ΣΤΑΤΙΣΤΙΚΗ

Γενικές έννοιες

Στατιστική

είναι ο κλάδος των μαθηματικών, ο οποίος ως έργο έχει τη συγκέντρωση στοιχείων, την ταξινόμησή τους και την παρουσίασή τους σε κατάλληλη μορφή, ώστε να μπορούν να αναλυθούν και να ερμηνευθούν για την εξυπηρέτηση διαφόρων σκοπών.

Πληθυσμός

είναι το σύνολο των αντικειμένων (έμψυχων ή άψυχων) για τα οποία συλλέγονται στοιχεία.

Άτομο

ονομάζεται κάθε στοιχείο ενός πληθυσμού ή ενός δείγματος.

Δείγμα

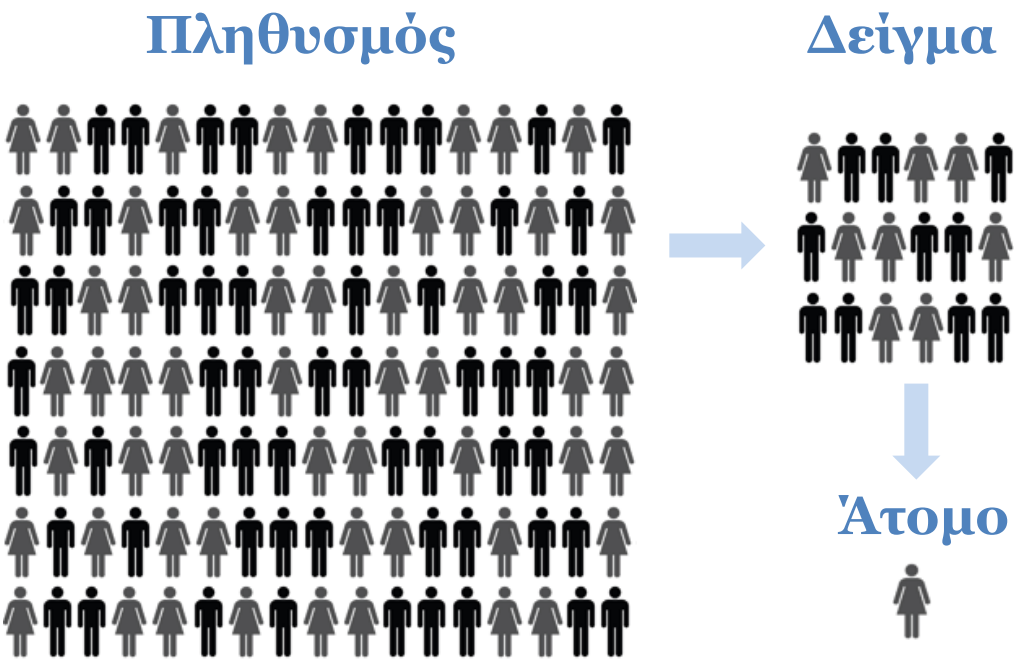
είναι ένα μέρος (υποσύνολο) του πληθυσμού, που είναι αντιπροσωπευτικό του πληθυσμού και από την εξέταση του οποίου βγάζουμε συμπεράσματα για ολόκληρο τον πληθυσμό.

Δειγματοληψία

είναι η εξέταση ενός δείγματος, κάποιου πληθυσμού.

Μέγεθος ή πλήθος (ν)

ενός πληθυσμού (ή ενός δείγματος) ονομάζεται το πλήθος των ατόμων του.



Μεταβλητή

είναι το χαρακτηριστικό ενός πληθυσμού, ως προς το οποίο αυτός εξετάζεται.

- Η μεταβλητές συμβολίζονται με κεφαλαία γράμματα π.χ. συνήθως το γράμμα Χ , ενώ οι τιμές τους με με τα αντίστοιχα πεζά κι ένα μικρό δείκτη αρίθμησης π.χ. x_1 , x_2 , x_3 , ... και γενικά x_i , όπου i = 1, 2, 3, ...

Είδη μεταβλητών

Οι μεταβλητές χωρίζονται σε δύο γενικές κατηγορίες :

Ποιοτικές

είναι εκείνες που δεν επιδέχονται μέτρηση, πχ. χρώμα ματιών, μόρφωση, θρήσκευμα, κλπ.

Ποσοτικές

είναι εκείνες που μπορούν να μετρηθούν, πχ. ύψος, μισθός, ώρες εργασίας, τιμή, κλπ.

Είδη ποσοτικών μεταβλητών

Με τη σειρά τους, οι ποσοτικές μεταβλητές χωρίζονται σε :

Διακριτές

δηλαδή, εκείνες στις οποίες κάθε άτομο του πληθυσμού μπορεί να πάρει μόνο διακεκριμένες τιμές, πχ. αριθμός παιδιών, μέρες διακοπών, κλπ.

Συνεχείς

είναι εκείνες, στις οποίες κάθε άτομο του πληθυσμού μπορεί να πάρει οποιαδήποτε πραγματική τιμή, που ανήκει σε διάστημα (ή ένωση διαστημάτων) πραγματικών αριθμών, πχ. ύψος, βάρος, κλπ.

Παρατηρήσεις μη ομαδοποιημένες

Συχνότητα ή απόλυτη συχνότητα (ν_i)

Συχνότητα ή απόλυτη συχνότητα μιας τιμής ονομάζεται το πλήθος των ατόμων του πληθυσμού (ή του δείγματος) για τα οποία η μεταβλητή παίρνει την τιμή αυτή.

- $\nu_1 + \nu_2 + \ldots + \nu_\kappa = \nu$ $\quad (\kappa \le \nu)$
- $0 \le \nu_i \le \nu$ $\quad i = 1, 2, \ldots, \kappa \quad (\kappa \le \nu)$

Σχετική συχνότητα (f_i)

Σχετική συχνότητα μιας τιμής ονομάζεται ο λόγος της συχνότητας προς το μέγεθος του δείγματος και συμβολίζεται με f_i .

$$f_i = \frac{\nu_i}{\nu} \qquad i = 1, 2, \ldots, \kappa \quad (\kappa \le \nu)$$

- $f_1 + f_2 + \ldots + f_\kappa = 1$ $\quad (\kappa \le \nu)$
- $f_1\,\% + f_2\,\% + \ldots + f_\kappa\,\% = 100\,\%$ $\quad (\kappa \le \nu)$
- $f_i\,\% = f_i \cdot 100$ $\quad (\kappa \le \nu)$
- $0 \le f_i \le 1$ $\quad (\kappa \le \nu)$

Αθροιστική συχνότητα (N_i)

Αθροιστική συχνότητα μιας τιμής ονομάζεται το άθροισμα των συχνοτήτων ν_i των τιμών που είναι μικρότερες ή ίσες με την τιμή αυτή.

- $N_i = \nu_1 + \nu_2 + \ldots + \nu_i$ $\quad i = 1, 2, \ldots$
- $N_1 = \nu_1$ $\quad i = 1, 2, \ldots$
- $N_{i+1} = N_i + \nu_{i+1}$ $\quad i = 1, 2, \ldots$

Αθροιστική σχετική συχνότητα (F_i)

Αθροιστική σχετική συχνότητα μιας τιμής ονομάζεται το άθροισμα των σχετικών συχνοτήτων f_i των τιμών που είναι μικρότερες ή ίσες με την τιμή αυτή.

- $F_i = f_1 + f_2 + \ldots + f_i$ $\quad i = 1, 2, \ldots$
- $F_i\,\% = F_i \cdot 100$ $\quad i = 1, 2, \ldots$
- $F_1 = f_1$ $\quad i = 1, 2, \ldots$
- $F_{i+1} = F_i + f_{i+1}$ $\quad i = 1, 2, \ldots$
- $F_i = \dfrac{N_i}{\nu}$ $\quad i = 1, 2, \ldots$

Παρατηρήσεις ομαδοποιημένες

κλάση

κλειστό ← $[\,\alpha\,,\,\beta\,)$ → ανοιχτό

κάτω άκρο, άνω άκρο

Πλάτος (c) μιας κλάσης [α, β)

ονομάζεται η διαφορά των άκρων της, δηλαδή :

$$c = \beta - \alpha$$

Κέντρο ή κεντρική τιμή (x_i) μιας κλάσης [α, β)

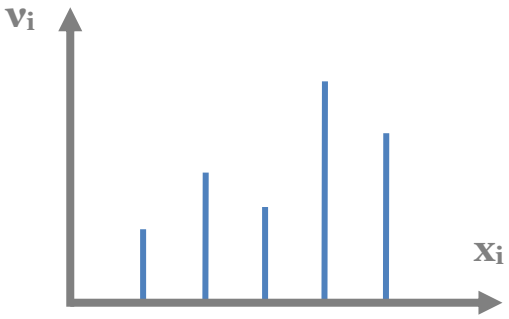
ονομάζεται το ημιάθροισμα των άκρων της, δηλαδή :

$$x_i = \frac{\alpha + \beta}{2}$$

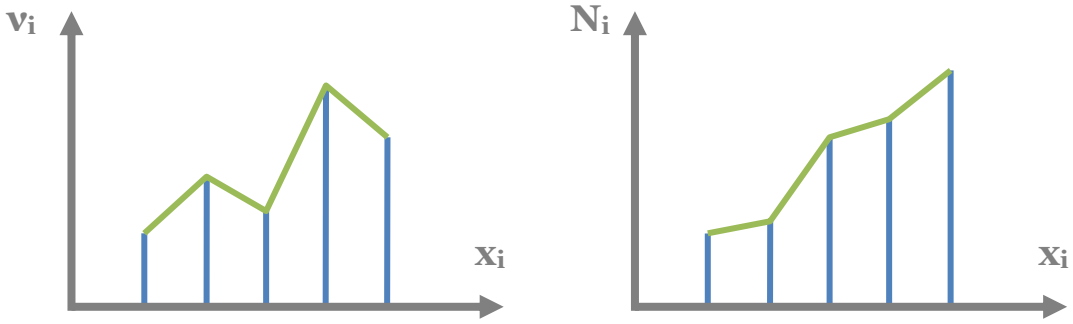
Διαγράμματα

Διάγραμμα συχνοτήτων

- Το διάγραμμα συχνοτήτων χρησιμοποιείται για **ποσοτικές** μεταβλητές.

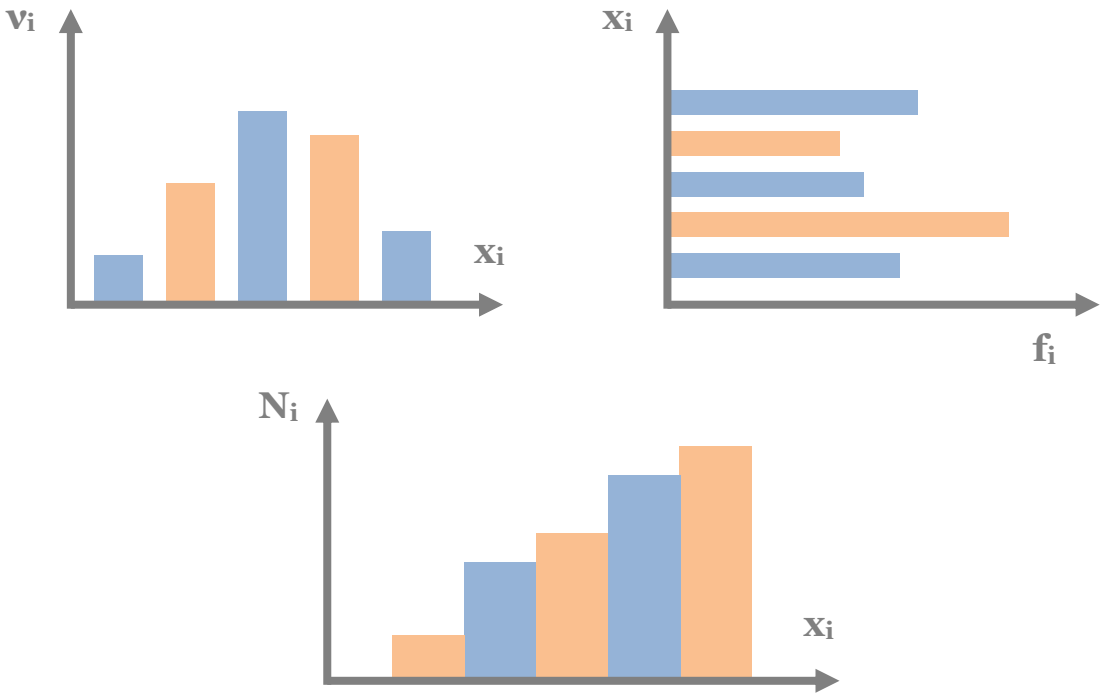


- Μπορούμε να φτιάξουμε διαγράμματα για οποιοδήποτε συχνότητα: απλή, σχετική, αθροιστική, σχετική αθροιστική ή τις αντίστοιχες %.
- Αν ενώσουμε τις κορυφές ενός διαγράμματος συχνοτήτων ή αθροιστικών συχνοτήτων, σχηματίζουμε τα αντίστοιχα **πολύγωνα συχνοτήτων**.



Ραβδόγραμμα

- Το ραβδόγραμμα χρησιμοποιείται για **ποιοτικές** μεταβλητές.
- Το ραβδόγραμμα μπορεί να είναι κάθετο ή οριζόντιο.



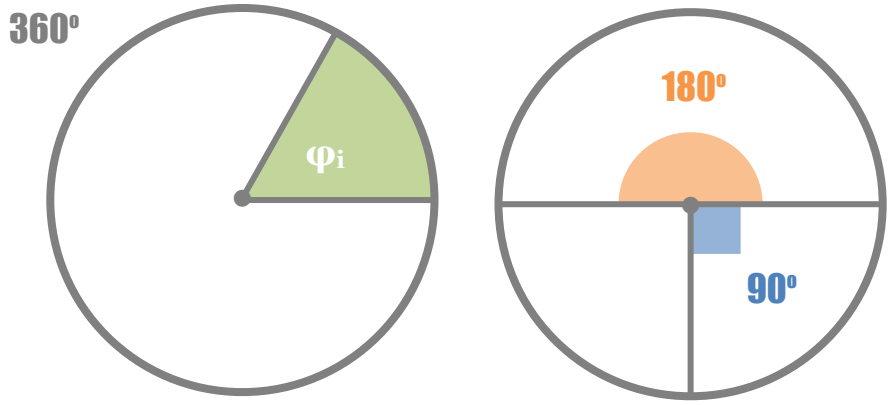
- Ο οριζόντιος άξονας δεν είναι άξονας με την αυστηρή έννοια, αλλά μια σειρά από διαδοχικές θέσεις. Γι' αυτό, η απόσταση ή το πάχος των ράβδων, που σχεδιάζουμε, είναι αυθαίρετο και δεν παίζει κανένα ουσιαστικό ρόλο.
- Τα ραβδογράμματα δεν έχουν πολύγωνο συχνοτήτων.

Κυκλικό διάγραμμα

- Το κυκλικό διάγραμμα χρησιμοποιείται τόσο για ποσοτικές, όσο και για ποιοτικές μεταβλητές.
- Τα τόξα α_i ενός κυκλικού διαγράμματος (ή οι αντίστοιχες επίκεντρες γωνίες) συνδέονται άμεσα με τη σχετική συχνότητα f_i , σύμφωνα με τη σχέση:

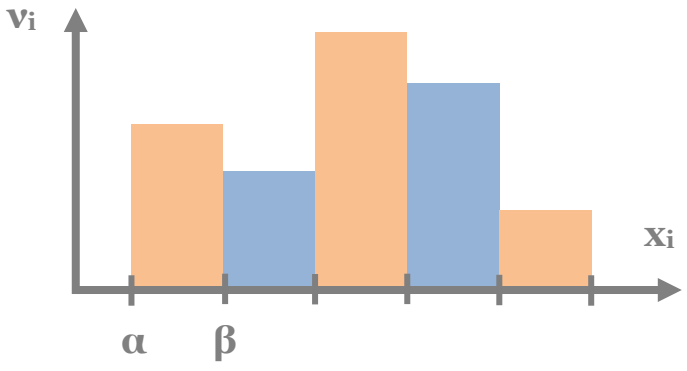
$$f_i = \frac{\alpha_i}{360} \quad \text{ή} \quad \alpha_i = 360 \cdot f_i$$

- Κάθε κύκλος είναι χωρισμένος σε 360° . Κάθε διάμετρος χωρίζει τον κύκλο σε δύο ημικύκλια, καθένα ίσο με 180° . Κάθε ορθή γωνία έχει μέτρο 90° .



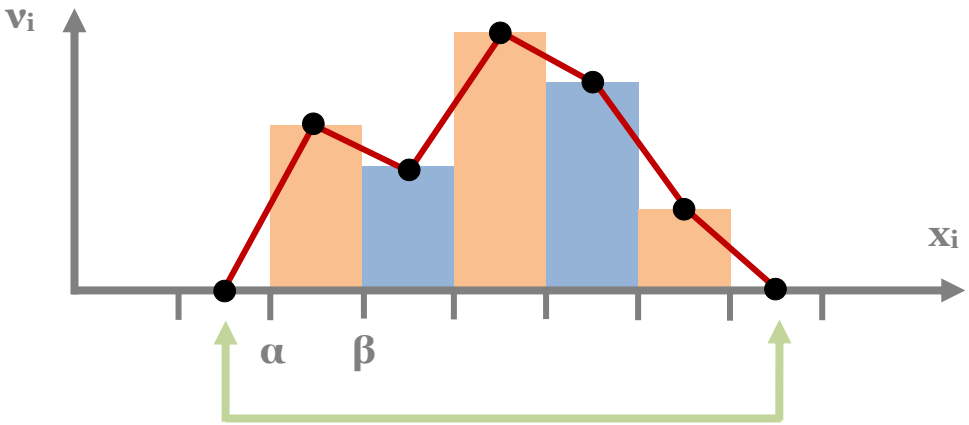
Ιστόγραμμα συχνοτήτων

- Το ιστόγραμμα χρησιμοποιείται για **ομαδοποιημένες** ποσοτικές παρατηρήσεις.



- Μπορούμε να φτιάξουμε ιστόγραμμα για οποιοδήποτε συχνότητα: απλή, σχετική, αθροιστική ή σχετική αθροιστική (ή τις αντίστοιχες %).
- Η κατασκευή του πολύγωνου συχνοτήτων, στην περίπτωση ιστογράμματος, είναι ελαφρά δυσκολότερη από το απλό διάγραμμα.

Πολύγωνο συχνοτήτων

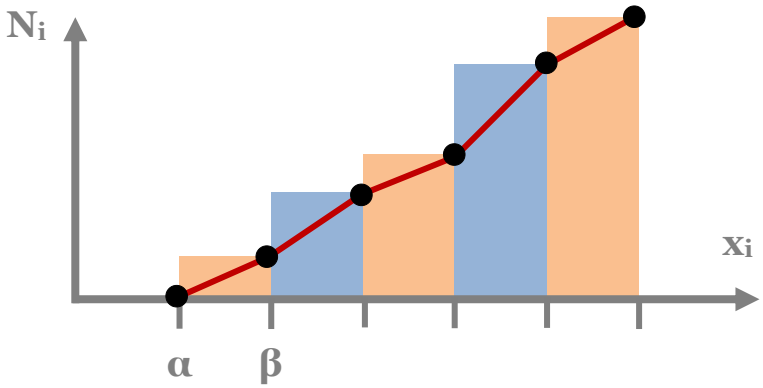


Χρησιμοποιούμε 2 βοηθητικές κλάσεις, επιπλέον.

- Ενώνουμε τα κέντρα των κλάσεων.
- Το εμβαδό που περικλείεται από το πολύγωνο συχνοτήτων και τον άξονα x΄x ισούται με το άθροισμα των εμβαδών όλων των ορθογωνίων.
- Το εμβαδό κάτω από το πολύγωνο συχνοτήτων (ν_i) ισούται με ν , ενώ το εμβαδό κάτω από το πολύγωνο σχετικών συχνοτήτων (f_i) ισούται με 1 . (*)

(*) Με την προϋπόθεση ότι χρησιμοποιήσουμε ως μονάδα μέτρησης το πλάτος c .

Πολύγωνο αθροιστικών συχνοτήτων



- Ενώνουμε διαγωνίως, τις δεξιές κορυφές των ορθογωνίων.

Μέτρα θέσης

Μέτρα θέσης μιας μεταβλητής ονομάζουμε τα παρακάτω μεγέθη :

- **Επικρατούσα τιμή (ή κορυφή)**
- **Διάμεσος**
- **Μέση τιμή (ή αριθμητικός μέσος)**
- **Σταθμικός μέσος**

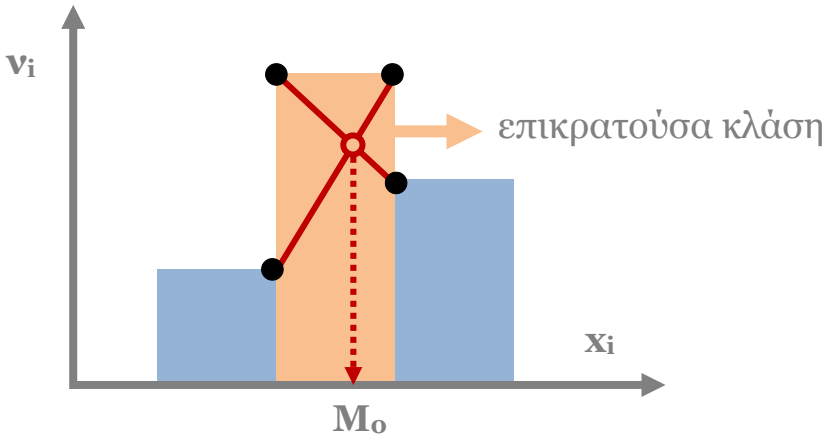
Επικρατούσα τιμή (M_o)

Επικρατούσα τιμή μια μεταβλητής ονομάζεται η τιμή της μεταβλητής με τη μεγαλύτερη συχνότητα.

- Είναι δυνατόν να υπάρχουν περισσότερες από μία επικρατούσες τιμές, στην περίπτωση που δύο ή περισσότερες τιμές έχουν τη μέγιστη συχνότητα. Τότε λέμε πως έχουμε **δικόρυφη** ή **πολυκόρυφη** κατανομή συχνοτήτων.
- Είναι δυνατόν να μην υπάρχει επικρατούσα τιμή, αν όλες οι παρατηρήσεις είναι διαφορετικές

Ομαδοποιημένες παρατηρήσεις

- **Επικρατούσα κλάση** ονομάζεται η κλάση με τη μεγαλύτερη συχνότητα.
- Στην περίπτωση ομαδοποιημένων παρατηρήσεων, η εύρεση της επικρατούσας τιμής γίνεται γραφικα, από το ιστόγραμμα συχνοτήτων :



Διάμεσος (δ)

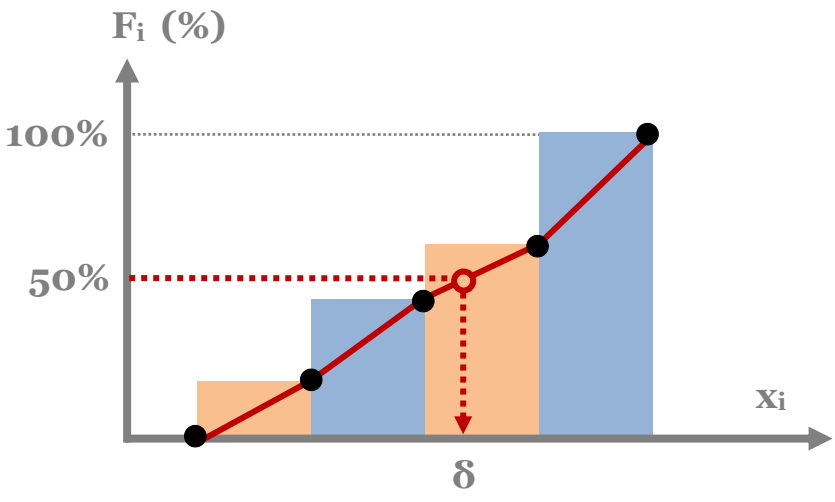
Διάμεσος ενός δείγματος ν παρατηρήσεων, οι οποίες έχουν διαταχθεί σε αύξουσα σειρά ονομάζεται :

- Η **μεσαία** παρατήρηση, αν το πλήθος ν είναι **περιττό**.
- Το **ημιάθροισμα** των δύο μεσαίων παρατηρήσεων, αν το πλήθος ν είναι **άρτιο**.

- Η διάμεσος και η θέση στην οποία την αναζητούμε είναι δύο διαφορετικά πράγματα.

Ομαδοποιημένες παρατηρήσεις

- Στην περίπτωση ομαδοποιημένων παρατηρήσεων, η εύρεση της διαμέσου γίνεται γραφικα, από το πολύγωνο, οποιουδήποτε αθροιστικού ιστογράμματος (N_i ή F_i) :



Μέση τιμή ($\bar{x}$)

Μέση τιμή ενός δείγματος μεγέθους ν ονομάζεται το πηλίκο του αθροίσματος των ν παρατηρήσεων, προς το πλήθος τους ν .

$$\bar{x} = \frac{\sum_{i=1}^{\nu} x_i}{\nu} = \frac{x_1 + x_2 + \ldots + x_\nu}{\nu}$$

- Προκειμένου για παρατηρήσεις ταξινομημένες σε κατανομή συχνοτήτων :

$$\bar{x} = \frac{\sum_{i=1}^{\kappa} x_i \cdot \nu_i}{\nu} = \frac{\nu_1 x_1 + \nu_2 x_2 + \ldots + \nu_\kappa x_\kappa}{\nu} \qquad (\kappa \le \nu)$$

- Ισχύει ακόμα :

$$\bar{x} = \sum_{i=1}^{\kappa} x_i \cdot f_i = x_1 f_1 + x_2 f_2 + \ldots + x_\kappa f_\kappa \qquad (\kappa \le \nu)$$

- Στην περίπτωση που κάθε τιμή x_i συμμετέχει στα δεδομένα, με διαφορετική **βαρύτητα /συντελεστή βαρύτητας (w_i)** , τότε αντί του αριθμητικού μέσου χρησιμοποιούμε τον τύπο του σταθμικού μέσου :

$$\bar{x} = \frac{\sum_{i=1}^{\kappa} x_i \cdot w_i}{\sum_{i=1}^{\kappa} w_i} = \frac{x_1 w_1 + x_2 w_2 + \ldots + x_\kappa w_\kappa}{w_1 + w_2 + \ldots + w_\kappa} \qquad (\kappa \le \nu)$$

Σύγκριση μέτρων θέσης

Μέση Τιμή

- ▲ Εξαρτάται απ' όλες τις τιμές.
- ▲ Εργαζόμαστε ευκολότερα θεωρητικά ή αλγεβρικά.
- ▼ Επηρεάζεται από ακραίες τιμές.

Διάμεσος

- ▲ Δεν επηρεάζεται από ακραίες τιμές.
- ▲ Εξαρτάται από το πλήθος όλων των τιμών.
- ▼ Ο υπολογισμός της παρουσιάζει δυσκολίες σε ορισμένες περιπτώσεις , όπως π.χ. σε συνεχή μεταβλητή.

Επικρατούσα Τιμή

- ▲ Δεν επηρεάζεται από ακραίες τιμές.
- ▲ Χρήσιμη σε ποιοτικά δεδομένα, όπου μέση τιμή και διάμεσος δεν έχουν νόημα.
- ▼ Εξαρτάται μόνο από τη μεγαλύτερη τιμή και αγνοεί τις υπόλοιπες.

Μέτρα διασποράς

Εύρος (R)

Εύρος των τιμών μιας μεταβλητής ονομάζεται η διαφορά της μικρότερης τιμής, από τη μεγαλύτερη :

$$R = x_{max} - x_{min}$$

Διακύμανση ή διασπορά (s^2)

Διακύμανση μιας μεταβλητής X , η οποία παίρνει ν το πλήθος τιμές x_i , i = 1, 2, 3, ... ν , με μέση τιμή $\bar{x}$ ονομάζεται το πηλίκο :

$$s^2 = \frac{\sum_{i=1}^{\nu} (\bar{x} - x_i)^2}{\nu} = \frac{(\bar{x} - x_1)^2 + (\bar{x} - x_2)^2 + \ldots + (\bar{x} - x_\nu)^2}{\nu}$$

- Αν οι τιμές της μεταβλητής X είναι ταξινομημένες σε πίνακα συχνοτήτων, με κ διαφορετικές τιμές, τότε:

$$s^2 = \frac{\sum_{i=1}^{\kappa} \nu_i (\bar{x} - x_i)^2}{\nu} = \frac{\nu_1(\bar{x} - x_1)^2 + \nu_2(\bar{x} - x_2)^2 + \ldots + \nu_\kappa(\bar{x} - x_\kappa)^2}{\nu}$$

- Αν η μέση τιμή $\bar{x}$ δεν είναι ακέραιος αριθμός, τότε προκειμένου να διευκολύνουμε τους υπολογισμούς, χρησιμοποιούμε τους εναλλακτικούς τύπους :

$$s^2 = \frac{1}{\nu}\left\{\sum_{i=1}^{\nu} x_i^2 - \frac{\left(\sum_{i=1}^{\nu} x_i\right)^2}{\nu}\right\} \quad \text{ή} \quad s^2 = \frac{1}{\nu}\left\{\sum_{i=1}^{\kappa} x_i^2 \nu_i - \frac{\left(\sum_{i=1}^{\kappa} x_i \nu_i\right)^2}{\nu}\right\}$$

Τυπική απόκλιση (s)

Τυπική απόκλιση **s** μιας μεταβλητής X, που παίρνει ν το πλήθος τιμές x_i, $i = 1, 2, \dots \nu$ με μέση τιμή $\bar{x}$ ονομάζεται η τετραγωνική ρίζα της διακύμανσης, δηλαδή:

$$\mathbf{s = \sqrt{s^2}}$$

Μεταβολές μέσης τιμής & τυπικής απόκλισης

Πρόσθεση σταθεράς $c \in \mathbb{R}$

Αν σε κάθε παρατήρηση x_i , ενός δείγματος $x_1, x_2, \dots, x_\nu$ με μέση τιμή $\bar{x}$ και τυπική απόκλιση s_x , **προσθέσουμε** ένα σταθερό αριθμό $c \in \mathbb{R}$, τότε για τη μέση τιμή $\bar{y}$ και την τυπική απόκλιση s_y των παρατηρήσεων $y_1, y_2, \dots, y_\nu$, που προκύπτουν, ισχύει ότι :

- $\mathbf{\bar{y} = \bar{x} + c}$
- $\mathbf{s_y = s_x}$

Πολλαπλασιασμός με σταθερά $c \in \mathbb{R}$

Αν κάθε παρατήρηση x_i , ενός δείγματος $x_1, x_2, \dots, x_\nu$ με μέση τιμή $\bar{x}$ και τυπική απόκλιση s_x , **πολλαπλασιαστεί** με ένα σταθερό αριθμό $c \in \mathbb{R}$, τότε για τη μέση τιμή $\bar{y}$ και την τυπική απόκλιση s_y των παρατηρήσεων $y_1, y_2, \dots, y_\nu$, που προκύπτουν, ισχύει ότι :

- $\mathbf{\bar{y} = c \cdot \bar{x}}$
- $\mathbf{s_y = |c| \cdot s_x}$

Μέτρα μεταβολής

Συντελεστής μεταβολής ή μεταβλητότητας (CV)

Συντελεστής μεταβλητότητας μιας ποσοτικής μεταβλητής X , η οποία παρουσιάζει μέση τιμή $\bar{x}$ και τυπική απόκλιση s , ονομάζεται το πηλίκο :

$$CV = \frac{s}{\bar{x}} \quad \text{ή} \quad CV = \frac{s}{\bar{x}} \cdot 100\%$$

- Αν **CV ≤ 10%** τότε ο πληθυσμός (ή το δείγμα) ονομάζεται **ομοιογενής** ή **ομογενής** . Αν **CV > 10%** τότε ο πληθυσμός (ή το δείγμα) είναι **ανομοιογενής** .
- Αν $\bar{x} < 0$, τότε χρησιμοποιούμε την $|\bar{x}|$. Δηλαδή :

$$CV = \frac{s}{|\bar{x}|} \quad \text{ή} \quad CV = \frac{s}{|\bar{x}|} \cdot 100\%$$

- Ο συντελεστής μεταβολής είναι ένα μέτρο **σχετικής διασποράς**.
- Ο συντελεστής μεταβολής είναι ανεξάρτητος απ' τις μονάδες μέτρησης.

Κανονική κατανομή

- Σε μια κανονική (ή περίπου κανονική) κατανομή, το εύρος ισούται με περίπου 6 τυπικές αποκλίσεις:

$$\mathbf{R \cong 6 \cdot s}$$

- Σε μια κανονική (ή περίπου κανονική) κατανομή, η μέση τιμή και η διάμεσος ταυτίζονται :

$$\mathbf{\bar{x} = \delta}$$

- κανονική (ή περίπου κανονική) κατανομή :
  - το **68 %** των παρατηρήσεων βρίσκεται στο διάστημα :

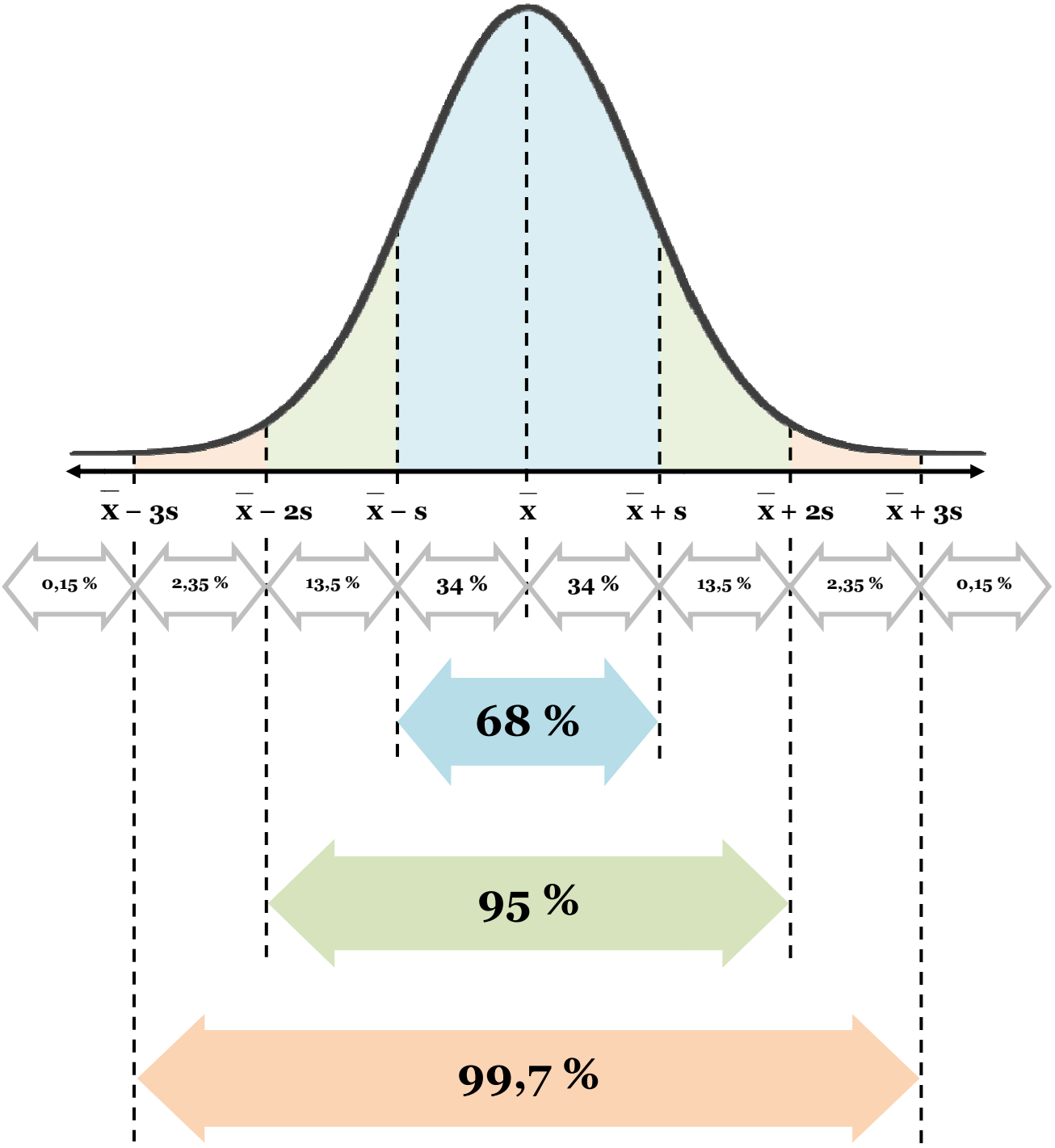
$$\mathbf{\left(\bar{x} - s , \bar{x} + s \right)}$$

  - το **95 %** των παρατηρήσεων βρίσκεται στο διάστημα :

$$\mathbf{\left(\bar{x} - 2s , \bar{x} + 2s \right)}$$

  - το **99,7 %** των παρατηρήσεων βρίσκεται στο διάστημα :

$$\mathbf{\left(\bar{x} - 3s , \bar{x} + 3s \right)}$$


$\bar{x} - 3s$
$\bar{x} - 2s$
$\bar{x} - s$
$\bar{x}$
$\bar{x} + s$
$\bar{x} + 2s$
$\bar{x} + 3s$
0,15 %
2,35 %
13,5 %
34 %
34 %
13,5 %
2,35 %
0,15 %
68 %
95 %
99,7 %