.2: Διάγραμμα διασκόρπισης των δεδομένων του χρόνου σε λεπτά μεταξύ διαδοχικών εκτοξεύσεων θερμού νερού και της διάρκειας σε λεπτά της κάθε έκρηξης του θερμοπίδακα Old Faithful του Εθνικού Πάρκου του Yellowstone στις ΗΠΑ.

```
-1.29917 -0.37689  0.03508  0.34909  1.19329

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.874016   0.160143  -11.70   <2e-16 ***
waiting      0.075628   0.002219   34.09   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4965 on 270 degrees of freedom
Multiple R-squared:  0.8115,Adjusted R-squared:  0.8108
F-statistic:  1162 on 1 and 270 DF,  p-value: < 2.2e-16
-----------------------------------------------------------------
```

Στα παραπάνω αποτελέσματα παρουσιάζονται

- κάποια περιγραφικά στατιστικά στοιχεία για τα υπόλοιπα, τα οποία μας επιτρέπουν να κάνουμε μια πρώτη, αδρή αξιολόγηση του μοντέλου (πεδίο: Residuals),
- ο πίνακας με τις εκτιμήσεις των συντελεστών του μοντέλου μαζί με τους ελέγχους t των υποθέσεων

$$\mathrm{H}_0 : \beta_i = 0 \quad \text{έναντι} \quad \mathrm{H}_1 : \beta_i \neq 0$$

  για τις παραμέτρους β_0 και β_1 (πεδίο: Coefficients), καθώς και
- κάποιες επιπλέον πληροφορίες, όπως η τιμή του συντελεστή προσδιορισμού R^2 και η τιμή του ελέγχου F.

Αρχικά, τα υπόλοιπα φαίνεται να παρουσιάζουν μια σχετική συμμετρία γύρω από το μηδέν, η οποία είναι συμβατή με την υπόθεση της κανονικότητας των τυχαίων σφαλμάτων με μέση τιμή 0. Από το πεδίο Coefficients έχουμε ότι η εκτιμώμενη ευθεία παλινδρόμησης είναι η

$$\hat{y} = -1.874016 + 0.075628x.$$

Στην προτελευταία στήλη του πεδίου Coefficients εμφανίζονται οι τιμές των στατιστικών συναρτήσεων

$$T_i = \frac{\hat{\beta}_i - \beta_i}{se(\hat{\beta}_i)} \sim t_{n-2},\ i = 1,2,$$

για τον έλεγχο της υπόθεσης $H_0 : \beta_i = 0$ έναντι της $H_1 : \beta_i \neq 0$.

Για τον υπολογισμό των τιμών αυτών των στατιστικών συναρτήσεων ελέγχων έχουν χρησιμοποιηθεί οι εκτιμήσεις, $se(\hat{\beta}_i)$, των τυπικών αποκλίσεων των εκτιμητριών $\hat{\beta}_i$, οι οποίες εμφανίζονται στη στήλη Std. Error. Οι τιμές των στατιστικών συναρτήσεων ελέγχων μπορούν να συγκριθούν με τις αντίστοιχες κρίσιμες τιμές από τους πίνακες της t κατανομής, έτσι ώστε να αποφανθούμε αν οι μηδενικές υποθέσεις απορρίπτονται ή όχι σε επίπεδο σημαντικότητας α. Εναλλακτικά, η R προσφέρει την τιμή του παρατηρούμενου επιπέδου σημαντικότητας - (p-value)- για καθέναν έλεγχο στην τελευταία στήλη υπό τον τίτλο Pr(>|t|). Για τα δεδομένα του προβλήματος παρατηρούμε ότι το παρατηρούμενο επίπεδο σημαντικότητας για καθέναν έλεγχο είναι μικρότερο από οποιαδήποτε συνηθισμένη τιμή του α (π.χ. 0.05 ή 0.01). Επομένως, μπορούμε να πούμε ότι οι μηδενικές υποθέσεις $\mathrm{H}_0 : \beta_i = 0,\ i = 1,2$, απορρίπτονται σε οποιοδήποτε συνηθισμένο επίπεδο σημαντικότητας.

Στην τελευταία ενότητα των αποτελεσμάτων η R εμφανίζει

- την εκτίμηση της τυπικής απόκλισης των τυχαίων σφαλμάτων (Residual standard error), η οποία ισούται με $s = 0.4965$,
- την τιμή του συντελεστή προσδιορισμού $R^2 = 0.8115$ (Multiple R-squared) και
- το παρατηρούμενο επίπεδο σημαντικότητας του ελέγχου F ($p - value :< 2.2e - 16$).

Η τυπική απόκλιση των τυχαίων σφαλμάτων είναι η τρίτη παράμετρος του μοντέλου (μετά από τις β_0 και β_1) και για αυτό παρατίθεται η εκτίμησή της, ενώ από την τιμή του συντελεστή προσδιορισμού συμπεραίνουμε ότι το 81.15% της μεταβλητότητας που παρατηρείται στη διάρκεια των εκρήξεων του θερμοπίδακα Old Faithful του Εθνικού Πάρκου του Yellowstone στις ΗΠΑ ερμηνεύεται από τον χρόνο που μεσολαβεί μεταξύ των διαδοχικών εκτοξεύσεων θερμού νερού. Με βάση τον έλεγχο F καταλήγουμε στο ίδιο συμπέρασμα με τον έλεγχο t για το β_1, δηλαδή ότι απορρίπτεται η μηδενική υπόθεση $\mathrm{H}_0 : \beta_1 = 0$ σε οποιοδήποτε συνηθισμένο επίπεδο σημαντικότητας.

Στα παραπάνω αποτελέσματα δεν παρουσιάζεται αναλυτικά ο πίνακας ανάλυσης διασποράς του μοντέλου, αλλά μόνο το παρατηρούμενο επίπεδο σημαντικότητας του ελέγχου F. Για να λάβουμε τον πίνακα ανάλυσης διασποράς, αρκεί να πληκτρολογήσουμε την ακόλουθη εντολή:

```
1 anova(linearMod)
```

η οποία επιστρέφει τον πίνακα ανάλυσης διασποράς με, πρακτικά, όλα τα στοιχεία. Η διαφορά με τον πίνακα που παρουσιάσαμε στην προηγούμενη ενότητα είναι η απουσία του ολικού αθροίσματος τετραγώνων SST και η σειρά των δύο πρώτων στηλών.

```
-------------------------------------------------------------
Analysis of Variance Table

Response: eruptions
           Df  Sum Sq Mean Sq F value    Pr(>F)
waiting     1 286.478 286.478  1162.1 < 2.2e-16 ***
Residuals 270  66.562   0.247
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
-------------------------------------------------------------
```

Επιπλέον, 95% διαστήματα εμπιστοσύνης για τις παραμέτρους β_0 και β_1 μπορούν να ληφθούν με την εντολή

```
1 confint(linearMod, level = 0.95)
```

η οποία επιστρέφει τα ακόλουθα 95% διαστήματα εμπιστοσύνης για τα β_0 και β_1.

```
                2.5 %      97.5 %
(Intercept) -2.18930436 -1.55872761
waiting      0.07126011  0.07999579
```

Από τα παραπάνω αποτελέσματα παρατηρούμε, παραδείγματος χάριν, ότι το μηδέν δεν ανήκει στο διάστημα εμπιστοσύνης για τον συντελεστή, β_1, της επεξηγηματικής μεταβλητής με συνέπεια και με αυτήν τη μέθοδο να απορρίπτεται, σε επίπεδο σημαντικότητας 0.05, η μηδενική υπόθεση $\mathrm{H}_0 : \beta_1 = 0$.

Κλείνοντας την ενότητα, θα πρέπει να αναφερθούμε στον τρόπο υπολογισμού με τη βοήθεια της R των διαστημάτων εμπιστοσύνης και πρόβλεψης για συγκεκριμένες τιμές της επεξηγηματικής μεταβλητής. Παραδείγματος χάριν με τις εντολές

```
1 new <- data.frame(waiting = c(50,70,90))
2
3 distConf <- predict(linearMod, newdata=new,interval = "confidence", level
     =0.95)
4 distPred <- predict(linearMod, newdata=new,interval = "prediction", level
     =0.95)
5 cbind(new,distConf)
6 cbind(new,distPred)
```

η R επιστρέφει τα 95% διαστήματα εμπιστοσύνης (distConf) και πρόβλεψης (distPred) για τις τιμές 50, 70 και 90 της επεξηγηματικής μεταβλητής. Πιο συγκεκριμένα, η εντολή cbind(new,distConf) επιστρέφει τον ακόλουθο πίνακα αποτελεσμάτων με τις σημειακές εκτιμήσεις (δίνονται υπό τη στήλη fit), τα κάτω και άνω όρια των 95% διαστημάτων εμπιστοσύνης (δίνονται υπό τις στήλες lwr και upr, αντίστοιχα) για την πραγματική μέση τιμή της διάρκειας των εκρήξεων του θερμοπίδακα Old Faithful του Εθνικού Πάρκου του Yellowstone στις ΗΠΑ, όταν ο χρόνος που μεσολαβεί μεταξύ των διαδοχικών εκτοξεύσεων θερμού νερού είναι ίσος με 50, 70 και 90 λεπτά.

```
------------------------------------------------------------
  waiting      fit      lwr      upr
1      50 1.907381 1.798550 2.016213
2      70 3.419940 3.360540 3.479341
3      90 4.932499 4.830151 5.034847
------------------------------------------------------------
```

Προκύπτει λοιπόν, για παράδειγμα, ότι είμαστε 95% σίγουροι ότι η πραγματική μέση τιμή της διάρκειας των εκρήξεων του θερμοπίδακα Old Faithful του Εθνικού Πάρκου του Yellowstone στις ΗΠΑ, όταν ο χρόνος που μεσολαβεί μεταξύ των διαδοχικών εκτοξεύσεων θερμού νερού είναι ίσος με 70 λεπτά, είναι μεγαλύτερη από 3.360540 και μικρότερη από 3.479341 λεπτά.

Η εντολή `cbind(new,distPred)` επιστρέφει τον ακόλουθο πίνακα αποτελεσμάτων με τις σημειακές πάλι εκτιμήσεις (δίνονται υπό τη στήλη fit), αλλά αυτήν τη φορά με τα κάτω και άνω όρια των 95% διαστημάτων πρόβλεψης (δίνονται υπό τις στήλες lwr και upr, αντίστοιχα) για μια νέα μέτρηση της διάρκειας των εκρήξεων του θερμοπίδακα Old Faithful, όταν ο χρόνος που μεσολαβεί μεταξύ των διαδοχικών εκτοξεύσεων θερμού νερού είναι ίσος με 50, 70 και 90 λεπτά.

```
------------------------------------------------------------
  waiting      fit       lwr      upr
1      50 1.907381 0.9238126 2.890950
2      70 3.419940 2.4406080 4.399273
3      90 4.932499 3.9496268 5.915372
------------------------------------------------------------
```

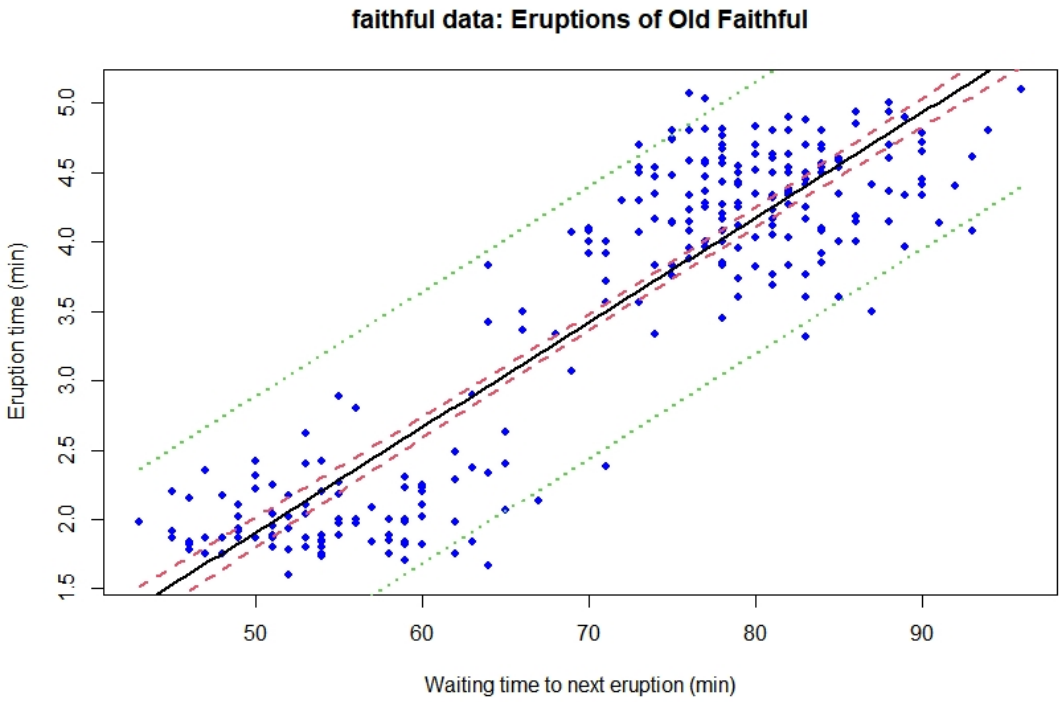
Προκύπτει επομένως, για παράδειγμα, ότι είμαστε 95% σίγουροι ότι μια νέα μέτρηση της διάρκειας των εκρήξεων του θερμοπίδακα Old Faithful, όταν ο χρόνος που μεσολαβεί μεταξύ των διαδοχικών εκτοξεύσεων θερμού νερού είναι ίσος με 70 λεπτά, θα είναι μεγαλύτερη από 2.4406080 και μικρότερη από 4.399273 λεπτά.

Παρατηρήστε ότι τα διαστήματα πρόβλεψης έχουν, όπως το αναμέναμε, μεγαλύτερο πλάτος από τα αντίστοιχα διαστήματα εμπιστοσύνης. Αυτό μπορεί να αποτυπωθεί και γραφικά σε όλο το εύρος των τιμών της επεξηγηματικής μεταβλητής με τη βοήθεια των παρακάτω εντολών:

```
new <- data.frame(waiting = seq(43, 96, 0.25))
pred.w.plim <- predict(linearMod, new, interval = "prediction")
pred.w.clim <- predict(linearMod, new, interval = "confidence")
matplot(new$waiting, cbind(pred.w.clim, pred.w.plim[,-1]),
        col =  c(1,2,2,3,3),lty = c(1,2,2,3,3),lwd=2, type = "l", add=TRUE)
```

Οι εντολές αυτές κατασκευάζουν το γράφημα του Σχήματος 12.2 προσθέτοντας στο προηγούμενο διάγραμμα διασκόρπισης (Σχήμα 12.3):

- την εκτιμώμενη ευθεία παλινδρόμησης (μαύρη γραμμή),
- τα 95% διαστήματα εμπιστοσύνης (κόκκινες γραμμές) και
- τα 95% διαστήματα πρόβλεψης (πράσινες γραμμές).



Σχήμα 12.3: Τα 95% διαστήματα εμπιστοσύνης και πρόβλεψης της διάρκειας σε λεπτά της κάθε έκρηξης του θερμοπίδακα Old Faithful του Εθνικού Πάρκου του Yellowstone στις ΗΠΑ σε σχέση με τον χρόνο σε λεπτά μεταξύ διαδοχικών εκτοξεύσεων θερμού νερού.

Από την εικόνα του γραφήματος είναι φανερό ότι:

- τα διαστήματα πρόβλεψης έχουν μεγαλύτερο πλάτος σε σχέση με τα διαστήματα πρόβλεψης της εμπιστοσύνης, και
- το πλάτος των διαστημάτων αυξάνεται καθώς απομακρυνόμαστε από τη μέση τιμή (αυτό εντοπίζεται πιο εύκολα παρατηρώντας τα 95% διαστήματα εμπιστοσύνης και συγκρίνοντας το κέντρο με τα άκρα του γραφήματος).

12.5 Ασκήσεις

Άσκηση 12.1 *Χρησιμοποιώντας τα δεδομένα της Άσκησης Αυτοαξιολόγησης 12.2 να κατασκευάσετε και να ερμηνεύσετε ένα 95% διάστημα εμπιστοσύνης για την παράμετρο* β_1 *.*

Άσκηση 12.2 *Για τα δεδομένα της Άσκησης Αυτοαξιολόγησης 12.2 να εξετάσετε με τη βοήθεια του ελέγχου* F *και σε επίπεδο σημαντικότητας 0.05 αν είναι σημαντική η παλινδρόμηση. Στη συνέχεια να υπολογίσετε τον συντελεστή προσδιορισμού και να δώσετε την ερμηνεία της τιμής του.*

Άσκηση 12.3 *Η χρήση εμφανούς σκυροδέματος κατασκευασμένου από ομοιόμορφα διαβαθμισμένο αδρανές υλικό και τσιμεντοκονίαμα ενδείκνυται σε περιοχές με έντονες βροχοπτώσεις λόγω των εξαιρετικών ιδιοτήτων αποστράγγισης που έχει.*

Για τον προσδιορισμό της σχέσης μεταξύ του y *=πορώδες του εδάφους (αντιπροσωπεύει το ποσοστό του όγκου του εδάφους το οποίο καταλαμβάνουν οι πόροι, όπου βρίσκεται η υγρή και η αέρια φάση του εδάφους και υπολογίζεται ως ο λόγος του όγκου των κενών πόρων προς τον συνολικό όγκο, σε %) και του* x *=ειδικού βάρους (σε* pcf*) του σκυροδέματος (ειδικό βάρος: ο λόγος του βάρους ενός σώματος προς τον όγκο αυτού) μελετήθηκαν 17 δοκίμια εμφανούς σκυροδέματος. Στη συνέχεια, παρουσιάζονται τα αποτελέσματα των μετρήσεων και τα συγκεντρωτικά στοιχεία τους:*

x	99.0	101.1	102.7	103.0	105.4	107.0	108.7	110.8	112.1
y	28.8	27.9	27.0	25.2	22.8	21.5	20.9	19.6	17.1
x	112.4	113.6	113.8	115.1	115.4	120.0	117.5	104.0	
y	18.9	16.0	16.7	13.0	13.6	10.8	11.1	24.8	

Δίνεται ότι

$\sum_{i=1}^{n} x_i = 1861.6$ $\sum_{i=1}^{n} y_i = 335.7$ $\sum_{i=1}^{n} (x_i - \bar{x})^2 = 615.83$ $\sum_{i=1}^{n} (y_i - \bar{y})^2 = 539.22$
$\sum_{i=1}^{n} x_i^2 = 204471.98$ $\sum_{i=1}^{n} y_i^2 = 7168.31$ $\sum_{i=1}^{n} x_i y_i = 36192.04$ $n = 17$
άθροισμα τετραγώνων υπολοίπων$= 13.33$.

α. Ποια από τις παρακάτω είναι η εκτίμηση $\hat{y} = b_0 + b_1 x$ *της ευθείας παλινδρόμησης* $\mathrm{E}(Y) = \beta_0 + \beta_1 x$*;*

- $\hat{y} = 120.941 - 0.924095x$
- $\hat{y} = 120.941 + 0.924095x$
- $\hat{y} = -81.4468 - 0.924095x$
- $\hat{y} = -81.4468 + 0.924095x$

β. Δώστε την ερμηνεία της εκτίμησης της ευθείας παλινδρόμησης.

γ. Δώστε την ερμηνεία του συντελεστή προσδιορισμού.

δ. Αν θέλουμε να εξετάσουμε τη σημαντικότητα της παλινδρόμησης, ποια/ποιες από τις παρακάτω υποθέσεις πρέπει να εξετάσουμε;

- $H_0 : \beta_0 = 0$ *έναντι* $H_1 : \beta_0 \neq 0$
- $H_0 : \hat{\beta}_0 = 0$ *έναντι* $H_1 : \hat{\beta}_0 \neq 0$
- $H_0 : \hat{\beta}_0 = \hat{\beta}_1$ *έναντι* $H_1 : \hat{\beta}_0 \neq \hat{\beta}_1$
- $H_0 : \beta_1 = 0$ *έναντι* $H_1 : \beta_1 \neq 0$
- $H_0 : \hat{\beta}_1 = 0$ *έναντι* $H_1 : \hat{\beta}_1 \neq 0$
- $H_0 : \beta_0 = \beta_1$ *έναντι* $H_1 : \beta_0 \neq \beta_1$

ε. Αν η τιμή του ελέγχου t για τη σημαντικότητα του μοντέλου ισούται με -24.32, τότε ποια/ποιες από τις παρακάτω προτάσεις είναι αληθής/αληθείς;

- *Η παλινδρόμηση είναι σημαντική σε επίπεδο σημαντικότητας 0.01.*
- *Η παλινδρόμηση δεν είναι σημαντική σε επίπεδο σημαντικότητας 0.01.*
- *Τα στοιχεία που μας δίνονται (στη συγκεκριμένη ερώτηση) δεν αρκούν, για να δώσουμε μια ξεκάθαρη απάντηση.*

στ. Δώστε την ερμηνεία ενός 96% διαστήματος εμπιστοσύνης για τη μέση τιμή του πορώδους εδάφους, όταν το ειδικό βάρος είναι $100pcf$.

Άσκηση 12.4 *Σε ένα σημείο ενός αυτοκινητόδρομου έχει εγκατασταθεί ένα σύστημα παρακολούθησης της μέσης ταχύτητας Y και του αριθμού X των αυτοκινήτων που διέρχονται από το σημείο κάθε ώρα. Τα συγκεντρωτικά στοιχεία που παρουσιάζονται στη συνέχεια αφορούν ένα τυχαία επιλεγμένο δείγμα 20 ωρών.*

$$n = 20 \qquad \sum_{i=1}^{n} x_i = 12090 \qquad \sum_{i=1}^{n} y_i = 1924.87$$

$$\sum_{i=1}^{n} x_i^2 = 7837148 \qquad \sum_{i=1}^{n} y_i^2 = 189187.08 \qquad \sum_{i=1}^{n} x_i y_i = 1118277.29$$

1. *Το ολικό άθροισμα τετραγώνων για το απλό γραμμικό μοντέλο ισούται με 3930.6, ενώ το άθροισμα τετραγώνων των υπολοίπων υπολογίστηκε ίσο με 48.3. Με τη βοήθεια των στοιχείων αυτών*
   - *i. εξετάστε αν σε επίπεδο σημαντικότητας 0.01 μπορούμε να ισχυριστούμε ότι η κλίση της ευθείας είναι διάφορη του μηδενός,*
   - *ii. υπολογίστε τον συντελεστή προσδιορισμού. Τι εκφράζει η τιμή αυτού του συντελεστή;*
2. *Εκτιμήστε (i) σημειακά και (ii) με ένα 95% διάστημα εμπιστοσύνης την αναμενόμενη μέση ταχύτητα των αυτοκινήτων, όταν από το σημείο διέρχονται 1000 αυτοκίνητα ανά ώρα. Ποια είναι η ερμηνεία του διαστήματος εμπιστοσύνης;*

Άσκηση 12.5 *Η κλίση του κεκλιμένου Πύργου της Πίζας συνεχίζει να αυξάνεται με τον χρόνο. Μάλιστα, σε μελέτη που πραγματοποιήθηκε τα έτη 1975-1987 (13 έτη) προσδιορίστηκε ότι η απόσταση (σε μέτρα) ενός συγκεκριμένου σημείου του από το σημείο όπου θα βρισκόταν, αν ο πύργος ήταν στη σωστή του θέση, σχετίζεται γραμμικά με το έτος που έγινε η μέτρηση. Πιο συγκεκριμένα, από τις μετρήσεις των 13 ετών εκτιμήθηκε η ακόλουθη ευθεία παλινδρόμησης*

$$y = 0.0009x + 1.1233$$

για τη μέση απόσταση του σημείου από το σημείο όπου θα βρισκόταν, αν ο πύργος ήταν στη σωστή του θέση κατά τη διάρκεια του έτους.

α. Ερμηνεύστε στο πλαίσιο του προβλήματος την εκτιμώμενη ευθεία παλινδρόμησης και τον συντελεστή της ανεξάρτητης μεταβλητής.

β. Το ολικό άθροισμα τετραγώνων υπολογίστηκε ίσο με 0.00015997, ενώ το άθροισμα τετραγώνων των υπολοίπων υπολογίστηκε ίσο με 0.00000192.

β1. Με βάση τα στοιχεία αυτά υπολογίστε τον συντελεστή προσδιορισμού και δώστε την ερμηνεία του.

β2. Ποια είναι η τιμή του συντελεστή γραμμικής συσχέτισης του Pearson; Αιτιολογήστε την απάντησή σας.

γ. Είναι στατιστικά σημαντική η παλινδρόμηση σε επίπεδο σημαντικότητας 0.05;

δ. Αν υποθέσουμε ότι η αύξηση της κλίσης του κεκλιμένου Πύργου της Πίζας συνεχίστηκε και τα επόμενα έτη με τον ίδιο ρυθμό (μια υπόθεση που συμβαδίζει με τα ιστορικά στοιχεία) και ότι δεν έγινε καμία παρέμβαση συντήρησης στον Πύργο από το 1987 μέχρι σήμερα (κάτι το οποίο στην πραγματικότητα δεν αληθεύει), τότε για να εκτιμήσουμε ένα διάστημα μέσα στο οποίο θα βρισκόταν το 2014 με μεγάλη εμπιστοσύνη (π.χ. 95%) η μέση απόσταση (σε μέτρα) του συγκεκριμένου σημείου του Πύργου από το σημείο που θα βρισκόταν, αν ο πύργος ήταν στη σωστή του θέση, θα έπρεπε να κατασκευάσουμε ένα διάστημα εμπιστοσύνης ή ένα διάστημα πρόβλεψης; Δικαιολογήστε πλήρως την απάντησή σας.

ε. Κατασκευάστε και ερμηνεύστε ένα 99% δ.ε. για τον συντελεστή του έτους.
(Δίνεται ότι $\sum_{i=1}^{n}(x_i - \bar{x})^2 = 182$ *και* $\sum_{i=1}^{n} x_i^2 = 51016875$*).*

Άσκηση 12.6 *Σε έρευνα αγοράς που έγινε σε 8 μεγάλες πόλεις των ΗΠΑ και αφορούσε τη μέση (διετή) κατανάλωση κρασιού ανά κάτοικο (σε φιάλες) ανάλογα με την τιμή πώλησής του (σε δολάρια), προέκυψαν τα ακόλουθα αποτελέσματα:*

Κατανάλωση (σε τεμάχια)	*75*	*86*	*82*	*67*	*80*	*89*	*74*	*95*
Τιμή πώλησης (σε $)	*37*	*31*	*32*	*38*	*34*	*30*	*36*	*28*

α. Με τη βοήθεια της R προσδιορίστε την εκτιμώμενη ευθεία παλινδρόμησης. Ερμηνεύστε στο πλαίσιο του προβλήματος την εκτιμώμενη ευθεία παλινδρόμησης και τον συντελεστή της ανεξάρτητης μεταβλητής.

β. Τι ποσοστό της ολικής μεταβλητότητας της μέσης διετούς κατανάλωσης κρασιού ερμηνεύεται από την παλινδρόμηση;

γ. Ελέγξτε αν η παλινδρόμηση είναι σημαντική σε επίπεδο σημαντικότητας $\alpha = 0.01$ *με χρήση κατάλληλης κρίσιμης περιοχής.*

δ. Βρείτε και ερμηνεύστε το 95% δ.ε. για τη μέση (διετή) κατανάλωση κρασιού αν η τιμή πώλησης της φιάλης διαμορφωθεί στα 33 δολάρια.

Άσκηση 12.7 *Υπό τις υποθέσεις του απλού γραμμικού μοντέλου, να αποδείξετε ότι οι εκτιμήτριες ελαχίστων τετραγώνων* $\hat{\beta}_0$ *και* $\hat{\beta}_1$ *είναι αμερόληπτες εκτιμήτριες των* β_0 *και* β_1*, αντίστοιχα. Τι παρατηρείτε σχετικά με την αναγκαιότητα της ικανοποίησης όλων των υποθέσεων;*

Άσκηση 12.8 *Υπό τις υποθέσεις του απλού γραμμικού μοντέλου, να αποδείξετε ότι*

$$Var(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

και

$$Var(\hat{\beta}_0) = \frac{\sigma^2 \sum_{i=1}^{n} x_i^2}{n\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

Τι παρατηρείτε σχετικά με την αναγκαιότητα της ικανοποίησης όλων των υποθέσεων;

Άσκηση 12.9 *Υπό τις υποθέσεις του απλού γραμμικού μοντέλου, να αποδείξετε ότι*

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right),$$

και

$$\hat{\beta}_0 \sim N\left(\beta_0, \frac{\sigma^2 \sum_{i=1}^n x_i^2}{n \sum_{i=1}^n (x_i - \bar{x})^2}\right).$$

Τι παρατηρείτε σχετικά με την αναγκαιότητα της ικανοποίησης όλων των υποθέσεων;

Άσκηση 12.10 *Υπό τις υποθέσεις του απλού γραμμικού μοντέλου, να αποδείξετε ότι το μέσο άθροισμα τετραγώνων των υπολοίπων είναι αμερόληπτη εκτιμήτρια της διακύμανσης των σφαλμάτων.*

Άσκηση 12.11 *Υπό τις υποθέσεις του απλού γραμμικού μοντέλου, να αποδείξετε ότι:*

$$Cov(\hat{\beta}_0, \hat{\beta}_1) = -\frac{\bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2}\sigma^2.$$

Πότε είναι ασυσχέτιστες οι εκτιμήτριες των παραμέτρων;

Άσκηση 12.12 *Να αποδείξετε ότι στην περίπτωση του απλού γραμμικού ελέγχου ο έλεγχος t και ο έλεγχος F είναι ισοδύναμοι.*

Άσκηση 12.13 *Να αποδείξετε ότι*

$$SSR = \hat{\beta}_1^2 \sum_{i=1}^n (x_i - \bar{x})^2 = \hat{\beta}_1 \left(\sum_{i=1}^n x_i y_i - \frac{\sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n} \right).$$

Άσκηση 12.14 *Να αποδείξετε ότι στο απλό γραμμικό μοντέλο* $R^2 = r_{xy}^2$, *όπου* R^2 *ο συντελεστής προσδιορισμού και* r_{xy}^2 *ο συντελεστής συσχέτισης του Pearson.*

Άσκηση 12.15 *Στο απλό γραμμικό μοντέλο παλινδρόμησης να δείξετε ότι*

$$\sum_{i=1}^n x_i \cdot (y_i - \hat{y}_i) = 0 \text{ και } \sum_{i=1}^n y_i = \sum_{i=1}^n \hat{y}_i.$$

Επιπλέον, να δείξετε ότι το σημείο $(\bar{x}, \bar{y})$ *είναι σημείο της εκτιμώμενης ευθείας παλινδρόμησης.*

Άσκηση 12.16 *Στο απλό γραμμικό μοντέλο παλινδρόμησης ποια συνθήκη πρέπει να ικανοποιείται έτσι ώστε ένας εκτιμητής του* β_1 *που είναι γραμμικός συνδυασμός των* Y_i *να είναι αμερόληπτος; Πότε αυτός έχει τη μικρότερη διακύμανση; Ποιο γνωστό αποτέλεσμα της βιβλιογραφίας προέκυψε;*

12.6 Απαντήσεις στις ασκήσεις αυτοαξιολόγησης

Λύση Άσκησης Αυτοαξιολόγησης 12.1

Από την εικόνα του διαγράμματος διασκόρπισης παρατηρούμε ότι υπάρχει μια αρνητική συσχέτιση μεταξύ των δύο μεταβλητών. Αυτό σημαίνει ότι η εκτίμηση του β_1 θα πρέπει να είναι αρνητική. Αυτό περιορίζει τις επιλογές μας στις δύο πρώτες ευθείες. Οι δύο αυτές ευθείες διαφοροποιούνται ως προς την εκτίμηση του σταθερού όρου. Στην πρώτη ευθεία η εκτιμώμενη τιμή του σταθερού όρου είναι θετική και ίση με 1.9699, ενώ στη δεύτερη ευθεία η εκτιμώμενη τιμή είναι αρνητική και ίση με -1.9699. Από την εικόνα πάλι του γραφήματος καταλήγουμε ότι η εκτιμώμενη ευθεία τέμνει τον κατακόρυφο άξονα ανάμεσα στο 1.95 και το 2 και, επομένως, η επιλογή μας είναι η ευθεία $\hat{y} = 1.9699 - 0.1962x$.
Για την ερμηνεία της ευθείας παλινδρόμησης μπορούμε να πούμε ότι:

- η εκτιμώμενη ευθεία αποτελεί την εκτίμησή μας για την αναμενόμενη τιμή της Y για οποιαδήποτε τιμή (μέσα στο εύρος των παρατηρούμενων τιμών) της X,
- η αναμενόμενη τιμή της Y μειώνεται κατά 0.1962 μονάδες για κάθε μοναδιαία αύξηση της X και, τέλος,
- η αναμενόμενη τιμή της Y ισούται με 1.9699 για $x = 0$.

Λύση Άσκησης Αυτοαξιολόγησης 12.2

Αφού η εκτίμηση ελαχίστων τετραγώνων του συντελεστή της επεξηγηματικής μεταβλητής ισούται με -0.085689, έχουμε ότι $\hat{\beta}_1 = -0.085687$. Επομένως, από τη σχέση (12.2), η εκτιμήτρια ελαχίστων τετραγώνων της παραμέτρου β_0 ισούται με:

$$\hat{\beta}_0 = \frac{1924.87}{20} - (-0.085687)\frac{12090}{20} = 148.041.$$

Συνεπώς, η εκτιμώμενη ευθεία παλινδρόμησης είναι η:

$$\hat{y} = 148.041 - 0.085687x,$$

η οποία εκφράζει την εκτίμηση της αναμενόμενης τιμής της μέσης ταχύτητας Y σε σχέση με τον αριθμό X των αυτοκινήτων που διέρχονται από το σημείο κάθε ώρα. Επιπροσθέτως, η τιμή $\hat{\beta}_1 = -0.085687$ εκφράζει την αναμενόμενη μεταβολή (μείωση) της μέσης ταχύτητας για καθένα επιπλέον αυτοκίνητο που διέρχεται από το σημείο κατά τη διάρκεια μίας ώρας. Η εκτίμηση του σταθερού όρου β_0 δεν έχει νόημα να ερμηνευτεί, αφού δεν γίνεται να έχουμε μέση ταχύτητα από μηδενικό αριθμό αυτοκινήτων.

Λύση Άσκησης Αυτοαξιολόγησης 12.3

Οι υποθέσεις που ζητείται να εξετάσουμε, σε επίπεδο σημαντικότητας 0.05, είναι οι

$$\mathrm{H}_0 : \beta_0 = 2600 \quad \text{έναντι} \quad \mathrm{H}_1 : \beta_1 \neq 2600.$$

Η στατιστική συνάρτηση του ελέγχου είναι η

$$\mathrm{T}_0 = \frac{\hat{\beta}_0 - \beta_0}{se(\hat{\beta}_0)},$$

η οποία υπό τη μηδενική υπόθεση εκφράζεται ως

$$\mathrm{T}_0 = \frac{\hat{\beta}_0 - 2600}{se(\hat{\beta}_0)}$$

και κατανέμεται σύμφωνα με την $t_{20-2} = t_{18}$ κατανομή. Η κρίσιμη περιοχή του ελέγχου είναι η

$$C = \{|T_0| > t_{18,0.05/2}\} = \{|T_0| > 2.101\}.$$

Η παρατηρούμενη τιμή $\mathrm{T}_{0,obs}$, της στατιστικής συνάρτησης ελέγχου T_0 για τα δεδομένα του παραδείγματος ισούται με

$$\mathrm{T}_{0,obs} = \frac{\hat{\beta}_0 - 2600}{se(\hat{\beta}_0)} = \frac{2627.8225 - 2600}{96.1061\sqrt{\frac{4677.69}{20 \cdot 1106.562}}} = 0.63,$$

καθώς

$$\sum_{i=1}^{n} (x_i - \bar{x})^2 = 1106.562.$$

Επομένως, αφού

$$|\mathrm{T}_{0,obs}| = 0.63 < 2.101 = t_{18,0.05/2}$$

η μηδενική υπόθεση $\mathrm{H}_0 : \beta_0 = 2600$ δεν απορρίπτεται σε επίπεδο σημαντικότητας 0.05.

Λύση Άσκησης Αυτοαξιολόγησης 12.4

Αρχικά, παρατηρούμε ότι ο συντελεστής γραμμικής συσχέτισης του Pearson πρέπει να είναι αρνητικός, αφού οι δύο μεταβλητές σχετίζονται αρνητικά μεταξύ τους. Αυτή η παρατήρηση περιορίζει τις επιλογές στη δεύτερη και στην τρίτη τιμή. Από τις δύο αυτές τιμές πιο συμβατή με τα δεδομένα του γραφήματος είναι η τρίτη, δηλαδή η -0.99352, που υποδηλώνει μια έντονη αρνητική συσχέτιση σαν αυτή που αποτυπώνεται στο γράφημα.

ΒΙΒΛΙΟΓΡΑΦΙΑ

Ελληνόγλωσση

Καρακώστας, Μ. Ξ. (2002). *Γραμμικά Μοντέλα. Παλινδρόμηση-Ανάλυση Διακύμανσης*. Πανεπιστήμιο Ιωαννίνων.

Καρώνη, Χ. και Οικονόμου, Π. (2017). *Στατιστικά Μοντέλα Παλινδρόμησης*. 2η έκδοση, Συμεών, Αθήνα.

Κούτρας Μ. και Ευαγγελάρας, Χ. (2010). *Ανάλυση Παλινδρόμησης. Θεωρία και Εφαρμογές*. Εκδόσεις Σταμούλη.

Κουτρουβέλης, Ι. Α. (2000). *Βασικά Εργαλεία και Μέθοδοι για τον Έλεγχο Ποιότητας: Πιθανότητες και Στατιστική ΙΙ (Τόμος Β΄)*. Πάτρα: Ελληνικό Ανοικτό Πανεπιστήμιο.

Ξενόγλωσση

Azzalini, A. and Bowman, A. W. (1990). A Look at Some Data on the Old Faithful Geyser. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *39*(3), pp. 357–365.

Draper, N. and Smith, H. (2014). *Applied Regression Analysis*. Wiley Series in Probability and Statistics. Wiley, New Jersey.

Rencher, A. C. and Schaalje, G. B. (2008). *Linear Models in Statistics*. Wiley, New Jersey.

Searle, S. and Gruber, M. (2016). *Linear Models*. Wiley Series in Probability and Statistics. Wiley, New Jersey.

ΚΕΦΑΛΑΙΟ 13

ΑΝΑΛΥΣΗ ΔΙΑΚΥΜΑΝΣΗΣ

Σύνοψη

Σε αυτό το κεφάλαιο γενικεύεται ο απλός έλεγχος t για τη διαφορά των μέσων τιμών δύο ανεξάρτητων πληθυσμών σε έναν έλεγχο για την ισότητα των μέσων τιμών k ανεξάρτητων πληθυσμών υπό τις υποθέσεις της κανονικότητας και της ισότητας των διασπορών των πληθυσμών. Επίσης, προτείνονται τεχνικές για επακόλουθες πολλαπλές ζευγαρωτές συγκρίσεις (post-hoc ανάλυση) στην περίπτωση απόρριψης της αρχικής μηδενικής υπόθεσης και στην προσπάθεια διερεύνησης περαιτέρω στατιστικά σημαντικών διαφορών μεταξύ των μέσων τιμών των πληθυσμών. Τέλος, επιγραμματικά παρουσιάζονται μεθοδολογίες ελέγχου της υπόθεσης της κοινής διασποράς των πληθυσμών, δηλαδή της υπόθεσης πάνω στην οποία στηρίζεται ο έλεγχος ισότητας των μέσων τιμών των k ανεξάρτητων πληθυσμών.

Προαπαιτούμενη γνώση: Κεφάλαια 9-11 του παρόντος συγγράμματος.

Προσδοκώμενα μαθησιακά αποτελέσματα: Όταν θα έχετε ολοκληρώσει τη μελέτη του κεφαλαίου αυτού, θα μπορείτε

- να ελέγχετε υποθέσεις για την ισότητα των μέσων τιμών k ανεξάρτητων πληθυσμών,
- να ελέγχετε την υπόθεση της ισότητας των διασπορών k ανεξάρτητων πληθυσμών,
- να υλοποιείτε πολλαπλές ζευγαρωτές συγκρίσεις για όλες τις δυνατές διαφορές των μέσων τιμών των k ανεξάρτητων πληθυσμών στην περίπτωση που απορρίπτεται η αρχική μηδενική υπόθεση της ισότητάς τους,
- να σχολιάζετε και να ερμηνεύετε τα αποτελέσματα των παραπάνω ελέγχων υποθέσεων και
- να χρησιμοποιείτε την R για να υλοποιείτε τους παραπάνω ελέγχους.

Οικονόμου, Π., Μαλεφάκη, Σ., & Μπατσίδης, Α. (2023). *Πιθανότητες - Στατιστική.*
[Προπτυχιακό εγχειρίδιο].

Γλωσσάριο επιστημονικών όρων

- Άθροισμα τετραγώνων μεταξύ των ομάδων (Sum of Squares Between groups - SSB)
- Άθροισμα τετραγώνων μέσα στις ομάδες (Sum of Squares Within groups - SSW)
- Μέθοδος Ελάχιστης Σημαντικής Διαφοράς
- Μέθοδος Bartlett
- Μέθοδος Cochran
- Μέθοδος Bonferroni-Holm
- Μέθοδος Bonferroni
- Μέθοδος Tukey
- Πολλαπλές συγκρίσεις
- Συνολικό άθροισμα τετραγώνων (Total Sum of Squares - SST)

13.1 Εισαγωγή

Στο Κεφάλαιο 11 του παρόντος συγγράμματος παρουσιάστηκαν έλεγχοι υποθέσεων που αφορούσαν την ισότητα των μέσων τιμών δύο ανεξάρτητων πληθυσμών. Ωστόσο, στην πράξη πολύ συχνά χρειάζεται να ελέγξουμε την ισότητα των μέσων, μ_i, $i = 1, \ldots, k$, περισσότερων των δύο πληθυσμών ($k > 2$). Πιο συγκεκριμένα, καλούμαστε να ελέγξουμε τις υποθέσεις:

$$H_0: \ \mu_1 = \mu_2 = \ldots = \mu_k,$$

κατά

$$H_1: \text{ υπάρχει τουλάχιστον ένα ζεύγος } (i,l), \text{ με } i,\ell \in \{1,\ldots,k\}, i \neq \ell, \text{τέτοιο ώστε } \mu_i \neq \mu_\ell.$$

Μία πρώτη απλοϊκή σκέψη για την αντιμετώπιση του παραπάνω προβλήματος θα ήταν να γίνουν όλοι οι δυνατοί έλεγχοι ανά δύο. Σε αυτήν την περίπτωση όμως αυξάνεται σημαντικά η πιθανότητα σφάλματος τύπου Ι, δηλαδή η πιθανότητα να απορρίψουμε τη μηδενική υπόθεση, ενώ αυτή στην πραγματικότητα είναι αληθής. Με άλλα λόγια αυξάνεται σημαντικά η πιθανότητα να υποστηρίξουμε ότι υπάρχει στατιστικά σημαντική διαφορά μεταξύ των μέσων των πληθυσμών, ενώ στην πραγματικότητα δεν υπάρχει. Για παράδειγμα, αν $k = 10$ και θέλουμε να πραγματοποιήσουμε όλα τα δυνατά ζευγάρια ελέγχων, συνολικά θα πρέπει να γίνουν $k(k-1)/2 = (10 \cdot 9)/2 = 45$ έλεγχοι. Αν το επίπεδο σημαντικότητας καθενός εκ των 45 αυτών ελέγχων είναι 0.05, τότε $0.05 \times 45 \approx 2$ έλεγχοι μπορεί να δώσουν στην τύχη στατιστικά σημαντική διαφορά μεταξύ των μέσων. Πιο συγκεκριμένα, η πιθανότητα να απορριφθεί η μηδενική υπόθεση σε έναν τουλάχιστον έλεγχο ισούται με $1 - (1-\alpha)^{45} = 1 - (1-0.05)^{45} = 0.9006$.

Με βάση τα παραπάνω είναι προφανές ότι μία νέα μεθοδολογία χρειάζεται για την υλοποίηση του παραπάνω ελέγχου υποθέσεων. Η μεθοδολογία αυτή είναι η **ανάλυση διασποράς κατά έναν παράγοντα** (One Way Anova), που θα παρουσιαστεί στην επόμενη ενότητα και αποτελεί ειδική περίπτωση μιας γενικότερης μεθοδολογίας που είναι γνωστή ως **ανάλυση διασποράς**.

Η **ανάλυση διασποράς ή διακύμανσης** (Analysis of Variance - ANOVA) είναι μια αρκετά δυναμική μεθοδολογία με σημαντικό ρόλο στη Στατιστική, που οφείλεται κυρίως στον Sir Ronald Aylmer Fisher (1890 - 1962), με πρωτοπόρα την εργασία του Fisher (1918). Η μεθοδολογία αυτή μπορεί να χρησιμοποιηθεί για τη μελέτη της επίδρασης μίας, δύο ή περισσότερων ανεξάρτητων μεταβλητών (παραγόντων) σε μια εξαρτημένη μεταβλητή, αλλά και για τη μελέτη της αλληλεπίδρασης των ανεξάρτητων παραγόντων μεταξύ τους σε σχέση με την εξαρτημένη μεταβλητή. Επισημαίνεται σε αυτό το σημείο ότι παρόλο που η ανάλυση διασποράς, υπό αυτήν την οπτική, είναι ισοδύναμη με την ανάλυση παλινδρόμησης, μια σημαντική διαφοροποίησή τους είναι ότι στη συνήθη μορφή της ανάλυσης διασποράς η εξαρτημένη μεταβλητή είναι ποσοτική και οι ανεξάρτητες μεταβλητές είναι ποιοτικές, ενώ στη συνήθη μορφή της ανάλυσης παλινδρόμησης οι ανεξάρτητες μεταβλητές είναι ποσοτικές. Στο πλαίσιο του παρόντος συγγράμματος, θα παρουσιάσουμε τη μεθοδολογία της ανάλυσης διασποράς για τη μελέτη της επίδρασης μιας ανεξάρτητης ποιοτικής μεταβλητής με k το πλήθος δυνατές τιμές σε μια εξαρτημένη μεταβλητή, δηλαδή θα περιοριστούμε στη μεθοδολογία που είναι γνωστή ως ανάλυση διασποράς κατά έναν παράγοντα με k επίπεδα[1].

[1] Για το μοντέλο της ανάλυσης διασποράς κατά δύο παράγοντες, καθώς και για μεγαλύτερη εμβάθυνση στο μοντέλο της ανάλυσης διασποράς και τη σύνδεσή του με το μοντέλο της παλινδρόμησης παραπέμπουμε, μεταξύ άλλων, στα συγγράμματα των Καρακώστας (2002), Rencher and Schaalje (2008) και Παπαϊωάννου και Λουκάς (2002).

13.2 Ανάλυση διασποράς κατά έναν παράγοντα

Όπως έχει ήδη αναφερθεί, θέλουμε να ελέγξουμε την ισότητα των μέσων τιμών k πληθυσμών (ομάδων), θέλουμε δηλαδή να ελέγξουμε τις υποθέσεις:

$$H_0: \mu_1 = \mu_2 = \ldots = \mu_k,$$

κατά

$$H_1: \text{υπάρχει τουλάχιστον ένα ζεύγος } (i,\ell), \text{ με } i,\ell \in \{1,\ldots,k\}, i \neq \ell, \text{τέτοιο ώστε } \mu_i \neq \mu_\ell,$$

όπου μ_i η μέση τιμή του i-οστού πληθυσμού, $i = 1,..,k$, με $k > 2$.

Για τον σκοπό αυτό, έστω $Y_{11},\ldots,Y_{1n_1}$ ένα τυχαίο δείγμα μεγέθους n_1 από τον πρώτο πληθυσμό, $Y_{21},\ldots,Y_{2n_2}$ ένα τυχαίο δείγμα μεγέθους n_2 από τον δεύτερο πληθυσμό και ούτω καθεξής και $Y_{k1},\ldots,Y_{kn_k}$ ένα τυχαίο δείγμα μεγέθους n_k από τον $k-$οστό πληθυσμό, με Y_{ij} να είναι η $j-$ οστή τιμή της μεταβλητής στο i-οστό δείγμα. Υποθέτουμε ότι τα k το πλήθος τυχαία δείγματα είναι ανεξάρτητα μεταξύ τους και οι κατανομές τους είναι κανονικές με κοινή διασπορά σ^2.

Με βάση τις παραπάνω υποθέσεις έχουμε ότι:

$$Y_{ij} \sim N(\mu_i, \sigma^2),\ i = 1,\ldots,k,\ j = 1,\ldots,n_i.$$

Οι προφανείς εκτιμητές των $\mu_i,\ i = 1,\ldots,k$, είναι οι k δειγματικοί μέσοι:

$$\bar{Y}_{i\cdot} = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij},\ i = 1,\ldots,k. \tag{13.1}$$

Επίσης, ο συνολικός δειγματικός μέσος (ο δειγματικός μέσος όλων των παρατηρήσεων) ισούται με:

$$\bar{Y}_{\cdot\cdot} = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{n_i} Y_{ij}, \tag{13.2}$$

όπου $N = \sum_{i=1}^{k} n_i$.

Από τη σχέση (13.1) έχουμε ότι:

$$\bar{Y}_{i\cdot} \cdot n_i = \sum_{j=1}^{n_i} Y_{ij}, \tag{13.3}$$

οπότε ο συνολικός μέσος που δίνεται στη σχέση (13.2) μπορεί να γραφεί και ως ένας γραμμικός συνδυασμός των δειγματικών μέσων του κάθε πληθυσμού. Πιο συγκεκριμένα, έχουμε:

$$\bar{Y}_{\cdot\cdot} = \frac{1}{N}\sum_{i=1}^{k} n_i \cdot \bar{Y}_{i\cdot} = \sum_{i=1}^{k}\frac{n_i}{N}\cdot\bar{Y}_{i\cdot} \tag{13.4}$$

Στο Κεφάλαιο 12 του παρόντος συγγράμματος (βλ. Ενότητα 12.3.3) είδαμε ότι η διάσπαση της ολικής διασποράς της εξαρτημένης μεταβλητής, που μετριέται με το συνολικό άθροισμα τετραγώνων, σε άθροισμα τετραγώνων από την παλινδρόμηση και σε άθροισμα τετραγώνων των υπολοίπων, μπορεί να χρησιμοποιηθεί για την κατασκευή ελέγχου για τη σημαντικότητα της παλινδρόμησης. Η μεθοδολογία της ανάλυσης διασποράς κατά έναν παράγοντα βασίζεται σε μια ανάλογη διάσπαση της ολικής μεταβλητότητας $\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij} - \overline{y}_{\cdot\cdot}\right)^2$ των δεδομένων μας, η οποία προσδιορίζεται στην πρόταση που ακολουθεί.

Πρόταση 13.1

Υπό τον παραπάνω συμβολισμό ισχύει ότι:

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij}-\bar{y}_{..}\right)^2=\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij}-\bar{y}_{i.}\right)^2+\sum_{i=1}^{k}n_i\left(\bar{y}_{i.}-\bar{y}_{..}\right)^2 \tag{13.5}$$

Απόδειξη Πρότασης 13.1

Αρχικά, παρατηρούμε ότι η απόκλιση μιας παρατήρησης y_{ij} από τον συνολικό μέσο $\bar{y}_{..}$ μπορεί να γραφτεί ως

$$y_{ij}-\bar{y}_{..}=\left(y_{ij}-\bar{y}_{i.}\right)+\left(\bar{y}_{i.}-\bar{y}_{..}\right),$$

δηλαδή ως το άθροισμα της απόκλισης κάθε παρατήρησης από τον δειγματικό μέσο του αντίστοιχου δείγματος και της απόκλισης κάθε δειγματικού μέσου από τον συνολικό δειγματικό μέσο. Έπειτα έχουμε

$$\begin{aligned}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij}-\bar{y}_{..}\right)^2&=\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij}-\bar{y}_{i.}+\bar{y}_{i.}-\bar{y}_{..}\right)^2\\&=\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left\{\left(y_{ij}-\bar{y}_{i.}\right)^2+\left(\bar{y}_{i.}-\bar{y}_{..}\right)^2+2\left(y_{ij}-\bar{y}_{i.}\right)\left(\bar{y}_{i.}-\bar{y}_{..}\right)\right\}\\&=\sum_{i=1}^{k}\left\{\sum_{j=1}^{n_i}\left(y_{ij}-\bar{y}_{i.}\right)^2+n_i\left(\bar{y}_{i.}-\bar{y}_{..}\right)^2+2\left(\bar{y}_{i.}-\bar{y}_{..}\right)\sum_{j=1}^{n_i}\left(y_{ij}-\bar{y}_{i.}\right)\right\}\\&=\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij}-\bar{y}_{i.}\right)^2+\sum_{i=1}^{k}n_i\left(\bar{y}_{i.}-\bar{y}_{..}\right)^2,\end{aligned}$$

καθώς $\sum_{j=1}^{n_i}\left(y_{ij}-\bar{y}_{i.}\right)=\sum_{j=1}^{n_i}y_{ij}-n_i\bar{y}_{i.}=0$ για κάθε $i=1,\ldots,k$.

Στην πραγματικότητα η σχέση (13.5) αναλύει το συνολικό άθροισμα τετραγώνων (Total Sum of Squares, SST), των αποκλίσεων των y_{ij} από το μέσο τους $\bar{y}_{..}$, δηλαδή τη συνολική διασπορά, σε δύο αθροίσματα τετραγώνων, το άθροισμα των τετραγώνων των αποκλίσεων των μέσων των k ομάδων από το συνολικό μέσο, που λέγεται άθροισμα τετραγώνων μεταξύ των ομάδων (Sum of Squares Between groups, SSB) και το άθροισμα των τετραγώνων των αποκλίσεων των y_{ij} από το μέσο της ομάδας στην οποία ανήκουν, το οποίο λέγεται άθροισμα τετραγώνων μέσα στις ομάδες (Sum of Squares Within groups, SSW). Επομένως, έχουμε ότι:

$$SST=SSW+SSB, \tag{13.6}$$

όπου

$$SST=\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij}-\bar{y}_{..}\right)^2, \tag{13.7}$$

$$SSB=\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(\bar{y}_{i.}-\bar{y}_{..}\right)^2=\sum_{i=1}^{k}n_i\left(\bar{y}_{i.}-\bar{y}_{..}\right)^2, \tag{13.8}$$

και

$$SSW=\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij}-\bar{y}_{i.}\right)^2. \tag{13.9}$$

Εύκολα μπορεί να αποδειχθεί (αφήνεται ως άσκηση για τον/την αναγνώστη/στρια) ότι μια ισοδύναμη και πιο βολική στους υπολογισμούς σχέση για το συνολικό άθροισμα τετραγώνων είναι η:

$$SST = \sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}^2 - N \cdot \overline{y}_{..}^2. \tag{13.10}$$

Παρατήρηση 13.1

Στο σημείο αυτό, θα πρέπει να τονίσουμε ότι τα SST, SSB και SSW στις παραπάνω σχέσεις είναι παρατηρηθείσες τιμές των αντίστοιχων τυχαίων μεταβλητών. Στη συνέχεια, θα χρησιμοποιείται το ίδιο σύμβολο και για τις δύο αυτές περιπτώσεις, καθώς κάτι τέτοιο απλοποιεί τους συμβολισμούς και δεν δημιουργεί ιδιαίτερα προβλήματα στην παρακολούθηση του κειμένου.

Στο επόμενο θεώρημα προσδιορίζονται κάποιες βασικές ιδιότητες των τ.μ. SSB και SSW που θα φανούν ιδιαίτερα χρήσιμες στην κατασκευή του υπό μελέτη ελέγχου.

Θεώρημα 13.1

Υπό την παραδοχή της κανονικότητας των πληθυσμών και της ισότητας των διασπορών τους ισχύουν τα εξής:

1. Τα SSB και SSW είναι ανεξάρτητες τυχαίες μεταβλητές.
2. $\frac{SSW}{\sigma^2} \sim \chi^2_{N-k}$.
3. Αν επιπλέον $\mu_1 = \cdots = \mu_k$, τότε $\frac{SSB}{\sigma^2} \sim \chi^2_{k-1}$.

Απόδειξη Θεωρήματος 13.1

1. Από τη σχέση (13.8) και λαμβάνοντας υπόψη ότι από τη σχέση (13.4) ο γενικός δειγματικός μέσος είναι μία συνάρτηση των δειγματικών μέσων των k ομάδων, συμπεραίνουμε ότι το SSB είναι συνάρτηση των δειγματικών μέσων.
   Επιπλέον, εύκολα προκύπτει από τη σχέση (13.9) ότι:

$$SSW = \sum_{i=1}^{k}(n_i - 1)S_i^2, \tag{13.11}$$

   όπου

$$S_i^2 = \sum_{j=1}^{n_i} \frac{\left(Y_{ij} - \overline{Y}_{i\cdot}\right)^2}{n_i - 1},\ i = 1, \ldots, k,$$

   είναι η δειγματική διασπορά του i-οστού δείγματος, $i = 1, \ldots, k$.
   Επομένως, το SSB είναι συνάρτηση των δειγματικών μέσων, ενώ το SSW είναι συνάρτηση των δειγματικών διασπορών. Επειδή σε κάθε τυχαίο δείγμα από την κανονική κατανομή ο δειγματικός μέσος και η δειγματική διασπορά είναι ανεξάρτητες τυχαίες μεταβλητές και καθώς τα k τυχαία δείγματα είναι ανεξάρτητα μεταξύ τους, συμπεραίνουμε ότι τα SSB και SSW είναι επίσης ανεξάρτητες τυχαίες μεταβλητές.
2. Από τη σχέση (13.11) άμεσα προκύπτει ότι:

$$\frac{SSW}{\sigma^2} = \sum_{i=1}^{k}\frac{(n_i - 1)S_i^2}{\sigma^2} = \sum_{i=1}^{k} V_i,$$

   όπου

$$V_i = \frac{(n_i - 1)S_i^2}{\sigma^2} \sim \chi^2_{n_i - 1},\ i = 1, \ldots, k.$$

Γνωρίζουμε ότι, όταν έχουμε ένα τυχαίο δείγμα μεγέθους n_i από κανονική κατανομή με διασπορά σ^2, τότε για τη δειγματική διασπορά S_i^2 ισχύει ότι:

$$V_i = \frac{(n_i - 1)S_i^2}{\sigma^2} \sim \chi^2_{n_i - 1}.$$

Οι k το πλήθος δειγματικές διασπορές είναι ανεξάρτητες τυχαίες μεταβλητές (αφού τα k το πλήθος δείγματα είναι ανεξάρτητα), επομένως το άθροισμα $\sum_{i=1}^{k} V_i$ θα έχει χι-τετράγωνο κατανομή, με βαθμούς ελευθερίας το άθροισμα των επιμέρους βαθμών ελευθερίας, δηλαδή $\sum_{i=1}^{k}(n_i - 1) = N - k$ βαθμούς ελευθερίας.

3. Αν οι k το πλήθος πληθυσμιακές μέσες τιμές είναι ίσες μεταξύ τους, τότε όλες οι παρατηρήσεις προέρχονται από την ίδια κανονική κατανομή. Η δειγματική διασπορά των N αυτών παρατηρήσεων ισούται με $S^2 = SST/(N-1)$. Επομένως, έχουμε ότι

$$W := \frac{(N-1)S^2}{\sigma^2} = \frac{SST}{\sigma^2} \sim \chi^2_{N-1}.$$

Επίσης, γνωρίζουμε ότι η ροπογεννήτρια της κατανομής χ^2_m ισούται με $(1-2t)^{-m/2}$, $t < 1/2$. Έστω τώρα $W_1 := SSB/\sigma^2$, $W_2 := SSW/\sigma^2$. Από το πρώτο μέρος της πρότασης ξέρουμε ότι αυτές είναι ανεξάρτητες τυχαίες μεταβλητές, ενώ από το δεύτερο ότι η W_2 ακολουθεί κατανομή χ^2_{N-k}. Επομένως, λαμβάνοντας υπόψη ότι $W = W_1 + W_2$ και ότι η ροπογεννήτρια αθροίσματος ανεξάρτητων τυχαίων μεταβλητών ισούται με το γινόμενο των ροπογεννητριών τους, για τις ροπογεννήτριες των W, W_1, W_2 θα ισχύει ότι:

$$M_W(t) = M_{W_1}(t) M_{W_2}(t)$$

ή, ισοδύναμα, ότι

$$(1-2t)^{-(N-1)/2} = M_{W_1}(t)(1-2t)^{-(N-k)/2}, \quad t < 1/2.$$

Από την τελευταία σχέση άμεσα προκύπτει ότι:

$$M_{W_1}(t) = (1-2t)^{-(k-1)/2},\ t < 1/2,$$

που ταυτίζεται με τη ροπογεννήτρια της κατανομής χ^2_{k-1}. Το συμπέρασμα προκύπτει λόγω της ιδιότητας του μονοσήμαντου των ροπογεννητριών.

Παρατήρηση 13.2

Στο σημείο αυτό, να σημειώσουμε ότι για την απόδειξη του πρώτου μέρους του Θεωρήματος 13.1, το μόνο που απαιτείται είναι οι κατανομές να είναι κανονικές.

Πρόταση 13.2

Υπό τις υποθέσεις του Θεωρήματος 13.1 και αν $\mu_1 = \ldots = \mu_k$, τότε

$$F = \frac{SSB/(k-1)}{SSW/(N-k)} \sim F_{k-1, N-k}.$$

Απόδειξη Πρότασης 13.2

Από το Θεώρημα 13.1 έχουμε ότι οι $W_1 = SSB/\sigma^2$ και $W_2 = SSW/\sigma^2$ είναι ανεξάρτητες τυχαίες μεταβλητές. Η δεύτερη έχει κατανομή χ^2_{N-k}, ενώ όταν $\mu_1 = \cdots = \mu_k$ η πρώτη έχει κατανομή χ^2_{k-1}. Επομένως, όταν $\mu_1 = \cdots = \mu_k$, τότε

$$F = \frac{SSB/(k-1)}{SSW/(N-k)} = \frac{W_1/(k-1)}{W_2/(N-k)} \sim F_{k-1,N-k}$$

από τον ορισμό της κατανομής F.

Από την προηγούμενη πρόταση συμπεραίνουμε ότι για να υλοποιήσουμε τον έλεγχο

$$H_0 : \mu_1 = \cdots = \mu_k \text{ κατά } H_1 : \text{ τουλάχιστον ένα } \mu_i \text{ διαφορετικό}$$

μπορούμε να χρησιμοποιήσουμε τη στατιστική συνάρτηση, $F = \frac{SSB/(k-1)}{SSW/(N-k)}$ καθώς έχει γνωστή κατανομή, πλήρως προσδιορισμένη, υπό τη μηδενική υπόθεση. Επιπλέον, θα απορρίπτεται η μηδενική υπόθεση για μεγάλες τιμές της F, αφού σε αυτήν την περίπτωση η μεταβλητότητα μεταξύ των ομάδων θα είναι πολύ μεγαλύτερη από τη μεταβλητότητα εντός των ομάδων, πράγμα που σημαίνει ότι υπάρχει διαφορά στους μέσους των ομάδων, αφού οι διασπορές είναι ίσες. Αν ο έλεγχος γίνεται σε επίπεδο σημαντικότητας α, τότε η κρίσιμη περιοχή του ελέγχου είναι η $F > F_{k-1,N-k,\alpha}$.

Οι τιμές των παραπάνω αθροισμάτων τετραγώνων, της τ.μ. F, καθώς και άλλα χρήσιμα στοιχεία του παραπάνω ελέγχου παρουσιάζονται σε έναν πίνακα που ονομάζεται πίνακας ανάλυσης διασποράς (ANOVA table), η γενική μορφή του οποίου παρατίθεται στη συνέχεια.

Μεταβλητότητα	Βαθμοί ελευθερίας	Άθροισμα τετραγώνων	Μέσο Άθροισμα τετραγώνων	F
Μεταξύ των ομάδων	$k-1$	$SSB = \sum_{i=1}^{k} n_i(\overline{y}_{i\cdot} - \overline{y}_{\cdot\cdot})^2$	$\frac{SSB}{k-1}$	$\frac{SSB/(k-1)}{SSW/(N-k)}$
Εντός των ομάδων	$N-k$	$SSW = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \overline{y}_{i\cdot})^2$	$\frac{SSW}{N-k}$	
Ολική	$N-1$	$SST = \sum_{i-1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \overline{y}_{\cdot\cdot})^2$	$\frac{SST}{N-1}$	

Παρατήρηση 13.3

Οι τιμές στη στήλη του μέσου αθροίσματος τετραγώνων προκύπτουν από τη διαίρεση των αθροισμάτων τετραγώνων με τους αντίστοιχους βαθμούς ελευθερίας. Από την άλλη, οι βαθμοί ελευθερίας εκφράζουν το πλήθος των τιμών στον τελικό υπολογισμό μιας στατιστικής συνάρτησης που είναι ελεύθερες να ποικίλλουν έτσι ώστε να λάβουμε ένα προκαθορισμένο αποτέλεσμα. Για να γίνει αυτό πιο κατανοητό, αρκεί να παρατηρήσουμε, για παράδειγμα, ότι, αν και το SST υπολογίζεται από $N = \sum_{i=1}^{k} n_i$ όρους της μορφής $\left(y_{ij} - \overline{y}_{\cdot\cdot}\right)^2$, δεν πρόκειται για N ξεχωριστές πληροφορίες, αφού $\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij} - \overline{y}_{\cdot\cdot}\right) = 0$.
Έτσι, το SST έχει $N-1$ βαθμούς ελευθερίας, διότι, αν γνωρίζουμε τις $N-1$ αποκλίσεις $\left(y_{ij} - \overline{y}_{\cdot\cdot}\right)$, τότε προφανώς και η εναπομείνασα απόκλιση καθορίζεται από τις υπόλοιπες και δεν προσθέτει καμία νέα πληροφορία.
Παρόμοια, το SSW έχει $N-k$ βαθμούς ελευθερίας, γιατί υπάρχουν N όροι και k περιορισμοί $\sum_{j=1}^{n_i}\left(y_{ij} - \overline{y}_{i\cdot}\right) = 0$, για $i = 1, \ldots, k$. Τέλος, το SSB επίσης έχει $k-1$ βαθμούς ελευθερίας, γιατί υπάρχουν

k όροι και η συνθήκη $\sum_{i=1}^{k} n_i \left(\overline{y}_{i.} - \overline{y}_{..}\right) = 0$.
Παρατηρήστε ότι το άθροισμα των βαθμών ελευθερίας του SSB και του SSW ισούται με τους βαθμούς ελευθερίας του SST.

Παράδειγμα 13.1: (Φουσκάκης, 2013)

Ένας γιατρός θέλει να συγκρίνει τρεις αγωγές που χρησιμοποιούνται σε ασθενείς που υποβάλλονται σε εγχείρηση καρδιάς ως προς το μέσο επίπεδο φολικού οξέος. Για τον λόγο αυτό, είκοσι δύο (22) ασθενείς που υποβάλλονται σε εγχείρηση καρδιάς χωρίζονται τυχαία σε 3 ομάδες, ως εξής:
► Ομάδα 1: αποτελείται από ασθενείς που έλαβαν μείγμα 50% νιτρώδους οξειδίου και 50% οξυγόνου για 24 ώρες.
► Ομάδα 2: αποτελείται από ασθενείς που έλαβαν μείγμα 50% νιτρώδους οξειδίου και 50% οξυγόνου μόνο κατά τη διάρκεια της εγχείρησης.
► Ομάδα 3: αποτελείται από ασθενείς που έλαβαν $35-50\%$ νιτρώδους οξειδίου και 50% οξυγόνου για 24 ώρες.
Στον παρακάτω πίνακα καταγράφεται η τιμή του φολικού οξέος (σε mg/l) στα ερυθροκύτταρα των ασθενών.

Ασθενής	Ομάδα 1	Ομάδα 2	Ομάδα 3
1	243	206	241
2	251	210	258
3	275	226	270
4	291	249	293
5	347	255	328
6	354	273	
7	380	285	
8	392	295	
9		309	

Υπάρχει, σε επίπεδο σημαντικότητας 0.05, στατιστικά σημαντική διαφορά στο μέσο επίπεδο φολικού οξέος (σε mg/l) στα ερυθροκύτταρα των τριών αγωγών;
Υπόδειξη: υποθέτουμε ότι η συγκέντρωση φολικού οξέος στα ερυθροκύτταρα με χρήση καθεμίας εκ των τριών αγωγών ακολουθεί κανονική κατανομή με κοινή διασπορά.

Λύση Παραδείγματος 13.1

Έστω μ_1, μ_2 και μ_3 οι μέσες τιμές του φολικού οξέος στα ερυθροκύτταρα του πρώτου, δεύτερου και τρίτου πληθυσμού ασθενών, αντίστοιχα. Οι υποθέσεις που θέλουμε να ελέγξουμε, σε επίπεδο σημαντικότητας 0.05, είναι:

$$H_0 : \mu_1 = \mu_2 = \mu_3,$$

κατά

$$H_1 : \text{ υπάρχει τουλάχιστον ένα ζεύγος } (i,\ell), \text{ με } i,\ell = 1,2,3, i \neq \ell, \text{τέτοιο ώστε } \mu_i \neq \mu_\ell.$$

Θα ακολουθήσουμε τη μεθοδολογία της ανάλυσης διακύμανσης κατά έναν παράγοντα, δημιουργώντας τον Πίνακα Ανάλυσης της Διασποράς. Σε όσα ακολουθούν y_{ij} είναι η j τιμή της συγκέντρωσης φολικού οξέος στα ερυθροκύτταρα στην i-οστή ομάδα ασθενών για $i = 1,2,3$ και $j = 1,\ldots,n_i$, όπου $n_1 = 8$, $n_2 = 9$ και $n_3 = 5$.
Αρχικά, θέλοντας να υπολογίσουμε το συνολικό άθροισμα τετραγώνων, υπολογίζουμε τον συνολικό

δειγματικό μέσο. Είναι

$$\bar{y}_{..} = \sum_{i=1}^{3}\sum_{j=1}^{n_i} y_{ij}/\mathrm{N} = \frac{243 + 251 + \cdots + 328}{8 + 9 + 5} = \frac{6231}{22} = 283.2273,$$

αφού $N = n_1 + n_2 + n_3 = 22$. Επομένως,

$$\begin{aligned} SST &= \sum_{i=1}^{3}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{..})^2 \\ &= (243 - 283.2273)^2 + \cdots + (328 - 283.2273)^2 = 55231.8636 \end{aligned}$$

ή

$$\begin{aligned} SST &= \sum_{i=1}^{3}\sum_{j=1}^{n_i} y_{ij}^2 - N \cdot \bar{y}_{..}^2 \\ &= (243^2 + \cdots + 328^2)^2 - 22(283.2273)^2 = 55231.8636. \end{aligned}$$

Στη συνέχεια, θέλοντας να υπολογίσουμε το συνολικό άθροισμα τετραγώνων μεταξύ των ομάδων, υπολογίζουμε τους δειγματικούς μέσους κάθε ομάδας ασθενών. Είναι:

$$\bar{y}_{1\cdot} = \frac{\sum_{j=1}^{n_1} y_{1j}}{n_1} = \frac{243 + \cdots + 392}{8} = \frac{2533}{8} = 316.625,$$

$$\bar{y}_{2\cdot} = \frac{\sum_{j=1}^{n_2} y_{2j}}{n_2} = \frac{206 + \cdots + 309}{9} = \frac{2308}{9} = 256.4444,$$

και

$$\bar{y}_{3\cdot} = \frac{\sum_{j=1}^{n_3} y_{3j}}{n_3} = \frac{241 + \cdots + 328}{5} = \frac{1390}{5} = 278.$$

Επομένως,

$$\begin{aligned} SSB &= \sum_{i=1}^{3} n_i(\bar{y}_{i\cdot} - \bar{y}_{..})^2 \\ &= 8 \cdot (316.625 - 283.2273)^2 + 9 \cdot (256.4444 - 283.2273)^2 + 5 \cdot (278 - 283.2273)^2 \\ &= 15515.7664. \end{aligned}$$

Άρα είναι:

$$SSW = SST - SSB = 55231.8636 - 15515.7664 = 39716.0972.$$

Με βάση τα παραπάνω προκύπτει ο ακόλουθος πίνακας ανάλυσης διασποράς

Μεταβλητότητα	Βαθμοί ελευθερίας	Άθροισμα τετραγώνων	Μέσο Άθροισμα τετραγώνων	F
Μεταξύ των ομάδων (SSB)	2	15515.7664	7757.883	3.711
Εντός των ομάδων (SSW)	19	39716.0972	2090.321	
Ολική (SST)	21	55231.8636	2630.089	

Από τον πίνακα της κατανομής F (βλ. τον Πίνακα Α΄.9, Παραρτήματος Α΄) για επίπεδο σημαντικότητας 0.05 έχουμε ότι $F_{2,19,0.05} = 3.52$. Καθώς $f = 3.711 > F_{2,19,0.05}$, απορρίπτεται σε επίπεδο σημαντικότητας 0.05 η μηδενική υπόθεση. Με βάση αυτά τα δεδομένα μπορούμε να ισχυριστούμε ότι το μέσο φολικό οξύ στα ερυθροκύτταρα τουλάχιστον μίας ομάδας ασθενών διαφέρει στατιστικά σημαντικά από το αντίστοιχο των υπολοίπων.

Εναλλακτικά, θα μπορούσαμε να λύσουμε την άσκηση χρησιμοποιώντας την R και τις ακόλουθες εντολές:

```
d1 <- data.frame(
  folic_acid=c(243,251,275,291,347,354,380,392,
               206,210,226,249,255,273,285,295,309,
               241,258,270,293,328),
  patient_group=c( rep("Group1",8), rep("Group2",9),rep("Group3",5) )
)
res.aov <- aov(formula =  folic_acid~patient_group,data = d1)
summary(res.aov)
```

Τα αποτελέσματα που λαμβάνουμε είναι τα εξής:

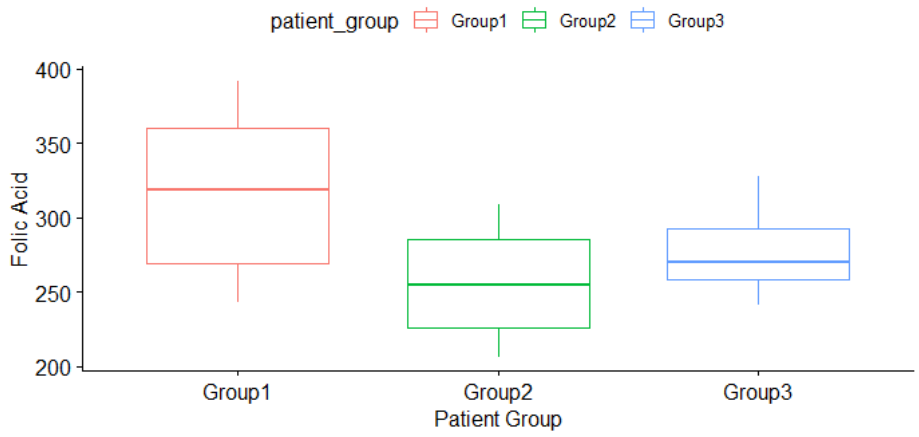
```
               Df Sum Sq Mean Sq F value Pr(>F)
patient_group   2  15516    7758   3.711 0.0436 *
Residuals      19  39716    2090
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Παρατηρούμε ότι $p - value = 0.0436 < 0.05$, άρα απορρίπτεται η μηδενική υπόθεση, όπως και πριν. Επομένως, με βάση αυτά τα δεδομένα δεν μπορούμε να ισχυριστούμε σε επίπεδο σημαντικότητας 0.05 ότι η μέση τιμή του φολικού οξέος στις τρεις ομάδες ασθενών είναι η ίδια.

Στο Παράδειγμα 13.1 προέκυψε ότι δεν μπορούμε να ισχυριστούμε ότι οι μέσες τιμές των τριών πληθυσμών είναι ίσες και ότι υπάρχει τουλάχιστον ένα ζεύγος (i,j), $i,j \in \{1,2,3\}$ με $i \neq j$, τέτοιο ώστε $\mu_i \neq \mu_j$. Ένα εύλογο ερώτημα που προκύπτει είναι ποια/-ες από τις μέσες τιμές διαφέρει/-ουν στατιστικά σημαντικά από τις υπόλοιπες. Θα μπορούσε κάποιος να σχηματίσει μία πρώτη άποψη για την απάντηση στο παραπάνω ερώτημα κατασκευάζοντας τα θηκογράμματα των συγκεντρώσεων φολικού οξέος στα ερυθροκύτταρα των ασθενών των τριών ομάδων. Τα προαναφερθέντα θηκογράμματα παρατίθενται στο Σχήμα 13.1 και προέκυψαν χρησιμοποιώντας τις ακόλουθες εντολές της R:

```
library("ggpubr")
ggboxplot(d1, x = "patient_group", y = "folic_acid",
          color = "patient_group",
          ylab = "Folic Acid", xlab = "Patient Group")
```

Από τα θηκογράμματα που παρατίθενται στο Σχήμα 13.1 μπορούμε να παρατηρήσουμε ότι το επίπεδο φολικού οξέος στα ερυθροκύτταρα της πρώτης ομάδας (κόκκινο χρώμα) μοιάζει να είναι υψηλότερο από αυτό των άλλων δύο ομάδων. Ωστόσο, θα θέλαμε να υπάρχουν στατιστικοί τρόποι ελέγχου των επιμέρους υποθέσεων με σκοπό να αποφασίσουμε ποια/-ες από τις μέσες τιμές διαφέρει/-ουν από τις υπόλοιπες και κατά αυτόν τον τρόπο να εντοπίσουμε τα επίπεδα του παράγοντα που επιδρούν στη μέση τιμή της εξαρτημένης μεταβλητής. Αυτό είναι το αντικείμενο των μεθοδολογιών που είναι γνωστές ως πολλαπλές συγκρίσεις και αποτελούν αντικείμενο μελέτης της επόμενης ενότητας.



Σχήμα 13.1: Θηκογράμματα των συγκεντρώσεων φολικού οξέος στα ερυθροκύτταρα των τριών ομάδων ασθενών του Παραδείγματος 13.1.

Άσκηση Αυτοαξιολόγησης 13.1

Όταν χρησιμοποιείται ορείχαλκος σε μια παραγωγή, ο συντελεστής ελαστικότητας Ε του υλικού είναι συχνά σημαντικός για τη λειτουργικότητα. Ο συντελεστής ελαστικότητας Ε μετριέται για 3 διαφορετικά κράματα ορείχαλκου και ελέγχονται 5 δείγματα από κάθε κράμα. Τα αποτελέσματα φαίνονται στον παρακάτω πίνακα, όπου ο μετρούμενος συντελεστής ελαστικότητας δίνεται σε GPa:

M1	M2	M3
82.5	82.7	92.2
83.7	81.9	106.8
80.9	78.9	104.6
85.2	83.6	94.5
80.8	78.6	100.7

Να ελέγξετε σε επίπεδο σημαντικότητας 0.05 αν υπάρχει διαφοροποίηση στον μέσο συντελεστή ελαστικότητας των τριών διαφορετικών τύπων κραμάτων. Υπόδειξη: να υποθέσετε ότι η μέτρηση του συντελεστή ελαστικότητας σε κάθε τύπο κράματος ακολουθεί κανονική κατανομή με κοινή διασπορά.

Παρατήρηση 13.4

Όσα αναφέρθηκαν σε αυτήν την ενότητα ισχύουν υπό την υπόθεση της κανονικότητας των πληθυσμών και της ισότητας των διασπορών τους. Αν η υπόθεση της κανονικότητας παραβιάζεται για κάποιον από τους πληθυσμούς και το μέγεθος του δείγματος από τον αντίστοιχο πληθυσμό είναι μεγαλύτερο του 30, η παραπάνω μεθοδολογία μπορεί να χρησιμοποιηθεί, λόγω του Κεντρικού Οριακού Θεωρήματος. Αν ωστόσο το μέγεθος του δείγματος είναι μικρότερο, τότε προσφεύγουμε είτε σε κάποιο μετασχηματισμό διόρθωσης της κανονικότητας είτε σε μη παραμετρικές μεθοδολογίες, όπως αυτή των Kruskal and Wallis (1952). Σχετικά με τον έλεγχο της ισότητας των διασπορών παραπέμπουμε στην Ενότητα 13.4.

13.3 Πολλαπλές συγκρίσεις

Έστω ότι στον έλεγχο ANOVA απορρίπτουμε τη μηδενική υπόθεση της ισότητας των k το πλήθος μέσων τιμών. Αυτό σημαίνει ότι μία τουλάχιστον από τις πληθυσμιακές μέσες τιμές μ_i, $i = 1, \ldots, k$ διαφέρει στατιστικά σημαντικά από κάποια/κάποιες από τις υπόλοιπες. Ένα εύλογο ερώτημα που προκύπτει είναι ποιο/ποια

ζεύγος/ζεύγη πληθυσμιακών μέσων τιμών διαφέρουν στατιστικά σημαντικά μεταξύ τους. Στην ενότητα αυτήν θα παρουσιαστούν στατιστικοί τρόποι ελέγχου της ισότητας μέσων τιμών κάθε επιμέρους ζεύγους των k το πλήθος πληθυσμιακών μέσων τιμών.

Μία προφανής τακτική είναι να καταφύγουμε στον έλεγχο όλων των ανά δύο διαφορών, δηλαδή να πραγματοποιήσουμε $\binom{k}{2} = \frac{k\cdot(k-1)}{2}$ το πλήθος ελέγχους. Σε όσα ακολουθούν, υπό την υπόθεση της κανονικότητας των k το πλήθος πληθυσμών και της ισότητας των διασπορών, θεωρούμε ότι θέλουμε να ελέγξουμε τις υποθέσεις:

$$H_0 : \mu_i = \mu_j \text{ κατά } H_1 : \mu_i \neq \mu_j, \text{για } i,j = 1,\dots,k, \text{ με } i \neq j,$$

δηλαδή αν ο μέσος του i-οστού πληθυσμού δεν διαφέρει στατιστικά σημαντικά από τον μέσο του j-οστού πληθυσμού. Στο πλαίσιο αυτό θα παρουσιαστούν συνοπτικά τα βασικά στοιχεία κάποιων μεθοδολογιών που έχουν προταθεί στη βιβλιογραφία για τον σκοπό αυτό.

13.3.1 Η μέθοδος της ελάχιστης σημαντικής διαφοράς

Καθώς μας ενδιαφέρει να συγκρίνουμε τις μέσες τιμές μ_i και μ_j, είναι λογικό να καταφύγουμε στον κλασικό έλεγχο t για τη διαφορά των μέσων με δύο ανεξάρτητα δείγματα και άγνωστες αλλά ίσες διασπορές. Επομένως, είναι λογικό να θεωρηθεί ως στατιστική συνάρτηση ελέγχου η συνάρτηση

$$\mathrm{T}_{ij}^{*} = \frac{\bar{Y}_{i\cdot} - \bar{Y}_{j\cdot}}{S_{ij,p}\sqrt{1/n_i + 1/n_j}},$$

όπου

$$S_{ij,p}^2 = \frac{(n_i - 1)S_i^2 + (n_j - 1)S_j^2}{n_i + n_j - 2},$$

με S_i^2, S_j^2 να είναι οι διασπορές του i-οστού και j-οστού δείγματος, αντίστοιχα. Σε μία τέτοια περίπτωση, σε επίπεδο σημαντικότητας α, η κρίσιμη περιοχή του ελέγχου θα είναι η

$$|t_{ij}^{*}| > t_{n_i+n_j-2,\alpha/2},$$

όπου t_{ij}^{*} η παρατηρούμενη τιμή της στατιστικής συνάρτησης T_{ij}^{*}.

Ωστόσο, στην παραπάνω στατιστική συνάρτηση ελέγχου η εκτίμηση της κοινής διασποράς των k πληθυσμών στηρίζεται στα δεδομένα των δύο επιμέρους δειγμάτων και όχι σε όλα τα διαθέσιμα δεδομένα από τα k το πλήθος δείγματα. Εναλλακτικά, λοιπόν, θα μπορούσαμε να εκτιμήσουμε την κοινή διασπορά των πληθυσμών i και j, χρησιμοποιώντας τον εκτιμητή

$$\hat{\sigma}^2 = \frac{SSW}{N - k} \tag{13.12}$$

ο οποίος στηρίζεται και στα k τυχαία δείγματα. Διαισθητικά είναι σαφές ότι ο εκτιμητής $\hat{\sigma}^2$ είναι καλύτερος από τον $S_{ij,p}^2$, αφού βασίζεται και στα k το πλήθος τυχαία δείγματα σε αντίθεση με τον $S_{ij,p}^2$, ο οποίος βασίζεται μόνο σε δύο δείγματα, το i-οστό και το j-οστό.

Επομένως, τα παραπάνω οδηγούν στη στατιστική συνάρτηση ελέγχου

$$\mathrm{T}_{ij} = \frac{\bar{Y}_{i\cdot} - \bar{Y}_{j\cdot}}{\hat{\sigma}\sqrt{1/n_i + 1/n_j}}. \tag{13.13}$$

Εύκολα μπορεί να αποδειχθεί (η απόδειξη αφήνεται ως άσκηση για τον/την αναγνώστη/στρια) ότι υπό τη μηδενική υπόθεση $H_0 : \mu_i = \mu_j$, ισχύει ότι $T_{ij} \sim t_{N-k}$. Επομένως, η μηδενική υπόθεση $H_0 : \mu_i = \mu_j$ κατά της $H_1 : \mu_i \neq \mu_j$ απορρίπτεται, σε επίπεδο σημαντικότητας α, αν

$$|t_{ij}| > t_{N-k,\alpha/2}$$

ή, ισοδύναμα, αν

$$\left|\bar{y}_{i\cdot} - \bar{y}_{j\cdot}\right| > t_{N-k,\alpha/2}\sqrt{\frac{SSW}{N-k}}\sqrt{1/n_i + 1/n_j}$$

Παρατηρήστε ότι, όταν όλα τα μεγέθη δείγματος είναι ίσα, τότε προκύπτει ότι η ποσότητα στο δεύτερο μέλος της παραπάνω κρίσιμης περιοχής είναι η ίδια για τον έλεγχο κάθε ζεύγους μέσων τιμών. Σε μία τέτοια περίπτωση η ποσότητα αυτή ονομάζεται και Ελάχιστη Σημαντική Διαφορά (Least Significant Difference, LSD).

Ισοδύναμα, θα μπορούσαμε να πραγματοποιήσουμε τον παραπάνω έλεγχο χρησιμοποιώντας το $100(1-\alpha)\%$ διάστημα εμπιστοσύνης για τη διαφορά $\mu_i - \mu_j$, οπότε η μηδενική υπόθεση απορρίπτεται σε επίπεδο σημαντικότητας α, αν το 0 δεν περιέχεται στο διάστημα εμπιστοσύνης

$$\left(\bar{y}_{i\cdot} - \bar{y}_{j\cdot} - t_{N-k,\alpha/2}\hat{\sigma}\sqrt{1/n_i + 1/n_j},\ \bar{y}_{i\cdot} - \bar{y}_{j\cdot} + t_{N-k,\alpha/2}\hat{\sigma}\sqrt{1/n_i + 1/n_j}\right). \tag{13.14}$$

Στην πράξη, οι παραπάνω έλεγχοι υλοποιούνται μέσω των διαστημάτων εμπιστοσύνης τους, δηλαδή κατασκευάζονται τα $k(k-1)/2$ διαστήματα εμπιστοσύνης για όλες τις δυνατές διαφορές και βλέπουμε ποια περιέχουν το μηδέν και ποια όχι.

Παράδειγμα 13.2

Σε συνέχεια του Παραδείγματος 13.1, εφόσον απορρίφθηκε η μηδενική υπόθεση της ισότητας του μέσου φολικού οξέος στα ερυθροκύτταρα των ασθενών των τριών ομάδων, χρησιμοποιήστε τη μέθοδο της ελάχιστης σημαντικής διαφοράς για να προχωρήσετε σε περαιτέρω πολλαπλές ζευγαρωτές συγκρίσεις όλων των ομάδων ασθενών ανά δύο, σε επίπεδο σημαντικότητας 0.05.

Λύση Παραδείγματος 13.2

Από το Παράδειγμα 13.1 έχουμε ότι $\bar{y}_{1\cdot} = 316.625$, $\bar{y}_{2\cdot} = 256.4444$, $\bar{y}_{3\cdot} = 278$, ενώ $\hat{\sigma}^2 = 2090.3209$, οπότε $\hat{\sigma} = \sqrt{2090.3209} = 45.72$. Στη συνέχεια, χρησιμοποιώντας τη σχέση (13.14), θα προσδιοριστούν τα διαστήματα εμπιστοσύνης για όλες τις δυνατές διαφορές για $\alpha = 0.05$, $N - k = 22 - 3 = 19$ και $t_{N-k,\alpha/2} = t_{19,0.025} = 2.093$.

► Ομάδα 1 - Ομάδα 2

Για τη σύγκριση των Ομάδων 1 και 2 έχουμε με αντικατάσταση στη σχέση (13.14):

$$\left(316.625 - 256.444 - 2.093 \cdot 45.72\sqrt{\frac{1}{8} + \frac{1}{9}},\ 316.625 - 256.444 + 2.093 \cdot 45.72\sqrt{\frac{1}{8} + \frac{1}{9}}\right),$$

οπότε, μετά από λίγη άλγεβρα, προκύπτει ότι το διάστημα εμπιστοσύνης για τη διαφορά $\mu_1 - \mu_2$ είναι το $(13.682, 106.679)$. Παρατηρούμε ότι το 0 δεν ανήκει στο 95% διάστημα εμπιστοσύνης για τη διαφορά $\mu_1 - \mu_2$, οπότε μπορούμε να ισχυριστούμε ότι υπάρχει στατιστικά σημαντική διαφορά στο μέσο επίπεδο φολικού οξέος των ασθενών της ομάδας 1 και της ομάδας 2, σε επίπεδο σημαντικότητας 0.05.

► Ομάδα 1 - Ομάδα 3

Για τη σύγκριση των Ομάδων 1 και 3 έχουμε με αντικατάσταση στη σχέση (13.14):

$$\left(316.625 - 278 - 2.093 \cdot 45.72\sqrt{\frac{1}{8} + \frac{1}{5}},\ 316.625 - 278 + 2.093 \cdot 45.72\sqrt{\frac{1}{8} + \frac{1}{5}}\right),$$

οπότε, μετά από λίγη άλγεβρα, προκύπτει ότι το διάστημα εμπιστοσύνης για τη διαφορά $\mu_1 - \mu_3$ είναι το $(-15.9285,\ 93.1785)$. Παρατηρούμε ότι το 0 ανήκει στο 95% διάστημα εμπιστοσύνης για τη διαφορά $\mu_1 - \mu_3$, οπότε μπορούμε να ισχυριστούμε ότι δεν υπάρχει στατιστικά σημαντική διαφορά στο μέσο επίπεδο φολικού οξέος των ασθενών της ομάδας 1 και της ομάδας 3, σε επίπεδο σημαντικότητας 0.05.

► Ομάδα 2 - Ομάδα 3

Για τη σύγκριση των Ομάδων 2 και 3 έχουμε με αντικατάσταση στη σχέση (13.14):

$$\left(256.444 - 278 - 2.093 \cdot 45.72\sqrt{\frac{1}{9}+\frac{1}{5}},\ 256.444 - 278 + 2.093 \cdot 45.72\sqrt{\frac{1}{9}+\frac{1}{5}}\right),$$

οπότε, μετά από λίγη άλγεβρα, προκύπτει ότι το διάστημα εμπιστοσύνης για τη διαφορά $\mu_1 - \mu_3$ είναι το $(-74.9306,\ 31.8195)$. Παρατηρούμε ότι το 0 ανήκει στο 95% διάστημα εμπιστοσύνης για τη διαφορά $\mu_2 - \mu_3$, οπότε μπορούμε να ισχυριστούμε ότι δεν υπάρχει στατιστικά σημαντική διαφορά στο μέσο επίπεδο φολικού οξέος των ασθενών της ομάδας 2 και της ομάδας 3, σε επίπεδο σημαντικότητας 0.05.

Για να υλοποιήσουμε την παραπάνω μέθοδο πολλαπλών συγκρίσεων μπορούμε να χρησιμοποιήσουμε τις παρακάτω εντολές της R :

```
d1 <- data.frame(
  folic_acid=c(243,251,275,291,347,354,380,392,
               206,210,226,249,255,273,285,295,309,
               241,258,270,293,328),
  patient_group=c( rep("Group1",8), rep("Group2",9),rep("Group3",5) )
)
anova <- aov(d1$folic_acid ~ d1$patient_group)
library(DescTools)
PostHocTest(anova, method = "lsd")
```

από όπου λαμβάνουμε τα παρακάτω αποτελέσματα

```
  Posthoc multiple comparisons of means : Fisher LSD
    95% family-wise confidence level

$`d1$patient_group`
                   diff     lwr.ci    upr.ci   pval
Group2-Group1 -60.18056 -106.67905 -13.68206 0.0139 *
Group3-Group1 -38.62500  -93.17847  15.92847 0.1548
Group3-Group2  21.55556  -31.81952  74.93063 0.4085

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Παρατηρούμε ότι η p-τιμή του ελέγχου των ομάδων 1 και 2 είναι $0.0139 < 0.05$, άρα απορρίπτεται η μηδενική υπόθεση, όπως και πριν. Επομένως, προκύπτει ότι υπάρχει στατιστικά σημαντική διαφορά στο μέσο επίπεδο φολικού οξέος στα ερυθροκύτταρα των ασθενών των ομάδων 1 και 2, κάτι που δεν ισχύει για τη σύγκριση του μέσου φολικού οξέος στα ερυθροκύτταρα των ασθενών των ομάδων 1 και 3 και 2 και 3. Η $p-$τιμή του ελέγχου των ομάδων 1 και 3 ισούται με $0.1548 > 0.05$ και των ομάδων 2 και 3 0.4085, οπότε και δεν μπορεί να απορριφθεί η μηδενική υπόθεση.

Πέρα από την παραπάνω προσέγγιση έχουν προταθεί πολλοί άλλοι τρόποι κατασκευής αυτών των διαστημάτων εμπιστοσύνης στη βιβλιογραφία. Στη συνέχεια θα αναφερθούμε σε κάποιους από αυτούς.

13.3.2 Η μέθοδος Bonferroni

Κατασκευάζοντας τα διαστήματα εμπιστοσύνης με τον τρόπο που περιγράψαμε στην προηγούμενη ενότητα, ο από κοινού συντελεστής εμπιστοσύνης των διαστημάτων μειώνεται σημαντικά. Πιο συγκεκριμένα, έστω ότι έχουμε τα διαστήματα εμπιστοσύνης I_{12} και I_{13} για τις διαφορές $\mu_1 - \mu_2$ και $\mu_1 - \mu_3$, αντίστοιχα, που το καθένα από αυτά έχει συντελεστή εμπιστοσύνης $1-\alpha$, τότε

$$P\left(\mu_1 - \mu_2 \in I_{12},\ \mu_1 - \mu_3 \in I_{13}\right) < P\left(\mu_1 - \mu_2 \in I_{12}\right) = 1 - \alpha.$$

Έστω λοιπόν ότι θέλουμε να κατασκευάσουμε $\mathrm{K} = k(k-1)/2$ διαστήματα εμπιστοσύνης για όλες τις δυνατές διαφορές $\mu_i - \mu_j$. Για ευκολία, συμβολίζουμε με $\delta_1, \ldots, \delta_K$ όλες τις δυνατές διαφορές των μέσων των k πληθυσμών και $I_1, \ldots, I_K$ τα αντίστοιχα τυχαία διαστήματα εμπιστοσύνης τους. Τότε

$$\begin{aligned} P(\delta_1 \in I_1 \cap \ldots \cap \delta_K \in I_K) &= 1 - P\left([\cap_{i=1}^{K}(\delta_i \in I_i)]'\right) = 1 - P\left(\cup_{i=1}^{K}(\delta_i \in I_i)'\right) \\ &= 1 - P\left(\cup_{i=1}^{K}(\delta_i \notin I_i)\right) \geq 1 - \sum_{i=1}^{K} P(\delta_i \notin I_i). \end{aligned}$$

Η τελευταία ανισότητα προκύπτει από την ανισότητα Bonferroni (Galambos and Simonelli, 1996), σύμφωνα με την οποία, αν $A_1, \ldots, A_K,$, $K \geq 2$, είναι K το πλήθος ενδεχόμενα ενός δειγματικού χώρου Ω, τότε

$$P\left(\cup_{i=1}^{K} A_i\right) \leq \sum_{i=1}^{K} P(A_i).$$

Επομένως, έχουμε ότι:

$$P\left(\cup_{i=1}^{K} A_i\right) \leq \sum_{i=1}^{K} P(A_i).$$

Αν τα διαστήματα εμπιστοσύνης $I_1, \cdots, I_K$ έχουν επίπεδο εμπιστοσύνης $1-\alpha$, τότε $P(\delta_i \notin I_i) = \alpha$, οπότε θα έχουμε

$$P\left(\cap_{i=1}^{K}(\delta_i \in I_i)\right) \geq 1 - K\alpha.$$

Αν λοιπόν επιθυμούμε να έχουμε κοινό συνολικό συντελεστή εμπιστοσύνης τουλάχιστον $1-\alpha$, θα πρέπει να κατασκευάσουμε το καθένα από αυτά τα K διαστήματα εμπιστοσύνης να είναι $100(1-\alpha/K)\%$ διάστημα εμπιστοσύνης. Για παράδειγμα, αν $k = 10$ και θέλουμε ο από κοινού συντελεστής εμπιστοσύνης να είναι τουλάχιστον 0.95, δηλαδή αν θέλουμε $\alpha = 0.05$, θα πρέπει συνολικά να κατασκευάσουμε $k(k-1)/2 = 10 \cdot 9/2 = 45$ $(1 - 0.05/45)\%$ διαστήματα εμπιστοσύνης.

Γενικότερα, σε αυτήν την περίπτωση τα K το πλήθος διαστήματα εμπιστοσύνης θα δίνονται από τη σχέση:

$$\left(\bar{y}_{i\cdot} - \bar{y}_{j\cdot} - t_{N-k,\alpha/(k(k-1))}\hat{\sigma}\sqrt{1/n_i + 1/n_j},\ \bar{y}_{i\cdot} - \bar{y}_{j\cdot} + t_{N-k,\alpha/(k(k-1))}\hat{\sigma}\sqrt{1/n_i + 1/n_j}\right). \tag{13.15}$$

Όπως ήδη έχουμε αναφέρει στο Κεφάλαιο 11 αυτού του βιβλίου, σε έναν έλεγχο υπόθεσης μπορούμε να αποφασίσουμε αν θα απορρίψουμε τη μηδενική υπόθεση στηριζόμενη στην p-τιμή του ελέγχου (έλεγχος σημαντικότητας). Σε συνέχεια των παραπάνω και λαμβάνοντας υπόψη και όσα αναφέρθηκαν στην εισαγωγή αυτού του κεφαλαίου για το πραγματικό επίπεδο σημαντικότητας των ελέγχων, θα μπορούσαμε να εφαρμόσουμε τη μέθοδο Bonferroni ορίζοντας ως επίπεδο σημαντικότητας του κάθε ελέγχου το α/K και έπειτα να συγκρίνουμε τις $\mathrm{K} = \frac{k(k-1)}{2}$ το πλήθος $p-$τιμές με αυτό. Με αυτήν τη διαδικασία το συνολικό επίπεδο σημαντικότητας του ελέγχου θα είναι το πολύ ίσο με α. Μία παραλλαγή αυτής της μεθόδου αποτελεί αντικείμενο της επόμενης υποενότητας.

Παράδειγμα 13.3

Σε συνέχεια του Παραδείγματος 13.1, εφόσον απορρίφθηκε η μηδενική υπόθεση της ισότητας του μέσου φολικού οξέος στα ερυθροκύτταρα των ασθενών των τριών ομάδων, χρησιμοποιήστε τη μέθοδο Bonferroni για να προχωρήσετε σε περαιτέρω πολλαπλές ζευγαρωτές συγκρίσεις όλων των ομάδων ασθενών ανά δύο, σε κοινό συνολικό επίπεδο εμπιστοσύνης τουλάχιστον 0.95.

Λύση Παραδείγματος 13.3

Προκειμένου να προβούμε στους ελέγχους της μορφής

$$H_0 : \mu_i = \mu_j \text{ κατά } H_1 : \mu_i \neq \mu_j, \text{για } i,j = 1,2,3 \text{ με } i \neq j,$$

σε συνολικό επίπεδο εμπιστοσύνης $1 - \alpha = 0.95$ χρησιμοποιούμε τη σχέση (13.15) για την κατασκευή των αντίστοιχων διαστημάτων εμπιστοσύνης, με $\alpha = 0.05$, $N - k = 22 - 3 = 19$ και $t_{N-k,\alpha/(k(k-1))} = t_{19,0.025/3} = 2.625$ (Πίνακας Α΄.4, Παραρτήματος Α΄). Οπότε, σύμφωνα με τα δεδομένα του Παραδείγματος 13.1, θα έχουμε:

► Ομάδα 1 - Ομάδα 2

Για τη σύγκριση των Ομάδων 1 και 2 έχουμε με αντικατάσταση στη σχέση (13.15):

$$\left(316.625 - 256.444 - 2.625 \cdot 45.72\sqrt{\frac{1}{8} + \frac{1}{9}},\ 316.625 - 256.444 + 2.625 \cdot 45.72\sqrt{\frac{1}{8} + \frac{1}{9}}\right),$$

οπότε, μετά από λίγη άλγεβρα, προκύπτει ότι το διάστημα εμπιστοσύνης για τη διαφορά $\mu_1 - \mu_2$ είναι το $(1.8614,\ 119.5)$. Παρατηρούμε ότι το 0 δεν ανήκει στο παραπάνω διάστημα εμπιστοσύνης για τη διαφορά $\mu_1 - \mu_2$, οπότε με τη μέθοδο Bonferonni μπορούμε να ισχυριστούμε ότι υπάρχει στατιστικά σημαντική διαφορά στο μέσο επίπεδο φολικού οξέος των ασθενών της ομάδας 1 και της ομάδας 2, σε επίπεδο σημαντικότητας 0.05.

► Ομάδα 1 - Ομάδα 3

Για τη σύγκριση των Ομάδων 1 και 3 έχουμε με αντικατάσταση στη σχέση (13.15):

$$\left(316.625 - 278 - 2.2,625 \cdot 45.72\sqrt{\frac{1}{8} + \frac{1}{5}},\ 316.625 - 278 + 2.625 \cdot 45.72\sqrt{\frac{1}{8} + \frac{1}{5}}\right),$$

οπότε, μετά από λίγη άλγεβρα, προκύπτει ότι το διάστημα εμπιστοσύνης για τη διαφορά $\mu_1 - \mu_3$ είναι το $(-29.80,\ 107.055)$. Παρατηρούμε ότι το 0 ανήκει στο παραπάνω διάστημα εμπιστοσύνης για τη διαφορά $\mu_1 - \mu_3$, οπότε με τη μέθοδο Bonferonni μπορούμε να ισχυριστούμε ότι δεν υπάρχει στατιστικά σημαντική διαφορά στο μέσο επίπεδο φολικού οξέος των ασθενών της ομάδας 1 και της ομάδας 3, σε επίπεδο σημαντικότητας 0.05.

► Ομάδα 2 - Ομάδα 3

Για τη σύγκριση των Ομάδων 2 και 3 έχουμε με αντικατάσταση στη σχέση (13.15):

$$\left(256.444 - 278 - 2.625 \cdot 45.72\sqrt{\frac{1}{9} + \frac{1}{5}},\ 256.444 - 278 + 2.625 \cdot 45.72\sqrt{\frac{1}{9} + \frac{1}{5}}\right),$$

οπότε, μετά από λίγη άλγεβρα, προκύπτει ότι το διάστημα εμπιστοσύνης για τη διαφορά $\mu_2 - \mu_3$ είναι το $(-88.5,\ 45.39)$. Παρατηρούμε ότι το 0 ανήκει στο παραπάνω διάστημα εμπιστοσύνης για τη διαφορά $\mu_2 - \mu_3$, οπότε με τη μέθοδο Bonferonni μπορούμε να ισχυριστούμε ότι δεν υπάρχει στατιστικά σημαντική διαφορά στο μέσο επίπεδο φολικού οξέος των ασθενών της ομάδας 2 και της ομάδας 3, σε επίπεδο σημαντικότητας 0.05.

Εναλλακτικά, θα μπορούσαμε να υλοποιήσουμε όλες τις δυνατές ζευγαρωτές συγκρίσεις μεταξύ των τριών ομάδων ασθενών με τη μέθοδο Bonferroni χρησιμοποιώντας τις παρακάτω εντολές της R:

```r
d1 <- data.frame(
  folic_acid=c(243,251,275,291,347,354,380,392,
               206,210,226,249,255,273,285,295,309,
               241,258,270,293,328),
  patient_group=c( rep("Group1",8), rep("Group2",9),rep("Group3",5) )
)
anova <- aov(d1$folic_acid ~ d1$patient_group)
library(DescTools)
PostHocTest(anova, method = "bonferroni")
```

Τότε λαμβάνουμε τα ακόλουθα αποτελέσματα

```
 Posthoc multiple comparisons of means : Bonferroni
    95% family-wise confidence level

$`d1$patient_group`
                    diff     lwr.ci   upr.ci   pval
Group2-Group1 -60.18056 -118.49975 -1.86136 0.0418 *
Group3-Group1 -38.62500 -107.04688 29.79688 0.4643
Group3-Group2  21.55556  -45.38835 88.49947 1.0000

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Παρατηρούμε ότι στατιστικά σημαντική διαφορά στο μέσο φολικό οξύ των ερυθροκυττάρων υπάρχει μόνο μεταξύ των ασθενών της ομάδας 1 και 2, αφού $p - value = 0.0418 < 0.05$ και, επομένως, καταλήγουμε στα ίδια αποτελέσματα με πριν.

13.3.3 Η μέθοδος Bonferroni-Holm

Ο Holm (1979) πρότεινε μία παραλλαγή της διαδικασίας που περιγράφθηκε στην προηγούμενη ενότητα. Πιο συγκεκριμένα, πρότεινε αρχικά τον υπολογισμό των $p-$τιμών των K το πλήθος ελέγχων και, στη συνέχεια, τη διάταξή τους σε αύξουσα σειρά, $p_{(1)} \leq \cdots \leq p_{(K)}$, όπου $p_{(i)}$ η i-οστή διατεταγμένη σε αύξουσα $p-$τιμή, οι οποίες αντιστοιχούν στις $H_{0(1)}, H_{0(2)}, \ldots, H_{0(K)}$ μηδενικές υποθέσεις. Σε αυτό το πλαίσιο, η μεθοδολογία που προτάθηκε και είναι γνωστή ως Bonferroni-Holm είναι η ακόλουθη.

Αν $p_{(1)} > \alpha/K$, τότε δεν απορρίπτεται καμία από τις παραπάνω μηδενικές υποθέσεις, ενώ διαφορετικά απορρίπτεται η $H_{0(1)}$ και πραγματοποιείται ο επόμενος έλεγχος.

Αν $p_{(2)} > \alpha/(K-1)$, τότε δεν απορρίπτεται καμία από τις υπόλοιπες $K-1$ μηδενικές υποθέσεις, ενώ διαφορετικά απορρίπτεται η $H_{0(2)}$ και συνεχίζουμε με αντίστοιχο τρόπο στον επόμενο έλεγχο.

Η διαδικασία συνεχίζει μέχρι την πρώτη $H_{0(i)}$ που δεν θα απορριφθεί, οπότε δεν απορρίπτονται και όσες έπονται. Διαφορετικά καταλήγουμε να ελέγξουμε και την $H_{0(K)}$ υπόθεση, η οποία, αν $p_{(K)} > \alpha$, δεν την απορρίπτουμε, διαφορετικά την απορρίπτουμε και αυτή.

Η μέθοδος θα γίνει περισσότερο κατανοητή μέσω του παραδείγματος που ακολουθεί.

Παράδειγμα 13.4

Θέλουμε να ελέγξουμε, σε επίπεδο σημαντικότητας 0.05, πέντε το πλήθος μηδενικές υποθέσεις, τις $H_{0,1}, \cdots, H_{0,5}$. Ποιες από τις παραπάνω μηδενικές υποθέσεις θα απορριφθούν με τη μέθοδο Bonferroni-Holm, αν οι $p-$τιμές των ελέγχων είναι 0.03, 0.07, 0.002, 0.011 και 0.052, αντίστοιχα;

Λύση Παραδείγματος 13.4

Οι $p-$τιμές των παραπάνω ελέγχων είναι 0.03, 0.07, 0.002, 0.011, 0.052. Αρχικά, τις διατάσσουμε κατ΄ αύξουσα σειρά και έχουμε 0.002, 0.011, 0.03, 0.052, 0.07. Επομένως, προκύπτει ότι:

$$H_{0,(1)} = H_{0,3}, H_{0,(2)} = H_{0,4}, H_{0,(3)} = H_{0,1}, H_{0,(4)} = H_{0,5}, H_{0,(5)} = H_{0,2}.$$

Αρχικά, συγκρίνουμε την ελάχιστη $p-$τιμή με το $0.05/5 = 0.01$. Επειδή $0.002 < 0.01$, η $H_{0,(1)}$ (δηλαδή η $H_{0,3}$) απορρίπτεται. Συνεχίζοντας, συγκρίνουμε την επόμενη $p-$ τιμή με το $0.05/4 = 0.0125$. Αφού $0.011 < 0.0125$ απορρίπτεται και η $H_{0,(2)}$ (δηλαδή η $H_{0,4}$). Στο τρίτο βήμα της διαδικασίας, συγκρίνουμε την επόμενη $p-$ τιμή με το $0.05/3 = 0.0167$. Επειδή $0.03 > 0.0167$, αυτή η μηδενική υπόθεση, καθώς και οι υπόλοιπες μηδενικές υποθέσεις δεν απορρίπτονται. Επομένως, δεν απορρίπτονται οι $H_{0,(3)}$ έως και την τελευταία, δηλαδή την $H_{0,(5)}$, δηλαδή δεν απορρίπτονται οι $H_{0,1}, H_{0,2}$ και $H_{0,5}$.
Συνοψίζοντας, η μέθοδος Bonferroni-Holm απέρριψε σε επίπεδο σημαντικότητας 0.05 τις υποθέσεις $H_{0,3}, H_{0,4}$ και δεν απέρριψε τις υποθέσεις $H_{0,1}, H_{0,2}$ και $H_{0,5}$.

Παρατήρηση 13.5

Σε αυτό το σημείο, θα πρέπει να επισημάνουμε ότι η κλασική μέθοδος Bonferroni θα απέρριπτε μόνο την υπόθεση $H_{0,3}$, αφού μόνο για αυτήν τη μηδενική υπόθεση η $p-$ τιμή του ελέγχου είναι μικρότερη του $\alpha/K = 0.05/5 = 0.01$.

Παράδειγμα 13.5

Σε συνέχεια του Παραδείγματος 13.1, εφόσον απορρίφθηκε η μηδενική υπόθεση της ισότητας του μέσου φολικού οξέος στα ερυθροκύτταρα των ασθενών των τριών ομάδων, χρησιμοποιήστε την R και τη μέθοδο Bonferroni-Holm, για να προχωρήσετε σε περαιτέρω πολλαπλές ζευγαρωτές συγκρίσεις όλων των ομάδων ασθενών ανά δύο, σε επίπεδο σημαντικότητας 0.05.

Λύση Παραδείγματος 13.5

Για την υλοποίηση της μεθόδου Bonferroni-Holm χρησιμοποιούμε τις εντολές:

```
d1 <- data.frame(
  folic_acid=c(243,251,275,291,347,354,380,392,206,210,226,
               249,255,273,285,295,309,241,258,270,293,328),
  patient_group=c( rep("Group1",8), rep("Group2",9),rep("Group3",5) )
)
pairwise.t.test(d1$folic_acid, d1$patient_group, p.adj = "holm")
```

Τότε λαμβάνουμε τα παρακάτω αποτελέσματα:

```
Pairwise comparisons using t tests with pooled SD

data:  d1$folic_acid and d1$patient_group

       Group1 Group2
Group2 0.042  -
Group3 0.310  0.408

P value adjustment method: holm
```

Παρατηρούμε ότι στατιστικά σημαντική διαφορά στο μέσο φολικό οξύ των ερυθροκυττάρων υπάρχει μόνο μεταξύ των ασθενών της ομάδας 1 και 2, αφού $p - value = 0.042 < 0.05$.

13.3.4 Η μέθοδος Tukey

Η μέθοδος Tukey (Tukey, 1977) χρησιμοποιείται κυρίως όταν τα μεγέθη των k το πλήθος τυχαίων δειγμάτων είναι ίσα, δηλαδή όταν έχουμε $n_1 = \ldots = n_k = n$, οπότε $N = n \cdot k$. Σε αυτήν την περίπτωση, θέλουμε να κατασκευάσουμε ένα διάστημα εμπιστοσύνης της μορφής

$$P\left(\cap_{1\le i<j\le K}\left(\bar{Y}_{i\cdot} - \bar{Y}_{j\cdot} - q\frac{\hat{\sigma}}{\sqrt{n}} \le \mu_i - \mu_j \le \bar{Y}_{i\cdot} - \bar{Y}_{j\cdot} + q\frac{\hat{\sigma}}{\sqrt{n}}\right)\right) = 1 - \alpha.$$

Το q εξαρτάται από το πλήθος των δειγμάτων k, το σύνολο των διαθέσιμων παρατηρήσεων N, καθώς και από τον επιθυμητό συντελεστή εμπιστοσύνης $1 - \alpha$. Αν συμβολίσουμε το επιθυμητό q ως $q_{k,N,\alpha}$, τότε οι πολλαπλές συγκρίσεις των μέσων, χρησιμοποιώντας τη μέθοδο του Tukey, βασίζονται στα παρακάτω διαστήματα εμπιστοσύνης:

$$\left(\bar{y}_{i\cdot} - \bar{y}_{j\cdot} - q_{k,N,\alpha}\hat{\sigma}\sqrt{1/n_i + 1/n_j},\ \bar{y}_{i\cdot} - \bar{y}_{j\cdot} + q_{k,N\alpha}\hat{\sigma}\sqrt{1/n_i + 1/n_j}\right).$$

Το σημείο $q_{k,N,p}$ είναι το $p-$ποσοστιαίο σημείο της κατανομής της τυχαίας μεταβλητής

$$\frac{\max\sqrt{n}\left(\bar{Y}_{i\cdot} - \mu_i\right) - \min\sqrt{n}\left(\bar{Y}_{i\cdot} - \mu_i\right)}{\hat{\sigma}}$$

και μπορεί να βρεθεί από ειδικούς πίνακες αντίστοιχους με αυτούς γνωστών κατανομών (κανονικής κατανομής, κατανομής t κ.ά.). Στο πλαίσιο του παρόντος συγγράμματος δεν παρατίθεται κάποιος πίνακας και η υλοποίηση του ελέγχου διεξάγεται με τη βοήθεια της R, με τον τρόπο που θα δούμε στο παράδειγμα που ακολουθεί.

Παράδειγμα 13.6

Σε συνέχεια του Παραδείγματος 13.1, εφόσον απορρίφθηκε η μηδενική υπόθεση της ισότητας του μέσου φολικού οξέος στα ερυθροκύτταρα των ασθενών των τριών ομάδων, χρησιμοποιήστε τη μέθοδο Tukey, για να προχωρήσετε σε περαιτέρω πολλαπλές ζευγαρωτές συγκρίσεις όλων των ομάδων ασθενών ανά δύο σε επίπεδο σημαντικότητας 0.05.

Λύση Παραδείγματος 13.6

Για την υλοποίηση της μεθόδου Tukey χρησιμοποιούμε την ακόλουθη εντολή:

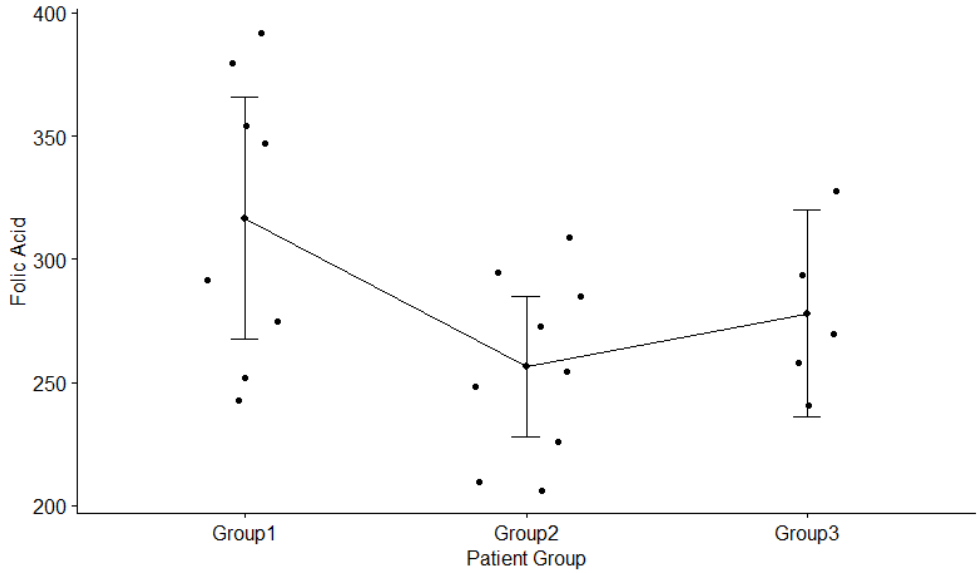
```
d1 <- data.frame(
  folic_acid=c(243,251,275,291,347,354,380,392,
               206,210,226,249,255,273,285,295,309,
               241,258,270,293,328),
  patient_group=c( rep("Group1",8), rep("Group2",9),rep("Group3",5) )
)
res.aov <- aov(formula =  folic_acid~patient_group,data = d1)
TukeyHSD(res.aov)
```

και λαμβάνουμε τα εξής αποτελέσματα:

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = folic_acid ~ patient_group, data = d1)

$patient_group
```



Σχήμα 13.2: Μέσες τιμές και τα αντίστοιχα 95% διαστήματα εμπιστοσύνης του επιπέδου φολικού οξέος στα ερυθροκύτταρα των τριών ομάδων ασθενών του Παραδείγματος 13.1.

```
                    diff        lwr      upr     p adj
Group2-Group1 -60.18056 -116.61904 -3.74207 0.0354792
Group3-Group1 -38.62500 -104.84037 27.59037 0.3214767
Group3-Group2  21.55556  -43.22951 86.34062 0.6802018
```

Παρατηρούμε ότι υπάρχει στατιστικά σημαντική διαφορά στο μέσο φολικό οξύ των ερυθροκυττάρων μόνο μεταξύ των ασθενών της ομάδας 1 και 2, αφού ισχύει ότι $p-$τιμή $= 0.0354792 < 0.05$.

Παρατήρηση 13.6

Συνοψίζοντας τα συμπεράσματα των Παραδειγμάτων 13.2, 13.3, 13.5 και 13.6, προκύπτει ότι και με τις τρεις μεθόδους καταλήγουμε στα ίδιο ακριβώς συμπέρασμα. Αυτό φυσικά δεν ισχύει γενικά, όπως ήδη έχουμε δει στο Παράδειγμα 13.4 (βλ. και την Παρατήρηση 13.5). Για καλύτερη κατανόηση των παραπάνω αποτελεσμάτων, μπορούμε να κατασκευάσουμε ένα γράφημα στο οποίο απεικονίζεται το μέσο φολικό οξύ στα ερυθροκύτταρα των ασθενών της κάθε ομάδας μαζί με τα αντίστοιχα 95% διαστήματα εμπιστοσύνης τους. Προκειμένου να κατασκευάσουμε το παραπάνω γράφημα στην R, χρησιμοποιούμε τις ακόλουθες εντολές:

```
ggline(d1, x = "patient_group", y = "folic_acid",
       add = c("mean_ci", "jitter"),
       ylab = "Folic Acid", xlab = "Patient Group")
```

Το αποτέλεσμα που προκύπτει δίνεται στο Σχήμα 13.2, στο οποίο αποτυπώνονται, πέρα από τις παρατηρήσεις σε κάθε ομάδα, και τα αντίστοιχα διαστήματα εμπιστοσύνης για τις μέσες τιμές.

Άσκηση Αυτοαξιολόγησης 13.2

Σε συνέχεια της Άσκησης Αυτοαξιολόγησης 13.1, εφόσον απορριφθεί η μηδενική υπόθεση της ισότητας του μέσου συντελεστή ελαστικότητας των τριών τύπων κραμάτων, χρησιμοποιήστε τις μεθόδους Bonferroni, Bonferroni-Holm και Tukey, για να προχωρήσετε σε περαιτέρω πολλαπλές ζευγαρωτές συγκρίσεις όλων των τύπων κραμάτων ανά δύο, σε επίπεδο σημαντικότητας 0.05. Υπόδειξη: να λύσετε την άσκηση χρησιμοποιώντας την R.

13.4 Έλεγχος ισότητας διασπορών

Όπως ήδη έχουμε αναφέρει σε προηγούμενη ενότητα αυτού του κεφαλαίου, μια σημαντική υπόθεση για τον έλεγχο ANOVA είναι η ισότητα των διασπορών των k πληθυσμών από τους οποίους έχουν επιλεχθεί τα k το πλήθος τυχαία δείγματα. Επομένως, η υπόθεση που πρέπει να ελέγξουμε είναι η εξής:

$$H_0 : \sigma_1^2 = \ldots = \sigma_k^2 \text{ κατά } H_1 : \text{ όχι η } H_0.$$

Υπό την υπόθεση της κανονικότητας των πληθυσμών, ο Bartlett (1937) απέδειξε ότι το τεστ πηλίκου πιθανοφανειών για τον παραπάνω έλεγχο δίνεται από τη σχέση:

$$B = \frac{\left((S_1^2)^{n_1-1}(S_2^2)^{n_2-1}\ldots(S_k^2)^{n_k-1}\right)^{1/(N-k)}}{S_p^2}$$

όπου $S_1^2, \ldots, S_k^2$ οι k το πλήθος δειγματικές διασπορές και

$$S_p^2 = \frac{\sum_{i=1}^{k}(n_i-1)S_i^2}{N-k}.$$

Επιπλέον, προσδιόρισε την ακριβή δειγματική κατανομή της στατιστικής συνάρτησης B υπό την υπόθεση ότι τα δείγματα είναι ίσου μεγέθους, ήτοι υπό την υπόθεση ότι $n_1 = n_2 = \ldots = n_k$. Σε αυτήν την περίπτωση, απορρίπτεται η μηδενική υπόθεση σε επίπεδο σημαντικότητας α, αν $b < b_{k,n,\alpha}$, όπου $b_{k,n,\alpha}$ το α ποσοστιαίο σημείο της κατανομής Bartlett. Αν τα μεγέθη των δειγμάτων είναι διαφορετικά, τότε η μηδενική υπόθεση απορρίπτεται, αν

$$b < b_{k,n_1,\ldots,n_k,\alpha}$$

όπου

$$b_{k,n_1,\ldots,n_k,\alpha} \approx \frac{n_1 b_{k,n_1,\alpha} + \ldots + n_k b_{k,n_k,\alpha}}{N}.$$

Εναλλακτικά, ο Bartlett εισήγαγε τη στατιστική συνάρτηση ελέγχου που δίνεται από τη σχέση[2]:

$$M = \frac{(N-k)\log(S_p^2) - \sum_{i=1}^{k}\log(S_i^2)}{1 + \frac{1}{3(k-1)}\left(\sum_{i=1}^{k}\left(\frac{1}{n_i-1}\right) - \frac{1}{N-k}\right)},$$

η οποία ακολουθεί, υπό τη μηδενική υπόθεση της ισότητας των πληθυσμιακών διασπορών, προσεγγιστικά χι-τετράγωνο κατανομή με $k-1$ βαθμούς ελευθερίας. Επομένως, καθώς η μηδενική υπόθεση απορρίπτεται για μεγάλες τιμές της στατιστικής συνάρτησης ελέγχου, η περιοχή απόρριψης της υπόθεσης της ομοσκεδαστικότητας σε προσεγγιστικό επίπεδο σημαντικότητας α είναι $M \geq \chi^2_{k-1,\alpha}$.

Ο έλεγχος αυτός υλοποιείται στην R μέσω της συνάρτησης `bartlett.test` της βιβλιοθήκης `stats`. Στη συνέχεια, δίνεται η υλοποίηση για τα δεδομένα του Παραδείγματος 13.1.

[2] Ο έλεγχος αυτός αποτελεί τροποποίηση του τεστ πηλίκου πιθανοφανειών σε μια προσπάθεια να επιτευχθεί καλύτερη προσέγγιση στη χι-τετράγωνο κατανομή με $k-1$ βαθμούς ελευθερίας.

```r
library(stats)
d1 <- data.frame(
  folic_acid=c(243,251,275,291,347,354,380,392,
               206,210,226,249,255,273,285,295,309,
               241,258,270,293,328),
  patient_group=c( rep("Group1",8), rep("Group2",9),rep("Group3",5) )
)
bartlett.test(d1$folic_acid ~ d1$patient_group,d1)
```

Τότε λαμβάνουμε τα εξής αποτελέσματα:

```
 Bartlett test of homogeneity of variances

data:  d1$folic_acid by d1$patient_group
Bartlett's K-squared = 2.0951, df = 2, p-value = 0.3508
```

Παρατηρούμε ότι η υπόθεση της ισότητας των πληθυσμιακών διακυμάνσεων δεν μπορεί να απορριφθεί σε κάποιο από τα συνήθη επίπεδα σημαντικότητας.

Παρατήρηση 13.7

Παρόλο που ο έλεγχος Bartlett χρησιμοποιείται συχνά, υπάρχουν και άλλοι έλεγχοι ισότητας των διασπορών k το πλήθος ανεξάρτητων πληθυσμών. Ένας από τους πιο δημοφιλείς είναι ο αποκαλούμενος έλεγχος Levene (Levene, 1960), ο οποίος είναι λιγότερο ευαίσθητος από τον έλεγχο Bartlett σε αποκλίσεις από την κανονικότητα.
Η σσε του έλεγχου Levene είναι η

$$B = \frac{(N-k)\sum_{i=1}^{k} n_i(\bar{Z}_{i\cdot} - \bar{Z}_{\cdot\cdot})^2)}{(k-1)\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Z_{ij} - \bar{Z}_{i\cdot})^2)}$$

όπου $Z_{ij} = |Y_{ij} - m_i|$, $i = 1,\dots,k$, $j = 1,\dots,n_i$ και m_i είτε (α) ο μέσος του i–οστού δείγματος, $m_i = \bar{Y}_{i\cdot}$ είτε (β) η διάμεσος του i–οστού δείγματος είτε (γ) ο περικομμένος μέσος (`trimmed mean`). Επίσης, ισχύει $\bar{Z}_{i\cdot} = \frac{\sum_{j=1}^{n_i} Z_{ij}}{n_i}$ ο μέσος των Z_{ij} της i-οστής ομάδας και $\bar{Z}_{\cdot\cdot} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i} Z_{ij}}{N}$ ο συνολικός μέσος των Z_{ij}.
Ο έλεγχος Levene υλοποιείται στην R μέσω της συνάρτησης `leveneTest` της βιβλιοθήκης `library(car)`.
Για το Παράδειγμα 13.1 η υλοποίηση επιτυγχάνεται μέσω της εντολής

```r
library(car)
d1 <- data.frame(
  folic_acid=c(243,251,275,291,347,354,380,392,
               206,210,226,249,255,273,285,295,309,
               241,258,270,293,328),
  patient_group=c( rep("Group1",8), rep("Group2",9),rep("Group3",5) )
)
leveneTest(d1$folic_acid ~ d1$patient_group,d1)
```

Τότε λαμβάνουμε τα εξής αποτελέσματα:

```
 Levene's Test for Homogeneity of Variance (center = median)
      Df F value  Pr(>F)
group  2  3.6413 0.04585 *
```

```
      19
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Παρατηρήστε ότι το συμπέρασμα για την υπόθεση της ισότητας των πληθυσμιακών διακυμάνσεων, σε επίπεδο σημαντικότητας 5%, είναι διαφορετικό με τη χρήση του ελέγχου του Levene από το αντίστοιχο με τη χρήση του ελέγχου του Bartlett. Κάτι τέτοιο είναι αναμενόμενο, καθώς διαφορετικοί έλεγχοι έχουν τη δυνατότητα εντοπισμού διαφορετικών αποκλίσεων από τη μηδενική υπόθεση.

13.5 Ασκήσεις

Άσκηση 13.1 *Να αποδείξετε τη σχέση (13.10).*

Άσκηση 13.2 *Να αποδείξετε ότι* $E\left(\frac{SSW}{N-k}\right) = \sigma^2$.

Άσκηση 13.3 *Να αποδειχθεί πλήρως ότι υπό τη μηδενική υπόθεση* $H_0 : \mu_i = \mu_j$, *ισχύει ότι* $T_{ij} \sim t_{N-k}$ *με* T_{ij} *τη στατιστική συνάρτηση που δίνεται στη σχέση (13.13).*

Άσκηση 13.4 *Να συμπληρώσετε τον πίνακα ANOVA που ακολουθεί*

Μεταβλητότητα	*Βαθμοί ελευθερίας*	*Άθροισμα τετραγώνων*	*Μέσο Άθροισμα τετραγώνων*	F
Μεταξύ των ομάδων	*4*			*6.40*
Εντός των ομάδων			*10.60*	
Ολική		*377.36*		

Άσκηση 13.5 *Σε έναν διαγωνισμό στατιστικής συμμετείχαν ομάδες των πέντε φοιτητών από τρεις διαφορετικές χώρες. Χρησιμοποιώντας τα δεδομένα που παρατίθενται στον πίνακα που ακολουθεί να ελέγξετε πλήρως, σε επίπεδο σημαντικότητας 0.05, αν υπάρχει στατιστικά σημαντική επίδραση του παράγοντα χώρα στην επίδοση στη στατιστική.*

Χώρα					
X1	*8*	*8*	*7*	*6*	*6*
X2	*6*	*9*	*9*	*7*	*6*
X3	*8*	*10*	*7*	*6*	*6*

Άσκηση 13.6 *Στον πίνακα που ακολουθεί καταγράφονται τα εβδομαδιαία έξοδα για βασικά προϊόντα σε ευρώ πέντε τυχαία επιλεγμένων νοικοκυριών από τρεις διαφορετικές χώρες. Να ελέγξετε πλήρως, σε επίπεδο σημαντικότητας 0.05, αν υπάρχει στατιστικά σημαντική επίδραση του παράγοντα χώρα στα εβδομαδιαία έξοδα.*

Χώρα					
X1	*108*	*115*	*113*	*111*	*120*
X2	*121*	*100*	*105*	*103*	*100*
X3	*108*	*99*	*111*	*110*	*113*

Άσκηση 13.7 *Τυχαία δείγματα μεγέθους έξι πελατών τριών διαφορετικών τύπων καταστημάτων επιλέχθηκαν και καταγράφεται η ηλικία τους.*

Τύπος	*Μέση ηλικία*
A	*47.2*
B	*26.8*
C	*29.8*

Επιπλέον δίνεται ότι $SSW = 1377$ *και* $SSB = 1446$. *Υπάρχει, σε επίπεδο σημαντικότητας 0.05, στατιστικά σημαντική διαφοροποίηση στη μέση ηλικία στους τρεις τύπους καταστημάτων;*

Άσκηση 13.8 *Ένας διατροφολόγος θέλει να διαπιστώσει αν υπάρχει διαφοροποίηση στη μέση απώλεια βάρους μεταξύ τριών διαφορετικών μεθόδων απώλειας βάρους. Για τον λόγο αυτό, εφάρμοσε τις τρεις διαφορετικές μεθόδους σε διαφορετικά άτομα και κατέγραψε τα ακόλουθα αποτελέσματα (απώλεια βάρους σε κιλά).*

Δίαιτα 1: 3, 6, 7, 4
Δίαιτα 2: 10, 12, 11, 14, 8, 6
Δίαιτα 3: 8, 3, 2, 5

Υπάρχει, σε επίπεδο σημαντικότητας 0.05, στατιστικά σημαντική διαφοροποίηση στη μέση απώλεια βάρους στους τρεις τύπους δίαιτας;

Άσκηση 13.9 *Μια φαρμακευτική εταιρεία θέλοντας να ελέγξει την επίδραση τριών διαφορετικών αγωγών στη χοληστερόλη επιλέγει τυχαία 15 άτομα και καταγράφει τα επίπεδα της χοληστερόλης τους μετά από έναν μήνα από την έναρξη της αγωγής. Τα αποτελέσματα δίνονται στον πίνακα που ακολουθεί.*

Αγωγή					
A1	*220*	*215*	*272*	*252*	*232*
A2	*215*	*240*	*220*	*250*	*280*
A3	*200*	*205*	*210*	*210*	*202*

Υπάρχει, σε επίπεδο σημαντικότητας 0.05, στατιστικά σημαντική διαφοροποίηση στη μέση χοληστερόλη στις τρεις αγωγές;

Άσκηση 13.10 *Μια ομάδα ερευνητών μελέτησε τη μόλυνση των νερών ενός ποταμού από βιομηχανικά απόβλητα. Ως μέρος αυτής της μελέτης συνέκρινε τη συγκέντρωση διαλυμένου οξυγόνου στα νερά του ποταμού σε τέσσερις περιοχές της κοίτης του, K1, K2, K3 και K4. Η περιοχή K1 βρίσκεται στις εκβολές του ποταμού, ενώ κοντά στην K3 ρίχνονται απόβλητα από παρακείμενη βιομηχανία. Από κάθε περιοχή οι ερευνητές πήραν, με βάση ένα σχέδιο τυχαίας δειγματοληψίας, πέντε δείγματα νερού. Στον πίνακα που ακολουθεί δίνονται οι συγκεντρώσεις διαλυμένου οξυγόνου (σε ppm) στις τέσσερις περιοχές.*

Περιοχή					
K1	*5.9*	*6.1*	*6.3*	*6.1*	*6.0*
K2	*6.3*	*6.6*	*6.4*	*6.4*	*6.5*
K3	*4.8*	*4.3*	*5.0*	*4.7*	*5.1*
K4	*6.0*	*6.2*	*6.1*	*5.8*	*5,7*

Συγκέντρωση διαλυμένου οξυγόνου (σε ppm)

Σε επίπεδο σημαντικότητας 0.05, υποστηρίζουν αυτά τα δεδομένα ότι μεταξύ των τεσσάρων περιοχών υπάρχουν στατιστικά σημαντικές διαφορές στη μέση συγκέντρωση διαλυμένου οξυγόνου;

Άσκηση 13.11 *Μετρήθηκε η ποσότητα χοληστερίνης που περιέχουν τέσσερα διαφορετικά είδη τροφίμων διαίτης, A1, A2, A3 και A4, αντίστοιχα. Από κάθε είδος επιλέχθηκε ένα τυχαίο δείγμα μεγέθους 3 και μετρήθηκε η ποσότητα χοληστερίνης (σε $milligrams$ ανά $170gr$). Οι μετρήσεις δίνονται στον πίνακα που ακολουθεί.*

Ποσότητα Χοληστερίνης (σε $milligrams/170gr$)

A1	*3.6*	*4.1*	*4.0*
A2	*3.1*	*3.2*	*3.9*
A3	*3.2*	*3.5*	*3.5*
A4	*3.5*	*3.8*	*3.8*

Σε επίπεδο σημαντικότητας 0.05, υποστηρίζουν αυτά τα δεδομένα ότι υπάρχουν στατιστικά σημαντικές διαφορές μεταξύ των μέσων ποσοτήτων χοληστερίνης στα τέσσερα είδη τροφίμων;

Άσκηση 13.12 *Όταν χρησιμοποιείται ορείχαλκος σε μια παραγωγή, ο συντελεστής ελαστικότητας E του υλικού είναι συχνά σημαντικός για τη λειτουργικότητα. Ο συντελεστής ελαστικότητας E μετριέται για 6 διαφορετικά κράματα ορείχαλκου και ελέγχονται 5 δείγματα από κάθε κράμα. Τα αποτελέσματα φαίνονται στον παρακάτω πίνακα, όπου ο μετρούμενος συντελεστής ελαστικότητας δίνεται σε* GPa :

M1	*M2*	*M3*	*M4*	*M5*	*M6*
82.5	*82.7*	*92.2*	*96.5*	*88.9*	*75.6*
83.7	*81.9*	*106.8*	*93.8*	*89.2*	*78.1*
80.9	*78.9*	*104.6*	*92.1*	*94.2*	*92.2*
95.2	*83.6*	*94.5*	*87.4*	*91.4*	*87.3*
80.8	*78.6*	*100.7*	*89.6*	*90.1*	*83.8*

1. *Είναι ο μέσος συντελεστής ελαστικότητας για τα έξι διαφορετικά κράματα ορείχαλκου ίσος, σε επίπεδο σημαντικότητας 0.05;*
2. *Μπορείτε να προτείνετε κάποιες επακόλουθες συγκρίσεις;*

13.6 Απαντήσεις στις ασκήσεις αυτοαξιολόγησης

Λύση Άσκησης Αυτοαξιολόγησης 13.1

Έστω μ_1, μ_2 και μ_3 οι πληθυσμιακές μέσες τιμές του συντελεστή ελαστικότητας του πρώτου, δεύτερου και τρίτου τύπου κράματος, αντίστοιχα. Οι υποθέσεις που θέλουμε να ελέγξουμε, σε επίπεδο σημαντικότητας 0.05, είναι:

$$H_0 : \mu_1 = \mu_2 = \mu_3,$$

κατά

$$H_1 : \text{ υπάρχει τουλάχιστον ένα ζεύγος } (i,\ell) \text{ με } i,\ell = 1,2,3, i \neq \ell, \text{τέτοιο ώστε} \mu_i \neq \mu_\ell.$$

Θα ακολουθήσουμε τη μεθοδολογία της ανάλυσης διακύμανσης κατά έναν παράγοντα, δημιουργώντας τον πίνακα ανάλυσης της διασποράς. Σε όσα ακολουθούν y_{ij} είναι η j τιμή του συντελεστή ελαστικότητας στον i-οστό τύπο κράματος, για $i = 1,2,3$ και $j = 1, \ldots, n_i$, όπου $n_1 = 5$, $n_2 = 5$ και $n_3 = 5$.
Αρχικά, θέλοντας να υπολογίσουμε το συνολικό άθροισμα τετραγώνων, υπολογίζουμε τον συνολικό δειγματικό μέσο, είναι

$$\bar{y}_{..} = \sum_{i=1}^{3}\sum_{j=1}^{5} y_{ij}/\mathrm{N} = \frac{1317.6}{15} = 87.84,$$

αφού $N = n_1 + n_2 + n_3 = 15$. Επομένως,

$$SST = \sum_{i-1}^{3}\sum_{j=1}^{5}(y_{ij} - \bar{y}_{..})^2 = (82.5 - 87.84)^2 + \cdots + (100.7 - 87.84)^2 = 1264.456.$$

Στη συνέχεια θέλοντας να υπολογίσουμε το συνολικό άθροισμα τετραγώνων μεταξύ των ομάδων, υπολογίζουμε τους δειγματικούς μέσους κάθε ομάδας ασθενών. Είναι:

$$\bar{y}_{1\cdot} = \frac{\sum_{j=1}^{5} y_{1j}}{n_1} = \frac{413.1}{5} = 82.62,$$

$$\bar{y}_{2\cdot} = \frac{\sum_{j=1}^{5} y_{2j}}{n_2} = \frac{405.7}{5} = 81.14,$$

και

$$\bar{y}_{3\cdot} = \frac{\sum_{j=1}^{5} y_{3j}}{n_3} = \frac{498.8}{5} = 99.76.$$

Επομένως,

$$SSB = \sum_{i=1}^{3} n_i(\bar{y}_{i\cdot} - \bar{y}_{..})^2 = 5 \cdot (82.62 - 87.84)^2 + 5 \cdot (81.14 - 87.84)^2$$

$$+ 5 \cdot (99.76 - 87.84)^2 = 1071.124.$$

Άρα είναι:

$$SSW = SST - SSB = 1264.456 - 1071.124 = 193.332.$$

Με βάση τα παραπάνω προκύπτει ο ακόλουθος πίνακας ανάλυσης διασποράς

Μεταβλητότητα	Βαθμοί ελευθερίας	Άθροισμα τετραγώνων	Μέσο Άθροισμα τετραγώνων	F
Μεταξύ των ομάδων (SSB)	2	1071.124	535.562	33.24201
Εντός των ομάδων (SSW)	12	193.332	16.111	
Ολική (SST)	14	1264.456		

Από τον πίνακα της κατανομής F για επίπεδο σημαντικότητας 0.05 έχουμε ότι $F_{2,12,0.05} = 3.89$. Καθώς $f = 33.24201 > 3.89 = F_{2,12,0.05}$, απορρίπτεται σε επίπεδο σημαντικότητας 0.05 η μηδενική υπόθεση. Με βάση αυτά τα δεδομένα μπορούμε να ισχυριστούμε ότι τουλάχιστον ενός τύπου κραμάτων ο μέσος συντελεστής ελαστικότητας διαφέρει στατιστικά σημαντικά από τον μέσο συντελεστή ελαστικότητας των κραμάτων των υπόλοιπων τύπων κραμάτων.
Εναλλακτικά, θα μπορούσαμε να λύσουμε την άσκηση χρησιμοποιώντας τις ακόλουθες εντολές της R:

```
d2 <- data.frame(
  elastic=c(82.5,83.7,80.9,85.2,80.8,
    82.7,81.9,78.9,83.6,78.6,
    92.2,106.8,104.6,94.5,100.7),
  group=c( rep("M1",5), rep("M2",5),rep("M3",5) )
)

res.aov2 <- aov(formula =  elastic~group,data = d2)
summary(res.aov2)
```

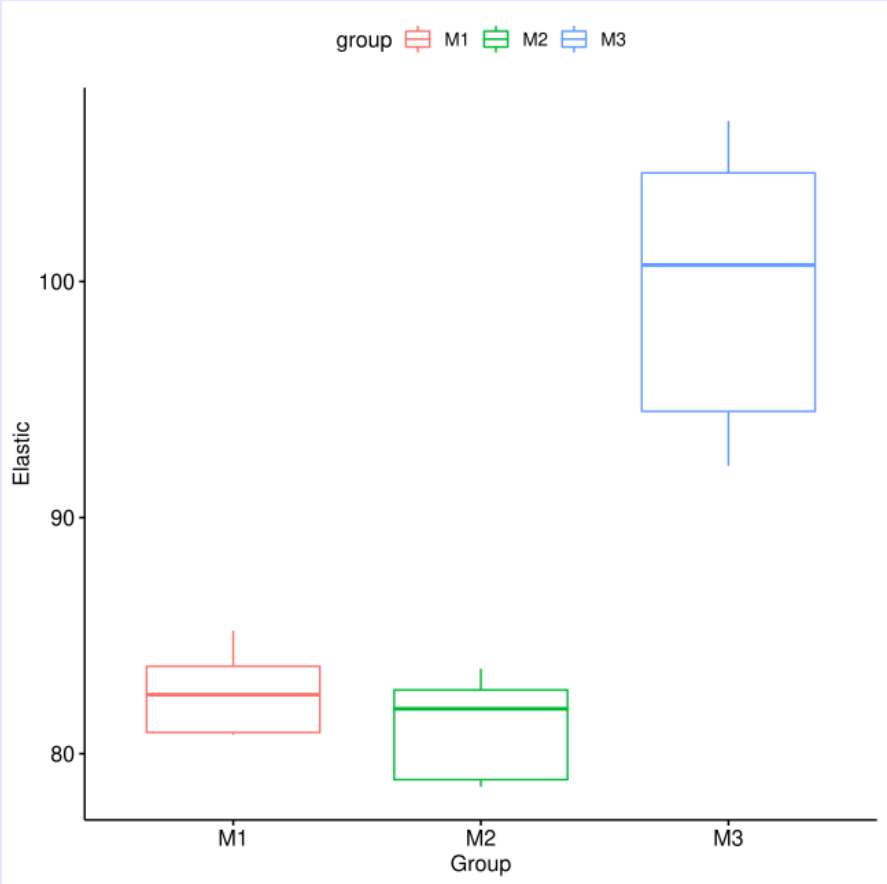
Τότε λαμβάνουμε τα παρακάτω αποτελέσματα

```
            Df Sum Sq Mean Sq F value   Pr(>F)
group        2 1071.1   535.6   33.24 1.28e-05 ***
Residuals   12  193.3    16.1
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Παρατηρούμε ότι $p - value = 1.28 \cdot 10^{-5} < 0.05$, άρα απορρίπτεται η μηδενική υπόθεση, όποτε υπάρχει στατιστικά σημαντική διαφορά σε επίπεδο σημαντικότητας 0.05 στον μέσο συντελεστή ελαστικότητας των τριών τύπων κραμάτων.
Κάποιος θα μπορούσε να σχηματίσει μία πρώτη άποψη για τη διαφορά του συντελεστή ελαστικότητας των τριών τύπων κραμάτων κατασκευάζοντας τα θηκογράμματά τους, χρησιμοποιώντας τις ακόλουθες εντολές της R:

```
ggboxplot(d2, x = "group", y = "elastic",
          color = "group",
          ylab = "Elastic", xlab = "Group")
```

Από τα θηκογράμματα του παρακάτω σχήματος μπορούμε να παρατηρήσουμε ότι ο συντελεστής ελαστικότητας της τρίτης ομάδας (μπλε χρώμα) μοιάζει να είναι υψηλότερος από αυτούς των άλλων δύο ομάδων.



Λύση Άσκησης Αυτοαξιολόγησης 13.2

Για να κάνουμε όλες τις δυνατές ζευγαρωτές συγκρίσεις μεταξύ των τριών τύπων κραμάτων χρησιμοποιώντας τις μεθόδους Bonferroni, Bonferroni-Holm και Tukey χρησιμοποιούμε τις παρακάτω εντολές της R :

```
1 pairwise.t.test(d2$elastic , d2$group , p.adj = "bonf")
2
3 pairwise.t.test(d2$elastic , d2$group , p.adj = "holm")
4
5 TukeyHSD(res.aov2)
```

Τα αποτελέσματα που προκύπτουν είναι τα ακόλουθα:

```
        Pairwise comparisons using t tests with pooled SD

data:  d2$elastic and d2$group

   M1      M2
M2 1       -
M3 6.1e-05 2.7e-05

P value adjustment method: bonferroni

Pairwise comparisons using t tests with pooled SD
```

```
data:  d2$elastic and d2$group

   M1      M2
M2 0.57    -
M3 4.1e-05 2.7e-05

P value adjustment method: holm

 Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = elastic ~ group, data = d2)

$group
       diff       lwr       upr     p adj
M2-M1 -1.48 -8.252591  5.292591 0.8316209
M3-M1 17.14 10.367409 23.912591 0.0000561
M3-M2 18.62 11.847409 25.392591 0.0000250
```

Παρατηρώντας τις $p-$τιμές και από τις τρεις μεθόδους καταλήγουμε ακριβώς στα ίδια συμπεράσματα: Ο μέσος συντελεστής ελαστικότητας του τρίτου τύπου κράματος διαφέρει στατιστικά σημαντικά από τον αντίστοιχο του δεύτερου και πρώτου τύπου κράματος, ενώ οι μέσοι συντελεστές ελαστικότητας του πρώτου και δεύτερου τύπου κράματος δεν διαφέρουν στατιστικά σημαντικά.

ΒΙΒΛΙΟΓΡΑΦΙΑ

Ελληνόγλωσση

Καρακώστας, Μ. Ξ. (2002). *Γραμμικά Μοντέλα. Παλινδρόμηση-Ανάλυση Διακύμανσης*. Πανεπιστήμιο Ιωαννίνων.

Παπαϊωάννου, Τ. και Λουκάς, Σ. (2002). *Εισαγωγή στη Στατιστική*. Εκδόσεις Σταμούλη, Αθήνα.

Φουσκάκης, Δ. (2013). *Ανάλυση Δεδομένων με χρήση της R*. Αθήνα: Εκδόσεις Τσότρας.

Ξενόγλωσση

Bartlett, M. S. (1937). Properties of sufficiency and statistical tests. *Proceedings of the Royal Statistical Society, Series A*, *160*, pp. 268–282.

Fisher, R. A. (1918). The Correlation Between Relatives on the Supposition of Mendelian Inheritance. *Philosophical Transactions of the Royal Society of Edinburgh*, *52*, pp. 399–433.

Galambos, J. and Simonelli, I. (1996). *Bonferroni-Type Inequalities with Applications*. Springer-Verlag.

Holm, S. (1979). A simple sequential rejective multiple test procedure. *Scandinavian Journal of Statistics*, *6*, pp. 65–70.

Kruskal, W. and Wallis, W. (1952). Use of ranks on one-criterion variance analysis. *J. Amer. Statist. Assoc.*, *47*, pp. 583–621.

Levene, H. (1960). Robust Tests for Equality of Variances. In: *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*. Palo Alto: Stanford University Press, pp. 278–292.

Rencher, A. C. and Schaalje, G. B. (2008). *Linear Models in Statistics*. Wiley, New Jersey.

Tukey, J. (1977). *Addison-Wesley Publishing Company Reading, Mass.-Menlo Park, Cal., hndon,Amsterdam, Don Mills, Ontario, Sydney*. Addison-Wesley Publishing Company.

Μέρος ΙΙΙ

ΠΑΡΑΡΤΗΜΑΤΑ

Οικονόμου, Π., Μαλεφάκη, Σ., & Μπατσίδης, Α. (2023). *Πιθανότητες - Στατιστική.*
[Προπτυχιακό εγχειρίδιο].

ΠΑΡΑΡΤΗΜΑ Α΄

ΠΙΝΑΚΕΣ ΚΑΤΑΝΟΜΩΝ

Σύνοψη

Σε αυτό το παράρτημα, δίνονται πίνακες που αφορούν

- τη διωνυμική κατανομή,
- την κατανομή Poisson,
- την τυπική κανονική κατανομή $N(0,1)$,
- την κατανομή t (Student),
- τη χ^2 κατανομή και
- την κατανομή F.

Προαπαιτούμενη γνώση: Να γνωρίζετε και να έχετε κατανοήσει την έννοια της αθροιστικής συνάρτησης κατανομής και του ποσοστιαίου σημείου.

Προσδοκώμενα μαθησιακά αποτελέσματα: Χρησιμοποιώντας το παράρτημα θα μπορείτε να υπολογίζετε πιθανότητες που αφορούν τυχαίες μεταβλητές που ακολουθούν διωνυμική, Poisson ή την κανονική κατανομή, καθώς και να προσδιορίζετε κρίσιμα σημεία με τη χι-τετράγωνο, την κατανομή t (Student) και την κατανομή F.

ΠΙΝΑΚΑΣ ΔΙΩΝΥΜΙΚΗΣ ΣΥΝΑΡΤΗΣΗΣ ΚΑΤΑΝΟΜΗΣ

Πίνακας Α΄.1: Αν $X \sim B(n,p)$, ο πίνακας δίνει την πιθανότητα $P(X \leqslant x)$.

n	x	$\binom{n}{x}$	0.01	0.05	0.10	0.15	0.20	0.25	0.30	1/3	0.35	0.40	0.45	0.49	0.50
5	0	1	0.9510	0.7738	0.5905	0.4437	0.3277	0.2373	0.1681	0.1317	0.1160	0.0778	0.0503	0.0345	0.0312
	1	5	0.9990	0.9774	0.9185	0.8352	0.7373	0.6328	0.5282	0.4609	0.4284	0.3370	0.2562	0.2002	0.1875
	2	10	1.0000	0.9988	0.9914	0.9734	0.9421	0.8965	0.8369	0.7901	0.7648	0.6826	0.5931	0.5187	0.5000
	3	10	1.0000	1.0000	0.9995	0.9978	0.9933	0.9844	0.9692	0.9547	0.9460	0.9130	0.8688	0.8248	0.8125
	4	5	1.0000	1.0000	1.0000	0.9999	0.9997	0.9990	0.9976	0.9959	0.9947	0.9898	0.9815	0.9718	0.9688
	5	1	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
10	0	1	0.9044	0.5987	0.3487	0.1969	0.1074	0.0563	0.0282	0.0173	0.0135	0.0060	0.0025	0.0012	0.0010
	1	10	0.9957	0.9139	0.7361	0.5443	0.3758	0.2440	0.1493	0.1040	0.0860	0.0464	0.0233	0.0126	0.0107
	2	45	0.9999	0.9885	0.9298	0.8202	0.6778	0.5256	0.3828	0.2991	0.2616	0.1673	0.0996	0.0621	0.0547
	3	120	1.0000	0.9990	0.9872	0.9500	0.8791	0.7759	0.6496	0.5593	0.5138	0.3823	0.2660	0.1888	0.1719
	4	210	1.0000	0.9999	0.9984	0.9901	0.9672	0.9219	0.8497	0.7869	0.7515	0.6331	0.5044	0.4018	0.3770
	5	252	1.0000	1.0000	0.9999	0.9986	0.9936	0.9803	0.9527	0.9234	0.9051	0.8338	0.7384	0.6474	0.6230
	6	210	1.0000	1.0000	1.0000	0.9999	0.9991	0.9965	0.9894	0.9803	0.9740	0.9452	0.8980	0.8440	0.8281
	7	120	1.0000	1.0000	1.0000	1.0000	0.9999	0.9996	0.9984	0.9966	0.9952	0.9877	0.9726	0.9520	0.9453
	8	45	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9996	0.9995	0.9983	0.9955	0.9909	0.9893
	9	10	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9997	0.9992	0.9990
	10	1	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
15	0	1	0.8601	0.4633	0.2059	0.0874	0.0352	0.0134	0.0047	0.0023	0.0016	0.0005	0.0001	0.0000	0.0000
	1	15	0.9904	0.8290	0.5490	0.3186	0.1671	0.0802	0.0353	0.0194	0.0142	0.0052	0.0017	0.0006	0.0005
	2	105	0.9996	0.9638	0.8159	0.6042	0.3980	0.2361	0.1268	0.0794	0.0617	0.0271	0.0107	0.0046	0.0037
	3	455	1.0000	0.9945	0.9444	0.8227	0.6482	0.4613	0.2969	0.2092	0.1727	0.0905	0.0424	0.0212	0.0176
	4	1365	1.0000	0.9994	0.9873	0.9383	0.8358	0.6865	0.5155	0.4041	0.3519	0.2173	0.1204	0.0690	0.0592
	5	3003	1.0000	0.9999	0.9978	0.9832	0.9389	0.8516	0.7216	0.6184	0.5643	0.4032	0.2608	0.1699	0.1509
	6	5005	1.0000	1.0000	0.9997	0.9964	0.9819	0.9434	0.8689	0.7970	0.7548	0.6098	0.4522	0.3316	0.3036
	7	6435	1.0000	1.0000	1.0000	0.9994	0.9958	0.9827	0.9500	0.9118	0.8868	0.7869	0.6535	0.5314	0.5000
	8	6435	1.0000	1.0000	1.0000	0.9999	0.9992	0.9958	0.9848	0.9692	0.9578	0.9050	0.8182	0.7233	0.6964
	9	5005	1.0000	1.0000	1.0000	1.0000	0.9999	0.9992	0.9963	0.9915	0.9876	0.9662	0.9231	0.8667	0.8491
	10	3003	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9993	0.9982	0.9972	0.9907	0.9745	0.9494	0.9408
	11	1365	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9997	0.9995	0.9981	0.9937	0.9855	0.9824
	12	455	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9997	0.9989	0.9971	0.9963
	13	105	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9996	0.9995
	14	15	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	15	1	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
20	0	1	0.8179	0.3585	0.1216	0.0388	0.0115	0.0032	0.0008	0.0003	0.0002	0.0000	0.0000	0.0000	0.0000
	1	20	0.9831	0.7358	0.3917	0.1756	0.0692	0.0243	0.0076	0.0033	0.0021	0.0005	0.0001	0.0000	0.0000
	2	190	0.9990	0.9245	0.6769	0.4049	0.2061	0.0913	0.0355	0.0176	0.0121	0.0036	0.0009	0.0003	0.0002
	3	1140	1.0000	0.9841	0.8670	0.6477	0.4114	0.2252	0.1071	0.0604	0.0444	0.0160	0.0049	0.0017	0.0013
	4	4845	1.0000	0.9974	0.9568	0.8298	0.6296	0.4148	0.2375	0.1515	0.1182	0.0510	0.0189	0.0076	0.0059
	5	15504	1.0000	0.9997	0.9887	0.9327	0.8042	0.6172	0.4164	0.2972	0.2454	0.1256	0.0553	0.0255	0.0207
	6	38760	1.0000	1.0000	0.9976	0.9781	0.9133	0.7858	0.6080	0.4793	0.4166	0.2500	0.1299	0.0688	0.0577
	7	77520	1.0000	1.0000	0.9996	0.9941	0.9679	0.8982	0.7723	0.6615	0.6010	0.4159	0.2520	0.1518	0.1316
	8	125970	1.0000	1.0000	0.9999	0.9987	0.9900	0.9591	0.8867	0.8095	0.7624	0.5956	0.4143	0.2814	0.2517
	9	167960	1.0000	1.0000	1.0000	0.9998	0.9974	0.9861	0.9520	0.9081	0.8782	0.7553	0.5914	0.4475	0.4119
	10	184756	1.0000	1.0000	1.0000	1.0000	0.9994	0.9961	0.9829	0.9624	0.9468	0.8725	0.7507	0.6229	0.5881
	11	167960	1.0000	1.0000	1.0000	1.0000	0.9999	0.9991	0.9949	0.9870	0.9804	0.9435	0.8692	0.7762	0.7483
	12	125970	1.0000	1.0000	1.0000	1.0000	1.0000	0.9998	0.9987	0.9963	0.9940	0.9790	0.9420	0.8867	0.8684
	13	77520	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9997	0.9991	0.9985	0.9935	0.9786	0.9520	0.9423
	14	38760	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9998	0.9997	0.9984	0.9936	0.9834	0.9793
	15	15504	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9997	0.9985	0.9954	0.9941
	16	4845	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9997	0.9990	0.9987
	17	1140	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9998
	18	190	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	19	20	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	20	1	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

ΠΙΝΑΚΑΣ ΣΥΝΑΡΤΗΣΗΣ ΚΑΤΑΝΟΜΗΣ POISSON

Πίνακας Α΄.2: Αν $X \sim \mathscr{P}(\lambda)$, ο πίνακας δίνει την πιθανότητα $P(X \leqslant x)$.

					λ					
x	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
0	0.9048	0.8187	0.7408	0.6703	0.6065	0.5488	0.4966	0.4493	0.4066	0.3679
1	0.9953	0.9825	0.9631	0.9384	0.9098	0.8781	0.8442	0.8088	0.7725	0.7358
2	0.9998	0.9989	0.9964	0.9921	0.9856	0.9769	0.9659	0.9526	0.9371	0.9197
3	1.0000	0.9999	0.9997	0.9992	0.9982	0.9966	0.9942	0.9909	0.9865	0.9810
4	1.0000	1.0000	1.0000	0.9999	0.9998	0.9996	0.9992	0.9986	0.9977	0.9963
5	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9998	0.9997	0.9994
6	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999
7	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

					λ					
x	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2.0
0	0.3329	0.3012	0.2725	0.2466	0.2231	0.2019	0.1827	0.1653	0.1496	0.1353
1	0.6990	0.6626	0.6268	0.5918	0.5578	0.5249	0.4932	0.4628	0.4337	0.4060
2	0.9004	0.8795	0.8571	0.8335	0.8088	0.7834	0.7572	0.7306	0.7037	0.6767
3	0.9743	0.9662	0.9569	0.9463	0.9344	0.9212	0.9068	0.8913	0.8747	0.8571
4	0.9946	0.9923	0.9893	0.9857	0.9814	0.9763	0.9704	0.9636	0.9559	0.9473
5	0.9990	0.9985	0.9978	0.9968	0.9955	0.9940	0.9920	0.9896	0.9868	0.9834
6	0.9999	0.9997	0.9996	0.9994	0.9991	0.9987	0.9981	0.9974	0.9966	0.9955
7	1.0000	1.0000	0.9999	0.9999	0.9998	0.9997	0.9996	0.9994	0.9992	0.9989
8	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9999	0.9998	0.9998
9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

					λ					
x	2.2	2.4	2.6	2.8	3.0	3.4	3.8	4.0	4.5	5.0
0	0.1108	0.0907	0.0743	0.0608	0.0498	0.0334	0.0224	0.0183	0.0111	0.0067
1	0.3546	0.3084	0.2674	0.2311	0.1991	0.1468	0.1074	0.0916	0.0611	0.0404
2	0.6227	0.5697	0.5184	0.4695	0.4232	0.3397	0.2689	0.2381	0.1736	0.1247
3	0.8194	0.7787	0.7360	0.6919	0.6472	0.5584	0.4735	0.4335	0.3423	0.2650
4	0.9275	0.9041	0.8774	0.8477	0.8153	0.7442	0.6678	0.6288	0.5321	0.4405
5	0.9751	0.9643	0.9510	0.9349	0.9161	0.8705	0.8156	0.7851	0.7029	0.6160
6	0.9925	0.9884	0.9828	0.9756	0.9665	0.9421	0.9091	0.8893	0.8311	0.7622
7	0.9980	0.9967	0.9947	0.9919	0.9881	0.9769	0.9599	0.9489	0.9134	0.8666
8	0.9995	0.9991	0.9985	0.9976	0.9962	0.9917	0.9840	0.9786	0.9597	0.9319
9	0.9999	0.9998	0.9996	0.9993	0.9989	0.9973	0.9942	0.9919	0.9829	0.9682
10	1.0000	1.0000	0.9999	0.9998	0.9997	0.9992	0.9981	0.9972	0.9933	0.9863
11	1.0000	1.0000	1.0000	1.0000	0.9999	0.9998	0.9994	0.9991	0.9976	0.9945
12	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9998	0.9997	0.9992	0.9980
13	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9997	0.9993
14	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9998

ΠΙΝΑΚΑΣ ΤΥΠΙΚΗΣ ΚΑΝΟΝΙΚΗΣ ΚΑΤΑΝΟΜΗΣ - $N(0,1)$.

Πίνακας Α΄.3: Αν $Z \sim N(0,1)$, ο πίνακας δίνει την πιθανότητα $P(Z \leqslant z)$.

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	0.50000	0.50399	0.50798	0.51197	0.51595	0.51994	0.52392	0.52790	0.53188	0.53586
0.1	0.53983	0.54380	0.54776	0.55172	0.55567	0.55962	0.56356	0.56749	0.57142	0.57535
0.2	0.57926	0.58317	0.58706	0.59095	0.59483	0.59871	0.60257	0.60642	0.61026	0.61409
0.3	0.61791	0.62172	0.62552	0.62930	0.63307	0.63683	0.64058	0.64431	0.64803	0.65173
0.4	0.65542	0.65910	0.66276	0.66640	0.67003	0.67364	0.67724	0.68082	0.68439	0.68793
0.5	0.69146	0.69497	0.69847	0.70194	0.70540	0.70884	0.71226	0.71566	0.71904	0.72240
0.6	0.72575	0.72907	0.73237	0.73565	0.73891	0.74215	0.74537	0.74857	0.75175	0.75490
0.7	0.75804	0.76115	0.76424	0.76730	0.77035	0.77337	0.77637	0.77935	0.78230	0.78524
0.8	0.78814	0.79103	0.79389	0.79673	0.79955	0.80234	0.80511	0.80785	0.81057	0.81327
0.9	0.81594	0.81859	0.82121	0.82381	0.82639	0.82894	0.83147	0.83398	0.83646	0.83891
1.0	0.84134	0.84375	0.84614	0.84849	0.85083	0.85314	0.85543	0.85769	0.85993	0.86214
1.1	0.86433	0.86650	0.86864	0.87076	0.87286	0.87493	0.87698	0.87900	0.88100	0.88298
1.2	0.88493	0.88686	0.88877	0.89065	0.89251	0.89435	0.89617	0.89796	0.89973	0.90147
1.3	0.90320	0.90490	0.90658	0.90824	0.90988	0.91149	0.91309	0.91466	0.91621	0.91774
1.4	0.91924	0.92073	0.92220	0.92364	0.92507	0.92647	0.92785	0.92922	0.93056	0.93189
1.5	0.93319	0.93448	0.93574	0.93699	0.93822	0.93943	0.94062	0.94179	0.94295	0.94408
1.6	0.94520	0.94630	0.94738	0.94845	0.94950	0.95053	0.95154	0.95254	0.95352	0.95449
1.7	0.95543	0.95637	0.95728	0.95818	0.95907	0.95994	0.96080	0.96164	0.96246	0.96327
1.8	0.96407	0.96485	0.96562	0.96638	0.96712	0.96784	0.96856	0.96926	0.96995	0.97062
1.9	0.97128	0.97193	0.97257	0.97320	0.97381	0.97441	0.97500	0.97558	0.97615	0.97670
2.0	0.97725	0.97778	0.97831	0.97882	0.97932	0.97982	0.98030	0.98077	0.98124	0.98169
2.1	0.98214	0.98257	0.98300	0.98341	0.98382	0.98422	0.98461	0.98500	0.98537	0.98574
2.2	0.98610	0.98645	0.98679	0.98713	0.98745	0.98778	0.98809	0.98840	0.98870	0.98899
2.3	0.98928	0.98956	0.98983	0.99010	0.99036	0.99061	0.99086	0.99111	0.99134	0.99158
2.4	0.99180	0.99202	0.99224	0.99245	0.99266	0.99286	0.99305	0.99324	0.99343	0.99361
2.5	0.99379	0.99396	0.99413	0.99430	0.99446	0.99461	0.99477	0.99492	0.99506	0.99520
2.6	0.99534	0.99547	0.99560	0.99573	0.99585	0.99598	0.99609	0.99621	0.99632	0.99643
2.7	0.99653	0.99664	0.99674	0.99683	0.99693	0.99702	0.99711	0.99720	0.99728	0.99736
2.8	0.99744	0.99752	0.99760	0.99767	0.99774	0.99781	0.99788	0.99795	0.99801	0.99807
2.9	0.99813	0.99819	0.99825	0.99831	0.99836	0.99841	0.99846	0.99851	0.99856	0.99861
3.0	0.99865	0.99869	0.99874	0.99878	0.99882	0.99886	0.99889	0.99893	0.99896	0.99900
3.1	0.99903	0.99906	0.99910	0.99913	0.99916	0.99918	0.99921	0.99924	0.99926	0.99929
3.2	0.99931	0.99934	0.99936	0.99938	0.99940	0.99942	0.99944	0.99946	0.99948	0.99950
3.3	0.99952	0.99953	0.99955	0.99957	0.99958	0.99960	0.99961	0.99962	0.99964	0.99965
3.4	0.99966	0.99968	0.99969	0.99970	0.99971	0.99972	0.99973	0.99974	0.99975	0.99976
3.5	0.99977	0.99978	0.99978	0.99979	0.99980	0.99981	0.99981	0.99982	0.99983	0.99983
3.6	0.99984	0.99985	0.99985	0.99986	0.99986	0.99987	0.99987	0.99988	0.99988	0.99989
3.7	0.99989	0.99990	0.99990	0.99990	0.99991	0.99991	0.99992	0.99992	0.99992	0.99992
3.8	0.99993	0.99993	0.99993	0.99994	0.99994	0.99994	0.99994	0.99995	0.99995	0.99995
3.9	0.99995	0.99995	0.99996	0.99996	0.99996	0.99996	0.99996	0.99996	0.99997	0.99997

ΠΙΝΑΚΑΣ t (Student)

Πίνακας Α΄.4: Αν $X \sim t_\nu$, ο πίνακας δίνει τα σημεία $t_{\nu,p}$, που είναι τέτοια ώστε $P(X \geqslant t_{\nu,p}) = p$.

	p												
ν	.4	.3	.25	.2	.15	.1	.075	.05	.025	.010	.005	.001	.0005
1	0.325	0.727	1.000	1.376	1.963	3.078	4.165	6.314	12.706	31.821	63.657	318.310	636.620
2	0.289	0.617	0.816	1.061	1.386	1.886	2.282	2.920	4.303	6.965	9.925	22.327	31.599
3	0.277	0.584	0.765	0.978	1.250	1.638	1.924	2.353	3.182	4.541	5.841	10.215	12.924
4	0.271	0.569	0.741	0.941	1.190	1.533	1.778	2.132	2.776	3.747	4.604	7.173	8.610
5	0.267	0.559	0.727	0.920	1.156	1.476	1.699	2.015	2.571	3.365	4.032	5.893	6.869
6	0.265	0.553	0.718	0.906	1.134	1.440	1.650	1.943	2.447	3.143	3.707	5.208	5.959
7	0.263	0.549	0.711	0.896	1.119	1.415	1.617	1.895	2.365	2.998	3.499	4.785	5.408
8	0.262	0.546	0.706	0.889	1.108	1.397	1.592	1.860	2.306	2.896	3.355	4.501	5.041
9	0.261	0.543	0.703	0.883	1.100	1.383	1.574	1.833	2.262	2.821	3.250	4.297	4.781
10	0.260	0.542	0.700	0.879	1.093	1.372	1.559	1.812	2.228	2.764	3.169	4.144	4.587
11	0.260	0.540	0.697	0.876	1.088	1.363	1.548	1.796	2.201	2.718	3.106	4.025	4.437
12	0.259	0.539	0.695	0.873	1.083	1.356	1.538	1.782	2.179	2.681	3.055	3.930	4.318
13	0.259	0.538	0.694	0.870	1.079	1.350	1.530	1.771	2.160	2.650	3.012	3.852	4.221
14	0.258	0.537	0.692	0.868	1.076	1.345	1.523	1.761	2.145	2.624	2.977	3.787	4.140
15	0.258	0.536	0.691	0.866	1.074	1.341	1.517	1.753	2.131	2.602	2.947	3.733	4.073
16	0.258	0.535	0.690	0.865	1.071	1.337	1.512	1.746	2.120	2.583	2.921	3.686	4.015
17	0.257	0.534	0.689	0.863	1.069	1.333	1.508	1.740	2.110	2.567	2.898	3.646	3.965
18	0.257	0.534	0.688	0.862	1.067	1.330	1.504	1.734	2.101	2.552	2.878	3.610	3.922
19	0.257	0.533	0.688	0.861	1.066	1.328	1.500	1.729	2.093	2.539	2.861	3.579	3.883
20	0.257	0.533	0.687	0.860	1.064	1.325	1.497	1.725	2.086	2.528	2.845	3.552	3.850
21	0.257	0.532	0.686	0.859	1.063	1.323	1.494	1.721	2.080	2.518	2.831	3.527	3.819
22	0.256	0.532	0.686	0.858	1.061	1.321	1.492	1.717	2.074	2.508	2.819	3.505	3.792
23	0.256	0.532	0.685	0.858	1.060	1.319	1.489	1.714	2.069	2.500	2.807	3.485	3.768
24	0.256	0.531	0.685	0.857	1.059	1.318	1.487	1.711	2.064	2.492	2.797	3.467	3.745
25	0.256	0.531	0.684	0.856	1.058	1.316	1.485	1.708	2.060	2.485	2.787	3.450	3.725
26	0.256	0.531	0.684	0.856	1.058	1.315	1.483	1.706	2.056	2.479	2.779	3.435	3.707
27	0.256	0.531	0.684	0.855	1.057	1.314	1.482	1.703	2.052	2.473	2.771	3.421	3.690
28	0.256	0.530	0.683	0.855	1.056	1.313	1.480	1.701	2.048	2.467	2.763	3.408	3.674
29	0.256	0.530	0.683	0.854	1.055	1.311	1.479	1.699	2.045	2.462	2.756	3.396	3.659
30	0.256	0.530	0.683	0.854	1.055	1.310	1.477	1.697	2.042	2.457	2.750	3.385	3.646
31	0.256	0.530	0.682	0.853	1.054	1.309	1.476	1.696	2.040	2.453	2.744	3.375	3.633
32	0.255	0.530	0.682	0.853	1.054	1.309	1.475	1.694	2.037	2.449	2.738	3.365	3.622
33	0.255	0.530	0.682	0.853	1.053	1.308	1.474	1.692	2.035	2.445	2.733	3.356	3.611
34	0.255	0.529	0.682	0.852	1.052	1.307	1.473	1.691	2.032	2.441	2.728	3.348	3.601
35	0.255	0.529	0.682	0.852	1.052	1.306	1.472	1.690	2.030	2.438	2.724	3.340	3.591
36	0.255	0.529	0.681	0.852	1.052	1.306	1.471	1.688	2.028	2.434	2.719	3.333	3.582
37	0.255	0.529	0.681	0.851	1.051	1.305	1.470	1.687	2.026	2.431	2.715	3.326	3.574
38	0.255	0.529	0.681	0.851	1.051	1.304	1.469	1.686	2.024	2.429	2.712	3.319	3.566
39	0.255	0.529	0.681	0.851	1.050	1.304	1.468	1.685	2.023	2.426	2.708	3.313	3.558
40	0.255	0.529	0.681	0.851	1.050	1.303	1.468	1.684	2.021	2.423	2.704	3.307	3.551
45	0.255	0.528	0.680	0.850	1.049	1.301	1.465	1.679	2.014	2.412	2.690	3.281	3.520
50	0.255	0.528	0.679	0.849	1.047	1.299	1.462	1.676	2.009	2.403	2.678	3.261	3.496
60	0.254	0.527	0.679	0.848	1.045	1.296	1.458	1.671	2.000	2.390	2.660	3.232	3.460
70	0.254	0.527	0.678	0.847	1.044	1.294	1.456	1.667	1.994	2.381	2.648	3.211	3.435
80	0.254	0.526	0.678	0.846	1.043	1.292	1.453	1.664	1.990	2.374	2.639	3.195	3.416
90	0.254	0.526	0.677	0.846	1.042	1.291	1.452	1.662	1.987	2.368	2.632	3.183	3.402
100	0.254	0.526	0.677	0.845	1.042	1.290	1.451	1.660	1.984	2.364	2.626	3.174	3.390
120	0.254	0.526	0.677	0.845	1.041	1.289	1.449	1.658	1.980	2.358	2.617	3.160	3.373
∞	0.253	0.524	0.674	0.842	1.036	1.282	1.440	1.645	1.960	2.326	2.576	3.090	3.291

ΠΙΝΑΚΑΣ χ^2

Πίνακας Α΄.5: Αν $X \sim \chi^2_\nu$, ο πίνακας δίνει τα σημεία $\chi^2_{\nu,p}$, που είναι τέτοια ώστε $P(X \geqslant \chi^2_{\nu,p}) = p$.

	p											
ν	.9995	.999	.995	.99	.975	.95	.925	.9	.85	.8	.7	0.6
1	$.0^6393$	$.0^5157$	$.0^4393$	$.0^3157$	0.001	0.004	0.009	0.016	0.036	0.064	0.148	0.275
2	0.001	0.002	0.010	0.020	0.051	0.103	0.156	0.211	0.325	0.446	0.713	1.022
3	0.015	0.024	0.072	0.115	0.216	0.352	0.472	0.584	0.798	1.005	1.424	1.869
4	0.064	0.091	0.207	0.297	0.484	0.711	0.897	1.064	1.366	1.649	2.195	2.753
5	0.158	0.210	0.412	0.554	0.831	1.145	1.394	1.610	1.994	2.343	3.000	3.655
6	0.299	0.381	0.676	0.872	1.237	1.635	1.941	2.204	2.661	3.070	3.828	4.570
7	0.485	0.598	0.989	1.239	1.690	2.167	2.528	2.833	3.358	3.822	4.671	5.493
8	0.710	0.857	1.344	1.646	2.180	2.733	3.144	3.490	4.078	4.594	5.527	6.423
9	0.972	1.152	1.735	2.088	2.700	3.325	3.785	4.168	4.817	5.380	6.393	7.357
10	1.265	1.479	2.156	2.558	3.247	3.940	4.446	4.865	5.570	6.179	7.267	8.295
11	1.587	1.834	2.603	3.053	3.816	4.575	5.124	5.578	6.336	6.989	8.148	9.237
12	1.934	2.214	3.074	3.571	4.404	5.226	5.818	6.304	7.114	7.807	9.034	10.182
13	2.305	2.617	3.565	4.107	5.009	5.892	6.524	7.042	7.901	8.634	9.926	11.129
14	2.697	3.041	4.075	4.660	5.629	6.571	7.242	7.790	8.696	9.467	10.821	12.078
15	3.108	3.483	4.601	5.229	6.262	7.261	7.969	8.547	9.499	10.307	11.721	13.030
16	3.536	3.942	5.142	5.812	6.908	7.962	8.707	9.312	10.309	11.152	12.624	13.983
17	3.980	4.416	5.697	6.408	7.564	8.672	9.452	10.085	11.125	12.002	13.531	14.937
18	4.439	4.905	6.265	7.015	8.231	9.390	10.205	10.865	11.946	12.857	14.440	15.893
19	4.912	5.407	6.844	7.633	8.907	10.117	10.965	11.651	12.773	13.716	15.352	16.850
20	5.398	5.921	7.434	8.260	9.591	10.851	11.732	12.443	13.604	14.578	16.266	17.809
21	5.896	6.447	8.034	8.897	10.283	11.591	12.504	13.240	14.439	15.445	17.182	18.768
22	6.404	6.983	8.643	9.542	10.982	12.338	13.282	14.041	15.279	16.314	18.101	19.729
23	6.924	7.529	9.260	10.196	11.689	13.091	14.065	14.848	16.122	17.187	19.021	20.690
24	7.453	8.085	9.886	10.856	12.401	13.848	14.853	15.659	16.969	18.062	19.943	21.652
25	7.991	8.649	10.520	11.524	13.120	14.611	15.645	16.473	17.818	18.940	20.867	22.616
26	8.538	9.222	11.160	12.198	13.844	15.379	16.441	17.292	18.671	19.820	21.792	23.579
27	9.093	9.803	11.808	12.879	14.573	16.151	17.241	18.114	19.527	20.703	22.719	24.544
28	9.656	10.391	12.461	13.565	15.308	16.928	18.045	18.939	20.386	21.588	23.647	25.509
29	10.227	10.986	13.121	14.256	16.047	17.708	18.853	19.768	21.247	22.475	24.577	26.475
30	10.804	11.588	13.787	14.953	16.791	18.493	19.664	20.599	22.110	23.364	25.508	27.442
31	11.389	12.196	14.458	15.655	17.539	19.281	20.478	21.434	22.976	24.255	26.440	28.409
32	11.979	12.811	15.134	16.362	18.291	20.072	21.295	22.271	23.844	25.148	27.373	29.376
33	12.576	13.431	15.815	17.074	19.047	20.867	22.115	23.110	24.714	26.042	28.307	30.344
34	13.179	14.057	16.501	17.789	19.806	21.664	22.938	23.952	25.586	26.938	29.242	31.313
35	13.787	14.688	17.192	18.509	20.569	22.465	23.763	24.797	26.460	27.836	30.178	32.282
36	14.401	15.324	17.887	19.233	21.336	23.269	24.591	25.643	27.336	28.735	31.115	33.252
37	15.020	15.965	18.586	19.960	22.106	24.075	25.421	26.492	28.214	29.635	32.053	34.222
38	15.644	16.611	19.289	20.691	22.878	24.884	26.254	27.343	29.093	30.537	32.992	35.192
39	16.273	17.262	19.996	21.426	23.654	25.695	27.089	28.196	29.974	31.441	33.932	36.163
40	16.906	17.916	20.707	22.164	24.433	26.509	27.926	29.051	30.856	32.345	34.872	37.134
45	20.137	21.251	24.311	25.901	28.366	30.612	32.140	33.350	35.290	36.884	39.585	41.995
50	23.461	24.674	27.991	29.707	32.357	34.764	36.397	37.689	39.754	41.449	44.313	46.864
60	30.340	31.738	35.534	37.485	40.482	43.188	45.016	46.459	48.759	50.641	53.809	56.620
70	37.467	39.036	43.275	45.442	48.758	51.739	53.748	55.329	57.844	59.898	63.346	66.396
80	44.791	46.520	51.172	53.540	57.153	60.391	62.568	64.278	66.994	69.207	72.915	76.188
90	52.276	54.155	59.196	61.754	65.647	69.126	71.460	73.291	76.195	78.558	82.511	85.993
100	59.896	61.918	67.328	70.065	74.222	77.929	80.412	82.358	85.441	87.945	92.129	95.808
120	75.467	77.755	83.852	86.923	91.573	95.705	98.464	100.624	104.037	106.806	111.419	115.465
150	99.463	102.113	109.142	112.668	117.985	122.692	125.827	128.275	132.137	135.263	140.457	145.000

ΠΙΝΑΚΑΣ χ^2

Πίνακας Α΄.6: Αν $X \sim \chi^2_\nu$, ο πίνακας δίνει τα σημεία $\chi^2_{\nu,p}$, που είναι τέτοια ώστε $P(X \geqslant \chi^2_{\nu,p}) = p$.

	p												
ν	.5	.4	.3	.2	.15	.1	.075	.05	.025	.01	.005	.001	.0005
1	0.455	0.708	1.074	1.642	2.072	2.706	3.170	3.841	5.024	6.635	7.879	10.828	12.116
2	1.386	1.833	2.408	3.219	3.794	4.605	5.181	5.991	7.378	9.210	10.597	13.816	15.202
3	2.366	2.946	3.665	4.642	5.317	6.251	6.905	7.815	9.348	11.345	12.838	16.266	17.730
4	3.357	4.045	4.878	5.989	6.745	7.779	8.496	9.488	11.143	13.277	14.860	18.467	19.997
5	4.351	5.132	6.064	7.289	8.115	9.236	10.008	11.070	12.833	15.086	16.750	20.515	22.105
6	5.348	6.211	7.231	8.558	9.446	10.645	11.466	12.592	14.449	16.812	18.548	22.458	24.103
7	6.346	7.283	8.383	9.803	10.748	12.017	12.883	14.067	16.013	18.475	20.278	24.322	26.018
8	7.344	8.351	9.524	11.030	12.027	13.362	14.270	15.507	17.535	20.090	21.955	26.124	27.868
9	8.343	9.414	10.656	12.242	13.288	14.684	15.631	16.919	19.023	21.666	23.589	27.877	29.666
10	9.342	10.473	11.781	13.442	14.534	15.987	16.971	18.307	20.483	23.209	25.188	29.588	31.420
11	10.341	11.530	12.899	14.631	15.767	17.275	18.294	19.675	21.920	24.725	26.757	31.264	33.137
12	11.340	12.584	14.011	15.812	16.989	18.549	19.602	21.026	23.337	26.217	28.300	32.909	34.821
13	12.340	13.636	15.119	16.985	18.202	19.812	20.897	22.362	24.736	27.688	29.819	34.528	36.478
14	13.339	14.685	16.222	18.151	19.406	21.064	22.180	23.685	26.119	29.141	31.319	36.123	38.109
15	14.339	15.733	17.322	19.311	20.603	22.307	23.452	24.996	27.488	30.578	32.801	37.697	39.719
16	15.338	16.780	18.418	20.465	21.793	23.542	24.716	26.296	28.845	32.000	34.267	39.252	41.308
17	16.338	17.824	19.511	21.615	22.977	24.769	25.970	27.587	30.191	33.409	35.718	40.790	42.879
18	17.338	18.868	20.601	22.760	24.155	25.989	27.218	28.869	31.526	34.805	37.156	42.312	44.434
19	18.338	19.910	21.689	23.900	25.329	27.204	28.458	30.144	32.852	36.191	38.582	43.820	45.973
20	19.337	20.951	22.775	25.038	26.498	28.412	29.692	31.410	34.170	37.566	39.997	45.315	47.498
21	20.337	21.991	23.858	26.171	27.662	29.615	30.920	32.671	35.479	38.932	41.401	46.797	49.011
22	21.337	23.031	24.939	27.301	28.822	30.813	32.142	33.924	36.781	40.289	42.796	48.268	50.511
23	22.337	24.069	26.018	28.429	29.979	32.007	33.360	35.172	38.076	41.638	44.181	49.728	52.000
24	23.337	25.106	27.096	29.553	31.132	33.196	34.572	36.415	39.364	42.980	45.559	51.179	53.479
25	24.337	26.143	28.172	30.675	32.282	34.382	35.780	37.652	40.646	44.314	46.928	52.620	54.947
26	25.336	27.179	29.246	31.795	33.429	35.563	36.984	38.885	41.923	45.642	48.290	54.052	56.407
27	26.336	28.214	30.319	32.912	34.574	36.741	38.184	40.113	43.195	46.963	49.645	55.476	57.858
28	27.336	29.249	31.391	34.027	35.715	37.916	39.380	41.337	44.461	48.278	50.993	56.892	59.300
29	28.336	30.283	32.461	35.139	36.854	39.087	40.573	42.557	45.722	49.588	52.336	58.301	60.735
30	29.336	31.316	33.530	36.250	37.990	40.256	41.762	43.773	46.979	50.892	53.672	59.703	62.162
31	30.336	32.349	34.598	37.359	39.124	41.422	42.948	44.985	48.232	52.191	55.003	61.098	63.582
32	31.336	33.381	35.665	38.466	40.256	42.585	44.131	46.194	49.480	53.486	56.328	62.487	64.995
33	32.336	34.413	36.731	39.572	41.386	43.745	45.311	47.400	50.725	54.776	57.648	63.870	66.403
34	33.336	35.444	37.795	40.676	42.514	44.903	46.488	48.602	51.966	56.061	58.964	65.247	67.803
35	34.336	36.475	38.859	41.778	43.640	46.059	47.663	49.802	53.203	57.342	60.275	66.619	69.199
36	35.336	37.505	39.922	42.879	44.764	47.212	48.835	50.998	54.437	58.619	61.581	67.985	70.588
37	36.336	38.535	40.984	43.978	45.886	48.363	50.005	52.192	55.668	59.893	62.883	69.346	71.972
38	37.335	39.564	42.045	45.076	47.007	49.513	51.173	53.384	56.896	61.162	64.181	70.703	73.351
39	38.335	40.593	43.105	46.173	48.126	50.660	52.338	54.572	58.120	62.428	65.476	72.055	74.725
40	39.335	41.622	44.165	47.269	49.244	51.805	53.501	55.758	59.342	63.691	66.766	73.402	76.095
45	44.335	46.761	49.452	52.729	54.810	57.505	59.287	61.656	65.410	69.957	73.166	80.077	82.876
50	49.335	51.892	54.723	58.164	60.346	63.167	65.030	67.505	71.420	76.154	79.490	86.661	89.561
60	59.335	62.135	65.227	68.972	71.341	74.397	76.411	79.082	83.298	88.379	91.952	99.607	102.695
70	69.334	72.358	75.689	79.715	82.255	85.527	87.680	90.531	95.023	100.425	104.215	112.317	115.578
80	79.334	82.566	86.120	90.405	93.106	96.578	98.861	101.879	106.629	112.329	116.321	124.839	128.261
90	89.334	92.761	96.524	101.054	103.904	107.565	109.969	113.145	118.136	124.116	128.299	137.208	140.782
100	99.334	102.946	106.906	111.667	114.659	118.498	121.017	124.342	129.561	135.807	140.169	149.449	153.167
120	119.334	123.289	127.616	132.806	136.062	140.233	142.965	146.567	152.211	158.950	163.648	173.617	177.603
150	149.334	153.753	158.577	164.349	167.962	172.581	175.602	179.581	185.800	193.208	198.360	209.265	213.613

ΠΙΝΑΚΑΣ F

Πίνακας Α΄.7: Αν $X \sim F_{\nu_1,\nu_2}$, ο πίνακας δίνει τα σημεία $F_{\nu_1,\nu_2,\alpha}$, που είναι τέτοια ώστε $P(X \geqslant F_{\nu_1,\nu_2,\alpha}) = \alpha$.

		ν_1															
ν_2	α	1	2	3	4	5	6	7	8	9	10	12	15	20	30	50	∞
2	0.1	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38	9.39	9.41	9.42	9.44	9.46	9.47	9.49
	0.05	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41	19.43	19.45	19.46	19.48	19.50
	0.025	38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.37	39.39	39.40	39.41	39.43	39.45	39.46	39.48	39.50
	0.01	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39	99.40	99.42	99.43	99.45	99.47	99.48	99.50
	0.005	198.50	199.00	199.17	199.25	199.30	199.33	199.36	199.37	199.39	199.40	199.42	199.43	199.45	199.47	199.48	199.50
	0.001	998.50	999.00	999.17	999.25	999.30	999.33	999.36	999.37	999.39	999.40	999.42	999.43	999.45	999.47	999.48	999.50
3	0.1	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24	5.23	5.22	5.20	5.18	5.17	5.15	5.13
	0.05	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.62	8.58	8.53
	0.025	17.44	16.04	15.44	15.10	14.88	14.73	14.62	14.54	14.47	14.42	14.34	14.25	14.17	14.08	14.01	13.90
	0.01	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35	27.23	27.05	26.87	26.69	26.50	26.35	26.13
	0.005	55.55	49.80	47.47	46.19	45.39	44.84	44.43	44.13	43.88	43.69	43.39	43.08	42.78	42.47	42.21	41.83
	0.001	167.03	148.50	141.11	137.10	134.58	132.85	131.58	130.62	129.86	129.25	128.32	127.37	126.42	125.45	124.66	123.47
4	0.1	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.94	3.92	3.90	3.87	3.84	3.82	3.80	3.76
	0.05	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.75	5.70	5.63
	0.025	12.22	10.65	9.98	9.60	9.36	9.20	9.07	8.98	8.90	8.84	8.75	8.66	8.56	8.46	8.38	8.26
	0.01	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55	14.37	14.20	14.02	13.84	13.69	13.46
	0.005	31.33	26.28	24.26	23.15	22.46	21.97	21.62	21.35	21.14	20.97	20.70	20.44	20.17	19.89	19.67	19.32
	0.001	74.14	61.25	56.18	53.44	51.71	50.53	49.66	49.00	48.47	48.05	47.41	46.76	46.10	45.43	44.88	44.05
5	0.1	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32	3.30	3.27	3.24	3.21	3.17	3.15	3.10
	0.05	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.50	4.44	4.36
	0.025	10.01	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68	6.62	6.52	6.43	6.33	6.23	6.14	6.02
	0.01	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05	9.89	9.72	9.55	9.38	9.24	9.02
	0.005	22.78	18.31	16.53	15.56	14.94	14.51	14.20	13.96	13.77	13.62	13.38	13.15	12.90	12.66	12.45	12.14
	0.001	47.18	37.12	33.20	31.09	29.75	28.83	28.16	27.65	27.24	26.92	26.42	25.91	25.39	24.87	24.44	23.79
6	0.1	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.96	2.94	2.90	2.87	2.84	2.80	2.77	2.72
	0.05	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.81	3.75	3.67
	0.025	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.52	5.46	5.37	5.27	5.17	5.07	4.98	4.85
	0.01	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.72	7.56	7.40	7.23	7.09	6.88
	0.005	18.63	14.54	12.92	12.03	11.46	11.07	10.79	10.57	10.39	10.25	10.03	9.81	9.59	9.36	9.17	8.88
	0.001	35.51	27.00	23.70	21.92	20.80	20.03	19.46	19.03	18.69	18.41	17.99	17.56	17.12	16.67	16.31	15.75
7	0.1	3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.72	2.70	2.67	2.63	2.59	2.56	2.52	2.47
	0.05	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.38	3.32	3.23
	0.025	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.82	4.76	4.67	4.57	4.47	4.36	4.28	4.14
	0.01	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.47	6.31	6.16	5.99	5.86	5.65
	0.005	16.24	12.40	10.88	10.05	9.52	9.16	8.89	8.68	8.51	8.38	8.18	7.97	7.75	7.53	7.35	7.08
	0.001	29.25	21.69	18.77	17.20	16.21	15.52	15.02	14.63	14.33	14.08	13.71	13.32	12.93	12.53	12.20	11.70
8	0.1	3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.56	2.54	2.50	2.46	2.42	2.38	2.35	2.29
	0.05	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.08	3.02	2.93
	0.025	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36	4.30	4.20	4.10	4.00	3.89	3.81	3.67
	0.01	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.67	5.52	5.36	5.20	5.07	4.86
	0.005	14.69	11.04	9.60	8.81	8.30	7.95	7.69	7.50	7.34	7.21	7.01	6.81	6.61	6.40	6.22	5.95
	0.001	25.41	18.49	15.83	14.39	13.48	12.86	12.40	12.05	11.77	11.54	11.19	10.84	10.48	10.11	9.80	9.33

ΠΙΝΑΚΑΣ F

Πίνακας Α΄.8: Αν $X \sim F_{\nu_1,\nu_2}$, ο πίνακας δίνει τα σημεία $F_{\nu_1,\nu_2,\alpha}$, που είναι τέτοια ώστε $P(X \geqslant F_{\nu_1,\nu_2,\alpha}) = \alpha$.

		ν_1															
ν_2	α	1	2	3	4	5	6	7	8	9	10	12	15	20	30	50	∞
9	0.1	3.36	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.44	2.42	2.38	2.34	2.30	2.25	2.22	2.16
	0.05	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.86	2.80	2.71
	0.025	7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	4.03	3.96	3.87	3.77	3.67	3.56	3.47	3.33
	0.01	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.11	4.96	4.81	4.65	4.52	4.31
	0.005	13.61	10.11	8.72	7.96	7.47	7.13	6.88	6.69	6.54	6.42	6.23	6.03	5.83	5.62	5.45	5.19
	0.001	22.86	16.39	13.90	12.56	11.71	11.13	10.70	10.37	10.11	9.89	9.57	9.24	8.90	8.55	8.26	7.81
10	0.1	3.29	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.35	2.32	2.28	2.24	2.20	2.16	2.12	2.06
	0.05	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.70	2.64	2.54
	0.025	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78	3.72	3.62	3.52	3.42	3.31	3.22	3.08
	0.01	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.71	4.56	4.41	4.25	4.12	3.91
	0.005	12.83	9.43	8.08	7.34	6.87	6.54	6.30	6.12	5.97	5.85	5.66	5.47	5.27	5.07	4.90	4.64
	0.001	21.04	14.91	12.55	11.28	10.48	9.93	9.52	9.20	8.96	8.75	8.45	8.13	7.80	7.47	7.19	6.76
11	0.1	3.23	2.86	2.66	2.54	2.45	2.39	2.34	2.30	2.27	2.25	2.21	2.17	2.12	2.08	2.04	1.97
	0.05	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.57	2.51	2.40
	0.025	6.72	5.26	4.63	4.28	4.04	3.88	3.76	3.66	3.59	3.53	3.43	3.33	3.23	3.12	3.03	2.88
	0.01	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.40	4.25	4.10	3.94	3.81	3.60
	0.005	12.23	8.91	7.60	6.88	6.42	6.10	5.86	5.68	5.54	5.42	5.24	5.05	4.86	4.65	4.49	4.23
	0.001	19.69	13.81	11.56	10.35	9.58	9.05	8.66	8.35	8.12	7.92	7.63	7.32	7.01	6.68	6.42	6.00
12	0.1	3.18	2.81	2.61	2.48	2.39	2.33	2.28	2.24	2.21	2.19	2.15	2.10	2.06	2.01	1.97	1.90
	0.05	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.47	2.40	2.30
	0.025	6.55	5.10	4.47	4.12	3.89	3.73	3.61	3.51	3.44	3.37	3.28	3.18	3.07	2.96	2.87	2.72
	0.01	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.16	4.01	3.86	3.70	3.57	3.36
	0.005	11.75	8.51	7.23	6.52	6.07	5.76	5.52	5.35	5.20	5.09	4.91	4.72	4.53	4.33	4.17	3.90
	0.001	18.64	12.97	10.80	9.63	8.89	8.38	8.00	7.71	7.48	7.29	7.00	6.71	6.40	6.09	5.83	5.42
13	0.1	3.14	2.76	2.56	2.43	2.35	2.28	2.23	2.20	2.16	2.14	2.10	2.05	2.01	1.96	1.92	1.85
	0.05	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.38	2.31	2.21
	0.025	6.41	4.97	4.35	4.00	3.77	3.60	3.48	3.39	3.31	3.25	3.15	3.05	2.95	2.84	2.74	2.60
	0.01	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	3.96	3.82	3.66	3.51	3.38	3.17
	0.005	11.37	8.19	6.93	6.23	5.79	5.48	5.25	5.08	4.94	4.82	4.64	4.46	4.27	4.07	3.91	3.65
	0.001	17.82	12.31	10.21	9.07	8.35	7.86	7.49	7.21	6.98	6.80	6.52	6.23	5.93	5.63	5.37	4.97
14	0.1	3.10	2.73	2.52	2.39	2.31	2.24	2.19	2.15	2.12	2.10	2.05	2.01	1.96	1.91	1.87	1.80
	0.05	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.31	2.24	2.13
	0.025	6.30	4.86	4.24	3.89	3.66	3.50	3.38	3.29	3.21	3.15	3.05	2.95	2.84	2.73	2.64	2.49
	0.01	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94	3.80	3.66	3.51	3.35	3.22	3.00
	0.005	11.06	7.92	6.68	6.00	5.56	5.26	5.03	4.86	4.72	4.60	4.43	4.25	4.06	3.86	3.70	3.44
	0.001	17.14	11.78	9.73	8.62	7.92	7.44	7.08	6.80	6.58	6.40	6.13	5.85	5.56	5.25	5.00	4.60
15	0.1	3.07	2.70	2.49	2.36	2.27	2.21	2.16	2.12	2.09	2.06	2.02	1.97	1.92	1.87	1.83	1.76
	0.05	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.25	2.18	2.07
	0.025	6.20	4.77	4.15	3.80	3.58	3.41	3.29	3.20	3.12	3.06	2.96	2.86	2.76	2.64	2.55	2.40
	0.01	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.67	3.52	3.37	3.21	3.08	2.87
	0.005	10.80	7.70	6.48	5.80	5.37	5.07	4.85	4.67	4.54	4.42	4.25	4.07	3.88	3.69	3.52	3.26
	0.001	16.59	11.34	9.34	8.25	7.57	7.09	6.74	6.47	6.26	6.08	5.81	5.54	5.25	4.95	4.70	4.31
16	0.1	3.05	2.67	2.46	2.33	2.24	2.18	2.13	2.09	2.06	2.03	1.99	1.94	1.89	1.84	1.79	1.72
	0.05	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.19	2.12	2.01
	0.025	6.12	4.69	4.08	3.73	3.50	3.34	3.22	3.12	3.05	2.99	2.89	2.79	2.68	2.57	2.47	2.32
	0.01	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.55	3.41	3.26	3.10	2.97	2.75
	0.005	10.58	7.51	6.30	5.64	5.21	4.91	4.69	4.52	4.38	4.27	4.10	3.92	3.73	3.54	3.37	3.11
	0.001	16.12	10.97	9.01	7.94	7.27	6.80	6.46	6.19	5.98	5.81	5.55	5.27	4.99	4.70	4.45	4.06

ΠΙΝΑΚΑΣ F

Πίνακας Α΄.9: Αν $X \sim F_{\nu_1,\nu_2}$, ο πίνακας δίνει τα σημεία $F_{\nu_1,\nu_2,\alpha}$, που είναι τέτοια ώστε $P(X \geqslant F_{\nu_1,\nu_2,\alpha}) = \alpha$.

		ν_1															
ν_2	α	1	2	3	4	5	6	7	8	9	10	12	15	20	30	50	∞
17	0.1	3.03	2.64	2.44	2.31	2.22	2.15	2.10	2.06	2.03	2.00	1.96	1.91	1.86	1.81	1.76	1.69
	0.05	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.15	2.08	1.96
	0.025	6.04	4.62	4.01	3.66	3.44	3.28	3.16	3.06	2.98	2.92	2.82	2.72	2.62	2.50	2.41	2.25
	0.01	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.46	3.31	3.16	3.00	2.87	2.65
	0.005	10.38	7.35	6.16	5.50	5.07	4.78	4.56	4.39	4.25	4.14	3.97	3.79	3.61	3.41	3.25	2.98
	0.001	15.72	10.66	8.73	7.68	7.02	6.56	6.22	5.96	5.75	5.58	5.32	5.05	4.78	4.48	4.24	3.85
18	0.1	3.01	2.62	2.42	2.29	2.20	2.13	2.08	2.04	2.00	1.98	1.93	1.89	1.84	1.78	1.74	1.66
	0.05	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.11	2.04	1.92
	0.025	5.98	4.56	3.95	3.61	3.38	3.22	3.10	3.01	2.93	2.87	2.77	2.67	2.56	2.44	2.35	2.19
	0.01	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.37	3.23	3.08	2.92	2.78	2.57
	0.005	10.22	7.21	6.03	5.37	4.96	4.66	4.44	4.28	4.14	4.03	3.86	3.68	3.50	3.30	3.14	2.87
	0.001	15.38	10.39	8.49	7.46	6.81	6.35	6.02	5.76	5.56	5.39	5.13	4.87	4.59	4.30	4.06	3.67
19	0.1	2.99	2.61	2.40	2.27	2.18	2.11	2.06	2.02	1.98	1.96	1.91	1.86	1.81	1.76	1.71	1.63
	0.05	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.07	2.00	1.88
	0.025	5.92	4.51	3.90	3.56	3.33	3.17	3.05	2.96	2.88	2.82	2.72	2.62	2.51	2.39	2.30	2.13
	0.01	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.30	3.15	3.00	2.84	2.71	2.49
	0.005	10.07	7.09	5.92	5.27	4.85	4.56	4.34	4.18	4.04	3.93	3.76	3.59	3.40	3.21	3.04	2.78
	0.001	15.08	10.16	8.28	7.27	6.62	6.18	5.85	5.59	5.39	5.22	4.97	4.70	4.43	4.14	3.90	3.51
20	0.1	2.97	2.59	2.38	2.25	2.16	2.09	2.04	2.00	1.96	1.94	1.89	1.84	1.79	1.74	1.69	1.61
	0.05	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.04	1.97	1.84
	0.025	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84	2.77	2.68	2.57	2.46	2.35	2.25	2.09
	0.01	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.23	3.09	2.94	2.78	2.64	2.42
	0.005	9.94	6.99	5.82	5.17	4.76	4.47	4.26	4.09	3.96	3.85	3.68	3.50	3.32	3.12	2.96	2.69
	0.001	14.82	9.95	8.10	7.10	6.46	6.02	5.69	5.44	5.24	5.08	4.82	4.56	4.29	4.00	3.77	3.38
21	0.1	2.96	2.57	2.36	2.23	2.14	2.08	2.02	1.98	1.95	1.92	1.87	1.83	1.78	1.72	1.67	1.59
	0.05	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.18	2.10	2.01	1.94	1.81
	0.025	5.83	4.42	3.82	3.48	3.25	3.09	2.97	2.87	2.80	2.73	2.64	2.53	2.42	2.31	2.21	2.04
	0.01	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31	3.17	3.03	2.88	2.72	2.58	2.36
	0.005	9.83	6.89	5.73	5.09	4.68	4.39	4.18	4.01	3.88	3.77	3.60	3.43	3.24	3.05	2.88	2.61
	0.001	14.59	9.77	7.94	6.95	6.32	5.88	5.56	5.31	5.11	4.95	4.70	4.44	4.17	3.88	3.64	3.26
22	0.1	2.95	2.56	2.35	2.22	2.13	2.06	2.01	1.97	1.93	1.90	1.86	1.81	1.76	1.70	1.65	1.57
	0.05	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.15	2.07	1.98	1.91	1.78
	0.025	5.79	4.38	3.78	3.44	3.22	3.05	2.93	2.84	2.76	2.70	2.60	2.50	2.39	2.27	2.17	2.00
	0.01	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.12	2.98	2.83	2.67	2.53	2.31
	0.005	9.73	6.81	5.65	5.02	4.61	4.32	4.11	3.94	3.81	3.70	3.54	3.36	3.18	2.98	2.82	2.55
	0.001	14.38	9.61	7.80	6.81	6.19	5.76	5.44	5.19	4.99	4.83	4.58	4.33	4.06	3.78	3.54	3.15
23	0.1	2.94	2.55	2.34	2.21	2.11	2.05	1.99	1.95	1.92	1.89	1.84	1.80	1.74	1.69	1.64	1.55
	0.05	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.13	2.05	1.96	1.88	1.76
	0.025	5.75	4.35	3.75	3.41	3.18	3.02	2.90	2.81	2.73	2.67	2.57	2.47	2.36	2.24	2.14	1.97
	0.01	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21	3.07	2.93	2.78	2.62	2.48	2.26
	0.005	9.63	6.73	5.58	4.95	4.54	4.26	4.05	3.88	3.75	3.64	3.47	3.30	3.12	2.92	2.76	2.48
	0.001	14.20	9.47	7.67	6.70	6.08	5.65	5.33	5.09	4.89	4.73	4.48	4.23	3.96	3.68	3.44	3.05
24	0.1	2.93	2.54	2.33	2.19	2.10	2.04	1.98	1.94	1.91	1.88	1.83	1.78	1.73	1.67	1.62	1.53
	0.05	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.11	2.03	1.94	1.86	1.73
	0.025	5.72	4.32	3.72	3.38	3.15	2.99	2.87	2.78	2.70	2.64	2.54	2.44	2.33	2.21	2.11	1.94
	0.01	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	3.03	2.89	2.74	2.58	2.44	2.21
	0.005	9.55	6.66	5.52	4.89	4.49	4.20	3.99	3.83	3.69	3.59	3.42	3.25	3.06	2.87	2.70	2.43
	0.001	14.03	9.34	7.55	6.59	5.98	5.55	5.23	4.99	4.80	4.64	4.39	4.14	3.87	3.59	3.36	2.97

ΠΙΝΑΚΑΣ F

Πίνακας Α΄.10: Αν $X \sim F_{\nu_1,\nu_2}$, ο πίνακας δίνει τα σημεία $F_{\nu_1,\nu_2,\alpha}$, που είναι τέτοια ώστε $P(X \geqslant F_{\nu_1,\nu_2,\alpha}) = \alpha$.

		ν_1															
ν_2	α	1	2	3	4	5	6	7	8	9	10	12	15	20	30	50	∞
25	0.1	2.92	2.53	2.32	2.18	2.09	2.02	1.97	1.93	1.89	1.87	1.82	1.77	1.72	1.66	1.61	1.52
	0.05	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.09	2.01	1.92	1.84	1.71
	0.025	5.69	4.29	3.69	3.35	3.13	2.97	2.85	2.75	2.68	2.61	2.51	2.41	2.30	2.18	2.08	1.91
	0.01	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22	3.13	2.99	2.85	2.70	2.54	2.40	2.17
	0.005	9.48	6.60	5.46	4.84	4.43	4.15	3.94	3.78	3.64	3.54	3.37	3.20	3.01	2.82	2.65	2.38
	0.001	13.88	9.22	7.45	6.49	5.89	5.46	5.15	4.91	4.71	4.56	4.31	4.06	3.79	3.52	3.28	2.89
26	0.1	2.91	2.52	2.31	2.17	2.08	2.01	1.96	1.92	1.88	1.86	1.81	1.76	1.71	1.65	1.59	1.50
	0.05	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.15	2.07	1.99	1.90	1.82	1.69
	0.025	5.66	4.27	3.67	3.33	3.10	2.94	2.82	2.73	2.65	2.59	2.49	2.39	2.28	2.16	2.05	1.88
	0.01	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09	2.96	2.81	2.66	2.50	2.36	2.13
	0.005	9.41	6.54	5.41	4.79	4.38	4.10	3.89	3.73	3.60	3.49	3.33	3.15	2.97	2.77	2.61	2.33
	0.001	13.74	9.12	7.36	6.41	5.80	5.38	5.07	4.83	4.64	4.48	4.24	3.99	3.72	3.44	3.21	2.82
27	0.1	2.90	2.51	2.30	2.17	2.07	2.00	1.95	1.91	1.87	1.85	1.80	1.75	1.70	1.64	1.58	1.49
	0.05	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20	2.13	2.06	1.97	1.88	1.81	1.67
	0.025	5.63	4.24	3.65	3.31	3.08	2.92	2.80	2.71	2.63	2.57	2.47	2.36	2.25	2.13	2.03	1.85
	0.01	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15	3.06	2.93	2.78	2.63	2.47	2.33	2.10
	0.005	9.34	6.49	5.36	4.74	4.34	4.06	3.85	3.69	3.56	3.45	3.28	3.11	2.93	2.73	2.57	2.29
	0.001	13.61	9.02	7.27	6.33	5.73	5.31	5.00	4.76	4.57	4.41	4.17	3.92	3.66	3.38	3.14	2.75
28	0.1	2.89	2.50	2.29	2.16	2.06	2.00	1.94	1.90	1.87	1.84	1.79	1.74	1.69	1.63	1.57	1.48
	0.05	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.12	2.04	1.96	1.87	1.79	1.65
	0.025	5.61	4.22	3.63	3.29	3.06	2.90	2.78	2.69	2.61	2.55	2.45	2.34	2.23	2.11	2.01	1.83
	0.01	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03	2.90	2.75	2.60	2.44	2.30	2.06
	0.005	9.28	6.44	5.32	4.70	4.30	4.02	3.81	3.65	3.52	3.41	3.25	3.07	2.89	2.69	2.53	2.25
	0.001	13.50	8.93	7.19	6.25	5.66	5.24	4.93	4.69	4.50	4.35	4.11	3.86	3.60	3.32	3.09	2.69
29	0.1	2.89	2.50	2.28	2.15	2.06	1.99	1.93	1.89	1.86	1.83	1.78	1.73	1.68	1.62	1.56	1.47
	0.05	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18	2.10	2.03	1.94	1.85	1.77	1.64
	0.025	5.59	4.20	3.61	3.27	3.04	2.88	2.76	2.67	2.59	2.53	2.43	2.32	2.21	2.09	1.99	1.81
	0.01	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09	3.00	2.87	2.73	2.57	2.41	2.27	2.03
	0.005	9.23	6.40	5.28	4.66	4.26	3.98	3.77	3.61	3.48	3.38	3.21	3.04	2.86	2.66	2.49	2.21
	0.001	13.39	8.85	7.12	6.19	5.59	5.18	4.87	4.64	4.45	4.29	4.05	3.80	3.54	3.27	3.03	2.64
30	0.1	2.88	2.49	2.28	2.14	2.05	1.98	1.93	1.88	1.85	1.82	1.77	1.72	1.67	1.61	1.55	1.46
	0.05	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.84	1.76	1.62
	0.025	5.57	4.18	3.59	3.25	3.03	2.87	2.75	2.65	2.57	2.51	2.41	2.31	2.20	2.07	1.97	1.79
	0.01	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.84	2.70	2.55	2.39	2.25	2.01
	0.005	9.18	6.35	5.24	4.62	4.23	3.95	3.74	3.58	3.45	3.34	3.18	3.01	2.82	2.63	2.46	2.18
	0.001	13.29	8.77	7.05	6.12	5.53	5.12	4.82	4.58	4.39	4.24	4.00	3.75	3.49	3.22	2.98	2.59
31	0.1	2.87	2.48	2.27	2.14	2.04	1.97	1.92	1.88	1.84	1.81	1.77	1.71	1.66	1.60	1.54	1.45
	0.05	4.16	3.30	2.91	2.68	2.52	2.41	2.32	2.25	2.20	2.15	2.08	2.00	1.92	1.83	1.75	1.61
	0.025	5.55	4.16	3.57	3.23	3.01	2.85	2.73	2.64	2.56	2.50	2.40	2.29	2.18	2.06	1.95	1.77
	0.01	7.53	5.36	4.48	3.99	3.67	3.45	3.28	3.15	3.04	2.96	2.82	2.68	2.52	2.36	2.22	1.98
	0.005	9.13	6.32	5.20	4.59	4.20	3.92	3.71	3.55	3.42	3.31	3.15	2.98	2.79	2.60	2.43	2.14
	0.001	13.20	8.70	6.99	6.07	5.48	5.07	4.77	4.53	4.34	4.19	3.95	3.71	3.45	3.17	2.94	2.54
32	0.1	2.87	2.48	2.26	2.13	2.04	1.97	1.91	1.87	1.83	1.81	1.76	1.71	1.65	1.59	1.53	1.44
	0.05	4.15	3.29	2.90	2.67	2.51	2.40	2.31	2.24	2.19	2.14	2.07	1.99	1.91	1.82	1.74	1.59
	0.025	5.53	4.15	3.56	3.22	3.00	2.84	2.71	2.62	2.54	2.48	2.38	2.28	2.16	2.04	1.93	1.75
	0.01	7.50	5.34	4.46	3.97	3.65	3.43	3.26	3.13	3.02	2.93	2.80	2.65	2.50	2.34	2.20	1.96
	0.005	9.09	6.28	5.17	4.56	4.17	3.89	3.68	3.52	3.39	3.29	3.12	2.95	2.77	2.57	2.40	2.11
	0.001	13.12	8.64	6.94	6.01	5.43	5.02	4.72	4.48	4.30	4.14	3.91	3.66	3.40	3.13	2.89	2.50

ΠΙΝΑΚΑΣ F

Πίνακας Α΄.11: Αν $X \sim F_{\nu_1,\nu_2}$, ο πίνακας δίνει τα σημεία $F_{\nu_1,\nu_2,\alpha}$ που είναι τέτοια ώστε $P(X \geqslant F_{\nu_1,\nu_2,\alpha}) = \alpha$.

		ν_1															
ν_2	α	1	2	3	4	5	6	7	8	9	10	12	15	20	30	50	∞
33	0.1	2.86	2.47	2.26	2.12	2.03	1.96	1.91	1.86	1.83	1.80	1.75	1.70	1.64	1.58	1.53	1.43
	0.05	4.14	3.28	2.89	2.66	2.50	2.39	2.30	2.23	2.18	2.13	2.06	1.98	1.90	1.81	1.72	1.58
	0.025	5.51	4.13	3.54	3.20	2.98	2.82	2.70	2.61	2.53	2.47	2.37	2.26	2.15	2.03	1.92	1.73
	0.01	7.47	5.31	4.44	3.95	3.63	3.41	3.24	3.11	3.00	2.91	2.78	2.63	2.48	2.32	2.18	1.93
	0.005	9.05	6.25	5.14	4.53	4.14	3.86	3.66	3.49	3.37	3.26	3.09	2.92	2.74	2.54	2.37	2.09
	0.001	13.04	8.58	6.88	5.97	5.38	4.98	4.67	4.44	4.26	4.10	3.87	3.62	3.36	3.09	2.85	2.46
34	0.1	2.86	2.47	2.25	2.12	2.02	1.96	1.90	1.86	1.82	1.79	1.75	1.69	1.64	1.58	1.52	1.42
	0.05	4.13	3.28	2.88	2.65	2.49	2.38	2.29	2.23	2.17	2.12	2.05	1.97	1.89	1.80	1.71	1.57
	0.025	5.50	4.12	3.53	3.19	2.97	2.81	2.69	2.59	2.52	2.45	2.35	2.25	2.13	2.01	1.90	1.72
	0.01	7.44	5.29	4.42	3.93	3.61	3.39	3.22	3.09	2.98	2.89	2.76	2.61	2.46	2.30	2.16	1.91
	0.005	9.01	6.22	5.11	4.50	4.11	3.84	3.63	3.47	3.34	3.24	3.07	2.90	2.72	2.52	2.35	2.06
	0.001	12.97	8.52	6.83	5.92	5.34	4.93	4.63	4.40	4.22	4.06	3.83	3.58	3.33	3.05	2.82	2.42
35	0.1	2.85	2.46	2.25	2.11	2.02	1.95	1.90	1.85	1.82	1.79	1.74	1.69	1.63	1.57	1.51	1.41
	0.05	4.12	3.27	2.87	2.64	2.49	2.37	2.29	2.22	2.16	2.11	2.04	1.96	1.88	1.79	1.70	1.56
	0.025	5.48	4.11	3.52	3.18	2.96	2.80	2.68	2.58	2.50	2.44	2.34	2.23	2.12	2.00	1.89	1.70
	0.01	7.42	5.27	4.40	3.91	3.59	3.37	3.20	3.07	2.96	2.88	2.74	2.60	2.44	2.28	2.14	1.89
	0.005	8.98	6.19	5.09	4.48	4.09	3.81	3.61	3.45	3.32	3.21	3.05	2.88	2.69	2.50	2.33	2.04
	0.001	12.90	8.47	6.79	5.88	5.30	4.89	4.59	4.36	4.18	4.03	3.79	3.55	3.29	3.02	2.78	2.38
36	0.1	2.85	2.46	2.24	2.11	2.01	1.94	1.89	1.85	1.81	1.78	1.73	1.68	1.63	1.56	1.51	1.40
	0.05	4.11	3.26	2.87	2.63	2.48	2.36	2.28	2.21	2.15	2.11	2.03	1.95	1.87	1.78	1.69	1.55
	0.025	5.47	4.09	3.50	3.17	2.94	2.78	2.66	2.57	2.49	2.43	2.33	2.22	2.11	1.99	1.88	1.69
	0.01	7.40	5.25	4.38	3.89	3.57	3.35	3.18	3.05	2.95	2.86	2.72	2.58	2.43	2.26	2.12	1.87
	0.005	8.94	6.16	5.06	4.46	4.06	3.79	3.58	3.42	3.30	3.19	3.03	2.85	2.67	2.48	2.30	2.01
	0.001	12.83	8.42	6.74	5.84	5.26	4.86	4.56	4.33	4.14	3.99	3.76	3.51	3.26	2.98	2.75	2.35
37	0.1	2.85	2.45	2.24	2.10	2.01	1.94	1.89	1.84	1.81	1.78	1.73	1.68	1.62	1.56	1.50	1.40
	0.05	4.11	3.25	2.86	2.63	2.47	2.36	2.27	2.20	2.14	2.10	2.02	1.95	1.86	1.77	1.68	1.54
	0.025	5.46	4.08	3.49	3.16	2.93	2.77	2.65	2.56	2.48	2.42	2.32	2.21	2.10	1.97	1.87	1.67
	0.01	7.37	5.23	4.36	3.87	3.56	3.33	3.17	3.04	2.93	2.84	2.71	2.56	2.41	2.25	2.10	1.85
	0.005	8.91	6.13	5.04	4.43	4.04	3.77	3.56	3.40	3.28	3.17	3.01	2.83	2.65	2.46	2.28	1.99
	0.001	12.77	8.37	6.70	5.80	5.22	4.82	4.53	4.30	4.11	3.96	3.73	3.48	3.23	2.95	2.72	2.32
38	0.1	2.84	2.45	2.23	2.10	2.01	1.94	1.88	1.84	1.80	1.77	1.72	1.67	1.61	1.55	1.49	1.39
	0.05	4.10	3.24	2.85	2.62	2.46	2.35	2.26	2.19	2.14	2.09	2.02	1.94	1.85	1.76	1.68	1.53
	0.025	5.45	4.07	3.48	3.15	2.92	2.76	2.64	2.55	2.47	2.41	2.31	2.20	2.09	1.96	1.85	1.66
	0.01	7.35	5.21	4.34	3.86	3.54	3.32	3.15	3.02	2.92	2.83	2.69	2.55	2.40	2.23	2.09	1.84
	0.005	8.88	6.11	5.02	4.41	4.02	3.75	3.54	3.39	3.26	3.15	2.99	2.82	2.63	2.44	2.27	1.97
	0.001	12.71	8.33	6.66	5.76	5.19	4.79	4.49	4.26	4.08	3.93	3.70	3.45	3.20	2.92	2.69	2.29
39	0.1	2.84	2.44	2.23	2.09	2.00	1.93	1.88	1.83	1.80	1.77	1.72	1.67	1.61	1.55	1.49	1.38
	0.05	4.09	3.24	2.85	2.61	2.46	2.34	2.26	2.19	2.13	2.08	2.01	1.93	1.85	1.75	1.67	1.52
	0.025	5.43	4.06	3.47	3.14	2.91	2.75	2.63	2.54	2.46	2.40	2.30	2.19	2.08	1.95	1.84	1.65
	0.01	7.33	5.19	4.33	3.84	3.53	3.30	3.14	3.01	2.90	2.81	2.68	2.54	2.38	2.22	2.07	1.82
	0.005	8.85	6.09	5.00	4.39	4.00	3.73	3.53	3.37	3.24	3.13	2.97	2.80	2.62	2.42	2.25	1.95
	0.001	12.66	8.29	6.63	5.73	5.16	4.76	4.46	4.23	4.05	3.90	3.67	3.43	3.17	2.90	2.66	2.26
40	0.1	2.84	2.44	2.23	2.09	2.00	1.93	1.87	1.83	1.79	1.76	1.71	1.66	1.61	1.54	1.48	1.38
	0.05	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92	1.84	1.74	1.66	1.51
	0.025	5.42	4.05	3.46	3.13	2.90	2.74	2.62	2.53	2.45	2.39	2.29	2.18	2.07	1.94	1.83	1.64
	0.01	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.66	2.52	2.37	2.20	2.06	1.80
	0.005	8.83	6.07	4.98	4.37	3.99	3.71	3.51	3.35	3.22	3.12	2.95	2.78	2.60	2.40	2.23	1.93
	0.001	12.61	8.25	6.59	5.70	5.13	4.73	4.44	4.21	4.02	3.87	3.64	3.40	3.14	2.87	2.64	2.23

ΠΙΝΑΚΑΣ F

Πίνακας Α΄.12: Αν $X \sim F_{\nu_1,\nu_2}$, ο πίνακας δίνει τα σημεία $F_{\nu_1,\nu_2,\alpha}$ που είναι τέτοια ώστε $P(X \geqslant F_{\nu_1,\nu_2,\alpha}) = \alpha$.

		ν_1															
ν_2	α	1	2	3	4	5	6	7	8	9	10	12	15	20	30	50	∞
45	0.1	2.82	2.42	2.21	2.07	1.98	1.91	1.85	1.81	1.77	1.74	1.70	1.64	1.58	1.52	1.46	1.35
	0.05	4.06	3.20	2.81	2.58	2.42	2.31	2.22	2.15	2.10	2.05	1.97	1.89	1.81	1.71	1.63	1.47
	0.025	5.38	4.01	3.42	3.09	2.86	2.70	2.58	2.49	2.41	2.35	2.25	2.14	2.03	1.90	1.79	1.59
	0.01	7.23	5.11	4.25	3.77	3.45	3.23	3.07	2.94	2.83	2.74	2.61	2.46	2.31	2.14	2.00	1.74
	0.005	8.71	5.97	4.89	4.29	3.91	3.64	3.43	3.28	3.15	3.04	2.88	2.71	2.53	2.33	2.16	1.85
	0.001	12.39	8.09	6.45	5.56	5.00	4.61	4.32	4.09	3.91	3.76	3.53	3.29	3.04	2.76	2.53	2.12
50	0.1	2.81	2.41	2.20	2.06	1.97	1.90	1.84	1.80	1.76	1.73	1.68	1.63	1.57	1.50	1.44	1.33
	0.05	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.03	1.95	1.87	1.78	1.69	1.60	1.44
	0.025	5.34	3.97	3.39	3.05	2.83	2.67	2.55	2.46	2.38	2.32	2.22	2.11	1.99	1.87	1.75	1.55
	0.01	7.17	5.06	4.20	3.72	3.41	3.19	3.02	2.89	2.78	2.70	2.56	2.42	2.27	2.10	1.95	1.68
	0.005	8.63	5.90	4.83	4.23	3.85	3.58	3.38	3.22	3.09	2.99	2.82	2.65	2.47	2.27	2.10	1.79
	0.001	12.22	7.96	6.34	5.46	4.90	4.51	4.22	4.00	3.82	3.67	3.44	3.20	2.95	2.68	2.44	2.03
55	0.1	2.80	2.40	2.19	2.05	1.95	1.88	1.83	1.78	1.75	1.72	1.67	1.61	1.55	1.49	1.43	1.31
	0.05	4.02	3.16	2.77	2.54	2.38	2.27	2.18	2.11	2.06	2.01	1.93	1.85	1.76	1.67	1.58	1.41
	0.025	5.31	3.95	3.36	3.03	2.81	2.65	2.53	2.43	2.36	2.29	2.19	2.08	1.97	1.84	1.72	1.51
	0.01	7.12	5.01	4.16	3.68	3.37	3.15	2.98	2.85	2.75	2.66	2.53	2.38	2.23	2.06	1.91	1.64
	0.005	8.55	5.84	4.77	4.18	3.80	3.53	3.33	3.17	3.05	2.94	2.78	2.61	2.42	2.23	2.05	1.73
	0.001	12.09	7.85	6.25	5.38	4.82	4.43	4.15	3.92	3.75	3.60	3.37	3.13	2.88	2.61	2.37	1.95
60	0.1	2.79	2.39	2.18	2.04	1.95	1.87	1.82	1.77	1.74	1.71	1.66	1.60	1.54	1.48	1.41	1.29
	0.05	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.65	1.56	1.39
	0.025	5.29	3.93	3.34	3.01	2.79	2.63	2.51	2.41	2.33	2.27	2.17	2.06	1.94	1.82	1.70	1.48
	0.01	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.50	2.35	2.20	2.03	1.88	1.60
	0.005	8.49	5.79	4.73	4.14	3.76	3.49	3.29	3.13	3.01	2.90	2.74	2.57	2.39	2.19	2.01	1.69
	0.001	11.97	7.77	6.17	5.31	4.76	4.37	4.09	3.86	3.69	3.54	3.32	3.08	2.83	2.55	2.32	1.89
65	0.1	2.78	2.39	2.17	2.03	1.94	1.87	1.81	1.77	1.73	1.70	1.65	1.59	1.53	1.47	1.40	1.28
	0.05	3.99	3.14	2.75	2.51	2.36	2.24	2.15	2.08	2.03	1.98	1.90	1.82	1.73	1.63	1.54	1.37
	0.025	5.26	3.91	3.32	2.99	2.77	2.61	2.49	2.39	2.32	2.25	2.15	2.04	1.93	1.80	1.68	1.46
	0.01	7.04	4.95	4.10	3.62	3.31	3.09	2.93	2.80	2.69	2.61	2.47	2.33	2.17	2.00	1.85	1.57
	0.005	8.44	5.75	4.69	4.11	3.73	3.46	3.26	3.10	2.98	2.87	2.71	2.54	2.36	2.16	1.98	1.65
	0.001	11.88	7.70	6.11	5.25	4.70	4.32	4.04	3.81	3.64	3.49	3.27	3.03	2.78	2.51	2.27	1.84
70	0.1	2.78	2.38	2.16	2.03	1.93	1.86	1.80	1.76	1.72	1.69	1.64	1.59	1.53	1.46	1.39	1.27
	0.05	3.98	3.13	2.74	2.50	2.35	2.23	2.14	2.07	2.02	1.97	1.89	1.81	1.72	1.62	1.53	1.35
	0.025	5.25	3.89	3.31	2.97	2.75	2.59	2.47	2.38	2.30	2.24	2.14	2.03	1.91	1.78	1.66	1.44
	0.01	7.01	4.92	4.07	3.60	3.29	3.07	2.91	2.78	2.67	2.59	2.45	2.31	2.15	1.98	1.83	1.54
	0.005	8.40	5.72	4.66	4.08	3.70	3.43	3.23	3.08	2.95	2.85	2.68	2.51	2.33	2.13	1.95	1.62
	0.001	11.80	7.64	6.06	5.20	4.66	4.28	3.99	3.77	3.60	3.45	3.23	2.99	2.74	2.47	2.23	1.79
75	0.1	2.77	2.37	2.16	2.02	1.93	1.85	1.80	1.75	1.72	1.69	1.63	1.58	1.52	1.45	1.38	1.25
	0.05	3.97	3.12	2.73	2.49	2.34	2.22	2.13	2.06	2.01	1.96	1.88	1.80	1.71	1.61	1.52	1.34
	0.025	5.23	3.88	3.30	2.96	2.74	2.58	2.46	2.37	2.29	2.22	2.12	2.01	1.90	1.76	1.65	1.42
	0.01	6.99	4.90	4.05	3.58	3.27	3.05	2.89	2.76	2.65	2.57	2.43	2.29	2.13	1.96	1.81	1.52
	0.005	8.37	5.69	4.63	4.05	3.67	3.41	3.21	3.05	2.93	2.82	2.66	2.49	2.31	2.10	1.92	1.59
	0.001	11.73	7.58	6.01	5.16	4.62	4.24	3.96	3.74	3.56	3.42	3.19	2.96	2.71	2.44	2.19	1.75
80	0.1	2.77	2.37	2.15	2.02	1.92	1.85	1.79	1.75	1.71	1.68	1.63	1.57	1.51	1.44	1.38	1.24
	0.05	3.96	3.11	2.72	2.49	2.33	2.21	2.13	2.06	2.00	1.95	1.88	1.79	1.70	1.60	1.51	1.32
	0.025	5.22	3.86	3.28	2.95	2.73	2.57	2.45	2.35	2.28	2.21	2.11	2.00	1.88	1.75	1.63	1.40
	0.01	6.96	4.88	4.04	3.56	3.26	3.04	2.87	2.74	2.64	2.55	2.42	2.27	2.12	1.94	1.79	1.49
	0.005	8.33	5.67	4.61	4.03	3.65	3.39	3.19	3.03	2.91	2.80	2.64	2.47	2.29	2.08	1.90	1.56
	0.001	11.67	7.54	5.97	5.12	4.58	4.20	3.92	3.70	3.53	3.39	3.16	2.93	2.68	2.41	2.16	1.72

ΠΙΝΑΚΑΣ F

Πίνακας Α΄.13: Αν $X \sim F_{\nu_1,\nu_2}$, ο πίνακας δίνει τα σημεία $F_{\nu_1,\nu_2,\alpha}$ που είναι τέτοια ώστε $P(X \geqslant F_{\nu_1,\nu_2,\alpha}) = \alpha$.

		ν_1															
ν_2	α	1	2	3	4	5	6	7	8	9	10	12	15	20	30	50	∞
85	0.1	2.77	2.37	2.15	2.01	1.92	1.84	1.79	1.74	1.71	1.67	1.62	1.57	1.51	1.44	1.37	1.24
	0.05	3.95	3.10	2.71	2.48	2.32	2.21	2.12	2.05	1.99	1.94	1.87	1.79	1.70	1.59	1.50	1.31
	0.025	5.21	3.85	3.27	2.94	2.72	2.56	2.44	2.35	2.27	2.20	2.10	1.99	1.87	1.74	1.62	1.38
	0.01	6.94	4.86	4.02	3.55	3.24	3.02	2.86	2.73	2.62	2.54	2.40	2.26	2.10	1.93	1.77	1.47
	0.005	8.31	5.64	4.59	4.01	3.63	3.37	3.17	3.01	2.89	2.79	2.62	2.45	2.27	2.07	1.88	1.54
	0.001	11.62	7.50	5.94	5.09	4.55	4.18	3.90	3.68	3.50	3.36	3.14	2.90	2.65	2.38	2.14	1.69
90	0.1	2.76	2.36	2.15	2.01	1.91	1.84	1.78	1.74	1.70	1.67	1.62	1.56	1.50	1.43	1.36	1.23
	0.05	3.95	3.10	2.71	2.47	2.32	2.20	2.11	2.04	1.99	1.94	1.86	1.78	1.69	1.59	1.49	1.30
	0.025	5.20	3.84	3.26	2.93	2.71	2.55	2.43	2.34	2.26	2.19	2.09	1.98	1.86	1.73	1.61	1.37
	0.01	6.93	4.85	4.01	3.53	3.23	3.01	2.84	2.72	2.61	2.52	2.39	2.24	2.09	1.92	1.76	1.46
	0.005	8.28	5.62	4.57	3.99	3.62	3.35	3.15	3.00	2.87	2.77	2.61	2.44	2.25	2.05	1.87	1.52
	0.001	11.57	7.47	5.91	5.06	4.53	4.15	3.87	3.65	3.48	3.34	3.11	2.88	2.63	2.36	2.11	1.66
95	0.1	2.76	2.36	2.14	2.00	1.91	1.84	1.78	1.74	1.70	1.67	1.62	1.56	1.50	1.43	1.36	1.22
	0.05	3.94	3.09	2.70	2.47	2.31	2.20	2.11	2.04	1.98	1.93	1.86	1.77	1.68	1.58	1.48	1.29
	0.025	5.19	3.84	3.26	2.92	2.70	2.54	2.42	2.33	2.25	2.19	2.08	1.98	1.86	1.72	1.60	1.36
	0.01	6.91	4.84	3.99	3.52	3.22	3.00	2.83	2.70	2.60	2.51	2.38	2.23	2.08	1.90	1.75	1.44
	0.005	8.26	5.61	4.56	3.98	3.60	3.34	3.14	2.98	2.86	2.76	2.59	2.42	2.24	2.04	1.85	1.50
	0.001	11.53	7.44	5.88	5.04	4.50	4.13	3.85	3.63	3.46	3.31	3.09	2.86	2.61	2.34	2.09	1.64
100	0.1	2.76	2.36	2.14	2.00	1.91	1.83	1.78	1.73	1.69	1.66	1.61	1.56	1.49	1.42	1.35	1.21
	0.05	3.94	3.09	2.70	2.46	2.31	2.19	2.10	2.03	1.97	1.93	1.85	1.77	1.68	1.57	1.48	1.28
	0.025	5.18	3.83	3.25	2.92	2.70	2.54	2.42	2.32	2.24	2.18	2.08	1.97	1.85	1.71	1.59	1.35
	0.01	6.90	4.82	3.98	3.51	3.21	2.99	2.82	2.69	2.59	2.50	2.37	2.22	2.07	1.89	1.74	1.43
	0.005	8.24	5.59	4.54	3.96	3.59	3.33	3.13	2.97	2.85	2.74	2.58	2.41	2.23	2.02	1.84	1.49
	0.001	11.50	7.41	5.86	5.02	4.48	4.11	3.83	3.61	3.44	3.30	3.07	2.84	2.59	2.32	2.08	1.62
110	0.1	2.75	2.35	2.13	2.00	1.90	1.83	1.77	1.73	1.69	1.66	1.61	1.55	1.49	1.42	1.35	1.20
	0.05	3.93	3.08	2.69	2.45	2.30	2.18	2.09	2.02	1.97	1.92	1.84	1.76	1.67	1.56	1.47	1.27
	0.025	5.16	3.82	3.24	2.90	2.68	2.53	2.40	2.31	2.23	2.17	2.07	1.96	1.84	1.70	1.58	1.33
	0.01	6.87	4.80	3.96	3.49	3.19	2.97	2.81	2.68	2.57	2.49	2.35	2.21	2.05	1.88	1.72	1.40
	0.005	8.21	5.56	4.52	3.94	3.57	3.30	3.11	2.95	2.83	2.72	2.56	2.39	2.21	2.00	1.82	1.46
	0.001	11.43	7.36	5.82	4.98	4.45	4.07	3.79	3.58	3.41	3.26	3.04	2.81	2.56	2.29	2.04	1.58
120	0.1	2.75	2.35	2.13	1.99	1.90	1.82	1.77	1.72	1.68	1.65	1.60	1.55	1.48	1.41	1.34	1.19
	0.05	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91	1.83	1.75	1.66	1.55	1.46	1.25
	0.025	5.15	3.80	3.23	2.89	2.67	2.52	2.39	2.30	2.22	2.16	2.05	1.94	1.82	1.69	1.56	1.31
	0.01	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47	2.34	2.19	2.03	1.86	1.70	1.38
	0.005	8.18	5.54	4.50	3.92	3.55	3.28	3.09	2.93	2.81	2.71	2.54	2.37	2.19	1.98	1.80	1.43
	0.001	11.38	7.32	5.78	4.95	4.42	4.04	3.77	3.55	3.38	3.24	3.02	2.78	2.53	2.26	2.02	1.54
130	0.1	2.74	2.34	2.13	1.99	1.89	1.82	1.76	1.72	1.68	1.65	1.60	1.54	1.48	1.40	1.33	1.18
	0.05	3.91	3.07	2.67	2.44	2.28	2.17	2.08	2.01	1.95	1.90	1.83	1.74	1.65	1.55	1.45	1.24
	0.025	5.14	3.80	3.22	2.89	2.67	2.51	2.39	2.29	2.21	2.15	2.05	1.94	1.82	1.68	1.55	1.30
	0.01	6.83	4.77	3.94	3.47	3.16	2.94	2.78	2.65	2.55	2.46	2.32	2.18	2.02	1.85	1.69	1.36
	0.005	8.16	5.52	4.48	3.90	3.53	3.27	3.07	2.92	2.79	2.69	2.53	2.36	2.17	1.97	1.78	1.41
	0.001	11.34	7.29	5.75	4.92	4.39	4.02	3.74	3.53	3.36	3.21	2.99	2.76	2.51	2.24	1.99	1.52
∞	0.1	2.71	2.30	2.08	1.94	1.85	1.77	1.72	1.67	1.63	1.60	1.55	1.49	1.42	1.34	1.26	1.00
	0.05	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.46	1.35	1.00
	0.025	5.02	3.69	3.12	2.79	2.57	2.41	2.29	2.19	2.11	2.05	1.94	1.83	1.71	1.57	1.43	1.00
	0.01	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32	2.18	2.04	1.88	1.70	1.52	1.00
	0.005	7.88	5.30	4.28	3.72	3.35	3.09	2.90	2.74	2.62	2.52	2.36	2.19	2.00	1.79	1.59	1.00
	0.001	10.83	6.91	5.42	4.62	4.10	3.74	3.47	3.27	3.10	2.96	2.74	2.51	2.27	1.99	1.73	1.00

ΠΑΡΑΡΤΗΜΑ Β΄

ΧΡΗΣΙΜΕΣ ΜΑΘΗΜΑΤΙΚΕΣ ΓΝΩΣΕΙΣ

Σύνοψη

Στο παράρτημα αυτό γίνεται παράθεση χρήσιμων μαθηματικών γνώσεων που είναι απαραίτητες στη μελέτη των Κεφαλαίων του βιβλίου.
Προαπαιτούμενη γνώση: Βασικές μαθηματικές γνώσεις.
Προσδοκώμενα μαθησιακά αποτελέσματα: Όταν θα έχετε ολοκληρώσει τη μελέτη του κεφαλαίου αυτού θα μπορείτε να κατανοήσετε καλύτερα κάποια αποτελέσματα των κυρίως κεφαλαίων αυτού του συγγράμματος.

Γλωσσάριο επιστημονικών όρων

- Αριθμητική πρόοδος
- Βήτα συνάρτηση
- Γάμμα συνάρτηση
- Γεωμετρική πρόοδος

Β΄.1 Αριθμητική και Γεωμετρική πρόοδος

Ορισμός Β΄.1

Μια ακολουθία ονομάζεται **αριθμητική πρόοδος**, αν η διαφορά δύο οποιωνδήποτε διαδοχικών όρων της είναι σταθερός αριθμός, δηλαδή ισχύει ότι $a_{n+1} - a_n = \omega$. Αν $\omega > 0$, λέμε ότι η αριθμητική πρόοδος είναι γνησίως αύξουσα, ενώ, αν $\omega < 0$, λέμε ότι η αριθμητική πρόοδος είναι γνησίως φθίνουσα. Αντίστροφα, αποδεικνύεται ότι, αν η οποιαδήποτε διαφορά δύο διαδοχικών όρων μιας ακολουθίας είναι συγκεκριμένη, τότε αυτή η ακολουθία είναι αριθμητική πρόοδος.

Έστω μια αριθμητική πρόοδος με πρώτο όρο το a_1. Τότε ισχύει

$$\sum_{i=1}^{n} a_i = \frac{n(a_1 + a_n)}{2}. \tag{Β΄.1}$$

Ορισμός Β΄.2

Μια ακολουθία ονομάζεται **γεωμετρική πρόοδος** αν κανένας όρος της δεν ισούται με το μηδέν και κάθε πηλίκο διαδοχικών όρων της είναι ίσο με μια μη μηδενική σταθερή ποσότητα, δηλαδή ισχύει ότι $\frac{a_{n+1}}{a_n} = \lambda \neq 0$. Αντίστροφα, αποδεικνύεται ότι, αν το οποιοδήποτε πηλίκο δύο διαδοχικών όρων μιας ακολουθίας είναι συγκεκριμένο, τότε αυτή η ακολουθία είναι γεωμετρική πρόοδος.

Έστω μια γεωμετρική πρόοδος με πρώτο όρο a_1 και λόγο λ. Τότε, προφανώς $a_2 = \lambda a_1$, $a_3 = \lambda a_2 = \lambda^2 a_1$ και με παρόμοιο σκεπτικό ισχύει η αναδρομική σχέση $a_r = \lambda^{r-1} a_1$. Αν το $\lambda \neq 1$, τότε το άθροισμα των k πρώτων όρων της γεωμετρικής προόδου με πρώτο όρο το a_1 ισούται με:

$$\sum_{r=1}^{k} a_r = \sum_{r=1}^{k} \left(\lambda^{r-1} a_1\right) = a_1 \frac{\lambda^k - 1}{\lambda - 1}, \tag{Β΄.2}$$

ενώ, αν η πρόοδος είναι φθίνουσα, δηλαδή αν $|\lambda| < 1$, τότε

$$\sum_{r=1}^{+\infty} a_r = \sum_{r=1}^{+\infty} \lambda^{r-1} a_1 = a_1 \frac{1}{1 - \lambda}. \tag{Β΄.3}$$

Παραγωγίζοντας τη σχέση (Β΄.3) μία φορά ως προς λ προκύπτει:

$$\sum_{r=2}^{+\infty} (r-1)\lambda^{r-2} a_1 = a_1 \frac{1}{(1-\lambda)^2}, \tag{Β΄.4}$$

ενώ από αυτήν, παραγωγίζοντας ως προς λ, έχουμε:

$$\sum_{r=3}^{+\infty} (r-1)(r-2)\lambda^{r-3} a_1 = a_1 \frac{2}{(1-\lambda)^3}. \tag{Β΄.5}$$

Β΄.2 Η Γάμμα και η Βήτα συνάρτηση

Στα μαθηματικά η Γάμμα συνάρτηση, η οποία συμβολίζεται με το γράμμα Γ του ελληνικού αλφάβητου, είναι μια συχνά χρησιμοποιούμενη συνάρτηση, που αποτελεί επέκταση της συνάρτησης του παραγοντικού στους μιγαδικούς αριθμούς, καθώς ορίζεται για κάθε μιγαδικό αριθμό με θετικό πραγματικό μέρος. Στην ενότητα αυτή, θα δοθούν ο ορισμός της Γάμμα συνάρτησης και κάποιες βασικές ιδιότητές της, αποφεύγοντας περισσότερες λεπτομέρειες.

Ορισμός Β΄.3

Η **Γάμμα συνάρτηση** ορίζεται στο πεδίο $\{a : Re(a) > 0\}$, όπου $Re(a)$ συμβολίζει το πραγματικό μέρος του αριθμού a και δίνεται από το γενικευμένο ολοκλήρωμα (γνωστό ως ολοκλήρωμα Euler δευτέρου είδους):

$$\Gamma(a) = \int_0^\infty t^{a-1}e^{-t}dt, \tag{Β΄.6}$$

ή, ισοδύναμα,

$$\Gamma(a) = 2\int_0^\infty e^{-t^2}t^{2a-1}dt. \tag{Β΄.7}$$

Από τη σχέση (Β΄.6) προκύπτει ότι $\Gamma(1) = 1$. Επιπρόσθετα, αποδεικνύεται ότι η Γάμμα συνάρτηση ικανοποιεί τη συναρτησιακή σχέση:

$$\Gamma(a+1) = a\Gamma(a), a > 0. \tag{Β΄.8}$$

Με συνδυασμό των παραπάνω αποτελεσμάτων προκύπτει ότι:

$$\Gamma(n+1) = n!, n \in \mathbb{N}. \tag{Β΄.9}$$

Λόγω της σχέσης (Β΄.9) η Γάμμα συνάρτηση θεωρείται επέκταση, γενίκευση του παραγοντικού. Για άλλες χρήσιμες σχέσεις για την Γάμμα συνάρτηση παραπέμπουμε τον/την αναγνώστη/στρια στον ιστότοπο https://mathworld.wolfram.com/GammaFunction.html (ημερομηνία προσπέλασης: 1/3/2022).

Στα μαθηματικά η Βήτα συνάρτηση, η οποία συμβολίζεται με το γράμμα B του ελληνικού αλφάβητου, είναι μια συχνά χρησιμοποιούμενη συνάρτηση, που είναι στενά συνδεδεμένη με τη Γάμμα συνάρτηση και τους διωνυμικούς συντελεστές. Στη συνέχεια, δίνεται ο ορισμός της Βήτα συνάρτησης.

Ορισμός Β΄.4

Η **Βήτα συνάρτηση** δίνεται από το ακόλουθο ολοκλήρωμα (γνωστό ως ολοκλήρωμα Euler πρώτου είδους):

$$B(x,y) = B(y,x) = \int_0^1 t^{x-1}(1-t)^{y-1}dt, \tag{Β΄.10}$$

για μιγαδικούς αριθμούς x, y τέτοιους, ώστε $Re(x) > 0$ και $Re(y) > 0$.

Από τη σχέση (Β΄.10) προκύπτουν, μετά από αλγεβρικές πράξεις, οι ακόλουθες χρήσιμες ιδιότητες:

$$B(x,y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}, \tag{Β΄.11}$$

ενώ, όταν x και y είναι θετικοί ακέραιοι,

$$B(x,y) = \frac{(x-1)!(y-1)!}{(x+y-1)!}. \tag{Β΄.12}$$

Για άλλες χρήσιμες σχέσεις παραπέμπουμε τον/την αναγνώστη/στρια στον ιστότοπο https://mathworld.wolfram.com/BetaFunction.html (ημερομηνία προσπέλασης: 1/3/2022).

Πρόταση Β΄.1

Ισχύει ότι:

$$\int_{-\infty}^{+\infty} \exp(-z^2)dz = \Gamma(0.5) = \sqrt{\pi}. \tag{Β΄.13}$$

Απόδειξη Θεωρήματος

Καθώς η συνάρτηση $\exp\left(-z^2\right)$ είναι άρτια, προκύπτει ότι:

$$\int_{-\infty}^{+\infty} \exp\left(-z^2\right) = 2\int_{0}^{+\infty} \exp\left(-z^2\right) dz$$

Κάνοντας την αλλαγή μεταβλητών $z = \sqrt{t}$, οπότε και $dz = 0.5t^{-0.5}dt$, έχουμε ότι:

$$2\int_{0}^{+\infty} \exp\left(-z^2\right) dz = \int_{0}^{+\infty} t^{0.5-1} \exp\left(-t\right) dt = \Gamma(0.5).$$

Μένει, επομένως να δείξουμε ότι $\Gamma(0.5) = \sqrt{\pi}$. Υπάρχουν διάφοροι τρόποι για την απόδειξη αυτού του αποτελέσματος. Ένας από αυτούς βασίζεται στη σχέση (Β΄.11). Ειδικότερα, για $x = 0.5$ και $y = 0.5$ προκύπτει ότι: $\Gamma(0.5) = \sqrt{B(0.5, 0.5)}$, άρα αρκεί να δείξουμε ότι $B(0.5, 0.5) = \pi$. Εξ ορισμού είναι, μετά από λίγη άλγεβρα:

$$B(0.5, 0.5) = \int_{0}^{1} \frac{1}{\sqrt{t(1-t)}} dt.$$

Θέτοντας $t = sin^2\theta$, είναι

$$\begin{aligned} \mathrm{B}(0.5, 0.5) &= \int_{0}^{\pi/2} \frac{2\sin\theta\cos\theta}{\sin\theta\sqrt{1-\sin^2\theta}} d\theta \\ &= \int_{0}^{\pi/2} \frac{2\sin\theta\cos\theta}{\sin\theta\cos\theta} d\theta \\ &= \int_{0}^{\pi/2} 2\theta d\theta = \pi \end{aligned}$$

και η απόδειξη ολοκληρώθηκε.

Η Γάμμα συνάρτηση, $\Gamma(a)$, γενικεύεται στην άνω (upper) και κάτω (lower) ελλιπή Γάμμα συνάρτηση (incomplete gamma function). Στα μαθηματικά αυτές οι συναρτήσεις είναι μορφές ειδικών συναρτήσεων, οι οποίες προκύπτουν στην επίλυση διάφορων μαθηματικών προβλημάτων ως συγκεκριμένα ολοκληρώματα. Ειδικότερα, η άνω ελλιπής (ατελής) Γάμμα συνάρτηση ορίζεται ως εξής:

Ορισμός Β΄.5

Η **άνω ελλιπής Γάμμα συνάρτηση** ορίζεται ως:

$$\Gamma(a, x) = \int_{x}^{\infty} t^{a-1}e^{-t}dt, \qquad \text{(Β΄.14)}$$

όπου $\{a : Re(a) > 0\}$, με $Re(a)$ να συμβολίζει το πραγματικό μέρος του αριθμού a. Επιπλέον, αν a είναι οποιοσδήποτε ακέραιος, έστω n, τότε ισχύει ότι:

$$\Gamma(n, x) = (n-1)!e^{-x}\sum_{k=0}^{n-1}\frac{x^k}{k!}. \qquad \text{(Β΄.15)}$$

Προφανώς $\Gamma(a, 0) = \Gamma(a)$. Στη συνέχεια, παραθέτουμε τον ορισμό της κάτω ελλιπούς Γάμμα συνάρτησης.

Ορισμός Β΄.6

Η **κάτω ελλιπής Γάμμα συνάρτηση** ορίζεται ως:

$$\gamma(a,x) = \int_0^x t^{a-1}e^{-t}dt, \tag{Β΄.16}$$

όπου $\{a : Re(a) > 0\}$, με $Re(a)$ να συμβολίζει το πραγματικό μέρος του αριθμού a και $|arg(x) < \pi|$. Επιπλέον, αν a είναι οποιοσδήποτε ακέραιος, έστω n, τότε ισχύει ότι:

$$\gamma(n,x) = (n-1)!\left(1 - e^{-x}\sum_{k=0}^{n-1}\frac{x^k}{k!}\right). \tag{Β΄.17}$$

Παρατήρηση Β΄.1

Στη γλώσσα προγραμματισμού R η συνάρτηση gamma(x) δίνει τη $\Gamma(x)$, ενώ μέσω του πακέτου pracma και των συναρτήσεων του `gammainc(x,a)` και `incgam(x,a)` είναι δυνατός ο υπολογισμός των $\gamma(a,x)$ και $\Gamma(a,x)$, αντίστοιχα. Εναλλακτικά, κάποιος θα μπορούσε να χρησιμοποιήσει και την ασκ της Γάμμα κατανομής και να υπολογίσει τις ατελείς Γάμμα συναρτήσεις ως εξής:

$$\gamma(a,x) = \texttt{pgamma(x, a)} * \texttt{gamma}(a)$$

και

$$\Gamma(a,x) = \texttt{pgamma(x, a, lower = FALSE)} * \texttt{gamma(a)}.$$

Όμοια, η μη πλήρης ή ελλιπής Βήτα συνάρτηση αποτελεί γενίκευση της Βήτα συνάρτησης.

Ορισμός Β΄.7

Η **μη πλήρης ή ελλιπής Βήτα συνάρτηση** αποτελεί γενίκευση της Βήτα συνάρτησης και ορίζεται ως:

$$B(x;a,b) = \int_0^x t^{a-1}(1-t)^{b-1}dt, \tag{Β΄.18}$$

ενώ η **κανονικοποιημένη ελλιπής Βήτα συνάρτηση** (regularized incomplete beta function) ορίζεται ως:

$$I_x(a,b) = \frac{B(x;a,b)}{B(a,b)}. \tag{Β΄.19}$$

Για $x = 1$, η ελλιπής Βήτα συνάρτηση ταυτίζεται με τη Βήτα συνάρτηση $B(a,b)$.

Παρατήρηση Β΄.2

Στη γλώσσα προγραμματισμού R η συνάρτηση `beta(a,b)` δίνει τη $B(a,b)$, ενώ γράφοντας `gamma(a) * gamma(b) * pbeta(x,a,b)` θα έχουμε τη $B(x;a,b)$.

ΕΥΡΕΤΗΡΙΟ

Τις τελευταίες δεκαετίες αυξάνεται συνεχώς η αναγνώριση του κυρίαρχου ρόλου της πιθανοθεωρίας και των στατιστικών μεθόδων στην επίλυση προβλημάτων σε διάφορα επιστημονικά πεδία. Για τον λόγο αυτόν, οι Πιθανότητες και η Στατιστική διδάσκονται σε πληθώρα τμημάτων της τριτοβάθμιας εκπαίδευσης, πέραν των Τμημάτων Μαθηματικών και Στατιστικής. Βασικός στόχος αυτού του βιβλίου είναι να εφοδιάσει αυτούς που διδάσκονται τα αντικείμενα αυτά για πρώτη φορά με όλο το απαραίτητο θεωρητικό υπόβαθρο που θα τους επιτρέπει να επιλέγουν και να εφαρμόζουν την κατάλληλη μεθοδολογία για την επίλυση προβλημάτων της επιστημονικής περιοχής τους. Στο παραπάνω πλαίσιο, το παρόν βιβλίο μπορεί να χρησιμοποιηθεί ως ένα εισαγωγικό σύγγραμμα σε μαθήματα Πιθανοτήτων ή Στατιστικής ή Πιθανοτήτων-Στατιστικής. Ειδικότερα, για τη διδασκαλία ενός εισαγωγικού μαθήματος Πιθανοτήτων κρίνεται επαρκής η ύλη των Κεφαλαίων 1-7 με πλήρη έμφαση στις αποδείξεις και στο θεωρητικό υπόβαθρο, ενώ για τη διδασκαλία ενός εισαγωγικού μαθήματος στη Στατιστική θεωρείται επαρκής η ύλη των Κεφαλαίων 8-13 με πλήρη έμφαση στα θεωρητικά αποτελέσματα. Τέλος, για τη διδασκαλία ενός εισαγωγικού μαθήματος Πιθανοτήτων και Στατιστικής ή για τη διδασκαλία του μαθήματος σε Τμήματα άλλα, πέραν των Μαθηματικών και Στατιστικής, προτείνεται η κάλυψη της ύλης των Κεφαλαίων 1-6 και 8-13, χωρίς να δίνεται έμφαση στις μαθηματικές αποδείξεις.

Το παρόν σύγγραμμα δημιουργήθηκε στο πλαίσιο του Έργου ΚΑΛΛΙΠΟΣ+	
Χρηματοδότης	Υπουργείο Παιδείας και Θρησκευμάτων, Προγράμματα ΠΔΕ, ΕΠΑ 2020-2025
Φορέας υλοποίησης	ΕΛΚΕ ΕΜΠ
Φορέας λειτουργίας	ΣΕΑΒ/Παράρτημα ΕΜΠ/Μονάδα Εκδόσεων
Διάρκεια 2ης Φάσης	2020-2023
Σκοπός	Η δημιουργία ακαδημαϊκών ψηφιακών συγγραμμάτων ανοικτής πρόσβασης (περισσότερων από 700) • Προπτυχιακών και μεταπτυχιακών εγχειριδίων • Μονογραφιών • Μεταφράσεων ανοικτών textbooks • Βιβλιογραφικών Οδηγών
Επιστημονικά Υπεύθυνος	Νικόλαος Μήτρου, Καθηγητής ΣΗΜΜΥ ΕΜΠ
ISBN: 9786185667856	**DOI:** http://dx.doi.org/10.57713/kallipos101

Το παρόν σύγγραμμα χρηματοδοτήθηκε από το Πρόγραμμα Δημοσίων Επενδύσεων του Υπουργείου Παιδείας