Πανεπιστήμιο Αιγαίου

Χωρική Ανάλυση

Ενότητα 2: Στοιχεία περιγραφικής στατιστικής

Κυριακίδης Φαίδων

Τμήμα Γεωγραφίας

Άδειες Χρήσης

Χρηματοδότηση

Το παρόν εκπαιδευτικό υλικό έχει αναπτυχθεί στα πλαίσια του εκπαιδευτικού έργου του διδάσκοντα.

Το έργο «**Ανοικτά Ακαδημαϊκά Μαθήματα στο Πανεπιστήμιο Αιγαίου**» έχει χρηματοδοτήσει μόνο τη αναδιαμόρφωση του εκπαιδευτικού υλικού.

Το έργο υλοποιείται στο πλαίσιο του Επιχειρησιακού Προγράμματος «Εκπαίδευση και Δια Βίου Μάθηση» και συγχρηματοδοτείται από την Ευρωπαϊκή Ένωση (Ευρωπαϊκό Κοινωνικό Ταμείο) και από εθνικούς πόρους.

Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Στοιχεία Μονομεταβλητής Περιγραφικής Στατιστικής

Φαίδων Κυριακίδης

Καθηγητής Τμήματος Γεωγραφίας

phkyriakidis@geo.aegean.gr

Πανεπιστήμιο Αιγαίου
Λόφος Πανεπιστημίου, 81100 Μυτιλήνη

Χωρική Ανάλυση

ΤΜΗΜΑ ΓΕΩΓΡΑΦΙΑΣ
ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

ΓΕΩΓΡΑΦΙΑ ΚΑΙ ΕΦΑΡΜΟΣΜΕΝΗ ΓΕΩΠΛΗΡΟΦΟΡΙΚΗ

Περιγραφική Στατιστική

Ορισμός

Η περιγραφική στατιστική αποτελεί ένα σύνολο μεθόδων και εργαλείων (γραφικών και ποσοτικών) για την αποτελεσματική παρουσίαση δεδομένων μέσω της σύνοψης, ομαδοποίησης και απεικόνισής τους

Μερικά εργαλεία

- πίνακες: ομαδοποίηση και ταξινόμηση δεδομένων...
- γραφήματα: ραβδογράμματα, θηκογράμματα...
- στατιστικά μέτρα: αριθμητικές ποσότητες που περιγράφουν κεντρική τάση ή θέση, μεταβλητότητα...

Στόχοι του μαθήματος αυτού

- σύντομη επισκόπηση των διαφόρων εννοιών της περιγραφικής στατιστικής
- εξοικίωση με τα αντίστοιχα εργαλεία που χρησιμοποιούνται στην περιγραφική στατιστική ανάλυση ποσοτικών δεδομένων

Ορισμοί

Πληθυσμός και δείγμα

- Πληθυσμός: Σύνολο όλων των δυνατών στοιχείων ή μετρήσεων που είναι (υποθετικά) δυνατό να παρατηρηθούν στη μελέτη μιας μεταβλητής
  π.χ., όλοι οι φοιτητές και φοιτήτριες της Ελλάδας
- Δείγμα: Υποσύνολο στοιχείων ή μετρήσεων από ένα πληθυσμό,
  π.χ., φοιτητές 4ου έτους, ή φοιτήτριες στη Βόρεια Ελλάδα

Μεταβλητές πληθυσμού

Μεταβλητές ή χαρακτηριστικά που περιγράφουν ή ορίζουν ένα πληθυσμό,
π.χ., ηλικία φοιτητών/φοιτητριών ή εισόδημα φοιτητών/φοιτητριών

Παράμετροι πληθυσμού και στατιστικά δείγματος

- Παράμετροι: Ποσοτικά μέτρα που περιγράφουν μια μεταβλητή πληθυσμού,
  π.χ., μέσος όρος ηλικίας φοιτητών/φοιτητριών
- Στατιστικά: Ποσοτικά μέτρα που περιγράφουν μια μεταβλητή δείγματος,
  π.χ., μέσος όρος ηλικίας όλων των φοιτητών/φοιτητριών του 4ου έτους

Ορισμοί (2)

Στατιστική δειγματοληψία

- διαδικασία λήψης ενός αντιπροσωπευτικού δείγματος από έναν πληθυσμό,
  π.χ., εξαγωγή δείγματος φοιτητών από όλους τους φοιτητές της Ελλάδας
- τυχαίο δείγμα: ένα δείγμα στο οποίο κάθε στοιχείο ή μονάδα του πληθυσμού έχει την ίδια πιθανότητα να συμπεριληφθεί στο δείγμα

Στατιστική εκτίμηση

βέλτιστη πρόταση σχετικά με την τιμή μιας παραμέτρου ενός πληθυσμού από ένα δείγμα

Έλεγχος υποθέσεως και στατιστική συμπερασματολογία

Διαδικασία με την οποία ελέγχουμε αν τα δεδομένα υποστηρίζουν μια υπόθεση που καθορίζει την τιμή (ή το εύρος τιμών) μιας παραμέτρου ενός πληθυσμού

Ιστόγραμμα Συχνοτήτων

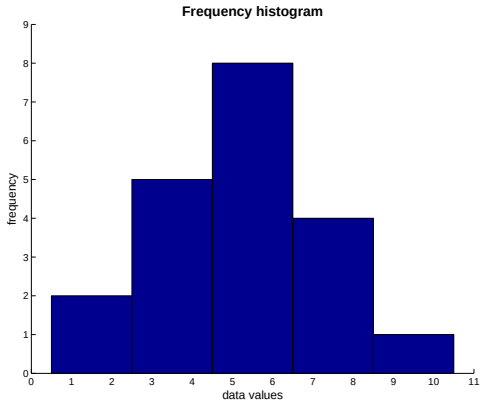
Παράδειγμα $N = 20$ τιμών ή μετρήσεων μιας μεταβλητής:

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

Παράδειγμα $K = 5$ (υποκειμενικά ορισμένων) διαστημάτων που διακριτοποιούν το εύρος των πιθανών τιμών της μεταβλητής, και αριθμός n_k (συχνότητα) μετρήσεων που εμπίπτουν σε κάθε διάστημα:

$(0.5 - 2.5]$	$(2.5 - 4.5]$	$(4.5 - 6.5]$	$(6.5 - 8.5]$	$(8.5 - 10.5]$
2	5	8	4	1

Το ραβδόγραμμα (bar graph) όπου το ύψος κάθε ράβδου αντοιστιχεί στη συχνότητα n_k των μετρήσεων σε κάθε διάστημα k, λέγεται **ιστόγραμμα** συχνοτήτων



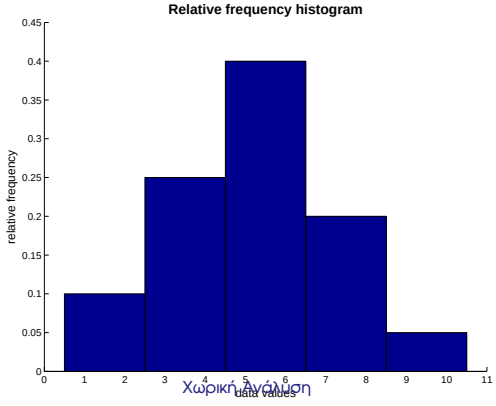
Ιστόγραμμα Σχετικών Συχνοτήτων

Σχετική συχνότητα f_k διαστήματος k = (αριθμός μετρήσεων n_k στο διάστημα k) / (συνολικός αριθμός μετρήσεων N) $\boxed{f_k = \frac{n_k}{N}}$

Παράδειγμα $K = 5$ (υποκειμενικά ορισμένων) διαστημάτων που διακριτοποιούν το εύρος των πιθανών τιμών μιας μεταβλητής, και σχετική συχνότητα f_k (ποσοστό) μετρήσεων που εμπίπτουν σε κάθε διάστημα k:

$(0.5 - 2.5]$	$(2.5 - 4.5]$	$(4.5 - 6.5]$	$(6.5 - 8.5]$	$(8.5 - 10.5]$
0.1	0.25	0.4	0.2	0.05

Το ραβδόγραμμα (bar graph) όπου το ύψος κάθε ράβδου αντοιστιχεί στη σχετική συχνότητα των μετρήσεων σε κάθε διάστημα, λέγεται **ιστόγραμμα σχετικών συχνοτήτων** – ίδιο σχήμα με το ιστόγραμμα συχνοτήτων αλλά με διαφορετική κλίμακα στον άξονα των Ψ



Ιστόγραμμα Σχετικών Συχνοτήτων (2)

Παράδειγμα $K = 5$ (υποκειμενικά ορισμένων) διαστημάτων που διακριτοποιούν το εύρος των πιθανών τιμών μιας μεταβλητής, και σχετική συχνότητα f_k (ποσοστό) μετρήσεων που εμπίπτουν σε κάθε διάστημα k:

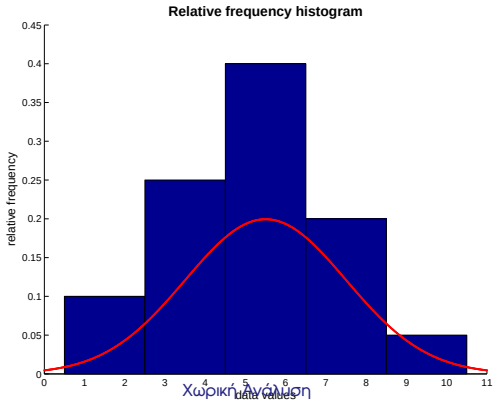
$(0.5 - 2.5]$	$(2.5 - 4.5]$	$(4.5 - 6.5]$	$(6.5 - 8.5]$	$(8.5 - 10.5]$
0.1	0.25	0.4	0.2	0.05

Συνολικό εμβαδόν ιστογράμματος = άθροισμα εμβαδού a_k ραβδών =

$$\sum_{k=1}^{K} a_k = \sum_{k=1}^{K} d_k \times f_k = d \times \sum_{k=1}^{K} f_k = d \neq 1$$

Πρόβλημα

Το ιστόγραμμα σχετικών συχνοτήτων δεν μπορεί να χρησιμοποιηθεί για τη σύγκριση εμπειρικών κατανομών με θεωρητικές κατανομές, όπως η κανονική κατανομή ή κατανομή του Gauss, που πάντα έχουν εμβαδό 1 κάτω από την καμπύλη πιθανοτήτων

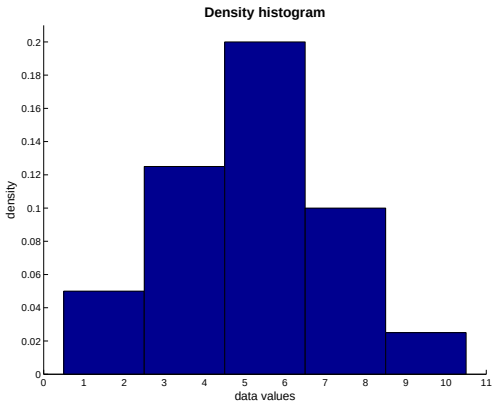


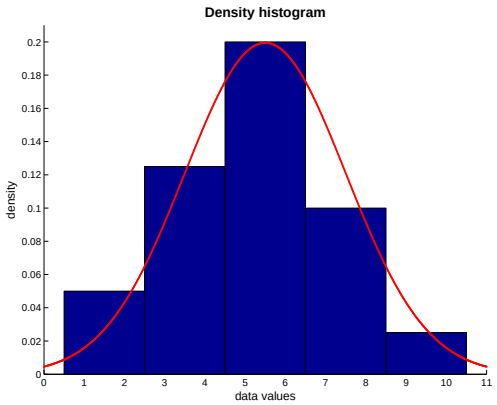
Ιστόγραμμα Πυκνοτήτων

Ορισμός

Ραβδόγραμμα (bar graph) όπου το ύψος κάθε ράβδου αντοιστιχεί στην πυκνότητα f_k/d_k ή f_k/d των μετρήσεων σε κάθε διάστημα k, λέγεται **ιστόγραμμα πυκνοτήτων**

ίδιο σχήμα με το ιστόγραμμα σχετικών συχνοτήτων με διαφορετική κλίμακα στον άξονα των Ψ, έτσι ώστε το συνολικό εμβαδό του ιστογράμματος να είναι 1, γεγονός που επιτρέπει τη σύγκριση μιας εμπειρικής κατανομής με μια θεωρητική

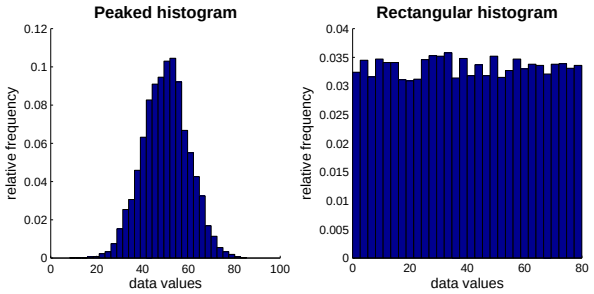




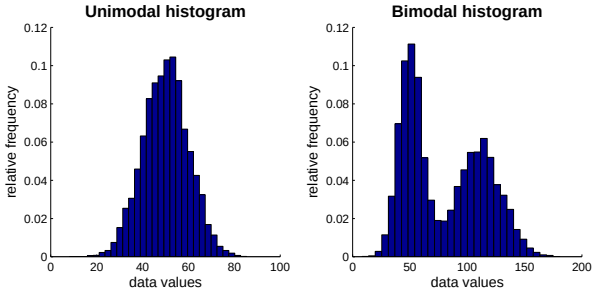
Η θεωρητική καμπύλη στα δεξιά απεικονίζει μια κανονική (Gauss) κατανομή, της οποίας οι παράμετροι, μέσος όρος και τυπική απόκλιση, έχουν υπολογιστεί από τα ίδια δεδομένα στα οποία βασίστηκε η κατασκευή του ιστογράμματος πυκνότητας

Χαρακτηριστικά Σχήματος Ιστογράμματος

Με ή όχι σαφή επικρατούσα τιμή (ή διάστημα συγκέντρωσης τιμών)



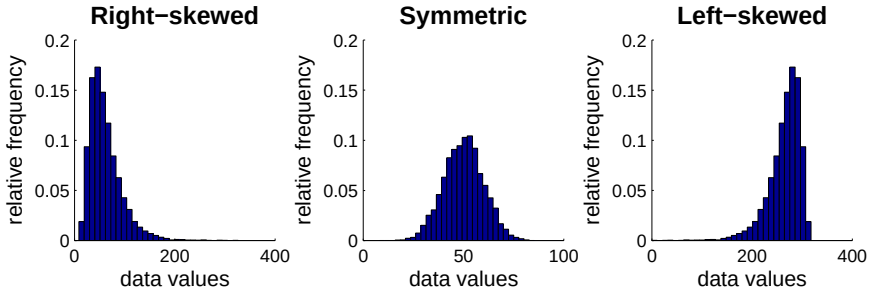
Με μία ή περισσότερες επικρατούσες τιμές (ή διάστημα συγκέντρωσης τιμών)



Όταν στο ιστόγραμμα εμφανίζονται περισσότερες της μίας επικρατούσες τιμές (modes), αυτό συνήθως υποδηλώνει την ύπαρξη περισσότερων από ένα πληθυσμό ή (λιγότερο συχνά) την δειγματοληψία με διαφορετικά όργανα μέτρησης

Χαρακτηριστικά Σχήματος Ιστογράμματος (2)

Ασυμμετρία



Αριστερά: κατανομή με τα κοίλα (ουρά) προς τα δεξιά – περισσότερες μικρές τιμές και πολύ λιγότερες μεγάλες τιμές – θετική ασυμμετρία. **Μέση:** συμμετρική κατανομή – περισσότερες 'μεσαίες τιμές' και πολύ λιγότερες μικρές και μεγάλες τιμές. **Δεξιά:** κατανομή με τα κοίλα (ουρά) προς τα αριστερά – λιγότερες μικρές τιμές και πολύ περισσότερες μεγάλες τιμές – αρνητική ασυμμετρία.

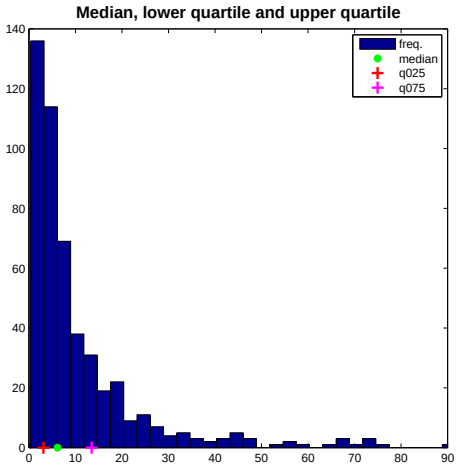
Για όλους τους τύπους ιστογραμμάτων

- το σχήμα του ιστογράμματος επηρεάζεται από τον αριθμό K και τα όρια (συνεπώς τη θέση και το πλάτος) των διαστημάτων που διακριτοποιούν το εύρος N τιμών μιας μεταβλητής
- συχνά χρησιμοποιούνται μη-αλληλεπικαλυπτόμενα διαστήματα με απλά ορισμένα όρια. Ένας κανόνας για την επιλογή του αριθμού των διαστημάτων είναι: $5 \times log_{10}(N)$

Ποσοστιαία Σημεία ή Ποσοστημόρια -- Quantiles

Ορισμός

Ως p-ποσοστημόριο αναφέρεται η τιμή x_p μιας μεταβλητής X που αντιστοιχεί σε ένα αθροιστικό ποσοστό p μετρήσεων στο δείγμα, οι οποίες είναι μικρότερες ή ίσες από την τιμή x_p



Κυριώτερα ποσοστιαία σημεία

- κατώτερο τεταρτημόριο $x_{0.25}$, διάμεσος $x_{0.5}$, ανώτερο τεταρτημόριο $x_{0.75}$
- εκατοστιαία σημεία, ή εκατοστημόρια – percentiles: $x_{0.01}, x_{0.02}, \ldots, x_{0.98}, x_{0.99}$
- δεκατημόρια – deciles: $x_{0.1}, x_{0.2}, \ldots, x_{0.8}, x_{0.9}$

Μέτρα Κεντρική Τάσης ή Θέσης

Μερικά στατιστικά κεντρικής τάσης

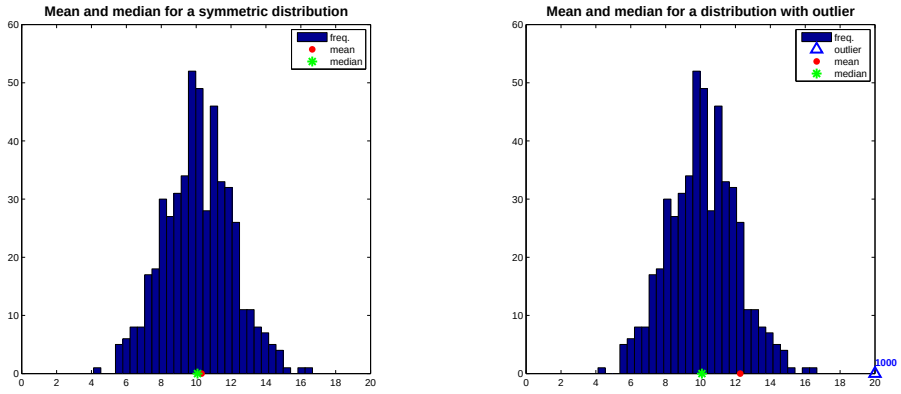
- διάμεσος: η τιμή, $\tilde{x}$ ή $\tilde{m}_X$, της μεταβλητής X που χωρίζει ένα (διατεταγμένο σε αύξουσα σειρά) δείγμα σε δύο ίσα μέρη, ή αλλιώς το 50ο ποσοστημόριο
- μέσος όρος: η αριθμητική μέση τιμή, $\bar{x}$ ή m_X, του δείγματος. Πιό αναλυτικά, $m_X = 1/N \sum_{i=1}^{N} x_i$ ή $m_X = \sum_{i=1}^{N} \frac{1}{N} x_i$, δηλαδή μέσος όρος είναι το ισοσταθμισμένο (με βάρος $1/N$) άθροισμα των τιμών του δείγματος
  Σημείωση: Ο μέσος όρος μπορεί να μην αντιστοιχεί σε μια πραγματοποιήσιμη τιμή
- επικρατούσα τιμή: η τιμή της μεταβλητής που εμφανίζεται πιο συχνά στο δείγμα

Σημειώσεις

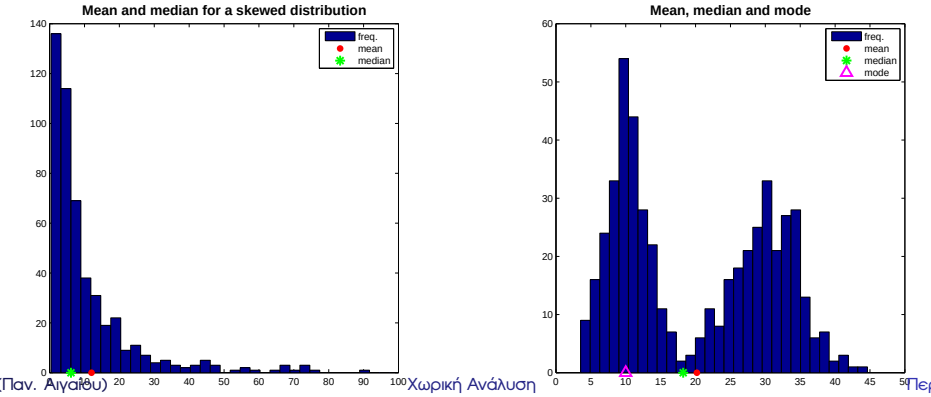
- όλα τα παραπάνω μέτρα έχουν μονάδες ίδιες με τις μονάδες της μεταβλητής
- η επιλογή του καταλληλότερου μέτρου κεντρικής τάσης εξαρτάται από το σχήμα της κατανομής του δείγματος. *Π.χ., ο διάμεσος προτιμάται για την περιγραφή της κεντρικής τάσης ενός δείγματος με μη συμμετρική κατανομή. Ο μέσος όρος είναι ευαίσθητος στην παρουσία ακραίων τιμών (outliers)*

Μέτρα Κεντρική Τάσης ή Θέσης: Παραδείγματα

Επίδραση ακραίων τιμών στο μέσος όρο και το διάμεσο



Επίδραση σχήματος κατανομής στα στατιστικά κεντρικής τάσης



Ιδιότητες Μέσου Όρου και Διάμεσου

Μέσος όρος

Η τιμή $\bar{x}$ της μεταβλητής X, για την οποία η συνολική τετραγωνική απόκλιση μεταξύ της τιμής αυτής και των N τιμών x_i είναι ελάχιστη:

$$\sum_{i=1}^{N}(x_i - \bar{x})^2 = \min$$

Όμως: $\sum_{i=1}^{N}(x_i - \bar{x}) = 0$

Διάμεσος

Η τιμή $\tilde{x}$ της μεταβλητής X, για την οποία η συνολική απόλυτη απόκλιση μεταξύ της τιμής αυτής και των N τιμών x_i είναι ελάχιστη:

$$\sum_{i=1}^{N}|x_i - \tilde{x}| = \min$$

Μέτρα Μεταβλητότητας ή Διασποράς

Μερικά στατιστικά διασποράς ενός δείγματος

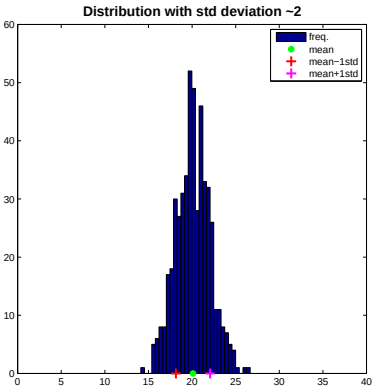
- εύρος: η διαφορά μεταξύ της μέγιστης κι ελάχιστης τιμής στο δείγμα: $x_{max} - x_{min}$
- ενδοτεταρτημοριακό πλάτος: η διαφορά μεταξύ του ανώτερου και κατώτερου τεταρτημορίου του δείγματος: $x_{0.75} - x_{0.25}$
- διακύμανση: η μέση τετραγωνική απόκλιση των τιμών του δείγματος από το μέσο όρο του δείγματος: $s_X^2 = 1/(N-1)\sum_{i=1}^{N}[x_i - m_X]^2$
- τυπική απόκλιση: η τετραγωνική ρίζα της διακύμανσης: $s_X = \sqrt{s_X^2}$
- mean absolute deviation (MAD): $1/(N-1)\sum_{i=1}^{N}|x_i - m_X|$
- συντελεστής μεταβλητότητας: πηλίκο τυπικής απόκλισης προς μέσο όρο: s_X/m_X

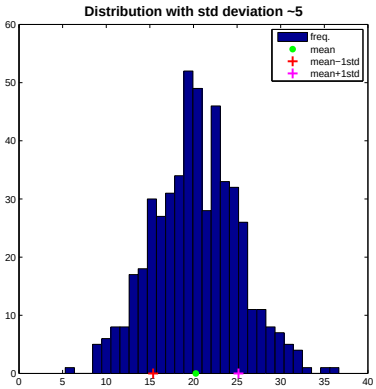
Σημειώσεις

- η διακύμανση s_X^2 έχει μονάδες μεταβλητής στο τετράγωνο...
- ο συντελεστής μεταβλητότητας s_X/m_X δεν έχει μονάδες και χρησιμοποιείται για τη σύγκριση μεταβλητότητας δειγμάτων με διαφορετικές μονάδες μέτρησης
- τα μέτρα μεταβλητότητας που βασίζονται σε ποσοστημόρια ή δεν εμπεριέχουν τετραγωνικές αποκλίσεις, είναι γενικά πιο εύρωστα από τα υπόλοιπα

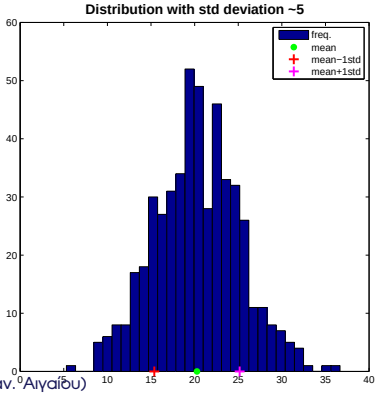
Μέτρα Μεταβλητότητας ή Διασποράς: Παραδείγματα

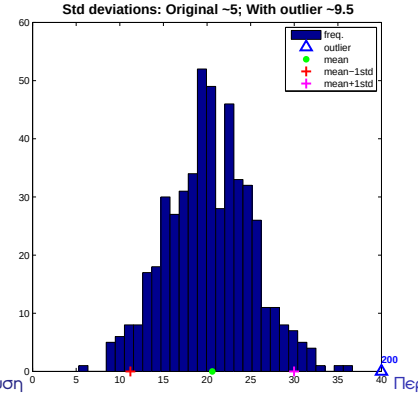
Διακύμανση συμμετρικών κατανομών





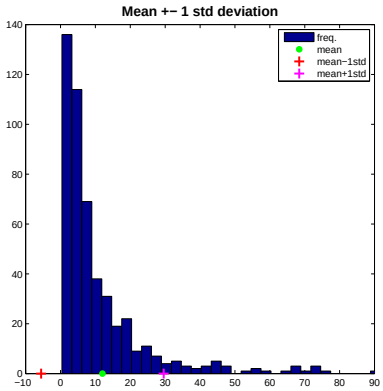
Επίδραση ακραίων τιμών στη διακύμανση

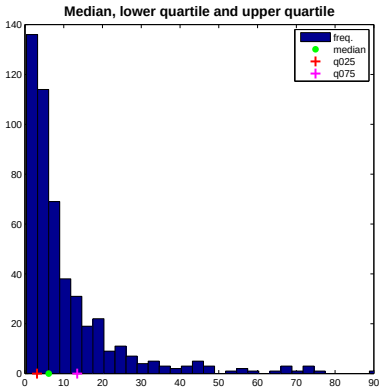




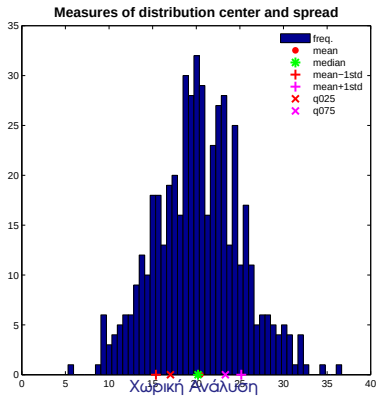
Μέτρα Μεταβλητότητας ή Διασποράς: Παραδείγματα

Περιγραφή διασποράς σε μή συμμετρικές κατανομές





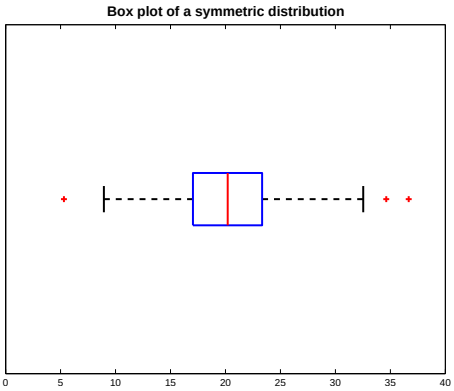
Περιγραφή κεντρικής τάσης και διασποράς

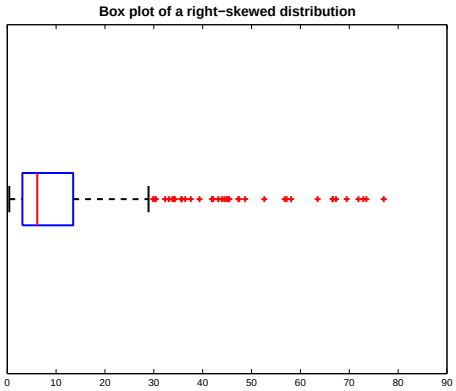


Θηκόγραμμα: Box plot

Τί απεικονίζεται στο γράφημα

- θήκη (box): κατώτερο τεταρτημόριο, διάμεσος, ανώτερο τεταρτημόριο
- μύστακες (whiskers), που εκτείνονται εκατέρωθεν των ορίων της θήκης σε μήκος 1.5 φορές του ενδοτεταρτημοριακού πλάτους
- ακραίες τιμές (outliers), δηλαδή τιμές πέραν των ορίων των μυστάκων



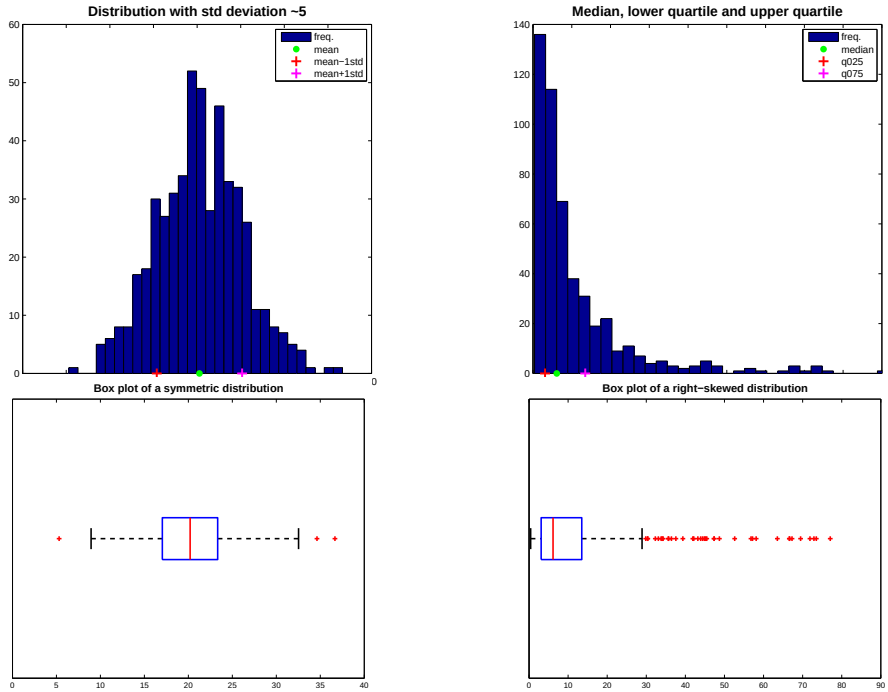


Χρησιμότητα

Συνοπτική παρουσίαση των κυριότερων ποσοστημορίων, της ασυμμετρίας, και των ακραίων τιμών ενός δείγματος

Θηκόγραμμα: Box plot (2)

Παραδείγματα ιστογραμμάτων και θηκογραμμάτων



Σημείωση: Όταν γνωρίζουμε την κατανομή (ιστόγραμμα) των μετρήσεων μπορούμε να κατασκευάσουμε το αντίστοιχο θηκόγραμμα. Το αντίστροφο όμως δεν είναι δυνατόν.

Αθροιστική Σχετική Συχνότητα

Ορισμός

Η εμπειρική αθροιστική σχετική συχνότητα $F(x)$ που αντιστοιχεί σε μια τιμή x είναι το **ποσοστό** των μετρήσεων στο δείγμα, οι οποίες έχουν τιμή μικρότερη (ή ίση) από x, ή αλλιώς: $F(x) = (1/N)\sum_{i=1}^{N} c(x_i)$, όπου $c(x_i) = 1$ αν $x_i \leq x$ και $c(x_i) = 0$ αν $x_i > x$

Χαρακτηριστικά

- κάθε αθροιστική σχετική συχνότητα $F(x)$ ανήκει στο διάστημα $[0, 1]$
- αν ισχύει $x < x'$ τότε $F(x) < F(x')$

Παράδειγμα $N = 20$ τιμών μιας μεταβλητής, **διατεταγμένες** κατά αύξουσα σειρά και αθροιστική σχετική συχνότητα $F(x_i)$ για κάθε τιμή x_i:

2	2	3	3	3	3	4	5	5	5	5	6	6	6	6	7	7	7	8	9
0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95	1.00

Οι αθροιστικές σχετικές συχνότητες που αντιστοιχούν σε μια αύξουσα διατεταγμένη σειρά N τιμών, δίνονται ως: $[1/N \quad 2/N \quad \ldots \quad (N-1)/N \quad N/N]$ ή πιο απλά ως: $([1 \quad 2 \quad \ldots \quad N-1 \quad N])/N$

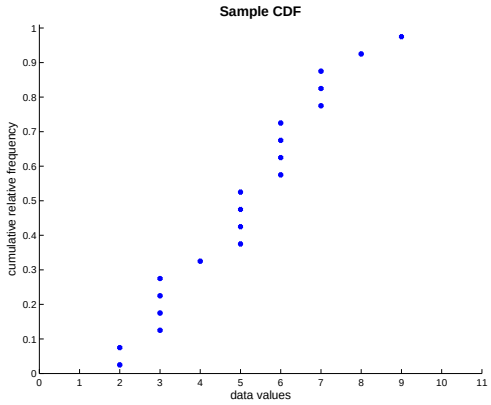
Συνάρτηση Αθροιστικής Κατανομής

Ορισμός

Η αντιστοίχιση των αθροιστικών σχετικών συχνοτήτων $F\{(x_i), i = 1, \ldots, N\}$ που ορίζονται για N μετρήσεις μιας μεταβλητής $\{x_i, i = 1, \ldots, N\}$ με τις τιμές αυτές διατεταγμένες σε αύξουσα σειρά, λέγεται συνάρτηση αθροιστικής κατανομής (cumulative distribution function -- CDF)

$N = 20$ τιμές μιας μεταβλητής, **διατεταγμένες** κατά αύξουσα σειρά και αθροιστική σχετική συχνότητα $F(x_i)$ για κάθε τιμή x_i:

2	2	3	3	3	3	4	5	5	5	5	6	6	6	6	7	7	7	8	9
0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95	1.00

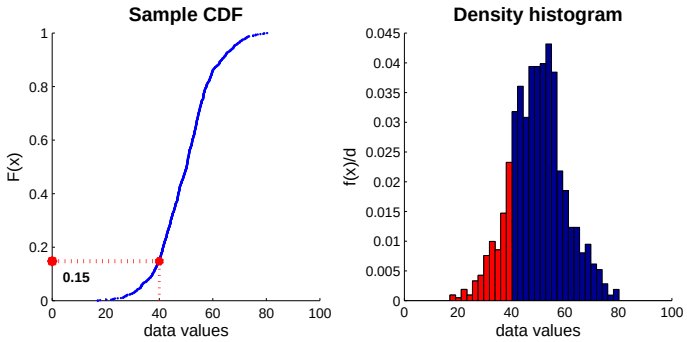


Εδώ οι αθροιστικές σχετικές συχνότητες υπολογίστηκαν ως: $([1\ \ 2\ \ \ldots\ \ N-1\ \ N] - 0.5)/N$, κανόνας που επιτρέπει μη μηδενική πιθανότητα εμφάνισης τιμών μεγαλύτερων από τη μέγιστη τιμή στο δείγμα

Συνάρτηση Αθροιστικής Κατανομής & Ιστόγραμμα Πυκνότητας

Αντιστοιχία

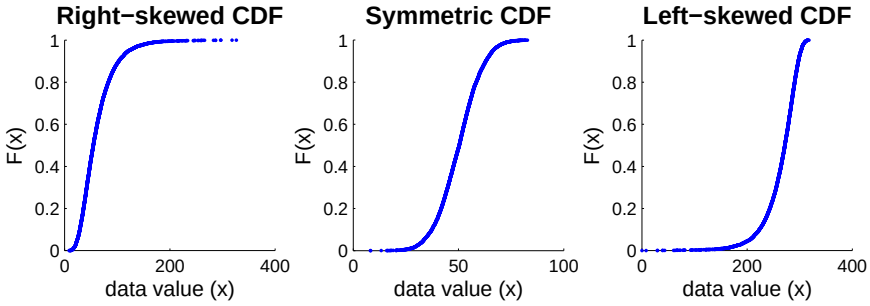
Ένα οποιοδήποτε σημείο $\{x, F(x)\}$ πάνω στην καμπύλη της συνάρτησης αθροιστικής κατανομής έχει 2 συντεταγμένες: την τιμή x της μεταβλητής (τετμημένη), και το ποσοστό $F(x)$ των μετρήσεων στο δείγμα που έχουν τιμή μικρότερη ή ίση του x (τεταγμένη). Το ποσοστό $F(x)$ αντιστοιχεί στο αθροιστικό εμβαδό εκείνων των ραβδών του ιστογράμματος πυκνότητας που βρίσκονται στα **αριστερά** της τιμής x.



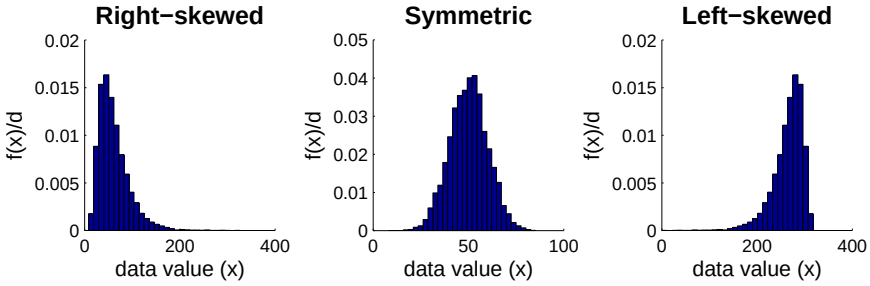
Στην περίπτωση ενός ιστογράμματος πυκνότητας με K ράβδους πλάτους d και ύψους $f(x_k)$, το ποσοστό $F(x)$ δίνεται ως: $F(x) = \sum_{k=1}^{J} f(x_k)/d$, όπου J είναι ο αριθμός των ραβδών με κεντρική τιμή $x_k \leq x$ και σχετική συχνότητα $f(x_k)$

Παραδείγματα Συναρτήσεων Αθροιστικής Κατανομής

Συνάρτηση Αθροιστικής Κατανομής



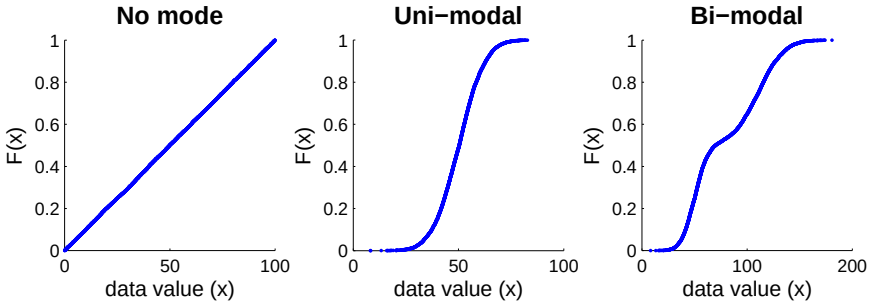
Ιστόγραμμα πυκνότητας



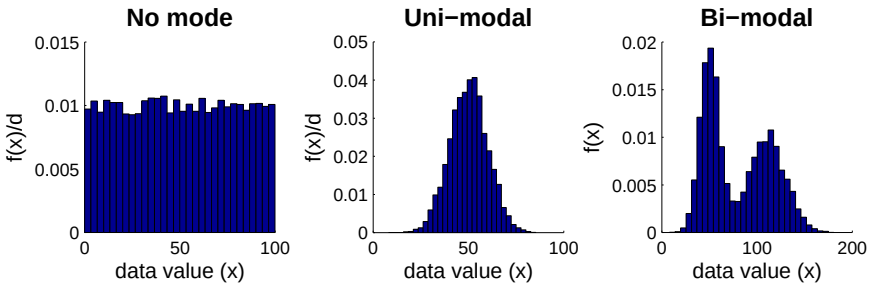
Αριστερά: κατανομή με θετική ασυμμετρία ή λοξότητα. **Μέση:** συμμετρική ή κωνοειδής κατανομή. **Δεξιά:** κατανομή με αρνητική ασυμμετρία ή λοξότητα.
Η αυξημένη πυκνότητα στο ιστόγραμμα εμφανίζεται ως μεγαλύτερη **κλίση** στη συνάρτηση αθροιστικής κατανομής.

Παραδείγματα Συναρτήσεων Αθροιστικής Κατανομής (2)

Συνάρτηση Αθροιστικής Κατανομής



Ιστόγραμμα πυκνότητας

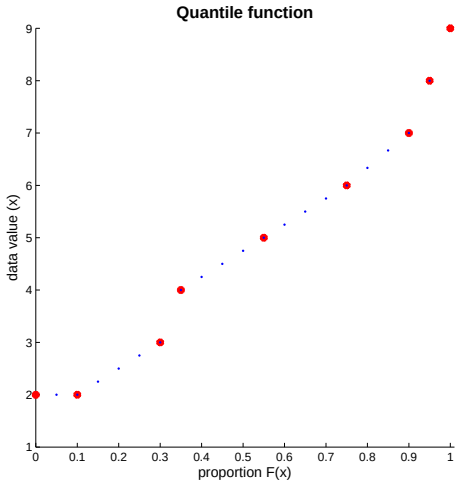


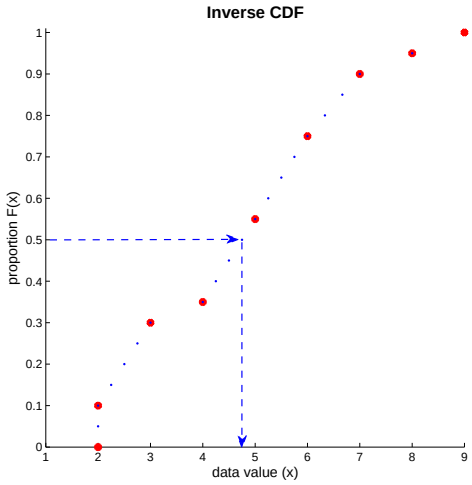
Αριστερά: κατανομή χωρίς σαφή επικρατούσα τιμή (mode). **Μέση:** κατανομή με μία σαφή επικρατούσα τιμή ή εύρος τιμών.
Δεξιά: κατανομή με δύο επικρατούσες τιμές ή εύρος τιμών.
Η αυξημένη πυκνότητα στο ιστόγραμμα εμφανίζεται ως μεγαλύτερη κλίση στη συνάρτηση αθροιστικής κατανομής.

Συνάρτηση Ποσοστημορίων -- Quantile Function

Ορισμοί

- Ποσοστημόριο: η τιμή x_p μιας μεταβλητής που αντιστοιχεί σε ένα αθροιστικό ποσοστό p
- η αντιστοιχία διάφορων αθροιστικών ποσοστών p με τα ποσοστημόριά τους x_p, λέγεται συνάρτηση ποσοστημορίων $x_p = Q(p)$
- η αντιστοιχία αυτή ονομάζεται και αντίστροφη συνάρτηση αθροιστικής κατανομής $x_p = F^{-1}(p)$





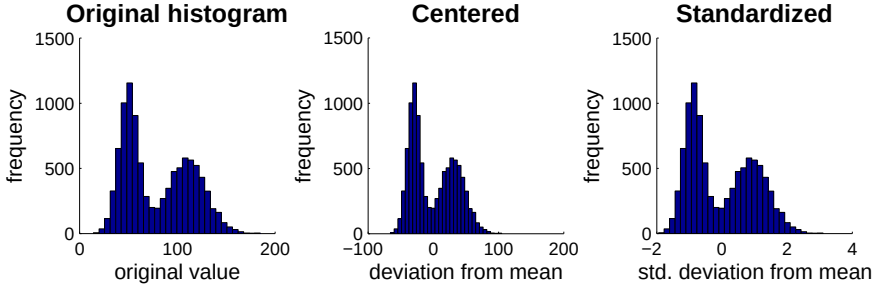
Κανονικοποίηση Μετρήσεων

Διαδικασία κανονικοποίησης

- μετατροπή άρχικών τιμών x_i δείγματος που ακολουθεί μια οποιαδήποτε κατανομή με μέσο όρο m_X και τυπική απόκλιση s_X σε τιμές με μέσο όρο 0 και τυπική απόκλιση 1 ως:

$$z_i = (x_i - m_X)/s_X$$

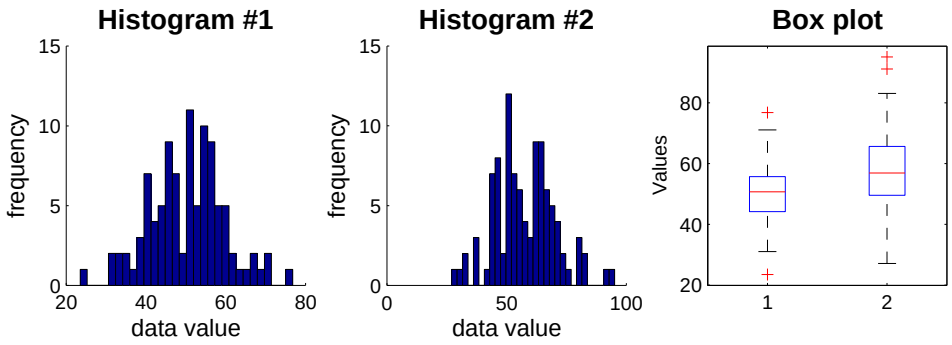
- οι τιμές $x_i - m_X$ είναι γνωστές και ως αποκλίσεις, έχουν μέσο όρο 0 και τυπική απόκλιση s_X και η διαδικασία υπολογισμού τους λέγεται επίσης και 'κεντροποίηση'
- η κανονικοποιημένη τιμή z_i εκφράζει την απόκλιση της αρχικης τιμής x_i από το μέσο όρο m_X σε ποσοστό επί της τυπικής απόκλισης s_X
- χρήσιμος μετασχηματισμός δεδομένων για τη σύγκριση κατανομών μετρήσεων με διαφορετικές μονάδες



Σύγκριση Κατανομών: Ιστογράμματα και Θηκογράμματα

Σκοπός και εργαλεία

- Σκοπός: Σύγκριση δύο ή περισσότερων εμπειρικών κατανομών
- Σημείωση: η εμπειρική κατανομή ενός δείγματος μπορεί να έχει κατασκευαστεί από διαφορετικό αριθμό μετρήσεων από ότι μια άλλη κατανομή – δεν μιλάμε για ζεύγη μετρήσεων στις ίδιες θέσεις παρατήρησης
- Ανάλυση: Χρήση ιστογραμμάτων και θηκογραμμάτων, σύγκριση στατιστικών: μέσων τιμών, διαμέσων, διακυμάνσεων...



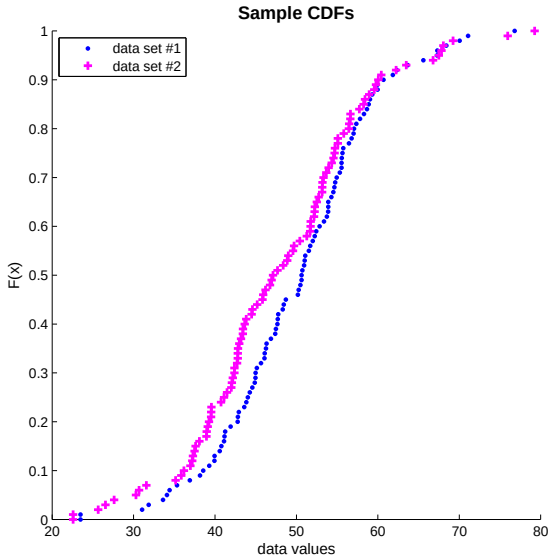
Προβλήματα

Δύσκολη απεικόνιση/σύγκριση στο ίδιο γράφημα, ιδιαίτερα στην περίπτωση πολλαπλών ιστογραμμάτων

Σύγκριση Κατανομών: Συναρτήσεις Αθροιστικών Κατανομών

Σκοπός και εργαλεία

- Σκοπός: Σύγκριση δύο ή περισσότερων εμπειρικών κατανομών
- Ανάλυση: Επικάλυψη εμπειρικών συναρτήσεων αθροιστικών σχετικών συχνοτήτων (CDF), μια για κάθε δείγμα, στο ίδιο γράφημα

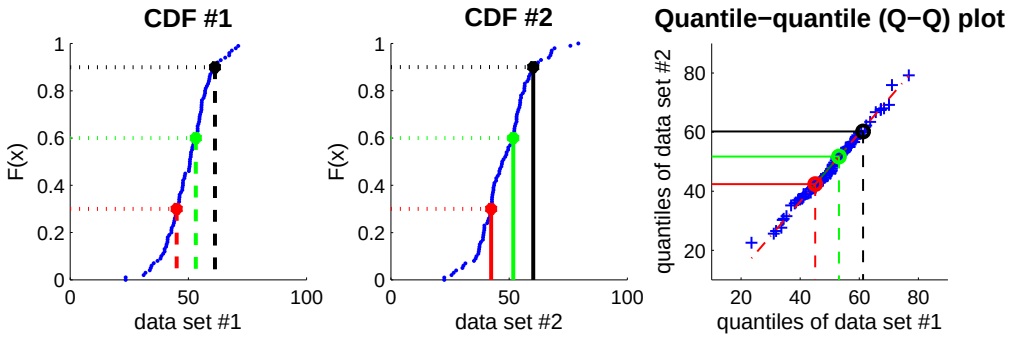


Εναλλακτικά χρησιμοποιείται το διασπορόγραμμα ποσοστημορίων (π.χ., δεκατημορίων) κάθε δείγματος, τα οποία αντιστοιχούν στα ίδια αθροιστικά ποσοστά

Γράφημα Ποσοστιαίων Σημείων Δύο Κατανομών: Q-Q plot

Διαδικασία κατασκευής γραφήματος

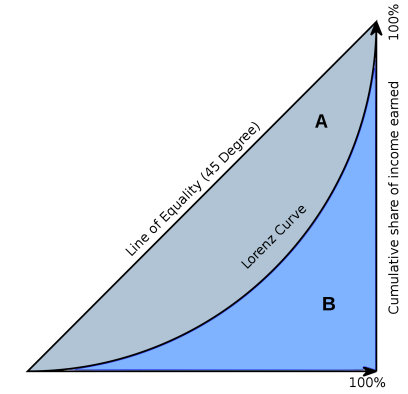
- διακριτοποίηση εύρους πιθανοτήτων σε N ίσα διαστήματα, π.χ., $[0.05 \;\; 0.10 \;\; \ldots \;\; 0.9 \;\; 0.95]$ και υπολογισμός των ποσοστιαίων σημείων κάθε δείγματος που αντιστοιχούν στις παραπάνω πιθανότητες. Π.χ., η πιθανότητα 0.3 μπορεί να οδηγεί στο 30 ποσοστιαίο σημείο 500 στο σύνολο Α και στο 30 ποσοστιαίο σημείο 550 στο σύνολο Β
- γραφική αντιστοίχιση των ποσοστιαίων σημείων που ορίζονται από το ίδιο ποσοστό, δηλαδή γράφημα των N ποσοστιαίων σημείων του ενός συνόλου δεδομένων με τα αντίστοιχα (ιδίου ποσοστού) ποσοστιαία σημεία του άλλου συνόλου δεδομένων
- χρήσιμο γράφημα για τη σύγκριση δύο κατανομών: αν τα σημεία του γραφήματος βρίσκονται κοντά στην γραμμή των 45 μοιρών, τότε οι δύο κατανομές έχουν παρόμοιο σχήμα, όχι μόνο παρόμοιο μέσο, διάμεσο, ή τεταρτημόρια



Η Καμπύλη του Lorenz

Ορισμός

- διακριτή καμπύλη που απεικονίζει το αθροιστικό ποσοστό των μετρήσεων (τεταγμένη) σε σχέση με το αντίστοιχο αθροιστικό ποσοστό του πλήθους των δεδομένων (τετμημένη) – *χρησιμοποιείται στα οικονομικά και στην οικολογία...*
- ένα σημείο στην καμπύλη του Lorenz παρουσιάζει, π.χ., το ποσοστό των ατόμων στο οποίο αντιστοιχεί ένα αθροιστικό ποσοστό του συνολικού εισόδηματος



Πηγή: Wikipedia

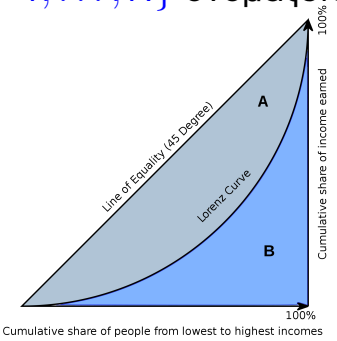
Οριακές μορφές

- γραμμή τέλειας ισότητας (perfect equality line): καθένας έχει το ίδιο εισόδημα
- γραμμή τέλειας ανισότητας: όλοι έχουν εισόδημα 0, εκτός από έναν που έχει το συνολικό εισόδημα του δείγματος

Η Καμπύλη του Lorenz (2)

Διαδικασία υπολογισμού

- διάταξη των μετρήσεων κατά αύξουσα σειρά $\{x_i^s, i = 1, \ldots, N\}$, όπου $x_i \leq x_{i+1}$
- σε κάθε (διατεταγμένη ή μη) μέτρηση x_i^s, αντιστοιχεί το ποσοστό $1/N$
- για την i-οστή διατεταγμένη μέτρηση x_i^s, υπολογίζονται: (1) το αντίστοιχο αθροιστικό ποσοστό $\sum_{j=1}^{i} x_j^s/T$, όπου $T = \sum_{j=1}^{N} x_j$ είναι το άθροισμα των N τιμών, και (2) το αντίστοιχο ποσοστό i/N των διατεταγμένων μετρήσεων
- η διακριτή καμπύλη που σχηματίζεται από τα $N + 1$ σημεία με συντεταγμένες $\{(0, 0), (i/N, \sum_{j=1}^{i} x_i/T), i = 1, \ldots, N\}$ ονομάζεται καμπύλη του Lorenz



Πηγή: Wikipedia

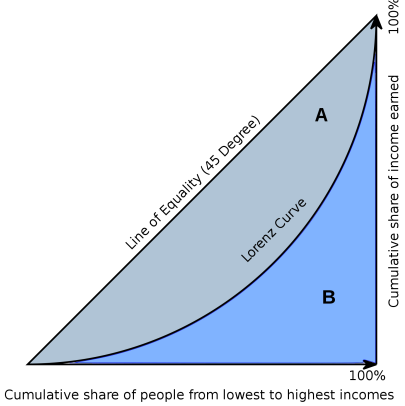
Σημείωση:

Όταν η κάθε μέτρηση x_i δεν αντιστοιχεί σε άτομο, αλλά σε κάποιο υποπληθυσμό, π.χ., οικισμό, ορίζονται N τιμές υποπληθυσμών $\{p_i, i = 1, \ldots, N\}$. Στην περίπτωση αυτή, το αθροιστικό ποσοστό ατόμων i/N αντικαθίσται από το αθροιστικό ποσοστό των υποπληθυσμών $\sum_{j=1}^{i} p_j$

Ο Δείκτης του Gini

Ορισμός

- δείκτης (στατιστικό) ανισότητας, G, που ορίζεται ως το πηλίκο του εμβαδού (A) της περιοχής μεταξύ της ευθείας της τέλειας ισότητας και της καμπύλης του Lorenz, προς το εμβαδό $(A + B)$ της περιοχής κάτω από την ευθεία τέλειας ισότητας
- ο δείκτης του Gini παίρνει τιμές από 0 – για την περίπτωση της τέλειας ισότητας, ως 1 – για την περίπτωση της τέλειας ανισότητας



Πηγή: Wikipedia

Υπολογισμός

Αφού $G = A/(A + B)$, και ισχύει $A + B = 0.5$, ο δείκτης Gini δίνεται από τις σχέσεις: $G = A/0.5 = 2A = 1 - 2B$, και συνήθως υπολογίζεται ως: $G = 1 - 2\int_0^1 L(p)\, dp$, όπου p είναι η τετμημένη κάθε σημείου της καμπύλης του Lorenz $L(p)$

Ο Δείκτης του Gini (2)

Σχετική μέση διαφορά

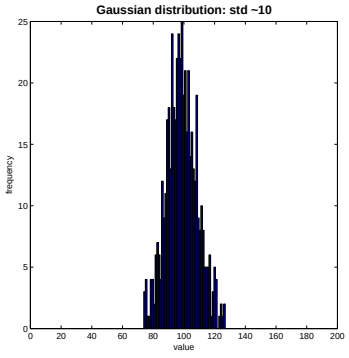
- Μέση διαφορά (MD): Μέση τιμή των απόλυτων διαφορών μεταξύ όλων των N^2 ζευγών μετρήσεων $\{(x_i, x_j), i = 1, \ldots, N, j = 1, \ldots, N\}$, που υπολογίζεται ως: $MD = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1}^{N} |x_i - x_j|$
- ένα στατιστικό διασποράς των μετρήσεων, όχι γύρω από το μέσο όρο του δείγματος $\bar{x}$, αλλά μεταξύ των ίδιων των μετρήσεων $|x_i - x_j|$
- Σχετική μέση διαφορά (RMD): πηλίκο μέσης διαφοράς (MD) προς το μέσο όρο του δείγματος ($\bar{x}$), για την απαλειφή μονάδων μέτρησης: $RMD = MD/\bar{x}$

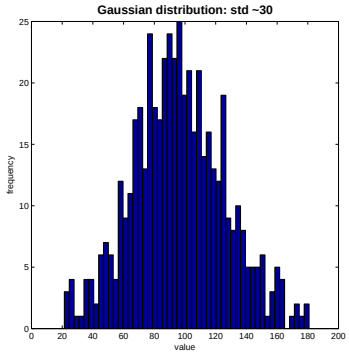
Σχετική μέση διαφορά και δείκτης του Gini

- ο δείκτης του Gini μπορεί να εκφραστεί και ως το μισό της σχετικής μέσης διαφοράς: $G = RMD/2$
- εναλλακτικός τρόπος ορισμού και υπολογισμού του δείκτη του Gini, χωρίς να υπεισέρχεται η έννοια της καμπύλης του Lorenz

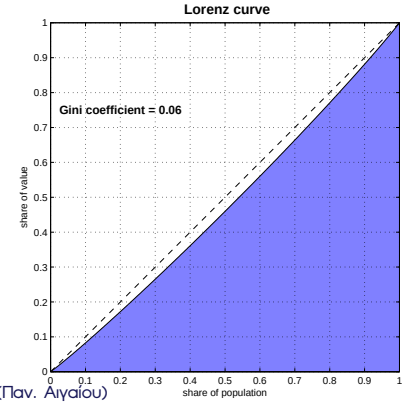
Καμπύλη του Lorenz και Δείκτης του Gini: Παραδείγματα

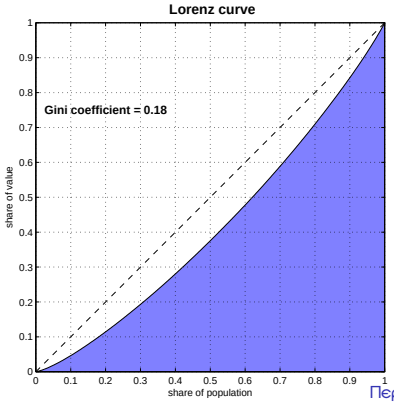
Κατανομές με διαφορετικές τυπικές αποκλίσεις





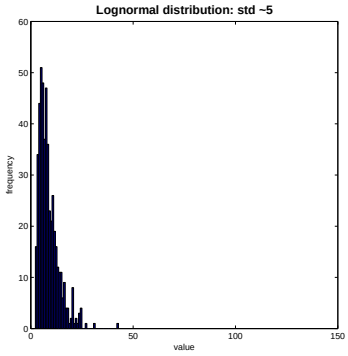
Καμπύλες του Lorenz και δείκτες του Gini

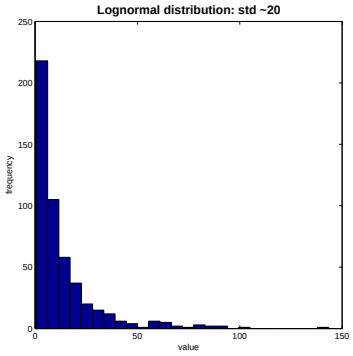




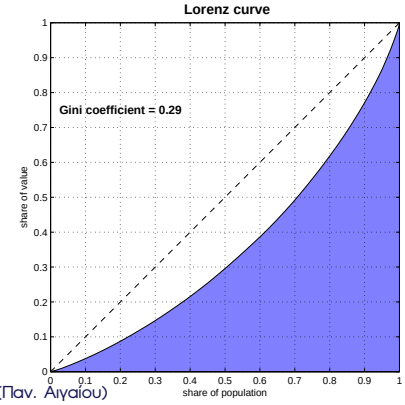
Καμπύλη του Lorenz και Δείκτης του Gini: Παραδείγματα (2)

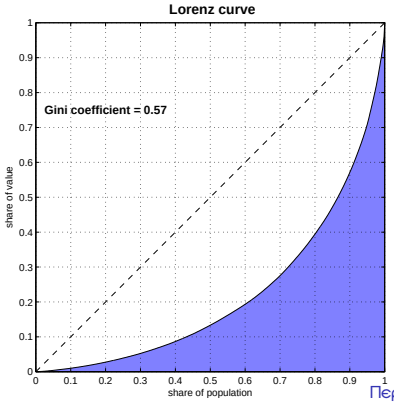
Κατανομές με διαφορετικές τυπικές αποκλίσεις





Καμπύλες του Lorenz και δείκτες του Gini





Ανακεφαλαίωση

Περιγραφική στατιστική

Ένα σύνολο μεθόδων και εργαλείων (γραφικών και ποσοτικών) για την αποτελεσματική παρουσίαση δεδομένων μέσω της σύνοψης, ομαδοποίησης και απεικόνισής τους

Μερικά εργαλεία

- ομαδοποίηση δεδομένων: μέσω της δειγματικής συνάρτησης πυκνότητας πιθανότητας και της δειγματικής συνάρτησης αθροιστικής κατανομής...
- γραφήματα: ραβδογράμματα, θηκογράμματα, γραφήματα ποσοστημορίων...
- στατιστικά μέτρα: αριθμητικές ποσότητες που περιγράφουν κεντρική τάση ή θέση, μεταβλητότητα ή διασπορά...
- περιγραφή ανισότητας ή "διαφορετικότητας": καμπύλη του Lorenz, δείκτης του Gini. *Εναλλακτικοί δείκτες:* εντροπία κατανομής, δείκτης του Theil, δείκτης του Atkinson...

Υπενθύμηση

Τα περιγραφικά στατιστικά και οι δείκτες αποτελούν περιληπτικά μέτρα μιας κατανομής. Πολύπλοκες κατανομές δεν μπορούν να περιγραφούν επαρκώς από ένα-δύο στατιστικά. Κατά συνέπεια, το εμπειρικό ιστόγραμμα, ή καλύτερα η εμπειρική συνάρτηση αθροιστικής κατανομής, πρέπει πάντα να υπολογίζονται και να μελετώνται...