Π. Ι. Νικήτας

ΕΙΣΑΓΩΓΗ ΣΤΗ ΣΤΑΤΙΣΤΙΚΗ ΑΝΑΛΥΣΗ ΠΕΙΡΑΜΑΤΙΚΩΝ ΔΕΔΟΜΕΝΩΝ

ΜΕ ΧΡΗΣΗ EXCEL ΚΑΙ SPSS

Θεσσαλονίκη 2013

Εκδόσεις ΣΙΜΩΝΗ
Όλγα Σιμώνη
Αρμενοπούλου 12, 54635, Θεσσαλονίκη
Τηλ./fax: 2310 246398
e-mail: 2000piga@gmail.com

Απαγορεύεται η μερική ή ολική έκδοση του παρόντος, επίσης και η με οποιονδήποτε τρόπο αναπαραγωγή του, καθώς και η φωτοτύπιση τμήματος ή ολόκληρου του βιβλίου, χωρίς την έγγραφη άδεια του εκδότη. Οι παραβάτες θα τιμωρούνται από το νόμο.

copyright: Παναγιώτης Νικήτας

ISBN: 978-960-98388-1

ΠΡΟΛΟΓΟΣ

Η ανάλυση, αξιολόγηση και παρουσίαση πειραματικών δεδομένων αποτελούν καθοριστικά στάδια της πειραματικής διαδικασίας, όταν το πείραμα σχετίζεται με ποσοτικές κυρίως μετρήσεις. Βασικό εργαλείο για το σκοπό αυτό είναι η *Στατιστική*. Ο κλάδος αυτός των Μαθηματικών έχει αναπτύξει μεθόδους για τη συνοπτική και αποτελεσματική παρουσίαση δεδομένων και επιπλέον τεχνικές για συσχέτιση, πρόβλεψη και για βελτίωση της ποιότητας των δεδομένων σχεδιάζοντας πειράματα. Επειδή τα δεδομένα μπορεί να είναι οποιουδήποτε τύπου, η Στατιστική μπορεί να εφαρμοστεί σχεδόν παντού: Στις φυσικές, βιολογικές, κοινωνικές και οικονομικές επιστήμες και στις επιχειρήσεις.

Κύριος στόχος αυτού του βοηθήματος είναι να δώσει στους φοιτητές τις βασικές γνώσεις που απαιτούνται για να γίνουν κατανοητές οι τεχνικές της Στατιστικής. Επιπλέον επιχειρείται να παρουσιαστεί μια σύγχρονη εισαγωγή στη Στατιστική σε πανεπιστημιακό επίπεδο, χωρίς την επιβάρυνση θεωρητικών αποδείξεων, αλλά με την εκτεταμένη εφαρμογή έτοιμων στατιστικών πακέτων. Για το σκοπό αυτό χρησιμοποιείται το *Excel* σε συνδυασμό με το πρόσθετο *ChemStat* και το *SPSS*.

Η επιλογή αυτή έγινε επειδή το *Excel* είναι ένα πρόγραμμα με μοναδικές δυνατότητες στην επεξεργασία και παρουσίαση πειραματικών δεδομένων, αλλά με περιορισμένες τεχνικές στατιστικών αναλύσεων. Την αδυναμία αυτή του *Excel* συμπληρώνει σε μεγάλο βαθμό το *ChemStat*. Το *ChemStat* είναι ένα αρχείο του *Excel* που περιέχει προγράμματα-μακροεντολές για βασικά στατιστικά προβλήματα που αφορούν τη χημεία. Επιπλέον έχει τη δυνατότητα άμεσης μετατροπής σε πρόσθετο (add-in) με αποτέλεσμα να μπορεί να ενσωματωθεί στη λειτουργία του *Excel*. Το αρχείο αυτό είναι ελεύθερα διαθέσιμο στο διαδίκτυο και περιγράφεται στο *Παράρτημα ΙΙΙ*. Σε ό,τι αφορά το *SPSS*, πρόκειται για ένα από τα πιο ευρέως διαδεδομένα στατιστικά προγράμματα για εφαρμογές της στατιστικής τόσο σε προπτυχιακό όσο και σε μεταπτυχιακό και ερευνητικό επίπεδο. Στις εφαρμογές που περιγράφονται χρησιμοποιείται η έκδοση *IBM Statistics 19*.

Θεσσαλονίκη 2013

ΠΕΡΙΕΧΟΜΕΝΑ

5. ΕΛΕΓΧΟΙ ΣΤΑΤΙΣΤΙΚΩΝ ΥΠΟΘΕΣΕΩΝ

6. ΕΛΕΓΧΟΙ ΣΤΑΤΙΣΤΙΚΩΝ ΥΠΟΘΕΣΕΩΝ ΣΕ ΕΝΑ ΔΕΙΓΜΑ

7. ΕΛΕΓΧΟΙ ΣΤΑΤΙΣΤΙΚΩΝ ΥΠΟΘΕΣΕΩΝ ΣΕ ΔΥΟ ΔΕΙΓΜΑΤΑ

8. ANOVA: ΑΝΑΛΥΣΗ ΔΙΑΣΠΟΡΑΣ

9. ΕΛΕΓΧΟΙ ΣΕ ΚΑΤΗΓΟΡΙΚΑ ΔΕΔΟΜΕΝΑ

10. ΠΡΟΣΑΡΜΟΓΗ ΚΑΙ ΣΥΣΧΕΤΙΣΗ

11. ΕΞΟΜΑΛΥΝΣΗ, ΠΑΡΑΓΩΓΙΣΗ, ΟΛΟΚΛΗΡΩΣΗ ΔΕΔΟΜΕΝΩΝ

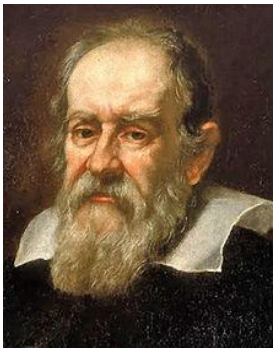
12. ΑΝΑΛΥΣΗ ΠΟΛΛΩΝ ΜΕΤΑΒΛΗΤΩΝ

Κεφάλαιο 1

ΠΕΙΡΑΜΑ ΚΑΙ ΑΒΕΒΑΙΟΤΗΤΑ

1.1 ΤΟ ΠΕΙΡΑΜΑ

Το πείραμα αποτελεί ουσιαστικό στοιχείο της *Επιστημονικής Μεθόδου*, δηλαδή της μεθόδου που πρέπει να ακολουθείται για την απόκτηση νέας γνώσης ή για τον έλεγχο ήδη αποκτηθείσας γνώσης. Εισηγητής του πειράματος ως απαραίτητου εργαλείου στην απόκτηση αξιόπιστης, δηλαδή επιστημονικής γνώσης, είναι ο *Γαλιλαίος*.



Galileo Galilei
(1564-1642)

Για να είναι έγκυρη η επιστημονική γνώση που προέρχεται από ένα πείραμα θα πρέπει:

α) Το πείραμα να μπορεί να επαναληφθεί από τον οποιοδήποτε ερευνητή.

β) Όταν εκτελείται κάτω από τις ίδιες ακριβώς συνθήκες να δίνει το ίδιο αποτέλεσμα.

Όμως, για λόγους που θα εξετάσουμε παρακάτω, η δεύτερη προϋπόθεση πρέπει να εξετάζεται με προσοχή και ευρύτητα.

1.1.1 ΠΕΙΡΑΜΑ ΚΑΙ ΑΒΕΒΑΙΟΤΗΤΑ

Τα πειράματα μπορεί να χωριστούν σε δύο κατηγορίες:

- ***Ποιοτικά*** – τα αποτελέσματα του πειράματος δεν εκφράζονται με αριθμούς.
- ***Ποσοτικά*** – τα αποτελέσματα του πειράματος παρουσιάζονται με αριθμούς.

Για παράδειγμα, στην κατηγορία των ποιοτικών πειραμάτων είναι όλες οι συνθέσεις οργανικών ή ανόργανων ενώσεων καθώς επίσης και τα πειράματα της Ποιοτικής Αναλυτικής Χημείας. Ποσοτικό είναι κάθε πείραμα που περιλαμβάνει μία ή περισσότερες μετρήσεις. Τυπικά ποσοτικά πειράματα είναι ο προσδιορισμός της συγκέντρωσης μιας ένωσης σε ένα σύστημα, η μέτρηση του pH ενός διαλύματος, ο προσδιορισμός της ειδικής ταχύτητας μιας αντίδρασης, κ.ο.κ.

Τα ποσοτικά πειράματα με τη σειρά τους χωρίζονται σε δύο υποκατηγορίες:

- ***Στοχαστικά*** και ***Αιτιοκρατικά***

Στα στοχαστικά πειράματα το αποτέλεσμα μιας μέτρησης δεν επαναλαμβάνεται. Παραδείγματα τέτοιων πειραμάτων είναι ο χρόνος διάσπασης ενός ραδιενεργού πυρήνα, ο χρόνος εξάχνωσης του νήματος μιας λυχνίας, η κίνηση Brown, κ.ά. Αντίθετα, σε ένα αιτιοκρατικό πείραμα το αποτέλεσμα εξαρτάται επακριβώς από τις συνθήκες του πειράματος. Αν προσδιορίζουμε το pH ενός διαλύματος ισχυρού οξέος με συγκέντρωση 0.001 M θα πρέπει να πάρουμε την τιμή 3. Όμως και σε αυτή την περίπτωση μπορούμε να δούμε στοιχεία αβεβαιότητας και μη επαναληψιμότητας των μετρήσεων. Έτσι, στο παράδειγμα του pH αν κάνουμε 5 μετρήσεις είναι μάλλον απίθανο να πάρουμε 5 φορές την τιμή 3. Το πιθανότερο είναι να πάρουμε αποτελέσματα που διαφέρουν στο δεύτερο δεκαδικό ψηφίο, όπως για παράδειγμα τα αποτελέσματα:

2.99 2.98 3.02 3.00 2.97

Συνεπώς, παρά το γεγονός ότι η συγκέντρωση των υδρογονοκατιόντων είναι σταθερή, λόγω τυχαίων σφαλμάτων που υπεισέρχονται στις μετρήσεις, δεν παίρνουμε μία τιμή pH.

Η αβεβαιότητα στις μετρήσεις του pH αφορά όλες σχεδόν τις πειραματικές μετρήσεις, με εξαίρεση ίσως τις μετρήσεις που σχετίζονται με καταμέτρηση γεγονότων. Συνεπώς, με την παραπάνω εξαίρεση, η αβεβαιότητα είναι *εγγενές χαρακτηριστικό* της πειραματικής μέτρησης. Θα πρέπει να διευκρινίσουμε ότι η αβεβαιότητα μπορεί να αφορά το τελευταίο ή τα τελευταία ψηφία της μέτρησης, μπορεί όμως να αφορά την άγνοιά μας για μεγαλύτερη ακρίβεια. Για παράδειγμα, έστω ότι ζυγίζουμε ένα βάρος με αναλυτικό ζυγό και παίρνουμε το αποτέλεσμα 1.2374 g. Ακόμη κι αν είμαστε σίγουροι για τα τέσσερα δεκαδικά ψηφία, σίγουρα αγνοούμε τι υπάρχει πέρα από αυτά.

Η πειραματική αβεβαιότητα ονομάζεται και *πειραματικό σφάλμα*. Θα πρέπει να γίνει συνείδηση στο νέο επιστήμονα και να θεωρεί εκ των ων ουκ άνευ ότι όταν έχει να παρουσιάσει τα αποτελέσματα των μετρήσεών του, δύο πράγματα δεν μπορεί να παραλείπει:

- Τις μονάδες μέτρησης, με την προϋπόθεση ότι υπάρχουν, δεδομένου ότι υπάρχουν φυσικές ποσότητες χωρίς μονάδες, π.χ. το pH, το μοριακό κλάσμα, η κατ' όγκο σύσταση, κ.ά.
- Το πειραματικό σφάλμα, το οποίο καθορίζει την ακρίβεια των μετρήσεων.

Χωρίς αυτές τις δύο πληροφορίες **ΚΑΜΙΑ** μέτρηση δεν έχει την παραμικρή αξία.

Παρατήρηση. Είναι προφανές ότι η αβεβαιότητα στην τιμή μιας φυσικής ποσότητας, όταν αυτή προσδιορίζεται κάτω από αυστηρά ελεγχόμενες συνθήκες, εξαρτάται από την ακρίβεια που θέλουμε να έχουν οι μετρήσεις. Αν για παράδειγμα είμαστε ικανοποιημένοι με ένα δεκαδικό ψηφίο, τότε όλες οι μετρήσεις του pH στο διάλυμα του ισχυρού οξέος με συγκέντρωση 0.001 M θα είναι ίσες με 3.0, με την προϋπόθεση ότι το πεχάμετρο έχει ακρίβεια τουλάχιστον δύο δεκαδικών ψηφίων.

1.1.2 ΤΥΧΑΙΕΣ ΜΕΤΑΒΛΗΤΕΣ

Εμπειρικά ονομάζουμε *τυχαία μεταβλητή* (τ.μ.) κάθε μεταβλητή της οποίας η τιμή σε μία μέτρηση δεν μπορεί να προβλεφθεί. Στη *Στατιστική* μια τυχαία μεταβλητή συμβολίζεται με ένα κεφαλαίο γράμμα, π.χ. Χ ή Υ, ενώ με μικρά γράμματα, x, y, συμβολίζουμε τις τιμές που παίρνουν οι μεταβλητές Χ και Υ. Στις θετικές επιστήμες τυχαίες μεταβλητές είναι οι ποσότητες που προσδιορίζουμε πειραματικά, όπως π.χ. το βάρος Β, η

συγκέντρωση c, το pH κ.ά. Όμως για λόγους απλότητας δε διαφοροποιούμε το σύμβολο μιας φυσικής μεταβλητής από την τιμή της. Έτσι το pH συμβολίζει και τον αρνητικό δεκαδικό λογάριθμο της συγκέντρωσης των υδρογονοκατιόντων και την τιμή που παίρνει η ποσότητα αυτή σε ένα διάλυμα.

1.1.3 ΠΕΙΡΑΜΑΤΙΚΑ ΣΦΑΛΜΑΤΑ

Τα πειραματικά σφάλματα μπορεί να χωριστούν σε δύο κατηγορίες: Σε ***συστηματικά*** (*systematic*) και σε ***τυχαία*** (*random*) σφάλματα.

Τα συστηματικά σφάλματα οφείλονται κυρίως σε κακή λειτουργία των οργάνων μέτρησης, στην ίδια τη μέθοδο μέτρησης και στην άγνοιά μας για την ύπαρξη κάποιου παράγοντα που επιδρά στα αποτελέσματα των μετρήσεων. Για παράδειγμα, η ύπαρξη νερού σε ένα υγροσκοπικό άλας μπορεί να γίνει πηγή σφαλμάτων αν το άλας χρησιμοποιηθεί για την παρασκευή διαλυμάτων. Τα τυχαία σφάλματα οφείλονται κατά κανόνα στην περιορισμένη ευαισθησία των οργάνων μέτρησης, ενώ σε ιδιαίτερα ευαίσθητα όργανα, τυχαίες, μη ελεγχόμενες διακυμάνσεις της θερμοκρασίας, πίεσης ή του περιβαλλοντικού θορύβου ελαττώνουν την επαναληψιμότητα των μετρήσεων και συνεπώς αυξάνουν το τυχαίο πειραματικό σφάλμα.

Είναι προφανές ότι όταν σχεδιάζουμε και εκτελούμε ένα πείραμα πρέπει να έχουμε τους ακόλουθους στόχους: Πρώτον να εξαλείψουμε κάθε συστηματικό σφάλμα και δεύτερον να εκτιμήσουμε την επίδραση που έχουν τα τυχαία σφάλματα, τόσο πάνω στην ακρίβεια των μετρήσεων, όσο και στους υπολογισμούς που γίνονται με βάση τις μετρήσεις αυτές.

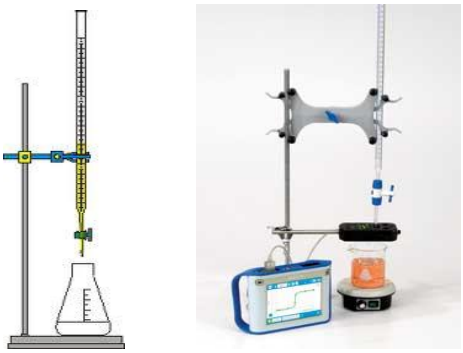
1.1.4 ΠΕΙΡΑΜΑΤΙΚΑ ΔΕΔΟΜΕΝΑ

Όταν εκτελούμε ένα ποσοτικό πείραμα κατά κανόνα μεταβάλουμε κατά βούληση την τιμή μιας ή περισσοτέρων φυσικών ποσοτήτων, ενώ ταυτόχρονα προσδιορίζουμε την τιμή μιας ή περισσοτέρων άλλων φυσικών ποσοτήτων. Οι φυσικές ποσότητες που μεταβάλλουμε κατά βούληση σε ένα πείραμα ονομάζονται ***ανεξάρτητες μεταβλητές*** (*independent variables*), ενώ οι φυσικές ποσότητες τις τιμές των οποίων προσδιορίζουμε πειραματικά ονομάζονται ***εξαρτημένες μεταβλητές*** (*dependent variables*).

Παράδειγμα 1.1

Στην πεχαμετρική τιτλοδότηση οξέος με βάση μεταβάλουμε τον όγκο

V της βάσης που προσθέτουμε με την προχοΐδα στο διάλυμα του οξέος και μετράμε το pH του διαλύματος. Δηλαδή μετά από κάθε προσθήκη ορισμένου όγκου βάσης, μετράμε με πεχάμετρο το pH του διαλύματος. Στο πείραμα αυτό ανεξάρτητη μεταβλητή είναι ο όγκος V της βάσης που προσθέτουμε και εξαρτημένη μεταβλητή είναι το pH του διαλύματος.



Σχήμα 1.1. Απλή (αριστερά) και πεχαμετρική (δεξιά) τιτλοδότηση

Παράδειγμα 1.2

Στην μελέτη της κινητικής της ιμβερτοποίησης του καλαμοσακχάρου

$$C_{12}H_{22}O_{11} + H_2O \quad \Leftrightarrow \quad C_6H_{12}O_6 + C_6H_{12}O_6$$

παρακολουθούμε τη μεταβολή της γωνίας α του πολωμένου φωτός με το χρόνο t. Συνεπώς μετράμε τη γωνία στροφής α σε τακτά χρονικά διαστήματα. Σε αυτές τις περιπτώσεις ο χρόνος t είναι πάντα η ανεξάρτητη μεταβλητή και συνεπώς η γωνία πόλωσης α θα είναι η εξαρτημένη μεταβλητή.

Σχήμα 1.2. Συσκευή πολωσιμετρίας

Παράδειγμα 1.3

Στην υγρή χρωματογραφία μπορεί να μεταβάλλεται η σύσταση του μίγματος των διαλυτών που περνούν από τη χρωματογραφική στήλη καθώς επίσης και η θερμοκρασία της στήλης. Έστω ότι σε μία μελέτη μεταβάλλουμε τη θερμοκρασία Τ της στήλης και την κατ' όγκο σύσταση φ ενός οργανικού διαλύτη. Σε κάθε ζεύγος (Τ, φ) προσδιορίζουμε τους χρόνους έκλουσης t_R τριών ενώσεων. Στο πείραμα αυτό έχουμε δύο ανεξάρτητες μεταβλητές, τις Τ και φ, και τρεις εξαρτημένες, τους χρόνους έκλουσης t_R των τριών ενώσεων.

Σχήμα 1.3. Διάταξη υγρής χρωματογραφίας

1.1.5 ΤΑΞΙΝΟΜΙΣΗ ΜΕΤΑΒΛΗΤΩΝ

Οι μεταβλητές ενός πειράματος μπορούν να ταξινομηθούν σε δύο μεγάλες κατηγορίες:

α) ***Ποσοτικές*** *(quantitative)*

β) ***Ποιοτικές*** ή ***κατηγορικές*** *(qualitative or categorical)*

Οι ποσοτικές μεταβλητές παίρνουν αριθμητικές τιμές και μπορούν να διακριθούν σε: ***Διακριτές*** *(discrete)* όταν παίρνουν μόνο ακέραιες τιμές. ***Συνεχείς*** *(continuous)* όταν παίρνουν τιμές από το σύνολο των πραγματικών αριθμών.

Οι ποιοτικές ή κατηγορικές μεταβλητές δεν εκφράζουν κάτι μετρήσιμο. Για παράδειγμα, η ποικιλία και ο τύπος ενός τροφίμου, η οσμή μιας ένωσης, ο τύπος μιας χρωματογραφικής στήλης είναι ποιοτικά ή κατηγορικά δεδομένα. Οι μεταβλητές αυτές διακρίνονται σε δύο υπο-

κατηγορίες:

α) ***Διατεταγμένες*** *(ordinal)*

β) ***Ονομαστικές*** *(nominal)*

Διατεταγμένη είναι μια ποιοτική μεταβλητή όταν οι κωδικοποιημένες τιμές που δίνουμε στη μεταβλητή αυτή καθορίζονται από μια διάταξη. Παράδειγμα, για την οσμή μιας ένωσης δημιουργούμε μια διατεταγμένη μεταβλητή που παίρνει τις εξής κωδικοποιημένες τιμές: 1 = άοσμη, 2 = ασθενής οσμή, 3 = έντονη οσμή. Ονομαστική είναι μια ποιοτική μεταβλητή όταν οι κωδικοποιημένες τιμές που δίνουμε κατηγοριοποιούν τα στοιχεία ενός συνόλου σε ομάδες. Παράδειγμα, 1 = άνδρας, 2 = γυναίκα ή Μ = άνδρας, F = γυναίκα.

1.2 ΠΑΡΟΥΣΙΑΣΗ ΠΕΙΡΑΜΑΤΙΚΩΝ ΔΕΔΟΜΕΝΩΝ

Δύο είναι οι βασικοί τρόποι παρουσίασης των πειραματικών δεδομένων και αποτελεσμάτων ενός ποσοτικού πειράματος. Με πίνακες ή/και με γραφικές παραστάσεις.

1.2.1 ΠΑΡΟΥΣΙΑΣΗ ΔΕΔΟΜΕΝΩΝ ΜΕ ΠΙΝΑΚΕΣ

Στους πίνακες τα δεδομένα διατάσσονται έτσι ώστε η κάθε στήλη (ή η κάθε γραμμή) να αντιστοιχεί σε μία μεταβλητή. Κατά κανόνα οι ανεξάρτητες μεταβλητές τοποθετούνται αριστερά, ενώ οι εξαρτημένες δεξιά. Στην πρώτη γραμμή κάθε στήλης τοποθετείται το σύμβολο της μεταβλητής και η μονάδα μέτρησης. Πάνω από τον πίνακα τοποθετείται λεζάντα με όλες τις σχετικές πληροφορίες που αφορούν τον πίνακα.

Πίνακας 1.1. Μεταβολή της πίεσης με τον όγκο 1 mole Cl_2 όταν T=300K.

V, dm^3	P, atm
7	3.3
10	2.4
15	1.6
20	1.2
25	0.9
30	0.7

Πίνακας 1.2. Μεταβολή της γωνίας στροφής α του πολωμένου φωτός με το χρόνο t κατά την ιμβερτοποίηση του καλαμοσακχάρου.

t, min	α, ^{0}
5	18.0
10	15.5
15	12.0
20	9.0
25	8.0
30	5.5
35	4.0
40	2.1
45	1.5
50	-0.2
55	-1.2
60	-2.5

Πίνακας 1.3. Χρόνοι έκλουσης t_R των ενώσεων Arg, Ser, Dopa σε συνάρτηση με τη θερμοκρασία Τ της στήλης και την κατ' όγκο σύσταση φ της μεθανόλης στο μίγμα των διαλυτών.

φ	Τ, ^{0}C	t_R, min		
		Arg	Ser	Dopa
0.25	15	3.22	5.15	13.16
0.30	15	2.31	3.68	7.24
0.40	15	1.73	2.28	3.03
0.25	35	2.75	4.25	9.34
0.30	35	2.13	3.20	5.63
0.40	35	1.64	2.11	2.64
0.25	75	2.23	3.25	5.86
0.30	75	1.83	2.55	3.87
0.40	75	1.54	1.87	2.20

Εναλλακτικά ο πίνακας αυτός μπορεί να γραφεί και ως

Πίνακας 1.4. Χρόνοι έκλουσης t_R σε min των ενώσεων Arg, Ser, Dopa σε συνάρτηση με τη θερμοκρασία Τ της στήλης και την κατ' όγκο σύσταση φ της μεθανόλης στο μίγμα των διαλυτών.

Τ, ^{0}C	15			35			75		
φ	0.25	0.30	0.40	0.25	0.30	0.40	0.25	0.30	0.40
Arg	3.22	2.31	1.73	2.75	2.13	1.64	2.23	1.83	1.54
Ser	5.15	3.68	2.28	4.25	3.20	2.11	3.25	2.55	1.87
Dopa	13.16	7.24	3.03	9.34	5.63	2.64	5.86	3.87	2.20

Παρατηρείστε ότι εδώ η μονάδα μέτρησης του χρόνου έκλουσης παρουσιάζεται στη λεζάντα του πίνακα και όχι στον ίδιο τον πίνακα.

Πίνακας 1.5. Μεταβολή της ποσότητας x/m του CH_3COOH που προσροφάται ανά γραμμάριο προσροφητικού από υδατικό διάλυμα σε συνάρτηση με τη συγκέντρωση c του CH_3COOH στο διάλυμα.

c, mol dm^{-3}	0.1	0.2	0.3	0.5	1.0
$(x/m)\times10^3$, mol g^{-1}	3.9	4.2	4.4	4.6	5.0

Στον πίνακα αυτόν το σύμβολο $(x/m)\times10^3$ σημαίνει ότι οι τιμές της μεταβλητής x/m έχουν πολλαπλασιαστεί επί 10^3. Άρα οι τιμές της δεύτερης γραμμής είναι 0.0039, 0.0042, …

1.2.2 ΠΑΡΟΥΣΙΑΣΗ ΜΕ ΓΡΑΦΙΚΕΣ ΠΑΡΑΣΤΑΣΕΙΣ

Οι γραφικές παραστάσεις πειραματικών δεδομένων έχουν ιδιαίτερο ενδιαφέρον στις θετικές επιστήμες, επειδή δίνουν εποπτικά τις ιδιότητες του φαινομένου από το οποίο προέρχονται τα πειραματικά δεδομένα. Υπάρχουν αρκετοί τύποι γραφικών παραστάσεων για την παρουσίαση πειραματικών δεδομένων. Από αυτούς ο βασικότερος τύπος είναι τα διαγράμματα διασποράς, ενώ ενδιαφέρον παρουσιάζουν και οι τύποι: ραβδογράμματα, ιστογράμματα και θηκογράμματα. Οι τρεις τελευταίοι τύποι αφορούν την παρουσίαση κυρίως στατιστικών δεδομένων και θα εξετασθούν στο επόμενο κεφάλαιο.

Στα *διαγράμματα διασποράς (scatter plots)* παρέχεται εποπτικά η

μεταβολή της εξαρτημένης μεταβλητής y ή των εξαρτημένων μεταβλητών y_1, y_2, … από την ανεξάρτητη μεταβλητή x. Οι πιο βασικοί κανόνες που χρησιμοποιούνται σε ένα διάγραμμα διασποράς είναι οι ακόλουθοι:

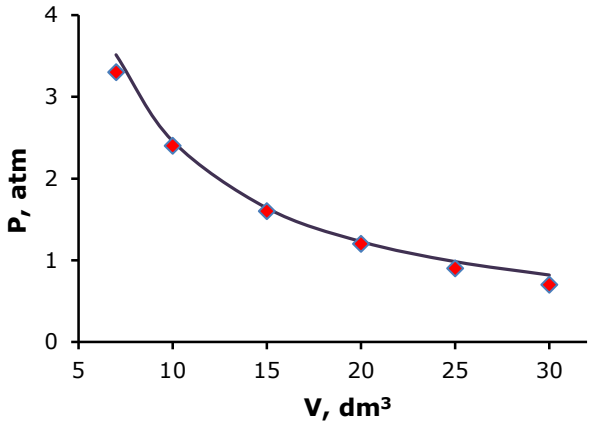
- Τα διαγράμματα διασποράς γίνονται πάντα σε χιλιοστομετρικό χαρτί, εκτός κι αν χρησιμοποιείται κατάλληλο πρόγραμμα γραφικών σε υπολογιστή. Τέτοια προγράμματα είναι το *Excel*, το *Origin*, αλλά και τα περισσότερα στατιστικά προγράμματα, όπως το *SPSS*.
- Στους άξονες των x και y σημειώνονται μόνο οι κλίμακες, ενώ δίπλα στους άξονες σημειώνεται η φυσική ποσότητα που παριστάνουν και η μονάδα μέτρησής της.
- Οι κλίμακες δεν είναι απαραίτητο να αρχίζουν από το μηδέν. Η γραφική παράσταση πρέπει να γεμίζει όλο το επίπεδο x - y και όχι ένα μικρό τμήμα του.
- Στα διαγράμματα διασποράς, τα πειραματικά δεδομένα εμφανίζονται με σημεία που μπορεί να συνδέονται ή να μη συνδέονται με γραμμές. Αντίθετα, αν έχουμε θεωρητικά δεδομένα, δηλαδή δεδομένα που προβλέπονται από κάποια εξίσωση, τότε αυτά σχεδιάζονται με συνεχή και ομαλή γραμμή.
- Τέλος, σε αντίθεση με τους πίνακες, η λεζάντα των γραφικών παραστάσεων τοποθετείται συνήθως κάτω από τη γραφική παράσταση και πρέπει να δίνει όλες τις απαραίτητες πληροφορίες.

Ακολουθούν παραδείγματα σωστών γραφικών παραστάσεων. Η σχεδίαση γραφικών παραστάσεων με το *Excel* περιγράφεται στο Παράρτημα Ι και με το *SPSS* στο Παράρτημα ΙΙ.

Παράδειγμα 1.4

Στο σχήμα 1.4 δίνεται η γραφική παράσταση των τιμών του παρακάτω πίνακα, όπου P(πειρ.) είναι οι πειραματικές τιμές της πίεσης ενός mole Cl_2 όταν T = 300K, ενώ οι τιμές P(υπολ.) έχουν υπολογιστεί με βάσει την καταστατική εξίσωση των ιδανικών αερίων, P = RT/V = 0.082·300/V.

V, dm^3	7	10	15	20	25	30
P(πειρ.), atm	3.3	2.4	1.6	1.2	0.9	0.7
P(υπολ.), atm	3.514	2.460	1.640	1.230	0.984	0.820



Σχήμα 1.4. Μεταβολή της πίεσης με τον όγκο 1 mol Cl_2 σε θερμοκρασία 300K. (♦) πειραματικά δεδομένα, (—) καμπύλη της ιδανικής συμπεριφοράς

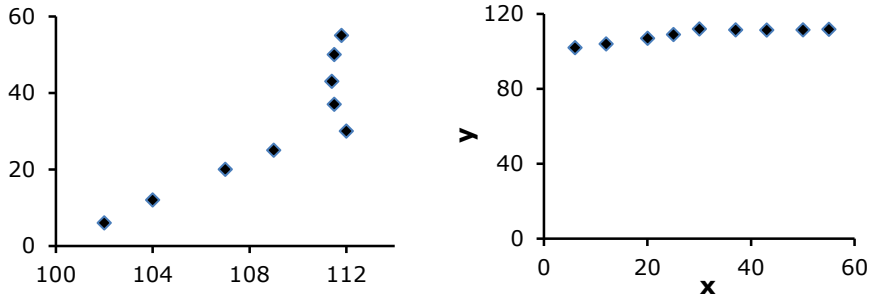
Παράδειγμα 1.5

Σε ένα πείραμα εξετάζεται η επίδραση της ποσότητας m σε g του KCl στον όγκο V ενός διαλύματός του. Συγκεκριμένα προστίθενται σε 100 mL νερού συγκεκριμένες ποσότητες KCl σε g. Το σύστημα αφήνεται να ισορροπήσει και ακολούθως το διάλυμα φιλτράρεται για να απομακρυνθεί το στερεό που υπάρχει και προσδιορίζεται ο όγκος του διαλύματος. Με τον τρόπο αυτό ελήφθησαν τα πειραματικά δεδομένα του παρακάτω πίνακα. Να γίνει η κατάλληλη γραφική παράσταση.

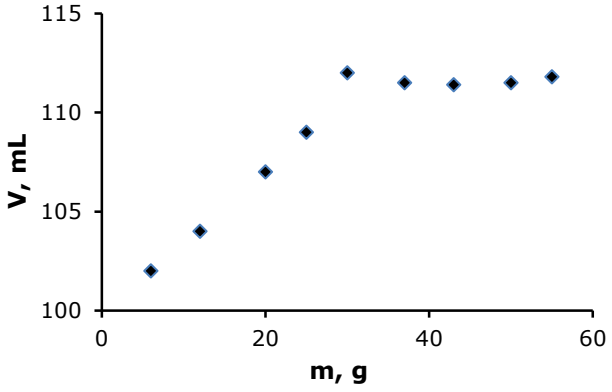
m, g	V, mL	m, g	V, mL
6	102.0	37	111.5
12	104.0	43	111.4
20	107.0	50	111.5
25	109.0	55	111.8
30	112.0		

Στο σχήμα 1.5 παρουσιάζονται δύο λανθασμένες γραφικές παραστάσεις. Στη γραφική παράσταση που είναι αριστερά οι άξονες έχουν τοποθετηθεί λανθασμένα και δεν υπάρχουν οι τίτλοι των αξόνων. Στην παράσταση δεξιά ο άξονας των y ξεκινά από το 0 με αποτέλεσμα τα πειραματικά σημεία να καταλαμβάνουν ένα μικρό μέρος της γραφικής

παράστασης. Αυτό έχει ως συνέπεια να μη φαίνονται οι λεπτομέρειες της μεταβολής του όγκου V με την ποσότητα του KCl. Επίσης δεν σημειώνεται η φυσική ποσότητα δίπλα από τους άξονες. Αντίθετα η σωστή γραφική παράσταση δίνεται στο σχήμα 1.6.



Σχήμα 1.5. Λανθασμένες γραφικές παραστάσεις μεταβολής του όγκου υδατικού διαλύματος KCl με την ποσότητα του προστιθέμενου KCl



Σχήμα 1.6. Μεταβολή του όγκου υδατικού διαλύματος KCl με την ποσότητα του προστιθέμενου KCl

ΑΣΚΗΣΕΙΣ

1.1. Να προσδιοριστεί ο τύπος των παρακάτω μεταβλητών:

Θερμοκρασία, Χρόνος, Συγκέντρωση, Πλήθος συμβάντων, Βάρος, Ημερήσια ελάχιστη θερμοκρασία μιας περιοχής, pH, Αριθμός δείγματος, Αριθμός μορίων

1.2. Να προσδιοριστεί ο τύπος των μεταβλητών στους παρακάτω πίνακες δεδομένων. Σε κάθε πίνακα ποια είναι η ανεξάρτητη και ποια η εξαρτημένη μεταβλητή;

Ποιότητα οίνου	3	5	1	2
Συγκέντρωση SO_2 σε ppm στον οίνο	2.7	4.5	0.8	1.9

Φύλο	φοιτητής	φοιτητής	φοιτήτρια	φοιτητής
Βαθμολογία	5	8	8	7

Χρόνος, μέρες	1	2	3	4
Απόδοση αντίδρασης, %	5	20	25	28

Καταλύτης	A	B	C	D
Συγκέντρωση προϊόντος, M	0.22	0.31	0.15	0.18

Δυναμικό απαριθμητή, V	700	750	800	850
Παλμοί απαριθμητή	10	880	3670	5850

1.3. Αν φ είναι η κατ' όγκο σύσταση ενός διαλύματος και d(πειρ.), d(υπολ.) η πειραματική και η υπολογιζόμενη θεωρητικά πυκνότητά του, να γίνουν οι γραφικές παραστάσεις των τιμών της πειραματικής και θεωρητικής πυκνότητας με το φ και οι επιμέρους γραφικές παραστάσεις d(πειρ.) - φ και d(υπολ.) - φ με βάση τις τιμές του παρακάτω πίνακα:

φ	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7
d(υπολ.), g/mL	0.996	0.986	0.973	0.957	0.937	0.915	0.89	0.863
d(πειρ.), g/mL	0.998	0.985	0.970	0.958	0.939	0.914	0.89	0.861

1.4. Να γίνει η γραφική παράσταση των τιμών του παρακάτω πίνακα:

x	y1	y2
1	1	0.0080
2	1.4	0.0070
3	1.7	0.0055
4	2.0	0.0045
5	2.2	0.0035
6	2.5	0.0030
7	2.6	0.0025
8	2.8	0.0020
9	3.0	0.0016
10	3.1	0.0013

1.5. Να γίνει η γραφική παράσταση των τιμών του παρακάτω πίνακα, όπου TSP είναι η συγκέντρωση αιωρούμενων σωματιδίων σε μg/m^3:

Ημερομηνία	TSP	Ημερομηνία	TSP	Ημερομηνία	TSP
17-NOV-2000	115	21-FEB-2001	170	09-JUN-2001	68
25-NOV-2000	27	11-MAR-2001	113	15-JUN-2001	53
03-DEC-2000	128	17-MAR-2001	130	21-JUN-2001	92
11-DEC-2000	67	23-MAR-2001	112	09-JUL-2001	111
27-DEC-2000	28	16-OCT-2001	148	21-JUL-2001	76
04-JAN-2001	116	06-NOV-2001	122	08-AUG-2001	116
12-JAN-2001	76	22-APR-2001	93	14-AUG-2001	52
20-JAN-2001	61	10-MAY-2001	12	20-AUG-2001	97
09-FEB-2001	319	16-MAY-2001	140	07-SEP-2001	63
15-FEB-2001	102	22-MAY-2001	132	19-SEP-2001	95

1.6. Να γίνουν οι γραφικές παραστάσεις με βάση τις τιμές των πινάκων 1.2, 1.3 και 1.5.

1.7. Στον παρακάτω πίνακα δίνεται η επίδραση της θερμοκρασίας Τ στη μοριακή θερμοχωρητικότητα C του αερίου O_2. Να γίνει η κατάλληλη γραφική παράσταση.

T, K	300	400	500	600	700	800	900	1000	1100	1200	1300
C, J/mol K	29.2	30.6	31.4	32.0	32.5	33.0	33.5	34.0	34.5	34.9	35.2

1.8. Να γίνουν οι γραφικές παραστάσεις των α) y = sinx στο διάστημα [0^o, 1080^o], β) y = sinx και y = cosx στο διάστημα [0^o, 1080^o], γ) y= cosx και y = $\tan^{-1}$(cosx) στο διάστημα [-10 rad, 10 rad].

Κεφάλαιο 2

ΠΕΡΙΓΡΑΦΙΚΗ ΣΤΑΤΙΣΤΙΚΗ

2.1 ΕΙΣΑΓΩΓΗ

Η *Στατιστική* είναι η επιστήμη που επιχειρεί να εξάγει γνώση χρησιμοποιώντας πειραματικά ή ακόμη και εμπειρικά δεδομένα που χαρακτηρίζονται σε μικρό ή μεγάλο βαθμό από τυχαιότητα και απροσδιοριστία. Στόχος της είναι να βελτιώσει την ποιότητα των δεδομένων, να προσδιορίσει συσχετίσεις μεταξύ διαφορετικών ομάδων δεδομένων ή μέσα στην ίδια ομάδα δεδομένων και επιπλέον να αποτελέσει ένα εργαλείο για πρόβλεψη με βάση ένα μικρό δείγμα δεδομένων και κατάλληλα στατιστικά μοντέλα. Για το λόγο αυτό η Στατιστική σήμερα αφορά σχεδόν όλους τους επιστημονικούς κλάδους.

Στα πρώτα της βήματα η Στατιστική άρχισε να αναπτύσσεται με τη συλλογή, οργάνωση και παρουσίαση δεδομένων με πίνακες και γραφικές παραστάσεις. Χρησιμοποιώντας σημερινή ορολογία, αναπτύχθηκε πρώτα η ***Περιγραφική Στατιστική*** *(Descriptive Statistics)*. Συγκεκριμένα, η *Περιγραφική Στατιστική* είναι εκείνος ο κλάδος της Στατιστικής που στόχο έχει την ανάπτυξη μεθόδων για τη συνοπτική και αποτελεσματική παρουσίαση των δεδομένων που προέρχονται από επαναλαμβανόμενες μετρήσεις σε ένα ή σε περισσότερα του ενός συστήματα. Για το σκοπό αυτό χρησιμοποιούνται α) αριθμητικά περιγραφικά μέτρα, β) πίνακες συχνοτήτων και γ) μέθοδοι γραφικής παρουσίασης των δεδομένων.

2.2 ΔΕΙΓΜΑ ΚΑΙ ΠΛΗΘΥΣΜΟΣ

Ονομάζουμε ***δείγμα*** *(sample)* μια συλλογή ομοειδών αποτελεσμάτων που προέρχονται από επαναλαμβανόμενες μετρήσεις ή παρατηρήσεις σε

ένα ή περισσότερα του ενός συστήματα. Το πλήθος όλων των δυνατών αποτελεσμάτων ονομάζεται ***πληθυσμός*** *(population)*.

Ο πληθυσμός μπορεί να έχει άπειρα ή πεπερασμένα στοιχεία. Για παράδειγμα, όταν προσδιορίζουμε την περιεκτικότητα σε λιπαρά μιας παρτίδας δοχείων γάλακτος επιλέγοντας πέντε τυχαία δοχεία, το πλήθος των πέντε τιμών των συγκεντρώσεων των λιπαρών στα δοχεία που επιλέχτηκαν αποτελεί ένα δείγμα. Ο πληθυσμός από τον οποίον προέρχεται το δείγμα αυτό είναι το πλήθος όλων των τιμών των συγκεντρώσεων των λιπαρών στα δοχεία γάλακτος της συγκεκριμένης παρτίδας. Αν τώρα για να προσδιορίσουμε τη συγκέντρωση των λιπαρών σε ένα δοχείο εκτελούμε 3 μετρήσεις για να πάρουμε το μέσο όρο τους, οι τρείς αυτές τιμές αποτελούν ένα άλλο δείγμα του οποίου ο πληθυσμός έχει άπειρο πλήθος τιμών, δεδομένου ότι μπορούμε να εκτελέσουμε, εν δυνάμει, άπειρο πλήθος μετρήσεων προσδιορισμού της συγκέντρωσης των λιπαρών σε κάθε δοχείο.

Είναι φανερό ότι ένα δείγμα είναι πάντα μικρότερο ή ίσο του πληθυσμού. Αν η επιλογή των στοιχείων του δείγματος από αυτά του πληθυσμού είναι τυχαία, έτσι ώστε κάθε στοιχείο του πληθυσμού να έχει την ίδια πιθανότητα να είναι στο δείγμα, τότε έχουμε ένα *τυχαίο δείγμα*.

2.3 ΑΡΙΘΜΗΤΙΚΑ ΠΕΡΙΓΡΑΦΙΚΑ ΜΕΤΡΑ

Έστω ένα δείγμα που περιέχει τις μετρήσεις x_1, x_2, ..., x_m κάποιας φυσικής ποσότητας. Υπάρχει μια πληθώρα μέτρων (δεικτών) για να χαρακτηρίσουμε τις διάφορες ιδιότητες του δείγματος αυτού. Τα πιο βασικά μέτρα δίνονται στον πίνακα 2.1, όπου ομαδοποιούνται σε δύο κατηγορίες: Μέτρα θέσης και μέτρα διασποράς. Τα μέτρα θέσης δίνουν πληροφορίες που σχετίζονται με τη θέση των δεδομένων του δείγματος αν τα τοποθετήσουμε σε έναν άξονα, ενώ τα μέτρα διασποράς ελέγχουν πόσο διασκορπισμένα είναι τα δεδομένα γύρω από κάποια κεντρική τιμή του δείγματος.

Τα μέτρα του πίνακα 2.1 ορίζονται ως εξής:

1) Η ***μέση τιμή*** *(mean ή average value)* του δείγματος τιμών $\{x_1, x_2, ..., x_m\}$ ορίζεται από τη σχέση

$$\bar{x} = (x_1 + x_2 + ... + x_m)/ m \qquad (2.1)$$

Η μέση τιμή συμβολίζεται και με <x> και είναι η τιμή γύρω από την οποία βρίσκονται συνήθως συγκεντρωμένες οι τιμές του δείγματος.

Πίνακας 2.1. Βασικά μέτρα (δείκτες) των ιδιοτήτων ενός δείγματος.

Μέτρα θέσης	Μέτρα διασποράς
Μέση τιμή (Mean)	Διασπορά (Variance)
Ισοσταθμισμένη μέση τιμή (Trimmed mean)	Τυπική απόκλιση (Standard deviation)
Διάμεσος (Median)	Τυπική απόκλιση μέσου (Standard error of the mean)
Κορυφή (Mode)	Εύρος ή Περιοχή (Range)
Πρώτο τεταρτημόριο (First quartile)	Μέγιστη/Ελάχιστη τιμή (Maximum/ Minimum)
Τρίτο τεταρτημόριο (Third quartile)	Ενδοτεταρτημοριακό εύρος (Interquartile range)

2) Η ***p% ισοσταθμισμένη μέση τιμή*** *(trimmed mean)* είναι η μέση τιμή του δείγματος που προκύπτει αφού αφαιρέσουμε το p% των μικρότερων και το p% των μεγαλύτερων τιμών του, όπου το p είναι συνήθως 5 στα μεγάλα δείγματα, ενώ σε μικρά δείγματα μπορούμε να χρησιμοποιήσουμε μεγαλύτερες τιμές, μέχρι και p = 30. Την ισοσταθμισμένη μέση τιμή τη χρησιμοποιούμε αν σε ένα δείγμα υπάρχουν μία ή περισσότερες *ακραίες τιμές*, δηλαδή τιμές που διαφέρουν σημαντικά από τις υπόλοιπες τιμές. Τότε η μέση τιμή δεν είναι η τιμή γύρω από την οποία συγκεντρώνονται οι τιμές του δείγματος.

Για παράδειγμα, έστω το δείγμα {2, 3, 5, 3, 6, 2, 4, 3, 55, 5}. Η μέση τιμή είναι 8.8 λόγω της ακραίας και πιθανόν εσφαλμένης τιμής 55. Είναι προφανές ότι οι τιμές του δείγματος δε βρίσκονται συγκεντρωμένες γύρω από την τιμή 8.8. Αν όμως αφαιρέσουμε τη μικρότερη τιμή, 2, και τη μεγαλύτερη, 55, παίρνουμε ως μέση τιμή την 3.875, που εκφράζει πράγματι τη μέση τιμή των υπόλοιπων τιμών του δείγματος. Στο παράδειγμα αυτό επιλέξαμε p = 10.

3) Η ***διάμεσος*** *(median)* είναι η "μεσαία" τιμή ενός δείγματος με την εξής έννοια. Οι μισές τιμές του δείγματος είναι μικρότερες ή ίσες με αυτή και οι υπόλοιπες μισές μεγαλύτερες ή ίσες.

Για παράδειγμα, έστω το δείγμα $\{x_1, x_2, x_3, x_4, x_5\}$, όπου οι τιμές x_i βαίνουν αυξανόμενες από το x_1 στο x_5. Η διάμεσος είναι η τιμή $x_{median} = x_3$. Αντίθετα στο δείγμα $\{x_1, x_2, x_3, x_4, x_5, x_6\}$, όπου και πάλι οι τιμές x_i βαίνουν αυξανόμενες από το x_1 στο x_6, η διάμεσος υπολογίζεται από τη σχέση $x_{median} = (x_3 + x_4)/2$. Έτσι, στο δείγμα {2, 3, 5, 3, 6, 2, 4, 3, 55, 5} η διάμεσος είναι ίση με 3.5.

Η διάμεσος δεν επηρεάζεται από ακραίες τιμές. Συνεπώς, για την περιγραφή δεδομένων που εμφανίζουν ακραίες τιμές προτιμάται ως μέτρο θέσης από τη μέση τιμή, η οποία, όπως είδαμε, επηρεάζεται πολύ από ακραίες τιμές. Φυσικά, όπως αναφέραμε, σε αυτή την περίπτωση μπορούμε να χρησιμοποιήσουμε και την p% ισοσταθμισμένη μέση τιμή.

4) Η ***κορυφή*** *(mode)* είναι η μέτρηση με τη μεγαλύτερη συχνότητα σε ένα δείγμα. Για παράδειγμα, στο δείγμα {2, 3, 5, 3, 6, 2, 4, 3} η κορυφή είναι η τιμή 3. Είναι προφανές ότι το μέτρο αυτό αφορά κυρίως δείγματα με διακριτές τιμές.

5) Το ***πρώτο, τρίτο τεταρτημόριο*** *(first, third quartile)* και το ***ενδοτεταρτημοριακό εύρος*** *(interquartile range)* ορίζονται ως εξής: Κάθε δείγμα $\{x_1, x_2, ..., x_m\}$ έχει τρία τεταρτημόρια (quartiles). Αν θεωρήσουμε ότι στο δείγμα αυτό οι τιμές αυξάνονται από το x_1 στο x_m, τότε το 25% των πρώτων τιμών είναι μικρότερες ή ίσες του πρώτου τεταρτημορίου, Q_1, ενώ το 75% των πρώτων τιμών είναι μικρότερες ή ίσες του τρίτου τεταρτημορίου, Q_3. Ως δεύτερο τεταρτημόριο, Q_2, θεωρείται η διάμεσος. Η διαφορά $Q_3 - Q_1$ ισούται με το ενδοτεταρτημοριακό εύρος.

Για παράδειγμα, έστω το δείγμα

{3.1 2 4.4 2.2 5.8 6 4 5.2 1.5 3 4 5 8 3.2 4 2 3.6 6 5.3 4.6}

Αν θέσουμε τις τιμές αυξανόμενες, παίρνουμε:

{1.5 2 2 2.2 3 3.1 3.2 3.6 4 4 4 4.4 4.6 5 5.2 5.3 5.8 6 6 8}

Το πρώτο τεταρτημόριο είναι η πέμπτη τιμή ($20\times0.25 = 5$), δηλαδή $Q_1 = 3$, το τρίτο τεταρτημόριο είναι η δέκατη πέμπτη τιμή ($20\times0.75 = 15$), δηλαδή $Q_3 = 5.2$, ενώ το ενδοτεταρτημοριακό εύρος ισούται με $Q_3 - Q_1 = 2.2$.

6) Η ***διασπορά*** ή ***διακύμανση*** *(variance)* ορίζεται από τη σχέση

$$\mathrm{Var}(x) \equiv s'^2 = \overline{(x-\bar{x})^2} = \frac{(x_1-\bar{x})^2+(x_2-\bar{x})^2+\cdots+(x_m-\bar{x})^2}{m} \qquad (2.2)$$

Δηλαδή η διασπορά s'^2 είναι η μέση τιμή της ποσότητας $(x - \bar{x})^2$. Ορθότερα η *δειγματική διασπορά* ορίζεται από τη σχέση

$$\mathrm{Var}(x) \equiv s^2 = \frac{(x_1-\bar{x})^2+(x_2-\bar{x})^2+\cdots+(x_m-\bar{x})^2}{m-1} \qquad (2.3)$$

επειδή αν πάρουμε μια πληθώρα δειγμάτων από τον ίδιο πληθυσμό και σε κάθε δείγμα υπολογίζουμε την s^2, τότε ο μέσος όρος όλων των s^2 αποδεικνύεται ότι ταυτίζεται με τη διασπορά του πληθυσμού.

Γενικά, κάθε παράμετρος δείγματος που έχει αυτή την ιδιότητα, δηλαδή η μέση τιμή της, αν πάρουμε πολλά δείγματα, να ισούται με την αντίστοιχη παράμετρο του πληθυσμού, ονομάζεται ***αβίαστη*** ή ***αμερόληπτη εκτιμήτρια*** *(unbiased estimator)*. Έτσι, ενώ η s^2 είναι αβίαστη εκτιμήτρια, δεν συμβαίνει το ίδιο με τη s'^2. Αβίαστη εκτιμήτρια είναι και η μέση τιμή $\bar{x}$ ενός δείγματος.

Η διασπορά, ανεξάρτητα από ποια σχέση ορίζεται, είναι ένα μέτρο που δείχνει πως διασπείρονται οι τιμές ενός δείγματος γύρω από τη μέση τιμή. Αν οι s'^2 και s^2 έχουν μεγάλες τιμές, τότε υπάρχει μεγάλη διασπορά στις τιμές του δείγματος.

7) Η ***τυπική απόκλιση*** *(standard deviation)* είναι η τετραγωνική ρίζα της διασποράς, δηλαδή είναι η ποσότητα s' ή s. Είναι φανερό ότι και η τυπική απόκλιση είναι μέτρο των αποκλίσεων των μετρήσεων x_i (i = 1, 2,..., m) από τη μέση τιμή $\bar{x}$. Η τυπική απόκλιση δεν είναι αβίαστη εκτιμήτρια.

Για να διακρίνουμε τη μέση τιμή $\bar{x}$, τη διασπορά s^2 και την τυπική απόκλιση s ενός δείγματος από αυτές του αντίστοιχου πληθυσμού, συμβολίζουμε τη μέση τιμή, τη διασπορά και την τυπική απόκλιση του πληθυσμού με μ, σ^2 και σ, αντίστοιχα.

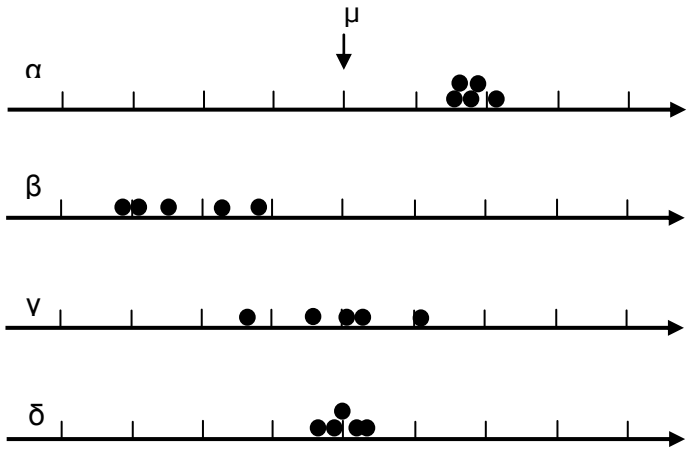
8) Η ***τυπική απόκλιση του μέσου*** *(standard error of the mean)* είναι η ποσότητα $s/\sqrt{m}$. Αν δημιουργήσουμε πολλά δείγματα από τον ίδιο πληθυσμό με m τιμές το κάθε ένα και σε κάθε δείγμα υπολογίσουμε τη μέση τιμή, τότε δημιουργούμε ένα νέο δείγμα, το δείγμα των μέσων τιμών των αρχικών δειγμάτων. Σε αυτό το δείγμα η ποσότητα $s/\sqrt{m}$ αποτελεί ένα μέτρο της τυπικής απόκλισης των τιμών του.

9) Το ***εύρος*** ή ***περιοχή*** *(range)* είναι η διαφορά μέγιστης μείον ελάχιστης τιμής του δείγματος.

2.4 ΑΞΙΟΛΟΓΗΣΗ ΤΙΜΩΝ ΔΕΙΓΜΑΤΟΣ

Στη διεθνή βιβλιογραφία με βάση τις τιμές $\bar{x}$, s και μ τα δεδομένα ενός δείγματος αξιολογούνται με τους αγγλικούς όρους *accuracy* και *precision*. Στα ελληνικά θα μπορούσαμε να αποδώσουμε τους όρους αυτούς με τις λέξεις ***ακρίβεια*** και ***πιστότητα***, αντίστοιχα. Οι μετρήσεις ενός δείγματος χαρακτηρίζονται από *μεγάλη ακρίβεια (high accuracy)* αν η

μέση τιμή του δείγματος $\bar{x}$ είναι πολύ κοντά στη μέση τιμή μ του πληθυσμού και από *μεγάλη πιστότητα (high precision)* όταν το s είναι πολύ μικρό. Είναι φανερό ότι μπορούμε να εκτιμήσουμε την *ακρίβεια* των μετρήσεων μόνο όταν γνωρίζουμε την τιμή μ του πληθυσμού.



Σχήμα 2.1. Δείγματα με διαφορετική ακρίβεια και πιστότητα

Με βάση τους παραπάνω ορισμούς, οι μετρήσεις των δειγμάτων α, β, γ και δ στο σχήμα 2.1 χαρακτηρίζονται ως εξής: α) Μεγάλης πιστότητας και μικρής ακρίβειας, β) μικρής ακρίβειας και μικρής πιστότητας, γ) μεγάλης ακρίβειας και μικρής πιστότητας, και δ) μεγάλης ακρίβειας και μεγάλης πιστότητας. Από το σχήμα αυτό γίνεται φανερό ότι στην περίπτωση (α) κάποιο συστηματικό σφάλμα επηρεάζει τις μετρήσεις και το ίδιο ισχύει στην περίπτωση (β). Επιπλέον στην περίπτωση (β) όπως και στη (γ) η σχετικά μεγάλη διασπορά των σημείων δείχνει ότι οι μετρήσεις έγιναν με όργανα ή με μέθοδο μικρής επαναληψιμότητας ή σε πειραματικές συνθήκες μη ελεγχόμενες. Τέλος, θα πρέπει να σημειώσουμε ότι επειδή στις περισσότερες περιπτώσεις δε γνωρίζουμε τη μέση τιμή μ του πληθυσμού, οι περιπτώσεις (α) και (δ) ουσιαστικά δεν ξεχωρίζουν. Έτσι αν έχουμε μόνο τα αποτελέσματα του δείγματος (α) είναι πολύ πιθανό να θεωρήσουμε (λανθασμένα), λόγω της πολύ καλής επαναληψιμότητας των μετρήσεων, ότι η μέση τιμή $\bar{x}$ του δείγματος είναι κοντά στη μέση τιμή μ του πληθυσμού. Αυτό δείχνει παραστατικά πόσο "επικίνδυνα" είναι τα συστηματικά σφάλματα όταν δεν γνωρίζουμε τις παραμέτρους του πληθυσμού.

2.5 ΠΙΝΑΚΕΣ ΣΥΧΝΟΤΗΤΩΝ

Έστω x_1, x_2, …, x_m οι τιμές μιας διακριτής μεταβλητής Χ σε ένα δείγμα. Ονομάζουμε *συχνότητα* (*frequency*) της τιμής x_i τον φυσικό αριθμό v_i που δείχνει πόσες φορές επαναλαμβάνεται η τιμή x_i στο δείγμα. Αν $v = v_1 + v_2 + … + v_m$, τότε ο λόγος

$$f_i = \frac{v_i}{v} \tag{2.4}$$

ονομάζεται *σχετική συχνότητα* της τιμής x_i. Τέλος, η *αθροιστική συχνότητα*, F_i, ορίζεται από τη σχέση

$$F_i = f_1 + f_2 + ... + f_i \quad \text{για } i = 1, 2, …, m \tag{2.5}$$

Όταν το πλήθος των τιμών του δείγματος είναι μεγάλο αλλά κυρίως όταν η μεταβλητή Χ είναι συνεχής, οπότε μπορεί να πάρει μια οποιαδήποτε τιμή στο πεδίο ορισμού της, οι συχνότητες δεν ορίζονται σε μια συγκεκριμένη τιμή x_i αλλά σε μια περιοχή τιμών της Χ, που ονομάζεται *κλάση* (*bin*). Συγκεκριμένα αν x_{min} και x_{max} είναι η ελάχιστη και η μέγιστη τιμή της μεταβλητής Χ στο δείγμα, διαιρούμε το διάστημα x_{max} - x_{min} σε k υποδιαστήματα μήκους $\Delta x = (x_{max} - x_{min})/k$, που ονομάζονται *κλάσεις* και σε κάθε κλάση υπολογίζουμε το σύνολο των τιμών του δείγματος που ανήκουν σε αυτή. Η ποσότητα αυτή, που προφανώς είναι ένας φυσικός αριθμός, είναι η *συχνότητα της κλάσης*. Αντίστοιχα ορίζονται η *σχετική* και η *αθροιστική συχνότητα μιας κλάσης*.

2.6 ΜΕΘΟΔΟΙ ΓΡΑΦΙΚΗΣ ΠΑΡΟΥΣΙΑΣΗΣ ΔΕΔΟΜΕΝΩΝ

Υπάρχουν αρκετοί τύποι γραφικών παραστάσεων για την παρουσίαση στατιστικών δεδομένων. Από αυτούς οι βασικοί τύποι είναι τα α) ραβδογράμματα, β) κυκλικά διαγράμματα, γ) ιστογράμματα και δ) θηκογράμματα. Οι δύο πρώτοι τύποι γραφικών παραστάσεων χρησιμοποιούνται συνήθως όταν η μεταβλητή Χ είναι ποιοτική, ενώ οι δύο τελευταίοι τύποι όταν έχουμε ποσοτικά δεδομένα.

α) *Ραβδόγραμμα* *(barchart)*

Το ραβδόγραμμα σχηματίζεται με βάση τον πίνακα συχνοτήτων μιας διακριτής μεταβλητής Χ. Στον οριζόντιο άξονα τοποθετούνται ισαπέχοντα τα στοιχεία του δείγματος και σε κάθε στοιχείο αντιστοιχούμε μια ορθογώνια στήλη με ύψος ίσο με τη συχνότητα του στοιχείου.

β) *Κυκλικό διάγραμμα* *(piechart)*

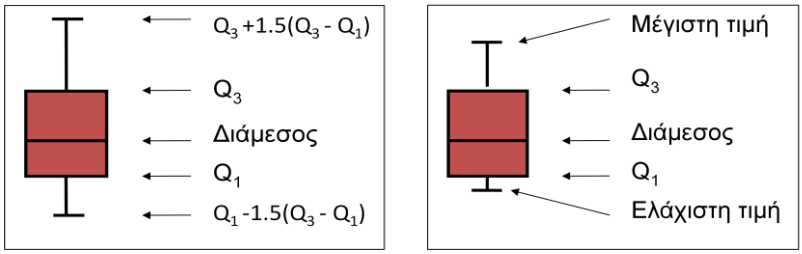
Το διάγραμμα αυτό είναι ένας κυκλικός δίσκος χωρισμένος σε κυκλικούς τομείς. Κάθε κυκλικός τομέας εκφράζει ένα στοιχείο του δείγματος και έχει εμβαδό ανάλογο προς τη συχνότητα του στοιχείου.

γ) *Ιστόγραμμα* *(histogram)*

Είναι αντίστοιχο του ραβδογράμματος, μόνο που στον οριζόντιο άξονα τοποθετούνται όχι τα στοιχεία του δείγματος αλλά οι κλάσεις που δημιουργούμε.

δ) *Θηκόγραμμα* *(boxplot)*

Το θηκόγραμμα απαρτίζεται από ένα ορθογώνιο με δύο κεραίες, η μία στην κάτω βάση του ορθογωνίου με σχήμα αντεστραμμένου Τ και η άλλη στην επάνω βάση του σε σχήμα Τ. Η κάτω βάση του ορθογωνίου βρίσκεται στο Q_1 και η επάνω στο Q_3. Η διάμεσος παριστάνεται με ένα ευθύγραμμο οριζόντιο τμήμα στο εσωτερικό του ορθογωνίου. Το μήκος των βάσεων του ορθογωνίου είναι αυθαίρετο. Οι κεραίες, *φράκτες (whiskers)*, εκτείνονται μέχρι τις τιμές $Q_3 + 1.5(Q_3 - Q_1)$ και $Q_1 - 1.5(Q_3 - Q_1)$, όπως στο σχήμα 2.2. Αν η μέγιστη ή η ελάχιστη τιμή του δείγματος βρίσκονται εντός των περιοχών αυτών, τότε οι φράκτες μετατοπίζονται στη μέγιστη ή στην ελάχιστη τιμή. Αν υπάρχουν ακραίες τιμές αυτές εμφανίζονται ως σημεία εκτός των φρακτών.



Σχήμα 2.2. Θηκογράμματα

2.7 ΠΕΡΙΓΡΑΦΙΚΗ ΣΤΑΤΙΣΤΙΚΗ ΣΤΟ EXCEL

Για τη στατιστική ανάλυση ενός δείγματος το *Excel* διαθέτει μια πλούσια βιβλιοθήκη συναρτήσεων και προγραμμάτων. Οι βασικές συναρτήσεις που σχετίζονται με την περιγραφική στατιστική δίνονται στον πίνακα 2.2. Για ειδικές αναλύσεις υπάρχει επίσης μια πληθώρα άλλων στατιστικών συναρτήσεων, που μπορούμε να τις βρούμε αν κάνουμε *κλικ* στο εικονίδιο *Εισαγωγή συνάρτησης (Insert Function),*  , στη *Γραμμή τύπων* (*Formula bar*), και επιλέξουμε *Στατιστικές* (*Statistical*) από το πλαίσιο λίστας *Επιλογή κατηγορίας (Select a category)*. Εναλλακτικά μπορούμε να επιλέξουμε μια στατιστική συνάρτηση από: *Τύποι (Formulas)* → *Βιβλιοθήκη συναρτήσεων (Function Library)* → *Περισσότερες συναρτήσεις (More Functions)* → *Στατιστική* (*Statistical*).

Πίνακας 2.2. Βασικές στατιστικές συναρτήσεις του *Excel*.

Συνάρτηση	Περιγραφή
AVERAGE	Υπολογίζει τη μέση τιμή.
TRIMMEAN(array;percent)	Αν percent = 0.05, υπολογίζει την ισοσταθμισμένη μέση τιμή αφού αφαιρεθεί το 2.5% των μικρότερων και το 2.5% των μεγαλύτερων τιμών.
MEDIAN	Υπολογίζει τη διάμεσο.
MODE	Υπολογίζει την κορυφή.
QUARTILE(array;quart)	Με quart = 1 υπολογίζει το πρώτο τεταρτημόριο, με quart = 2 τη διάμεσο και με quart = 3 το τρίτο τεταρτημόριο.
VAR	Υπολογίζει τη διασπορά.
STDEV	Υπολογίζει την τυπική απόκλιση.

Παρατήρηση. Στο *Excel* δεν έχει σημασία αν εισάγουμε (πληκτρολογούμε) τις συναρτήσεις με μικρά ή κεφαλαία γράμματα.

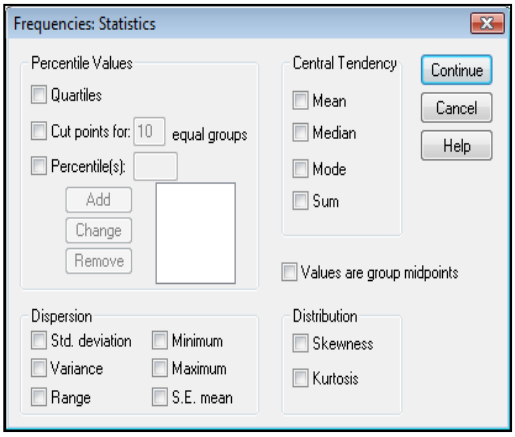
Εκτός όμως από τις συναρτήσεις αυτές, το *Excel* διαθέτει ένα επιπλέον ισχυρό εργαλείο ανάλυσης δεδομένων. Είναι το *Ανάλυση Δεδομένων* (*Data Analysis*), που βρίσκεται στη λωρίδα *Δεδομένα* (*Data),* στο πλαίσιο *Ανάλυση* (*Analysis)*. Αν δεν υπάρχει εκεί ή αν δεν υπάρχει το πλαίσιο *Ανάλυση (Analysis)*, πρέπει να το εγκαταστήσουμε. Για το σκοπό αυτό κάνουμε *κλικ* στο εικονίδιο *Κουμπί Office* (*Office Button)*  αν

έχουμε την έκδοση *Excel 2007* ή στο *Αρχείο (File)* στο *Excel 2010*, *κλικ* στο *Επιλογές του Excel* (*Excel Options)* και *κλικ* στο *Πρόσθετα* (*Add-ins)*. Στο παράθυρο διαλόγου που ανοίγει εισάγουμε στο πλαίσιο κειμένου *Διαχείριση* (*Manage)* το *Πρόσθετα του Excel* (*Excel Add-ins)* και κάνουμε *κλικ* στο *Μετάβαση* (*Go).* Τότε εμφανίζεται μια λίστα εντολών που πρέπει να το περιέχει με το όνομα *Πακέτο εργαλείων ανάλυσης (Analysis ToolPak*). Το επιλέγουμε με απλό *κλικ*. Ενδέχεται στο σημείο αυτό να ζητηθεί η εγκατάσταση του *Πακέτου εργαλείων ανάλυσης.* Απαντάμε καταφατικά, οπότε η *Ανάλυση Δεδομένων* (*Data Analysis*) θα εμφανίζεται πλέον στο πλαίσιο *Ανάλυση (Analysis)* της λωρίδας *Δεδομένα* (*Data)*. Οι δυνατότητες που έχει και ο τρόπος χρησιμοποίησής τους θα εξεταστούν παρακάτω.

Σε ό,τι αφορά τις δυνατότητες γραφικών παραστάσεων, με το *Excel* κάνουμε εύκολα ραβδογράμματα και κυκλικά διαγράμματα, σχετικά δύσκολα ιστογράμματα, ενώ δεν υπάρχουν θηκογράμματα. Θηκογράμματα αλλά και ιστογράμματα γίνονται εύκολα με το *SPSS* ή το *ChemStat*.

2.8 ΠΕΡΙΓΡΑΦΙΚΗ ΣΤΑΤΙΣΤΙΚΗ ΣΤΟ SPSS

Για να υπολογίσουμε τα περιγραφικά μέτρα μιας μεταβλητής με το *SPSS* πηγαίνουμε: *Analyze* → *Descriptive Statistics* → *Frequencies* και στο παράθυρο διαλόγου *Frequencies* επιλέγουμε με *κλικ* τη μεταβλητή που θέλουμε να αναλύσουμε και με *κλικ* στο βέλος  τη μεταφέρουμε στο πάνελ *Variable(s)*. Κάνουμε *κλικ* στο *Statistics* και επιλέγουμε τα μέτρα που θέλουμε από το παράθυρο *Frequencies: Statistics* (σχήμα 2.3).



Σχήμα 2.3. Παράθυρο διαλόγου *Frequencies:Statistics*

Παρατηρούμε ότι μπορούμε να υπολογίσουμε όλα τα στατιστικά μέτρα του πίνακα 2.1 με εξαίρεση την *ισοσταθμισμένη μέση τιμή (trimmed mean)*. Η ποσότητα αυτή υπολογίζεται με τη διαδικασία *Explore*, δηλαδή από *Analyze* → *Descriptive Statistics* → *Explore*.

Παράδειγμα 2.1

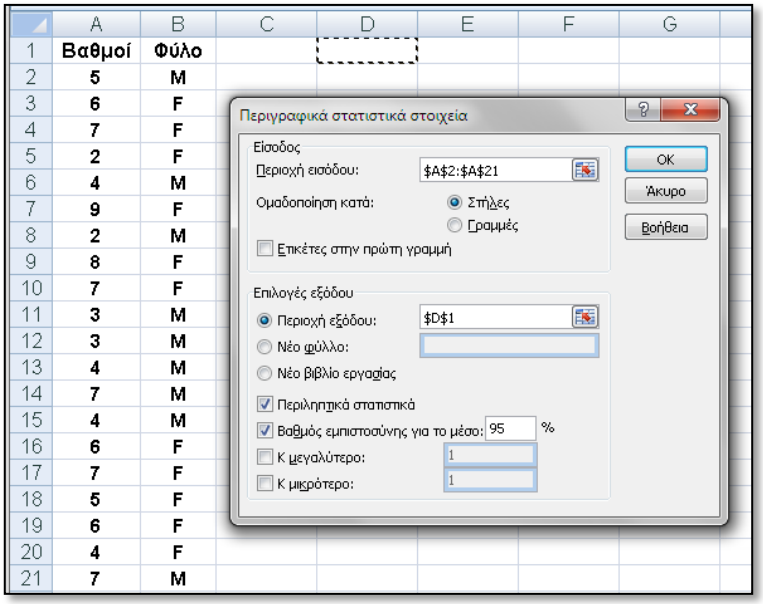
Σε ένα δείγμα 20 φοιτητών καταγράφηκαν οι βαθμοί στο μάθημα των μαθηματικών και δίνονται στον πίνακα 2.3. Στον ίδιο πίνακα είναι και το φύλο των φοιτητών. Να υπολογιστούν τα περιγραφικά στατιστικά μέτρα του δείγματος, να γίνει ο πίνακας συχνοτήτων, το αντίστοιχο ραβδόγραμμα και το κυκλικό διάγραμμα. Επιπλέον να γίνει το ραβδόγραμμα που να δείχνει τις συχνότητες των βαθμών που πήραν ξεχωριστά οι φοιτητές και οι φοιτήτριες και τέλος τα θηκογράμματα της βαθμολογίας φοιτητών-φοιτητριών.

Πίνακας 2.3. Βαθμολογία φοιτητών (M) και φοιτητριών (F).

5	6	7	2	4	9	2	8	7	3
M	F	F	F	M	F	M	F	F	M
3	4	7	4	6	7	5	6	4	7
M	M	M	M	F	F	F	F	F	M

❖ **Ανάλυση στο Excel**

Μεταφέρουμε τα δεδομένα σε φύλλο του *Excel*, όπως φαίνεται στο σχήμα 2.4, και πηγαίνουμε *Δεδομένα* (*Data*) → *Ανάλυση Δεδομένων* (*Data Analysis)* → *Περιγραφικά Στατιστικά (Descriptive Statistics)*. Στο αντίστοιχο παράθυρο διαλόγου που εμφανίζεται εισάγουμε την περιοχή A2:A21, ορίζουμε το κελί εξόδου των αποτελεσμάτων, επιλέγουμε το *Περιληπτικά στατιστικά (Summary statistics)*. Για την εισαγωγή των δεδομένων κάνουμε *κλικ* στο πλαίσιο κειμένου *Περιοχή εισόδου (Input Range)* και με το ποντίκι επιλέγουμε την περιοχή των δεδομένων (σχήμα 2.4). Κάνοντας *κλικ* στο *OK* παίρνουμε τα αποτελέσματα που δίνονται στο σχήμα 2.5.



Σχήμα 2.4. Παράθυρο εισόδου δεδομένων για τον υπολογισμό περιγραφικών μέτρων στο *Excel*

	A	B	C	D	E	F
1	**Βαθμοί**	**Φύλο**		*Στήλη1*		
2	5	M				
3	6	F		Μέσος	5,3	
4	7	F		Τυπικό σφάλμα	0,447802	
5	2	F		Διάμεσος	5,5	
6	4	M		Επικρατούσα τιμή	7	
7	9	F		Μέση απόκλιση τετραγώνου	2,00263	
8	2	M		Διακύμανση	4,010526	
9	8	F		Κύρτωση	-0,89998	
10	7	F		Ασυμμετρία	-0,0664	
11	3	M		Εύρος	7	
12	3	M		Ελάχιστο	2	
13	4	M		Μέγιστο	9	
14	7	M		Άθροισμα	106	
15	4	M		Πλήθος	20	
16	6	F		Βαθμός εμπιστοσύνης(95,0%)	0,93726	
17	7	F				
18	5	F				
19	6	F				

Σχήμα 2.5. Αποτελέσματα περιγραφικών μέτρων στο *Excel*

Στον πίνακα του σχήματος 2.5, *Μέση απόκλιση τετραγώνου* είναι η *Τυπική απόκλιση* και *Επικρατούσα τιμή* είναι η *Κορυφή*. Το *Τυπικό σφάλμα* είναι η ποσότητα $s/\sqrt{m}$, δηλαδή η *Τυπική απόκλιση του μέσου (Standard error of mean).*

Αν θέλουμε να υπολογίσουμε επιλεκτικά κάποιο μέτρο, π.χ. τη μέση τιμή, σε ένα κελί πληκτρολογούμε την αντίστοιχη συνάρτηση, *=AVERAGE(*, με το ποντίκι επιλέγουμε την περιοχή A2:A21, οπότε η περιοχή αυτή εισέρχεται στο όρισμα της συνάρτησης, και πατάμε *Enter*. Θα πάρουμε την τιμή 5.3.

Για τον πίνακα συχνοτήτων εργαζόμαστε ως εξής: Σε ένα φύλλο εργασίας εισάγουμε πάλι τα δεδομένα της άσκησης με τις διαδικασίες *Αντιγραφή-Επικόλληση*. Ακολούθως εισάγουμε τους τίτλους *Βαθμός, Συχνότητα*, *Σχετική συχνότητα* και *Αθροιστική συχνότητα* στην περιοχή C3:F3, όπως φαίνεται στο σχήμα 2.6. Εφόσον οι βαθμοί είναι από το 2 έως το 9, στην περιοχή C5:C12 εισάγουμε αυτούς τους αριθμούς και επιλέγουμε με το ποντίκι την περιοχή D5:D12, στην οποία θα υπολογίσουμε τη συχνότητα των βαθμών χρησιμοποιώντας τη συνάρτηση *FREQUENCY*.

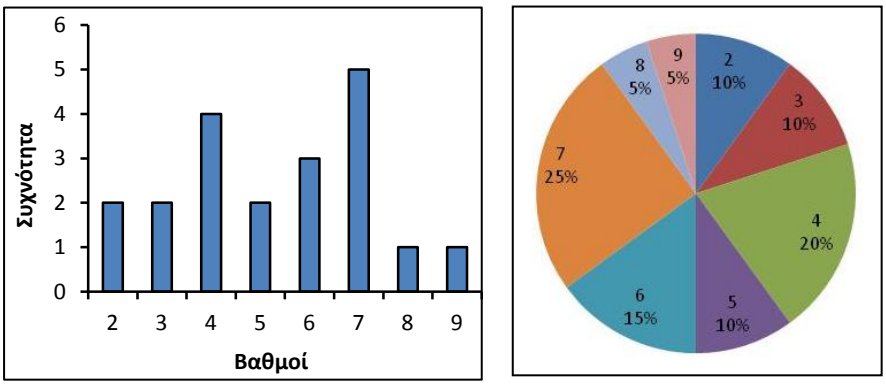
	A	B	C	D	E	F
1	**Βαθμοί**	**Φύλο**				
2	5	M				
3	6	F	**Βαθμός**	**Συχνότητα**	**Σχετική συχνότητα**	**Αθροιστική συχνότητα**
4	7	F				
5	2	F	2	2	0,1	0,1
6	4	M	3	2	0,1	0,2
7	9	F	4	4	0,2	0,4
8	2	M	5	2	0,1	0,5
9	8	F	6	3	0,15	0,65
10	7	F	7	5	0,25	0,9
11	3	M	8	1	0,05	0,95
12	3	M	9	1	0,05	1
13	4	M				
14	7	M				
15	4	M	sum=	20		
16	6	F				
17	7	F				
18	5	F				
19	6	F				
20	4	F				
21	7	M				

Σχήμα 2.6. Φύλλο εργασίας με υπολογισμούς συχνοτήτων

Για το σκοπό αυτό κάνουμε *κλικ* στο εικονίδιο *Εισαγωγή Συνάρτησης* και επιλέγουμε τη συνάρτηση *FREQUENCY*. Στο πλαίσιο διαλόγου που ανοίγει εισάγουμε με το ποντίκι την περιοχή A2:A21 στο *Data_array* και την περιοχή C5:C12 στο *Bins_array*. Επειδή η συνάρτηση αυτή έχει έξοδο μια περιοχή δεν κάνουμε *κλικ* στο *OK* αλλά πρώτα πατάμε *Ctrl+Shift* και με πατημένα αυτά τα πλήκτρα κάνουμε *κλικ* στο *OK*.

Συνεχίζουμε υπολογίζοντας το άθροισμα των συχνοτήτων στο D15, στο E5 πληκτρολογούμε = D5/D$15 και συμπληρώνουμε με τη διαδικασία της αυτόματης συμπλήρωσης μέχρι το E12. Τέλος η αθροιστική συχνότητα υπολογίζεται αν πληκτρολογήσουμε =E5 στο F5, =E6+F5 στο F6 και συμπληρώσουμε με τη διαδικασία της αυτόματης συμπλήρωσης μέχρι το F12. Θα πάρουμε την εικόνα του σχήματος 2.6.

Για το ραβδόγραμμα επιλέγουμε την περιοχή D5:D12 και πηγαίνουμε *Εισαγωγή (Insert)* → *Στήλες (Column)* → *Στήλη τμημάτων (Clustered column)*. Η γραφική παράσταση που σχηματίζεται δεν έχει τον σωστό άξονα των x. Κάνουμε δεξί *κλικ* στο γράφημα και από την αναδυόμενη λίστα επιλέγουμε *Επιλογή δεδομένων (Select Data)*. Στο πλαίσιο διαλόγου που ανοίγει κάνουμε *κλικ* στο κουμπί *Επεξεργασία (Edit)* που βρίσκεται δεξιά και στο νέο πλαίσιο διαλόγου εισάγουμε στο πλαίσιο *Περιοχή ετικετών άξονα (Axis label range)* την περιοχή C5:C12.



Σχήμα 2.7. Ραβδόγραμμα και κυκλικό διάγραμμα του παραδείγματος 2.1

Για το κυκλικό διάγραμμα εργαζόμαστε ακριβώς όπως και για το ραβδόγραμμα, επιλέγοντας *Πίτα (Pie)* αντί *Στήλη τμημάτων (Clustered column).* Επιπλέον από το *Διατάξεις γραφήματος (Charts Layouts)* επιλέγουμε το *Διάταξη 1 (Layout 1).* Και τα δύο διαγράμματα, κυρίως όμως

το ραβδόγραμμα, απαιτούν περαιτέρω μορφοποίηση ώστε να αποκτήσουν τα επιθυμητά χαρακτηριστικά, π.χ. αυτά του σχήματος 2.7.

Για να γίνει το ραβδόγραμμα με τις συχνότητες των βαθμών που πήραν ξεχωριστά οι φοιτητές και οι φοιτήτριες εργαζόμαστε ως εξής. Σε ένα νέο φύλλο εργασίας εισάγουμε πάλι τα δεδομένα της άσκησης με τις διαδικασίες *Αντιγραφή-Επικόλληση*. Ακολούθως εισάγουμε τους τίτλους *Φοιτητές, Φοιτήτριες, Βαθμός, Συχνότητα Φοιτητών, Συχνότητα Φοιτητριών*, όπως φαίνεται στο σχήμα 2.8. Οι βαθμοί από το 2 έως το 9 εισάγονται στην περιοχή E5:E12.

	A	B	C	D	E	F	G
1	**Βαθμοί**	**Φύλο**	**Φοιτητές**	**Φοιτήτριες**			
2	5	M	5			**Συχνότητα**	**Συχνότητα**
3	6	F		6	**Βαθμός**	**Φοιτητών**	**Φοιτητριών**
4	7	F		7			
5	2	F		2	2	1	1
6	4	M	4		3	2	0
7	9	F		9	4	3	1
8	2	M	2		5	1	1
9	8	F		8	6	0	3
10	7	F		7	7	2	3
11	3	M	3		8	0	1
12	3	M	3		9	0	1
13	4	M	4				
14	7	M	7				
15	4	M	4				
16	6	F		6			
17	7	F		7			
18	5	F		5			
19	6	F		6			
20	4	F		4			
21	7	M	7				

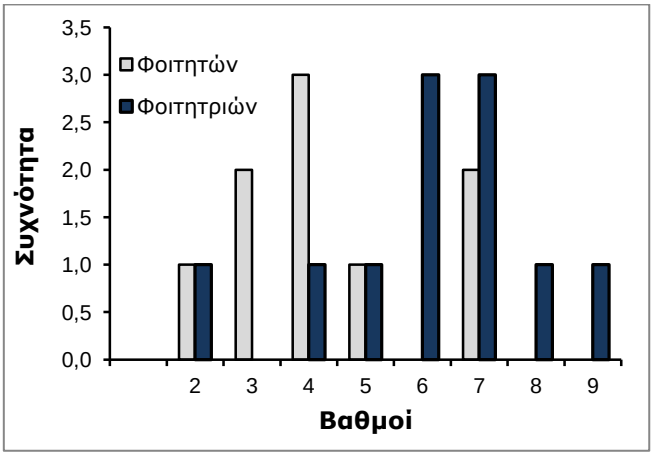
Σχήμα 2.8. Φύλλο εργασίας με υπολογισμούς συχνοτήτων

Στη συνέχεια στο κελί C2 πληκτρολογούμε =IF(B2="M";A2;""), πατάμε *Enter* και εφαρμόζουμε τη διαδικασία της αυτόματης συμπλήρωσης μέχρι το κελί C21. Με βάση τον τύπο IF(B2="M";A2;""), όταν ισχύει B2="M", δηλαδή όταν στο κελί B2 υπάρχει το M, τότε στο κελί C2 εμφανίζεται το περιεχόμενο του κελιού A2. Αντίθετα, όταν δεν ισχύει B2="M", τότε στο C2 εφαρμόζεται η εντολή "" και συνεπώς το C2 παραμένει κενό. Επαναλαμβάνουμε τη διαδικασία αυτή στο κελί D2 πληκτρολογώντας τον τύπο =IF(B2="F";A2;"") και εφαρμόζοντας τη διαδικασία της αυτόματης συμπλήρωσης μέχρι το κελί D21.

Για να προσδιορίσουμε τώρα τις συχνότητες που αντιστοιχούν στους φοιτητές και μετά στις φοιτήτριες, επιλέγουμε με το ποντίκι την περιοχή

F5:F12, κάνουμε *κλικ* στο εικονίδιο f_x και επιλέγουμε τη συνάρτηση *FREQUENCY*. Στο πλαίσιο διαλόγου που ανοίγει εισάγουμε με το ποντίκι στο πλαίσιο *Data_array* την περιοχή C2:C21 και στο πλαίσιο *Bins_array* την περιοχή E5:E12. Πατάμε *Ctrl+Shift* και με πατημένα αυτά τα πλήκτρα κάνουμε *κλικ* στο *OK*. Με τον ίδιο τρόπο υπολογίζουμε στην περιοχή G5:G12 τις συχνότητες που αντιστοιχούν στις φοιτήτριες.

Για το ραβδόγραμμα επιλέγουμε την περιοχή F5:G12 και πηγαίνουμε *Εισαγωγή (Insert)* → *Στήλες (Column)* → *Στήλη τμημάτων (Clustered column)*. Διορθώνουμε και πάλι τον άξονα των x, όπως στο σχήμα 2.7, και με κατάλληλη μορφοποίηση παίρνουμε το σχήμα 2.9.

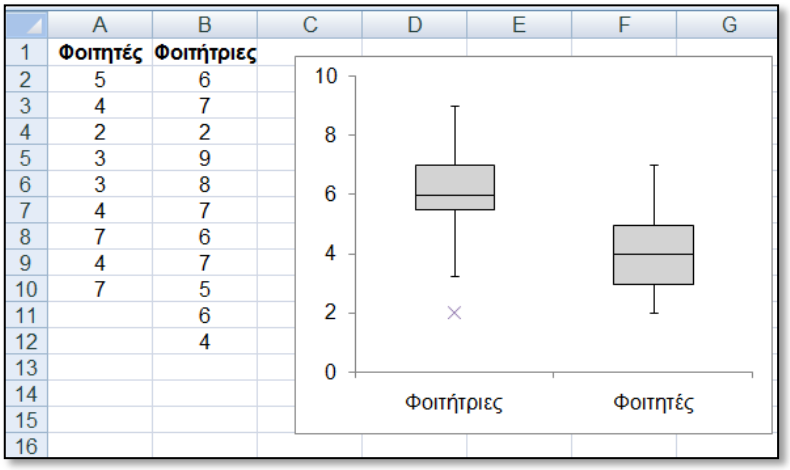


Σχήμα 2.9. Ραβδόγραμμα συχνοτήτων των βαθμών που πήραν ξεχωριστά οι φοιτητές και οι φοιτήτριες

Θηκογράμματα στο *Excel* δεν γίνονται. Μπορεί όμως να χρησιμοποιηθεί το *ChemStat* με την προϋπόθεση ότι εργαζόμαστε σε έκδοση του *Excel 2007* ή *2010*. Σε αυτή την περίπτωση με τις διαδικασίες *Αντιγραφή-Επικόλληση Τιμών* μεταφέρουμε τις στήλες C και D του φύλλου εργασίας του σχήματος 2.8 σε ένα νέο φύλλο και διαγράφουμε τα κενά κελιά, όπως στο σχήμα 2.10. Ακολούθως πηγαίνουμε *Πρόσθετα (Add-ins)* → *ChemStat* → *Graphs* → *Boxplot* και συμπληρώνουμε κατάλληλα τα παράθυρα που ανοίγουν.

Συγκεκριμένα, στο δεύτερο παράθυρο που ανοίγει εισάγουμε τον αριθμό 2, που είναι ο αριθμός των δειγμάτων και στο επόμενο εισάγουμε το

δείγμα των φοιτητριών επειδή αυτό είναι το μεγαλύτερο δείγμα. Το δείγμα εισάγεται ΧΩΡΙΣ τον τίτλο του, δηλαδή εισάγουμε την περιοχή B2:B12. Ακολούθως εισάγουμε τον τίτλο του δείγματος με *κλικ* στο B1, μετά εισάγουμε το δείγμα των φοιτητών, A2:A10, και τέλος τον τίτλο του δευτέρου δείγματος κάνοντας *κλικ* στο A1. Στο τελευταίο παράθυρο κάνουμε *κλικ* με το ποντίκι σε ένα κελί που είναι το κελί εξόδου των αποτελεσμάτων. Το πρόγραμμα εξάγει στο φύλλο εργασίας όλες τις ποσότητες που είναι απαραίτητες για το σχεδιασμό των θηκογραμμάτων και με βάση αυτές τα κατασκευάζει. Στο συγκεκριμένο παράδειγμα παίρνουμε τα θηκογράμματα του σχήματος 2.10.



Σχήμα 2.10. Θηκογράμματα βαθμολογίας φοιτητριών και φοιτητών του παραδείγματος 2.1

❖ Ανάλυση στο SPSS

Μεταφέρουμε τα δεδομένα σε ένα φύλλο του *SPSS* και ονομάζουμε τη στήλη των βαθμών grades και τη στήλη με το φύλο sex. Ανάλογα με την έκδοση του *SPSS* είναι πιθανόν για να εισάγουμε την αλφαριθμητική μεταβλητή sex να πρέπει πρώτα να τη δηλώσουμε ως *String* στο παράθυρο *Variable View.*

Για να αναλύσουμε τα δεδομένα πηγαίνουμε *Analyze* → *Descriptive Statistics* → *Frequencies*. Στο παράθυρο διαλόγου που ανοίγει επιλέγουμε τη στήλη (μεταβλητή) grades με τους βαθμούς κάνοντας *κλικ* στη μεταβλητή αυτή και τη μεταφέρουμε στο πάνελ *Variable(s)*. Ακολούθως

ενεργοποιούμε την επιλογή *Display frequency tables*, κάνουμε *κλικ* στο κουμπί *Statistics* και επιλέγουμε τα μέτρα που θέλουμε. Τέλος, από το κουμπί *Charts* μπορούμε να επιλέξουμε και τη γραφική παράσταση των συχνοτήτων. Αυτή μπορεί να είναι ή ραβδόγραμμα ή κυκλικό γράφημα ή ιστόγραμμα (*Bar charts*, *Pie charts*, *Histograms*). Δυστυχώς δεν μπορούμε να επιλέξουμε ταυτόχρονα ραβδόγραμμα και κυκλικό γράφημα. Έτσι επιλέγουμε αρχικά ραβδόγραμμα και ξανα-εκτελούμε την ίδια διαδικασία επιλέγοντας κυκλικό γράφημα. Επίσης, στον άξονα των y μπορούμε να επιλέξουμε ή απλές *συχνότητες* (*Frequencies*) ή *εκατοστιαίες συχνότητες* (*Percentages*). Τα αποτελέσματα που παίρνουμε δίνονται στα σχήματα 2.11 και 2.12.

Statistics

grades

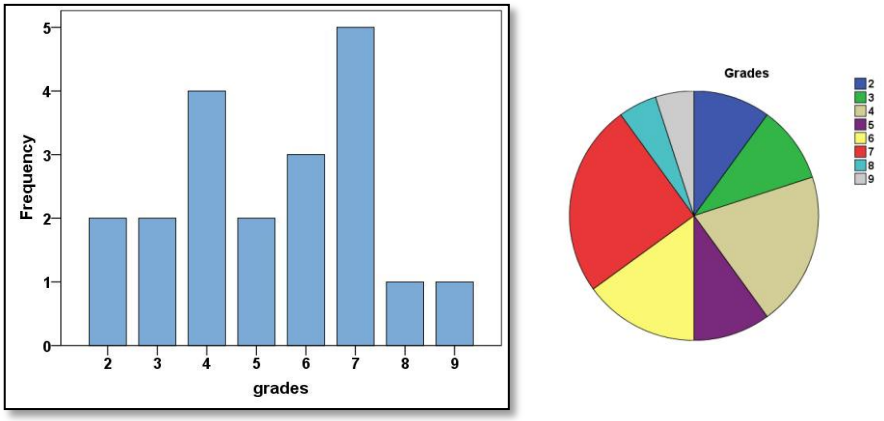
N	Valid	20
	Missing	0
Mean		5,30
Median		5,50
Std. Deviation		2,003

grades

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	2	2	10,0	10,0	10,0
	3	2	10,0	10,0	20,0
	4	4	20,0	20,0	40,0
	5	2	10,0	10,0	50,0
	6	3	15,0	15,0	65,0
	7	5	25,0	25,0	90,0
	8	1	5,0	5,0	95,0
	9	1	5,0	5,0	100,0
	Total	20	100,0	100,0	

Σχήμα 2.11. Πίνακες αποτελεσμάτων του *SPSS*

Παρατηρούμε ότι η ανάλυση συχνοτήτων και τα αντίστοιχα ραβδογράμματα και κυκλικά διαγράμματα γίνονται ιδιαίτερα εύκολα με το *SPSS*. Αυτό είναι αναμενόμενο δεδομένου ότι το *SPSS* έχει σχεδιαστεί για αυτή τη δουλειά, ενώ το *Excel* έχει δημιουργηθεί για πολύ ευρύτερες χρήσεις.



Σχήμα 2.12. Ραβδόγραμμα και κυκλικό διάγραμμα του παραδείγματος 2.1

Σε ό,τι αφορά το ραβδόγραμμα που δείχνει τις συχνότητες των βαθμών που πήραν ξεχωριστά οι φοιτητές και οι φοιτήτριες, εργαζόμαστε ως εξής: Επειδή πρέπει να δώσουμε στο πρόγραμμα την πληροφορία ότι οι τιμές της μεταβλητής grades σχετίζονται με αυτές της μεταβλητής sex, πηγαίνουμε *Data* → *Split File* και επιλέγουμε το *Organize output by groups*. Ακολούθως κάνουμε *κλικ* στη μεταβλητή sex, τη μεταφέρουμε στο πάνελ *Groups based on* και ολοκληρώνουμε με *κλικ* στο *OK*.

Τώρα συνεχίζουμε όπως και στην προηγούμενη περίπτωση. Δηλαδή πηγαίνουμε *Analyze* → *Descriptive Statistics* → *Frequencies*, στο παράθυρο διαλόγου που ανοίγει επιλέγουμε τη μεταβλητή grades και τη μεταφέρουμε στο πάνελ *Variable(s)*. Ενεργοποιούμε την επιλογή *Display frequency tables* και απενεργοποιούμε όλες τις προηγούμενες επιλογές για μέτρα και διαγράμματα. Με *κλικ* στο *OK* θα πάρουμε τους πίνακες συχνοτήτων των βαθμών των φοιτητριών και φοιτητών (σχήμα 2.13).

Για να κάνουμε το κοινό ραβδόγραμμα φοιτητριών και φοιτητών πρέπει πρώτα να επαναφέρουμε τη μεταβλητή grades στην προηγούμενη κατάσταση, δηλαδή πριν την ομαδοποίηση των τιμών της. Έτσι πηγαίνουμε *Data* → *Split File* και επιλέγουμε τώρα το *Analyze all cases*. Για το ραβδόγραμμα πηγαίνουμε *Graphs* → *Legacy Dialogs* → *Bar* και επιλέγουμε τον τύπο *Clustered* και *Summaries for groups of cases* (σχήμα 2.14). Στο παράθυρο διαλόγου που ανοίγει μεταφέρουμε τη μεταβλητή grades στο πλαίσιο *Category Axis* και τη sex στο *Define Clusters by*. Η γραφική παράσταση που παίρνουμε μετά από μορφοποίηση δίνεται στο σχήμα 2.15.

grades

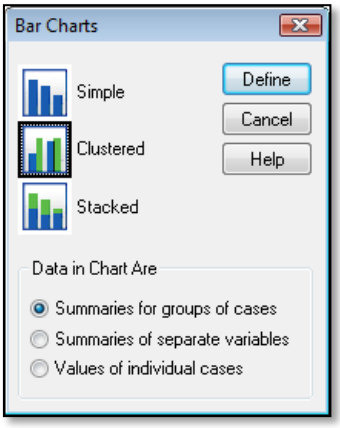
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	2	1	9,1	9,1	9,1
	4	1	9,1	9,1	18,2
	5	1	9,1	9,1	27,3
	6	3	27,3	27,3	54,5
	7	3	27,3	27,3	81,8
	8	1	9,1	9,1	90,9
	9	1	9,1	9,1	100,0
	Total	11	100,0	100,0	

sex = F

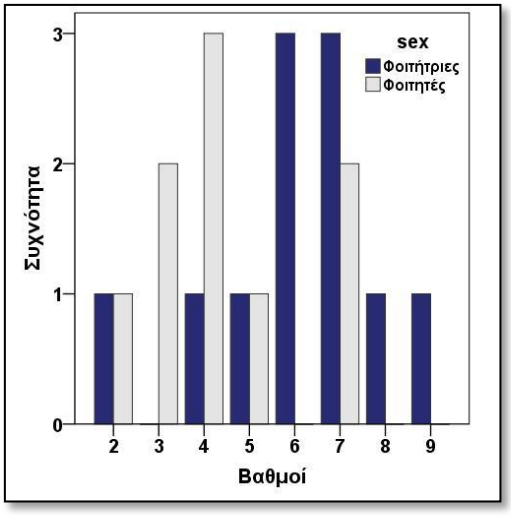
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	2	1	11,1	11,1	11,1
	3	2	22,2	22,2	33,3
	4	3	33,3	33,3	66,7
	5	1	11,1	11,1	77,8
	7	2	22,2	22,2	100,0
	Total	9	100,0	100,0	

sex = M

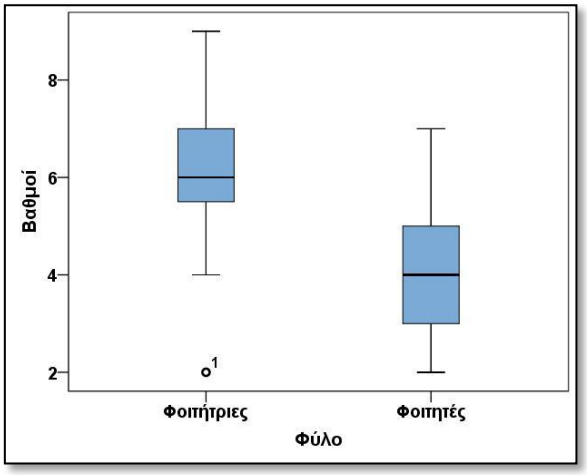
Σχήμα 2.13. Πίνακες συχνοτήτων των βαθμών των φοιτητριών (επάνω) και των φοιτητών (κάτω)



Σχήμα 2.14. Επιλογή τύπου ραβδογράμματος



Σχήμα 2.15. Ραβδόγραμμα κοινό φοιτητριών και φοιτητών



Σχήμα 2.16. Θηκογράμματα βαθμολογίας φοιτητριών και φοιτητών

Τέλος, για τα θηκογράμματα της βαθμολογίας φοιτητών-φοιτητριών πηγαίνουμε *Graphs* → *Legacy Dialogs* → *Boxplot* και επιλέγουμε τον τύπο *Simple* και *Summaries for groups of cases*. Στο παράθυρο διαλόγου που ανοίγει μεταφέρουμε τη μεταβλητή grades στο πλαίσιο *Variable* και τη

μεταβλητή sex στο *Category Axis*. Η γραφική παράσταση που παίρνουμε δίνεται στο σχήμα 2.16. Οι καλύτερες επιδόσεις των φοιτητριών φαίνονται σε όλα τα διαγράμματα αλλά κυρίως στα θηκογράμματα.

Παράδειγμα 2.2

Ο πίνακας 2.4 περιέχει 15 τιμές δυναμικού ενός στοιχείου Weston, που καταγράφηκαν σε ηλεκτρονικό υπολογιστή με τη βοήθεια κάρτας A/D. Να γίνει το ιστόγραμμα και το θηκόγραμμα των τιμών.

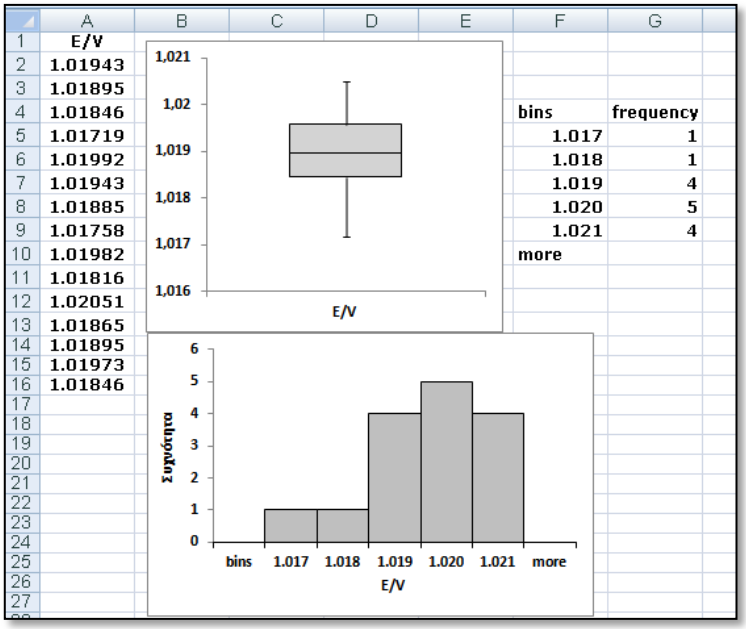
Πίνακας 2.4. Τιμές δυναμικού (σε Volt) στοιχείου Weston.

1.019434	1.017187	1.018848	1.018164	1.018945
1.018945	1.019922	1.017578	1.020508	1.019727
1.018457	1.019434	1.019824	1.018652	1.018457

❖ **Ανάλυση στο ChemStat**

Για να εργαστούμε στο *ChemStat* μεταφέρουμε τα δεδομένα σε μία στήλη του *Excel* (σχήμα 2.17) και ακολούθως πηγαίνουμε *Πρόσθετα (Add-ins)* → *ChemStat* → *Graphs* → *Boxplot*. Συμπληρώνουμε κατάλληλα τα παράθυρα που ανοίγουν και παίρνουμε το θηκόγραμμα του σχήματος 2.17.

Για το ιστόγραμμα εργαζόμαστε ανάλογα, δηλαδή πηγαίνουμε *Πρόσθετα (Add-ins)* → *ChemStat* → *Graphs* → *Histogram* και συμπληρώνουμε κατάλληλα τα παράθυρα που ανοίγουν. Συγκεκριμένα στο πρώτο πλαίσιο εισαγωγής δεδομένων εισάγουμε με το ποντίκι ΜΟΝΟ τις τιμές του δείγματος, στο επόμενο πλαίσιο αφήνουμε τη μονάδα ώστε στο γράφημα να χρησιμοποιηθούν *Συχνότητες*, ακολούθως μπορούμε να εισάγουμε το πλήθος των κλάσεων ή να αφήσουμε το πρόγραμμα να υπολογίσει αυτό το πλήθος, στη συνέχεια ορίζουμε το κελί εξόδου των αποτελεσμάτων και στο τελευταίο πλαίσιο επιλέγουμε αν το πρόγραμμα θα προσθέσει την καμπύλη της *κανονικής (normal)* κατανομής ή της *λογαριθμοκανονικής (log-normal)* κατανομής. Οι κατανομές αυτές θα εξεταστούν στο επόμενο κεφάλαιο. Το πρόγραμμα υπολογίζει τις κλάσεις και τις συχνότητες, τα δεδομένα αυτά εξάγονται στο φύλλο εργασίας και με βάση αυτά κατασκευάζεται το ιστόγραμμα.



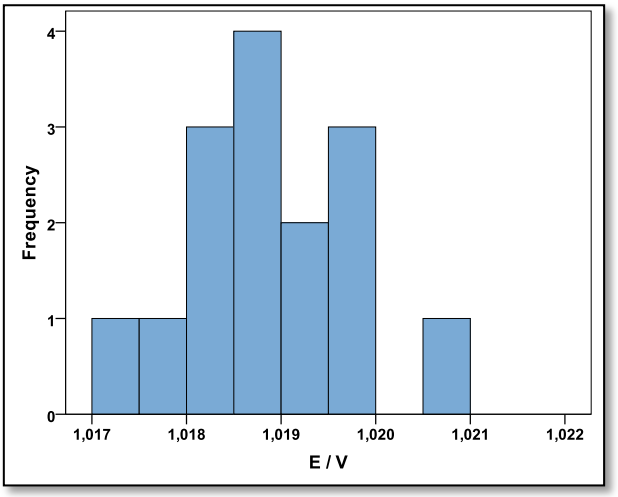
Σχήμα 2.17. Θηκόγραμμα και ιστόγραμμα του δείγματος των τιμών του δυναμικού στο *Excel*

Το ιστόγραμμα του παραδείγματος δίνεται επίσης στο σχήμα 2.17. Τρεις είναι οι βασικές μορφοποιήσεις που απαιτούν τα ιστογράμματα του *ChemStat*. Πρώτον αυξάνουμε το πλάτος των στηλών. Κάνουμε δεξί *κλικ* σε μία στήλη, επιλέγουμε στη λίστα που εμφανίζεται το *Μορφοποίηση σειράς δεδομένων (Format Data Series)* και στο παράθυρο που ανοίγει ρυθμίζουμε το πλάτος από το *Επιλογές σειράς (Series Options)* → *Πλάτος ανοίγματος (Gap Width)*. Δεύτερον ελαττώνουμε τα δεκαδικά του άξονα των x. Αυτό γίνεται αν μορφοποιήσουμε τους αριθμούς στην περιοχή των *κλάσεων (bins)*, δηλαδή στην περιοχή F5:F9 στο παράδειγμα του σχήματος 2.17. Τέλος, εισάγουμε τίτλους στους άξονες από τη λωρίδα εντολών *Διάταξη* (*Layout)*, όπως σε όλα τα γραφήματα του *Excel*.

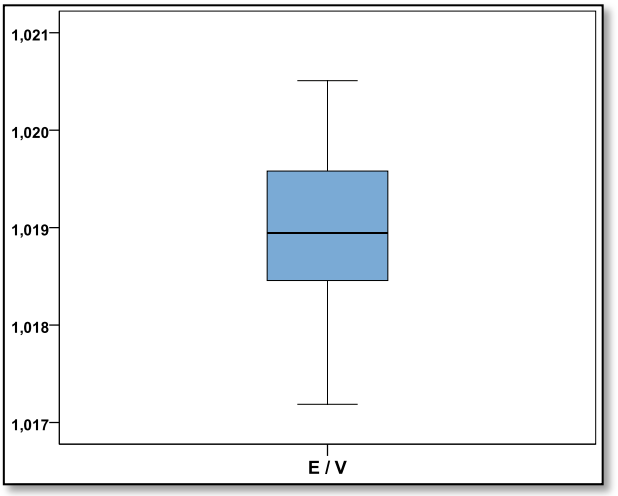
❖ Ανάλυση στο SPSS

Μεταφέρουμε τα δεδομένα σε μία στήλη ενός φύλλου του *SPSS*, την ονομάζουμε έστω E, και στο *Label* πληκτρολογούμε E / V. Ακολούθως πηγαίνουμε *Graphs* → *Legacy Dialogs* → *Histogram* και στο παράθυρο διαλόγου που ανοίγει μεταφέρουμε τη μεταβλητή E στο πλαίσιο *Variable*

και πατάμε *OK*. Η γραφική παράσταση που παίρνουμε μετά από μορφοποίηση δίνεται στο σχήμα 2.18.



Σχήμα 2.18. Ιστόγραμμα του δείγματος των τιμών του δυναμικού στο *SPSS*



Σχήμα 2.19. Θηκόγραμμα του δείγματος των τιμών του δυναμικού στο *SPSS*

Παρατηρούμε ότι τα δύο ιστογράμματα, του *ChemStat* και του *SPSS*, είναι αρκετά διαφορετικά. Αυτό παρατηρείται πάντα όταν έχουμε μικρά δείγματα και διαφορετικό αριθμό κλάσεων στα ιστογράμματα.

Για το θηκόγραμμα πηγαίνουμε *Graphs* → *Legacy Dialogs* → *Boxplot* και επιλέγουμε *Simple* και *Summaries of separate variable*. Στο παράθυρο διαλόγου που ανοίγει μεταφέρουμε τη μεταβλητή Ε στο πλαίσιο *Boxes Represent* και πατάμε *OK*. Η γραφική παράσταση μετά από μορφοποίηση δίνεται στο σχήμα 2.19.

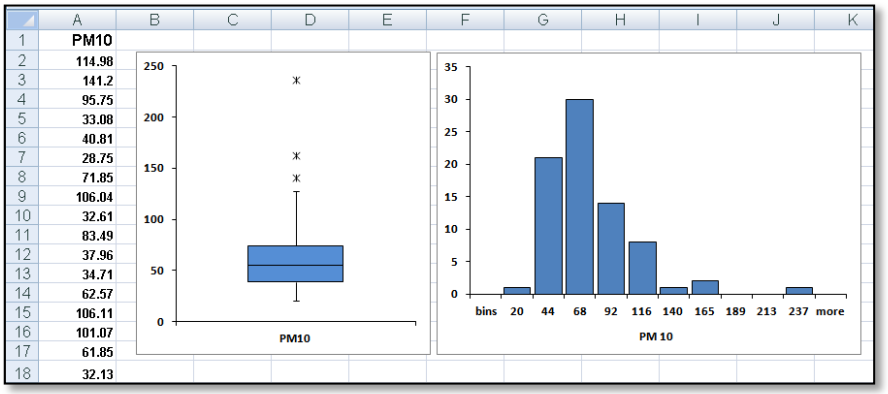
Παράδειγμα 2.3

Στον πίνακα 2.5 δίνονται οι αέριοι ρύποι στον σταθμό της Κοζάνης κατά τη διάρκεια 2006-2007. Να γίνει το ιστόγραμμα και το θηκόγραμμα των μετρήσεων.

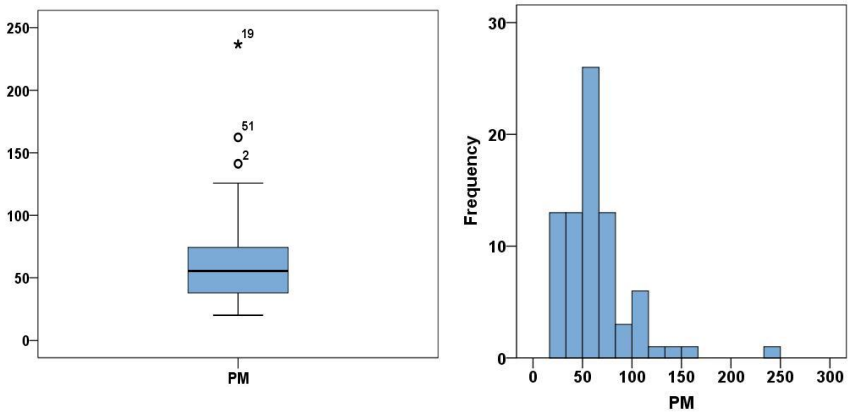
Πίνακας 2.5. Τιμές ρύπων στον σταθμό της Κοζάνης.

Ημερομηνία	PM10 μg/m^3	Ημερομηνία	PM10 μg/m^3	Ημερομηνία	PM10 μg/m^3	Ημερομηνία	PM10 μg/m^3
15/12/2006	114.98	13/2/2007	62.94	13/6/2007	37.60	20/8/2007	49.52
17/12/2006	141.20	17/2/2007	47.14	15/6/2007	69.07	22/8/2007	73.04
19/12/2006	95.75	19/2/2007	41.37	17/6/2007	33.08	24/8/2007	80.99
21/12/2006	33.08	21/2/2007	65.56	19/6/2007	73.58	26/8/2007	58.22
23/12/2006	40.81	23/2/2007	70.44	21/6/2007	76.11	28/8/2007	33.63
25/12/2006	28.75	1/3/2007	101.26	13/7/2007	31.82	1/9/2007	76.86
27/12/2006	71.85	3/3/2007	55.94	15/7/2007	52.79	3/9/2007	53.33
29/12/2006	106.04	5/3/2007	34.69	17/7/2007	125.77	9/9/2007	20.07
31/12/2006	32.61	7/3/2007	52.79	19/7/2007	82.62	11/9/2007	49.18
2/1/2007	83.49	9/3/2007	79.35	21/7/2007	107.03	13/9/2007	22.78
4/1/2007	37.96	11/3/2007	53.78	25/7/2007	162.36	17/9/2007	51.88
6/1/2007	34.71	17/3/2007	61.77	27/7/2007	56.77	19/9/2007	74.31
8/1/2007	62.57	19/3/2007	77.03	31/7/2007	98.17	21/9/2007	30.73
10/1/2007	106.11	21/3/2007	37.85	2/8/2007	54.23	23/9/2007	27.84
12/1/2007	101.07	23/3/2007	68.56	4/8/2007	58.93	25/9/2007	62.92
14/1/2007	61.85	25/3/2007	32.20	6/8/2007	30.19	27/9/2007	52.97
5/2/2007	32.13	27/3/2007	54.68	8/8/2007	54.06	29/9/2007	54.06
7/2/2007	53.94	29/3/2007	62.58	10/8/2007	49.36	1/10/2007	54.96
9/2/2007	236.76	31/3/2007	59.76	12/8/2007	29.83		
11/2/2007	51.79	11/6/2007	56.42	14/8/2007	37.42		

◆ Εργαζόμαστε όπως και στο προηγούμενο παράδειγμα. Συνεπώς μεταφέρουμε τα δεδομένα μόνο των ρύπων, δηλαδή τις τιμές του PM10 σε μία στήλη του *Excel/SPSS* και κατασκευάζουμε τα γραφήματα από *Πρόσθετα → ChemStat → Graphs → Histogram (ή Boxplot)* στο *Excel* (σχήμα 2.20) ή από *Graphs → Legacy Dialogs → Histogram (ή Boxplot)* στο *SPSS* (σχήμα 2.21).



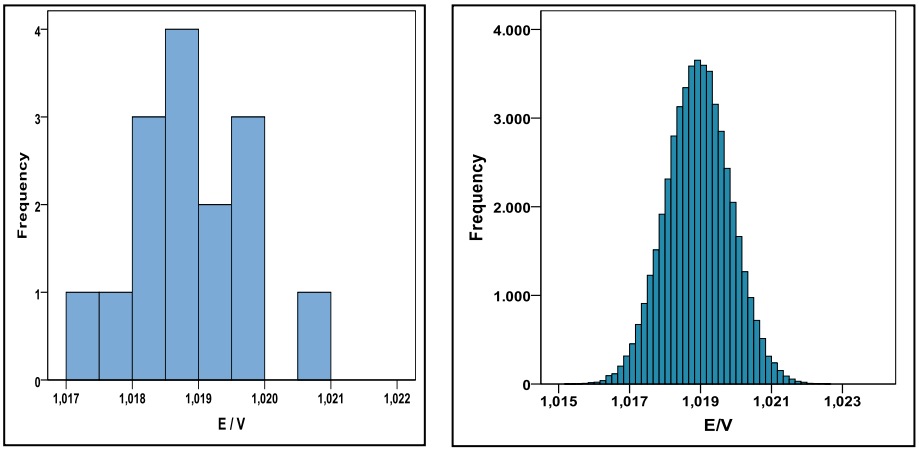
Σχήμα 2.20. Θηκόγραμμα και ιστόγραμμα του δείγματος των ρύπων στο *Excel*



Σχήμα 2.21. Θηκόγραμμα και ιστόγραμμα του δείγματος των ρύπων στο *SPSS*

Στα σχήματα 2.18 και 2.20, 2.21 βλέπουμε ότι οι τιμές του δυναμικού και των ρύπων ακολουθούν διαφορετικές ***κατανομές*** *(distributions).* Οι τιμές του δυναμικού φαίνεται να κατανέμονται συμμετρικά γύρω από τη μέση τιμή, ενώ αντίθετα η κατανομή των ρύπων είναι ασύμμετρη. Η συμμετρία ή ασυμμετρία της κατανομής των δεδομένων φαίνεται επίσης χαρακτηριστικά στα θηκογράμματα (σχήματα 2.19 - 2.21), τα οποία επηρεάζονται λιγότερο από το πλήθος των τιμών του δείγματος.

Τα ιστογράμματα μας δείχνουν εποπτικά πως κατανέμονται οι τιμές ενός δείγματος. Όμως όταν το δείγμα είναι σχετικά μικρό, η εικόνα μπορεί να είναι πλασματική. Η επίδραση του μεγέθους του δείγματος στην ποιότητα ενός ιστογράμματος φαίνεται στο σχήμα 2.22. Σε αυτό δίνονται τα ιστογράμματα δύο δειγμάτων με 15 (αριστερά) και 50000 (δεξιά) τιμές δυναμικού.



Σχήμα 2.22. Ιστογράμματα δείγματος 15 (αριστερά) και 50000 (δεξιά) τιμών δυναμικού ενός στοιχείου Weston

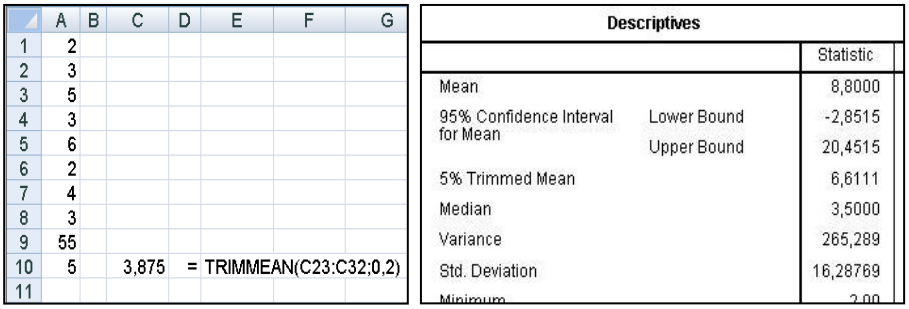
Παράδειγμα 2.4

Να υπολογιστεί η ισοσταθμισμένη μέση τιμή του δείγματος {2, 3, 5, 3, 6, 2, 4, 3, 55, 5}.

◆ Αν στο *Excel* χρησιμοποιήσουμε τη συνάρτηση *TRIMMEAN* με *percent* ίσο με 0.2 (20%), τότε αποκόπτονται δύο τιμές, η μικρότερη, το 2,

και η μεγαλύτερη, το 55. Έτσι προκύπτει η τιμή 3.875 για τον ισοσταθμισμένο μέσο (σχήμα 2.23-αριστερά).

Στο *SPSS* πηγαίνουμε *Analyze* → *Descriptive Statistics* → *Explore* και στο πλαίσιο *Statistics* επιλέγουμε *Descriptive*. Το πρόγραμμα προσδιορίζει την ισοσταθμισμένη μέση τιμή στο 5% που αντιστοιχεί σε μισή τιμή (10*0.05 = 0.5). Έτσι αντικαθιστά τη μικρότερη και τη μεγαλύτερη τιμή, 2 και 55, με το ημιάθροισμά τους, (2+55)/2 = 28.5, με αποτέλεσμα να παίρνουμε το αποτέλεσμα 6.611 (σχήμα 2.23-δεξιά).



	A	B	C	D	E	F	G
1	2						
2	3						
3	5						
4	3						
5	6						
6	2						
7	4						
8	3						
9	55						
10	5		3,875	= TRIMMEAN(C23:C32;0,2)			
11							

Descriptives

		Statistic
Mean		8,8000
95% Confidence Interval for Mean	Lower Bound	-2,8515
	Upper Bound	20,4515
5% Trimmed Mean		6,6111
Median		3,5000
Variance		265,289
Std. Deviation		16,28769
Minimum		2,00

Σχήμα 2.23. Αποτελέσματα υπολογισμού ισοσταθμισμένης μέσης τιμής στο *Excel* (αριστερά) και στο *SPSS* (δεξιά)

ΑΣΚΗΣΕΙΣ

2.1. Ποιο είναι το καλύτερο μέτρο θέσης για τα δεδομένα στα παρακάτω δείγματα:

α) Αέριοι ρύποι ($\mu g/m^3$): 70 36 114 27 26 31 52 77 38 66 121

β) Τιμές pH ενός διαλύματος: 3.11 3.05 3.08 3.21 3.15 3.12

γ) Τιμές χοληστερόλης LDL (mg/dL): 146 155 161 182 158 164 140

δ) Τιμές φρουκτόζης σε μήλα (g/L): 51 44 48 42 38 40 58

2.2. Χωρίς υπολογισμούς, ποιό από τα παρακάτω δείγματα έχει την μεγαλύτερη τυπική απόκλιση;

α) 98 99 100 101 102, β) 2 4 6 8 10, γ) 2 10

2.3. Η συγκέντρωση της γλυκόζης σε δύο διαφορετικές παρτίδες σύκων (σε g ανά 100 g φρούτου) δίνεται στους παρακάτω πίνακες. Να γίνουν τα ιστογράμματα και τα θηκογράμματα και να συζητηθούν ως προς την κατανομή των δεδομένων, τη διασπορά των τιμών και την ύπαρξη ακραίων τιμών.

31 30 30 30 30 29 31 20	26 35 23 27 26 25 33 27
30 31 29 29 30 31 29 19	26 25 29 30 28 24 28 26
29 30 28 29 29 29 29 24	23 21 24 32 27 23 24 27
28 29 27 29 28 29 29 24	24 23 31 32 35 23 23 25
20 29 22 25 20 19 28 22	28 28 29 29 30 27 30 29
24 22 23 25 26 21 22 18	25 24 25 24 29 26 28 37
20 16 23 26 22 18 21 20	29 31 23 26 29 31 26 25

2.4. Στους παρακάτω πίνακες δίνονται οι αέριοι ρύποι (σε $\mu g/m^3$) στην πλατεία της Αγίας Σοφίας και στο Κορδελιό. Να προσδιοριστούν τα διάφορα περιγραφικά μέτρα, να γίνουν τα θηκογράμματα και να συζητηθεί ποια περιοχή έχει τη μεγαλύτερη ρύπανση.

Αγία Σοφία		Κορδελιό	
81.67	78.33	114.98	37.96
48.75	37.92	141.20	34.71
40.83	16.67	95.75	62.57
29.17	29.17	33.08	106.11
20.42	56.67	40.81	101.07
50.42	50.42	28.75	61.85
28.75	36.25	71.85	32.13
117.08	16.25	106.04	53.94
108.75	45.83	32.61	236.76
55.00	75.00	83.49	51.79

2.5. Στον παρακάτω πίνακα δίνεται η μέση τιμή των συγκεντρώσεων της σουρκόζης (Su), γλυκόζης (Gl), φρουκτόζης (Fr) και σορβιτόλης (So) σε τρεις ποικιλίες μήλων. Να γίνει το ραβδόγραμμα και τα κυκλικά γραφήματα.

Ποικιλία	c, g/L			
	Su	Gl	Fr	So
A	27	10	44	4
B	10	21	49	5
C	13	18	57	6

Κεφάλαιο 3

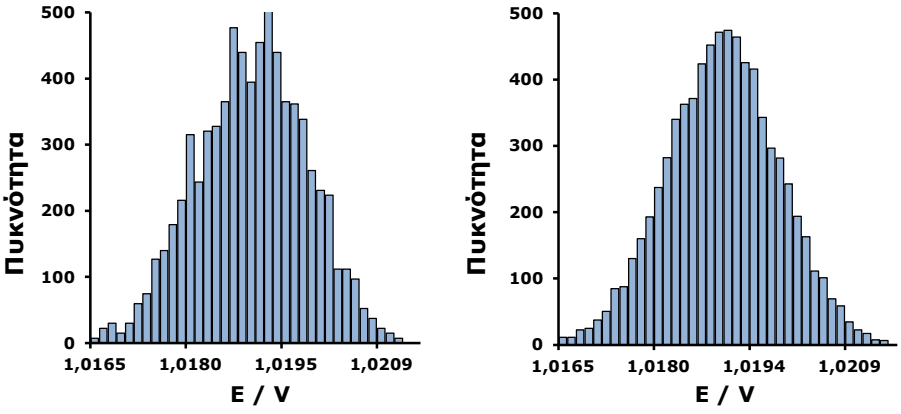
ΣΥΝΑΡΤΗΣΕΙΣ ΚΑΤΑΝΟΜΗΣ

3.1 Η ΕΝΝΟΙΑ ΤΗΣ ΣΥΝΑΡΤΗΣΗΣ ΚΑΤΑΝΟΜΗΣ

Όπως αναφέρθηκε, τα ιστογράμματα δείχνουν εποπτικά τον τρόπο με τον οποίον κατανέμονται οι τιμές ενός δείγματος. Επιπλέον μπορούν να αποκτήσουν ενδιαφέρουσες ιδιότητες αν χρησιμοποιήσουμε στον άξονα των y αντί για τη *συχνότητα (Frequency)* την *πυκνότητα (Density)*, που ορίζεται ως

$$\textit{πυκνότητα} = \textit{συχνότητα}/(\mathrm{m}\Delta x) \qquad (3.1)$$

όπου m είναι το πλήθος των τιμών του δείγματος και Δx το μήκος κάθε κλάσης. Αυτή η δυνατότητα υπάρχει στο *ChemStat* αλλά όχι στο *SPSS*.

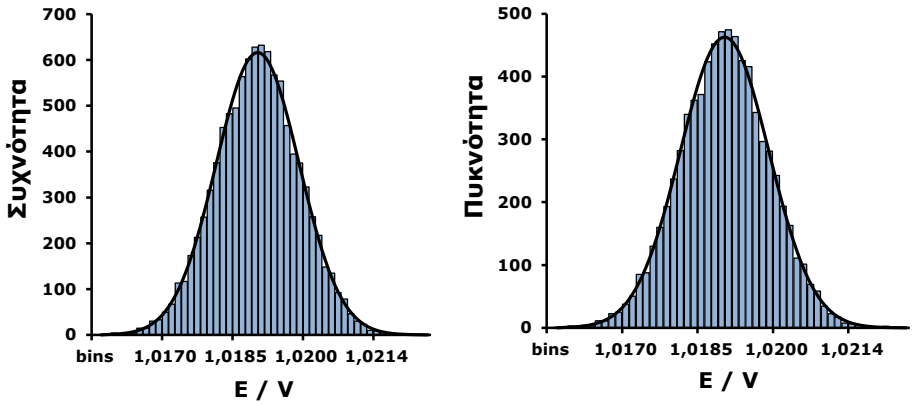


Σχήμα 3.1. Ιστογράμματα δείγματος 1000 (αριστερά) και 10000 (δεξιά) τιμών δυναμικού ενός στοιχείου Weston όταν στον άξονα των y χρησιμοποιείται η πυκνότητα

Χρησιμοποιώντας την *πυκνότητα* τα ιστογράμματα αποκτούν τις ακόλουθες τρεις βασικές ιδιότητες: α) Παύουν να εξαρτώνται από τον αριθμό των τιμών του δείγματος. Δηλαδή, όλα τα δείγματα με μετρήσεις της ίδιας μεταβλητής που έγιναν κάτω από τις ίδιες πειραματικές συνθήκες είναι ακριβώς ίδια (σχήμα 3.1), αν και τα ιστογράμματα δειγμάτων με λίγες τιμές είναι συνήθως παραμορφωμένα. β) Το εμβαδό τους, δηλαδή το εμβαδόν όλων των ορθογωνίων που σχηματίζουν το ιστόγραμμα γίνεται ίσο με 1. γ) Το εμβαδόν του ιστογράμματος στο διάστημα [a, b] δίνει την πιθανότητα που έχει η μεταβλητή του ιστογράμματος να βρίσκεται στο διάστημα αυτό:

$$E_{ab} = P(a \leq X \leq b) \qquad (3.2)$$

Όταν δοθεί ένα ιστόγραμμα, ανεξάρτητα αν στον άξονα των y έχουμε *συχνότητα* ή *πυκνότητα*, πάντα μπορούμε να βρούμε μια συνάρτηση y = f(x) που να το περιγράφει, όπως φαίνεται στα επόμενα σχήματα.



Σχήμα 3.2. Ιστογράμματα δείγματος 10000 τιμών δυναμικού ενός στοιχείου Weston όταν στον άξονα των y χρησιμοποιείται η συχνότητα (δεξιά) και η πυκνότητα (αριστερά) και η καμπύλη που τα περιγράφει

Όταν όμως το ιστόγραμμα εκφράζεται σε *πυκνότητα*, τότε η συνάρτηση y = f(x) που το περιγράφει αποκτά προφανώς τις ιδιότητες του ιστογράμματος. Συγκεκριμένα: α) η συνάρτηση y = f(x) είναι ανεξάρτητη από το πλήθος των τιμών του δείγματος και είναι πάντα θετική, β) το ολοκλήρωμα της f(x) από -∞ έως +∞ ισούται με τη μονάδα και γ) το ολοκλήρωμα της f(x) από a έως b ισούται με την πιθανότητα που έχει η

μεταβλητή X να βρίσκεται στο διάστημα [a, b]. Δηλαδή η συνάρτηση y = f(x) έχει τις ακόλουθες ιδιότητες:

α) $f(x) \geq 0$ (3.3)

β) $\int_{-\infty}^{\infty} f(x)dx = 1$ (3.4)

γ) $P(a \leq X \leq b) = \int_{a}^{b} f(x)dx$ (3.5)

Κάθε συνεχής συνάρτηση που έχει τις παραπάνω ιδιότητες ονομάζεται ***συνάρτηση πυκνότητας πιθανότητας*** *(probability density function)*. Στα συγκεκριμένα παραδείγματα αποδεικνύεται ότι η συνάρτηση που περιγράφει τα δεδομένα των δειγμάτων με τιμές δυναμικού είναι η

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} \quad (3.6)$$

με μέση τιμή μ = 1.01894 και τυπική απόκλιση σ = 0.0009 (σχήμα 3.2). Η σχέση αυτή είναι η συνάρτηση πυκνότητας πιθανότητας της ***κανονικής κατανομής*** *(normal distribution)*. Αντίθετα το περιβαλλοντικό δείγμα του πίνακα 2.5 περιγράφεται από τη συνάρτηση

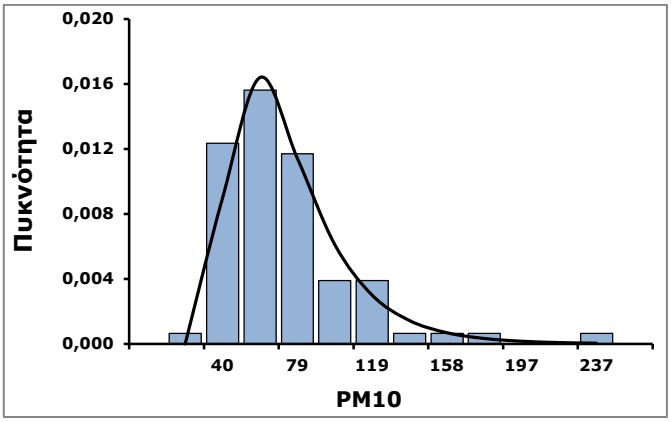
$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-(\ln x-\mu)^2/2\sigma^2} \quad (3.7)$$

με μέση τιμή μ = 4.04 και τυπική απόκλιση σ = 0.463 (σχήμα 3.3). Η σχέση αυτή είναι η συνάρτηση πυκνότητας πιθανότητας της ***λογαριθμοκανονικής κατανομής*** *(log-normal distribution)*.

Όταν γνωρίζουμε τη συνάρτηση πυκνότητας πιθανότητας μπορούμε να υπολογίσουμε την *αθροιστική συνάρτηση κατανομής (cumulative distribution function)* ή απλά τη ***συνάρτηση κατανομής*** *(distribution function)* από τη σχέση

$$F(x) = \int_{-\infty}^{x} f(t)dt = P(X \leq x) \quad (3.8)$$

Συνεπώς η *συνάρτηση κατανομής* μας δίνει την πιθανότητα μια μεταβλητή X να έχει τιμή μικρότερη ή ίση από την τιμή x.



Σχήμα 3.3. Ιστόγραμμα του περιβαλλοντικού δείγματος όταν στον άξονα των y χρησιμοποιείται η πυκνότητα. Η καμπύλη έχει σχεδιαστεί με τη βάση τη συνάρτηση (3.7) χρησιμοποιώντας μ = 4.04 και σ = 0.463

3.2 ΣΥΝΑΡΤΗΣΕΙΣ ΚΑΤΑΝΟΜΗΣ ΣΕ ΔΙΑΚΡΙΤΕΣ ΜΕΤΑΒΛΗΤΕΣ

Οι παραπάνω ορισμοί και σχέσεις αφορούν συνεχείς μεταβλητές, ενώ για διακριτές μεταβλητές ισχύουν τα ακόλουθα. Μία συνάρτηση f(x) είναι ***συνάρτηση πιθανότητας*** *(probability function)* της τυχαίας διακριτής μεταβλητής X που παίρνει τιμές x_1, x_2, x_3, …, όταν η τιμή της $f(x_i)$ για κάθε i = 1, 2, 3, … μας δίνει την πιθανότητα η τυχαία μεταβλητή X να πάρει την τιμή x_i. Δηλαδή

$$f(x_i) = P(X = x_i) \quad \text{για κάθε } i = 1, 2, 3, \ldots \tag{3.9}$$

Η *αθροιστική συνάρτηση κατανομής* ή απλά ***συνάρτηση κατανομής*** της διακριτής τυχαίας μεταβλητής X είναι μια συνάρτηση F(x) για την οποία ισχύει

$$F(x_i) = P(X \leq x_i) \quad \text{για κάθε } i = 1, 2, 3, \ldots \tag{3.10}$$

όπου $P(X \leq x_i)$ δηλώνει την πιθανότητα η X να πάρει μια οποιαδήποτε τιμή μικρότερη ή ίση της x_i. Αν είναι γνωστή η f(x), τότε ισχύει

$$F(x_i) = f(x_1) + f(x_2) + \ldots + f(x_i) \tag{3.11}$$

3.3 ΒΑΣΙΚΕΣ ΚΑΤΑΝΟΜΕΣ

Υπάρχει μια πολύ μεγάλη ποικιλία κατανομών από τις οποίες οι κυριότερες είναι:

3.3.1 ΔΙΑΚΡΙΤΕΣ ΚΑΤΑΝΟΜΕΣ

❖ **ΔΙΩΝΥΜΙΚΗ ΚΑΤΑΝΟΜΗ**, B(n,p)

Έστω ότι εκτελούμε ένα πείραμα τύχης n φορές και κάθε φορά ένα γεγονός μπορεί να πραγματοποιηθεί με πιθανότητα p ή να μην πραγματοποιηθεί με πιθανότητα q = 1 - p. Η πιθανότητα να πραγματοποιηθεί το γεγονός αυτό ακριβώς x φορές, δηλαδή να έχουμε x επιτυχίες και n - x αποτυχίες, δίνεται από τη συνάρτηση πιθανότητας της *διωνυμικής κατανομής*

$$f(x) = P(X = x) = \binom{n}{x} p^x q^{n-x} = \frac{n!}{x!(n-x)!} p^x q^{n-x} \qquad (3.12)$$

Για τη μέση τιμή και τη διασπορά αυτής της κατανομής ισχύει μ = np και σ = npq.

Παράδειγμα 3.1

Σε ένα διαγώνισμα δίνονται 50 ερωτήσεις του τύπου σωστό-λάθος. Αν ένας φοιτητής απαντά στην τύχη, ποια η πιθανότητα να περάσει το μάθημα όταν η βάση είναι το 25; Τι θα συμβεί αν υπάρχει τετραπλή επιλογή σε κάθε ερώτηση;

◆ Αν X είναι η τυχαία μεταβλητή που εκφράζει το πλήθος των σωστών απαντήσεων, τότε ζητάμε να υπολογίσουμε την πιθανότητα $P(X \geq 25)$ όταν η X ακολουθεί τη *διωνυμική* κατανομή με n = 50 και p = q = 1/2 στην πρώτη περίπτωση. Έχουμε

$$P(X \geq 25) = 1 - P(X < 25) =$$

$$1 - \sum_{i=0}^{24} P(X = i) = 1 - \sum_{i=0}^{24} \frac{50!}{i!(50-i)!} 0.5^i 0.5^{50-i} = 1 - \sum_{i=0}^{24} \frac{0.5^{50}\, 50!}{i!(50-i)!}$$

Το άθροισμα αυτό υπολογίζεται εύκολα αν χρησιμοποιήσουμε το *Excel* και τη συνάρτηση *FACT(n)* για τον υπολογισμό του n!. Παίρνουμε

$$P(X \geq 25) = 0.5561$$

Δηλαδή ο φοιτητής, αν και αδιάβαστος, έχει την πολύ σημαντική πιθανότητα 55.6% να περάσει το μάθημα.

Όταν υπάρχουν 4 επιλογές σε κάθε ερώτηση, η πιθανότητα να απαντήσει σωστά σε μια ερώτηση ίση με 1/4 και λανθασμένα ίση με 3/4. Συνεπώς

$$P(X\geq 25) = 1 - P(X<25) = 1 - \sum_{i=0}^{24} P(X=i) = 1 - \sum_{i=0}^{24} \frac{50!}{i!(50-i)!}(1/4)^i(3/4)^{50-i}$$

από την οποία παίρνουμε

$$P(X\geq 25) = 0.00012$$

Δηλαδή, έχει μηδενική πλέον πιθανότητα να περάσει το μάθημα.

❖ **ΚΑΤΑΝΟΜΗ POISSON**, P(λ)

Πρόκειται για την κατανομή των σπάνιων γεγονότων και χρησιμοποιείται όταν θέλουμε να μετρήσουμε τον αριθμό των "συμβάντων" στη μονάδα μέτρησης, όπου η μονάδα μέτρησης μπορεί να είναι χρόνος, εμβαδόν, όγκος, πληθυσμός, κ.ά. Ένα γεγονός θεωρείται σπάνιο αν η πιθανότητά του είναι μικρότερη από 1/10. Τότε η πιθανότητα να πραγματοποιηθεί το γεγονός αυτό ακριβώς x φορές δίνεται από τη συνάρτηση πιθανότητας της *κατανομής Poisson*

$$f(x) = P(X=x) = e^{-\lambda}\frac{\lambda^x}{x!} \qquad (3.13)$$

Για τη μέση τιμή και τη διασπορά της κατανομής ισχύει: μ = λ και σ = λ, όπου ο λ είναι πάντα θετικός αριθμός και ίσος με τον μέσο αριθμό περιστατικών που εμφανίζονται κατά τη διάρκεια της μέτρησης. Για παράδειγμα, εάν τα γεγονότα εμφανίζονται κατά μέσον όρο 4 φορές ανά λεπτό και ενδιαφερόμαστε για τον αριθμό των γεγονότων που εμφανίζονται σε ένα διάστημα 10 λεπτών, τότε θα χρησιμοποιήσουμε την κατανομή *Poisson* με λ = 10·4 = 40.

Παράδειγμα 3.2

Έστω ότι ένα νέο φάρμακο έχει πιθανότητα 0.001 να δημιουργήσει σοβαρές παρενέργειες. Αν χορηγηθεί σε 2000 ασθενείς ποια η πιθανότητα να παρουσιαστούν παρενέργειες το πολύ σε 3 ασθενείς;

◆ Επειδή η εμφάνιση μιας παρενέργειας είναι σπάνιο γεγονός, θα εφαρμόσουμε την κατανομή *Poisson*. Ο αναμενόμενος αριθμός

περιστατικών με σοβαρές παρενέργειες είναι 0.001·2000 = 2 και συνεπώς θα έχουμε λ = 2. Έστω τώρα X η τυχαία μεταβλητή που δείχνει πόσες φορές παρουσιάζονται παρενέργειες. Θα έχουμε

$$P(X \le 3) = P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) =$$

$$e^{-2}\frac{2^0}{0!} + e^{-2}\frac{2^1}{1!} + e^{-2}\frac{2^2}{2!} + e^{-2}\frac{2^3}{3!} = 0.135 + 0.271 + 0.271 + 0.18$$

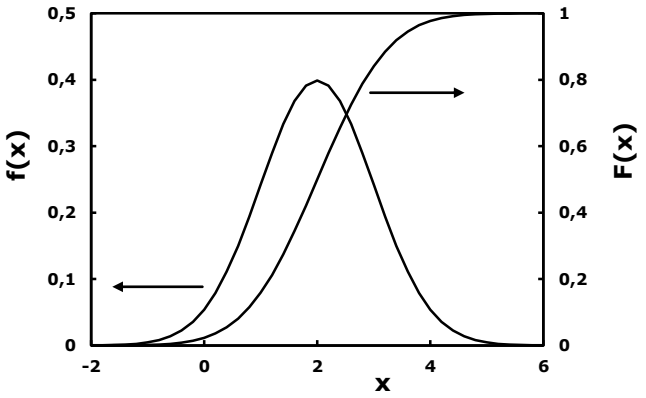
και τελικά

$$P(X \le 3) = 0.857$$

3.3.2 ΣΥΝΕΧΕΙΣ ΚΑΤΑΝΟΜΕΣ

❖ **ΚΑΝΟΝΙΚΗ ΚΑΤΑΝΟΜΗ,** $N(\mu,\sigma^2)$

Πρόκειται για μια από τις σημαντικότερες κατανομές με εφαρμογές στη θεωρία πειραματικών σφαλμάτων, δεδομένου ότι τα τυχαία σφάλματα στις πειραματικές μετρήσεις ακολουθούν συνήθως αυτή την κατανομή. Όπως έχει αναφερθεί, η συνάρτηση πυκνότητας πιθανότητας δίνεται από τη σχέση (3.6). Η γραφική παράσταση της συνάρτησης αυτής καθώς επίσης και της αντίστοιχης αθροιστικής συνάρτησης κατανομής F(x) δίνεται στο σχήμα 3.4.



Σχήμα 3.4. Γραφική παράσταση της κανονικής κατανομής όταν μ=2 και σ=1

Όταν μ = 0 και σ = 1, τότε η σχέση (3.6) ανάγεται στην

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \qquad (3.14)$$

που είναι η ***τυπικά κανονική κατανομή*** (*standard normal distribution*) και συμβολίζεται με N(0,1). Αποδεικνύεται εύκολα ότι αν η τυχαία μεταβλητή X ακολουθεί την κανονική κατανομή $N(\mu,\sigma^2)$, τότε η τυχαία μεταβλητή Z = (X-μ)/σ ακολουθεί την N(0,1) κατανομή.

Αποδεικνύεται επίσης ότι αν δύο ανεξάρτητες τυχαίες μεταβλητές, X_1 και X_2, ακολουθούν κανονικές κατανομές, $N(\mu_1,\sigma_1^2)$ και $N(\mu_2,\sigma_2^2)$, τότε η τυχαία μεταβλητή $X = X_1 \pm X_2$ ακολουθεί την κανονική κατανομή $N(\mu,\sigma^2)$ με $\mu = \mu_1 \pm \mu_2$ και $\sigma^2 = \sigma_1^2 + \sigma_2^2$. Το θεώρημα αυτό επεκτείνεται και σε περισσότερες από δύο τυχαίες μεταβλητές όταν αυτές προστίθενται.

❖ ΛΟΓΑΡΙΘΜΟΚΑΝΟΝΙΚΗ ΚΑΤΑΝΟΜΗ

Πρόκειται για την κατανομή που έχει συνάρτηση πυκνότητας πιθανότητας τη σχέση (3.7). Όπως αναφέρθηκε, παρουσιάζει ιδιαίτερο ενδιαφέρον επειδή ισχύει κατά κανόνα σε περιβαλλοντικά και βιολογικά δείγματα. Αν λογαριθμίσουμε τις τιμές ενός δείγματος που ακολουθεί την *λογαριθμοκανονική (log-normal) κατανομή*, το δείγμα που θα προκύψει ακολουθεί την *κανονική* κατανομή.

❖ ΚΑΤΑΝΟΜΗ t ή STUDENT, t_v

Πρόκειται για μια πολύ χρήσιμη κατανομή με εκτεταμένες εφαρμογές, κυρίως στον προσδιορισμό διαστημάτων εμπιστοσύνης και στη σύγκριση των μέσων τιμών δύο δειγμάτων, όπως θα δούμε στα επόμενα κεφάλαια. Η κατανομή *Student* ή απλά η κατανομή t με ν βαθμούς ελευθερίας έχει συνάρτηση πυκνότητας πιθανότητας την

$$f_V(x) = \frac{\Gamma(\frac{v+1}{2})}{\sqrt{v\pi}\,\Gamma(v/2)} \left(1 + \frac{x^2}{v}\right)^{-(v+1)/2} \qquad (3.15)$$

όπου Γ(p) είναι η συνάρτηση *γάμμα (gamma)*:

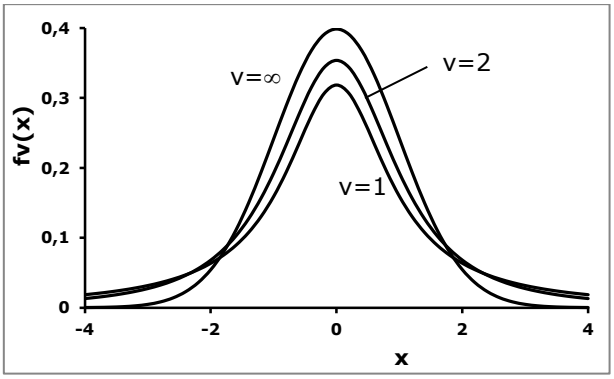
$$\Gamma(p) = \int_0^\infty u^{p-1} e^{-u}\, du \qquad (3.16)$$

Η κατανομή έχει το παράξενο όνομα *Student* (= *Σπουδαστής*) για

τον ακόλουθο λόγο. Ο μαθηματικός *William Gosset* (1876-1937) ήταν υπάλληλος της ζυθοποιίας *Guiness* όταν μελέτησε την κατανομή αυτή. Δεδομένου ότι το ιρλανδικό ζυθοποιείο δεν του επέτρεψε τη δημοσίευση των ερευνητικών του αποτελεσμάτων, τα δημοσίευσε το 1908 με το ψευδώνυμο σπουδαστής, επειδή ήταν φοιτητής.

William Sealy Gosset
(1876-1937)

Η γραφική παράσταση της συνάρτησης σε σύγκριση με την κανονική κατανομή δίνεται στο σχήμα 3.5. Η καμπύλη της κατανομής *student* με $\nu \geq 30$ πρακτικά ταυτίζεται με την καμπύλη της τυπικά κανονικής κατανομής.



Σχήμα 3.5. Γραφική παράσταση της κατανομής *student* όταν $\nu = 1, 2, \infty$. Η καμπύλη με $\nu = \infty$ ταυτίζεται με την τυπικά κανονική κατανομή

❖ ΚΑΤΑΝΟΜΗ Χι-ΤΕΤΡΑΓΩΝΟ, χ^2_v

Έστω ότι οι ανεξάρτητες τυχαίες μεταβλητές Z_1, Z_2, …, Z_v ακολουθούν την κανονική κατανομή με μέση τιμή 0 και διασπορά 1. Τότε η τυχαία μεταβλητή

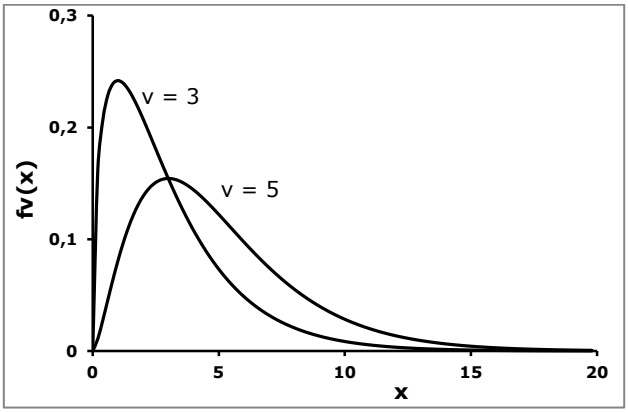
$$\chi^2 = Z_1^2 + Z_2^2 + \cdots + Z_v^2 \qquad (3.17)$$

ακολουθεί μια κατανομή με συνάρτηση πυκνότητας πιθανότητας:

$$f_v(x) = \begin{cases} \dfrac{x^{(v/2)-1} e^{-x/2}}{2^{v/2}\,\Gamma(v/2)} & x > 0 \\ 0 & x \le 0 \end{cases} \qquad (3.18)$$

Η κατανομή αυτή ονομάζεται *χι-τετράγωνο* με ν βαθμούς ελευθερίας και έχει: μ = ν και $\sigma^2 = 2v$. Όπως η κατανομή *student*, έτσι και η χ^2_v βρίσκει εφαρμογές στον έλεγχο στατιστικών υποθέσεων, κυρίως κατηγορικών δεδομένων (κεφάλαιο 9). Η γραφική παράσταση της συνάρτησης όταν ν = 3 και ν = 5 δίνεται στο σχήμα 3.6.

Αποδεικνύεται ότι αν οι ανεξάρτητες τυχαίες μεταβλητές, X_1, X_2, …, X_n ακολουθούν η κάθε μία κατανομή χ^2 με v_1, v_2, …, v_n βαθμούς ελευθερίας, αντίστοιχα, τότε η τυχαία μεταβλητή $Z = X_1 + X_2 + \ldots + X_n$ ακολουθεί επίσης την κατανομή χ^2 με $v = v_1 + v_2 + \ldots + v_n$ βαθμούς ελευθερίας.



Σχήμα 3.6. Γραφική παράσταση της κατανομής χ^2 όταν ν = 3 και ν = 5

❖ ΚΑΤΑΝΟΜΗ F ή FISHER, F_{v_1,v_2}

Αν Y_1 και Y_2 είναι ανεξάρτητες τυχαίες μεταβλητές που ακολουθούν την κατανομή χ^2 με v_1 και v_2 βαθμούς ελευθερίας, αντίστοιχα, τότε το κλάσμα

$$W = \frac{Y_1 / v_1}{Y_2 / v_2} \tag{3.19}$$

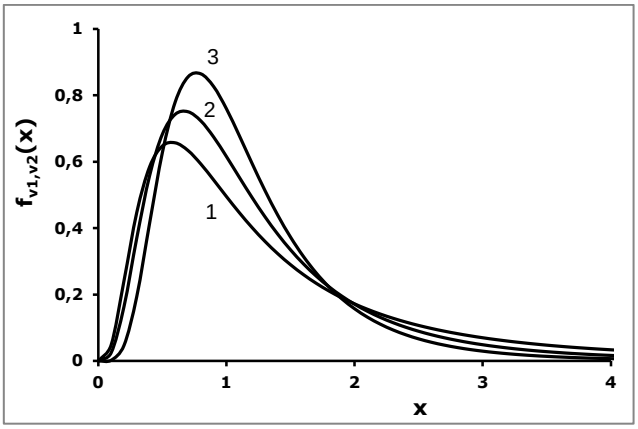
είναι τυχαία μεταβλητή που ακολουθεί μία κατανομή την οποία μελέτησε ο *Sir R. A. Fisher* και η οποία προς τιμή του συμβολίζεται με F και ονομάζεται κατανομή *Fischer* ή απλά κατανομή F με v_1 και v_2 βαθμούς ελευθερίας. Βρίσκει εφαρμογές κυρίως στον έλεγχο στατιστικών υποθέσεων που σχετίζονται με διασπορές, θέμα που εξετάζεται στο κεφάλαιο 8.

Sir Ronald Fisher
(1890-1962)

Η συνάρτηση πυκνότητας πιθανότητας δίνεται από τη σχέση

$$f_{v_1,v_2}(x) = \frac{\Gamma(\frac{v_1+v_2}{2}) \cdot \left(\frac{v_1}{v_2}\right)^{v_1/2}}{\Gamma(v_1/2)\Gamma(v_2/2)} \left(1+\frac{v_1 x}{v_2}\right)^{-(v_1+v_2)/2} x^{(v_1-2)/2} \tag{3.20}$$

Ενδεικτικές γραφικές παραστάσεις της συνάρτησης F δίνονται στο σχήμα 3.7.



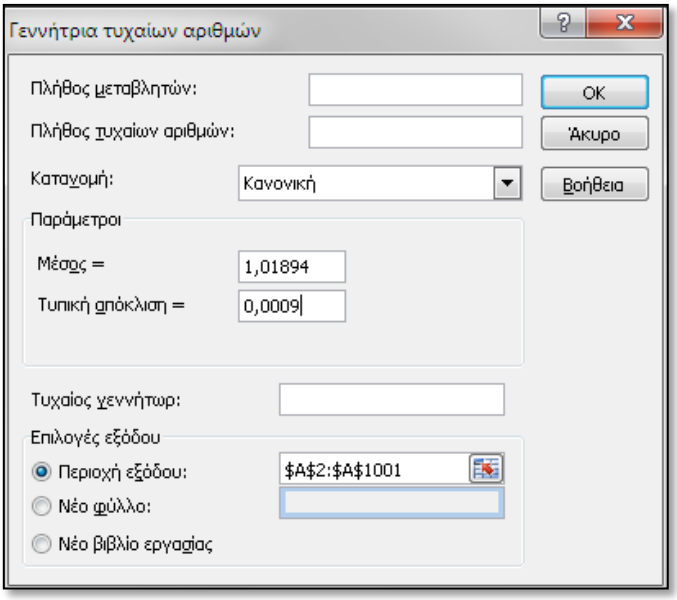
Σχήμα 3.7. Γραφική παράσταση της κατανομής F όταν (1) v_1=10, v_2=5, (2) v_1=10, v_2=10 και (3) v_1=15, v_2=15

Παράδειγμα 3.3

Να γίνουν τα ιστογράμματα δύο δειγμάτων με 1000 τιμές που προσομοιώνουν α) το δείγμα τιμών του δυναμικού του στοιχείου Weston του παραδείγματος 2.2 και β) το περιβαλλοντικό δείγμα του παραδείγματος 2.3.

◆ Για να προσομοιώσουμε ένα δείγμα θα πρέπει να χρησιμοποιήσουμε μια **γεννήτρια τυχαίων αριθμών** που να έχει τη δυνατότητα δημιουργίας αριθμών που να ακολουθούν μια ορισμένη κατανομή. Αυτή η δυνατότητα υπάρχει στο *Excel*. Το δείγμα τιμών δυναμικού ακολουθεί την κανονική κατανομή. Χρησιμοποιώντας τις συναρτήσεις *AVERAGE* και *STDEV* υπολογίζουμε τη μέση τιμή και την τυπική απόκλιση των τιμών του δείγματος του παραδείγματος 2.2. Παίρνουμε τις τιμές <x> = 1.01894 και s = 0.0009. Ακολούθως, σε ένα φύλλο εργασίας πηγαίνουμε *Δεδομένα (Data)* → *Ανάλυση (Analysis)* → *Ανάλυση δεδομένων (Data Analysis)* και στη λίστα που εμφανίζεται επιλέγουμε *Γεννήτρια τυχαίων αριθμών (Random number generator)*. Στο παράθυρο που ανοίγει επιλέγουμε την *Κανονική (Normal)* κατανομή και εισάγουμε τους αριθμούς <x> = 1.01894 και s = 0.0009 στα αντίστοιχα πεδία (σχήμα 3.8). Κάνουμε *κλικ* στο *Περιοχή εξόδου (Output Range)* και με το ποντίκι επιλέγουμε μια περιοχή-στήλη με 1000 κελιά. Πατώντας *OK* δημιουργείται στη επιλεγμένη περιοχή ένα δείγμα με 1000 τιμές που ακολουθούν την κανονική κατανομή με <x>

= 1.01894 και s = 0.0009 και συνεπώς προσομοιώνει το δείγμα τιμών του δυναμικού του στοιχείου Weston του παραδείγματος 2.2.

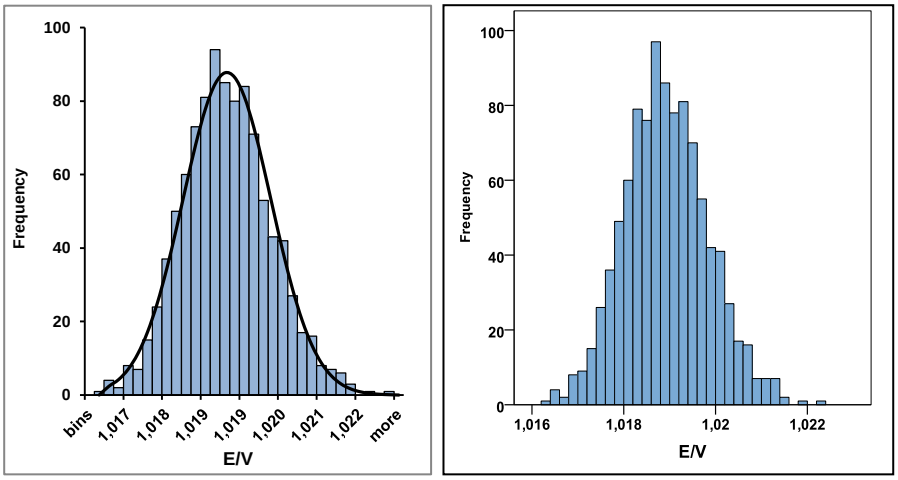


Σχήμα 3.8. Συμπλήρωση πλαισίου για δημιουργία δείγματος από 1000 τιμές που ακολουθούν την κανονική κατανομή με <x> =1.01894 και s = 0.0009

Για να κατασκευάσουμε το ιστόγραμμα στο *Excel* πηγαίνουμε *Πρόσθετα (Add-ins)* → *ChemStat* → *Graphs* → *Histogram.* Στο πρώτο πλαίσιο εισαγωγής δεδομένων εισάγουμε με το ποντίκι ΜΟΝΟ τις τιμές του δείγματος, στο επόμενο πλαίσιο αφήνουμε τη μονάδα ώστε στο γράφημα να χρησιμοποιηθούν *Συχνότητες*, ακολούθως εισάγουμε τον αριθμό 30 για το πλήθος των κλάσεων, στη συνέχεια ορίζουμε το κελί εξόδου των αποτελεσμάτων και στο τελευταίο πλαίσιο εισάγουμε τον αριθμό 1 για να σχεδιάσει το πρόγραμμα και την καμπύλη της κανονικής κατανομής. Το πρόγραμμα υπολογίζει τις κλάσεις και τις συχνότητες, τα δεδομένα αυτά εξάγονται στο φύλλο εργασίας και με βάση αυτά κατασκευάζεται το ιστόγραμμα. Αυτό μετά από κατάλληλη μορφοποίηση, όπως περιγράφεται στο παράδειγμα 2.2, δίνεται στο σχήμα 3.9-αριστερά.

Για να κάνουμε το ιστόγραμμα στο *SPSS*, μεταφέρουμε τα δεδομένα σε μία στήλη του *SPSS* και ακολουθούμε την πορεία *Graphs* → *Legacy*

Dialogs → *Histogram.* Η γραφική παράσταση που παίρνουμε μετά από μορφοποίηση δίνεται στο σχήμα 3.9-δεξιά.

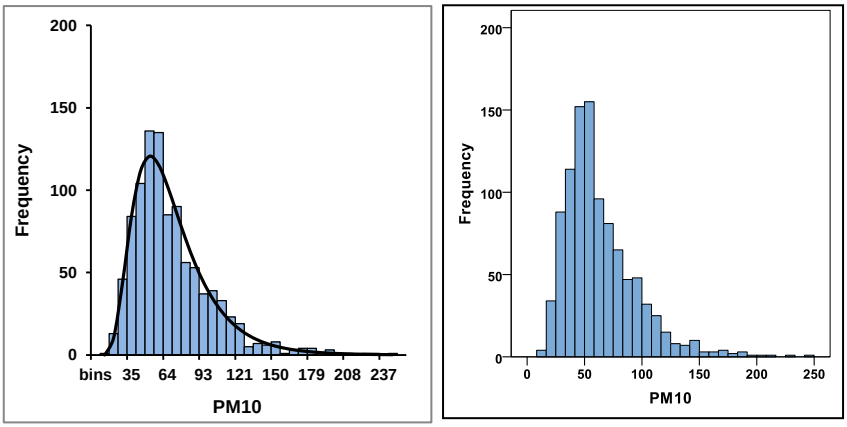


Σχήμα 3.9. Προσομοιωμένο ιστόγραμμα κανονικού δείγματος κατασκευασμένο με το *ChemStat* (αριστερά) και το *SPSS* (δεξιά)

Η γεννήτρια τυχαίων αριθμών του *Excel* δεν διαθέτει τη λογαριθμοκανονική (log-normal) κατανομή. Έτσι για να προσομοιώσουμε το περιβαλλοντικό δείγμα λαμβάνουμε υπόψη ότι αν λογαριθμήσουμε τις τιμές ενός δείγματος που ακολουθεί την λογαριθμοκανονική κατανομή, το δείγμα που θα προκύψει ακολουθεί την κανονική κατανομή. Συνεπώς, μεταφέρουμε τις τιμές PM10 του Πίνακα 2.5 σε μία στήλη του *Excel*, έστω στην A2:A79, και στη διπλανή στήλη υπολογίζουμε τους λογαρίθμους των τιμών της πρώτης στήλης χρησιμοποιώντας τη συνάρτηση *LN*. Στη συνέχεια με τις συναρτήσεις *AVERAGE* και *STDEV* υπολογίζουμε τη μέση τιμή και την τυπική απόκλιση των τιμών του νέου δείγματος. Παίρνουμε τις τιμές <x> = 4.03925 και s = 0.463 και με βάση αυτές τις τιμές δημιουργούμε ένα κανονικό δείγμα 1000 τιμών, έστω στην περιοχή C2:C1001. Τέλος, από το δείγμα αυτό δημιουργούμε το προσομοιωμένο περιβαλλοντικό δείγμα με αντι-λογαρίθμηση των τιμών του κανονικού δείγματος. Έτσι αν το κανονικό δείγμα είναι στην περιοχή C2:C1001, στο D2 πληκτρολογούμε =EXP(C2), πατάμε *Enter* και συμπληρώνουμε την περιοχή D3:D1001 με τη διαδικασία της αυτόματης συμπλήρωσης.

Τα ιστογράμματα που κατασκευάζονται με το *ChemStat* και το *SPSS*

δίνονται στο σχήμα 3.10. Στο ιστόγραμμα του *ChemStat* έχει προστεθεί και η καμπύλη της λογαριθμοκανονικής κατανομής εισάγοντας την τιμή 2 στο τελευταίο πλαίσιο κατασκευής ιστογραμμάτων.



Σχήμα 3.10. Προσομοιωμένο ιστόγραμμα περιβαλλοντικού δείγματος κατασκευασμένο με το *ChemStat* (αριστερά) και το *SPSS* (δεξιά)

ΑΣΚΗΣΕΙΣ

3.1. Να σχηματισθούν 4 δείγματα με 10, 100, 1000 και 10000 τιμές από δεδομένα που ακολουθούν την κανονική κατανομή με μέση τιμή 0 και τυπική απόκλιση 1 και ακολούθως να γίνουν τα ιστογράμματα και τα θηκογράμματά τους. Τι συμπεράσματα προκύπτουν;

3.2. Να σχηματισθούν 4 δείγματα με 10, 100, 1000 και 10000 τιμές από δεδομένα που ακολουθούν την λογαριθμοκανονική (log-normal) κατανομή με μ = 4 και σ = 0.5. Ακολούθως να γίνουν τα ιστογράμματα και τα θηκογράμματά τους. Τι συμπεράσματα προκύπτουν;

3.3. Να σχηματισθούν 10 δείγματα με 15 τιμές το κάθε ένα από δεδομένα που ακολουθούν την κανονική κατανομή με μέση τιμή 1 και τυπική απόκλιση 0.1. Σε αυτά να υπολογιστούν η μέση τιμή, η διασπορά, η τυπική απόκλιση και η διάμεσος και να εξετασθεί αν συγκλίνουν τα αποτελέσματα στα 10 αυτά δείγματα.

Κεφάλαιο 4

ΣΤΑΤΙΣΤΙΚΕΣ ΕΚΤΙΜΗΣΕΙΣ

4.1 ΓΕΝΙΚΑ

Όταν εκτελούμε ένα πείραμα δεν μας ενδιαφέρει άμεσα ούτε η μέση τιμή $\bar{x}$ ούτε η τυπική απόκλιση s των μετρήσεων στο δείγμα μας. Εκείνο που μας ενδιαφέρει είναι να προσδιορίσουμε ή έστω να εκτιμήσουμε τη μέση τιμή μ του πληθυσμού από τον οποίο προέρχεται το δείγμα και αντίστοιχα την τυπική απόκλιση σ. Επειδή ο πληθυσμός περιέχει το σύνολο των δυνατών μετρήσεων, η μέση τιμή μ του πληθυσμού είναι εξορισμού η *πραγματική τιμή* της μεταβλητής που θέλουμε να προσδιορίσουμε.

Τα $\bar{x}$ και s^2 είναι μια πρώτη εκτίμηση των μ και σ^2, δεδομένου ότι εμπίπτουν στην κατηγορία των *αβίαστων* ή *αμερόληπτων εκτιμητριών*. Όπως αναφέρθηκε, κάθε παράμετρος δείγματος που η μέση τιμή της, αν πάρουμε πολλά δείγματα, ισούται με την αντίστοιχη παράμετρο του πληθυσμού, ονομάζεται αβίαστη ή αμερόληπτη εκτιμήτρια. Συνεπώς, απουσία άλλης πληροφορίας οι ποσότητες $\bar{x}$ και s^2 μπορούν να χρησιμοποιηθούν ως μια πρώτη εκτίμηση για τις άγνωστες τιμές του μέσου όρου μ και της διασποράς σ^2 ενός πληθυσμού. Δεν συμβαίνει το ίδιο και με την τυπική απόκλιση s που δεν είναι μια αβίαστη εκτιμήτρια της τυπικής απόκλισης σ του πληθυσμού. Η εκτίμηση των μ και σ^2 από τα $\bar{x}$ και s^2 ονομάζεται ***σημειακή εκτίμηση*** *(point estimation)*.

Είναι προφανές ότι όταν υπάρχει μεγάλο πειραματικό σφάλμα, η χρήση αβίαστων εκτιμητριών δεν είναι ικανοποιητική, γιατί αν εκτελέσουμε μια άλλη σειρά πειραματικών μετρήσεων της μεταβλητής Χ, τότε θα προκύψει μια άλλη μέση τιμή $\bar{x}$ και μια άλλη τιμή για τη διασπορά s^2. Για το λόγο αυτό συνήθως αποφεύγεται η *σημειακή εκτίμηση* και προτιμάται η ***εκτίμηση διαστήματος*** *(interval estimation)*. Συγκεκριμένα ονομάζεται ***P% διάστημα εμπιστοσύνης*** (*confidence interval*) μιας παραμέτρου θ του πληθυσμού, το διάστημα [a, b] μέσα στο οποίο αναμένεται να υπάρχει

η θ με πιθανότητα P%. Συνήθως ως P λαμβάνουμε την τιμή 95. Δηλαδή συνηθίζεται να υπολογίζουμε το 95% διάστημα εμπιστοσύνης των διαφόρων παραμέτρων, π.χ. μ, σ^2 ή σ, ενός πληθυσμού.

Η πιθανότητα P% γράφεται ως: P = 100(1 - α). Σε αυτή την περίπτωση η ποσότητα α = 1 – P/100 εκφράζει τον *κίνδυνο σφάλματος*, δηλαδή την πιθανότητα που έχει η παράμετρος θ του πληθυσμού να βρίσκεται έξω από το P% διάστημα εμπιστοσύνης.

4.2 ΒΑΣΙΚΟ ΘΕΩΡΗΜΑ

Αν υπολογίσουμε τη μέση τιμή σε πολλά δείγματα της ίδιας μεταβλητής θα πάρουμε διαφορετικές και μη προβλεπόμενες τιμές $\bar{x}$. Αυτό δείχνει ότι η ποσότητα $\bar{x}$ είναι επίσης μια *τυχαία μεταβλητή*, δηλαδή μεταβλητή της οποίας την τιμή δεν μπορούμε να προβλέψουμε, και προφανώς το ίδιο ισχύει για την τυπική απόκλιση s και για κάθε συνάρτηση των $\bar{x}$ και s.

Θεώρημα 4.1. Σε δείγματα που αποτελούνται από m τιμές και προέρχονται από κανονικό πληθυσμό η τυχαία μεταβλητή

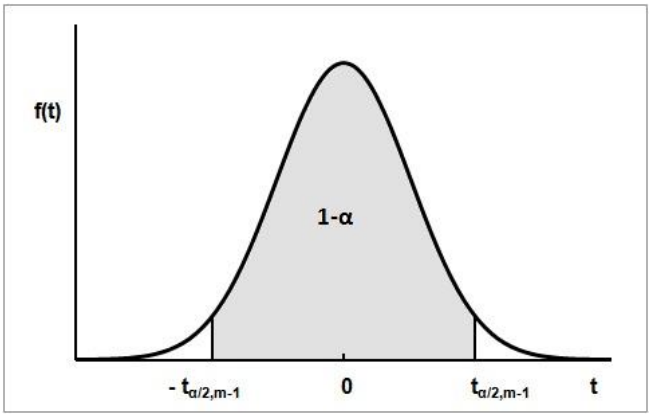
$$t = \frac{\bar{x} - \mu}{s / \sqrt{m}} \qquad (4.1)$$

ακολουθεί την κατανομή *student* με m-1 βαθμούς ελευθερίας.

4.3 ΔΙΑΣΤΗΜΑ ΕΜΠΙΣΤΟΣΥΝΗΣ ΜΕΣΗΣ ΤΙΜΗΣ

Σύμφωνα με το θεώρημα 4.1 η μεταβλητή $t = (\bar{x} - \mu)\sqrt{m} / s$ ακολουθεί την κατανομή *student* με m-1 βαθμούς ελευθερίας με την προϋπόθεση ότι οι τιμές του δείγματος ακολουθούν την κανονική κατανομή. Στο σχήμα 4.1 δίνεται η γραφική παράσταση της κατανομής *student* με m-1 βαθμούς ελευθερίας και έστω ότι το εμβαδόν κάτω από την καμπύλη του τόπου που έχει γκρι χρώμα είναι 1 – α. Έστω επίσης ότι τα σημεία που οριοθετούν αυτόν τον συμμετρικό χώρο είναι το $-t_{\alpha/2,m-1}$ και το $t_{\alpha/2,m-1}$.

Αν ζητάμε να υπολογίσουμε το P% διάστημα εμπιστοσύνης, το P είναι γνωστό και το ίδιο ισχύει για το α = 1 - P/100. Τότε το σημείο $t_{\alpha/2,m-1}$ υπολογίζεται ή από στατιστικούς πίνακες ή πολύ πιο απλά χρησιμοποιώντας τη συνάρτηση =TINV(α;m-1) του *Excel*. Επομένως όταν δοθεί το P, το σημείο $t_{\alpha/2,m-1}$ υπολογίζεται και από τις ιδιότητες των συναρτήσεων κατανομής παίρνουμε



Σχήμα 4.1. Καμπύλη της κατανομής *student* με m-1 βαθμούς ελευθερίας

$$P(-t_{α/2,m-1} < t < t_{α/2,m-1}) = 1 - α \Rightarrow P(-t_{α/2,m-1} < \frac{\bar{x} - μ}{s/\sqrt{m}} < t_{α/2,m-1}) = 1-α \Rightarrow$$

$$P(-t_{α/2,m-1}\frac{s}{\sqrt{m}} < \bar{x} - μ < t_{α/2,m-1}\frac{s}{\sqrt{m}}) = 1-α \Rightarrow$$

$$P(\bar{x} - t_{α/2,m-1}\frac{s}{\sqrt{m}} < μ < \bar{x} + t_{α/2,m-1}\frac{s}{\sqrt{m}}) = 1-α$$

Άρα αν το δείγμα ακολουθεί την κανονική κατανομή, τότε το P% = 100(1-α) διάστημα εμπιστοσύνης για τη μέση τιμή του πληθυσμού είναι το:

$$\left(\bar{x} - t_{α/2,m-1}\frac{s}{\sqrt{m}}, \bar{x} + t_{α/2,m-1}\frac{s}{\sqrt{m}}\right) \quad (4.2)$$

Το παραπάνω διάστημα μπορεί να γραφεί και ως $\bar{x} \pm b$, όπου

$$b = t_{α/2,m-1}\frac{s}{\sqrt{m}} \quad (4.3)$$

Παράδειγμα 4.1

Να υπολογιστεί το 95% διάστημα εμπιστοσύνης για τη μέση τιμή του πληθυσμού από τον οποίον προέρχεται το δείγμα τιμών pH:

5.173 5.182 5.201 5.175 5.189 5.179

❖ **Ανάλυση στο Excel**

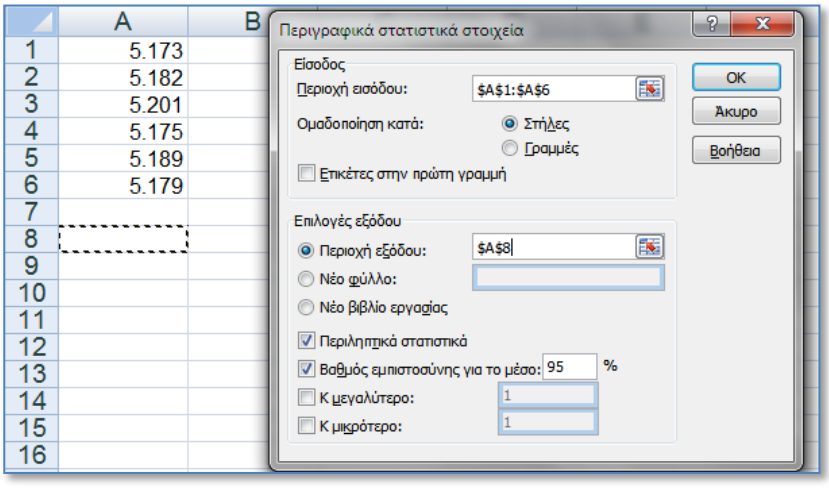
Εισάγουμε τα δεδομένα σε μία στήλη του *Excel*, έστω στην περιοχή A1:A6. Η τιμή b υπολογίζεται από *Δεδομένα (Data)* → *Ανάλυση Δεδομένων (Data Analysis)* → *Περιγραφικά στατιστικά στοιχεία (Descriptive Statistics)*. Στο παράθυρο διαλόγου που ανοίγει εισάγουμε την περιοχή δεδομένων, A1:A6, επιλέγουμε το *Περιληπτικά στατιστικά (Summary Statistics)* και το 95% διάστημα εμπιστοσύνης από το *Βαθμός εμπιστοσύνης για το μέσο (Confidence Level for Mean)* (σχήμα 4.2). Παίρνουμε τα αποτελέσματα του σχήματος 4.3.

Με βάση τα αποτελέσματα αυτά, το 95% διάστημα εμπιστοσύνης για τη μέση τιμή μ μπορεί να γραφεί ως 5.183167 ± 0.010914. Όμως τα πολλά δεκαδικά ψηφία είναι χωρίς σημασία. Έτσι στρογγυλοποιούμε την τιμή 0.010914 στο πρώτο ή σπάνια (συντηρητική επιλογή) στα δύο πρώτα σημαντικά δεκαδικά ψηφία. Δηλαδή, στην τιμή 0.01 ή σπάνια στην 0.011. Με βάση τις τιμές αυτές στρογγυλοποιούμε την τιμή 5.183167 στα δύο, 5.18, ή στα τρία, 5.183, δεκαδικά ψηφία. Έτσι, το 95% διάστημα εμπιστοσύνης για τη μέση τιμή μ μπορεί να γραφεί ως

5.18 ± 0.01 και συντηρητικά ως 5.183 ± 0.011

Συνεπώς, η πραγματική τιμή του pH βρίσκεται με πιθανότητα 95% στο διάστημα 5.18 ± 0.01. Εναλλακτικά γράφουμε ότι το pH του διαλύματος είναι:

pH = 5.18 ± 0.01



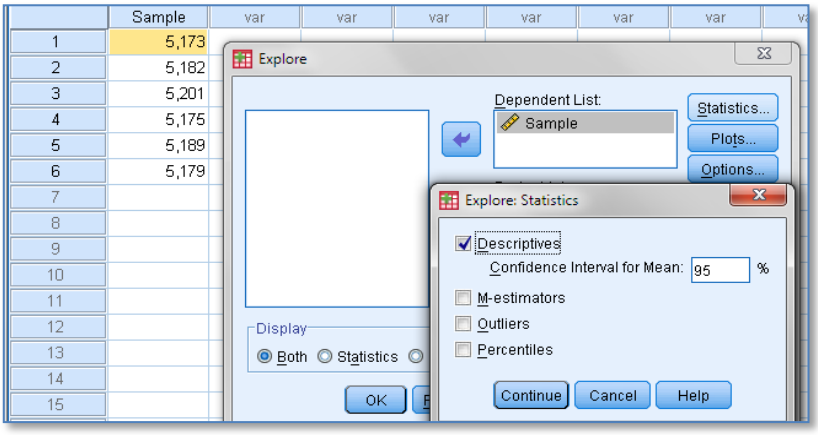
Σχήμα 4.2. Υπολογισμός διαστήματος εμπιστοσύνης στο *Excel*

	A	B
8	*Στήλη1*	
9		
10	Μέσος	5.183167
11	Τυπικό σφάλμα	0.004246
12	Διάμεσος	5.1805
13	Επικρατούσα τιμή	#Δ/Υ
14	Μέση απόκλιση τετραγώνου	0.0104
15	Διακύμανση	0.000108
16	Κύρτωση	0.793927
17	Ασυμμετρία	1.12018
18	Εύρος	0.028
19	Ελάχιστο	5.173
20	Μέγιστο	5.201
21	Άθροισμα	31.099
22	Πλήθος	6
23	Βαθμός εμπιστοσύνης(95.0%)	0.010914

Σχήμα 4.3. Πίνακας περιγραφικών μέτρων με το 95% διάστημα εμπιστοσύνης στο *Excel*

❖ **Ανάλυση στο SPSS**

Στο *SPSS* ακολουθούμε την πορεία *Analyze* → *Descriptive Statistics* → *Explore*. Στο παράθυρο διαλόγου που ανοίγει εισάγουμε το δείγμα στο πλαίσιο *Dependent List* και από το *Statistics* επιλέγουμε το *Descriptives* και το 95% διάστημα εμπιστοσύνης στο *Confidence Interval for Mean* (σχήμα 4.4). Από τον πίνακα αποτελεσμάτων *Descriptives* παίρνουμε το κάτω και το επάνω όριο του διαστήματος εμπιστοσύνης, [5.17225, 5.19408], που ταυτίζεται με το αποτέλεσμα του *Excel*.



Σχήμα 4.4. Βήματα υπολογισμού διαστήματος εμπιστοσύνης στο *SPSS*

Παράδειγμα 4.2

Για τον έλεγχο μιας νέας αναλυτικής μεθόδου προσδιορισμού της σεληνιοουρίας στο νερό, παρασκευάστηκε ένα διάλυμα της ουσίας αυτής με συγκέντρωση 40 ng/mL και ακολούθως έγιναν 5 μετρήσεις της συγκέντρωσης της σεληνιοουρίας. Να εξετασθεί αν η μέθοδος είναι αξιόπιστη λαμβάνοντας υπόψη ότι ελήφθησαν τα αποτελέσματα:

40.3, 40.2, 40.2, 40.0, 40.3

◆ Εργαζόμενοι όπως και στο προηγούμενο παράδειγμα βρίσκουμε ότι το 95% διάστημα εμπιστοσύνης για τη μέση τιμή του πληθυσμού είναι

$$40.20 \pm 0.15 = [40.05,\ 40.35]$$

Όμως εφόσον γνωρίζουμε ότι η συγκέντρωση της σεληνιοουρίας στο πρότυπο διάλυμα είναι 40 ng/mL, ισχύει μ = 40. Παρατηρούμε ότι η τιμή αυτή, μ = 40, είναι έξω από το 95% διάστημα εμπιστοσύνης [40.05, 40.35]. Επομένως είναι πιθανόν να υπάρχει κάποιο συστηματικό σφάλμα ή στην παρασκευή του πρότυπου διαλύματος της σεληνιοουρίας ή στην προτεινόμενη αναλυτική μέθοδο, ενώ υπάρχει και μια μικρή πιθανότητα του 5% αυτό να οφείλεται σε τυχαίους λόγους.

Το θέμα αυτό επανεξετάζεται με διαφορετικό τρόπο στο επόμενο κεφάλαιο.

Θα πρέπει να τονιστεί ότι ο υπολογισμός των διαστημάτων εμπιστοσύνης προϋποθέτει ότι τα δεδομένα του δείγματος ακολουθούν την ***κανονική κατανομή***. Ο έλεγχος της κανονικότητας γίνεται με τα κριτήρια *Anderson-Darling*, *Kolmogorov-Smirnov* και *Shapiro-Wilk*, όπως αναλύεται στο κεφάλαιο 6.2.

4.4 ΜΕΤΑΔΟΣΗ ΣΦΑΛΜΑΤΩΝ

Όταν τα αποτελέσματα ενός πειράματος χρησιμοποιούνται για να υπολογιστεί η τιμή μιας φυσικοχημικής ποσότητας z, τότε τα πειραματικά σφάλματα στα αρχικά δεδομένα μεταδίδονται και επιδρούν στην τελική τιμή της z. Έστω ότι οι μεταβλητές x_1, x_2, ..., x_n αντιστοιχούν σε φυσικές ποσότητες που προκύπτουν πειραματικά με *απόλυτο σφάλμα* Δx_1, Δx_2, ..., Δx_n, αντίστοιχα, όπου

$$\Delta x_i = |x_i - \mu_i| \tag{4.4}$$

Έστω επίσης ότι οι μεταβλητές x_1, x_2, ..., x_n ακολούθως χρησιμοποιούνται για να υπολογιστεί η τιμή της φυσικοχημικής ποσότητας z με τη βοήθεια της εξίσωσης $z = f(x_1, x_2, ..., x_m)$. Σε αυτή την περίπτωση το σφάλμα Δz στην τιμή του z, που οφείλεται στα πειραματικά σφάλματα Δx_1, Δx_2, ..., Δx_m, μπορεί να εκτιμηθεί από τη σχέση

$$\Delta z \approx \frac{\partial f}{\partial x_1}\Delta x_1 + \frac{\partial f}{\partial x_2}\Delta x_2 + \cdots + \frac{\partial f}{\partial x_n}\Delta x_n \tag{4.5}$$

που προκύπτει από το γεγονός ότι το ολικό διαφορικό dz της συνάρτησης $z = f(x_1, x_2, ..., x_m)$ δίνει κατά προσέγγιση τη μεταβολή Δz της τιμής της συνάρτησης z όταν οι μεταβολές Δx_1, Δx_2, ..., Δx_m είναι πολύ μικρές.

Αν η παράγωγος $\partial f/\partial x_i$ είναι αρνητικός αριθμός, τότε η συνεισφορά του σφάλματος Δx_i στο σφάλμα Δz, δηλαδή ο όρος $(\partial f/\partial x_i)\Delta x_i$, είναι αρνητική, ενώ το αντίστροφο συμβαίνει όταν η παράγωγος $\partial f/\partial x_i$ είναι θετικός αριθμός. Αλλά εξαιτίας του τυχαίου χαρακτήρα των πειραματικών σφαλμάτων, η συνεισφορά κάθε σφάλματος Δx_1, Δx_2, ..., Δx_n στην υπολογιζόμενη τιμή του z μπορεί να είναι θετική ή αρνητική με την ίδια πιθανότητα. Για το λόγο αυτό, για την εκτίμηση του σφάλματος στη μεταβλητή z χρησιμοποιείται η τροποποιημένη σχέση

$$\Delta z = \left|\frac{\partial f}{\partial x_1}\right|\Delta x_1 + \left|\frac{\partial f}{\partial x_2}\right|\Delta x_2 + \cdots + \left|\frac{\partial f}{\partial x_n}\right|\Delta x_n \tag{4.6}$$

που δίνει το *μέγιστο σφάλμα* που μπορεί να συμβεί στη μεταβλητή z.

Οι ποσότητες $100\Delta z/z$ και $100\Delta x_i/x_i$ ονομάζονται *εκατοστιαία σφάλματα*. Συνεπώς η σχέση (4.6) μπορεί να χρησιμοποιηθεί για τον υπολογισμό του εκατοστιαίου σφάλματος στη μεταβλητή z όταν είναι

γνωστά τα εκατοστιαία σφάλματα στις ανεξάρτητες μεταβλητές x_1, x_2, ..., x_n.

Στη στατιστική αντί της έννοιας του σφάλματος, που δε σχετίζεται άμεσα με κάποια κατανομή, χρησιμοποιείται η τυπική απόκλιση. Αποδεικνύεται ότι αν οι μεταβλητές x_1, x_2, ..., x_n είναι ανεξάρτητες, τότε η τυπική απόκλιση στη μεταβλητή z μπορεί να υπολογιστεί από τη σχέση

$$s_z = \sqrt{\left(\frac{\partial f}{\partial x_1}\right)^2 s_1^2 + \left(\frac{\partial f}{\partial x_2}\right)^2 s_2^2 + \cdots + \left(\frac{\partial f}{\partial x_n}\right)^2 s_n^2} \qquad (4.7)$$

Αν οι μεταβλητές x_1, x_2, ..., x_n δεν είναι ανεξάρτητες, τότε η παραπάνω σχέση γενικεύεται στην

$$s_z = \sqrt{\sum_{i=1}^{n}\left(\frac{\partial f}{\partial x_i}\right)^2 s_i^2 + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\left(\frac{\partial f}{\partial x_i}\right)\left(\frac{\partial f}{\partial x_j}\right) s_{ij}} \qquad (4.8)$$

όπου s_{ij} είναι η συνδιασπορά των μεταβλητών x_i και x_j. Η ***συνδιασπορά*** (*covariance*) δύο μεταβλητών x και y ορίζεται από τη σχέση

$$s_{xy} = \sum_{i=1}^{m}\frac{(x_i - \bar{x})(y_i - \bar{y})}{m-1} \qquad (4.9)$$

Αποδεικνύεται ότι αν δύο μεταβλητές είναι ανεξάρτητες, τότε η συνδιασπορά τους είναι ίση με μηδέν, $s_{xy} = 0$.

Παράδειγμα 4.3

Το ειδικό βάρος ρ ενός υγρού μπορεί να υπολογιστεί από τη σχέση ορισμού του, ρ = B/V, όπου B είναι το βάρος του υγρού που έχει όγκο V. Αν V = 10 cm^3 και B = 9.5254 g με σφάλματα ΔV = 0.02 cm^3 και ΔB = 0.0001 g, ποιο είναι το εκατοστιαίο σφάλμα στον υπολογισμό του ρ;

◆ Επειδή ρ = ρ(B, V), έχουμε

$$\Delta\rho = \left|\frac{\partial\rho}{\partial B}\right|\Delta B + \left|\frac{\partial\rho}{\partial V}\right|\Delta V = \frac{1}{V}\Delta B + \frac{B}{V^2}\Delta V = \frac{\rho}{B}\Delta B + \frac{\rho}{V}\Delta V \Rightarrow$$

$$100\Delta\rho/\rho = 100\Delta B/B + 100\Delta V/V$$

Συνεπώς

$$100\Delta\rho/\rho = 0.00105 + 0.2 \approx 0.2 \text{ g/cm}^3$$

Αυτό σημαίνει ότι για το ειδικό βάρος έχουμε: $\rho = B/V = 9.5254/10 = 0.95254$ g/cm^3 με αβεβαιότητα $\Delta\rho = 0.2\rho/100 = 0.002$ g/cm^3. Άρα το τελικό αποτέλεσμα μπορεί να γραφεί ως

$$\rho = 0.95254 \pm \Delta\rho/2 \text{ g/cm}^3 \Rightarrow \rho = 0.953 \pm 0.001 \text{ g/cm}^3$$

Παράδειγμα 4.4

Αν γνωρίζουμε από ανεξάρτητες μετρήσεις του όγκου της φιάλης ότι αυτός είναι $V = 10$ cm^3 με τυπική απόκλιση $s_V = 0.015$ cm^3, ενώ για το βάρος έχουμε $B = 9.5254$ g με $s_B = 0.0003$ g, ποιο είναι το ειδικό βάρος του υγρού και ποια η τυπική του απόκλιση;

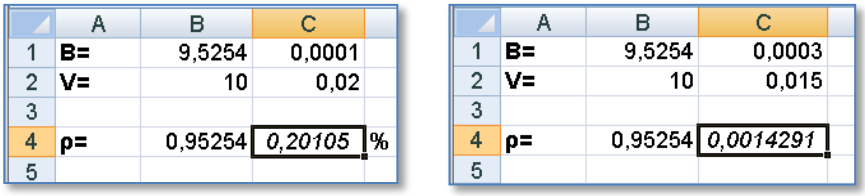
◆ Για το ειδικό βάρος έχουμε: $\rho = B/V = 9.5254/10 = 0.95254$ g/cm^3. Για την τυπική απόκλιση εφαρμόζουμε τη σχέση (4.7) και παίρνουμε

$$s_\rho = \sqrt{\left(\frac{\partial\rho}{\partial B}\right)^2 s_B^2 + \left(\frac{\partial\rho}{\partial V}\right)^2 s_V^2} = \sqrt{\left(\frac{\rho}{B}\right)^2 s_B^2 + \left(\frac{\rho}{V}\right)^2 s_V^2} = 0.0014 \text{ g/cm}^3$$

Συνεπώς η ακρίβεια των μετρήσεων είναι στο τρίτο δεκαδικό ψηφίο και επομένως το αποτέλεσμα πρέπει να γραφεί ως

$$\rho = 0.953 \pm 0.001 \text{ g/cm}^3$$

Ένας πολύ εύκολος τρόπος για να προσδιορίζουμε τη μετάδοση σφάλματος, δηλαδή το εκατοστιαίο σφάλμα $100\Delta z/z$ ή την τυπική απόκλιση σ_z, είναι με τη χρήση του προγράμματος *Propagation* του *ChemStat*. Για το σκοπό αυτό εργαζόμαστε ως εξής. Σε ένα φύλλο του *Excel* διευθετούμε τα δεδομένα του προβλήματος όπως φαίνεται στο σχήμα 4.5. Συγκεκριμένα οι μεταβλητές B και V πρέπει να είναι σε συνεχόμενα κελιά σε μία στήλη ή σε μία γραμμή και το ίδιο να ισχύει για τα σφάλματα ΔB και ΔV ή τις τυπικές αποκλίσεις s_B και s_V. Στο σχήμα 4.5 ο υπολογισμός του ρ γίνεται στο κελί B4, στο οποίο πληκτρολογούμε τον τύπο =B1/B2.



	A	B	C
1	**B=**	9,5254	0,0001
2	**V=**	10	0,02
3			
4	**ρ=**	0,95254	*0,20105* %
5			

	A	B	C
1	**B=**	9,5254	0,0003
2	**V=**	10	0,015
3			
4	**ρ=**	0,95254	*0,0014291*
5			

Σχήμα 4.5. Διευθέτηση δεδομένων και προσδιορισμός του 100Δρ/ρ (αριστερά) και της s_ρ (δεξιά) με το πρόγραμμα *Propagation*

Για να υπολογίσουμε τώρα το εκατοστιαίο σφάλμα 100Δρ/ρ ή την τυπική απόκλιση s_ρ πηγαίνουμε *Πρόσθετα (Add-ins)* → *ChemStat* → *Error Propagation*. Στο πρώτο πλαίσιο που ανοίγει επιλέγουμε με 0 τον υπολογισμό της s_ρ και με 1 τον υπολογισμό του 100Δρ/ρ. Ακολούθως, στο επόμενο πλαίσιο εισάγουμε με το ποντίκι την περιοχή των μεταβλητών B και V, δηλαδή την B1:B2. Κάνουμε *κλικ* στο *OK* και στο τρίτο πλαίσιο που ανοίγει εισάγουμε την περιοχή C1:C2. Τέλος, όταν εμφανίζεται το τελευταίο πλαίσιο κάνουμε *κλικ* στο κελί B4, που βρίσκεται ο τύπος της συνάρτησης. Με *κλικ* στο *OK* παίρνουμε στο κελί C4 την τιμή του εκατοστιαίου σφάλματος 100Δρ/ρ = 0.2% (σχήμα 4.5-αριστερά) ή της τυπικής απόκλισης s_ρ = 0.0014291 (σχήμα 4.5-δεξιά).

ΑΣΚΗΣΕΙΣ

4.1. Να σχηματισθούν 7 δείγματα με m = 15, 30, 50, 100, 200, 500, και 1000 τιμές από δεδομένα που ακολουθούν την κανονική κατανομή με μέση τιμή 1 και τυπική απόκλιση 0.1. Στα δείγματα αυτά να υπολογιστούν η μέση τιμή, η διασπορά, η τυπική απόκλιση και τα 95% διαστήματα εμπιστοσύνης. Να γίνει η γραφική παράσταση του εύρους των 95% διαστημάτων εμπιστοσύνης σε συνάρτηση με τη ρίζα του m και να συζητηθεί το αποτέλεσμα.

4.2. Χρησιμοποιώντας τη μέθοδο της δενδροχρονολόγησης προέκυψε το ακόλουθο δείγμα τιμών σε έτη:

1252, 1188, 1305, 1273, 1195, 1231, 1287, 1267

Να προσδιοριστεί το 95% διάστημα εμπιστοσύνης της μέσης τιμής.

4.3. Για τον έλεγχο μιας νέας αναλυτικής μεθόδου προσδιορισμού φθορίου στο θαλασσινό νερό, παρασκευάστηκε ένα διάλυμα φθορίου με συγκέντρωση 30 ng/mL και ακολούθως έγιναν 5 μετρήσεις της συγκέντρωσής του που έδωσαν τα αποτελέσματα:

30.1, 30.2, 29.8, 30.0, 30.3

Να προσδιοριστεί το 95% διάστημα εμπιστοσύνης για τη μέση τιμή και με βάση αυτό να εξετασθεί αν υπάρχει πρόβλημα αξιοπιστίας της μεθόδου.

4.4. 0.1 M διάλυμα οξέος χρησιμοποιείται για την ογκομέτρηση 10 mL διαλύματος βάσης συγκέντρωσης 0.1 M και σε 5 μετρήσεις καταναλώθηκαν οι ακόλουθοι όγκοι οξέος σε mL:

9.88, 10.18, 10.23, 10.39, 10.21

Να προσδιοριστεί το 95% διάστημα εμπιστοσύνης για τη μέση τιμή και με βάση αυτό να εξετασθεί αν υπάρχει πρόβλημα αξιοπιστίας της ογκομέτρησης.

4.5. Μπορούν να υπολογιστούν διαστήματα εμπιστοσύνης στα δείγματα της άσκησης 2.4;

4.6. Σε ένα διάλυμα έγιναν οι ακόλουθες μετρήσεις τιμών pH: 2.21, 2.25, 2.20, 2.23, 2.18. Να υπολογιστεί η συγκέντρωση των υδρογονοκατιόντων και η τυπική της απόκλιση.

4.7. Για τον υπολογισμό της οπτικής πυκνότητας D ενός διαλύματος έγιναν τρεις μετρήσεις της διαπερατότητας T: 0.495, 0.500, 0.498. Να υπολογιστεί η οπτική πυκνότητα D = log(1/T) και η τυπική της απόκλιση.

4.8. Η εσωτερική διάμετρος d και το ύψος h δύο κυλινδρικών δεξαμενών είναι d = 32.12 ± 5 cm και h = 24.45 ± 3 cm. Να υπολογιστεί ο συνολικός όγκος των δύο δεξαμενών και το σφάλμα υπολογισμού.

4.9. Μια ορισμένη ποσότητα οξυγόνου εισάγεται σε φιάλη όγκου 2.0 L σε θερμοκρασία 25 ^{o}C και πίεση 0.78 atm. Οι αβεβαιότητες στις μετρήσεις θερμοκρασίας, όγκου και πίεσης είναι: 1 ^{o}C, 0.01 L και 0.05 atm, αντίστοιχα. Υποθέτοντας ότι το αέριο είναι ιδανικό να υπολογιστεί ο αριθμός των moles του αερίου και το σφάλμα υπολογισμού.

4.10. Η σταθερά ταχύτητας μιας αντίδρασης δίνεται από τη σχέση του *Arrhenius* $k = A \cdot e^{-E/RT}$, όπου A είναι μια σταθερά, E η ενέργεια ενεργοποίησης και T η θερμοκρασία. Αν $A = (4.3 \pm 0.6) \times 10^{8}$ L/(mol·s) και E = 64.9 ± 0.5 kJ/mol, να υπολογιστεί η τιμή της σταθεράς ταχύτητας k και το σφάλμα της στην θερμοκρασία 25.0 ± 0.1 ^{o}C;

Κεφάλαιο 5

ΕΛΕΓΧΟΙ ΣΤΑΤΙΣΤΙΚΩΝ ΥΠΟΘΕΣΕΩΝ

5.1 ΣΤΑΤΙΣΤΙΚΕΣ ΥΠΟΘΕΣΕΙΣ

Οι έλεγχοι στατιστικών υποθέσεων αποτελούν αν όχι την ουσία της στατιστικής το πιο ενδιαφέρον τμήμα της. Ένας έλεγχος στατιστικών υποθέσεων είναι στην πραγματικότητα μια μέθοδος για να πάρουμε στατιστικές αποφάσεις χρησιμοποιώντας πειραματικά δεδομένα. Για να λάβουμε στατιστικές αποφάσεις είναι απαραίτητο να κάνουμε *στατιστικές υποθέσεις*, όμως δεν είμαστε ελεύθεροι να κάνουμε όποια στατιστική υπόθεση θέλουμε. Ανάλογα με το πρόβλημα που εξετάζουμε οι στατιστικές υποθέσεις που μπορούμε να κάνουμε είναι αυστηρά προσδιορισμένες.

Για παράδειγμα, έστω ότι έχουμε δύο δείγματα με μέσες τιμές $\bar{x}_1$ και $\bar{x}_2$. Το ερώτημα που γεννιέται είναι αν οι μέσες τιμές $\bar{x}_1$ και $\bar{x}_2$ παρουσιάζουν ή όχι στατιστικά σημαντική διαφορά. Οι μέσες τιμές $\bar{x}_1$ και $\bar{x}_2$ θα είναι στατιστικά ίσες αν τα δείγματα προέρχονται από πληθυσμούς που έχουν ίσες μέσες τιμές, $\mu_1 = \mu_2$ και διαφορετικές αν ισχύει γενικά $\mu_1 \neq \mu_2$. Συνεπώς οι στατιστικές υποθέσεις που μπορούμε να κάνουμε είναι:

α) Τα δείγματα προέρχονται από πληθυσμούς με $\mu_1 = \mu_2$

β) Τα δείγματα προέρχονται από πληθυσμούς με $\mu_1 \neq \mu_2$ ή $\mu_1 > \mu_2$ ή $\mu_1 < \mu_2$

Πολλές φορές καλούμαστε να αποφασίσουμε αν τα δεδομένα ενός δείγματος προέρχονται από πληθυσμό που ακολουθεί την κανονική κατανομή, οπότε και τα ίδια ακολουθούν την κανονική κατανομή. Σε αυτή την περίπτωση οι στατιστικές υποθέσεις που μπορούμε να διατυπώσουμε είναι:

α) Το δείγμα προέρχεται από κανονικό πληθυσμό

β) Το δείγμα δεν προέρχεται από κανονικό πληθυσμό

5.2 Η ΜΗΔΕΝΙΚΗ ΥΠΟΘΕΣΗ

Από τα παραπάνω παραδείγματα γίνεται φανερό ότι διατυπώνουμε δύο υποθέσεις. Από τις υποθέσεις αυτές η μία ονομάζεται ***μηδενική υπόθεση*** *(null hypothesis)* και συμβολίζεται με H_0, ενώ η εναλλακτική της μηδενικής συμβολίζεται με H_1. Έτσι στο πρώτο παράδειγμα η μηδενική υπόθεση διατυπώνεται ως

H_0: Τα δείγματα προέρχονται από πληθυσμούς με ίσες μέσες τιμές

Για την εναλλακτική της, την H_1, έχουμε τις ακόλουθες δυνατότητες:

H_1: Τα δείγματα προέρχονται από πληθυσμούς με $\mu_1 \neq \mu_2$

ή H_1: Τα δείγματα προέρχονται από πληθυσμούς με $\mu_1 > \mu_2$

ή H_1: Τα δείγματα προέρχονται από πληθυσμούς με $\mu_1 < \mu_2$

Στο δεύτερο παράδειγμα η μηδενική υπόθεση διατυπώνεται ως

H_0: Το δείγμα προέρχεται από κανονικό πληθυσμό

με εναλλακτική την

H_1: Το δείγμα δεν προέρχεται από κανονικό πληθυσμό

Δυστυχώς η Στατιστική δεν μπορεί να υπολογίσει την πιθανότητα να ισχύει η υπόθεση H_0 ή η H_1. Στη Στατιστική ελέγχουμε τη μηδενική υπόθεση και γι αυτόν τον έλεγχο ορίζουμε ένα ***επίπεδο*** ή μια ***στάθμη σημαντικότητας*** *(significant level)* που είναι η πιθανότητα με την οποία δεχόμαστε να κάνουμε λάθος απορρίπτοντας τη μηδενική υπόθεση ενώ αυτή είναι σωστή. Η πιθανότητα αυτή συμβολίζεται με το ελληνικό γράμμα α. Συνήθως θέτουμε α = 0.05, ενώ υπάρχουν περιπτώσεις όπου θέτουμε α = 0.01 ή α = 0.1. Είναι προφανές ότι όταν θέσουμε α = 0.05 και ο στατιστικός έλεγχος, όπως οι έλεγχοι που θα εξετάσουμε παρακάτω, μας οδηγήσει στην απόρριψη της μηδενικής υπόθεσης, τότε είμαστε 95% βέβαιοι ότι πήραμε τη σωστή απόφαση και συνεπώς η πιθανότητα να έχουμε κάνει λάθος είναι 0.05 ή 5%.

Κάθε στατιστικός έλεγχος έχει δύο δυνατά αποτελέσματα: Να απορριφθεί η μηδενική υπόθεση ή να μην απορριφθεί. Όταν απορρίπτεται η μηδενική υπόθεση, αποδεχόμαστε την εναλλακτική της. Σε αυτή την απόφαση υπάρχει μια πιθανότητα ίση με α να έχουμε κάνει λάθος. Αν όμως

ο στατιστικός έλεγχος οδηγήσει στη μη απόρριψη της μηδενικής υπόθεσης, δεν σημαίνει ότι πρέπει να αποδεχθούμε την υπόθεση αυτή. Ο λόγος που δεν αποδεχόμαστε τη μηδενική υπόθεση είναι επειδή η λέξη αποδοχή υπονοεί ότι η υπόθεση αυτή είναι σωστή. Όμως, αν αποδεχτούμε τη H_0, τότε δεν μπορούμε σε καμιά περίπτωση να εκτιμήσουμε τον κίνδυνο να έχουμε κάνει λάθος. **Για το λόγο αυτό στη στατιστική αξία έχει μόνο όταν απορρίπτεται η μηδενική υπόθεση**.

Για την απόρριψη της μηδενικής υπόθεσης κάθε στατιστικός έλεγχος υπολογίζει μια πιθανότητα που συμβολίζεται με p-value. Η πιθανότητα αυτή, που θα οριστεί πιο αυστηρά στην ενότητα 5.5, είναι ουσιαστικά το ***ελάχιστο επίπεδο σημαντικότητας*** στο οποίο μπορεί να απορριφθεί η μηδενική υπόθεση. Όταν η τιμή p-value είναι μικρότερη του α, η μηδενική υπόθεση απορρίπτεται και αποδεχόμαστε την εναλλακτική, H_1. Όταν όμως η p-value είναι μεγαλύτερη του α, τότε υπάρχει "αποτυχία απόρριψης" της μηδενικής υπόθεσης, με την έννοια ότι τα δεδομένα δεν είναι επαρκή ώστε να μας οδηγήσουν στο να αποδεχθούμε την H_1 αντί για τη H_0.

Αν με βάση τα στατιστικά δεδομένα απορρίψουμε μια υπόθεση που είναι αληθινή, τότε λέμε ότι κάνουμε ένα ***σφάλμα τύπου I***. Αντίθετα αν δεχθούμε μια λανθασμένη υπόθεση ως σωστή, τότε κάνουμε ένα ***σφάλμα τύπου II***. Δυστυχώς όταν προσπαθούμε να περιορίσουμε ένα σφάλμα τύπου I αυξάνουμε την πιθανότητα να κάνουμε ένα σφάλμα τύπου II. Η μόνη περίπτωση να ελαττώσουμε την πιθανότητα να κάνουμε σφάλμα τύπου I και II είναι να αυξήσουμε το μέγεθος των δειγμάτων.

5.3 ΜΟΝΟΠΛΕΥΡΟΙ ΚΑΙ ΔΙΠΛΕΥΡΟΙ ΕΛΕΓΧΟΙ

Όταν η εναλλακτική υπόθεση, H_1, διατυπώνεται με το σύμβολο $\neq$, τότε ο έλεγχος ονομάζεται ***δίπλευρος*** *(two-tailed)*. Αν η εναλλακτική υπόθεση, H_1, διατυπώνεται με το σύμβολο > ή το <, ο έλεγχος ονομάζεται ***μονόπλευρος*** *(one-tailed)*.

5.4 ΠΑΡΑΔΕΙΓΜΑ ΕΛΕΓΧΟΥ ΤΗΣ ΜΗΔΕΝΙΚΗΣ ΥΠΟΘΕΣΗΣ

Για να δούμε πώς ελέγχεται η μηδενική υπόθεση θα επανεξετάσουμε το παράδειγμα 4.2 του προηγούμενου κεφαλαίου. Σύμφωνα με αυτό, για τον έλεγχο μιας νέας αναλυτικής μεθόδου προσδιορισμού της σεληνιοουρίας στο νερό παρασκευάστηκε ένα διάλυμα της ουσίας αυτής με συγκέντρωση 40 ng/mL και ακολούθως έγιναν 5 μετρήσεις της συγκέντρωσης της σεληνιοουρίας. Ζητείται να εξετασθεί αν η μέθοδος είναι

αξιόπιστη λαμβάνοντας υπόψη ότι ελήφθησαν τα αποτελέσματα:

$$40.3,\ 40.2,\ 40.2,\ 40.0,\ 40.3$$

◆ Υπάρχουν δύο τρόποι για να ελέγξουμε τη μηδενική υπόθεση σε αυτό το πρόβλημα. Χρησιμοποιώντας διαστήματα εμπιστοσύνης ή συγκρίνοντας την τιμή της τυχαίας μεταβλητής t που υπολογίζεται από τη σχέση (4.1) με μια κρίσιμη τιμή. Και στις δύο περιπτώσεις πρέπει πρώτα να ορίσουμε α) τη μηδενική υπόθεση και την εναλλακτική της και β) τη στάθμη σημαντικότητας με βάση την οποία θα εξετάσουμε αν απορρίπτεται η μηδενική υπόθεση.

Στο συγκεκριμένο παράδειγμα ορίζουμε τις υποθέσεις H_0 και H_1 ως:

$$H_0: \mu = 40, \quad H_1: \mu \neq 40$$

Η μηδενική υπόθεση σημαίνει ότι το δείγμα προέρχεται από πληθυσμό με μέση τιμή μ = 40, δεδομένου ότι αυτή είναι η συγκέντρωση της σεληνιοουρίας στο πρότυπο διάλυμα. Σε ό,τι αφορά τη στάθμη σημαντικότητας, αυτή, όπως αναφέρθηκε, συνήθως ορίζεται στο α = 0.05, που σημαίνει ότι δεχόμαστε να κάνουμε ένα λάθος με πιθανότητα 5% απορρίπτοντας τη μηδενική υπόθεση ενώ αυτή είναι σωστή.

Α τρόπος

Υπολογίζουμε το 95% διάστημα εμπιστοσύνης για τη μέση τιμή και αν η τιμή μ = 40 είναι έξω από αυτό απορρίπτουμε τη μηδενική υπόθεση, επειδή αν δεν έχει συμβεί κάποιο συστηματικό λάθος, τότε η πιθανότητα το μ = 40 να είναι έξω από το 95% διάστημα εμπιστοσύνης για τη μέση τιμή είναι 5%. Στο συγκεκριμένο παράδειγμα είδαμε ότι το 95% διάστημα εμπιστοσύνης είναι:

$$40.2 \pm 0.15 = [40.05,\ 40.35]$$

Άρα με κίνδυνο να κάνουμε λάθος με πιθανότητα 5% μπορούμε να δεχθούμε ότι υπάρχει κάποιο συστηματικό σφάλμα ή στην παρασκευή του πρότυπου διαλύματος της σεληνιοουρίας ή στην προτεινόμενη αναλυτική μέθοδο με αποτέλεσμα οι τιμές του δείγματος να είναι στατιστικά μετατοπισμένες σε μεγαλύτερες τιμές από την αναμενόμενη τιμή μ = 40.

Ας υπολογίσουμε τώρα γενικά το P% διάστημα εμπιστοσύνης διευθετώντας τα δεδομένα και τους υπολογισμούς όπως στο σχήμα 5.1. Υπόψη ότι σύμφωνα με τη θεωρία που εξετάσαμε στο προηγούμενο κεφάλαιο, το P% = 100(1 - α) διάστημα εμπιστοσύνης μπορεί να γραφεί ως $\bar{x} \pm b$, όπου $b = t_{\alpha/2,m-1} s/\sqrt{m}$. Στη σχέση αυτή το $t_{\alpha/2,m-1}$

υπολογίζεται στο *Excel* από τον τύπο: =TINV(α;m-1).

	A	B	C	D	E	F	G
1	40,3	m=	5				
2	40,2	s=	0,12247	=	STDEV(A1:A5)		
3	40,2	<x>=	40,2	=	AVERAGE(A1:A5)		
4	40,0	a=	0,05				
5	40,3	b=	0,1521	=	TINV(C4;C1-1)*C2/SQRT(C1)		
6		<x>-b=	40,048				
7		<x>+b=	40,352				

	A	B	C	D	E	F	G
1	40,3	m=	5				
2	40,2	s=	0,12247	=	STDEV(A1:A5)		
3	40,2	<x>=	40,2	=	AVERAGE(A1:A5)		
4	40,0	a=	0,03				
5	40,3	b=	0,1806	=	TINV(C4;C1-1)*C2/SQRT(C1)		
6		<x>-b=	40,019				
7		<x>+b=	40,381				

	A	B	C	D	E	F	G
1	40,3	m=	5				
2	40,2	s=	0,12247	=	STDEV(A1:A5)		
3	40,2	<x>=	40,2	=	AVERAGE(A1:A5)		
4	40,0	a=	0,022				
5	40,3	b=	0,1993	=	TINV(C4;C1-1)*C2/SQRT(C1)		
6		<x>-b=	40,001				
7		<x>+b=	40,399				

Σχήμα 5.1. Διαδοχικοί προσδιορισμοί P% διαστημάτων εμπιστοσύνης στο *Excel* με α = 0.05, 0.03 και 0.022 από επάνω προς τα κάτω

Στο σχήμα 5.1-επάνω υπολογίζεται το 95% διάστημα εμπιστοσύνης, δεδομένου ότι α = 1 - P/100 = 0.05. Παρατηρούμε ότι το α που υπεισέρχεται στους υπολογισμούς μέσω της συνάρτησης $t_{\alpha/2,m-1}$ είναι ουσιαστικά η στάθμη σημαντικότητας για την απόρριψη της μηδενικής υπόθεσης. Αν ελαττώσουμε το α στην τιμή 0.03, τότε το P% = 100(1 - 0.03) = 97% διάστημα εμπιστοσύνης είναι το (σχήμα 5.1-μέσο)

$$40.2 \pm 0.18 = [40.02,\ 40.38]$$

όπου και πάλι το μ = 40 είναι έξω από αυτό το διάστημα. Συνεπώς μπορούμε να απορρίψουμε τη μηδενική υπόθεση με κίνδυνο λάθους 3%.

Αν συνεχίσουμε να ελαττώνουμε την τιμή του α θα παρατηρήσουμε

ότι υπάρχει ένα όριο, που είναι η τιμή α = 0.022 (σχήμα 5.1-κάτω). Κάτω από αυτό το όριο η μέση τιμή μ = 40 εισέρχεται μέσα στο αντίστοιχο P% διάστημα εμπιστοσύνης και συνεπώς η μηδενική υπόθεση δεν απορρίπτεται. Όπως αναφέρθηκε, η τιμή αυτή του α συμβολίζεται συνήθως με p-value και είναι η *ελάχιστη στάθμη σημαντικότητας* στην οποία μπορεί να απορριφθεί η μηδενική υπόθεση.

Συνεπώς καταλήγουμε στο συμπέρασμα ότι αν δεν έχει γίνει λάθος στην παρασκευή του πρότυπου διαλύματος της σεληνιοουρίας, τότε η μέθοδος εισάγει κάποιο συστηματικό σφάλμα και η πιθανότητα να κάνουμε λάθος σε αυτό το συμπέρασμα είναι 2.2%.

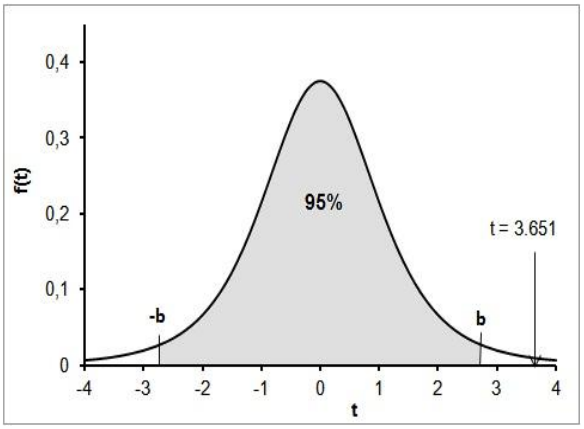
Β τρόπος

Υπολογίζουμε την τιμή μιας ***στατιστικής συνάρτησης ελέγχου*** *(test statistic)* που στο συγκεκριμένο πρόβλημα είναι η συνάρτηση t:

$$t = \frac{\bar{x} - \mu}{s}\sqrt{m} = \frac{40.2 - 40}{0.12247}\sqrt{5} = 3.651$$

Όπως έχει αναφερθεί, θεώρημα (4.1), η μεταβλητή t ακολουθεί την κατανομή *student* με 5-1 = 4 βαθμούς ελευθερίας. Για να κατανοήσουμε τι σημαίνει ότι t = 3.651 και τι ότι η μεταβλητή t ακολουθεί την κατανομή *student* με 4 βαθμούς ελευθερίας, έστω το σχήμα 5.2 που δίνεται η γραφική παράσταση της κατανομής *student* με 4 βαθμούς ελευθερίας.

Μπορεί να αποδειχθεί ότι στο σχήμα αυτό όταν b = 2.776, τότε το εμβαδόν κάτω από την καμπύλη από -b μέχρι b είναι 0.95. Αυτό σημαίνει ότι αν δημιουργούμε δείγματα με 5 τιμές συγκεντρώσεων από το διάλυμα των 40 ng/mL και η μέθοδος προσδιορισμού της συγκέντρωσης δεν εισάγει κάποιο συστηματικό σφάλμα, όταν δηλαδή ισχύει η μηδενική υπόθεση, τότε η τιμή του t που υπολογίζουμε σε κάθε δείγμα έχει πιθανότητα 95% να βρίσκεται στο διάστημα [-2.776, 2.776]. Αν όμως η μέθοδος εισάγει κάποιο συστηματικό σφάλμα, π.χ. υπερεκτιμά τη συγκέντρωση της ουσίας στο διάλυμα, τότε οι μέσες τιμές των δειγμάτων θα είναι συστηματικά μεγαλύτερες από το 40 και οι υπολογιζόμενες τιμές του t θα βρίσκονται με μεγάλη πιθανότητα πέρα του 2.776. Φυσικά υπάρχει και μια πιθανότητα του 5% η τιμή του t να είναι έξω από το διάστημα [-2.776, 2.776] λόγω τυχαίων παραγόντων.



Σχήμα 5.2. Γραφική παράσταση της κατανομής student με 4 βαθμούς ελευθερίας. Ισχύει b = 2.776

Επομένως μια τιμή του t μεγαλύτερη από το 2.776 (ή μικρότερη από το -2.776) μπορεί να σημαίνει την ύπαρξη συστηματικού σφάλματος, ενώ υπάρχει και μια πιθανότητα 5% αυτό να οφείλεται σε τυχαίους λόγους. Επειδή όμως η πιθανότητα να μην υπάρχει συστηματικό σφάλμα είναι μόνο 5%, δηλαδή πολύ μικρή, συμπεραίνουμε ότι στις μετρήσεις υπεισέρχεται συστηματικό σφάλμα, έχοντας φυσικά κατά νου ότι υπάρχει και μια πιθανότητα 5% να κάνουμε λάθος. Συνεπώς απορρίπτουμε τη μηδενική υπόθεση, γνωρίζοντας ότι μπορεί να σφάλουμε με μέγιστη πιθανότητα 5%.

Η τιμή b = 2.776 ονομάζεται ***κρίσιμη τιμή*** *(critical value)*, συνήθως συμβολίζεται με t(cr), εξαρτάται από το επίπεδο σημαντικότητας α και το πλήθος των τιμών m του δείγματος και στο *Excel* υπολογίζεται από τον τύπο: =TINV(α; m-1).

Επομένως, από τα παραπάνω προκύπτει ότι όταν $|t| \geq t(cr)$, απορρίπτουμε τη μηδενική υπόθεση. Είναι προφανές ότι για οποιαδήποτε τιμή του |t| που είναι μεγαλύτερη ή ίση με t(cr) = b = 2.776 (όταν m = 5) η μηδενική υπόθεση θα απορρίπτεται σε επίπεδο σημαντικότητας 0.05. Αν τώρα μετατοπίσουμε το b από το 2.776 στο 3.0, το εμβαδόν κάτω από την καμπύλη και στο διάστημα [-3, 3] αποδεικνύεται ότι είναι 0.96 και συνεπώς το εμβαδόν που είναι έξω από το διάστημα [-3, 3] γίνεται 0.04. Αλλά και σε αυτή την περίπτωση ισχύει $|t| = 3.651 \geq 3.00$ και επομένως η μηδενική υπόθεση απορρίπτεται, όμως τώρα η απόρριψη γίνεται σε επίπεδο σημαντικότητας α = 0.04.

Συνεχίζουμε να μετατοπίζουμε το b μέχρι την τιμή του t = 3.651. Τώρα το εμβαδόν κάτω από την καμπύλη από -3.651 έως 3.651 γίνεται 0.978 και συνεπώς το εμβαδόν έξω από το διάστημα [-3.651, 3.651]

γίνεται 0.022. Προφανώς και σε αυτή την πολύ οριακή περίπτωση, επειδή ισχύει $|t| = 3.651 \geq 3.651 = t(cr)$ (πάντα για $m = 5$), η μηδενική υπόθεση μπορεί να απορριφθεί στο ελάχιστο επίπεδο σημαντικότητας $\alpha = 0.022$. Συνεπώς, καταλήγουμε και πάλι στο συμπέρασμα ότι η νέα αναλυτική μέθοδος εισάγει κάποιο συστηματικό σφάλμα και η πιθανότητα να κάνουμε λάθος σε αυτό το συμπέρασμα είναι 0.022, δηλαδή 2.2%.

Παρατήρηση 1. Στο παράδειγμα που εξετάζουμε ισχύει p-value = 0.022 και επειδή η p-value είναι η *ελάχιστη στάθμη σημαντικότητας* στην οποία μπορεί να απορριφθεί η μηδενική υπόθεση, καταλήγουμε στο συμπέρασμα ότι μπορούμε να απορρίψουμε τη μηδενική υπόθεση με πιθανότητα σφάλματος 2.2%.

Αν και η ερμηνεία αυτή της p-value είναι ιδιαίτερα φιλική, η πλειοψηφία των στατιστικολόγων προτιμά την ακόλουθη πιο γενική προσέγγιση. Όταν p-value < α απορρίπτουμε τη μηδενική υπόθεση και στην απόφαση αυτή είμαστε τόσο πιο σίγουροι όσο πιο μικρή είναι η τιμή p-value σε σχέση με την τιμή του α.

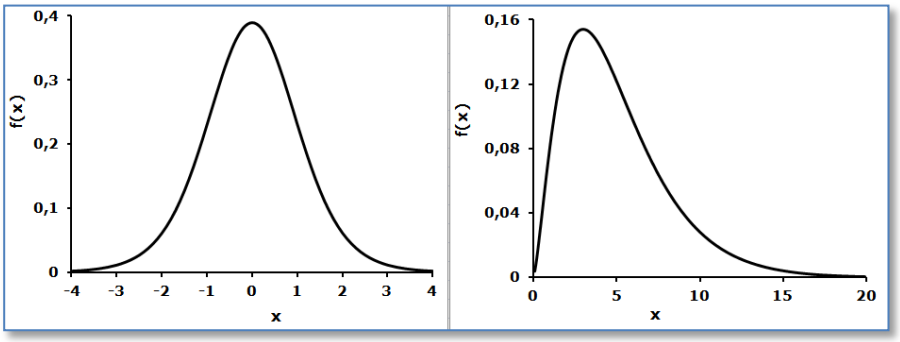
Παρατήρηση 2. Βασική υπόθεση για να ισχύει η παραπάνω ανάλυση είναι ότι το δείγμα των συγκεντρώσεων της σεληνιοουρίας ακολουθεί την κανονική κατανομή. Η υπόθεση αυτή πρέπει να ελεγχθεί με τα κριτήρια που θα εξεταστούν στο επόμενο κεφάλαιο.

5.5 ΓΕΝΙΚΕΥΣΗ

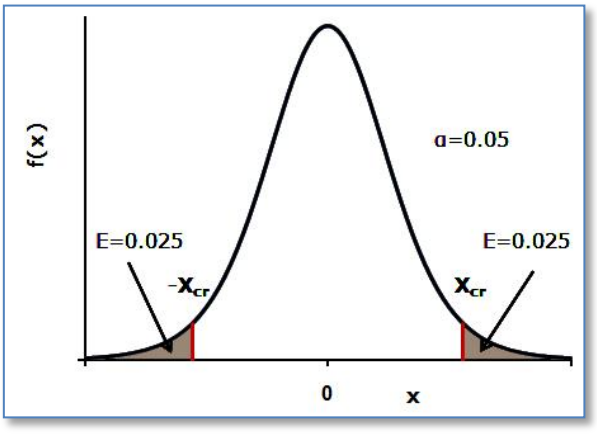
Για να ελέγξουμε μία στατιστική υπόθεση πρώτα ορίζουμε τη μηδενική υπόθεση H_0 και την εναλλακτική της H_1. Σε κάθε μηδενική υπόθεση αντιστοιχεί τουλάχιστον μία ***στατιστική συνάρτηση ελέγχου*** *(test statistic)*, έστω η Χ. Αυτή ακολουθεί μια συγκεκριμένη κατανομή που μπορεί να είναι συμμετρική ή ασύμμετρη, όπως στα παραδείγματα του σχήματος 5.3.

Για τον έλεγχο της μηδενικής υπόθεση θέτουμε πάντα μια ***στάθμη σημαντικότητας*** *(significant level)*, α, που είναι η μέγιστη πιθανότητα με την οποία δεχόμαστε να κάνουμε λάθος απορρίπτοντας τη μηδενική υπόθεση ενώ αυτή είναι σωστή. Με βάση τη στάθμη σημαντικότητας α υπολογίζεται η ***κρίσιμη τιμή*** *(critical value)* της στατιστικής συνάρτησης ελέγχου, X_{cr}, ως εξής. Σε δίπλευρο έλεγχο, X_{cr} είναι εκείνη η τιμή της μεταβλητής Χ για την οποία το εμβαδόν Ε που είναι κάτω από την καμπύλη της κατανομής στο διάστημα $[X_{cr}, \infty)$ είναι ίσο με α/2 (σχήμα 5.4). Σε

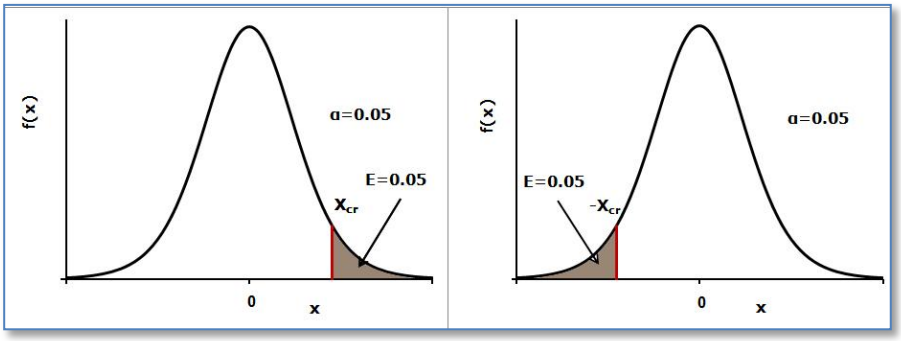
μονόπλευρο έλεγχο, X_{cr} είναι εκείνη η τιμή της X για την οποία το εμβαδόν E που είναι κάτω από την καμπύλη της κατανομής στο διάστημα $[X_{cr}, \infty)$ ή στο διάστημα, $(-\infty, -X_{cr}]$ είναι ίσο με α (σχήμα 5.5).



Σχήμα 5.3. Γραφική παράσταση της κατανομής *student* με 10 βαθμούς ελευθερίας (αριστερά) και της χ^2 με 5 βαθμούς ελευθερίας (δεξιά)

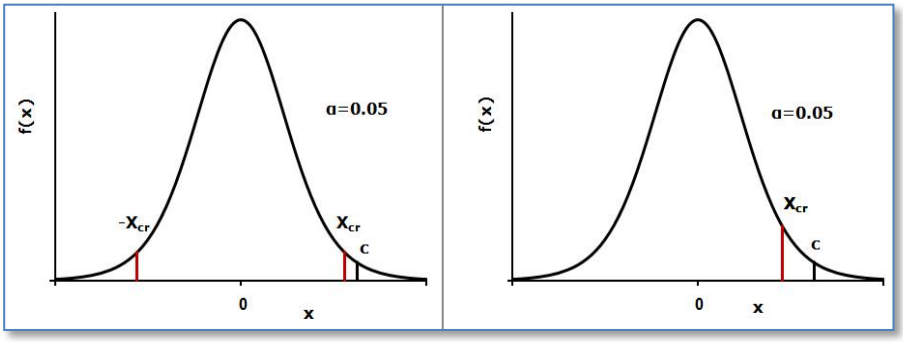


Σχήμα 5.4. Ορισμός της κρίσιμης τιμής X_{cr} σε δίπλευρο έλεγχο όταν η X ακολουθεί συμμετρική κατανομή



Σχήμα 5.5. Ορισμός της κρίσιμης τιμής X_{cr} σε μονόπλευρο έλεγχο όταν η X ακολουθεί συμμετρική κατανομή

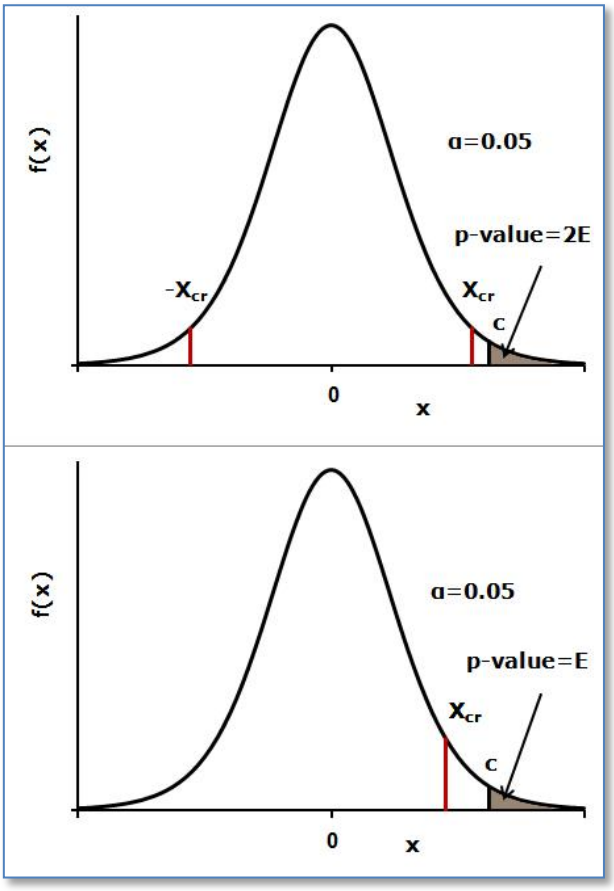
Ακολούθως με βάση τις τιμές του δείγματος ή των δειγμάτων που εμπλέκονται στη μηδενική υπόθεση υπολογίζουμε την τιμή της μεταβλητής X και έστω X = c. Αν η τιμή του c είναι μεγαλύτερη της κρίσιμης τιμής X_{cr} ή μικρότερη της $-X_{cr}$, η μηδενική υπόθεση απορρίπτεται, επειδή η πιθανότητα να συμβαίνει αυτό λόγω τυχαίων παραγόντων είναι μικρότερη από α (σχήμα 5.6). Υπόψη ότι από τις ιδιότητες των συναρτήσεων κατανομής προκύπτει ότι το ολοκλήρωμα μιας συνάρτησης κατανομής από a έως b και συνεπώς το εμβαδόν στο ίδιο διάστημα ισούται με την πιθανότητα που έχει η μεταβλητή της συνάρτησης κατανομής να βρίσκεται στο διάστημα [a, b].



Σχήμα 5.6. Σύγκριση X_{cr} και c για την απόρριψη της μηδενικής υπόθεσης σε δίπλευρο (αριστερά) και μονόπλευρο (δεξιά) έλεγχο

Στα σύγχρονα στατιστικά προγράμματα αντί για τη σύγκριση των τιμών X_{cr} και c υπολογίζεται η πιθανότητα p-value και συγκρίνεται με την

στάθμη σημαντικότητας α. Σε δίπλευρο έλεγχο η p-value είναι δύο φορές το εμβαδόν που είναι κάτω από την καμπύλη της κατανομής στο διάστημα [c, ∞), δηλαδή p-value = P(|X| ≥ c) = 2E, ενώ σε μονόπλευρο έλεγχο ισχύει p-value = P(X ≥ c) = E. Εποπτικά, ο ορισμός της p-value δίνεται στο σχήμα 5.7, από το οποίο εύκολα προκύπτει ότι η p-value είναι ουσιαστικά το *ελάχιστο επίπεδο σημαντικότητας* στο οποίο μπορεί να απορριφθεί η μηδενική υπόθεση, αρκεί να λάβουμε υπόψη ότι το α σε δίπλευρο έλεγχο είναι δύο φορές το εμβαδόν στο διάστημα [X_{cr}, ∞) και μία φορά το εμβαδόν αυτό σε μονόπλευρο έλεγχο.



Σχήμα 5.7. Ορισμός της p-value σε δίπλευρο (επάνω) και σε μονόπλευρο (κάτω) έλεγχο

Από την παραπάνω ανάλυση προκύπτει ότι ισχύει πάντα:

p-value (δίπλευρος έλεγχος) = 2p-value (μονόπλευρος έλεγχος)

Όταν η τιμή p-value είναι μικρότερη του α, τότε ισχύει $c \geq X_{cr}$ ή $c \leq -X_{cr}$ και συνεπώς η μηδενική υπόθεση μπορεί να απορριφθεί και μάλιστα σε επίπεδο σημαντικότητας α = p-value. Όταν η τιμή p-value είναι μεγαλύτερη του α, η μηδενική υπόθεση δεν μπορεί να απορριφθεί, όμως ταυτόχρονα δεν γίνεται αποδεκτή επειδή δεν μπορούμε σε καμιά περίπτωση να εκτιμήσουμε τον κίνδυνο να έχουμε κάνει λάθος. Συνεπώς μόνο από την απόρριψη της μηδενικής υπόθεσης μπορούμε να πάρουμε χρήσιμες πληροφορίες, επειδή τότε μπορούμε να δεχθούμε ότι ισχύει η εναλλακτική υπόθεση.

5.6 ΠΑΡΑΜΕΤΡΙΚΟΙ ΚΑΙ ΜΗ ΠΑΡΑΜΕΤΡΙΚΟΙ ΕΛΕΓΧΟΙ

Οι κυριότεροι στατιστικοί έλεγχοι έχουν ως βασική προϋπόθεση ότι για να ακολουθεί η στατιστική συνάρτηση ελέγχου Χ μια ορισμένη κατανομή πρέπει οι τιμές του δείγματος ή των δειγμάτων που μετέχουν στον έλεγχο να ακολουθούν την κανονική κατανομή. Αν η προϋπόθεση αυτή δεν ισχύει, η στατιστική ανάλυση είναι πολύ πιθανό να οδηγήσει σε εσφαλμένα συμπεράσματα με ανεξέλεγκτο μέγεθος σφάλματος.

Επομένως είναι εμφανής η ανάγκη για στατιστικούς ελέγχους που να επιτρέπουν την ανάλυση δειγμάτων με τιμές για τις οποίες δε γνωρίζουμε αν ακολουθούν και ποια κατανομή. Έτσι ως ***μη-παραμετρικοί στατιστικοί έλεγχοι*** *ή* ***δοκιμασίες*** *(non-parametric tests)* ορίζονται γενικά εκείνοι οι έλεγχοι, στους οποίους δεν απαιτούνται παραδοχές σχετικά με την κατανομή που ακολουθούν οι τιμές των δειγμάτων. Σε αντίθεση, οι στατιστικοί έλεγχοι που στηρίζονται στην παραδοχή ότι οι τιμές των δειγμάτων προέρχονται από πληθυσμό ή πληθυσμούς που ακολουθούν μια συγκεκριμένη κατανομή, συνήθως την κανονική κατανομή, ονομάζονται ***παραμετρικοί έλεγχοι*** *(parametric tests)*.

Η κύρια διαφοροποίηση των μη παραμετρικών ελέγχων από τους παραμετρικούς είναι ότι στη μη παραμετρική στατιστική δεν αναλύουμε τα δείγματα καθαυτά, αλλά δημιουργούμε και αναλύουμε ***βαθμούς*** *(ranks)* ή ***ροές*** *(runs)* δεδομένων. Αν $\{x_1, x_2, ..., x_m\}$ είναι ένα τυχαίο δείγμα με ποσοτικά δεδομένα, ονομάζουμε βαθμό της τιμής x_i τον αριθμό r_i των δεδομένων του δείγματος που είναι μικρότερα ή ίσα με το x_i. Η έννοια της ροής (run) σχετίζεται με ακολουθίες που σχηματίζονται από σύμβολα δύο ειδών, όπως για παράδειγμα η ακολουθία:

+ + + o o + o o o o + + +

Στην ακολουθία αυτή έχουμε πέντε ροές: '+++', 'oo', '+', 'oooo' και '+++' με μήκη 3, 2, 1, 4 και 3. Τα ποσοτικά δεδομένα ενός δείγματος μετατρέπονται εύκολα σε ροές αν σε κάθε τιμή μικρότερη από τη διάμεσο αντιστοιχίσουμε το σύμβολο + ή όποιο άλλο σύμβολο θέλουμε και σε κάθε τιμή μεγαλύτερη από τη διάμεσο αντιστοιχίσουμε το σύμβολο - ή κάποιο άλλο. Τιμές ίσες με τη διάμεσο παραλείπονται.

Η αντικατάσταση των τιμών του δείγματος με βαθμούς ή ακολουθίες ροών μας επιτρέπει να αγνοούμε την κατανομή των τιμών του δείγματος. Αυτό όμως έχει το μειονέκτημα ότι σε αρκετές μη παραμετρικές δοκιμασίες οι πληροφορίες που παίρνουμε να είναι λιγότερες από τις αντίστοιχες των παραμετρικών δοκιμασιών. Επίσης επειδή στους βαθμούς ή στις ροές δεν υπάρχει η έννοια της μέσης τιμής, στους μη-παραμετρικούς ελέγχους οι υποθέσεις H_0 και H_1 δεν περιέχουν τη μέση τιμή. Τέλος, θα πρέπει να τονιστεί ότι αν μία στατιστική υπόθεση ελέγχεται τόσο με παραμετρικό όσο και με μη-παραμετρικό έλεγχο, ο παραμετρικός έλεγχος είναι ισχυρότερος του μη-παραμετρικού όταν όμως πληρούνται οι προϋποθέσεις εφαρμογής του παραμετρικού ελέγχου.

Δυνατότητες άμεσης εφαρμογής μη-παραμετρικών ελέγχων διαθέτουν τα προγράμματα *ChemStat* και *SPSS*. Στο *Excel* η εφαρμογή τους είναι σχετικά χρονοβόρα. Για το λόγο αυτό μόνο ενδεικτικές εφαρμογές μη-παραμετρικών ελέγχων στο *Excel* δίνονται στο Παράρτημα IV.

Τέλος, στην κατηγορία των μη-παραμετρικών ελέγχων μπορεί να θεωρηθεί ότι ανήκουν οι έλεγχοι που βασίζονται στις μεθόδους ***Bootstrap*** και ***Monte-Carlo με αντιμεταθέσεις*** *(permutations)*. Οι μέθοδοι αυτές περιγράφονται Κεφάλαιο 7.5.

Κεφάλαιο 6

ΕΛΕΓΧΟΙ ΣΤΑΤΙΣΤΙΚΩΝ ΥΠΟΘΕΣΕΩΝ ΣΕ ΕΝΑ ΔΕΙΓΜΑ

6.1 ΓΕΝΙΚΑ

Οι κυριότεροι στατιστικοί έλεγχοι που μπορούν να γίνουν σε ένα δείγμα είναι ο έλεγχος της κανονικότητας των τιμών του δείγματος, η ύπαρξη ακραίων τιμών και ο έλεγχος της μέσης τιμής που ήδη εξετάσαμε ως παράδειγμα στατιστικού ελέγχου στο κεφάλαιο 5.4. Από τους ελέγχους αυτούς, οι έλεγχοι κανονικότητας και ακραίων τιμών ανήκουν στην κατηγορία των μη παραμετρικών ελέγχων, ενώ ο έλεγχος της μέσης τιμής δείγματος είναι κατά βάση παραμετρικός αν και υπάρχουν μη παραμετρικές τεχνικές.

6.2 ΕΛΕΓΧΟΣ ΤΗΣ ΚΑΝΟΝΙΚΟΤΗΤΑΣ

Ο έλεγχος της κανονικότητας των τιμών ενός δείγματος πρέπει να είναι ο πρώτος και ίσως ο βασικότερος έλεγχος για μια σωστή ανάλυση των δεδομένων ενός πειράματος με βάση τη στατιστική. Όλα τα στατιστικά προγράμματα έχουν τους βασικούς ελέγχους, που είναι τα κριτήρια *Shapiro-Wilk* και *Kolmogorov-Smirnov* με ισχυρότερο το πρώτο κριτήριο. Εναλλακτικά μπορεί να χρησιμοποιηθεί ο έλεγχος *Anderson-Darling*, που είναι επίσης ένας ισχυρός έλεγχος, απαιτεί όμως δείγματα με πλήθος τιμών μεγαλύτερο από 6.

Στο κριτήριο *Shapiro-Wilk* οι m τιμές του δείγματος διατάσσονται

κατά αύξουσα σειρά και ακολούθως υπολογίζεται η ποσότητα:

$$W = \frac{(\sum_{i=1}^{k} a_i(x_{m-i+1} - x_i))^2}{\sum_{i=1}^{m}(x_i - \bar{x})^2} \qquad (6.1)$$

όπου k = m/2 ή k = (m-1)/2 ανάλογα αν ο m είναι άρτιος ή περιττός, ενώ οι σταθερές a_i λαμβάνονται από πίνακες.

Σύμφωνα με το κριτήριο των *Kolmogorov-Smirnov* υπολογίζεται η μέγιστη απόλυτη απόκλιση μεταξύ των πειραματικών τιμών της εκατοστιαίας αθροιστικής συχνότητας, F_O, και των αντίστοιχων θεωρητικών τιμών της κανονικής κατανομής, F_E:

$$d = \max|F_O - F_E| \qquad (6.2)$$

Τέλος, στο κριτήριο *Anderson-Darling* οι τιμές x_i του δείγματος κανονικοποιούνται με βάση τη σχέση $z_i = (x_i - \bar{x})/s$ και ακολούθως υπολογίζεται η στατιστική συνάρτηση:

$$A^2 = -m - \frac{1}{m}\sum(2i-1)[\ln F(z_i) + \ln(1 - F(z_{n+1-i}))] \qquad (6.3)$$

όπου F είναι η αθροιστική συνάρτηση της κανονικής κατανομής.

Οι στατιστικές μεταβλητές W, d και A^2 ακολουθούν ειδικές κατανομές που επιτρέπουν να υπολογιστεί η πιθανότητα p-value, που καθορίζει αν θα απορρίψουμε ή όχι τη μηδενική υπόθεση:

H_0: Το δείγμα προέρχεται από κανονικό πληθυσμό

με εναλλακτική

H_1: Το δείγμα δεν προέρχεται από κανονικό πληθυσμό

Από τους παραπάνω ελέγχους, τα κριτήρια *Kolmogorov-Smirnov* και *Shapiro-Wilk* χρησιμοποιούνται στο *SPSS*, ενώ ο έλεγχος *Anderson-Darling* στο *ChemStat*. Στο *Excel* δεν υπάρχει δυνατότητα άμεσης εφαρμογής κάποιου ελέγχου. Σε αυτή την περίπτωση μπορούμε σχετικά εύκολα να κατασκευάσουμε το *διάγραμμα Q-Q*, με την προϋπόθεση όμως το δείγμα έχει πάνω από 10 ίσως και πάνω από 20 τιμές. Το διάγραμμα αυτό περιγράφεται στο παράδειγμα 6.2. Διαγράμματα Q-Q μπορεί να γίνουν και στο *ChemStat* και στο *SPSS*.

Παράδειγμα 6.1

Να εξετασθεί η κανονικότητα των τιμών του δείγματος του παραδείγματος 4.2:

40.3, 40.2, 40.2, 40.0, 40.3

◆ Επειδή το δείγμα είναι μικρό δεν θα εξεταστούν διαγράμματα Q-Q.

❖ **Ανάλυση στο ChemStat**

Ο έλεγχος της κανονικότητας με το πρόγραμμα *ChemStat* γίνεται από *Πρόσθετα (Add-ins)* → *ChemStat* → *Normality Test*. Στο πρώτο παράθυρο που ανοίγει πληκτρολογούμε τον αριθμό των δειγμάτων που θέλουμε να ελέγξουμε, δηλαδή τον αριθμό 1, και στο δεύτερο εισάγουμε με το ποντίκι την περιοχή τιμών του δείγματος, που πρέπει να βρίσκονται σε μία στήλη. Αν ελέγχουμε περισσότερα του ενός δείγματα, ανοίγουν διαδοχικά παράθυρα για την είσοδο κάθε δείγματος χωριστά. Σε αυτή την περίπτωση πρώτο εισάγουμε το μεγαλύτερο δείγμα. Στο τελευταίο παράθυρο ορίζουμε το κελί εξόδου των αποτελεσμάτων, που ουσιαστικά είναι το επάνω και αριστερά κελί του πίνακα των αποτελεσμάτων.

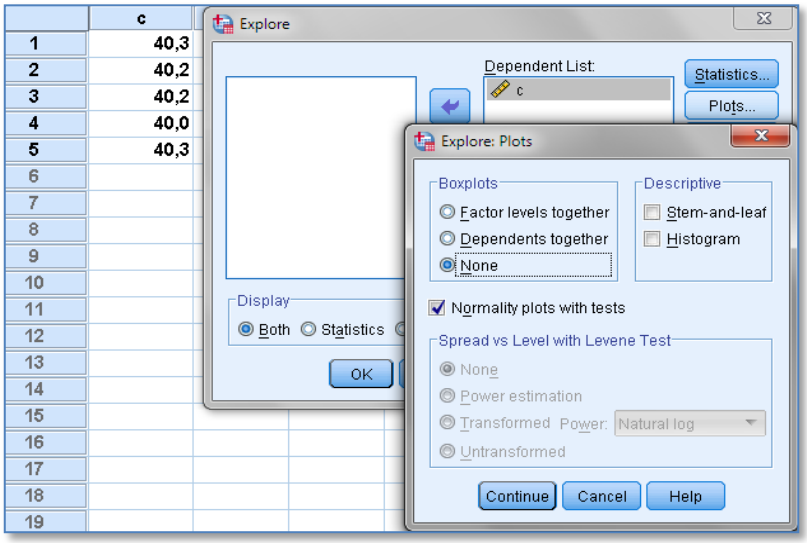
40,3	Anderson-Darling Normality Test			
40,2				
40,2	Sample	p-value	Comments	
40,0	1	0,1354	Normality may be assumed	Samples must have more than 6 data
40,3				

Σχήμα 6.1. Αποτελέσματα του ελέγχου *Anderson-Darling* σε δείγμα με λιγότερες από 6 τιμές

Στο παράδειγμα που εξετάζουμε παίρνουμε τα αποτελέσματα που δίνονται στο σχήμα 6.1. Παρατηρούμε ότι το πρόγραμμα εφαρμόζει το κριτήριο *Anderson-Darling* και μας προειδοποιεί ότι το δείγμα έχει λιγότερες από 6 τιμές. Έτσι με κίνδυνο το κριτήριο να μην είναι απόλυτα αξιόπιστο στο συγκεκριμένο παράδειγμα, το πρόγραμμα υπολογίζει την τιμή p-value = 0.1354 > 0.05. Συνεπώς, σε επίπεδο σημαντικότητας α = 0.05, δεν μπορούμε να απορρίψουμε τη μηδενική υπόθεση και να αποδεχθούμε την εναλλακτική υπόθεση H_1 ότι το δείγμα δεν είναι κανονικό. Άρα σε επίπεδο σημαντικότητας α = 0.05 δεν διαπιστώνονται στατιστικά σημαντικές αποκλίσεις από την κανονικότητα.

❖ **Ανάλυση στο SPSS**

Στο SPSS εισάγουμε τα δεδομένα σε μία στήλη, την οποία ονομάζουμε έστω c, πηγαίνουμε *Analyze* → *Descriptive Statistics* → *Explore* και στο παράθυρο που ανοίγει επιλέγουμε τη μεταβλητή c και τη μεταφέρουμε στο πλαίσιο *Dependent List.* Συνεχίζουμε με *κλικ* στο κουμπί *Plots* και στο νέο πλαίσιο που εμφανίζεται επιλέγουμε *Normality plots with tests*, ενώ απενεργοποιούμε το *Stem-and-leaf* και τα *Boxplots* (σχήμα 6.2). Με *κλικ* στο *Continue* και μετά στο *OK* παίρνουμε, ανάμεσα στα άλλα, τον πίνακα αποτελεσμάτων του σχήματος 6.3.



Σχήμα 6.2. Εισαγωγή δεδομένων για έλεγχο κανονικότητας στο *SPSS*

Στο *SPSS* η πιθανότητα p-value συμβολίζεται με Sig. Παρατηρούμε ότι η τιμή Sig. = 0.146 από τον έλεγχο *Shapiro-Wilk* και η τιμή Sig. = 0.161 από τον έλεγχο *Kolmogorov-Smirnov* είναι μεγαλύτερες από 0.05 και συνεπώς σε επίπεδο σημαντικότητας α = 0.05 δεν διαπιστώνονται στατιστικά σημαντικές αποκλίσεις από την κανονικότητα. Αυτό δεν σημαίνει ότι δεν υπάρχουν αποκλίσεις από την κανονικότητα. Μπορεί να υπάρχουν και να μην έχουν εντοπιστεί ή να μην υπάρχουν. Ο έλεγχος δεν μας οδηγεί σε κάποια βεβαιότητα. Πάντως όταν δεν διαπιστώνονται στατιστικά σημαντικές αποκλίσεις από την κανονικότητα, δηλαδή όταν p-value ή Sig. > 0.05, τότε μπορούμε να χρησιμοποιήσουμε το δείγμα σε παραμετρικούς ελέγχους.

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
c	,300	5	,161	,833	5	,146

a. Lilliefors Significance Correction

Σχήμα 6.3. Αποτελέσματα ελέγχου κανονικότητας στο *SPSS*

Βασική παρατήρηση. Μετρήσεις που γίνονται στο ίδιο σύστημα κάτω από σταθερές και ελεγχόμενες συνθήκες, όπως είναι μετρήσεις pH, συγκέντρωσης, ιξώδους, πυκνότητας κ.ο.κ, κατά κανόνα ακολουθούν την κανονική κατανομή. Η παρατήρηση αυτή είναι ισχυρότερη από τα αποτελέσματα του ελέγχου της κανονικότητας με βάση τα κριτήρια που έχουν αναφερθεί παραπάνω.

Αντίθετα πρέπει να είμαστε ιδιαίτερα προσεκτικοί στην κανονικότητα δειγμάτων που προέρχονται από μετρήσεις σε διαφορετικά συστήματα ή στο ίδιο σύστημα αλλά κάτω από μεταβαλλόμενες συνθήκες. Τέτοιες μετρήσεις αφορούν περιβαλλοντικά και βιολογικά δείγματα, δείγματα τροφίμων κ.ά.

Παράδειγμα 6.2

Να εξετασθεί η κανονικότητα των 20 πρώτων τιμών του δείγματος του πίνακα 2.5 και όλων των τιμών του πίνακα 2.4. Επιπλέον να γίνουν τα διαγράμματα Q-Q.

◆ Για τις 20 πρώτες τιμές του πίνακα 2.5 οι έλεγχοι *Kolmogorov-Smirnov*, *Shapiro-Wilk* και *Anderson-Darling* δίνουν τις τιμές p-value = 0.122, 0.002 και 0.011, αντίστοιχα, ενώ για τις τιμές του πίνακα 2.4 τα κριτήρια αυτά δίνουν p-value = 0.2, 0.973 και 0.913. Συνεπώς το πρώτο δείγμα είναι πολύ πιθανό να παρουσιάζει αποκλίσεις από την κανονικότητα αν και το κριτήριο *Kolmogorov-Smirnov*, που όμως, όπως αναφέρθηκε, δεν θεωρείται ισχυρό, οδηγεί στο αντίθετο συμπέρασμα. Στο δεύτερο δείγμα και σε επίπεδο σημαντικότητας α = 0.05 δεν διαπιστώνονται στατιστικά σημαντικές αποκλίσεις από την κανονικότητα (p-value > 0.05).

Σε ό,τι αφορά τα *διαγράμματα Q-Q* θα εξετάσουμε πρώτα πως κατασκευάζονται στο *Excel* και μετά πως δημιουργούνται τέτοια διαγράμματα με το *ChemStat* και το *SPSS*.

❖ **Ανάλυση στο Excel**

Για να κατασκευάσουμε το *διάγραμμα Q-Q* τοποθετούμε τις τιμές του δείγματος σε μία στήλη, έστω τη B, όπως φαίνεται στο σχήμα 6.4, και ακολούθως τις κατατάσσουμε κατά αύξουσα σειρά χρησιμοποιώντας το εργαλείο *Ταξινόμηση από το Α προς το Ω* (*Sort A to Z*) από το *Κεντρική* (*Home) → Επεξεργασία (Editing) → Ταξινόμηση & φιλτράρισμα (Sort & Filter)*. Στα κελιά B23 και B24 υπολογίζουμε τη μέση τιμή και την τυπική απόκλιση των τιμών του δείγματος. Με βάση αυτές τις τιμές στη στήλη C υπολογίζουμε τις κανονικοποιημένες τιμές του δείγματος χρησιμοποιώντας τη σχέση $z = (x - \bar{x})/s$. Στη επόμενη στήλη υπολογίζουμε την *Αθροιστική Συχνότητα* (*Cumulative Frequency*) κάθε τιμής, που είναι ο αριθμός των τιμών του δείγματος που είναι μικρότερος ή ίσος από αυτή την τιμή. Για λόγους απλότητας και επειδή σε συνεχή δεδομένα καμία τιμή στην πραγματικότητα δεν είναι ίδια με κάποια άλλη τιμή, στη στήλη D μπορούμε να πληκτρολογήσουμε τους αριθμούς 1, 2, 3, …, 20. Τέλος, στη στήλη E υπολογίζεται η *Εκατοστιαία Αθροιστική Συχνότητα* από τη σχέση

% Αθροιστική Συχνότητα = (*Αθροιστική Συχνότητα-0.5)*/m

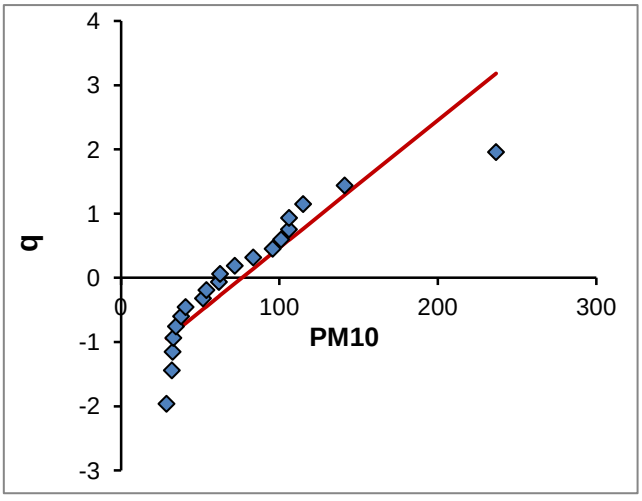
Στο σχήμα 6.4 η ποσότητα αυτή συμβολίζεται με FO. Όμως ισχύει % *Αθροιστική Συχνότητα* = F(x), όπου F(x) είναι η αθροιστική συνάρτηση κατανομής. Συνεπώς στη στήλη F μπορούμε να υπολογίσουμε τις τιμές z που αντιστοιχούν στις τιμές της % *Αθροιστικής Συχνότητας* όταν οι τιμές αυτές ακολουθούν την τυπικά κανονική κατανομή. Συμβολίζουμε τις τιμές αυτές με q. Για να τις υπολογίσουμε, πληκτρολογούμε στο F2 τον τύπο =NORMSINV(E2) και συμπληρώνουμε τη στήλη με τη διαδικασία της αυτόματης συμπλήρωσης.

Με βάση τις τιμές των z και q κατασκευάζουμε το *διάγραμμα Q-Q* της μεταβολής των q, z ως προς PM10, το οποίο μορφοποιούμε έτσι ώστε η μεταβολή του z ως προς PM10 να αναπαριστάνεται με μία ευθεία, η οποία εκφράζει την περίπτωση που τα δεδομένα ακολουθούν ιδανικά την κανονική κατανομή (σχήμα 6.5).

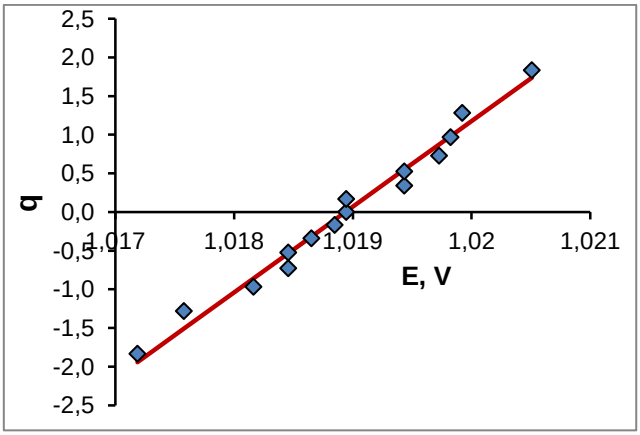
Παρατηρούμε ότι τα σημεία q παρουσιάζουν σημαντικές και χαρακτηριστικές αποκλίσεις από την ευθεία. Αυτό είναι μια σαφής ένδειξη ότι το δείγμα δεν πρέπει να είναι κανονικό. Αντίθετα, στο σχήμα 6.6 που δίνεται το *διάγραμμα Q-Q* των τιμών του δεύτερου δείγματος, τα σημεία βρίσκονται κοντά στην ευθεία ή οι όποιες αποκλίσεις από την ευθεία αυτή είναι τυχαίες, γεγονός που αποτελεί μια καλή ένδειξη ότι τα δεδομένα ακολουθούν την κανονική κατανομή.

	A	B	C	D	E	F
1	**PM10**	**PM10↓**	**z**	**ΑΣ**	**FO**	**q**
2	114,98	28,75	-0,94514	1	0,025	-1,95996
3	141,2	32,13	-0,87806	2	0,075	-1,43953
4	95,75	32,61	-0,86853	3	0,125	-1,15035
5	33,08	33,08	-0,8592	4	0,175	-0,93459
6	40,81	34,71	-0,82685	5	0,225	-0,75542
7	28,75	37,96	-0,76235	6	0,275	-0,59776
8	71,85	40,81	-0,70579	7	0,325	-0,45376
9	106,04	51,79	-0,48787	8	0,375	-0,31864
10	32,61	53,94	-0,44521	9	0,425	-0,18912
11	83,49	61,85	-0,28822	10	0,475	-0,06271
12	37,96	62,57	-0,27393	11	0,525	0,062707
13	34,71	71,85	-0,08976	12	0,575	0,189118
14	62,57	83,49	0,141257	13	0,625	0,318639
15	106,11	95,75	0,384574	14	0,675	0,453762
16	101,07	101,07	0,490157	15	0,725	0,59776
17	61,85	106,04	0,588794	16	0,775	0,755415
18	32,13	106,11	0,590183	17	0,825	0,934589
19	53,94	114,98	0,766221	18	0,875	1,150349
20	236,76	141,2	1,286595	19	0,925	1,439531
21	51,79	236,76	3,183119	20	0,975	1,959964
22						
23	**mean=**	76,3725				
24	**stdev=**	50,38689				

Σχήμα 6.4. Προσδιορισμός ποσοτήτων q και z



Σχήμα 6.5. *Διάγραμμα Q-Q* μεταβολής του q με το PM10 και η ευθεία z ως προς PM10 για έλεγχο της κανονικότητας του δείγματος των 20 πρώτων τιμών του δείγματος του πίνακα 2.5



Σχήμα 6.6. *Διάγραμμα Q-Q* για έλεγχο της κανονικότητας των τιμών του πίνακα 2.4

Πάντως αν και το κριτήριο είναι ποιοτικό και απαιτεί αρκετή εμπειρία για να μη γίνονται λανθασμένες εκτιμήσεις, τα σχήματα 6.5 και 6.6 σε συνδυασμό με τις αντίστοιχες τιμές p-value δεν αφήνουν καμία αμφιβολία ότι το πρώτο δείγμα είναι μη κανονικό και το δεύτερο κανονικό.

❖ **Ανάλυση στο ChemStat και στο SPSS**

Για να κατασκευάσουμε το *διάγραμμα Q-Q* στο *ChemStat* πηγαίνουμε *Πρόσθετα* → *ChemStat* → *Normality plots* → *Univariate Q-Q plot* και εισάγουμε τις τιμές x του δείγματος στο σχετικό πλαίσιο. Το πρόγραμμα υπολογίζει τις τιμές q και z, τις οποίες μαζί με τις τιμές x παρουσιάζει σε τρεις διαδοχικές στήλες. Με βάση αυτές κατασκευάζουμε το *διάγραμμα Q-Q* όπως και στην περίπτωση του *Excel*.

Στο *SPSS* το *διάγραμμα Q-Q* κατασκευάζεται ταυτόχρονα με τον έλεγχο της κανονικότητας. Θα πρέπει όμως να παρατηρήσουμε ότι ενδέχεται να παρουσιαστούν μικρές διαφοροποιήσεις ανάμεσα στα διαγράμματα του *SPSS* και του *ChemStat*/*Excel*. Αυτό συμβαίνει κυρίως όταν υπάρχουν στο δείγμα τιμές που επαναλαμβάνονται. Τότε το *SPSS* χρησιμοποιεί μια διαφορετική διαδικασία υπολογισμού της *Αθροιστικής Συχνότητας*.

6.3 ΕΛΕΓΧΟΙ ΥΠΟΘΕΣΕΩΝ ΓΙΑ ΑΚΡΑΙΕΣ ΤΙΜΕΣ

Μια τιμή ενός δείγματος είναι ***ακραία*** *(outlier)* αν διαφέρει σημαντικά από τις υπόλοιπες τιμές. Μια τέτοια τιμή μπορεί να οφείλεται σε πειραματικό λάθος και συνεπώς θα πρέπει να αφαιρεθεί, μπορεί όμως να υποδηλώνει την ύπαρξη κάποιου σημαντικού φαινομένου ή παράγοντα που οφείλουμε να προσδιορίσουμε. Δυστυχώς, στις περισσότερες περιπτώσεις είναι αδύνατο να προσδιορίσουμε τι ακριβώς συμβαίνει. Αν πάντως οφείλεται σε πειραματικό σφάλμα, αντί να απορρίψουμε την τιμή αυτή, είναι καλό να διορθώσουμε την αιτία του πειραματικού σφάλματος και να επαναλάβουμε όλες τις μετρήσεις από την αρχή.

Παρά τη γενική αυτή σύσταση, στη στατιστική έχουν αναπτυχθεί κριτήρια για τον εντοπισμό μιας ακραίας τιμής. Ένα απλό τεστ για τον έλεγχο μιας ακραίας τιμής είναι το ακόλουθο. Υπολογίζουμε την ποσότητα

$$G = |\text{ακραία τιμή} - \bar{x}|/s \qquad (6.4)$$

όπου $\bar{x}$ και s είναι η μέση τιμή και η τυπική απόκλιση του δείγματος περιλαμβανομένης και της ακραίας τιμής και ελέγχουμε αν η τιμή της G είναι μεγαλύτερη από 2.576 ≈ 2.6. Αν τα δεδομένα ακολουθούν την κανονική κατανομή, η G έχει πιθανότητα μικρότερη από 0.01 να είναι μεγαλύτερη από 2.6 και συνεπώς είναι πολύ πιθανό να αντιστοιχεί σε ακραία τιμή.

Ένας άλλος απλός έλεγχος για την ύπαρξη ακραίων τιμών είναι με τα θηκογράμματα, δεδομένου ότι στο θηκόγραμμα οι ακραίες τιμές σημειώνονται έξω από τους φράκτες.

Ο αυστηρός όμως έλεγχος απαιτεί τη χρήση του κριτηρίου του ***Grubbs***, το οποίο συνιστά η *ISO (International Organization for Standardization)*. Σύμφωνα με το κριτήριο του *Grubbs* υπολογίζεται η τιμή της συνάρτηση G από τη σχέση (6.4) και η πιθανότητα p-value για μονόπλευρο έλεγχο υπολογίζεται από τη λύση της εξίσωσης

$$G = \frac{m-1}{\sqrt{m}} \sqrt{\frac{t^2_{p/m,m-2}}{m-2+t^2_{p/m,m-2}}} \qquad (6.5)$$

όπου $t_{p/m,m-2}$ είναι το σημείο εκείνο για το οποίο ισχύει ότι το εμβαδόν κάτω από την καμπύλη της κατανομής *student* με m-2 βαθμούς ελευθερίας στο διάστημα $[t_{p/m,m-2}, \infty)$ είναι ίσο με (p-value)/m.

Στο *Excel* μπορούμε να λύσουμε την εξίσωση (6.5) ως προς $t_{p/m,m-2}$

χρησιμοποιώντας το πρόγραμμα *Αναζήτηση στόχου (Goal Seek)*, όπως περιγράφεται στο Παράρτημα Ι, ενότητα Ι.7. Τότε η p-value υπολογίζεται από τον τύπο: = m*TDIST($t_{p/m,m-2}$;m-2;1).

Παράδειγμα 6.3

Να εξετασθεί αν υπάρχουν ακραίες τιμές στα δείγματα:

Δ1: 4.5, 4.4, 4.2, 3.4, 4.3, 4.2, 4.4, 4.3

Δ2: 50.3, 50.4, 50.2, 50.0, 50.3

Δ3: 0.28, 0.304, 0.310, 0.301

◆ Στο *ChemStat* εισάγουμε τα δεδομένα σε τρεις στήλες, όπως στο σχήμα 6.7, και πηγαίνουμε *Πρόσθετα* → *ChemStat* → *Test for Outliers.* Το πρόγραμμα δέχεται μόνο ένα δείγμα κάθε φορά και ελέγχει ως ακραία τιμή, την τιμή που απέχει περισσότερο από τη μέση τιμή. Έτσι, στο δείγμα Δ1 ελέγχεται ως ακραία η τιμή 3.4. Η μηδενική υπόθεση και η εναλλακτική της διατυπώνονται ως εξής:

H_0: Η τιμή 3.4 δεν είναι ακραία - H_1: Η τιμή 3.4 είναι ακραία

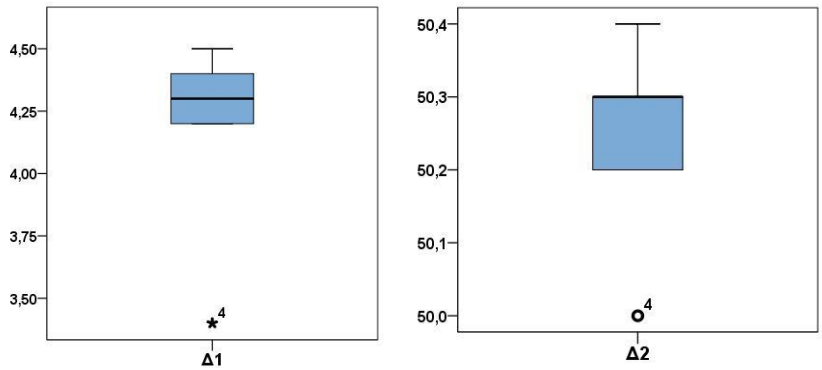
	A	B	C	D	E	F	G
1	Δ1	Δ2	Δ3	Δ1:	Grubbs test for outliers		
2	4,5	50,3	0,280				
3	4,4	50,4	0,304		p-value=	1E-13	
4	4,2	50,2	0,310		3,4 may be an outlier		
5	3,4	50,0	0,301				
6	4,3	50,3		Δ2:	Grubbs test for outliers		
7	4,2						
8	4,4				p-value=	0,115	
9	4,3				50 may be not an outlier		
10							
11				Δ3:	Grubbs test for outliers		
12							
13					p-value=	0,083	
14					0,28 may be not an outlier		

Σχήμα 6.7. Διευθέτηση δεδομένων και αποτελέσματα κριτηρίου Grubbs

Τα αποτελέσματα δίνονται στην περιοχή E3:F4 του σχήματος 6.7, όπου παρατηρούμε ότι η τιμή 3.4 φαίνεται να είναι ακραία. Αντίθετα δεν θα πρέπει να θεωρηθούν ακραίες οι τιμές 50 και 0.28 με βάση τις τιμές p-value = 0.115 και 0.083 και στάθμη σημαντικότητας την α = 0.05.

Αν χρησιμοποιήσουμε το *Excel* για το δείγμα Δ2, τότε έχουμε G = 1.5825 και από την επίλυση της (6.5) παίρνουμε $t_{p/m,m-2}$ = 3.2863. Συνεπώς p-value = 5*TDIST(3.2863;3;1) = 0.1155. Όπως αναμένεται, η τιμή αυτή ταυτίζεται με την αντίστοιχη του *ChemStat* στο σχήμα 6.7.

Είναι ενδιαφέρον ότι αν κάνουμε τα θηκογράμματα των δειγμάτων Δ1 και Δ2 (σχήμα 6.8), παρατηρούμε ότι και η τιμή 50 εμφανίζεται να είναι ακραία τιμή. Όμως σε αυτή την περίπτωση πιο ισχυρός θεωρείται ο έλεγχος *Grubbs* και γι αυτό δεχόμαστε ότι το 50 δεν είναι ακραία τιμή ή πιο αυστηρά τα στοιχεία που έχουμε δεν μας επιτρέπουν να αποφανθούμε ότι το 50 είναι ακραία τιμή. Ένα μεγαλύτερο δείγμα θα βοηθούσε σε ένα πιο αξιόπιστο συμπέρασμα.



Σχήμα 6.8. Θηκογράμματα των δειγμάτων Δ1 και Δ2 του προβλήματος 6.3

6.4 ΕΛΕΓΧΟΣ ΜΕΣΗΣ ΤΙΜΗΣ ΔΕΙΓΜΑΤΟΣ

Όπως ήδη έχουμε αναφέρει, ο έλεγχος αυτός εξετάζει αν η μέση τιμή $\bar{x}$ ενός δείγματος είναι στατιστικά ίδια με τη μέση τιμή μ_0 του πληθυσμού από τον οποίον προέρχεται το δείγμα, που σημαίνει ότι εξετάζει αν το δείγμα προέρχεται από πληθυσμό με μέση τιμή $\mu = \mu_0$. Συνεπώς η μηδενική υπόθεση διατυπώνεται ως

H_0: $\mu = \mu_0$ με εναλλακτική H_1: $\mu \neq \mu_0$ ή H_1: $\mu > \mu_0$ ή H_1: $\mu < \mu_0$

Για τον έλεγχο χρησιμοποιούμε τη μεταβλητή

$$t = \frac{\bar{x} - \mu_0}{s}\sqrt{m} \qquad (6.6)$$

η οποία ακολουθεί την κατανομή *student* με m-1 βαθμούς ελευθερίας με την προϋπόθεση ότι το δείγμα είναι κανονικό. Ο έλεγχος αυτός της μέσης τιμής ενός δείγματος καθώς επίσης και κάθε έλεγχος που βασίζεται στην κατανομή *student* ονομάζεται ***έλεγχος t*** *(t test)*.

Εκτός από τον έλεγχο t, η μηδενική υπόθεση H_0: $\mu = \mu_0$ μπορεί να ελεγχθεί με υπολογιστικές τεχνικές, όπως είναι οι μέθοδοι ***Bootstrap*** και ***Monte-Carlo με αντιμεταθέσεις*** *(permutations)*. Οι μέθοδοι αυτές περιγράφονται στο Κεφάλαιο 7.5 και εφαρμόζονται συμπληρωματικά της παραμετρικής μεθόδου όταν υπάρχουν αμφιβολίες ως προς τις προϋποθέσεις εφαρμογής της παραμετρικής μεθόδου.

Παράδειγμα 6.4

Να επανεξετασθεί με το *ChemStat*, το *Excel* και το *SPSS* το πρόβλημα 4.2, αν δηλαδή το δείγμα με τιμές

40.3, 40.2, 40.2, 40.0, 40.3

έχει μέση τιμή που δεν διαφέρει στατιστικά σημαντικά από το 40.

◆ Στο παράδειγμα 6.1 για το δείγμα αυτό δεν διαπιστώσαμε στατιστικά σημαντικές αποκλίσεις από την κανονικότητα και επομένως μπορούμε να χρησιμοποιήσουμε τον *έλεγχο t* για τη μέση τιμή δείγματος ορίζοντας τη μηδενική υπόθεση ως

H_0: $\mu = 40$ με εναλλακτική H_1: $\mu \neq 40$

❖ **Ανάλυση στο ChemStat**

Το πρόγραμμα *ChemStat* έχει τη δυνατότητα εκτέλεσης του *ελέγχου t της μέσης τιμής ενός δείγματος* και επιπλέον μπορεί να εφαρμόσει για τον ίδιο σκοπό τις μεθόδους *Bootstrap* και *Monte-Carlo* με αντιμεταθέσεις. Για να χρησιμοποιήσουμε το *ChemStat* πηγαίνουμε *Πρόσθετα* → *ChemStat* → *One Sample tests*, στο πρώτο πλαίσιο που ανοίγει εισάγουμε με το ποντίκι την περιοχή τιμών του δείγματος, στο δεύτερο πλαίσιο πληκτρολογούμε την τιμή 40, εφόσον μ = 40, και στο τρίτο ορίζουμε τον αριθμό των επαναλήψεων (Iterations) που θα χρησιμοποιηθούν στις μεθόδους *Bootstrap* και *Monte-Carlo*. Αν δεν θέλουμε να εφαρμοστούν αυτές οι μέθοδοι εισάγουμε τον αριθμό 1. Στο τελευταίο πλαίσιο ορίζουμε το κελί εξόδου των αποτελεσμάτων. Το πρόγραμμα εκτός από τον έλεγχο της

μέσης τιμής του δείγματος εκτελεί και έλεγχο κανονικότητας του δείγματος με το κριτήριο *Anderson-Darling*.

Στο παράδειγμα που εξετάζουμε παίρνουμε τα αποτελέσματα του σχήματος 6.9. Στο επάνω μέρος δίνονται τα αποτελέσματα του κριτήριου *Anderson-Darling*, ακολούθως δίνεται η μέση τιμή και η διασπορά του δείγματος και μετά παρουσιάζονται οι τιμές t και p-value του παραμετρικού ελέγχου. Τέλος, και εφόσον ζητηθεί, παρουσιάζονται οι τιμές p-value ως p(permutation) και p(bootstrap) των μεθόδων *Monte-Carlo* και *Bootstrap* και ο συνολικός χρόνος εκτέλεσης του προγράμματος.

40,3	**One Sample tests**				
40,2					
40,2	**Anderson-Darling Normality test:**				
40	Sample too small. Normality Results may be unreliable				
40,3	p-value=	0,135428	Normality may be assumed		
	Parametric t-test				
	Mean=	40,2	Var=	0,015	
	2 tailed t-test:				
	t-value=	3,651484			
	p-value=	0,021743	Null hypothesis, mean = 40, maybe rejected at level 0.05		
	2 tailed Permutation/Bootstrap tests				
	Iterations=	10000			
	p(permutation)=	0,1237	Null hypothesis, mean = 40 may be assumed at level 0.05		
	p(bootstrap)=	0,060368	Null hypothesis, mean = 40 may be assumed at level 0.05		
	Elapsed time =	0,002 min			

Σχήμα 6.9. Αποτελέσματα ελέγχου μέσης τιμής δείγματος με το πρόγραμμα *One Sample tests* του *ChemStat*

Το πρόγραμμα με βάση τις τιμές p-value μας ενημερώνει σε κάθε έλεγχο αν απορρίπτεται ή δεν απορρίπτεται η μηδενική υπόθεση σε επίπεδο σημαντικότητας 0.05. Παρατηρούμε ότι στο συγκεκριμένο παράδειγμα η παραμετρική μέθοδος δίνει αποτελέσματα διαφορετικά των μεθόδων *Monte-Carlo* με αντιμεταθέσεις και *Bootstrap*. Αυτό εμφανίζεται συχνά σε μικρά δείγματα όταν οι τιμές της p-value είναι κοντά στο όριο των στατιστικών αποφάσεων (α = 0.05).

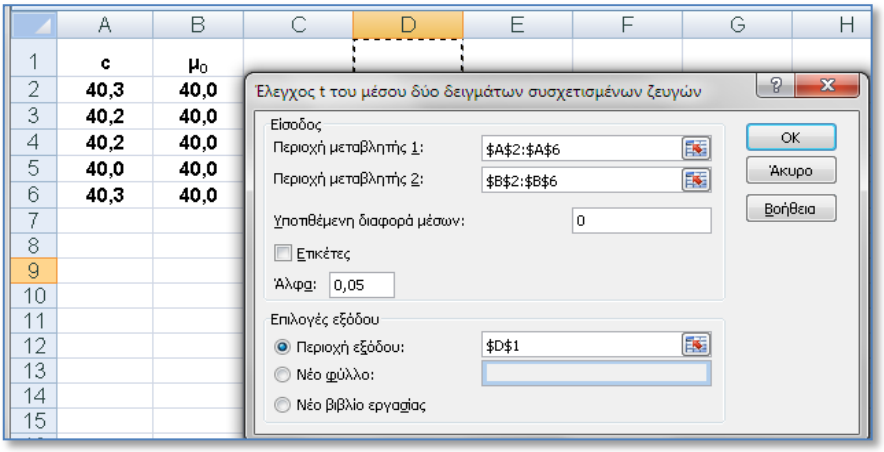
Στο συγκεκριμένο παράδειγμα δεν έχουμε αμφιβολίες ως προς τις προϋποθέσεις εφαρμογής του *παραμετρικού ελέγχου t*, δεδομένου ότι το δείγμα είναι κανονικό επειδή περιέχει τιμές μιας φυσικής ποσότητας που έγιναν κάτω από σταθερές και ελεγχόμενες συνθήκες. Επιπλέον και ο

ισχυρός έλεγχος *Shapiro-Wilk* που εξετάστηκε στο παράδειγμα 6.1 βρίσκεται σε συμφωνία με αυτή την παρατήρηση. Συνεπώς μπορούμε να αποδεχθούμε τα αποτελέσματά του, δεδομένου ότι οι παραμετρικοί έλεγχοι είναι πάντα ισχυρότεροι όλων των άλλων ελέγχων.

Συνεπώς με κίνδυνο να κάνουμε λάθος με πιθανότητα 5% η μηδενική υπόθεση μπορεί να απορριφθεί που σημαίνει ότι με κίνδυνο λάθους 5% η μέση τιμή του δείγματος είναι στατιστικά διαφορετική από το 40. Επιπλέον και επειδή η p-value είναι το *ελάχιστο επίπεδο σημαντικότητας* στο οποίο απορρίπτεται η μηδενική υπόθεση, ουσιαστικά μπορούμε να απορρίψουμε τη μηδενική υπόθεση και στο επίπεδο σημαντικότητας α = 0.022. Συνεπώς μπορούμε να συμπεράνουμε ότι η μέθοδος εισάγει κάποιο συστηματικό σφάλμα ή κάποιο σφάλμα έγινε στην παρασκευή του διαλύματος των 40 ng/mL, αποδεχόμενοι μια μέγιστη πιθανότητα λάθους 2.2%.

❖ Ανάλυση στο Excel

Στο *Excel* μπορούμε να εκτελέσουμε τον παραμετρικό έλεγχο μέσης τιμής δείγματος έμμεσα ως εξής. Πληκτρολογούμε τα δεδομένα του δείγματος σε μία στήλη και σε μία γειτονική στήλη εισάγουμε την τιμή μ_0 δίπλα σε κάθε τιμή του δείγματος, όπως στο σχήμα 6.10. Με τον τρόπο αυτό δημιουργούμε ένα ζεύγος δειγμάτων και ελέγχουμε αν υπάρχει στατιστικά σημαντική διαφορά μεταξύ των δύο αυτών δειγμάτων. Οι βάσεις αυτού του ελέγχου εξετάζονται στο Κεφάλαιο 7.3.



Σχήμα 6.10. Διευθέτηση και είσοδος δεδομένων στο *Excel*

Για να εφαρμόσουμε τώρα τον έλεγχο ζευγών δειγμάτων πηγαίνουμε *Δεδομένα* (*Data*) → *Ανάλυση* (*Analysis*) → *Ανάλυση δεδομένων* (*Data Analysis*) → *Έλεγχος t του μέσου δύο δειγμάτων συσχετιζόμενων ζευγών (t-Test:Paired Two Sample for Means)* και συμπληρώνουμε το παράθυρο διαλόγου που ανοίγει όπως στο σχήμα 6.10. Με *κλικ* στο *OK* παίρνουμε τον πίνακα αποτελεσμάτων του σχήματος 6.11. Ο έλεγχος που κάνουμε είναι δίπλευρος και συνεπώς η τιμή p-value που μας ενδιαφέρει στον πίνακα αυτόν είναι η P(T<=t) δίπλευρη = 0.0217, που ταυτίζεται με την p-value του *ChemStat*.

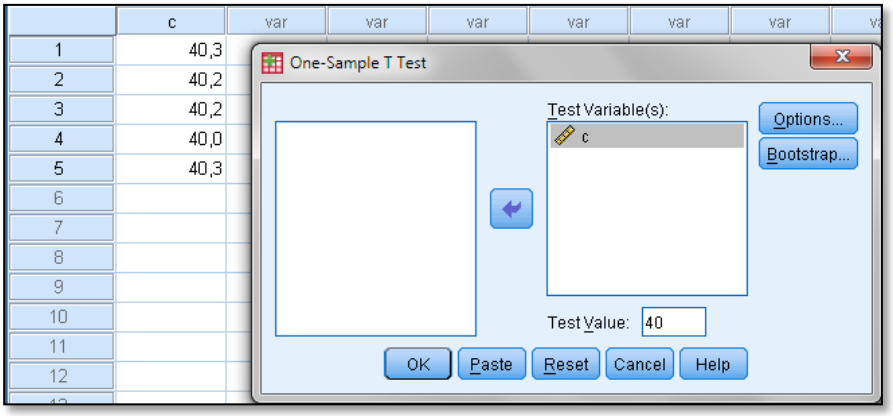
Έλεγχος t του μέσου δύο δειγμάτων συσχετισμένων ζευγών		
	Μεταβλητή 1	Μεταβλητή 2
Μέσος	40,2	40
Διακύμανση	0,015	0
Μέγεθος δείγματος	5	5
Συσχέτιση Pearson	#ΔΙΑΙΡ/0!	
Υποτιθέμενη διαφορά μέσων	0	
βαθμοί ελευθερίας	4	
t	3,6515	
P(T<=t) μονόπλευρη	0,0109	
t κρίσιμο, μονόπλευρο	2,1318	
P(T<=t) δίπλευρη	0,0217	
t κρίσιμο, δίπλευρο	2,7764	

Σχήμα 6.11. Αποτελέσματα ελέγχου μέσης τιμής δείγματος με το *Excel*

❖ Ανάλυση στο SPSS

Στο *SPSS* πηγαίνουμε *Analyze* → *Compare Means* → *One-Sample T Test*. Στο παράθυρο που ανοίγει επιλέγουμε τη μεταβλητή c και τη μεταφέρουμε στο πλαίσιο *Test Variable(s)* και στο *Test Value* πληκτρολογούμε την τιμή 40 (σχήμα 6.12). Μπορούμε να εκτελέσουμε και τη μέθοδο Bootstrap από το αντίστοιχο κουμπί. Όμως, όπως έχουμε αναφέρει, δε συντρέχουν λόγοι εφαρμογής της στο συγκεκριμένο πρόβλημα.

Με *κλικ* στο *OK* παίρνουμε τον πίνακα αποτελεσμάτων του σχήματος 6.13, στον οποίο παρατηρούμε ότι η τιμή του Sig., που είναι η αντίστοιχη του p-value, είναι 0.022. Όπως αναμένεται, το αποτέλεσμα αυτό ταυτίζεται με το αντίστοιχο του *ChemStat* και του *Excel*.



Σχήμα 6.12. Συμπλήρωση του παράθυρου διαλόγου για τον έλεγχο μέσης τιμής δείγματος στο *SPSS*

One-Sample Test

	Test Value = 40					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
c	3,651	4	,022	,2000	,048	,352

Σχήμα 6.13. Βασικός πίνακας αποτελεσμάτων στο *SPSS* για τον έλεγχο μέσης τιμής δείγματος

Παράδειγμα 6.5

Αν η αναλυτική μέθοδος του προηγούμενου παραδείγματος έδινε τα παρακάτω αποτελέσματα

40.4, 40.7, 39.5, 39.6, 41.0

τι συμπέρασμα θα προέκυπτε για την αξιοπιστία της μεθόδου;

◆ Εργαζόμενοι όπως και στο προηγούμενο παράδειγμα παίρνουμε τα αποτελέσματα του σχήματος 6.14, από τα οποία προκύπτει ότι τώρα δεν μπορούμε να απορρίψουμε τη μηδενική υπόθεση H_0: μ = 40, δεδομένου ότι ισχύει p-value = 0.4653 > 0.05.

40,4	**One Sample tests**				
40,7					
39,5	**Anderson-Darling Normality test:**				
39,6	Sample too small. Normality Results may be unreliable				
41	p-value=	0,405843	Normality may be assumed		
	Parametric t-test				
	Mean=	40,24	Var=	0,443	
	2 tailed t-test:				
	t-value=	0,806296			
	p-value=	0,465278	Null hypothesis, mean = 40, may be assumed at level 0.05		
	2 tailed Permutation/Bootstrap tests				
	Iterations=	10000			
	p(permutation)=	0,507	Null hypothesis, mean = 40 may be assumed at level 0.05		
	p(bootstrap)=	0,431033	Null hypothesis, mean = 40 may be assumed at level 0.05		
	Elapsed time =	0,002 min			

Σχήμα 6.14. Πίνακας αποτελεσμάτων του προβλήματος 6.5

Αυτό σημαίνει ότι πιθανόν η μέθοδος να μην εισάγει κάποιο συστηματικό σφάλμα, πρέπει όμως για να βεβαιωθούμε ότι αυτό πράγματι ισχύει να ελέγξουμε και άλλα δείγματα της μεθόδου και κυρίως δείγματα με μεγάλο αριθμό τιμών. Δηλαδή, όπως έχει τονιστεί, το γεγονός ότι δεν μπορούμε να απορρίψουμε τη μηδενική υπόθεση δεν σημαίνει αυτόματα ότι ισχύει, ιδιαίτερα όταν το δείγμα είναι μικρό.

Παράδειγμα 6.6

Να εξετασθεί αν παρατηρείται στατιστικά σημαντική διάσπαση μιας ουσίας όταν σε ένα διάλυμά της με συγκέντρωση 10 mg/L έγιναν μετά από μια μέρα 10 μετρήσεις που έδωσαν το δείγμα:

9.4 9.5 9.7 10.2 10.1 10 8.9 9.4 10.1 8.8

◆ Οι υποθέσεις H_0 και H_1 στο συγκεκριμένο πρόβλημα είναι

$$H_0: \mu = 10 \quad \text{και} \quad H_1: \mu < 10$$

επειδή θέλουμε να ελέγξουμε αν διασπάται η ουσία και συνεπώς αν η μέση τιμή μ είναι μικρότερη από 10.

9,4	**One Sample tests**				
9,5					
9,7	**Anderson-Darling Normality test:**				
10,2					
10,1	p-value=	0,336049	Normality may be assumed		
10	**Parametric t-test**				
8,9	Mean=	9,61	Var=	0,249889	
9,4	2 tailed t-test:				
10,1	t-value=	2,467125			
8,8	p-value=	0,035738	Null hypothesis, mean = 10, maybe rejected at level 0.05		
	2 tailed Permutation/Bootstrap tests				
	Iterations=	10000			
	p(permutation)=	0,0368	Null hypothesis, mean = 10 maybe rejected at level 0.05		
	p(bootstrap)=	0,0447	Null hypothesis, mean = 10 maybe rejected at level 0.05		
	Elapsed time =	0,003 min			

Σχήμα 6.15. Πίνακας αποτελεσμάτων του προβλήματος 6.6 με το πρόγραμμα *One Sample tests* του *ChemStat*

Στο σχήμα 6.15 δίνονται τα αποτελέσματα του προγράμματος *One Sample tests* του *ChemStat*. Παρατηρούμε ότι το δείγμα δεν παρουσιάζει στατιστικά σημαντικές διαφοροποιήσεις από την κανονικότητα και συνεπώς μπορούμε να εφαρμόσουμε τον έλεγχο t της μέσης τιμής. Όμως επειδή το πρόγραμμα εκτελεί πάντα δίπλευρο έλεγχο, πρέπει την τιμή p-value που παίρνουμε να τη διαιρέσουμε δια 2. Συνεπώς έχουμε p-value = 0.036/2 = 0.01787 < 0.05 που δείχνει ότι η πιθανότητα η διαφορά της μέσης τιμής του δείγματος $\bar{x}$ = 9.61 από την τιμή μ = 10 να οφείλεται στην τύχη είναι μικρότερη από 1.8%. Έτσι καταλήγουμε στο συμπέρασμα ότι το διάλυμα της ουσίας δεν είναι σταθερό.

ΑΣΚΗΣΕΙΣ

6.1. Να δημιουργηθούν 5 δείγματα των 15 αριθμών από έναν πληθυσμό που ακολουθεί την κανονική κατανομή με μ = 0.5 και σ = 1. Ακολούθως να εξετασθεί αν τα δείγματα αυτά προέρχονται από τον πληθυσμό N(0,1).

6.2. Να δημιουργηθούν 5 δείγματα των 150 αριθμών από έναν πληθυσμό που ακολουθεί την κανονική κατανομή με μ = 0.5 και σ = 1 και ακολούθως να εξετασθεί αν τα δείγματα αυτά προέρχονται από τον πληθυσμό N(0,1). Τι συμπέρασμα βγάζετε από τις ασκήσεις 6.1 και 6.2;

6.3. Να επανεξετασθούν οι ασκήσεις 4.3 και 4.4 με *έλεγχο t μέσης τιμής*.

6.4. Να ελεγχθεί η κανονικότητα των τιμών του πίνακα 2.5 και ακολούθως να ελεγχθεί η κανονικότητα των 10 πρώτων τιμών. Τι συμπέρασμα βγάζετε;

6.5. Δημιουργήστε δείγματα ανά 10 τιμές και μετά ανά 20 τιμές του πίνακα 2.5 και σε κάθε δείγμα ελέγξτε την κανονικότητα. Τι συμπέρασμα βγάζετε;

6.6. Ένας κατασκευαστής ισχυρίζεται ότι η περιεκτικότητα σε Mg ενός προϊόντος του είναι 0.137 %w/w. Για να ελέγξετε αυτόν τον ισχυρισμό κάνατε 6 μετρήσεις προσδιορισμού του Mg και πήρατε τα αποτελέσματα:

0.129, 0.133, 0.136, 0.130, 0.128 και 0.131 %w/w

Είναι δικαιολογημένος ο ισχυρισμός του κατασκευαστή;

6.7. Να εξετασθεί αν υπάρχουν ακραίες τιμές στα παρακάτω δείγματα:

α) 22.5, 22.1, 22.8, 25.8, 26.6, 23.0, 23.8, 23.5

β) 0.10, 0.11, 0.10, 0.15, 0.11, 0.11, 0.10

γ) 0.215, 0.196, 0.222, 0.224, 0.221, 0.223

δ) 0.10, 0.11, 0.10, 0.15

Κεφάλαιο 7

ΕΛΕΓΧΟΙ ΣΤΑΤΙΣΤΙΚΩΝ ΥΠΟΘΕΣΕΩΝ ΣΕ ΔΥΟ ΔΕΙΓΜΑΤΑ

7.1 ΓΕΝΙΚΑ

Δύο δείγματα μπορεί να είναι ανεξάρτητα ή να σχηματίζουν ζεύγη. Ο όρος ***ανεξάρτητα δείγματα*** *(independent samples)* χρησιμοποιείται με την έννοια ότι το κάθε δείγμα έχει ληφθεί ανεξάρτητα από το άλλο. Αντίθετα, δύο δείγματα σχηματίζουν ένα ***ζεύγος*** *(paired samples)* ή είναι ***εξαρτώμενα*** *(related samples)* αν υπάρχει μία ένα προς ένα αντιστοιχία μεταξύ των τιμών των δύο δειγμάτων. Για παράδειγμα, αν προσδιορίσουμε τη συγκέντρωση ενός αντιδραστηρίου σε διαφορετικά σκευάσματα με δύο διαφορετικές μεθόδους θα πάρουμε δύο εξαρτώμενα δείγματα. Επομένως το πλήθος των τιμών σε δύο ανεξάρτητα δείγματα μπορεί να είναι διαφορετικό, ενώ σε ζεύγη δειγμάτων είναι υποχρεωτικά το ίδιο.

Οι στατιστικοί έλεγχοι που μπορούμε να κάνουμε όταν έχουμε δύο δείγματα είναι να εξετάσουμε αν οι διαφορές στις μέσες τιμές ή/και στις διασπορές των τιμών των δειγμάτων είναι στατιστικά ίσες ή όχι. Επίσης μπορούμε να ελέγξουμε αν τα δείγματα προέρχονται ή όχι από τον ίδιο πληθυσμό.

7.2 ΑΝΕΞΑΡΤΗΤΑ ΔΕΙΓΜΑΤΑ

7.2.1 ΠΑΡΑΜΕΤΡΙΚΟΣ ΕΛΕΓΧΟΣ ΤΩΝ ΜΕΣΩΝ ΤΙΜΩΝ

Ο έλεγχος αυτός μας επιτρέπει να διαπιστώσουμε αν οι διαφορές στις μέσες τιμές δύο δειγμάτων είναι στατιστικά ίσες ή όχι. Στηρίζεται στην κατανομή *student* και γι αυτό το λόγο ο έλεγχος ονομάζεται και εδώ

έλεγχος t *(t-test)*. Υπάρχουν δύο τέτοιοι έλεγχοι ανάλογα με το αν τα δείγματα προέρχονται από πληθυσμούς με ίσες ή διαφορετικές τυπικές αποκλίσεις. Και στις δύο περιπτώσεις η μηδενική υπόθεση και η εναλλακτική της διατυπώνονται ως

$$H_0: \mu_1 = \mu_2$$

και

$$H_1: \mu_1 \neq \mu_2 \quad \text{ή} \quad H_1: \mu_1 > \mu_2 \quad \text{ή} \quad H_1: \mu_1 < \mu_2$$

Δείγματα από πληθυσμούς με ίσες τυπικές αποκλίσεις

Έστω δύο δείγματα με m_1 και m_2 τιμές που προέρχονται από πληθυσμούς με ίσες τυπικές αποκλίσεις ($\sigma_1 = \sigma_2$). Για να ελέγξουμε τη μηδενική υπόθεση $H_0: \mu_1 = \mu_2$, υπολογίζουμε τη στατιστική συνάρτηση ελέγχου

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s\sqrt{1/m_1 + 1/m_2}} \tag{7.1}$$

όπου $\bar{x}_1$ και $\bar{x}_2$ είναι οι μέσες τιμές των δύο δειγμάτων, ενώ η τυπική απόκλιση s υπολογίζεται από τις διασπορές s_1^2 και s_2^2 των δειγμάτων με βάση τη σχέση

$$s^2 = \frac{(m_1 - 1)s_1^2 + (m_2 - 1)s_2^2}{m_1 + m_2 - 2} \tag{7.2}$$

Αποδεικνύεται ότι η μεταβλητή t ακολουθεί την κατανομή *student* με $m_1 + m_2 - 2$ βαθμούς ελευθερίας με την προϋπόθεση ότι τα δείγματα προέρχονται από κανονικούς πληθυσμούς. Η πορεία ελέγχου εξετάζεται με παραδείγματα που δίνονται παρακάτω.

Δείγματα από πληθυσμούς με διαφορετικές τυπικές αποκλίσεις

Αν δύο δείγματα με m_1 και m_2 τιμές προέρχονται από πληθυσμούς με διαφορετικές τυπικές αποκλίσεις ($\sigma_1 \neq \sigma_2$), τότε για να ελέγξουμε τη μηδενική υπόθεση $H_0: \mu_1 = \mu_2$, υπολογίζουμε τη στατιστική συνάρτηση ελέγχου

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_1^2/m_1 + s_2^2/m_2}} \tag{7.3}$$

Αν τα δείγματα προέρχονται από κανονικούς πληθυσμούς, η μεταβλητή αυτή ακολουθεί την κατανομή *student* με βαθμούς ελευθερίας που

υπολογίζονται από την στρογγυλοποίηση της παρακάτω παράστασης

$$\nu \approx \frac{(s_1^2/m_1 + s_2^2/m_2)^2}{\frac{s_1^4}{m_1^2(m_1-1)} + \frac{s_2^4}{m_2^2(m_2-1)}} \qquad (7.4)$$

7.2.2 ΜΗ ΠΑΡΑΜΕΤΡΙΚΟΣ ΕΛΕΓΧΟΣ

Ο έλεγχος αυτός χρησιμοποιείται όταν έχουμε αμφιβολίες αν τουλάχιστον ένα από τα δείγματα που εξετάζουμε προέρχεται από κανονικό πληθυσμό. Τα βασικότερα κριτήρια (έλεγχοι) που χρησιμοποιούνται είναι των *Kolmogorov-Smirnov*, *Wald-Wolfowitz* και *Mann-Whitney*. Από αυτά ισχυρότερο είναι το κριτήριο *Mann-Whitney*. Εναλλακτικά ή και συμπληρωματικά μπορεί να χρησιμοποιηθούν οι μέθοδοι *Bootstrap* και *Monte-Carlo με αντιμεταθέσεις*, που περιγράφονται στην Ενότητα 7.5.

Στον μη παραμετρικό έλεγχο εξετάζεται αν τα δείγματα προέρχονται ή όχι από τον ίδιο πληθυσμό. Συνεπώς η μηδενική υπόθεση εκφράζεται ως

H_0: Τα δείγματα προέρχονται από τον ίδιο πληθυσμό

με εναλλακτική σε δίπλευρο έλεγχο την

H_1: Όχι η H_0

Σε μονόπλευρο έλεγχο η εναλλακτική μπορεί να διατυπωθεί ως εξής:

H_1: Ο πληθυσμός του πρώτου δείγματος έχει κατανομή μετατοπισμένη σε μεγαλύτερες (ή μικρότερες) τιμές σε σχέση με τον πληθυσμό του δεύτερου δείγματος

Η διατύπωση αυτή είναι ισοδύναμη με την

$$H_1: d_1 > d_2 \quad \text{ή} \quad H_1: d_1 < d_2$$

όπου d_1, d_2 είναι οι διάμεσοι των πληθυσμών από τους οποίους προέρχονται τα δείγματα.

Κριτήριο Mann-Whitney

Για την εφαρμογή του κριτηρίου αυτού ενώνουμε τα δύο δείγματα, 1 και 2, σε ένα ενιαίο με $m_1 + m_2$ στοιχεία. Διατάσουμε τα στοιχεία του ενιαίου δείγματος σε αύξουσα σειρά και σημειώνουμε τους αντίστοιχους

βαθμούς για κάθε τιμή. Ακολούθως υπολογίζουμε τα αθροίσματα των βαθμών των δειγμάτων 1 και 2 στο ενιαίο δείγμα. Έστω ότι αυτά είναι R_1 και R_2, αντίστοιχα. Με βάση αυτές τις τιμές υπολογίζουμε τις ποσότητες

$$U_1 = m_1m_2 + m_1(m_1+1)/2 - R_1 \quad \text{και} \quad U_2 = m_1m_2 + m_2(m_2+1)/2 - R_2 \qquad (7.5)$$

και επιλέγουμε την ελάχιστη από αυτές

$$U = \min(U_1, U_2) \qquad (7.6)$$

Για σχετικά μεγάλα δείγματα η U ακολουθεί ασυπτωματικά την κανονική κατανομή με

$$\mu = \frac{m_1m_2}{2} \quad \text{και} \quad \sigma^2 = \frac{m_1m_2(m_1+m_2-1)}{12} \qquad (7.7)$$

Ενδεικτική εφαρμογή του κριτηρίου στο *Excel* δίνεται στο Παράρτημα IV.

Παρατήρηση. Ονομάζουμε ***δεσμό*** *(tie)* όταν μια τιμή εμφανίζεται περισσότερο από μια φορά στο δείγμα. Όταν υπάρχουν δεσμοί επηρεάζουν την τιμή της υπολογιζόμενης πιθανότητας p-value και για το λόγο αυτό έχουν προταθεί τροποποιήσεις στην εφαρμογή των μη παραμετρικών κριτηρίων. Οι τροποποιήσεις αυτές συζητούνται στο Παράρτημα IV.

Παράδειγμα 7.1

Δύο διαφορετικές μέθοδοι προσδιορισμού του αζώτου σε αλεύρι σιταριού έδωσαν τα αποτελέσματα των δειγμάτων 1 και 2 στον παρακάτω πίνακα (σε g ανά 100 g αλεύρου). Να εξετασθεί αν οι μέσες τιμές των δύο δειγμάτων παρουσιάζουν στατιστικά σημαντική απόκλιση σε επίπεδο σημαντικότητας 0.05.

Πίνακας 7.1. Δείγματα παραδείγματος 7.1

Δείγμα 1	1.92	1.68	2.06	2.31	2.29	1.82
	2.23	1.55	1.94	2.27	1.72	
Δείγμα 2	2.25	1.95	2.39	2.36	1.95	2.34
	2.29	2.31	1.85			

◆ Για να λύσουμε το πρόβλημα αυτό, πρώτα ελέγχουμε αν τα δείγματα

ακολουθούν την κανονική κατανομή, αν την ακολουθούν εξετάζουμε αν παρουσιάζουν ή δεν παρουσιάζουν στατιστικά σημαντικές διαφορές στις διασπορές και ανάλογα με το αποτέλεσμα των διασπορών επιλέγεται ο κατάλληλος *έλεγχος t* με βάση τη σχέση (7.1) ή (7.3). Στο *ChemStat* και οι τρεις παραπάνω έλεγχοι γίνονται ταυτόχρονα. Επιπλέον το πρόγραμμα εκτελεί τον μη παραμετρικό έλεγχο *Mann-Whitney*, δηλαδή τον έλεγχο που πρέπει να γίνει όταν τα δείγματα δεν είναι κανονικά, και εφόσον ζητηθεί υπολογίζει την p-value που αντιστοιχεί στον παραμετρικό και στον μη παραμετρικό έλεγχο με τη μέθοδο *Monte-Carlo με αντιμεταθέσεις*. Στο *Excel* υπάρχουν μόνο οι παραμετρικοί έλεγχοι για τις διασπορές και τις μέσες τιμές. Στο *SPSS* υπάρχουν όλοι οι έλεγχοι, όμως ο καθένας από αυτούς γίνεται ξεχωριστά. Επιπλέον, το *SPSS* έχει ως εναλλακτική επιλογή για τον παραμετρικό έλεγχο τη μέθοδο *Bootstrap*, ενώ για τον μη παραμετρικό έλεγχο χρησιμοποιεί ως εναλλακτική επιλογή τη μέθοδο *Monte-Carlo με αντιμεταθέσεις* ή σε μικρά δείγματα τον *Ακριβή (Exact)* υπολογισμό της p-value με τη μέθοδο των αντιμεταθέσεων. Τέλος, θα πρέπει να τονιστεί ότι ο έλεγχος των διασπορών εξετάζεται στην ενότητα 7.4.

❖ Ανάλυση στο ChemStat

Για να χρησιμοποιήσουμε το *ChemStat* μεταφέρουμε τα δείγματα του Πίνακα 7.1 σε δύο στήλες ενός φύλλου του *Excel* και πηγαίνουμε *Πρόσθετα* → *ChemStat* → *Two Samples tests* → *Independent Samples*. Στα παράθυρα που ανοίγουν εισάγουμε διαδοχικά τα δύο δείγματα (τις τιμές τους - χωρίς τους τίτλους) και το κελί εξόδου των αποτελεσμάτων. Το πρόγραμμα αρχικά εκτελεί τον έλεγχο κανονικότητας των δειγμάτων, τον έλεγχο διασπορών, τον παραμετρικό έλεγχο μέσων τιμών και τον έλεγχο *Mann-Whitney*. Ακολούθως εμφανίζει το πλαίσιο *Monte-Carlo permutation tests*, στο οποίο ορίζουμε τον αριθμό των επαναλήψεων (Iterations) που θα χρησιμοποιηθούν στη μέθοδο *Monte-Carlo με αντιμεταθέσεις*. Αν δεν θέλουμε να εφαρμοστεί αυτή η μέθοδος, εισάγουμε τον αριθμό 1. Παίρνουμε τον πίνακα αποτελεσμάτων του σχήματος 7.1.

Στον έλεγχο της κανονικότητας με το κριτήριο *Anderson-Darling* φαίνεται ότι το δεύτερο δείγμα να μην είναι κανονικό. Επειδή όμως η απόκλιση από την κανονικότητα δεν φαίνεται να είναι σημαντική, το p-value = 0.022 είναι κοντά στο α = 0.05, και κυρίως επειδή τα δείγματα προέρχονται από μετρήσεις σε ένα συγκεκριμένο σύστημα κάτω από σταθερές και ελεγχόμενες συνθήκες, μπορούμε να εφαρμόσουμε τον *έλεγχο t* για τις μέσες τιμές.

Two Independent Samples tests - 2 tailed				
Anderson-Darling Normality test:				
p1-value=	0,416406	Sample1-Normality may be assumed		
p2-value=	0,021986	Sample2-Deviations from normality may be assumed		
Mean1=	1,980909		Mean2=	2,187778
Var1=	0,073129		Var2=	0,043769
Test of variances with F test				
p-value=	0,478754	Equality of variances may be assumed		
2 tailed parametric t test:				
	If equality of variances may be assumed			
t-value=	1,877721			
p-value=	0,07672	Null hypothesis, mean1=mean2, may be assumed at level 0.05		
	If equality of variances may be rejected			
t-value=	1,928107			
p-value=	0,069769	Null hypothesis, mean1=mean2, may be assumed at level 0.05		
2 tailed Permutation test				
p(permut.)=	0,0629	Equal variances are assumed		
2 tailed Mann-Witney Non-Parametric test				
U-value=	22	Z=	-2,09164	
p(asymp.)=	0,036471	Null hypothesis may be rejected at level 0.05		
2 tailed Mann-Whitney Permutation test				
p(permut.)=	0,0343	Null hypothesis, d = 0, may be rejected at level 0.05		
Monte-Carlo it	10000			
Elapsed time =	0,034 min			

Σχήμα 7.1. Πίνακας αποτελεσμάτων του προβλήματος 7.1

Για τον έλεγχο των διασπορών των δύο δειγμάτων χρησιμοποιείται ο *έλεγχος F* (Κεφάλαιο 7.4). Από την τιμή p-value = 0.479 > 0.05 προκύπτει ότι σε επίπεδο σημαντικότητας α = 0.05 δεν διαπιστώνονται στατιστικά σημαντικές διαφοροποιήσεις στις διασπορές των δύο δειγμάτων και συνεπώς ο *έλεγχος t* των μέσων τιμών πρέπει να στηριχθεί στη σχέση (7.1). Ο έλεγχος αυτός γίνεται κάτω από τον τίτλο "If equality of variances may be assumed". Παρατηρούμε ότι p-value = 0.0767 και επομένως σε επίπεδο σημαντικότητας 0.05 η μηδενική υπόθεση δεν μπορεί να απορριφθεί. Η τιμή αυτή επιβεβαιώνεται από τη μέθοδο *Monte-Carlo με αντιμεταθέσεις*, που δίνει p-value = p(permut.) = 0.063. Αυτό σημαίνει ότι δεν διαπιστώνονται στατιστικά σημαντικές διαφοροποιήσεις στις μέσες τιμές

των δύο δειγμάτων και συνεπώς οι μέθοδοι χημικής ανάλυσης που χρησιμοποιήθηκαν δεν φαίνεται να δίνουν σημαντικά διαφορετικά αποτελέσματα.

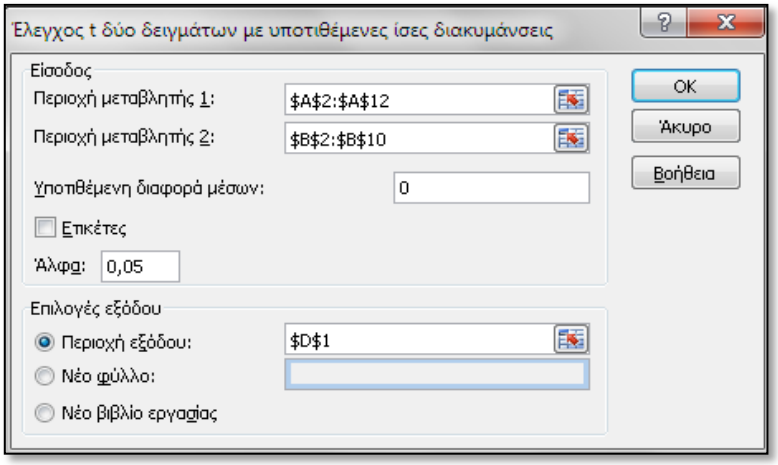
Τέλος, θα πρέπει να προσέξουμε ότι ο μη παραμετρικός έλεγχος με το κριτήριο *Mann-Whitney* και η παραλλαγή του με τη μέθοδο *Monte-Carlo με αντιμεταθέσεις* δείχνουν ότι η μηδενική υπόθεση μπορεί να απορριφθεί (p-value = 0.036 και 0.034, αντίστοιχα). Δηλαδή ο παραμετρικός *έλεγχος t* και ο μη-παραμετρικός έλεγχος με το κριτήριο *Mann-Whitney* δίνουν αντιφατικά αποτελέσματα. Αυτό συμβαίνει όταν η τιμή του p είναι κοντά στο όριο του 0.05 και δείχνει ότι σε τέτοιες περιπτώσεις θα πρέπει να επαναλάβουμε τις μετρήσεις που οδηγούν στα δείγματα αυξάνοντας των αριθμό των τιμών κάθε δείγματος.

Παρατήρηση. Σε μικρά δείγματα ο μη παραμετρικός έλεγχος με το κριτήριο *Mann-Whitney* είναι ασθενέστερος της παραλλαγής του με τη μέθοδο *Monte-Carlo με αντιμεταθέσεις*, επειδή η U ακολουθεί ασυπτωματικά την κανονική κατανομή με μ και σ από τις σχέσεις (7.7). Αυτό σημαίνει ότι όσο μεγαλύτερα είναι τα δείγματα, τόσο καλύτερη είναι η σύγκλιση της U στην κανονική κατανομή.

❖ Ανάλυση στο Excel

Όπως αναφέρθηκε, στο *Excel* υπάρχουν μόνο οι παραμετρικοί έλεγχοι για τις διασπορές και τις μέσες τιμές. Ο στατιστικός έλεγχος διασπορών περιγράφεται στην Ενότητα 7.4 - παράδειγμα 7.5, όπου αποδεικνύεται ότι στο παράδειγμα που εξετάζουμε μπορούμε να υποθέσουμε την ισότητα των διασπορών.

Με βάση αυτή την πληροφορία πηγαίνουμε *Δεδομένα* (*Data)* → *Ανάλυση* (*Analysis)* → *Ανάλυση δεδομένων* (*Data Analysis)* → *Έλεγχος t δύο δειγμάτων με υποτιθέμενες ίσες διακυμάνσεις* (*t-Test:Two-Sample Assuming Equal Variances*) και συμπληρώνουμε το παράθυρο διαλόγου που ανοίγει όπως στο σχήμα 7.2, με την προϋπόθεση ότι οι τιμές των δειγμάτων έχουν εισαχθεί στις περιοχές A2:A12 και B2:B10, αντίστοιχα. Με *κλικ* στο *OK* παίρνουμε τον πίνακα του σχήματος 7.3. Ο έλεγχος που κάνουμε είναι δίπλευρος και συνεπώς η τιμή p που μας ενδιαφέρει στον πίνακα αυτόν είναι η p = 0.0767. Όπως αναμένεται, η τιμή αυτή ταυτίζεται με την αντίστοιχη τιμή του *ChemStat*.



Σχήμα 7.2. Εισαγωγή δεδομένων για έλεγχο t δύο δειγμάτων με υποτιθέμενες ίσες διακυμάνσεις στο *Excel*

	A	B	C	D	E	F	G
1	Δείγμα 1	Δείγμα 2		Έλεγχος t δύο δειγμάτων με υποτιθέμενες ίσες διακυμάνσεις			
2	1,92	2,25					
3	1,68	1,95			*Μεταβλητή 1*	*Μεταβλητή 2*	
4	2,06	2,39		Μέσος	1,9809	2,1878	
5	2,31	2,36		Διακύμανση	0,0731	0,0438	
6	2,29	1,95		Μέγεθος δείγματος	11	9	
7	1,82	2,34		Διάμεση διακύμανση	0,0601		
8	2,23	2,29		Υποτιθέμενη διαφορά μ	0		
9	1,55	2,31		βαθμοί ελευθερίας	18		
10	1,94	1,85		t	-1,8777		
11	2,27			P(T<=t) μονόπλευρη	0,0384		
12	1,72			t κρίσιμο, μονόπλευρο	1,7341		
13				P(T<=t) δίπλευρη	0,0767		
14				t κρίσιμο, δίπλευρο	2,1009		

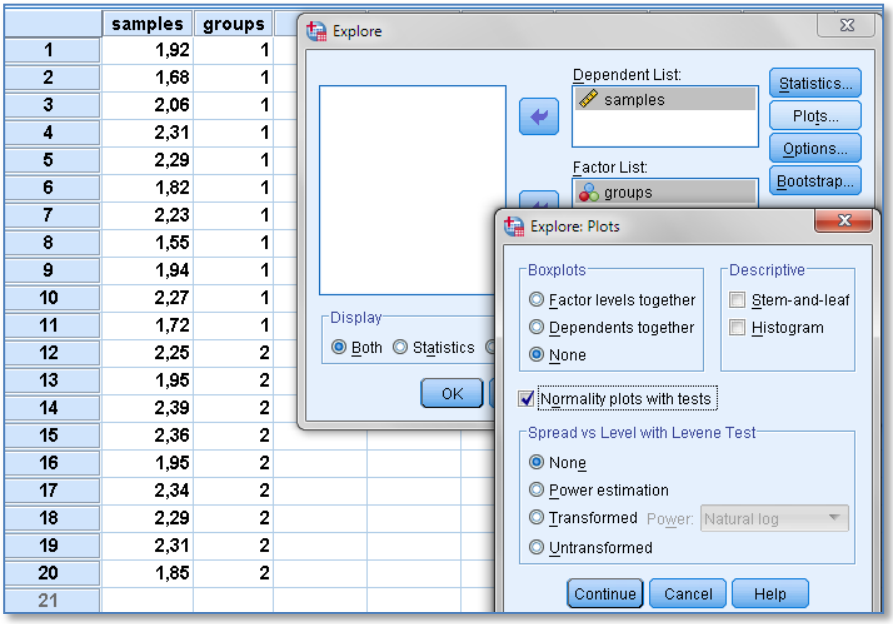
Σχήμα 7.3. Δεδομένα και αποτελέσματα από τον στατιστικό έλεγχο t δύο δειγμάτων με υποτιθέμενες ίσες διακυμάνσεις

❖ Ανάλυση στο SPSS

Τοποθετούμε τα δύο δείγματα σε μία στήλη, την οποία ονομάζουμε έστω samples, και στη διπλανή, έστω την groups, σημειώνουμε 1 δίπλα στις τιμές του πρώτου δείγματος και 2 δίπλα στις τιμές του δεύτερου δείγματος (σχήμα 7.4). Στο *SPSS* αυτή η διευθέτηση χρησιμοποιείται πάντα

όταν έχουμε δύο ή περισσότερα ανεξάρτητα δείγματα.

Ακολούθως ελέγχουμε την κανονικότητα των δειγμάτων από *Analyze → descriptive Statistics → Explore*. Στο παράθυρο που ανοίγει μεταφέρουμε τη στήλη με τα δείγματα στο πλαίσιο *Dependent List* και τη στήλη με τους αριθμούς 1, 2 στο *Factor List.* Κάνουμε *κλικ* στο κουμπί *Plots* και στο νέο πλαίσιο που εμφανίζεται επιλέγουμε *Normality plots with test*, ενώ απενεργοποιούμε το *Stem-and-leaf* και τα *Boxplots* (σχήμα 7.4). Με *κλικ* στο *Continue* και μετά στο *OK*, παίρνουμε τα αποτελέσματα του σχήματος 7.5.



Σχήμα 7.4. Διευθέτηση δειγμάτων στο *SPSS* και εισαγωγή τους για έλεγχο της κανονικότητας

Παρατηρούμε ότι με βάση τα κριτήρια *Kolmogorov-Smirnov* και *Shapiro-Wilk*, όπως και με το κριτήριο *Anderson-Darling*, το δεύτερο δείγμα φαίνεται να παρουσιάζει αποκλίσεις από την κανονικότητα, τουλάχιστον σε επίπεδο σημαντικότητας 0.05.

Για τον παραμετρικό έλεγχο πηγαίνουμε *Analyze → Compare Means → Independent-Samples T Test* και στο παράθυρο που ανοίγει μεταφέρουμε τη στήλη με τα δείγματα στο πλαίσιο *Test Variable(s)* και τη

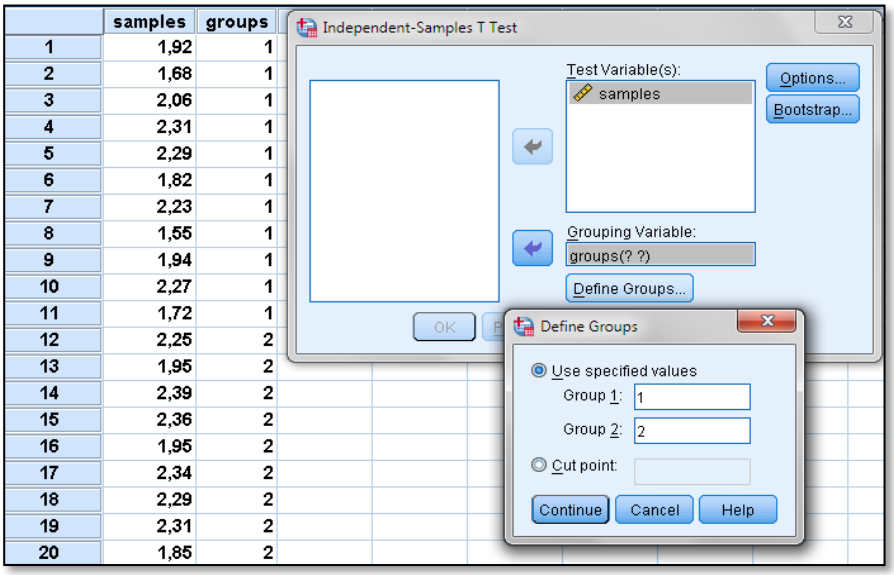
στήλη με τους αριθμούς 1, 2 στο *Grouping Variable.* Στη συνέχεια κάνουμε *κλικ* στο *Define Groups* και στο νέο παράθυρο εισάγουμε την τιμή 1 στο *Group 1*, την τιμή 2 στο *Group 2* και ολοκληρώνουμε με *κλικ* στο *Continue* και μετά στο *OK* (σχήμα 7.6).

Tests of Normality

		Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	groups	Statistic	df	Sig.	Statistic	df	Sig.
samples	1	,185	11	,200*	,921	11	,327
	2	,284	9	,035	,814	9	,029

a. Lilliefors Significance Correction
*. This is a lower bound of the true significance.

Σχήμα 7.5. Αποτελέσματα ελέγχου κανονικότητας στο *SPSS*



Σχήμα 7.6. Διευθέτηση δειγμάτων στο *SPSS* και εισαγωγή τους για έλεγχο μέσων τιμών

Τμήμα του βασικού πίνακα αποτελεσμάτων δίνεται στο σχήμα 7.7. Παρατηρούμε ότι το *SPSS* εκτελεί τον έλεγχο *Levene* για την ισότητα των διασπορών των δειγμάτων και ανεξάρτητα από το αποτέλεσμα αυτού του

ελέγχου παρουσιάζει τα αποτελέσματα του ελέγχου των μέσων τιμών και όταν οι διασπορές των πληθυσμών από τους οποίους προέρχονται τα δείγματα είναι ίσες και όταν αυτές είναι διαφορετικές. Η πρώτη γραμμή αναφέρεται όταν μπορούμε να υποθέσουμε ισότητα των διασπορών, ενώ η δεύτερη όταν οι διασπορές διαφέρουν. Όπως αναμένεται οι τιμές του Sig. (2-tailed) ταυτίζονται με τις αντίστοιχες τιμές του p-value του *ChemStat* στον πίνακα του σχήματος 7.1.

Independent Samples Test

	Levene's Test for Equality of Variances		t-test for Equality of Means		
	F	Sig.	t	df	Sig. (2-tailed)
Equal variances assumed	,939	,345	-1,878	18	,077
Equal variances not assumed			-1,928	17,965	,070

Σχήμα 7.7. Τμήμα αποτελεσμάτων στο *SPSS* για έλεγχο μέσων τιμών

Test Statistics[b]

	samples
Mann-Whitney U	22,000
Wilcoxon W	88,000
Z	-2,092
Asymp. Sig. (2-tailed)	,036
Exact Sig. [2*(1-tailed Sig.)]	,038[a]
Exact Sig. (2-tailed)	,036
Exact Sig. (1-tailed)	,018
Point Probability	,002

a. Not corrected for ties.

b. Grouping Variable: groups

Σχήμα 7.8. Πίνακας αποτελεσμάτων του κριτηρίου *Mann-Whitney*

Για τον μη παραμετρικό έλεγχο πηγαίνουμε *Analyze* → *Non parametric Tests* → *Legacy Dialogs* → *2 Independent Samples* και συμπληρώνουμε το πλαίσιο που ανοίγει όπως και το αντίστοιχο του παραμετρικού ελέγχου, σχήμα 7.6. Εδώ επιπλέον επιλέγουμε το κριτήριο *Mann-Whitney*, αν δεν είναι προεπιλεγμένο. Επίσης, επειδή τα δείγματα

είναι μικρά, κάνουμε *κλικ* στο κουμπί *Exact* και στο πλαίσιο *Exact Tests* επιλέγουμε *Exact* ή *Monte Carlo*. Παίρνουμε τα αποτελέσματα του σχήματος 7.8: U = 22 και Asymp. Sig. (2-tailed) = 0.036, που ταυτίζονται με τα αντίστοιχα του πίνακα 7.1. Ο *Ακριβής (Exact)* έλεγχος δίνει την ίδια τιμή Exact Sig. (2-tailed) = 0.036.

Παράδειγμα 7.2

Ο Rayleigh παρασκεύασε άζωτο χρησιμοποιώντας δύο διαφορετικές τεχνικές. Στην πρώτη πήρε μια ορισμένη ποσότητα αέρα από την οποία απομάκρυνε τους υδρατμούς, το O_2 και το CO_2. Ακολούθως ζύγισε ένα συγκεκριμένο όγκο του αερίου και πήρε τα αποτελέσματα (σε g) του δείγματος 1 στον πίνακα 7.2. Στη δεύτερη τεχνική παρασκεύασε χημικά καθαρό N_2 και πήρε τα αποτελέσματα (σε g) του δείγματος 2 στον ίδιο πίνακα. Να εξετασθεί αν οι μέσες τιμές των δύο δειγμάτων παρουσιάζουν στατιστικά σημαντική απόκλιση.

Πίνακας 7.2. Αποτελέσματα Raleigh.

Δείγμα 1	2.31017	2.30986	2.3101	2.31001
	2.31024	2.3101	2.31028	
Δείγμα 2	2.30143	2.29816	2.30182	2.2989
	2.29869	2.2994	2.29849	2.29889

◆ Πρόκειται για πρόβλημα ίδιο με το προηγούμενο, μόνο που εδώ τα δύο δείγματα προέρχονται από πληθυσμούς με διαφορετικές διασπορές, όπως εύκολα δείχνει ο έλεγχος F των διασπορών που γίνεται στο παράδειγμα 7.5.

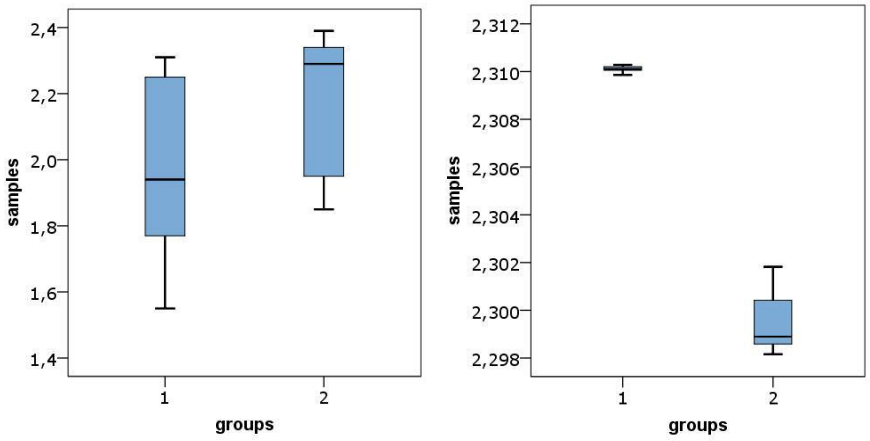
Το *ChemStat* μας δίνει τον πίνακα αποτελεσμάτων του σχήματος 7.9, όπου παρατηρούμε ότι όλοι οι έλεγχοι, παραμετρικοί και μη παραμετρικοί, δίνουν τιμές p-value πολύ μικρές, μικρότερες από 0.001. Συνεπώς η πιθανότητα οι διαφορές στις μέσες τιμές των δύο δειγμάτων να οφείλονται σε τυχαίους παράγοντες είναι αμελητέα. Ακριβώς γι αυτό το λόγο ο Rayleigh υποπτεύθηκε ότι κάποιο συστηματικό σφάλμα ήταν υπεύθυνο για τις στατιστικά πολύ σημαντικές διαφορές των δύο δειγμάτων. Έτσι υπέθεσε ότι ο αέρας περιέχει εκτός από υδρατμούς N_2, O_2 και CO_2 και κάποιο άλλο (αδρανές) αέριο. Η υπόθεση αυτή τον οδήγησε τελικά στην ανακάλυψη του Αργού.

Two Independent Samples tests - 2 tailed				
Anderson-Darling Normality test:				
p1-value=	0,740875	Sample1-Normality may be assumed		
p2-value=	0,025156	Sample2-Deviations from normality may be assumed		
Mean1=	2,310109		Mean2=	2,299473
Var1=	2,03E-08		Var2=	1,9E-06
Test of variances with F test				
p-value=	2,13E-05	Equality of variances may be rejected		
2 tailed parametric t test:				
	If equality of variances may be assumed			
t-value=	20,21372			
p-value=	3,32E-11	Null hypothesis, mean1=mean2, may be rejected at level 0.05		
	If equality of variances may be rejected			
t-value=	21,68022			
p-value=	1,12E-07	Null hypothesis, mean1=mean2, may be rejected at level 0.05		
2 tailed Permutation test				
p(permut.)=	0,0001	Equal variances are not assumed		
2 tailed Mann-Witney Non-Parametric test				
U-value=	0	Z=	-3,24327	
p(asymp.)=	0,001182	Null hypothesis may be rejected at level 0.05		
2 tailed Mann-Whitney Permutation test				
p(permut.)=	0,0002	Null hypothesis, d = 0, may be rejected at level 0.05		
Monte-Carlo it	10000			
Elapsed time =	0,026 min			

Σχήμα 7.9. Πίνακας αποτελεσμάτων για τον έλεγχο μέσων τιμών

Τέλος, μια καλή εικόνα για τη σχέση δύο ή περισσότερων δειγμάτων παρέχουν τα θηκογράμματα. Στο σχήμα 7.10 δίνονται τα θηκογράμματα των δειγμάτων του παραδείγματος 7.1 (αριστερά) και 7.2 (δεξιά). Για να τα κατασκευάσουμε στο *SPSS* διευθετούμε τα δείγματα κάθε παραδείγματος όπως στο παράδειγμα 7.1, σχήμα 7.4, και πηγαίνουμε *Graphs* → *Legacy Dialogs* → *Boxplot*. Στο πλαίσιο που ανοίγει επιλέγουμε *Simple* και *Summaries for group of cases*. Πατάμε *Define* και στο νέο παράθυρο που ανοίγει εισάγουμε τη μεταβλητή samples στο πλαίσιο *Variable* και την groups στο *Category Axis*.

Παρατηρούμε ότι η διαφορά των δειγμάτων του παραδείγματος 7.2 δεν αφήνει αμφιβολίες ότι προέρχονται από διαφορετικούς πληθυσμούς, ενώ αντίθετα τα δείγματα του παραδείγματος 7.1 είναι πιθανόν να προέρχονται από τον ίδιο πληθυσμό.



Σχήμα 7.10. Θηκογράμματα των δειγμάτων του παραδείγματος 7.1 (αριστερά) και 7.2 (δεξιά)

Παράδειγμα 7.3

Στον πίνακα 7.3 δίνονται οι αέριοι ρύποι σε TSP στους σταθμούς Αγία Σοφία και Κορδελιό τους δύο πρώτους μήνες του 2007. Να ελεγχθεί που υπάρχει στατιστικά μικρότερη ρύπανση.

◆ Επειδή τα δείγματα είναι περιβαλλοντικά, ο έλεγχος της κανονικότητας είναι απαραίτητος. Όμως επειδή πρόκειται για σχετικά μικρά δείγματα, ανεξάρτητα τι θα δείξει ο έλεγχος αυτός, θα πρέπει η μελέτη να στηριχθεί κυρίως σε μη παραμετρικούς ελέγχους.

❖ **Ανάλυση στο ChemStat**

Αν εφαρμόσουμε το *ChemStat* όπως στα προηγούμενα παραδείγματα, παίρνουμε τον πίνακα αποτελεσμάτων του σχήματος 7.11. Στον έλεγχο της κανονικότητας με το κριτήριο *Anderson-Darling* και σε επίπεδο σημαντικότητας α = 0.05 δεν διαπιστώνονται στατιστικά σημαντικές αποκλίσεις από την κανονικότητα. Όμως, όπως ήδη αναφέρθηκε, το

σχετικά μικρό τους μέγεθος και το γεγονός ότι πρόκειται για περιβαλλοντικά δείγματα πρέπει να μας κάνει επιφυλακτικούς. Έτσι αν και ο παραμετρικός έλεγχος για τις μέσες τιμές δείχνει ότι η μηδενική υπόθεση δεν μπορεί να απορριφθεί (p-value = 0.085), τον προσπερνάμε και εξετάζουμε τα αποτελέσματα των μη παραμετρικών ελέγχων.

Πίνακας 7.3. Αέριοι ρύποι σε TSP στους σταθμούς Αγία Σοφία και Κορδελιό.

Ημερομηνία	Αγία Σοφία	Ημερομηνία	Κορδελιό
10/1/2007	61.67	2/1/2007	83.49
22/1/2007	48.75	4/1/2007	37.96
26/1/2007	40.83	6/1/2007	34.71
28/1/2007	29.17	8/1/2007	62.57
30/1/2007	20.42	10/1/2007	106.11
1/2/2007	50.42	12/1/2007	101.07
5/2/2007	28.75	14/1/2007	61.85
7/2/2007	97.08	5/2/2007	32.13
9/2/2007	108.75	7/2/2007	53.94
11/2/2007	55.00	9/2/2007	136.76
13/2/2007	58.33	11/2/2007	51.79
15/2/2007	37.92	13/2/2007	62.94
17/2/2007	16.67	17/2/2007	47.14
19/2/2007	29.17	19/2/2007	41.37
23/2/2007	36.67	21/2/2007	65.56
		23/2/2007	70.44

Στον πίνακα του σχήματος 7.11 παρατηρούμε ότι ο μη παραμετρικός έλεγχος με το κριτήριο *Mann-Whitney* και η παραλλαγή του με τη μέθοδο *Monte-Carlo με αντιμεταθέσεις* δείχνουν ότι η μηδενική υπόθεση πρέπει να απορριφθεί (p-value = 0.032 και 0.031, αντίστοιχα). Αυτό σημαίνει ότι πρέπει να υπάρχει στατιστικά σημαντική διαφορά στην αέρια ρύπανση των δύο περιοχών. Για να δούμε σε ποια περιοχή η ρύπανση είναι μικρότερη, μπορούμε να χρησιμοποιήσουμε θηκογράμματα ή πιο απλά να υπολογίσουμε τις διαμέσους των δειγμάτων των δύο περιοχών. Στο συγκεκριμένο πρόβλημα είναι καλύτερα να χρησιμοποιήσουμε διαμέσους αντί για μέσες τιμές, για να αποφύγουμε την επίδραση τυχών ακραίων τιμών στις μέσες τιμές. Χρησιμοποιώντας τη συνάρτηση *MEDIAN*, παίρνουμε:

Αγία Σοφία: d = 40.83

Κορδελιό: d = 62.21

Συνεπώς μπορούμε να συμπεράνουμε ότι στην Αγία Σοφία η αέρια ρύπανση φαίνεται να είναι μικρότερη από την ρύπανση του Κορδελιού, τουλάχιστον στους δύο πρώτους μήνες του 2007.

Two Independent Samples tests - 2 tailed				
Anderson-Darling Normality test:				
p1-value=	0,062292	Sample1-Normality may be assumed		
p2-value=	0,110994	Sample2-Normality may be assumed		
Mean1=	47,97222		Mean2=	65,61616
Var1=	682,338		Var2=	828,6293
Test of variances with F test				
p-value=	0,721449	Equality of variances may be assumed		
2 tailed parametric t test:				
	If equality of variances may be assumed			
t-value=	1,783135			
p-value=	0,085036	Null hypothesis, mean1=mean2, may be assumed at level 0.05		
	If equality of variances may be rejected			
t-value=	1,788904			
p-value=	0,084082	Null hypothesis, mean1=mean2, may be assumed at level 0.05		
2 tailed Permutation test				
p(permut.)=	0,0793	Equal variances are assumed		
2 tailed Mann-Witney Non-Parametric test				
U-value=	66	Z=	-2,13475	
p(asymp.)=	0,032781	Null hypothesis may be rejected at level 0.05		
2 tailed Mann-Whitney Permutation test				
p(permut.)=	0,0312	Null hypothesis, d = 0, may be rejected at level 0.05		
Monte-Carlo i	10000			
Elapsed time	0,053 min			

Σχήμα 7.11. Πίνακας αποτελεσμάτων για τον έλεγχο μέσων τιμών

❖ Ανάλυση στο SPSS

Τοποθετούμε τα δύο δείγματα σε μία στήλη, την οποία ονομάζουμε έστω samples, και στη διπλανή, έστω την groups, σημειώνουμε 1 δίπλα στις τιμές του πρώτου δείγματος και 2 δίπλα στις τιμές του δεύτερου δείγματος. Επιπλέον πηγαίνουμε στο παράθυρο *Variable View* και από το

κελί της στήλης *Values* που αντιστοιχεί στη γραμμή groups ανοίγουμε το παράθυρο διαλόγου *Value Labels*. Στο πλαίσιο *Value* πληκτρολογούμε 1, στο *Label* πληκτρολογούμε AgiaSofia και κάνουμε *κλικ* στο *Add*. Συνεχίζουμε και στο πλαίσιο *Value* πληκτρολογούμε 2, στο *Label* πληκτρολογούμε Kordelio και κάνουμε *κλικ* στο *Add* και *κλικ* στο *OK*.

Ακολούθως ελέγχουμε την κανονικότητα των δειγμάτων από *Analyze → descriptive Statistics → Explore*. Στον πίνακα αποτελεσμάτων (σχήμα 7.12) παρατηρούμε ότι με το ισχυρό κριτήριο των *Shapiro-Wilk* στο πρώτο δείγμα διαπιστώνονται στατιστικά σημαντικές αποκλίσεις από την κανονικότητα. Το πιθανότερο πάντως είναι και τα δύο δείγματα να μην είναι κανονικά.

Tests of Normality

		Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	group	Statistic	df	Sig.	Statistic	df	Sig.
TSP	AgiaSofia	,167	15	,200*	,879	15	,046
	Kordelio	,188	16	,133	,900	16	,080

a. Lilliefors Significance Correction
*. This is a lower bound of the true significance.

Σχήμα 7.12. Αποτελέσματα από τον έλεγχο κανονικότητας των δύο δειγμάτων στο *SPSS*

Για τον μη παραμετρικό έλεγχο πηγαίνουμε *Analyze → Non parametric Tests → Legacy Dialogs → 2 Independent Samples* και συμπληρώνουμε τα πλαίσια που ανοίγουν όπως στο παράδειγμα 7.1. Παίρνουμε τα αποτελέσματα του σχήματος 7.13, όπου παρατηρούμε ότι σε πλήρη συμφωνία με το *ChemStat*, ο ασυμπτωτικός και ο ακριβής έλεγχος δίνουν πρακτικά ταυτόσημα αποτελέσματα: Asymp. Sig. (2-tailed) = 0.033 και Exact Sig. (2-tailed) = 0.032, που δείχνουν ότι τα δείγματα είναι πιθανόν να προέρχονται από διαφορετικούς πληθυσμούς.

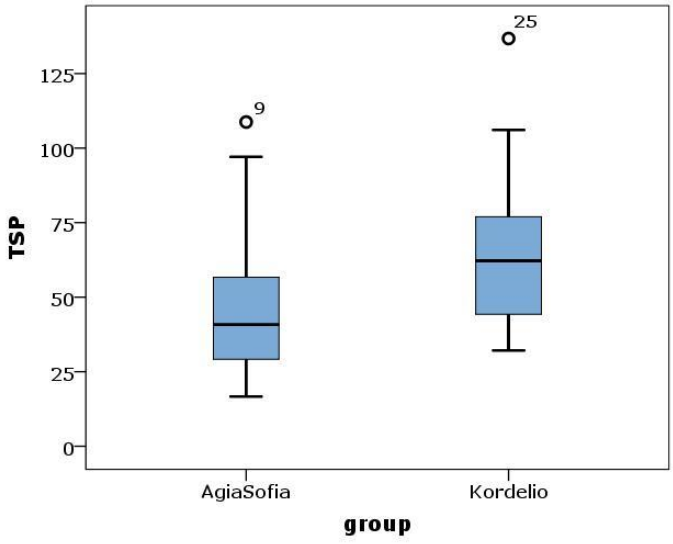
Για να ελέγξουμε σε πιο σταθμό είναι μικρότερη η ρύπανση κάνουμε τα θηκογράμματα, από τα οποία προκύπτει αβίαστα ότι η μικρότερη ρύπανση είναι στην Αγία Σοφία (σχήμα 7.14).

Test Statistics[b]

	TSP
Mann-Whitney U	66,000
Wilcoxon W	186,000
Z	-2,135
Asymp. Sig. (2-tailed)	,033
Exact Sig. [2*(1-tailed Sig.)]	,033[a]
Exact Sig. (2-tailed)	,032
Exact Sig. (1-tailed)	,016
Point Probability	,001

a. Not corrected for ties.
b. Grouping Variable: group

Σχήμα 7.13. Πίνακας αποτελεσμάτων του κριτηρίου *Mann-Whitney*



Σχήμα 7.14. Θηκογράμματα ρύπων

7.3 ΣΥΓΚΡΙΣΕΙΣ ΖΕΥΓΩΝ ΔΕΙΓΜΑΤΩΝ

Όταν δύο δείγματα σχηματίζουν ένα ***ζεύγος*** το κύριο ερώτημα που εγείρεται είναι αν υπάρχει στατιστικά σημαντική διαφορά μεταξύ των τιμών των ζευγών των δύο δειγμάτων. Αν δεν υπάρχει στατιστικά σημαντική διαφορά, θα πρέπει ο μέσος όρος των διαφορών μεταξύ των αντίστοιχων τιμών των δύο δειγμάτων, $\bar{x}_D$, να είναι μηδέν. Εναλλακτικά, θα πρέπει η διάμεσος d_D των διαφορών $x_i - y_i$ των τιμών του ενός δείγματος $\{x_1, x_2, \ldots, x_m\}$ από τις αντίστοιχες τιμές του άλλου δείγματος $\{y_1, y_2, \ldots, y_m\}$ να είναι ίση με μηδέν. Συνεπώς οι υποθέσεις H_0 και H_1 που ελέγχουμε μπορούν να διατυπωθούν ως εξής:

Παραμετρικός έλεγχος:

$$H_0: \mu_d = 0 \quad \text{και} \quad H_1: \mu_d \neq 0 \ \text{ή} \ H_1: \mu_d > 0 \ \text{ή} \ H_1: \mu_d < 0 \qquad (7.8)$$

Μη παραμετρικός έλεγχος:

$$H_0: d = 0 \quad \text{και} \quad H_1: d \neq 0 \ \text{ή} \ H_1: d > 0 \ \text{ή} \ H_1: d < 0 \qquad (7.9)$$

όπου μ_d και d είναι η μέση τιμή και η διάμεσος του πληθυσμού από τον οποίον προέρχονται οι διαφορές $x_i - y_i$.

Για να ελέγξουμε την υπόθεση (7.8) υπολογίζουμε τη μεταβλητή

$$t = \frac{\bar{x}_D \sqrt{m}}{s_d} \qquad (7.10)$$

όπου s_d είναι η τυπική απόκλιση του δείγματος που προκύπτει από τη διαφορά των ζευγών τιμών των αρχικών δειγμάτων. Αποδεικνύεται ότι η παραπάνω μεταβλητή ακολουθεί την κατανομή *student* με m-1 βαθμούς ελευθερίας με την προϋπόθεση ότι το δείγμα των διαφορών των τιμών ακολουθεί την κανονική κατανομή.

Για τον έλεγχο της μη παραμετρικής μηδενικής υπόθεσης (7.9) μπορεί να χρησιμοποιηθεί το κριτήριο του *Wilcoxon*.

Κριτήριο Wilcoxon

Αν $\{x_1, x_2, \ldots, x_m\}$ και $\{y_1, y_2, \ldots, y_m\}$ είναι τα δύο δείγματα, πρώτα υπολογίζουμε τις διαφορές $d_i = x_i - y_i$ που τις διατάσσουμε σε αύξουσα σειρά των απόλυτων τιμών $|d_i|$ και σε κάθε τιμή $|d_i|$ αντιστοιχούμε το βαθμό της. Ακολούθως σε κάθε βαθμό αντιστοιχούμε το πρόσημο του d_i και υπολογίζουμε το άθροισμα T^+ των βαθμών των θετικών d_i και το αντίστοιχο

άθροισμα T^- των βαθμών των αρνητικών d_i. Τέλος υπολογίζουμε την τιμή

$$W = \min(T^+, T^-) \qquad (7.11)$$

Στους υπολογισμούς δεν λαμβάνονται υπόψη μηδενικές διαφορές, $d_i = 0$.

Σε μεγάλα δείγματα η τυχαία μεταβλητή W ακολουθεί ασυμπτωτικά την κανονική κατανομή με

$$\mu = \frac{m(m+1)}{4} \quad \text{και} \quad \sigma^2 = \frac{m(m+1)(2m+1)}{24} \qquad (7.12)$$

Frank Wilcoxon
(1892-1965)
Χημικός - Στατιστικολόγος

Παράδειγμα 7.4

Στον πίνακα 7.4 δίνονται οι συγκεντρώσεις σε αυθαίρετες μονάδες του ενεργού συστατικού σε 8 δισκία που προσδιορίστηκαν με δύο διαφορετικές αναλυτικές τεχνικές. Να ελεγχθεί αν οι τεχνικές δίνουν στατιστικά το ίδιο αποτέλεσμα.

Πίνακας 7.4. Δεδομένα ως ζεύγη τιμών.

Δείγμα 1	4.6	3.4	3.1	2.8	1.5	4.1	2.5	3.3
Δείγμα 2	4.3	3.5	3.3	3	1.1	4.5	2.8	3.1

❖ **Ανάλυση στο ChemStat**

Τα δείγματα τοποθετούνται σε δύο στήλες και πηγαίνουμε *Πρόσθετα → ChemStat → Two Samples tests → Paired Samples.* Στα παράθυρα που ανοίγουν εισάγουμε διαδοχικά τις τιμές των δύο δειγμάτων και το κελί

εξόδου των αποτελεσμάτων. Όπως και στον έλεγχο ανεξάρτητων δειγμάτων, το πρόγραμμα όταν ολοκληρώσει τον παραμετρικό έλεγχο και τον έλεγχο *Wilcoxon* εμφανίζει το πλαίσιο *Monte-Carlo permutation tests* για να ορίσουμε τον αριθμό των επαναλήψεων (Iterations) που θα χρησιμοποιηθούν στη μέθοδο *Monte-Carlo*. Παίρνουμε τον πίνακα αποτελεσμάτων του σχήματος 7.15.

Two Dependent Samples (paired-samples) tests - 2 tailed		
Anderson-Darling Normality test:		
p1-value=	0,894924	Sample1-Normality may be assumed
p2-value=	0,292449	Sample2-Normality may be assumed
pd-value=	0,299706	SampleOfDifferences-Normality may be assumed
2 tailed Parametric t-test:		
t-value	0,356753	
p-value=	0,731788	Null hypothesis, mean1=mean2, may be assumed at level 0.05
2 tailed Permutation test		
p(permut.)=	0,6432	Null hypothesis, mean1=mean2, may be assumed at level 0.05
2 tailed Wilcoxon Non-Parametric test		
W-value=	16	Z= -0,28214
p(asump.)=	0,777838	Null hypothesis may be assumed at level 0.05
2 tailed Wilcoxon Permutation test		
p(permut.)=	0,8017	Null hypothesis, d = 0, may be assumed at level 0.05
Monte-Carlo iter	10000	
Elapsed time =	0,012 min	

Σχήμα 7.15. Πίνακας αποτελεσμάτων για τον έλεγχο μέσων τιμών ζεύγους δειγμάτων

Παρατηρούμε ότι ο έλεγχος της κανονικότητας δίνει τις τιμές p-value = 0.895 για το πρώτο και 0.292 για το δεύτερο, που δείχνουν ότι τα δείγματα είναι πιθανόν να προέρχονται από κανονικούς πληθυσμούς, εκείνο όμως που ενδιαφέρει είναι η κανονικότητα του δείγματος των διαφορών των τιμών των δύο δειγμάτων. Για το δείγμα αυτό ισχύει p-value = pd-value = 0.2997 > 0.05 και συνεπώς σε επίπεδο σημαντικότητας α = 0.05 δεν διαπιστώνονται στατιστικά σημαντικές αποκλίσεις από την κανονικότητα, αν και το πολύ μικρό μέγεθος του δείγματος θα πρέπει να μας οδηγεί στην εξέταση και των αποτελεσμάτων των μη παραμετρικών μεθόδων.

Ο *έλεγχος t* για τη μέση τιμή των διαφορών των τιμών των δύο δειγμάτων δίνει p = 0.7318 που σημαίνει ότι η μηδενική υπόθεση δεν μπορεί να απορριφθεί. Το ίδιο συμπέρασμα προκύπτει και από τις τιμές p-value όλων των άλλων ελέγχων. Συνεπώς σε επίπεδο σημαντικότητας α = 0.05 δεν διαπιστώνονται στατιστικά σημαντικές διαφοροποιήσεις στα αποτελέσματα των δύο μεθόδων. Θα πρέπει όμως και πάλι να τονιστεί ότι το γεγονός ότι δεν διαπιστώθηκαν στον συγκεκριμένο έλεγχο στατιστικά σημαντικές διαφοροποιήσεις στα αποτελέσματα των δύο μεθόδων δεν σημαίνει ότι δεν υπάρχουν. Για το λόγο αυτό ο έλεγχος θα πρέπει να επαναληφθεί με μεγαλύτερα δείγματα.

❖ Ανάλυση στο Excel

Στο *Excel* πηγαίνουμε *Δεδομένα* (*Data*) → *Ανάλυση* (*Analysis*) → *Ανάλυση δεδομένων* (*Data Analysis*) → *Έλεγχος t του μέσου δύο δειγμάτων συσχετιζόμενων ζευγών (t-Test:Paired Two Sample for Means).* Τα αποτελέσματα που παίρνουμε δίνονται στο σχήμα 7.16 και ταυτίζονται με τα αντίστοιχα του *ChemStat*.

Δείγμα 1	Δείγμα 2
4,6	4,3
3,4	3,5
3,1	3,3
2,8	3
1,5	1,1
4,1	4,5
2,5	2,8
3,3	3,1

Έλεγχος t του μέσου δύο δειγμάτων συσχετισμένων ζευγών

	Μεταβλητή 1	*Μεταβλητή 2*
Μέσος	3,1625	3,2
Διακύμανση	0,9084	1,0886
Μέγεθος δείγματος	8	8
Συσχέτιση Pearson	0,9597	
Υποτιθέμενη διαφορά μ	0	
βαθμοί ελευθερίας	7	
t	-0,3568	
P(T<=t) μονόπλευρη	0,3659	
t κρίσιμο, μονόπλευρο	1,8946	
P(T<=t) δίπλευρη	0,7318	
t κρίσιμο, δίπλευρο	2,3646	

Σχήμα 7.16. Πίνακας αποτελεσμάτων για τον έλεγχο της μέσης τιμής ζεύγους δειγμάτων στο *Excel*

❖ Ανάλυση στο SPSS

Στην περίπτωση σύγκρισης ζεύγους δειγμάτων στο *SPSS*, τα δείγματα τοποθετούνται σε δύο στήλες και όχι σε μία όπως στον έλεγχο ανεξάρτητων δειγμάτων. Για τον έλεγχο της κανονικότητας δημιουργούμε

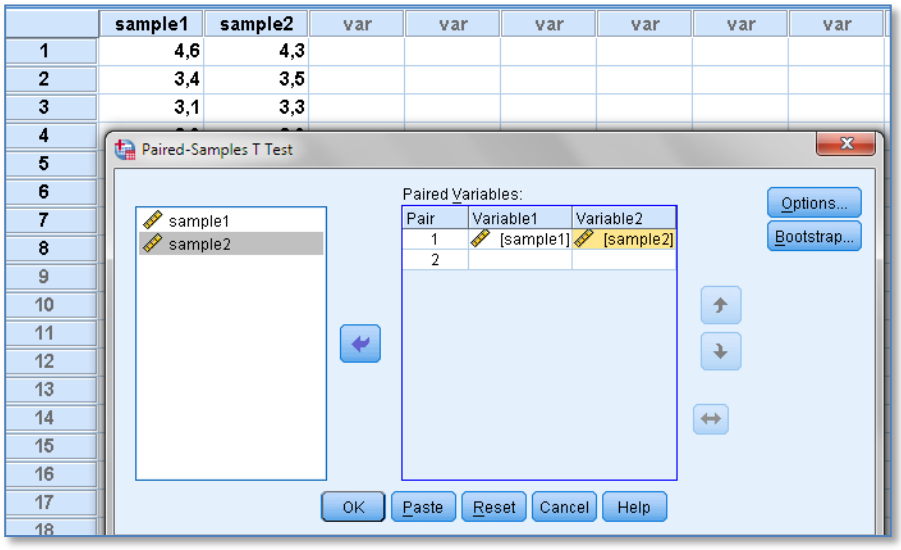
ακόμη μία στήλη με τις διαφορές των τιμών των δύο δειγμάτων. Ο έλεγχος της κανονικότητας αυτής της μεταβλητής οδηγεί σε τιμές Sig. > 0.05 (σχήμα 7.17) και συνεπώς μπορούμε να εφαρμόσουμε τον παραμετρικό έλεγχο. Όμως για λόγους σύγκρισης θα εφαρμόσουμε και τον μη παραμετρικό έλεγχο, που άλλωστε ως συμπληρωματικό έλεγχο μπορούμε να τον εφαρμόζουμε πάντα. Για τον *έλεγχο t* της μέσης τιμής ακολουθούμε τη διαδικασία *Analyze* → *Compare Means* → *Paired-Samples T Test* και συμπληρώνουμε το πλαίσιο που ανοίγει όπως στο σχήμα 7.18. Τα αποτελέσματα που παίρνουμε δίνονται στο σχήμα 7.19 και ταυτίζονται με τα αντίστοιχα του *ChemStat*.

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
SampleOfDifferences	,208	8	,200*	,908	8	,340

a. Lilliefors Significance Correction
*. This is a lower bound of the true significance.

Σχήμα 7.17. Αποτελέσματα από τον έλεγχο κανονικότητας του δείγματος των διαφορών των τιμών του ζεύγους δειγμάτων



Σχήμα 7.18. Διευθέτηση δειγμάτων και είσοδος δεδομένων στο *SPSS*

Paired Samples Test

Paired Differences				t	df	Sig. (2-tailed)
Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
		Lower	Upper			
,29731	,10511	-,28606	,21106	-,357	7	,732

Σχήμα 7.19. Τμήμα του πίνακα αποτελεσμάτων ελέγχου μέσης τιμής σε ζεύγος δειγμάτων στο *SPSS*

Για τον μη παραμετρικό έλεγχο με το κριτήριο Wilcoxon πηγαίνουμε *Analyze* → *Nonparametric Tests* → *Legacy Dialogs* → *2 Related Samples*. Στο παράθυρο που ανοίγει κάνουμε *κλικ* και στις δύο μεταβλητές που αντιπροσωπεύουν τα δείγματα και τις μεταφέρουμε στο πλαίσιο *Test Pairs*. Επιλέγουμε το *Wilcoxon* και επειδή τα δείγματα είναι μικρά, κάνουμε *κλικ* στο κουμπί *Exact* και στο πλαίσιο *Exact Tests* επιλέγουμε *Exact* ή *Monte Carlo*. Με *κλικ* στο *OK* παίρνουμε τον πίνακα αποτελεσμάτων του σχήματος 7.20. Παρατηρούμε και εδώ ότι τα αποτελέσματα πρακτικά ταυτίζονται με τα αντίστοιχα του *ChemStat*.

Test Statistics[b]

	sample2 - sample1
Z	-,282[a]
Asymp. Sig. (2-tailed)	,778
Exact Sig. (2-tailed)	,828
Exact Sig. (1-tailed)	,414
Point Probability	,051

a. Based on negative ranks.
b. Wilcoxon Signed Ranks Test

Σχήμα 7.20. Αποτελέσματα από τον έλεγχο *Wilcoxon*

7.4 ΕΛΕΓΧΟΙ ΥΠΟΘΕΣΕΩΝ ΓΙΑ ΔΙΑΣΠΟΡΕΣ

Στον έλεγχο διασπορών η μηδενική υπόθεση διατυπώνεται ως

$$H_0 : \sigma_1^2 = \sigma_2^2 \tag{7.13}$$

με εναλλακτικές τις

$$H_1 : \sigma_1^2 \neq \sigma_2^2 \quad \text{ή} \quad H_1 : \sigma_1^2 > \sigma_2^2 \quad \text{ή} \quad H_1 : \sigma_1^2 < \sigma_2^2 \tag{7.14}$$

Στη στατιστική έχουν αναπτυχθεί αρκετοί έλεγχοι για την παραπάνω μηδενική υπόθεση. Ο πιο απλός είναι ο *έλεγχος F*, στον οποίον ως στατιστική συνάρτηση ελέγχου χρησιμοποιείται η μεταβλητή

$$F = s_1^2 / s_2^2 \tag{7.15}$$

όπου τα δείγματα 1 και 2 επιλέγονται έτσι ώστε να ισχύει $s_1^2 \geq s_2^2$. Η μεταβλητή F όταν τα δείγματα προέρχονται από κανονικό πληθυσμό ακολουθεί την κατανομή F με m_1-1 και m_2-1 βαθμούς ελευθερίας. Συνεπώς, στο *Excel* αφού υπολογίσουμε την τιμή της F, η p-value υπολογίζεται με τον τύπο: =2*FDIST(F;m_1-1;m_2-1).

Άλλοι έλεγχοι, που όμως χρησιμοποιούν πιο πολύπλοκες στατιστικές συναρτήσεις ελέγχου, είναι ο έλεγχος *Bartlett*, ο έλεγχος *Levene* και η παραλλαγή του ελέγχου *Levene* που είναι ο έλεγχος *Brown – Forsythe*. Βασικό πλεονέκτημα αυτών των ελέγχων είναι ότι δεν απαιτούν τα δεδομένα να ακολουθούν την κανονική κατανομή και επιπλέον μπορούν να χρησιμοποιηθούν για τον έλεγχο της *ισότητας (ομοιογένειας)* της διασποράς σε περισσότερα από δύο δείγματα. Από τους ελέγχους *Levene* και *Brown – Forsythe* θα πρέπει ο *Levene* να προτιμάται σε κανονικά δείγματα, επειδή χρησιμοποιεί τη μέση τιμή, και ο *Brown – Forsythe* σε μη κανονικά επειδή βασίζεται στη διάμεσο.

Παρατήρηση. Ο παραμετρικός έλεγχος μέσων τιμών ελέγχει μόνο τη διαφορά στις μέσες τιμές, ενώ ο μη παραμετρικός ελέγχει αν τα δείγματα προέρχονται από τον ίδιο πληθυσμό. Είναι όμως προφανές ότι αν δύο κανονικά δείγματα προέρχονται από πληθυσμούς που έχουν ίσες μέσες τιμές και ίσες διασπορές, τότε θα προέρχονται από τον ίδιο κανονικό πληθυσμό.

Παράδειγμα 7.5

Να εξετασθεί αν οι διαφορές των διασπορών στα δείγματα των παραδειγμάτων 7.1 και 7.2 είναι στατιστικά σημαντικές.

❖ **Ανάλυση στο ChemStat**

Όπως είδαμε, στο *ChemStat* ο έλεγχος των διασπορών γίνεται ταυτόχρονα με τον έλεγχο των μέσων τιμών. Εναλλακτικά όμως μπορούμε να πάμε *Πρόσθετα* → *ChemStat* → *Test of Variances*. Στο πρώτο παράθυρο που ανοίγει πληκτρολογούμε το πλήθος των δειγμάτων, στα επόμενα εισάγουμε τις περιοχές που είναι οι τιμές κάθε ενός δείγματος χωριστά ξεκινώντας από το μεγαλύτερο δείγμα και στο τελευταίο παράθυρο εισάγουμε το κελί εξόδου των αποτελεσμάτων. Όταν ελέγχονται δύο δείγματα το πρόγραμμα εκτελεί τους ελέγχους *Brown – Forsythe, Levene* και *έλεγχο F*, ενώ όταν υπάρχουν περισσότερα δείγματα εκτελούνται μόνο οι έλεγχοι *Brown – Forsythe* και *Levene*. Τα αποτελέσματα δίνονται στο σχήμα 7.21 όπου παρατηρούμε τα ακόλουθα:

Tests of variance - Παράδειγμα 7.1		
Brown - Forsythe's test		
p-value=	0,329125	Equality of variances may be assumed
Levene's test		
p-value=	0,34536	Equality of variances may be assumed
F test		
p-value=	0,478754	Equality of variances may be assumed
Tests of variance - Παράδειγμα 7.2		
Brown - Forsythe's test		
p-value=	0,086879	Equality of variances may be assumed
Levene's test		
p-value=	0,005252	Deviations from Equality of variances may be assumed
F test		
p-value=	2,13E-05	Deviations from Equality of variances may be assumed

Σχήμα 7.21. Αποτελέσματα από τον έλεγχο διασπορών

Στο παράδειγμα 7.1 όλοι οι έλεγχοι δείχνουν ότι η μηδενική υπόθεση, $H_0 : \sigma_1^2 = \sigma_2^2$, δεν μπορεί να απορριφθεί και συνεπώς σε επίπεδο σημαντικότητας α = 0.05 δεν διαπιστώνονται στατιστικά σημαντικές

διαφοροποιήσεις στις διασπορές των δειγμάτων. Σε ό,τι αφορά τα δείγματα του παραδείγματος 7.2, οι έλεγχοι δίνουν διαφορετικά αποτελέσματα, αν και σύμφωνα με τους ελέγχους *F* και *Levene* η μηδενική υπόθεση μπορεί να απορριφθεί. Αντίθετα, η μηδενική υπόθεση δεν μπορεί να απορριφθεί με βάση τον έλεγχο *Brown – Forsythe* που δίνει p-value = 0.087. Όμως, όπως έχει ήδη αναφερθεί, επειδή τα δείγματα προέρχονται από κανονικό πληθυσμό, ο έλεγχος *Levene* (που στηρίζεται στη μέση τιμή) σε αυτή την περίπτωση ίσως να είναι καλύτερος. Έτσι αν στηριχθούμε σε αυτό το δεδομένο και στο αποτέλεσμα του ελέγχου *F*, απορρίπτουμε τη μηδενική υπόθεση και δεχόμαστε ότι τα δείγματα έχουν στατιστικά σημαντικές διασπορές. Αυτό μάλιστα φαίνεται χαρακτηριστικά στο σχήμα 7.10-δεξιά.

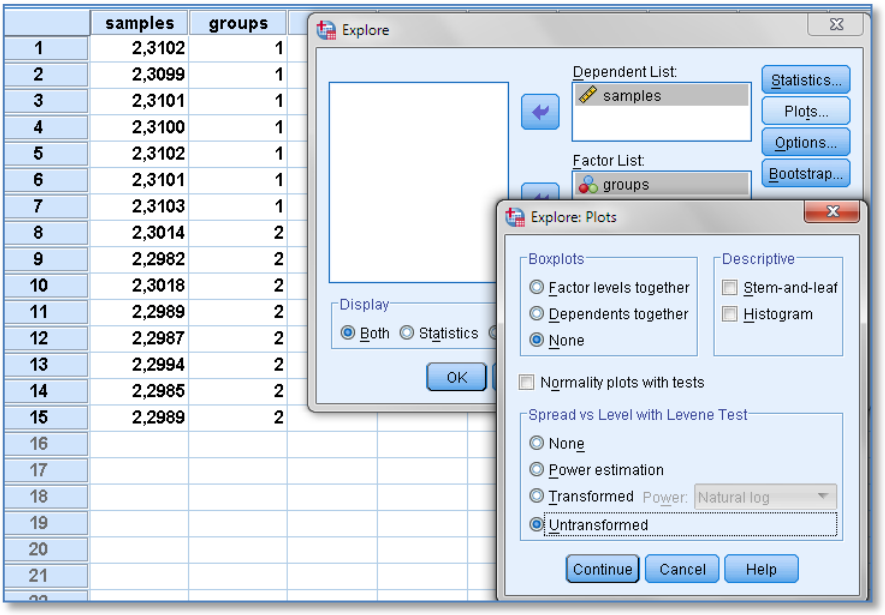
❖ Ανάλυση στο Excel

Στο *Excel*, όπως ήδη αναφέραμε, η p-value για δίπλευρο έλεγχο μπορεί να υπολογιστεί με τον τύπο: =2*FDIST(F;m_1-1;m_2-1). Εναλλακτικά πηγαίνουμε *Δεδομένα* (*Data)* → *Ανάλυση* (*Analysis)* → *Ανάλυση δεδομένων* (*Data Analysis)* και επιλέγουμε *Έλεγχος F των διακυμάνσεων δύο δειγμάτων (F test Two-Sample for Variances)*. Το πρόγραμμα εκτελεί μόνο μονόπλευρο (one tail) έλεγχο με βάση το *κριτήριο F*. Συνεπώς, για να προσδιορίσουμε την πιθανότητα p-value σε δίπλευρο έλεγχο διπλασιάζουμε την τιμή p-value του μονόπλευρου ελέγχου. Τα αποτελέσματα που παίρνουμε ταυτίζονται με τα αντίστοιχα του πίνακα του σχήματος 7.21.

❖ Ανάλυση στο SPSS

Στο *SPSS* δεν υπάρχει ο έλεγχος *F*, αλλά οι έλεγχοι *Levene* και *Brown – Forsythe*. Για να γίνουν οι έλεγχοι αυτοί, έστω για τα δείγματα του παραδείγματος 7.2, διευθετούμε τα δεδομένα όπως στο σχήμα 7.22. Ακολούθως πηγαίνουμε *Analyze* → *Descriptive Statistics* → *Explore* και στο παράθυρο που ανοίγει μεταφέρουμε τη στήλη με τα δείγματα, samples, στο πλαίσιο *Dependent List* και τη στήλη groups στο *Factor List.* Στη συνέχεια κάνουμε *κλικ* στο κουμπί *Plots* και στο νέο παράθυρο επιλέγουμε *Untrasformed* και απενεργοποιούμε τα *Boxplots* και το *Stem-and-leaf* (σχήμα 7.22).

Με *κλικ* στο *Continue* και στο *OK* παίρνουμε τα αποτελέσματα του σχήματος 7.23, που ταυτίζονται με τα αντίστοιχα του σχήματος 7.21. Σημειώστε ότι στο σχήμα αυτό ο έλεγχος *Levene* αντιστοιχεί στο *Based on Mean* και ο έλεγχος *Brown - Forsythe* στο *Based on Median*.



Σχήμα 7.22. Διευθέτηση δειγμάτων στο *SPSS* και εισαγωγή τους για έλεγχο διασπορών

Test of Homogeneity of Variance

		Levene Statistic	df1	df2	Sig.
samples	Based on Mean	11,202	1	13	,005
	Based on Median	3,430	1	13	,087
	Based on Median and with adjusted df	3,430	1	7,073	,106
	Based on trimmed mean	9,360	1	13	,009

Σχήμα 7.23. Αποτελέσματα από τους ελέγχους *Levene* (*Based on Mean*) και *Brown - Forsythe* (*Based on Median*)

Παράδειγμα 7.6

Έστω διάλυμα σεληνιοουρίας με συγκέντρωση 50 ng/mL. Δύο αναλυτικές μέθοδοι προσδιορισμού της σεληνιοουρίας έδωσαν τα ακόλουθα αποτελέσματα:

Πρώτη μέθοδος: 50.4, 50.7, 49.1, 49.5, 51.0

Δεύτερη μέθοδος: 50.3, 50.2, 49.8, 49.6, 50.1

Ποια μέθοδος είναι καλύτερη;

◆ Για να αντιμετωπίσουμε τέτοιου είδους προβλήματα κάνουμε τους ακόλουθους ελέγχους:

1) Ελέγχουμε την κανονικότητα των τιμών των δύο δειγμάτων. Αν τα δείγματα είναι κανονικά προχωρούμε στα επόμενα βήματα, αν και όπως έχει αναφερθεί, η κανονικότητα δειγμάτων μετρήσεων που έγιναν στο ίδιο σύστημα κάτω από σταθερές και ελεγχόμενες συνθήκες πρέπει να θεωρείται δεδομένη.
2) Ελέγχουμε για κάθε δείγμα τις υποθέσεις
$$H_0: \mu = 50 \quad \text{και} \quad H_1: \mu \neq 50$$
3) Αν για κάποιο από τα δείγματα η μηδενική υπόθεση απορρίπτεται, η αναλυτική μέθοδος που οδήγησε στις τιμές αυτού του δείγματος απορρίπτεται, επειδή πιθανότατα εισάγει κάποιο συστηματικό σφάλμα.
4) Αν και για τα δύο δείγματα η υπόθεση $H_0: \mu = 50$ δεν απορρίπτεται, τότε είναι πιθανόν και οι δύο μέθοδοι να μην εισάγουν συστηματικά σφάλματα. Σε αυτή την περίπτωση ελέγχουμε τις υποθέσεις:
$$H_0: \sigma_1^2 = \sigma_2^2 \text{ και } H_1: \sigma_1^2 > \sigma_2^2 \text{ ή } H_1: \sigma_1^2 < \sigma_2^2$$
5) Αν και για τα δύο δείγματα η μηδενική υπόθεση $H_0: \sigma_1^2 = \sigma_2^2$ δεν μπορεί να απορριφθεί, τότε είναι πιθανόν και οι δύο μέθοδοι να είναι εξίσου ικανοποιητικές, δεδομένου ότι τα δείγματά τους έχουν μέση τιμή που δεν φαίνεται να διαφοροποιείται στατιστικά από την τιμή της πρότυπης συγκέντρωσης, ενώ δεν διαπιστώνεται διαφοροποίηση και στις διασπορές των τιμών των δύο δειγμάτων. Φυσικά το συμπέρασμα αυτό ισχύει με την προϋπόθεση ότι τα δείγματα είναι σχετικά μεγάλα.
6) Αν απορριφθεί η μηδενική υπόθεση υπέρ της $H_1: \sigma_1^2 < \sigma_2^2$, τότε η πρώτη μέθοδος υπερτερεί της δεύτερης, επειδή η διασπορά των μετρήσεών της είναι μικρότερη. Προφανώς αν η μηδενική υπόθεση απορριφθεί υπέρ της $H_1: \sigma_1^2 > \sigma_2^2$, τότε υπερτερεί η δεύτερη μέθοδος.

Ο έλεγχος *Anderson-Darling* δεν δείχνει κάποιο πρόβλημα με την κανονικότητα και ο έλεγχος της μέσης τιμής δείχνει ότι η μηδενική υπόθεση $H_0: \mu = 50$ δεν μπορεί να απορριφθεί και για τα δύο δείγματα (σχήμα 7.24, όπου συνοψίζονται τα αποτελέσματα μόνο του παραμετρικού

ελέγχου). Συνεπώς είναι πιθανόν και οι δύο μέθοδοι να μην εισάγουν συστηματικά σφάλματα αν και στο συμπέρασμα αυτό θα πρέπει να είμαστε επιφυλακτικοί επειδή τα δείγματα είναι μικρά. Θα πρέπει επομένως να ελέγξουμε τις υποθέσεις:

$$H_0: \sigma_1^2 = \sigma_2^2 \quad \text{και} \quad H_1: \sigma_1^2 > \sigma_2^2 \quad \text{ή} \quad H_1: \sigma_1^2 < \sigma_2^2$$

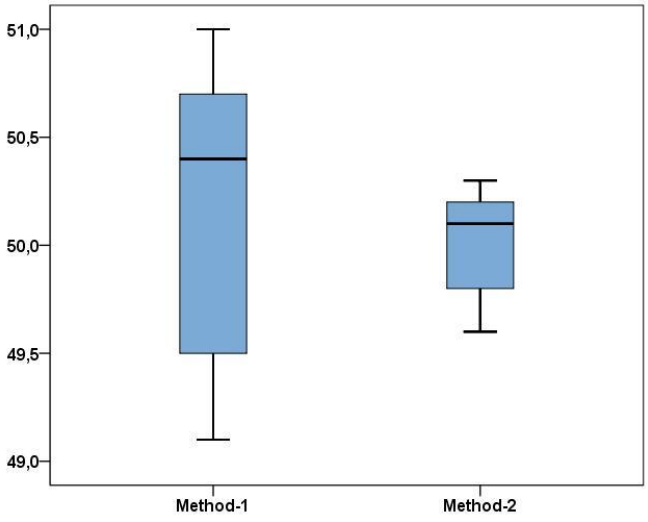
One Sample tests - sample 1					
Anderson-Darling Normality test:					
Sample too small. Normality Results may be unreliable					
p-value=	0,508711	Normality may be assumed			
Parametric t-test					
Mean=	50,14		Var=	0,653	
2 tailed t-test:					
t-value=	0,387397				
p-value=	0,718192	Null hypothesis, mean = 50, may be assumed at level 0.05			
One Sample tests - sample 2					
Anderson-Darling Normality test:					
Sample too small. Normality Results may be unreliable					
p-value=	0,526934	Normality may be assumed			
Parametric t-test					
Mean=	50		Var=	0,085	
2 tailed t-test:					
t-value=	0				
p-value=	1	Null hypothesis, mean = 50, may be assumed at level 0.05			
Tests of variance					
Brown - Forsythe's test					
p-value=	0,13764	Equality of variances may be assumed			
Levene's test					
p-value=	0,016266	Deviations from Equality of variances may be assumed			
F test					
p-value=	0,073482	Equality of variances may be assumed			

Σχήμα 7.24. Αποτελέσματα στατιστικών ελέγχων του παραδείγματος 7.6

Πηγαίνουμε *Πρόσθετα* → *ChemStat* → *Test of Variances* και συμπληρώνουμε κατάλληλα τα πλαίσια που εμφανίζονται. Οι έλεγχοι δίνουν τα αποτελέσματα του σχήματος 7.24. Όμως πρέπει να προσέξουμε, επειδή το πρόγραμμα υπολογίζει την τιμή p-value που αντιστοιχεί σε δίπλευρο έλεγχο. Συνεπώς θα πρέπει οι τιμές 0.1376, 0.01627 και 0.07348 στον πίνακα του σχήματος 7.24 να διαιρεθούν δια 2.

Παρατηρούμε και εδώ μια διαφοροποίηση των κριτηρίων. Με το κριτήριο *F* (p-value = 0.07348/2 = 0.0367) και τον έλεγχο *Levene* (p-value = 0.01627/2 = 0.008) υπάρχει στατιστικά σημαντική διαφορά στις διασπορές. Μάλιστα επειδή η διασπορά των τιμών του πρώτου δείγματος είναι μεγαλύτερη από τη διασπορά των τιμών του δευτέρου (0.653 > 0.085) μπορούμε να καταλήξουμε στο συμπέρασμα ότι η δεύτερη μέθοδος είναι καλύτερη. Αντίθετα, ο έλεγχος *Brown – Forsythe,* που στο συγκεκριμένο πρόβλημα πρέπει να θεωρηθεί ασθενέστερος λόγω της κανονικότητας των δειγμάτων, δείχνει ότι η μηδενική υπόθεση H_0: $\sigma_1^2 = \sigma_2^2$ δεν μπορεί να απορριφθεί σε επίπεδο σημαντικότητας 0.05 (p-value = 0.1376/2 = 0.069 >0.05).

Η διαφοροποίηση των δύο μεθόδων και η καλύτερη επίδοση της δεύτερης φαίνεται εποπτικά μέσω των θηκογραμμάτων των δειγμάτων τους στο σχήμα 7.25. Από το σχήμα αυτό προκύπτει ότι η τιμή μ = 50 βρίσκεται κοντά στο μέσο όρο των τιμών των δύο δειγμάτων, ενώ η διασπορά των τιμών του δεύτερου δείγματος είναι σαφώς πιο περιορισμένη γύρω από τη μέση τιμή του.



Σχήμα 7.25. Θηκογράμματα των τιμών των δειγμάτων των δύο αναλυτικών μεθόδων

7.5 ΥΠΟΛΟΓΙΣΤΙΚΕΣ ΣΤΑΤΙΣΤΙΚΕΣ ΤΕΧΝΙΚΕΣ

Οι υπολογιστικές στατιστικές τεχνικές περιλαμβάνουν κυρίως τις μεθόδους *Bootstrap* και *Monte-Carlo με αντιμεταθέσεις (permutations)*. Όταν το σύνολο των αντιμεταθέσεων σε έναν έλεγχο μπορεί να προσδιοριστεί επακριβώς, τότε η μέθοδος *Monte-Carlo* ανάγεται στην *Ακριβή (Exact)* μέθοδο.

Οι μέθοδοι αυτές εφαρμόζονται πάντα συμπληρωματικά τόσο των παραμετρικών μεθόδων όταν υπάρχουν αμφιβολίες ως προς τις προϋποθέσεις εφαρμογής των, όσο και των μη παραμετρικών μεθόδων όταν έχουμε μικρά δείγματα και συνεπώς αμφιβολίες ως προς την ασυμπτωτική συμπεριφορά τους.

7.5.1 Η ΜΕΘΟΔΟΣ BOOTSTRAP

Η μέθοδος *bootstrap* βασίζεται στη δημιουργία N (> 1000) δειγμάτων που έχουν ίδιο μέγεθος με το αρχικό δείγμα. Τα δείγματα αυτά δημιουργούνται με δειγματοληψία με επανατοποθέτηση από το αρχικό δείγμα. Σύμφωνα με αυτόν τον τύπο δειγματοληψίας, κάθε νέο δείγμα με m τιμές δημιουργείται επιλέγοντας με τυχαίο τρόπο m τιμές του αρχικού δείγματος, όπου κάθε τιμή μπορεί να μην επιλεγεί ή να επιλεγεί μία ή περισσότερες φορές, όπως στον πίνακα 7.5.

Σε στατιστικό έλεγχο ενός δείγματος, η στατιστική συνάρτηση ελέγχου υπολογίζεται σε κάθε ένα νέο δείγμα. Από τις τιμές αυτές και την τιμή της στατιστικής συνάρτησης ελέγχου στο αρχικό δείγμα υπολογίζεται η p-value ως εξής. Έστω X_i η τιμή της στατιστικής συνάρτησης ελέγχου στο νέο δείγμα i και c η τιμή της στο αρχικό δείγμα. Τότε σε δίπλευρο έλεγχο ισχύει

$$\text{p-value} = \frac{2 \cdot N(|X_i| \geq |c|)}{N} \qquad (7.16)$$

όπου $N(|X_i| \geq |c|)$ είναι το πλήθος των περιπτώσεων όπου το $|X_i|$ είναι μεγαλύτερο ή ίσο του |c|. Σε μονόπλευρο έλεγχο ισχύει η (7.16) χωρίς τον συντελεστή 2.

Για παράδειγμα, έστω το δείγμα {40.3, 40.2, 40.2, 40.0, 40.3} του παραδείγματος 6.4. Για να εξετάσουμε με τη *μέθοδο Bootstrap* τη μηδενική υπόθεση μ = 40 πρώτα ορίζουμε την κατάλληλη στατιστική συνάρτηση ελέγχου. Στο συγκεκριμένο πρόβλημα αυτή μπορεί να οριστεί από τη σχέση

$$t = \frac{\bar{x} - \bar{x}_0}{s}\sqrt{m} \tag{7.17}$$

όπου $\bar{x}_0$ είναι η μέση τιμή του αρχικού δείγματος, και $\bar{x}$ και s είναι η μέση τιμή και η τυπική απόκλιση κάθε νέου δείγματος που δημιουργείται από το αρχικό με δειγματοληψία με επανατοποθέτηση. Για παράδειγμα, πέντε τέτοια δείγματα δίνονται στον Πίνακα 7.5. Η απόλυτη τιμή του t συγκρίνεται με την απόλυτη τιμή του $t = t_0$ στο αρχικό δείγμα που υπολογίζεται από τη σχέση

$$t_0 = \frac{\bar{x}_0 - \mu_0}{s_0}\sqrt{m} = \frac{40.2 - 40}{0.122}\sqrt{10} \tag{7.18}$$

και η p-value υπολογίζεται από τη σχέση (7.16) όπου τώρα $N(|X_i| \geq |c|) = N(|t| \geq |t_0|)$ είναι το πλήθος των περιπτώσεων όπου το |t| είναι μεγαλύτερο ή ίσο του $|t_0|$.

Η μέθοδος επεκτείνεται άμεσα σε στατιστικούς ελέγχους δύο ή περισσοτέρων δειγμάτων.

Πίνακας 7.5. Δείγματα με δειγματοληψία με επανατοποθέτηση από το αρχικό δείγμα {40.3, 40.2, 40.2, 40.0, 40.3 }.

Δείγμα 1	40.0, 40.0, 40.3, 40.2, 40.3
Δείγμα 2	40.2, 40.3, 40.3, 40.3, 40.0
Δείγμα 3	40.2, 40.3, 40.2, 40.3, 40.0
Δείγμα 4	40.3, 40.2, 40.0, 40.0, 40.0
Δείγμα 5	40.0, 40.2, 40.3, 40.2, 40.2

7.5.2 Η ΜΕΘΟΔΟΣ MONTE-CARLO ΜΕ ΑΝΤΙΜΕΤΑΘΕΣΕΙΣ

Η μέθοδος αυτή εφαρμόζεται σε στατιστικούς ελέγχους που περιλαμβάνουν δύο ή περισσότερα δείγματα. Έστω ότι ελέγχουμε τη μηδενική υπόθεση H_0: $\mu_1 = \mu_2$ σε δύο δείγματα με m_1 και m_2 τιμές, αντίστοιχα. Τα βήματα που ακολουθούμε είναι τα εξής:

1. Υπολογίζεται η τιμή της στατιστικής συνάρτησης ελέγχου στα αρχικά δείγματα, έστω από τη σχέση (7.1), και έστω ότι αυτή είναι ίση με c.
2. Τα δείγματα ενώνονται σε ένα ενιαίο δείγμα με $m_1 + m_2$ τιμές.
3. Στο δείγμα αυτό οι τιμές αντιμετατίθενται με τρόπο τυχαίο και επιλέγονται m_1 τιμές για τη δημιουργία του πρώτου δείγματος, ενώ οι υπόλοιπες m_2 τιμές συγκροτούν το δεύτερο δείγμα.

4. Στα δύο νέα αυτά δείγματα υπολογίζεται η τιμή της στατιστικής συνάρτησης ελέγχου t πάλι από τη σχέση (7.1) και η απόλυτη τιμή της συγκρίνεται με την απόλυτη τιμή της c.
5. Τα βήματα 3-4 επαναλαμβάνονται N φορές, όπου το N κυμαίνεται μεταξύ 1000 και 10000, και καταμετρείται το πλήθος $N(|t| \geq |c|)$, δηλαδή το πλήθος των περιπτώσεων όπου $|t| \geq |c|$.
6. Σε δίπλευρο έλεγχο η p-value υπολογίζεται από τη σχέση (7.16), όπου $N(|X_i| \geq |c|) = N(|t| \geq |c|)$.

Όταν τα δείγματα είναι μικρά οι αντιμεταθέσεις μπορούν να προσδιοριστούν επακριβώς. Για παράδειγμα, αυτή η δυνατότητα υπάρχει στο *SPSS*. Τότε η μέθοδος ονομάζεται *Ακριβής* (*Exact*). Εναλλακτικά οι αντιμεταθέσεις γίνονται χρησιμοποιώντας τυχαίους αριθμούς, οπότε η μέθοδος ονομάζεται *Monte-Carlo με αντιμεταθέσεις* (*permutations*).

Στους στατιστικούς ελέγχους που αφορούν δύο δείγματα οι στατιστικές συναρτήσεις ελέγχου που συνήθως χρησιμοποιούνται είναι αυτές που μελετήθηκαν παραπάνω, δηλαδή είναι οι συναρτήσεις που ορίζονται από τις σχέσεις (7.1), (7.3), (7.6), (7.10), (7.11) και (7.15). Επομένως ως στατιστική συνάρτηση ελέγχου μπορεί να χρησιμοποιηθεί η συνάρτηση του παραμετρικού ή του μη παραμετρικού ελέγχου ή να εξεταστούν και οι δύο αυτές περιπτώσεις. Η τελευταία επιλογή υπάρχει στο *ChemStat*.

Ενδιαφέρον παρουσιάζει η περίπτωση του στατιστικού ελέγχου της μέσης τιμής ενός δείγματος, που εξετάστηκε στο προηγούμενο κεφάλαιο. Αυστηρά δεν υπάρχει δυνατότητα εφαρμογής της μεθόδου των *αντιμεταθέσεων* σε ένα δείγμα. Όμως ο έλεγχος αυτός είναι ισοδύναμος με τον αντίστοιχο στατιστικό έλεγχο σε ζεύγος δειγμάτων, όπου το ένα δείγμα είναι το αρχικό δείγμα με m τιμές και το άλλο είναι ένα δείγμα αποτελούμενο από m τιμές ίσες με μ_0. Συνεπώς ως στατιστική συνάρτηση ελέγχου μπορεί να χρησιμοποιηθεί η σχέση (7.1) ή και η σχέση (7.11).

ΠΡΟΣΟΧΗ. Η εφαρμογή της μεθόδου *Monte-Carlo με αντιμεταθέσεις* μπορεί να είναι ιδιαίτερα χρονοβόρος όταν τα δείγματα είναι μεγάλα ή υπάρχουν πολλά δείγματα ή στην περίπτωση κατηγορικών δεδομένων (Κεφάλαιο 9) όταν υπάρχουν κελιά με μεγάλες τιμές. Γι αυτό το λόγο είναι καλό να ξεκινάμε την εφαρμογή της μεθόδου με μικρό αριθμό επαναλήψεων N έτσι, ώστε αφού ελέγξουμε το χρόνο εκτέλεσης, να μπορέσουμε να εκτιμήσουμε τον χρόνο εκτέλεσης όταν θα αυξήσουμε τις επαναλήψεις σε 1000 ή ακόμη καλύτερα σε 10000.

ΑΣΚΗΣΕΙΣ

7.1. Να δημιουργηθούν 2 δείγματα με 4 τιμές το κάθε ένα από έναν πληθυσμό που ακολουθεί την κανονική κατανομή με σ = 1 και μ = 0 το πρώτο και μ = 0.5 δεύτερο. Ακολούθως να εξετασθεί αν υπάρχει στατιστικά σημαντική διαφορά μεταξύ των δειγμάτων.

7.2. Να επαναληφθεί η άσκηση 7.1 αυξάνοντας το πλήθος των τιμών των δειγμάτων σε α) 10, β) 50, γ) 100 και δ) 1000. Τι συμπέρασμα βγάζεται από τις ασκήσεις 7.1 και 7.2;

7.3. Δύο γεωλογικοί σχηματισμοί συγκρίνονται ανάλογα με την εκατοστιαία περιεκτικότητά τους σε μέταλλα. Τι συμπέρασμα προκύπτει με βάση τα αποτελέσματα του παρακάτω πίνακα;

Σχηματισμός 1:	8.8	12.2	5.8	8.7	9.8	6.3	14.8
Σχηματισμός 2:	5.9	7.8	4.5	4.9	4.1		

7.4. Σε υπερλιπιδεμικό άτομο χορηγήθηκε για ορισμένο χρονικό διάστημα εικονικό φάρμακο (Placebo) και προσδιορίστηκε το ποσό της χοληστερίνης σε g/L. Ακολούθως χορηγήθηκε αντιχοληστερινικό φάρμακο για ίδιο χρονικό διάστημα και προσδιορίστηκε πάλι το ποσό της χοληστερίνης σε g/L. Τα αποτελέσματα που ελήφθησαν ήταν τα ακόλουθα:

Εικονικό φάρμακο: 3.65, 4.20, 4.05, 4.35, 3.95, 4.20, 4.65

Αντιχοληστερινικό φάρμακο: 2.40, 2.65, 2.75, 3.10, 2.70

Είναι στατιστικά σημαντική η δράση του φάρμακου κατά της χοληστερίνης;

7.5. Για να ελέγξει δύο πεχάμετρα, Α και Β, ένας χημικός παρασκεύασε ένα διάλυμα με pH = 7 και πήρε τις ακόλουθες μετρήσεις. Τι συμπέρασμα προκύπτει για τα δύο όργανα;

A	6.95	6.9	7.02	6.89	7.05	6.97	6.99	7.01	7.03	6.83
B	7.1	7.05	6.93	6.87	7.07	7.01	7.03	7.06	6.95	6.97

7.6. Οκτώ διαλύματα ενός αντιδραστηρίου τιτλοδοτούνται με δύο διαφορετικούς τρόπους και λαμβάνονται τα παρακάτω αποτελέσματα σε mg/L. Να εξετασθεί εάν οι δύο αυτοί μέθοδοι δίνουν στατιστικά ισότιμα αποτελέσματα.

Διάλυμα	1	2	3	4	5	6	7	8
Μέθοδος 1	0.15	0.27	0.18	0.55	0.17	0.43	0.31	0.22
Μέθοδος 2	0.18	0.23	0.17	0.52	0.19	0.40	0.29	0.21

7.7. Στον παρακάτω πίνακα συγκρίνεται η ατομική απορρόφηση (AAS) με τη φασματομετρία για τον προσδιορισμό ασβεστίου στο αίμα. Οι τιμές του πίνακα είναι μετρήσεις της συγκέντρωσης του ασβεστίου σε mg/L στο ίδιο διάλυμα. Υπάρχει στατιστικά σημαντική διαφορά μεταξύ των δύο τεχνικών;

AAS	109	101	103	110	97	99	
Φασματομετρία	91	104	98	100	114	116	93

7.8. Εξαιτίας της μόλυνσης του περιβάλλοντος η συγκέντρωση του CO_2 αυξάνεται συνεχώς. Για να εξεταστεί η επίδραση αυτής της αύξησης στα φυτά έγινε η ακόλουθη μελέτη. Ένα δασύλλιο χωρίστηκε σε τρία τμήματα και κάθε τμήμα σε δύο επιμέρους τμήματα που σχημάτιζαν ένα ζεύγος. Στο ένα τμήμα του ζεύγους διοχετευόταν με ειδική συσκευή επιπλέον CO_2, ενώ στο άλλο δεν γινόταν καμία παρέμβαση. Μετά από 1 χρόνο προσδιορίστηκε η μέση αύξηση του ύψους των δένδρων ως κλάσμα του αρχικού τους ύψους και τα αποτελέσματα δίνονται στον παρακάτω πίνακα. Υπάρχει στατιστικά σημαντική επίδραση του CO_2 στο ύψος των φυτών;

Ζεύγη	Χωρίς διοχέτευση CO_2	Με διοχέτευση CO_2
1	0.06	0.08
2	0.05	0.06
3	0.04	0.06

7.9. Για να ελεγχθεί μια νέα μέθοδος προσδιορισμού του αζώτου στο αίμα παρασκευάζονται 6 δείγματα και προσδιορίζεται σε αυτά η συγκέντρωση του αζώτου σε mg/L με τη νέα μέθοδο και με μία πρότυπη μέθοδο. Τα αποτελέσματα δίνονται στον παρακάτω πίνακα. Τι συμπέρασμα βγάζετε για την προτεινόμενη μέθοδο;

Δείγμα	1	2	3	4	5	6
Νέα μέθοδος	122	166	88	120	155	123
Πρότυπη μέθοδος	125	160	89	113	151	120

7.10. Στον επόμενο πίνακα δίνεται η βαθμολογία με κλίμακα 0-100 του κοινού σε δύο προϊόντα. Συγκεκριμένα χρησιμοποιήθηκαν 20 εθελοντές στους οποίους δόθηκαν 10 προϊόντα τύπου Α και 10 τύπου Β. Κάθε εθελοντής βαθμολογούσε 1 μόνο προϊόν. Να συγκριθούν οι προτιμήσεις του κοινού.

Προϊόν Α:	50	95	83	65	55	55	85	73	85	58
Προϊόν Β:	45	51	55	70	74	45	76	54	50	75

7.11. Η %w/w περιεκτικότητα σε ασβέστιο ενός ορυκτού προσδιορίζεται με δύο διαφορετικές μεθόδους και τα αποτελέσματα δίνονται στον επόμενο πίνακα. Να συγκριθούν οι μέθοδοι.

Μέθοδος 1:	0.025	0.039	0.022	0.029	0.028
Μέθοδος 2:	0.037	0.027	0.042	0.029	0.037

7.12. Στο επόμενο πίνακα δίνεται η περιεκτικότητα %w/w σε Ti δύο διαφορετικών μεταλλευμάτων. Να συγκριθούν τα μεταλλεύματα ως προς τη συγκέντρωση του Ti.

Δείγμα 1:	0.015	0.015	0.017	0.014	0.016
Δείγμα 2:	0.024	0.019	0.018	0.025	0.020

7.13. Η περιεκτικότητα σε mg/L σε ασβέστιο στα ούρα ενός ατόμου προσδιορίζεται με επαναλαμβανόμενες μετρήσεις σε δύο διαδοχικές ημέρες και τα αποτελέσματα δίνονται στον παρακάτω πίνακα. Παρατηρείται στατιστικά σημαντική αύξηση του ασβεστίου τη δεύτερη ημέρα;

Ημέρα 1:	244	209	230	227
Ημέρα 2:	257	259	234	240

7.14. Το ετήσιο βροχομετρικό ύψος σε mm σε δύο γεωγραφικές περιοχές δίνεται στον παρακάτω πίνακα. Να εξετασθεί αν υπάρχει στατιστικά σημαντική διαφοροποίηση στις τιμές.

Περιοχή Α		Περιοχή Β	
102	178	115	117
210	78	135	302
99	102	330	280
109	85	95	205

7.15. Για να ελεγχθεί η υπόθεση ότι η συγκέντρωση της βιταμίνης C σε ειδικά αρτοσκευάσματα, που παρέχονται από προγράμματα διεθνούς βοήθειας για τη διατροφή παιδιών του τρίτου κόσμου, ελαττώνεται με το χρόνο, παρασκευάστηκαν 9 αρτοσκευάσματα και προσδιορίστηκε η συγκέντρωση της C (σε mg ανά 100 g) αμέσως μετά την παρασκευή τους και μετά από 6 μήνες. Τι συμπέρασμα προκύπτει με βάση τα αποτελέσματα του παρακάτω πίνακα:

t = 0	53	33	45	35	48	36	39	55	46
t = 6 μήνες	50	30	43	35	48	35	39	48	42

Κεφάλαιο 8

ANOVA : ΑΝΑΛΥΣΗ ΔΙΑΣΠΟΡΑΣ

8.1 ΕΙΣΑΓΩΓΗ

Η ***Ανάλυση Διασποράς*** *(ANOVA - ANalysis Of VAriance)* είναι ουσιαστικά η επέκταση των ελέγχων υποθέσεων για μέσες τιμές σε περισσότερα από δύο δείγματα. Η ανάπτυξη της μεθόδου οφείλεται στον θεμελιωτή της σύγχρονης στατιστικής επιστήμης, τον άγγλο *Sir Ronald Aylmer Fisher* (1890-1962).

Αν και υπάρχουν πολλές παραλλαγές της μεθόδου, μπορούμε να διακρίνουμε δύο περιπτώσεις: Τη *Μονο-παραγοντική Ανάλυση Διασποράς (One-way ANOVA)* και τη *Δι-παραγοντική Ανάλυση Διασποράς (Two-way ANOVA)*. Η δεύτερη έχει επίσης δύο υποπεριπτώσεις: την ανάλυση χωρίς αλληλεπιδράσεις ή με αλληλεπιδράσεις. Οι δύο αυτές υποπεριπτώσεις ονομάζονται και ανάλυση διασποράς χωρίς επαναλήψεις ή με επαναλήψεις.

Όπως στους περισσότερους στατιστικούς ελέγχους, η *ANOVA* μπορεί να γίνει με παραμετρικούς ή/και με μη παραμετρικούς ελέγχους. Όμως οι μη παραμετρικοί έλεγχοι είναι σχετικά περιορισμένοι. Συγκεκριμένα στη μονο-παραγοντική ανάλυση διασποράς χρησιμοποιείται το κριτήριο των *Kruskal-Wallis*, ενώ το κριτήριο *Friedman* μπορεί να χρησιμοποιηθεί στη δι-παραγοντική ανάλυση διασποράς. Στο κεφάλαιο αυτό θα εξεταστούν πρώτα οι παραμετρικοί και ακολούθως οι μη παραμετρικοί έλεγχοι.

8.2 ΜΟΝΟΠΑΡΑΓΟΝΤΙΚΗ ΑΝΑΛΥΣΗ ΔΙΑΣΠΟΡΑΣ

Στη μονοπαραγοντική ανάλυση διασποράς, η *ANOVA* μας δίνει τη δυνατότητα να ελέγξουμε την υπόθεση ότι οι μέσες τιμές διαφόρων δειγμάτων είναι ίσες. Συνεπώς η μηδενική υπόθεση που καλούμαστε να εξετάσουμε είναι:

$$H_0: \mu_1 = \mu_2 = \mu_3 = \ldots$$

με εναλλακτική ότι τουλάχιστον μία από τις παραπάνω ισότητες δεν ισχύει.

Εκ πρώτης όψεως φαίνεται παράδοξο το γεγονός ότι, όπως υποδηλώνει το όνομα *ANOVA*, η μέθοδος χρησιμοποιεί διασπορές για να διακρίνει στατιστικά σημαντικές διαφορές στις μέσες τιμές δειγμάτων. Το παράδοξο εξηγείται από το γεγονός ότι μπορούμε να ορίσουμε περισσότερες από μία διασπορές, από τις οποίες άλλες εξαρτώνται και άλλες δεν εξαρτώνται από τις τιμές των μέσων όρων των δειγμάτων.

Έστω ότι έχουμε τα ακόλουθα δείγματα, που για λόγους απλότητας έχουν όλα το ίδιο μέγεθος:

Δείγμα 1	x_{11}	x_{12}	. . .	x_{1m}	$\bar{x}_1$	s_1^2
Δείγμα 2	x_{21}	x_{22}	. . .	x_{2m}	$\bar{x}_2$	s_2^2
. . .			. . .			
Δείγμα n	x_{n1}	x_{n2}	. . .	x_{nm}	$\bar{x}_n$	s_n^2

Σε αυτή την περίπτωση μπορούμε να ορίσουμε τις ακόλουθες εκφράσεις διασποράς:

α) ***Διασπορά μεταξύ των δειγμάτων*** *(between-groups variance)*

$$s_b^2 = \frac{\sum_{i=1}^{n} m_i(\bar{x}_i - \bar{\bar{x}})^2}{n-1} = \frac{m\sum_{i=1}^{n}(\bar{x}_i - \bar{\bar{x}})^2}{n-1} \qquad (8.1)$$

όπου $\bar{\bar{x}}$ είναι η μέση τιμή όλων των τιμών και $m_1 = m_2 = \ldots = m_n = m$

β) ***Διασπορά μέσα στα δείγματα*** *(within-groups variance)*

$$s_w^2 = \frac{\sum_{i=1}^{n}(m_i - 1)s_i^2}{\sum_{i=1}^{n}(m_i - 1)} = \frac{\sum_{i=1}^{n} s_i^2}{n} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m}(x_{ij} - \bar{x}_j)^2}{n(m-1)} \qquad (8.2)$$

Αποδεικνύεται ότι με την προϋπόθεση ότι όλα τα δείγματα προέρχονται από τον ίδιο πληθυσμό ισχύει

$$< s_w^2 > = < s_b^2 > = \sigma^2 \qquad (8.3)$$

και συνεπώς ασυμπτωτικά θα ισχύει $F = s_b^2 / s_w^2 = 1$.

Όμως αν ισχύει η ομοιογένεια της διασποράς, δηλαδή $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma^2$, αλλά τα δείγματα ανήκουν σε πληθυσμούς με διαφορετικές μέσες τιμές, τότε η διασπορά μέσα στα δείγματα δεν επηρεάζεται από την διαφορετικότητα των μέσων τιμών, ενώ αντίθετα η διασπορά μεταξύ των δειγμάτων αυξάνει επειδή αυξάνει η ποσότητα $(\bar{x}_i - \bar{x})^2$. Συνεπώς όταν υπάρχουν δείγματα με στατιστικά σημαντικές διαφορές στις μέσες τιμές τους, η συνάρτηση $F = s_b^2 / s_w^2$ αυξάνει. Επιπλέον αποδεικνύεται ότι η

$$F = s_b^2 / s_w^2 \qquad (8.4)$$

ακολουθεί την κατανομή F ή Fisher με n-1 και n(m-1) βαθμούς ελευθερίας. Συνεπώς μπορούμε να προσδιορίσουμε την κρίσιμη τιμή της F καθώς και την τιμή της πιθανότητας p-value που σχετίζεται με τη μηδενική υπόθεση ότι όλα τα δείγματα προέρχονται από πληθυσμούς με ίσες μέσες τιμές.

8.2.1 ΠΡΟΫΠΟΘΕΣΕΙΣ ΕΦΑΡΜΟΓΗΣ ΤΗΣ ΜΕΘΟΔΟΥ

Για να είναι επιτρεπτή η εφαρμογή της μεθόδου πρέπει να πληρούνται οι εξής τρεις προϋποθέσεις:

α) Θα πρέπει να υπάρχει *ομοιογένεια της διασποράς* (*homogeneity of variance*), $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma^2$.

β) Τα δείγματα πρέπει να προέρχονται από κανονικούς πληθυσμούς. Μικρές αποκλίσεις από την κανονική κατανομή δεν επηρεάζουν τα αποτελέσματα της μεθόδου.

γ) Τα δείγματα πρέπει να είναι ανεξάρτητα.

Για την εφαρμογή της μεθόδου δεν είναι απαραίτητο τα δείγματα να έχουν το ίδιο πλήθος τιμών.

8.2.2 ΠΟΛΛΑΠΛΟΙ ΕΛΕΓΧΟΙ

Όταν συγκρίνουμε την τιμή $F = s_b^2 / s_w^2$ με την κρίσιμη τιμή της ή όταν υπολογίζουμε με βάση την F την πιθανότητα p-value, μπορούμε να διαπιστώσουμε αν υπάρχουν στατιστικά σημαντικές διαφορές μεταξύ των δειγμάτων. Δε διευκρινίζεται όμως μεταξύ ποιών δειγμάτων εντοπίζονται αυτές οι διαφορές. Για το σκοπό αυτό έχουν αναπτυχθεί αρκετά κριτήρια, όπως τα κριτήρια *Tukey*, *Sidak*, *LSD*, *Dyncan*, *Dunnett* κ.ά. καθώς επίσης

και οι διορθώσεις *Bonferroni* και *Holm-Bonferroni*. Από αυτά συνήθως προτιμούνται το κριτήριο *Tukey* και η διόρθωση *Holm-Bonferroni*.

Στη διόρθωση *Holm-Bonferroni* γίνονται πρώτα όλοι οι έλεγχοι των δειγμάτων ανά δύο χρησιμοποιώντας τον *έλεγχο t*, που εξετάσαμε στο κεφάλαιο 7.2, με τις ακόλουθες τροποποιήσεις. Ως συνάρτηση στατιστικού ελέγχου χρησιμοποιείται η t από τη σχέση (7.1), όπου όμως η τυπική απόκλιση υπολογίζεται από τη διασπορά μέσα στα δείγματα, σχέση (8.2). Με αυτή την τροποποίηση η μεταβλητή t της σχέσης (7.1) ακολουθεί την κατανομή *student* με nm-n βαθμούς ελευθερίας.

Από τους ελέγχους των δειγμάτων ανά δύο καταγράφονται οι τιμές p-value. Όμως αυτές πρέπει να διορθωθούν για τον ακόλουθο λόγο. Έστω ότι εφαρμόζουμε τον έλεγχο t στα δείγματα 1 και 2 με επίπεδο σημαντικότητας α = 0.05 και μετά στο ίδιο επίπεδο σημαντικότητας ελέγχουμε τα δείγματα 1 και 3. Η πιθανότητα στον πρώτο έλεγχο να απορριφθεί σωστά η μηδενική υπόθεση είναι 0.95 και η ίδια πιθανότητα ισχύει για τον δεύτερο έλεγχο. Συνεπώς σύμφωνα με τον πολλαπλασιαστικό κανόνα των πιθανοτήτων, η πιθανότητα να απορριφθεί σωστά η μηδενική υπόθεση στους δύο ελέγχους ελαττώνεται από 0.95 σε

$$0.95 \cdot 0.95 = 0.9025$$

Άρα η πιθανότητα να απορριφθεί εσφαλμένα η μηδενική υπόθεση σε έναν από τους δύο ελέγχους αυξάνεται από 0.05 σε 1-0.9025 = 0.0978. Αν έχουμε 5 δείγματα απαιτούνται 10 έλεγχοι των δειγμάτων ανά δύο. Σε αυτή την περίπτωση η πιθανότητα να απορριφθεί εσφαλμένα η μηδενική υπόθεση σε έναν από τους 10 ελέγχους αυξάνεται από 0.05 σε $1-0.95^{10} = 0.4012$. Δηλαδή η πιθανότητα απόρριψης μιας σωστής μηδενικής υπόθεσης σε έναν από τους 10 ελέγχους αυξάνει υπερβολικά από το 5% στο 40.1%.

Για τη διόρθωση των τιμών p-value με τη μέθοδο *Holm-Bonferroni* οι τιμές p-value κατατάσσονται από τη μικρότερη στη μεγαλύτερη,

$$\text{p-value}(1) < \text{p-value}(2) < \text{p-value}(3) < \ldots < \text{p-value}(N)$$

και ακολούθως ξεκινώντας από τη μικρότερη πολλαπλασιάζονται επί N, N-1, N-2, …, 1, όπου N είναι το πλήθος των τιμών p-value. Ο πολλαπλασιασμός αυτός οδηγεί στις διορθωμένες τιμές p-value, τις p-HB. Πιο αυστηρά οι διορθωμένες τιμές προκύπτουν με βάση τις σχέσεις:

$$\text{p-HB}(1) = \min\{N \times \text{p-value}(1), 1\}$$

$$\text{p-HB}(j) = \min\{\max[\text{p-HB}(j-1), (N-j+1) \times \text{p-value}(j)], 1\},\ j = 2,3,\ldots,N$$

όπου min(a, b) είναι η ελάχιστη των τιμών a και b και max(c, d) η μέγιστη

των τιμών c και d.

Στο *ChemStat* το πρόγραμμα *Holm-Bonferroni* εκτελεί τη διόρθωση *Holm-Borferroni* σε έναν πίνακα με τιμές p-value, με την προϋπόθεση ότι αυτές περιβάλλονται αριστερά από μία στήλη με αλφαριθμητικά και στο επάνω μέρος από μία γραμμή με τις ονομασίες των μεταβλητών.

Σε ό,τι αφορά το κριτήριο *Tukey*, και εδώ γίνονται όλοι οι έλεγχοι των δειγμάτων ανά δύο, όμως χρησιμοποιώντας μια παραλλαγή της *κατανομής t* που ονομάζεται *τυποποιημένη κατά student κατανομή εύρους (Studentized range distribution)*.

Παράδειγμα 8.1

Έστω ότι θέλουμε να συγκρίνουμε 5 καταλυτικά συστήματα συγκράτησης υδρογονανθράκων για βενζινοκινητήρες. Για το σκοπό αυτό προσδιορίστηκε η συγκέντρωση σε ppm των υδρογονανθράκων που εκπέμπονται και τα αποτελέσματα δίνονται στον πίνακα 8.1. Να εξετασθεί αν υπάρχουν στατιστικά σημαντικές διαφορές μεταξύ των καταλυτικών συστημάτων σε επίπεδο σημαντικότητας α = 0.05.

Πίνακας 8.1. Τιμές συγκέντρωσης υδρογονανθράκων σε ppm.

Καταλυτικό σύστημα	Συγκέντρωση					
1	100	95	98	90	92	98
2	85	88	86	90	88	90
3	82	80	77	85	83	79
4	90	85	88	82	86	84
5	90	95	89	85	88	89

❖ **Ανάλυση στο ChemStat**

Για να εφαρμόσουμε την *ANOVA* στο *ChemStat* εισάγουμε τα δείγματα όπως φαίνεται στο σχήμα 8.1 και ακολούθως πηγαίνουμε

Πρόσθετα → *ChemStat* → *ANOVA* → *ANOVA parametric*

Στο πρώτο παράθυρο που ανοίγει πληκτρολογούμε τον αριθμό 1 για να επιλεγεί η μονοπαραγοντική *ANOVA*, στο δεύτερο παράθυρο που μας ενημερώνει για τη διευθέτηση των δειγμάτων πατάμε *OK*, στο τρίτο παράθυρο εισάγουμε τον αριθμό 5 που είναι το πλήθος των δειγμάτων και

στα επόμενα πέντε παράθυρα εισάγουμε διαδοχικά τις τιμές των δειγμάτων. Στο τελευταίο παράθυρο κάνουμε *κλικ* στο κελί εξόδου των αποτελεσμάτων.

Δ1	Δ2	Δ3	Δ4	Δ5
100	85	82	90	90
95	88	80	85	95
98	86	77	88	89
90	90	85	82	85
92	88	83	86	88
98	90	79	84	89

ONE WAY ANOVA

Anderson-Darling Normality Tests

p-value(1)=	0,51551416
p-value(2)=	0,41445331
p-value(3)=	0,94640692
p-value(4)=	0,94199172
p-value(5)=	0,26983742

Homogeneity of variances Levene's test

p-value=	0,58441248	Equality of variances may be assumed

F test for ANOVA

p-value=	4,4464E-07	Null hypothesis, m1=m2=m3=..., may be rejected

Pairwise comparisons

pair	p-uncorrectec	HB correction
pair 12	0,0001986	0,00139
pair 13	1,3791E-08	1,38E-07
pair 14	1,054E-05	9,49E-05
pair 15	0,00175476	0,008774
pair 23	0,00067201	0,004032
pair 24	0,26676109	0,533522
pair 25	0,40233356	0,533522
pair 34	0,01103819	0,044153
pair 35	7,4431E-05	0,000595
pair 45	0,05788319	0,17365

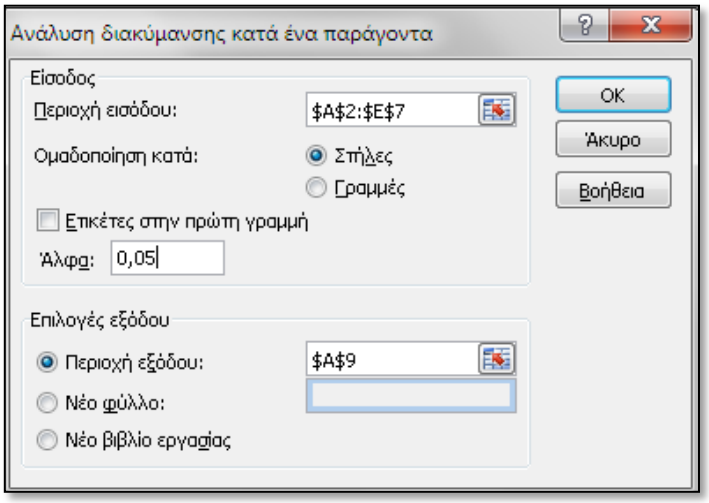
Σχήμα 8.1. Διευθέτηση δειγμάτων και πίνακας αποτελεσμάτων

Όπως παρατηρούμε στον πίνακα αποτελεσμάτων (σχήμα 8.1), το πρόγραμμα εκτελεί όλους τους δυνατούς ελέγχους. Από αυτούς προκύπτει ότι όλοι οι έλεγχοι κανονικότητας οδηγούν σε τιμές p-value > 0.05 και το ίδιο ισχύει για τον έλεγχο της ομοιογένειας της διασποράς (p-value =

0.584). Συνεπώς οι προϋποθέσεις εφαρμογής της μεθόδου φαίνεται να υπάρχουν. Η τιμή p-value = 4.446×10^{-7} δείχνει ότι υπάρχουν στατιστικά σημαντικές διαφορές μεταξύ των δειγμάτων, δεδομένου ότι οι διαφορές των μέσων τιμών των δειγμάτων έχουν πρακτικά μηδαμινή πιθανότητα να οφείλονται σε τυχαίους παράγοντες. Τέλος, από τους ανά δύο ελέγχους με τη διόρθωση *Holm-Bonferroni* προκύπτει ότι υπάρχουν στατιστικά σημαντικές διαφορές μεταξύ των δειγμάτων 1-2, 1-3, 1-4, 1-5, 2-3, 3-4 και 3-5.

❖ Ανάλυση στο Excel

Στο *Excel* μπορούμε να εφαρμόσουμε *ANOVA* μόνο αν όλα τα δείγματα έχουν το ίδιο πλήθος τιμών, όπως στο παράδειγμα που εξετάζουμε. Υποθέτοντας ότι ισχύουν οι προϋποθέσεις εφαρμογής της *ANOVA* διατάσουμε τα δεδομένα όπως στο σχήμα 8.1 και ακολούθως πηγαίνουμε *Δεδομένα* (*Data*) → *Ανάλυση (Analysis)* → *Ανάλυση δεδομένων* (*Data Analysis*) → *Ανάλυση διακύμανσης κατά ένα παράγοντα* (*Anova: Single Factor*) και συμπληρώνουμε το παράθυρο που ανοίγει όπως φαίνεται στο σχήμα 8.2, με την προϋπόθεση ότι τα δείγματα τοποθετούνται στις στήλες από Α έως Ε και από τη γραμμή 2 έως την 7. Τα αποτελέσματα που παίρνουμε δίνονται στο σχήμα 8.3.



Σχήμα 8.2. Εισαγωγή δεδομένων στο *Excel* για την εφαρμογή *μονοπαραγοντικής ανάλυσης διασποράς*

Παρατηρούμε ότι το *Excel* υπολογίζει την τιμή p-value (τιμή-P) = 4.45×10^{-7} που ταυτίζεται με την αντίστοιχη του *ChemStat*, όμως δεν εκτελεί πολλαπλούς ελέγχους. Συνεπώς δεν διευκρινίζεται μεταξύ ποιών δειγμάτων εντοπίζονται στατιστικά σημαντικές διαφορές. Σε αυτή την περίπτωση, αν δεν υπάρχει άλλη εναλλακτική λύση, μπορούμε να προχωρήσουμε σε ανά δύο ελέγχους με το *Έλεγχος t δύο δειγμάτων με υποτιθέμενες ίσες διακυμάνσεις* (*t-Test:Two-Sample Assuming Equal Variances*), όπως στο παράδειγμα 7.1, και να χρησιμοποιήσουμε τη διόρθωση *Holm-Bonferroni* στις τιμές p-value.

Ανάλυση διακύμανσης κατά ένα παράγοντα

ΣΥΜΠΕΡΑΣΜΑ

Ομάδες	*Πλήθος*	*Άθροισμα*	*Μέσος όρος*	*Διακύμανση*
Στήλη 1	6	573	95,5	15,1
Στήλη 2	6	527	87,833	4,166667
Στήλη 3	6	486	81	8,4
Στήλη 4	6	515	85,833	8,166667
Στήλη 5	6	536	89,333	10,66667

ΑΝΑΛΥΣΗ ΔΙΑΚΥΜΑΝΣΗΣ

Προέλευση διακύμανσης	*SS*	*βαθμοί ελευθερίας*	*MS*	*F*	*τιμή-P*	*κριτήριο F*
Μεταξύ ομάδων	670,2	4	167,55	18,01613	4,45E-07	2,75871
Μέσα στις ομάδες	232,5	25	9,3			
Σύνολο	902,7	29				

Σχήμα 8.3. Πίνακας αποτελεσμάτων *μονοπαραγοντικής ανάλυσης διασποράς* στο *Excel*

❖ Ανάλυση στο SPSS

Αν χρησιμοποιήσουμε το *SPSS*, πρέπει να ελέγξουμε πρώτα την κανονικότητα των δειγμάτων. Έτσι, πρώτα εισάγουμε τα δεδομένα σε ένα φύλλο του *SPSS* όπως φαίνεται στο σχήμα 8.4. Δηλαδή όλα τα δείγματα εισάγονται σε μια στήλη, που την ονομάζουμε samples, ενώ σε μια άλλη στήλη, την groups, ένας δείκτης με τιμές από 1 έως 5 δηλώνει σε ποιο δείγμα αντιστοιχεί κάθε τιμή της στήλης των δειγμάτων.

Ακολούθως ελέγχουμε την κανονικότητα από *Analyze* → *Descriptive Statistics* → *Explore*. Στο παράθυρο διαλόγου που ανοίγει εισάγουμε τη μεταβλητή samples στο πλαίσιο *Dependent List*, τη μεταβλητή groups στο πλαίσιο *Factor List*, στο πάνελ *Display* επιλέγουμε *Plots* και κάνουμε *κλικ* στο κουμπί *Plots*. Στο νέο παράθυρο που ανοίγει επιλέγουμε μόνο το

Normality plots with tests και ολοκληρώνουμε κάνοντας *κλικ* στο *Continue* και μετά στο *OK*. Στο παράθυρο των αποτελεσμάτων παίρνουμε τον πίνακα του σχήματος 8.5, όπου παρατηρούμε ότι και τα δύο κριτήρια, *Kolmogorov-Smirnov* και *Shapiro-Wilk*, δίνουν τιμές Sig. > 0.05.

	samples	groups
1	100	1
2	95	1
3	98	1
4	90	1
5	92	1
6	98	1
7	85	2
8	88	2
9	86	2
10	90	2
11	88	2
12	90	2
13	82	3
14	80	3
15	77	3
16	85	3

	samples	groups
15	77	3
16	85	3
17	83	3
18	79	3
19	90	4
20	85	4
21	88	4
22	82	4
23	86	4
24	84	4
25	90	5
26	95	5
27	89	5
28	85	5
29	88	5
30	89	5

Σχήμα 8.4. Διευθέτηση δειγμάτων για έλεγχο κανονικότητας και για εφαρμογή μονοπαραγοντικής *ANOVA* στο *SPSS*

Tests of Normality

		Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	groups	Statistic	df	Sig.	Statistic	df	Sig.
samples	1	,240	6	,200*	,930	6	,579
	2	,199	6	,200*	,903	6	,393
	3	,135	6	,200*	,988	6	,985
	4	,143	6	,200*	,989	6	,987
	5	,252	6	,200*	,916	6	,480

a. Lilliefors Significance Correction

*. This is a lower bound of the true significance.

Σχήμα 8.5. Αποτελέσματα ελέγχου κανονικότητας των δειγμάτων

Για να εφαρμόσουμε τώρα τη μέθοδο της *Ανάλυσης Διασποράς* πηγαίνουμε *Analyze* → *Compare Means* → *One-way ANOVA*, ανοίγει το

παράθυρο *One-Way ANOVA* και μεταφέρουμε τη στήλη των δειγμάτων, samples, στο πλαίσιο *Dependent List*, τη μεταβλητή groups στο *Factor* και από το *Options* επιλέγουμε το *Homogeneity of variance test* για να ελέγξουμε την ομοιογένεια της διασποράς. Τέλος, για να διευκρινίσουμε μεταξύ ποιών δειγμάτων εντοπίζονται στατιστικά σημαντικές διαφορές, κάνουμε *κλικ* στο κουμπί *Post Hoc* και στο αντίστοιχο παράθυρο διαλόγου επιλέγουμε τον έλεγχο πολλαπλών συγκρίσεων του *Tukey*. Ολοκληρώνουμε με *κλικ* στο *Continue* και μετά στο *OK*.

Στο παράθυρο των αποτελεσμάτων ενδιαφέρον παρουσιάζουν οι πίνακες των σχημάτων 8.6 – 8.8. Ο πρώτος πίνακας του σχήματος 8.6 αφορά την ομοιογένεια της διασποράς και η τιμή, Sig. = 0.584 > 0.05, δείχνει ότι η μηδενική υπόθεση, H_0: $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2$, ότι δηλαδή τα δείγματα προέρχονται από πληθυσμούς με την ίδια διασπορά, δεν μπορεί να απορριφθεί.

Ο επόμενος πίνακας του σχήματος 8.6 είναι ο κύριος πίνακας αποτελεσμάτων της μεθόδου στο *SPSS*. Η τιμή Sig. < 0.001 δείχνει ότι υπάρχουν στατιστικά σημαντικές διαφορές μεταξύ των μέσων τιμών των δειγμάτων.

Test of Homogeneity of Variances

samples

Levene Statistic	df1	df2	Sig.
,723	4	25	,584

ANOVA

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	670,200	4	167,550	18,016	,000
Within Groups	232,500	25	9,300		
Total	902,700	29			

Σχήμα 8.6. Πίνακες αποτελεσμάτων *μονοπαραγοντικής ανάλυσης διασποράς* στο *SPSS*

Τέλος, από τους πίνακες με τους πολλαπλούς ελέγχους, σχήματα 8.7-8.8, παρατηρούμε ότι με βάση το κριτήριο του *Tukey* στατιστικά σημαντικές διαφορές υπάρχουν μεταξύ των δειγμάτων 1-2, 1-3, 1-4, 1-5, 2-3 και 3-5. Σε αντίθεση με τον έλεγχο *Holm-Bonferroni* στο *ChemStat* δεν υπάρχει στατιστικά σημαντική διαφορά στα δείγματα 3-4.

Multiple Comparisons

samples
Tukey HSD

(I) groups	(J) groups	Mean Difference (I-J)	Std. Error	Sig.	95% CI Lower Bound	95% CI Upper Bound
1	2	7,667*	1,761	,002	2,50	12,84
	3	14,500*	1,761	,000	9,33	19,67
	4	9,667*	1,761	,000	4,50	14,84
	5	6,167*	1,761	,014	1,00	11,34
2	1	-7,667*	1,761	,002	-12,84	-2,50
	3	6,833*	1,761	,006	1,66	12,00
	4	2,000	1,761	,786	-3,17	7,17
	5	-1,500	1,761	,911	-6,67	3,67
3	1	-14,500*	1,761	,000	-19,67	-9,33
	2	-6,833*	1,761	,006	-12,00	-1,66
	4	-4,833	1,761	,075	-10,00	,34
	5	-8,333*	1,761	,001	-13,50	-3,16
4	1	-9,667*	1,761	,000	-14,84	-4,50
	2	-2,000	1,761	,786	-7,17	3,17
	3	4,833	1,761	,075	-,34	10,00
	5	-3,500	1,761	,301	-8,67	1,67
5	1	-6,167*	1,761	,014	-11,34	-1,00
	2	1,500	1,761	,911	-3,67	6,67
	3	8,333*	1,761	,001	3,16	13,50
	4	3,500	1,761	,301	-1,67	8,67

*. The mean difference is significant at the 0.05 level.

Σχήμα 8.7. Αποτελέσματα ελέγχων πολλαπλών συγκρίσεων με το κριτήριο *Tukey*

samples

Tukey HSD[a]

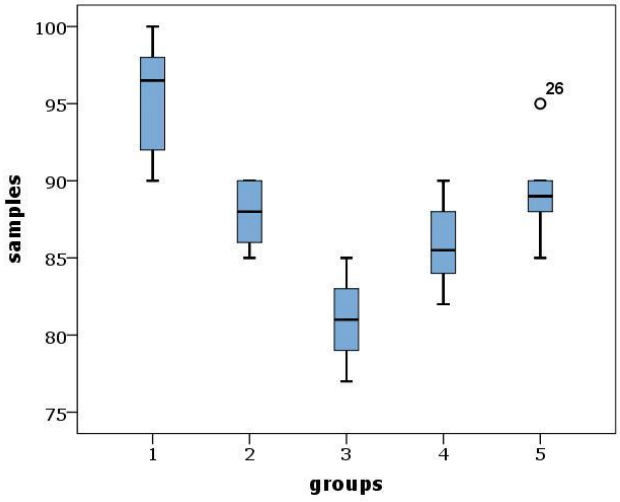
groups	N	Subset for alpha = 0.05 1	2	3
3	6	81,00		
4	6	85,83	85,83	
2	6		87,83	
5	6		89,33	
1	6			95,50
Sig.		,075	,301	1,000

Σχήμα 8.8. Πίνακας ομογενών υποσυνόλων

Στον πίνακα του σχήματος 8.8 βλέπουμε ότι τα δείγματα χωρίζονται σε τρεις κατηγορίες με παρόμοιες συμπεριφορές. Η μία περιλαμβάνει τα

δείγματα 3, 4, η άλλη τα δείγματα 2, 4, 5 και η τρίτη μόνο τα δείγμα 1. Λαμβάνοντας υπόψη τις μέσες τιμές των δειγμάτων, γίνεται φανερό ότι στατιστικά καλύτερη συμπεριφορά δείχνουν οι καταλύτες 3 και 4 και τη χειρότερη ο 1.

Τα συμπεράσματα αυτά οπτικοποιούνται στα θηκογράμματα του σχήματος 8.9.



Σχήμα 8.9. Θηκογράμματα παραδείγματος 8.1

8.3 ΔΙΠΑΡΑΓΟΝΤΙΚΗ ΑΝΑΛΥΣΗ ΔΙΑΣΠΟΡΑΣ

Όπως έχει αναφερθεί, η *Δι-παραγοντική Ανάλυση Διασποράς* (*Two-way ANOVA*) διακρίνεται σε δύο υποπεριπτώσεις: την ανάλυση χωρίς αλληλεπιδράσεις/επαναλήψεις και την ανάλυση με αλληλεπιδράσεις/ επαναλήψεις. Και στις δύο περιπτώσεις έχουμε ένα σύνολο τιμών x_{ik} που επηρεάζονται από δύο παράγοντες. Για παράδειγμα, η σταθερότητα και συνεπώς η συγκέντρωση ενός αντιδραστηρίου μπορεί να εξαρτάται από το χρόνο και τη θερμοκρασία.

Για την ανάλυση διασποράς με αλληλεπιδράσεις/επαναλήψεις απαιτείται να υπάρχουν σε κάθε τιμή του πρώτου παράγοντα περισσότερες από μια τιμές του δεύτερου παράγοντα. Αν δεν υπάρχουν επαναλήψεις τιμών στους δύο παράγοντες, έχουμε την ανάλυση διασποράς χωρίς αλληλεπιδράσεις/επαναλήψεις.

Γενικά ο πίνακας δεδομένων έχει την παρακάτω μορφή.

Δεύτερος παράγοντας	Πρώτος παράγοντας				
	1	2	. . .	m	
1	x_{11}	x_{12}	. . .	x_{1m}	$\bar{x}_{1r}$
2	x_{21}	x_{22}	. . .	x_{2m}	$\bar{x}_{2r}$
. . .			. . .		. . .
n	x_{n1}	x_{n2}	. . .	x_{nm}	$\bar{x}_{nr}$
	$\bar{x}_{1c}$	$\bar{x}_{2c}$		$\bar{x}_{nc}$	

Στα δείγματα αυτά και εφόσον δεν υπάρχουν επαναλήψεις τιμών στους δύο παράγοντες μπορούμε να ορίσουμε τις ακόλουθες διασπορές:

α) ***Μεταξύ γραμμών***: $$s_r^2 = \frac{m}{n-1}\sum_{i=1}^{n}(\bar{x}_{ir} - \bar{x})^2 \qquad (8.5)$$

β) ***Μεταξύ στηλών***: $$s_c^2 = \frac{n}{m-1}\sum_{k=1}^{m}(\bar{x}_{kc} - \bar{x})^2 \qquad (8.6)$$

γ) ***Τυχαία***: $$s_e^2 = \frac{1}{(n-1)(m-1)}\sum_{i,k}(x_{ik} - \bar{x}_{kc} - \bar{x}_{ic} + \bar{x})^2 \qquad (8.7)$$

και τις στατιστικές συναρτήσεις

$$F_r = \frac{s_r^2}{s_e^2} \quad \text{και} \quad F_c = \frac{s_c^2}{s_e^2} \qquad (8.8)$$

Έτσι, στην *Ανάλυση Διασποράς χωρίς Αλληλεπιδράσεις/Επαναλήψεις* ελέγχουμε δύο μηδενικές υποθέσεις:

$H_0^{(1)}$: Οι μέσες τιμές των πληθυσμών από τους οποίους προέρχονται οι τιμές των γραμμών είναι στατιστικά ίσες

$H_0^{(2)}$: Οι μέσες τιμές των πληθυσμών από τους οποίους προέρχονται οι τιμές των στηλών είναι στατιστικά ίσες

Στην *Ανάλυση Διασποράς με Αλληλεπιδράσεις/Επαναλήψεις*, ελέγχουμε τρεις μηδενικές υποθέσεις:

$H_0^{(1)}$: Οι μέσες τιμές των πληθυσμών από τους οποίους προέρχονται οι τιμές των γραμμών είναι στατιστικά ίσες

$H_0^{(2)}$: Οι μέσες τιμές των πληθυσμών από τους οποίους προέρχονται οι τιμές των στηλών είναι στατιστικά ίσες

$H_0^{(3)}$: Δεν υπάρχει αλληλεπίδραση μεταξύ των παραγόντων

όπου για τον έλεγχό τους χρησιμοποιούνται στατιστικές συναρτήσεις ανάλογες των σχέσεων (8.8).

Για την εφαρμογή της *δι-παραγοντικής ανάλυσης διασποράς* απαιτείται τα δεδομένα να ακολουθούν έστω και προσεγγιστικά την κανονική κατανομή σε κάθε *κυψέλη* που σχηματίζεται από τους δύο παράγοντες. Η έννοια της *κυψέλης* θα δοθεί στο παρακάτω παράδειγμα. Επίσης θα πρέπει σε κάθε κυψέλη να υπάρχει ομοιογένεια της διασποράς. Όμως επειδή οι κυψέλες περιέχουν κατά κανόνα έναν πολύ μικρό αριθμό τιμών και στην περίπτωση της δι-παραγοντικής *ANOVA* χωρίς αλληλεπιδράσεις μόνο μία τιμή, οι έλεγχοι αυτοί ή δεν γίνονται ή δεν είναι ιδιαίτερα αξιόπιστοι. Έτσι η κύρια πρόνοια που λαμβάνεται υπόψη είναι να μην υπάρχουν στα δείγματα πολύ ακραίες τιμές.

Παράδειγμα 8.2

Έστω ότι μελετάμε την επίδραση της θερμοκρασίας και του pH στην ανάπτυξη ενός βακτηρίου σε φιάλες που περιέχουν γλυκόζη ως υπόστρωμα. Η μελέτη της ανάπτυξης του βακτηρίου γίνεται με μετρήσεις οπτικής πυκνότητας και τα αποτελέσματα που ελήφθησαν δίνονται στον πίνακα 8.2. Να εξετασθεί πόσο στατιστικά σημαντική είναι η επίδραση της θερμοκρασίας και του pH στην ανάπτυξη του βακτηρίου.

Πίνακας 8.2. Δεδομένα παραδείγματος 8.2.

T / ^{0}C	pH = 5	pH = 6	pH = 7
25	10	20	40
30	15	25	45
35	20	30	55
40	15	22	40

◆ Πριν προχωρήσουμε στην εφαρμογή της διπαραγοντικής *ANOVA* θα ορίσουμε την *κυψέλη* σε έναν πίνακα τιμών που επηρεάζονται από δύο παράγοντες. Στη μονοπαραγοντική *ANOVA* δημιουργούμε τη μεταβλητή-διάνυσμα που περιέχει όλα τα δείγματα και τη μεταβλητή-παράγοντας που

ξεχωρίζει τα διάφορα δείγματα στη μεταβλητή-διάνυσμα. Για παράδειγμα, στο σχήμα 8.1 η μεταβλητή samples είναι μια μεταβλητή-διάνυσμα που περιέχει όλα τα δείγματα, τα οποία διακρίνονται μεταξύ τους λόγω της μεταβλητής groups που είναι η μεταβλητή-παράγοντας.

Στη διπαραγοντική *ANOVA* τα δείγματα στη μεταβλητή-διάνυσμα εξαρτώνται από δύο παράγοντες. Οι δύο αυτοί παράγοντες μπορούν να ενοποιηθούν σε έναν μέσω της διαδικασίας που φαίνεται σχηματικά στον πίνακα 8.3. Από τη διαδικασία αυτή προκύπτει ότι αν ο παράγοντας 2 δεν παρουσιάζει επαναλήψεις, όπως στη διπαραγοντική ανάλυση διασποράς χωρίς επαναλήψεις, τότε ο ενιαίος παράγοντας, δηλαδή η μεταβλητή-παράγοντας περιέχει τους αριθμούς 1, 2, 3, 4, … και συνεπώς δεν ξεχωρίζει δείγματα ή έστω τμήματα δειγμάτων στη μεταβλητή-διάνυσμα. Αν όμως ο παράγοντας 2 παρουσιάζει τρεις επαναλήψεις, τότε η μεταβλητή-παράγοντας περιέχει τους αριθμούς 1, 1, 1, 2, 2, 2, 3, 3, 3, 4, 4, 4, … και αρχίζει να ξεχωρίζει τμήματα δειγμάτων των τριών τιμών στη μεταβλητή-διάνυσμα. Τα τμήματα αυτά ονομάζονται *κυψέλες*.

Πίνακας 8.3. Ενοποίηση παραγόντων, Π1 και Π2, όταν ο Παράγοντας 2 δεν παρουσιάζει επαναλήψεις (αριστερά) και παρουσιάζει τρεις επαναλήψεις (δεξιά).

Π1	Π2	Ενοποίηση	Μεταβλητή-παράγοντας	Π1	Π2	Ενοποίηση	Μεταβλητή-παράγοντας
1	1	(1,1)	1	1	1	(1,1)	1
1	2	(1,2)	2	1	1	(1,1)	1
2	1	(2,1)	3	1	1	(1,1)	1
2	2	(2,2)	4	1	2	(1,2)	2
				1	2	(1,2)	2
				1	2	(1,2)	2
				2	1	(2,1)	3
				2	1	(2,1)	3
				2	1	(2,1)	3
				2	2	(2,2)	4
				2	2	(2,2)	4
				2	2	(2,2)	4

Είναι προφανές ότι στη διπαραγοντική ανάλυση διασποράς χωρίς επαναλήψεις οι κυψέλες αποτελούνται από μία μόνο τιμή και συνεπώς σε αυτή την περίπτωση δεν μπορεί να εφαρμοστεί ούτε ο έλεγχος της κανονικότητας ούτε της ομοιογένειας της διασποράς. Οι έλεγχοι αυτοί

μπορεί να γίνουν στη διπαραγοντική ανάλυση διασποράς με επαναλήψεις, αλλά για να είναι αξιόπιστοι θα πρέπει ο αριθμός των επαναλήψεων να είναι μεγάλος.

❖ Ανάλυση στο ChemStat

Για να εφαρμόσουμε διπαραγοντική *ANOVA* χωρίς επαναλήψεις με το *ChemStat* διευθετούμε τα δεδομένα όπως στο σχήμα 8.10 και πηγαίνουμε

Πρόσθετα → *ChemStat* → *ANOVA* → *ANOVA parametric*

Στο πρώτο παράθυρο που ανοίγει πληκτρολογούμε τον αριθμό 2 για να επιλεγεί η διπαραγοντική *ANOVA*, στο δεύτερο παράθυρο που αφορά στη διευθέτηση των δειγμάτων πατάμε *OK*, στο τρίτο εισάγουμε με το ποντίκι την περιοχή B2:D5 (και OXI A2:D5) των τιμών των δειγμάτων και στο τελευταίο παράθυρο εισάγουμε το κελί εξόδου των αποτελεσμάτων, έστω F1.

	A	B	C	D	E	F	G	H	I	J
1	**T / ^{0}C**	**pH = 5**	**pH = 6**	**pH = 7**		**TWO-WAY ANOVA without interactions**				
2	**25**	10	20	40						
3	**30**	15	25	45		**F(columns) test for ANOVA**				
4	**35**	20	30	55		p-value=	3,1E-06	Null hypothesis among colum		
5	**40**	15	22	40						
6						**F(rows) test for ANOVA**				
7						p-value=	0,00267	Null hypothesis among rows		
8										
9						variances, s(error), s(rows), s(columns)				
10						4,6389	76,3056	944,083		
11						Factors effect may be assumed additive				
12										
13						**Pairwise comparisons among Columns**				
14						pair	p-uncorre	HB correction		
15						pair 12	0,0009	0,0009		
16						pair 13	1,1E-06	3,3E-06		
17						pair 23	9,7E-06	1,9E-05		
18										
19						**Comparisons among Rows**				
20						pair 12	0,02944	0,08833		
21						pair 13	0,00057	0,00339		
22						pair 14	0,23282	0,36042		
23						pair 23	0,00906	0,03625		
24						pair 24	0,18021	0,36042		
25						pair 34	0,00182	0,00909		

Σχήμα 8.10. Διευθέτηση δειγμάτων και πίνακας αποτελεσμάτων

Στον πίνακα των αποτελεσμάτων παρατηρούμε ότι το πρόγραμμα εκτελεί μόνο τους βασικούς ελέγχους της διπαραγοντικής *ANOVA*. Παρατηρούμε επίσης ότι σε επίπεδο σημαντικότητας α = 0.05 η επίδραση τόσο της θερμοκρασίας (p-value = 0.0027) όσο και του pH (p-value = 3×10^{-6}) είναι στατιστικά σημαντικές. Πιο συγκεκριμένα, από τους πολλαπλούς ελέγχους προκύπτει ότι στατιστικά σημαντικές διαφορές παρατηρούνται ανάμεσα στη επίδραση της θερμοκρασίας των 35 ^{0}C και τις υπόλοιπες, ενώ στατιστικά σημαντικές διαφορές παρουσιάζονται και στα τρία δείγματα του pH.

Στη γραμμή 10 του σχήματος 8.10 δίνονται οι τιμές των διασπορών s_e^2 = 4.64, s_r^2 = 76.31 και s_c^2 = 944.08. Παρατηρούμε ότι τιμή της s_e^2 = 4.6 είναι πολύ μικρή σε σύγκριση με τις τιμές 76.3 και 944.1 των s_r^2 και s_c^2, αντίστοιχα. Αυτό ουσιαστικά δείχνει ότι δεν υπάρχει αλληλεπίδραση μεταξύ της θερμοκρασίας και του pH και συνεπώς δικαιολογείται η εφαρμογή της απλής διπαραγοντικής *ANOVA*. Αν υπήρχε αλληλεπίδραση μεταξύ Τ και pH θα έπρεπε να επανασχεδιάσουμε το πείραμα ώστε να εφαρμόσουμε δι-παραγοντική *ANOVA* με επαναλήψεις. Σε αυτή την περίπτωση θα πρέπει να πάρουμε δύο ή περισσότερες μετρήσεις σε κάθε θερμοκρασία ή pH, όπως φαίνεται στο επόμενο παράδειγμα.

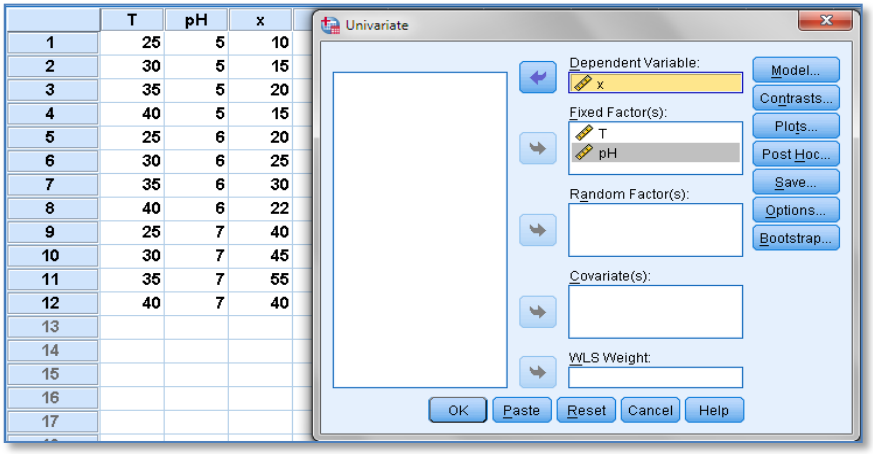
❖ Ανάλυση στο Excel

Στο *Excel* πηγαίνουμε *Δεδομένα* (*Data*) → *Ανάλυση (Analysis)* → *Ανάλυση δεδομένων* (*Data Analysis)* → *Ανάλυση διακύμανσης δύο παραγόντων χωρίς αλληλεπιδράσεις* (*Anova:Two-Factor Without Replication*) και στο παράθυρο που ανοίγει εισάγουμε <u>όλη την περιοχή</u> B2:D5 στην *Περιοχή εισόδου* (*Input Range*), εφόσον έχουμε διευθετήσει τα δεδομένα όπως στο σχήμα 8.10. Στον πίνακα αποτελεσμάτων παίρνουμε και πάλι τις τιμές p-value = 0.0027 για την επίδραση της θερμοκρασίας και p-value = 3.07×10^{-6} για την επίδραση του pH.

Όμως, όπως σε όλους τους σχετικούς ελέγχους με το *Excel*, δεν υπάρχουν πολλαπλές συγκρίσεις. Απουσία άλλης εναλλακτικής λύσης, μπορούμε να προχωρήσουμε και εδώ σε ελέγχους ανά δύο δειγμάτων χρησιμοποιώντας το *Έλεγχος t δύο δειγμάτων με υποτιθέμενες ίσες διακυμάνσεις* (*t-Test: Two-Sample Assuming Equal Variances*) και διόρθωση *Holm-Bonferroni* στις τιμές p-value.

❖ Ανάλυση στο SPSS

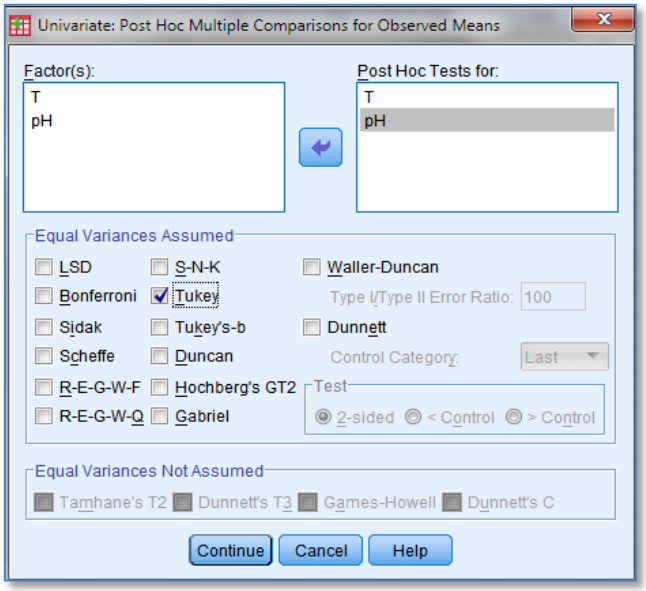
Για να εφαρμόσουμε τη διπαραγοντική *ANOVA* στο *SPSS* διευθετούμε τα δεδομένα όπως στο σχήμα 8.11 και πηγαίνουμε *Analyze* → *General Linear Model* → *Univariate*. Στο παράθυρο που ανοίγει μεταφέρουμε τις μεταβλητές x, T και pH στα πλαίσια *Dependent Variable* και *Fixed Factor(s)* όπως φαίνεται στο σχήμα 8.11 και κάνουμε *κλικ* στο *Model*. Στο παράθυρο διαλόγου που ανοίγει κάνουμε τις ακόλουθες ενέργειες: Επιλέγουμε *Custom*, στο πλαίσιο *Built Term(s)* επιλέγουμε *Main effects*, μεταφέρουμε τις μεταβλητές T και pH στο πλαίσιο *Model* και ενεργοποιούμε την επιλογή *Include intercept in the model*. Τέλος, για πολλαπλές συγκρίσεις μπορούμε να κάνουμε *κλικ* στο κουμπί *Post Hoc* και να επιλέξουμε το κριτήριο *Tukey* αφού μεταφέρουμε τις μεταβλητές T και pH στο πλαίσιο *Post Hoc Tests for* (σχήμα 8.12).



Σχήμα 8.11. Εισαγωγή δεδομένων στο *SPSS* για εφαρμογή *διπαραγοντικής ανάλυσης διασποράς*

Ο βασικός πίνακας αποτελεσμάτων, πίνακας *Tests of Between-Subjects Effects* (σχήμα 8.13), δείχνει ότι και ο αντίστοιχος πίνακας του *ChemStat*: Η τιμή Sig. = 0.003 σημαίνει ότι η πιθανότητα να οφείλονται σε τυχαίους παράγοντες οι διαφορές που προκαλούνται από την επίδραση της θερμοκρασίας είναι μικρότερη από 0.3%, ενώ για το pH η πιθανότητα αυτή είναι ουσιαστικά μηδενική, Sig. < 0.001. Επίσης, όπως ήδη έχει αναφερθεί, η πολύ μικρή τιμή της τυχαίας διασποράς s_e^2 = 4.64 σε

σύγκριση με τις τιμές s_r^2 = 76.31 και s_c^2 = 944.08 δείχνει ότι δεν υπάρχει αλληλεπίδραση μεταξύ της θερμοκρασίας και του pH.



Σχήμα 8.12. Επιλογή ελέγχων πολλαπλών συγκρίσεων

Tests of Between-Subjects Effects

Dependent Variable:x

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	2117,083[a]	5	423,417	91,275	,000
Intercept	9464,083	1	9464,08	2040,162	,000
T	228,917	3	76,306	16,449	,003
pH	1888,167	2	944,083	203,515	,000
Error	27,833	6	4,639		
Total	11609,000	12			
Corrected Total	2144,917	11			

a. R Squared = ,987 (Adjusted R Squared = ,976)

Σχήμα 8.13. Πίνακες αποτελεσμάτων *διπαραγοντικής ανάλυσης διασποράς χωρίς αλληλεπιδράσεις*

Τέλος, στους πίνακες ελέγχων πολλαπλών συγκρίσεων με το

κριτήριο *Tukey* (σχήμα 8.14) παρατηρούμε ότι, σε πλήρη συμφωνία με το *ChemStat*, στατιστικά σημαντικές διαφορές παρατηρούνται ανάμεσα στη επίδραση της θερμοκρασίας των 35 ^{0}C και τις υπόλοιπες, ενώ στατιστικά σημαντικές διαφορές παρουσιάζονται και στα τρία δείγματα του pH.

Όπως πάντα μια καλή εποπτική εικόνα των διαφορών παίρνουμε από τα θηκογράμματα που δίνονται στο σχήμα 8.15.

Multiple Comparisons

x
Tukey HSD

(I) T	(J) T	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interv. Lower Bound	95% Confidence Interv. Upper Bound
25	30	-5,00	1,759	,104	-11,09	1,09
	35	-11,67*	1,759	,002	-17,75	-5,58
	40	-2,33	1,759	,581	-8,42	3,75
30	25	5,00	1,759	,104	-1,09	11,09
	35	-6,67*	1,759	,034	-12,75	-,58
	40	2,67	1,759	,484	-3,42	8,75
35	25	11,67*	1,759	,002	5,58	17,75
	30	6,67*	1,759	,034	,58	12,75
	40	9,33*	1,759	,007	3,25	15,42
40	25	2,33	1,759	,581	-3,75	8,42
	30	-2,67	1,759	,484	-8,75	3,42
	35	-9,33*	1,759	,007	-15,42	-3,25

Based on observed means.
The error term is Mean Square(Error) = 4,639.
*. The mean difference is significant at the 0,05 level.

x

Tukey HSD[a,b]

T	N	Subset 1	Subset 2
25	3	23,33	
40	3	25,67	
30	3	28,33	
35	3		35,00
Sig.		,104	1,000

Means for groups in homogeneous subsets are displayed.
Based on observed means.
The error term is Mean Square(Error) = 4,639.
a. Uses Harmonic Mean Sample Size = 3,000.
b. Alpha = 0,05.

Σχήμα 8.14. Αποτελέσματα ελέγχων πολλαπλών συγκρίσεων με το κριτήριο *Tukey*

Multiple Comparisons

x

Tukey HSD

(I) pH	(J) pH	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
5	6	-9,25*	1,523	,002	-13,92	-4,58
	7	-30,00*	1,523	,000	-34,67	-25,33
6	5	9,25*	1,523	,002	4,58	13,92
	7	-20,75*	1,523	,000	-25,42	-16,08
7	5	30,00*	1,523	,000	25,33	34,67
	6	20,75*	1,523	,000	16,08	25,42

Based on observed means.
The error term is Mean Square(Error) = 4,639.
*. The mean difference is significant at the 0,05 level.

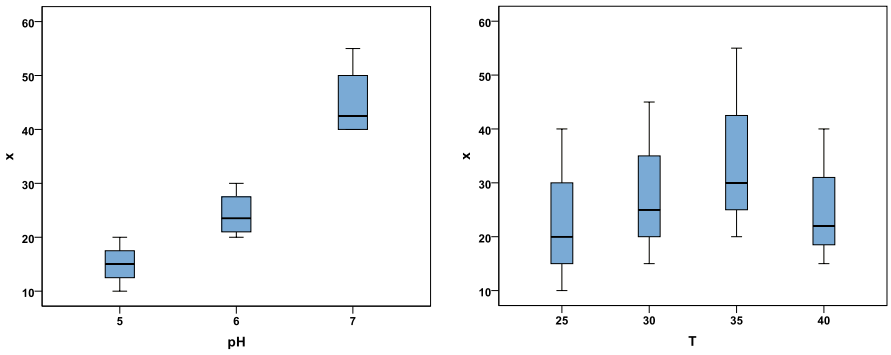
x

Tukey HSD[a,b]

pH	N	Subset 1	Subset 2	Subset 3
5	4	15,00		
6	4		24,25	
7	4			45,00
Sig.		1,000	1,000	1,000

Means for groups in homogeneous subsets are displayed.
Based on observed means.
The error term is Mean Square(Error) = 4,639.
a. Uses Harmonic Mean Sample Size = 4,000.
b. Alpha = 0,05.

Σχήμα 8.14 (συνέχεια). Αποτελέσματα ελέγχων πολλαπλών συγκρίσεων με το κριτήριο *Tukey*



Σχήμα 8.15. Θηκογράμματα επίδρασης του pH (αριστερά) και της θερμοκρασίας (δεξιά)

Παράδειγμα 8.3

Σε ένα πείραμα μελέτης της επίδρασης της θερμοκρασίας και του pH στην εκατοστιαία απόδοση μιας χημικής αντίδρασης ελήφθησαν τα αποτελέσματα που δίνονται στον πίνακα 8.4. Να εξετασθεί αν η επίδραση της θερμοκρασίας και του pH στην απόδοση της αντίδρασης είναι στατιστικά σημαντική.

Πίνακας 8.4. Επίδρασης θερμοκρασίας και pH στην εκατοστιαία απόδοση χημικής αντίδρασης.

T / ^{0}C	pH = 5	pH = 6	pH = 7
25	18	18	22
25	15	20	28
25	15	20	25
30	21	23	23
30	20	27	28
30	22	25	25
35	24	27	20
35	25	33	25
35	22	31	28
40	24	20	24
40	28	24	26
40	25	25	29

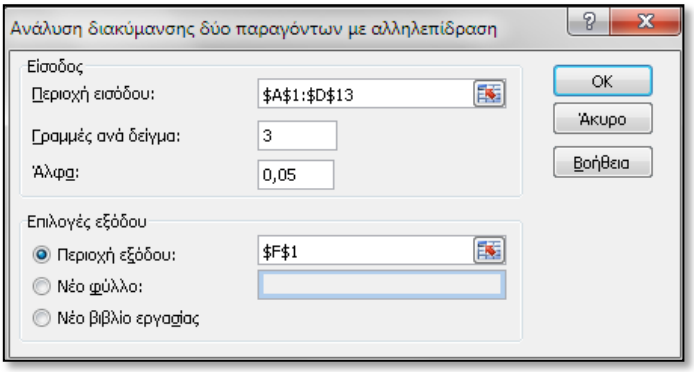
◆ Παρατηρούμε ότι πρόκειται για τυπική περίπτωση δι-παραγοντικής *ANOVA* με επαναλήψεις, δεδομένου ότι υπάρχουν τρεις μετρήσεις σε κάθε θερμοκρασία. Για να την εφαρμόσουμε θα χρησιμοποιήσουμε το *Excel* και το *SPSS*, δεδομένου ότι δεν υπάρχει αντίστοιχο πρόγραμμα στο *ChemStat*.

❖ **Ανάλυση στο Excel**

Υποθέτοντας ότι ισχύουν οι προϋποθέσεις εφαρμογής της δι-παραγοντικής *ANOVA* με επαναλήψεις διατάξουμε τα δεδομένα όπως στο σχήμα 8.16 και ακολούθως πηγαίνουμε *Δεδομένα* (*Data)* → *Ανάλυση (Analysis)* → *Ανάλυση δεδομένων* (*Data Analysis)* → *Ανάλυση διακύμανσης δύο παραγόντων με αλληλεπιδράσεις* (*Anova:Two-Factor With Replication*), όπου συμπληρώνουμε το παράθυρο που ανοίγει όπως φαίνεται στο σχήμα 8.17. Παρατηρούμε ότι στο παράθυρο που ανοίγει εισάγουμε όλη την περιοχή A1:D13 στην *Περιοχή εισόδου* (*Input Range*).

	A	B	C	D
1	**T / ^{0}C**	**pH = 5**	**pH = 6**	**pH = 7**
2	**25**	18	18	22
3	**25**	15	20	28
4	**25**	15	20	25
5	**30**	21	23	23
6	**30**	20	27	28
7	**30**	22	25	25
8	**35**	24	27	20
9	**35**	25	33	25
10	**35**	22	31	28
11	**40**	24	20	24
12	**40**	28	24	26
13	**40**	25	25	29

Σχήμα 8.16. Διευθέτηση δειγμάτων στο *Excel*



Σχήμα 8.17. Εισαγωγή δεδομένων στο *Excel* για την εφαρμογή της *διπαραγοντικής ανάλυσης διασποράς με επαναλήψεις*

Τμήμα του πίνακα των αποτελεσμάτων δίνεται στο σχήμα 8.18. Στον πίνακα αυτό η λέξη *Δείγμα* αναφέρεται στη θερμοκρασία και η λέξη *Στήλες* στο pH. Παρατηρούμε ότι:

α) Η επίδραση της θερμοκρασίας και του pH στην απόδοση της αντίδρασης είναι στατιστικά πολύ σημαντική (p-value = 0.00014 και p-value = 0.0028, αντίστοιχα).

β) Υπάρχει μια πολύ σημαντική αλληλεπίδραση μεταξύ T και pH (p-value = 0.0023). Αυτό σημαίνει ότι η ανταπόκριση της αντίδρασης στο pH εξαρτάται από τη θερμοκρασία και αντίστροφα.

ΑΝΑΛΥΣΗ ΔΙΑΚΥΜΑΝΣΗΣ						
Προέλευση διακύμανσης	*SS*	*βαθμοί ελευθερίας*	*MS*	*F*	*τιμή-P*	*κριτήριο F*
Δείγμα	183,42	3	61,139	10,431	0,00014	3,00879
Στήλες	88,667	2	44,333	7,564	0,00284	3,40283
Αλληλεπίδραση	170	6	28,333	4,8341	0,0023	2,50819
Μέσα σε	140,67	24	5,8611			
Σύνολο	582,75	35				

Σχήμα 8.18. Τμήμα πίνακα αποτελεσμάτων της *διπαραγοντικής ανάλυσης διασποράς με επαναλήψεις* στο *Excel*

❖ Ανάλυση στο SPSS

Για να αναλύσουμε τα δεδομένα με το *SPSS*, τα τοποθετούμε στο φύλλο εργασίας όπως φαίνεται στο σχήμα 8.19 και ακολουθούμε την πορεία *Analyze* → *General Linear Model* → *Univariate*. Στο παράθυρο που ανοίγει μεταφέρουμε τη μεταβλητή x στο πλαίσιο *Dependent Variable* και τις Τ και pH στο *Fixed Factor(s)*. Στο *Options* επιλέγουμε το *Homogeneity tests* και στο *Model* επιλέγουμε *Full Factorial* (σε αντίθεση με την προηγούμενη περίπτωση). Τέλος, στο *Post Hoc* μεταφέρουμε τους δύο παράγοντες, Τ και pH, στο πάνελ *Post Hoc Tests for* και επιλέγουμε τον έλεγχο *Tukey*. Τα βασικά αποτελέσματα που παίρνουμε δίνονται στους πίνακες των σχημάτων 8.20 και 8.21.

Παρατηρούμε ότι ο πίνακας 8.20 ταυτίζεται ουσιαστικά με τον αντίστοιχο πίνακα του *Excel*, ενώ από τους πίνακες με τις πολλαπλές συγκρίσεις προκύπτει ότι στατιστικά σημαντικές διαφορές υπάρχουν μεταξύ των τιμών Τ: 25-30, 25-35, 25-40 και pH: 5-6 και 5-7.

Ένα ενδιαφέρον διάγραμμα με θηκογράμματα μπορεί να γίνει αν ενοποιήσουμε τους παράγοντες Τ και pH σε έναν με τιμές 25-5, 25-5, 25-5, 30-5, 30-5,3 0-5, 35-5, … , 40-7. Στο σχήμα 8.22 παρατηρούμε ότι την καλύτερη απόδοση έχει ο συνδυασμός Τ = 35 και pH = 6. Ακριβώς το ίδιο διάγραμμα γίνεται αν ακολουθήσουμε την πορεία *Graphs* → *Legacy Dialogs* → *Boxplot*. Στο πλαίσιο που ανοίγει επιλέγουμε *Clustered* και *Summaries for group of cases*. Πατάμε *Define* και στο νέο παράθυρο που ανοίγει εισάγουμε τη μεταβλητή x στο πλαίσιο *Variable*, την Τ στο *Category Axis* και την pH στο *Define Clusters by*.

	T	pH	x
1	25	5	18
2	25	5	15
3	25	5	15
4	30	5	21
5	30	5	20
6	30	5	22
7	35	5	24
8	35	5	25
9	35	5	22
10	40	5	24
11	40	5	28
12	40	5	25
13	25	6	18
14	25	6	20
15	25	6	20
16	30	6	23
17	30	6	27
18	30	6	25
19	35	6	27

	T	pH	x
18	30	6	25
19	35	6	27
20	35	6	33
21	35	6	31
22	40	6	20
23	40	6	24
24	40	6	25
25	25	7	22
26	25	7	28
27	25	7	25
28	30	7	23
29	30	7	28
30	30	7	25
31	35	7	20
32	35	7	25
33	35	7	28
34	40	7	24
35	40	7	26
36	40	7	29

Σχήμα 8.19. Διευθέτηση δεδομένων σε φύλλο εργασίας του *SPSS*

Levene's Test of Equality of Error Variances[a]

Dependent Variable:x

F	df1	df2	Sig.
,842	11	24	,603

Tests of Between-Subjects Effects

Dependent Variable:x

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	442,083[a]	11	40,189	6,857	,000
Intercept	20306,250	1	20306,25	3464,573	,000
T	183,417	3	61,139	10,431	,000
pH	88,667	2	44,333	7,564	,003
T * pH	170,000	6	28,333	4,834	,002
Error	140,667	24	5,861		
Total	20889,000	36			
Corrected Total	582,750	35			

a. R Squared = ,759 (Adjusted R Squared = ,648)

Σχήμα 8.20. Πίνακες αποτελεσμάτων της *διπαραγοντικής ανάλυσης διασποράς με αλληλεπιδράσεις*

Multiple Comparisons

x

Tukey HSD

(I) T	(J) T	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
25	30	-3,67*	1,141	,018	-6,81	-,52
	35	-6,00*	1,141	,000	-9,15	-2,85
	40	-4,89*	1,141	,001	-8,04	-1,74
30	25	3,67*	1,141	,018	,52	6,81
	35	-2,33	1,141	,200	-5,48	,81
	40	-1,22	1,141	,710	-4,37	1,93
35	25	6,00*	1,141	,000	2,85	9,15
	30	2,33	1,141	,200	-,81	5,48
	40	1,11	1,141	,766	-2,04	4,26
40	25	4,89*	1,141	,001	1,74	8,04
	30	1,22	1,141	,710	-1,93	4,37
	35	-1,11	1,141	,766	-4,26	2,04

Based on observed means.

The error term is Mean Square(Error) = 5,861.

*. The mean difference is significant at the ,05 level.

x

Tukey HSD[a,b]

T	N	Subset 1	Subset 2
25	9	20,11	
30	9		23,78
40	9		25,00
35	9		26,11
Sig.		1,000	,200

Σχήμα 8.21. Αποτελέσματα ελέγχων πολλαπλών συγκρίσεων με το κριτήριο *Tukey*

Multiple Comparisons

x

Tukey HSD

(I) pH	(J) pH	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
5	6	-2,83*	,988	,022	-5,30	-,37
	7	-3,67*	,988	,003	-6,13	-1,20
6	5	2,83*	,988	,022	,37	5,30
	7	-,83	,988	,680	-3,30	1,63
7	5	3,67*	,988	,003	1,20	6,13
	6	,83	,988	,680	-1,63	3,30

Based on observed means.

The error term is Mean Square(Error) = 5,861.

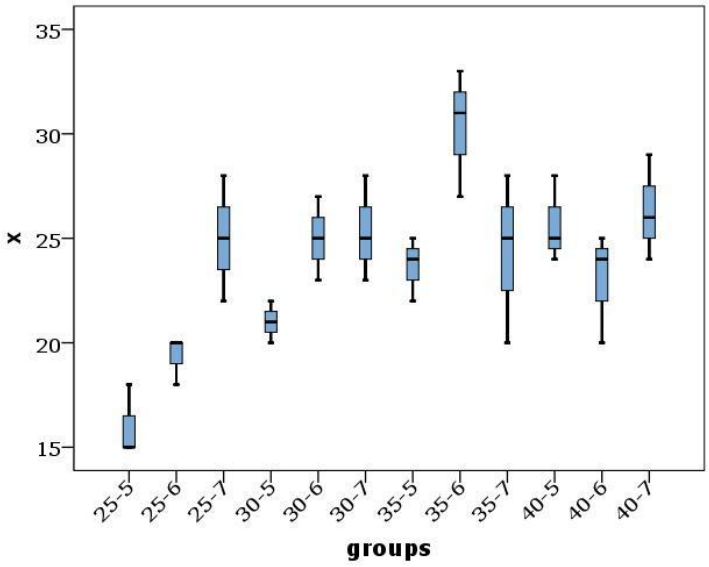
*. The mean difference is significant at the ,05 level.

x

Tukey HSD[a,b]

pH	N	Subset 1	Subset 2
5	12	21,58	
6	12		24,42
7	12		25,25
Sig.		1,000	,680

Σχήμα 8.21 (συνέχεια). Αποτελέσματα ελέγχων πολλαπλών συγκρίσεων με το κριτήριο *Tukey*



Σχήμα 8.22 Θηκογράμματα ως προς τον ενοποιημένο παράγοντα T-pH

8.4 ΜΗ ΠΑΡΑΜΕΤΡΙΚΗ ΑΝΑΛΥΣΗ ΔΙΑΣΠΟΡΑΣ

Αυστηρά η μη παραμετρική *ANOVA* εφαρμόζεται όταν δεν ισχύουν ή όταν έχουμε αμφιβολίες για την ισχύ των προϋποθέσεων εφαρμογής της παραμετρικής *ANOVA*. Ιδιαίτερα όταν τα δείγματα δεν προέρχονται από μετρήσεις στο ίδιο σύστημα κάτω από σταθερές συνθήκες ή/και όταν έχουμε πολύ μικρά δείγματα, τότε είναι απαραίτητο να εφαρμόζουμε τη μη παραμετρική ανάλυση διασποράς είτε αυτόνομα είτε συμπληρωματικά της παραμετρικής *ANOVA*.

8.4.1 ΜΟΝΟΠΑΡΑΓΟΝΤΙΚΗ ΑΝΑΛΥΣΗ - Κριτήριο Kruskal-Wallis

Η μηδενική υπόθεση εκφράζεται ως:

H_0: Όλα τα δείγματα προέρχονται από τον ίδιο πληθυσμό

με εναλλακτική

H_1: Όχι η H_0

Για το έλεγχο της μηδενικής υπόθεσης δημιουργείται ένα ενιαίο δείγμα από τα επιμέρους δείγματα και υπολογίζεται η ποσότητα

$$H = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{R_i^2}{m_i} - 3(n+1) \qquad (8.9)$$

όπου k είναι το πλήθος των δειγμάτων, m_i είναι το μέγεθος του δείγματος i, $n = m_1 + m_2 + \ldots + m_k$ και R_i είναι το άθροισμα των βαθμών του δείγματος i στο ενιαίο δείγμα μεγέθους n.

Αποδεικνύεται ότι η μεταβλητή H ακολουθεί ασυμπτωτικά την κατανομή χ^2 με k-1 βαθμούς ελευθερίας και η ιδιότητα αυτή χρησιμοποιείται για τον υπολογισμό του p-value.

Παρατήρηση. Η σχέση (8.9) ισχύει με την προϋπόθεση ότι δεν υπάρχουν *δεσμοί (ties)* ή είναι πολύ λίγοι. Διαφορετικά απαιτούνται διορθώσεις.

Παράδειγμα 8.4

Στον πίνακα του Σχήματος 8.23 δίνονται οι αέριοι ρύποι σε TSP ($\mu g/m^3$) σε τέσσερες περιοχές, Κοζάνη, Πετρανά, Πολύμυλο και Κλείτο. Να εξεταστεί αν υπάρχουν στατιστικά σημαντικές διαφοροποιήσεις της ρύπανσης σε αυτές τις περιοχές.

	A	B	C	D
1	**KOZANI**	**PETRANA**	**POLYMYLOS**	**KLITOS**
2	115	85	70	241
3	27	38	56	120
4	128	11	114	295
5	67	49	27	168
6	28	47	78	95
7	116	21	36	187
8	76	23	91	322
9	61	13	77	124
10	190	18	121	258
11	102	52	51	164
12	170	67	77	281
13	113	57	38	336
14	130	75	66	171
15	112	58	114	160
16	148	93	135	75
17	122	88	135	208
18	93	8	195	112
19	12	19	52	62
20	140	15	22	
21	132	26	67	
22		49	40	
23		15	123	
24			59	

Σχήμα 8.23. Αέριοι ρύποι σε TSP σε Κοζάνη, Πετρανά, Πολύμυλο, Κλείτο

❖ **Ανάλυση στο ChemStat**

Στο *ChemStat* πηγαίνουμε *Πρόσθετα* → *ChemStat* → *ANOVA* → *ANOVA non parametric*. Στο πρώτο παράθυρο εισάγουμε τον αριθμό 1 για να εφαρμοστεί το κριτήριο *Kruskal-Wallis*, στο επόμενο πατάμε *OK* εφόσον τα δεδομένα έχουν διευθετηθεί όπως στο σχήμα 8.23, στο τρίτο παράθυρο εισάγουμε τον αριθμό 4 που είναι το πλήθος των δειγμάτων και στα επόμενα τέσσερα παράθυρα εισάγουμε διαδοχικά τις τιμές των δειγμάτων. Στο τελευταίο παράθυρο εισάγουμε το κελί εξόδου των αποτελεσμάτων.

Το πρόγραμμα εκτελεί τον έλεγχο *Kruskal-Wallis* και επιπλέον εκτελεί ανά δύο ελέγχους μεταξύ όλων των δειγμάτων με δύο τεχνικές: α) Τις σχέσεις Bewick et al. [Crit. Care 8 (2004) 196] και β) προσεγγιστικά χρησιμοποιώντας τον έλεγχο *Mann-Whitney* με *Holm-Bonferroni* διόρθωση. Από τους ελέγχους αυτούς ισχυρότερος είναι ο πρώτος. Μετά την εκτέλεση, εμφανίζεται το πλαίσιο *Kruskal-Wallis Monte-Carlo test*, στο οποίο ορίζουμε τον αριθμό των επαναλήψεων που θα χρησιμοποιηθούν στη μέθοδο *Monte-Carlo με αντιμεταθέσεις* ή εισάγουμε τον αριθμό 1 αν δεν θέλουμε να εφαρμοστεί αυτή η μέθοδος.

Στον πίνακα των αποτελεσμάτων (σχήμα 8.24) παρατηρούμε ότι υπάρχουν στατιστικά σημαντικές διαφορές μεταξύ των δειγμάτων, δεδομένου ότι η πιθανότητα οι διαφορές αυτές να οφείλονται στην τύχη είναι αμελητέα (p-value=9.67×10^{-9}). Οι στατιστικά σημαντικές διαφορές υπάρχουν μεταξύ όλων των ζευγών με εξαίρεση το ζεύγος 1-3.

Kruskal-Wallis one-way ANOVA - 2 tailed			
H=	40,19976		
p(asymp.)=	9,67E-09	Null hypothesis may be rejected at level 0.05	
Monte-Carlo ite	10000		
p(permut.)=	0	Null hypothesis may be rejected at level 0.05	
Pairwise comparisons		**Approximate comparisons**	
		with Mann-Witney test and Holm-Bonferroni correction	
Pairs	p-values	p-uncorrec	p-HB correction
Pair1- 2	7,46E-07	4,03E-05	0,000202
Pair1- 3	0,056098	0,07952	0,07952
Pair1- 4	0,002116	0,003002	0,006004
Pair2- 3	0,000584	0,00131	0,003931
Pair2- 4	1,02E-12	2,57E-07	1,54E-06
Pair3- 4	1,76E-06	4,65E-05	0,000202
Elapsed time =	0,167 min		

Σχήμα 8.24. Πίνακας αποτελεσμάτων στο *ChemStat*

❖ Ανάλυση στο SPSS

Εισάγουμε τα δεδομένα στο φύλλο εργασίας όπως και στην περίπτωση της παραμετρικής *Ανάλυσης Διασποράς* και ακολουθούμε τη διαδικασία *Analyze → Nonparametric Tests → Legacy Dialogs → K Independent Samples*. Στο παράθυρο που εμφανίζεται μεταφέρουμε τη μεταβλητή των δειγμάτων, έστω samples, στο πλαίσιο *Test Variable List* και τη μεταβλητή που ορίζει τα δείγματα, έστω groups, στο *Grouping Variable*. Επίσης επιλέγουμε το κριτήριο *Kruskal-Wallis* αν αυτό δεν είναι προεπιλεγμένο. Κάνουμε *κλικ* στο *Define Groups* και στο πλαίσιο που εμφανίζεται εισάγουμε την τιμή 1 στο *Minimum* και την τιμή 4 στο *Maximum*, εφόσον χρησιμοποιήσαμε τους ακέραιους 1, 2, 3, 4 για να διακρίνουμε τα δείγματα. Επίσης κάνουμε *κλικ* στο κουμπί *Exact* και επιλέγουμε *Monte Carlo* με 10000 επαναλήψεις *(Number of samples)*. Με *κλικ* στο *Continue* και στο *OK* παίρνουμε τον πίνακα αποτελεσμάτων *Test Statistics* (σχήμα 8.25).

Test Statistics[b,c]

			samples
Chi-Square			40,200
df			3
Asymp. Sig.			,000
Monte Carlo Sig.	Sig.		,000[a]
	99% Confidence Interval	Lower Bound	,000
		Upper Bound	,000

a. Based on 10000 sampled tables with starting seed 2000000.
b. Kruskal Wallis Test
c. Grouping Variable: groups

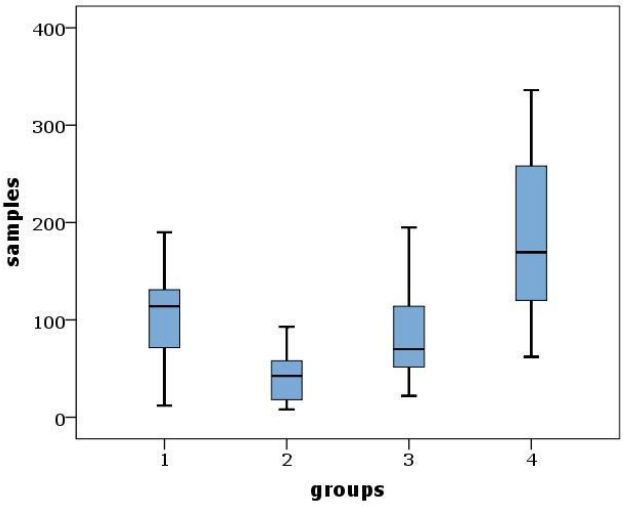
Σχήμα 8.25. Πίνακας αποτελεσμάτων κριτηρίου *Kruskal-Wallis*

Παρατηρούμε ότι υπάρχουν στατιστικά σημαντικές διαφορές μεταξύ των δειγμάτων, δεδομένου ότι η πιθανότητα οι διαφορές αυτές να οφείλονται στην τύχη είναι μικρότερη από 0.001 (Sig. = 0.000). Δυστυχώς δεν γνωρίζουμε μεταξύ ποιων δειγμάτων εντοπίζονται οι διαφορές. Γι αυτό στο *SPSS* προχωρούμε στην κατασκευή των θηκογραμμάτων.

Πηγαίνουμε από *Graphs* και επιλέγουμε *Boxplot*. Στο πλαίσιο που ανοίγει επιλέγουμε *Simple* και *Summaries for group of cases* και πατάμε *Define*. Στο παράθυρο που εμφανίζεται εισάγουμε τη μεταβλητή samples στο *Variable* και τη groups στο *Category Axis*. Η γραφική παράσταση που προκύπτει δίνεται στο σχήμα 8.26, όπου παρατηρούμε ότι μάλλον το

δεύτερο δείγμα είναι αυτό που διαφοροποιείται από τα υπόλοιπα.

Πάντως, αν συνδυάσουμε τα θηκογράμματα με τα αποτελέσματα του *ChemStat* στο σχήμα 8.24 προκύπτει ότι η στατιστικά μικρότερη αέρια ρύπανση υπάρχει στα Πετρανά, η μεγαλύτερη στον Κλείτο, ενώ Κοζάνη και Πολύμυλος έχουν παρόμοια ρύπανση.



Σχήμα 8.26. Θηκογράμματα των δειγμάτων του πίνακα 8.5

8.4.2 ΔΙΠΑΡΑΓΟΝΤΙΚΗ ΑΝΑΛΥΣΗ - Κριτήριο Friedman

Αυστηρά το κριτήριο *Friedman* εφαρμόζεται σε πολλά εξαρτώμενα δείγματα και συνεπώς μπορεί να εφαρμοστεί και σε δείγματα που οι τιμές τους καθορίζονται από δύο παράγοντες. Η μηδενική και η εναλλακτική της υπόθεση ορίζονται όπως και στην μονοπαραγοντική ανάλυση. Σε αντίθεση με το κριτήριο *Kruskal-Wallis*, εδώ δεν απαιτείται η δημιουργία ενιαίου δείγματος. Τα δεδομένα τοποθετούνται σε πίνακα της μορφής 8.5 και υπολογίζεται η συνάρτηση F:

$$F = \frac{12}{mk(k+1)} \sum_{i=1}^{k} R_i^2 - 3m(k+1) \qquad (8.10)$$

όπου k είναι το πλήθος των δειγμάτων, m είναι το μέγεθος των δειγμάτων κοινό σε όλα τα δείγματα και R_i είναι το άθροισμα των βαθμών του i δείγματος (= 1, 2, …, k), όπου οι βαθμοί της τιμής x_{ji} υπολογίζονται στη

γραμμή j του πίνακα 8.5. Αυστηρά η σχέση (8.10) ισχύει όταν δεν υπάρχουν δεσμοί ή είναι πολύ λίγοι. Διαφορετικά, όπως και στην περίπτωση της (8.9), απαιτούνται διορθώσεις.

Με βάση την τιμή F υπολογίζεται η πιθανότητα p-value δεδομένου ότι η F ακολουθεί ασυμπτωτικά την κατανομή χ^2 με k-1 βαθμούς ελευθερίας.

Πίνακας 8.5. Πίνακας για διπαραγοντική ανάλυση διασποράς.

Δεύτερος παράγοντας	Πρώτος παράγοντας			
	1	2	. . .	k
1	x_{11}	x_{12}	. . .	x_{1k}
2	x_{21}	x_{22}	. . .	x_{2k}
. . .			. . .	
m	x_{m1}	x_{m2}	. . .	x_{mk}

Παρατήρηση. Η σχέση (8.10) ουσιαστικά ελέγχει την μηδενική υπόθεση ότι τα δείγματα προέρχονται από τον ίδιο πληθυσμό και συνεπώς ο πρώτος παράγοντας δεν επιδρά στα δείγματα. Για να ελέγξουμε αν ισχύει το ίδιο με τον δεύτερο παράγοντα, αντιμεταθέτουμε τις γραμμές με τις στήλες στον πίνακα 8.5 και ξανα-εφαρμόζουμε τη σχέση (8.10).

Παράδειγμα 8.5

Μία τεχνική καταστροφής ρύπων ελέγχεται κάτω από διαφορετικές θερμοκρασίες και τύπους καταλυτών και στον πίνακα 8.6 δίνεται η επί τοις εκατό ποσότητα των ρύπων που καταστρέφονται. Να εξετασθεί αν η επίδραση του τύπου του καταλύτη και της θερμοκρασίας είναι στατιστικά σημαντικές στην τεχνική αυτή.

Πίνακας 8.6. Δεδομένα του παραδείγματος 8.5.

	Καταλύτης Α	Καταλύτης Β	Καταλύτης C
T = 25 °C	90	80	60
T = 35 °C	90	75	60
T = 45 °C	85	70	65
T = 55 °C	80	75	55

❖ Ανάλυση στο ChemStat

Για την εφαρμογή του κριτηρίου *Friedman* με το *ChemStat* εργαζόμαστε όπως και στο προηγούμενο παράδειγμα με μόνη διαφορά ότι στο πρώτο παράθυρο που ανοίγει εισάγουμε τον αριθμό 2.

	A	B	C	D
1		**K-A**	**K-B**	**K-C**
2	**T = 25 °C**	90	80	60
3	**T = 35 °C**	90	75	60
4	**T = 45 °C**	85	70	65
5	**T = 55 °C**	80	75	55

F	G	H	I	J	K	L
Friedman two-way Non-parametric ANOVA - 2 tailed						
Test of VARIABLES (columns)						
F=	8					
p-value=	0,018316	Null hypothesis may be rejected at level 0.05				
Monte-Carlo iter	10000					
p(permut.)=	0,0046	Null hypothesis may be rejected at level 0.05				
Pairwise comparisons						
Pair1- 2	P-value=	0				
Pair1- 3	P-value=	0				
Pair2- 3	P-value=	0				
Test of CASES (rows)						
F=	3,666667					
p-value=	0,299781	Null hypothesis may be assumed at level 0.05				
Monte-Carlo iter	10000					
p(permut.)=	0,3949	Null hypothesis may be assumed at level 0.05				
Pairwise comparisons						
Pair1- 2	P-value=	0,6149				
Pair1- 3	P-value=	0,3297				
Pair1- 4	P-value=	0,0998				
Pair2- 3	P-value=	0,6149				
Pair2- 4	P-value=	0,207				
Pair3- 4	P-value=	0,4108				

Σχήμα 8.27. Διευθέτηση δεδομένων και πίνακας αποτελεσμάτων στο *ChemStat*

Από τον πίνακα των αποτελεσμάτων (σχήμα 8.27) προκύπτει ότι η H_0 μπορεί να απορριφθεί <u>όταν τα δείγματα ορίζονται κατά στήλες</u> (p-value = 0.018 < 0.05). Συνεπώς υπάρχουν στατιστικά σημαντικές διαφορές όταν μελετάμε την επίδραση του τύπου του καταλύτη. Αντίθετα, η επίδραση της

θερμοκρασίας σε επίπεδο σημαντικότητας α = 0.05 δεν διαπιστώνεται να είναι στατιστικά σημαντική (p-value = 0.2998). Επίσης παρατηρούμε ότι μεταξύ όλων των ζευγών των καταλυτών υπάρχουν στατιστικά σημαντικές διαφορές, ενώ δεν φαίνεται να υπάρχει ούτε ένα ζεύγος θερμοκρασιών που τα δείγματά του να διαφοροποιούνται στατιστικά.

❖ Ανάλυση στο SPSS

Εισάγουμε τα δεδομένα στο φύλλο εργασίας όπως στο σχήμα 8.28 και ακολουθούμε τη διαδικασία *Analyze → Nonparametric Tests → Legacy Dialogs → K Related Samples.* Στο παράθυρο που ανοίγει μεταφέρουμε τις μεταβλητές K_A, K_B, K_C στο πλαίσιο *Tests Variable*. Στο πάνελ *Test Type* επιλέγουμε το *Friedman* (αν δεν είναι επιλεγμένο) και από το κουμπί *Exact* επιλέγουμε *Exact*. Με *κλικ* στο *OK* παίρνουμε τον πίνακα αποτελεσμάτων του σχήματος 8.29. Παρατηρούμε ότι Asymp. Sig. = 0.018, που ταυτίζεται με την αντίστοιχη τιμή του *ChemStat*, ενώ ο ακριβής υπολογισμός δίνει Exact Sig.= 0.005.

	K_A	K_B	K_C
1	90	80	60
2	90	75	60
3	85	70	65
4	80	75	55

Σχήμα 8.28. Διευθέτηση δεδομένων σε φύλλο εργασίας του *SPSS*

Test Statistics[a]

N	4
Chi-Square	8,000
df	2
Asymp. Sig.	,018
Exact Sig.	,005
Point Probability	,005

a. Friedman Test

Σχήμα 8.29. Πίνακας αποτελεσμάτων ελέγχου *Friedman* για την επίδραση του καταλύτη

Για να δούμε την επίδραση της θερμοκρασίας κάνουμε τις γραμμές στήλες και τις στήλες γραμμές ως εξής: Από το *Data → Transpose* στο

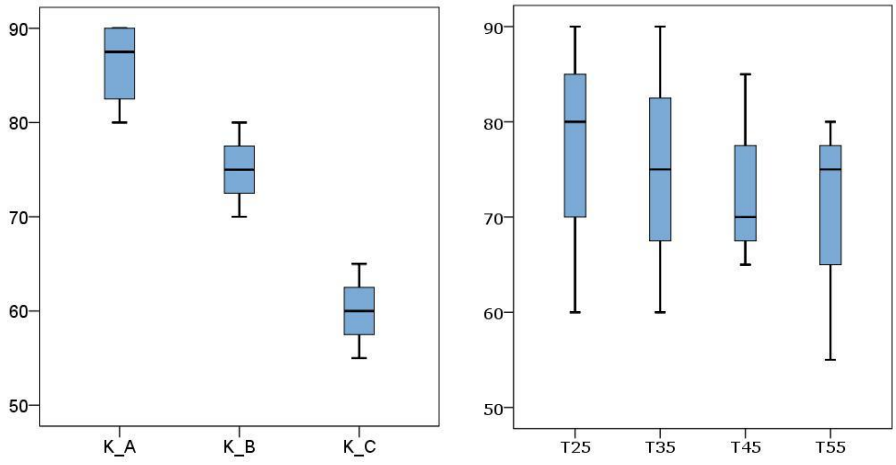
παράθυρο που ανοίγει μεταφέρουμε τις μεταβλητές K_A, K_B, K_C στο πλαίσιο *Variable(s)* και κάνουμε *κλικ* στο *OK*. Τα δεδομένα αναστρέφονται σε ένα νέο φύλλο εργασίας που ανοίγει αυτόματα. Στο νέο αυτό φύλλο δίνουμε στις νέες μεταβλητές τα ονόματα T25, T35, T45 και T55 και επαναλαμβάνουμε την παραπάνω διαδικασία *Analyze* → *Nonparametric Tests* → *K Related Samples* εισάγοντας τώρα τις νέες μεταβλητές στο πλαίσιο *Test Variables*. Τα αποτελέσματα που παίρνουμε δίνονται στον παρακάτω πίνακα, όπου παρατηρούμε και πάλι ότι ταυτίζονται με τα αντίστοιχα του *ChemStat*.

Test Statistics[a]

N	3
Chi-Square	3,667
df	3
Asymp. Sig.	,300
Exact Sig.	,389
Point Probability	,146

a. Friedman Test

Σχήμα 8.30. Πίνακας αποτελεσμάτων ελέγχου *Friedman* για την επίδραση της θερμοκρασίας



Σχήμα 8.31. Θηκογράμματα των δειγμάτων του πίνακα 8.6

Η σημαντική επίδραση του καταλύτη σε αντίθεση με αυτή της θερμοκρασίας φαίνεται χαρακτηριστικά στα θηκογράμματα του σχήματος 8.31. Για να τα κατασκευάσουμε ακολουθούμε τη γνωστή διαδικασία *Graphs* → *Legacy Dialogs* → *Boxplot* και επιλέγουμε τον τύπο *Simple* και *Summaries of separate variables*. Στο παράθυρο διαλόγου που ανοίγει μεταφέρουμε τις μεταβλητές K_A, K_B, K_C στο πλαίσιο *Boxes Represent* και πατάμε *OK*. Επαναλαμβάνουμε την παραπάνω διαδικασία με τις μεταβλητές T25, T35, T45 και T55.

ΑΣΚΗΣΕΙΣ

8.1. Μετρήθηκε η ποσότητα πρωτεΐνης (g/100 mL) στο αίμα ατόμων που διαβιούν στις περιοχές Α, Β, Γ και βρέθηκε:

Α	7.6	7.0	7.4	7.6	7.7	7.6	8.1		
Β	7.7	7.6	7.0	6.7	7.3	7.1	7.5	7.2	
Γ	8.0	7.9	7.1	7.7	8.2	8.3	7.2	7.4	6.3

Λαμβάνοντας υπόψη ότι η γεωγραφική διαφοροποίηση οδηγεί σε διαφορετικές συνθήκες διαβίωσης, μπορούμε να ισχυριστούμε ότι η ποσότητα της πρωτεΐνης στο αίμα είναι η ίδια και στις τρεις περιοχές;

8.2. Για να ελεγχθεί η σταθερότητα υδατικών διαλυμάτων τυροσίνης παρασκευάστηκε ένα αρχικό διάλυμα ορισμένης συγκέντρωσης και από το διάλυμα αυτό σε ορισμένα χρονικά διαστήματα γινόταν δειγματοληψία και προσδιορισμός της τυροσίνης με 5 επαναλαμβανόμενες μετρήσεις. Τα αποτελέσματα που ελήφθησαν δίνονται στον παρακάτω πίνακα. Υπάρχουν στατιστικά σημαντικές διαφορές μεταξύ των δειγμάτων;

Χρόνος, t	Συγκεντρώσεις τυροσίνης, mg/L				
0	10.2	10.0	10.1	9.9	10.1
6 h	10.1	10.1	10.3	10.0	9.8
12 h	9.7	9.5	9.9	9.6	9.8
24 h	9.0	9.2	9.4	9.2	9.3

8.3. Οι συγκεντρώσεις των αέριων σωματιδίων σε mg/m^3 σε πέντε διαφορετικές πόλεις και σε 4 χρονικές περιόδους δίνονται στον παρακάτω πίνακα. Να εξετασθεί αν παρατηρούνται στατιστικά σημαντικές διαφοροποιήσεις;

	Μήνας			
Πόλη	Δεκέμβριος	Μάρτιος	Ιούνιος	Αύγουστος
Α	94	87	78	72
Β	79	77	67	64
Γ	110	93	77	73
Δ	67	64	62	66
Ε	80	58	48	42

8.4. Για να συγκριθούν τέσσερα φάρμακα με υποτασικές ιδιότητες χορηγείται το κάθε ένα σε μια ομάδα 5 ατόμων και καταγράφεται η αρτηριακή πίεση του κάθε ατόμου σε mmHg. Τα αποτελέσματα δίνονται στον παρακάτω πίνακα. Να εξετασθεί αν υπάρχουν στατιστικά σημαντικές διαφορές μεταξύ των φαρμάκων.

Φάρμακο	Αρτηριακή πίεση ασθενούς				
1	160	150	160	170	180
2	160	130	170	160	150
3	130	170	130	150	130
4	130	150	130	120	140

8.5. Το ετήσιο βροχομετρικό ύψος σε mm στα νησιά του Ιονίου πελάγους και του Βορείου και Νοτίου Αιγαίου δίνεται στον παρακάτω πίνακα. Να εξετασθεί αν οι διαφοροποιήσεις που φαίνονται είναι στατιστικά σημαντικές.

Ιόνιο	Βόρειο Αιγαίο	Νότιο Αιγαίο
700	590	295
1150	710	720
1350	920	460
990	605	510
880	660	730
1560		440
685		410
		295

8.6. Στον παρακάτω πίνακα δίνονται οι ποσότητες σκουριάς σε g ως απώλεια βάρους μετά την αφαίρεσή της σε 4 κομμάτια σιδήρου που υπέστησαν τρεις διαφορετικές κατεργασίες Α, Β και Γ. Να εξετασθεί αν υπάρχει στατιστικά σημαντική διαφορά των κατεργασιών.

Α	3	5	4	4
Β	4	2	3	3
Γ	6	4	5	5

8.7. Σε μία περιβαλλοντική μελέτη προσδιορίστηκε η συγκέντρωση του μολύβδου σε ppm σε συνάρτηση με την απόσταση από μία βιομηχανική μονάδα ανακύκλωσης συσσωρευτών μολύβδου και το βάθος από τη επιφάνεια του εδάφους. Τα αποτελέσματα δίνονται στο επόμενο πίνακα. Να εξετασθεί αν υπάρχουν στατιστικά σημαντικές διαφοροποιήσεις της συγκέντρωσης του μολύβδου σε σχέση με το βάθος και την απόσταση.

Απόσταση, km	Βάθος, m		
	0	0.5	1
1	48	46	45
2	30	30	25
3	20	18	15
4	10	10	7
5	5	5	3

8.8. Να εξετασθεί η επίδραση της θερμοκρασίας και του τύπου του καταλύτη στην εκατοστιαία απόδοση μιας χημικής αντίδρασης με βάση τα δεδομένα του παρακάτω πίνακα.

T / ^{0}C	Καταλύτης Α	Καταλύτης Β	Καταλύτης Γ
25	12	18	27
25	11	20	27
25	13	21	33
30	15	27	33
30	18	27	33
30	16	33	27
35	25	34	28
35	24	35	29
35	28	37	35
40	28	37	38
40	25	37	34
40	27	36	32

8.9. Επανεξετάστε το παράδειγμα 8.3 με μονοπαραγοντική ANOVA χρησιμοποιώντας τον ενοποιημένο παράγοντα T-pH και προσδιορίστε τις συνθήκες που οδηγούν στη μέγιστη απόδοση της χημικής αντίδρασης.

Κεφάλαιο 9

ΕΛΕΓΧΟΙ ΣΕ ΚΑΤΗΓΟΡΙΚΑ ΔΕΔΟΜΕΝΑ

9.1 ΓΕΝΙΚΑ

Οι παραμετρικοί και μη παραμετρικοί έλεγχοι που εξετάσαμε στα προηγούμενα κεφάλαια αφορούν ποσοτικά δεδομένα. Πολλές φορές όμως είναι απαραίτητο να αναλύσουμε δεδομένα του τύπου κατηγορικά. Τα κατηγορικά δεδομένα προκύπτουν όταν με βάση κάποιο ποιοτικό ή και ποσοτικό κριτήριο ταξινομούμε τα δεδομένα σε κατηγορίες. Για παράδειγμα, έστω ότι μελετάμε μια ομάδα νέων ανθρώπων αποτελούμενη από αγόρια και κορίτσια. Η μεταβλητή "φύλο" είναι μια κατηγορική μεταβλητή με τιμές στο συγκεκριμένο παράδειγμα "αγόρι" και "κορίτσι".

Στο παράδειγμα αυτό τα δεδομένα κατηγοριοποιούνται ως προς ένα χαρακτηριστικό, το φύλο. Μπορούν όμως τα δεδομένα μιας έρευνας να κατηγοριοποιηθούν ως προς δύο χαρακτηριστικά. Σε αυτή την περίπτωση μπορούμε να γενικεύσουμε και να δεχθούμε ότι στα δεδομένα υπάρχουν δύο ποιοτικές μεταβλητές, Α και Β, των οποίων οι τιμές είναι οι κατηγορίες A_1, A_2, ..., A_R και B_1, B_2, ..., B_C, αντίστοιχα. Τέτοια δεδομένα δίνονται συνήθως με τη μορφή του παρακάτω πίνακα, που ονομάζεται ***πίνακας συνάφειας*** *(contigency table)* των μεταβλητών Α και Β.

	Μεταβλητή Β			
Μεταβλητή Α	B_1	B_2	...	B_C
A_1	n_{11}	n_{12}	...	n_{1C}
A_2	n_{21}	n_{22}	...	n_{2C}
...	...	...	...	...
A_R	n_{R1}	n_{R2}	...	n_{RC}

Στον πίνακα αυτόν n_{ij} είναι το πλήθος των μετρήσεων/παρατηρήσεων που αντιστοιχούν ταυτόχρονα στις κατηγορίες A_i και B_j.

Οι στατιστικοί έλεγχοι που εφαρμόζονται σε έναν πίνακα συνάφειας είναι συνήθως η μη παραμετρική ***δοκιμασία της ανεξαρτησίας*** *(test of independence)* και η ***ακριβής δοκιμασία του Fisher*** *(Fisher's exact test)*.

9.2 ΔΟΚΙΜΑΣΙΑ ΤΗΣ ΑΝΕΞΑΡΤΗΣΙΑΣ

Με τη μη παραμετρική *δοκιμασία της ανεξαρτησίας* ελέγχουμε αν δύο κατηγορικές μεταβλητές είναι ανεξάρτητες. Ο έλεγχος βασίζεται στην κατανομή χ^2 (chi-square) ως εξής: Υπολογίζουμε τη συνάρτηση χ^2 του *Pearson*

$$\chi^2 = \sum_i \sum_j \frac{(n_{ij} - E_{ij})^2}{E_{ij}} \qquad (9.1)$$

όπου n_{ij} είναι οι τιμές του πίνακα συνάφειας και E_{ij} οι αναμενόμενες συχνότητες, οι οποίες υπολογίζονται από τον τύπο:

$$E_{ij} = (\text{άθροισμα } i \text{ γραμμής}) \times (\text{άθροισμα } j \text{ στήλης}) / (\text{γενικό άθροισμα } n)$$

Η συνάρτηση που υπολογίζεται από την παραπάνω σχέση ακολουθεί την κατανομή χ^2 με (R-1)(C-1) βαθμούς ελευθερίας όταν στον πίνακα συνάφειας έχουμε R γραμμές και C στήλες. Η μηδενική υπόθεση που ελέγχουμε είναι:

H_0: Τα χαρακτηριστικά Α και Β είναι ανεξάρτητα μεταξύ τους

με εναλλακτική την

H_1: Τα χαρακτηριστικά Α και Β είναι εξαρτημένα

Ένα σημείο που πρέπει να προσέξουμε στον έλεγχο της ανεξαρτησίας δύο μεταβλητών είναι το ποσοστό των κελιών με αναμενόμενη συχνότητα μικρότερη του 5. Αν το ποσοστό αυτό ξεπερνά το 20%, τότε ο έλεγχος μπορεί να μην είναι αξιόπιστος. Σε αυτές τις περιπτώσεις μπορεί να χρησιμοποιηθεί η μέθοδος *Monte-Carlo με αντιμεταθέσεις*.

Παράδειγμα 9.1

Σε ένα πείραμα μελετήθηκε η συγκέντρωση φυτοφαρμάκων σε εδάφη με διαφορετική ικανότητα απορροής των υδάτων. Ακολούθως οι συγκεντρώσεις κατηγοριοποιήθηκαν σε Α, Β, C, όπου Α αντιστοιχεί σε συγκεντρώσεις πάνω από το επιτρεπτό όριο, Β είναι συγκεντρώσεις μεταξύ επιτρεπτού ορίου και ελάχιστου ορίου ανίχνευσης και C συγκεντρώσεις κάτω από το όριο ανίχνευσης. Επίσης κατηγοριοποιήθηκε η ικανότητα απορροής των υδάτων με βάση την κλίση του εδάφους σε Κακή, Μέτρια και Καλή. Στον πίνακα 9.1 δίνονται τα αποτελέσματα του πειράματος. Να εξετασθεί αν η απορροή έχει στατιστικά σημαντική επίδραση στη συγκέντρωση των φυτοφαρμάκων στο έδαφος.

Πίνακας 9.1. Πίνακας συνάφειας 3×3.

Συγκέντρωση φυτοφαρμάκων	Επίδοση		
	Κακή	Μέτρια	Καλή
A	25	15	9
B	8	12	10
C	6	6	12

❖ **Ανάλυση στο ChemStat**

Διευθετούμε τα δεδομένα όπως στο σχήμα 9.1 και στη συνέχεια πηγαίνουμε *Πρόσθετα* → *ChemStat* → *Tables* → *Chi-square test.* Συμπληρώνουμε κατάλληλα τα παράθυρα που ανοίγουν και παίρνουμε τα αποτελέσματα του σχήματος 9.1.

Παρατηρούμε ότι p-value = 0.029, ενώ και η αντίστοιχη τιμή από τη μέθοδο *Monte-Carlo με αντιμεταθέσεις* είναι πρακτικά ίδια, p(permut.) = 0.028. Συνεπώς η μηδενική υπόθεση μπορεί να απορριφθεί και επομένως μπορούμε να συμπεράνουμε ότι εδάφη με καλύτερη απορροή υδάτων έχουν στατιστικά μικρότερη συγκέντρωση φυτοφαρμάκων.

❖ **Ανάλυση στο Excel**

Στο *Excel* θα πρέπει να εκτελέσουμε όλες τις πράξεις αναλυτικά. Δηλαδή, να υπολογίσουμε αναμενόμενες συχνότητες, το άθροισμα χ^2 και τέλος να προσδιορίσουμε την p-value χρησιμοποιώντας τον τύπο: =CHIDIST(χ^2;(C-1)(R-1)). Όλη αυτή η πορεία δίνεται στο σχήμα 9.2.

	A	B	C	D
1	Συγκέντρωση φυτοφαρμάκων	Επίδοση		
2		Κακή	Μέτρια	Καλή
3	A	25	15	9
4	B	8	12	10
5	C	6	6	12
6				
7	**Chi_Squared Test**			
8	Chi-Squared =	10,78119		
9	p-value=	0,029136		
10	The examined factor appears to be significant			
11	MC iterations=	10000		
12	p(permut.)=	0,027997		
13	The examined factor appears to be significant			
14				
15	Elapsed time =	0,128 min		

Σχήμα 9.1. Διευθέτηση δεδομένων και πίνακας αποτελεσμάτων στο *ChemStat*

	A	B	C	D	E
1	Συγκέντρωση	Επίδοση			
2	φυτοφαρμάκων	Κακή	Μέτρια	Καλή	Σύνολα
3	A	25	15	9	49
4	B	8	12	10	30
5	C	6	6	12	24
6	Σύνολα	39	33	31	103
7					
8		Αναμενόμενες τιμές			
9		18,5534	15,69903	14,74757	
10		11,35922	9,61165	9,029126	
11		9,087379	7,68932	7,223301	
12					
13		Τιμές $(n_{ij}-E_{ij})^2/E_{ij}$			
14		2,23995	0,031126	2,240002	
15		0,993411	0,593469	0,104395	
16		1,048917	0,371139	3,158785	
17					
18	χ^2 =	10,78119			
19	p-value=	0,029136			

Σχήμα 9.2. Υπολογισμός p-value στο *Excel* για τη δοκιμασία της ανεξαρτησίας

Συγκεκριμένα, πρώτα υπολογίζουμε στη γραμμή 6 τα αθροίσματα των στηλών, στη στήλη Ε τα αθροίσματα των γραμμών και στο κελί Ε6 το ολικό άθροισμα των τιμών. Ακολούθως στην περιοχή B9:D11 υπολογίζουμε τις αναμενόμενες συχνότητες ως εξής. Στο κελί B9 πληκτρολογούμε τον τύπο =$E3*B$6/E6 και συμπληρώνουμε με τη διαδικασία της αυτόματης συμπλήρωσης όλη την περιοχή B9:D11. Επίσης, στην περιοχή B14:D16 υπολογίζουμε τους προσθετέους του αθροίσματος χ^2, δηλαδή τις ποσότητες $(n_{ij} - E_{ij})^2/E_{ij}$. Για το σκοπό αυτό στο κελί B14 πληκτρολογούμε τον τύπο =(B3-B9)^2/B9 και συμπληρώνουμε με τη διαδικασία της αυτόματης συμπλήρωσης όλη την περιοχή B14:D16. Τέλος, στο κελί B18 υπολογίζουμε το χ^2 με τον τύπο =SUM(B14:D16) και στο B19 την πιθανότητα p-value χρησιμοποιώντας τη σχέση =CHIDIST(B18;2*2). Όπως αναμένεται το αποτέλεσμα ταυτίζεται με το αντίστοιχο το *ChemStat*.

❖ Ανάλυση στο SPSS

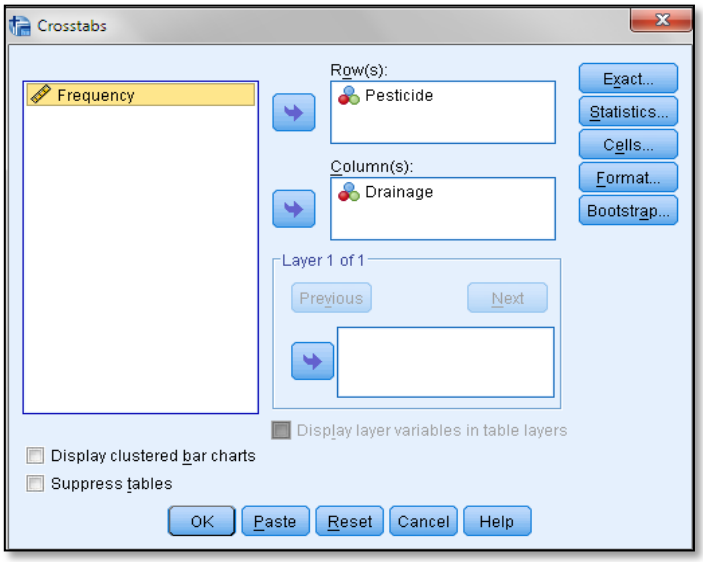
Στο *SPSS* εισάγουμε τα δεδομένα όπως στο σχήμα 9.3. Παρατηρούμε ότι οι συχνότητες τοποθετούνται σε μία στήλη και το ίδιο συμβαίνει για τις μεταβλητές που αφορούν τη συγκέντρωση φυτοφαρμάκων (Pesticide) και την ικανότητα απορροής υδάτων (Drainage). Η μεταβλητή Pesticide παίρνει τρεις τιμές, 1, 2 και 3, που αντιστοιχούν στις κατηγορίες Α, Β και C και η μεταβλητή Drainage παίρνει επίσης τις τιμές 1, 2, 3 που αντιστοιχούν στις κατηγορίες Κακή, Μέτρια και Καλή.

	Pesticide	Drainage	Frequency
1	1	1	25
2	1	2	15
3	1	3	9
4	2	1	8
5	2	2	12
6	2	3	10
7	3	1	6
8	3	2	6
9	3	3	12

Σχήμα 9.3. Διευθέτηση δεδομένων στο *SPSS*

Ακολούθως σταθμίζουμε τα δεδομένα από *Data* → *Weight Cases*, όπου στο παράθυρο διαλόγου που ανοίγει κάνουμε *κλικ* στο *Weight cases by* και μεταφέρουμε στο αντίστοιχο πεδίο τη μεταβλητή Frequency. Τέλος, πηγαίνουμε *Analyze* → *Descriptive Statistics* → *Crosstabs* και στο

παράθυρο που εμφανίζεται μεταφέρουμε τη μεταβλητή Pesticide στο πάνελ *Row(s)* και τη μεταβλητή Drainage στο πάνελ *Column(s)* (Σχήμα 9.4). Στη συνέχεια κάνουμε *κλικ* στο *Statistics* και επιλέγουμε *Chi-square* και *κλικ* στο *Exact* όπου επιλέγουμε *Monte-Carlo*. Με την τελευταία επιλογή το πρόγραμμα εφαρμόζει τη μέθοδο *Monte-Carlo με αντιμεταθέσεις*. Ολοκληρώνουμε με *κλικ* στο *Continue* και στη συνέχεια *κλικ* στο *OK*.



Σχήμα 9.4. Είσοδος δεδομένων στο *SPSS* για τη δοκιμασία της ανεξαρτησίας

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Monte Carlo Sig. (2-sided)		
				Sig.	99% C I	
					Lower Bound	Upper Bound
Pearson Chi-Square	10,781[a]	4	,029	,030[b]	,026	,034
Likelihood Ratio	10,639	4	,031	,036[b]	,032	,041
Fisher's Exact Test	10,343			,034[b]	,029	,039
Linear-by-Linear Ass	8,837[c]	1	,003	,003[b]	,002	,005
N of Valid Cases	103					

a. 0 cells (,0%) have expected count less than 5.

Σχήμα 9.5. Τμήμα του πίνακα αποτελεσμάτων στο *SPSS* για τη δοκιμασία της ανεξαρτησίας

Ο βασικός πίνακας αποτελεσμάτων δίνεται στο σχήμα 9.5. Στον πίνακα αυτόν ελέγχουμε μόνο την πρώτη γραμμή. Η τιμή Asymp. Sig. = 0.029 ταυτίζεται με την αντίστοιχη του *ChemStat* και πρακτικά το ίδιο ισχύει για την τιμή p-value από τη μέθοδο *Monte-Carlo με αντιμεταθέσεις*, Sig. = 0.030.

9.3 Η ΑΚΡΙΒΗΣ ΔΟΚΙΜΑΣΙΑ ΤΟΥ FISHER

Η *ακριβής δοκιμασία του Fisher* χρησιμοποιείται κυρίως σε πίνακες συνάφειας 2×2 όταν υπάρχουν τιμές μικρότερες του 5. Εναλλακτικά μπορεί να χρησιμοποιηθεί και η μέθοδος *Monte-Carlo με αντιμεταθέσεις*.

Πίνακας 9.2. Πίνακας συνάφειας 2×2.

	Μεταβλητή Β	
Μεταβλητή Α	B_1	B_2
A_1	α	β
A_2	γ	δ

Αποδεικνύεται ότι αν έχουμε ένα πίνακα συνάφειας της μορφής του 9.2 και με την προϋπόθεση ότι οι μεταβλητές Α και Β δεν σχετίζονται (ισχύει η μηδενική υπόθεση), τότε η πιθανότητα Ρ στα κελιά του πίνακα να υπάρχουν οι αριθμοί α, β, γ, δ, όπως στον πίνακα 9.2, είναι

$$P = \frac{(\alpha+\beta)!(\alpha+\gamma)!(\gamma+\delta)!(\beta+\delta)!}{\alpha!\beta!\gamma!\delta!(\alpha+\beta+\gamma+\delta)!}$$

Η παραπάνω πιθανότητα ισχύει με την επιπλέον προϋπόθεση ότι τα αθροίσματα α+β, α+γ, γ+δ, β+δ είναι σταθερά. Αν τώρα υπολογίσουμε τις πιθανότητες Ρ ως προς όλες τις δυνατές διατάξεις αριθμών στον πίνακα 2×2 έτσι ώστε τα αθροίσματα α+β, α+γ, γ+δ, β+δ να παραμένουν πάντα σταθερά, τότε το άθροισμα όλων των Ρ μας δίνει την πιθανότητα p-value και συνεπώς μπορούμε να ελέγξουμε την μηδενική υπόθεση.

Παράδειγμα 9.2

Θέλουμε να μελετήσουμε αν η προεγχειρητική χρήση αντιβιοτικών μειώνει τις μετεγχειρητικές λοιμώξεις με βάση τα δεδομένα του Πίνακα 9.3, στα οποία καταγράφονται οι συχνότητες λοίμωξης 31 ασθενών.

Πίνακας 9.3. Πίνακας συνάφειας του παραδείγματος 9.2.

	Μετεγχειρητικές λοιμώξεις	
Προεγχειρητική χρήση αντιβιοτικών	Ναι	Όχι
Ναι	5	10
Όχι	12	4

◆ Το παράδειγμα είναι ίδιο με το προηγούμενο όμως τώρα πρέπει να χρησιμοποιηθεί η *ακριβής δοκιμασία του Fisher*, επειδή σε ένα κελί υπάρχει ο αριθμός 4 < 5. Στα προγράμματα που εξετάζουμε, δυνατότητες εφαρμογής αυτού του ελέγχου υπάρχουν μόνο στο *ChemStat* και στο *SPSS*.

❖ **Ανάλυση στο ChemStat**

Εργαζόμαστε όπως στο προηγούμενο παράδειγμα, όμως τώρα επιλέγουμε *Fisher exact test*. Παίρνουμε τον παρακάτω πίνακα αποτελεσμάτων, στον οποίο η τιμή p-value = 0.032 δείχνει ότι η μηδενική υπόθεση μπορεί να απορριφθεί, δηλαδή η προεγχειρητική χρήση αντιβιοτικών φαίνεται να μειώνει στατιστικά σημαντικά τις μετεγχειρητικές λοιμώξεις.

	A	B	C	D
1	**Προεγχειρητική χρήση**	**Μετεγχειρητικές λοιμώξεις**		
2	**αντιβιοτικών**	**Ναι**	**Όχι**	
3	**Ναι**	5	10	
4	**Όχι**	12	4	
5				
6	**Fisher exact test**			
7	p-value:	0,031952		
8	The examined factor appears to be significant			
9				

Σχήμα 9.6. Πίνακας αποτελεσμάτων στο *ChemStat*

Είναι ενδιαφέρον να εξετάσουμε το παράδειγμα αυτό και με τον έλεγχο χ^2. Πηγαίνουμε *Πρόσθετα* → *ChemStat* → *Tables* → *Chi-square test* και στα παράθυρα που ανοίγουν επιλέγουμε και τον έλεγχο *Monte-Carlo με αντιμεταθέσεις* με 10000 επαναλήψεις. Στα αποτελέσματα του

σχήματος 9.7 παρατηρούμε ότι η μέθοδος *Monte-Carlo* δίνει πρακτικά ταυτόσημα αποτελέσματα με αυτά της *ακριβούς δοκιμασίας του Fisher* (p-value = 0.032), ενώ αντίθετα ο απλός έλεγχος χ^2 δίνει μια αρκετά διαφορετική τιμή, p-value = 0.0198.

Chi_Squared Test	
Chi-Squared =	5,427171
p-value=	0,019826
The examined factor appears to be significant	
MC iterations=	10000
p(permut.)=	0,032297
The examined factor appears to be significant	

Σχήμα 9.7. Πίνακας αποτελεσμάτων των ελέγχων χ^2 και *Monte-Carlo* στο *ChemStat*

❖ Ανάλυση στο SPSS

Στο *SPSS* εργαζόμαστε ακριβώς όπως στο προηγούμενο παράδειγμα. Δηλαδή εισάγουμε τα δεδομένα σε τρεις στήλες-μεταβλητές, όπου η μία μεταβλητή αφορά τη χρήση αντιβιοτικών (Antibiotics), η άλλη τις μολύνσεις (Infections) και η τρίτη τις συχνότητες (Frequency). Οι μεταβλητές Antibiotics και Infections παίρνουν και οι δύο τις τιμές 1 (Ναι) και 2 (Όχι). Αυτή η διευθέτηση των δεδομένων δίνεται στο σχήμα 9.8.

	Antibiotics	Infections	Frequency
1	1	1	5
2	1	2	10
3	2	1	12
4	2	2	4

Σχήμα 9.8. Εισαγωγή δεδομένων στο *SPSS*

Ακολούθως σταθμίζουμε τα δεδομένα από *Data* → *Weight Cases* και στη συνέχεια ακολουθούμε τη διαδικασία *Analyze* → *Descriptive Statistics* → *Crosstabs.* Στο παράθυρο που εμφανίζεται μεταφέρουμε τη μεταβλητή Antibiotics στο πάνελ *Row(s)* και τη μεταβλητή Infections στο πάνελ *Column(s)*. Κάνουμε *κλικ* στο *Statistics*, επιλέγουμε *Chi-square* και

ολοκληρώνουμε με *κλικ* στο *Continue* και στη συνέχεια *κλικ* στο *OK*. Παίρνουμε τους πίνακες αποτελεσμάτων του σχήματος 9.9. Η τιμή Exact Sig.(2-sided) = 0.032 όπως αναμένεται ταυτίζεται με την αντίστοιχη του *ChemStat* και δείχνει ότι η προεγχειρητική χρήση αντιβιοτικών επηρεάζει στατιστικά σημαντικά τις μετεγχειρητικές λοιμώξεις.

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	5,427[a]	1	,020		
Continuity Correction[b]	3,875	1	,049		
Likelihood Ratio	5,594	1	,018		
Fisher's Exact Test				,032	,024
Linear-by-Linear Association	5,252	1	,022		
N of Valid Cases	31				

a. 0 cells (,0%) have expected count less than 5. The minimum expected count is 6,77.
b. Computed only for a 2x2 table

Σχήμα 9.9. Πίνακας αποτελεσμάτων στο *SPSS*

ΑΣΚΗΣΕΙΣ

9.1. Στον παρακάτω πίνακα καταγράφεται το πλήθος των ειδών που είναι ανθεκτικά στη μόλυνση ποταμών. Να εξετάσετε αν το πλήθος αυτό είναι στατιστικά διαφορετικό στους τρεις ποταμούς.

Ανθεκτικά είδη στη μόλυνση	Ποταμός 1	Ποταμός 2	Ποταμός 3
Ναι	2	12	5
Όχι	15	6	6

9.2. Τα επίπεδα του αρσενικού προσδιορίστηκαν σε 87 θέσεις τεσσάρων περιοχών και τα αποτελέσματα δίνονται στον παρακάτω πίνακα. Να εξεταστεί αν η κατανομή του αρσενικού διαφοροποιείται στις τέσσερες αυτές περιοχές.

Επίπεδα αρσενικού πάνω από το επιτρεπτό όριο	Α	Β	Γ	Δ
Ναι	1	5	5	12
Όχι	15	12	15	22

9.3. Η παρουσία πτητικών οργανικών ενώσεων προσδιορίστηκε σε επιφανειακά και σε υπόγεια ύδατα, όπως φαίνεται στον παρακάτω πίνακα. Υπάρχει στατιστικά σημαντική εξάρτηση της μόλυνσης από τη θέση των υδάτων;

Θέση υδάτων	Παρουσία πτητικών οργανικών ενώσεων	
	Ναι	Όχι
Επιφανειακά	32	5
Υπόγεια	55	28

9.4. Η βαθμολογία σε ένα μάθημα μπορεί να κατηγοριοποιηθεί ως προς την επίδοση: Κακή (0-4), Μέτρια (5-6), Καλή (7-8) και Άριστη (9-10). Μπορεί όμως να κατηγοριοποιηθεί και με βάση τη συχνότητα των παρακολουθήσεων του μαθήματος: Πάντα, Συχνή και Σπάνια. Στον επόμενο πίνακα έχουν κατηγοριοποιηθεί με βάση αυτά τα δύο χαρακτηριστικά οι βαθμοί 130 φοιτητών σε ένα μάθημα. Κάθε αριθμός αντιστοιχεί σε μια συχνότητα. Για παράδειγμα, ο αριθμός 20 σημαίνει ότι 20 φοιτητές που παρακολουθούσαν ανελλιπώς το μάθημα πήραν βαθμούς 5 ή 6. Να εξετασθεί αν η επίδοση είναι στατιστικά ανεξάρτητη από την παρακολούθηση του μαθήματος.

Συχνότητα παρακολουθήσεων του μαθήματος	Επίδοση			
	Κακή	Μέτρια	Καλή	Άριστη
Πάντα	11	20	11	5
Συχνά	10	15	10	3
Σπάνια	24	15	5	1

9.5. Δίνονται οι συχνότητες άθλησης 26 φοιτητών. Να εξετασθεί αν υπάρχει συσχέτιση ανάμεσα στο φύλο και στην άθληση αυτής της ομάδας.

	Φύλο	
Γυμναστήριο	Φοιτητής	Φοιτήτρια
Ναι	1	11
Όχι	8	6

Κεφάλαιο 10

ΠΡΟΣΑΡΜΟΓΗ ΚΑΙ ΣΥΣΧΕΤΙΣΗ

10.1 ΓΕΝΙΚΑ

Σε ένα μεγάλο αριθμό προβλημάτων έχουμε πειραματικά δεδομένα της γενικής μορφής (x_i, y_i), όπου $i = 1, 2, ..., m$, και απαιτείται να προσδιορίσουμε τη συνάρτηση που τα περιγράφει, δηλαδή τη συνάρτηση για την οποία ισχύει $y_i \approx f(x_i)$. Η διαδικασία εύρεσης της συνάρτησης αυτής ονομάζεται ***προσαρμογή καμπύλης*** *(curve fitting)* ή ***παλινδρόμηση*** *(regression)* και είναι ιδιαίτερα χρήσιμη επειδή επιτρέπει να αντικαθίσταται ένας πίνακας δεδομένων από μια απλή εξίσωση. Επιπλέον επιτρέπει να επαληθεύεται η ισχύς ενός φυσικού νόμου και να γίνονται μαθηματικές πράξεις, όπως παραγώγιση και ολοκλήρωση των δεδομένων, που συχνά οδηγούν στον υπολογισμό άλλων χρήσιμων ιδιοτήτων του συστήματος. Η μέθοδος γενικεύεται στην περίπτωση που έχουμε δεδομένα της μορφής $(x_{1i}, x_{2i}, \ldots, x_{ni}, y_i)$ και αναζητούμε τη συνάρτηση για την οποία ισχύει $y_i \approx f(x_{1i}, x_{2i}, \ldots, x_{ni})$.

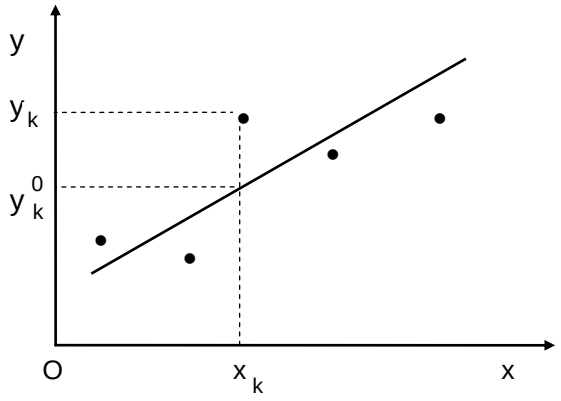
Οι μέθοδοι που χρησιμοποιούνται για να προσδιοριστεί η συνάρτηση προσαρμογής βασίζονται κυρίως στη μέθοδο των ***ελαχίστων τετραγώνων*** *(least squares)* που περιγράφεται αμέσως παρακάτω. Η μέθοδος αυτή αναπτύχτηκε για πρώτη φορά το 1805 από τον *Legendre* και ακολούθως το 1809 από τον *Gauss* για να προσδιορίσουν τις τροχιές ουράνιων σωμάτων γύρω από τον ήλιο. Ο μάλλον ατυχής όρος παλινδρόμηση για την προσαρμογή καμπύλης με τη μέθοδο των ελαχίστων τετραγώνων χρησιμοποιήθηκε από τον *Francis Galton*, ξάδελφο του *Charles Darwin*, το 1885 για να περιγράψει ένα βιολογικό φαινόμενο. Το φαινόμενο ήταν ότι τα ύψη των παιδιών ψηλών γονέων τείνουν να *παλινδρομήσουν* προς έναν κανονικό μέσο όρο.

10.2 Η ΜΕΘΟΔΟΣ ΤΩΝ ΕΛΑΧΙΣΤΩΝ ΤΕΤΡΑΓΩΝΩΝ

Στην πιο απλή της μορφή η μέθοδος των ελαχίστων τετραγώνων προσδιορίζει την *καλύτερη ευθεία* που περνά μέσα από πειραματικά σημεία (x_i, y_i), όπου $i = 1, 2, ..., m$. Για να αποφευχθεί ο υποκειμενικός παράγοντας ο *Gauss* όρισε το ακόλουθο κριτήριο για την καλύτερη ευθεία. Έστω (x_k, y_k) ένα τυχαίο πειραματικό σημείο. Τότε στην τιμή $x = x_k$ αντιστοιχεί η πειραματική τιμή y_k και η αντίστοιχη τιμή y_k^0 πάνω στην ευθεία που φέρουμε μέσα από τα πειραματικά σημεία (σχήμα 10.1). Η διαφορά $y_k - y_k^0$ ονομάζεται ***υπόλοιπο*** *(residual)* και συμβολίζεται με d_k. Ορίζουμε τώρα ως καλύτερη ευθεία που περνά μέσα από τα σημεία (x_i, y_i) εκείνη την ευθεία για την οποία το άθροισμα των τετραγώνων των υπολοίπων είναι ελάχιστο. Δηλαδή όταν

$$S = d_1^2 + d_2^2 + \cdots + d_m^2 = \sum_{i=1}^{m} d_i^2 = \text{ελάχιστο} \qquad (10.1)$$

Το κριτήριο αυτό γενικεύεται στην περίπτωση οποιασδήποτε καμπύλης. Έτσι η καλύτερη καμπύλη που περνά μέσα από τα πειραματικά σημεία (x_i, y_i) είναι εκείνη για την οποία ισχύει και πάλι η σχέση (10.1). Η καμπύλη αυτή ονομάζεται *καμπύλη ελαχίστων τετραγώνων*.



Σχήμα 10.1. Ευθεία μέσα από πειραματικά σημεία

Η μέθοδος των ελαχίστων τετραγώνων εφαρμόζεται εύκολα όταν η θεωρητική καμπύλη που περιγράφει τα πειραματικά σημεία δίνεται από μια συνάρτηση της γενικής μορφής

$$y = b_0 + b_1\varphi_1(x) + b_2\varphi_2(x) + \cdots + b_p\varphi_p(x) \qquad (10.2)$$

όπου οι συναρτήσεις $\varphi_k(x)$ δεν περιέχουν άγνωστες σταθερές. Όταν έχουμε $\varphi_k(x) = x^k$, τότε η συνάρτηση (10.2) ανάγεται σε πολυώνυμο βαθμού p:

$$y = b_0 + b_1x + b_2x^2 + \cdots + b_px^p \qquad (10.3)$$

Η εξίσωση (10.2) ονομάζεται εξίσωση ή συνάρτηση προσαρμογής ή ακόμη και ***μοντέλο προσαρμογής***, ενώ οι συντελεστές b_0, b_1, ..., b_p ονομάζονται ***προσαρμόσιμοι παράμετροι*** (*adjustable parameters*).

10.3 ΥΠΟΛΟΓΙΣΜΟΣ ΠΡΟΣΑΡΜΟΣΙΜΩΝ ΠΑΡΑΜΕΤΡΩΝ

Είναι φανερό ότι με τη μέθοδο των ελαχίστων τετραγώνων προσδιορίζονται οι συντελεστές b_0, b_1, ..., b_p της (10.2) έτσι, ώστε να ισχύει το κριτήριο (10.1). Επομένως η ποσότητα που πρέπει να ελαχιστοποιηθεί είναι η

$$S = \sum_{i=1}^{m}\left\{y_i - b_0 - b_1\varphi_1(x_i) - \cdots - b_p\varphi_p(x_i)\right\}^2 \qquad (10.4)$$

Έστω η απλή περίπτωση που η συνάρτηση προσαρμογής είναι η ευθεία

$$y = a + bx \qquad (10.5)$$

Τότε η (10.4) γίνεται

$$S = \sum_{i=1}^{m}(y_i - a - bx_i)^2 \qquad (10.6)$$

Επειδή το S είναι συνάρτηση των a και b, οι αναγκαίες συνθήκες για την ύπαρξη τοπικού ελάχιστου της S είναι: $\partial S/\partial a = 0$ και $\partial S/\partial b = 0$, από τις οποίες παίρνουμε το σύστημα των εξισώσεων

$$\left.\begin{aligned} am + b\sum x_i &= \sum y_i \\ a\sum x_i + b\sum x_i^2 &= \sum x_iy_i \end{aligned}\right\} \qquad (10.7)$$

Αν επιλύσουμε το σύστημα αυτό ως προς a και b παίρνουμε

$$a = \frac{(\sum y_i)(\sum x_i^2) - (\sum x_i)(\sum x_i y_i)}{m\sum x_i^2 - (\sum x_i)^2} \quad \text{και} \quad b = \frac{m\sum x_i y_i - (\sum x_i)(\sum y_i)}{m\sum x_i^2 - (\sum x_i)^2} \qquad (10.8)$$

Συνεπώς οι τιμές των προσαρμόσιμων παραμέτρων a και b μπορούν να υπολογιστούν από τις παραπάνω σχέσεις με βάση τα πειραματικά δεδομένα (x_i, y_i).

Αντίστοιχες σχέσεις υπάρχουν για τη γενική περίπτωση που η συνάρτηση προσαρμογής δίνεται από τη σχέση (10.2).

10.4 ΧΡΗΣΗ ΠΙΝΑΚΩΝ

Το σύστημα (10.7) με μορφή πινάκων μπορεί να γραφεί ως εξής: Έστω οι πίνακες **Y**, **X** και **b**

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ & \vdots \\ 1 & x_m \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} a \\ b \end{pmatrix} \qquad (10.9)$$

Το γινόμενο των πινάκων **X′Y** είναι

$$\mathbf{X'Y} = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_m \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix} = \begin{pmatrix} \sum y_i \\ \\ \sum x_i y_i \end{pmatrix}$$

Σε ό,τι αφορά το γινόμενο **(X′X)b** έχουμε

$$(\mathbf{X'X})\mathbf{b} =$$

$$\begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_m \end{pmatrix} \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ & \vdots \\ 1 & x_m \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} m & \sum x_i \\ \\ \sum x_i & \sum x_i^2 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} am + b\sum x_i \\ \\ a\sum x_i + b\sum x_i^2 \end{pmatrix}$$

Άρα το σύστημα (10.7) μπορεί να γραφεί ως

$$(\mathbf{X'X})\mathbf{b} = \mathbf{X'Y} \qquad (10.10)$$

Επομένως οι παράμετροι a και b υπολογίζονται από τη σχέση

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \qquad (10.11)$$

όπου $\mathbf{X}'$ είναι ο *ανάστροφος* (*transpose*) του $\mathbf{X}$ και $(\mathbf{X}'\mathbf{X})^{-1}$ είναι ο *αντίστροφος* (*inverse*) του $\mathbf{X}'\mathbf{X}$.

Αποδεικνύεται ότι η σχέση (10.11) ισχύει και για τη γενική περίπτωση που το μοντέλο προσαρμογής δίνεται από τη σχέση (10.2), όπου ο πίνακας $\mathbf{X}$ ορίζεται από τη σχέση:

$$\mathbf{X} = \begin{pmatrix} 1 & \varphi_1(x_1) & \varphi_2(x_1) & \cdots & \varphi_p(x_1) \\ 1 & \varphi_1(x_2) & \varphi_2(x_2) & \cdots & \varphi_p(x_2) \\ & \vdots & \vdots & & \vdots \\ 1 & \varphi_1(x_m) & \varphi_2(x_m) & \cdots & \varphi_p(x_m) \end{pmatrix}$$

10.5 ΣΥΝΤΕΛΕΣΤΗΣ ΣΥΣΧΕΤΙΣΗΣ ΚΑΙ ΤΥΠΙΚΕΣ ΑΠΟΚΛΙΣΕΙΣ

Για να ελέγξουμε αν η ευθεία ή γενικότερα η συνάρτηση (10.2) που προκύπτει με βάση τη μέθοδο των ελαχίστων τετραγώνων περνά πράγματι μέσα από τα πειραματικά σημεία και πόσο καλά τα περιγράφει, υπολογίζουμε το *συντελεστή συσχέτισης* (*correlation coefficient*) R από την τετραγωνική ρίζα της σχέσης

$$R^2 = 1 - \frac{\sum (y_{\theta i} - y_i)^2}{\sum (y_i - \bar{y})^2} \qquad (10.12)$$

όπου η άθροιση γίνεται από i = 1 μέχρι i = m, $\bar{y}$ είναι η μέση τιμή των πειραματικών τιμών y_i και $y_{\theta i}$ είναι οι θεωρητικές τιμές του y, δηλαδή αυτές που υπολογίζονται από τη σχέση (10.2). Όταν τα πειραματικά σημεία βρίσκονται ακριβώς επάνω στη θεωρητική καμπύλη, τότε τα y_i ταυτίζονται με τα $y_{\theta i}$ και συνεπώς ο συντελεστής συσχέτισης R παίρνει την τιμή 1. Όσο όμως αυξάνουν οι αποκλίσεις των πειραματικών σημείων από τη θεωρητική καμπύλη, δηλαδή οι αποκλίσεις των y_i από τα $y_{\theta i}$, τόσο η τιμή του R αποκλίνει από τη μονάδα και τείνει στο μηδέν. Συνεπώς ο συντελεστής συσχέτισης R είναι ένα μέτρο που δείχνει αν η θεωρητική καμπύλη περιγράφει ικανοποιητικά ή μη ικανοποιητικά τα πειραματικά σημεία. Κατά κανόνα δεχόμαστε ότι η προσαρμογή των πειραματικών σημείων από μια θεωρητική καμπύλη είναι ικανοποιητική όταν το R είναι μεγαλύτερο από 0.95.

Εκτός από το συντελεστή συσχέτισης, ιδιαίτερα χρήσιμη είναι η

τυπική απόκλιση *(standard deviation)* s_y των πειραματικών σημείων από την καμπύλη. Ορίζεται από τη σχέση

$$s_y = \sqrt{\sum (y_{\theta i} - y_i)^2/(m-p-1)} = \sqrt{\sum d_i^2/(m-p-1)} \qquad (10.13)$$

και είναι ένα μέτρο της διασποράς των πειραματικών σημείων γύρω από τη θεωρητική καμπύλη. Η s_y μπορεί να χρησιμοποιηθεί για την επιλογή του καλύτερου μοντέλου προσαρμογής, δεδομένου ότι όταν στα πειραματικά δεδομένα (x_i, y_i) προσαρμόζονται περισσότερα από ένα μοντέλα, τότε καλύτερο θεωρείται εκείνο που αντιστοιχεί στη μικρότερη τιμή της s_y.

Σε ό,τι αφορά τις τυπικές αποκλίσεις των συντελεστών a και b ή γενικότερα των b_0, b_1, ..., b_p, αποδεικνύεται ότι αν λ_{kk} (k = 0, 1, 2, ..., p) είναι τα διαγώνια στοιχεία του πίνακα $(\mathbf{X'X})^{-1}$ τότε η τυπική απόκλιση του b_k δίνεται από τη σχέση

$$s_{b_k} = s_y \sqrt{\lambda_{kk}} \qquad (10.14)$$

Η έννοια της τυπικής απόκλισης της σταθεράς b_k είναι ταυτόσημη με την έννοια της τυπικής απόκλισης σε ένα δείγμα. Επίσης αποδεικνύεται ότι το *P% διάστημα εμπιστοσύνης* για τις σταθερές b_k είναι:

$$b_k \pm t_{m-p-1,\alpha/2} s_{b_k} \qquad (10.15)$$

όπου στο *Excel* η ποσότητα $t_{m-p-1,\alpha/2}$ μπορεί να υπολογιστεί με τον τύπο =TINV(α;m-p-1).

Ενδιαφέρον παρουσιάζει και η *συνδιασπορά* μεταξύ των συντελεστών, έστω των b_k και b_j, που δίνεται από τη σχέση

$$s_{b_k b_j} = s_y^2 \lambda_{kj} \qquad (10.16)$$

Η έννοια στης συνδιασποράς έχει δοθεί στο κεφάλαιο 4.4, ενώ μια ενδιαφέρουσα εφαρμογή παρουσιάζεται στο παράδειγμα 10.5.

10.6 ΑΡΙΘΜΟΣ ΠΡΟΣΑΡΜΟΣΙΜΩΝ ΠΑΡΑΜΕΤΡΩΝ

Όταν το μοντέλο προσαρμογής, δηλαδή η σχέση (10.2), είναι γνωστή επειδή εκφράζει κάποιο φυσικοχημικό νόμο, τότε ο αριθμός των προσαρμόσιμων παραμέτρων b_0, b_1, ..., b_p είναι γνωστός και η μέθοδος των ελαχίστων τετραγώνων βοηθά στον προσδιορισμό τους. Υπάρχουν όμως περιπτώσεις όπου το μοντέλο προσαρμογής είναι μια εμπειρική

σχέση, όπως για παράδειγμα ένα πολυώνυμο βαθμού p. Σε αυτή την περίπτωση ανακύπτει το πρόβλημα του προσδιορισμού του βαθμού του πολυωνύμου, δηλαδή ποιος είναι ο βέλτιστος βαθμός του πολυωνύμου που πρέπει να χρησιμοποιηθεί για την προσαρμογή. Το ίδιο πρόβλημα μπορεί να προκύψει και στην περίπτωση που το μοντέλο προσαρμογής εκφράζει κάποιο φυσικοχημικό νόμο, αλλά θέλουμε να απορρίψουμε τις παραμέτρους που είναι στατιστικά μη σημαντικές.

Το πρόβλημα αυτό λύνεται εύκολα αν γνωρίζουμε τις τιμές p-value των προσαρμόσιμων παραμέτρων. Έστω ότι γνωρίζουμε την p-value της σταθεράς b_k. Η τιμή αυτή αφορά τον έλεγχο της μηδενικής υπόθεσης $H_0: b_k = 0$ με εναλλακτική την $H_1: b_k \neq 0$. Συνεπώς, όταν p-value < 0.05 η μηδενική υπόθεση απορρίπτεται και συνεπώς η παράμετρος b_k είναι στατιστικά σημαντική. Ένας λιγότερο αυστηρός έλεγχος είναι ο ακόλουθος. Υπολογίζουμε την απόλυτη τιμή $|t_k|$ της μεταβλητής $t_k = b_k / s_{b_k}$ και τη συγκρίνουμε με το 2. Αν ισχύει $|t_k| > 2$, η παράμετρος b_k είναι στατιστικά σημαντική. Ο έλεγχος αυτός προέρχεται από το γεγονός ότι όταν ισχύει η μηδενική υπόθεση $H_0: b_k = 0$, η μεταβλητή t_k ακολουθεί την κατανομή *Student* με m-p+1 βαθμούς ελευθερίας. Από αυτήν την ιδιότητα προκύπτει ότι σε επίπεδο σημαντικότητας $\alpha = 0.05$ και σε σχετικά μεγάλα δείγματα η κρίσιμη τιμή του t_k είναι κοντά στο 2.

Με βάση τα παραπάνω εργαζόμαστε ως εξής: Έστω ότι το μοντέλο προσαρμογής είναι δεδομένο και απλά θέλουμε να απορρίψουμε τις στατιστικά μη σημαντικές παραμέτρους του. Σε αυτή την περίπτωση υπολογίζουμε τις τιμές p-value όλων των παραμέτρων και εξετάζουμε για ποιες παραμέτρους ισχύει p-value > 0.05. Απορρίπτουμε αυτή που παρουσιάζει τη μεγαλύτερη απόκλιση από το 0.05 και ξανα-εφαρμόζουμε τη μέθοδο των ελαχίστων τετραγώνων, όπου προφανώς τώρα το μοντέλο προσαρμογής δεν περιέχει τον συντελεστή b_k που απορρίψαμε. Ελέγχουμε αν υπάρχουν συντελεστές με p-value > 0.05, απορρίπτουμε πάλι αυτόν που παρουσιάζει τη μεγαλύτερη απόκλιση από το 0.05 και η όλη διαδικασία συνεχίζεται μέχρι να παραμείνουν μόνο στατιστικά σημαντικοί συντελεστές. Η τεχνική αυτή εφαρμόζεται αυτόματα στο πρόγραμμα *LS Significant* του *ChemStat* και στο *Linear* του *SPSS.*

Αν θέλουμε να προσδιορίσουμε τον βέλτιστο βαθμό του πολυωνύμου προσαρμογής, συνήθως ξεκινάμε με ένα πολυώνυμο μικρού βαθμού, δηλαδή με τη γραμμική εξίσωση $y = a + bx$ και αυξάνουμε το βαθμό του πολυωνύμου κάθε φορά κατά ένα, δηλαδή $y = a + bx + cx^2$, $y = a + bx + cx^2 + dx^3$, … Σε κάθε πολυώνυμο ελέγχουμε αν οι σταθερές του είναι στατιστικά σημαντικές (p-value < 0.05). Στο πρώτο πολυώνυμο που θα

διαπιστώσουμε ότι έστω και μια σταθερά του δεν είναι στατιστικά σημαντική σταματάμε και επιλέγουμε ως καλύτερο πολυώνυμο το αμέσως προηγούμενο. Η διαδικασία αυτή γίνεται αυτόματα στο πρόγραμμα *LS Optimum Polynomial* του *ChemStat* χρησιμοποιώντας την επιλογή *Automatic forward*. Αν στο ίδιο πρόγραμμα επιλεγεί *Automatic backward* τότε ακολουθείται η αντίστροφη διαδικασία. Συγκεκριμένα, το πρόγραμμα ξεκινά με ένα πολυώνυμο 10 βαθμού, εφόσον οι τιμές x_i είναι περισσότερες από 11, ελέγχει τις τιμές p-value των συντελεστών και αν έστω και ένας είναι μεγαλύτερος από 0.05 ελαττώνει τον βαθμό του πολυωνύμου κατά ένα, ελέγχει τις τιμές p-value των συντελεστών και η όλη διαδικασία συνεχίζεται μέχρι να προσδιοριστεί το πολυώνυμο που έχει μόνο στατιστικά σημαντικούς συντελεστές.

Το πρόγραμμα *LS Optimum Polynomial* έχει την επιπλέον επιλογή *Manual* που δίνει όλες τις δυνατές πληροφορίες στον χρήστη για να επιλέξει αυτός το βέλτιστο πολυώνυμο. Συγκεκριμένα παρέχει τις τιμές p-value των συντελεστών των πολυωνύμων με βαθμούς 1 , 2, …, 10 και τις αντίστοιχες τιμές s_y σε δύο περιπτώσεις: όταν ο σταθερός όρος b_0 είναι μηδέν και όταν αυτός είναι διάφορος του μηδενός. Για την επιλογή του βέλτιστου πολυωνύμου διαγράφουμε όλα τα πολυώνυμα που έχουν έστω και έναν συντελεστή με p-value > 0.05 και επιλέγουμε το βέλτιστο από τα εναπομείναντα. Αυτό είναι το πολυώνυμο με την ελάχιστη τιμή s_y.

10.7 ΠΡΟΫΠΟΘΕΣΕΙΣ ΕΦΑΡΜΟΓΗΣ ΤΩΝ ΕΛΑΧΙΣΤΩΝ ΤΕΤΡΑΓΩΝΩΝ

Η εφαρμογή της μεθόδου των ελαχίστων τετραγώνων δεν απαιτεί τα δεδομένα να ακολουθούν την κανονική κατανομή. Παρόλα αυτά για να έχουν νόημα τα αποτελέσματα της προσαρμογής θα πρέπει να ελέγχεται η ύπαρξη ακραίων τιμών. Επίσης η προσαρμογή των δεδομένων θεωρείται αποδεκτή αν το διάγραμμα των υπολοίπων, δηλαδή το διάγραμμα μεταβολής των υπολοίπων με το x, δείχνει μια ομοιόμορφη κατανομή των υπολοίπων γύρω από τον άξονα των x.

Ενώ η εφαρμογή των ελαχίστων τετραγώνων δεν απαιτεί υποθέσεις για την κατανομή των δεδομένων, ο υπολογισμός διαστημάτων εμπιστοσύνης ή ο έλεγχος σημαντικότητας των προσαρμόσιμων παραμέτρων απαιτεί τα υπόλοιπα να ακολουθούν την κανονική κατανομή. Συνεπώς, αν θέλουμε να έχουμε αξιόπιστα διαστήματα εμπιστοσύνης ή αξιόπιστους ελέγχους σχετικά με τη σημαντικότητα των προσαρμόσιμων παραμέτρων, θα πρέπει να ελέγχουμε την κανονικότητα των υπολοίπων.

10.8 ΠΡΟΣΑΡΜΟΓΗ ΕΥΘΕΙΑΣ

Στην προσαρμογή ευθείας η συνάρτηση προσαρμογής είναι η σχέση y = a + bx. Μπορεί όμως η συνάρτηση προσαρμογής να μην είναι ευθεία, αλλά να μετασχηματίζεται σε ευθεία με απλή μετατροπή. Για παράδειγμα, η συνάρτηση $y=a\cdot e^{b/x}$ με λογαρίθμιση δίνει $\ln y = \ln a + b/x$, η οποία ανάγεται στην ευθεία $Y = A + BX$ αρκεί να θέσουμε $Y = \ln y$, $X = 1/x$, $A = \ln a$ και $B = b$. Στον πίνακα 10.1 συνοψίζονται μερικές βασικές περιπτώσεις αναγωγής μη γραμμικών συναρτήσεων σε γραμμικές της μορφής $Y = A + BX$.

Πίνακας 10.1. Μετατροπή συναρτήσεων σε γραμμικές της γενικής μορφής $Y = A + BX$.

Συνάρτηση	Συντελεστές της $Y = A + BX$
$y = a + b/x$	$Y = y$, $X = 1/x$, $A = a$, $B = b$
$y = a\cdot x^b$	$Y = \ln y$, $X = \ln x$, $A = \ln a$, $B = b$
$y = a + b\sqrt{x}$	$Y = y$, $X = \sqrt{x}$, $A = a$, $B = b$
$y = a\cdot e^{b/x}$	$Y = \ln y$, $X = 1/x$, $A = \ln a$, $B = b$
$y = a\cdot e^{bx}$	$Y = \ln y$, $X = x$, $A = \ln a$, $B = b$
$y = 1/(ax+b)$	$Y = 1/y$, $X = x$, $A = b$, $B = a$
$y = x/(ax+b)$	$Y = 1/y$, $X = 1/x$, $A = a$, $B = b$
$y = 1/(ax+b)^2$	$Y = 1/\sqrt{y}$, $X = x$, $A = b$, $B = a$

10.9 ΠΑΡΑΔΕΙΓΜΑΤΑ ΠΡΟΣΑΡΜΟΓΗΣ

Στην ενότητα αυτή θα εξετάσουμε μέσα από παραδείγματα μερικές από τις εφαρμογές που έχει η μέθοδος των ελαχίστων τετραγώνων. Για να την εφαρμόσουμε όμως γρήγορα και απλά απαιτείται το κατάλληλο πρόγραμμα. Τέτοια προγράμματα υπάρχουν στο *Excel*, *ChemStat* και *SPSS*.

Στο *Excel* τα ελάχιστα τετράγωνα μπορεί να εφαρμοστούν α) απ' ευθείας στη γραφική παράσταση με την εντολή *Προσθήκη γραμμής τάσης* (*Add Trendline*), β) με τη συνάρτηση *LINEST* και γ) με το πρόγραμμα *Παλινδρόμηση* (*Regression*). Από αυτές τις επιλογές η απλούστερη είναι η *Προσθήκη γραμμής τάσης*, όμως δεν δίνει στατιστικά στοιχεία. Η συνάρτηση *LINEST* δεν έχει αυτό το μειονέκτημα, αλλά τα αποτελέσματά της είναι δυσανάγνωστα, ενώ το πρόγραμμα *Παλινδρόμηση* είναι ένα σχετικά πλήρες πρόγραμμα για ελάχιστα τετράγωνα.

Το *SPSS* διαθέτει πολλές δυνατότητες εφαρμογών της μεθόδου των ελαχίστων τετραγώνων μέσα από *Analyze* → *Regression*, όπου επιλέγουμε μια συγκεκριμένη μέθοδο εφαρμογής, συνήθως *Curve Estimation* ή *Linear* ή *Nonlinear*. Το κυριότερο μειονέκτημα του *SPSS* είναι η σχετική δυσκολία με την οποία γίνονται οι γραφικές παραστάσεις.

Το *ChemStat* διαθέτει μια πλούσια βιβλιοθήκη προγραμμάτων για γενικές εφαρμογές της μεθόδου των ελαχίστων τετραγώνων αλλά και ειδικές που αφορούν τη Χημεία. Τα προγράμματα αυτά βρίσκονται στα πάνελ *Regression* και *Calibration* και λόγω της απλότητας της εφαρμογής τους χρησιμοποιούνται κατά προτεραιότητα στα παραδείγματα που ακολουθούν. Για λόγους όμως πληρότητας υπάρχουν και ενδεικτικές εφαρμογές με το *Excel* και το *SPSS*.

Παράδειγμα 10.1 – Έλεγχος φυσικού νόμου

Στον παρακάτω πίνακα δίνεται η μεταβολή της πίεσης, P, με τον όγκο, V, ενός mole Cl_2 όταν T = 300 K. Να εξετασθεί αν ισχύει η καταστατική εξίσωση των ιδανικών αερίων.

Πίνακας 10.2. Πειραματικά δεδομένα εκτόνωσης αερίου.

P, atm	3.5	2.3	1.7	1.2	0.9	0.7
V, dm^3	7	10	15	20	25	30

◆ Σε όλα τα προβλήματα ελαχίστων τετραγώνων το πρώτο που εξετάζουμε είναι ποια είναι η ανεξάρτητη μεταβλητή. Συνήθως ανεξάρτητη μεταβλητή είναι αυτή που μεταβάλουμε κατά βούληση σε ένα πείραμα. Όμως σε προβλήματα ελαχίστων τετραγώνων προτιμούμε ως ανεξάρτητη μεταβλητή να επιλέγουμε εκείνη που μετράμε με τη μεγαλύτερη ακρίβεια. Πάντως αν υπάρχει αμφιβολία, τότε επιλέγουμε ως ανεξάρτητη μεταβλητή εκείνη που μεταβάλουμε κατά βούληση. Έτσι στο συγκεκριμένο παράδειγμα επιλέγουμε ως ανεξάρτητη μεταβλητή τον όγκο.

Σε ό,τι αφορά τον έλεγχο της ισχύος μιας εξίσωσης φυσικών μεγεθών με βάση πειραματικά δεδομένα, τη μετατρέπουμε, αν μπορούμε, σε γραμμική και ελέγχουμε τη γραμμικότητα των πειραματικών σημείων σε κατάλληλο γράφημα. Στο συγκεκριμένο παράδειγμα η εξίσωση μετατρέπεται εύκολα σε γραμμική, επειδή ισχύει

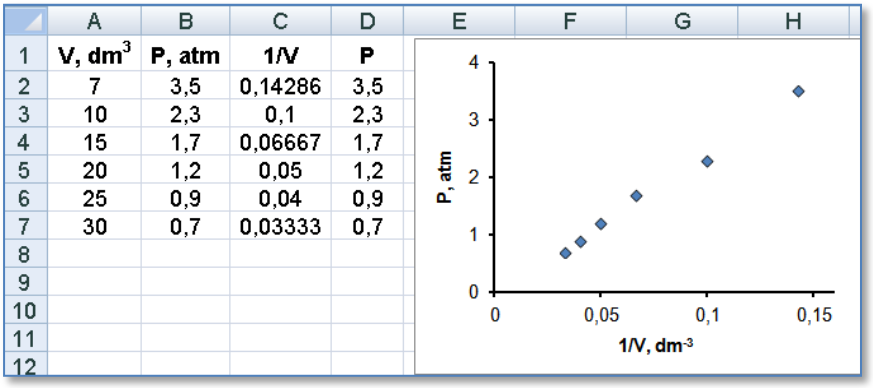
$$PV = RT \Rightarrow P = RT(1/V)$$

Συνεπώς για να ισχύει η καταστατική εξίσωση των ιδανικών αερίων πρέπει α) η γραφική παράσταση του P ως προς το 1/V να είναι γραμμική, β) η ευθεία των ελαχίστων τετραγώνων να διέρχεται από την αρχή των αξόνων και γ) η κλίση αυτής της ευθείας να είναι στατιστικά ίση με

$$RT = 0.0082*300 = 24.6 \text{ dm}^3 \text{ atm}$$

Στο σημείο αυτό καλό είναι να κάνουμε τη γραφική παράσταση του P ως προς το 1/V, ώστε να έχουμε μια πρώτη οπτική εικόνα των δεδομένων. Για το σκοπό αυτό εργαζόμαστε ως εξής:

(1) Σε ένα φύλλο εργασίας του *Excel* εισάγουμε τα αρχικά δεδομένα στις δύο πρώτες στήλες, έστω στην περιοχή A2:B7, στην επόμενη στήλη υπολογίζουμε τις τιμές 1/V χρησιμοποιώντας τον τύπο =1/A2 και τη διαδικασία της αυτόματης συμπλήρωσης και τέλος στη στήλη D επαναλαμβάνουμε τις τιμές P, επειδή για να κάνουμε εύκολα γραφικές παραστάσεις στο *Excel* πρέπει η ανεξάρτητη μεταβλητή να είναι αριστερά της εξαρτημένης (σχήμα 10.2).



Σχήμα 10.2. Πειραματικά σημεία και η γραφική παράσταση P ως προς 1/V

(2) Κάνουμε τη γραφική παράσταση της μεταβολής του P με το *1/*V, επιλέγοντας την περιοχή C2:D7 και κάνοντας *κλικ* στο *Εισαγωγή → Γραφήματα → Διασπορά → Διασπορά μόνο με δείκτες (Insert → Charts → Scatter → Scatter with only Markers)*.

(3) Μορφοποιούμε το διάγραμμα, κυρίως σε ότι αφορά τους άξονες. Για να αλλάξουμε την κλίμακα του άξονα των x κάνουμε δεξί *κλικ* πάνω σε έναν αριθμό του άξονα και στον κατάλογο εντολών που ανοίγει επιλέγουμε

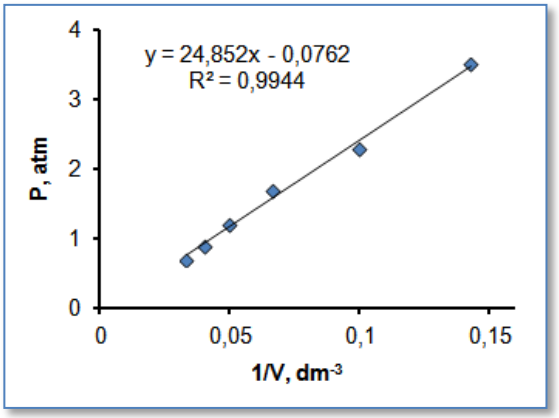
το *Μορφοποίηση Άξονα* (*Format Axis)*. Ακολούθως στο *Επιλογές Άξονα* (*Axis Options)* συμπληρώνουμε στα αντίστοιχα πλαίσια κειμένου την ελάχιστη και μέγιστη τιμή, αφού προηγουμένως καταστήσουμε ενεργές τις επιλογές *Σταθερό* (*Fixed)*. Επίσης εισάγουμε τίτλους αξόνων. Για το σκοπό αυτό κάνουμε *κλικ* πάνω στη γραφική παράσταση και από τη λωρίδα εντολών *Διάταξη* (*Layout),* πηγαίνουμε στο πάνελ *Τίτλοι* (*Labels)* και από το εικονίδιο *Τίτλοι Άξονα* (*AxisTitles)* επιλέγουμε να τοποθετήσουμε έναν οριζόντιο και έναν κάθετο τίτλο στους άξονες (σχήμα 10.2).

Από τη γραφική παράσταση παρατηρούμε ότι πράγματι το P μεταβάλλεται γραμμικά ως προς το 1/V και συνεπώς μπορούμε να περάσουμε να εφαρμόσουμε τη μέθοδο των ελαχίστων τετραγώνων για να προσδιορίσουμε τις σταθερές του γραμμικού μοντέλου y = a + bx. Στο μοντέλο αυτό το a πρέπει να είναι στατιστικά ίσο με 0 και η κλίση b στατιστικά ίση με 24.6.

❖ Ανάλυση στο Excel – Προσθήκη γραμμής τάσης

Για να προσθέσουμε την ευθεία των ελαχίστων τετραγώνων, κάνουμε δεξί *κλικ* σε ένα από τα πειραματικά σημεία και στη λίστα επιλογών που εμφανίζεται επιλέγουμε το *Προσθήκη γραμμής τάσης* (*Add Trendline*). Ανοίγει το αντίστοιχο παράθυρο διαλόγου και από το *Επιλογές γραμμής τάσης* (*Trendline Options)* επιλέγουμε το *Γραμμική* (*Linear*) και κάνουμε *κλικ* στο *Προβολή εξίσωσης στο γράφημα* (*Display equation on chart*) και *Προβολή τιμής R-τετράγωνο στο γράφημα* (*Display R-squared on chart*). Αν θέλαμε να κάνουμε προσαρμογή στο μοντέλο y = bx, θα έπρεπε να ενεργοποιήσουμε την επιλογή *Ορισμός σημείου τομής (Set Intercept)* και στο αντίστοιχο πεδίο να πληκτρολογήσουμε 0 αν η τιμή αυτή δεν είναι προεπιλεγμένη.

Στο σχήμα 10.3 παρατηρούμε ότι η ευθεία των ελαχίστων τετραγώνων είναι η y = 24.852x – 0.0762 με R^2 = 0.9944. Παρατηρούμε επίσης ότι πράγματι η μεταβολή του P με το 1/V είναι γραμμική, πρέπει όμως να ελέγξουμε αν η τεταγμένη στην αρχή, a = -0.0762, είναι στατιστικά ίση με το μηδέν και η κλίση, b = 24.852, στατιστικά ίση με 24.6 dm^3 atm. Το πρόγραμμα *Προσθήκη γραμμής τάσης* δεν παρέχει στατιστικά στοιχεία. Για το σκοπό αυτό μπορούμε να χρησιμοποιήσουμε το πρόγραμμα *Παλινδρόμηση*.



Σχήμα 10.3. Η γραφική παράσταση P ως προς 1/V και η ευθεία των ελαχίστων τετραγώνων

❖ **Ανάλυση στο Excel – Παλινδρόμηση**

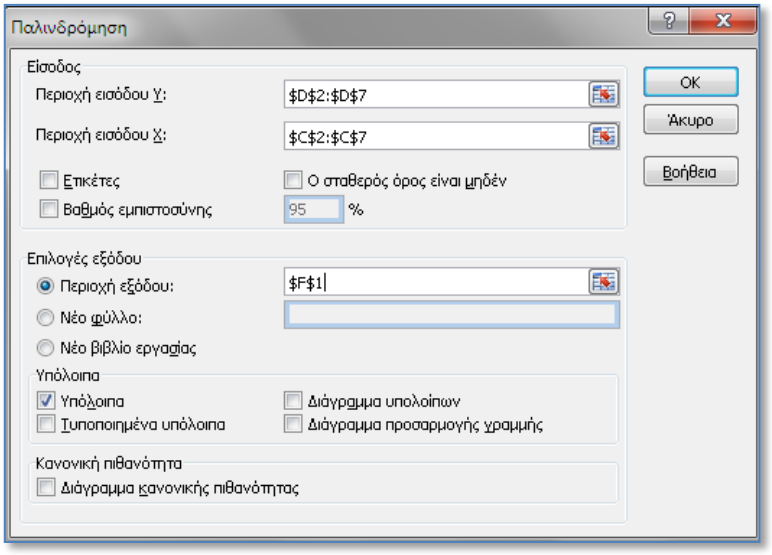
Για να εφαρμόσουμε το πρόγραμμα *Παλινδρόμηση* (*Regression*) χρησιμοποιούμε τη διευθέτηση των δεδομένων που υπάρχει στο σχήμα 10.2 και συνεχίζουμε με τα παρακάτω βήματα :

(1) Από το *Δεδομένα → Ανάλυση* (*Data → Analysis)* επιλέγουμε το *Ανάλυση Δεδομένων* (*Data Analysis)* και στον κατάλογο που ανοίγει κάνουμε *διπλό κλικ* στο *Παλινδρόμηση* (*Regression*).

(2) Στην οθόνη παρουσιάζεται ένα μεγάλο παράθυρο διαλόγου με τίτλο *Παλινδρόμηση*, που συμπληρώνεται ως εξής (σχήμα 10.4): Κάνουμε *κλικ* στο πλαίσιο κειμένου που υπάρχει στην *Περιοχή εισόδου Y* (*Input Y Range*) και με το ποντίκι επιλέγουμε την περιοχή τιμών του y, δηλαδή την D2:D7, και με τον ίδιο τρόπο εισάγουμε την περιοχή τιμών του x, C2:C7, στην *Περιοχή εισόδου X* (*Input X Range*).

(3) Αφήνουμε ως *Βαθμό εμπιστοσύνης* (*Confidence Level*) το 95% για τα διαστήματα εμπιστοσύνης των προσαρμόσιμων παραμέτρων. Επίσης δεν ενεργοποιούμε την επιλογή *Ο σταθερός όρος είναι μηδέν (Constant is Zero),* επειδή θέλουμε να ελέγξουμε αν πράγματι ο όρος αυτός είναι μηδέν.

(4) Ορίζουμε την έξοδο, έστω το κελί F1, αφού προηγουμένως έχουμε κάνει *κλικ* στο κουμπί επιλογής *Περιοχή εξόδου* (*Output Range*). Τέλος, κάνουμε *κλικ* στο κουμπί επιλογής *Υπόλοιπα* (*Residuals*) και *κλικ* στο *ΟΚ*.



Σχήμα 10.4. Συμπλήρωση παραθύρου *Παλινδρόμηση*

Τα αποτελέσματα του προγράμματος *Παλινδρόμηση* εμφανίζονται στην οθόνη υπό μορφή πίνακα από τη στήλη F μέχρι την N και από την πρώτη γραμμή μέχρι την τριακοστή (σχήμα 10.5). Ο πίνακας αυτός παρέχει έναν υπερβολικά μεγάλο αριθμό πληροφοριών στατιστικού περιεχομένου, που όμως πολλές μας είναι ουσιαστικά άχρηστες. Οι χρήσιμες πληροφορίες είναι αυτές που αναφέρονται στην τιμή των προσαρμόσιμων παραμέτρων, των τυπικών τους αποκλίσεων και των παραμέτρων R^2 και s_y. Αν συμβολίσουμε την ευθεία προσαρμογής ως $y = a + bx$, από τον πίνακα παίρνουμε τα αποτελέσματα:

$$a = -0.076 \pm 0.077, \quad b = 24.85 \pm 0.94$$

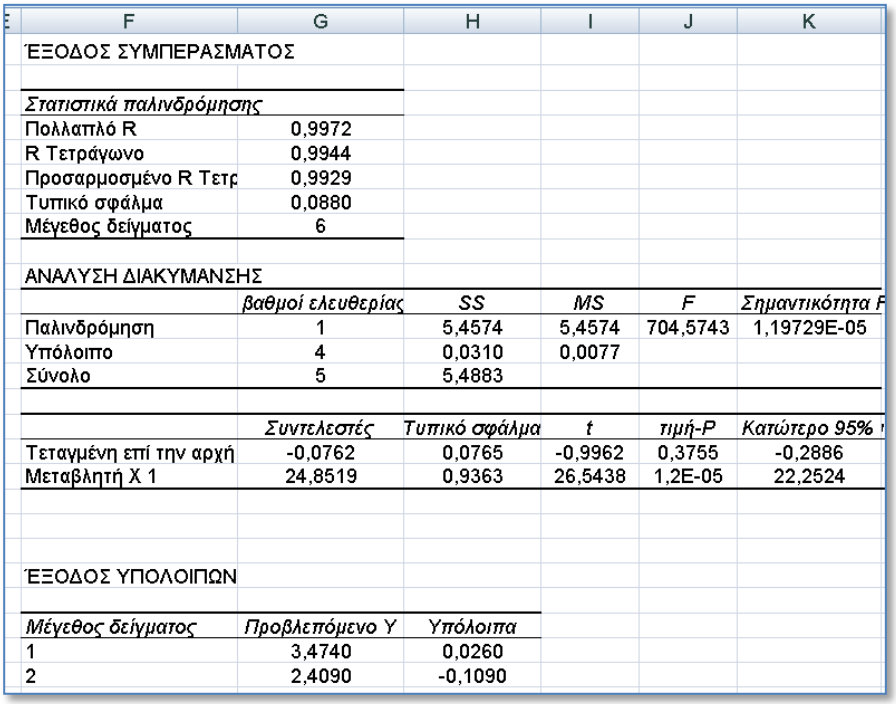
ή ορθότερα

$$a = -0.08 \pm 0.08, \quad b = 24.9 \pm 0.9$$

και

$$R^2 = 0.9944, \quad s_y = 0.088$$

Ενδιαφέρον παρουσιάζουν και οι τιμές p-value για τις σταθερές a και b. Αυτές βρίσκονται στη στήλη τιμή-P. Παρατηρούμε ότι για τη σταθερά a ισχύει p-value = 0.376 > 0.05. Άρα η σταθερά αυτή είναι στατιστικά μη σημαντική και μπορεί να απαλειφθεί. Αντίθετα η σταθερά b είναι στατιστικά σημαντική δεδομένου ότι p-value = 1.2×10^{-5} < 0.05.



ΕΞΟΔΟΣ ΣΥΜΠΕΡΑΣΜΑΤΟΣ

Στατιστικά παλινδρόμησης	
Πολλαπλό R	0,9972
R Τετράγωνο	0,9944
Προσαρμοσμένο R Τετρ	0,9929
Τυπικό σφάλμα	0,0880
Μέγεθος δείγματος	6

ΑΝΑΛΥΣΗ ΔΙΑΚΥΜΑΝΣΗΣ

	βαθμοί ελευθερίας	*SS*	*MS*	*F*	*Σημαντικότητα F*
Παλινδρόμηση	1	5,4574	5,4574	704,5743	1,19729E-05
Υπόλοιπο	4	0,0310	0,0077		
Σύνολο	5	5,4883			

	Συντελεστές	*Τυπικό σφάλμα*	*t*	*τιμή-P*	*Κατώτερο 95%*
Τεταγμένη επί την αρχή	-0,0762	0,0765	-0,9962	0,3755	-0,2886
Μεταβλητή X 1	24,8519	0,9363	26,5438	1,2E-05	22,2524

ΕΞΟΔΟΣ ΥΠΟΛΟΙΠΩΝ

Μέγεθος δείγματος	*Προβλεπόμενο Y*	*Υπόλοιπα*
1	3,4740	0,0260
2	2,4090	-0,1090

Σχήμα 10.5. Τμήμα της οθόνης με αποτελέσματα του προγράμματος *Παλινδρόμηση*

Συνεπώς θα πρέπει να ξανα-εφαρμόσουμε τη μέθοδο των ελαχίστων τετραγώνων χρησιμοποιώντας τώρα το μοντέλο y = bx και στη συνέχεια να ελέγξουμε αν η κλίση b είναι στατιστικά ίση με 24.6 dm^3 atm. Για το σκοπό αυτό ακολουθούμε και πάλι την παραπάνω πορεία εφαρμογής του προγράμματος *Παλινδρόμηση*, όμως τώρα ενεργοποιούμε την επιλογή *Ο σταθερός όρος είναι μηδέν (Constant is Zero).* Από τον πίνακα αποτελεσμάτων παίρνουμε:

$$b = 24.028 \pm 0.44 \quad \text{ή ορθότερα } b = 24.0 \pm 0.4$$

ενώ από τον ίδιο πίνακα αποτελεσμάτων παίρνουμε ότι το 95% διάστημα εμπιστοσύνης της κλίσης b είναι το [22.9, 25.2]. Παρατηρούμε ότι η θεωρητική τιμή 24.6 είναι μέσα στο διάστημα αυτό και συνεπώς όλα τα τεστ συγκλίνουν στο συμπέρασμα ότι τα πειραματικά δεδομένα περιγράφονται από την καταστατική εξίσωση των ιδανικών αερίων.

❖ **Ανάλυση στο ChemStat**

Στο *ChemStat* τα δεδομένα πρέπει να βρίσκονται σε στήλες με τη στήλη των τιμών του x αριστερά και του y δεξιά. Οι συντελεστές προσαρμογής, οι τυπικές τους αποκλίσεις, τα διαστήματα εμπιστοσύνης, οι τιμές |t| και p-value και ο πίνακας συνδιασποράς (variance-covariance) των συντελεστών προσαρμογής παρουσιάζονται κάτω από την περιοχή των δεδομένων x-y. Δεξιά της στήλης y παρουσιάζονται οι υπολογιζόμενες τιμές του y και δεξιά αυτής της στήλης παρουσιάζονται τα υπόλοιπα. Για να εφαρμόσουμε το πρόγραμμα πηγαίνουμε

Πρόσθετα → *ChemStat* → *Regression* → *LS Polynomial*

Στα πλαίσια που ανοίγουν εισάγουμε τις τιμές της ανεξάρτητης μεταβλητής, δηλαδή την περιοχή C2:C7, ορίζουμε ως βαθμό πολυωνύμου το 1, αφήνουμε ως διάστημα εμπιστοσύνης το 95%, ακολούθως στο παράθυρο με το μήνυμα "If constant term is zero enter 0 else enter 1" πληκτρολογούμε 1 για να έχει σταθερό όρο η ευθεία προσαρμογής, ενώ στο συγκεκριμένο πρόβλημα δεν είναι απαραίτητο να πάρουμε τον πίνακα της *συνδιασποράς*.

	A	B	C	D	E	F
1	**V, dm^3**	**P, atm**	**1/V**	**P**	y(calc)	d=y-y(calc)
2	7	3,5	0,14286	3,5	3,4740	0,0260
3	10	2,3	0,1	2,3	2,4090	-0,1090
4	15	1,7	0,06667	1,7	1,5806	0,1194
5	20	1,2	0,05	1,2	1,1664	0,0336
6	25	0,9	0,04	0,9	0,9179	-0,0179
7	30	0,7	0,03333	0,7	0,7522	-0,0522
8						
9				c0	c1	
10			c(i) =	-0,0762	24,8519	
11			St.Dev.=	0,0765	0,9363	
12			95% =	0,2124	2,5991	
13			|t| =	0,9962	26,5438	
14			p-value =	0,3755	1,2E-05	
15			r =	0,9972		
16			sy =	0,0880		
17			Type of fitting polynomial:			
18			y = c0 + c1*x + c2*x^2 + ...			

Σχήμα 10.6. Αποτελέσματα από την προσαρμογή στην ευθεία των ελαχίστων τετραγώνων με το πρόγραμμα *LS polynomial*

Στον πίνακα των αποτελεσμάτων (σχήμα 10.6) το μοντέλο προσαρμογής συμβολίζεται γενικά με y = c0 + c1*x + c2*x^2 + … Συνεπώς, η ευθεία των ελαχίστων τετραγώνων είναι η

$$y = -0.0762 + 24.8519x$$

όπου η σταθερά -0.0762 είναι στατιστικά μη σημαντική, επειδή p-value = 0.3755 > 0.05 που σημαίνει ότι η μηδενική υπόθεση H_0:a = 0 δεν μπορεί να απορριφθεί.

Συνεπώς πρέπει να προσαρμόσουμε τα δεδομένα στο μοντέλο y = bx και στη συνέχεια να ελέγξουμε αν η κλίση είναι στατιστικά ίση με 24.6 dm^3 atm. Έτσι επαναλαμβάνουμε την προηγούμενη διαδικασία, αλλά στο παράθυρο με το μήνυμα "If constant term is zero enter 0 else enter 1" πληκτρολογούμε 0. Παίρνουμε τον πίνακα 10.7 από τον οποίο προκύπτει ότι η ευθεία των ελαχίστων τετραγώνων είναι η

$$y = (24.0 \pm 0.4)\, x$$

Επιπλέον η τιμή 24.6 είναι μέσα στο 95% διάστημα εμπιστοσύνης της κλίσης b που είναι το [24.03-1.13=22.9, 24.03+1.13=25.2]. Άρα και πάλι συμπεραίνουμε ότι τα πειραματικά δεδομένα περιγράφονται από την καταστατική εξίσωση των ιδανικών αερίων.

	c1
c(i) =	24.0284
St.Dev.=	0.43936
95% =	1.12917
|t| =	54.689
p-value =	3.9E-08
r =	0.99647
sy =	0.08794
Type of fitting polynomial:	
y = c1*x + c2*x^2 + ...	

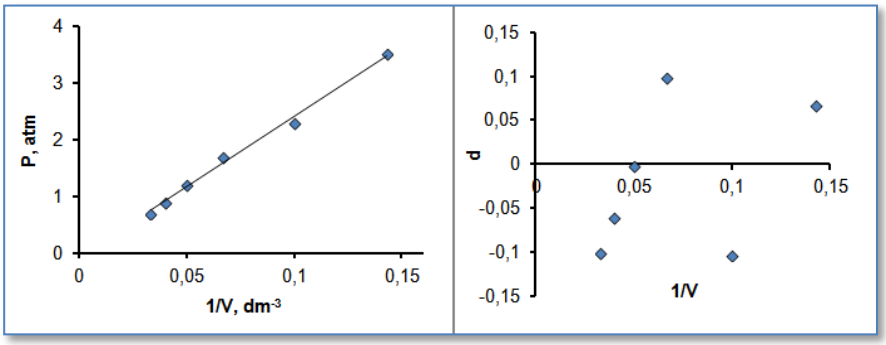
Σχήμα 10.7. Αποτελέσματα *ChemStat* από την προσαρμογή στην ευθεία y = bx

Για να προσθέσουμε την ευθεία των ελαχίστων τετραγώνων στο γράφημα του σχήματος 10.2 επιλέγουμε την περιοχή E2:E7, την αντιγράφουμε με Ctrl + C, κάνουμε *κλικ* στο εσωτερικό του γραφήματος του σχήματος 10.2 και επικολλάμε τα σημεία της στήλης E με Ctrl + V. Ακολούθως, επιλέγουμε ένα από αυτά τα σημεία, κάνουμε δεξί *κλικ* και επιλέγουμε *Μορφοποίηση σειράς δεδομένων*. Στο παράθυρο που ανοίγει πηγαίνουμε στο *Επιλογή δείκτη* και ενεργοποιούμε το *Κανένας*, ενώ από το *Χρώμα* και *Στυλ γραμμής* καθορίζουμε την εμφάνιση της ευθείας των ελαχίστων τετραγώνων (σχήμα 10.8-αριστερά).

Τέλος, θα πρέπει να αναφερθεί ότι όλοι οι παραπάνω έλεγχοι σημαντικότητας ισχύουν με την προϋπόθεση ότι τα υπόλοιπα ακολουθούν την κανονική κατανομή και είναι ομοιόμορφα διεσπαρμένα εκατέρωθεν του

άξονα των x. Το διάγραμμα των υπολοίπων γίνεται εύκολα αν επιλέξουμε την περιοχή C2:C7;F2:F7 και κάνουμε *κλικ* στο *Εισαγωγή* → *Γραφήματα* → *Διασπορά* → *Διασπορά μόνο με δείκτες* (σχήμα 10.8-δεξιά). Αν και τα σημεία είναι λίγα, παρατηρούμε μια σχετικά ομοιόμορφη διασπορά γύρω από τον άξονα των x.

Σε ό,τι αφορά την κανονικότητα των υπολοίπων, δηλαδή των τιμών της περιοχής F2:F7, ο έλεγχος με το κριτήριο *Shapiro-Wilk* στο *SPSS* δίνει στην τιμή p-value = 0.978 > 0.05 και δείχνει ότι δεν διαπιστώνονται στατιστικά σημαντικές αποκλίσεις από την κανονικότητα.

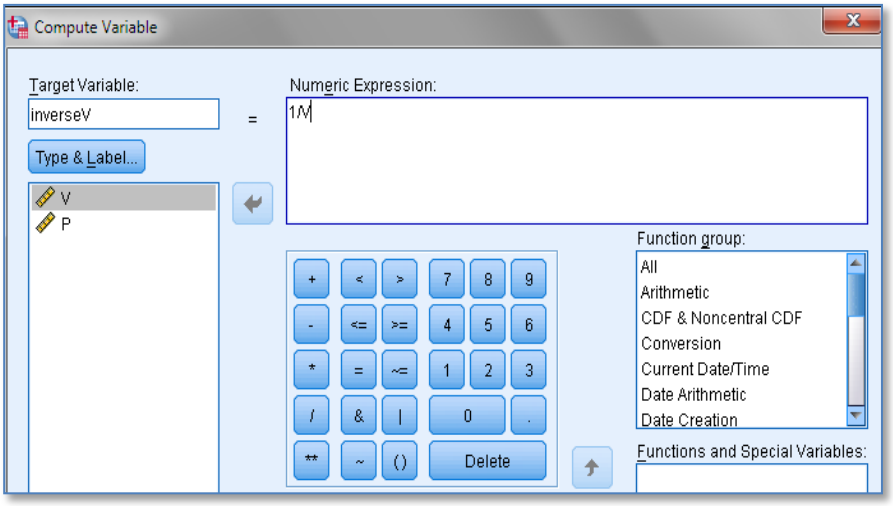


Σχήμα 10.8. Η ευθεία των ελαχίστων τετραγώνων (αριστερά) και το διάγραμμα υπολοίπων (δεξιά)

❖ Ανάλυση στο SPSS

Τοποθετούμε τα δεδομένα V και P σε δύο στήλες και δημιουργούμε μια νέα στήλη/μεταβλητή με όνομα inverseV και τιμές 1/V ως εξής. Πηγαίνουμε *Transform* → *Compute Variable* και συμπληρώνουμε το παράθυρο που ανοίγει όπως στο σχήμα 10.9. Ακολούθως από το βασικό παράθυρο *Variable View* μορφοποιούμε τις μεταβλητές και συμπληρώνουμε τη στήλη *Label* όπως φαίνεται στο σχήμα 10.10.

Μετά τη δημιουργία της μεταβλητής 1/V πηγαίνουμε *Analyze* → *Regression* → *Curve Estimation* και εισάγουμε τη μεταβλητή P στο πλαίσιο *Dependent(s)* και τη μεταβλητή 1/V στο *Independent Variable*, όπως φαίνεται στο σχήμα 10.11. Επίσης επιλέγουμε *Display ANOVA table, Linear, Plot models* και *Include constant in equation*. Είναι προφανές ότι αν το μοντέλο προσαρμογής δεν είναι γραμμικό μπορούμε να επιλέξουμε ένα άλλο μοντέλο από αυτά που διαθέτει το πρόγραμμα, π.χ. *Quadratic (πολυώνυμο δευτέρου βαθμού)*, *Cubic (πολυώνυμο τρίτου βαθμού)*, κ.ο.κ.



Σχήμα 10.9. Τμήμα παραθύρου δημιουργίας της μεταβλητής inverseV με τιμές 1/V

File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons

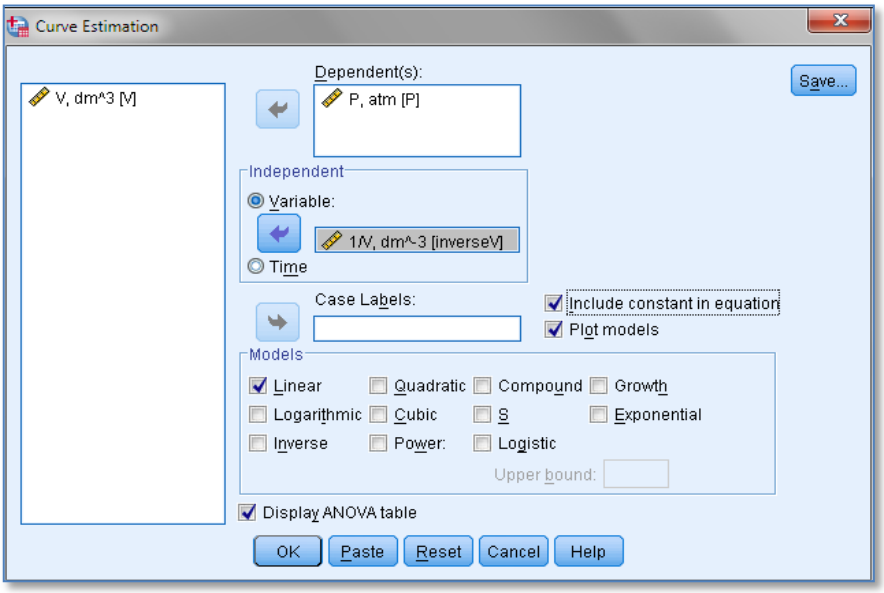
	Name	Type	Width	Decimals	Label
1	V	Numeric	8	0	V, dm^3
2	P	Numeric	8	1	P, atm
3	inverseV	Numeric	8	4	1/V, dm^-3
4					
5					

Σχήμα 10.10. Μορφοποίηση μεταβλητών και συμπλήρωση της στήλης *Label*

Με *κλικ* στο *OK* παίρνουμε αρκετούς πίνακες από τους οποίους ενδιαφέρον έχει ο *Coefficients*, που δίνεται στο σχήμα 10.12. Ο πίνακας αυτός μας δίνει τις τιμές a και b του μοντέλου προσαρμογής (y = a + bx) στη στήλη B, στη διπλανή στήλη είναι οι τυπικές αποκλίσεις αυτών των σταθερών και στην τελευταία στήλη οι τιμές p-value ως Sig. Όπως αναμένεται, τα αποτελέσματα ταυτίζονται με αυτά που προσδιορίστηκαν με το *Excel* και το *ChemStat*.

Εφόσον ο σταθερός όρος εμφανίζεται να μην είναι στατιστικά σημαντικός (Sig. = 0.375 > 0.05) επαναλαμβάνουμε την παραπάνω διαδικασία, όμως στο παράθυρο επιλογής μοντέλου προσαρμογής, *Curve Estimation*, απενεργοποιούμε την επιλογή *Include constant in equation*.

Από τον πίνακα *Coefficients* προκύπτει ότι στο απλό μοντέλο y = bx ισχύει b = 24.0 ± 0.4 (σχήμα 10.13). Όμως το πρόγραμμα δεν υπολογίζει το 95% διάστημα εμπιστοσύνης των προσαρμόσιμων παραμέτρων και συνεπώς δεν μπορούμε να είμαστε βέβαιοι ότι η τιμή 24.6 είναι συμβατή με την κλίση b = 24.0 ± 0.4.



Σχήμα 10.11. Παράθυρο επιλογής μοντέλου προσαρμογής στο *SPSS*

Coefficients

	Unstandardized Coefficients		Standardized Coefficients		
	B	Std. Error	Beta	t	Sig.
1/V	24,844	,932	,997	26,653	,000
(Constant)	-,076	,076		-,998	,375

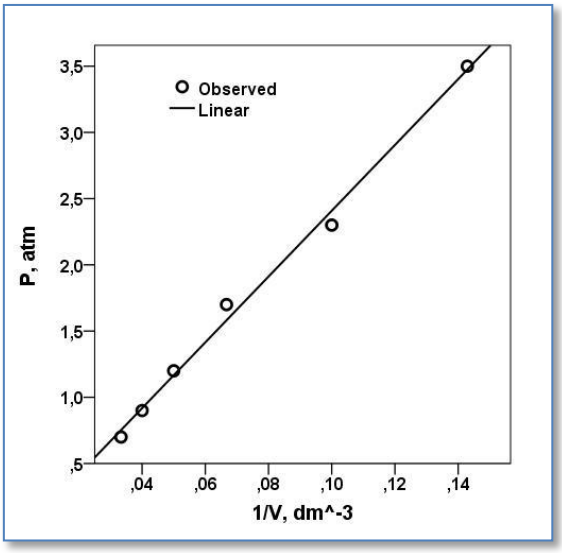
Σχήμα 10.12. Πίνακας προσαρμόσιμων παραμέτρων

Για να κάνουμε τη γραφική παράσταση P ως προς 1/V με την ευθεία των ελαχίστων τετραγώνων πρέπει να έχουμε ενεργοποιήσει την επιλογή *Plot Models* στο παράθυρο *Curve Estimation* (σχήμα 10.11). Όμως η μορφοποίησή της είναι εξαιρετικά δύσκολη (σχήμα 10.14).

Coefficients

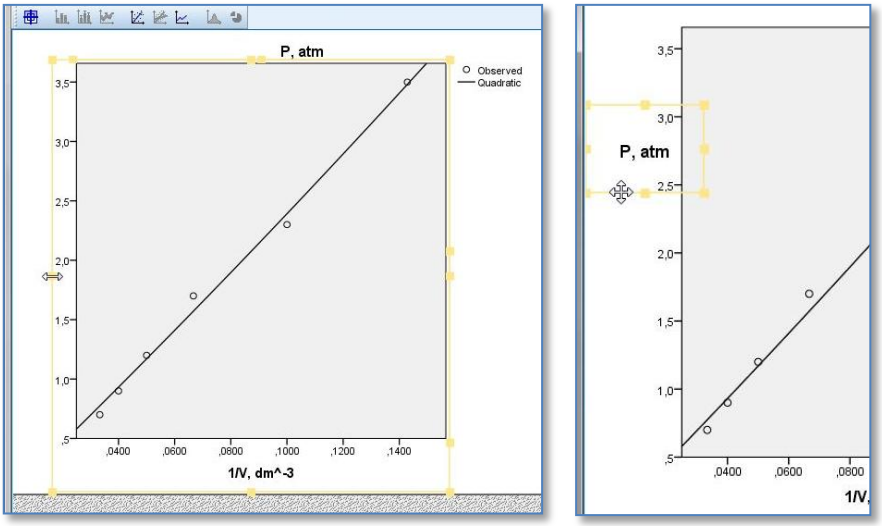
	Unstandardized Coefficients		Standardized Coefficients		
	B	Std. Error	Beta	t	Sig.
1/V, dm^-3	24,028	,439	,999	54,689	,000

Σχήμα 10.13. Πίνακας προσαρμόσιμων παραμέτρων στο μοντέλο y= bx



Σχήμα 10.14. Η γραφική παράσταση P ως προς 1/V και η ευθεία των ελαχίστων τετραγώνων

Για να μορφοποιήσουμε γενικά μια γραφική παράσταση κάνουμε διπλό *κλικ* επάνω της, οπότε ανοίγει ο επεξεργαστής γραφικών παραστάσεων (*Chart Editor*) για να κάνουμε τις μεταβολές που θέλουμε. Στη συγκεκριμένη γραφική παράσταση θα πρέπει πρώτα να μετακινήσουμε τη γραφική παράσταση δεξιά, ώστε να δημιουργήσουμε χώρο για τον τίτλο του κάθετου άξονα. Για το σκοπό αυτό κάνουμε *κλικ* στο κάτω μέρος της γραφικής παράστασης, έξω από τον άξονα των x, ώστε να επιλεγεί η γραφική παράσταση. Ακολούθως φέρνουμε τον κέρσορα σε μία από τις κεντρικές *λαβές* που εμφανίζονται στην αριστερή κάθετη γραμμή επιλογής και όταν ο κέρσορας μετατραπεί σε διπλό βέλος τον σέρνουμε προς τα δεξιά με συνεχώς πατημένο το πλήκτρο του ποντικιού μέχρι να δημιουργηθεί ο επιθυμητός χώρος (σχήμα 10.15-αριστερά).

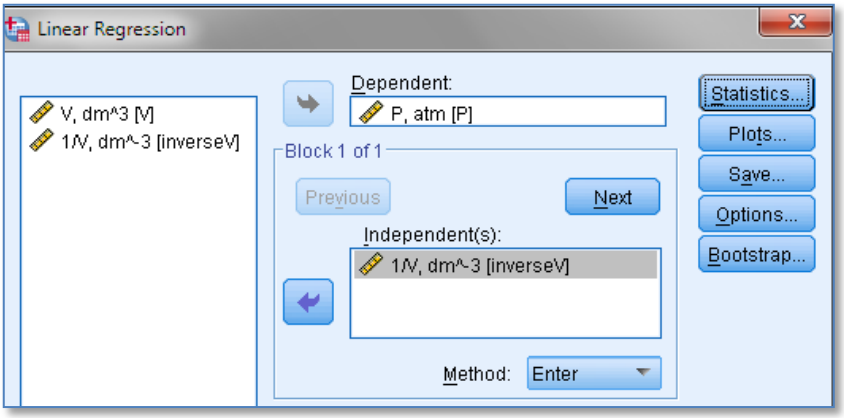


Σχήμα 10.15. Βήματα μορφοποίησης γραφικής παράστασης

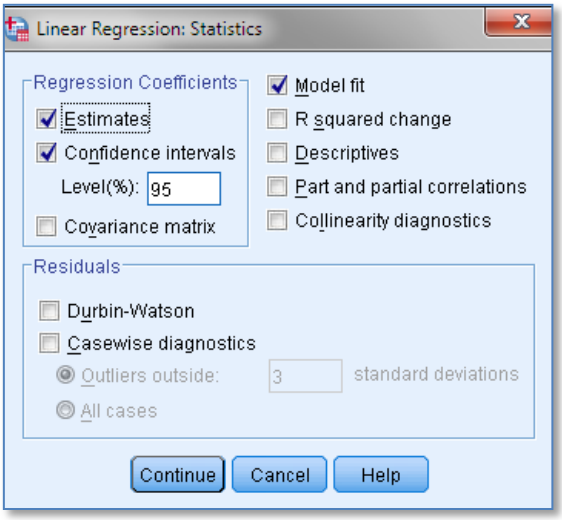
Ακολούθως επιλέγουμε το πλαίσιο του τίτλου "P, atm", το περιορίζουμε γύρω από τον τίτλο και το μετακινούμε προς τον χώρο που δημιουργήσαμε (σχήμα 10.15-δεξιά). Για να περιστρέψουμε τον τίτλο, πηγαίνουμε στο πλαίσιο *Properties* στην καρτέλα *Text Layout* και επιλέγουμε *Bottom up*. Κάθε φορά που κάνουμε μια αλλαγή πρέπει να πατάμε το κουμπί *Apply*. Από τη καρτέλα *Text Style* επιλέγουμε την κατάλληλη γραμματοσειρά και πατάμε *Apply*. Γενικά, για να μορφοποιήσουμε ένα οποιοδήποτε σχεδιαστικό στοιχείο κάνουμε *κλικ* στο στοιχείο αυτό και το μορφοποιούμε από το πλαίσιο *Properties*. Ιδιαίτερο πρόβλημα υπάρχει αν θέλουμε να τοποθετήσουμε δείκτες ή εκθέτες. Αυτό μπορεί να γίνει μόνο με την εισαγωγή του δείκτη/εκθέτη σε πλαίσιο κειμένου από *Options* → *Text Box* και μεταφορά του πλαισίου στη θέση του δείκτη/εκθέτη. Για να φύγουμε από τον επεξεργαστή κάνουμε *κλικ* στο εικονίδιο ή πηγαίνουμε *File* → *Close.*

Όπως αναφέραμε, το πρόγραμμα *Curve Estimation* δεν προσδιορίζει διαστήματα εμπιστοσύνης για τις προσαρμόσιμες παραμέτρους. Αυτό μπορεί να γίνει από *Analyze* → *Regression* → *Linear*. Στο παράθυρο *Linear Regression* που ανοίγει εισάγουμε τη μεταβλητή P στο πλαίσιο *Dependent*, τη μεταβλητή 1/V στο *Independent(s)* και στο *Method* επιλέγουμε *Enter*, όπως φαίνεται στο σχήμα 10.16. Στο *Method* υπάρχουν ενδιαφέρουσες επιλογές που θα τις εξετάσουμε στο παράδειγμα 10.4. Συνεχίζουμε με *κλικ*

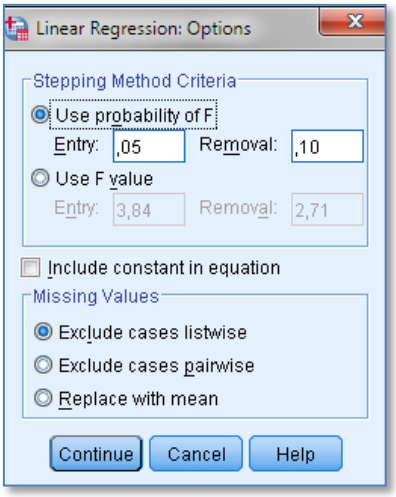
στο *Statistics* και στο πλαίσιο που ανοίγει ορίζουμε τα διαστήματα εμπιστοσύνης για τις προσαρμόσιμες παραμέτρους ενεργοποιώντας την επιλογή *Confidence intervals* με *Level 95* (σχήμα 10.17). Επίσης από το κουμπί *Options* στο βασικό παράθυρο της μεθόδου απενεργοποιούμε την επιλογή *Include constant in equation* (σχήμα 10.18).



Σχήμα 10.16. Τμήμα παραθύρου εισαγωγής δεδομένων στο πρόγραμμα *Linear* του *SPSS*



Σχήμα 10.17. Επιλογή διαστημάτων εμπιστοσύνης



Σχήμα 10.18. Επιλογή σταθερού όρου μηδέν στο μοντέλο προσαρμογής

Το πρόγραμμα έχει δυνατότητες γραφικών παραστάσεων από το κουμπί *Plots* στο βασικό παράθυρο τη μεθόδου, όμως καμία από αυτές δεν είναι αντίστοιχη του σχήματος 10.14. Όταν ολοκληρώσουμε όλες τις ρυθμίσεις πατάμε *OK* και πηγαίνουμε στον πίνακα *Coefficients* (σχήμα 10.19). Σε αυτόν υπάρχουν οι τιμές των προσαρμόσιμων παραμέτρων, οι τυπικές τους αποκλίσεις, οι τιμές p-value και τα διαστήματα εμπιστοσύνης κάθε παραμέτρου. Παρατηρούμε ότι για την κλίση b το 95% διάστημα εμπιστοσύνης είναι [22.9, 25.2], σε πλήρη συμφωνία με τα αποτελέσματα του *Excel* και του *ChemStat*.

Coefficients

Model	Unstandardized Coefficients		t	Sig.	95,0% Confidence Interval for B	
	B	Std. Error			Lower Bound	Upper Bound
1/V, dm^-3	24,028	,439	54,689	,000	22,899	25,158

Σχήμα 10.19. Τμήμα του πίνακα των προσαρμόσιμων παραμέτρων στο μοντέλο y= bx

Παράδειγμα 10.2 – Έλεγχος φυσικού νόμου

Στον πίνακα 10.3 δίνεται η μεταβολή της ποσότητας x/m του CH_3COOH που προσροφάται ανά γραμμάριο προσροφητικού από υδατικό διάλυμα σε συνάρτηση με τη συγκέντρωση c του CH_3COOH. Να εξετάσετε αν ισχύει η εξίσωση

$$x/m = K\, c^{1/n} \text{ (ισόθερμη Freudlich)}$$

και αν ισχύει να προσδιορίστε τις σταθερές K και n.

Πίνακας 10.3. Πειραματικά δεδομένα προσρόφησης CH_3COOH.

c, mol dm^{-3}	0.1	0.2	0.3	0.5	1.0
$(x/m)\times10^3$, mol g^{-1}	3.9	4.2	4.4	4.6	5.0

◆ Όπως έχει αναφερθεί στο κεφάλαιο 1 (Πίνακας 1.5) το σύμβολο $(x/m)\times10^3$ σημαίνει ότι οι τιμές 3.9, 4.2, … έχουν πολλαπλασιαστεί επί 1000. Συνεπώς οι τιμές του πίνακα 10.3 που αντιστοιχούν στο x/m πρέπει να διαιρεθούν δια του 1000. Επιπλέον η εξίσωση $x/m = K\, c^{1/n}$ γίνεται γραμμική με λογαρίθμιση

$$\ln(x/m) = \ln K + (1/n)\ln c$$

Με βάση αυτές τις παρατηρήσεις μεταφέρουμε τον πίνακα 10.3 στο *Excel* και στη συνέχεια δημιουργούμε τρεις νέες στήλες, τις x/m, lnc και ln(x/m), όπου στα κελιά C2, D2 και E2 εισάγουμε τους τύπους =B2/1000, =ln(A2) και =ln(C2), αντίστοιχα, και συμπληρώνουμε την περιοχή C2:E6 με τη διαδικασία της αυτόματης συμπλήρωσης (σχήμα 10.20). Ακολούθως πηγαίνουμε

Πρόσθετα → *ChemStat* → *Regression* → *LS polynomial*

συμπληρώνουμε κατάλληλα τα παράθυρα που ανοίγουν και παίρνουμε τα αποτελέσματα του σχήματος 10.20. Η γραφική παράσταση των δεδομένων με την ευθεία των ελαχίστων τετραγώνων δίνεται στο σχήμα 10.21.

Παρατηρούμε ότι

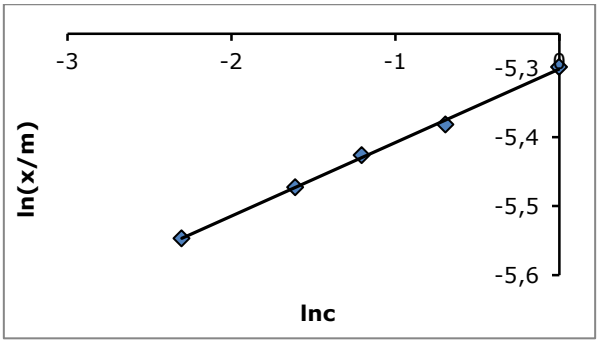
$$\ln K = -5.301 \pm 0.004 \text{ και } 1/n = 0.107 \pm 0.003$$

και συνεπώς

$$K = e^{-5.301} = 0.00499 \text{ και } n = 1/0.107 = 9.346$$

	A	B	C	D	E	F	G
1	c, mol dm^{-3}	(x/m)*10^3, mol g^{-1}	x/m	ln c	ln(x/m)	y(calc)	d=y-y(calc)
2	0,1	3,9	0,0039	-2,303	-5,5468	-5,5468	2,578E-05
3	0,2	4,2	0,0042	-1,609	-5,4727	-5,47287	0,0001977
4	0,3	4,4	0,0044	-1,204	-5,4262	-5,42962	0,003468
5	0,5	4,6	0,0046	-0,693	-5,3817	-5,37513	-0,006569
6	1	5	0,005	0	-5,2983	-5,30119	0,002877
7							
8					c0	c1	
9				c(i) =	-5,3012	0,10667	
10				St.Dev.=	0,00368	0,00262	
11				95% =	0,0117	0,00835	
12				|t| =	1441,26	40,646	
13				p-value	7,4E-10	3,3E-05	
14				r =	0,99909		
15				sy =	0,0046		
16				Type of fitting polynomial:			
17				y = c0 + c1*x + c2*x^2 + ...			

Σχήμα 10.20. Διευθέτηση δεδομένων και αποτελέσματα από την εφαρμογή του προγράμματος *LS polynomial*



Σχήμα 10.21. Διάγραμμα μεταβολής του ln(x/m) με το lnc

Σε ό,τι αφορά τις τυπικές αποκλίσεις των K και n έχουμε: Εφόσον K = e^b = f(b), από τη σχέση (4.7) παίρνουμε

$$s_K = \sqrt{\left(\frac{\partial f}{\partial b}\right)^2 s_b^2} = \sqrt{\left(e^b\right)^2 s_b^2} = \sqrt{K^2 s_b^2} = 1.8\times10^{-5}$$

Όμως πιο απλά οι πράξεις αυτές γίνονται με το πρόγραμμα *Propagation*. Έτσι στα κελιά G9, G10 πληκτρολογούμε K = και n =,

αντίστοιχα, στο κελί H9 εισάγουμε τον τύπο =EXP(E9) και στο κελί H10 τον τύπο =1/F9. Ακολούθως πηγαίνουμε *Πρόσθετα* → *ChemStat* → *Error Propagation,* στο πρώτο πλαίσιο πατάμε *OK* ώστε να υπολογιστούν τυπικές αποκλίσεις, στο δεύτερο κάνουμε *κλικ* στο κελί E9, στο τρίτο στο κελί E10 και στο τελευταίο κάνουμε *κλικ* στο H9. Παίρνουμε την τιμή 1.834×10^{-5} για την τυπική απόκλιση του K. Με τον ίδιο τρόπο υπολογίζουμε την τυπική απόκλιση του n ίση με 0.231 (σχήμα 10.22). Άρα τελικά παίρνουμε:

$$K = 0.00499 \pm 0.00002 \text{ και } n = 9.4 \pm 0.2$$

D	E	F	G	H	I
ln c	**ln(x/m)**	**y(calc)**	**d=y-y(calc)**		
-2,303	-5,5468	-5,5468	2,578E-05		
-1,609	-5,4727	-5,47287	0,0001977		
-1,204	-5,4262	-5,42962	0,003468		
-0,693	-5,3817	-5,37513	-0,006569		
0	-5,2983	-5,30119	0,002877		
	c0	c1			
c(i) =	-5,3012	0,10667	K=	0,004986	*1,8E-05*
St.Dev.=	0,00368	0,00262	n=	9,374961	*0,23065*
95% =	0,0117	0,00835			
|t| =	1441,26	40,646			
p-value	7,4E-10	3,3E-05			
r =	0,99909				
sy =	0,0046				

Σχήμα 10.22. Υπολογισμός τυπικών αποκλίσεων των σταθερών K και n

Παράδειγμα 10.3 – Προσαρμογή σε καμπύλη, πρόβλεψη τιμών

Στον πίνακα 10.4 δίνεται η επίδραση της θερμοκρασίας T στη μοριακή θερμοχωρητικότητα C του αερίου O_2. Να εξετασθεί αν τα πειραματικά δεδομένα περιγράφονται από την εξίσωση $C = a + bT + cT^{-2}$. Επιπλέον να υπολογιστεί η θερμοχωρητικότητα όταν T = 550 και 670 K.

Πίνακας 10.4. Επίδραση της θερμοκρασίας T στη θερμοχωρητικότητα C του αερίου O_2.

T, K	300	400	500	600	700	800	900	1000	1100	1200	1300
C, J mol^{-1} K^{-1}	29.2	30.6	31.4	32.0	32.5	33.0	33.5	34.0	34.5	34.9	35.2

◆ Παρατηρούμε ότι η συνάρτηση προσαρμογής είναι η

$$y = b_0 + b_1\varphi_1(x) + b_2\varphi_2(x)$$

όπου $\varphi_1(x) = T$ και $\varphi_2(x) = T^{-2}$. Θα εξετάσουμε το παράδειγμα αυτό και με τα τρία προγράμματα, *ChemStat*, *Excel* και *SPSS*.

❖ **Ανάλυση στο ChemStat**

Για να προσδιορίσουμε τους συντελεστές b_0, b_1, b_2 αλλά και για να κάνουμε πρόβλεψη της C όταν T = 550 και 670 K διευθετούμε τα δεδομένα όπως στο σχήμα 10.23. Συγκεκριμένα δημιουργούμε με τη σειρά τις στήλες y και x(s), όπου y = C είναι η στήλη της εξαρτημένης μεταβλητής και x(s) είναι οι στήλες των ποσοτήτων $\varphi_1(x) = T$ και $\varphi_2(x) = T^{-2}$.

	A	B	C	D
1	**T, K**	**C, J/mol K**	**T**	**T^{-2}**
2	300	29,2	300	1,1111E-05
3	400	30,6	400	0,00000625
4	500	31,4	500	0,000004
5	600	32	600	2,7778E-06
6	700	32,5	700	2,0408E-06
7	800	33	800	1,5625E-06
8	900	33,5	900	1,2346E-06
9	1000	34	1000	0,000001
10	1100	34,5	1100	8,2645E-07
11	1200	34,9	1200	6,9444E-07
12	1300	35,2	1300	5,9172E-07
13			550	3,3058E-06
14			670	2,2277E-06

Σχήμα 10.23. Διευθέτηση δεδομένων

Μετά τη διευθέτηση των δεδομένων πηγαίνουμε

Πρόσθετα → *ChemStat* → *Regression* → *LS MultiLinear*

και συμπληρώνουμε κατάλληλα τα παράθυρα που ανοίγουν. Συγκεκριμένα στο πλαίσιο "Select the entire range of y, w(if exists), x(s) values" εισάγουμε την περιοχή B2:D14, στο πλαίσιο " Select the range of y values" την περιοχή B2:B12 και στο επόμενο πλαίσιο πατάμε *OK* επειδή δεν χρησιμοποιούμε *συντελεστές βαρύτητας* (*weighting factors*). Στο τελευταίο πλαίσιο πατάμε επίσης *OK*, επειδή η συνάρτηση προσαρμογής έχει σταθερό όρο, και παίρνουμε τα αποτελέσματα του σχήματος 10.24. Η γραφική παράσταση των δεδομένων με την καμπύλη των ελαχίστων τετραγώνων

δίνεται στο σχήμα 10.25, ενώ το διάγραμμα των υπολοίπων δίνεται στο σχήμα 10.26.

	A	B	C	D	E	F	G
1	**T, K**	**C, J/mol K**	**T**	**T^{-2}**	**y(calc)**	**d=y-y(calc)**	
2	300	29,2	300	1,1111E-05	29,22675	-0,02675	
3	400	30,6	400	0,00000625	30,54359	0,05641	
4	500	31,4	500	0,000004	31,37247	0,027533	
5	600	32	600	2,7778E-06	32,00927	-0,00927	
6	700	32,5	700	2,0408E-06	32,55539	-0,05539	
7	800	33	800	1,5625E-06	33,05317	-0,05317	
8	900	33,5	900	1,2346E-06	33,52285	-0,02285	
9	1000	34	1000	0,000001	33,97508	0,024921	
10	1100	34,5	1100	8,2645E-07	34,41591	0,084093	
11	1200	34,9	1200	6,9444E-07	34,84897	0,05103	
12	1300	35,2	1300	5,9172E-07	35,27656	-0,07656	
13			550	3,3058E-06	31,7064	sy(calc) =	0,028095
14			670	2,2277E-06	32,39795		0,026225
15							
16			c0	c1	c2		
17		c(i) =	30,078	0,00408394	-186881		
18		St.Dev.=	0,1083	0,00010049	10371,74		
19		|t| =	277,62	40,6384911	18,01831		
20		p-value =	3E-17	1,4797E-10	9,24E-08		
21		sy =	0,0584				

Σχήμα 10.24. Αποτελέσματα του προγράμματος *LS MultiLinear*

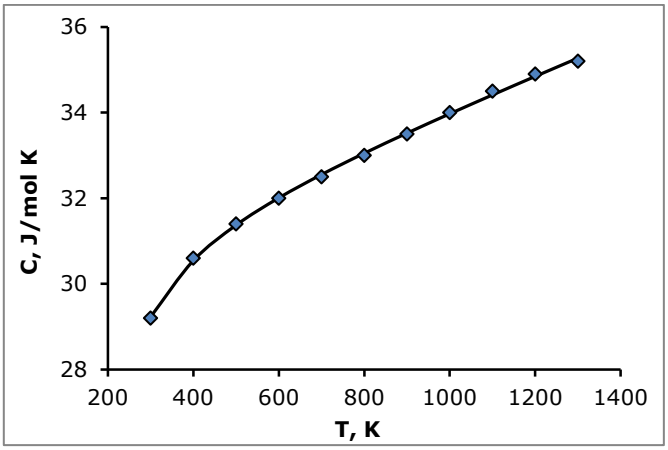
Από τον πίνακα του σχήματος 10.24 παίρνουμε τα αποτελέσματα:

$b_0 = 30.1 \pm 0.1, \quad b_1 = 0.0041 \pm 0.0001, \quad b_2 = -186881 \pm 10372$

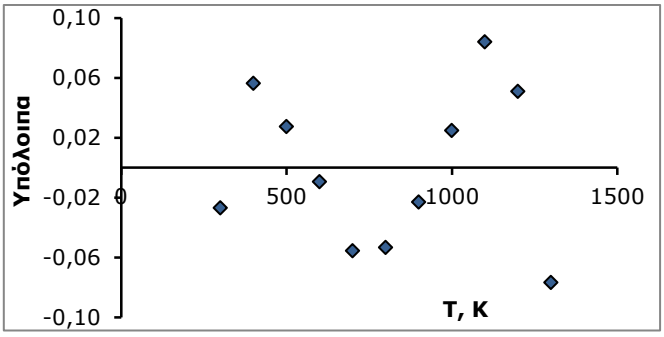
Παρατηρούμε επίσης ότι όλες οι σταθερές είναι στατιστικά σημαντικές. Επιπλέον, από το διάγραμμα του σχήματος 10.25, το διάγραμμα των υπολοίπων καθώς επίσης και από την τιμή του s_y προκύπτει ότι έχουμε μια απόλυτα ικανοποιητική προσαρμογή. Επομένως το μοντέλο προσαρμογής είναι το:

$$C = 30.1 + 0.0041\ T - 186881\ T^{-2}$$

Τέλος, από τον πίνακα του σχήματος 10.24 παίρνουμε ότι η θερμοχωρητικότητα όταν T = 550 είναι C = 31.71 ± 0.03 $J\,mol^{-1}\,K^{-1}$, ενώ όταν T = 670 K έχουμε C = 32.40 ± 0.03 $J\,mol^{-1}\,K^{-1}$.



Σχήμα 10.25. Διάγραμμα μεταβολής του C με το T



Σχήμα 10.26. Διάγραμμα υπολοίπων

❖ **Ανάλυση στο Excel**

Διευθετούμε τα δεδομένα όπως και στην περίπτωση του *ChemStat* στο σχήμα 10.23 με μόνη διαφοροποίηση ότι δεν προσθέτουμε τις γραμμές 13 και 14 με τις τιμές T = 550 και 670 και τις αντίστοιχες τιμές T^{-2}. Ακολούθως από *Δεδομένα* → *Ανάλυση* (*Data* → *Analysis)* πηγαίνουμε στο *Ανάλυση Δεδομένων* (*Data Analysis)* και στον κατάλογο που ανοίγει κάνουμε διπλό *κλικ* στο *Παλινδρόμηση* (*Regression*). Στο παράθυρο διαλόγου με τίτλο *Παλινδρόμηση* που παρουσιάζεται κάνουμε *κλικ* στο πλαίσιο κειμένου που υπάρχει στην *Περιοχή εισόδου Y* (*Input Y Range*) και με το ποντίκι επιλέγουμε την περιοχή τιμών του y, δηλαδή την B2:B12. Για να εισάγουμε τις τιμές των x(s) κάνουμε *κλικ* στο πλαίσιο *Περιοχή εισόδου*

X (*Input X Range*) και με το ποντίκι επιλέγουμε την περιοχή C2:D12. **Δηλαδή ως x επιλέγουμε τις δύο στήλες με τιμές T και T^{-2}**. Στη συνέχεια ορίζουμε την έξοδο, κάνουμε *κλικ* στο κουμπί επιλογής *Υπόλοιπα* (*Residuals*) και κάνουμε *κλικ* στο *OK*.

Από τον πίνακα αποτελεσμάτων παίρνουμε:

	Συντελεστές	*Τυπικό σφάλμα*	*t*	*τιμή-P*
Τεταγμένη επί την αρχή	30,0780	0,1083	277,6190	3,173E-17
Μεταβλητή X 1	0,0041	0,0001	40,6385	1,48E-10
Μεταβλητή X 2	-186881,2797	10371,7	-18,0183	9,235E-08

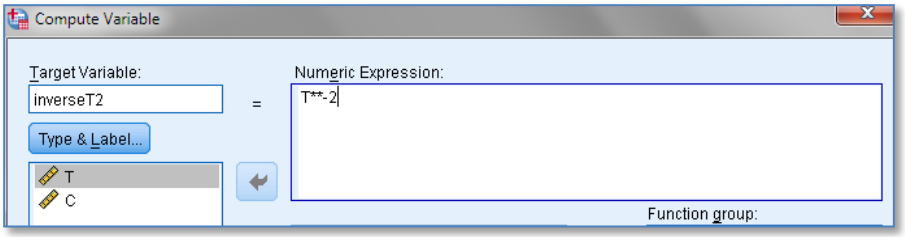
δηλαδή αποτελέσματα που ταυτίζονται με αυτά του *ChemStat*.

Σε ό,τι αφορά την πρόβλεψη, μόνη δυνατότητα που υπάρχει είναι να εφαρμόσουμε τη σχέση $C = 30.078 + 0.0041\ T - 186881.2797\ T^{-2}$ με T = 550 και 670 K.

Παρατήρηση. Για να κάνουμε τις γραφικές παραστάσεις των σχημάτων 10.25 και 26 με βάση τα αποτελέσματα του προγράμματος *Παλινδρόμηση*, μπορούμε να χρησιμοποιήσουμε τα αρχικά δεδομένα T, C και τα αποτελέσματα που το πρόγραμμα παρέχει στις στήλες *Προβλεπόμενο Y* και *Υπόλοιπα*.

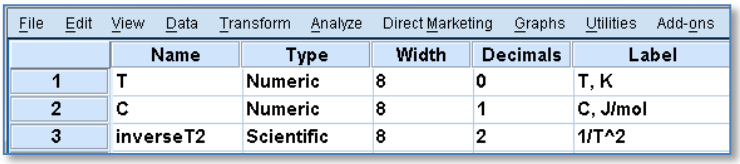
❖ Ανάλυση στο SPSS

Τοποθετούμε τα δεδομένα T και C σε δύο στήλες και δημιουργούμε μια νέα στήλη/μεταβλητή με όνομα έστω inverseT2 και τιμές T^{-2} ως εξής. Πηγαίνουμε *Transform* → *Compute Variable* και συμπληρώνουμε το παράθυρο που ανοίγει όπως στο σχήμα 10.27. Εκείνο που πρέπει να προσέξουμε είναι ότι για να υψώσουμε σε δύναμη στο *SPSS* χρησιμοποιούμε το σύμβολο ** και όχι το ^.



Σχήμα 10.27. Δημιουργία της μεταβλητής inverseT2 με τιμές T^{-2}

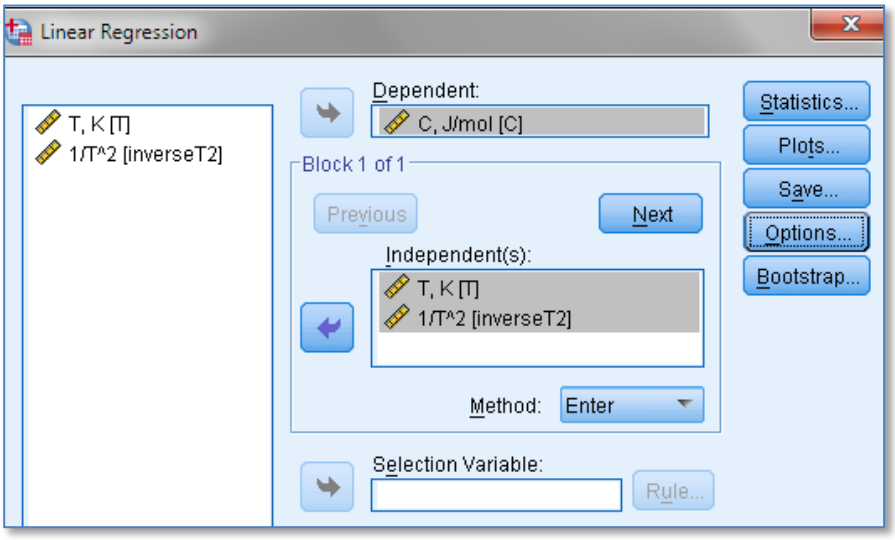
Ακολούθως από το βασικό παράθυρο *Variable View* μορφοποιούμε τις μεταβλητές και συμπληρώνουμε τη στήλη *Label* όπως φαίνεται στο σχήμα 10.28.


File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons

	Name	Type	Width	Decimals	Label
1	T	Numeric	8	0	T, K
2	C	Numeric	8	1	C, J/mol
3	inverseT2	Scientific	8	2	1/T^2

Σχήμα 10.28. Μορφοποίηση μεταβλητών και συμπλήρωση της στήλης *Label*

Για να εφαρμόσουμε τώρα τη μέθοδο των ελαχίστων τετραγώνων πηγαίνουμε *Analyze* → *Regression* → *Linear* και συμπληρώνουμε το παράθυρο που ανοίγει όπως στο σχήμα 10.29. Από το κουμπί *Statistics* μπορούμε να ορίσουμε τα διαστήματα εμπιστοσύνης για τις προσαρμόσιμες παραμέτρους, ενώ από το κουμπί *Options* ενεργοποιούμε την επιλογή *Include constant in equation*.



Σχήμα 10.29. Παράθυρο εισαγωγής δεδομένων

Ο πίνακας αποτελεσμάτων *Coefficients* (σχήμα 10.30) ταυτίζεται στα αντίστοιχα στοιχεία του με τους πίνακες του *ChemStat* και του *Excel*. Και εδώ η πρόβλεψη σε διάφορες τιμές Τ γίνεται με την άμεση εφαρμογή της σχέσης $C = 30.078 + 0.0041\ T - 186881.2797\ T^{-2}$.

Τέλος, επειδή το πρόγραμμα δεν κάνει γραφικές παραστάσεις αντίστοιχες των σχημάτων 10.25 και 26, μπορούμε στο βασικό παράθυρο *Linear Regression* (σχήμα 10.29) να κάνουμε *κλικ* στο κουμπί Save και να ενεργοποιήσουμε τις επιλογές *Predicted values unstandardized* και *Residuals unstandardized*. Με αυτές τις επιλογές δημιουργούνται δύο νέες στήλες με τις προβλεπόμενες τιμές και τα υπόλοιπα που μαζί με τις αρχικές μεταβλητές Τ και C μας δίνουν τη δυνατότητα να κατασκευάσουμε τις γραφικές παραστάσεις που θέλουμε. Θα πρέπει πάντως να τονιστεί ότι επειδή η μορφοποίηση τους είναι δύσκολη στο *SPSS*, είναι πολύ ποιο εύκολο να πάρουμε από το *SPSS* τα δεδομένα και να τα μεταφέρουμε σε κάποιο άλλο σχεδιαστικό πρόγραμμα, όπως το *Excel* ή το *Origin*.

Coefficients

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
(Constant)	30,078	,108		277,619	,000
T, K	,004	,000	,719	40,638	,000
1/T^2	-186881,280	10371,743	-,319	-18,018	,000

Σχήμα 10.30. Τμήμα του πίνακα προσαρμόσιμων παραμέτρων

Παράδειγμα 10.4 – Προσδιορισμός φυσικού νόμου

Η ποσότητα y του νερού σε kg που εξατμίζεται σε μία μέρα από 1 m^2 εδάφους σε συνθήκες σχετικής άπνοιας ενδέχεται να εξαρτάται από τη μέγιστη (T_1) και την ελάχιστη (T_2) θερμοκρασία του εδάφους, τη μέγιστη (T_3) και την ελάχιστη (T_4) θερμοκρασία του αέρα και τη μέση υγρασία (h) σύμφωνα με τα δεδομένα του πίνακα 10.5. Να προσδιοριστεί το μοντέλο:

$$y = b_0 + b_1T_1 + b_2T_2 + b_3T_3 + b_4T_4 + b_5h$$

◆ Το πρόβλημα μπορεί να αντιμετωπιστεί ακριβώς όπως και το προηγούμενο. Παρουσιάζει όμως την ακόλουθη ιδιαιτερότητα. Ορισμένες σταθερές b_i είναι στατιστικά μη σημαντικές.

Πίνακας 10.5. Δεδομένα εξάρτησης της ποσότητας y του νερού που εξατμίζεται από 1 m^2 εδάφους από τις θερμοκρασίες T_1, T_2, T_3, T_4 (σε ^oC) και τη μέση εκατοστιαία υγρασία h

y	T_1	T_2	T_3	T_4	h
3.9	25	20	27	22	55
3.7	30	15	31	15	60
2.7	24	14	26	17	60
5.4	33	25	30	27	70
7.3	36	31	33	26	80
2.1	28	18	32	25	60
3	22	14	25	16	55
2.4	20	15	26	19	50
1.8	10	4	12	7	45
1.8	11	3	15	7	40
1.3	5	0	8	3	40
1.6	4	0	5	0	40
1.6	12	3	15	4	50
1.7	15	6	15	9	55
2.7	25	17	28	20	60

❖ **Ανάλυση στο ChemStat**

Διευθετούμε τα δεδομένα όπως στο σχήμα 10.31, πηγαίνουμε *Πρόσθετα* → *ChemStat* → *Regression* → *LS MultiLinear* και συμπληρώνουμε κατάλληλα τα πλαίσια που ανοίγουν. Τα αποτελέσματα που παίρνουμε δείχνουν ότι τα δεδομένα του πίνακα μπορούν να περιγραφούν από το μοντέλο

$$y = 4.6 + 0.22T_1 + 0.37T_2 - 0.2T_3 - 0.21T_4 - 0.06h \qquad (10.17)$$

Όμως ο τελευταίος όρος φαίνεται να είναι οριακά στατιστικά μη σημαντικός, δεδομένου ότι p-value = 0.069 > 0.05. Συνεπώς μπορούμε να απορρίψουμε τον όρο αυτό. Σε αυτή την περίπτωση διαγράφουμε τη στήλη Η και ξανα-εφαρμόζουμε το πρόγραμμα *LS MultiLinear*. Τα αποτελέσματα της προσαρμογής δίνονται στον πίνακα 10.32, όπου παρατηρούμε ότι τώρα ο συντελεστής της μεταβλητής T_1 είναι στατιστικά μη σημαντικός.

	A	B	C	D	E	F	G	H
1	**y**	**T1**	**T2**	**T3**	**T4**	**H**	**y(calc)**	**d=y-y(calc)**
2	3,9	25	20	27	22	55	4,177	-0,277
3	3,7	30	15	31	15	60	3,810	-0,110
4	2,7	24	14	26	17	60	2,688	0,012
5	5,4	33	25	30	27	70	5,212	0,188
6	7,3	36	31	33	26	80	7,077	0,223
7	2,1	28	18	32	25	60	2,186	-0,086
8	3	22	14	25	16	55	2,962	0,038
9	2,4	20	15	26	19	50	2,375	0,025
10	1,8	10	4	12	7	45	1,703	0,097
11	1,8	11	3	15	7	40	1,276	0,524
12	1,3	5	0	8	3	40	1,066	0,234
13	1,6	4	0	5	0	40	2,062	-0,462
14	1,6	12	3	15	4	50	1,505	0,095
15	1,7	15	6	15	9	55	1,913	-0,213
16	2,7	25	17	28	20	60	2,989	-0,289
17								
18		c0	c1	c2	c3	c4	c5	
19	c(i) =	4,6322	0,2199	0,367	-0,196	-0,21	-0,062	
20	St.Dev.=	1,2322	0,0787	0,051	0,0598	0,045	0,0299	
21	|t| =	3,7593	2,795	7,253	3,2817	4,659	2,0622	
22	p-value =	0,0045	0,0209	5E-05	0,0095	0,001	0,0692	
23	sy =	0,3116						

Σχήμα 10.31. Διευθέτηση δεδομένων και αποτελέσματα του προγράμματος *LS MultiLinear*

18		c0	c1	c2	c3	c4	
19	c(i) =	2,1584	0,1038	0,341	-0,127	-0,19	
20	St.Dev.=	0,3242	0,0633	0,056	0,0569	0,051	
21	|t| =	6,6584	1,6406	6,044	2,229	3,798	
22	p-value =	6E-05	0,1319	1E-04	0,0499	0,003	
23	sy =	0,3588					

Σχήμα 10.32. Αποτελέσματα του προγράμματος *LS MultiLinear* όταν αφαιρεθεί ο τελευταίος όρος

Επομένως ξανα-εφαρμόζουμε το *LS MultiLinear* χωρίς της μεταβλητή T_1. Δηλαδή διαγράφουμε τη στήλη με τη μεταβλητή T1, μεταφέρουμε τις στήλες T_2, T_3 και T_4 μία θέση αριστερά και εφαρμόζουμε το *LS MultiLinear*. Αν ήταν και ο συντελεστής της μεταβλητής T_3 στατιστικά μη σημαντικός, τότε θα απορρίπταμε μόνο τον όρο με τη μεγαλύτερη τιμή p-value. Από τα αποτελέσματα της προσαρμογής που δίνονται στον επόμενο πίνακα παρατηρούμε ότι και πάλι υπάρχει στατιστικά μη σημαντικός συντελεστής και αυτός είναι ο συντελεστής της μεταβλητής T_3.

18		c0	c1	c2	c3
19	c(i) =	2,204131	0,410329	-0,04863	-0,23051
20	St.Dev.=	0,346909	0,039969	0,033378	0,048923
21	|t| =	6,35363	10,2661	1,457046	4,711709
22	p-value =	5,42E-05	5,68E-07	0,173051	0,000638
23	sy =	0,385359			

Συνεπώς, αφού αφαιρέσουμε και τη μεταβλητή T_3 εφαρμόζουμε για μία ακόμη φορά το *LS MultiLinear* και παίρνουμε:

18		c0	c1	c2
19	c(i) =	1,806964	0,400636	-0,2683
20	St.Dev.=	0,224394	0,041215	0,043379
21	|t| =	8,052635	9,720714	6,185158
22	p-value =	3,52E-06	4,86E-07	4,69E-05
23	sy =	0,402988		

Παρατηρούμε ότι τώρα όλοι οι όροι είναι στατιστικά σημαντικοί και συνεπώς η εξάτμιση του νερού από το έδαφος μπορεί να περιγραφεί από τη σχέση

$$y = 1.8070 + 0.4006T_2 - 0.2683T_4$$

ή ορθότερα, λαμβάνοντας υπόψη τις τυπικές αποκλίσεις, από το μοντέλο

$$y = 1.8 + 0.4T_2 - 0.27T_4 \qquad (10.18)$$

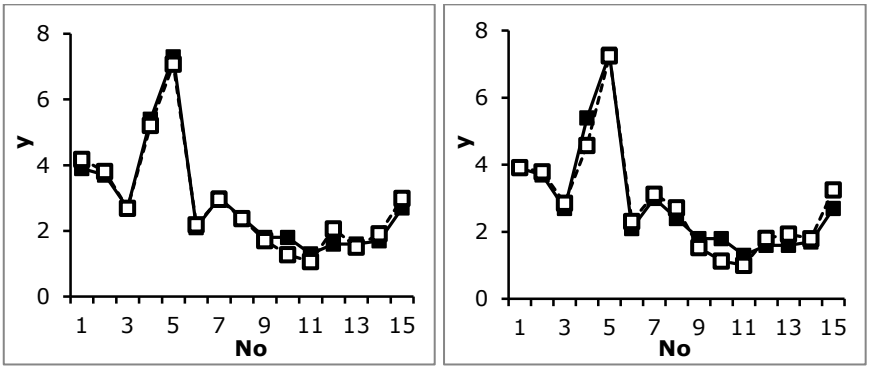
Στο ίδιο αποτέλεσμα, αλλά αμέσως, καταλήγουμε αν στα αρχικά δεδομένα εφαρμόσουμε το πρόγραμμα *LS Significant* από *Πρόσθετα* → *ChemStat* → *Regression* → *LS Significant* (σχήμα 10.33).

Είναι ενδιαφέρον ότι το αρχικό μοντέλο της σχέσης (10.17) έχει 7 όρους, ενώ το μοντέλο με στατιστικά σημαντικούς όρους έχει μόνο 3 όρους. Στο σχήμα 10.34 συγκρίνονται οι πειραματικές τιμές y με τις αντίστοιχες θεωρητικές που υπολογίστηκαν με βάση τα δύο μοντέλα. Παρατηρούμε ότι η αφαίρεση των στατιστικά μη σημαντικών όρων έχει πολύ μικρή επίδραση στην υπολογιζόμενη τιμή του y.

Τέλος, πρέπει και πάλι να τονίσουμε ότι εφόσον χρησιμοποιούμε στατιστικά στοιχεία για την τελική επιλογή του μοντέλου, θα πρέπει να ελεγχθεί η κανονικότητα των υπολοίπων. Ο έλεγχος με το κριτήριο *Anderson-Darling* δείχνει ότι στα υπόλοιπα δεν διαπιστώνονται στατιστικά σημαντικές αποκλίσεις από την κανονικότητα, δεδομένου ότι p-value = 0.886 > 0.05.

	A	B	C	D	E	F	G	H
1	**y**	**T1**	**T2**	**T3**	**T4**	**H**	**y(calc)**	**d=y-y(calc)**
2	3,9	25	20	27	22	55	3,917	-0,017
3	3,7	30	15	31	15	60	3,792	-0,092
4	2,7	24	14	26	17	60	2,855	-0,155
5	5,4	33	25	30	27	70	4,579	0,821
6	7,3	36	31	33	26	80	7,251	0,049
7	2,1	28	18	32	25	60	2,311	-0,211
8	3	22	14	25	16	55	3,123	-0,123
9	2,4	20	15	26	19	50	2,719	-0,319
10	1,8	10	4	12	7	45	1,531	0,269
11	1,8	11	3	15	7	40	1,131	0,669
12	1,3	5	0	8	3	40	1,002	0,298
13	1,6	4	0	5	0	40	1,807	-0,207
14	1,6	12	3	15	4	50	1,936	-0,336
15	1,7	15	6	15	9	55	1,796	-0,096
16	2,7	25	17	28	20	60	3,252	-0,552
17								
18		c-const	c2	c4				
19	c(i) =	1,807	0,4006	-0,268				
20	St.Dev.=	0,2244	0,0412	0,043				
21	|t| =	8,0526	9,7207	6,185				
22	p-value =	4E-06	5E-07	5E-05				
23	sy =	0,403						

Σχήμα 10.33. Διευθέτηση δεδομένων και αποτελέσματα του προγράμματος *LS Significant*



Σχήμα 10.34. Σύγκριση πειραματικών (■) με θεωρητικές (□) τιμές y που υπολογίστηκαν από τη σχέση (10.17) αριστερά και (10.18) δεξιά

❖ Ανάλυση στο Excel

Το πρόβλημα αντιμετωπίζεται με το πρόγραμμα *Παλινδρόμηση* (*Regression*), το οποίο εφαρμόζεται όπως και το *LS Multilinear*. Δηλαδή, πρώτα εφαρμόζουμε το *Παλινδρόμηση* στο πλήρες μοντέλο. Συνεπώς στο πλαίσιο κειμένου που υπάρχει στην *Περιοχή εισόδου Y* (*Input Y Range*) εισάγουμε με το ποντίκι την περιοχή των τιμών του y, δηλαδή την A2:A16, αν έχουμε τη διευθέτηση του σχήματος 10.33. Στο πλαίσιο *Περιοχή εισόδου X* (*Input X Range*) εισάγουμε με το ποντίκι την περιοχή B2:F16, δηλαδή ως x επιλέγουμε όλες τις στήλες με τις ανεξάρτητες μεταβλητές. Όπως και με το *LS Multilinear* ελέγχουμε στα αποτελέσματα για όρους b_i που έχουν p-value > 0.05. Αν υπάρχουν, διαγράφουμε τη στήλη που αντιστοιχεί στον όρο με τη μεγαλύτερη τιμή p-value και αν χρειάζεται μετατοπίζουμε τις στήλες έτσι ώστε να μην παρεμβάλλονται κενές στήλες ανάμεσα στις στήλες με τις ανεξάρτητες μεταβλητές. Ακολούθως ξανα-εφαρμόζουμε το πρόγραμμα *Παλινδρόμηση* και η όλη διαδικασία συνεχίζεται μέχρι να παραμείνουν μόνο στατιστικά σημαντικοί συντελεστές.

Το τελικό αποτέλεσμα που παίρνουμε είναι σε πλήρη συμφωνία με το αντίστοιχο του *ChemStat*:

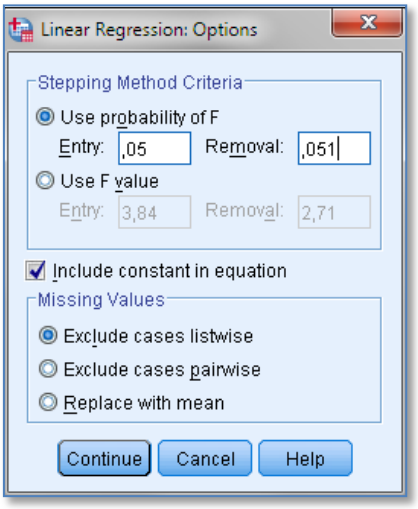
	Συντελεστές	*Τυπικό σφάλμα*	*t*	*τιμή-P*
Τεταγμένη επί την αρχή	1,8070	0,2244	8,0526	3,52E-06
Μεταβλητή X 1	0,4006	0,0412	9,7207	4,86E-07
Μεταβλητή X 2	-0,2683	0,0434	-6,1852	4,69E-05

❖ Ανάλυση στο SPSS

Στο *SPSS* πηγαίνουμε *Analyze* → *Regression* → *Linear* και εισάγουμε τη μεταβλητή y στο πλαίσιο *Dependent* και τις μεταβλητές T_1, T_2, T_3, T_4 και h στο *Independent(s)*. Από το *Save* επιλέγουμε *Residuals Unstandardized* και από το *Options* επιλέγουμε το *Include constant in equation* και διορθώνουμε την τιμή στο *Removal* σε 0.051 (σχήμα 10.35). Θα πρέπει να διευκρινιστεί ότι το *SPSS* χρησιμοποιεί έναν ειδικό *έλεγχο F* για να προσδιορίσει το μοντέλο με τους στατιστικά σημαντικούς όρους.

Τέλος, στο *Method* επιλέγουμε τη μέθοδο που θα χρησιμοποιηθεί για τον υπολογισμό των σταθερών του μοντέλου. Όταν επιλέγουμε *Enter* το πρόγραμμα υπολογίζει όλες τις σταθερές, στην περίπτωση που εξετάζουμε τις σταθερές b_0, b_1, b_2, b_3, b_4, b_5 και h. Αν επιλέξουμε *Backward* το πρόγραμμα αρχικά υπολογίζει όλες τις σταθερές και μετά αρχίζει να αφαιρεί

μία-μία τις στατιστικά μη σημαντικές, όπως κάναμε παραπάνω στο *Excel*. Με την επιλογή *Forward* το πρόγραμμα πρώτα εισάγει τον σταθερό όρο και μετά τη σταθερά που αντιστοιχεί στη μεταβλητή που έχει τη μεγαλύτερη συσχέτιση με την εξαρτημένη μεταβλητή. Εξετάζεται αν είναι στατιστικά σημαντική και μετά το πρόγραμμα εισάγει την επόμενη μεταβλητή με την καλύτερη συσχέτιση με την εξαρτημένη μεταβλητή κ.ο.κ. Τέλος, η επιλογή *Stepwise* είναι συνδυασμός των μεθόδων *Backward* και *Forward*. Γενικά οι μέθοδοι *Stepwise*, *Forward* και *Backward* χρησιμοποιούνται για να πάρουμε μόνο τους στατιστικά σημαντικούς όρους, ενώ η *Enter* όλες τις σταθερές. Δυστυχώς και οι τρεις μέθοδοι, *Stepwise*, *Forward* και *Backward*, δεν δίνουν πάντα το ίδιο αποτέλεσμα, οπότε καλούμαστε να επιλέξουμε εμείς τη μέθοδο με άλλα κριτήρια. Ένα από αυτά είναι η μικρότερη τιμή της s_y ή η φυσική σημασία των όρων του συμμετέχουν στο μοντέλο.



Σχήμα 10.35. Συμπλήρωση πλαισίου *Linear Regression: Options*

Αν στο παράδειγμα που εξετάζουμε επιλέξουμε το *Backward* παίρνουμε τον πίνακα αποτελεσμάτων του σχήματος 10.36. Στον πίνακα αυτόν βλέπουμε όλα τα βήματα μέχρι να φτάσουμε στο τελικό αποτέλεσμα που δίνεται στο πάνελ 4. Παρατηρούμε ότι το τελικό μοντέλο ταυτίζεται με αυτό που προσδιορίσαμε παραπάνω χρησιμοποιώντας το *ChemStat* και το *Excel*.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	4,632	1,232		3,759	,004
	T1	,220	,079	1,339	2,795	,021
	T2	,367	,051	2,087	7,253	,000
	T3	-,196	,060	-1,103	-3,282	,010
	T4	-,209	,045	-1,129	-4,659	,001
	h	-,062	,030	-,421	-2,062	,069
2	(Constant)	2,158	,324		6,658	,000
	T1	,104	,063	,632	1,641	,132
	T2	,341	,056	1,938	6,044	,000
	T3	-,127	,057	-,713	-2,229	,050
	T4	-,193	,051	-1,044	-3,798	,003
3	(Constant)	2,204	,347		6,354	,000
	T2	,410	,040	2,333	10,266	,000
	T3	-,049	,033	-,273	-1,457	,173
	T4	-,231	,049	-1,245	-4,712	,001
4	(Constant)	1,807	,224		8,053	,000
	T2	,401	,041	2,278	9,721	,000
	T4	-,268	,043	-1,450	-6,185	,000

a. Dependent Variable: y

Σχήμα 10.36. Πίνακας αποτελεσμάτων *SPSS* με τη μέθοδο *Linear* και επιλογή *backward*

Τέλος, για την κανονικότητα των υπολοίπων ελέγχουμε τη στήλη με τα υπόλοιπα που δημιουργεί το πρόγραμμα όταν από το *Save* επιλέγουμε *Residuals Unstandardized*. Ο πίνακας κανονικότητας αυτής της μεταβλητής δίνεται στο σχήμα 10.37 και δείχνει ότι η στατιστική ανάλυση είναι αξιόπιστη.

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Unstandardized Residual	,197	15	,120	,910	15	,134

a. Lilliefors Significance Correction

Σχήμα 10.37. Πίνακας ελέγχου της κανονικότητας των υπολοίπων

Παράδειγμα 10.5 – Βαθμονόμηση με γραμμική προσαρμογή

Έξη πρότυπα υδατικά διαλύματα φλουορεσκεΐνης αναλύονται με φθορισμομετρία και η ένταση φθορισμού καταγράφεται σε συνάρτηση με τη συγκέντρωση της φλουορεσκεΐνης στον παρακάτω πίνακα

c, pg/mL	0	2	4	6	8	10
Ένταση	2	7	13	17	23	27

Σε ένα άγνωστης συγκέντρωσης διάλυμα γίνονται 3 μετρήσεις που έδωσαν τις τιμές έντασης φθορισμού 8, 8.3 και 8.1. Να προσδιοριστούν:

α) Η καμπύλη αναφοράς των μετρήσεων

β) Η συγκέντρωση της φλουορεσκεΐνης στο διάλυμα της άγνωστης συγκέντρωσης

γ) Το ελάχιστο όριο ανίχνευσης της φλουορεσκεΐνης

◆ Σε πολλές ενόργανες διατάξεις η ένταση του σήματος y δεν ταυτίζεται με τη μετρούμενη ποσότητα x. Ονομάζουμε ***βαθμονόμηση*** (*calibration*) τον προσδιορισμό με τη μέθοδο των ελαχίστων τετραγώνων της ευθείας $y = a + bx$ και σε σπάνιες περιπτώσεις της καμπύλης $y = a + bx + cx^2$ που επιτρέπει να προσδιοριστεί η τιμή του x όταν δίνεται το y. Η γραφική παράσταση του y ως προς x ονομάζεται ***καμπύλη αναφοράς*** (*calibration curve*).

Αν κάνουμε τη γραφική παράσταση y = Ένταση ως προς x = c προκύπτει ή ορθότερα φαίνεται ότι αυτή είναι μια ευθεία γραμμή. Όμως για να είμαστε απόλυτα βέβαιοι προσδιορίζουμε το βέλτιστο πολυώνυμο προσαρμογής από *Πρόσθετα → ChemStat → Regression → LS Optimum Polynomial* χρησιμοποιώντας την επιλογή *Automatic backward search*. Παίρνουμε ότι το βέλτιστο πολυώνυμο προσαρμογής είναι η ευθεία

$$y = a + bx = 2.19 + 2.53\ x$$

Με βάση τη σχέση αυτή μπορούμε τώρα να υπολογίσουμε τη συγκέντρωση x_0 της φλουορεσκεΐνης στο διάλυμα που έδωσε τις τιμές έντασης 8, 8.3 και 8.1. Η μέση τιμή αυτών των μετρήσεων είναι $\langle y_0 \rangle = 8.1333$ και συνεπώς: $x_0 = (8.1333 - 2.19)/2.53 = 2.349$ pg/mL. Για να αποκτήσουμε περισσότερες πληροφορίες, όπως την ακρίβεια υπολογισμού αυτής της τιμής και το ελάχιστο όριο ανίχνευσης της φλουορεσκεΐνης μπορούμε να χρησιμοποιήσουμε το πρόγραμμα *Calibration*.

Τοποθετούμε τα δεδομένα x – y σε δύο στήλες, έστω στην περιοχή A2:B7, τις τιμές 8, 8.3, 8.1 σε μία στήλη, έστω στην περιοχή E2:E4 όπως στο σχήμα 10.38, και ενεργοποιούμε το πρόγραμμα *Calibration*. Συγκεκριμένα, πηγαίνουμε *Πρόσθετα* → *ChemStat* → *Calibration* → *Calibration*, στο πρώτο πλαίσιο αφήνουμε το 0 επειδή η καμπύλη αναφοράς είναι γραμμική, στο πλαίσιο "Enter x values" εισάγουμε την περιοχή A2:A7, στο πλαίσιο "Enter cell or Range of y0 value(s)" εισάγουμε την περιοχή E2:E4, στο επόμενο πλαίσιο πατάμε *OK* αν θέλουμε το 95% διάστημα εμπιστοσύνης για την προβλεπόμενη τιμή της συγκέντρωσης της φλουορεσκεΐνης, στο πλαίσιο που ακολουθεί αφήνουμε την τιμή 1 εφόσον η καμπύλη αναφοράς έχει σταθερό όρο, και τέλος ορίζουμε ως κελί εξόδου το E6. Τα αποτελέσματα που παίρνουμε δίνονται επίσης στο σχήμα 10.38.

	A	B	C	D	E	F	G	H	I	J
1	**x**	**y**	**y(calc)**		**y0**					
2	0	2	2,1905		8					
3	2	7	7,2476		8,3					
4	4	13	12,305		8,1					
5	6	17	17,362							
6	8	23	22,419		Prediction:					
7	10	27	27,476		<y0> =	8,1333	stdev(y0) =	0,1528		
8					x0 =	2,3503	stdev(x0) =	0,1303	95% =	0,362
9		c0	c1							
10	c(i) =	2,1905	2,5286		Limit of Detection:					
11	St.Dev.=	0,4088	0,0675		x(EOA) =	0,6702	stdev(EOA	0,0179	95% =	0,046
12	|t| =	5,3583	37,454							
13	p-values =	0,0059	3E-06							
14	r =	0,9986								
15	sy =	0,5648								

Σχήμα 10.38. Αποτελέσματα βαθμονόμησης με το πρόγραμμα *Calibration*

Παρατηρούμε ότι το πρόγραμμα υπολογίζει τη μέση τιμή των 8, 8.3, 8.1, $\langle y_0 \rangle$ = 8.1333, την τιμή x_0 = (8.1333 – 2.1905)/2.5286 = 2.3503 που αντιστοιχεί στη μέση τιμή των 8.1333, την τυπική απόκλιση του x_0 και το 95% διάστημα εμπιστοσύνης:

$$x_0 = 2.3503 \pm 0.1303 = 2.4 \pm 0.1 \text{ pg/mL}$$

ενώ το 95% διάστημα εμπιστοσύνης είναι

$$x_0 = 2.3503 \pm 0.362 = 2.4 \pm 0.4 \text{ pg/mL}$$

Επιπλέον το πρόγραμμα υπολογίζει και το ελάχιστο όριο ανίχνευσης. Όταν η καμπύλη αναφοράς είναι ευθεία, y = a + bx, και αφορά κάποια αναλυτική τεχνική, οπότε y είναι το σήμα του οργάνου ανίχνευσης και x η συγκέντρωση της ουσίας που ανιχνεύουμε, τότε το ***ελάχιστο όριο***

ανίχνευσης *(limit of detection)*, x_{EOA}, ορίζεται ως εξής: Είναι η τιμή του x που υπολογίζεται από την y = a + bx αν θέσουμε y = a + 3s_y. Συνεπώς

$$x_{EOA} = 3s_y/b$$

Στο παράδειγμα που μελετάμε ισχύει

$$x_{EOA} = 3*0.565/2.529 = 0.67 \pm 0.02 \text{ pg/mL}$$

Η θεωρία που βρίσκεται πίσω από την τυπική απόκλιση στην τιμή του x_0 είναι η εξής. Η τυπική απόκλιση του x_0 θα μπορούσε να υπολογιστεί από τη σχέση (4.8) που στη συγκεκριμένη περίπτωση γράφεται ως

$$s_{x_0} = \sqrt{\left(\frac{\partial f}{\partial b}\right)^2 s_b^2 + \left(\frac{\partial f}{\partial a}\right)^2 s_a^2 + 2\left(\frac{\partial f}{\partial b}\right)\left(\frac{\partial f}{\partial a}\right)s_{ba} + \left(\frac{\partial f}{\partial \bar{y}}\right)^2 s_{\bar{y}}^2} \qquad (10.19)$$

όπου s_{ba} είναι η συνδιασπορά των b και a, $\bar{y}$ είναι η μέση τιμή των τιμών απορρόφησης 8, 8.3, 8.1 και $s_{\bar{y}}$ είναι η τυπική τους απόκλιση. Στη σχέση αυτή ισχύει

$$x_0 = f(b, a, \bar{y}) \Rightarrow x_0 = (\bar{y} - a)/b \qquad (10.20)$$

Επομένως

$$\frac{\partial f}{\partial b} = \frac{\partial x_0}{\partial b} = -\frac{\bar{y} - a}{b^2}, \quad \frac{\partial f}{\partial a} = -\frac{1}{b} \quad \text{και} \quad \frac{\partial f}{\partial \bar{y}} = \frac{1}{b} \qquad (10.21)$$

και τελικά

$$s_{x_0} = \frac{1}{b}\sqrt{\left(\frac{\bar{y} - a}{b}\right)^2 s_b^2 + s_a^2 + 2\left(\frac{\bar{y} - a}{b}\right)s_{ba} + s_{\bar{y}}^2} \qquad (10.22)$$

Παρατηρούμε ότι για να εφαρμόσουμε τη σχέση αυτή απαιτείται η συνδιασπορά s_{ba}. Όπως είδαμε στη θεωρία των ελαχίστων τετραγώνων, αυτή υπολογίζεται από τη σχέση (10.16). Το πρόγραμμα *LS Polynomial* έχει τη δυνατότητα υπολογισμού του πίνακα *διασπορές-συνδιασπορές* (*variance-covariance matrix*), δηλαδή του πίνακα του οποίου τα διαγώνια στοιχεία είναι οι διασπορές και τα υπόλοιπα στοιχεία είναι οι συνδιασπορές. Εμφανίζεται όταν εισάγουμε τη μονάδα στο παράθυρο διαλόγου με το μήνυμα: "If you want the Covariance Matrix, enter 1". Η τιμή της συνδιασποράς στο συγκεκριμένο παράδειγμα είναι s_{ba} = -0.0228. Με βάση αυτή την τιμή και τη σχέση (10.22) παίρνουμε $s_{x_0} = 0.13$ που ταυτίζεται με αυτή που υπολογίζει το πρόγραμμα *Calibration*.

Παράδειγμα 10.6 – Βαθμονόμηση με προσαρμογή δευτέρου βαθμού

Ο επόμενος πίνακας δίνει τη μεταβολή της επιφάνειας y χρωματογραφικών κορυφών σε συνάρτηση με την εκατοστιαία συγκέντρωση, x, της αιθανόλης σε υδατικά διαλύματα. Να προσδιοριστεί η καμπύλη αναφοράς και με βάση αυτή να υπολογιστεί η συγκέντρωση της αιθανόλης σε δείγμα στο οποίο έγιναν τρεις μετρήσεις με y_0 = 45, 45.1 και 45.2.

x	10	20	30	40	50	60	70	80	90
y	8.2	15.9	22.7	31.5	39.8	49.4	59.7	70.6	83.6

◆ Αν εφαρμόσουμε το πρόγραμμα *LS Optimum Polynomial* διαπιστώνουμε ότι το βέλτιστο πολυώνυμο είναι τρίτου βαθμού χωρίς σταθερό όρο. Επειδή όμως δεν είναι δυνατόν να χρησιμοποιήσουμε τέτοιο πολυώνυμο, ελέγχουμε ποιο πολυώνυμο δευτέρου βαθμού είναι το καλύτερο. Για το σκοπό αυτό εφαρμόζουμε πάλι το πρόγραμμα *LS Optimum Polynomial* χρησιμοποιώντας την επιλογή *Manual*. Το πρόγραμμα δίνει τα αποτελέσματα του σχήματος 10.39.

Στο σχήμα αυτό με γκρι σημειώνουμε τα πολυώνυμα που απορρίπτονται επειδή έχουν τουλάχιστον μία προσαρμόσιμη παράμετρο στατιστικά μη σημαντική. Παρατηρούμε ότι τα πολυώνυμα που μπορούν να χρησιμοποιηθούν είναι ένα τρίτου βαθμού με σταθερό όρο μηδέν (s_y = 0.37) και τα πολυώνυμα δευτέρου και πρώτου βαθμού με σταθερό όρο 0 ή διάφορο του μηδενός. Στο σχήμα 10.39 τα πολυώνυμα με σταθερό όρο 0 δίνονται στο επάνω τμήμα (constant = 0), ενώ αυτά με σταθερό όρο διάφορο του 0 δίνονται στο κάτω τμήμα (constant <> 0). Παρατηρούμε ότι αν απορρίψουμε το πολυώνυμο τρίτου βαθμού, το επόμενο καλύτερο πολυώνυμο είναι ένα δευτέρου βαθμού με σταθερό όρο διάφορο του μηδενός, δεδομένου ότι αυτό έχει τη μικρότερη τιμή s_y = 0.50.

Συνεπώς επιλέγουμε για βαθμονόμηση ένα πολυώνυμο δευτέρου βαθμού με σταθερό όρο. Έτσι αφού διευθετήσουμε τα δεδομένα όπως στο σχήμα 10.40, πηγαίνουμε *Πρόσθετα* → *ChemStat* → *Calibration* → *Calibration*, στο πρώτο πλαίσιο εισάγουμε το 1 για να δηλώσουμε ότι η καμπύλη αναφοράς είναι δευτέρου βαθμού, στο πλαίσιο "Enter x values" εισάγουμε την περιοχή A2:A10, στο πλαίσιο "Enter cell or Range of y0 value(s)" εισάγουμε την περιοχή E2:E4, στο επόμενο πλαίσιο πατάμε *OK* αν θέλουμε το 95% διάστημα εμπιστοσύνης για την προβλεπόμενη τιμή της συγκέντρωσης, και τέλος ορίζουμε ως κελί εξόδου το F1. Το αποτέλεσμα

που παίρνουμε είναι:

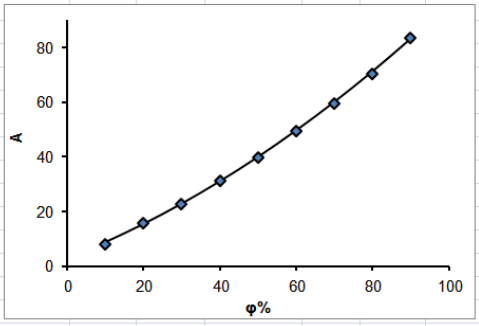
$$x_0 = 55.49 \pm 0.280 = 55.5 \pm 0.3 \ \%v/v$$

p	sy	p-values when Constant = 0							
1	3.087	4.4E-11							
2	0.944	1.2E-08	5E-05						
3	0.368	3E-08	0.0452	0					
4	0.376	9.6E-06	0.1917	0.17	0.43				
5	0.315	0.00019	0.0838	0.1	0.137	0.15			
6	0.361	0.00774	0.6044	0.81	0.985	0.91	0.824		
7	0.442	0.1073	0.8669	0.95	0.997	0.99	0.993	0.999	
8	0.445	0.78223	0.5505	0.52	0.513	0.51	0.507	0.506	0.506

p	sy	p-values when Constant <> 0								
1	2.523	0.06088	2E-08							
2	0.502	0.00496	1E-06	0						
3	0.337	0.20437	7E-05	0.93	0.034					
4	0.369	0.33807	0.0202	0.72	0.893	0.7				
5	0.353	0.68669	0.0821	0.38	0.337	0.34	0.323			
6	0.423	0.71006	0.4105	0.66	0.683	0.72	0.758	0.794		
7	0.33	0.35409	0.3176	0.35	0.356	0.36	0.366	0.369	0.372	
8	0.33	0.35409	0.3176	0.35	0.356	0.36	0.366	0.369	0.372	0

Σχήμα 10.39. Επιλογή του καλύτερου πολυωνύμου προσαρμογής

	A	B	C	D	E	F	G	H	I	J	K
1	φ%	A	y(calc)		y0	Prediction:					
2	10	8,2	8,6933		45	<y0> =	45,10	stdev(y0) =	0,10		
3	20	15,9	15,37		45,1	x0 =	55,49	stdev(x0) =	0,28	95% =	0,68
4	30	22,7	22,794		45,2						
5	40	31,5	30,966								
6	50	39,8	39,886								
7	60	49,4	49,553								
8	70	59,7	59,968								
9	80	70,6	71,13								
10	90	83,6	83,04								
11											
12		c0	c1	c2							
13	c(i) =	2,7643	0,5555	0,0037							
14	St.Dev.=	0,6394	0,0294	0,0003							
15	|t| =	4,3234	18,923	13,056							
16	p-values =	0,005	1E-06	1E-05							
17	r =	0,9999									
18	sy =	0,5025									



Σχήμα 10.40. Αποτελέσματα βαθμονόμησης με το πρόγραμμα *Calibration*

Παράδειγμα 10.7 – Εσωτερική βαθμονόμηση

Στη μέθοδο της ***εσωτερικής βαθμονόμησης*** (*standard addition internal calibration*) μετρείται το δείγμα με την άγνωστη συγκέντρωση, c_0, με κατάλληλη αναλυτική τεχνική και έστω ότι δίνει σήμα y_0. Στη συνέχεια με βάση το αρχικό δείγμα δημιουργούνται Ν διαλύματα ίσου όγκου με το αρχικό, τα οποία εκτός από την άγνωστη συγκέντρωση c_0 περιέχουν επιπλέον και διαφορετικές αλλά γνωστές συγκεντρώσεις του αναλύτη, c_i. Σε κάθε διάλυμα μετρείται το σήμα y_i. Από την γραμμική προσαρμογή του y ως προς c προσδιορίζεται η άγνωστη συγκέντρωση, c_0, του αναλύτη. Σε μια παραλλαγή της μεθόδου, τα Ν διαλύματα δημιουργούνται με την προσθήκη διαφορετικών όγκων V_i από πρότυπο διάλυμα του αναλύτη συγκέντρωσης c_{std}. Η άγνωστη συγκέντρωση, c_0, προσδιορίζεται από τη γραμμική προσαρμογή του y ως προς V. Η μέθοδος διευκρινίζεται στο παράδειγμα αυτό και στο επόμενο.

Έστω ότι θέλουμε το προσδιορίσουμε την άγνωστη συγκέντρωση Ag σε ένα δείγμα. Δημιουργούμε 6 διαλύματα ίσου όγκου στα οποία έχουμε προσθέσει γνωστές ποσότητες Ag, 5, 10, 15, 20, 25 και 30 μg/mL, όπως φαίνεται στον επόμενο πίνακα στον οποίον δίνεται η μεταβολή της *Απορρόφησης* (*Absorbance*) των διαλυμάτων αυτών σε συνάρτηση με τη συγκέντρωση c_i του προστιθέμενου Ag. Ποια η άγνωστη συγκέντρωση c_0 του Ag στο δείγμα;

Ag, μg/mL	0	5	10	15	20	25	30
Απορρόφηση	0.20	0.35	0.50	0.60	0.75	0.90	1.0

◆ Επειδή το σήμα που καταγράφεται είναι ανάλογο της ολικής συγκέντρωσης του αναλύτη, θα έχουμε

$$y_i = k(c_0 + c_i) = kc_0 + kc_i$$

Συνεπώς αν κάνουμε τη γραφική παράσταση y_i ως προς c_i και προσδιορίσουμε την ευθεία των ελαχίστων τετραγώνων, y = a + bc, θα έχουμε

$$a = kc_0 \quad \text{και} \quad b = k \quad \Rightarrow \quad c_0 = a/b$$

Για την τυπική απόκλιση και τα διαστήματα εμπιστοσύνης της c_0 ισχύουν ότι και στα προηγούμενα παραδείγματα.

Η παραπάνω ανάλυση γίνεται άμεσα αν χρησιμοποιήσουμε το πρόγραμμα *Addition*. Έτσι διευθετούμε τα δεδομένα όπως στο σχήμα

10.41, πηγαίνουμε *Πρόσθετα* → *ChemStat* → *Calibration* → *Standard Addition*, στο πρώτο πλαίσιο εισάγουμε την τιμή 0 για να δηλώσουμε ότι η ανεξάρτητη μεταβλητή περιέχει τις συγκεντρώσεις c_i, στο πλαίσιο "Enter x values" εισάγουμε την περιοχή A2:A8, στο επόμενο πλαίσιο πατάμε *OK* αν θέλουμε το 95% διάστημα εμπιστοσύνης για την προβλεπόμενη τιμή της συγκέντρωσης του Ag, και τέλος ορίζουμε ως κελί εξόδου το E2. Τα αποτελέσματα που παίρνουμε δίνονται επίσης στο σχήμα 10.41 και δείχνουν ότι

$$c_0 = 7.9333 \pm 0.5778 = 7.9 \pm 0.6 \text{ μg/mL}$$

	A	B	C	D	E	F	G	H
1	**Ag, μg/mL**	**Absorbance**	**y(calc)**					
2	0	0,2	0,2125		**Prediction:**			
3	5	0,35	0,34643		C =	7,93333	stdev(C) =	0,5778
4	10	0,5	0,48036				95% =	1,4852
5	15	0,6	0,61429					
6	20	0,75	0,74821					
7	25	0,9	0,88214					
8	30	1	1,01607					
9								
10		c0	c1					
11	c(i) =	0,2125	0,02679					
12	St.Dev.=	0,011151782	0,00062					
13	|t| =	19,0552503	43,3013					
14	p-values =	7,33669E-06	1,2E-07					
15	r =	0,998669327						
16	sy =	0,016366342						

Σχήμα 10.41. Αποτελέσματα βαθμονόμησης με το πρόγραμμα *Addition*

Παράδειγμα 10.8 – Εσωτερική βαθμονόμηση

Για τον προσδιορισμό της άγνωστης συγκέντρωσης c_0 του Mn^{2+} σε ένα διάλυμα δημιουργούνται 6 διαλύματα. Κάθε διάλυμα περιέχει 25 mL του αρχικού δείγματος και 0, 1, 2, 3, 4 και 5 mL, αντίστοιχα, από ένα πρότυπο διάλυμα Mn^{2+} συγκέντρωσης c_{std} = 95 mg/L. Όλα τα διαλύματα αραιώνονται στα 100 mL και μετρείται η απορρόφηση. Προσδιορίστηκαν οι τιμές απορρόφησης: 0.11, 0.20, 0.30, 0.38, 0.45, 0.55. Ποια η άγνωστη συγκέντρωση c_0 του Mn^{2+};

◆ Για το σήμα απορρόφησης ισχύει

$$y_i = k(c_0 + c_i) = kc_0 + k(c_{std}/V_0)V_i$$

όπου V_0 = 25 mL. Συνεπώς αν κάνουμε τη γραφική παράσταση y_i - V_i και προσδιορίσουμε την ευθεία των ελαχίστων τετραγώνων, y = a + bV, τώρα θα έχουμε

$$a = kc_0 \quad \text{και} \quad b = k\, c_{std}/V_0 \quad \Rightarrow \quad c_0 = (a/b)c_{std}/V_0$$

Για την άμεση εφαρμογή της μεθόδου, διευθετούμε τα δεδομένα όπως στο σχήμα 10.42 και πηγαίνουμε *Πρόσθετα* → *ChemStat* → *Calibration* → *Standard Addition*, όπου στο πρώτο πλαίσιο αφήνουμε την τιμή 1 για να δηλώσουμε ότι η ανεξάρτητη μεταβλητή περιέχει όγκους, V_i, στο πλαίσιο "Enter x values" εισάγουμε την περιοχή A5:A10, στα δύο επόμενα πλαίσια εισάγουμε τα κελιά B1 και B2 με τις τιμές των c_{std} και V_0, αντίστοιχα, στο επόμενο πλαίσιο πατάμε *OK* για το 95% διάστημα εμπιστοσύνης για την προβλεπόμενη τιμή της συγκέντρωσης του Mn^{2+}, και τέλος ορίζουμε ως κελί εξόδου το E4. Τα αποτελέσματα που παίρνουμε δίνονται στο σχήμα 10.42 και δείχνουν ότι

$$c_0 = 5.058 \pm 0.398 = 5.1 \pm 0.4 \text{ mg/L}$$

	A	B	C	D	E	F	G	H
1	**c_{std}=**	95						
2	**V_0=**	25						
3								
4	**V, mL**	**Absorbance**	**y(calc)**		**Prediction:**			
5	0	0,11	0,11524		C =	5,05831	stdev(C) =	0,398
6	1	0,2	0,20181				95% =	1,106
7	2	0,3	0,28838					
8	3	0,38	0,37495					
9	4	0,45	0,46152					
10	5	0,55	0,5481					
11								
12		c0	c1					
13	c(i) =	0,1152381	0,08657					
14	St.Dev.=	0,00655	0,00216					
15	|t| =	17,5936035	40,0165					
16	p-values =	6,1297E-05	2,3E-06					
17	r =	0,99875337						
18	sy =	0,00905012						

Σχήμα 10.42. Αποτελέσματα βαθμονόμησης με το πρόγραμμα *Addition*

Παρατήρηση. Όπως αναφέρθηκε, το πρόγραμμα *Standard Addition* εφαρμόζεται μόνο όταν η καμπύλη προσαρμογής είναι ευθεία.

Παράδειγμα 10.9 – Σημείο τομής ευθειών

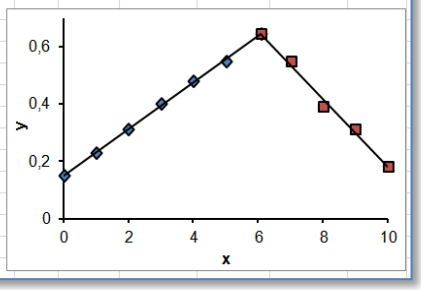
Αν η άγνωστη συγκέντρωση προσδιορίζεται από το σημείο τομής δύο ευθειών, τότε μπορεί να χρησιμοποιηθεί το πρόγραμμα *Intersection of lines*. Έστω δύο ευθείες που ορίζονται από τα σημεία: x1 = 0, 1, 2, 3, 4, 5 και y1 = 0.15, 0.23, 0.31, 0.40, 0.48, 0.55 η πρώτη και x2 = 7, 8, 9, 10 και y2 = 0.55, 0.39, 0.31, 0.18 η δεύτερη. Να προσδιοριστεί το σημείο τομής τους.

◆ Διευθετούμε τα σημεία όπως στο σχήμα 10.43, όπου αφήνουμε τουλάχιστον μια στήλη κενή ανάμεσα στα δύο σετ δεδομένων. Ακολούθως πηγαίνουμε *Πρόσθετα* → *ChemStat* → *Calibration* → *Intersection of lines*, όπου στο πλαίσιο "Enter x1 values" εισάγουμε την περιοχή A2:A7, στο επόμενο πλαίσιο "Enter x2 values" εισάγουμε την περιοχή D2:D5, στο επόμενο πλαίσιο πατάμε *OK* για το 95% διάστημα εμπιστοσύνης για την τιμή του σημείου τομής, και τέλος ορίζουμε ως κελί εξόδου έστω το H1. Τα αποτελέσματα που παίρνουμε δίνονται στο σχήμα 10.43 και δείχνουν ότι το σημείο τομής x_0 είναι το

$$x_0 = 6.09 \pm 0.14 = 6.1 \pm 0.1$$

και αντιστοιχεί σε $y_0 = 0.64$.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	**x1**	**y1**	**y(calc)**	**x2**	**y2**	**y(calc)**		**Prediction:**					
2	0	0,15	0,150	7	0,55	0,536		x =	6,088	stdev(x) =	0,1396	95% (x) =	0,342
3	1	0,23	0,232	8	0,39	0,417		y =	0,644				
4	2	0,31	0,313	9	0,31	0,298							
5	3	0,4	0,394	10	0,18	0,179							
6	4	0,48	0,475	6,088	0,6445	0,644496							
7	5	0,55	0,556										
8	6,08827	0,6445	0,644										
9		c0	c1		c0	c1							
10	c(i) =	0,1505	0,081		1,369	-0,119							
11	St.Dev.=	0,0038	0,001		0,089	0,010							
12	|t| =	39,562	64,590		15,437	11,504							
13	p-values =	2E-06	3,4E-07		0,004	0,007							
14	r =	0,9995			0,993								
15	sy =	0,0053			0,023								
16													



Σχήμα 10.43. Αποτελέσματα προσδιορισμού της τομής δύο ευθειών

Για να κάνουμε τη γραφική παράσταση με τις ευθείες τεμνόμενες, μεταφέρουμε το σημείο τομής (6.088, 0.6445) στα κελιά A8, C8 και D6, F6 και κάνουμε τη γραφική παράσταση που μετά από κατάλληλη μορφοποίηση μπορεί να είναι όπως στο σχήμα 10.43.

10.10 ΜΗ ΓΡΑΜΜΙΚΗ ΠΡΟΣΑΡΜΟΓΗ

Η μέθοδος των ελαχίστων τετραγώνων επιτρέπει την προσαρμογή μιας εξίσωσης σε πειραματικά δεδομένα μόνο όταν η εξίσωση έχει ή μπορεί να πάρει τη μορφή της εξίσωσης (10.2), όπου οι συναρτήσεις $\varphi_k(x)$ δεν περιέχουν άγνωστες σταθερές. Αν περιέχουν άγνωστες σταθερές, τότε η μέθοδος αυτή δεν μπορεί να εφαρμοσθεί. Για παράδειγμα, δεν υπάρχει κανένας τρόπος να προσαρμοσθεί η εξίσωση $y = a + be^{cx}$ σε πειραματικά δεδομένα (x_i, y_i) με βάση τη μεθοδολογία των ελαχίστων τετραγώνων.

Σε αυτές τις περιπτώσεις είμαστε υποχρεωμένοι να καταφύγουμε σε εξειδικευμένες μεθόδους στατιστικής ανάλυσης. Οι μέθοδοι αυτές, που ονομάζονται μέθοδοι ***μη γραμμικής προσαρμογής*** *(non-linear fitting)*, στηρίζονται επίσης στην ελαχιστοποίηση του αθροίσματος S των τετραγώνων των υπολοίπων, σχέση (10.1). Ενώ όμως στη γραμμική προσαρμογή η ελαχιστοποίηση του S οδηγεί σε ένα γραμμικό σύστημα που επιλύεται εύκολα, στη μη γραμμική προσαρμογή απαιτούνται προχωρημένες τεχνικές αριθμητικής ανάλυσης.

Το *Excel* διαθέτει ένα ιδιαίτερα ισχυρό πρόγραμμα, το *Επίλυση* (*Solver*), που επιτρέπει να αντιμετωπίζουμε εύκολα προβλήματα μη γραμμικής προσαρμογής. Το πρόγραμμα αυτό βρίσκεται στη λωρίδα *Δεδομένα (Data)* στο πλαίσιο *Ανάλυση (Analysis)*. Αντίστοιχο πρόγραμμα υπάρχει στο *SPSS* που ενεργοποιείται από *Analyze* → *Regression* → *Nonlinear*. Για να δούμε πώς λειτουργούν αυτά τα προγράμματα θα εξετάσουμε το παρακάτω παράδειγμα.

Παράδειγμα 10.10

Στη χρωματογραφία υψηλής πίεσης (HPLC) ο χρόνος συγκράτησης t ενός ασθενούς οργανικού οξέος στη στήλη εξαρτάται από το pH του διαλύματος. Αποδεικνύεται θεωρητικά ότι η εξάρτηση αυτή περιγράφεται από τη σχέση

$$t = \frac{k_0 + k_{-1} 10^{(pH - pK_a)}}{1 + 10^{(pH - pK_a)}} \qquad (10.23)$$

όπου K_a είναι η σταθερά διάστασης του οξέος και k_0, k_{-1} είναι σταθερές που εξαρτώνται τόσο από το οξύ, όσο και από τα χαρακτηριστικά της χρωματογραφικής στήλης. Να προσαρμοσθεί η σχέση αυτή στα πειραματικά δεδομένα του πίνακα 10.6, στον οποίο δίνεται η επίδραση του pH στο χρόνο συγκράτησης του 3,4-διυδροξυ-φαινυλακετοξικού οξέος (DOPAC), και να προσδιοριστούν το pK_a και οι σταθερές k_0, k_{-1}.

Πίνακας 10.6. Επίδραση του pH στο χρόνο συγκράτησης του DOPAC.

pH	1	2	3	4	5	6	7	8	9
t, min	28.6	28.5	26.8	10.5	5.5	3.1	2.9	2.8	2.8

❖ **Ανάλυση στο Excel**

Ανοίγουμε ένα φύλλο εργασίας και προχωρούμε στα παρακάτω βήματα:

(1) Στην περιοχή A1:A3 πληκτρολογούμε τους τίτλους pK_a, k_0, k_{-1} και στην B1:B3 εισάγουμε τις αρχικές τιμές αυτών των σταθερών, έστω αυθαίρετα τις 4, 5, 1. Στην περιοχή A5:D5 γράφουμε τις επικεφαλίδες pH, t,min, t(υπολ.), SR και εισάγουμε τα πειραματικά δεδομένα του πίνακα στις στήλες A6:A14 και B6:B14.

(2) Στο κελί C6 πληκτρολογούμε τη σχέση (10.23), δηλαδή

=(B$2+B$3*10^(A6-B$1))/(1+10^(A6-B$1))

Στο D6 εισάγουμε τον τύπο =(B6-C6)^2, δηλαδή το *τετράγωνο των υπολοίπων* (*Squared Residuals - SR*) και συμπληρώνουμε την περιοχή C6:D14 με τη διαδικασία της αυτόματης συμπλήρωσης.

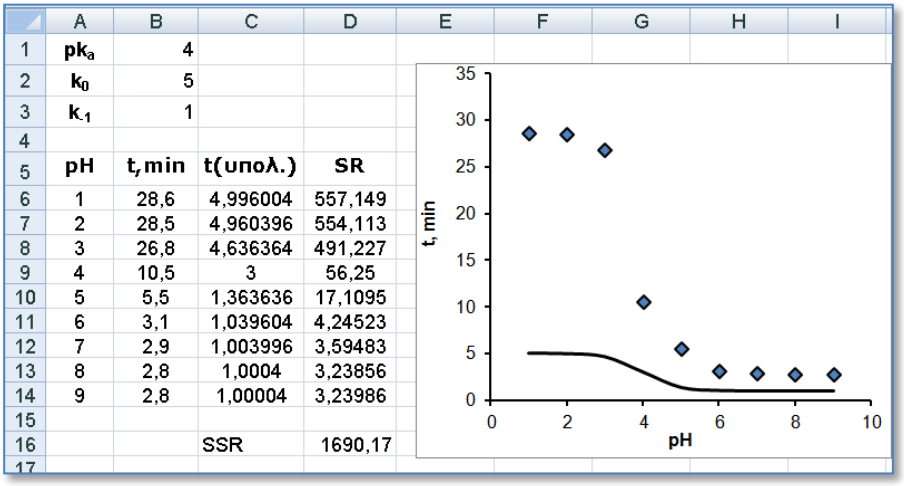
(3) Στο C16 εισάγουμε τον τίτλο SSR= και στο D16 προσδιορίζουμε το *άθροισμα των τετραγώνων των υπολοίπων* (*Sum of Squared Residuals - SSR*). Συνεπώς εισάγουμε τον τύπο =SUM(D6:D14).

(4) Κάνουμε τη γραφική παράσταση των t,min και t(υπολ.) σε συνάρτηση με το pH, όπως φαίνεται στο σχήμα (10.44).

(5) Από το *Δεδομένα (Data)* → *Ανάλυση (Analysis)* κάνουμε *κλικ* στο *Επίλυση* (*Solver*) και συμπληρώνουμε το παράθυρο διαλόγου που εμφανίζεται όπως στο σχήμα 10.45. Συγκεκριμένα, στο *Ορισμός στόχου* (*Set Objective*) συμπληρώνουμε το κελί D16 όπου βρίσκεται το άθροισμα των τετραγώνων των υπολοίπων. Στο *Σε* (*To*) κάνουμε *κλικ* στο *Ελάχιστο* (*min*) και στο *Με αλλαγή μεταβλητών κελιών* (*By Changing Variable Cells*) εισάγουμε την περιοχή B1:B3, δηλαδή την περιοχή στην οποία υπάρχουν οι αρχικές τιμές των σταθερών pK_a, k_0, k_{-1}. Συνεπώς ζητάμε από το *Επίλυση* να κάνει ελάχιστο τον αριθμό που βρίσκεται στο κελί D16 με αλλαγή των τιμών που υπάρχουν στην περιοχή B1:B3. Επίσης ως μέθοδο επίλυσης επιλέγουμε *Μη γραμμικό GRG* (*GRG Nonlinear*).

(6) Στο σημείο αυτό είναι χρήσιμο να κάνουμε τις ακόλουθες ρυθμίσεις. Πατάμε στο κουμπί *Επιλογές* (*Options*) και στην καρτέλα *Όλες οι*

μέθοδοι (*All Solving Methods*) ενεργοποιούμε τη *Χρήση αυτόματης κλίμακας* (*Use Automatic Scaling*) και διορθώνουμε την τιμή στο πεδίο *Ακρίβεια περιορισμού* (*Constraint Precision*) σε 0.0000000001. Επίσης, στην καρτέλα *Μη γραμμικό GRG (GRG Nonlinear)* ενεργοποιούμε στο πάνελ *Παράγωγοι* (*Derivatives*) το *Κεντρική* (*Central*) και διορθώνουμε την τιμή στο πεδίο *Σύγκλιση* (*Convergence*) σε 0.0000000001.



	A	B	C	D
1	pk_a	4		
2	k_0	5		
3	k_{-1}	1		
4				
5	**pH**	**t,min**	**t(υπολ.)**	**SR**
6	1	28,6	4,996004	557,149
7	2	28,5	4,960396	554,113
8	3	26,8	4,636364	491,227
9	4	10,5	3	56,25
10	5	5,5	1,363636	17,1095
11	6	3,1	1,039604	4,24523
12	7	2,9	1,003996	3,59483
13	8	2,8	1,0004	3,23856
14	9	2,8	1,00004	3,23986
15				
16			SSR	1690,17
17				

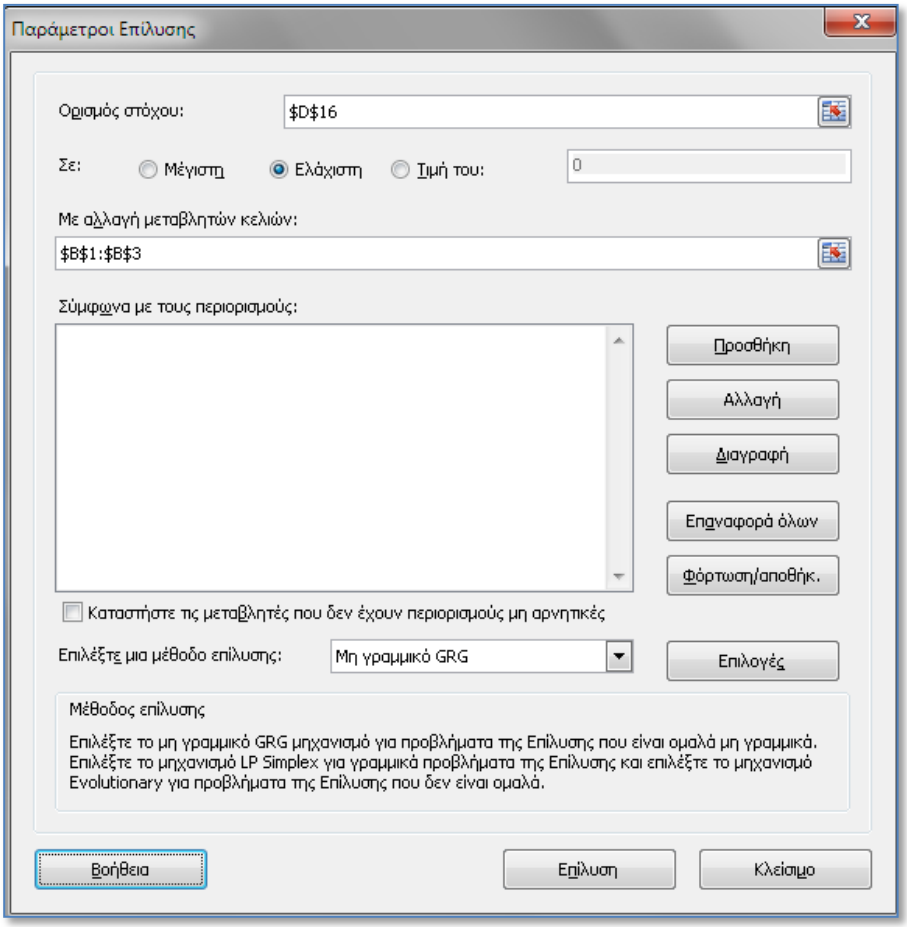
Σχήμα 10.44. Διευθέτηση δεδομένων και γραφική παράσταση

(7) Κάνοντας *κλικ* στο κουμπί *Επίλυση* (*Solve*) το πρόγραμμα αρχίζει να προσεγγίζει τις ζητούμενες τιμές και μετά από ένα σύντομο χρονικό διάστημα εμφανίζει τα αποτελέσματα στην περιοχή B1:B3, διαγράφοντας τα προηγούμενα. Παράλληλα εμφανίζεται ένα πλαίσιο διαλόγου, στο οποίο μας δίνεται η επιλογή ή να κρατήσουμε τα νέα αποτελέσματα ή να επιστρέψουμε στις παλιές αρχικές τιμές.

Οι τιμές που παίρνουμε στην περιοχή B1:B3 είναι

$$pK_a = 3.7007, \quad k_0 = 29.3922 \quad \text{και} \quad k_{-1} = 3.0079$$

Επιπλέον στο D16 παρατηρούμε ότι το άθροισμα των τετραγώνων των υπολοίπων γίνεται 7.272, πολύ μικρότερο από την αρχική τιμή 1690.2. Όμως το στοιχείο που δείχνει ότι η προσαρμογή είναι καλή είναι το διάγραμμα μεταβολής των πειραματικών και θεωρητικών τιμών του χρόνου συγκράτησης t με το pH (σχήμα 10.46).



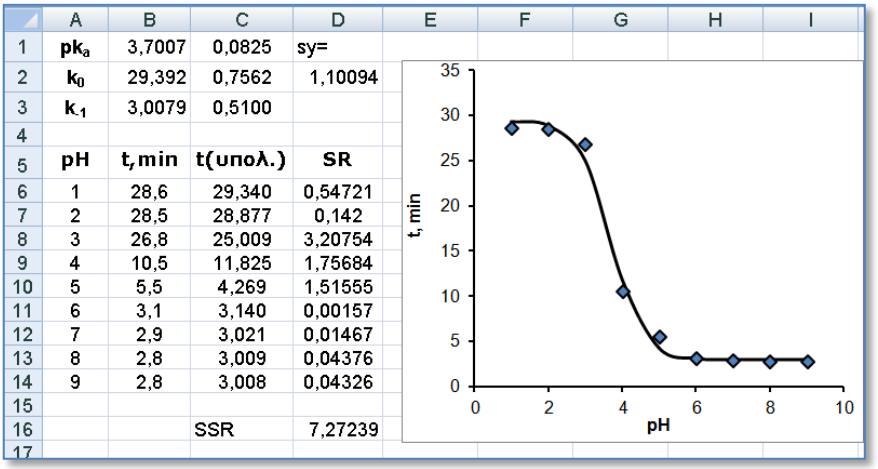
Σχήμα 10.45. Συμπλήρωση των παραμέτρων του προγράμματος *Επίλυση*

Παρατήρηση 1. Στη μη γραμμική προσαρμογή η επιλογή των αρχικών τιμών παίζει καθοριστικό ρόλο στην επίλυση του προβλήματος. Αν διαπιστώσουμε ότι το πρόγραμμα αδυνατεί να προσδιορίσει την καμπύλη των ελαχίστων τετραγώνων, αλλάζουμε τις αρχικές τιμές και ξανατρέχουμε το *Επίλυση*.

Παρατήρηση 2. Ένα από τα μειονεκτήματα του προγράμματος *Επίλυση* είναι ότι δεν υπολογίζει τις τυπικές αποκλίσεις των προσαρμόσιμων σταθερών. Αυτό μπορεί να γίνει με το πρόγραμμα *Regression* → *Solver Errors* του *ChemStat*. Όταν εκτελέσουμε το πρόγραμμα αυτό, στο πρώτο

παράθυρο εισάγουμε την περιοχή B1:B3 των προσαρμόσιμων παραμέτρων, στο δεύτερο το κελί D16 του αθροίσματος των τετραγώνων των υπολοίπων και στο τρίτο την περιοχή των υπολογιζόμενων τιμών y, την C6:C9. Οι τυπικές αποκλίσεις εμφανίζονται στα κελιά C1:C3, ενώ στο κελί D2 παρουσιάζεται η τιμή της s_y. Παίρνουμε:

$$pK_a = 3.70 \pm 0.08, \quad k_0 = 29.4 \pm 0.8 \quad \text{και} \quad k_{-1} = 3.0 \pm 0.5$$



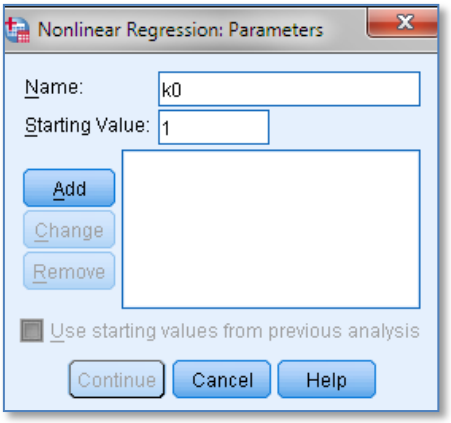
	A	B	C	D
1	pk_a	3,7007	0,0825	sy=
2	k_0	29,392	0,7562	1,10094
3	k_{-1}	3,0079	0,5100	
4				
5	pH	t,min	t(υπολ.)	SR
6	1	28,6	29,340	0,54721
7	2	28,5	28,877	0,142
8	3	26,8	25,009	3,20754
9	4	10,5	11,825	1,75684
10	5	5,5	4,269	1,51555
11	6	3,1	3,140	0,00157
12	7	2,9	3,021	0,01467
13	8	2,8	3,009	0,04376
14	9	2,8	3,008	0,04326
15				
16			SSR	7,27239
17				

Σχήμα 10.46. Τελικά αποτελέσματα μετά την εφαρμογή της *Επίλυσης* και του *Solver Errors*

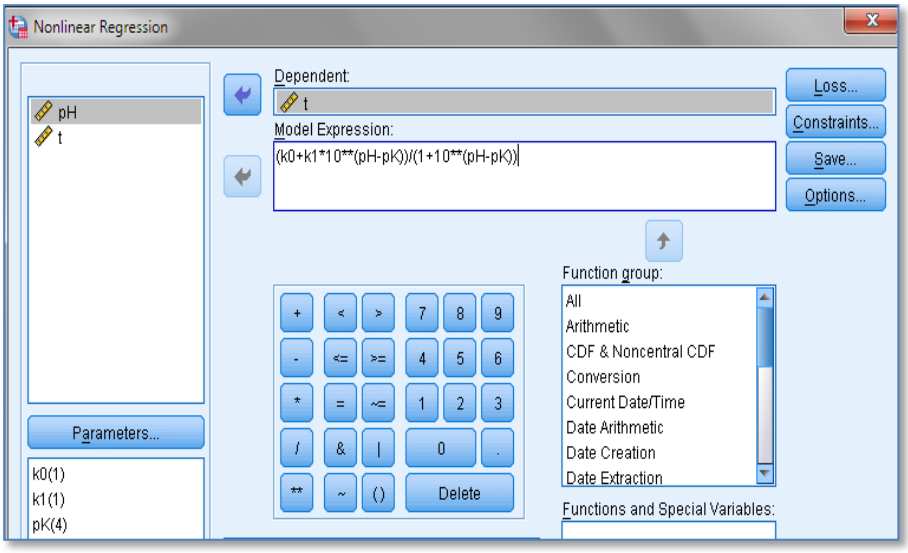
❖ Ανάλυση στο SPSS

Τα βήματα που ακολουθούμε στο *SPSS* για μη γραμμική προσαρμογή είναι τα ακόλουθα:

(1) Πηγαίνουμε *Analyze* → *Regression* → *Nonlinear* και στο παράθυρο *Nonlinear Regression* που ανοίγει εισάγουμε τη μεταβλητή t στο πλαίσιο *Dependent* και στο πάνελ *Parameters* εισάγουμε τις προσαρμόσιμες σταθερές, που για λόγους απλότητας θα συμβολίζουμε τώρα με k0, k1 pK, μαζί με τις αρχικές τους τιμές, π.χ. 1, 1, 4, ως εξής: Κάνουμε *κλικ* στο κουμπί *Parameters*, στο πλαίσιο που εμφανίζεται εισάγουμε στο πεδίο *Name* την πρώτη μεταβλητή, k0, στο πεδίο *Starting Value* την αρχική της τιμή, 1 (σχήμα 10.47) και ολοκληρώνουμε με *Add*. Η σταθερά k0 θα περάσει στο πάνελ *Parameters*. Με τον ίδιο τρόπο εισάγουμε και τις υπόλοιπες σταθερές (σχήμα 10.48).



Σχήμα 10.47. Εισαγωγή προσαρμόσιμων παραμέτρων με αρχικές τιμές



Σχήμα 10.48. Εισαγωγή του μη γραμμικού μοντέλου στο *SPSS*

(2) Στο πλαίσιο *Model Expression* πληκτρολογούμε το μαθηματικό μοντέλο, δηλαδή (k0+k1*10**(pH-pK))/(1+10**(pH-pK)). Για να εισάγουμε στον τύπο μια προσαρμόσιμη σταθερά μπορούμε να κάνουμε διπλό *κλικ* στη σταθερά, στο πάνελ *Parameters*. Ανάλογα μπορούμε να εισάγουμε τις μεταβλητές.

(3) Αν χρειάζεται να εισάγουμε συναρτήσεις, ln, sin κ.ο.κ., μπορούμε ή να τις πληκτρολογήσουμε ή να τις εισάγουμε από το πάνελ *Functions* χρησιμοποιώντας το βέλος.

(4) Με *κλικ* στο *OK* παίρνουμε τα αποτελέσματα του πίνακα στο σχήμα 10.49, που ταυτίζονται με αυτά του *Excel*. Αν δεν υπάρξει σύγκλιση, τότε επαναλαμβάνουμε τη διαδικασία με διαφορετικές αρχικές τιμές. Πηγαίνουμε *Analyze* → *Regression* → *Nonlinear*, κάνουμε *κλικ* στο *Parameters* και *κλικ* στη σταθερά που θέλουμε να αλλάξουμε την αρχική τιμή της. Αλλάζουμε την τιμή στο πλαίσιο *Starting Value* και κάνουμε *κλικ* στο *Change*. Έτσι μπορούμε να αλλάξουμε όσες αρχικές τιμές θέλουμε και να δούμε αν το πρόγραμμα συγκλίνει ή όχι σε λύση. Για παράδειγμα αν χρησιμοποιήσουμε ως αρχικές τιμές k0 = 0, k1 = 1, pK = 4, το μοντέλο δε συγκλίνει και αυτό φαίνεται από την πολύ μεγάλη τιμή του αθροίσματος των τετραγώνων των υπολοίπων που είναι 1158.4, ενώ στη σύγκλιση έχουμε 7.27.

Το πλεονέκτημα του *SPSS* είναι ότι υπολογίζονται οι τυπικές αποκλίσεις των προσαρμόσιμων σταθερών. Συνεπώς μπορούμε να υπολογίσουμε τη μεταβλητή t = (Estimate)/(Std. Error) και να εκτιμήσουμε αν υπάρχουν στατιστικά μη σημαντικές προσαρμόσιμες παράμετροι, δεδομένου ότι τότε ισχύει |t| < 2.

Parameter Estimates

			95% Confidence Interval	
Parameter	Estimate	Std. Error	Lower Bound	Upper Bound
k0	29,392	,756	27,542	31,243
k1	3,008	,510	1,760	4,256
pK	3,701	,083	3,499	3,903

Σχήμα 10.49. Πίνακας αποτελεσμάτων μη γραμμικής προσαρμογής του *SPSS*

Το μειονέκτημα του *SPSS* σχετίζεται με τη γραφική παράσταση. Στο *Excel* η εύκολη γραφική παράσταση μας επιτρέπει να έχουμε εποπτική εικόνα της σύγκλισης του μοντέλου. Στο *SPSS* για να κάνουμε τη γραφική παράσταση πρέπει να αποθηκεύσουμε από το *Save* τις προβλεπόμενες τιμές στο φύλλο εργασίας (*Save* → *Predicted values*) και μετά να κάνουμε τη γραφική παράσταση. Αυτή όμως μορφοποιείται εξαιρετικά δύσκολα.

10.11 ΣΥΣΧΕΤΙΣΗ

Με τον όρο ***συσχέτιση*** *(correlation)* εννοούμε την ύπαρξη γραμμικής και μόνο γραμμικής συσχέτισης μεταξύ δύο μεταβλητών, x και y. Συνεπώς είναι ένα πρόβλημα που συνδέεται άμεσα με τα ελάχιστα τετράγωνα. Αν προσδιορίσουμε την ευθεία των ελαχίστων τετραγώνων, y = a + bx, και διαπιστώσουμε ότι η κλίση b είναι στατιστικά διάφορη του μηδενός, p-value(b) < 0.05 ή $|t_b| > 2$, τότε οι μεταβλητές a και b σχετίζονται γραμμικά. Φυσικά αυτός ο έλεγχος προϋποθέτει ότι τα υπόλοιπα ακολουθούν την κανονική κατανομή.

Εναλλακτικά ο έλεγχος της συσχέτισης γίνεται με τον συντελεστή α) *Pearson* και β) *Spearman*. Ο πρώτος έλεγχος είναι παραμετρικός, ενώ ο δεύτερος μη παραμετρικός. Όπως σε όλους σχεδόν τους ελέγχους, συμπληρωματικά μπορεί να χρησιμοποιηθεί και η μέθοδος *Monte-Carlo με αντιμεταθέσεις*.

Συντελεστής συσχέτισης του Pearson

Ο *συντελεστής Pearson* (*product moment correlation coefficient*), r, υπολογίζεται από τη σχέση

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{(m-1)s_x s_y} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} \qquad (10.24)$$

και παίρνει τιμές στο διάστημα από –1 έως 1. Αρνητικές τιμές του r σημαίνει ότι όταν η μεταβλητή x αυξάνει, η y ελαττώνεται και το αντίστροφο, r = 0 σημαίνει έλλειψη *γραμμικής συσχέτισης* και r > 0 σημαίνει ότι όταν η μια μεταβλητή αυξάνει, αυξάνει και η άλλη.

Θα πρέπει πάντως να τονιστεί και πάλι ότι ο συντελεστής *Pearson* χρησιμοποιείται μόνο για να ελέγξουμε αν δύο τυχαίες συνεχείς μεταβλητές σχετίζονται γραμμικά. Για να ελέγξουμε αν ο συντελεστής r είναι στατιστικά ίσος ή διάφορος του μηδενός χρησιμοποιούμε τη στατιστική συνάρτηση ελέγχου

$$t = \frac{r\sqrt{m-2}}{\sqrt{1-r^2}} \qquad (10.25)$$

που ακολουθεί ασυμπτωματικά την κατανομή *student* με m-2 βαθμούς ελευθερίας με την προϋπόθεση ότι οι δύο μεταβλητές προέρχονται από πληθυσμό που ακολουθεί τη *δισδιάστατη κανονική κατανομή (bivariate normal distribution)*:

$$f(x,y) = \frac{1}{2\sigma_1\sigma_2\pi} e^{-(x-\mu_1)^2/2\sigma_1^2-(y-\mu_2)^2/2\sigma_2^2} \quad (10.26)$$

Η *δισδιάστατη κανονική κατανομή* και η επέκτασή της η *πολυμεταβλητή κανονική κατανομή* παίζουν καθοριστικό ρόλο σε στατιστικούς ελέγχους πολλών μεταβλητών. Παρόλα αυτά είναι ελάχιστα τα στατιστικά προγράμματα που περιέχουν έλεγχο *διμεταβλητής* ή *πολυμεταβλητής κανονικότητας*. Στο *ChemStat* υπάρχει δυνατότητα γραφικού ελέγχου της διμεταβλητής κανονικότητας των μεταβλητών (x, y) με το *διάγραμμα chi-square* που περιγράφεται στο βιβλίο *Applied Multivariate Analysis* των Johnson και Wichern. Πάντως, ως γενικός κανόνας θα πρέπει να είναι ότι σε "ύποπτα" δείγματα, δηλαδή δείγματα που δεν είναι κανονικά ή/και παρουσιάζουν ακραίες τιμές, να χρησιμοποιείται κυρίως ο συντελεστής του *Spearman* ή και ο συντελεστής *Pearson* αλλά με υπολογισμό της p-value με τη μέθοδο *Monte-Carlo με αντιμεταθέσεις*.

Συντελεστής συσχέτισης του Spearman

Ο συντελεστής *Spearman* είναι κατάλληλος και για συνεχείς και για διακριτές μεταβλητές όπως επίσης και για σειριακές μεταβλητές. Υπολογίζεται από τη σχέση

$$\rho = 1 - \frac{6\sum_{i=1}^{m} d_i^2}{m(m^2-1)} \quad (10.27)$$

όπου d_i είναι η διαφορά $r_{x_i} - r_{y_i}$, r_{x_i} είναι ο βαθμός της τιμής x_i της μεταβλητής x, r_{y_i} είναι ο βαθμός της τιμής y_i της μεταβλητής y και m είναι το πλήθος των τιμών x_i ή y_i στο δείγμα. Για να ελέγξουμε αν ο συντελεστής ρ είναι στατιστικά ίσος με μηδέν χρησιμοποιούμε τη συνάρτηση

$$t_s = \rho\sqrt{\frac{m-2}{1-\rho^2}} \quad (10.28)$$

που ακολουθεί ασυμπτωματικά την κατανομή *student* με m-2 βαθμούς ελευθερίας χωρίς προϋποθέσεις σχετικά με την κατανομή των μεταβλητών x και y.

Παράδειγμα 10.11

Στον πίνακα 10.7 δίνονται οι τιμές των αέριων ρύπων (σε TSP - total suspended particulate) σε σταθμό της Κοζάνης σε συνάρτηση με το χρόνο κατά τη διάρκεια του 2005. Μπορούμε να βγάλουμε το συμπέρασμα ότι παρατηρείται αύξηση των ρύπων με το χρόνο;

◆ Η μηδενική υπόθεση που ελέγχουμε είναι

$$H_0: r \text{ ή } \rho = 0 \text{ με εναλλακτική } H_1: r \text{ ή } \rho > 0$$

Παρατηρούμε ότι έχουμε μονόπλευρο έλεγχο. Αυτή είναι η κυρίαρχη περίπτωση στους ελέγχους συσχέτισης και γι αυτό το λόγο στο *ChemStat* υπολογίζεται πάντα η p-value του μονόπλευρου ελέγχου. Αν θέλουμε την p-value του δίπλευρου ελέγχου πολλαπλασιάζουμε την τιμή της p-value επί 2. Στο *SPSS* υπάρχει η δυνατότητα επιλογής μονόπλευρου ή δίπλευρου ελέγχου.

Πίνακας 10.7. Μεταβολή της συγκέντρωσης των αέριων ρύπων στην Κοζάνη.

t, μήνες	0	0.27	0.53	0.90	1.43	1.67	1.93	2.13	2.97	3.1
TSP	115	27	128	67	28	116	76	61	102	170
t, μήνες	3.77	4	4.2	5.2	5.87	6.07	6.27	6.77	7	7.2
TSP	113	130	112	140	132	68	53	92	111	76
t, μήνες	7.87	8.2	8.83	9.07	9.27	9.83	10.2	10.73	11	11.17
TSP	116	52	97	63	95	142	174	148	122	93

❖ **Ανάλυση στο ChemStat**

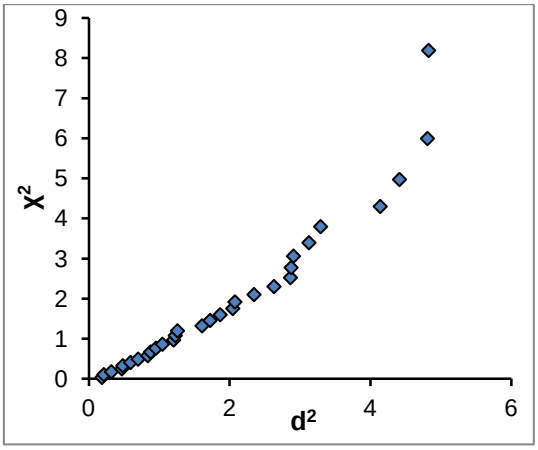
Στο *ChemStat* μπορούμε να ξεκινήσουμε με έναν έλεγχο της *διμεταβλητής κανονικότητας* των μεταβλητών (t, TSP). Για το σκοπό αυτό τοποθετούμε τα δεδομένα σε δύο στήλες και πηγαίνουμε

Πρόσθετα → *ChemStat* → *Normality plots* → *Bivariate chi-square plot*

Στα παράθυρα που ανοίγουν εισάγουμε διαδοχικά τα δύο δείγματα και κάνουμε *κλικ* στο κελί εξόδου. Το πρόγραμμα δημιουργεί δύο στήλες δεδομένων με τίτλους d^2 και x^2. Το d^2 παριστάνει το τετράγωνο μιας

γενικευμένης απόστασης, d_i, του σημείου (x_i, y_i) από το σημείο $(\bar{x}, \bar{y})$, που ονομάζεται και *απόσταση Mahalanobis*. Το x^2 είναι μια τιμή αντίστοιχη της τιμής q στο διάγραμμα Q-Q (παράδειγμα 6.2), με μόνη διαφορά ότι η q υπολογίζεται με την τυπικά κανονική κατανομή, ενώ το x^2 με την κατανομή χ^2 με 2 βαθμούς ελευθερίας. Όταν τα δεδομένα ακολουθούν τη δισδιάστατη κανονική κατανομή το *διάγραμμα chi-square* του χ^2 = x^2 ως προς d^2 = d^2 είναι γραμμικό, διαφορετικά παρατηρούνται συστηματικές αποκλίσεις από τη γραμμικότητα.

Συνεπώς με βάση τις τιμές d^2 = d^2 και χ^2 = x^2 κάνουμε το αντίστοιχο διάγραμμα και ελέγχουμε τη γραμμικότητα των σημείων του. Στο παράδειγμα που εξετάζουμε το διάγραμμα αυτό δίνεται στο σχήμα 10.50, όπου παρατηρούμε μια συστηματική απόκλιση από τη γραμμικότητα στις μεγάλες τιμές d^2. Αυτό δείχνει ότι κατά πάσα πιθανότητα τα δεδομένα δεν προέρχονται από πληθυσμό με τιμές (x, y) που ακολουθούν τη δισδιάστατη κανονική κατανομή και συνεπώς η ανάλυση θα πρέπει να στηριχθεί στον συντελεστή *Spearman*.



Σχήμα 10.50. Διάγραμμα chi-square για έλεγχο *διμεταβλητής κανονικότητας* των δεδομένων του πίνακα 10.7

Για να προσδιορίσουμε τώρα τους συντελεστές συσχέτισης πηγαίνουμε

Πρόσθετα → *ChemStat* → *Correlations* → *Bivariate*

και στα παράθυρα που ανοίγουν εισάγουμε διαδοχικά τα δύο δείγματα,

ορίζουμε το πλήθος των επαναλήψεων για τη μέθοδο *Monte-Carlo με αντιμεταθέσεις* και ορίζουμε το κελί εξόδου των αποτελεσμάτων.

Στον πίνακα των αποτελεσμάτων του σχήματος 10.51 παρατηρούμε ότι το πρόγραμμα ελέγχει την κανονικότητα του κάθε δείγματος χωριστά και την ύπαρξη ακραίων τιμών, ώστε να αξιολογηθεί και με αυτή τη διαδικασία η αξιοπιστία των αποτελεσμάτων του συντελεστή *Pearson*. Παρατηρούμε ότι και τα δύο δείγματα εμφανίζονται να είναι κανονικά, ενώ δεν παρουσιάζονται ακραίες τιμές. Θα μπορούσε συνεπώς κάποιος να υποθέσει ότι πιθανόν να ακολουθούν τη δισδιάστατη κανονική κατανομή. Κι όμως, όπως διαπιστώσαμε παραπάνω, αυτό πιθανότατα δεν συμβαίνει.

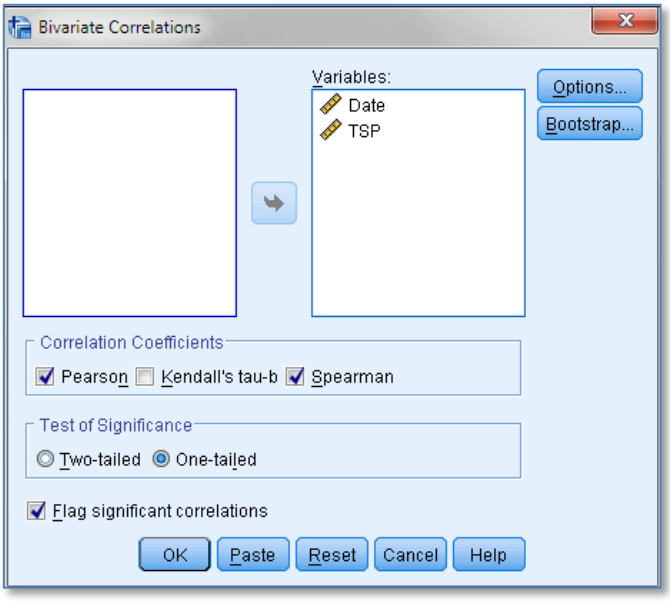
Στο συγκεκριμένο παράδειγμα παρατηρούμε ότι ο συντελεστής *Spearman* δείχνει μια θετική συσχέτιση του χρόνου με τη ρύπανση όμως σε όλους τους ελέγχους έχουμε p-value > 0.05. Συνεπώς η μηδενική υπόθεση δεν μπορεί να απορριφθεί, γεγονός που σημαίνει ότι δεν υπάρχουν επαρκή δεδομένα που να μας οδηγούν στο συμπέρασμα ότι η ρύπανση αυξάνει με το χρόνο.

DATE	TSP			
0,00	115	**Bivariate Correlation test:**		
0,27	27	**Anderson-Darling Normality test:**		
0,53	128	p1-value=	0,247193	Sample1-Normality may be assumed
0,90	67	p2-value=	0,804904	Sample2-Normality may be assumed
1,43	28	**Outliers:**		
1,67	116	sample 1:	may be no outliers	
1,93	76	sample 2:	may be no outliers	
2,13	61	**Pearson Correlation test - 1 tailed**		
2,97	102	r=	0,261598	
3,10	170	p-value=	0,081296	Null hypothesis r = 0 may be assumed at level 0.05
3,77	113	MC iterations=	10000	
4,00	130	p(permut.)=	0,08095	Null hypothesis r = 0 may be assumed at level 0.05
4,20	112	**Spearman correlation Non-Parametric test - 1 tailed**		
5,20	140	r=	0,218291	
5,87	132	p-value=	0,123255	Null hypothesis r = 0 may be assumed at level 0.05
6,07	68	MC iterations=	10000	
6,27	53	p(permut.)=	0,1236	Null hypothesis r = 0 may be assumed at level 0.05
6,77	92			
7,00	111	Elapsed time =	0,294 min	
7,20	76			

Σχήμα 10.51. Τμήμα δεδομένων και πίνακας αποτελεσμάτων για *γραμμική συσχέτιση* μεταβλητών

❖ **Ανάλυση στο Excel**

Στο *Excel* μπορούμε να υπολογίσουμε μόνο τον συντελεστή r από *Δεδομένα* → *Ανάλυση* → *Ανάλυση Δεδομένων* → *Συσχέτιση* (*Data* → *Analysis* → *Data Analysis* → *Correlation)*. Στο παράθυρο που ανοίγει και στο πεδίο *Περιοχή εισόδου* (*Input Range*) εισάγουμε όλη την περιοχή που είναι τα δύο δείγματα και ορίζουμε το κελί εξόδου των αποτελεσμάτων. Παίρνουμε την τιμή r = 0.2616. Για να υπολογίσουμε την πιθανότητα p-value για μονόπλευρο έλεγχο, πρώτα υπολογίζουμε την τιμή t από τη σχέση (10.25), t = 0.2616*SQRT(30-2)/(SQRT(1-0.2616^2)) = 1.434191, και ακολούθως την πιθανότητα p-value με βάση τον τύπο =TDIST(t;m-2;1) = TDIST(1.434191;28;1) = 0.081296.



Σχήμα 10.52. Παράθυρο επιλογής ελέγχων στο *SPSS*

❖ **Ανάλυση στο SPSS**

Στο *SPSS* εισάγουμε τα δεδομένα σε δύο στήλες, με τίτλους έστω Date και TSP, και πηγαίνουμε *Analyze* → *Correlate* → *Bivariate*, όπου στο παράθυρο που εμφανίζεται μεταφέρουμε τις δύο μεταβλητές στο πλαίσιο *Variables* (σχήμα 10.52). Ακολούθως επιλέγουμε τους συντελεστές *Pearson* και *Spearman* και στο πάνελ *Test of Significance* επιλέγουμε τον

έλεγχο *One-tailed*. Με *κλικ* στο *OK* παίρνουμε τα αποτελέσματα του σχήματος 10.53. Όπως αναμένεται, παίρνουμε αποτελέσματα ταυτόσημα με αυτά του *ChemStat*.

Correlations

		Date	TSP
Date	Pearson Correlation	1	,261
	Sig. (1-tailed)		,081
	N	30	30
TSP	Pearson Correlation	,261	1
	Sig. (1-tailed)	,081	
	N	30	30

Correlations

			Date	TSP
Spearman's rho	Date	Correlation Coefficient	1,000	,218
		Sig. (1-tailed)	.	,123
		N	30	30
	TSP	Correlation Coefficient	,218	1,000
		Sig. (1-tailed)	,123	.
		N	30	30

Σχήμα 10.53. Πίνακες αποτελεσμάτων για γραμμική συσχέτιση μεταβλητών στο *SPSS*

Παράδειγμα 10.12

Για να ελεγχθεί αν η ποιότητα του οίνου εξαρτάται από την περιεκτικότητά του σε SO_2, εξετάστηκαν 10 διαφορετικά κρασιά. Η ποιότητά τους βαθμολογήθηκε από γευσιγνώστη στην κλίμακα 0 – 5, όπου 5 είναι το άριστα. Η συγκέντρωση του SO_2 προσδιορίστηκε με ανάλυση με έγχυση του δείγματος σε συνεχή ροή (*flow injection analysis*) σε ppm. Τα αποτελέσματα δίνονται στον πίνακα 10.8. Μπορούμε να συμπεράνουμε ότι η ποιότητα του οίνου καθορίζεται από το SO_2;

◆ Επειδή η μεταβλητή *Ποιότητα* είναι κατηγορική, δεν μπορούμε να εφαρμόσουμε το κριτήριο του *Pearson*. Επομένως ο έλεγχος γίνεται μόνο με τα κριτήριο *Spearman* ή/και τη μέθοδο *Monte-Carlo*. Συνεπώς η υπόθεση που ελέγχουμε είναι

$$H_0: \rho = 0 \text{ με εναλλακτική } H_1: \rho > 0$$

Πίνακας 10.8. Εξάρτηση της ποιότητας του οίνου από τη περιεκτικότητά του σε SO_2.

Ποιότητα	3	4	0	1	5	5	2	4	5	4
SO_2, ppm	2.7	2.5	1.2	0.8	4.5	3.5	1.9	3.8	3.1	4.8

Από τον πίνακα αποτελεσμάτων του *ChemStat* στο σχήμα 10.54 παρατηρούμε ότι p-value(Spearman) = 0.006 και p-value(permutations) =0.0072 και συνεπώς μπορούμε να συμπεράνουμε ότι η αύξηση της συγκέντρωσης του SO_2 στον οίνο αυξάνει την ποιότητα της γεύσης του.

Quality	SO2,ppm	Bivariate Correlation test:		
3	2,7	**Anderson-Darling Normality test:**		
4	2,5	p1-value=	0,120885	Sample1-Normality may be assumed
0	1,2	p2-value=	0,963935	Sample2-Normality may be assumed
1	0,8	**Outliers:**		
5	4,5	sample 1:	may be no outliers	
5	3,5	sample 2:	may be no outliers	
2	1,9			
4	3,8	**Pearson Correlation test - 1 tailed**		
5	3,1	r=	0,831382	
4	4,8	p-value=	0,001435	Null hypothesis r = 0 may be rejected at level 0.05
		MC iterations=	10000	
		p(permut.)=	0,0012	Null hypothesis r = 0 may be rejected at level 0.05
		Spearman correlation Non-Parametric test - 1 tailed		
		r=	0,751785	
		p-value=	0,006077	Null hypothesis r = 0 may be rejected at level 0.05
		MC iterations=	10000	
		p(permut.)=	0,0072	Null hypothesis r = 0 may be rejected at level 0.05
		Elapsed time =	0,098 min	

Σχήμα 10.54. Πίνακας αποτελεσμάτων του *ChemStat*

Παράδειγμα 10.13

Για να ελεγχθεί αν υπάρχει συσχέτιση της κρεατινίνης με την ουρία στον ορό του αίματος προσδιορίστηκαν οι συγκεντρώσεις τους σε 15 υγιείς ενήλικες και τα αποτελέσματα που ελήφθησαν δίνονται στον πίνακα 10.9. Τι συμπέρασμα προκύπτει;

◆ Αν μας ενδιαφέρει να εξετάσουμε γενικά αν υπάρχει συσχέτιση μεταξύ της κρεατινίνης και της ουρίας στον ορό του αίματος, η μηδενική

και η εναλλακτική της μπορεί να διατυπωθούν ως

$$H_0: r \text{ ή } \rho = 0 \text{ με εναλλακτική } H_1: r \text{ ή } \rho \neq 0$$

Δηλαδή πρέπει να εκτελέσουμε έναν δίπλευρο και όχι μονόπλευρο έλεγχο.

Πίνακας 10.9. Συγκεντρώσεις κρεατινίνης και ουρίας στον ορό αίματος.

Κρεατινίνη mg/dL	Ουρία mg/dL
0.6	17
0.58	16.5
0.88	18.1
1.25	17.5
1.1	16.6
0.93	16.3
1.2	18.2
0.76	16.3
0.85	18
0.89	15.2
1.23	17.3
0.71	12.8
0.77	14.7
0.82	13.5
0.54	11.2

Creatinine	urea			
0,6	17	**Bivariate Correlation test:**		
0,58	16,5	**Anderson-Darling Normality test:**		
0,88	18,1	p1-value=	0,38289	Sample1-Normality may be assumed
1,25	17,5	p2-value=	0,088824	Sample2-Normality may be assumed
1,1	16,6	**Outliers:**		
0,93	16,3	sample 1:	may be no outliers	
1,2	18,2	sample 2:	may be no outliers	
0,76	16,3	**Pearson Correlation test - 1 tailed**		
0,85	18	r=	0,559149	
0,89	15,2	p-value=	0,015117	Null hypothesis r = 0 may be rejected at level 0.05
1,23	17,3	MC iterations=	10000	
0,71	12,8	p(permut.)=	0,0148	Null hypothesis r = 0 may be rejected at level 0.05
0,77	14,7	**Spearman correlation Non-Parametric test - 1 tailed**		
0,82	13,5	r=	0,575514	
0,54	11,2	p-value=	0,012389	Null hypothesis r = 0 may be rejected at level 0.05
		MC iterations=	10000	
		p(permut.)=	0,0131	Null hypothesis r = 0 may be rejected at level 0.05

Σχήμα 10.55. Πίνακας αποτελεσμάτων του *ChemStat*

Ο πίνακας αποτελεσμάτων του *ChemStat* δίνεται στο σχήμα 10.55. Με βάση την παραπάνω παρατήρηση όλες οι τιμές της p-value πρέπει να διπλασιαστούν. Παρατηρούμε ότι r = 0.559 και ρ = 0.576 με τιμές p-value από 0.024 μέχρι 0.03. Συνεπώς υπάρχει στατιστικά σημαντική συσχέτιση της κρεατινίνης με την ουρία στον ορό του αίματος, πιθανότητα επειδή και οι δύο σχετίζονται με τη λειτουργία των νεφρών.

Η ύπαρξη συσχέτισης μεταξύ δύο μεταβλητών δεν συνεπάγεται μια σχέση *αιτίας – αιτιατού*. Όμως στην έρευνα μπορεί να δώσει το έναυσμα για να διερευνηθεί η ύπαρξη μιας τέτοιας σχέσης.

10.12 ΜΕΡΙΚΗ ΣΥΣΧΕΤΙΣΗ

Σε πολλά προβλήματα, κυρίως κλινικής χημείας, υπάρχει μια αλληλοσυσχέτιση πολλών μεταβλητών. Η ***μερική συσχέτιση*** (*partial correlation*) χρησιμοποιείται προκειμένου να εξετάσουμε τη συσχέτιση δύο μεταβλητών, ενώ συγχρόνως κρατάμε σταθερή την επίδραση μιας ή περισσοτέρων άλλων μεταβλητών.

Στην απλή περίπτωση του ελέγχου των μεταβλητών A, B, C, όταν εξετάζουμε τη συσχέτιση των A, B θεωρώντας τη C σταθερή, ο *συντελεστής μερικής συσχέτισης* ορίζεται από τη σχέση

$$r_{ABC} = \frac{r_{AB} - r_{AC}r_{BC}}{\sqrt{(1 - r_{AC}^2)(1 - r_{BC}^2)}} \qquad (10.29)$$

όπου r_{AB}, r_{AC} και r_{BC} είναι οι συντελεστές συσχέτισης *Pearson* ή *Spearman* των μεταβλητών A-B, A-C και B-C. Για τον έλεγχο σημαντικότητας χρησιμοποιούμε τη στατιστική συνάρτηση ελέγχου

$$t = \frac{r_{ABC}\sqrt{m - 3}}{\sqrt{1 - r_{ABC}^2}} \qquad (10.30)$$

που ακολουθεί την κατανομή *student* με m-3 βαθμούς ελευθερίας.

Οι σχέσεις (10.29) και (10.30) μπορεί να επεκταθούν για τη γενική περίπτωση που ελέγχουμε τη συσχέτιση δύο μεταβλητών A και B θεωρώντας k άλλες μεταβλητές, C, D, …, ως σταθερές. Το πρόβλημα της μερικής συσχέτισης απλοποιείται με βάση την ακόλουθη παρατήρηση. Ο συντελεστής συσχέτισης της σχέσης (10.29) προκύπτει εναλλακτικά ως εξής. Εφαρμόζουμε τη μέθοδο των ελαχίστων τετραγώνων χρησιμο-

ποιώντας ως y τη μεταβλητή A και ως x τη C και υπολογίζουμε τα υπόλοιπα, έστω A_r. Ακολούθως εφαρμόζουμε τη μέθοδο των ελαχίστων τετραγώνων χρησιμοποιώντας ως y τη B και ως x πάλι τη C και υπολογίζουμε τα υπόλοιπα, έστω B_r. Ο συντελεστής συσχέτισης r_{ABC} προκύπτει από τη σχέση (10.24) ή (10.27) αν χρησιμοποιήσουμε ως μεταβλητές x και y τις A_r και B_r, αντίστοιχα. Η μέθοδος αυτή επεκτείνεται άμεσα στη γενική περίπτωση που ελέγχουμε τη μερική συσχέτιση των A και B θεωρώντας k άλλες μεταβλητές ως σταθερές.

Παράδειγμα 10.14

Στον πίνακα 10.10 παρέχεται η μεταβολή της κακής χοληστερόλης (LDL) με την ηλικία και τον δείκτη μάζας σώματος (BMI - Body Mass Index). Να εξεταστεί αν ο δείκτης μάζας σώματος επηρεάζει τη συσχέτιση της LDL με την ηλικία.

Πίνακας 10.10. Μεταβολή της LDL με την ηλικία και τον δείκτη μάζας σώματος (BMI).

Ηλικία Έτη	LDL mg/dL	BMI kg/m^2	Ηλικία Έτη	LDL mg/dL	BMI kg/m^2
22	146	26	46	162	20
26	131	18	50	156	19
28	140	20	53	165	23
30	136	19	55	170	27
31	145	22	57	168	27
35	140	16	59	182	17
36	155	19	62	177	24
38	143	22	64	157	25
42	173	28	68	165	31
45	165	14			

◆ Για να εξετάσουμε αν ο δείκτης μάζας σώματος επηρεάζει την συσχέτιση της LDL με την ηλικία πρέπει να προσδιορίσουμε τον συντελεστή μερικής συσχέτισης των μεταβλητών Ηλικία και LDL όταν ο δείκτης μάζας σώματος, BMI, κρατείται σταθερός.

❖ **Ανάλυση στο ChemStat**

Τοποθετούμε τα δεδομένα σε τρεις στήλες με τίτλους age, LDL, BMI και αρχικά ελέγχουμε την απλή συσχέτιση των μεταβλητών age-LDL και LDL-BMI. Τα αποτελέσματα δίνονται στον πίνακα του σχήματος 10.56.

Λόγω της φύσης των δεδομένων πρέπει να εμπιστευτούμε κυρίως τον έλεγχο *Spearman* αν και όλοι οι έλεγχοι οδηγούν στο ίδιο συμπέρασμα σε ό,τι αφορά τη συσχέτιση της LDL με την ηλικία. Παρατηρούμε ότι υπάρχει μια ισχυρή θετική συσχέτιση αυτών των δύο μεταβλητών που σημαίνει ότι με την αύξηση της ηλικίας αυξάνει η LDL στο αίμα. Σε ό,τι αφορά τη συσχέτιση των μεταβλητών LDL και BMI, αυτή είναι θετική και αν λάβουμε υπόψη τον έλεγχο *Spearman* η συσχέτιση είναι στατιστικά σημαντική.

Bivariate Correlation test: age vs. LDL		
Pearson Correlation test - 1 tailed		
r=	0,789825	
p-value=	2,88E-05	Null hypothesis r = 0 may be rejected at level 0.05
p(permut.)=	0	Null hypothesis r = 0 may be rejected at level 0.05
Spearman correlation Non-Parametric test - 1 tailed		
r=	0,776266	
p-value=	4,66E-05	Null hypothesis r = 0 may be rejected at level 0.05
p(permut.)=	0,00005	Null hypothesis r = 0 may be rejected at level 0.05
Bivariate Correlation test: LDL vs. BMI		
Pearson Correlation test - 1 tailed		
r=	0,33353	
p-value=	0,081438	Null hypothesis r = 0 may be assumed at level 0.05
p(permut.)=	0,081	Null hypothesis r = 0 may be assumed at level 0.05
Spearman correlation Non-Parametric test - 1 tailed		
r=	0,411376	
p-value=	0,040072	Null hypothesis r = 0 may be rejected at level 0.05
p(permut.)=	0,04075	Null hypothesis r = 0 may be rejected at level 0.05

Σχήμα 10.56. Συγκεντρωτικός πίνακας αποτελεσμάτων για απλή συσχέτιση των μεταβλητών age-LDL και LDL-BMI

Για τη *μερική συσχέτιση* πηγαίνουμε *Πρόσθετα* → *ChemStat* → *Correlations* → *Partial Correlation* και στα παράθυρα που ανοίγουν εισάγουμε διαδοχικά τα τρία δείγματα, το κάθε ένα ΜΑΖΙ με τον τίτλο του, ορίζουμε το πλήθος των επαναλήψεων για τη μέθοδο *Monte-Carlo με αντιμεταθέσεις* και ορίζουμε το κελί εξόδου των αποτελεσμάτων.

Στον πίνακα των αποτελεσμάτων του σχήματος 10.57 παρατηρούμε ότι εξακολουθεί να υπάρχει μια ισχυρή θετική συσχέτιση μεταξύ LDL και ηλικίας. Συνεπώς αν και το βάρος επηρεάζει την LDL δεν έχει πρακτικά σημαντική επίδραση στη συσχέτιση της LDL με τη ηλικία, δεδομένου ότι ο συντελεστής *Spearman* ελαττώνεται ελάχιστα από την τιμή $\rho = 0.776$ όταν αγνοούμε τον παράγοντα βάρος στην τιμή $\rho = 0.735$ όταν αφαιρούμε την επίδραση του βάρους.

Partial correlation based on regression:		
Partial correlation of age versus LDL controlling by BMI		
Pearson Correlation test - 1 tailed		
r=	0,759693	
p-value=	0,000127	Null hypothesis r = 0 may be rejected at level 0.05
MC iterations=	10000	
p(permut.)=	0	Null hypothesis r = 0 may be rejected at level 0.05
Spearman correlation Non-Parametric test - 1 tailed		
r=	0,735018	
p-value=	0,000255	Null hypothesis r = 0 may be rejected at level 0.05
MC iterations=	10000	
p(permut.)=	0,0001	Null hypothesis r = 0 may be rejected at level 0.05
Results based on recursive formulas when the controlling variables are less than 5		
r(Pearson)=	0,759693	
p-value=	0,000127	
r(Spearman)=	0,735018	
p-value=	0,000255	

Σχήμα 10.57. Πίνακας αποτελεσμάτων για μερική συσχέτιση των μεταβλητών age-LDL όταν η επίδραση της BMI διατηρείται σταθερή

❖ Ανάλυση στο SPSS

Το *SPSS* παρουσιάζει το σημαντικό μειονέκτημα ότι εκτελεί μόνο τον παραμετρικό έλεγχο *Pearson*. Για την εφαρμογή αυτού του ελέγχου εισάγουμε τα δεδομένα σε τρεις στήλες, με τίτλους έστω age, LDL και BMI και πηγαίνουμε *Analyze* → *Correlate* → *Partial*. Μεταφέρουμε τις μεταβλητές age και LDL των οποίων τη συσχέτιση θέλουμε να εξετάσουμε στο πλαίσιο *Variables*, και τη μεταβλητή ελέγχου, BMI, στο πλαίσιο *Controlling for*. Τα αποτελέσματα δίνονται στον πίνακα *Correlations* και ταυτίζονται με τα αντίστοιχα του *ChemStat* του ελέγχου *Pearson*.

Correlations

Control Variables			age	LDL
BMI	age	Correlation	1,000	,760
		Significance (1-tailed)	.	,000
		df	0	16
	LDL	Correlation	,760	1,000
		Significance (1-tailed)	,000	.
		df	16	0

Σχήμα 10.58. Πίνακας αποτελεσμάτων στο *SPSS* για μερική συσχέτιση των μεταβλητών age-LDL όταν η επίδραση της BMI διατηρείται σταθερή

ΑΣΚΗΣΕΙΣ

10.1. Στον παρακάτω πίνακα δίνονται δύο σειρές σημείων, (x, y_1) και (x, y_2). Να ελεγχθεί η γραμμικότητά τους με τη μέθοδο των ελαχίστων τετραγώνων. Να γίνουν τα διαγράμματα των υπολοίπων. Τι συμπεράσματα βγάζετε;

x	y_1	y_2	x	y_1	y_2	x	y_1	y_2	x	y_1	y_2
0.5		-6	3.0	9.5	13	5.5	21.3	27	8.0	39.5	34
1.0	4.5	0	3.5	11.4	9	6.0	24.5	27	8.5	43.5	41
1.5	5.4	5	4.0	13.5	14	6.5	27.9	30	9.0	48.5	46
2.0	6.5	3	4.5	15.8	21	7.0	31.5	30	9.5		42
2.5	7.9	8	5.0	18.5	25	7.5	35.5	38			

10.2. Στον πίνακα 10.4 δίνεται η επίδραση της θερμοκρασίας Τ στη μοριακή θερμοχωρητικότητα C του αερίου O_2. Επειδή πολλές φορές η θερμοχωρητικότητα C εκφράζεται και ως πολυώνυμο δευτέρου βαθμού, να γίνει προσαρμογή των δεδομένων στο μοντέλο $C = a + bT + cT^2$. Είναι το μοντέλο αυτό καλύτερο από το μοντέλο $C = a + bT + cT^{-2}$ που εξετάστηκε στο παράδειγμα 10.3;

10.3. Η αποσύνθεση του N_2O_5 σε O_2 και NO_2 όταν είναι σε διάλυμα CCl_4 περιγράφεται από την απλή εξίσωση

$$c = c_0 e^{-kt}$$

όπου c είναι η συγκέντρωση του N_2O_5 στο διάλυμα σε χρόνο t, c_0 είναι η τιμή του c όταν t = 0 και k η ειδική ταχύτητα. Στον πίνακα που ακολουθεί δίνεται η μεταβολή του c με τον χρόνο t όταν η αποσύνθεση του N_2O_5 γίνεται στους 45 ^{0}C. Να υπολογιστούν οι σταθερές c_0 και k χρησιμοποιώντας α) τη μέθοδο των ελαχίστων τετραγώνων και β) το πρόγραμμα *Επίλυση*. Επίσης να υπολογιστούν οι τυπικές αποκλίσεις και τα 95 % διαστήματα εμπιστοσύνης των c_0 και k.

t, min	5	10	20	30	40	50	60	70	80
c, M	1.70	1.35	0.90	0.70	0.40	0.35	0.20	0.15	0.07

10.4. Η κινητική της θερμικής ισομερίωσης του δικυκλο[2.1.1]εξανίου περιγράφεται από τη θεωρητική εξίσωση

$$y = \exp\{-a\ t \exp[-b(1/T - 1/620)]\}$$

όπου y είναι το κλάσμα του αντιδρώντος που παραμένει αμετάβλητο μετά από χρόνο t (σε min) στη θερμοκρασία πειράματος T και a, b είναι σταθερές. Η πειραματική μελέτη του συστήματος αυτού έγινε από τους R. Srinivasan και A. Levi και ένα μέρος των μετρήσεων δίνεται στον παρακάτω πίνακα. Να υπολογιστούν οι σταθερές a και b της εξίσωσης.

t, min	15	30	45.1	60	90	120	150
y(T=600K)				0.949		0.900	
y(T=620K)	0.938	0.877	0.827	0.787	0.696		0.582
y(T=639K)	0.808	0.655		0.425	0.309		

10.5. Δέκα φοιτητές που έδωσαν το μάθημα των μαθηματικών ρωτήθηκαν για τις ώρες που αφιέρωσαν στο μάθημα εβδομαδιαίως. Στον επόμενο πίνακα δίνονται τα στοιχεία αυτά μαζί με τον βαθμό που πήραν.

Ώρες	3	4	2	3.5	4.5	10	0	3.5	7	1
Βαθμός	4	6.5	4	5	8	9.5	1	3	6	2

Εξετάστε α) αν σχετίζονται οι ώρες μελέτης με το βαθμό, β) το βαθμό που θα πάρει ένας φοιτητής αν διαβάζει 5 ώρες τη βδομάδα, και γ) πόσες ώρες πρέπει να διαβάζει για να πάρει 5.

10.6. Σε μία μελέτη καθαρισμού αποβλήτων που χαρακτηρίζονται από υψηλή στάθμη απαιτούμενου βιοχημικού οξυγόνου και στερεής ύλης, εξετάζεται το ποσοστό (%) μείωσης του απαιτούμενου οξυγόνου σε συνάρτηση με το ποσοστό (%) μείωσης των στερεών. Η μελέτη οδήγησε στον παρακάτω πίνακα. Εξετάστε αν σχετίζονται οι δύο παραπάνω μεταβλητές και αν σχετίζονται προβλέψτε το ποσοστό μείωσης του απαιτούμενου οξυγόνου όταν το ποσοστό μείωσης των στερεών είναι 10%.

Στερεά ύλη	3	7	11	15	18	27
Απαιτούμενο οξυγόνο	5	11	21	16	16	28

10.7. Έστω τα δεδομένα του παρακάτω πίνακα που δίνουν την εξάρτηση του σημείου βρασμού του νερού με το υψόμετρο. Να προσαρμοστούν τα δεδομένα αυτά σε ευθεία και να βρεθεί το υψόμετρο όταν η θερμοκρασία βρασμού του νερού είναι 97 ^{0}C. Μπορούμε να υπολογίσουμε διαστήματα εμπιστοσύνης;

t, ^{0}C	90.2	90.1	92.1	92.4	93	93.2	93.8	93.9	94.1
h, m	3055	3055	2435	2335	2160	2090	1900	1865	1855
t, ^{0}C	94	95.3	95.8	98.6	98.1	99.2	99.9	100	
h, m	1860	1480	1020	440	655	280	45	0	

10.8. Η μεταβολή της εντροπίας S του CO_2 περιγράφεται από τη σχέση $\Delta S = S - S_{300} = aT + bT^{-2} + c\ln T$, όπου S_{300} είναι η εντροπία του αερίου στη θερμοκρασία των 300 K. Να προσδιοριστούν οι σταθερές a, b και c με βάση τις τιμές του παρακάτω πίνακα.

T, K	300	400	500	600	700	800	900	1000	1100
ΔS, kJ $K^{-1}mol^{-1}$	0	0.89	1.78	2.67	3.55	4.44	5.32	6.21	7.09

10.9. Να προσδιοριστεί η καμπύλη αναφοράς με βάση τα παρακάτω δεδομένα που αφορούν τη συγκέντρωση τανίνης σε φυτά. Η συγκέντρωση προσδιορίζεται με ατομική απορρόφηση. Ακολούθως να υπολογιστεί η συγκέντρωση τανίνης σε δείγμα που έδωσε απορρόφηση 0.211.

Απορρόφηση	0.084	0.183	0.326	0.464	0.643
c, mg/mL	0.123	0.288	0.562	0.921	1.42

10.10. Εννέα πρότυπα υδατικά διαλύματα Ag αναλύονται με ατομική απορρόφηση και η απορρόφηση καταγράφεται σε συνάρτηση με τη συγκέντρωση του Ag στον παρακάτω πίνακα

c, ng/mL	0	2	4	6	8	10	12	14	16
Απορρόφηση	0.003	0.05	0.11	0.15	0.21	0.25	0.29	0.35	0.40

Σε δύο άγνωστης συγκέντρωσης διαλύματα γίνονται μία μέτρηση στο πρώτο που έδωσε απορρόφηση 0.132 και 3 μετρήσεις στο δεύτερο που έδωσαν τις τιμές απορρόφησης 0.314, 0.328 και 0.312. Να προσδιοριστούν:

α) Η καμπύλη αναφοράς των μετρήσεων.

β) Οι συγκεντρώσεις του Ag στα διαλύματα.

γ) Το ελάχιστο όριο ανίχνευσης του Ag.

10.11. Για τον προσδιορισμό του Fe σε δείγματα αυγών, τα δείγματα, μετά από κατάλληλη επεξεργασία, ομογενοποιήθηκαν με άλεση και σε ίσες ποσότητες των 10 g του ομογενοποιημένου υλικού προστίθενται 0, 0.2, 0.5, 1.0 και 1.5 μg Fe. Τα δείγματα αυτά διαλυτοποιούνται και μετράται η απορρόφησή τους με φασματοφωτομετρία ατομικής απορρόφησης. Να προσδιοριστεί η άγνωστη συγκέντρωση του σιδήρου λαμβάνοντας υπόψη ότι ελήφθησαν οι ακόλουθες τιμές απορρόφησης: 0.009, 0.012, 0.021, 0.052, 0.08.

10.12. Η άγνωστη συγκέντρωση μολύβδου σε υδατικό διάλυμα προσδιορίζεται με εσωτερική βαθμονόμηση. Δημιουργούνται 6 διαλύματα όγκου 50 mL το κάθε ένα, που περιέχουν 0, 0.5, 1.0, 1.5, 2.0 και 2.5 mL ενός διαλύματος καδμίου συγκέντρωσης 10.0 mg/L. Ακολούθως με *αναδιαλυτική βολταμμετρία (strippng voltametry)* προσδιορίστηκε ο λόγος R των ρευμάτων κορυφής που οφείλονται στον μόλυβδο και στο κάδμιο. Ελήφθησαν οι τιμές: 0.86, 1.11, 1.44, 1.74, 2.04, 2.33, αντίστοιχα. Να προσδιοριστεί η άγνωστη συγκέντρωση του μολύβδου.

10.13. Στο παράδειγμα 10.4 να εξετασθεί αν υπάρχει συσχέτιση μεταξύ της ποσότητας που εξατμίζεται y και της μέσης υγρασίας h όταν οι υπόλοιπες μεταβλητές είναι σταθερές. Επίσης να εξετασθεί η πιθανή μερική συσχέτιση μεταξύ της μεταβλητής y και της μέγιστης θερμοκρασία του αέρα.

10.14. Τέσσερα πρότυπα υδατικά διαλύματα Ag αναλύονται με ατομική απορρόφηση και η απορρόφηση καταγράφεται σε συνάρτηση με τη συγκέντρωση του Ag στον παρακάτω πίνακα

c, μg/mL	0	5	10	20
Απορρόφηση	0.003	0.05	0.09	0.18

Να προσδιοριστεί η άγνωστη συγκέντρωση που έδωσε απορρόφηση 0.121 και το ελάχιστο όριο ανίχνευσης του Ag.

10.15. Στο παράδειγμα 10.11 αντικαταστήστε το ζεύγος τιμών (8.83, 97) με το (8.83, 250) και επαναλάβατε την ανάλυση. Τι συμπεράσματα προκύπτουν;

Κεφάλαιο 11

ΕΞΟΜΑΛΥΝΣΗ, ΠΑΡΑΓΩΓΙΣΗ, ΟΛΟΚΛΗΡΩΣΗ ΔΕΔΟΜΕΝΩΝ

11.1 ΕΞΟΜΑΛΥΝΣΗ ΔΕΔΟΜΕΝΩΝ

Όπως ήδη έχει αναφερθεί, τα αποτελέσματα ενός ποσοτικού πειράματος περιέχουν πάντα σφάλματα με αποτέλεσμα να αποκλίνουν λίγο ή πολύ από την αναμενόμενη θεωρητική συμπεριφορά. Για τον ίδιο λόγο αν έχουμε πειραματικά δεδομένα της γενικής μορφής (x_i, y_i), όπου $i = 1, 2, ..., m$, και κάνουμε τη γραφική παράσταση της μεταβολής του y με το x θα παρατηρήσουμε μια μικρή ή μεγάλη διασπορά των πειραματικών σημείων γύρω από τη θεωρητική καμπύλη που τα περιγράφει. Αυτή η διασπορά ονομάζεται συνήθως *θόρυβος* (*noise*) και, όπως αναφέρθηκε, οφείλεται στα τυχαία σφάλματα που υπεισέρχονται στις μετρήσεις.

Ένας από τους στόχους σε ένα πείραμα είναι ο περιορισμός του θορύβου, επειδή συνήθως επηρεάζει την ανάλυση των αποτελεσμάτων του πειράματος. Σε ένα δεύτερο επίπεδο ο θόρυβος μπορεί να περιοριστεί κατά την ανάλυση των πειραματικών αποτελεσμάτων με μαθηματικές μεθόδους. Η διαδικασία αυτή ονομάζεται ***εξομάλυνση δεδομένων*** και σχετίζεται στενά με τις μεθόδους προσαρμογής καμπύλης. Για παράδειγμα, αν έχουμε προσδιορίσει με τη μέθοδο των ελαχίστων τετραγώνων τη συνάρτηση που περιγράφει τα πειραματικά δεδομένα (x_i, y_i), δηλαδή τη συνάρτηση για την οποία ισχύει $y_i \approx f(x_i)$, τότε η συνάρτηση αυτή θα δίνει τις *ομαλές* τιμές του y, δηλαδή τις τιμές y_i στις οποίες έχει περιοριστεί ο θόρυβος.

Παράδειγμα 11.1

Το N_2O_5 σε διάλυμα με CCl_4 διασπάται σε O_2 και NO_2. Στον πίνακα 11.1 δίνεται η μεταβολή της συγκέντρωσης c του N_2O_5 με τον χρόνο t όταν η αποσύνθεση γίνεται στους 45 ^{0}C. Λαμβάνοντας υπόψη ότι η διάσπαση του N_2O_5 είναι αντίδραση πρώτης τάξης, να εξομαλυνθούν τα πειραματικά δεδομένα.

Πίνακας 11.1. Μεταβολή της συγκέντρωσης c του N_2O_5 με τον χρόνο t.

t, min	c, M	t, min	c, M
5	1.70	45	0.35
10	1.35	50	0.35
15	1.15	55	0.25
20	0.90	60	0.20
25	0.80	65	0.20
30	0.70	70	0.15
35	0.55	75	0.10
40	0.40		

◆ Εφόσον η θερμική διάσπαση του N_2O_5 είναι αντίδραση πρώτης τάξης, ισχύει

$$c = c_0 e^{-kt} \qquad (11.1)$$

όπου c_0 είναι η συγκέντρωση του N_2O_5 σε χρόνο μηδέν (t = 0) και k είναι η *ειδική ταχύτητα* της αντίδρασης. Επομένως αν με ελάχιστα τετράγωνα προσδιορίσουμε τις σταθερές c_0 και k, τότε η παραπάνω εξίσωση θα δίνει τις ομαλές τιμές της συγκέντρωσης. Με βάση αυτή την παρατήρηση εργαζόμαστε ως εξής:

(1) Ανοίγουμε ένα φύλλο εργασίας του *Excel*, στο C1 εισάγουμε τον τίτλο ΘΕΡΜΙΚΗ ΔΙΑΣΠΑΣΗ ΠΕΝΤΟΞΕΙΔΙΟΥ ΤΟΥ ΑΖΩΤΟΥ, στα κελιά A3, A4 πληκτρολογούμε τους τίτλους c_0 = και k =, αντίστοιχα, στα κελιά A6, B6 εισάγουμε τους τίτλους t, min και c, M και στην περιοχή A7:B21 εισάγουμε τις τιμές του πίνακα 11.1.

(2) Κάνουμε τη γραφική παράσταση της μεταβολής του c με το t και προχωρούμε με δεξί *κλικ* επάνω σε ένα από τα πειραματικά σημεία. Στη λίστα επιλογών που εμφανίζεται επιλέγουμε *Προσθήκη γραμμής τάσης* (*Add Trendline*) και από το *Trendline Options* επιλέγουμε *Εκθετικός* (*Exponential*). Κάνουμε *κλικ* στα *Προβολή εξίσωσης στο γράφημα* (*Display equation on chart*) και *Προβολή τιμής R-τετράγωνο στο γράφημα* (*Display R-squared on chart*).

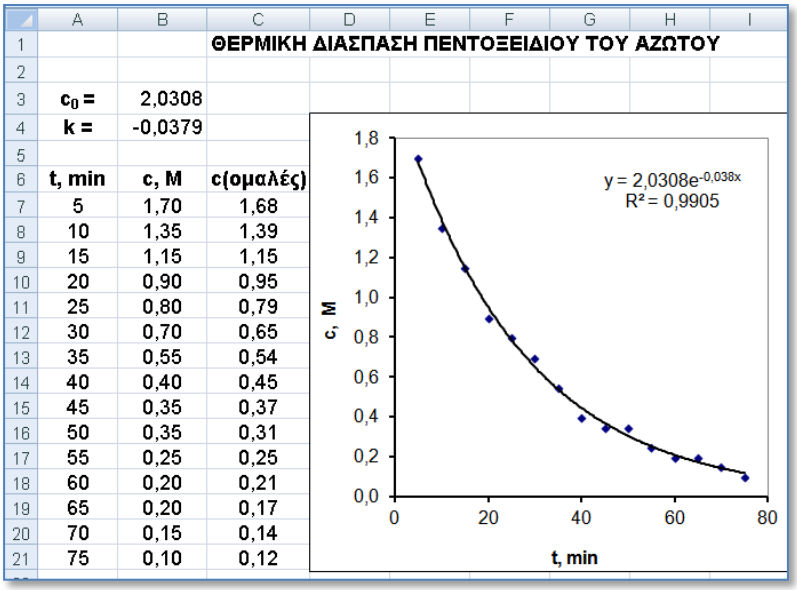
(3) Παρατηρούμε ότι η εξίσωση των ελαχίστων τετραγώνων είναι η

$$c = 2.0308\ e^{-0.0379\ t}$$

με R^2 = 0.9905. Από την τιμή αυτή και την εικόνα της γραφικής παράστασης προκύπτει ότι η παραπάνω εξίσωση περιγράφει απόλυτα ικανοποιητικά τα πειραματικά δεδομένα και συνεπώς μπορεί να

χρησιμοποιηθεί για την εξομάλυνσή τους (σχήμα 11.1).

(4) Για την εξομάλυνση τώρα των πειραματικών δεδομένων, στο κελί B3 εισάγουμε τον αριθμό 2.0308, στο B4 το –0.0379, στο C6 τον τίτλο c(ομαλές) και στο C7 πληκτρολογούμε τον τύπο =B$3*EXP(B$4*A7). Πατάμε *Enter* και συμπληρώνουμε την περιοχή C8:C21 με την διαδικασία της αυτόματης συμπλήρωσης. Στην περιοχή C7:C21 θα εμφανιστούν οι *ομαλές τιμές* των πειραματικών δεδομένων.



	A	B	C
1			ΘΕΡΜΙΚΗ ΔΙΑΣΠΑΣΗ ΠΕΝΤΟΞΕΙΔΙΟΥ ΤΟΥ ΑΖΩΤΟΥ
2			
3	c_0 =	2,0308	
4	k =	-0,0379	
5			
6	t, min	c, M	c(ομαλές)
7	5	1,70	1,68
8	10	1,35	1,39
9	15	1,15	1,15
10	20	0,90	0,95
11	25	0,80	0,79
12	30	0,70	0,65
13	35	0,55	0,54
14	40	0,40	0,45
15	45	0,35	0,37
16	50	0,35	0,31
17	55	0,25	0,25
18	60	0,20	0,21
19	65	0,20	0,17
20	70	0,15	0,14
21	75	0,10	0,12

Σχήμα 11.1. Τμήμα της οθόνης του υπολογιστή στο Παράδειγμα 11.1

Παρατήρηση. Ενώ χρησιμοποιούμε τον όρο *εξομάλυνση δεδομένων*, που υπονοεί την εξομάλυνση και των τιμών x και των τιμών y, στην πραγματικότητα εξομαλύνουμε μόνο τις τιμές y θεωρώντας ότι ο θόρυβος στις τιμές x είναι αμελητέος.

Παράδειγμα 11.2

Στον πίνακα 11.2 δίνεται η κορυφή ενός χρωματογραφήματος που έχει αρκετό θόρυβο. Να εξομαλυνθούν τα πειραματικά δεδομένα λαμβάνοντας υπόψη ότι η κορυφή αυτή περιγράφεται από την εξίσωση

$$y = a e^{-(x-m)^2 / s} \qquad (11.2)$$

όπου a, s και m είναι σταθερές.

Πίνακας 11.2. Τιμές (x, y) που περιγράφουν μια χρωματογραφική κορυφή.

x	y	x	y
2	-0.001	11	0.168
3	0.002	12	0.150
4	0.005	13	0.086
5	0.001	14	0.038
6	0.034	15	0.021
7	0.093	16	-0.010
8	0.136	17	0.016
9	0.189	18	-0.010
10	0.195	19	-0.004

◆ Εφόσον η παραπάνω εξίσωση δεν μπορεί να μετατραπεί σε γραμμική, μπορούμε να χρησιμοποιήσουμε το πρόγραμμα *Επίλυση* (*Solver*) του *Excel*. Για το σκοπό αυτό ανοίγουμε ένα φύλλο εργασίας, στο D1 γράφουμε τον τίτλο ΕΞΟΜΑΛΥΝΣΗ ΔΕΔΟΜΕΝΩΝ ΜΕ ΤΟ ΠΡΟΓΡΑΜΜΑ ΕΠΙΛΥΣΗ και προχωρούμε στα παρακάτω βήματα:

(1) Στην περιοχή A2:A4 πληκτρολογούμε τους τίτλους a =, s =, m = και στην B2:B4 εισάγουμε τις αρχικές τιμές αυτών των σταθερών, έστω αυθαίρετα τις 1, 5, 5.

(2) Στην περιοχή A6:D6 γράφουμε τους τίτλους x, y, y(ομαλ.) και SR και εισάγουμε τα πειραματικά δεδομένα του πίνακα 11.2 στις στήλες A7:A24 και B7:B24 (σχήμα 11.2).

(3) Στο κελί C7 πληκτρολογούμε τον τύπο

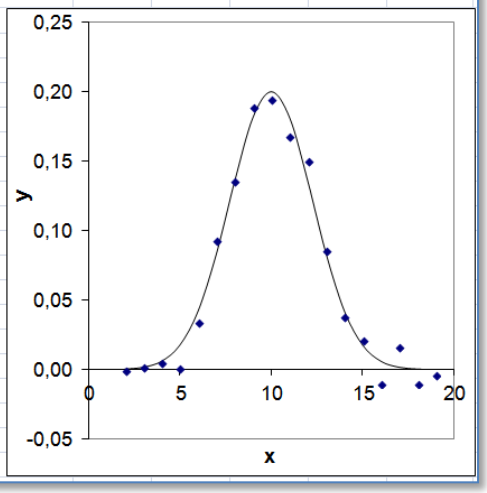
```
=B$2*EXP(-((A7-B$4)^2)/B$3)
```

που εκφράζει την εξίσωση (11.2). Στο D7 εισάγουμε τον τύπο =(B7-C7)^2 και συμπληρώνουμε την περιοχή C7:D24 με τη διαδικασία της αυτόματης συμπλήρωσης.

(4) Στο D4 προσδιορίζουμε το άθροισμα των τετραγώνων των υπολοίπων (SSR) και συνεπώς εισάγουμε τον τύπο =SUM(D7:D24).

(5) Από το *Δεδομένα (Data)* → *Ανάλυση (Analysis)* κάνουμε *κλικ* στο *Επίλυση* (*Solver*) και συμπληρώνουμε το παράθυρο διαλόγου που εμφανίζεται. Συγκεκριμένα, στο *Ορισμός στόχου* (*Set Objective*) συμπληρώνουμε το κελί D4, όπου βρίσκεται το άθροισμα των τετραγώνων των υπολοίπων. Στο *Σε* (*To*) κάνουμε *κλικ* στο *Ελάχιστο* (*min*) και στο *Με αλλαγή μεταβλητών κελιών* (*By Changing Variable Cells*) εισάγουμε την περιοχή B2:B4, δηλαδή την περιοχή στην οποία υπάρχουν οι αρχικές τιμές των σταθερών a, s, m.

	A	B	C	D
1				ΕΞΟΜΑΛΥΝΣΗ ΜΕ ΤΟ ΠΡΟΓΡΑΜΜΑ ΕΠΙΛΥΣΗ
2	**a =**	0,20039		
3	**s =**	10,2369		**SSR =**
4	**m =**	10,003		0,00153
5				
6	**x**	**y**	**y(ομαλ.)**	**SR**
7	2	-0,001	0,000	1,9E-06
8	3	0,002	0,002	1,1E-07
9	4	0,005	0,006	8,7E-07
10	5	0,001	0,017	0,00027
11	6	0,034	0,042	6,2E-05
12	7	0,093	0,083	9,9E-05
13	8	0,136	0,135	3,4E-07
14	9	0,189	0,182	5,4E-05
15	10	0,195	0,200	2,9E-05
16	11	0,168	0,182	0,00019
17	12	0,15	0,136	0,0002
18	13	0,086	0,083	7,1E-06
19	14	0,038	0,042	1,7E-05
20	15	0,021	0,017	1,2E-05
21	16	-0,01	0,006	0,00026
22	17	0,016	0,002	0,00021
23	18	-0,01	0,000	0,00011
24	19	-0,004	0,000	1,7E-05



Σχήμα 11.2. Τμήμα της οθόνης του υπολογιστή στο Παράδειγμα 11.2

(6) Κάνοντας *κλικ* στο κουμπί *Επίλυση* (*Solve*) το πρόγραμμα προσεγγίζει τις ζητούμενες τιμές των σταθερών που εμφανίζονται στην περιοχή B2:B4, ενώ στην περιοχή C7:C24 υπολογίζονται οι θεωρητικές τιμές του y με βάση την εξίσωση (11.2). *Επομένως στην περιοχή αυτή είναι οι ομαλοποιημένες τιμές του y*.

Στο σχήμα 11.2 δίνεται τμήμα της οθόνης του υπολογιστή, όπου παρουσιάζονται τα αποτελέσματα του παραδείγματος 11.2 καθώς επίσης και η γραφική παράσταση των πειραματικών (σημεία) και ομαλών (συ-

νεχής γραμμή) τιμών του y σε συνάρτηση με το x.

11.2 ΠΑΡΑΓΩΓΙΣΗ ΔΕΔΟΜΕΝΩΝ

Αν έχουμε μια σειρά πειραματικών δεδομένων (x_i, y_i), i = 1, 2 ,…, m και θέλουμε να υπολογίσουμε την παράγωγο dy/dx σε κάθε ένα από αυτά τα σημεία, μπορούμε να χρησιμοποιήσουμε μια από τις παρακάτω τεχνικές:

1. Προσδιορίζουμε την καμπύλη των ελαχίστων τετραγώνων y = f(x) και υπολογίζουμε την παράγωγο στο σημείο x_i από τη σχέση $y_i' = f'(x_i)$, δηλαδή παραγωγίζοντας τη συνάρτηση προσαρμογής y = f(x).
2. Αν η διαφορά x_{i+1} - x_i είναι πολύ μικρή μπορούν να χρησιμοποιηθούν για τον υπολογισμό της παραγώγου οι σχέσεις

$$f'(x_i) = \frac{f(x_{i+1}) - f(x_i)}{x_{i+1} - x_i} \quad \text{ή} \quad f'(x_i) = \frac{f(x_{i+1}) - f(x_{i-1})}{x_{i+1} - x_{i-1}} \qquad (11.3)$$

που προκύπτουν άμεσα από τον ορισμό της παραγώγου. Από τις σχέσεις αυτές καλύτερα αποτελέσματα δίνει η δεύτερη με τη <u>βασική όμως προϋπόθεση ότι οι τιμές του y έχουν ομαλοποιηθεί πριν την εφαρμογή της σχέσης αυτής</u>.

Θα πρέπει επίσης να τονιστεί ότι η πράξη της παραγώγισης, ιδιαίτερα με βάση τις σχέσεις (11.3), δεν είναι πάντα αξιόπιστη και γι αυτό θα πρέπει να γίνεται με πολύ προσοχή.

Παράδειγμα 11.3

Στον πίνακα που ακολουθεί δίνεται η μεταβολή του συντελεστή επιφανειακής τάσης γ υδατικών διαλυμάτων 2-βουτανόλης θερμοκρασίας 25 ^{0}C σε συνάρτηση με το ln x, όπου x είναι το μοριακό κλάσμα της βουτανόλης στο διάλυμα. Να προσδιοριστεί η συγκέντρωση της βουτανόλης στην επιφάνεια του διαλύματος αν γνωρίζουμε ότι η επιφανειακή συγκέντρωση Γ, η επιφανειακή τάση γ και το μοριακό κλάσμα x συνδέονται με την *εξίσωση του Gibbs*:

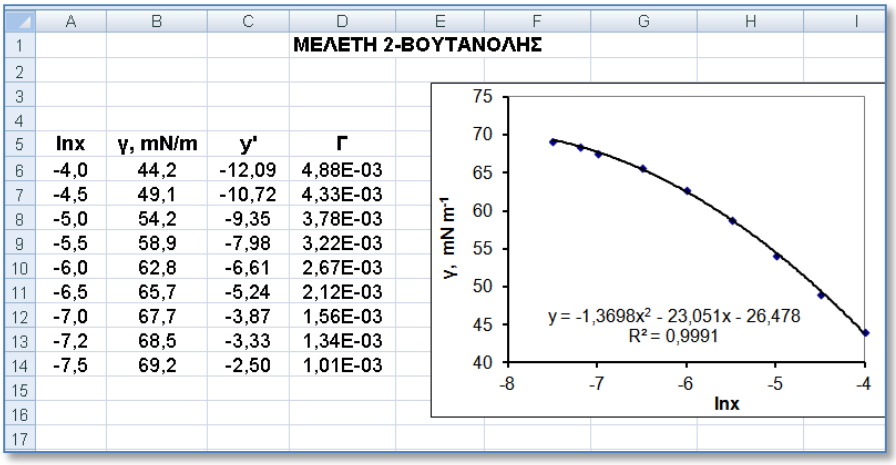
$$\Gamma = -\frac{1}{RT}\left(\frac{\partial \gamma}{\partial \ln x}\right)_T \qquad (11.4)$$

Πίνακας 11.3. Μεταβολή του συντελεστή επιφανειακής τάσης γ (σε mN/m) υδατικών διαλυμάτων 2-βουτανόλης σε συνάρτηση με το νεπέριο λογάριθμο του μοριακού της κλάσματος στο διάλυμα.

- lnx	4.0	4.5	5.0	5.5	6.0	6.5	7.0	7.2	7.5
γ	44.2	49.1	54.2	58.9	62.8	65.7	67.7	68.5	69.2

◆ Μεταφέρουμε τα δεδομένα του παραπάνω πίνακα σε ένα φύλλο εργασίας και κάνουμε τη γραφική παράσταση της μεταβολής του γ με το lnx (σχήμα 11.3). Ακολούθως προσδιορίζουμε την καμπύλη ελαχίστων τετραγώνων δευτέρου βαθμού που περνά μέσα από τα πειραματικά σημεία χρησιμοποιώντας την εντολή *Προσθήκη γραμμής τάσης* (*Add Trendline*) ή το πρόγραμμα *LS Polynomial* του *ChemStat* ή ακόμη καλύτερα το *LS Optimum Polynomial* του *ChemStat*. Η εξίσωση των ελαχίστων τετραγώνων που παίρνουμε είναι η

$$y = -1.3698x^2 - 23.051x - 26.478$$



Σχήμα 11.3. Τμήμα της οθόνης του υπολογιστή στο Παράδειγμα 11.3

Για την πρώτη παράγωγο έχουμε y′ = -2.7396x – 23.051. Συνεπώς στο κελί C6 πληκτρολογούμε τον τύπο =-2.7396*A6-23.051, πατάμε *Enter* και συμπληρώνουμε την περιοχή C7:C14 με τη διαδικασία της αυτόματης συμπλήρωσης. Για τον υπολογισμό του Γ σε mmol/m^2 πληκτρολογούμε στο

D6 τον τύπο =-C6/(8.314*298), πατάμε *Enter* και συμπληρώνουμε την περιοχή D7:D14 επίσης με τη διαδικασία της αυτόματης συμπλήρωσης. Επειδή η καμπύλη των ελαχίστων τετραγώνων y = $-1.3698x^2$ - 23.051x – 26.478 περιγράφει ικανοποιητικά τα πειραματικά δεδομένα αναμένεται τα αποτελέσματα της παραγώγισης να είναι αξιόπιστα (σχήμα 11.3).

Επειδή η παραγώγιση δεν είναι πάντα αξιόπιστη πράξη, μια τεχνική που συνήθως ακολουθείται είναι η εξής. Προσδιορίζουμε την καμπύλη ελαχίστων τετραγώνων δευτέρου βαθμού στα 5 πρώτα σημεία (x_1, x_2, x_3, x_4, x_5) ή στα 7 πρώτα σημεία (x_1, x_2, x_3, x_4, x_5, x_6, x_7) και με βάση την εξίσωσή της υπολογίζουμε τις παραγώγους στα σημεία αυτά. Και στις δύο περιπτώσεις κρατάμε τις παραγώγους των τριών κεντρικών σημείων και απορρίπτουμε τις παραγώγους στα ακραία σημεία, επειδή αυτές υπόκεινται κατά κανόνα σε μεγαλύτερο σφάλμα. Δηλαδή υπολογίζουμε τις παραγώγους στα σημεία x_2, x_3, x_4 αν δουλεύουμε με 5 σημεία ή στα x_3, x_4, x_5 αν δουλεύουμε με 7 σημεία. Ακολούθως προσδιορίζουμε την καμπύλη ελαχίστων τετραγώνων δευτέρου βαθμού στα σημεία (x_4, x_5, x_6, x_7, x_8) ή στα (x_4, x_5, x_6, x_7, x_8, x_9, x_{10}) και υπολογίζουμε τις παραγώγους στα σημεία x_5, x_6, x_7 ή στα x_6, x_7, x_8, αντίστοιχα. Με τον τρόπο αυτό συνεχίζουμε μέχρι να καλύψουμε όλα τα σημεία.

11.3 ΟΛΟΚΛΗΡΩΣΗ ΔΕΔΟΜΕΝΩΝ

Η ολοκλήρωση μιας σειράς πειραματικών δεδομένων (x_i, y_i), δηλαδή ο υπολογισμός του εμβαδού μεταξύ του άξονα των x και της καμπύλης των πειραματικών σημείων, μπορεί να γίνει με μια από τις παρακάτω μεθόδους:

1. Ολοκληρώνοντας την καμπύλη των ελαχίστων τετραγώνων είτε αναλυτικά είτε με μεθόδους αριθμητικής ανάλυσης.
2. Χρησιμοποιώντας τον ***κανόνα του τραπεζίου*** *(trapezoidal rule).* Σύμφωνα με αυτόν το εμβαδόν I_i μεταξύ των σημείων x_i και x_{i+1} προσεγγίζεται από τη σχέση

$$I_i = (y_{i+1} + y_i) * (x_{i+1} - x_i)/2 \qquad (11.5)$$

Αν τα πειραματικά δεδομένα έχουν αρκετό θόρυβο καλό είναι να ομαλοποιούνται πρώτα και μετά να εφαρμόζεται ο κανόνας του *τραπεζίου*, αν και η ολοκλήρωση δεν είναι ιδιαίτερα ευαίσθητη στον πειραματικό θόρυβο.

Παράδειγμα 11.4

Να ολοκληρωθεί η καμπύλη του παραδείγματος 11.2.

◆ Θα λύσουμε το παράδειγμα αυτό και με τις δύο παραπάνω μεθόδους. Στο παράδειγμα 11.2 είχαμε αποδείξει, χρησιμοποιώντας το πρόγραμμα *Επίλυση* του *Excel*, ότι η καμπύλη των ελαχίστων τετραγώνων μπορεί να είναι η εξίσωση (11.2) με a = 0.20039, m = 10.003 και s = 10.2369. Συνεπώς το ζητούμενο ολοκλήρωμα είναι

$$I = \int_{2}^{19} ae^{-(x-m)^2/s}dx = a\int_{-\infty}^{+\infty} e^{-(x-m)^2/s}dx = a\sqrt{s\pi} = 1.13640$$

Θα εφαρμόσουμε τώρα τον κανόνα του τραπεζίου τόσο στα αρχικά δεδομένα, y, όσο και στα ομαλοποιημένα, y(ομαλ.). Για το σκοπό αυτό ανοίγουμε το φύλλο εργασίας του παραδείγματος 11.2, στα κελιά E6 και F6 εισάγουμε τους τίτλους I(ομαλ.) και I(αρχ.) και στα κελιά E7 και F7 τον αριθμό 0 και στα δύο. Ακολούθως στο κελί E8 πληκτρολογούμε τη σχέση =(C8+C7)*(A8-A7)/2+E7, ενώ στο κελί F8 τη σχέση =(B8+B7)*(A8-A7)/2+F7. Οι σχέσεις αυτές υπολογίζουν με τον κανόνα του τραπεζίου το εμβαδόν μεταξύ των σημείων x = 2 και x = 3 και προσθέτουν το εμβαδόν που υπάρχει πριν από τα σημεία αυτά, που στην προκειμένη περίπτωση είναι μηδέν. Είναι προφανές ότι η πρώτη σχέση υπολογίζει το ολοκλήρωμα χρησιμοποιώντας ομαλές τιμές, ενώ η δεύτερη τα αρχικά δεδομένα. Με τη διαδικασία της αυτόματης συμπλήρωσης συμπληρώνουμε όλη την περιοχή E9:F24.

Στο σχήμα 11.4 παρατηρούμε ότι η τελική τιμή του ολοκληρώματος είναι 1.1361 αν χρησιμοποιήσουμε τις ομαλοποιημένες τιμές του y και 1.1115 αν χρησιμοποιήσουμε τις αρχικές τιμές. Η διαφορά μεταξύ των δύο αυτών τιμών είναι 2.16%, δηλαδή αρκετά μικρή αν λάβουμε υπόψη τον μεγάλο πειραματικό θόρυβο. Θα πρέπει επίσης να σημειώσουμε ότι η τιμή π.χ. 0.7584, που αντιστοιχεί σε x = 11, δίνει την τιμή του ορισμένου ολοκληρώματος από x = 2 μέχρι x = 11.

	A	B	C	D	E	F
1		**ΟΛΟΚΛΗΡΩΣΗ ΚΑΜΠΥΛΗΣ**				
2	**a =**	0,20039				
3	**s =**	10,2369		**SSR =**	**I(curve)=**	
4	**m =**	10,003		0,00153	1,13640	
5						
6	**x**	**y**	**y(ομαλ.)**	**SR**	**I(ομαλ.)**	**I(αρχ.)**
7	2	-0,001	0,000	1,9E-06	0	0
8	3	0,002	0,002	1,1E-07	0,0010	0,0005
9	4	0,005	0,006	8,7E-07	0,0048	0,0040
10	5	0,001	0,017	0,00027	0,0165	0,0070
11	6	0,034	0,042	6,2E-05	0,0461	0,0245
12	7	0,093	0,083	9,9E-05	0,1086	0,0880
13	8	0,136	0,135	3,4E-07	0,2178	0,2025
14	9	0,189	0,182	5,4E-05	0,3763	0,3650
15	10	0,195	0,200	2,9E-05	0,5673	0,5570
16	11	0,168	0,182	0,00019	0,7584	0,7385
17	12	0,15	0,136	0,0002	0,9172	0,8975
18	13	0,086	0,083	7,1E-06	1,0268	1,0155
19	14	0,038	0,042	1,7E-05	1,0895	1,0775
20	15	0,021	0,017	1,2E-05	1,1193	1,1070
21	16	-0,01	0,006	0,00026	1,1310	1,1125
22	17	0,016	0,002	0,00021	1,1348	1,1155
23	18	-0,01	0,000	0,00011	1,1358	1,1185
24	19	-0,004	0,000	1,7E-05	1,1361	1,1115

Σχήμα 11.4. Αποτελέσματα ολοκλήρωσης πειραματικών δεδομένων

ΑΣΚΗΣΕΙΣ

11.1. Τα πειραματικά σημεία του παρακάτω πίνακα περιγράφονται από την γενική εξίσωση $y = a/[a^2+(x-b)^2]$. Να εξομαλυνθούν, να παραγωγιστούν και να ολοκληρωθούν χρησιμοποιώντας όλες τις δυνατές τεχνικές.

x	y	x	y	x	y	x	y
0.5	-0.01	5.5	0.01	10.5	0.80	15.5	0.00
1.0	0.01	6.0	0.02	11.0	0.60	16.0	0.02
1.5	0.03	6.5	0.06	11.5	0.30	16.5	0.03
2.0	0.03	7.0	0.08	12.0	0.25	17.0	0.02
2.5	0.05	7.5	0.10	12.5	0.10	17.5	0.01
3.0	0.00	8.0	0.20	13.0	0.20	18.0	0.00
3.5	0.02	8.5	0.30	13.5	0.10	18.5	0.01
4.0	0.05	9.0	0.55	14.0	0.10	19.0	0.00
4.5	0.01	9.5	0.90	14.5	0.03	19.5	0.00
5.0	0.02	10.0	1.00	15.0	0.07	20.0	0.01

11.2. Η μεταβολή της θερμοχωρητικότητας υπό σταθερή πίεση 1 atm, C_p, ενός mole υδρογόνου με τη θερμοκρασία δίνεται στον επόμενο πίνακα. Λαμβάνοντας υπόψη ότι η μεταβολή της εντροπίας ΔS για μεταβολή της θερμοκρασίας από T_1 σε T_2 υπό σταθερή πίεση δίνεται από τη σχέση

$$\Delta S = \int_{T_1}^{T_2} (C_p / T)\, dT$$

να υπολογιστεί η ΔS για μεταβολή της θερμοκρασίας από 300 K σε 500 K.

T, K	C_p, J/mol K	T, K	C_p, J/mol K	T, K	C_p, J/mol K
300	28.30	370	28.60	440	28.85
310	28.35	380	28.60	450	28.90
320	28.40	390	28.65	460	28.90
330	28.45	400	28.70	470	28.95
340	28.45	410	28.70	480	29.00
350	28.50	420	28.80	490	29.00
360	28.55	430	28.80	500	29.10

11.3. Στον παρακάτω πίνακα δίνεται η μεταβολή του συντελεστή διεπιφανειακής τάσης γ (σε mN/m) υδατικών διαλυμάτων ακετονιτριλίου σε επαφή με δωδεκάνιο σε συνάρτηση με το μοριακό κλάσμα x του ακετονιτριλίου στο διάλυμα. Τα δεδομένα αντιστοιχούν σε θερμοκρασία 25 ^{0}C. Να προσδιοριστεί η συγκέντρωση του ακετονιτριλίου στη διεπιφάνεια διαλύματος – δωδεκανίου με βάση τη σχέση (11.4).

x	γ, mN/m	x	γ, mN/m
0.0017	43.1	0.0769	20.3
0.0033	41.2	0.1237	14.4
0.0085	39.5	0.1786	11.2
0.0173	33.6	0.2443	10
0.0361	28.9		

11.4. Στον παρακάτω πίνακα δίνεται η μεταβολή της χωρητικότητας C του ηλεκτροδίου του Hg με το δυναμικό E σε ηλεκτρολυτικά διαλύματα φορμαμιδίου παρουσία 5 διαφορετικών συγκεντρώσεων τριφαινυλο-φωσφινοξειδίου (TPO). Να υπολογιστεί το ηλεκτροδιακό φορτίο q σε συνάρτηση με το δυναμικό E σε κάθε συγκέντρωση TPO λαμβάνοντας υπόψη ότι C = dq/dE και ότι στο δυναμικό E = -1.8 V οι τιμές του φορτίου είναι ανεξάρτητες της συγκέντρωσης του TPO. Ακολούθως να γίνουν οι γραφικές παραστάσεις C – E και q – E.

E, V	C, F/m^2					E, V	C, F/m^2				
	c1	c2	c3	c4	c5		c1	c2	c3	c4	c5
-0.22	1.80	1.84	1.84	1.87	1.88	-1.03	1.65	1.36	0.98	0.75	0.57
-0.26	1.74	1.76	1.77	1.80	1.80	-1.08	1.64	1.51	1.24	0.91	0.63
-0.31	1.63	1.66	1.65	1.70	1.72	-1.13	1.62	1.59	1.51	1.25	0.72
-0.36	1.55	1.57	1.58	1.62	1.70	-1.18	1.59	1.61	1.64	1.79	0.91
-0.41	1.50	1.52	1.51	1.57	1.84	-1.22	1.56	1.60	1.65	2.12	1.31
-0.46	1.45	1.47	1.47	1.54	3.48	-1.27	1.51	1.59	1.62	2.00	2.27
-0.50	1.41	1.44	1.45	1.55	1.82	-1.32	1.47	1.55	1.58	1.79	3.15
-0.55	1.40	1.43	1.43	1.73	1.01	-1.37	1.43	1.50	1.54	1.67	2.46
-0.60	1.39	1.44	1.45	1.67	0.75	-1.42	1.40	1.46	1.49	1.58	1.92
-0.65	1.40	1.47	1.51	0.84	0.63	-1.46	1.39	1.42	1.44	1.52	1.67
-0.70	1.43	1.52	1.62	0.68	0.56	-1.51	1.37	1.40	1.43	1.48	1.55
-0.74	1.46	1.55	1.06	0.62	0.52	-1.56	1.36	1.39	1.40	1.45	1.48
-0.79	1.51	1.27	0.72	0.59	0.50	-1.61	1.36	1.38	1.39	1.39	1.45
-0.84	1.55	0.90	0.66	0.58	0.50	-1.66	1.36	1.38	1.39	1.43	1.43
-0.89	1.59	0.86	0.66	0.60	0.51	-1.70	1.37	1.38	1.38	1.42	1.41
-0.94	1.62	0.94	0.71	0.62	0.52	-1.75	1.37	1.37	1.37	1.40	1.40
-0.98	1.65	1.12	0.80	0.67	0.54	-1.80	1.37	1.37	1.37	1.37	1.37

Κεφάλαιο 12

ΑΝΑΛΥΣΗ ΠΟΛΛΩΝ ΜΕΤΑΒΛΗΤΩΝ

12.1 ΓΕΝΙΚΑ

Η ανάπτυξη των υπολογιστών έχει ως αποτέλεσμα να σωρεύουμε πληθώρα δεδομένων, που στις περισσότερες περιπτώσεις μπορούν να εκφραστούν με τη μορφή του παρακάτω γενικού πίνακα.

Πίνακας 12.1. Πίνακας δεδομένων πολλών μεταβλητών

	Ιδιότητα 1	Ιδιότητα 2	**. . . . .**	Ιδιότητα m
Δείγμα 1	a_{11}	a_{12}	. . .	a_{1m}
Δείγμα 2	a_{21}	a_{22}	. . .	a_{2m}
. . . .	. . .	. . .	. . .	. . .
Δείγμα n	a_{n1}	a_{n2}	. . .	a_{nm}

Το κύριο ερώτημα που γεννιέται σε αυτές τις περιπτώσεις είναι αν υπάρχουν σχέσεις μεταξύ των δειγμάτων ή καλύτερα *ομάδες (clusters)* δειγμάτων με παρόμοιες ιδιότητες και αν υπάρχουν ποιες είναι αυτές. Στο ερώτημα αυτό απάντηση προσπαθεί να δώσει η ***Ανάλυση Πολλών Μεταβλητών*** *(Multivariate Analysis)*.

Οι τεχνικές που έχουν αναπτυχθεί για να προσδιοριστεί αν υπάρχουν ομάδες δειγμάτων σε ένα πίνακα δεδομένων πολλών μεταβλητών είναι πολλές. Εδώ θα εξετασθούν οι μέθοδοι: α) *Ανάλυση σε Κύριες Συνιστώσες (Principal Component Analysis – PCA)*, β) *Ανάλυση σε Ομάδες (Cluster Analysis – CA)*, γ) *Γραμμική Διαχωριστική Ανάλυση (Linear Discriminant Analysis – LDA)*, και δ) *Ανάλυση Διασποράς Πολλών Μεταβλητών (Multivariate Analysis of Variance – MANOVA)*.

Για την εφαρμογή των μεθόδων *PCA* και *CA* δεν απαιτείται καμία παραδοχή σχετικά με τη μορφή των πληθυσμιακών κατανομών των

δεδομένων. Αντίθετα η εφαρμογή των μεθόδων *LDA* και *MANOVA* προϋποθέτει, ανάμεσα στα άλλα, τα δεδομένα να ακολουθούν την ***πολυμεταβλητή κανονική κατανομή*** *(multivariate normal distribution)*. Δηλαδή την κανονική κατανομή που έχει όχι μία ανεξάρτητη μεταβλητή, όπως η σχέση (3.6), αλλά δύο, όπως η σχέση (10.26) ή και περισσότερες.

Κριτήρια για την *πολυμεταβλητή κανονικότητα* των δειγμάτων, αν και υπάρχουν, δεν είναι διαθέσιμα στο *ChemStat* και στο *SPSS*, με εξαίρεση το *διάγραμμα chi-square* που υπάρχει στο *ChemStat* ως κριτήριο της *διμεταβλητής κανονικότητας*. Ένας ικανοποιητικός έλεγχος της *πολυμεταβλητής κανονικότητας* είναι να ελέγξουμε την *κανονικότητα* όλων των δειγμάτων και να κάνουμε ελέγχους *διμεταβλητής κανονικότητας* σε ζεύγη δειγμάτων. Πάντως οι μέθοδοι *LDA* και *MANOVA* εφαρμόζονται και όταν υπάρχουν αποκλίσεις από την κανονικότητα αρκεί να μην υπάρχουν πολύ ακραίες τιμές.

12.2 ΑΝΑΛΥΣΗ ΣΕ ΚΥΡΙΕΣ ΣΥΝΙΣΤΩΣΕΣ (PCA)

12.2.1 ΠΕΡΙΓΡΑΦΙΚΗ ΕΙΣΑΓΩΓΗ

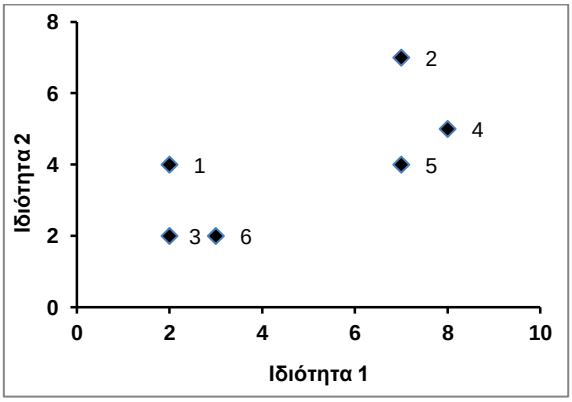
Για να μπορέσουμε να εξετάσουμε αν σε έναν πίνακα δεδομένων υπάρχουν ομάδες ομοειδών δεδομένων (clusters) θα πρέπει να ελαττώσουμε τις *διαστάσεις του πίνακα*, δηλαδή να ελαττώσουμε τις στήλες του σε δύο ή τρεις. Τότε μπορούμε εύκολα με το κατάλληλο γράφημα να δούμε την ύπαρξη ή μη ομάδων.

Πίνακας 12.2. Πίνακας τριών διαστάσεων.

	Ιδιότητα 1	Ιδιότητα 2	Ιδιότητα 3
Δείγμα 1	2	4	3
Δείγμα 2	7	7	5
Δείγμα 3	2	2	3
Δείγμα 4	8	5	7
Δείγμα 5	7	4	6
Δείγμα 6	3	2	2

Για παράδειγμα, έστω ότι έχουμε τον πίνακα 12.2, που είναι ένας πίνακας τριών διαστάσεων. Αν περιορίσουμε τη μελέτη στις δύο πρώτες μεταβλητές, ιδιότητα 1 και ιδιότητα 2, είναι εύκολο να διαπιστώσουμε την ύπαρξη ομάδων αρκεί να κάνουμε τη γραφική παράσταση *Ιδιότητα 2* ως προς *Ιδιότητα 1*. Από το σχήμα 12.1 βλέπουμε ότι υπάρχουν δύο ομάδες

δεδομένων, όπου η μία περιλαμβάνει τα δείγματα 1, 3 και 6 και η άλλη τα δείγματα 2, 4 και 5.

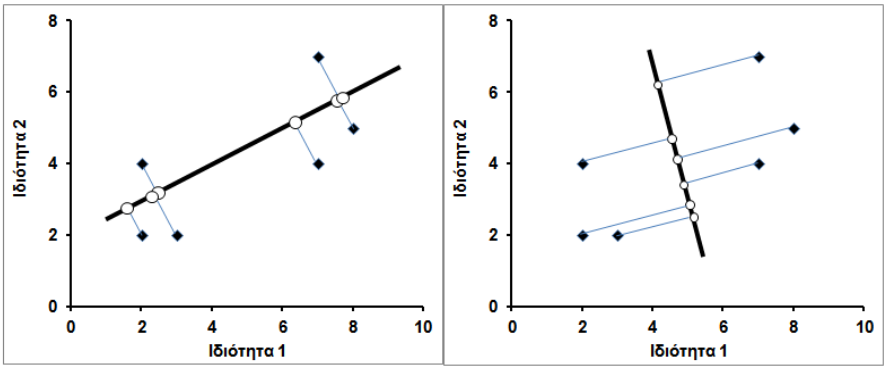


Σχήμα 12.1. Γραφική παράσταση τιμών του πίνακα 12.2

Ας υποθέσουμε τώρα ότι για μια "παραξενιά της φύσης" δεν μπορούμε να δούμε σε ένα δισδιάστατο γράφημα παρά μόνο σε μονοδιάστατα γραφήματα. Τότε για να μπορέσουμε να εξετάσουμε αν υπάρχουν ομάδες σημείων στο παραπάνω παράδειγμα πρέπει να εργαστούμε ως εξής: Φέρνουμε έναν άξονα (μια ευθεία) μέσα από τα σημεία της γραφικής παράστασης του σχήματος 12.1 και προβάλουμε τα σημεία αυτά πάνω στον άξονα. Ανάλογα με το πώς φέρνουμε τον άξονα μπορούμε να πάρουμε τις δύο ακραίες περιπτώσεις του σχήματος 12.2.

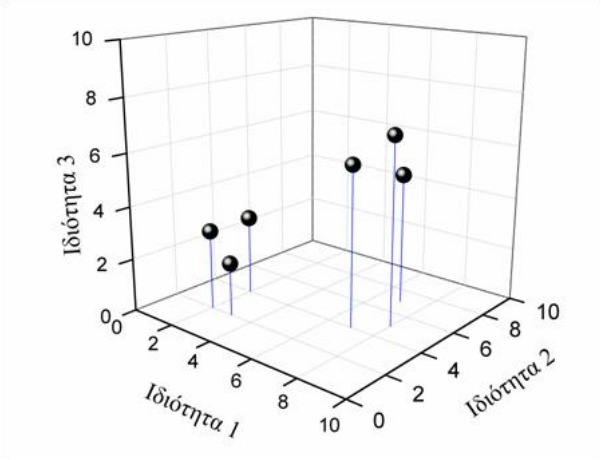
Παρατηρούμε ότι αν φέρουμε τον άξονα κατά μήκος της μεγαλύτερης διασποράς των σημείων, τότε η προβολή των σημείων στον άξονα διατηρεί την βασική τους ιδιότητα να ανήκουν σε διαφορετικές και διακριτές ομάδες (σχήμα 12.2-αριστερά). Αντίθετα, αν ο άξονας είναι κατά μήκος της μικρότερης διασποράς των σημείων, τότε τα σημεία προβαλλόμενα στον άξονα δεν ξεχωρίζουν σε ομάδες (σχήμα 12.2-δεξιά).

Επομένως αν θέλουμε να μετατρέψουμε έναν δισδιάστατο πίνακα μετρήσεων σε μονοδιάστατο, θα πρέπει να φέρουμε έναν άξονα κατά μήκος της μεγαλύτερης διασποράς των σημείων και να προβάλουμε τα σημεία στον άξονα. Ο άξονας αυτός ονομάζεται *PC1* ή *πρώτη κύρια συνιστώσα*.



Σχήμα 12.2. Προβολή των σημείων του σχήματος 12.1 σε δύο διαφορετικούς άξονες

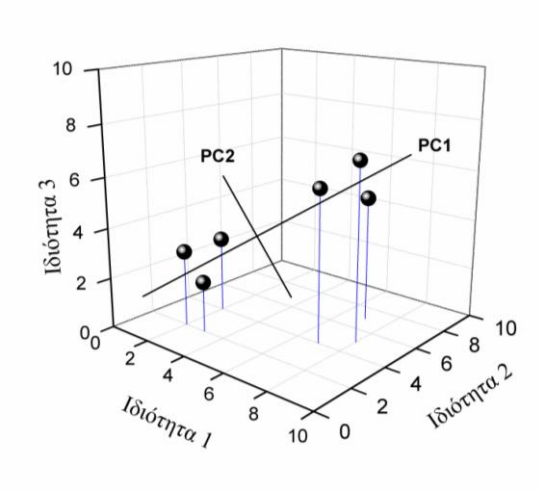
Θα εξετάσουμε τώρα και τις τρεις μεταβλητές του πίνακα 12.2. Η τρισδιάστατη γραφική παράσταση των τιμών του πίνακα αυτού δίνεται στο σχήμα 12.3, όπου παρατηρούμε ότι και εδώ υπάρχουν δύο ομάδες σημείων.



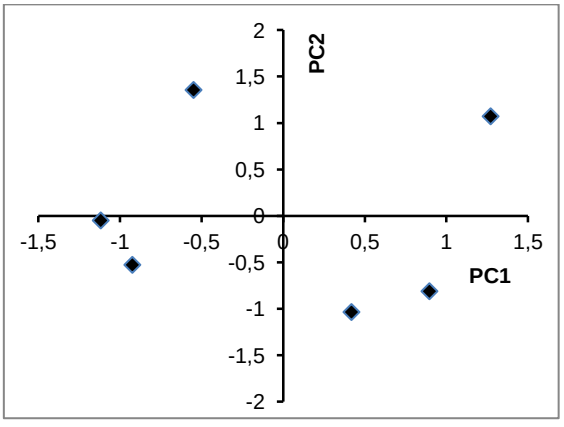
Σχήμα 12.3. Γραφική παράσταση των τιμών του πίνακα 12.2

Αν τώρα θέλουμε να ελαττώσουμε τους άξονες σε δύο, εργαζόμαστε όπως και προηγουμένως. Φέρνουμε έναν άξονα, τον PC1 στο σχήμα 12.4, μέσα από τα σημεία της γραφικής παράστασης του σχήματος και κατά μήκος της μεγαλύτερης διασποράς των σημείων. Ακολούθως

φέρνουμε ένα δεύτερο άξονα, τον PC2, που είναι κάθετος στον PC1 και τον περιστρέφουμε κάθετα προς τον PC1, έτσι ώστε και αυτός να είναι κατά μήκος της μεγαλύτερης διασποράς των σημείων ως προς τη διεύθυνσή του. Οι δύο αυτοί άξονες ορίζουν ένα επίπεδο στο οποίο προβάλουμε όλα τα σημεία του σχήματος 12.4. Η εικόνα που παίρνουμε δίνεται στο σχήμα 12.5 και ονομάζεται *διάγραμμα αποτελεσμάτων (Score plot)*. Παρατηρούμε ότι και στο διάγραμμα αυτό υπάρχουν οι δύο ομάδες σημείων του αρχικού σχήματος 12.3.



Σχήμα 12.4. Άξονες PC1, PC2



Σχήμα 12.5. Διάγραμμα *Αποτελεσμάτων (score plot)*

Η αναγωγή ενός δισδιάστατου πίνακα σε μονοδιάστατο ή ενός τρισδιάστατου σε δισδιάστατο δεν έχει καμία πρακτική αξία. Ενδιαφέρον παρουσιάζει μόνο η αναγωγή πινάκων με διαστάσεις μεγαλύτερες από τρία σε πίνακες με δύο ή τρεις διαστάσεις. Αυτό γίνεται με τον ίδιο τρόπο που γίνεται η αναγωγή ενός τρισδιάστατου πίνακα σε δισδιάστατο, η βάση όμως της αναγωγής αυτής είναι η θεωρία των πινάκων, όπως περιγράφεται συνοπτικά παρακάτω.

12.2.2 ΟΙ ΜΑΘΗΜΑΤΙΚΕΣ ΒΑΣΕΙΣ ΤΗΣ PCA

Έστω $\mathbf{X}_{n\times p} = [x_{ij}]$, i = 1, 2, …, n και j = 1, 2, …, p, ένας πίνακας n×p τιμών, δηλαδή n γραμμών και p στηλών. Η μαθηματική πορεία που ακολουθείται για να ελαττωθούν οι διαστάσεις του πίνακα περιλαμβάνει α) την προκατεργασία του πίνακα, β) την διάσπασή του με βάση το θεώρημα ***svd*** *(singular value decomposition)* και γ) την περιστροφή των αξόνων.

1) Προκατεργασία του πίνακα X

Συνήθως πριν από την εφαρμογή της *PCA* ο αρχικός πίνακας **X** μετασχηματίζεται σε έναν νέο πίνακα $\mathbf{Z} = [z_{ij}]$, που συνήθως είναι ο *πίνακας συσχέτισης (correlation matrix)* του **X**. Δηλαδή ο πίνακας

$$\mathbf{Z} = \begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ r_{21} & r_{22} & \cdots & r_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ r_{n1} & r_{n2} & \cdots & r_{nn} \end{pmatrix}$$

όπου r_{ij} είναι ο συντελεστής συσχέτισης μεταξύ των τιμών του δείγματος i και του δείγματος j του πίνακα 12.1. Ο συντελεστής αυτός υπολογίζεται από τη σχέση (10.24). Ισχύει $r_{11} = r_{22} = \cdots = r_{ii} = 1$ και $r_{ij} = r_{ji}$. Ο μετασχηματισμός του πίνακα **X** στον πίνακα **Z** είναι ιδιαίτερα χρήσιμος όταν τα δεδομένα του πίνακα **X** έχουν διαφορετικές κλίμακες.

2) Το θεώρημα svd

Η μέθοδος *PCA* στηρίζεται στο θεώρημα *svd (singular value decomposition)*, σύμφωνα με το οποίο κάθε πίνακας **Z** διασπάται στους εξής τρεις πίνακες:

$$\mathbf{Z}_{n\times p} = \mathbf{U}_{n\times p}\mathbf{W}_{p\times p}\mathbf{V'}_{p\times p} = \mathbf{S}_{n\times p}\mathbf{V'}_{p\times p} \qquad (12.1)$$

όπου το σύμβολο **′** δηλώνει τον *ανάστροφο* πίνακα, δηλαδή τον πίνακα που προκύπτει όταν οι γραμμές του αρχικού γίνουν στήλες και οι στήλες γραμμές. Ο πίνακας $\mathbf{W}_{p\times p}$ είναι ένας διαγώνιος πίνακας της μορφής

$$\mathbf{W} = \begin{pmatrix} w_{11} & 0 & \cdots & 0 \\ 0 & w_{22} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & w_{pp} \end{pmatrix} \tag{12.2}$$

Στον πίνακα αυτόν οι τιμές w_{11}, w_{22}, …, w_{pp} διατάσσονται με φθίνουσα σειρά και ο w_{11} είναι πάντα θετικός αριθμός. Ο πίνακας **V** με διαστάσεις p×p ονομάζεται ***πίνακας φορτώσεων*** *(loading matrix)*, ενώ ο **S** με διαστάσεις n×p ονομάζεται ***πίνακας αποτελεσμάτων*** *(score matrix)*. Συνήθως οι στήλες του **S** διαιρούνται δια της τυπικής τους απόκλισης s_j, ενώ ταυτόχρονα οι στήλες του **V** πολλαπλασιάζονται επί s_j, έτσι ώστε να μη μεταβληθεί το γινόμενο **SV′**.

Οι πίνακες **U**, **V** είναι ορθογώνιοι, δηλαδή ισχύουν οι σχέσεις:

$$\mathbf{U'U} = \mathbf{V'V} = \mathbf{I} \tag{12.3}$$

όπου **I** είναι ο *μοναδιαίος* πίνακας. Λόγω αυτής της ιδιότητας έχουμε

$$\mathbf{Z'Z} = \mathbf{V\,W\,U'\,U\,W\,V'} = \mathbf{V\,W^2\,V'} \;\Rightarrow\; \mathbf{W^2} = \mathbf{V'(Z'Z)V} \tag{12.4}$$

Επίσης για τον **S** έχουμε τη σχέση

$$\mathbf{S = ZV} \tag{12.5}$$

3) Ελάττωση των διαστάσεων

Αν οι q τελευταίες διαγώνιες τιμές του πίνακα **W** είναι πολύ μικρές, μπορούμε να κρατήσουμε στους πίνακες **U**, **W**, **V** μόνο τις k = p - q πρώτες στήλες. Τότε έχουμε

$$\mathbf{Z}_{n\times p} = \mathbf{U^*}_{n\times k}\mathbf{W^*}_{k\times k}\mathbf{V^{*\prime}}_{k\times p} + \mathbf{E}_{n\times p}$$

$$\mathbf{Z}_{n\times p} = \mathbf{S^*}_{n\times k}\mathbf{V^{*\prime}}_{k\times p} + \mathbf{E}_{n\times p} \tag{12.6}$$

όπου ο ***πίνακας σφάλματος*** *(error matrix)* $\mathbf{E}_{n\times p}$ μας δίνει το σφάλμα που προκύπτει όταν αντικαθιστούμε τον αρχικό πίνακα **Z** από το γινόμενο $\mathbf{S^*V^{*\prime}}$. Συνεπώς

$$\mathbf{Z} \approx \mathbf{S}^*\mathbf{V}^{*\prime} \tag{12.7}$$

Ο πίνακας $\mathbf{V}^*$ με διαστάσεις p×k ονομάζεται επίσης *πίνακας φορτώσεων*, ενώ ο $\mathbf{S}^* = \mathbf{U}^*\mathbf{W}^*$ με διαστάσεις n×k ονομάζεται και αυτός *πίνακας αποτελεσμάτων*. Ο πίνακας $\mathbf{S}^*$ έχει τις ίδιες γραμμές με τον αρχικό πίνακα **Z** αλλά λιγότερες στήλες, k.

Αποδεικνύεται ότι στον πίνακα $\mathbf{S}^*$ η πρώτη στήλη αντιστοιχεί στον άξονα PC1, η δεύτερη στον PC2, κ.ο.κ. Επίσης επειδή η πρώτη στήλη αντιστοιχεί στη μεγαλύτερη τιμή w_{11} ο άξονας PC1 αντιστοιχεί στη μεγαλύτερη δυνατή διασπορά των σημείων, ο PC2 αντιστοιχεί στη δεύτερη μεγαλύτερη διασπορά των σημείων και ακολουθούν οι υπόλοιποι άξονες.

Λόγω αυτών των ιδιοτήτων ο πίνακας $\mathbf{S}^*$ διατηρεί σε μεγάλο βαθμό όλες τις πληροφορίες του αρχικού πίνακα **X** ή **Z** σχετικά με την ύπαρξη ή μη ομάδων. Για το λόγο αυτό ο πίνακας $\mathbf{S}^*$ μπορεί να χρησιμοποιηθεί σε άλλες στατιστικές αναλύσεις αντί για τους αρχικούς πίνακες **X** ή **Z**. Μάλιστα αυτή η αντικατάσταση είναι σε ορισμένες περιπτώσεις αναγκαία. Για παράδειγμα, αν ο πίνακας **X** έχει περισσότερες στήλες από γραμμές, δεν μπορούν να εφαρμοστούν η *MANOVA* και η *Γραμμική Διαχωριστική Ανάλυση (LDA)*. Σε αυτή την περίπτωση εφαρμόζουμε *PCA* ώστε να ελαττώσουμε τις διαστάσεις του **X** και εφαρμόζουμε *MANOVA* ή *LDA* στον πίνακα $\mathbf{S}^*$, στον οποίο φροντίζουμε ο αριθμός των στηλών να είναι μικρότερος των γραμμών.

Παρατήρηση. Τα τετράγωνα των τιμών w_{jj} ονομάζονται ***ιδιοτιμές*** *(eigenvalues)* και συνήθως οι q τελευταίες διαγώνιες τιμές του πίνακα **W** θεωρούνται μικρές όταν $(w_{jj})^2 < 1$, δηλαδή όταν οι ιδιοτιμές είναι μικρότερες από τη μονάδα.

4) Μια διαφορετική εφαρμογή της PCA - Γενικό παράδειγμα

Η *PCA* είναι μια αρκετά ισχυρή μέθοδος για τη διευκρίνιση της ύπαρξης *ομάδων (clusters)* σε ένα πίνακα δεδομένων. Εκτός όμως από αυτή την εφαρμογή, η *PCA* χρησιμοποιείται για να διαπιστώσουμε αν υπάρχουν παράγοντες που επιδρούν στα δεδομένα ενός πίνακα. Εδώ περιγράφουμε γενικά αυτή τη δυνατότητα, ενώ συγκεκριμένη εφαρμογή δίνεται στο παράδειγμα 12.2.

Ένας πίνακας με περιβαλλοντικά δεδομένα έχει τη γενική μορφή του πίνακα 12.3. Αν εφαρμόσουμε την *PCA* και κρατήσουμε τις k πρώτες στήλες, συνήθως αυτές που αντιστοιχούν σε ιδιοτιμές μεγαλύτερες της μονάδας, θα πάρουμε τον *πίνακα φορτώσεων* (πίνακας 12.4) και τον

πίνακα αποτελεσμάτων (πίνακας 12.5). Συνηθίζεται οι άξονες PC1, PC2, …, PCk, που ονομάζονται και *παράγοντες*, να θεωρούνται ότι αντιστοιχούν σε πηγές ρύπανσης και στόχος είναι να προσδιοριστούν οι πιθανές πραγματικές πηγές ρύπων που αντιστοιχούν στους παραπάνω άξονες. Αυτό γίνεται κυρίως από τον *πίνακα φορτώσεων* εξετάζοντας ποιοι ρύποι κυριαρχούν σε κάθε άξονα PC1, PC2, …, PCk. Έτσι από τους ρύπους που κυριαρχούν συμπεραίνουμε τις πηγές που προκαλούν τους ρύπους.

Πίνακας 12.3. Πίνακας περιβαλλοντικών δεδομένων.

	Ρύπος 1	Ρύπος 2	. . .	Ρύπος p
Δείγμα 1	a_{11}	a_{12}	. . .	a_{1p}
Δείγμα 2	a_{21}	a_{22}	. . .	a_{2p}
. . . .	. . .	. . .	. . .	. . .
Δείγμα n	a_{n1}	a_{n2}	. . .	a_{np}

Πίνακας 12.4. Πίνακας φορτώσεων (loading matrix).

	PC1	PC2	. . .	PCk
Ρύπος 1	v_{11}	v_{12}	. . .	v_{1k}
Ρύπος 2	v_{21}	v_{22}	. . .	v_{2k}
. . . .	. . .	. . .	. . .	. . .
Ρύπος p	v_{p1}	v_{p2}	. . .	v_{pk}

Πίνακας 12.5. Πίνακας αποτελεσμάτων (score matrix).

	PC1	PC2	. . .	PCk
Δείγμα 1	u_{11}	u_{12}	. . .	u_{1k}
Δείγμα 2	u_{21}	u_{22}	. . .	u_{2k}
. . . .	. . .	. . .	. . .	. . .
Δείγμα n	u_{n1}	u_{n2}	. . .	u_{nk}

5) Περιστροφή

Η παραπάνω διαδικασία απλοποιείται με κατάλληλη περιστροφή των αξόνων. Η *περιστροφή (rotation)* εφαρμόζεται συνήθως στον πίνακα $\mathbf{V}^*$ και είναι ο πολλαπλασιασμός του πίνακα αυτού με έναν ορθογώνιο πίνακα **R** τέτοιον ώστε ο νέος πίνακας που προκύπτει

$$\mathbf{V}_r^* = \mathbf{V}^*\mathbf{R} \qquad (12.8)$$

να έχει συγκεκριμένες ιδιότητες. Για παράδειγμα, με βάση το ***κριτήριο varimax***, οι στήλες του $\mathbf{V}_r^*$ πρέπει να περιέχουν όσο το δυνατό περισσότερες τιμές 0 ή κοντά στο 0. Όταν περιστρέφουμε τον **V***, τότε αλλάζει και ο **S*** και ο **W*** δεδομένου ότι ισχύει

$$\mathbf{S}_r^* = \mathbf{Z}\,\mathbf{V}_r^* \tag{12.9}$$

και

$$\mathbf{W}_r^{*2} = \mathbf{V}_r^{*T}\,(\mathbf{Z}^T\,\mathbf{Z})\,\mathbf{V}_r^* \tag{12.10}$$

12.2.3 PCA ΜΕ ΣΤΑΤΙΣΤΙΚΑ ΠΡΟΓΡΑΜΜΑΤΑ

Στην ενότητα αυτή θα εξετάσουμε μέσα από δύο παραδείγματα την εφαρμογή της *Ανάλυση σε Κύριες Συνιστώσες (PCA)* τόσο στη διερεύνηση ομάδων όσο και στον προσδιορισμό των παραγόντων που επιδρούν στα δεδομένα ενός πίνακα.

Παράδειγμα 12.1

Στον πίνακα του σχήματος 12.6 δίνονται τα αποτελέσματα της χημικής ανάλυσης ειδωλίων της ίδιας χρονολογικής περιόδου που βρέθηκαν σε τρεις διαφορετικές περιοχές A, B, C. Να γίνει το *διάγραμμα αποτελεσμάτων (score plot)* και να εξαχθούν συμπεράσματα σχετικά με την προέλευση των ειδωλίων.

	A	B	C	D	E	F
1	**Location**	**Al**	**Fe**	**Mg**	**Si**	**Ca**
2	A	7,1	7,2	1,2	12	5
3	A	7,8	6,8	0,8	10	5,2
4	A	8	7	1,1	11,5	5,8
5	A	7,9	7,4	1	10,3	6,2
6	A	8,2	6,6	1	11,9	5
7	B	6,2	6,7	1,2	12,5	6,2
8	B	6,9	7,1	0,8	13	5,8
9	B	5,5	7,4	0,5	11,5	6,9
10	B	7,2	6,7	0,2	11,5	6,2
11	B	6,5	6,2	0,4	12,8	6,7
12	B	6,9	6,8	0,7	12,1	7
13	C	7,1	7	1,1	10	6,2
14	C	7,5	6,6	1,5	11,1	5,8
15	C	6,6	6,2	1,2	11	5,5
16	C	8,2	6,7	0,9	10,3	6
17	C	8	7,1	1	10,9	6,1

Σχήμα 12.6. Χημική σύσταση σε Al, Fe, Mg, Si, Ca (σε αυθαίρετες μονάδες) ειδωλίων από τις περιοχές A, B, C

❖ **Ανάλυση στο ChemStat**

Για να λύσουμε το πρόβλημα αυτό με το *ChemStat* εργαζόμαστε ως εξής: Ακολουθούμε την πορεία *Πρόσθετα* → *ChemStat* → *PCA*, στο πρώτο πλαίσιο που ανοίγει πατάμε *OK* εφόσον έχουμε διευθετήσει τα δεδομένα όπως στο σχήμα 12.6, στο δεύτερο εισάγουμε την περιοχή B1:F7, δηλαδή εισάγουμε και τους τίτλους των μεταβλητών, στο τρίτο και τέταρτο πλαίσιο πατάμε μόνο *OK* για να εφαρμοστεί η περιστροφή *Varimax* και να πάρουμε μόνο τους άξονες που έχουν ιδιοτιμές μεγαλύτερες από 1, και τέλος ορίζουμε το κελί εξόδου των αποτελεσμάτων.

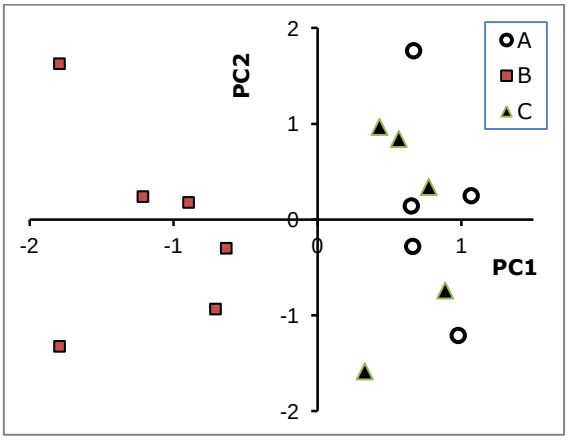
Στα αποτελέσματα παρατηρούμε ότι υπάρχουν μόνο δύο άξονες με ιδιοτιμές μεγαλύτερες από 1. Πηγαίνουμε στον πίνακα αποτελεσμάτων μετά από περιστροφή (Score Matrix-rotated, σχήμα 12.7) και με βάση αυτόν κατασκευάζουμε το διάγραμμα αποτελεσμάτων του σχήματος 12.8 ως εξής.

	Score Matrix-rotated	
1	0,716512	-0,06127
2	0,940232	0,567291
3	0,577316	0,337787
4	0,094206	1,884264
5	1,303352	-0,84832
6	-0,38604	-1,10491
7	-0,51	-0,47778
8	-2,205	0,998555
9	-0,90605	-0,10335
10	-1,29675	-1,80843
11	-1,22665	-0,14291
12	0,113111	1,054956
13	1,072834	-0,42991
14	0,801257	-1,40533
15	0,631858	0,563561
16	0,279833	0,9758

Σχήμα 12.7. Ο πίνακας *αποτελεσμάτων* μετά από περιστροφή

Επιλέγουμε τις πέντε πρώτες τιμές που αντιστοιχούν στους άξονες PC1 και PC2 (συνεπώς αποφεύγουμε τη στήλη με τις τιμές 1, 2, 3, …) και κάνουμε το διάγραμμα διασποράς αυτών των τιμών. Ακολούθως κάνουμε δεξί *κλικ* στο εσωτερικό του διαγράμματος και από την αναδυόμενη λίστα επιλέγουμε *Επιλογή δεδομένων (Select Data)*. Στο πλαίσιο που εμφανίζεται πατάμε στο *Επεξεργασία (Edit)* και εισάγουμε στο *Όνομα σειράς (Series name)* το γράμμα Α που συμβολίζει τη θέση των πέντε πρώτων ειδωλίων.

Πατάμε *OK* και ακολούθως *Προσθήκη (Add)* για να προσθέσουμε νέα δεδομένα στο διάγραμμα. Στο πλαίσιο που ανοίγει εισάγουμε το γράμμα Β στο *Όνομα σειράς*, στο *Τιμές σειράς X (Series X values)* την περιοχή τιμών PC1 που αντιστοιχούν στις επόμενες 6 τιμές PC1, δηλαδή τις -0.386, -0.51, …, -1.2267, και ακολούθως αφού διαγράψουμε το "={1}" εισάγουμε στο *Τιμές σειράς Y (Series Y values)* την περιοχή με τις αντίστοιχες τιμές PC2. Με τον ίδιο τρόπο συνεχίζουμε και εισάγουμε τις τιμές που αντιστοιχούν στην περιοχή C.



Σχήμα 12.8. Διάγραμμα *αποτελεσμάτων (score plot)*

Στο σχήμα 12.8 παρατηρούμε ότι τα σημεία της περιοχής Β σχηματίζουν μια ξεχωριστή *ομάδα (cluster)*, ενώ τα σημεία των περιοχών Α και C σχηματίζουν μαζί μια άλλη ομάδα. Πιθανές ερμηνείες είναι ότι η πηγή αργίλου που χρησιμοποιούσαν οι κάτοικοι της περιοχής Β ήταν διαφορετική από αυτή των περιοχών Α και C και επιπλέον οι κάτοικοι της Β δεν είχαν ανταλλαγές με τους κατοίκους των περιοχών Α και C, τουλάχιστον ως προς τα ειδώλια. Σε ό,τι αφορά τους κατοίκους των περιοχών Α και C ή είχαν κοινή πηγή αργίλου ή ανταλλαγές μεταξύ τους.

❖ Ανάλυση στο SPSS

Για να λύσουμε το πρόβλημα αυτό με το *SPSS* εργαζόμαστε ως εξής: Διευθετούμε τα δεδομένα όπως και στο *Excel*, σχήμα 12.9, και ακολουθούμε την πορεία *Analyze* → *Data Reduction* → *Factor.* Στο

παράθυρο που ανοίγει εισάγουμε τις μεταβλητές Al, Fe, Mg, Si, Ca στο πλαίσιο *Variables*, στο *Rotation* επιλέγουμε ως μέθοδο περιστροφής την *Varimax* και στο *Extraction* επιλέγουμε ως μέθοδο την *Principal Components*. Επίσης από το *Extraction* επιλέγουμε το *Correlation Matrix* για προκατεργασία και είσοδο δεδομένων και το *Eigenvalues over 1* για να πάρουμε μόνο τους άξονες που έχουν ιδιοτιμές μεγαλύτερες από 1. Τέλος, από το *Scores* επιλέγουμε το *Save as variables*. Με αυτή την επιλογή οι τιμές των PC1, PC2, … που αντιστοιχούν σε ιδιοτιμές μεγαλύτερες από 1 αποθηκεύονται στο φύλλο εργασίας με τίτλους FAC1_1, FAC2_1, …. Με *κλικ* στο *OK* δημιουργούνται οι στήλες FAC1_1, FAC2_1, όπως φαίνεται στο σχήμα 12.9, και επιπλέον παίρνουμε μια σειρά από άλλα αποτελέσματα που σε αυτό το παράδειγμα τα αγνοούμε.

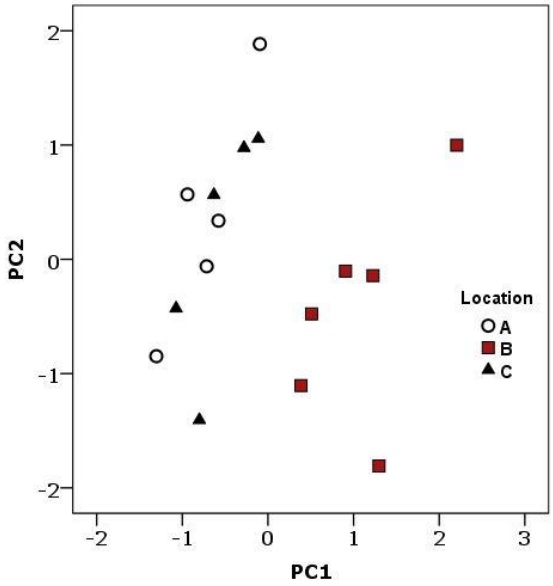
	Location	Al	Fe	Mg	Si	Ca	FAC1_1	FAC2_1
1	A	7,1	7,2	1,2	12,0	5,0	-,71651	-,06127
2	A	7,8	6,8	,8	10,0	5,2	-,94023	,56729
3	A	8,0	7,0	1,1	11,5	5,8	-,57732	,33779
4	A	7,9	7,4	1,0	10,3	6,2	-,09421	1,88426
5	A	8,2	6,6	1,0	11,9	5,0	-1,30335	-,84832
6	B	6,2	6,7	1,2	12,5	6,2	,38604	-1,10491
7	B	6,9	7,1	,8	13,0	5,8	,51000	-,47778
8	B	5,5	7,4	,5	11,5	6,9	2,20500	,99855
9	B	7,2	6,7	,2	11,5	6,2	,90605	-,10335
10	B	6,5	6,2	,4	12,8	6,7	1,29675	-1,80843
11	B	6,9	6,8	,7	12,1	7,0	1,22665	-,14291
12	C	7,1	7,0	1,1	10,0	6,2	-,11311	1,05496
13	C	7,5	6,6	1,5	11,1	5,8	-1,07283	-,42991
14	C	6,6	6,2	1,2	11,0	5,5	-,80126	-1,40533
15	C	8,2	6,7	,9	10,3	6,0	-,63186	,56356
16	C	8,0	7,1	1,0	10,9	6,1	-,27983	,97580

Σχήμα 12.9. Δεδομένα του παραδείγματος 12.1 και οι στήλες FAC1_1, FAC2_1 μετά την εκτέλεση του προγράμματος *Factor* στο *SPSS*

Για να κάνουμε τώρα το διάγραμμα αποτελεσμάτων ακολουθούμε τη διαδικασία *Graphs* → *Legacy Dialogs* → *Scatter/Dot* και στο πρώτο παράθυρο που ανοίγει επιλέγουμε *Simple Scatter* και συνεχίζουμε με *κλικ* στο *Define*. Στο νέο παράθυρο μεταφέρουμε τη μεταβλητή *REGR factor score 1* στο πλαίσιο *X Axis*, τη μεταβλητή *REGR factor score 2* στο *Y Axis* και τη μεταβλητή Location στο *Set Markers by*. Με αυτόν τον τρόπο η κάθε

περιοχή, A, B, C, θα έχει διαφορετικό σύμβολο. Με *κλικ* στο *OK* παίρνουμε (μετά από κατάλληλη μορφοποίηση) το σχήμα 12.10.

Παρατηρούμε και πάλι ότι τα σημεία της περιοχής B σχηματίζουν μια ξεχωριστή ομάδα, ενώ τα σημεία των περιοχών A και C ομαδοποιούνται μαζί.



Σχήμα 12.10. Διάγραμμα *αποτελεσμάτων (score plot)* στο *SPSS*

Παρατήρηση 1. Ανάλογα με το πρόβλημα είναι δυνατόν να δημιουργηθούν στο φύλλο εργασίας περισσότερες από δύο στήλες, FAC1_1, FAC2_1, FAC3_1, … Το διάγραμμα *αποτελεσμάτων (score plot)* γίνεται πάντα μεταξύ των δύο πρώτων στηλών.

Παρατήρηση 2. Αν στο διάγραμμα αποτελεσμάτων δεν ξεχωρίσουν ομάδες δοκιμάζουμε διαφορετικές μεθόδους περιστροφής. Δηλαδή ξανα-εφαρμόζουμε τη μέθοδο και από το *Rotation* ή δεν επιλέγουμε καμία μέθοδο περιστροφής (None) ή δοκιμάζουμε τις άλλες μεθόδους, Quartimax, Equamax, Promax.

Παράδειγμα 12.2

Στον πίνακα του σχήματος 12.11 δίνονται τα αποτελέσματα της χημικής ανάλυσης 25 αέριων δειγμάτων μιας περιοχής σε ng/m^3. Να προσδιοριστούν με *PCA* οι κύριες πηγές ρύπανσης της περιοχής.

◆ Για να αντιμετωπίσουμε τέτοιου είδους προβλήματα χρησιμοποιώντας τη μέθοδο *PCA*, προσδιορίζουμε τον *πίνακα φορτώσεων* μετά από περιστροφή και αφού έχουμε απορρίψει άξονες με ιδιοτιμές μικρότερες από 1. Στους κύριους άξονες που παραμένουν ελέγχουμε τους ρύπους που υπάρχουν στον καθένα και συμπεραίνουμε για τις βασικές πηγές ρύπανσης.

	A	B	C	D	E	F	G	H	I	J	K
1	**Al**	**Si**	**S**	**Cl**	**Ca**	**Fe**	**Zn**	**As**	**Br**	**Sb**	**Pb**
2	816,8	2453,2	639,2	101,4	20325,3	1692,1	57,0	16,0	27,7	17,5	63,4
3	109,9	337,7	765,1	378,7	4426,9	292,4	34,7	13,6	16,6	9,2	41,8
4	549,2	1553,4	1207,5	85,7	18791,5	1366,0	41,7	16,5	23,1	8,6	49,0
5	391,0	1149,0	473,9	98,7	13002,3	946,9	32,4	7,3	14,7	4,9	34,8
6	281,0	765,7	336,7	433,3	5886,5	619,5	23,2	14,4	22,3	5,1	50,5
7	529,0	1725,8	522,1	306,2	20088,4	1302,9	61,9	22,6	39,7	2,0	95,5
8	591,7	1755,0	1180,1	297,6	23343,8	1673,3	66,9	27,2	42,2	11,3	90,8
9	170,6	510,4	1184,2	115,9	4401,2	379,7	30,3	8,9	14,1	2,8	33,0
10	1832,5	4469,8	4655,0	459,3	41850,7	3859,2	152,8	31,7	77,8	4,1	129,4
11	474,8	1410,2	816,4	186,4	16270,3	969,1	55,2	11,5	17,2	6,9	50,2
12	1008,9	2311,1	1663,1	345,1	19868,8	2427,7	60,7	17,0	21,6	4,5	51,3
13	387,6	1091,6	607,9	105,5	11416,3	813,7	34,2	9,7	12,8	8,9	33,0
14	659,7	2031,8	530,7	266,2	23240,9	1555,0	58,3	16,1	20,8	7,5	54,2
15	541,0	1525,8	582,5	197,0	18386,9	1371,8	48,3	11,2	17,0	5,8	34,1
16	415,4	1191,0	316,9	101,6	9015,1	803,7	22,7	11,0	16,7	5,8	46,3
17	127,0	358,8	329,4	16,5	3924,4	330,0	18,7	17,6	29,2	0,9	41,5
18	498,8	1422,6	773,6	79,8	13742,1	1099,4	37,8	7,8	8,7	3,1	17,9
19	1172,2	3142,2	553,0	137,3	15594,5	1965,3	44,7	8,3	6,4	4,8	23,1
20	356,1	1093,3	524,5	54,9	9776,6	914,9	38,7	7,9	10,7	12,3	29,3
21	291,6	846,4	382,4	72,1	7471,4	815,7	28,7	10,4	13,0	2,3	45,1
22	183,1	547,8	226,7	93,9	7361,6	642,9	23,0	7,8	10,9	7,9	36,5
23	495,0	1409,2	542,0	104,9	12085,3	1308,0	39,4	8,4	7,7	7,5	23,1
24	443,2	1166,7	508,7	70,3	8593,1	1121,4	49,8	4,5	8,6	13,7	23,0
25	889,2	2429,7	1684,3	104,2	17258,3	1889,8	58,8	8,1	10,5	3,3	34,4
26	455,9	1152,6	1119,0	73,4	11224,2	1192,7	41,0	12,9	6,0	8,6	22,7
27											

Σχήμα 12.11. Αποτελέσματα της χημικής σύστασης 25 αέριων δειγμάτων

❖ **Ανάλυση στο ChemStat**

Όπως και στο προηγούμενο παράδειγμα, πηγαίνουμε *Πρόσθετα* → *ChemStat* → *PCA*, στο πρώτο πλαίσιο που ανοίγει πατάμε *OK*, στο δεύτερο εισάγουμε όλη την περιοχή A1:K26, στο τρίτο και τέταρτο πλαίσιο πατάμε μόνο *OK* για να εφαρμοστεί η περιστροφή *Varimax* και να πάρουμε μόνο τους άξονες που έχουν ιδιοτιμές μεγαλύτερες από 1, και τέλος ορίζουμε το

κελί εξόδου των αποτελεσμάτων. Με *κλικ* στο *OK* παίρνουμε, ανάμεσα στα άλλα, τον πίνακα του σχήματος 12.12, που είναι ο *πίνακας φορτώσεων* μετά την περιστροφή. Θα αναλύσουμε τον πίνακα αυτόν αφού πρώτα δούμε την ίδια διαδικασία στο *SPSS*.

	Loading Matrix-rotated		
Al	0,960	0,219	-0,034
Si	0,951	0,215	0,020
S	0,742	0,454	-0,159
Cl	0,196	0,762	-0,105
Ca	0,823	0,476	0,091
Fe	0,942	0,291	0,011
Zn	0,808	0,530	0,057
As	0,299	0,907	-0,003
Br	0,399	0,876	-0,043
Sb	0,010	-0,064	0,992
Pb	0,350	0,902	0,017

Σχήμα 12.12. Ο πίνακας *φορτώσεων (loading matrix)* μετά από περιστροφή στο *ChemStat*

❖ **Ανάλυση στο SPSS**

Πηγαίνουμε *Analyze* → *Data Reduction* → *Factor* και στο παράθυρο που ανοίγει εισάγουμε όλες τις μεταβλητές στο πλαίσιο *Variables*. Στο *Rotation* επιλέγουμε την *Varimax* και στο *Extraction* επιλέγουμε την *Principal Components*, το *Correlation Matrix* και το *Eigenvalues over 1*. Στο *Scores* δεν είναι απαραίτητο να επιλέξουμε το *Save as variables*. Όταν ολοκληρώσουμε τις επιλογές αυτές και κάνουμε *κλικ* στο *OK* παίρνουμε, ανάμεσα στα άλλα, τον πίνακα του σχήματος 12.13, που ονομάζεται *Rotated Component Matrix* και είναι ο *πίνακας φορτώσεων* μετά την περιστροφή.

Παρατηρούμε ότι τα δύο προγράμματα, *ChemStat* και *SPSS*, δίνουν πίνακες με ακριβώς ίδιες τιμές, όμως είναι δυνατόν σε κάποιες στήλες οι τιμές να έχουν διαφορετικό πρόσημο στα δύο προγράμματα. Αυτό δεν αλλοιώνει τα τελικά αποτελέσματα. Παρατηρούμε επίσης ότι μετά την περιστροφή υπάρχουν τρεις κύριοι άξονες που θα πρέπει να εξετάσουμε αν μπορούμε να τους συσχετίσουμε με πηγές ρύπανσης. Για το σκοπό αυτό ελέγχουμε τους ρύπους, δηλαδή τα στοιχεία, που υπάρχουν στον κάθε άξονα.

Αν ξεκινήσουμε από τον τρίτο άξονα, παρατηρούμε ότι υπάρχει μόνο

ένα στοιχείο σε σημαντική ποσότητα, το Sb, γεγονός που δείχνει ή την ύπαρξη κάποιας βιομηχανίας-βιοτεχνίας που το εκπέμπει ή το πιθανότερο η παρουσία του να οφείλεται σε καύση σκουπιδιών. Άρα θα πρέπει να γίνει έλεγχος της περιοχής για να διευκρινιστεί πλήρως η αιτία του Sb. Στον δεύτερο άξονα η κύρια συνεισφορά προέρχεται από τα στοιχεία Cl, As, Br, Pb, που προέρχονται από τα καυσαέρια των αυτοκινήτων. Άρα η δεύτερη πηγή ρύπανσης είναι τα αυτοκίνητα. Τέλος, στον πρώτο άξονα κυριαρχούν τα στοιχεία Al, Si, Ca, Fe, ενώ σημαντική είναι και η παρουσία των στοιχείων S και Zn. Τα στοιχεία Al, Si, Ca, Fe σχετίζονται με τη σκόνη που αιωρείται και που σε σημαντικό βαθμό οφείλεται πάλι στην κυκλοφορία των αυτοκινήτων. Συνεπώς ο πρώτος άξονας οφείλεται κυρίως στην αιωρούμενη σκόνη, ενώ η παρουσία των S και Zn δείχνει και μια συμπληρωματική πηγή ρύπανσης που πρέπει να διευκρινιστεί με έλεγχο της περιοχής.

Rotated Component Matrix[a]

	Component		
	1	2	3
Al	,960	,219	-,034
Si	,951	,215	,020
S	,742	,454	-,159
Cl	,196	,762	-,105
Ca	,823	,476	,091
Fe	,942	,291	,011
Zn	,808	,530	,056
As	,299	,907	-,003
Br	,399	,877	-,042
Sb	,009	-,064	,992
Pb	,349	,902	,017

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

Σχήμα 12.13. Ο πίνακας *φορτώσεων (loading matrix)* μετά την περιστροφή στο *SPSS*

Η παραπάνω ανάλυση δείχνει τα πλεονεκτήματα αλλά και τα μειονεκτήματα της μεθόδου. Το πλεονέκτημά της είναι ότι με τρόπο απλό προσδιορίζει τις κύριες πηγές ρύπανσης. Το μειονέκτημα είναι ότι ο προσδιορισμός αυτός είναι ποιοτικός και πολλές φορές δεν είναι πλήρης. Έτσι, στο παράδειγμα που εξετάζουμε ο πρώτος άξονας φαίνεται να περιέχει δύο διαφορετικές πηγές.

12.3 ΑΝΑΛΥΣΗ ΣΕ ΟΜΑΔΕΣ (CA)

Η ***Ανάλυση σε Ομάδες*** *(Cluster Analysis - CA)* περιλαμβάνει μεθόδους που διαχωρίζουν τα δείγματα του πίνακα 12.1 σε ομάδες με παρόμοιες ιδιότητες κατά τρόπο ασυνεχή. Η δημιουργία των ομάδων μπορεί να γίνεται με τρόπο διαδοχικό ενώνοντας στην ομάδα ένα δείγμα κάθε φορά ή με μη διαδοχικό τρόπο ελέγχοντας πολλά δείγματα ταυτόχρονα. Οι μέθοδοι που ανήκουν στην πρώτη κατηγορία ονομάζονται ***Ιεραρχικές*** *(Hierarchical)*, ενώ αυτές της δεύτερης κατηγορίας ονομάζονται ***Μη ιεραρχικές*** *(Non-hierarchical)*.

Η ομαδοποίηση μεταξύ των δειγμάτων γίνεται συνήθως με βάση την *Ευκλείδεια* απόστασή τους, d, που για τα σημεία (a_{11}, a_{12}, a_{13}, …, a_{1m}) και (a_{21}, a_{22}, a_{23}, …, a_{2m}) ορίζεται από τη σχέση

$$d = \sqrt{(a_{11} - a_{21})^2 + (a_{12} - a_{22})^2 + \cdots + (a_{1m} - a_{2m})^2} \qquad (12.11)$$

Μπορεί όμως να χρησιμοποιηθούν και άλλα μέτρα για την απόσταση των σημείων, όπως είναι για παράδειγμα το τετράγωνο της ευκλείδειας απόστασης και ο συντελεστής Pearson.

12.3.1 ΙΕΡΑΡΧΙΚΗ ΑΝΑΛΥΣΗ ΣΕ ΟΜΑΔΕΣ

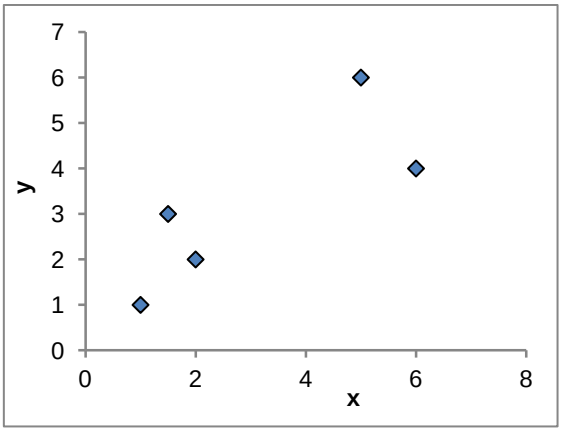
Για να δούμε πως εφαρμόζεται η σχέση (12.11) στην ομαδοποίηση δεδομένων με βάση την *Ιεραρχική Ανάλυση σε Ομάδες*, έστω το πολύ απλό παράδειγμα των δεδομένων του πίνακα 12.6. Τα σημεία αυτού του πίνακα δίνονται εποπτικά στο σχήμα 12.14, όπου εύκολα μπορούμε να δούμε τις ομαδοποιήσεις τους. Για να εφαρμοστεί η μέθοδος *Ιεραρχική Ανάλυση σε Ομάδες* πρώτα υπολογίζουμε τις ευκλείδειες αποστάσεις των σημείων. Οι αποστάσεις αυτές δίνονται στον επόμενο πίνακα 12.7 ταξινομημένες από τη μικρότερη προς τη μεγαλύτερη.

Η πορεία που ακολουθούμε τώρα για την ομαδοποίηση των δεδομένων του πίνακα 12.6 είναι η εξής. Η πρώτη ομάδα που δημιουργείται είναι μεταξύ των δειγμάτων 2 και 3 που έχουν την μικρότερη ευκλείδεια απόσταση, d=1.12. Στη συνέχεια βλέπουμε ότι η επόμενη μικρότερη απόσταση είναι μεταξύ των δειγμάτων 1 και 2. Επειδή το 2 ήδη ανήκει στην πρώτη ομάδα, ενώνουμε τα δείγματα 1, 2 και 3 σε μια νέα ενιαία ομάδα. Η επόμενη μικρότερη απόσταση είναι μεταξύ των δειγμάτων 1 και 3, αλλά αυτά ήδη βρίσκονται στην ομάδα που δημιουργήσαμε. Έτσι περνάμε στη απόσταση d=2.24 μεταξύ των δειγμάτων 4 και 5 και

δημιουργούμε μεταξύ των δειγμάτων αυτών μια δεύτερη ομάδα. Τέλος, ενοποιούμε όλα τα δείγματα σε μια ενιαία ομάδα.

Πίνακας 12.6. Πίνακας δύο διαστάσεων

	Ιδιότητα 1	Ιδιότητα 2
Δείγμα 1	1	1
Δείγμα 2	1.5	3
Δείγμα 3	2	2
Δείγμα 4	5	6
Δείγμα 5	6	4

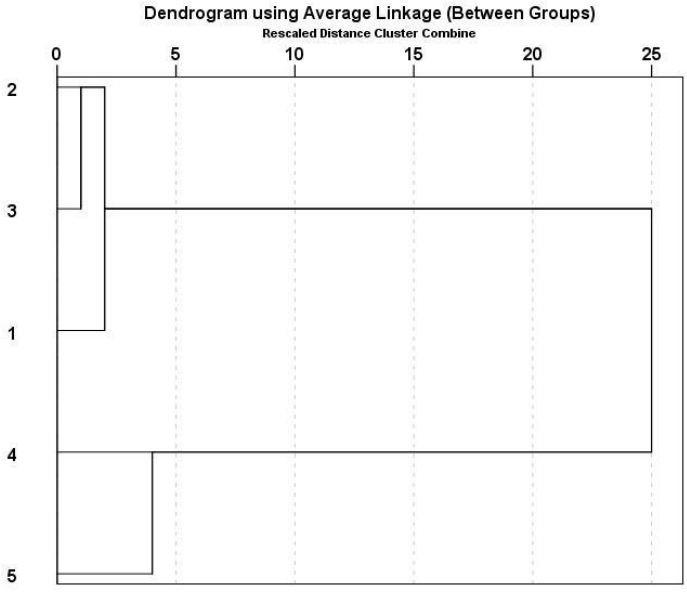


Σχήμα 12.14. Γραφική παράσταση των σημείων του πίνακα 12.6

Πίνακας 12.7. Ευκλείδειες αποστάσεις των σημείων του πίνακα 12.6.

Δείγματα	d	Δείγματα	d
2-3	1.12	2-5	4.61
1-3	1.41	3-4	5
1-2	2.06	2-4	5.61
4-5	2.24	1-5	5.83
3-5	4.47	1-4	6.4

Τα διαδοχικά αυτά βήματα ομαδοποίησης των δεδομένων φαίνονται παραστατικά στο ***δενδρόγραμμα*** *(dendrogram)* που δίνεται στο σχήμα 12.15. Παρατηρούμε, όπως είδαμε και παραπάνω, ότι πρώτα ομαδοποιούνται και συνεπώς παρουσιάζουν ομοιότητες τα δείγματα 2 και 3. Αυτά ως ομάδα παρουσιάζουν ομοιότητα με το 1 και συνεπώς μπορούν να δημιουργήσουν μια ευρύτερη ομάδα. Παρατηρούμε επίσης ότι αυτή η ομάδα διαφοροποιείται από την άλλη ομάδα, των δειγμάτων 4 και 5.



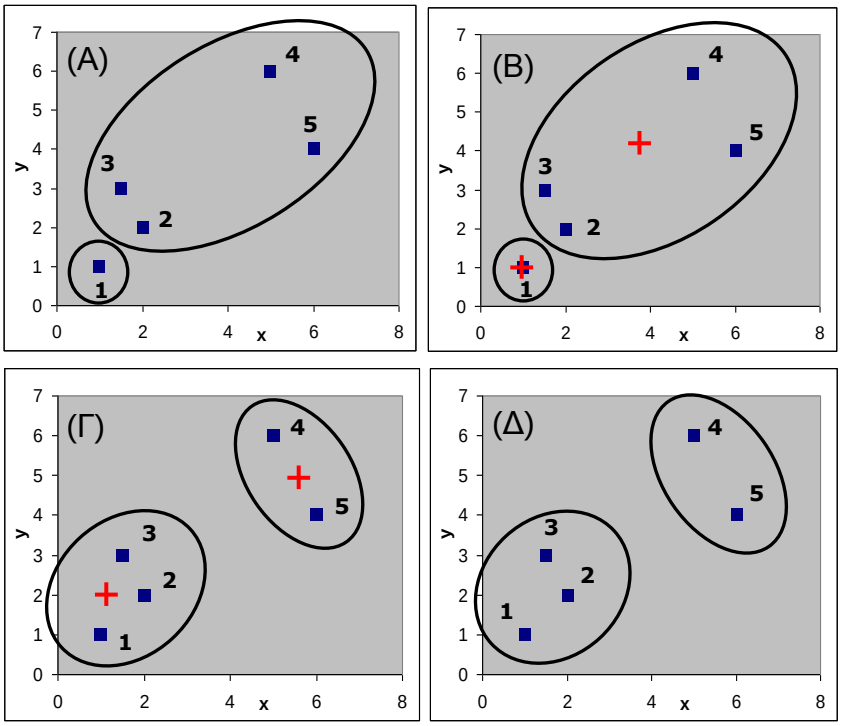
Σχήμα 12.15. Δενδρόγραμμα που αντιστοιχεί στις ομαδοποιήσεις των σημείων του σχήματος 12.14

12.3.2 ΜΗ ΙΕΡΑΡΧΙΚΗ ΑΝΑΛΥΣΗ ΣΕ ΟΜΑΔΕΣ

Στις μεθόδους *Μη Ιεραρχικής Ανάλυσης σε Ομάδες* συνήθως πρέπει να γνωρίζουμε το αριθμό των ομάδων και αυτό είναι ένα σημαντικό μειονέκτημά τους. Η απλούστερη από αυτές προσδιορίζει τις ομάδες ως εξής. Έστω ότι θέλουμε να διαχωρίσουμε τα στοιχεία του πίνακα 12.6 σε δύο ομάδες. Τα βήματα που ακολουθούμε είναι:

1. Επιλέγουμε τα δύο πρώτα δείγματα, (1, 1) και (1.5, 3), και υπολογίζουμε τις αποστάσεις των υπολοίπων δειγμάτων από αυτά.

Αυτός ο υπολογισμός έχει ήδη γίνει και τα αποτελέσματα δίνονται στον πίνακα 12.7, όπου παρατηρούμε ότι όλα τα δείγματα 3, 4 και 5 είναι πιο κοντά στο 2 από ότι στο ένα. Συνεπώς δημιουργούνται δύο ομάδες. Η μία περιλαμβάνει μόνο το δείγμα 1 και η άλλη τα υπόλοιπα δείγματα 2, 3, 4 και 5 (σχήμα 12.16Α).



Σχήμα 12.16. Διαδοχικά στάδια προσδιορισμού ομάδων (clusters)

2. Υπολογίζονται τα ***κέντρα βάρους*** *(centroids)* των δύο ομάδων:

$$\bar{x}_1 = (1, 1) \text{ και}$$

$$\bar{x}_2 = \left(\frac{1.5+2+5+6}{4}, \frac{3+2+6+4}{4}\right) = (3.625, 3.75)$$

Τα κέντρα βάρους δίνονται στο σχήμα 12.16Β με κόκκινο σταυρό. Προφανώς η ομάδα με το ένα δείγμα έχει κέντρο βάρους το ίδιο το

σημείο του δείγματος (1, 1).

3. Υπολογίζουμε τις αποστάσεις των κέντρων βάρους $\bar{x}_1$ = (1, 1) και $\bar{x}_2$ = (3.625, 3.75) από όλα τα δείγματα. Τα αποτελέσματα δίνονται στον πίνακα 12.8, όπου παρατηρούμε ότι τώρα τα δείγματα 1, 2, 3 δημιουργούν μια ομάδα και τα υπόλοιπα 4, 5 μια άλλη (σχήμα 12.16Γ).

Πίνακας 12.8. Ευκλείδειες αποστάσεις των δειγμάτων από τα κέντρα βάρους (1, 1) και (3.625, 3.75)

Δείγμα	d από $\bar{x}_1$	d από $\bar{x}_2$
1	0	3.802
2	2.062	2.253
3	1.414	2.388
4	6.403	2.637
5	4.610	2.388

4. Υπολογίζονται τα κέντρα βάρους των δύο νέων ομάδων:

$$\bar{x}_1 = (\frac{1+1.5+2}{3}, \frac{1+3+2}{3}) = (1.5, 2) \text{ και}$$

$$\bar{x}_2 = (\frac{5+6}{2}, \frac{6+4}{2}) = (5.5, 5)$$

Τα κέντρα βάρους δίνονται στο σχήμα 12.16Γ με κόκκινο σταυρό.

Πίνακας 12.9. Ευκλείδειες αποστάσεις των δειγμάτων από τα κέντρα βάρους (1.5, 2) και (5.5, 5)

Δείγμα	d από $\bar{x}_1$	d από $\bar{x}_2$
1	1.118	6.021
2	1.000	4.472
3	0.500	4.610
4	5.315	1.118
5	4.924	1.118

5. Υπολογίζουμε πάλι τις αποστάσεις των νέων κέντρων βάρους, $\bar{x}_1$ = (1.5, 2) και $\bar{x}_2$ = (5.5, 5), από όλα τα δείγματα και τα αποτελέσματα δίνονται στον πίνακα 12.9, όπου παρατηρούμε ότι δεν δημιουργούνται άλλες ομάδες. Άρα η διαδικασία ομαδοποίησης τελειώνει εδώ με τη δημιουργία των ομάδων με δείγματα 1, 2, 3 η μία και 4, 5 η άλλη (σχήμα 12.16Δ).

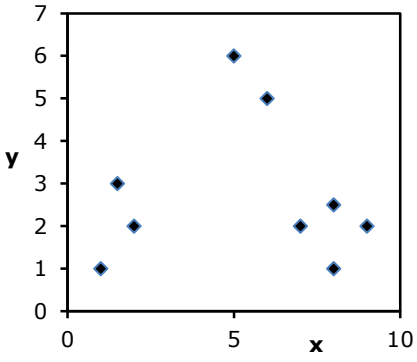
12.3.3 CA ΜΕ ΣΤΑΤΙΣΤΙΚΑ ΠΡΟΓΡΑΜΜΑΤΑ

Από τα προγράμματα που εξετάζουμε το *Excel* και το *ChemStat* δεν έχουν δυνατότητες εφαρμογής των μεθόδων *CA*. Συνεπώς θα χρησιμοποιήσουμε μόνο το *SPSS*.

Παράδειγμα 12.3

Να προσδιοριστούν με μεθόδους *CA* οι ομάδες που σχηματίζουν τα x, y δεδομένα του πίνακα του σχήματος 12.17.

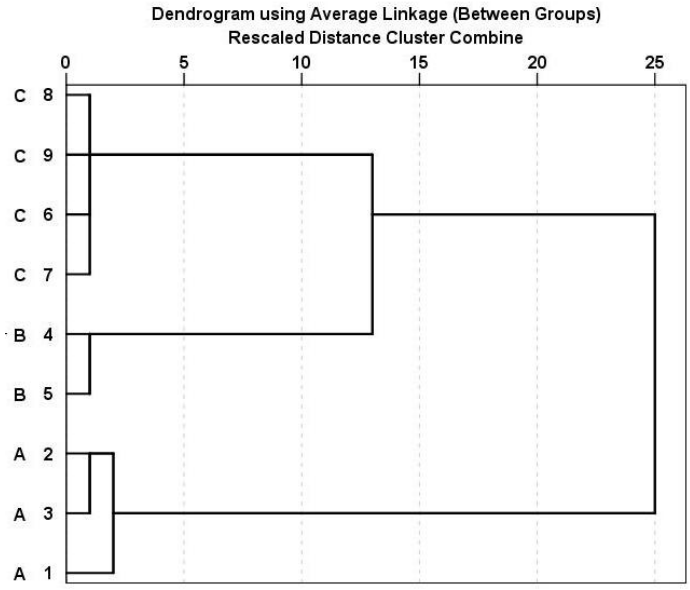
	g1	g2	x	y
1	1	A	1,0	1,0
2	2	A	1,5	3,0
3	3	A	2,0	2,0
4	4	B	5,0	6,0
5	5	B	6,0	5,0
6	6	C	8,0	1,0
7	7	C	9,0	2,0
8	8	C	7,0	2,0
9	9	C	8,0	2,5



Σχήμα 12.17. Πίνακας δεδομένων και γραφική τους παράσταση

◆ Στον πίνακα του σχήματος 12.17 η στήλη g1 αριθμεί τα δείγματα, ενώ η g2 τα διαχωρίζει σε ομάδες, επειδή τα σημεία x, y έχουν προεπιλεγεί έτσι ώστε να σχηματίζουν τρεις διακριτές ομάδες, όπως φαίνεται από το σχήμα δίπλα στον πίνακα. Για να εφαρμόσουμε στο *SPSS* την ιεραρχική μέθοδο διαχωρισμού σε ομάδες ακολουθούμε την πορεία *Analyze* → *Classify* → *Hierarchical Cluster* και στο παράθυρο που ανοίγει μεταφέρουμε τις μεταβλητές x, y στο πλαίσιο *Variables*, τη μεταβλητή g2 στο *Label*

Cases by και επιλέγουμε στο πάνελ *Cluster* το *Cases* για να δηλώσουμε ότι θα αναζητηθούν ομαδοποιήσεις μεταξύ των δειγμάτων (γραμμών) και όχι μεταξύ των μεταβλητών (στηλών). Στο *Plots* επιλέγουμε *Dendrogram*, ενώ από το *Methods* επιλέγουμε το *Between-groups linkage* για τη μέθοδο σχηματισμού των ομάδων-clusters και το *Euclidean* ως μέτρο των αποστάσεων μεταξύ των σημείων. Σε πολύπλοκα δείγματα οι επιλογές αυτές μπορεί να οδηγήσουν σε διαφορετικά δενδρογράμματα και αυτό είναι ένα μειονέκτημα της μεθόδου. Ολοκληρώνουμε με *κλικ* στο *Continue* και στο *OK*. Το δενδρόγραμμα που παίρνουμε δίνεται στο σχήμα 12.18 και δείχνει καθαρά τις τρεις ομάδες δειγμάτων που έχουμε δημιουργήσει στον πίνακα δεδομένων.



Σχήμα 12.18. Δενδρόγραμμα που αντιστοιχεί στις ομάδες του πίνακα του σχήματος 12.17

Για να εφαρμόσουμε τη *Μη Ιεραρχική Μέθοδο k-μέσοι (k-means clustering)*, πηγαίνουμε *Analyze → Classify → K-Means Cluster* και στο παράθυρο που ανοίγει εισάγουμε τις μεταβλητές x, y στο πλαίσιο *Variables*, στο *Number of Clusters* εισάγουμε την τιμή 3 και από το *Save* επιλέγουμε το *Cluster membership.* Με αυτή την επιλογή στο φύλλο εργασίας θα εμφανιστεί μια νέα μεταβλητή, η QCL_1, που θα δείχνει τις σχηματιζόμενες

ομάδες.

Στον πίνακα του σχήματος 12.19 παρατηρούμε ότι η μεταβλητή QCL_1 ταυτίζεται με τη μεταβλητή g2 και συνεπώς το πρόγραμμα ανίχνευσε σωστά τις ομάδες του πίνακα του σχήματος 12.17. Αν όμως δώσουμε στο *Number of Clusters* την τιμή 4, τότε θα πάρουμε τα αποτελέσματα της μεταβλητής QCL_2. Παρατηρούμε ότι το πρόγραμμα τώρα αναγκάζεται να δημιουργήσει 4 ομάδες. Το παράδειγμα αυτό δείχνει το μειονέκτημα της μεθόδου *k-μέσοι (k-means clustering)*. Όταν δεν γνωρίζουμε τον αριθμό των ομάδων σε έναν πίνακα δεδομένων και θέλουμε να τον προσδιορίσουμε με τις διάφορες μεθόδους που εξετάζουμε, η μέθοδος αυτή δεν βοηθάει καθόλου.

	g1	g2	x	y	QCL_1	QCL_2
1	1	A	1,0	1,0	1	1
2	2	A	1,5	3,0	1	1
3	3	A	2,0	2,0	1	1
4	4	B	5,0	6,0	2	4
5	5	B	6,0	5,0	2	4
6	6	C	8,0	1,0	3	3
7	7	C	9,0	2,0	3	2
8	8	C	7,0	2,0	3	2
9	9	C	8,0	2,5	3	2
10						

Σχήμα 12.19. Αποτελέσματα της μεθόδου *k-means clustering* όταν k = 3 (QCL_1) και k = 4 (QCL_2)

Παράδειγμα 12.4

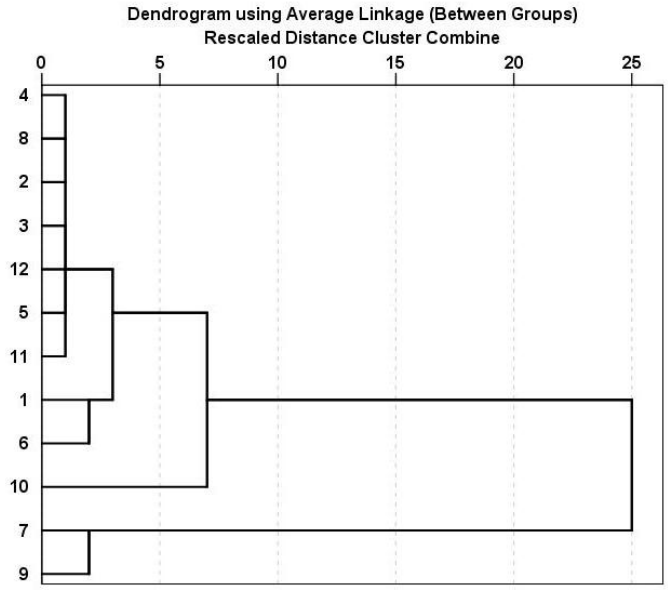
Στον πίνακα του σχήματος 12.20 δίνεται η ποσότητα των διατροφικών συστατικών 12 δημητριακών προϊόντων. Στον πίνακα αυτόν x1 είναι η ποσότητα σε πρωτεΐνες (g), x2 σε υδατάνθρακες (g), x3 σε λίπος (g), x4 οι θερμίδες (kJ) και x5 η ποσότητα σε βιταμίνη A (%). Να εξετασθεί η ύπαρξη ομάδων παρόμοιων δημητριακών προϊόντων.

◆ Εργαζόμαστε όπως και στο προηγούμενο παράδειγμα. Στο δενδρόγραμμα του σχήματος 12.21 παρατηρούμε ότι τα δημητριακά 7 και 9 σχηματίζουν μια ξεχωριστή ομάδα, πιθανόν λόγω της αυξημένης βιταμίνης A που περιέχουν. Το δημητριακό 10 διαφοροποιείται από τα υπόλοιπα ίσως επειδή έχει ελάχιστη βιταμίνη A και λίγες θερμίδες. Επίσης διαφοροποίηση παρατηρείται και στα δημητριακά 1 και 6 πιθανόν λόγω των

πολλών θερμίδων.

	x1	x2	x3	x4	x5
1	8	17	0	120	5
2	4	22	1	110	22
3	1	25	1	100	23
4	5	19	0	105	25
5	1	28	0	105	20
6	4	22	0	125	20
7	2	23	1	115	90
8	4	20	0	105	22
9	4	20	0	100	100
10	1	12	1	75	2
11	2	21	1	95	15
12	3	23	0	98	20
13					

Σχήμα 12.20. Πίνακας δεδομένων

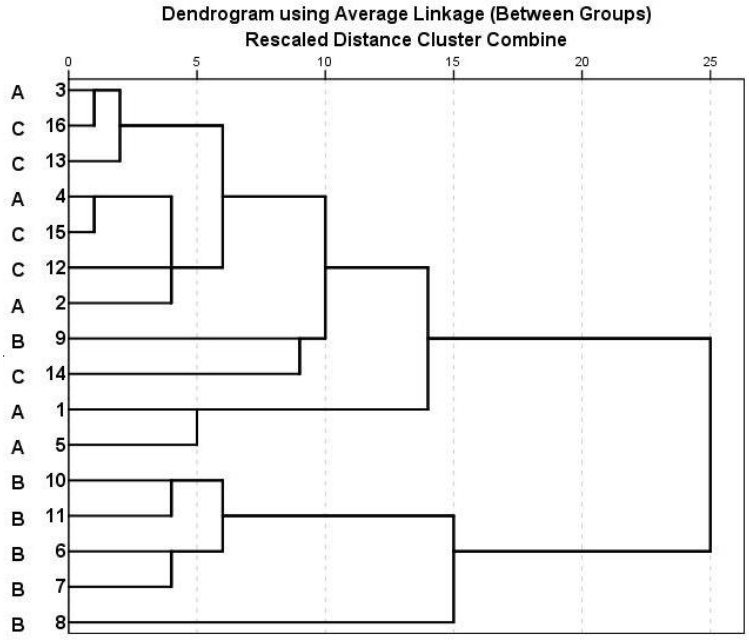


Σχήμα 12.21. Δενδρόγραμμα των τιμών του πίνακα του σχήματος 12.20

Παράδειγμα 12.5

Να γίνει το δενδρόγραμμα των δεδομένων του παραδείγματος 12.1.

◆ Εργαζόμαστε όπως στο παράδειγμα 12.3 και παίρνουμε το δενδρόγραμμα του σχήματος 12.22. Παρά την πολυπλοκότητα του δενδρογράμματος μπορούμε να διακρίνουμε ότι η ομάδα Β ξεχωρίζει από τις ομάδες Α και C που αλληλο-εμπλέκονται. Η εικόνα αυτή βρίσκεται σε πλήρη συμφωνία με τα αποτελέσματα της *PCA* και συγκεκριμένα με τα σχήματα 12.8 και 12.10.

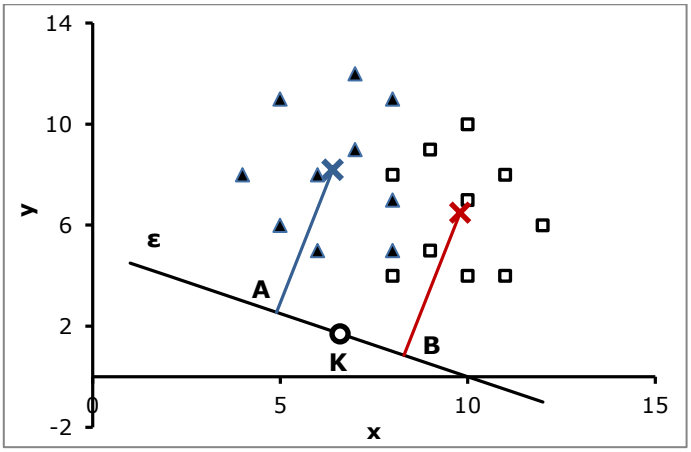


Σχήμα 12.22. Δενδρόγραμμα των τιμών του παραδείγματος 12.1

12.4 ΓΡΑΜΜΙΚΗ ΔΙΑΧΩΡΙΣΤΙΚΗ ΑΝΑΛΥΣΗ (LDA)

Η ***Γραμμική Διαχωριστική Ανάλυση*** *(Linear Discriminant Analysis - LDA)* είναι μια στατιστική μέθοδος που μας επιτρέπει να βρούμε σε ποια κατηγορία ανήκουν ένα ή περισσότερα δείγματα, με την προϋπόθεση ότι υπάρχει μια βάση δεδομένων με κατηγορίες στις οποίες μπορούν να ανήκουν αυτά. Όπως έχει ήδη αναφερθεί, για να εφαρμοστεί η *LDA* πρέπει οι μεταβλητές στις ομάδες να ακολουθούν την *πολυμεταβλητή κανονική κατανομή* και να υπάρχει *ομοιογένεια της διασποράς*. Όμως γενικά η μέθοδος εφαρμόζεται χωρίς ιδιαίτερους ελέγχους αρκεί να μην υπάρχουν πολύ ακραίες τιμές.

Αν και υπάρχουν διαφορετικές μαθηματικές προσεγγίσεις, η αρχή της μεθόδου φαίνεται εποπτικά στο σχήμα 12.23 για την περίπτωση που έχουμε μόνο δύο ομάδες δεδομένων με δύο μεταβλητές, x, y, η κάθε μία. Στο σχήμα αυτό η μία ομάδα, έστω *ομάδα 1*, αποτελείται από τα τρίγωνα και η άλλη, *ομάδα 2*, από τα τετράγωνα. Πρώτα υπολογίζουμε τα κέντρα βάρους των δύο ομάδων που στο σχήμα 12.23 σημειώνονται με τα σημεία **x**. Στόχος τώρα της μεθόδου είναι να προσδιορίσει μία ευθεία (ε) τέτοια που η απόσταση των προβολών των κέντρων βάρους των δύο ομάδων σε αυτή να είναι η μέγιστη δυνατή. Στο σχήμα 12.23 η απόσταση αυτή είναι η ΑΒ. Ακολούθως προσδιορίζεται το κέντρο Κ του τμήματος ΑΒ. Είναι προφανές ότι κάθε σημείο που βρίσκεται αριστερά του Κ θα ανήκει στην *ομάδα 1* και κάθε σημείο που είναι δεξιά θα ανήκει στην *ομάδα 2*.



Σχήμα 12.23. Αρχή της *LDA*

Θα πρέπει να σημειώσουμε ότι η μέθοδος όχι μόνο κατατάσσει τα σημεία σε ομάδες, αλλά μπορεί να προσδιορίσει και την πιθανότητα με την οποία γίνεται αυτή η κατάταξη, δεδομένου ότι τα σημεία ακολουθούν την *πολυμεταβλητή κανονική κατανομή*.

12.4.1 LDA ΜΕ ΣΤΑΤΙΣΤΙΚΑ ΠΡΟΓΡΑΜΜΑΤΑ

Παράδειγμα 12.6

Στον πίνακα 12.10 δίνεται η συγκέντρωση σε g/L της σουρκόζης, γλυκόζης, φρουκτόζης και σορβιτόλης στο χυμό μήλων που προέρχονται από τρεις διαφορετικές ποικιλίες. Με βάση τα δεδομένα του πίνακα να προσδιοριστεί η ποικιλία στην οποία ανήκουν τρία μήλα με συγκεντρώσεις (15, 20, 55, 3.5), (20, 11, 42, 5) και (27, 8, 45, 4) σε σουρκόζη, γλυκόζη, φρουκτόζη και σορβιτόλη, αντίστοιχα.

Πίνακας 12.10. Πίνακας δεδομένων.

Ποικιλία	Σουρκόζη	Γλυκόζη	Φρουκτόζη	Σορβιτόλη
A	18	5	39	4
A	22	10	42	3
A	27	12	45	2.5
A	29	7	46	3
A	33	12	38	5
B	5	25	45	3.5
B	15	20	44	3.5
B	16	18	49	4.8
B	12	19	47	3
B	10	21	46	3.3
B	15	22	49	4.5
C	5	15	51	4.5
C	15	22	55	4.8
C	14	23	65	5
C	10	12	59	5.2
C	12	24	53	4.5
C	16	16	52	4.9

❖ **Ανάλυση στο ChemStat**

Για να χρησιμοποιήσουμε το *ChemStat* διευθετούμε τα δεδομένα όπως στο σχήμα 12.24, όπου έχουμε αλλάξει τη μεταβλητή *Ποικιλία* με την *group* που παίρνει τις τιμές 0, 1, 2, … Θα πρέπει να προσέξουμε ότι για την εφαρμογή της μεθόδου στο *ChemStat* οι ομάδες ορίζονται ΜΟΝΟ με αριθμούς που πρέπει να αρχίζουν από το 0. Ακολούθως πηγαίνουμε

Πρόσθετα → *ChemStat* → *Discriminant Analysis* και στο πρώτο πλαίσιο εισαγωγής δεδομένων εισάγουμε την περιοχή A2:A18 που περιέχει μόνο τις τιμές της μεταβλητής group και στο επόμενο εισάγουμε την περιοχή B2:E21 των τιμών των μεταβλητών. Στο τελευταίο πλαίσιο αφήνουμε την τιμή 1 ώστε το πρόγραμμα να εφαρμόσει την τεχνική *cross-validation*, η οποία ουσιαστικά ελέγχει πόσο αξιόπιστα είναι τα δεδομένα ώστε να εφαρμοστεί η *LDA*.

Τα αποτελέσματα εμφανίζονται δεξιά της τελευταίας στήλης των δεδομένων και συνεπώς η περιοχή αυτή πρέπει να είναι ελεύθερη. Παρατηρούμε ότι η μέθοδος ταξινομεί σωστά το 88.2% των δειγμάτων και προβλέπει ότι το πρώτο άγνωστο δείγμα ανήκει στην ποικιλία C, ενώ τα άλλα δύο στην A. Τέλος, ελέγχουμε την τιμή που εκφράζει το ποσοστό της σωστής ταξινόμησης με την τεχνική *cross-validation*. Αυτό πρέπει να είναι όσο το δυνατό υψηλότερο, συνήθως μεγαλύτερο από 80% για να εφαρμοστεί με επιτυχία η μέθοδος.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	**group**	**Su**	**Gl**	**Fr**	**So**	**Discriminatior**		**Classification summary**					
2	0	18	5	39	4	0							
3	0	22	10	42	3	0		88,23% of original grouped cases correctly classified					
4	0	27	12	45	2,5	0		88,23% of cross-validated grouped cases correctly classified.					
5	0	29	7	46	3	0							
6	0	33	12	38	5	0							
7	1	5	25	45	3,5	1							
8	1	15	20	44	3,5	1							
9	1	16	18	49	4,8	2							
10	1	12	19	47	3	1							
11	1	10	21	46	3,3	1							
12	1	15	22	49	4,5	1							
13	2	5	15	51	4,5	2							
14	2	15	22	55	4,8	2							
15	2	14	23	65	5	2							
16	2	10	12	59	5,2	2							
17	2	12	24	53	4,5	1							
18	2	16	16	52	4,9	2							
19		15	20	55	3,5	2							
20		20	11	42	5	0							
21		27	8	45	4	0							

Σχήμα 12.24. Διευθέτηση δεδομένων στο *ChemStat* και αποτελέσματα του προγράμματος *Discriminant Analysis*

❖ Ανάλυση στο SPSS

Μεταφέρουμε τα δεδομένα στο *SPSS*, όπως φαίνεται στο σχήμα 12.25, όπου τα άγνωστα δείγματα τοποθετούνται στο τέλος των στηλών *Su, Gl, Fr, So*. Ακολούθως δημιουργούμε μια νέα στήλη με όνομα group, της οποίας η μεταβλητή παίρνει τις τιμές 1 όταν αντιστοιχεί στην ποικιλία A, 2

στην B και 3 στην C. Μετά τη διευθέτηση των δεδομένων, πηγαίνουμε *Analyze* → *Classify* → *Discriminant* και στο παράθυρο που ανοίγει μεταφέρουμε τις μεταβλητές *Su, Gl, Fr, So* στο πλαίσιο *Independent* και τη μεταβλητή *group* στο *Grouping Variable*. Με *κλικ* στο *Define Range* εισάγουμε στο *Minimum* την τιμή 1 και στο *Maximum* την τιμή 3. Κάνουμε *κλικ* στο *Define* και στο *Save* επιλέγουμε *Predicted group membership* και *Probabilities of group membership Plots*. Επίσης στο *Classify* επιλέγουμε το *Summary table* και *Leave-one-out classification* και ολοκληρώνουμε με *κλικ* στο *Continue* και στο *OK*.

Από τους πίνακες που παίρνουμε ενδιαφέρον παρουσιάζει ο τελευταίος, ο *Classification Results* (σχήμα 12.26). Επίσης το πρόγραμμα στο φύλλο εργασίας δημιουργεί 4 νέες στήλες με τίτλους Dis_1, Dis1_1, Dis2_1 και Dis3_1 (σχήμα 12.25).

	Var	Su	Gl	Fr	So	group	Dis_1	Dis1_1	Dis2_1	Dis3_1
1	A	18	5	39	4,0	1	1	1,00	,00	,00
2	A	22	10	42	3,0	1	1	1,00	,00	,00
3	A	27	12	45	2,5	1	1	1,00	,00	,00
4	A	29	7	46	3,0	1	1	1,00	,00	,00
5	A	33	12	38	5,0	1	1	1,00	,00	,00
6	B	5	25	45	3,5	2	2	,00	,99	,01
7	B	15	20	44	3,5	2	2	,00	1,00	,00
8	B	16	18	49	4,8	2	2	,00	,75	,25
9	B	12	19	47	3,0	2	2	,00	1,00	,00
10	B	10	21	46	3,3	2	2	,00	1,00	,00
11	B	15	22	49	4,5	2	2	,00	,89	,11
12	C	5	15	51	4,5	3	3	,00	,05	,95
13	C	15	22	55	4,8	3	3	,00	,04	,96
14	C	14	23	65	5,0	3	3	,00	,00	1,00
15	C	10	12	59	5,2	3	3	,00	,00	1,00
16	C	12	24	53	4,5	3	3	,00	,18	,82
17	C	16	16	52	4,9	3	3	,00	,18	,82
18		15	20	55	3,5	.	2	,00	,63	,37
19		20	11	42	5,0	.	1	,83	,17	,00
20		27	8	45	4,0	.	1	1,00	,00	,00

Σχήμα 12.25. Διευθέτηση δεδομένων στο *SPSS* και αποτελέσματα του προγράμματος *Discriminant*

Στον πίνακα *Classification Results* αξιολογείται αν πράγματι τα αρχικά δεδομένα σχηματίζουν τρεις διακριτές κατηγορίες. Παρατηρούμε ότι

και εδώ το ποσοστό της σωστής ταξινόμησης με την τεχνική *cross-validation* είναι 88.2%.

Classification Results[b,c]

			Predicted Group Membership			
		group	1	2	3	Total
Original	Count	1	5	0	0	5
		2	0	6	0	6
		3	0	0	6	6
		Ungrouped cases	2	1	0	3
	%	1	100,0	,0	,0	100,0
		2	,0	100,0	,0	100,0
		3	,0	,0	100,0	100,0
		Ungrouped cases	66,7	33,3	,0	100,0
Cross-validated[a]	Count	1	5	0	0	5
		2	0	5	1	6
		3	0	1	5	6
	%	1	100,0	,0	,0	100,0
		2	,0	83,3	16,7	100,0
		3	,0	16,7	83,3	100,0

a. Cross validation is done only for those cases in the analysis. In cross validation, each case is classified by the functions derived from all cases other than that case.
b. 100,0% of original grouped cases correctly classified.
c. 88,2% of cross-validated grouped cases correctly classified.

Σχήμα 12.26. Πίνακας αξιολόγησης αποτελεσμάτων

Από τις νέες στήλες, η Dis_1 μας δίνει την πρόβλεψη του προγράμματος για κάθε δείγμα, ενώ στις επόμενες στήλες είναι η εκτιμούμενη πιθανότητα ένα δείγμα να ανήκει στην ποικιλία Α (στήλη Dis1_1) ή στη Β (στήλη Dis2_1) ή στη C (στήλη Dis3_1). Για τα άγνωστα δείγματα ισχύουν τα ακόλουθα: Το πρώτο μήλο είναι με πιθανότητα 0.63 της ποικιλίας Β, το δεύτερο με πιθανότητα 0.83 ανήκει στην ποικιλία Α και το τρίτο ανήκει επίσης στην ποικιλία Α αλλά με πιθανότητα 1. Παρατηρούμε ότι υπάρχει μια διαφοροποίηση των αποτελεσμάτων σε σχέση με αυτά του *ChemStat*. Το *SPSS* φαίνεται να διαθέτει έναν ισχυρότερο αλγόριθμο για την εφαρμογή της μεθόδου και συνεπώς θα πρέπει να προτιμάται του *ChemStat*.

Όταν υπάρχουν αμφιβολίες για την εφαρμογή της μεθόδου, μπορεί να χρησιμοποιηθεί η μη ιεραρχική μέθοδος *k-Means Clustering*, επειδή

γνωρίζουμε τον αριθμό των ομάδων. Μια εφαρμογή της μεθόδου αυτής στο πρόβλημα που εξετάζουμε δίνει τα αποτελέσματα του σχήματος 12.27. Στον πίνακα αυτού του σχήματος το πρόγραμμα χρησιμοποιεί τον αριθμό 2 για την ποικιλία Α, το 3 για τη Β και το 1 για τη C. Παρατηρούμε ότι τα αποτελέσματα είναι πιο κοντά στα αντίστοιχα του ChemStat.

Θα πρέπει πάντως να τονιστεί ότι η *LDA* και η *k-Means Clustering* χρησιμοποιούν διαφορετικά μέτρα για τις αποστάσεις. Η *Μη Ιεραρχική Ανάλυση* χρησιμοποιεί την απλή ευκλείδεια απόσταση, ενώ η *Διαχωριστική Ανάλυση* χρησιμοποιεί την απόσταση *Mahalanobis* που αποδίδει πιο σωστά τις αποστάσεις μεταξύ ομάδων σημείων. Για το λόγο αυτό η *Γραμμική Διαχωριστική Ανάλυση* υπερτερεί της *Μη Ιεραρχικής Ανάλυσης* στον διαχωρισμό των ομάδων.

	Var	Su	Gl	Fr	So	QCL_1
1	A	18	5	39	4,0	2
2	A	22	10	42	3,0	2
3	A	27	12	45	2,5	2
4	A	29	7	46	3,0	2
5	A	33	12	38	5,0	2
6	B	5	25	45	3,5	3
7	B	15	20	44	3,5	3
8	B	16	18	49	4,8	3
9	B	12	19	47	3,0	3
10	B	10	21	46	3,3	3
11	B	15	22	49	4,5	3
12	C	5	15	51	4,5	3
13	C	15	22	55	4,8	1
14	C	14	23	65	5,0	1
15	C	10	12	59	5,2	1
16	C	12	24	53	4,5	3
17	C	16	16	52	4,9	1
18		15	20	55	3,5	1
19		20	11	42	5,0	2
20		27	8	45	4,0	2

Σχήμα 12.27. Αποτελέσματα της μεθόδου *k-Means Clustering*

12.5 ΑΝΑΛΥΣΗ ΔΙΑΣΠΟΡΑΣ ΠΟΛΛΩΝ ΜΕΤΑΒΛΗΤΩΝ (MANOVA)

Με τη μονοπαραγοντική ανάλυση διασποράς (*ANOVA*) εξετάζουμε αν υπάρχουν στατιστικά σημαντικές διαφορές μεταξύ των μέσων τιμών τριών ή περισσοτέρων δειγμάτων. Η ***Ανάλυση Διασποράς Πολλών Μεταβλητών*** *(Multivariate Analysis of Variance – MANOVA)* επεκτείνει αυτή τη δυνατότητα και εξετάζει την ύπαρξη στατιστικά σημαντικών διαφορών μεταξύ ομάδων δειγμάτων. Ως επέκταση της μονοπαραγοντικής ανάλυσης διασποράς, η εφαρμογή της *MANOVA* προϋποθέτει την *ομοιογένεια της διασποράς* και οι μεταβλητές στις ομάδες να ακολουθούν την *πολυμεταβλητή κανονική κατανομή*.

Όπως ισχύει με όλες τις παραμετρικές μεθόδους, στη *MANOVA* υπολογίζεται η τιμή μιας στατιστικής συνάρτησης ελέγχου και με βάση αυτή την τιμή και την κατανομή που ακολουθεί η συνάρτηση ελέγχου, υπολογίζεται η τιμή p-value. Η στατιστική συνάρτηση ελέγχου με μορφή πίνακα γράφεται γενικά ως

$$\mathbf{A} = \mathbf{H} \cdot \mathbf{E}^{-1} \qquad (12.12)$$

όπου οι πίνακες **H** και **E** περιέχουν στοιχεία που είναι ανάλογα των όρων των αθροισμάτων των σχέσεων (8.1) και (8.2), αντίστοιχα. Από τη σχέση (12.12) προκύπτουν περισσότεροι του ενός έλεγχοι. Οι πιο βασικοί είναι οι έλεγχοι *Pillai, Wilks, Hotelling* και *Roy*, που κατά κανόνα δίνουν πρακτικά ταυτόσημα αποτελέσματα.

Παράδειγμα 12.7

Να εφαρμοστεί η *Ανάλυση Διασποράς Πολλών Μεταβλητών* στα δεδομένα του παραδείγματος 12.1.

◆ Από τη μελέτη του παραδείγματος αυτού με τη μέθοδο *PCA* έχουμε διαπιστώσει ότι τα δείγματα της ομάδας Β διαφοροποιούνται από αυτά των ομάδων Α και C που σχηματίζουν μια ενιαία ομάδα. Έτσι έχει ενδιαφέρον να δούμε αν αυτό το συμπέρασμα επιβεβαιώνεται με την *MANOVA*.

❖ **Ανάλυση στο ChemStat**

Για να εφαρμόσουμε τη *MANOVA* στο *ChemStat* διευθετούμε τα δεδομένα όπως στον πίνακα του σχήματος 12.28. Δηλαδή στην πρώτη στήλη, που τώρα πρέπει να υπάρχει υποχρεωτικά, τοποθετείται μια

μεταβλητή με τιμές που αρχίζουν από 1 και συνεχίζουν 2, 3, …, για να εισάγουμε στο πρόγραμμα τις ομάδες που σχηματίζουν τα δεδομένα των στηλών που είναι δεξιά της πρώτης στήλης. Ακολούθως πηγαίνουμε *Πρόσθετα* → *ChemStat* → *MANOVA* → *MANOVA parametric*. Το πρώτο πλαίσιο που ανοίγει μας πληροφορεί για τη σωστή διευθέτηση των δεδομένων. Πατάμε *OK* και όταν ανοίγει το δεύτερο πλαίσιο κάνουμε *κλικ* στο κελί A2. Στο επόμενο πλαίσιο ορίζουμε το κελί εξόδου των αποτελεσμάτων.

	A	B	C	D	E	F	G	H	I	J	K	L
1	**Location**	**Al**	**Fe**	**Mg**	**Si**	**Ca**		**MANOVA RESULTS**				
2	1	7,1	7,2	1,2	12	5						
3	1	7,8	6,8	0,8	10	5,2		**Homogeneity tests (Levene)**				
4	1	8	7	1,1	11,5	5,8		p (var1):	0,508			
5	1	7,9	7,4	1	10,3	6,2		p (var2):	0,9318			
6	1	8,2	6,6	1	11,9	5		p (var3):	0,2287			
7	2	6,2	6,7	1,2	12,5	6,2		p (var4):	0,074			
8	2	6,9	7,1	0,8	13	5,8		p (var5):	0,1105			
9	2	5,5	7,4	0,5	11,5	6,9		**Wilks' test**				
10	2	7,2	6,7	0,2	11,5	6,2		p-value=	0,0004			
11	2	6,5	6,2	0,4	12,8	6,7		Wilks' L value	0,0485			
12	2	6,9	6,8	0,7	12,1	7		**Pillai's test**				
13	3	7,1	7	1,1	10	6,2		p-value=	0,0023			
14	3	7,5	6,6	1,5	11,1	5,8		Pillai's value=	1,3774			
15	3	6,6	6,2	1,2	11	5,5		**Pairwise comparisons:**				
16	3	8,2	6,7	0,9	10,3	6			Pillai's va	p-value	p-HB	corrected
17	3	8	7,1	1	10,9	6,1		Pair: 1 - 2	8,1848	0,0188	0,038	
18								Pair: 1 - 3	2,1663	0,2369	0,237	
19								Pair: 2 - 3	13,465	0,0063	0,019	

Σχήμα 12.28. Διευθέτηση δειγμάτων και πίνακες αποτελεσμάτων παραμετρικής *MANOVA*

Στους πίνακες αποτελεσμάτων ο πρώτος έλεγχος αφορά την ομοιογένεια της διασποράς (σχήμα 12.28). Παρατηρούμε ότι p-value > 0.05 για όλες τις μεταβλητές και συνεπώς μπορούμε να δεχθούμε ότι υπάρχει ομοιογένεια της διασποράς. Άρα η πρώτη προϋπόθεση για εφαρμογή της μεθόδου ισχύει. Η δεύτερη προϋπόθεση, ο έλεγχος της πολυμεταβλητής κανονικότητας δεν είναι υποχρεωτικός με την προϋπόθεση ότι δεν υπάρχουν πολύ ακραίες τιμές. Τέτοιες τιμές δεν φαίνεται να υπάρχουν στα δεδομένα που αναλύουμε και συνεπώς τα αποτελέσματα του σχήματος 12.28 που αφορούν την εφαρμογή της *MANOVA* φαίνεται να είναι έγκυρα.

Το πρόγραμμα εκτελεί δύο ελέγχους, τον έλεγχο *Wilk* και το έλεγχο

Pillai. Και οι δύο αυτοί έλεγχοι δίνουν τιμές p-value πολύ μικρότερες του 0.05 και συνεπώς μεταξύ των ομάδων 1, 2 και 3 υπάρχουν στατιστικά σημαντικές διαφορές. Για να δούμε μεταξύ ποιών ομάδων υπάρχουν αυτές οι διαφορές, ελέγχουμε τα αποτελέσματα των συγκρίσεων ανά ζεύγη, στην περιοχή K17:K19 στο σχήμα 12.28, όπου υπάρχουν οι τιμές p-value με βάση τη διόρθωση *Holm-Bonferroni*. Παρατηρούμε ότι η διαφορά μεταξύ των ομάδων 1-2 (A-B) και 2-3 (B-C) είναι στατιστικά σημαντική, ενώ η διαφορά μεταξύ των ομάδων 1-3 (A-C) δεν είναι στατιστικά σημαντική. Συνεπώς σε πλήρη συμφωνία με τα αποτελέσματα της *PCA*, στατιστικά σημαντικές διαφορές υπάρχουν μεταξύ των ομάδων 1 και 2 και μεταξύ των 2 και 3.

Στο ίδιο συμπέρασμα καταλήγουμε αν εφαρμόσουμε μη παραμετρική *MANOVA* χρησιμοποιώντας τη μέθοδο *Monte-Carlo με αντιμεταθέσεις*. Ακολουθούμε την ίδια διαδικασία, δηλαδή πηγαίνουμε *ChemStat* → *MANOVA* → *MANOVA non-parametric*, πρέπει όμως στο σχετικό πλαίσιο να ορίσουμε τον αριθμό των επαναλήψεων (permutations) που χρησιμοποιούνται στον αλγόριθμο της μεθόδου. Ο ελάχιστος αριθμός επαναλήψεων πρέπει να είναι 1000. Όμως θα πρέπει να προσέχουμε επειδή η εκτέλεση της μεθόδου απαιτεί σχετικά μεγάλους χρόνους, ιδιαίτερα όταν τα δείγματα είναι μεγάλα.

Τα αποτελέσματα που παίρνουμε δίνονται στο σχήμα 12.29. Το πρόγραμμα εφαρμόζει δύο μεθόδους, εκ των οποίων η μία στηρίζεται στις διαμέσους των δειγμάτων και η άλλη στις μέσες τιμές. Παρατηρούμε ότι οι δύο μέθοδοι δίνουν συγκλίνοντα μεταξύ τους αποτελέσματα και αποτελέσματα που συγκλίνουν με αυτά της παραμετρικής *MANOVA*.

NP MANOVA (with Medians) RESULTS			NP MANOVA (with Means) RESULTS		
Iterations =		5000	Iterations =		5000
All groups	p =	0.007	All groups	p =	0.0066
Pairwise comparisons:			Pairwise comparisons:		
	p-value	p-HB corrected		p-value	p-HB corrected
Pair : 1 - 2	0.004	0.0114	Pair : 1 - 2	0.002	0.006
Pair : 1 - 3	0.171	0.171	Pair : 1 - 3	0.181	0.1814
Pair : 2 - 3	0.004	0.0084	Pair : 2 - 3	0.002	0.004
Elapsed time =		2.859 min	Elapsed time =		3.186 min

Σχήμα 12.29. Πίνακες αποτελεσμάτων μη παραμετρικής *MANOVA*

❖ Ανάλυση στο SPSS

Στο *SPSS* δεν υπάρχει η δυνατότητα εφαρμογής μη-παραμετρικής *MANOVA*. Για να εφαρμόσουμε παραμετρική *MANOVA* διευθετούμε τα δεδομένα όπως στον πίνακα του σχήματος 12.9 και πηγαίνουμε *Analyze* → *Generalized Linear Model* → *Multivariate*, όπου στο παράθυρο που ανοίγει μεταφέρουμε τις μεταβλητές Al, Fe, Mg, Si, Ca στο πλαίσιο *Dependent Variables* και τη μεταβλητή *Location* στο *Fixed Factor(s)*. Από το *Options* επιλέγουμε να γίνει έλεγχος της ομοιογένειας της διασποράς κάνοντας *κλικ* στο *Homogeneity tests* και από το *Model* επιλέγουμε το *Full Factorial* και ενεργοποιούμε το *Include intercept in the model*. Από το *Post Hoc* μπορούμε να επιλέξουμε πολλαπλούς ελέγχους, αλλά αυτοί περιλαμβάνουν και ελέγχους μεταξύ των μεταβλητών και των ομάδων, οδηγώντας σε έναν μάλλον πολύπλοκο πίνακα αποτελεσμάτων.

Levene's Test of Equality of Error Variances[a]

	F	df1	df2	Sig.
Al	,714	2	13	,508
Fe	,071	2	13	,932
Mg	1,656	2	13	,229
Si	3,203	2	13	,074
Ca	2,622	2	13	,110

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.
a. Design: Intercept + Place

Multivariate Tests[c]

Effect		Value	F	Hyp. df	Error df	Sig.
Intercept	Pillai's Trace	,999	3228,486[a]	5,00	9,00	,000
	Wilks' Lambda	,001	3228,486[a]	5,00	9,00	,000
	Hotelling's Trace	1793,60	3228,486[a]	5,00	9,00	,000
	Roy's Largest Root	1793,60	3228,486[a]	5,00	9,00	,000
Location	Pillai's Trace	1,377	4,425	10,0	20,0	,002
	Wilks' Lambda	,049	6,370[a]	10,0	18,0	,000
	Hotelling's Trace	10,825	8,660	10,0	16,0	,000
	Roy's Largest Root	9,942	19,885[b]	5,00	10,0	,000

a. Exact statistic

b. The statistic is an upper bound on F that yields a lower bound on the significance level.
c. Design: Intercept + Location

Σχήμα 12.30. Πίνακες αποτελεσμάτων *MANOVA* στο *SPSS*

Όταν ολοκληρώσουμε τις επιλογές και κάνουμε *κλικ* στο *OK* παίρνουμε αρκετούς πίνακες αποτελεσμάτων, από τους οποίους οι σημαντικότεροι είναι ο πίνακας ελέγχου της ομοιογένειας της διασποράς με το κριτήριο *Levene* και ο πίνακας *Multivariate Tests* που δείχνει αν υπάρχουν στατιστικά σημαντικές διαφορές μεταξύ των μέσων τιμών (σχήμα 12.30). Παρατηρούμε στον πρώτο πίνακα ότι Sig. > 0.05 για όλες τις μεταβλητές και συνεπώς δεν διαπιστώνονται στατιστικά σημαντικές διαφοροποιήσεις στις διασπορές. Στον πίνακα *Multivariate Tests* πηγαίνουμε στο πάνελ *Location* όπου παρατηρούμε ότι όλοι οι έλεγχοι που χρησιμοποιεί το *SPSS* δείχνουν Sig. < 0.02, δηλαδή ότι υπάρχουν στατιστικά σημαντικές διαφορές μεταξύ των ομάδων των δειγμάτων του πίνακα του σχήματος 12.9.

Για να δούμε μεταξύ ποιών ομάδων υπάρχουν αυτές οι διαφορές, εφαρμόζουμε τη *MANOVA* στα δείγματα των ομάδων A-B, B-C και A-C, δηλαδή αφαιρούμε μία ομάδα δειγμάτων κάθε φορά και εφαρμόζουμε *MANOVA* στα υπόλοιπα δείγματα. Τα αποτελέσματα δίνονται στα σχήματα 12.31-12.33. Στο σημείο αυτό θα πρέπει να προσέξουμε να μην αυξηθεί η πιθανότητα λάθους στους ελέγχους αυτούς λόγω πολλαπλών ελέγχων, όπως αναλύθηκε στην ενότητα 8.2.2. Έτσι τοποθετούμε τις τιμές Sig. των πινάκων των σχημάτων 12.31-33 σε φθίνουσα σειρά: Sig.(A-C) = 0.237, Sig. (A-B) = 0.019, Sig. (B-C) = 0.006, και τις πολλαπλασιάζουμε επί 1, 2 και 3, αντίστοιχα. Παίρνουμε τις διορθωμένες τιμές: Sig.(A-C) = 0.237, Sig. (A-B) = 0.038, Sig. (B-C) = 0.018, που ταυτίζονται με τα αποτελέσματα του *ChemStat*.

Multivariate Tests[b]

Effect		Value	F	Hyp. df	Error df	Sig.
Intercept	Pillai's Trace	1,000	2466,486[a]	5,00	5,00	,000
	Wilks' Lambda	,000	2466,486[a]	5,00	5,00	,000
	Hotelling's Trace	2466,486	2466,486[a]	5,00	5,00	,000
	Roy's Largest Root	2466,486	2466,486[a]	5,00	5,00	,000
Location	Pillai's Trace	,891	8,185[a]	5,00	5,00	,019
	Wilks' Lambda	,109	8,185[a]	5,00	5,00	,019
	Hotelling's Trace	8,185	8,185[a]	5,00	5,00	,019
	Roy's Largest Root	8,185	8,185[a]	5,00	5,00	,019

a. Exact statistic
b. Design: Intercept + Location

Σχήμα 12.31. Πίνακας αποτελεσμάτων *MANOVA* για τις ομάδες A-B

Multivariate Tests[b]

Effect		Value	F	Hyp. df	Error df	Sig.
Intercept	Pillai's Trace	1,000	2543,306[a]	5,00	5,00	,000
	Wilks' Lambda	,000	2543,306[a]	5,00	5,00	,000
	Hotelling's Trace	2543,30	2543,306[a]	5,00	5,00	,000
	Roy's Largest Root	2543,30	2543,306[a]	5,00	5,00	,000
Location	Pillai's Trace	,931	13,465[a]	5,00	5,00	,006
	Wilks' Lambda	,069	13,465[a]	5,00	5,00	,006
	Hotelling's Trace	13,465	13,465[a]	5,00	5,00	,006
	Roy's Largest Root	13,465	13,465[a]	5,00	5,00	,006

a. Exact statistic
b. Design: Intercept + Location

Σχήμα 12.32. Πίνακας αποτελεσμάτων *MANOVA* για τις ομάδες B-C

Multivariate Tests[b]

Effect		Value	F	Hyp. df	Error df	Sig.
Intercept	Pillai's Trace	,999	1064,206[a]	5,00	4,00	,000
	Wilks' Lambda	,001	1064,206[a]	5,00	4,00	,000
	Hotelling's Trace	1330,25	1064,206[a]	5,00	4,00	,000
	Roy's Largest Root	1330,25	1064,206[a]	5,00	4,00	,000
Location	Pillai's Trace	,730	2,166[a]	5,00	4,00	,237
	Wilks' Lambda	,270	2,166[a]	5,00	4,00	,237
	Hotelling's Trace	2,708	2,166[a]	5,00	4,00	,237
	Roy's Largest Root	2,708	2,166[a]	5,00	4,00	,237

a. Exact statistic
b. Design: Intercept + Location

Σχήμα 12.33. Πίνακας αποτελεσμάτων *MANOVA* για τις ομάδες A-C

Συνεπώς, σε πλήρη συμφωνία με τα αποτελέσματα της *PCA* και της *CA*, στατιστικά σημαντικές διαφορές υπάρχουν κυρίως μεταξύ των ομάδων B και C και μεταξύ των B και A.

ΑΣΚΗΣΕΙΣ

12.1. Στον παρακάτω πίνακα δίνεται η συγκέντρωση σε mg/kg τεσσάρων στοιχείων σε δείγματα ρυζιού που ανήκουν σε δύο ποικιλίες (A, B), δύο τύπους (P=polished, U=unpolished) και έχουν καλλιεργηθεί σε τρεις περιοχές A, B, C. Να εξετασθεί αν υπάρχουν διαφοροποιήσεις (ομάδες) ανάλογα με την ποικιλία ή τον τύπο ή την περιοχή.

Πίνακας 12.11. Πίνακας δεδομένων άσκησης 12.1.

Ποικιλία	Τύπος	Περιοχή	Mg	Ca	P	K
A	U	A	650	60	2382	2084
A	U	A	629	58	2234	1955
A	U	A	591	54	2099	1836
A	P	B	264	24	938	821
A	P	B	268	24	952	833
A	P	B	255	23	908	794
A	U	B	509	47	1810	1583
A	P	B	211	19	750	656
B	U	C	739	68	2627	2298
B	U	C	632	58	2244	1963
B	P	C	466	43	1657	1450
B	P	A	359	33	1276	1117
B	U	C	725	66	2576	2254
B	U	C	736	67	2615	2288
B	P	C	439	40	1560	1365
B	P	A	345	31	1224	1071

12.2. Να εφαρμοστεί η *Ανάλυση σε Ομάδες* για να προσδιοριστούν οι ποικιλίες οίνου που υπάρχουν στον πίνακα 12.12 και ακολούθως να χρησιμοποιηθούν οι υπόλοιπες τεχνικές της *Ανάλυσης Πολλών Μεταβλητών* για να ελεγχθούν τα αποτελέσματα.

12.3. Στον πίνακα 12.13 δίνονται η μέγιστη (T1) και η ελάχιστη (T2) θερμοκρασία του αέρα, η μέγιστη (T3) και η ελάχιστη (T4) θερμοκρασία του εδάφους, η μέγιστη (H1) και η ελάχιστη (H2) υγρασία 15 πόλεων. Να εφαρμοστούν οι διάφορες τεχνικές της *Ανάλυσης Πολλών Μεταβλητών* για να προσδιοριστούν ομάδες πόλεων με παρόμοια χαρακτηριστικά.

Πίνακας 12.12. Πίνακας δεδομένων άσκησης 12.2.

Ποικιλία οίνου	Στάχτη g/L	Μαλικό οξύ g/L	Mg g/L	Φαινόλες g/L	Αλκοόλη %v/v
1	2.52	1.79	96	2.6	14.24
2	2.08	1.87	90	3.1	13.21
3	2.21	2.10	118	2.8	14.08
4	2.13	1.27	100	2.3	14.34
5	2.26	1.95	93	2.5	14.52
6	2.40	1.33	127	3.3	14.22
7	2.53	1.87	94	2.7	14.78
8	2.60	2.21	114	3.0	13.72
9	2.14	1.97	102	2.3	13.76
10	2.29	2.11	103	2.8	14.57
11	2.00	2.00	100	3.0	14.50
12	1.91	1.57	109	2.5	13.47
13	2.22	1.65	91	2.7	13.94
14	2.46	1.66	99	2.6	14.57
15	2.34	2.01	118	2.7	13.92

Πίνακας 12.13. Πίνακας δεδομένων άσκησης 12.3.

Πόλη	T1	T2	T3	T4	H1	H2
1	28	18	29	15	90	40
2	28	18	30	16	90	30
3	26	18	28	17	90	40
4	27	19	28	18	90	50
5	28	20	31	20	85	45
6	23	18	25	19	85	70
7	22	18	25	20	85	70
8	23	19	28	20	85	70
9	28	20	31	21	85	63
10	30	22	32	24	93	63
11	31	22	32	24	93	55
12	32	23	34	24	93	50
13	31	22	34	23	93	55
14	24	18	30	18	95	40
15	27	20	30	20	95	60

12.4. Για να εξετασθεί η επίδραση της γεωγραφικής περιοχής και έμμεσα οι διατροφικές συνήθειες στα επίπεδα λιπιδίων στο αίμα, προσδιορίστηκαν οι συγκεντρώσεις σε mg/L της κακής (LDL) και της καλής (HDL) χοληστερόλης και των τριγλυκεριδίων ατόμων από τέσσερες περιοχές. Τα αποτελέσματα δίνονται στους επόμενους δύο πίνακες. Να ενοποιηθούν οι πίνακες και να εξεταστούν με όλες τις δυνατές τεχνικές της *Ανάλυσης Πολλών Μεταβλητών*.

Ορεινή περιοχή			Ηπειρωτική περιοχή		
HDL	LDL	τριγλυκερίδια	HDL	LDL	τριγλυκερίδια
40	190	220	45	160	170
35	181	280	42	175	155
44	200	250	45	180	200
46	175	195	50	160	180
52	180	190	45	185	160

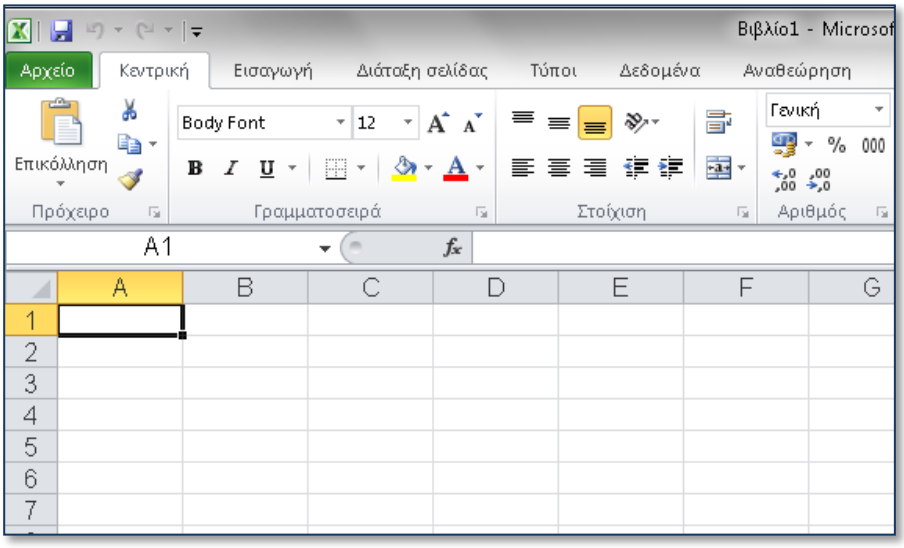
Παραθαλάσσια περιοχή			Νησιά		
HDL	LDL	τριγλυκερίδια	HDL	LDL	τριγλυκερίδια
55	100	145	60	75	120
45	120	160	55	95	145
65	80	95	42	110	150
60	95	100	48	130	110
55	110	110	55	110	100

Παράρτημα Ι

ΕΙΣΑΓΩΓΗ ΣΤΟ EXCEL

Ι.1 ΦΥΛΛΑ ΕΡΓΑΣΙΑΣ ΤΟΥ EXCEL

Όταν ανοίγουμε την ελληνική έκδοση του προγράμματος *Excel* 2010, στην οθόνη του υπολογιστή εμφανίζεται η εικόνα του σχήματος Ι.1. Τα κύρια στοιχεία που βλέπουμε στην εικόνα αυτή είναι τα ακόλουθα:



Σχήμα Ι.1. Τμήμα αρχικής οθόνης του *Excel*

Στο επάνω μέρος της οθόνης υπάρχει η *γραμμή τίτλου (title bar)*, που στα αριστερά της υπάρχει η *γραμμή εργαλείων γρήγορης πρόσβασης (quick access toolbar)*. Σε αυτήν υπάρχουν, με μορφή εικόνας, εντολές για να κάνουμε γρήγορα διάφορες ενέργειες στα δεδομένα του φύλλου

εργασίας. Στο σχήμα Ι.1 το εικονίδιο της δισκέτας είναι για να αποθηκεύουμε το βιβλίο εργασίας, δηλαδή το αρχείο του *Excel*, ενώ τα δύο άλλα εικονίδια είναι για να αναιρέσουμε ή να μην αναιρέσουμε μια ενέργεια που έχουμε κάνει στο φύλλο εργασίας. Δεξιά της γραμμής εργαλείων γρήγορης πρόσβασης βρίσκεται το εικονίδιο που δείχνει ένα βέλος προς τα κάτω,  . Με αυτό μπορούμε να προσθέσουμε περισσότερα εικονίδια στη γραμμή αυτή. Στην έκδοση *Excel* του 2007 και στο αριστερό άκρο της *γραμμής τίτλου* υπάρχει το εικονίδιο *Κουμπί Office* (*Office Button)*  , από το οποίο ανοίγει ένας κατάλογος εντολών που επιτρέπουν το άνοιγμα άλλων αρχείων, την αποθήκευση ή την εκτύπωση δεδομένων κ.ο.κ. Στην έκδοση *Excel 2010* οι ενέργειες αυτές γίνονται από το εικονίδιο *Αρχείο (File*) 

.

Κάτω από τη γραμμή τίτλου υπάρχει η *γραμμή μενού (menu bar)*, η οποία περιέχει τις λέξεις: *Αρχείο, Κεντρική, Εισαγωγή, Διάταξη σελίδας, Τύποι, Δεδομένα, Αναθεώρηση, Προβολή, Προγραμματιστής, Πρόσθετα (File, Home, Insert, Page Layout, Formulas, Data, Review, View, Developer, Add-Ins)*. Αν στην έκδοση του 2007 δεν υπάρχει η λέξη *Προγραμματιστής (Developer)*, την εισάγουμε με την ακόλουθη διαδικασία: Κάνουμε *κλικ* στο *Κουμπί Office*, *κλικ* στο *Επιλογές του Excel (Excel Options)* και επιλέγουμε *Εμφάνιση καρτέλας "Προγραμματιστής" στην κορδέλα (Show Developer tab in the Ribbon)*. Με *κλικ* στο *ΟΚ* η καρτέλα *Προγραμματιστής* εμφανίζεται στη γραμμή μενού. Όταν κάνουμε *κλικ* σε μία από τις λέξεις που υπάρχουν στη γραμμή μενού ανοίγει μια καρτέλα με μορφή λωρίδας ή κορδέλας με εικονίδια που εκτελούν συγκεκριμένες εντολές. Αν φέρουμε το δείκτη σε ένα από τα εικονίδια αυτά, μετά από κάποια δευτερόλεπτα εμφανίζεται μια μικρή πινακίδα, που με μορφή τίτλου μας πληροφορεί τι κάνει το συγκεκριμένο εικονίδιο. Παρατηρούμε επίσης ότι τα εικονίδια είναι ομαδοποιημένα ανάλογα με τις εντολές που εκτελούν μέσα σε πλαίσια, η ονομασία των οποίων υπάρχει στο κάτω μέρος κάθε πλαισίου.

Η *γραμμή τύπων* (*formula bar*) που ακολουθεί κάτω από τη λωρίδα των εικονιδίων αποτελείται από δύο πλαίσια, ένα μικρό αριστερά που ονομάζεται *πλαίσιο ονόματος* και πιθανόν να περιέχει τη λέξη Α1 και ένα μεγάλο δεξιά που είναι κενό. Σε αυτό εμφανίζεται το περιεχόμενο κάθε κελιού το οποίο επιλέγουμε με το ποντίκι.

Κάτω από τη γραμμή τύπων υπάρχει το κυρίως *φύλλο εργασίας* (*woksheet*). Αποτελείται από ένα μεγάλο αριθμό κενών *κελιών* (*cells*), που σχηματίζουν έναν δισδιάστατο πίνακα. Ο πίνακας αυτός στα αριστερά του καθώς επίσης και στο επάνω μέρος του περιβάλλεται από μια σειρά

ορθογώνιων παραλληλογράμμων με λατινικά γράμματα ή αριθμούς. Τα γράμματα και οι αριθμοί επιτρέπουν να ορίζουμε επακριβώς τις συντεταγμένες κάθε κελιού.

Σε ένα φύλλο εργασίας μπορούμε να εισάγουμε δεδομένα, να εκτελέσουμε μαθηματικές ή λογικές πράξεις και τέλος να παρουσιάσουμε τα τελικά αποτελέσματα με τη μορφή πινάκων ή/και διαγραμμάτων. Συνεπώς, τα φύλλα εργασίας επιτρέπουν την άμεση και εύκολη ανάλυση των δεδομένων ενός πειράματος.

Στην κάτω γραμμή και αριστερά υπάρχουν πινακίδες με τις λέξεις *Φύλλο1 (Sheet1*), *Φύλλο2 (Sheet2*), …. Αν κάνουμε απλό *κλικ* στο *Φύλλο2*, τότε μετακινούμαστε από το *Φύλλο1* στο *Φύλλο2*. Τέλος, αν κάνουμε *κλικ* στο εικονίδιο *Εισαγωγή φύλλου εργασίας* (*Insert Worksheet*),  , που υπάρχει δεξιά της τελευταίας πινακίδας, μπορούμε να εισάγουμε νέα φύλλα. Έτσι *ένα αρχείο του Excel είναι ένα βιβλίο με πολλά φύλλα, στα οποία μπορούμε να αναλύσουμε τα δεδομένα μας*.

Ι.2 ΚΕΛΙΑ ΚΑΙ ΠΕΡΙΟΧΕΣ

Όπως αναφέρθηκε, το κυρίως φύλλο εργασίας αποτελείται από ένα μεγάλο αριθμό κελιών. Κάθε *κελί* (*cell*) ταυτοποιείται από τις συντεταγμένες του (τη διεύθυνσή του) που είναι ένα λατινικό γράμμα συνοδευόμενο από έναν αριθμό: το γράμμα της στήλης και τον αριθμό της γραμμής που ανήκει. Ένα κελί γίνεται *ενεργό* αν κάνουμε *κλικ* επάνω σε αυτό. Τότε το κελί περιβάλλεται από ένα πλαίσιο και η διεύθυνσή του εμφανίζεται στο αριστερό παράθυρο της γραμμής τύπων.

Μια συλλογή από κελιά ονομάζεται *περιοχή* (*range*). Έστω ότι μια περιοχή περιλαμβάνει τα κελιά από A3 μέχρι A6, από B4 μέχρι B6 και από D1 μέχρι D7. Η περιοχή αυτή συμβολίζεται με A3:A6;B4:B6;D1:D7. Μπορεί όμως, ανάλογα με τις ρυθμίσεις του υπολογιστή η ίδια περιοχή να συμβολίζεται με A3:A6,B4:B6,D1:D7. Ο συμβολισμός απλοποιείται όταν η περιοχή είναι ορθογώνια. Έτσι η περιοχή που περιλαμβάνει τα κελιά από A2 μέχρι A7, από B2 μέχρι B7 και από C2 μέχρι C7 συμβολίζεται με A2:C7. Η επιλογή μιας περιοχής μπορεί να γίνει ως εξής:

Έστω ότι θέλουμε να επιλέξουμε την περιοχή A2:C7. Κάνουμε *κλικ* στο A2 και χωρίς να ελευθερώσουμε το πλήκτρο του ποντικιού φέρνουμε το δείκτη του στο C7. Στο σημείο αυτό ελευθερώνουμε το πλήκτρο του ποντικιού. Όλη η περιοχή, με εξαίρεση το πρώτο κελί, θα αποκτήσει ένα απαλό μοβ φόντο, δείχνοντας ότι έχει επιλεγεί. Εναλλακτικά, κάνουμε απλό *κλικ* στο A2, πατάμε το πλήκτρο *Schift* στο πληκτρολόγιο και έχοντας το

πλήκτρο αυτό συνεχώς πατημένο φέρνουμε το δείκτη του ποντικιού στο C7 όπου και ξανακάνουμε *κλικ*. Ελευθερώνοντας το πλήκτρο *Schift* η περιοχή A2:C7 έχει επιλεγεί. Πιο πολύπλοκες περιοχές μπορούν να επιλεγούν με τη βοήθεια των πλήκτρων *Schift* και *Ctrl*.

Σε ένα κελί μπορούμε να εισάγουμε μια λέξη, έναν αριθμό ή ένα μαθηματικό τύπο. Για να εισάγουμε μια λέξη ή έναν αριθμό απλά πληκτρολογούμε τη λέξη ή τον αριθμό στο πληκτρολόγιο, αφού πρώτα έχουμε επιλέξει το κελί. Η εισαγωγή ολοκληρώνεται πατώντας *Enter* ή κάνοντας *κλικ* σε ένα οποιοδήποτε άλλο κελί ή τέλος με *κλικ* στο εικονίδιο ✓ που εμφανίζεται στη γραμμή τύπων. Για να ξεχωρίζουν οι λέξεις από τους αριθμούς, στα κελιά οι λέξεις στοιχίζονται αριστερά και οι αριθμοί δεξιά.

Για να εισάγουμε έναν μαθηματικό τύπο, πρώτα πληκτρολογούμε το = και ακολούθως τον μαθηματικό τύπο. Για παράδειγμα, αν στο A2 θέλουμε να κάνουμε την πράξη (1.8×3.14)/2.25, πρώτα επιλέγουμε το κελί A2, ακολούθως πληκτρολογούμε την έκφραση =1.8*3.14/2.25 και τέλος πατάμε *Enter* ή κάνουμε *κλικ* στο εικονίδιο ✓ που υπάρχει στη γραμμή τύπων. Ο αριθμός 2.512 θα σχηματιστεί στο A2.

Η εισαγωγή αριθμών ή λέξεων σε ένα κελί ολοκληρώνεται πατώντας *Enter* ή με *κλικ* στο εικονίδιο ✓ ή με *κλικ* σε ένα οποιοδήποτε άλλο κελί του φύλλου εργασίας. Για να ολοκληρωθεί όμως η εισαγωγή ενός μαθηματικού τύπου, υπάρχει μια σημαντική διαφοροποίηση. Μπορούμε να πατήσουμε *Enter* ή να κάνουμε *κλικ* στο εικονίδιο ✓, δεν κάνουμε όμως *κλικ* σε ένα άλλο κελί. Αν προχωρήσουμε σε αυτή την ενέργεια, τότε το κελί αυτό (η διεύθυνσή του) εισέρχεται στον τύπο που πληκτρολογούμε.

Παρατήρηση 1. Στο *Excel*, ανάλογα με τις ρυθμίσεις του υπολογιστή (*Πίνακας ελέγχου* → *Τοπικές ρυθμίσεις*), μπορεί να χρησιμοποιείται ή η τελεία ή η υποδιαστολή στους δεκαδικούς αριθμούς. Αν έχει επιλεγεί η τελεία, τότε το *Excel* θεωρεί το 1,8 (με κόμμα) ως λέξη και δεν επιτρέπει μαθηματικές πράξεις. Επίσης σε αριθμούς με εκθετική μορφή χρησιμοποιείται το E αντί για το 10 και ακολουθεί ο εκθέτης γραμμένος κανονικά. Έτσι, ο αριθμός 1.6×10^{-19} γράφεται ως 1.6E-19.

Παρατήρηση 2. Όταν πληκτρολογούμε μια λέξη, έναν αριθμό ή μια μαθηματική παράσταση σε ένα κελί, στη γραμμή τύπων εμφανίζονται τρία εικονίδια, ✕ ✓ f_x . Το εικονίδιο f_x χρησιμοποιείται για να εισάγουμε μια συνάρτηση, το ✓ είναι ισοδύναμο με το *Enter*, ενώ το ✕ χρησιμοποιείται όταν θέλουμε να αναιρέσουμε διορθώσεις που έχουμε κάνει στον μαθηματικό τύπο.

Ι.3 ΧΡΗΣΗ ΤΟΥ ΦΥΛΛΟΥ ΕΡΓΑΣΙΑΣ ΓΙΑ ΥΠΟΛΟΓΙΣΜΟΥΣ

Όπως ήδη έχουμε αναφέρει, ένα φύλλο εργασίας μπορεί να χρησιμοποιηθεί ως αριθμομηχανή, αρκεί σε ένα κελί να εισάγουμε τη μαθηματική παράσταση που θέλουμε να υπολογίσουμε. Στο *Excel* οι πέντε βασικές πράξεις, τα αντίστοιχα σύμβολά τους και η σειρά εκτέλεσής τους είναι:

Πράξη	**Σύμβολο**	**Προτεραιότητα**
Δύναμη	^	1
Πολλαπλασιασμός	*	2
Διαίρεση	/	2
Πρόσθεση	+	3
Αφαίρεση	-	3

Η σειρά εκτέλεσης των πράξεων ισχύει μόνο όταν δεν υπάρχουν παρενθέσεις. Έτσι, αν επιλέξουμε το Α1 και πληκτρολογήσουμε τη μαθηματική παράσταση =4*3^2-6, με *Enter* ή *κλικ* στο εικονίδιο ✓ θα πάρουμε το αποτέλεσμα 30, δεδομένου ότι πρώτα εκτελείται η πράξη 3^2 = 9, ακολούθως η 4*9 = 36 και τέλος η αφαίρεση 36-6 = 30.

Αν υπάρχουν παρενθέσεις, τότε πρώτα γίνονται οι πράξεις μέσα στις παρενθέσεις με τη σειρά εκτέλεσης που αναφέρεται στον παραπάνω πίνακα, μέχρι να προκύψει παράσταση χωρίς παρενθέσεις. Έτσι, αν στο Α1 πληκτρολογήσουμε τη μαθηματική παράσταση =(4*3)^2-6, με *Enter* θα πάρουμε το αποτέλεσμα 138, επειδή πρώτα εκτελείται η πράξη 4*3 = 12, ακολούθως η 12^2 = 144 και τέλος η αφαίρεση 144-6 = 138. Αν υπάρχουν πολλαπλές παρενθέσεις ο υπολογιστής αρχίζει τις πράξεις από τις εσωτερικές παρενθέσεις προς τα έξω.

Τέλος, για πράξεις με την ίδια προτεραιότητα, ο υπολογιστής τις αρχίζει από αριστερά προς τα δεξιά. Έτσι η παράσταση =2/3*6 δίνει το αποτέλεσμα 4, δεδομένου ότι πρώτα γίνεται η διαίρεση 2/3 και το αποτέλεσμα της διαίρεσης πολλαπλασιάζεται επί 6.

Παράδειγμα Ι.1

Έστω ότι θέλουμε να υπολογίσουμε την τιμή των παραστάσεων

α) $\frac{1+2.7^2}{1-2.7/4}$ β) $\frac{1+3.14}{1+\frac{1}{1+3.14}}$ γ) $3^{2\cdot 2.7}$

◆ Όταν έχουμε να υπολογίσουμε κλάσματα καλό είναι να βάζουμε τόσο τον αριθμητή όσο και τον παρανομαστή μέσα σε παρενθέσεις και προφανώς το σύμβολο της διαίρεσης "/" ανάμεσά τους. Επίσης αν υπάρχουν εκθέτες με μαθηματικές πράξεις, τους βάζουμε μέσα σε παρένθεση. Έτσι έχουμε:

α) Η παράσταση μπορεί να γραφεί ως $(1 + 2.7^2)/(1 – 2.7/4)$ και συνεπώς σε κάποιο κελί πληκτρολογούμε τον τύπο =(1+2.7^2)/(1-2.7/4). Παίρνουμε το αποτέλεσμα: 25.507692.

β) Ο παρανομαστής μπορεί να γραφεί ως 1 + 1 / (1 + 3.14) και επομένως η παράσταση γράφεται ως (1 + 3.14) / (1 + 1 / (1 + 3.14)), που είναι και η τελική έκφραση που πρέπει να πληκτρολογήσουμε. Ισούται με 3.334553.

γ) Εδώ πρέπει να πληκτρολογήσουμε =3^(2*2.7). Ο υπολογιστής πρώτα υπολογίζει την ποσότητα μέσα στην παρένθεση, 2*2.7 = 5.4, και μετά υψώνει στη δύναμη, 3^5.4 = 377.09847.

Εκτός από τις πέντε βασικές πράξεις, το *Excel* διαθέτει μια πλούσια βιβλιοθήκη έτοιμων συναρτήσεων. Οι πιο χρήσιμες από αυτές δίνονται στον πίνακα Ι.1. Για να χρησιμοποιήσουμε μια συνάρτηση του *Excel* αρκεί να την πληκτρολογήσουμε.

Πίνακας Ι.1. Βασικές μαθηματικές συναρτήσεις του *Excel*.

Συνάρτηση	Ορισμός - περιγραφή
ABS(x)	= |x|
DEGREES(x)	Μετατρέπει το x από ακτίνια σε βαθμούς
EXP(x)	= exp(x) = e^x
FACT(n)	= n!
INT(x)	Στρογγυλεύει τον αριθμό x στο μικρότερο ακέραιο. Π.χ. INT(1.8) = 1, INT(-1.8) = -2
LN(x)	= ln(x)
LOG10(x)	Υπολογίζει το δεκαδικό λογάριθμο του x
MOD(x;y)	Υπολογίζει το υπόλοιπο της διαίρεσης x/y
PI()	= π
RADIANS(x)	Μετατρέπει το x από βαθμούς σε ακτίνια
ROUND(x;n)	Στρογγυλεύει τον αριθμό x σε n δεκαδικά ψηφία Π.χ. ROUND(3.14159;2) = 3.14
SQRT(x)	= $\sqrt{x}$

Παρατήρηση. Στο *Excel* δεν έχει σημασία αν ο τύπος γράφεται με κεφαλαία ή μικρά γράμματα.

ΠΡΟΣΟΧΗ. Ανάλογα με τις ρυθμίσεις, οι μεταβλητές στους τύπους χωρίζονται ή με το ερωτηματικό, όπως ROUND(x;n) ή με κόμμα, ROUND(x,n). Όταν πληκτρολογούμε έναν τύπο, εμφανίζεται ένα πλαίσιο που μας ενημερώνει για τα ορίσματα του τύπου και τον τρόπο που αυτά διαχωρίζονται.

Παράδειγμα Ι.2

Έστω ότι θέλουμε να υπολογίσουμε την παράσταση

$$\frac{\ln|\pi - 4|}{2 + \frac{e}{(1+2\pi)^3}}$$

◆ Από τις συναρτήσεις του πίνακα Ι.1 προκύπτει ότι στο *Excel* η απόλυτη τιμή της διαφοράς π - 4 εκφράζεται ως ABS(PI()-4) και η συνάρτηση ln είναι η LN. Συνεπώς ο αριθμητής πρέπει να γραφεί ως LN(ABS(PI()-4)). Σε ότι αφορά τον παρανομαστή, αυτός γράφεται ως 2+EXP(1)/((1+2*PI())^3). Η έκφραση αυτή πρέπει να μπει σε παρένθεση και να διαιρεθεί με τον αριθμητή. Επομένως για τον υπολογισμό του αρχικού κλάσματος επιλέγουμε ένα κελί και πληκτρολογούμε την έκφραση

=LN(ABS(PI()-4))/(2+EXP(1)/((1+2*PI())^3))

Πατώντας *Enter* ή με *κλικ* στο εικονίδιο ✓ παίρνουμε –0.07607.

Σε ένα κελί μπορούμε να εισάγουμε μια συνάρτηση του *Excel* χωρίς να την πληκτρολογήσουμε ως εξής: Επιλέγουμε πρώτα το κελί στο οποίο θα εισαχθεί και ακολούθως κάνουμε *κλικ* στο εικονίδιο *Insert Function* f_x που υπάρχει στη λωρίδα *Τύποι (Formulas)*. Εναλλακτικά κάνουμε *κλικ* στο εικονίδιο f_x στη γραμμή τύπων. Τότε ανοίγει ένα παράθυρο διαλόγου που περιέχει το σύνολο των συναρτήσεων του *Excel*, ομαδοποιημένες σε *Μαθηματικές & Τριγωνομετρικές* (*Math & Trig*), *Στατιστικές* (*Statistical*), *Λογικές* (*Logical*) κ.ο.κ. Αν επιλέξουμε μια από αυτές με απλό *κλικ*, τότε στο ίδιο παράθυρο βλέπουμε μια σύντομη περιγραφή των ιδιοτήτων της. Με *κλικ* στο *OK* ανοίγει ένα νέο παράθυρο διαλόγου για να εισάγουμε το όρισμα της συνάρτησης. Στο παράθυρο αυτό βλέπουμε επίσης μια πιο αναλυτική περιγραφή της συνάρτησης.

Ι.4 Η ΔΙΑΔΙΚΑΣΙΑ ΤΗΣ ΑΥΤΟΜΑΤΗΣ ΣΥΜΠΛΗΡΩΣΗΣ

Όταν έχουμε έναν πίνακα πειραματικών δεδομένων είμαστε πολλές φορές υποχρεωμένοι να επαναλάβουμε την ίδια σειρά πράξεων με όλες τις τιμές του πίνακα, γεγονός που είναι και κουραστικό και χρονοβόρο. Για να ξεπεραστεί αυτό το πρόβλημα στο *Excel* υπάρχει μια διαδικασία, που ονομάζεται *Αυτόματη Συμπλήρωση* (*AutoFill*). Η διαδικασία αυτή γίνεται εύκολα κατανοητή αν εξετάσουμε τα ακόλουθα απλά παραδείγματα:

Παράδειγμα Ι.3

Να δημιουργηθεί μια στήλη με τιμές από το 0.5 στο 10 με βήμα 0.1 και στη διπλανή στήλη να υπολογιστεί το τετράγωνο αυτών των τιμών.

◆ Σε ένα φύλλο εργασίας, στο κελί Α1 εισάγουμε τον τίτλο x και στο κελί Β1 τον τίτλο $y=x^2$. Πληκτρολογούμε στο κελί Α2 την πρώτη τιμή 0.5 και κάνουμε *κλικ* επάνω σε αυτή την τιμή. Ακολούθως από το *Κεντρική (Home)* → *Επεξεργασία (Editing)* κάνουμε *κλικ* στο εικονίδιο *Συμπλήρωση (Fill)* και μετά στην εντολή *Σειρά (Series)*. Στο παράθυρο διαλόγου που ανοίγει επιλέγουμε *Στήλες (Columns), Αριθμητική πρόοδος (Linear)* και εισάγουμε τις τιμές 0.1 ως *Τιμή βήματος (Step Value)* και 10 ως *Τελική τιμή (Stop value)*. Με *κλικ* στο *OK* παρατηρούμε ότι η περιοχή Α2:Α97 γεμίζει με τις τιμές που θέλουμε. Ακολούθως στο κελί Β2 πληκτρολογούμε τον τύπο =A2^2 και πατάμε *Enter*. Στο Β2 σχηματίζεται η τιμή του x^2 = 0.25 που αντιστοιχεί σε x = 0.5. Με προσοχή φέρνουμε το δείκτη του ποντικιού στην κάτω και δεξιά γωνία του κελιού Β2. Παρατηρούμε ότι ο δείκτης μετατρέπεται σε ένα μαύρο σταυρό. Πιέζουμε το αριστερό πλήκτρο του ποντικιού και κρατώντας το πατημένο φέρνουμε το δείκτη στο Β97. Αν τώρα ελευθερώσουμε το πλήκτρο του ποντικιού, η περιοχή Β2:Β97 γεμίζει με τιμές, που είναι οι τιμές του x^2. Η διαδικασία αυτή ονομάζεται *αυτόματη συμπλήρωση*.

Εναλλακτικά η διαδικασία της αυτόματης συμπλήρωσης μπορεί να εφαρμοστεί ως εξής: Όταν στην κάτω και δεξιά γωνία του κελιού Β2 ο δείκτης του ποντικιού μετατραπεί σε μαύρο σταυρό κάνουμε διπλό *κλικ*, οπότε η περιοχή Β2:Β97 γεμίζει με τιμές του x^2.

Παράδειγμα Ι.4

Στο προηγούμενο παράδειγμα να δημιουργηθεί μια νέα στήλη με τιμές α/x, όπου η τιμή της σταθεράς α θα είναι στο κελί Ε1. Για δοκιμή να δοθεί η τιμή α = π.

◆ Στο κελί D1 πληκτρολογούμε τον τίτλο a = και στο E1 εισάγουμε την τιμή του π πληκτρολογώντας =PI(). Ακολούθως στο C1 εισάγουμε τον τίτλο z, στο C2 πληκτρολογούμε τον τύπο =E1/A2, πατάμε *Enter* και εφαρμόζουμε τη διαδικασία της αυτόματης συμπλήρωσης. Δηλαδή, επιλέγουμε το κελί C2, φέρνουμε τον κέρσορα στην κάτω και δεξιά γωνία αυτού του κελιού και όταν αυτός μετατραπεί σε μαύρο σταυρό κάνουμε διπλό *κλικ*. Παρατηρούμε ότι ενώ η πρώτη τιμή είναι σωστή, 6.283, όλες οι υπόλοιπες τιμές είναι 0.

Η αιτία αυτού του λάθους φαίνεται αμέσως αν κάνουμε *κλικ* στο C3. Τότε στη γραμμή τύπων εμφανίζεται ο μαθηματικός τύπος =E2/A3, που είναι εσφαλμένος, επειδή αντί για E1 έχει E2 που είναι 0. Συνεπώς για να πάρουμε σωστά αποτελέσματα θα πρέπει το κελί E1 ή ορθότερα η διεύθυνσή του, να παραμένει αμετάβλητη κατά τη διαδικασία της αυτόματης συμπλήρωσης. Για να γίνει αυτό θα πρέπει η διεύθυνση του κελιού να συμβολίζεται κάπως διαφορετικά, ώστε το πρόγραμμα να τη ξεχωρίζει από κάποιο κελί που μεταβάλλεται κατά τη διαδικασία της αυτόματης συμπλήρωσης. Στο *Excel* για να δηλώσουμε ότι ένα κελί δεν μεταβάλλεται στη διαδικασία της αυτόματης συμπλήρωσης πρέπει να γράψουμε δεξιά και αριστερά του γράμματος που ορίζει τη στήλη στην οποία βρίσκεται το σύμβολο $. Έτσι στο παράδειγμα που εξετάζουμε, αντικαθιστούμε στο μαθηματικό τύπο το E1 με E1. Η διεύθυνση αυτή του κελιού ονομάζεται *απόλυτη*. Αν κάνουμε αυτή τη διόρθωση, θα πάρουμε τελικά τις σωστές τιμές του a/x στην περιοχή C2:C97.

Παρατήρηση. Στην απόλυτη διεύθυνση E1 το σύμβολο $ πριν από το E σημαίνει ότι η στήλη E παραμένει σταθερή και το σύμβολο $ πριν από το 1 σημαίνει ότι η γραμμή 1 παραμένει σταθερή. Συνεπώς αν χρησιμοποιηθεί ο επιμέρους συμβολισμός $E1, τότε κατά τη διαδικασία της αυτόματης συμπλήρωσης παραμένει σταθερή η στήλη E ενώ μπορεί να μεταβάλλεται η γραμμή. Τέλος, με E$1 κρατάμε σταθερή τη γραμμή 1 και μπορούμε να μεταβάλλουμε τη στήλη, εφόσον εφαρμόζουμε τη διαδικασία της αυτόματης συμπλήρωσης οριζοντίως.

Σημείωση. Όταν πληκτρολογούμε έναν τύπο που περιέχει κελιά, π.χ. =A1^2, δε χρειάζεται να γράψουμε A1. Για ευκολία πληκτρολογούμε το =, μετά κάνουμε *κλικ* στο κελί A1 και συνεχίζουμε πληκτρολογώντας ^2. Αυτό ισχύει γενικά, δηλαδή όταν πληκτρολογούμε ένα μαθηματικό τύπο και κάνουμε *κλικ* σε ένα κελί ή επιλέξουμε με το ποντίκι μια περιοχή, τότε το κελί ή η περιοχή εισέρχεται στο μαθηματικό τύπο.

Ι.5 ΠΡΑΞΕΙΣ ΜΕ ΣΤΗΛΕΣ ΔΕΔΟΜΕΝΩΝ

Στα μεγάλα προτερήματα του *Excel* είναι και η δυνατότητα που έχει να κάνει εύκολα πράξεις με στήλες δεδομένων. Εκτός από τη διαδικασία της αυτόματης συμπλήρωσης που επιτρέπει τέτοιου είδους πράξεις, υπάρχουν ειδικές συναρτήσεις με όρισμα μια περιοχή κελιών. Για παράδειγμα, τέτοιες συναρτήσεις είναι οι *AVERAGE*, *STDEV*, *SUM*, *PRODUCT* και *COUNT* που υπολογίζουν τη μέση τιμή, την τυπική απόκλιση, το άθροισμα, το γινόμενο και το πλήθος, αντίστοιχα, των τιμών μιας περιοχής καθώς επίσης και οι *MAX* και *MIN* που υπολογίζουν τη μέγιστη και ελάχιστη τιμή που υπάρχει σε μια περιοχή.

Έστω ότι θέλουμε να υπολογίσουμε το άθροισμα των τιμών που υπάρχουν στην περιοχή A2:A97 του προηγούμενου παραδείγματος και το αποτέλεσμα να εμφανιστεί στο A100. Κάνουμε *κλικ* στο A100, πληκτρολογούμε =SUM(A2:A97) και πατάμε *Enter*. Το αποτέλεσμα, 504, θα εμφανιστεί αμέσως. Εναλλακτικά, μπορούμε να γράψουμε μόνο =SUM(και με το δείκτη του ποντικιού να επιλέξουμε την περιοχή A2:A97. Η περιοχή αυτή θα περάσει αυτόματα στο όρισμα της συνάρτησης και, χωρίς να είναι απαραίτητο να κλείσουμε την παρένθεση, με *Enter* παίρνουμε το τελικό αποτέλεσμα.

Ι.6 ΠΡΑΞΕΙΣ ΜΕ ΠΙΝΑΚΕΣ

Για πράξεις με πίνακες το *Excel* έχει τις ακόλουθες συναρτήσεις: *MDETERM* που υπολογίζει την ορίζουσα ενός τετραγωνικού πίνακα, *MINVERSE* για τον υπολογισμό του αντίστροφου πίνακα και *MMULT* για τον πολλαπλασιασμό δύο πινάκων. Το κύριο χαρακτηριστικό των συναρτήσεων *MINVERSE* και *MMULT* είναι ο τρόπος που εισάγονται.

Για να εισάγουμε στο φύλλο εργασίας μια συνάρτηση της οποίας η τιμή είναι ένας πίνακας, πρώτα επιλέγουμε την περιοχή όπου θα εμφανιστούν οι τιμές της συνάρτησης. Προφανώς η περιοχή αυτή πρέπει να έχει τόσα κελιά, όσα και τα στοιχεία του πίνακα που θα προκύψει από την εφαρμογή της συνάρτησης. Ακολούθως πληκτρολογούμε το = και τη συνάρτηση, π.χ. =MINVERSE(, και εισάγουμε με το ποντίκι στο όρισμα της συνάρτησης την κατάλληλη περιοχή. Όταν γίνουν όλες αυτές οι διαδικασίες **δεν πατάμε *Enter*, αλλά πρώτα τα πλήκτρα *Ctrl* και *Shift* και έχοντας πατημένα αυτά τα δύο πλήκτρα πατάμε *Enter*. Αυτό ισχύει για όλες τις συναρτήσεις που η έξοδός τους (η τιμή τους) είναι ένας πίνακας**.

Παράδειγμα Ι.5

Έστω ότι θέλουμε να επιλύσουμε το σύστημα:

$$x_1 + (1/2)x_2 + (1/3)x_3 = 1$$
$$(1/2)x_1 + (1/3)x_2 + (1/4)x_3 = 0$$
$$(1/3)x_1 + (1/4)x_2 + (1/5)x_3 = 0$$

◆ Το παραπάνω σύστημα μπορεί να γραφεί με τη χρήση πινάκων ως **AX** = **B**, όπου οι πίνακες **A**, **B** και **X** είναι οι

$$\mathbf{A} = \begin{pmatrix} 1 & 1/2 & 1/3 \\ 1/2 & 1/3 & 1/4 \\ 1/3 & 1/4 & 1/5 \end{pmatrix} \qquad \mathbf{B} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \qquad \mathbf{X} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}$$

Τώρα το σύστημα λύνεται πολύ απλά αν πολλαπλασιάσουμε και τα δύο μέλη της σχέσης **AX** = **B** επί τον αντίστροφο πίνακα του **A**, δηλαδή επί $\mathbf{A^{-1}}$. Παίρνουμε

$$\mathbf{A^{-1}AX} = \mathbf{A^{-1}B} \quad \Rightarrow \quad \mathbf{X} = \mathbf{A^{-1}B}$$

διότι $\mathbf{A^{-1}AX} = \mathbf{IX} = \mathbf{X}$, όπου **I** είναι ο μοναδιαίος πίνακας.

Επομένως για την επίλυση ενός γραμμικού συστήματος εξισώσεων αρκεί να προσδιορίσουμε τον αντίστροφο πίνακα του πίνακα των συντελεστών των αγνώστων και να τον πολλαπλασιάσουμε επί τον πίνακα των σταθερών όρων. Στο σημεία αυτό πρέπει να γνωρίζουμε ότι γενικά ένας πίνακας **X** με n γραμμές και m στήλες μπορεί να πολλαπλασιαστεί μόνο επί έναν πίνακα **Y** με m γραμμές και k στήλες και το αποτέλεσμα είναι ένας πίνακας **Z** με n γραμμές και k στήλες.

Με βάση αυτές τις παρατηρήσεις, σε ένα φύλλο εργασίας στο A3 γράφουμε *ΠΙΝΑΚΑΣ **A***, στο E3 εισάγουμε το γράμμα ***B*** και στο G3 το γράμμα ***X***. Ακολούθως στην περιοχή A4:C6 εισάγουμε τους συντελεστές των αγνώστων και στην περιοχή E4:E6 τους σταθερούς όρους. Καλό είναι η σταθερά 1/3 να εισαχθεί με το μαθηματικό τύπο =1/3. Τέλος, στο A8 εισάγουμε τον τίτλο *ΑΝΤΙΣΤΡΟΦΟΣ ΤΟΥ **A***. Επειδή ο αντίστροφος ενός τετραγωνικού πίνακα τρίτης τάξης είναι επίσης ένας τετραγωνικός πίνακας τρίτης τάξης, επιλέγουμε με το ποντίκι την περιοχή A9:C11. Ακολούθως πληκτρολογούμε την έκφραση =MINVERSE(και με το ποντίκι επιλέγουμε την περιοχή A4:C6. Στο σημείο αυτό έχοντας πατημένα τα πλήκτρα *Ctrl* και *Shift* πατάμε *Enter*. Η περιοχή A9:C11 γεμίζει με τιμές, που είναι ο αντίστροφος του πίνακα **A** που υπάρχει στην περιοχή A4:C6.

Θα πρέπει τώρα να προχωρήσουμε στον πολλαπλασιασμό του

αντίστροφου πίνακα $\mathbf{A}^{-1}$ επί τον πίνακα **B**. Σύμφωνα με τον κανόνα πολλαπλασιασμού πινάκων που αναφέραμε, το γινόμενο ενός τετραγωνικού πίνακα τρίτης τάξης επί έναν πίνακα στήλης με τρία στοιχεία είναι πίνακας μιας στήλης με τρία στοιχεία. Έτσι, επιλέγουμε την περιοχή G4:G6 και πληκτρολογούμε =MMULT(. Για να εισάγουμε το όρισμα ή πληκτρολογούμε A9:C11;E4:E6 ή επιλέγουμε αυτή την περιοχή με το ποντίκι ως εξής. Κάνουμε *κλικ* στο A9 και με συνεχώς πατημένο το αριστερό πλήκτρο του ποντικιού φέρνουμε τον κέρσορα στο C11 και αφήνουμε το ποντίκι. Πατάμε το πλήκτρο *Ctrl* και με πατημένο το πλήκτρο αυτό φέρνουμε τον κέρσορα στο E4, κάνουμε *κλικ* και με πατημένο το αριστερό πλήκτρο του ποντικιού σέρνουμε τον κέρσορα στο E6 όπου και αφήνουμε το ποντίκι. Το πλήκτρο *Ctrl* μπορούμε να το ελευθερώσουμε μετά το *κλικ* στο E4. Η περιοχή A9:C11;E4:E6 εισάγεται στη συνάρτηση. Όταν ολοκληρωθεί αυτή η διαδικασία, με πατημένα τα πλήκτρα *Ctrl* και *Shift* πατάμε *Enter*. Η λύση του συστήματος εμφανίζεται στην περιοχή G4:G6. Στο σχήμα Ι.2 δίνεται η εικόνα του φύλλου εργασίας μετά τη λύση του συστήματος.

	A	B	C	D	E	F	G
1	**ΕΠΙΛΥΣΗ ΓΡΑΜΜΙΚΟΥ ΣΥΣΤΗΜΑΤΟΣ**						
2							
3	ΠΙΝΑΚΑΣ **A**				**B**		**X**
4	1	0,5	0,333333		1		9
5	0,5	0,333333	0,25		0		-36
6	0,333333	0,25	0,2		0		30
7							
8	ΑΝΤΙΣΤΡΟΦΟΣ ΤΟΥ **A**						
9	9	-36	30				
10	-36	192	-180				
11	30	-180	180				

Σχήμα Ι.2. Επίλυση του γραμμικού συστήματος στο Παράδειγμα Ι.5

Ι.7 ΕΠΙΛΥΣΗ ΕΞΙΣΩΣΕΩΝ

Οι ρίζες μιας εξίσωσης f(x) = 0, δηλαδή οι τιμές του x για τις οποίες ισχύει η ισότητα, μπορούν να εκτιμηθούν με ικανοποιητική ακρίβεια αν κάνουμε τη γραφική παράσταση της συνάρτησης. Για πολύ μεγαλύτερη ακρίβεια μπορούμε να χρησιμοποιήσουμε το πρόγραμμα *Αναζήτηση στόχου (Goal Seek)* του *Excel* από το *Δεδομένα* (*Data*) → *Εργαλεία Δεδομένων (Data Tools)* → *Ανάλυση Πιθανοτήτων (What-If Analysis)* → *Αναζήτηση στόχου* (*Goal Seek)*.

Παράδειγμα Ι.6

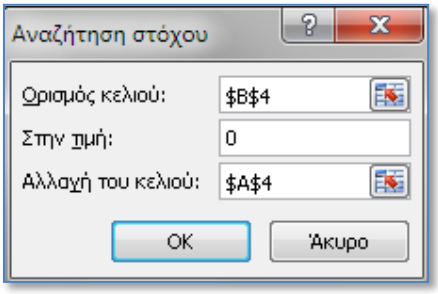
Να προσδιοριστεί η πραγματική ρίζα της εξίσωσης $x^5 - 100 = 0$.

◆ Εισάγουμε στο Α1 τον τίτλο *Επίλυση της x^5-100=0* και στο Α3 και Β3 τους τίτλους ρ και y, αντίστοιχα. Επίσης στο Α4 εισάγουμε μια πρώτη εκτίμηση της ρίζας, έστω 1, και στο Β4 πληκτρολογούμε τον τύπο = Α4^5-100. Στο σημείο αυτό θα έχουμε την εικόνα του σχήματος Ι.3.

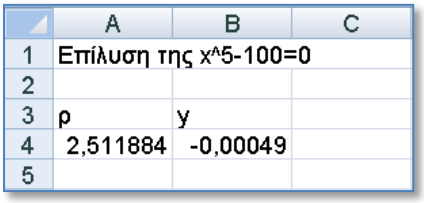
	A	B	C
1	Επίλυση της x^5-100=0		
2			
3	ρ	y	
4	1	-99	
5			

Σχήμα Ι.3. Πρώτο βήμα για τον υπολογισμό της πραγματικής ρίζας της εξίσωσης $x^5 - 100 = 0$ με το πρόγραμμα *Αναζήτηση στόχου*

Ακολούθως από *Δεδομένα* → *Εργαλεία Δεδομένων* → *Ανάλυση Πιθανοτήτων* → *Αναζήτηση στόχου* ενεργοποιούμε το *Αναζήτηση στόχου (Goal Seek)*, οπότε και ανοίγει το αντίστοιχο παράθυρο, το οποίο συμπληρώνουμε όπως στο σχήμα Ι.4. Ουσιαστικά ζητάμε από το πρόγραμμα να επιτύχει την τιμή 0 στο κελί Β4 αλλάζοντας τις τιμές στο κελί Α4. Είναι προφανές ότι όταν αυτό επιτευχθεί, στο κελί Α4 θα είναι η ζητούμενη ρίζα. Με *κλικ* στο *ΟΚ* ανοίγει ένα νέο παράθυρο που μας πληροφορεί για την τιμή που έχει βρεθεί στο κελί Β4. Με *κλικ* στο *ΟΚ* παίρνουμε την εικόνα του σχήματος Ι.5. Η τιμή ρ = 2.511884 είναι η πραγματική ρίζα της εξίσωσης $x^5 - 100 = 0$.



Σχήμα Ι.4. Συμπλήρωση παραθύρου *Αναζήτηση στόχου*



	A	B	C
1	Επίλυση της x^5-100=0		
2			
3	ρ	y	
4	2,511884	-0,00049	
5			

Σχήμα Ι.5. Τμήμα της οθόνης μετά τον υπολογισμό της ρίζας της εξίσωσης $x^5 - 100 = 0$

Παρατήρηση. Από την τιμή -0.00049 στο κελί Β4 του σχήματος Ι.5 παρατηρούμε ότι δε μηδενίζεται η τιμή της συνάρτησης f(ρ) = 0 όταν ρ = 2.511884. Αυτό οφείλεται στην ακρίβεια της μεθόδου προσδιορισμού της ρίζας. Αν θέλουμε να αυξήσουμε την ακρίβεια, πηγαίνουμε *Αρχείο* ή *Κουμπί Office* → *Επιλογές (του Excel)* → *Τύποι* και στο πλαίσιο κειμένου *Μέγιστη μεταβολή (Maximum Change)* πληκτρολογούμε 1e-15.

Ι.8 ΓΡΑΦΙΚΕΣ ΠΑΡΑΣΤΑΣΕΙΣ

Η πορεία που ακλουθούμε για την κατασκευή μιας απλής γραφικής παράστασης περιγράφεται αναλυτικά στο παρακάτω παράδειγμα.

Παράδειγμα Ι.7

Στον πίνακα Ι.3 δίνεται η μεταβολή των πειραματικών τιμών της πίεσης, P(πειρ.), με τον όγκο, V, 1 mole CO_2 στη θερμοκρασία των 300 K. Στον ίδιο πίνακα υπάρχουν και οι θεωρητικές τιμές της πίεσης, P(υπολ.), που προσδιορίστηκαν με βάση την καταστατική εξίσωση των ιδανικών αερίων. Να γίνει η γραφική παράσταση της μεταβολής των πειραματικών και θεωρητικών τιμών της πίεσης με τον όγκο.

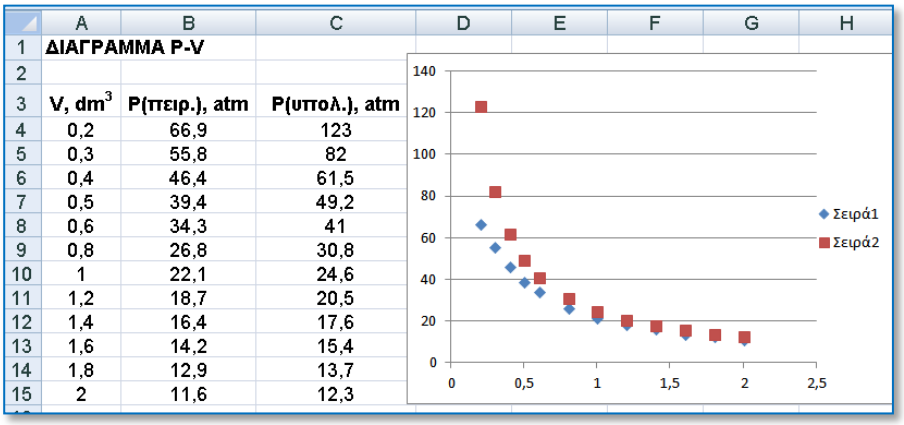
◆ Ανοίγουμε ένα φύλλο εργασίας, στο Α1 εισάγουμε τον τίτλο ΔΙΑΓΡΑΜΜΑ P – V και στα κελιά Α3, Β3, C3 πληκτρολογούμε αντίστοιχα "V, dm^3", "P(πειρ.), atm" και "P(υπολ.), atm". Ακολούθως, εισάγουμε τις τιμές του πίνακα Ι.3 στην περιοχή Α4:C15.

Για την κατασκευή της γραφικής παράστασης επιλέγουμε την περιοχή Α4:C15, κάνουμε *κλικ* στο *Εισαγωγή* (*Insert)* και από το *Γραφήματα* (*Charts*) → *Διασπορά* (*Scatter)* επιλέγουμε το *Διασπορά μόνο με δείκτες* (*Scatter with only Markers)*. Τότε παίρνουμε τη γραφική

παράσταση του σχήματος Ι.6. Το *Excel* αυτόματα χρησιμοποιεί ως ανεξάρτητη μεταβλητή τη στήλη που είναι αριστερά και ως εξαρτημένες τις στήλες που είναι δεξιά. Παρατηρούμε ότι η γραφική παράσταση που προκύπτει δεν έχει τίτλους στους άξονες και γενικά θέλει μορφοποίηση. Τα βήματα που ακολουθούμε για τη μορφοποίηση είναι τα εξής:

Πίνακας Ι.3. Μεταβολή της πειραματικής P(πειρ.) και θεωρητικής P(υπολ.) πίεσης με τον όγκο V ενός mole CO_2 στη θερμοκρασία των 300 K.

V dm^3	P(πειρ.) atm	P(υπολ.) atm	V dm^3	P(πειρ.) atm	P(υπολ.) atm
0.2	66.9	123.0	1.0	22.1	24.6
0.3	55.8	82.0	1.2	18.7	20.5
0.4	46.4	61.5	1.4	16.4	17.6
0.5	39.4	49.2	1.6	14.2	15.4
0.6	34.3	41.0	1.8	12.9	13.7
0.8	26.8	30.8	2.0	11.6	12.3



Σχήμα Ι.6. Τμήμα της οθόνης του υπολογιστή με τη γραφική παράσταση του Παραδείγματος Ι.7

(1) Για να εισάγουμε τίτλους αξόνων κάνουμε *κλικ* επάνω στη γραφική παράσταση και από τη λωρίδα εντολών *Διάταξη* (*Layout)* πηγαίνουμε στο πάνελ *Τίτλοι* (*Labels).* Από το εικονίδιο *Τίτλοι Άξονα*

(*AxisTitles)* επιλέγουμε αρχικά να τοποθετήσουμε έναν οριζόντιο τίτλο από το *Τίτλος πρωτεύοντα οριζόντιου άξονα (Primary Horizontal Axis Title)* → *Τίτλος κάτω από τον άξονα (Title Below Axis)*. Ακολούθως πληκτρολογούμε "V, dm3", επιλέγουμε το 3 και με δεξί *κλικ* από το *Γραμματοσειρά* (*Font)* κάνουμε *κλικ* στο *Εκθέτης* (*Superscript)*. Για τον άξονα των y εργαζόμαστε ανάλογα. Δηλαδή, από το *Τίτλοι Άξονα* (*AxisTitles)* επιλέγουμε *Τίτλος πρωτεύοντα κατακόρυφου άξονα (Primary Vertical Axis Title)* → *Περιστρεμμένος τίτλος (Rotated Title)*, γράφουμε τον τίτλο "P, atm" και τον μορφοποιούμε κατά βούληση.

(2) Η γραμματοσειρά των αριθμών που υπάρχουν στους άξονες, όπως επίσης το μέγεθος και το στυλ τους, μορφοποιούνται αν πρώτα κάνουμε *κλικ* σε έναν αριθμό, π.χ. του άξονα των x. Η μορφοποίηση μπορεί να γίνει από το πάνελ *Γραμματοσειρά (Font)* της κορδέλας *Κεντρική (Home)*. Αν κάνουμε *κλικ* σε έναν αριθμό, ακολούθως δεξί *κλικ* και στον κατάλογο εντολών που ανοίγει επιλέξουμε *Μορφοποίηση άξονα* (*Format Axis)*, τότε μπορούμε να μορφοποιήσουμε τον ίδιο τον άξονα. Δηλαδή να επιλέξουμε την κλίμακα, το χρώμα του άξονα, το πάχος γραμμής, τα σημεία της κλίμακας κ.ά.

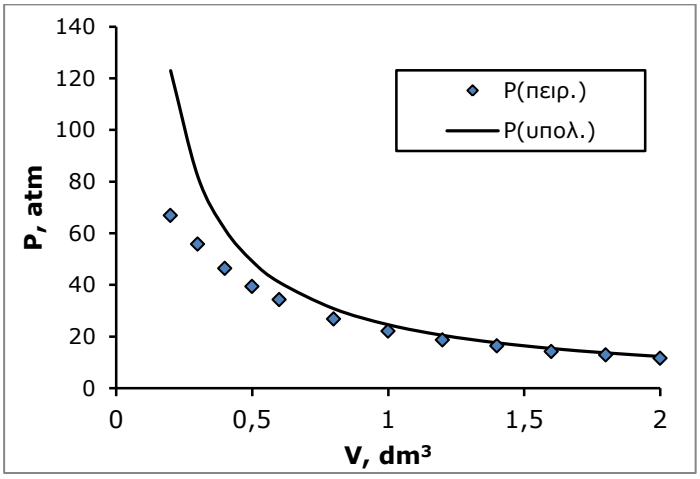
(3) Παρατηρούμε ότι το διάγραμμα αποτελείται μόνο από σημεία. Αν θέλουμε τα θεωρητικά δεδομένα να παρουσιάζονται με συνεχή καμπύλη, χωρίς σημεία, τότε εργαζόμαστε ως εξής. Κάνουμε απλό *κλικ* επάνω σε ένα από τα σημεία που αντιστοιχούν σε θεωρητικά δεδομένα. Τότε επιλέγονται όλα τα σημεία των θεωρητικών δεδομένων και με δεξί *κλικ* ανοίγει ένας κατάλογος εντολών από τις οποίες επιλέγουμε τη *Μορφοποίηση σειράς δεδομένων* (*Format Data Series)*, η οποία οδηγεί στο αντίστοιχο παράθυρο διαλόγου. Σε αυτό κάνουμε τις ακόλουθες ενέργειες. Από το *Επιλογές δείκτη* (*Market Options)* επιλέγουμε το *Κανένας* (*None)*, από το *Χρώμα γραμμής* (*Line Color)* το *Συμπαγής γραμμή* (*Solid line)* καθώς και το επιθυμητό χρώμα της γραμμής και τέλος από το *Στυλ γραμμής* (*Line Style)* επιλέγουμε την *Ομαλή γραμμή* (*Smoothed line)* και το πάχος της γραμμής από το *Πλάτος* (*Width)*. Ολοκληρώνουμε αυτές τις ενέργειες με *κλικ* στο *Κλείσιμο* (*Close)*.

(4) Για να διαγράψουμε τις οριζόντιες γραμμές κάνουμε *κλικ* σε μια από αυτές, οπότε επιλέγονται όλες. Διαγράφονται πατώντας το πλήκτρο *Delete*.

(5) Για να μορφοποιήσουμε τη λεζάντα που στο σχήμα Ι.6 έχει τους τίτλους Σειρά1 και Σειρά2 με τα αντίστοιχα σύμβολα, εργαζόμαστε ως εξής. Κάνουμε δεξί *κλικ* στην *Περιοχή σχεδίασης* (*Plot area*), *κλικ* στο *Επιλογή*

δεδομένων (*Select Data)* και στο παράθυρο που ανοίγει κάνουμε *κλικ* στο *Επεξεργασία* (*Edit).* Στο νέο παράθυρο που ανοίγει κάνουμε *κλικ* στο πλαίσιο κειμένου *Όνομα σειράς* (*Series Name)* και πληκτρολογούμε τον τίτλο Ρ(πειρ.). Με *κλικ* στο *ΟΚ* επιστρέφουμε στο προηγούμενο παράθυρο, στο οποίο κάνουμε *κλικ* στο Σειρά2, *κλικ* στο *Επεξεργασία* και στο παράθυρο που ανοίγει πληκτρολογούμε τον τίτλο Ρ(υπολ.) στο πλαίσιο κειμένου *Όνομα σειράς*. Ολοκληρώνουμε με δύο διαδοχικά *κλικ* στο *ΟΚ*. Η λεζάντα μπορεί να μπει σε πλαίσιο αν πρώτα την επιλέξουμε με απλό *κλικ* και ακολούθως με δεξί *κλικ* επιλέξουμε το *Μορφοποίηση υπομνήματος* (*Format Legend)*.

Το αποτέλεσμα των παραπάνω ενεργειών δίνεται στο σχήμα Ι.7.



Σχήμα Ι.7. Μορφοποιημένη η γραφική παράσταση του σχήματος Ι.6

Παρατήρηση 1. Αν θέλουμε να κάνουμε μόνο το διάγραμμα μεταβολής των πειραματικών τιμών της πίεσης με τον όγκο, πρέπει να επιλέξουμε την περιοχή A3:B15, ενώ για να κατασκευάσουμε μόνο το διάγραμμα μεταβολής των θεωρητικών τιμών της πίεσης με τον όγκο, πρέπει να επιλέξουμε την περιοχή A3:A15;C3:C15.

Παρατήρηση 2. Αν σε μία γραφική παράσταση θέλουμε να προσθέσουμε μια νέα σειρά δεδομένων (x, y) εργαζόμαστε ως εξής. Κάνουμε δεξί *κλικ* στην *Περιοχή σχεδίασης* (*Plot area*), *κλικ* στο *Επιλογή δεδομένων* (*Select*

Data) και στο παράθυρο που ανοίγει κάνουμε *κλικ* στο *Προσθήκη* (*Add)* και συμπληρώνουμε κατάλληλα τα συνδυαστικά πλαίσια κειμένου που υπάρχουν. Συγκεκριμένα, στο πλαίσιο κειμένου Όνο*μα σειράς* (*Series Name)* συμπληρώνουμε αν θέλουμε το όνομα των δεδομένων που θα προσθέσουμε, ακολούθως κάνουμε *κλικ* στο πλαίσιο κειμένου *Τιμές σειράς X* (*Series X Values)* και με το ποντίκι επιλέγουμε την περιοχή που υπάρχουν οι νέες τιμές x και τέλος διαγράφουμε το περιεχόμενο που υπάρχει στο πλαίσιο κειμένου *Τιμές σειράς Y (Series Y Values)* και με το ποντίκι επιλέγουμε την περιοχή που υπάρχουν οι νέες τιμές y. Με δύο διαδοχικά *κλικ* στο *OK* ολοκληρώνεται η προσθήκη των νέων δεδομένων στη γραφική παράσταση.

Παράρτημα ΙΙ

ΕΙΣΑΓΩΓΗ ΣΤΟ SPSS

ΙΙ.1 ΓΕΝΙΚΑ

Το στατιστικό πρόγραμμα *SPSS* πήρε το όνομά του στην αρχή από τα αρχικά *Statistical Package for the Social Sciences* και αργότερα από το *Statistical Product and Service Solutions*. Είναι ένα από τα καλύτερα στατιστικά πακέτα που μπορεί να χρησιμοποιηθεί για τη στατιστική ανάλυση τόσο κοινωνικοοικονομικών δεδομένων όσο και δεδομένων θετικών επιστημών.

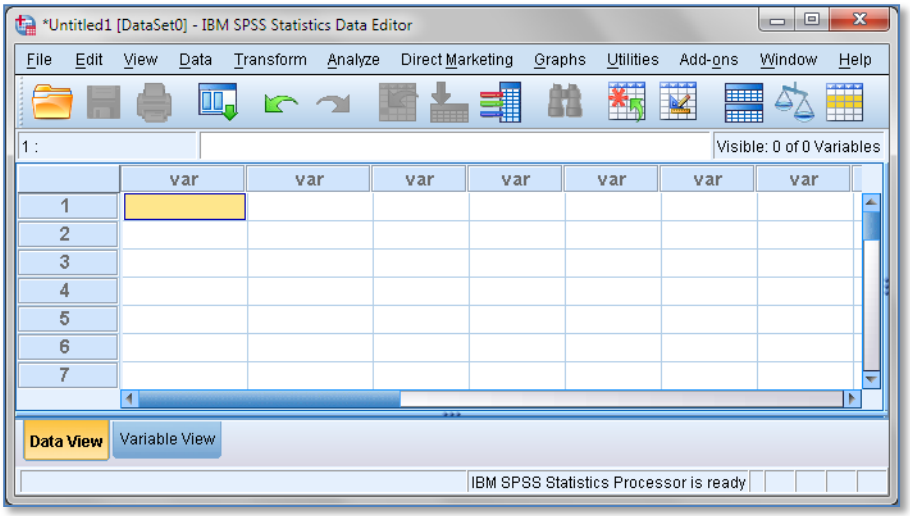
ΙΙ.2 ΦΥΛΛΑ ΕΡΓΑΣΙΑΣ ΤΟΥ SPSS

Στο *SPSS* υπάρχουν δύο βασικά παράθυρα: το παράθυρο δεδομένων, που είναι ο *Data Editor*, και το παράθυρο αποτελεσμάτων, που είναι ο *Viewer*. Ο *Data Editor* είναι ένα *φύλλο εργασίας*, στο οποίο καταχωρούμε τα δεδομένα που θέλουμε να αναλύσουμε. Στην πραγματικότητα ο *Data Editor* αποτελείται από δύο παράθυρα: Το *Data View* (σχήμα ΙΙ.1) και το *Variable View* (σχήμα ΙΙ.2). Στο πρώτο εισάγουμε τα δεδομένα που θα αναλύσουμε και στο δεύτερο τα μορφοποιούμε. Οι οριζόντιες γραμμές στο *Data View* ονομάζονται *Cases (Περιπτώσεις)* και είναι αριθμημένες με αύξουσα σειρά. Στο φύλλο αυτό οι στήλες αντιστοιχούν στις *Variables (Μεταβλητές)*.

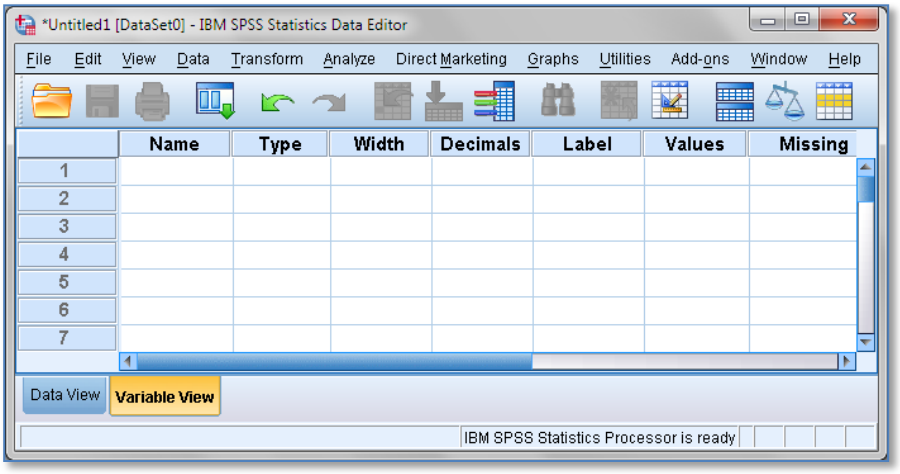
Ο *Viewer* είναι το αρχείο αποτελεσμάτων. Στο αριστερό του παράθυρο, στο *Output*, εμφανίζονται οι στατιστικές πράξεις που έχουν γίνει και στο δεξιό τα στατιστικά αποτελέσματα (σχήμα ΙΙ.3).

Όπως όλα τα παράθυρα των *Windows*, ο *Data Editor* και ο *Viewer* έχουν στο πάνω μέρος τη *γραμμή τίτλου (title bar)*. Κάτω από τη γραμμή τίτλου υπάρχει η *γραμμή μενού (menu bar)*, η οποία στον *Data Editor* περιέχει τις επιλογές: *File, Edit, View, Data, Transform, Analyse, DirectMarketing, Graphs, Utilities, Add-ons, Window, Help*. Οι ίδιες λέξεις

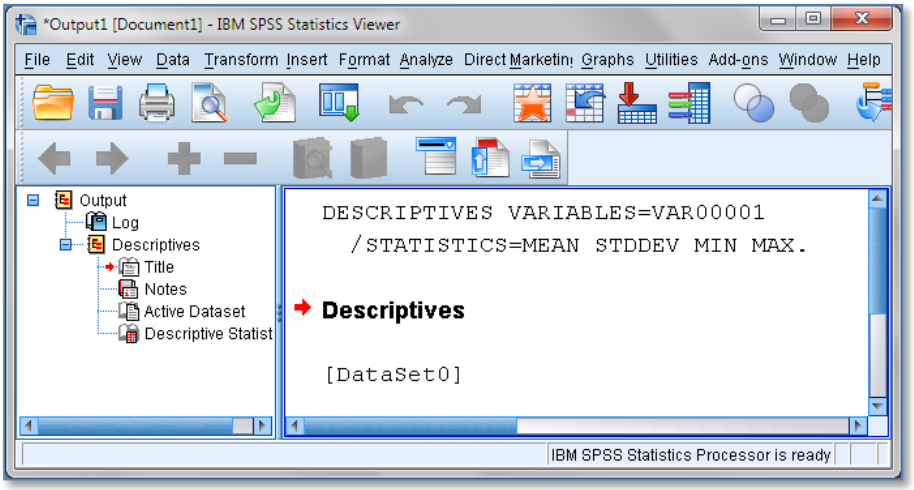
υπάρχουν και στον *Viewer*, όπου όμως υπάρχουν επιπλέον και οι λέξεις *Insert* και *Format*. Οι επιλογές αυτές επιτρέπουν να κάνουμε διάφορες ενέργειες στα δεδομένα του φύλλου εργασίας ή στο ίδιο το φύλλο εργασίας. Πολύ περιληπτικά για τις βασικότερες επιλογές ισχύει:



Σχήμα ΙΙ.1. Το παράθυρο *Data View* για εισαγωγή δεδομένων



Σχήμα ΙΙ.2. Το παράθυρο *Variable View* για μορφοποίηση δεδομένων



Σχήμα ΙΙ.3. Ο *SPSS Viewer* για την παρουσίαση των αποτελεσμάτων

- *File*: Μπορούμε να ανοίξουμε ένα νέο αρχείο *(New)*, ή ένα παλαιό *(Open)*, να αποθηκεύσουμε ένα αρχείο *(Save)*, να εκτυπώσουμε *(Print)*, κ.ο.κ.
- *Edit*: Μπορούμε να τροποποιήσουμε ή να αντιγράψουμε τμήματα του αρχείου δεδομένων.
- *View*: Επιτρέπει να προσαρμόζουμε τα διάφορα στοιχεία του παραθύρου *(Toolbars*, *Fonts*, *Grid lines*, *…)* ανάλογα με τις επιλογές μας.
- *Data*: Μπορούμε να πραγματοποιήσουμε καθολικές αλλαγές στα δεδομένα.
- *Transform*: Χρησιμοποιείται για να πραγματοποιήσουμε αλλαγές στις μεταβλητές.
- *Analyze*: Χρησιμοποιείται για τη στατιστική ανάλυση των δεδομένων.
- *Graphs*: Χρησιμοποιείται για δημιουργία γραφικών παραστάσεων.
- *Window*: Η επιλογή αυτή δίνει τη δυνατότητα να μεταβούμε σε κάποιο άλλο ενεργό παράθυρο.
- *Help*: Προσφέρει διάφορα είδη βοήθειας.

Κάτω από τη γραμμή μενού υπάρχει η *γραμμή εργαλείων (toolbars)*, η οποία περιέχει με μορφή εικόνας ή σχήματος εντολές που ήδη βρίσκονται

στη γραμμή μενού.

Τέλος, θα πρέπει να αναφέρουμε ότι τα φύλλα εργασίας του *SPSS* διαφέρουν σημαντικά από τα αντίστοιχα του *Excel*. Οι δυνατότητες μαθηματικών υπολογισμών είναι εξαιρετικά μικρές και περιορίζονται μόνο σε πράξεις που σχετίζονται με τις στήλες (μεταβλητές) των δεδομένων.

ΙΙ.3 ΚΑΤΑΧΩΡΙΣΗ ΔΕΔΟΜΕΝΩΝ ΣΤΟ ΦΥΛΛΟ ΕΡΓΑΣΙΑΣ

Ο απλούστερος τρόπος καταχώρησης δεδομένων σε ένα φύλλο εργασίας είναι με απ' ευθείας πληκτρολόγηση των δεδομένων στο *Data View*. Επίσης μπορούμε να μεταφέρουμε δεδομένα από ένα φύλλο του *Excel* σε φύλλο του *SPSS* ως εξής. Επιλέγουμε τα δεδομένα στο φύλλο του *Excel*, τα αντιγράφουμε με Ctrl+C και τα επικολλούμε στο φύλλο του *SPSS* με Ctrl+V. Για να είναι όμως δυνατή αυτή η μεταφορά θα πρέπει και τα δύο προγράμματα να χρησιμοποιούν το ίδιο σύμβολο για τα δεκαδικά (κόμμα ή τελεία). Με την ίδια διαδικασία μπορούμε να μεταφέρουμε δεδομένα από ένα txt αρχείο στο *SPSS*, με την προϋπόθεση όμως ότι οι στήλες των δεδομένων στο txt αρχείο χωρίζονται με tab.

Παρατήρηση. Ανάλογα με την έκδοση του *SPSS* για να εισάγουμε μια αλφαριθμητική μεταβλητή σε ένα φύλλο εργασίας, ίσως να χρειαστεί πρώτα να τη δηλώσουμε στο *Variable View* ως *String*.

Μετά την εισαγωγή των δεδομένων, συνήθως τα μορφοποιούμε από το *Variable View* ως εξής:

- Κάνουμε *κλικ* στο *Variable View* και στην πρώτη στήλη, *Name*, πληκτρολογούμε τις επικεφαλίδες που θέλουμε να έχουν οι στήλες (Μεταβλητές) στο *Data View*.
- Στη δεύτερη στήλη, *Type*, προσδιορίζουμε τον τύπο των μεταβλητών. Αν κάνουμε *κλικ* σε ένα κελί αυτής της στήλης, τότε στα δεξιά του κελιού εμφανίζεται ένα μικρό ορθογώνιο χρώματος γκρι. Με *κλικ* στο ορθογώνιο αυτό εμφανίζεται ένα παράθυρο διαλόγου που μας επιτρέπει να επιλέξουμε τον τύπο της μεταβλητής. Έχουμε τις ακόλουθες επιλογές: *Numeric, Comma, Dot, Scientific notation, Date, Dollar, Custom currency* και *String*. *Numeric* είναι οι αριθμητικές μεταβλητές. *Comma* είναι μια αριθμητική μεταβλητή όταν οι χιλιάδες προσδιορίζονται με κόμμα ενώ τα δεκαδικά με τελεία, π.χ. 5,012.6. *Dot* είναι μια αριθμητική μεταβλητή όταν οι χιλιάδες προσδιορίζονται με τελεία και τα δεκαδικά με κόμμα, π.χ. 5.012,6. *Scientific notation*

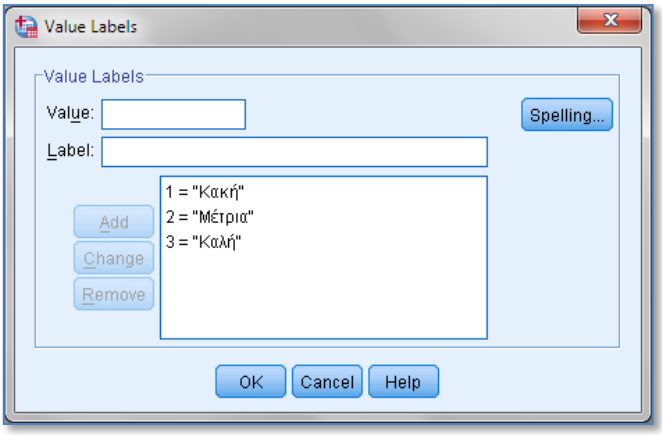
δηλώνει ότι θα χρησιμοποιηθεί επιστημονική παρουσίαση της αριθμητικής μεταβλητής, π.χ. 9.12E2 αντί για 912 ή 9.12E-2 αντί για 0.0912. Το *Date* χρησιμοποιείται για να εισάγουμε ημερομηνίες και το *Dollar* όταν αναφερόμαστε σε δολάρια. Σε περίπτωση άλλων νομισμάτων χρησιμοποιούμε το *Custom currency*. Τέλος, *String* είναι μια αλφαριθμητική μεταβλητή, δηλαδή μια μεταβλητή που περιλαμβάνει γράμματα ή γράμματα και αριθμούς, π.χ. male, m, f, f1, κ.ο.κ.

- Στην τρίτη στήλη, *Width*, καθορίζεται το εύρος μιας αλφαριθμητικής μεταβλητής και στην τέταρτη στήλη, *Decimals*, τα δεκαδικά των αριθμητικών μεταβλητών.

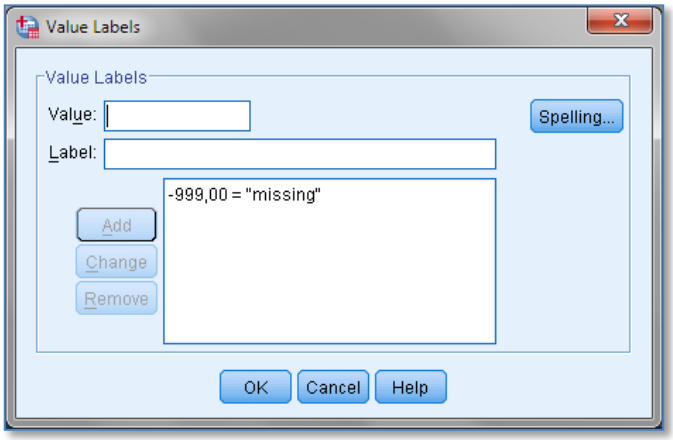
- Στη στήλη *Label* (*Ετικέτες*) μπορούμε να δώσουμε ένα διαφορετικό τίτλο σε μια μεταβλητή για τον ακόλουθο λόγο. Έστω ότι έχουμε ένα δείγμα τιμών δυναμικού και θέλουμε να έχει τον τίτλο E, V. Αυτόν τον τίτλο δε μπορούμε να τον εισάγουμε στην πρώτη στήλη, τη στήλη *Name*, επειδή στη στήλη *Name* δεν επιτρέπεται το σύμβολο ",". Αντίθετα στην στήλη *Label* μπορούμε να χρησιμοποιήσουμε τίτλους με πληθώρα συμβόλων.

- Στη στήλη *Values* δίνουμε πληροφορίες για τις τιμές της μεταβλητής. Πιο συγκεκριμένα ισχύουν τα ακόλουθα. Η προεπιλογή είναι *None* και αφορά τις περισσότερες μεταβλητές. Έστω όμως για παράδειγμα μια μεταβλητή, π.χ. η quality που παίρνει τις τιμές 1, 2 και 3 ανάλογα με την ποιότητα ενός τροφίμου και συγκεκριμένα έστω ότι quality = 1 δηλώνει Κακή, 2 Μέτρια και 3 Καλή. Σε αυτή την περίπτωση, για δική μας πληροφόρηση, κάνουμε *κλικ* στο κελί της στήλης *Values* που αντιστοιχεί στη μεταβλητή quality και *κλικ* στο μικρό γκρι ορθογώνιο, οπότε ανοίγει το παράθυρο διαλόγου του σχήματος ΙΙ.4. Στο πλαίσιο *Value* πληκτρολογούμε 1, στο *Label* πληκτρολογούμε Κακή και κάνουμε *κλικ* στο *Add*. Η έκφραση 1 = "Κακή" εισέρχεται στο μεγάλο ορθογώνιο πλαίσιο. Συνεχίζουμε εισάγοντας την τιμή 2 στο *Value*, τη λέξη Μέτρια στο *Label* και πάλι *κλικ* στο *Add*. Με αυτόν τον τρόπο στο τέλος θα πάρουμε την εικόνα το σχήματος ΙΙ.4.
- Επίσης στο *Values* σημειώνουμε τυχόν απούσες τιμές μιας μεταβλητής. Στο *SPSS* δεν επιτρέπεται να υπάρχουν κενά κελιά. Γι αυτό χρησιμοποιούμε μια συγκεκριμένη τιμή, π.χ. την τιμή –1 ή -999, για να δηλώσουμε ότι δεν γνωρίζουμε την τιμή σε αυτό το κελί. Έτσι αν σε μια μεταβλητή υπάρχουν μία ή περισσότερες τιμές που λείπουν, κάνουμε *κλικ* στο *Values* που αντιστοιχεί στη μεταβλητή αυτή και

ανοίγουμε το παράθυρο διαλόγου *Value Labels*. Στο πλαίσιο *Value* πληκτρολογούμε -999 ή όποια άλλη τιμή θέλουμε, στο *Label* πληκτρολογούμε missing και κάνουμε *κλικ* στο *Add*, όπως φαίνεται στο σχήμα ΙΙ.5.



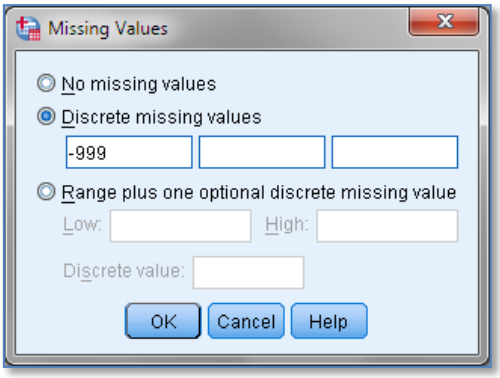
Σχήμα ΙΙ.4. Το παράθυρο διαλόγου *Value Labels* για τη μεταβλητή quality



Σχήμα ΙΙ.5. Το παράθυρο διαλόγου *Value Labels* για μεταβλητή που περιέχει απούσες τιμές

- Οι απούσες τιμές πρέπει να εισάγονται οπωσδήποτε στη επόμενη

στήλη *Missing* για χρήση στη στατιστική επεξεργασία που θα επιλέξουμε. Αυτό γίνεται αν κάνουμε *κλικ* στο μικρό γκρι ορθογώνιο που εμφανίζεται στα κελιά αυτής της στήλης. Τότε ανοίγει το παράθυρο διαλόγου του σχήματος ΙΙ.6. Τα τρία πλαίσια που υπάρχουν κάτω από το *Discrete missing values* δείχνουν ότι μπορούμε να χρησιμοποιήσουμε μέχρι και τρεις διαφορετικές τιμές για να δηλώσουμε στο πρόγραμμα ποιες τιμές θεωρούνται ως απούσες τιμές.



Σχήμα ΙΙ.6. Το παράθυρο διαλόγου *Missing Values*

- Στη στήλη *Columns* καθορίζουμε το πλάτος που θα έχει η στήλη μιας μεταβλητής.
- Η στήλη *Align* καθορίζει τη στοίχιση των τιμών μιας μεταβλητής στη στήλη της με επιλογές *Left (αριστερά)*, *Right (δεξιά)* και *Center (κέντρο)*.
- Η στήλη *Measure* καθορίζει το μέτρο των τιμών μιας μεταβλητής. Στο *SPSS* οι μεταβλητές χωρίζονται σε *Scale (κλιμακωτές)*, *Nominal (ονομαστικές)* και *Ordinal (σειριακές)*. *Scale* είναι οι τιμές μιας μεταβλητής όταν μπορούμε να υπολογίσουμε τις διαφορές που υπάρχουν μεταξύ δύο οποιοδήποτε τιμών. Στη βιβλιογραφία οι μεταβλητές *scale* χωρίζονται σε *ratio (αναλογία)* και σε *interval (διάστημα)*. Η ουσιαστική διαφορά ανάμεσα στις δύο αυτές υποκατηγορίες είναι ότι στις μεταβλητές *interval* το μηδέν δεν έχει πραγματική σημασία και μπορεί να μην υπάρχει. Για παράδειγμα, οι τιμές θερμοκρασίας στην κλίμακα Κελσίου είναι *interval* επειδή 0 ^{o}C δεν σημαίνει απουσία θερμοκρασίας. Αντίθετα η κλίμακα Kelvin είναι

ratio επειδή το μηδέν σε αυτή την κλίμακα σημαίνει το απόλυτο μηδέν της θερμοκρασίας. *Nominal* είναι οι τιμές μιας μεταβλητής όταν δεν έχουν καμιά σειρά ή σχέση μεταξύ τους. Για παράδειγμα, μια μεταβλητή που δηλώνει το φύλο και παίρνει τις τιμές f (γυναίκα) και m (άνδρας) είναι *Nominal*. Σε αυτή την περίπτωση μπορούμε αντί για f και m να χρησιμοποιήσουμε τους αριθμούς 1 και 2, αντίστοιχα. Και πάλι οι τιμές 1 και 2 θα είναι *Nominal*. Τέλος, *Ordinal* είναι οι τιμές μιας μεταβλητής όταν υποδηλώνουν μια σειριακή σχέση. Τυπικό παράδειγμα είναι οι αριθμοί 1, 2, 3, … όταν εκφράζουν σειρά επιτυχίας. Στη βιβλιογραφία χρησιμοποιείται και ο όρος *categorical* για να υποδηλώσει τιμές nominal ή/και ordinal.

ΙΙ.4 ΓΡΑΦΙΚΕΣ ΠΑΡΑΣΤΑΣΕΙΣ

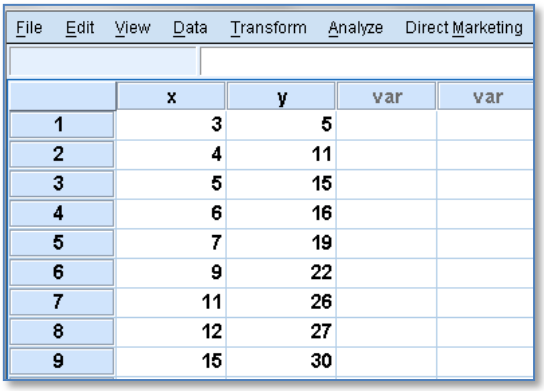
Το *SPSS* έχει πλούσια βιβλιοθήκη για γραφικές παραστάσεις που σχετίζονται με στατιστικά προβλήματα. Υστερεί όμως σε ευελιξία ως προς τη μορφοποίηση των γραφικών σε σχέση με το *Excel*. Για παράδειγμα, έστω ότι έχουμε τα x, y δεδομένα του πίνακα ΙΙ.1 και θέλουμε να κάνουμε τη γραφική τους παράσταση.

Πίνακας ΙΙ.1. Παράδειγμα δεδομένων x, y.

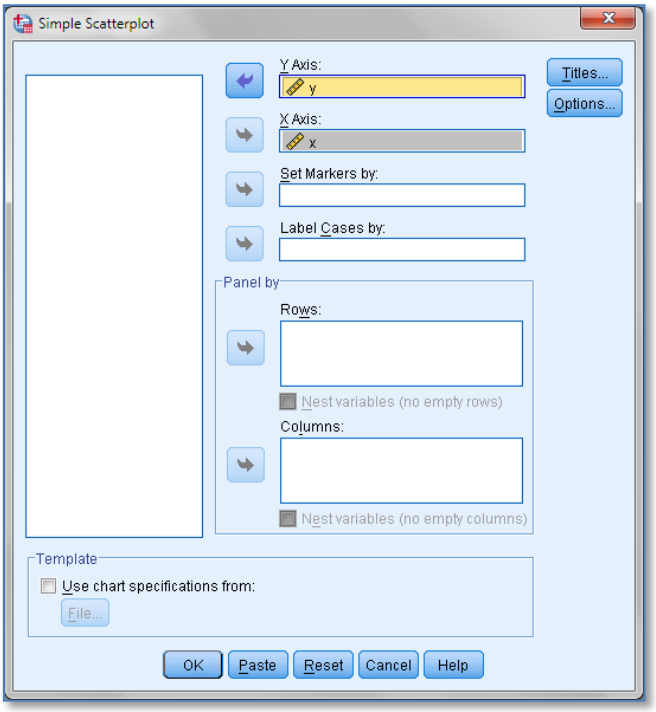
x	3	4	5	6	7	9	11	12	15
y	5	11	15	16	19	22	26	27	30

Μεταφέρουμε τα δεδομένα αυτά σε ένα φύλλο του *SPSS*, σε δύο στήλες με τίτλους x και y, αντίστοιχα, όπως στο σχήμα ΙΙ.7, και πηγαίνουμε *Graphs* → *Legacy Dialogs* → *Scatter/Dot*. Στο πλαίσιο διαλόγου που ανοίγει επιλέγουμε *Simple Scatter* και κάνουμε *κλικ* στο *Define*, οπότε εμφανίζεται το παράθυρο διαλόγου του σχήματος ΙΙ.8. Στο παράθυρο αυτό κάνουμε *κλικ* στο εικονίδιο που αντιπροσωπεύει τη στήλη x και ακολούθως κάνουμε *κλικ* στο βελάκι δίπλα στο πλαίσιο *X Axis* ώστε τα δεδομένα x να εισαχθούν στο σωστό πλαίσιο που αφορά τον άξονα των x. Με τον ίδιο τρόπο εισάγουμε τα δεδομένα y στο πλαίσιο *Y Axis* (σχήμα ΙΙ.8). Με *κλικ* στο *OK* παίρνουμε τη γραφική παράσταση του σχήματος ΙΙ.9.

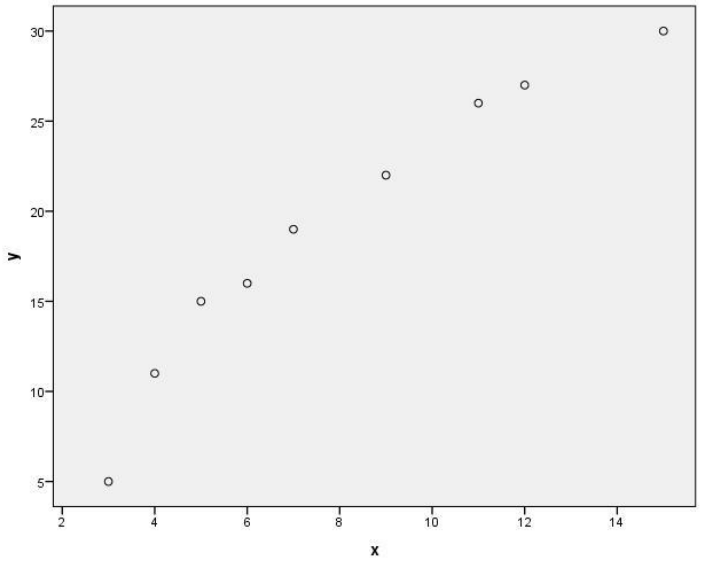
Συνήθως οι γραφικές παραστάσεις απαιτούν μορφοποίηση έτσι, ώστε οι αριθμοί και οι τίτλοι των αξόνων να έχουν το κατάλληλο μέγεθος και οι κλίμακες και τα σύμβολα να έχουν τα επιθυμητά χαρακτηριστικά. Για παράδειγμα, στο σχήμα ΙΙ.9 οι αριθμοί στους άξονες είναι πολύ μικροί.



Σχήμα ΙΙ.7. Εισαγωγή δεδομένων στο *SPSS*



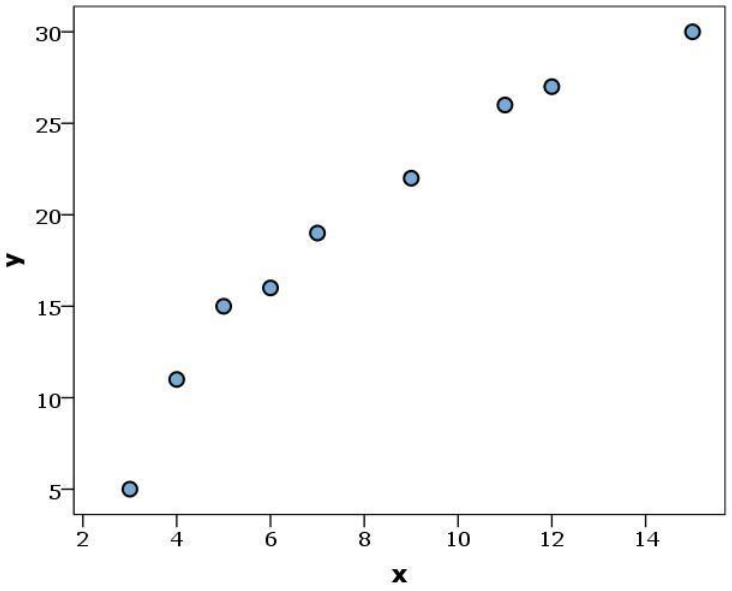
Σχήμα ΙΙ.8. Εισαγωγή δεδομένων στα πεδία X Axis και Y Axis



Σχήμα ΙΙ.9. Γραφική παράσταση των x, y δεδομένων

Για να μορφοποιήσουμε μια γραφική παράσταση κάνουμε διπλό *κλικ* επάνω της, οπότε ανοίγει ο επεξεργαστής γραφικών παραστάσεων (*Chart Editor*) για να κάνουμε τις μεταβολές που θέλουμε. Για να φύγουμε από τον επεξεργαστή κάνουμε *κλικ* στο εικονίδιο  ή πηγαίνουμε *File* → *Close*. Όταν είμαστε στον επεξεργαστή γραφικών παραστάσεων και κάνουμε *κλικ* σε έναν από τους αριθμούς της κλίμακας ενός άξονα, επιλέγονται όλοι οι αριθμοί και ταυτόχρονα ανοίγει ένα πλαίσιο διαλόγου, το *Properties*, στο οποίο μπορούμε να αλλάξουμε τη γραμματοσειρά, το μέγεθος, τον τύπο τους, το χρώμα τους, αλλά και τον αριθμό των δεκαδικών καθώς επίσης την κλίμακα του άξονα και τη μορφή του. Κάθε φορά που κάνουμε μια αλλαγή πρέπει να πατάμε το κουμπί *Apply*.

Με τον ίδιο τρόπο αν κάνουμε *κλικ* σε ένα σύμβολο που παριστάνει τα x, y δεδομένα στη γραφική παράσταση, επιλέγονται όλα και από το πλαίσιο *Properties* μπορούμε να τα μορφοποιήσουμε κατάλληλα (σχήμα ΙΙ.10).



Σχήμα ΙΙ.10. Μορφοποιημένη γραφική παράσταση των x, y δεδομένων

Αν σε μία γραφική παράσταση θέλουμε να εμφανίζονται δύο ή περισσότερες σειρές δεδομένων, (x_1, y_1), (x_2, y_2), … εργαζόμαστε ως εξής. Πηγαίνουμε *Graphs* → *Legacy Dialogs* → *Scatter/Dot* και στο πλαίσιο διαλόγου που ανοίγει επιλέγουμε *Overlay Scatter*. Ακολούθως εισάγουμε κάθε ζεύγος μεταβλητών (x_1, y_1), (x_2, y_2), … σε μία γραμμή στο πάνελ *Pairs,* στο παράθυρο *Overlay Scatterplot*, με προσοχή ώστε οι ανεξάρτητες μεταβλητές να εισέρχονται στο πλαίσιο *X Variable* και οι εξαρτημένες στο *Y Variable*.

ΙΙ.5 ΑΠΟΘΗΚΕΥΣΗ ΑΡΧΕΙΩΝ

Για να αποθηκεύσουμε ένα αρχείο του *SPSS* ακολουθούμε τα γνωστά βήματα των *Windows*: Κάνουμε *κλικ* στο *File (Αρχείο)*, που υπάρχει στη γραμμή μενού, *κλικ* στο *Save as (Αποθήκευση ως)* και συμπληρώνουμε κατάλληλα το παράθυρο διαλόγου που θα εμφανιστεί. Το αρχείο δεδομένων, ο *Data Editor*, αποθηκεύεται με το χαρακτηριστικό .sav. Αντίθετα, ένα αρχείο αποτελεσμάτων αποθηκεύεται με το χαρακτηριστικό .spv.

Είναι ενδιαφέρον ότι τα αρχεία του *SPSS* μπορούν να αποθηκευτούν

και ως αρχεία του *Excel*. Ένα αρχείο δεδομένων αποθηκεύεται ως αρχείο του *Excel* αν στο παράθυρο διαλόγου *Save Data As*, που ανοίγει μέσω *File → Save as*, επιλέξουμε την έκδοση του *Excel,* π.χ. *Excel 2007 through 2010*. Αφού εισάγουμε το όνομα του αρχείου στο πλαίσιο *File name* και κάνουμε *κλικ* στο *Save*, από το αρχείο του *SPSS* αποθηκεύεται μόνο το περιεχόμενο του φύλλου εργασίας που είναι στο *Data View*. Μετά την αποθήκευση εμφανίζεται ένα αρχείο αποτελεσμάτων που δίνει πληροφορίες για τις μεταβλητές που αποθηκεύθηκαν. Μπορούμε να το διαγράψουμε.

Η αποθήκευση αρχείων αποτελεσμάτων ως αρχεία του *Excel* ακολουθεί μια διαφορετική πορεία. Η απλή μεταφορά πινάκων αποτελεσμάτων του *SPSS* σε ένα φύλλο του *Excel* ή στο *Word* γίνεται πολύ απλά αν κάνουμε δεξί *κλικ* πάνω στο πίνακα, από τη λίστα που εμφανίζεται επιλέξουμε *Copy* ή *Copy Special* και ακολούθως αντιγράψουμε τον πίνακα στο φύλλο του *Excel* με απλή ή ειδική επικόλληση.

Τέλος, ενδιαφέρον παρουσιάζει η αποθήκευση γραφικών ως αρχεία εικόνων. Για το σκοπό αυτό κάνουμε *κλικ* στο γράφημα που θέλουμε να αποθηκεύσουμε, συνεχίζουμε με δεξί *κλικ* και στη λίστα που ανοίγει επιλέγουμε *Export*. Τότε εμφανίζεται ένα παράθυρο διαλόγου, στο οποίο κάνουμε τις ακόλουθες επιλογές: Από το πάνελ *Document → Type* επιλέγουμε *None (Graphics Only)*. Από το πάνελ *Graphics* επιλέγουμε τον τύπο της εικόνας, *TIF*, *JPG*,… και τέλος από το *Browse* επιλέγουμε τη διεύθυνση αποθήκευσης της εικόνας.

Παράρτημα III

CHEMSTAT

III.1 ΓΕΝΙΚΑ

Το *ChemStat* είναι ένα αρχείο του *Excel* που περιέχει προγράμματα-μακροεντολές για βασικά στατιστικά προβλήματα που αφορούν τη χημεία. Επιπλέον έχει τη δυνατότητα άμεσης μετατροπής σε *Add-in (Πρόσθετο)* με αποτέλεσμα να μπορεί να ενσωματωθεί στη λειτουργία του *Excel*. Το *ChemStat* διατίθεται ελεύθερα από την ιστοσελίδα

http://www.chem.auth.gr/index.php?st=56

στην οποία μπορούμε να πάμε από την ιστοσελίδα του Χημικού Τμήματος του Α.Π.Θ. (http://www.chem.auth.gr/index.php) ακολουθώντας την πορεία:

ΕΡΓΑΣΤΗΡΙΑ → ΦΥΣΙΚΗΣ ΧΗΜΕΙΑΣ → Νικήτας Παναγιώτης → CHEMSTAT

Όταν στο *ChemStat* επιλέγεται η εκτέλεση ενός προγράμματος συνήθως εμφανίζονται οδηγίες για τη σωστή εκτέλεσή του. Οι οδηγίες δίνονται στα αγγλικά, επειδή, ανάλογα με τις ρυθμίσεις ενός υπολογιστή, ενδέχεται να παρουσιαστούν προβλήματα με την ανάγνωση ελληνικών γραμματοσειρών καθιστώντας άχρηστες τις οδηγίες. Οδηγίες παρέχονται επίσης και στα φύλλα εργασίας του *ChemStat*.

III.2 ΔΥΝΑΤΟΤΗΤΕΣ ΤΟΥ CHEMSTAT

Τα στατιστικά προγράμματα *(Μακροεντολές - Macros)* που υπάρχουν στο *ChemStat* είναι τα παρακάτω, όπου το όνομα κάθε μακροεντολής δίνεται σε παρένθεση, ενώ ο τίτλος της στο μενού του *ChemStat* δίνεται με έντονη γραφή:

1. ***Error Propagation*** *(Propagation)*. Υπολογίζει το εκατοστιαίο σφάλμα ή την τυπική απόκλιση μιας ποσότητας z που έχει υπολογιστεί με βάση μια

σχέση $z = f(x_1, x_2, ..., x_n)$ όταν γνωρίζουμε τα εκατοστιαία σφάλματα ή τις τυπικές αποκλίσεις ή τον πίνακα συνδιασποράς (variance-covariance matrix) των $x_1, x_2, ..., x_n$. Οι τιμές των μεταβλητών $x_1, x_2, ..., x_n$ πρέπει να βρίσκονται στην ίδια στήλη ή γραμμή σε συνεχόμενα κελιά και το ίδιο πρέπει να ισχύει για τα εκατοστιαία σφάλματα ή τις τυπικές αποκλίσεις τους. Ο πίνακας συνδιασποράς των $x_1, x_2, ..., x_n$ χρησιμοποιείται όταν οι μεταβλητές αυτές δεν είναι ανεξάρτητες. Για να λειτουργήσει σωστά το πρόγραμμα αυτό πρέπει οι μεταβλητές $x_1, x_2, ..., x_n$ και οι τυπικές τους αποκλίσεις να είναι στα κελιά με τη μορφή αριθμού (και όχι ως αποτέλεσμα κάποιας συνάρτησης του *Excel*).

2. ***LS Polynomial*** *(LS_Polynomial)*. Το πρόγραμμα αυτό προσαρμόζει με τη μέθοδο των ελαχίστων τετραγώνων ένα πολυώνυμο βαθμού p σε δεδομένα x-y. Τα δεδομένα πρέπει να βρίσκονται σε στήλες με τη στήλη των τιμών του x αριστερά και του y δεξιά. Οι συντελεστές προσαρμογής, οι τυπικές τους αποκλίσεις, τα διαστήματα εμπιστοσύνης, οι τιμές της μεταβλητής t και ο πίνακας συνδιασποράς των συντελεστών προσαρμογής παρουσιάζονται κάτω από την περιοχή των δεδομένων x-y. Δεξιά της στήλης y παρουσιάζονται οι υπολογιζόμενες τιμές του y και δεξιά αυτής της στήλης παρουσιάζονται τα υπόλοιπα.

3. ***LS MultiLinear*** *(LS_MultiLinear)*. Προσαρμόζει με τη μέθοδο των ελαχίστων τετραγώνων την εξίσωση $y = c_0 + c_1x_1 + c_2x_2 + ... + c_px_p$ στα δεδομένα $y, x_1, x_2, ..., x_p$. Η διευθέτηση των δεδομένων σε στήλες γίνεται ως εξής. Η στήλη των τιμών του y είναι αριστερά, δεξιά της είναι η στήλη με τους συντελεστές βαρύτητας, w, εφόσον υπάρχουν, και δεξιά αυτής της στήλης τοποθετούνται οι στήλες με τις τιμές των $x_1, x_2, ..., x_p$. Τα αποτελέσματα εμφανίζονται όπως και στο προηγούμενο πρόγραμμα. Αν η στήλη του y είναι μικρότερη των στηλών $x_1, x_2, ..., x_p$, τότε το πρόγραμμα υπολογίζει το y και την τυπική του απόκλιση στις επιπλέον τιμές των x_i.

4. ***LS Significant*** *(LS_Significant)*. Προσδιορίζει μόνο τους στατιστικά σημαντικούς όρους $c_0, c_1, c_2, ... , c_p$ της εξίσωσης $y = c_0 + c_1x_1 + c_2x_2 + ... + c_px_p$ όταν αυτή προσαρμόζεται στα δεδομένα $y, x_1, x_2, ..., x_p$. Για τη διευθέτηση των δεδομένων και την παρουσίαση των αποτελεσμάτων ισχύει ότι και στο πρόγραμμα *LS Multilinear*. Σε αντίθεση με το *LS Multilinear*, δεν κάνει πρόβλεψη τιμών y σε επιπλέον τιμές x_i.

5. ***LS Optimum Polynomial*** *(LS_Optimum_Polynomial)*. Προσδιορίζει με τη μέθοδο των ελαχίστων τετραγώνων το βέλτιστο πολυώνυμο προσαρμογής στα δεδομένα x-y. Η διευθέτηση των δεδομένων γίνεται όπως στο *LS Polynomial*. Υπάρχουν τρεις επιλογές: *Automatic backward,*

Automatic forward και *Manual.*

6. ***LS Orthogonal polynomial*** *(LS_Ortho)*. Προσαρμόζει με τη μέθοδο των ελαχίστων τετραγώνων ένα ορθογώνιο πολυώνυμο βαθμού p στα δεδομένα x-y και υπολογίζει τις προβλεπόμενες τιμές y και τις αντίστοιχες παραγώγους y'. Η διευθέτηση των δεδομένων γίνεται όπως στο *LS Polynomial*. Δεξιά της στήλης y παρουσιάζονται οι υπολογιζόμενες τιμές του y και δεξιά αυτής της στήλης οι παράγωγοι y'.

7. ***Solver Errors*** *(Solver_Errors)*. Προσδιορίζει τις τυπικές αποκλίσεις των προσαρμόσιμων παραμέτρων που έχουν υπολογιστεί προηγουμένως με το πρόγραμμα *Επίλυση (Solver)* του *Excel*. Οι προσαρμόσιμες παράμετροι πρέπει να βρίσκονται στην ίδια στήλη και σε διαδοχικά κελιά και το ίδιο να ισχύει για τις υπολογιζόμενες τιμές της εξαρτημένης μεταβλητής. Οι τυπικές αποκλίσεις παρουσιάζονται δεξιά της στήλης των προσαρμόσιμων παραμέτρων και δεξιά των τυπικών αποκλίσεων παρουσιάζεται η τιμή s_y.

8. ***Calibration*** *(Calibration)*. Όταν η καμπύλη αναφοράς είναι ευθεία το πρόγραμμα υπολογίζει την τιμή του x_0 που αντιστοιχεί στην τιμή y_0 ή στο μέσο όρο των τιμών (y_{10}, y_{20}, y_{30}, ...), την τυπική απόκλιση του x_0 και το αντίστοιχο p% διάστημα εμπιστοσύνης. Επίσης υπολογίζει το ελάχιστο όριο ανίχνευσης και την τυπική του απόκλιση. Όταν η καμπύλη αναφοράς είναι παραβολή, το πρόγραμμα υπολογίζει μόνο την τιμή του x_0, την τυπική του απόκλιση και το p% διάστημα εμπιστοσύνης. Τα δεδομένα πρέπει να βρίσκονται σε στήλες με τη στήλη των τιμών του x αριστερά και του y δεξιά. Επίσης οι τιμές (y_{10}, y_{20}, y_{30}, ...) πρέπει να είναι σε μία στήλη και να παρεμβάλλονται τουλάχιστον δύο κενές στήλες μεταξύ αυτής της στήλης και της στήλης των τιμών y.

9. ***Standard Addition*** *(Addition)*. Το πρόγραμμα αυτό χρησιμοποιείται για τον προσδιορισμό της άγνωστης συγκέντρωσης (c), της τυπικής της απόκλισης, stdev(c), και του αντίστοιχου διαστήματος εμπιστοσύνης, p%(c), με την τεχνική *Standard Addition*. Τα δεδομένα πρέπει να βρίσκονται σε στήλες με τη στήλη των τιμών του x αριστερά και του y δεξιά.

10. ***Intersection of lines*** *(Intersection)*. Υπολογίζει το σημείο τομής (x) δύο ευθειών, την τυπική απόκλιση του x, stdev(x), και το αντίστοιχο διάστημα εμπιστοσύνης, p%(x).

11. ***Normality Test*** *(NormalityTest)*. Ελέγχει την κανονικότητα των τιμών ενός ή περισσοτέρων δειγμάτων με το κριτήριο *Anderson-Darling*. Τα δείγματα πρέπει να είναι σε στήλες με ελάχιστο πλήθος τιμών το 6. Όταν

υπάρχουν πολλά δείγματα, πρώτο εισάγεται το μεγαλύτερο. Επιπλέον υπάρχει δυνατότητα γραφικών ελέγχων με τα προγράμματα ***Univariate Q-Q plot*** *(PlotsQQ_PP)* και ***Bivariate chi-square plot*** *(ChiSquarePlot)*. Με το πρώτο ελέγχουμε τη *μονοδιάστατη κανονικότητα* και με το δεύτερο τη *δισδιάστατη κανονικότητα*.

12. ***Test for Outliers*** *(Outliers)*. Ελέγχει αν σε ένα δείγμα η τιμή που απέχει περισσότερο από τη μέση τιμή είναι ακραία χρησιμοποιώντας το κριτήριο του *Grubbs*. Οι τιμές του δείγματος πρέπει να είναι σε μία στήλη.

13. ***One Sample Tests*** *(OneSampleTests)*. Ελέγχει την κανονικότητα των τιμών ενός δείγματος με το κριτήριο *Anderson-Darling* και ακολούθως ελέγχει αν η μέση τιμή του δείγματος είναι στατιστικά ίση με την τιμή που θέλουμε να συγκρίνουμε. Υπάρχει ο παραμετρικός έλεγχος t και η δυνατότητα μη παραμετρικών ελέγχων με τις μεθόδους *Monte-Carlo με αντιμεταθέσεις* και *Bootstrap*.

14. ***Test of Variances*** *(Test_Variances)*. Ελέγχει την ομοιογένεια της διασποράς μεταξύ δύο ή περισσοτέρων δειγμάτων χρησιμοποιώντας τα κριτήρια *F*, *Levene* και *Brown-Forsythe*.

15. ***Independent Samples*** *(TwoIndependentSamples)*. Εκτελεί όλους τους δυνατούς ελέγχους, παραμετρικούς και μη παραμετρικούς, μεταξύ δύο ανεξάρτητων δειγμάτων. Συγκεκριμένα στον παραμετρικό έλεγχο χρησιμοποιείται ο έλεγχος t, ενώ στον μη παραμετρικό έλεγχο το κριτήριο *Mann-Whitney*. Επιπλέον υπάρχει η επιλογή της μεθόδου *Monte-Carlo με αντιμεταθέσεις*.

16. ***Paired Samples*** *(PairedSamples)*. Εκτελεί όλους τους δυνατούς ελέγχους, παραμετρικούς και μη παραμετρικούς, μεταξύ δύο δειγμάτων που σχηματίζουν ζεύγος. Στον παραμετρικό έλεγχο χρησιμοποιείται ο έλεγχος t και στον μη παραμετρικό έλεγχο το κριτήριο *Wilcoxon*, ενώ υπάρχει και η δυνατότητα εφαρμογής της μεθόδου *Monte-Carlo με αντιμεταθέσεις*.

17. ***Bivariate*** *(Correlation).* Ελέγχει αν δύο δείγματα σχετίζονται γραμμικά. Ο έλεγχος είναι και παραμετρικός (*Pearson correlation coefficient*) και μη παραμετρικός (*Spearman correlation coefficient*) με δυνατότητα εφαρμογής της μεθόδου *Monte-Carlo με αντιμεταθέσεις*.

18. ***Partial Correlation*** *(PartialCorrelation).* Εκτελεί μερική συσχέτιση δύο δειγμάτων ως προς μία ή περισσότερες ρυθμιστικές μεταβλητές.

19. ***Mantel Tests*** *(Mantel).* Εκτελεί τους στατιστικούς ελέγχους *Mantel* και *partial Mantel*.

20. ***Distances*** *(Distances).* Υπολογίζει τους πίνακες αποστάσεων *Mahalanobis*, *Euclidean* και *Jaccard* μεταξύ ομάδων δειγμάτων.

21. ***Chi-square test*** *(Chi_square).* Ελέγχει αν σε έναν πίνακα με κατηγορικά δεδομένα υπάρχει ή όχι επίδραση στατιστικά σημαντική του παράγοντα που επηρεάζει τις τιμές του πίνακα.

22. ***Fisher exact test*** *(Exact_Fisher).* Ελέγχει αν σε έναν πίνακα 2x2 με κατηγορικά δεδομένα υπάρχει ή όχι επίδραση στατιστικά σημαντική του παράγοντα που επηρεάζει τις τιμές του πίνακα.

23. ***Holm Bonferroni correction*** *(Holm_Bonferroni).* Εφαρμόζει τη διόρθωση *Holm-Bonferroni* σε έναν πίνακα με τιμές p-value. Ο πίνακας αυτός πρέπει να έχει στα αριστερά του μια στήλη με τίτλους και το ίδιο πρέπει να συμβαίνει με την πρώτη γραμμή.

24. ***ANOVA parametric*** *(ANOVA_parametric).* Εκτελεί παραμετρική μονοπαραγοντική *ANOVA* και διπαραγοντική *ANOVA* χωρίς αλληλεπιδράσεις. Επιπλέον εκτελεί ανά δύο (pairwise) ελέγχους μεταξύ όλων των δυνατών δειγμάτων χρησιμοποιώντας τη διόρθωση *Holm-Bonferroni*.

25. ***ANOVA non-parametric*** *(NP_ANOVA).* Εκτελεί μη παραμετρική *ANOVA* μονοπαραγοντική (*Kruskal-Wallis*) και διπαραγοντική (*Friedman*). Επιπλέον εκτελεί ανά δύο ελέγχους μεταξύ όλων των δυνατών δειγμάτων.

26. ***MANOVA parametric*** *(MANOVA).* Εκτελεί παραμετρική *MANOVA* σε όλες τις ομάδες και ακολούθως συγκρίσεις των ομάδων ανά δύο.

27. ***MANOVA non-parametric*** *(NP_MANOVA).* Εκτελεί δύο παραλλαγές μη παραμετρικής *MANOVA*.

28. ***PCA*** *(PCA).* Εκτελεί *Ανάλυση σε Κύριες Συνιστώσες (Principal Component Analysis)*.

29. ***Discriminant Analysis*** *(**LDA**).* Εκτελεί *Γραμμική Διαχωριστική Ανάλυση*.

Επιπλέον από το *Charts* υπάρχει η δυνατότητα κατασκευής *θηκογραμμάτων (boxplots)* και *ιστογραμμάτων (histograms)* με την προϋπόθεση ότι εργαζόμαστε στο *Excel* 2007 ή 2010.

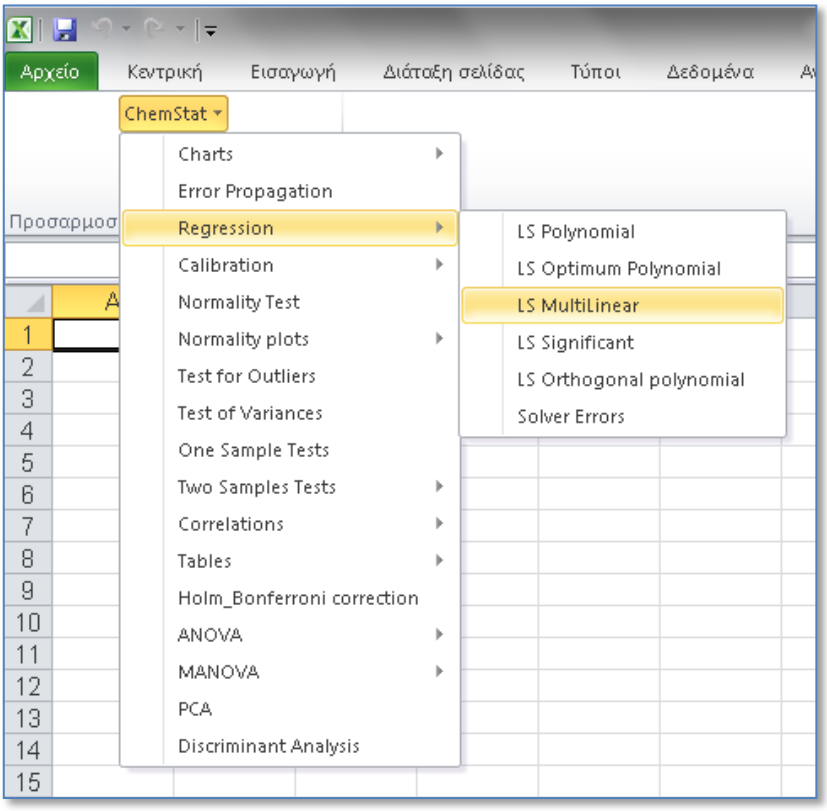
Τα παραπάνω προγράμματα εμφανίζονται ομαδοποιημένα σε αναδυόμενη λίστα αν κάνουμε *κλικ* στο κουμπί *ChemStat* της λωρίδας *Πρόσθετα* (σχήμα ΙΙΙ.1) με την προϋπόθεση ότι έχουμε εγκαταστήσει το *ChemStat* ως πρόσθετο, όπως περιγράφεται παρακάτω.

III.3 ΧΡΗΣΗ ΤΟΥ CHEMSTAT

Για να χρησιμοποιηθεί το *ChemStat* μπορούμε να εργαστούμε ως εξής:

α) Με βάση το αρχείο *ChemStat* δημιουργούμε ένα αρχείο *Add-in* *(Πρόσθετο)*

- Πρώτα αποθηκεύουμε το αρχείο *ChemStat* που κατεβάζουμε από το διαδίκτυο στην έκδοση του *Excel* που χρησιμοποιούμε, δηλαδή ως *ChemStat.xls* αν χρησιμοποιούμε το *Excel 2003* ή *ChemStat.xlsm* με δυνατότητα μακροεντολών αν χρησιμοποιούμε το *Excel* 2007 ή 2010.



Σχήμα III.1 Ομαδοποίηση προγραμμάτων στο *ChemStat*

- Ακολούθως ανοίγουμε το αρχείο που αποθηκεύσαμε στο προηγούμενο βήμα και το αποθηκεύουμε ως *Πρόσθετο του Excel* και συγκεκριμένα ως *ChemStat.xla* αν χρησιμοποιούμε το *Excel* 2003 ή *ChemStat.xlsam* αν χρησιμοποιούμε το *Excel* 2007/10. Το αρχείο αποθηκεύεται στο χώρο αποθήκευσης των *Πρόσθετων (Add ins).*

- Κλείνουμε όλα τα αρχεία του *Excel* και ανοίγουμε ένα νέο. 1) Στο *Excel* 2003 πηγαίνουμε *Εργαλεία → Πρόσθετα (Tools → Add ins)* και στο παράθυρο που ανοίγει ενεργοποιούμε την επιλογή *ChemStat*. 2) Στο *Excel* 2007/2010 πηγαίνουμε *Κουμπί Office ή Αρχείο → Επιλογές του Excel → Πρόσθετα (Office Button ή File → Excel Options → Add ins)* και κάνουμε *κλικ στο Μετάβαση (Go)* με την προϋπόθεση ότι στο αντίστοιχο πλαίσιο υπάρχει η επιλογή *Πρόσθετα του Excel (Excel add-ins). Σ*το παράθυρο που ανοίγει ενεργοποιούμε την επιλογή *ChemStat*.

- Στο *Excel* 2003 εμφανίζεται μια μπάρα με το κουμπί *ChemStat*, από το οποίο εμφανίζεται μια αναδυόμενη λίστα με όλες τις μακροεντολές. 2) Στο *Excel* 2007/10 το *ChemStat* με τη λίστα των μακροεντολών εμφανίζεται στη λωρίδα *Πρόσθετα* (σχήμα ΙΙΙ.1).

β) Εργαζόμαστε στο τρέχον βιβλίο εργασίας έχοντας ανοικτό το αρχείο *ChemStat*

Σε αυτή την περίπτωση έχουμε τη δυνατότητα άμεσης εκτέλεσης των μακροεντολών. Όμως, επειδή οι μακροεντολές ενεργούν μερικές φορές ως ιοί, στην αρχική εγκατάσταση του *Excel* η δυνατότητα εκτέλεσης των μακροεντολών είναι συνήθως απενεργοποιημένη. Για το λόγο αυτό για να εκτελέσουμε κάποιο από τα προγράμματα του *ChemStat* πρέπει να προηγηθούν οι ακόλουθες ενέργειες:

1) *Excel* 2003. Πηγαίνουμε *Εργαλεία → Μακροεντολές → Ασφάλεια (Tools → Macro → Security)* και επιλέγουμε *Μεσαία (Medium)*. Ακολούθως για να επιλέξουμε και να τρέξουμε κάποιο πρόγραμμα πηγαίνουμε *Εργαλεία → Μακροεντολές → Μακροεντολή (Tools → Macro → Macros)* και στο παράθυρο που ανοίγει κάνουμε διπλό *κλικ* στο πρόγραμμα που επιλέγουμε.

2) *Excel* 2007. Πηγαίνουμε *Κουμπί Office → Επιλογές του Excel → Δημοφιλείς (Office Button → Excel Options → Popular)* και επιλέγουμε *Εμφάνιση καρτέλας "Προγραμματιστής" στην κορδέλα (Show Developer tab in the Ribbon).* Ακολούθως πηγαίνουμε *Προγραμματιστής → Ασφάλεια μακροεντολών (Developer → Macros Security)* και επιλέγουμε *Ενεργοποίηση όλων των μακροεντολών (Enable all macros).* Τέλος, πηγαίνουμε

Προγραμματιστής → Μακροεντολές (Developer → Macros) και στο παράθυρο που ανοίγει κάνουμε διπλό *κλικ* στο πρόγραμμα που επιθυμούμε.

3) *Excel* 2010. Ο *Προγραμματιστής* εμφανίζεται από *Αρχείο → Επιλογές → Προσαρμογή κορδέλας (File → Options → Customize Ribbon)* όπου επιλέγουμε *Προγραμματιστής (Developer).* Οι άλλες ενέργειες είναι ίδιες με αυτές του *Excel* 2007.

Παράρτημα IV

ΕΦΑΡΜΟΓΗ ΜΗ-ΠΑΡΑΜΕΤΡΙΚΩΝ ΕΛΕΓΧΩΝ ΣΤΟ EXCEL

IV.1 ΒΑΘΜΟΙ ΚΑΙ ΔΕΣΜΟΙ

Οι περισσότεροι έλεγχοι στη μη-παραμετρική στατιστική στηρίζονται στη μετατροπή των αρχικών δεδομένων σε βαθμούς, όπου σε ένα δείγμα ονομάζουμε ***βαθμό*** (*rank*) της τιμής x το πλήθος r των τιμών του δείγματος που είναι μικρότερα ή ίσα με το x. Η μετατροπή των τιμών ενός δείγματος σε βαθμούς είναι άμεση, αρκεί να προσέξουμε την περίπτωση που στο δείγμα υπάρχουν δεσμοί. Ονομάζουμε ***δεσμό*** (*tie*) όταν μια τιμή εμφανίζεται περισσότερο από μια φορά στο δείγμα.

Για να υπολογίσουμε τους βαθμούς τέτοιων τιμών έχουν προταθεί διάφοροι τρόποι, εκ των οποίων ο επικρατέστερος χρησιμοποιεί σε κάθε δεσμό τους μέσους όρους των βαθμών που υπάρχουν στο δεσμό. Για παράδειγμα, στο δείγμα {2, 2, 5, 5, 5, 11} υπάρχουν δύο δεσμοί με δύο τιμές ο πρώτος και τρεις ο δεύτερος. Συνεπώς, οι βαθμοί των τιμών είναι {1.5, 1.5, 4, 4, 4, 6}, διότι θεωρούμε ότι στο δεσμό (2, 2) αντιστοιχούν οι βαθμοί 1 και 2 με μέσο όρο 1.5 και στο δεσμό (5, 5, 5) οι βαθμοί 3, 4, 5 με μέσο όρο 4.

Στο *Excel* η μετατροπή τιμών σε βαθμούς μπορεί να γίνει με τη συνάρτηση *RANK*. Όμως η συνάρτηση αυτή θεωρεί ως βαθμό της τιμής x σε ένα δείγμα το πλήθος των τιμών του δείγματος που είναι μικρότερα από το x συν ένα. Έτσι το δείγμα {2, 2, 5, 5, 5, 11} μετατρέπεται στο {1, 1, 3, 3, 3, 6}. Σε αυτή την περίπτωση θα πρέπει να επεμβαίνουμε και να διορθώνουμε τους βαθμούς των δεσμών με βάση τη σύμβαση του μέσου όρου που αναφέραμε παραπάνω.

IV.2 ΚΡΙΤΗΡΙΟ MANN-WHITNEY

Για την εφαρμογή του κριτηρίου αυτού ενώνουμε τα δύο δείγματα με πλήθος τιμών m_1 και m_2, αντίστοιχα, και υπολογίζουμε τα αθροίσματα των βαθμών των δειγμάτων στο ενιαίο δείγμα. Αν αυτά είναι R_1 και R_2, αντίστοιχα, υπολογίζουμε τις ποσότητες

$$U_1 = m_1m_2 + m_1(m_1+1)/2 - R_1 \quad \text{και} \quad U_2 = m_1m_2 + m_2(m_2+1)/2 - R_2$$

και επιλέγουμε ως συνάρτηση στατιστικού ελέγχου την ελάχιστη από αυτές $U = \min(U_1, U_2)$. Για σχετικά μεγάλα δείγματα η U ακολουθεί ασυπτωματικά την κανονική κατανομή με

$$\mu = \frac{m_1m_2}{2} \quad \text{και} \quad \sigma^2 = \frac{m_1m_2(m_1+m_2-1)}{12}$$

Αν υπάρχουν δεσμοί, η τυπική απόκλιση σ πρέπει να υπολογιστεί από τη σχέση

$$\sigma^2 = \frac{m_1m_2}{n(n-1)}\left[\frac{n^3-n}{12} - \sum_{j=1}^{g}\frac{t_j^3-t_j}{12}\right]$$

όπου $n = m_1+m_2$, g είναι το πλήθος των δεσμών και t_j το πλήθος των βαθμών που υπάρχουν στο δεσμό j = 1, 2, …, g.

Παράδειγμα IV.1

Στην άσκηση 7.14 δίνεται το ετήσιο βροχομετρικό ύψος σε mm σε δύο γεωγραφικές περιοχές και ζητείται να εξετασθεί αν υπάρχει στατιστικά σημαντική διαφοροποίηση στις τιμές.

Περιοχή Α		Περιοχή Β	
102	178	115	117
210	78	135	302
99	102	330	280
109	85	95	205

◆ Μεταφέρουμε τα δεδομένα του παραπάνω πίνακα σε ένα φύλλο του *Excel* και συγκεκριμένα τοποθετούμε σε μία στήλη, έστω στην Α τις τιμές των δύο δειγμάτων κάτω από τον τίτλο *samples*, ενώ στη διπλανή στήλη δημιουργούμε τη μεταβλητή *groups* με τιμές 1 για κάθε τιμή της στήλης *samples* που ανήκει στην περιοχή Α και 2 όταν ανήκει στην περιοχή Β.

Ακολούθως επιλέγουμε την περιοχή A2:A17, πηγαίνουμε *Κεντρική (Home) → Ταξινόμηση & φιλτράρισμα (Sort & Filter)* και επιλέγουμε *Ταξινόμηση από το μικρότερο προς το μεγαλύτερο (Sort A to Z)* και *Επέκταση της επιλογής (Expand the Selection)*. Οι δύο πρώτες στήλες θα εμφανίζονται όπως στο σχήμα IV.1.

	A	B	C	D	E	F	G	H
1	**samples**	**groups**	**Βαθμοί**	**Βαθμοί 1**	**Βαθμοί 2**			
2	78	1	1	1				
3	85	1	2	2				
4	95	2	3		3		U=	13
5	99	1	4	4			μ=	32
6	102	1	5.5	5.5			σ=	8.944272
7	102	1	5.5	5.5			σ(corrected)=	9.514901
8	109	1	7	7			Z=	-2.12426
9	115	2	8		8		Z(corrected)=	-1.99687
10	117	2	9		9			
11	135	2	10		10		p-value=	0.033648
12	178	1	11	11			p(corrected)=	0.04584
13	205	2	12		12			
14	210	1	13	13				
15	280	2	14		14			
16	302	2	15		15			
17	330	2	16		16			
18								
19			R_i=	49	87			
20			U_i=	51	13			

Σχήμα IV.1. Πίνακας ανάλυσης δειγμάτων με το κριτήριο *Mann-Whitney*

Στη στήλη C υπολογίζουμε τους βαθμούς των τιμών του ενιαίου δείγματος. Συνεπώς, στο C2 εισάγουμε τον τύπο =RANK(A2;A$2:A$17;1), πατάμε *Enter* και συμπληρώνουμε όλη τη στήλη με τη διαδικασία της αυτόματης συμπλήρωσης. Για λόγους που ήδη έχουμε εξηγήσει, η συνάρτηση *RANK* αντιστοιχεί στις δύο τιμές 102 που υπάρχουν στις γραμμές 6 και 7 τους βαθμούς 5 και 5. Τους διορθώνουμε σε 5.5 και 5.5, όπως φαίνεται στο παραπάνω σχήμα.

Στο κελί D1 εισάγουμε τον τίτλο "Βαθμοί 1", στο D2 τον τύπο = IF(B2=1;C2;"") και συμπληρώνουμε τη στήλη με τη διαδικασία της αυτόματης συμπλήρωσης. Με τη συνάρτηση αυτή επιλέγουμε από τους βαθμούς που υπάρχουν στη στήλη C μόνο εκείνους που αντιστοιχούν στο πρώτο δείγμα. Με τον ίδιο τρόπο, χρησιμοποιώντας στο E2 τον τύπο

=IF(B2=2;C2;"") συμπληρώνουμε τη στήλη Ε, στο κελί Ε1 της οποίας έχουμε εισάγει τον τίτλο "Βαθμοί 2".

Το άθροισμα των τιμών στις δύο αυτές στήλες μας δίνει R_1 = 49 και R_2 = 87, από τις οποίες προκύπτει ότι U_1 = 51 και U_2 = 13. Συνεπώς U = 13. Για τη μέση τιμή μ έχουμε μ = 8*8/2 = 32, ενώ η μη διορθωμένη τυπική απόκλιση υπολογίζεται με τον τύπο =SQRT(8*8*(16-1)/12) και ισούται με 8.944272. Για τον υπολογισμό της διορθωμένης τιμής σ λαμβάνουμε υπόψη ότι το πλήθος των δεσμών είναι 1 με δύο βαθμούς, δηλαδή g = 1 με t_1 = 2. Χρησιμοποιώντας αυτές τις τιμές παίρνουμε σ =SQRT((8*8/16/15)*((16^3-16)/12-(2^3-2)/12)) = 9.514901.

Εφόσον η συνάρτηση U ακολουθεί ασυμπτωτικά την κανονική κατανομή, η μεταβλητή Z = (U-μ)/σ θα ακολουθεί την τυπικά κανονική κατανομή. Οι τιμές της Z, χωρίς διόρθωση και με διόρθωση της σ, είναι Z = -2.12426 και Z(corrected) = -1.99687. Από τις τιμές αυτές υπολογίζονται οι τιμές p-value χρησιμοποιώντας τη συνάρτηση =2*NORMSDIST(Z). Παίρνουμε: p-value = 0.033648 και p(corrected) = 0.04584. Η τελευταία τιμή βρίσκεται σε πλήρη ταύτιση με τις αντίστοιχες τιμές που δίνουν τα προγράμματα *ChemStat* και *SPSS*.

IV.3 ΚΡΙΤΗΡΙΟ WILCOXON

Για την εφαρμογή του κριτηρίου υπολογίζουμε πρώτα τις διαφορές d_i μεταξύ των τιμών των δύο δειγμάτων, ακολούθως υπολογίζουμε τις απόλυτες τιμές $|d_i|$ και σε κάθε τιμή $|d_i|$ αντιστοιχούμε το βαθμό της. Στη συνέχεια σε κάθε βαθμό αντιστοιχούμε το πρόσημο του d_i και υπολογίζουμε το άθροισμα T^+ των βαθμών των θετικών d_i και το αντίστοιχο άθροισμα T^- των βαθμών των αρνητικών d_i. Ως στατιστική συνάρτηση ελέγχου επιλέγεται η ελάχιστη των τιμών T^+, T^-, δηλαδή η συνάρτηση $W = \min(T^+, T^-)$. Στους υπολογισμούς δεν λαμβάνονται υπόψη μηδενικές διαφορές, d_i = 0.

Σε μεγάλα δείγματα η τυχαία μεταβλητή W ακολουθεί ασυμπτωτικά την κανονική κατανομή με

$$\mu = \frac{m(m+1)}{4} \quad \text{και} \quad \sigma^2 = \frac{m(m+1)(2m+1)}{24}$$

όπου m είναι το πλήθος των τιμών $d_i \neq 0$. Αν υπάρχουν δεσμοί η τυπική απόκλιση σ πρέπει να υπολογιστεί από τη σχέση

$$\sigma^2 = \frac{m(m+1)(2m+1)}{24} - \sum_{j=1}^{g} \frac{t_j^3 - t_j}{48}$$

Παράδειγμα IV.2

Στην άσκηση 7.15 ζητείται, με βάση τα δεδομένα του παρακάτω πίνακα, να ελεγχθεί η υπόθεση ότι η συγκέντρωση της βιταμίνης C σε ειδικά αρτοσκευάσματα, που παρέχονται από προγράμματα διεθνούς βοήθειας για τη διατροφή παιδιών του τρίτου κόσμου, ελαττώνεται με το χρόνο.

t = 0	53	33	45	35	48	36	39	55	46
t = 6 μήνες	50	30	43	35	48	35	39	48	42

◆ Η ανάλυση των δεδομένων δίνεται στο σχήμα IV.2. Σύμφωνα με τη διευθέτηση στο σχήμα αυτό, τα δείγματα τοποθετούνται στις στήλες A και B, στη στήλη C υπολογίζεται η διαφορά τιμών μεταξύ των δύο πρώτων στηλών και στη D η απόλυτη τιμή των διαφορών, ενώ διαγράφονται διαφορές ίσες με 0. Στη στήλη E υπολογίζονται οι βαθμοί των τιμών |d|. Για το σκοπό αυτό, στο E2 εισάγουμε τον τύπο =RANK(D2;D$2:D$10;1), πατάμε *Enter* και συμπληρώνουμε όλη τη στήλη με τη διαδικασία της αυτόματης συμπλήρωσης.

Όμως η εικόνα που θα πάρουμε δεν είναι αυτή του σχήματος IV.2. Συγκεκριμένα, στις τιμές |d| = 1, 1, 1 η *RANK* αντιστοιχεί τους βαθμούς 1, 1, 1 και στις τιμές 3, 3 τους βαθμούς 5, 5. Επειδή στο δεσμό (1, 1, 1) θεωρούμε ότι αντιστοιχούν οι βαθμοί 1, 2, 3 με μέσο όρο 2, διορθώνουμε τους βαθμούς της στήλης E μετατρέποντας τις τιμές 1, 1, 1 σε 2, 2, 2. Επίσης στο δεσμό (3, 3) θεωρούμε ότι αντιστοιχούν οι βαθμοί 5 και 6 με μέσο όρο 5.5. Συνεπώς διορθώνουμε και τους βαθμούς 5, 5 σε 5.5, 5.5, όπως φαίνεται στο σχήμα IV.2.

Στη συνέχεια εισάγουμε στα κελιά F1 και G1 τους τίτλους "T+" και "T-", αντίστοιχα, ενώ στα κελιά F2 και G2 τους τύπους = IF(C2>0;E2;"") και IF(C2<0;E2;"") και συμπληρώνουμε τις στήλες με τη διαδικασία της αυτόματης συμπλήρωσης. Με τις συναρτήσεις αυτές επιλέγουμε τους βαθμούς των θετικών d_i στη στήλη F και τους βαθμούς των αρνητικών d_i στη στήλη G. Στα κελιά F12 και G12 υπολογίζονται τα αθροίσματα T^+ και T^- χρησιμοποιώντας τη συνάρτηση *SUM*. Από αυτά προκύπτει ότι W = 4.

	A	B	C	D	E	F	G	H	I	J
1	**sample1**	**sample2**	**d**	**|d|**	**Βαθμοί**	**T+**	**T-**			
2	53	50	3	3	5.5	5.5				
3	33	30	3	3	5.5	5.5			W=	4
4	45	43	2	2	4	4			μ=	18
5	35	36	-1	1	2		2		σ=	7.141428
6	48	48	0		#Δ/Υ				σ(corrected)=	7.097535
7	36	35	1	1	2	2			Z=	-1.96039
8	39	40	-1	1	2		2		Z(corrected)=	-1.97252
9	55	48	7	7	8	8				
10	46	42	4	4	7	7			p-value=	0.04995
11									p(correctd)=	0.048551
12					**sum=**	32	4			

Σχήμα IV.2. Πίνακας ανάλυσης δειγμάτων με το κριτήριο *Wilcoxon*

Για τη μέση τιμή μ έχουμε μ = 8*9/4 = 18, ενώ η μη διορθωμένη τυπική απόκλιση υπολογίζεται με τον τύπο = SQRT(8*9*(16+1)/24) και ισούται με 7.141428. Για τον υπολογισμό της διορθωμένης τιμής σ λαμβάνουμε υπόψη ότι το πλήθος των δεσμών είναι 2 με 3 και 2 βαθμούς, αντίστοιχα. Δηλαδή g = 2 με t_1 = 3 και t_2 = 2. Χρησιμοποιώντας αυτές τις τιμές παίρνουμε σ = SQRT(8*9*(16+1)/24-(3^3-3)/48-(2^3-2)/48) = 7.097535.

Με βάση τις παραπάνω τιμές υπολογίζουμε τη μεταβλητή Z = (W-μ)/σ που ακολουθεί την τυπικά κανονική κατανομή. Οι τιμές της Z, χωρίς διόρθωση και με διόρθωση της σ, είναι Z = -1.96039 και Z(corrected) = -1.97252, αντίστοιχα. Όπως και στο προηγούμενο κριτήριο, οι τιμές p-value υπολογίζονται χρησιμοποιώντας τη συνάρτηση =2*NORMSDIST(Z). Παίρνουμε: p-value = 0.04995 και p(corrected) = 0.048551, με την τελευταία τιμή να βρίσκεται σε πλήρη ταύτιση με τις αντίστοιχες τιμές που δίνουν τα προγράμματα *ChemStat* και *SPSS*.

IV.4 ΚΡΙΤΗΡΙΟ KRUSKAL-WALLIS

Ενώνουμε όλα τα δείγματα σε ένα και υπολογίζουμε την ποσότητα

$$H = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{R_i^2}{m_i} - 3(n+1)$$

όπου k είναι το πλήθος των δειγμάτων, m_i είναι το μέγεθος του δείγματος i, $n = m_1 + m_2 + \ldots + m_k$ και R_i είναι το άθροισμα των βαθμών του δείγματος i στο ενιαίο δείγμα μεγέθους n. Αν υπάρχουν δεσμοί η συνάρτηση H διορθώνεται με βάση τη σχέση

$$H(\text{corrected}) = \frac{H}{1 - \sum_{j=1}^{g} \frac{t_j^3 - t_j}{n^3 - n}}$$

Οι συναρτήσεις H και H(corrected) ακολουθούν ασυμπτωτικά την κατανομή χ^2 με k-1 βαθμούς ελευθερίας.

Παράδειγμα IV.3

Στην άσκηση 8.5 το ετήσιο βροχομετρικό ύψος σε mm στα νησιά του Ιονίου πελάγους και του Βορείου και Νοτίου Αιγαίου δίνεται στον παρακάτω πίνακα και ζητείται να εξετασθεί αν οι διαφοροποιήσεις που φαίνονται στον πίνακα είναι στατιστικά σημαντικές.

Ιόνιο	Βόρειο Αιγαίο	Νότιο Αιγαίο
700	590	295
1150	710	720
1350	920	460
990	605	510
880	660	730
1560		440
685		410
		295

◆ Ουσιαστικά ακολουθούμε τα βήματα εφαρμογής του κριτηρίου *Mann-Whitney*, όπως στο παράδειγμα IV.1. Συγκεκριμένα σε ένα φύλλο του *Excel* τοποθετούμε στη στήλη A τις τιμές των τριών δειγμάτων κάτω από τον τίτλο *samples*, ενώ στη διπλανή στήλη δημιουργούμε τη μεταβλητή *groups* με τιμές 1, 2 και 3 που σημειώνουν σε ποια γεωγραφική περιοχή ανήκει κάθε τιμή του ενιαίου δείγματος samples. Ακολούθως επιλέγουμε την περιοχή A2:A21, πηγαίνουμε *Κεντρική (Home)* → *Ταξινόμηση & φιλτράρισμα (Sort & Filter)* και επιλέγουμε *Ταξινόμηση από το μικρότερο προς το μεγαλύτερο (Sort A to Z)* και *Επέκταση της επιλογής (Expand the Selection)*. Συνεχίζουμε υπολογίζοντας στη στήλη C τους βαθμούς των τιμών του ενιαίου δείγματος. Συγκεκριμένα στο C2 εισάγουμε τον τύπο =RANK(A2;A$2:A$21;1), πατάμε *Enter* και συμπληρώνουμε όλη τη στήλη με τη διαδικασία της αυτόματης συμπλήρωσης. Διορθώνουμε μόνο τους δύο πρώτους βαθμούς σε 1.5 και 1.5, όπως φαίνεται στο σχήμα IV.3.

	A	B	C	D	E	F	G	H	I
1	**samples**	**groups**	**R**	**R1**	**R2**	**R3**			
2	295	3	1.5			1.5			
3	295	3	1.5			1.5		m1=	7
4	410	3	3			3		m2=	5
5	440	3	4			4		m3=	8
6	460	3	5			5		n=	20
7	510	3	6			6			
8	590	2	7		7			H=	10.06776
9	605	2	8		8			H(corrected)=	10.07533
10	660	2	9		9				
11	685	1	10	10				p-value=	0.006514
12	700	1	11	11				p(corrected)=	0.006489
13	710	2	12		12				
14	720	3	13			13			
15	730	3	14			14			
16	880	1	15	15					
17	920	2	16		16				
18	990	1	17	17					
19	1150	1	18	18					
20	1350	1	19	19					
21	1560	1	20	20					
22									
23			sum=	110	52	48			

Σχήμα IV.3. Πίνακας ανάλυσης δειγμάτων με το κριτήριο *Kruskal-Wallis*

Στις τρεις επόμενες στήλες επιλέγουμε από τους βαθμούς που υπάρχουν στη στήλη C τους βαθμούς που αντιστοιχούν σε κάθε αρχικό δείγμα. Έτσι, στα κελιά D1, E1, F1 εισάγουμε τους τίτλους "R1", "R2", "R3", στα κελιά D2, E2, F2 πληκτρολογούμε τους τύπους =IF(B2=1;C2;""), =IF(B2=2;C2;""), =IF(B2=3;C2;"") και συμπληρώνουμε την περιοχή D2:F21 με τη διαδικασία της αυτόματης συμπλήρωσης.

Τα αθροίσματα των τιμών στις στήλες D, E, F είναι $R_1 = 110$, $R_2 = 52$ και $R_3 = 48$. Επιπλέον έχουμε $m_1 = 7$, $m_2 = 5$, $m_3 = 8$ και $n = 20$. Από τις τιμές αυτές προκύπτει ότι H = 10.06776. Για τη διορθωμένη H λαμβάνουμε υπόψη ότι g = 1 με $t_1 = 2$ και παίρνουμε

H(corrected) = 10.06776/(1-(2^3-2)/(20^3-20)) = 10.07533

Εφόσον οι συναρτήσεις H και H(corrected) ακολουθούν ασυμπτωτικά την κατανομή χ^2 με k-1 = 2 βαθμούς ελευθερίας, οι τιμές p-value μπορούν να υπολογιστούν χρησιμοποιώντας τη συνάρτηση =CHIDIST(H;2) Παίρνουμε: p-value = 0.006514 και p(corrected) = 0.006489 (σχήμα IV.3). Όπως αναμένεται και εδώ, η τελευταία τιμή βρίσκεται σε πλήρη ταύτιση με τις αντίστοιχες τιμές των προγραμμάτων *ChemStat* και *SPSS*.

IV.5 ΚΡΙΤΗΡΙΟ FRIEDMAN

Η στατιστική συνάρτηση ελέγχου είναι η μεταβλητή F:

$$F = \frac{12}{mk(k+1)} \sum_{i=1}^{k} R_i^2 - 3m(k+1)$$

όπου k είναι το πλήθος των δειγμάτων, m είναι το μέγεθος των δειγμάτων κοινό σε όλα τα δείγματα και R_i είναι το άθροισμα των βαθμών του i (=1, 2, …, k), όπου όμως οι βαθμοί υπολογίζονται ανά γραμμή. Αν υπάρχουν δεσμοί η συνάρτηση F διορθώνεται χρησιμοποιώντας τη σχέση

$$F(\text{corrected}) = \frac{F}{1 - \sum_{j=1}^{g} \frac{t_j^3 - t_j}{n(k^3 - k)}}$$

Οι συναρτήσεις F και F(corrected) ακολουθούν ασυμπτωτικά την κατανομή χ^2 με k-1 βαθμούς ελευθερίας.

Παράδειγμα IV.4

Στο παράδειγμα 8.5 ελέγχθηκε μία τεχνική καταστροφής ρύπων κάτω από διαφορετικές συνθήκες καταλυτών και θερμοκρασίας και με βάση τις τιμές του παρακάτω πίνακα εξετάσθηκε αν η επίδραση αυτών των δύο παραγόντων στην επί τοις εκατό ποσότητα των ρύπων που καταστρέφονται είναι στατιστικά σημαντική.

	Καταλύτης Α	Καταλύτης Β	Καταλύτης C
T = 25 °C	90	80	60
T = 35 °C	90	75	60
T = 45 °C	85	70	65
T = 55 °C	80	75	55

◆ Θα επανεξετάσουμε αναλυτικά μόνο την επίδραση του τύπου του καταλύτη. Συνεπώς μεταφέρουμε τα δεδομένα σε τρεις στήλες, όπως στο σχήμα IV.4 και προσδιορίζουμε τους βαθμούς κάθε τιμής στην αντίστοιχη γραμμή. Δηλαδή, οι βαθμοί υπολογίζονται σε κάθε γραμμή δεδομένων. Συγκεκριμένα εισάγουμε στο κελί D2 τον τύπο =RANK(A2;$A2:$C2;1) και εφαρμόζουμε τη διαδικασία της αυτόματης συμπλήρωσης πρώτα οριζοντίως κατά μήκος της γραμμής 2 και ακολούθως καθέτως μέχρι να συμπληρωθεί

όλη η περιοχή D2:F5.

	A	B	C	D	E	F
1	**Καταλύτης Α**	**Καταλύτης**	**Καταλύτης C**	**R(A)**	**R(B)**	**R(C)**
2	90	80	60	3	2	1
3	90	75	60	3	2	1
4	85	70	65	3	2	1
5	80	75	55	3	2	1
6						
7			sum=	12	8	4
8						
9	m=	4				
10	k=	3				
11						
12	F=	8				
13	p-value=	0,01832				
14						

Σχήμα IV.4. Πίνακας ανάλυσης δειγμάτων με το κριτήριο *Friedman*

Από τις τιμές των βαθμών υπολογίζουμε τα αθροίσματα R_i και παίρνουμε: $R_1 = 12$, $R_2 = 8$ και $R_3 = 4$. Επειδή στο συγκεκριμένο πρόβλημα δεν έχουμε δεσμούς, η F υπολογίζεται από τη σχέση

$$F = \frac{12}{4*3*(3+1)}\{12^2+8^2+4^2\} - 3*4(3+1) = 8$$

Εφόσον η συνάρτηση F ακολουθεί ασυμπτωτικά την κατανομή χ^2 με k-1 = 2 βαθμούς ελευθερίας, η τιμή p-value υπολογίζεται από τον τύπο =CHIDIST(F;2). Παίρνουμε p-value = 0.0183, σε συμφωνία με τα αποτελέσματα στους πίνακες των σχημάτων 8.27 και 8.29.

Για να μελετήσουμε την επίδραση της θερμοκρασίας αντιμεταθέτουμε τα δεδομένα του πίνακα των μετρήσεων και επαναλαμβάνουμε την παραπάνω πορεία.

IV.6 ΣΥΝΤΕΛΕΣΤΗΣ SPEARMAN

Απουσία δεσμών ο συντελεστής Spearman υπολογίζεται από τη σχέση (10.27). Μια γενικότερη σχέση που ισχύει και όταν υπάρχουν δεσμοί είναι η

$$\rho = \frac{\sum (r_{x_i} - \bar{r}_x)(r_{y_i} - \bar{r}_y)}{\sqrt{\sum (r_{x_i} - \bar{r}_x)^2 \sum (r_{y_i} - \bar{r}_y)^2}}$$

που ουσιαστικά είναι ο συντελεστής συσχέτισης των μεταβλητών r_{x_i} και r_{y_i}, όπου r_{x_i} είναι ο βαθμός της τιμής x_i της μεταβλητής x, r_{y_i} είναι ο βαθμός της τιμής y_i της μεταβλητής y και $\bar{r}_x$, $\bar{r}_y$ είναι οι μέσες τιμές των τιμών r_{x_i} και r_{y_i}, αντίστοιχα. Για να ελέγξουμε αν ο συντελεστής ρ είναι στατιστικά σημαντικός χρησιμοποιούμε τη συνάρτηση

$$t_s = \rho\sqrt{\frac{m-2}{1-\rho^2}}$$

που ακολουθεί ασυμπτωματικά την κατανομή t με m-2 βαθμούς ελευθερίας.

Παράδειγμα IV.5

Στην άσκηση 10.6 ζητείται να εξεταστεί, με βάση τα δεδομένα του παρακάτω πίνακα, η συσχέτιση μεταξύ του ποσοστού μείωσης του απαιτούμενου οξυγόνου σε συνάρτηση με το ποσοστό μείωσης των στερεών σε μία διάταξη καθαρισμού αποβλήτων που χαρακτηρίζονται από υψηλή στάθμη απαιτούμενου βιοχημικού οξυγόνου και στερεής ύλης.

Στερεά ύλη	3	7	11	15	18	27
Απαιτούμενο οξυγόνο	5	11	21	16	16	28

◆ Για να εξετάσουμε αν σχετίζονται οι δύο παραπάνω μεταβλητές δημιουργούμε τον πίνακα ανάλυσης δεδομένων του σχήματος IV.5. Δηλαδή εισάγουμε τις μεταβλητές σε δύο στήλες και σε δύο γειτονικές στήλες υπολογίζουμε τους βαθμούς των τιμών των δύο μεταβλητών. Για να υπολογίσουμε τώρα τον συντελεστή *Spearman*, μπορούμε να χρησιμοποιήσουμε τη συνάρτηση *CORREL* του *Excel*. Συγκεκριμένα σε ένα

οποιοδήποτε κελί πληκτρολογούμε τον τύπο =CORREL(C2:C7;D2:D7) και πατάμε *Enter*. Θα πάρουμε την τιμή ρ = 0.8117 (σχήμα IV.5).

	A	B	C	D
1	**Στερεά ύλη**	**Απαιτούμενο οξυγόνο**	r_x	r_y
2	3	5	1	1
3	7	11	2	2
4	11	21	3	5
5	15	16	4	3,5
6	18	16	5	3,5
7	27	28	6	6
8				
9	ρ=	0,81167945		
10	ts=	2,779233335		
11	p-value=	0,024928793		
12				

Σχήμα IV.5. Πίνακας ανάλυσης συσχέτισης δειγμάτων με το κριτήριο *Spearman*

Η θετική τιμή του ρ δείχνει ότι όταν αυξάνονται τα στερεά αυξάνεται και το οξυγόνο και συνεπώς το ποσοστό του απαιτούμενου οξυγόνου μειώνεται με τη μείωση του ποσοστού των στερεών. Το ερώτημα είναι αν αυτή η συσχέτιση είναι στατιστικά σημαντική.

Για να προσδιορίσουμε την τιμή p-value πρέπει πρώτα να υπολογίσουμε την ποσότητα t_s. Αν η τιμή του ρ έχει υπολογιστεί στο κελί B9, τότε για τον υπολογισμό του t_s πληκτρολογούμε σε ένα οποιοδήποτε κελί =B9*SQRT(4/(1-B9^2)). Παίρνουμε την τιμή t_s = 2.7792. Επειδή η μεταβλητή t_s ακολουθεί ασυμπτωματικά την κατανομή t με m-2 = 4 βαθμούς ελευθερίας, σε μονόπλευρο έλεγχο η p-value μπορεί να υπολογιστεί με τον τύπο: =TDIST(t_s;4;1). Παίρνουμε p-value = 0.0249, από την οποία προκύπτει ότι υπάρχει στατιστικά σημαντική μείωση του ποσοστού του απαιτούμενου οξυγόνου με τη μείωση του ποσοστού των στερεών.

ΒΙΒΛΙΟΓΡΑΦΙΑ

1. J. N. Miller, J. C. Miller, *Statistics and Chemometrics for Analytical Chemistry*, Prentice Hall, 2000.
2. D. Massart, B. Vandeginste, L. Buydens, S.de Jong, P. Lewi, J. Smeyers-Verbeke, *Handbook of Chemometrics and Qualimetrics*, Elsevier, 1997.
3. N. Draper, H. Smith, *Applied Regression Analysis*, Wiley, 1998.
4. S. Chatterjee, B. Price, *Regression Analysis by Example*, Wiley, 1977,
5. R. Johnson, D. Wichern, *Applied Multivariate Analysis*, Prentice-Hall, 1988.
6. W. Conover, *Practical Nonoparametric Statistics*, Wiley, 1980.
7. R. deLevie, *Advanced Excel for Scientific Data Analysis*, Oxford, 2008.
8. Π. Νικήτας*, Excel: Εισαγωγή και Εφαρμογές*, University Studio Press, 2009.
9. Φ. Κολυβά-Μαχαίρα, Ε. Μπόρα-Σέντα, *Στατιστική : Θεωρία και Εφαρμογές* , Ζήτη, 1996. 1996.
10. Κ.Β. Μπαγιάτης, Φ. Κολυβά-Μαχαίρα, *Μαθηματική Στατιστική*, Θεσσαλονίκη, 1985.
11. M. Spiegel, *Πιθανότητες και Στατιστική*, McGraw-Hill, 1977.
12. Α. Ανθεμίδης, Α. Βουλγαρόπουλος, Γ. Ζαχαριάδης, Ι. Στράτης, *Ποσοτική Χημική Ανάλυση. Αρχές και Εργαστηριακές Ασκήσεις*, Ζήτη, Θεσσαλονίκη 2012, σελ. 118-126.

ΕΥΡΕΤΗΡΙΟ

P

Σ

Τ - Ω