Στατιστική Ι

Γιώργος Τσιρογιάννης

Τμήμα Διοίκησης Επιχειρήσεων Αγροτικών Προϊόντων και Τροφίμων,

Πανεπιστήμιο Πατρών

Ακαδημαϊκό έτος: 2021-2022

Διάλεξη 10η

Περιγραφική στατιστική

- Ποσοτικές μεταβλητές

9.1-9.1.3.1

Πληθυσμός και Τυχαίο Δείγμα

- **Απλό στοιχείο ή πειραματική/δειγματοληπτική μονάδα** είναι κάθε υποκείμενο επί του οποίου μετράμε/παρατηρούμε την τιμή μιας τυχαίας μεταβλητής
- Η κατανομή των τιμών μιας τυχαίας μεταβλητής ονομάζεται πληθυσμός ή **στατιστικός πληθυσμός.**
- **Τυχαίο δείγμα μεγέθους ν από έναν πληθυσμό**, δηλαδή, από την κατανομή των τιμών μιας τυχαίας μεταβλητής Χ, ονομάζουμε ν ανεξάρτητες τυχαίες μεταβλητές $\mathrm{X}_1, \mathrm{X}_2, \mathrm{X}_3, \ldots, \mathrm{X}_\nu$ που παίρνουν τιμές από τον πληθυσμό αυτό, που ακολουθούν δηλαδή την ίδια κατανομή, αυτήν της τ.μ. Χ.
- Οι συγκεκριμένες τιμές $\mathrm{x}_1, \mathrm{x}_2, \mathrm{x}_3, \ldots, \mathrm{x}_\nu$, της Χ που έχουμε διαθέσιμες για επεξεργασία μετά τη λήψη του δείγματος αποτελούν μια **πραγματοποίηση των $\boldsymbol{X_1, X_2, X_3, \ldots, X_\nu}$ και ονομάζονται δεδομένα** ή παρατηρήσεις.

Παράδειγμα

- Η πτυχιακή εργασία ενός φοιτητή αφορούσε στα άνθη μιας συγκεκριμένης ποικιλίας ενός φυτού που καλλιεργείται στο νομό Κοζάνης.
- Στο πλαίσιο αυτής της μελέτης, ο φοιτητής μέτρησε, μεταξύ άλλων, τον αριθμό των πετάλων σε 115 άνθη της συγκεκριμένης ποικιλίας που επέλεξε τυχαία από καλλιέργειες του νομού Κοζάνης.
- Τα αποτελέσματα αυτών των μετρήσεων φαίνονται στον πίνακα

7	5	8	7	5	5	6	6	5	7	5	5	5	9	6	8	5
5	5	6	6	5	5	6	5	9	6	5	5	7	6	6	7	5
7	5	5	6	6	5	6	5	6	5	5	5	5	6	6	5	5
8	5	5	5	5	6	5	5	5	6	5	5	6	5	5	5	6
7	5	7	5	5	8	5	5	5	6	5	10	5	6	5	5	6
5	7	5	5	5	9	5	5	7	5	5	5	5	6	7	5	5
6	5	6	5	7	5	10	5	6	5	5	5	8				

Παράδειγμα

- Η τυχαία μεταβλητή Χ που μελέτησε ο φοιτητής εκφράζει τον αριθμό των πετάλων του άνθους της συγκεκριμένης ποικιλίας φυτών
- Τα 115 άνθη που επέλεξε τυχαία από τις καλλιέργειες, αποτελούν τις 115 δειγματοληπτικές μονάδες (απλά στοιχεία) από τις οποίες αντίστοιχα πήρε τις 115 τιμές $x_1, x_2, x_3, \ldots, x_{115}$ η τμ Χ.
- Ο πίνακας αποτελεί το τυχαίο δείγμα.

Παράδειγμα

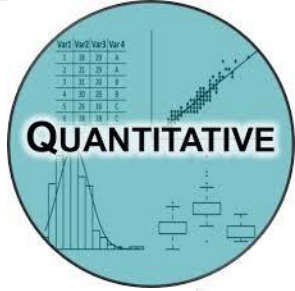
- Ο πληθυσμός που μελέτησε ο φοιτητής, με βάση το τυχαίο δείγμα τιμών που πήρε από αυτόν, είναι η κατανομή των τιμών της Χ.
- Αποτελείται από όλους τους αριθμούς πετάλων που αντιστοιχούν σε όλα τα άνθη όλων των φυτών της συγκεκριμένης ποικιλίας στο νομό Κοζάνης.
- Δείγμα είναι το σύνολο των 115 η $x_1, x_2, x_3, \ldots, x_{115}$ τιμών που παίρνει η τμ Χ.

Στόχος περιγραφικής στατιστικής

- Η περιγραφικής στατιστική μας προσφέρει μεθόδους επεξεργασίας των δεδομένων για να μπορέσουμε, πριν προχωρήσουμε σε επαγωγικά συμπεράσματα για τον πληθυσμό από τον οποίο προέρχονται, να περιγράψουμε και να κατανοήσουμε την κατανομή των δεδομένων (εμπειρική κατανομή).

Ποσοτικές Μεταβλητές (quantitative analysis)

Ποσοτικές Μεταβλητές



- Παίρνουν μόνο αριθμητικές τιμές και διακρίνονται σε συνεχείς (continuous) και διακριτές (discrete).
  - Πίνακας συχνοτήτων
  - Γραφική παρουσίαση κατανομής συχνοτήτων
    - Σημειόγραμμα
    - Ραβδόγραμμα
    - Διάγραμμα συχνοτήτων
    - Κυκλικό διάγραμμα συχνοτήτων
    - Ιστόγραμμα συχνοτήτων
    - Πολύγωνο συχνοτήτων
    - Θηκόγραμμα
- Αριθμητικά περιγραφικά μέτρα
  - Μέτρα Θέσης/Κεντρικής Τάσης
  - Μέτρα Μεταβλητότητας/Διασποράς
  - Μέτρα Λοξότητας (skewness) και Κύρτωσης (kurtosis)

Πίνακας (κατανομής) συχνοτήτων

- Μας δείχνει ποιες τιμές της μεταβλητής που μελετάμε και πόσο συχνά η κάθε μια εμφανίσθηκαν στο δείγμα.
- Συνήθως αποτελείται από τρεις στήλες:
  - πρώτη στήλη καταγράφονται σε αύξουσα σειρά οι k διαφορετικές τιμές της τμ X
  - η συχνότητα (frequency) εμφάνισης, ν_i , κάθε τιμής , y_i i = 1,2,…, k , δηλαδή, πόσες φορές εμφανίσθηκε η αντίστοιχη τιμή, y_i , στο δείγμα
  - η σχετική συχνότητα (relative frequency) εμφάνισης, f_i , κάθε τιμής y_i

$$f_i = \frac{\nu_i}{\nu} \text{ ή } f_i = \frac{\nu_i}{\nu} \cdot 100\%$$

Πίνακας (κατανομής) συχνοτήτων

- Πολλές φορές συμπληρώνεται από την:
- αθροιστική συχνότητα (cumulative frequency), N_i , κάθε τιμής y_i , που ορίζεται ως το άθροισμα των συχνοτήτων όλων των τιμών που είναι μικρότερες ή ίσες της y_i
- αθροιστική σχετική συχνότητα (cumulative relative frequency), F_i , κάθε τιμής y_i , που ορίζεται ως το άθροισμα των σχετικών συχνοτήτων όλων των τιμών που είναι μικρότερες ή ίσες της y_i

Παράδειγμα

- Η πτυχιακή εργασία ενός φοιτητή αφορούσε στα άνθη μιας συγκεκριμένης ποικιλίας ενός φυτού που καλλιεργείται στο νομό Κοζάνης.
- Στο πλαίσιο αυτής της μελέτης, ο φοιτητής μέτρησε, μεταξύ άλλων, τον αριθμό των πετάλων σε 115 άνθη της συγκεκριμένης ποικιλίας που επέλεξε τυχαία από καλλιέργειες του νομού Κοζάνης.
- Τα αποτελέσματα αυτών των μετρήσεων φαίνονται στον πίνακα

7	5	8	7	5	5	6	6	5	7	5	5	5	9	6	8	5
5	5	6	6	5	5	6	5	9	6	5	5	7	6	6	7	5
7	5	5	6	6	5	6	5	6	5	5	5	5	6	6	5	5
8	5	5	5	5	6	5	5	5	6	5	5	6	5	5	5	6
7	5	7	5	5	8	5	5	5	6	5	10	5	6	5	5	6
5	7	5	5	5	9	5	5	7	5	5	5	5	6	7	5	5
6	5	6	5	7	5	10	5	6	5	5	5	8				

Παράδειγμα (συν)

7	5	8	7	5	5	6	6	5	7	5	5	5	9	6	8	5
5	5	6	6	5	5	6	5	9	6	5	5	7	6	6	7	5
7	5	5	6	6	5	6	5	6	5	5	5	5	6	6	5	5
8	5	5	5	5	6	5	5	5	6	5	5	6	5	5	5	6
7	5	7	5	5	8	5	5	5	6	5	10	5	6	5	5	6
5	7	5	5	5	9	5	5	7	5	5	5	5	6	7	5	5
6	5	6	5	7	5	10	5	6	5	5	5	8				

y_i	ν_i	f_i	N_i	F_i
5	67	0.5826	67	0.5826
6	26	0.2261	93	0.8087
7	12	0.1043	105	0.9130
8	5	0.0435	110	0.9565
9	3	0.0261	113	0.9826
10	2	0.0174	115	1.0000
Σύνολα	**115**	**1.0000**		

Πιο συχνές τιμές

καλύπτουν το 91%

105 από τα 115 δείγματα
έχουν 5 έως 7 φύλλα

Γενικό παρατήρηση: ένα πολύ μεγάλο ποσοστό των τιμών του δείγματος συγκεντρώνεται στο αριστερό άκρο της κατανομής και ότι οι συχνότητες φθίνουν αυξανομένου του αριθμού των πετάλων

Παράδειγμα

- Επελέγησαν τυχαία 20 οικογένειες από το σύνολο των οικογενειών που κατοικούν μόνιμα στην επαρχία Γορτυνίας και για κάθε μια από αυτές καταγράφηκε ο αριθμός παιδιών της οικογένειας.

Αριθμός παιδιών οικογένειας u_i
0
1
0
2
2
2
3
2
4
1
1
2
3
4
1
2
2
2
2
2

Παράδειγμα (συν.)

Αριθμός παιδιών οικογένειας u_i
0
1
0
2
2
2
3
2
4
1
1
2
3
4
1
2
2
2
2
2

Πίνακας (κατανομής) συχνοτήτων

y_i	ν_i	f_i	N_i	F_i
0	2	0.1	2	0.1
1	4	0.2	6	0.3
2	10	0.5	16	0.8
3	2	0.1	18	0.9
4	2	0.1	20	1.0
Σύνολα	**20**	**1.0**		

Γενική παρατήρηση: η τιμή 2 παρουσιάζει τη μεγαλύτερη συχνότητα και ότι αριστερά αυτής της τιμής οι συχνότητες αυξάνουν αυξανομένου του αριθμού των παιδιών, ενώ δεξιά αυτής της τιμής, οι συχνότητες φθίνουν αυξανομένου του αριθμού των παιδιών

Παράδειγμα (ομαδοποίησης)

- Καταγράφεται για κάθε μια από 50 τυχαία επιλεγμένες γαλακτοπαραγωγές αγελάδες, ο χρόνος X (σε μήνες), από την πρώτη εκδήλωση μιας συγκεκριμένης ασθένειας από την οποία είχαν προσβληθεί, μέχρι την επανεμφάνισή της.

2.1	1.7	0.8	0.8	4.1	8.7	1.4	2.9	1.9	2.7
4.4	2.2	5.5	7.0	1.8	0.2	1.0	0.9	4.0	0.7
2.0	6.5	0.7	4.3	0.2	5.6	2.4	1.4	1.3	1.2
0.5	3.9	7.4	3.3	8.8	0.3	2.0	5.7	0.8	2.6
9.9	1.6	2.8	1.0	0.6	1.3	0.8	5.9	0.9	0.4

Παράδειγμα (ομαδοποίησης)

Για το 68% βλέπουμε επανεμφάνισης τη ασθένειας σε λιγότερο από 3 μήνες

2.1	1.7	0.8	0.8	4.1	8.7	1.4	2.9	1.9	2.7
4.4	2.2	5.5	7.0	1.8	0.2	1.0	0.9	4.0	0.7
2.0	6.5	0.7	4.3	0.2	5.6	2.4	1.4	1.3	1.2
0.5	3.9	7.4	3.3	8.8	0.3	2.0	5.7	0.8	2.6
9.9	1.6	2.8	1.0	0.6	1.3	0.8	5.9	0.9	0.4

Χρόνος Επανεμφάνισης	ν_i	f_i	N_i	F_i
[0.0 1.5)	21	0.42	21	0.42
[1.5 3.0)	13	0.26	34	0.68
[3.0 4.5)	6	0.12	40	0.80
[4.5 6.0)	4	0.08	44	0.88
[6.0 7.5)	3	0.06	47	0.94
[7.5 9.0)	2	0.04	49	0.98
[9.0 10.5)	1	0.02	50	1.00
Σύνολα	**50**	**1.00**		

Πόσες κλάσεις;

$$k = 1 + 3.32 \cdot \log_{10}(\nu)$$

Εύρος διαστημάτων ίσου μήκους:

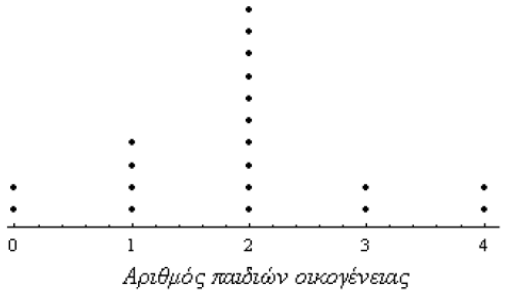
$$r = \frac{R}{k} = \frac{x_{\max} - x_{\min}}{k}$$

Γραφική παρουσίαση κατανομής συχνοτήτων

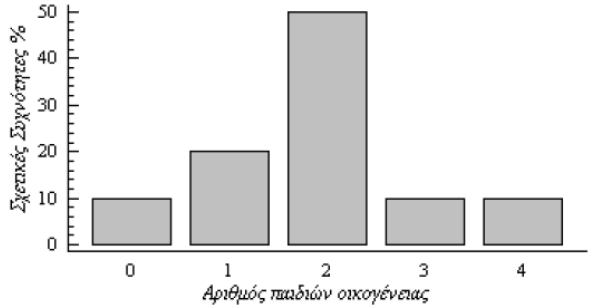
Αριθμός παιδιών οικογένειας u_i
0
1
0
2
2
2
3
2
4
1
1
2
3
4
1
2
2
2
2
2

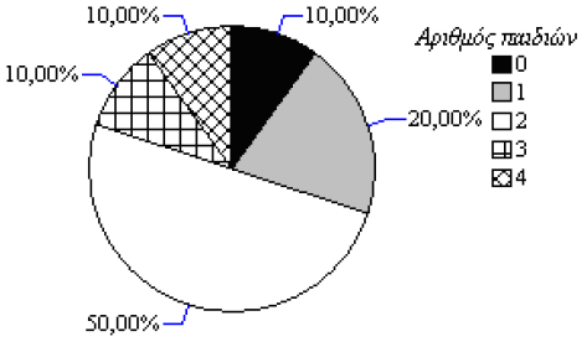
y_i	ν_i	f_i	N_i	F_i
0	2	0.1	2	0.1
1	4	0.2	6	0.3
2	10	0.5	16	0.8
3	2	0.1	18	0.9
4	2	0.1	20	1.0
Σύνολα	**20**	**1.0**		

Σημειόγραμμα



Ραβδόγραμμα



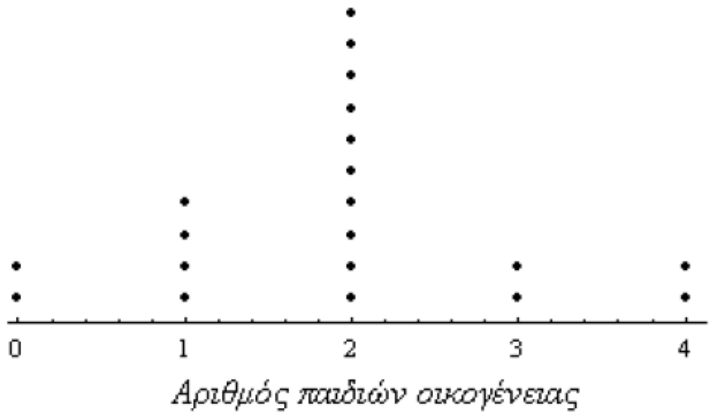


Κυκλικό διάγραμμα

Διάγραμμα σχετικών συχνοτήτων

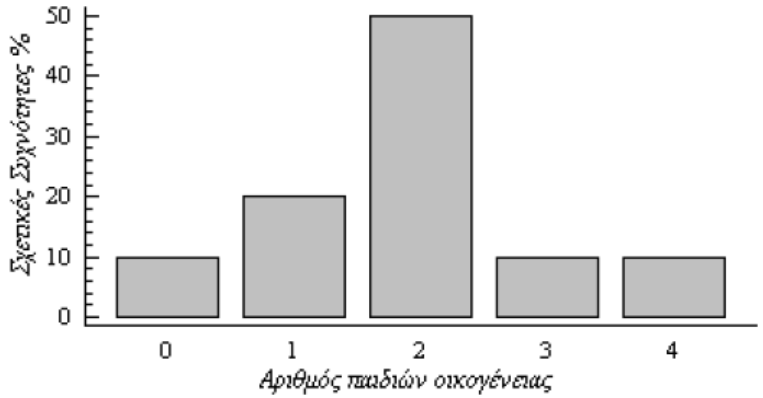
Σημειόγραμμα

- Στο ***σημειόγραμμα***, απεικονίζουμε τα ***δεδομένα*** ως κουκίδες στις αντίστοιχες θέσεις ενός οριζόντιου άξονα.



Ραβδόγραμμα

- Στο ραβδόγραμμα, απεικονίζουμε τις συχνότητες ή τις σχετικές συχνότητες ως ύψη ορθογωνίων που σχεδιάζουμε στις αντίστοιχες θέσεις του οριζόντιου άξονα. Τα ορθογώνια έχουν ίδιο πλάτος βάσης που επιλέγουμε αυθαίρετα.



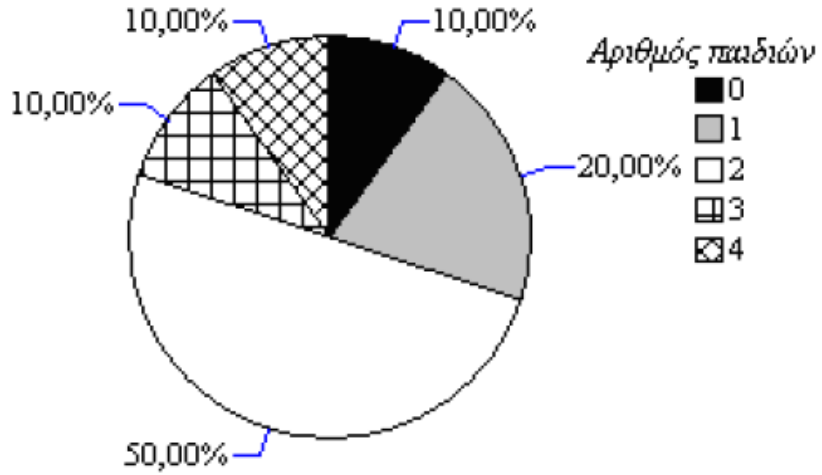
Διάγραμμα συχνοτήτων

- Στο διάγραμμα συχνοτήτων, απεικονίζουμε τις συχνότητες (αντίστοιχα τις σχετικές συχνότητες) των διαφορετικών τιμών στο ραβδόγραμμα, με τη διαφορά ότι στις θέσεις χαράσσουμε κάθετα ευθύγραμμα τμήματα αντί ορθογωνίων.

Κυκλικό διάγραμμα

- Πρόκειται για έναν κυκλικό δίσκο χωρισμένο σε k κυκλικούς τομείς, έναν για κάθε y_i, τα τόξα των οποίων, έστω φ_i, είναι ανάλογα με τις αντίστοιχες συχνότητες και σχετικές συχνότητες.

$$\varphi_i = \nu_i \cdot \frac{360^0}{\nu} = 360^0 \cdot f_i, \quad i = 1, 2, \ldots, k$$

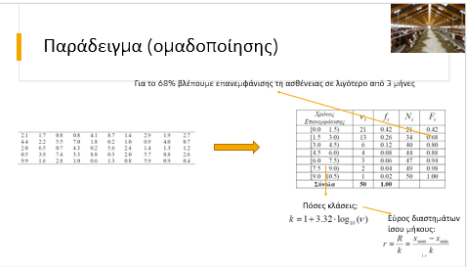


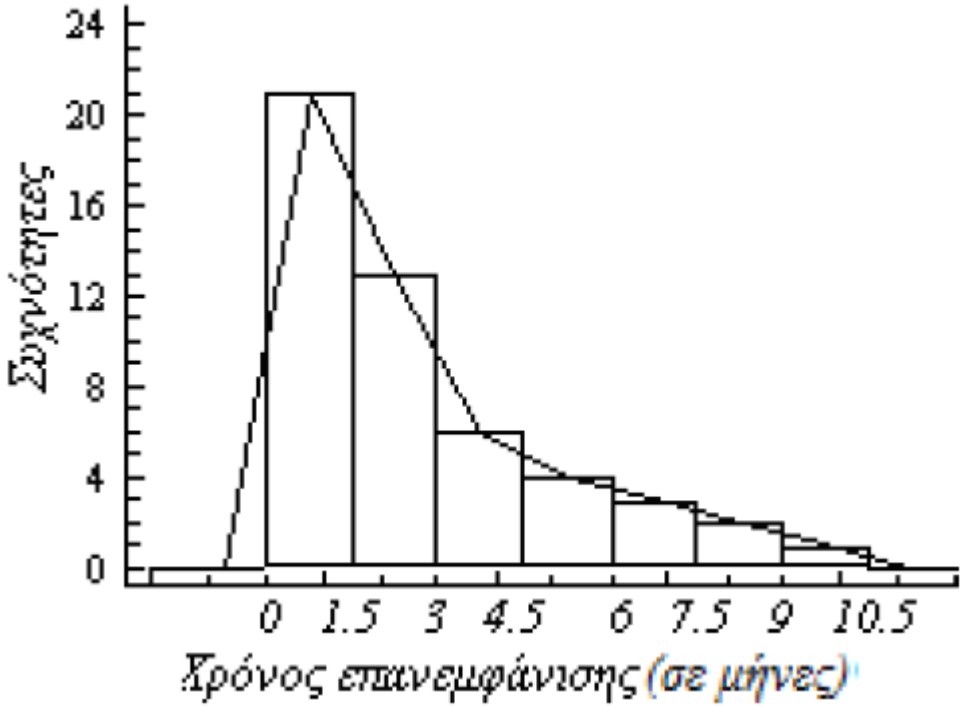
Ιστόγραμμα

- Τα ιστογράμματα κατασκευάζονται για τη γραφική παρουσίαση/αναπαράσταση της πληροφορίας που παίρνουμε από έναν πίνακα συχνοτήτων ομαδοποιημένων δεδομένων.
- Δομή: σε ορθογωνίου συστήματος αξόνων σημειώνουμε ορθογώνια έτσι ώστε, το εμβαδόν του να είναι ίσο με τη συχνότητα της αντίστοιχης κλάσης αν πρόκειται για το ιστόγραμμα συχνοτήτων

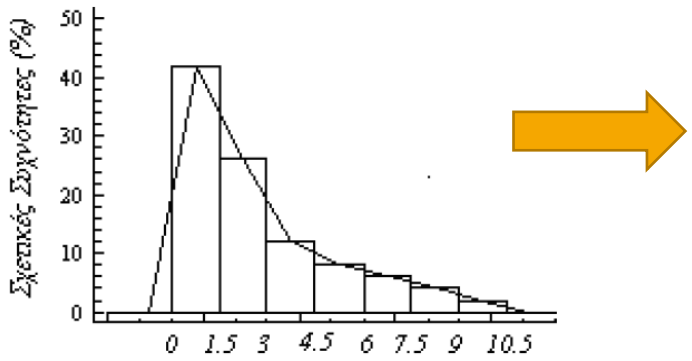
Παράδειγμα (συν.)

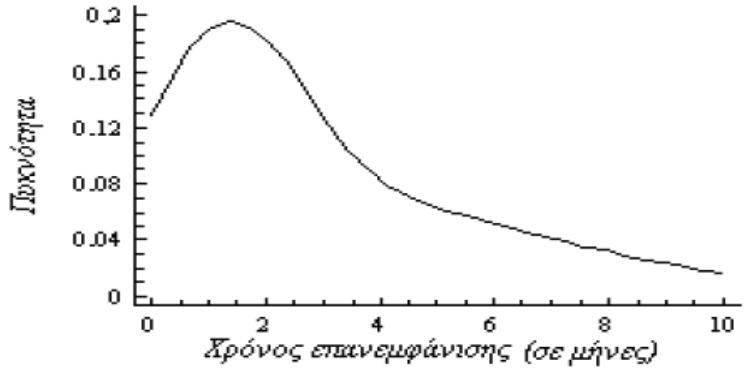
Χρόνος Επανεμφάνισης	ν_i	f_i	N_i	F_i
[0.0 1.5)	21	0.42	21	0.42
[1.5 3.0)	13	0.26	34	0.68
[3.0 4.5)	6	0.12	40	0.80
[4.5 6.0)	4	0.08	44	0.88
[6.0 7.5)	3	0.06	47	0.94
[7.5 9.0)	2	0.04	49	0.98
[9.0 10.5)	1	0.02	50	1.00
Σύνολα	**50**	**1.00**		

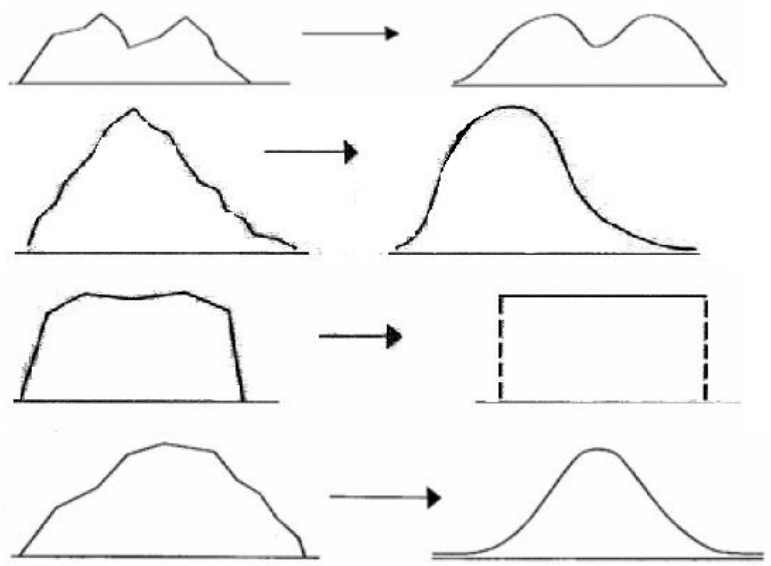




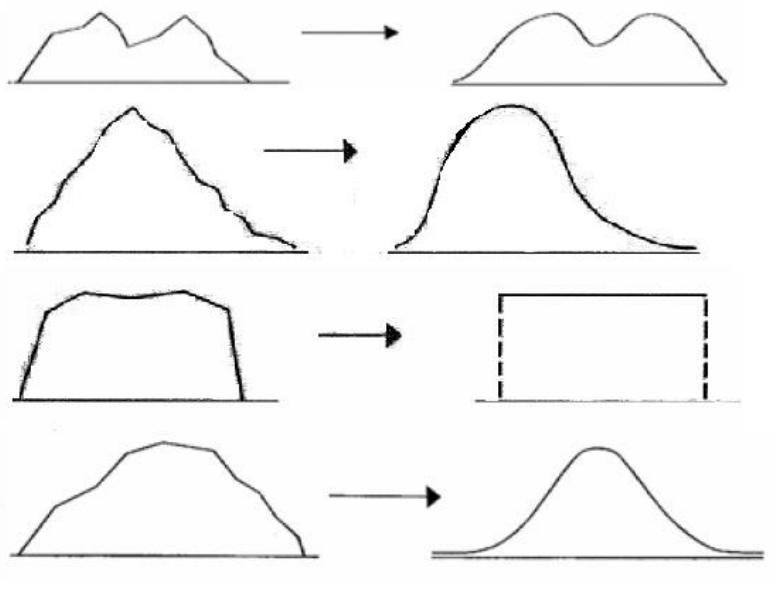
Από τα πολύγωνα στις καμπύλες τού πληθυσμού





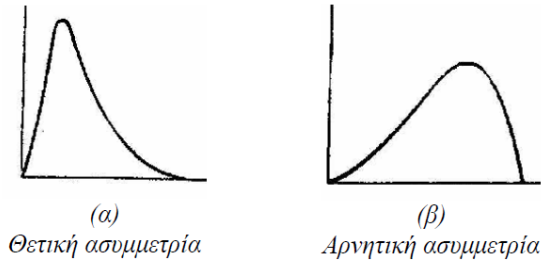


Σχόλια



καμία ή και περισσότερες κορυφές

συμμετρικές ή να είναι λοξές/ασύμμετρες



μεσόκυρτες, λεπτόκυρτες,* και *πλατύκυρτες

Αριθμητικά περιγραφικά μέτρα

Αριθμητικά περιγραφικά μέτρα (numerical descriptive measures)

- Ποσοτικά μεγέθη που βοηθούν στην περιγραφή της κατανομής ενός δείγματος ή στην περιγραφή ενός πληθυσμού
- Γνώριμα από την θεωρία πιθανοτήτων:
  - Αν αφορούν έναν πληθυσμό, ονομάζονται **παράμετροι (parameters)**
  - Αν αφορούν ένα δείγμα από έναν πληθυσμό ονομάζονται **στατιστικά (statistics)**

Διαφορές

- Οι παράμετροι ενός πληθυσμού είναι συγκεκριμένοι/μοναδικοί αριθμοί (γνωστοί είτε άγνωστοι).
- Τα στατιστικά, για συγκεκριμένη πραγματοποίηση $\boldsymbol{x_1, x_2, x_3, \ldots, x_\nu}$ ενός δείγματος $\boldsymbol{X_1, X_2, X_3, \ldots, X_\nu}$ μπορούμε να τα υπολογίσουμε και επομένως μας είναι γνωστά
- Σε μια άλλη πραγματοποίηση του δείγματος η τιμή τους μεταβάλλεται, δηλαδή, τα στατιστικά είναι τυχαίες μεταβλητές.
- Κάθε στατιστικό, **είναι τυχαία μεταβλητή** και για αυτό συμβολίζεται με κεφαλαίο γράμμα όπως οι τυχαίες μεταβλητές

Μέτρα θέσης/Κεντρικής τάσης

- Έστω $\boldsymbol{X_1, X_2, X_3, \ldots, X_\nu}$ ένα τυχαίο δείγμα από έναν πληθυσμό (από την κατανομή μιας τυχαίας μεταβλητή Χ) και $\boldsymbol{x_1, x_2, x_3, \ldots, x_\nu}$ μια πραγματοποίησή του.
- Έστω επίσης, $\boldsymbol{y_1, y_2, y_3, \ldots, y_k}$ οι k διαφορετικές μεταξύ τους τιμές από τις $\boldsymbol{x_1, x_2, x_3, \ldots, x_\nu}$.

Δειγματικός μέσος (sample mean/arithmetic mean/average)

- Ορίζεται από τον τύπο: $\bar{X} = \frac{1}{\nu}\sum_{i=1}^{\nu} X_i$
- Η συγκεκριμένη τιμή του $\bar{X}$, που υπολογίζεται για μια πραγματοποίηση $\boldsymbol{x_1, x_2, x_3, \ldots, x_\nu}$ του τυχαίου δείγματος $\boldsymbol{X_1, X_2, X_3, \ldots, X_\nu}$ συμβολίζεται με $\bar{x}$, δηλαδή $\bar{x} = \frac{1}{\nu}\sum_{i=1}^{\nu} x_i$
- Ευαίσθητο σε *ακραίες-έκτροπες (outlying* τιμες)
  - αποκρύπτει τις έκτροπες τιμές

Παράδειγμα απόκρυψης

- Ο ιδιοκτήτης μιας μικρής επιχείρησης που απασχολεί πέντε εργαζομένους ισχυρίσθηκε ότι οι εργαζόμενοι στην επιχείρησή του είναι πολύ καλά αμειβόμενοι αφού ο μέσος μισθός τους είναι 2000€.
- Οι μισθοί των εργαζομένων ήταν 400, 400, 500, 700 και 8000 € αντίστοιχα!
- Ο μισθός των 8000 € ήταν του manager και συνιδιοκτήτη.

Παράδειγμα

y_i	ν_i	f_i	N_i	F_i
0	2	0.1	2	0.1
1	4	0.2	6	0.3
2	10	0.5	16	0.8
3	2	0.1	18	0.9
4	2	0.1	20	1.0
Σύνολα	**20**	**1.0**		

$$\bar{x} = \frac{\sum_{i=1}^{k} \nu_i y_i}{\nu} = \frac{38}{20} = 1.9 \text{ παιδιά}$$

Ιδιότητες

Το άθροισμα των αποκλίσεων των τιμών $x_1, x_2, \ldots, x_\nu$ από τον *δειγματικό μέσο* $\bar{x}$ είναι 0. Δηλαδή,

$$\sum_{i=1}^{\nu}(x_i - \bar{x}) = \sum_{i=1}^{k}(y_i - \bar{x})\nu_i = 0$$

Το άθροισμα $\sum_{i=1}^{\nu}(x_i - \lambda)^2$ γίνεται ***ελάχιστο*** αν και μόνο αν $\lambda = \bar{x}$

Αν $t_i = x_i + \beta$ τότε $\bar{t} = \bar{x} + \beta$. Δηλαδή, αν στις τιμές $x_1, x_2, \ldots, x_\nu$ του δείγματος προσθέσουμε μια σταθερή ποσότητα β (θετική ή αρνητική), τότε η ο *δειγματικός μέσος* αυξάνεται (ή μειώνεται) κατά την ίδια ποσότητα.

Ιδιότητες

Αν $t_i = \alpha x_i$ τότε $\bar{t} = \alpha \bar{x}$. Δηλαδή, αν οι τιμές $x_1, x_2, \ldots, x_\nu$ του δείγματος πολλαπλασιασθούν με την ίδια ποσότητα α, τότε ο *δειγματικός μέσος* πολλαπλασιάζεται με την ίδια ποσότητα.

Αν $t_i = \alpha x_i + \beta$ τότε $\bar{t} = \alpha \bar{x} + \beta$

Κριτική

Πλεονεκτήματα

- Είναι πολύ απλό στον υπολογισμό.
- Είναι πολύ χρήσιμο στον έλεγχο ποιότητας.
- Μπορεί να χρησιμοποιηθεί για την εκτίμηση της τυπικής απόκλισης.

Μειονεκτήματα

- Επηρεάζεται από ακραίες τιμές.
- Μπορεί να μην αντιστοιχεί σε δυνατή τιμή της μεταβλητής.
- Δεν υπολογίζεται για ποιοτικά δεδομένα.

Σταθμικός μέσος

- Στις περιπτώσεις που τα δεδομένα $x_1, x_2, x_3, \ldots, x_\nu$, έχουν διαφορετική αξία/βάρος $w_1, w_2, w_3, \ldots, w_\nu$ αντίστοιχα, υπολογίζεται ο σταθμικός μέσος (weighted mean) ο οποίος ορίζεται με τον τύπο

$$\bar{x}_w = \frac{\sum_{i=1}^{\nu} w_i x_i}{\sum_{i=1}^{\nu} w_i}$$

Παράδειγμα

- Ένας οδηγός φορτηγού διανομής τροφίμων, αγόρασε σε μια ημέρα πετρέλαιο από τρία διαφορετικά πρατήρια. Από το πρώτο αγόρασε 6 λίτρα προς 0.75 € το λίτρο, από το δεύτερο 12 λίτρα προς 0.84 € το λίτρο και από το τρίτο 5 λίτρα προς 0.76 € το λίτρο.

Παράδειγμα

$$\bar{x}_w = \frac{\sum_{i=1}^{\nu} w_i x_i}{\sum_{i=1}^{\nu} w_i} = \frac{6 \cdot 0.75 + 12 \cdot 0.84 + 5 \cdot 0.76}{6 + 12 + 5} = 0.799\,€ \text{ ανά λίτρο.}$$

Μέσος των μέσων k δειγμάτων

- Ο μέσος των μέσων k δειγμάτων μεγέθους $\nu_1, \nu_2, \nu_3, \ldots, \nu_\kappa$ αντίστοιχα, είναι

$$\bar{x} = \frac{\sum_{i=1}^{k} \nu_i \bar{x}_i}{\sum_{i=1}^{k} \nu_i}$$

Παράδειγμα

- Αν το μέσο ύψος 10 φοιτητών είναι 170 cm και το μέσο ύψος 5 φοιτητριών είναι 160 cm τότε το μέσο ύψος φοιτητών και φοιτητριών είναι

Παράδειγμα (συν)

$$\bar{x} = \frac{\sum_{i=1}^{2} \nu_i \bar{x}_i}{\sum_{i=1}^{2} \nu_i} = \frac{10 \cdot 170 + 5 \cdot 160}{15} = 166.7\, cm$$

Ισοσταθμισμένος μέσος (trimmed mean)

- Αν από τον υπολογισμό του δειγματικού μέσου θέλουμε να παραλείψουμε τις ακραίες τιμές, μπορούμε να δημιουργήσουμε έναν ισοσταθμισμένο μέσο (trimmed mean) θέτοντας στον σταθμικό μέσο, βάρος 0 για τις ακραίες τιμές που θέλουμε να παραληφθούν και βάρος 1 για όλες τις υπόλοιπες.

Δειγματική Κορυφή ή Επικρατούσα τιμή

- Η κορυφή (mode) της κατανομής του δείγματος είναι η τιμή του δείγματος με τη μεγαλύτερη συχνότητα και συμβολίζεται M_0
- Γραμμική ιδιότητα: αν $t_i = ax_i + \beta$, δηλαδή αν γίνει γραμμικός μετασχηματισμός των ν $x_1, x_2, x_3, \ldots, x_\nu$ τότε και η κορυφή τους, έστω M_{ox}, μετασχηματίζεται αντίστοιχα, δηλαδή, για την κορυφή M_{ot} έχουμε $M_{0t} = \alpha M_{0x} + \beta$

Κριτική

Πλεονεκτήματα

- Υπολογίζεται εύκολα.
- Είναι εύκολα κατανοητή.
- Υπολογίζεται και από ελλιπή δεδομένα.
- Δεν επηρεάζεται από ακραίες τιμές.
- Υπολογίζεται και για ποιοτικά δεδομένα.

Μειονεκτήματα

- Δε χρησιμοποιούνται όλες οι τιμές για τον υπολογισμό της.
- Στη στατιστική συμπερασματολογία έχει περιορισμένη σημασία.
- Δεν ορίζεται πάντα μονοσήμαντα και επίσης μπορεί να μην υπάρχει.

Δειγματική διάμεσος

- Η διάμεσος δ (median) της κατανομής του δείγματος είναι ένας αριθμός για τον οποίο ισχύει ότι το πολύ 50% των τιμών του δείγματος (των παρατηρήσεων) είναι μικρότερες από αυτόν και επίσης το πολύ 50% των τιμών του δείγματος είναι μεγαλύτερες από αυτόν.
- Εκφράζει την κεντρική θέση της κατανομής του δείγματος και για αυτό στη βιβλιογραφία συναντάται και ως μέσος θέσης (position average).
- Αν το μέγεθος του δείγματος ν , είναι αριθμός περιττός $\delta = x_{(\frac{\nu+1}{2})}$
- Αν είναι άρτιος $\delta = \frac{x_{(\frac{\nu}{2})} + x_{(\frac{\nu}{2}+1)}}{2}$

Παράδειγμα

- Έστω οι παρατηρήσεις 5, 2, 9, 6, 11.
  - Τις διατάσσουμε σε αύξουσα σειρά 2, 5, 6, 9, 11
  - Η διάμεσος τιμή είναι αυτή που βρίσκεται στη θέση $0.5(5 + 1) = 3$
  - $\delta = 6$
- Έστω οι παρατηρήσεις 2, 5, 6, 27, 11, 9.
  - Τις διατάσσουμε σε αύξουσα σειρά 2, 5, 6, 9, 11, 27.
  - Επειδή ο αριθμός $0.5(6 + 1) = 3.5$ δεν είναι ακέραιος, η διάμεσος τιμή είναι το ημιάθροισμα της 3ης και της 4ης παρατήρησης,
  - $\delta = (6 + 9)/2 = 7.5$

Κριτική

Πλεονεκτήματα	Μειονεκτήματα
• Είναι εύκολα κατανοητή. • Δεν επηρεάζεται από ακραίες τιμές. • Ο υπολογισμός της είναι απλός. • Είναι μοναδική.	• Δε χρησιμοποιούνται όλες οι τιμές για τον υπολογισμό της. • Δεν υπολογίζεται για ποιοτικά δεδομένα.

Επειδή η διάμεσος δεν επηρεάζεται όπως ο μέσος από ακραίες τιμές, για την περιγραφή παρατηρήσεων που εμφανίζουν ακραίες τιμές προτιμάται ως μέτρο θέσης από τον μέσο.

p-ποσοστιαίο σημείο

- Το p-ποσοστιαίο σημείο p_x είναι ένας αριθμός για τον οποίο ισχύει ότι το πολύ 100 p% των τιμών του δείγματος είναι μικρότερες από αυτόν και το πολύ 100(1− p)% των τιμών του δείγματος είναι μεγαλύτερες από αυτόν

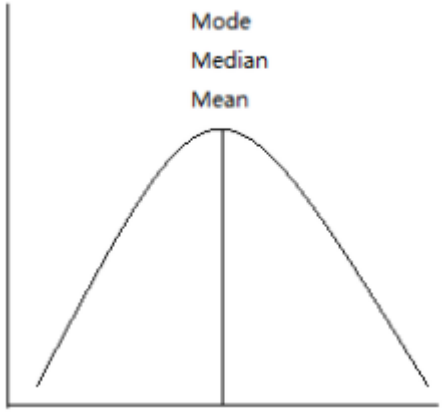
εκατοστημόρια (percentiles) $x_{0.01}, x_{0.02}, \ldots, x_{0.99}$

δεκατημόρια (deciles) $x_{0.1}, x_{0.2}, \ldots, x_{0.9}$

τεταρτημόρια (quartiles) $x_{0.25} = Q_1, \; x_{0.5} = Q_2 = \delta, \; x_{0.75} = Q_3$

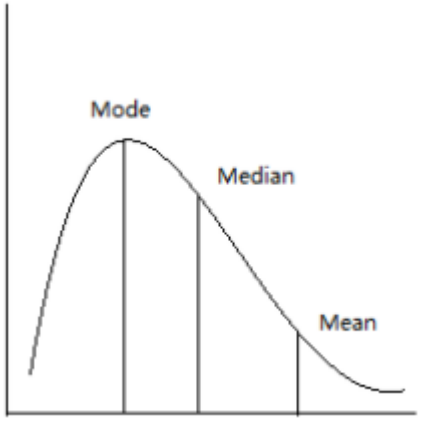
Σχετική θέση δειγματικού μέσου, δειγματικής κορυφής και δειγματικής διαμέσου

- Όταν η καμπύλη συχνοτήτων της κατανομής του δείγματος είναι συμμετρική ισχύει $\bar{x} = \delta = M_0$



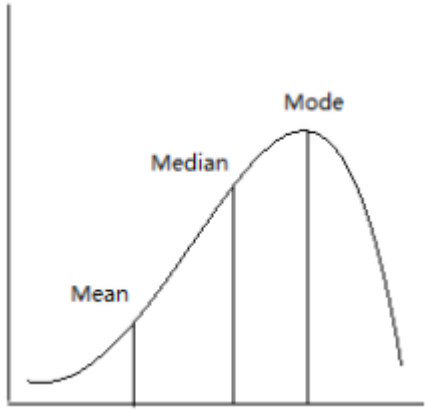
Σχετική θέση δειγματικού μέσου, δειγματικής κορυφής και δειγματικής διαμέσου

- Όταν η καμπύλη συχνοτήτων της κατανομής του δείγματος παρουσιάζει θετική ασυμμετρία ισχύει $\bar{x} > \delta > M_0$



Σχετική θέση δειγματικού μέσου, δειγματικής κορυφής και δειγματικής διαμέσου

- Όταν η καμπύλη συχνοτήτων της κατανομής του δείγματος παρουσιάζει αρνητική ασυμμετρία ισχύει $\bar{x} < \delta < M_0$



Υπάρχουν κατανομές που αποτελούν εξαίρεση των παραπάνω κανόνων.

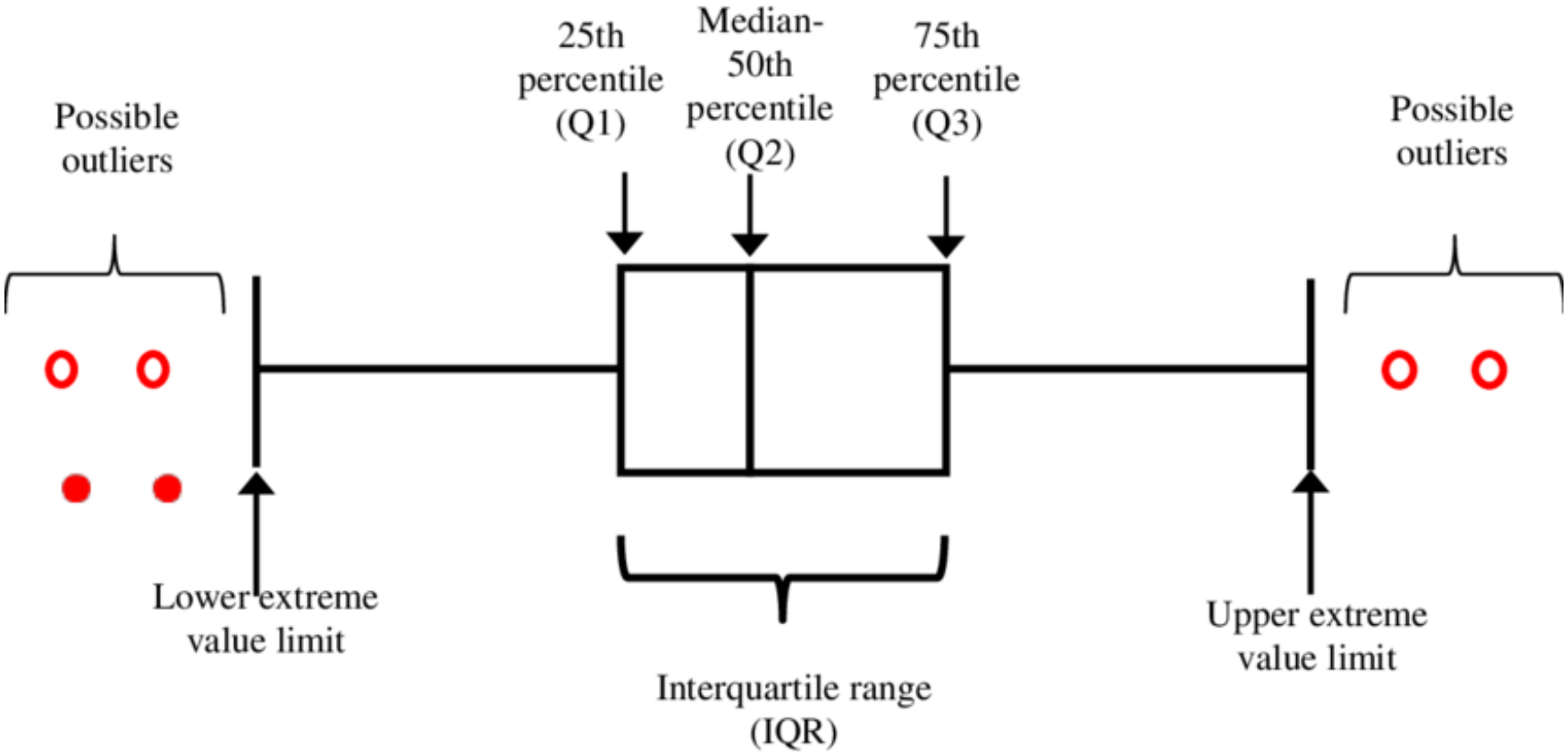
Box plot (θηκόγραμμα)

- Ορθογώνιο με δύο κεραίες (whiskers) το οποίο κατασκευάζεται ως εξής:
  - η κάτω βάση του ορθογωνίου βρίσκεται στο Q_1 και η πάνω στο Q_3
  - η διάμεσος, $\delta = Q_2$, αναπαριστάνεται με ένα οριζόντιο ευθύγραμμο τμήμα μέσα στο ορθογώνιο και στην κατάλληλη θέση
  - Το πλάτος των βάσεων του ορθογωνίου καθορίζεται αυθαίρετα
  - Η πάνω κεραία έχει τη μορφή Τ και εκτείνεται μέχρι την πάνω οριακή τιμή
  - Η κάτω κεραία έχει τη μορφή Τ και εκτείνεται μέχρι την κάτω οριακή τιμή

Box plot (θηκόγραμμα)

- Άνω οριακή τιμή (εναλλακτικά):
- *μέγιστη* τιμή του δείγματος
- η μεγαλύτερη τιμή του δείγματος που είναι μικρότερη ή ίση από το *ανώτερο εσωτερικό φράγμα* $Q_3 + 1.5(Q_3 - Q_1)$
- μεγαλύτερη τιμή του δείγματος που είναι μικρότερη ή ίση από το ανώτερο εξωτερικό φράγμα $Q_3 + 3(Q_3 - Q_1)$
- Οι τιμές εκτός των ορίων σημειώνονται ως σποραδικα σημεία

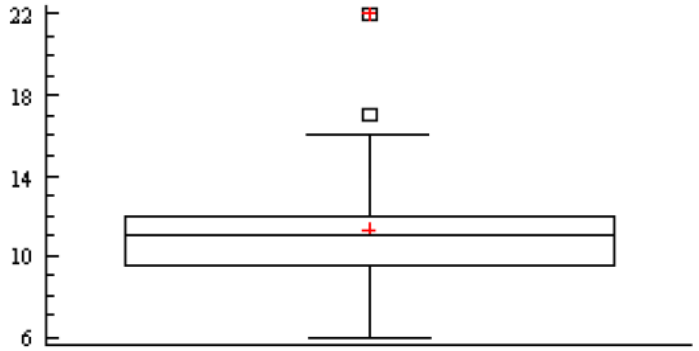
Box plot (θηκόγραμμα)

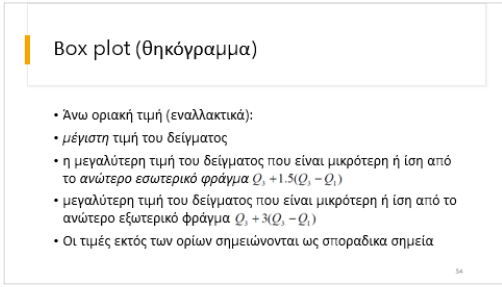


Παράδειγμα

- Έστω το δείγμα 15, 11, 11, 11, 22, 9, 11, 7, 11, 12, 12, 16, 8, 11, 15, 9, 10, 14, 9, 10, 11, 10, 6, 17, 11, 10, 8, 11 που μετρά το ύψος βροχής του Φεβρουαρίου. Να υπολογισθεί το box plot

Παράδειγμα





- Ανώτερο εσωτερικό φράγμα:

$$Q_1 = 9.25,\ Q_3 = 12 \quad \kappa\alpha\iota \quad \delta = 11 \quad Q_3 + 1.5(Q_3 - Q_1) = 12 + 1.5(12 - 9.25) = 16.125$$

- Άρα η πάνω οριακή τιμή είναι η παρατήρηση που είναι ίση με 16 (η μεγαλύτερη παρατήρηση που είναι ίση ή μικρότερη από 16.125)
- Κατώτερο εσωτερικό φράγμα: $Q_1 - 1.5(Q_3 - Q_1) = 9.25 - 1.5(12 - 9.25) = 5.125$
- άρα η *κάτω οριακή τιμή* είναι η παρατήρηση που είναι ίση με 6 *(η μικρότερη παρατήρηση που είναι ίση ή μεγαλύτερη από 5.125)*