ΠΕΡΙΓΡΑΦΙΚΗ ΣΤΑΤΙΣΤΙΚΗ

Περιγραφική στατιστική είναι ο κλάδος της στατιστικής που ασχολείται με την οργάνωση και τη συνοπτική παρουσίαση των δεδομένων, την παράστασή τους με γραφήματα και τον υπολογισμό περιγραφικών μέτρων.

Πληθυσμός είναι το σύνολο των τιμών μιας μεταβλητής. Ο πληθυσμός αποτελεί το άγνωστο μέρος της Στατιστικής. Σκοπός της Στατιστικής είναι η εξαγωγή συμπερασμάτων για τον πληθυσμό βάσει του δείγματος.

Το **δείγμα** ορίζεται ως ένα υποσύνολο του πληθυσμού. **Τυχαίο δείγμα** είναι το δείγμα που εκλέγεται κατά τέτοιο τρόπο ώστε όλα τα μέλη του πληθυσμού να έχουν ίση πιθανότητα να συμπεριληφθούν στο δείγμα.

Όπως ειπώθηκε και στη Θεωρία Πιθανοτήτων οι τυχαίες μεταβλητές συμβολίζονται με τα κεφαλαία γράμματα X, Y, Z,…., ενώ οι τιμές που παίρνουν με τα μικρά γράμματα $x_1, x_2, \dots, x_\nu$ ή $y_1, y_2, \dots, y_\nu$ ή $z_1, z_2, \dots, z_k$. Διακρίνονται σε **ποιοτικές** όταν εκφράζουν ποιοτικά χαρακτηριστικά ενός πληθυσμού (π.χ. το χρώμα των ανθέων ενός φυτού, το φύλο, η εθνικότητα, το επάγγελμα ενός ατόμου, κ.λ.π.) και **ποσοτικές** όταν μπορούν να μετρηθούν (π.χ. το ύψος ενός φυτού, ο αριθμός των σπόρων, το βάρος ενός ζώου, κ.λ.π.). Επίσης μια ποσοτική μεταβλητή μπορεί να είναι **διακριτή,** εάν παίρνει μεμονωμένες/διακριτές τιμές (π.χ. 1, 2, 3,…) και το σύνολο των τιμών της μπορεί να είναι πεπερασμένο ή απείρως αριθμήσιμο ή **συνεχής** εάν παίρνει τιμές σε ένα υποσύνολο των πραγματικών αριθμών, δηλαδή σε ένα διάστημα (α, β) με $-\infty \leq \alpha < \beta \leq \infty$. Το πλήθος των βακτηριδίων στη δειγματοληπτική πλάκα, ο αριθμός των γεννήσεων ή των θανάτων που συμβαίνουν σε μια κτηνοτροφική μονάδα, ο αριθμός των ημερών βροχής σε έναν μήνα ή σε ένα έτος σε μια συγκεκριμένη περιοχή είναι διακριτές ποσοτικές μεταβλητές, ενώ το ύψος, το βάρος, η θερμοκρασία είναι συνεχείς ποσοτικές μεταβλητές.

Έστω $x_1, x_2, \dots, x_\nu$ οι παρατηρήσεις ενός δείγματος και $y_1, y_2, \dots, y_k$ $(k \leq \nu)$ οι διαφορετικές τιμές των παρατηρήσεων που εμφανίστηκαν στο δείγμα. Στη συνέχεια για κάθε $y_i, i = 1,2, \dots, k$ θα συμβολίζουμε με $\boldsymbol{\nu_i}$ τη **συχνότητά** της (*πόσες φορές εμφανίστηκε*), με $\boldsymbol{f_i}$ τη **σχετική συχνότητά** της, $f_i = \nu_i/\nu$, με $\boldsymbol{N_i}$ την **αθροιστική συχνότητά** της (*το άθροισμα των συχνοτήτων των τιμών που είναι* $\leq y_i$) και με $\boldsymbol{F_i}$ την **αθροιστική σχετική συχνότητά** της (*το άθροισμα των σχετικών συχνοτήτων των τιμών που είναι* $\leq y_i$).

Αριθμητικά περιγραφικά μέτρα

Α) ***Μέτρα κεντρικής τάσης ή θέσης***

i) Μέσος όρος ή (δειγματική) μέση τιμή ή μέση τιμή του δείγματος $\bar{x}$

$$\bar{x} = \frac{1}{\nu}\sum_{i=1}^{\nu} x_i = \frac{1}{\nu}\sum_{i=1}^{\kappa} \nu_i y_i$$

Όταν τα δεδομένα είναι *ομαδοποιημένα σε k κλάσεις*, τα y_i είναι οι κεντρικές τιμές των κλάσεων.

Ο υπολογισμός του μέσου όρου είναι απλός, χρησιμοποιούνται όλες οι τιμές του δείγματος για τον υπολογισμό του και επίσης αξιοποιείται στην στατιστική συμπερασματολογία. Τα μειονεκτήματά του είναι ότι *επηρεάζεται από ακραίες τιμές*, ενδέχεται να μην αντιστοιχεί σε δυνατή τιμή της μεταβλητής και δεν υπολογίζεται για ποιοτικά δεδομένα.

ii) Διάμεσος δ

Για να προσδιορίσουμε τη διάμεσο παρατάσσουμε το δείγμα σε αύξουσα διάταξη.

Εάν το δείγμα είναι περιττού πλήθους, η διάμεσος είναι η μεσαία παρατήρηση, ενώ εάν το δείγμα είναι αρτίου πλήθους η διάμεσος ορίζεται ως το ημιάθροισμα των δύο μεσαίων παρατηρήσεων.

$$\delta = \begin{cases} x_{\left(\frac{\nu+1}{2}\right)} & \text{εάν το μέγεθος του δείγματος ν είναι περιττού πλήθους} \\ \dfrac{x_{\left(\frac{\nu}{2}\right)} + x_{\left(\frac{\nu}{2}+1\right)}}{2} & \text{εάν το μέγεθος του δείγματος ν είναι αρτίου πλήθους} \end{cases}$$

Εάν τα δεδομένα είναι *ομαδοποιημένα σε κλάσεις*, για τον υπολογισμό της διαμέσου προσδιορίζουμε καταρχάς την κλάση μέσα στην οποία βρίσκεται η διάμεσος. Στη συνέχεια εφαρμόζοντας τον παρακάτω τύπο υπολογίζουμε την τιμή της:

$$\delta = L_i + (0.5\nu - N_{i-1})\frac{c}{\nu_i}$$

όπου:

L_i: το κάτω άκρο της κλάσης μέσα στην οποία βρίσκεται η διάμεσος

ν: το μέγεθος του δείγματος

N_{i-1}: η αθροιστική συχνότητα της προηγούμενης κλάσης από αυτήν που βρίσκεται η διάμεσος

c: το εύρος των κλάσεων

ν_i: η συχνότητα της κλάσης μέσα στην οποία βρίσκεται η διάμεσος

Ο υπολογισμός της διαμέσου είναι απλός, *δεν επηρεάζεται από ακραίες τιμές* και η τιμή της είναι μοναδική. Δεν χρησιμοποιούνται όλες οι τιμές του δείγματος για τον υπολογισμό της και δεν υπολογίζεται για ποιοτικά δεδομένα.

iii) Επικρατούσα τιμή ή κορυφή M_0

Είναι η τιμή με την μεγαλύτερη συχνότητα.

Εάν τα δεδομένα είναι *ομαδοποιημένα σε κλάσεις* για τον υπολογισμό της επικρατούσας τιμής προσδιορίζουμε καταρχάς την *επικρατούσα κλάση*, την κλάση δηλαδή με τη μεγαλύτερη συχνότητα και στη συνέχεια υπολογίζουμε την επικρατούσα τιμή από τον τύπο:

$$M_0 = L_i + \frac{\Delta_1}{\Delta_1 + \Delta_2} \cdot c$$

όπου:

L_i: το κάτω άκρο της επικρατούσας κλάσης

$\Delta_1 = \nu_i - \nu_{i-1}$, όπου ν_i η συχνότητα της επικρατούσας κλάσης και ν_{i-1} η συχνότητα της προηγούμενης κλάσης

$\Delta_2 = \nu_i - \nu_{i+1}$, όπου ν_i η συχνότητα της επικρατούσας κλάσης και ν_{i+1} η συχνότητα της επόμενης κλάσης

c: το εύρος των κλάσεων

Ο υπολογισμός της επικρατούσας τιμής είναι απλός, *δεν επηρεάζεται από ακραίες τιμές* και *υπολογίζεται για ποιοτικά δεδομένα.* Δεν χρησιμοποιούνται όλες οι τιμές του δείγματος για τον υπολογισμό της, *δεν είναι μοναδική και επίσης μπορεί να μην υπάρχει*. Επιπλέον η σημασία της στην στατιστική συμπερασματολογία είναι περιορισμένη.

Β) Μέτρα μεταβλητότητας

***i)* Δειγματική διασπορά ή δειγματική διακύμανση s^2**

$$s^2 = \frac{1}{\nu - 1}\sum_{i=1}^{\nu}(x_i - \bar{x})^2 = \frac{1}{\nu - 1}\left(\sum_{i=1}^{\nu} x_i^2 - \nu\bar{x}^2\right) = \frac{1}{\nu - 1}\sum_{i=1}^{k}(y_i - \bar{x})^2\nu_i = \frac{1}{\nu - 1}\left(\sum_{i=1}^{k}\nu_i y_i^2 - \nu\bar{x}^2\right)$$

Ένα από τα μειονεκτήματα της δειγματικής διακύμανσης είναι ότι *δεν εκφράζεται στην ίδια μονάδα μέτρησης με τα δεδομένα*. Γι' αυτό συνήθως ως μέτρο μεταβλητότητας χρησιμοποιούμε την τυπική απόκλιση.

ii) Δειγματική τυπική απόκλιση $\boldsymbol{s = \sqrt{s^2}}$

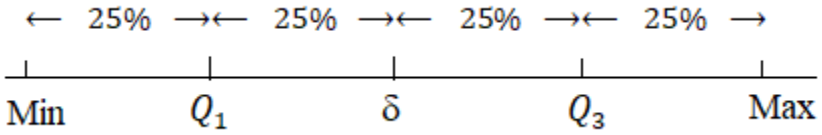
Για τον υπολογισμό της δειγματικής τυπικής απόκλισης χρησιμοποιούνται όλες οι τιμές του δείγματος, εκφράζεται στην ίδια μονάδα μέτρησης με τα δεδομένα και έχει μεγάλη σημασία της στην στατιστική συμπερασματολογία.

iii) Συντελεστής μεταβλητότητας CV

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

Εάν ο συντελεστής μεταβλητότητας CV < 10% το δείγμα θεωρείται *ομοιογενές*, δηλαδή έχει μικρή μεταβλητότητα. Επίσης ο CV μπορεί να χρησιμοποιηθεί ως *μέτρο σύγκρισης της μεταβλητότητας* μεταξύ δύο ή περισσοτέρων δειγμάτων με διαφορετικούς μέσους όρους.

iv) Ενδοτεταρτημοριακό εύρος $\boldsymbol{Q = Q_3 - Q_1}$



Για τον υπολογισμό του ενδοτεταρτημοριακού εύρους απαιτείται ο υπολογισμός των Q_1 και Q_3, δηλαδή των 25 και 75–ποσοστιαίων σημείων. Το *25–ποσοστιαίο σημείο* (Q_1) είναι το σημείο εκείνο, για το οποίο ισχύει ότι το πολύ 25% των τιμών του δείγματος είναι μικρότερες από αυτό και το πολύ 75% των τιμών του δείγματος είναι μεγαλύτερες από αυτό. Αντίστοιχα το *75–ποσοστιαίο σημείο* (Q_3) είναι το σημείο εκείνο, για το οποίο ισχύει ότι το πολύ 75% των τιμών του δείγματος είναι μικρότερες από αυτό και το πολύ 25% των τιμών του δείγματος είναι μεγαλύτερες από αυτό. Τα Q_1, $Q_2 = \delta$ και Q_3 λέγονται και τεταρτημόρια ή τεταρτοτόμοι, καθώς τέμνουν την κατανομή των δεδομένων σε τέσσερα μέρη. Εντός του ενδοτεταρτημοριακού εύρους βρίσκονται τα μισά δεδομένα του δείγματος, που είναι πιο κοντά στην κεντρική τιμή (διάμεσο), δηλαδή βρίσκεται το 50% των *μεσαίων* παρατηρήσεων.

Για να προσδιορίσουμε το *1ο τεταρτημόριο* (Q_1) και το *3ο τεταρτημόριο* (Q_3) παρατάσσουμε το δείγμα σε αύξουσα διάταξη και στη συνέχεια δουλεύουμε όπως στη διάμεσο.

Για τον προσδιορισμό των Q_1 και Q_3 σε ομαδοποιημένα δεδομένα προσδιορίζουμε καταρχάς τις κλάσεις μέσα στις οποίες βρίσκονται τα Q_1 και Q_3 και στη συνέχεια με τους παρακάτω τύπους υπολογίζουμε τις τιμές τους:

$$Q_1 = L_i + (0.25\nu - N_{i-1})\frac{c}{\nu_i} \qquad\qquad Q_3 = L_i + (0.75\nu - N_{i-1})\frac{c}{\nu_i}$$

όπου:

L_i: το κάτω άκρο της κλάσης μέσα στην οποία βρίσκεται το Q_1 ή το Q_3 αντίστοιχα

ν: το μέγεθος του δείγματος

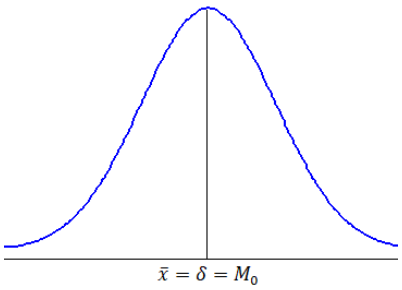
N_{i-1}: η αθροιστική συχνότητα της προηγούμενης κλάσης από αυτήν που βρίσκεται το Q_1 ή το Q_3

c: το εύρος των κλάσεων

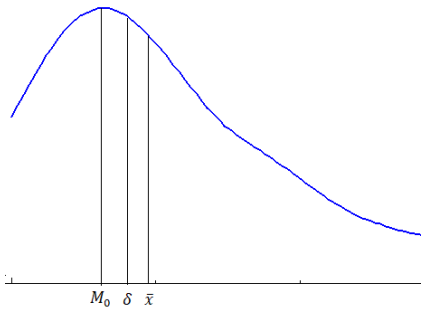
ν_i: η συχνότητα της κλάσης μέσα στην οποία βρίσκεται το Q_1 ή το Q_3αντίστοιχα

Συμμετρική ή λοξή κατανομή δεδομένων

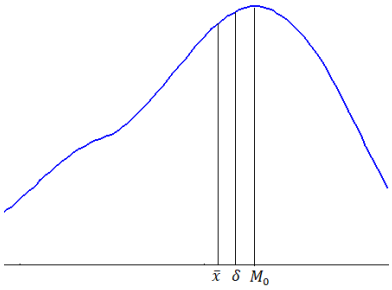
- Όταν σε κάποιο δείγμα $\bar{x} = \delta = M_0$, η καμπύλη συχνοτήτων της κατανομής του δείγματος είναι *συμμετρική*.
- Όταν σε κάποιο δείγμα $\bar{x} > \delta > M_0$, η καμπύλη συχνοτήτων της κατανομής του δείγματος παρουσιάζει *θετική ασυμμετρία*.
- Όταν σε κάποιο δείγμα $\bar{x} < \delta < M_0$, η καμπύλη συχνοτήτων της κατανομής του δείγματος παρουσιάζει *αρνητική ασυμμετρία*.



α) Συμμετρική κατανομή



β) Λοξή κατανομή με θετική ασυμμετρία

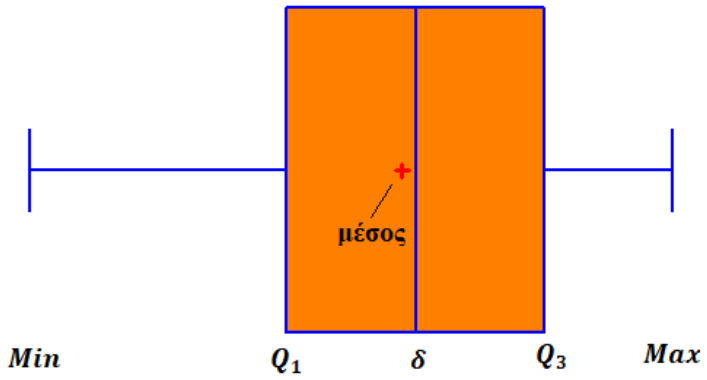


γ) Λοξή κατανομή με αρνητική ασυμμετρία

Θηκόγραμμα

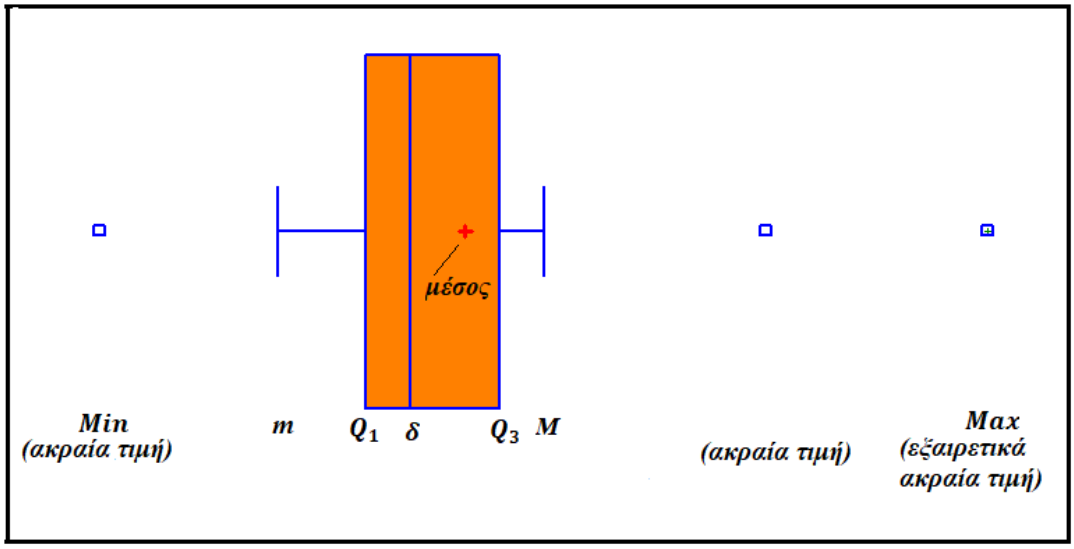
Στη συνέχεια δίνονται οι δύο τύποι του θηκογράμματος:

Α)



Στο θηκόγραμμα αναπαρίστανται *η ελάχιστη τιμή* του δείγματος (*Min*), το *1ο τεταρτημόριο* (Q_1), η *διάμεσος δ* (μπλε κάθετη γραμμή), το *3ο τεταρτημόριο* (Q_3), η *μέγιστη τιμή* του δείγματος *Max*, καθώς και ο *μέσος όρος* (+).

Β)



Στη 2η μορφή του θηκογράμματος καταγράφονται και οι *ακραίες* τιμές και / ή οι *εξαιρετικά ακραίες* τιμές του δείγματος. Μια τιμή του δείγματος θεωρείται **ακραία τιμή** εάν είναι μικρότερη από $Q_1 - 1.5Q = Q_1 - 1.5(Q_3 - Q_1)$ ή εάν είναι μεγαλύτερη από $Q_3 + 1.5Q = Q_3 + 1.5(Q_3 - Q_1)$. Επίσης μια τιμή του δείγματος θεωρείται **εξαιρετικά ακραία τιμή** εάν είναι μικρότερη από $Q_1 - 3Q = Q_1 - 3(Q_3 - Q_1)$ ή εάν είναι μεγαλύτερη από $Q_3 + 3Q = Q_3 + 3(Q_3 - Q_1)$. Το αριστερό άκρο ***m*** είναι η μικρότερη τιμή του

δείγματος, που είναι μεγαλύτερη ή ίση με $Q_1 - 1.5Q = Q_1 - 1.5(Q_3 - Q_1)$, ενώ το δεξιό άκρο ***M*** είναι η μεγαλύτερη τιμή του δείγματος, που είναι μικρότερη ή ίση από $Q_3 + 1.5Q = Q_3 + 1.5(Q_3 - Q_1)$.

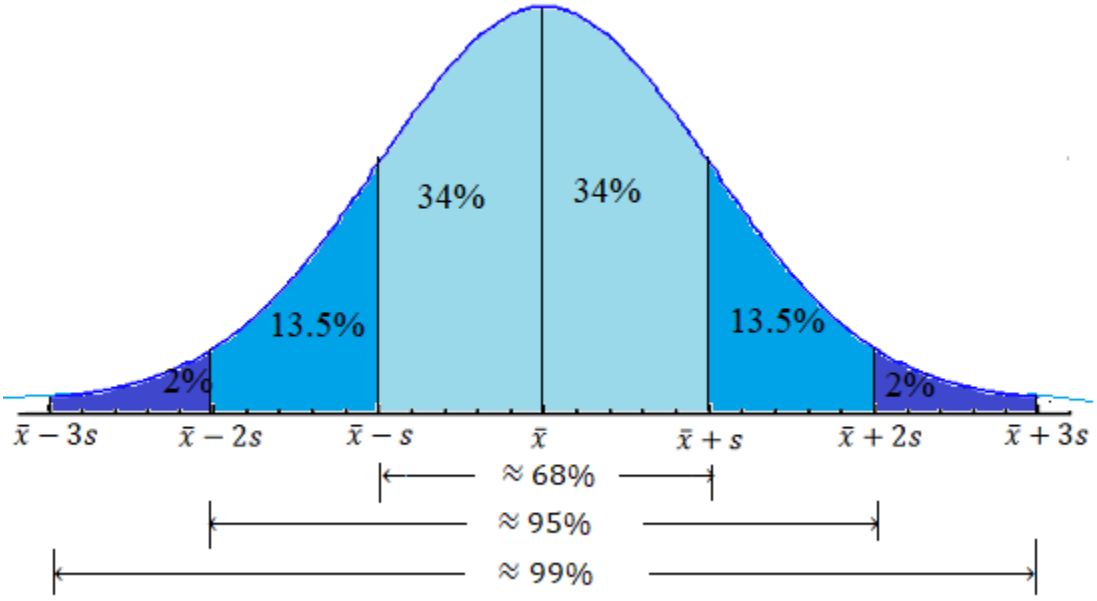
Εμπειρικός κανόνας

Αν η κατανομή του δείγματος έχει κωδωνοειδή μορφή, δηλαδή προσομοιάζει με μια κανονική κατανομή:

i) στο διάστημα $(\bar{x} - s,\ \bar{x} + s)$ βρίσκεται περίπου το 68% των παρατηρήσεων

ii) στο διάστημα $(\bar{x} - 2s,\ \bar{x} + 2s)$ βρίσκεται περίπου το 95% των παρατηρήσεων

iii) στο διάστημα $(\bar{x} - 3s,\ \bar{x} + 3s)$ βρίσκεται περίπου το 99% των παρατηρήσεων



Μέτρα κεντρικής τάσης και μεταβλητότητας γραμμικού μετασχηματισμού των δεδομένων

Έστω $x_1, x_2, \dots, x_v$ οι τιμές ενός δείγματος, οι οποίες μετασχηματίζονται σε $y_1, y_2, \dots, y_v$ σύμφωνα με τον γραμμικό μετασχηματισμό $y_i = \alpha x_i + \beta,\ \ i = 1,2, \dots, v$. Τότε τα αριθμητικά περιγραφικά μέτρα θέσης και μεταβλητότητας μετασχηματίζονται ως εξής:

- $\bar{y} = \alpha\bar{x} + \beta$
- $\delta_y = \alpha\delta_x + \beta$
- $M_{0y} = \alpha M_{0x} + \beta$
- $s_y^2 = \alpha^2 s_x^2$
- $s_y = |\alpha| s_x$

Ειδικότερα εάν ο γραμμικός μετασχηματισμός είναι της μορφής:

$$z_i = \frac{x_i - \bar{x}}{s} = \frac{1}{s}x_i - \frac{\bar{x}}{s}$$

όπου $\bar{x}$ και s η μέση τιμή και η τυπική απόκλιση των αρχικών μετρήσεων $x_1, x_2, \dots, x_\nu$

με $\alpha = \frac{1}{s}$ και $\beta = -\frac{\bar{x}}{s}$, τότε η μέση τιμή $\bar{z}$ και η τυπική απόκλιση s_z των μετασχηματισμένων δεδομένων είναι:

$$\bar{z} = \alpha\bar{x} + \beta = \frac{1}{s}\bar{x} - \frac{\bar{x}}{s} = 0 \qquad \text{και} \qquad s_z = |\alpha| s_x = \left|\frac{1}{s}\right| s_x = 1\,, \qquad \text{καθώς} \quad s = s_x$$

Επομένως εάν έχουμε οποιαδήποτε δεδομένα $x_1, x_2, \dots, x_\nu$ και τα μετασχηματίσουμε ως εξής: $z_i = \frac{x_i - \bar{x}}{s}$

τότε για τη μέση τιμή και την τυπική απόκλιση των μετασχηματισμένων δεδομένων ισχύει: $\bar{z} = 0$ και $s_z = 1$.

Λύσεις ασκήσεων από το **φυλλάδιο 1 – Ασκήσεις περιγραφικής στατιστικής**

1. Μετρήθηκε η ποσότητα νατρίου που περιέχεται στο κασέρι συνήθους τύπου που παράγει μια γνωστή γαλακτοβιομηχανία. Τα αποτελέσματα εννέα σχετικών μετρήσεων που πήρε ένας φοιτητής του Γ.Π.Α. σε κασέρι που επέλεξε τυχαία από εννέα παρτίδες παραγωγής της γαλακτοβιομηχανίας ήταν (σε milligrams/100gr): 340 300 340 320 320 290 330 320 310. **α)** Να υπολογίσετε και να ερμηνεύσετε τα μέτρα *κεντρικής τάσης* και *μεταβλητότητας* της κατανομής του δείγματος, **β)** Να κατασκευάσετε το *θηκόγραμμα* της κατανομής του δείγματος.

Αριθμητικά περιγραφικά μέτρα

Α) ***Μέτρα κεντρικής τάσης ή θέσης***

i) Μέσος όρος ή μέση τιμή $\bar{x}$

$$\bar{x} = \frac{1}{\nu}\sum_{i=1}^{\nu} x_i = \frac{x_1 + x_2 + \cdots + x_\nu}{\nu} = \frac{340 + 300 + \cdots + 310}{9} = 318.89$$

ii) Διάμεσος δ

Για να προσδιορίσουμε τη διάμεσο παρατάσσουμε το δείγμα σε αύξουσα διάταξη:

290 300 310 320 320 320 330 340 340

↑

Επειδή *ν=9* (περιττού πλήθους δείγμα), διάμεσος είναι η μεσαία παρατήρηση, δηλαδή

$\delta = x_{\left(\frac{\nu+1}{2}\right)} = x_{(5)} = 320$

iii) Επικρατούσα τιμή ή κορυφή M_0

Είναι η τιμή με την μεγαλύτερη συχνότητα, δηλαδή $M_0 = 320$

Β) Μέτρα μεταβλητότητας

i) Διασπορά ή διακύμανση s^2

$$s^2 = \frac{1}{\nu-1}\sum_{i=1}^{\nu}(x_i - \bar{x})^2 = \frac{1}{\nu-1}\left(\sum_{i=1}^{\nu} x_i^2 - \nu\bar{x}^2\right)$$

$$s^2 = \frac{1}{\nu-1}\sum_{i=1}^{\nu}(x_i - \bar{x})^2 = \frac{(340-318.89)^2 + (300-318.89)^2 + \cdots + (310-318.89)^2}{9-1} = 285.31$$

$$s^2 = \frac{1}{\nu-1}\left(\sum_{i=1}^{\nu} x_i^2 - \nu\bar{x}^2\right) = \frac{1}{9-1}[(340^2 + 300^2 + \cdots + 310^2) - 9\cdot 318.89^2]$$

$$= \frac{917500 - 915217.5}{8} = 285.31\ (\text{mg}/100\text{gr})^2$$

Ένα από τα μειονεκτήματα της διακύμανσης είναι ότι δεν εκφράζεται στην ίδια μονάδα μέτρησης με τα δεδομένα. Γι' αυτό συνήθως χρησιμοποιούμε την τυπική απόκλιση, που εκφράζεται στην ίδια μονάδα μέτρησης.

ii) Τυπική απόκλιση $s = \sqrt{s^2}$

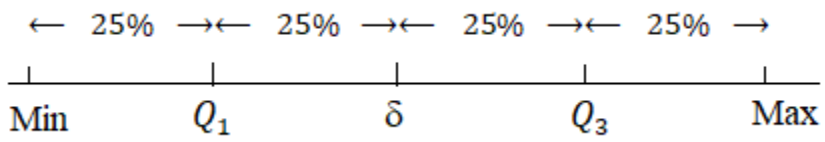
$s = \sqrt{s^2} = \sqrt{285.31}$=16.89 mg/100gr

iii) Συντελεστής μεταβλητότητας CV

$$CV = \frac{s}{\bar{x}}\cdot 100\% = \frac{16.89}{318.89}\cdot 100\% = 5.30\%$$

Καθώς ο $CV < 10\%$ το δείγμα θεωρείται ομοιογενές, δηλαδή έχει μικρή μεταβλητότητα.

iv) Ενδοτεταρτημοριακό εύρος $Q = Q_3 - Q_1$



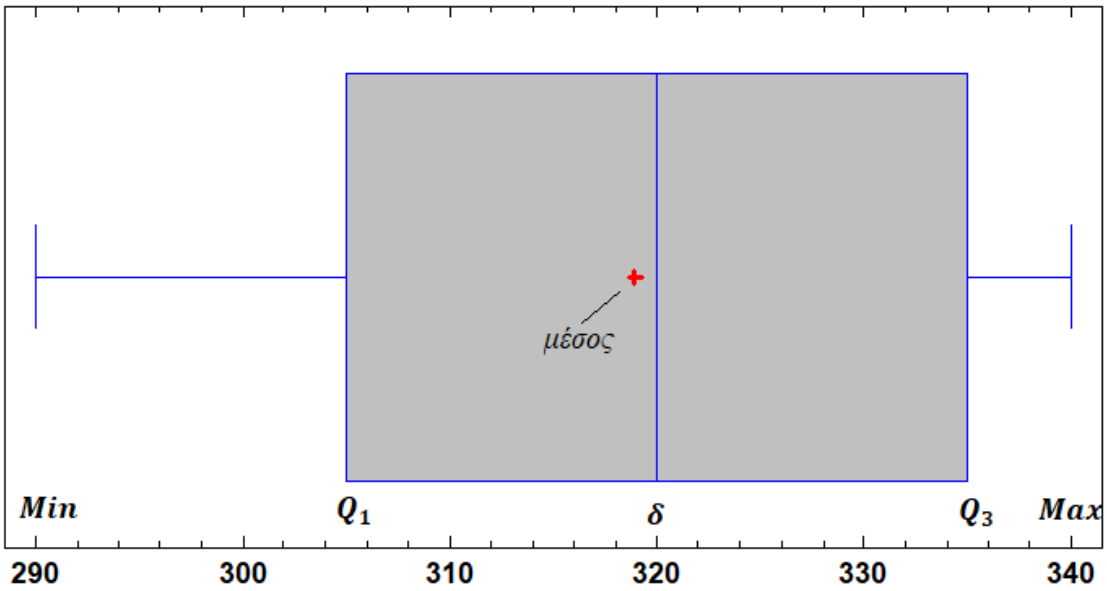
290 300 310 320 320 320 330 340 340

↔ ↑ ↔

Min Q_1 δ Q_3 Max

Επομένως

$$Q_1 = \frac{300+310}{2} = 305\ , \quad Q_3 = \frac{330+340}{2} = 335 \quad και \quad Q = Q_3 - Q_1 = 335 - 305 = 30$$



2. (Συνέχεια της άσκησης 1). Ο φοιτητής μελέτησε την ποσότητα νατρίου στο κασέρι τύπου light της ίδιας γαλακτοβιομηχανίας. Τα αποτελέσματα οκτώ μετρήσεων ήταν: 300 300 310 290 280 280 285 275. Να συγκρίνετε την κατανομή αυτού του δείγματος με την κατανομή του δείγματος της προηγούμενης άσκησης (ως προς την κεντρική τάση, τη μεταβλητότητα και τη λοξότητα).

Θα υπολογίσουμε καταρχάς τα αριθμητικά περιγραφικά μέτρα και στη συνέχεια θα συγκρίνουμε τις κατανομές των δύο δειγμάτων.

Α) ***Μέτρα κεντρικής τάσης ή θέσης***

i) Μέσος όρος ή μέση τιμή $\bar{x}$

$$\bar{x} = \frac{1}{\nu}\sum_{i=1}^{\nu} x_i = \frac{x_1 + x_2 + \cdots + x_\nu}{\nu} = \frac{300 + 300 + \cdots + 275}{8} = \frac{2320}{8} = 290$$

ii) Διάμεσος δ

Για να προσδιορίσουμε τη διάμεσο παρατάσσουμε το δείγμα σε αύξουσα διάταξη:

275 280 280 285 290 300 300 310

↔

Επειδή $\nu=8$ (αρτίου πλήθους δείγμα), η διάμεσος υπολογίζεται ως το ημιάθροισμα των δύο μεσαίων παρατηρήσεων, δηλαδή:

$$\delta = \frac{x_{(\frac{\nu}{2})} + x_{(\frac{\nu}{2}+1)}}{2} = \frac{x_{(4)} + x_{(5)}}{2} = \frac{285 + 290}{2} = 287.5$$

iii) Επικρατούσα τιμή ή κορυφή M_0

Είναι η τιμή με την μεγαλύτερη συχνότητα. Το παραπάνω δείγμα έχει δύο επικρατούσες τιμές, τις

$M_0 = 280$ και $M_0 = 300$ με συχνότητα 2.

Β) Μέτρα μεταβλητότητας

i) Διασπορά ή διακύμανση s^2

$$s^2 = \frac{1}{\nu - 1}\sum_{i=1}^{\nu}(x_i - \bar{x})^2 = \frac{1}{\nu - 1}\left(\sum_{i=1}^{\nu} x_i^2 - \nu\bar{x}^2\right)$$

$$s^2 = \frac{1}{\nu - 1}\sum_{i=1}^{\nu}(x_i - \bar{x})^2 = \frac{(300 - 290)^2 + (300 - 290)^2 + \cdots + (275 - 290)^2}{8 - 1}$$

$$= \frac{10^2 + 10^2 + \cdots + (-15)^2}{7} = \frac{1050}{7} = 150\ (\text{mg/100gr})^2$$

Όπως είπαμε και στην προηγούμενη άσκηση, ένα από τα μειονεκτήματα της διακύμανσης s^2 είναι ότι δεν εκφράζεται στην ίδια μονάδα μέτρησης με τα δεδομένα. Γι' αυτό συνήθως χρησιμοποιούμε την τυπική απόκλιση, που εκφράζεται στην ίδια μονάδα μέτρησης.

ii) Τυπική απόκλιση $s = \sqrt{s^2}$

$s = \sqrt{s^2} = \sqrt{150}$=12.25 mg/100gr

iii) Συντελεστής μεταβλητότητας CV

$$CV = \frac{s}{\bar{x}} \cdot 100\% = \frac{12.25}{290} \cdot 100\% = 4.22\%$$

Καθώς ο $CV < 10\%$ το δείγμα θεωρείται ομοιογενές, δηλαδή έχει μικρή μεταβλητότητα.

iv) Ενδοτεταρτημοριακό εύρος $Q = Q_3 - Q_1$

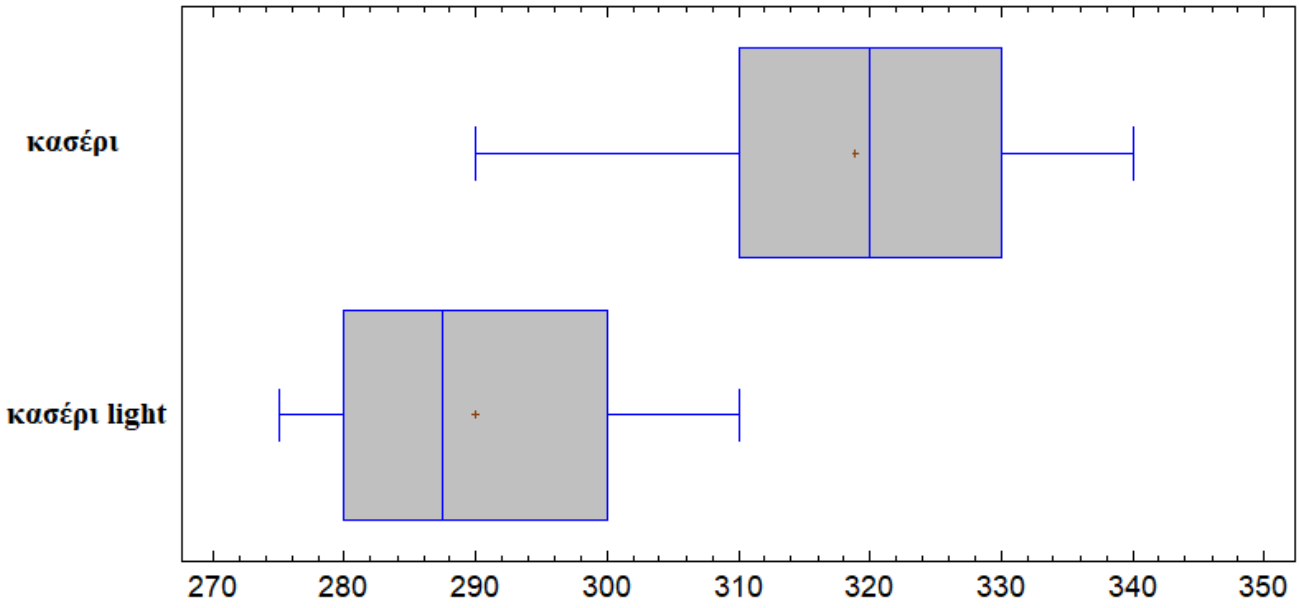
Παρατάσσουμε το δείγμα σε αύξουσα διάταξη:

275 280 280 285 290 300 300 310

Min $\leftrightarrow Q_1$ $\leftrightarrow \delta$ $\leftrightarrow Q_3$ Max

Επομένως

$$Q_1 = \frac{280+280}{2} = 280\ , \quad Q_3 = \frac{300+300}{2} = 300 \quad και \quad Q = Q_3 - Q_1 = 300 - 280 = 20$$



Παρατηρούμε ότι όλα τα αριθμητικά περιγραφικά μέτρα κεντρικής τάσης ή θέσης του 1ου δείγματος έχουν μεγαλύτερες τιμές σε σχέση με τα αντίστοιχα μέτρα του 2ου δείγματος (κασέρι light). Επομένως το κασέρι τύπου light έχει μικρότερη περιεκτικότητα νατρίου αναφορικά με το πλήρες κασέρι. Επιπλέον όλα τα αριθμητικά περιγραφικά μέτρα μεταβλητότητας του 1ου δείγματος έχουν μεγαλύτερες τιμές σε σχέση με τα αντίστοιχα μέτρα του 2ου δείγματος (κασέρι light), όπως και για τους συντελεστές μεταβλητότητας ισχύει: CV_1=5.30% > 4.22% = CV_2. Συνεπώς το 2ο δείγμα (κασέρι light) έχει μικρότερη μεταβλητότητα. Όσον αφορά τη λοξότητα των δύο δειγμάτων έχουμε:

1ο δείγμα (κασέρι πλήρες) $\bar{x}$ = 318.89 < δ = M_0 = 320 επομένως έχουμε μικρή αρνητική ασυμμετρία.

2ο δείγμα (κασέρι light) $\bar{x}$ = 290 > δ =287.5 επομένως έχουμε μικρή θετική ασυμμετρία.

3. Για τα παρακάτω δεδομένα να υπολογιστούν η μέση τιμή, η διάμεσος, η επικρατούσα τιμή, η διασπορά, η τυπική απόκλιση, ο συντελεστής μεταβλητότητας και το ενδοτεταρτημοριακό εύρος. Να κατασκευαστεί επίσης το θηκόγραμμα της κατανομής του δείγματος.

6 6 7 7 3 6 6 6 7 7 7 3 6 5 7 7 1 7 7 7 6 6 5 7 7
6 7 4 7 6 6 5 6 7 6 5 7 4 7 5 7 5 7 4 7 5 7 7 7 5
7 7 2 7 4 7 7 7 5 7 7 7 7 4 7

Έστω y_i, $i = 1, 2, \ldots, 7$ οι διαφορετικές τιμές του δείγματος. Τότε χρησιμοποιώντας τις συχνότητες ν_i έχουμε τον παρακάτω πίνακα:

$\mathbf{y_i}$	1	2	3	4	5	6	7	**Άθροισμα**
$\boldsymbol{\nu_i}$	1	1	2	5	9	13	34	**65**
$\boldsymbol{\nu_i y_i}$	1	2	6	20	45	78	238	**390**
$\boldsymbol{N_i}$	1	2	4	9	18	31	65	
$\boldsymbol{\nu_i y_i^2}$	1	4	18	80	225	468	1666	**2462**

Αριθμητικά περιγραφικά μέτρα

Α) ***Μέτρα κεντρικής τάσης ή θέσης***

i) Μέσος όρος ή μέση τιμή $\bar{x}$

$$\bar{x} = \frac{1}{\nu}\sum_{i=1}^{k} \nu_i y_i = \frac{390}{65} = 6$$

ii) Διάμεσος δ

Εφόσον *ν=65* (περιττού πλήθους δείγμα) η διάμεσος θα είναι η μεσαία παρατήρηση, $\delta = x_{\left(\frac{\nu+1}{2}\right)} = x_{(33)}$ όταν το δείγμα παραταχθεί σε αύξουσα διάταξη. Για τον προσδιορισμό της διαμέσου μας βοηθάει η αθροιστική συχνότητα N_i. Στο συγκεκριμένο δείγμα βέβαια εύκολα καταλαβαίνουμε ότι η διάμεσος είναι $\delta = x_{(33)} = 7$, εφόσον οι 31 πρώτες τιμές είναι ≤ 6 και από την 32η και μετά είναι όλα 7.

iii) Επικρατούσα τιμή ή κορυφή M_0

Είναι η τιμή με την μεγαλύτερη συχνότητα, δηλαδή $M_0 = 7$ με συχνότητα 34.

Β) Μέτρα μεταβλητότητας

***i)* Διασπορά ή διακύμανση s^2**

$$s^2 = \frac{1}{\nu-1}\sum_{i=1}^{\nu}(x_i - \bar{x})^2 = \frac{1}{\nu-1}\left(\sum_{i=1}^{\nu} x_i^2 - \nu\bar{x}^2\right) = \frac{1}{\nu-1}\sum_{i=1}^{k}(y_i - \bar{x})^2\nu_i = \frac{1}{\nu-1}\left(\sum_{i=1}^{k}\nu_i y_i^2 - \nu\bar{x}^2\right)$$

Αφού έχουμε τις συχνότητες ν_i, για τον υπολογισμό της διακύμανσης μπορούμε να χρησιμοποιήσουμε τους δύο τελευταίους τύπους. Επομένως:

$$s^2 = \frac{1}{\nu - 1}\sum_{i=1}^{k}(y_i - \bar{x})^2\nu_i = \frac{(1-6)^2 \cdot 1 + (2-6)^2 \cdot 1 + \cdots + (7-6)^2 \cdot 34}{65-1} = \frac{122}{64} = 1.91 \qquad \text{ή}$$

$$s^2 = \frac{1}{\nu - 1}\left(\sum_{i=1}^{k}\nu_i y_i^2 - \nu\bar{x}^2\right) = \frac{2462 - 65 \cdot 6^2}{64} = \frac{122}{64} = 1.91$$

ii) Τυπική απόκλιση $\boldsymbol{s = \sqrt{s^2}}$

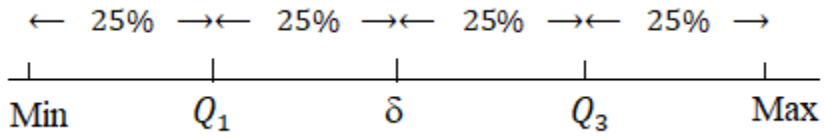
$$\boldsymbol{s = \sqrt{s^2}} = \sqrt{1.91} = 1.38$$

iii) Συντελεστής μεταβλητότητας CV

$$CV = \frac{s}{\bar{x}} \cdot 100\% = \frac{1.38}{6} \cdot 100\% = 23\%$$

Καθώς ο $CV > 10\%$ το δείγμα δεν θεωρείται ομοιογενές.

iv) Ενδοτεταρτημοριακό εύρος $\boldsymbol{Q = Q_3 - Q_1}$

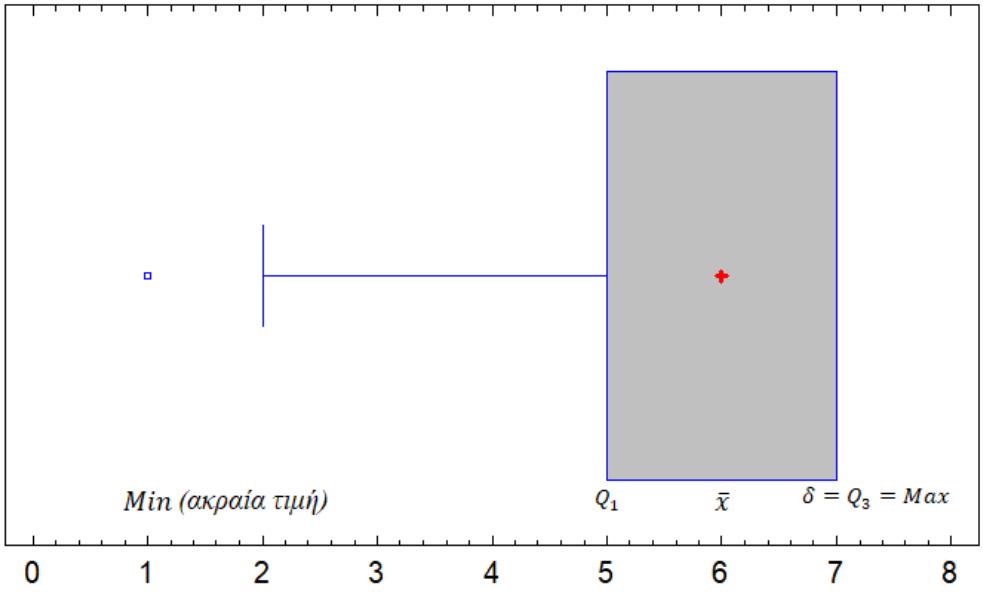


Για τον προσδιορισμό των Q_1 και Q_3 μας βοηθάει η αθροιστική συχνότητα N_i. Στο συγκεκριμένο δείγμα εύκολα καταλαβαίνουμε ότι:

$Q_1 = \frac{x_{(15)} + x_{(16)}}{2} = \frac{5+5}{2} = 5$ και $Q_3 = \frac{x_{(48)} + x_{(49)}}{2} = \frac{7+7}{2} = 7$.

Επομένως $Q = Q_3 - Q_1 = 7 - 5 = 2$

Στη συνέχεια κατασκευάζουμε το θηκόγραμμα, στο οποίο η ελάχιστη τιμή 1 του δείγματος καταγράφεται ως ακραία τιμή, καθώς είναι μικρότερη από: $Q_1 - 1.5Q = 5 - 1.5 \cdot 2 = 5 - 3 = 2$



5. Μετρήσαμε με ηλεκτρονικό μετρητή τον αριθμό των βακτηρίων σε 1cm^3 ενός υγρού. Πήραμε 100cm^3 του υγρού και είχαμε τις ακόλουθες μετρήσεις:

Αριθμός βακτηρίων	0	1	2	3	4
Πλήθος cm^3	12	21	32	25	10

Να υπολογιστούν η μέση τιμή, η διάμεσος, η επικρατούσα τιμή, η διασπορά, η τυπική απόκλιση, ο συντελεστής μεταβλητότητας και το ενδοτεταρτημοριακό εύρος των μετρήσεων.

Έστω y_i, $i = 1, 2, \ldots, 5$ οι διαφορετικές τιμές του δείγματος (αριθμός βακτηρίων) και ν_i οι αντίστοιχες συχνότητες (πλήθος cm^3). Ο παρακάτω πίνακας θα μας βοηθήσει στον υπολογισμό των αριθμητικών περιγραφικών μέτρων:

$\mathbf{y_i}$	0	1	2	3	4	**Άθροισμα**
$\boldsymbol{\nu_i}$	12	21	32	25	10	**100**
$\boldsymbol{\nu_i y_i}$	0	21	64	75	40	**200**
$\boldsymbol{N_i}$	12	33	65	90	100	
$\boldsymbol{\nu_i y_i^2}$	0	21	128	225	160	**534**

Α) ***Μέτρα κεντρικής τάσης ή θέσης***

i) Μέσος όρος ή μέση τιμή $\bar{x}$

$$\bar{x} = \frac{1}{\nu}\sum_{i=1}^{k} \nu_i y_i = \frac{200}{100} = 2$$

ii) Διάμεσος δ

Εφόσον *ν=100* (αρτίου πλήθους δείγμα) η διάμεσος θα είναι το ημιάθροισμα των δύο μεσαίων παρατηρήσεων, όταν το δείγμα παραταχθεί σε αύξουσα διάταξη, δηλαδή:

$$\delta = \frac{x_{(\frac{\nu}{2})} + x_{(\frac{\nu}{2}+1)}}{2} = \frac{x_{(50)} + x_{(51)}}{2} = \frac{2+2}{2} = 2$$

Για τον προσδιορισμό της διαμέσου μας βοηθάει η αθροιστική συχνότητα N_i. Στο συγκεκριμένο δείγμα εύκολα καταλαβαίνουμε ότι $x_{(50)} = x_{(51)} = 2$.

iii) Επικρατούσα τιμή ή κορυφή M_0

Είναι η τιμή με την μεγαλύτερη συχνότητα, δηλαδή $M_0 = 2$ με συχνότητα 32.

Παρατηρούμε ότι $\bar{x} = \delta = M_0 = 2$, επομένως η κατανομή του δείγματος είναι συμμετρική.

Β) Μέτρα μεταβλητότητας

i) **Διασπορά ή διακύμανση** $\mathbf{s^2}$

$$s^2 = \frac{1}{\nu-1}\sum_{i=1}^{\nu}(x_i - \bar{x})^2 = \frac{1}{\nu-1}\left(\sum_{i=1}^{\nu} x_i^2 - \nu\bar{x}^2\right) = \frac{1}{\nu-1}\sum_{i=1}^{k}(y_i - \bar{x})^2\nu_i = \frac{1}{\nu-1}\left(\sum_{i=1}^{k} \nu_i y_i^2 - \nu\bar{x}^2\right)$$

Αφού έχουμε τις συχνότητες ν_i, για τον υπολογισμό της διακύμανσης μπορούμε να χρησιμοποιήσουμε κάποιον από τους δύο τελευταίους τύπους. Επομένως:

$$s^2 = \frac{1}{\nu-1}\sum_{i=1}^{k}(y_i - \bar{x})^2\nu_i = \frac{(0-2)^2\cdot 12 + (1-2)^2\cdot 21 + \cdots + (4-2)^2\cdot 10}{100-1} = \frac{134}{99} = 1.35 \qquad \text{ή}$$

$$s^2 = \frac{1}{\nu-1}\left(\sum_{i=1}^{k} \nu_i y_i^2 - \nu\bar{x}^2\right) = \frac{534 - 100\cdot 2^2}{100-1} = \frac{134}{99} = 1.35$$

ii) Τυπική απόκλιση $\boldsymbol{s = \sqrt{s^2}}$

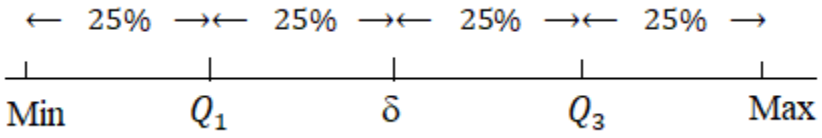
$\boldsymbol{s = \sqrt{s^2}} = \sqrt{1.35} = 1.16$

iii) Συντελεστής μεταβλητότητας CV

$$CV = \frac{s}{\bar{x}} \cdot 100\% = \frac{1.16}{2} \cdot 100\% = 50\%$$

Καθώς ο CV > >10% το δείγμα έχει μεγάλη μεταβλητότητα.

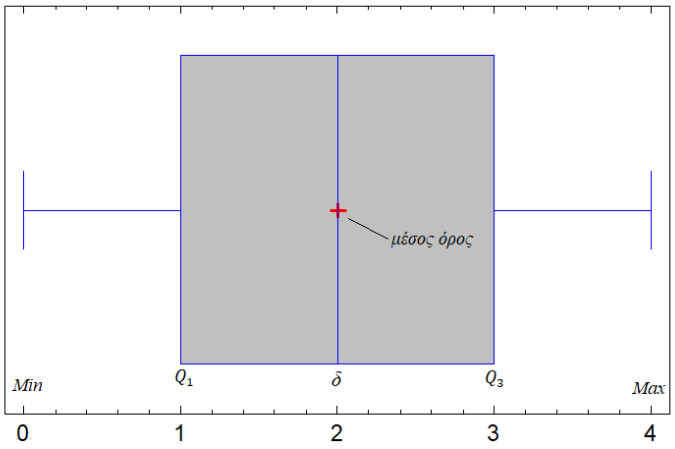
iv) Ενδοτεταρτημοριακό εύρος $Q = Q_3 - Q_1$



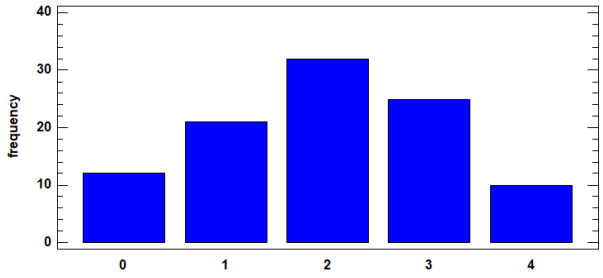
Για τον προσδιορισμό των Q_1 και Q_3 μας βοηθάει η αθροιστική συχνότητα N_i. Στο συγκεκριμένο δείγμα των 100 παρατηρήσεων εύκολα καταλαβαίνουμε ότι:

$Q_1 = \frac{x_{(25)}+x_{(26)}}{2} = \frac{1+1}{2} = 1$ και $Q_3 = \frac{x_{(75)}+x_{(76)}}{2} = \frac{3+3}{2} = 3$.

Επομένως $Q = Q_3 - Q_1 = 3 - 1 = 2$



α) Θηκόγραμμα



β) Ραβδόγραμμα

Η κατανομή του δείγματος είναι συμμετρική και δεν έχουμε ακραίες τιμές.

6. Η απόδοση σε γάλα (lt/24h) μιας προβατίνας που έχει γεννήσει, υπολογίζεται ζυγίζοντας το νεογνό πριν και μετά τον θηλασμό. Πήραμε 19 δείγματα γάλακτος και τα αποτελέσματα ήταν:

2.4 2.7 1.8 3.2 3.4 2.6 3.2 3.4 4.1 2.8 2.9 3.9 4.2 3.6 2.8 3.4 3.7 3.5 2.7

α) Να ομαδοποιήσετε τις παρατηρήσεις σε 5 κλάσεις με πλάτος 0.5 η κάθε μία και αριστερό άκρο της πρώτης κλάσης το 1.75. **β)** Να υπολογίσετε τη μέση τιμή, τη διάμεσο, την επικρατούσα τιμή, τη διασπορά, την τυπική απόκλιση, τον συντελεστή μεταβλητότητας και το ενδοτεταρτημοριακό εύρος των ομαδοποιημένων μετρήσεων, **γ)** Να κατασκευάσετε επίσης το ιστόγραμμα συχνοτήτων και το θηκόγραμμα των μετρήσεων.

α)

Κλάσεις	Κέντρο κλάσης y_i	Συχνότητα ν_i	$\nu_i y_i$	$\nu_i {y_i}^2$	Αθροιστική συχνότητα N_i
1.75 – 2.25	2	1	2	4	1
2.25 – 2.75	2.5	4	10	25	5
2.75 – 3.25	3	5	15	45	10
3.25 – 3.75	3.5	6	21	73.5	16
3.75 – 4.25	4	3	12	48	19
Άθροισμα		**19**	**60**	**195.5**	

β) Αριθμητικά περιγραφικά μέτρα (για δεδομένα ομαδοποιημένα)

1. Μέτρα κεντρικής τάσης ή θέσης

i) Μέσος όρος ή μέση τιμή $\bar{x}$

$$\bar{x} = \frac{1}{\nu}\sum_{i=1}^{k} \nu_i y_i = \frac{60}{19} = 3.16$$

ii) Διάμεσος δ

Όταν τα δεδομένα είναι ομαδοποιημένα σε κλάσεις προσδιορίζουμε καταρχάς την κλάση μέσα στην οποία βρίσκεται η διάμεσος. Σε αυτό μας βοηθάει η αθροιστική συχνότητα N_i. Στο συγκεκριμένο δείγμα των 19 παρατηρήσεων (περιττό πλήθος δείγματος) εύκολα καταλαβαίνουμε ότι $\delta = x_{\left(\frac{\nu+1}{2}\right)} = x_{(10)}$.

Προσδιορίζουμε την κλάση μέσα στην οποία βρίσκεται η $x_{(10)}$, δηλαδή η $10^{η}$ παρατήρηση, όταν το δείγμα είναι σε αύξουσα διάταξη. Με τη βοήθεια της αθροιστικής συχνότητας N_i βλέπουμε ότι η διάμεσος βρίσκεται στην $3^{η}$ κλάση (2.75 – 3.25). Στη συνέχεια εφαρμόζοντας τον παρακάτω τύπο υπολογίζουμε την τιμή της:

$$\delta = L_i + (0.5\nu - N_{i-1})\frac{c}{\nu_i}$$

όπου:

L_i: το κάτω άκρο της κλάσης μέσα στην οποία βρίσκεται η διάμεσος

ν: το μέγεθος του δείγματος

N_{i-1}: η αθροιστική συχνότητα της προηγούμενης κλάσης από αυτήν που βρίσκεται η διάμεσος

c: το εύρος των κλάσεων

ν_i: η συχνότητα της κλάσης μέσα στην οποία βρίσκεται η διάμεσος

Επομένως :

$$\delta = L_i + (0.5\nu - N_{i-1})\frac{c}{\nu_i} = 2.75 + (0.5 \cdot 19 - 5)\frac{0.5}{5} = 2.75 + 0.45 = 3.2$$

iii) Επικρατούσα τιμή ή κορυφή M_0

Όταν τα δεδομένα είναι ομαδοποιημένα σε κλάσεις προσδιορίζουμε καταρχάς την *επικρατούσα κλάση*, την κλάση δηλαδή με τη μεγαλύτερη συχνότητα και στη συνέχεια υπολογίζουμε την επικρατούσα τιμή από τον τύπο:

$$M_0 = L_i + \frac{\Delta_1}{\Delta_1 + \Delta_2} \cdot c$$

όπου:

L_i: το κάτω άκρο της επικρατούσας κλάσης

$\Delta_1 = \nu_i - \nu_{i-1}$, όπου ν_i η συχνότητα της επικρατούσας κλάσης και ν_{i-1} η συχνότητα της προηγούμενης κλάσης

$\Delta_2 = \nu_i - \nu_{i+1}$, όπου ν_i η συχνότητα της επικρατούσας κλάσης και ν_{i+1} η συχνότητα της επόμενης κλάσης

c: το εύρος των κλάσεων

Η επικρατούσα κλάση, δηλαδή η κλάση με τη μεγαλύτερη συχνότητα είναι η 4η κλάση (3.25-3.75) με συχνότητα 6. Εφαρμόζοντας τον παραπάνω τύπο υπολογίζουμε την επικρατούσα τιμή M_0:

$$M_0 = L_i + \frac{\Delta_1}{\Delta_1 + \Delta_2} \cdot c = 3.25 + \frac{6-5}{(6-5)+(6-3)} \cdot 0.5 = 3.25 + \frac{0.5}{4} = 3.25 + 0.125 = 3.375$$

2. Μέτρα μεταβλητότητας

***i)* Διασπορά ή διακύμανση s^2**

Όταν τα δεδομένα είναι ομαδοποιημένα σε κλάσεις η διασπορά υπολογίζεται από τον τύπο:

$$s^2 = \frac{1}{\nu - 1}\left(\sum_{i=1}^{k} \nu_i y_i^2 - \nu \bar{x}^2\right) = \frac{195.5 - 19 \cdot 3.16^2}{19 - 1} = \frac{5.77}{18} = 0.32$$

ii) Τυπική απόκλιση $\boldsymbol{s = \sqrt{s^2}}$

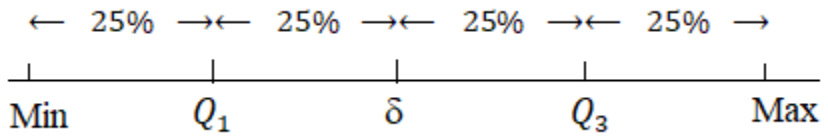
$\boldsymbol{s = \sqrt{s^2}} = \sqrt{0.32} = 0.566$

iii) Συντελεστής μεταβλητότητας CV

$$CV = \frac{s}{\bar{x}} \cdot 100\% = \frac{0.566}{3.16} \cdot 100\% = 17.91\%$$

Καθώς ο $CV > 10\%$ το δείγμα δεν θεωρείται ομοιογενές.

iv) Ενδοτεταρτημοριακό εύρος $\boldsymbol{Q = Q_3 - Q_1}$



Για τον προσδιορισμό των Q_1 και Q_3 σε ομαδοποιημένα δεδομένα δουλεύουμε όπως στη διάμεσο. Δηλαδή προσδιορίζουμε καταρχάς τις κλάσεις μέσα στις οποίες βρίσκονται τα Q_1 και Q_3 και στη συνέχεια με τους παρακάτω τύπους υπολογίζουμε τις τιμές τους.

$$Q_1 = L_i + (0.25\nu - N_{i-1})\frac{c}{\nu_i} \qquad\qquad Q_3 = L_i + (0.75\nu - N_{i-1})\frac{c}{\nu_i}$$

όπου:

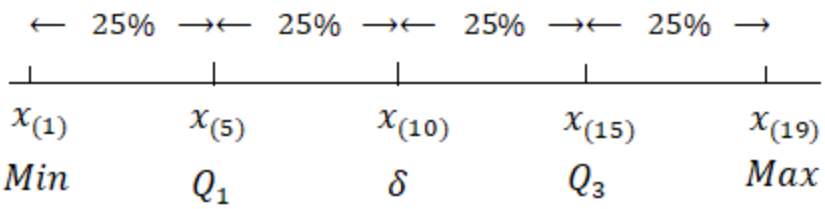
L_i: το κάτω άκρο της κλάσης μέσα στην οποία βρίσκεται το Q_1 ή το Q_3 αντίστοιχα

ν: το μέγεθος του δείγματος

N_{i-1}: η αθροιστική συχνότητα της προηγούμενης κλάσης από αυτήν που βρίσκεται το Q_1 ή το Q_3

c: το εύρος των κλάσεων

ν_i: η συχνότητα της κλάσης μέσα στην οποία βρίσκεται το Q_1 ή το Q_3αντίστοιχα



Στο συγκεκριμένο δείγμα των 19 παρατηρήσεων εύκολα καταλαβαίνουμε ότι $Q_1 = x_{(5)}$ και $Q_3 = x_{(15)}$.

$(0.25 \cdot 19 = 4.75 \rightarrow 5\eta \rightarrow Q_1 = x_{(5)} \quad \text{και} \quad 0.75 \cdot 19 = 14.25 \rightarrow 15\eta \rightarrow Q_3 = x_{(15)})$

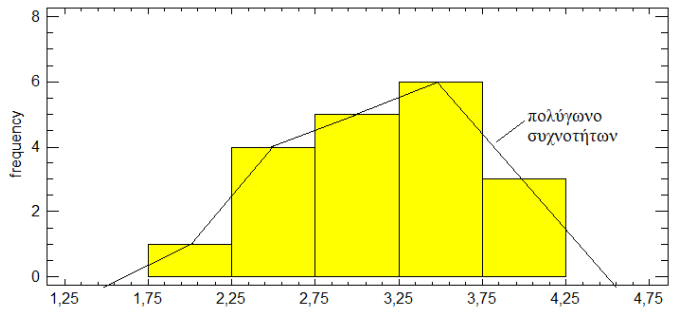
Για τον προσδιορισμό των κλάσεων μέσα στις οποίες βρίσκονται τα Q_1 και Q_3 μας βοηθάει η αθροιστική συχνότητα N_i. Το Q_1, δηλαδή η 5η παρατήρηση βρίσκεται στην 2η κλάση (2.25 – 2.75), ενώ το Q_3, δηλαδή η 15η παρατήρηση βρίσκεται στην 4η κλάση (3.25 – 3.75). Στη συνέχεια εφαρμόζοντας τους αντίστοιχους τύπους υπολογίζουμε τις τιμές τους:

$$Q_1 = L_i + (0.25\nu - N_{i-1})\frac{c}{\nu_i} = 2.25 + (0.25 \cdot 19 - 1)\frac{0.5}{4} = 2.25 + 0.47 = 2.72$$

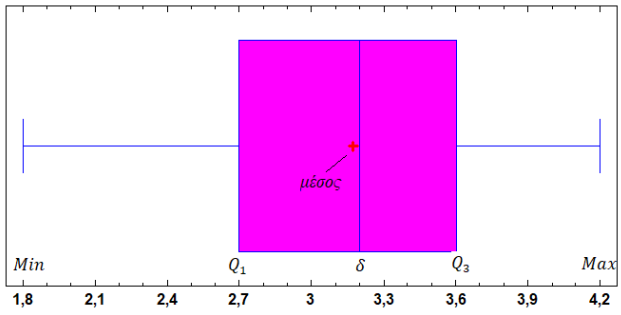
$$Q_3 = L_i + (0.75\nu - N_{i-1})\frac{c}{\nu_i} = 3.25 + (0.75 \cdot 19 - 10)\frac{0.5}{6} = 3.25 + 0.35 = 3.60$$

Επομένως $Q = Q_3 - Q_1 = 3.6 - 2.72 = 0.88$

γ)



i) Ιστόγραμμα συχνοτήτων



ii) Θηκόγραμμα

7. Σε ένα πείραμα, μετρήθηκε το μήκος 100 φύλλων ενός φυτού. Οι μετρήσεις είχαν ως εξής:

Μήκος (cm)	(0-4]	(4-8]	(8-12]	(12-16]	(16-20]
Πλήθος φύλλων	51	20	16	4	9

α) Υπολογίστε τη μέση τιμή, τη διάμεσο, τη διασπορά, την τυπική απόκλιση και το ενδοτεταρτημορικό εύρος των παραπάνω μετρήσεων. Κατασκευάστε πρόχειρα το ιστόγραμμα συχνοτήτων. Είναι συμμετρική η κατανομή των παραπάνω δεδομένων;

β) Οι παραπάνω μετρήσεις $x_1, x_2, \dots, x_{100}$ κατόπιν μετασχηματίστηκαν ως εξής: $y_i = 0.2x_i + 6$. Να βρεθούν η μέση τιμή, η διασπορά και ο συντελεστής μεταβλητότητας των μετασχηματισμένων παρατηρήσεων y_i.

γ) Εάν οι αρχικές μετρήσεις $x_1, x_2, \dots, x_{100}$ μετασχηματιστούν ως εξής: $z_i = \frac{x_i - \bar{x}}{s}$, όπου $\bar{x}$ και s η μέση τιμή και η τυπική απόκλιση των αρχικών μετρήσεων, να υπολογιστούν η μέση τιμή, η διασπορά και η τυπική απόκλιση των z_i.

Κλάσεις	Κέντρο κλάσης y_i	Συχνότητα ν_i	$\nu_i y_i$	$\nu_i {y_i}^2$	Αθροιστική συχνότητα N_i
(0-4]	2	51	102	204	51
(4-8]	6	20	120	720	71
(8-12]	10	16	160	1600	87
(12-16]	14	4	56	784	91
(16-20]	18	9	162	2916	100
Άθροισμα		**100**	**600**	**6224**	

Αριθμητικά περιγραφικά μέτρα (για δεδομένα ομαδοποιημένα)

Α) ***Μέτρα κεντρικής τάσης ή θέσης***

i) Μέσος όρος ή μέση τιμή $\bar{x}$

$$\bar{x} = \frac{1}{\nu}\sum_{i=1}^{k} \nu_i y_i = \frac{600}{100} = 6$$

ii) Διάμεσος δ

Όταν τα δεδομένα είναι ομαδοποιημένα σε κλάσεις προσδιορίζουμε καταρχάς την κλάση μέσα στην οποία βρίσκεται η διάμεσος. Σε αυτό μας βοηθάει η αθροιστική συχνότητα N_i. Στο συγκεκριμένο δείγμα των 100 παρατηρήσεων (άρτιο πλήθος δείγματος) εύκολα καταλαβαίνουμε ότι $\delta = \frac{x_{(50)} + x_{(51)}}{2}$. Επειδή στην 1η κλάση βρίσκονται οι 51 μικρότερες μετρήσεις, η διάμεσος θα βρίσκεται στην 1η κλάση (0-4].

Στη συνέχεια εφαρμόζοντας τον παρακάτω τύπο υπολογίζουμε την τιμή της:

$$\delta = L_i + (0.5\nu - N_{i-1})\frac{c}{\nu_i} = 0 + (0.5 \cdot 100 - 0)\frac{4}{51} = 0 + \frac{200}{51} = 3.92$$

iii) Επικρατούσα τιμή ή κορυφή M_0

Όταν τα δεδομένα είναι ομαδοποιημένα σε κλάσεις προσδιορίζουμε καταρχάς την *επικρατούσα κλάση*, την κλάση δηλαδή με τη μεγαλύτερη συχνότητα. Εδώ είναι η 1η κλάση (0-4] με συχνότητα 51. Στη συνέχεια υπολογίζουμε την επικρατούσα τιμή από τον τύπο:

$$M_0 = L_i + \frac{\Delta_1}{\Delta_1 + \Delta_2} \cdot c = 0 + \frac{51 - 0}{(51 - 0) + (51 - 20)} \cdot 4 = 2.49$$

Παρατηρούμε ότι: $\bar{x} = 6 > \delta = 3.92 > M_0 = 2.49$ επομένως έχουμε έντονη θετική ασυμμετρία.

Β) Μέτρα μεταβλητότητας

i) Διασπορά ή διακύμανση s^2

Όταν τα δεδομένα είναι ομαδοποιημένα σε κλάσεις η διασπορά υπολογίζεται από τον τύπο:

$$s^2 = \frac{1}{\nu - 1}\left(\sum_{i=1}^{k} \nu_i y_i^2 - \nu\bar{x}^2\right) = \frac{6224 - 100 \cdot 6^2}{100 - 1} = \frac{2624}{99} = 26.5$$

Όπως έχουμε ξαναπεί ένα μειονέκτημα της διακύμανσης είναι ότι δεν εκφράζεται στην ίδια μονάδα μέτρησης με τα δεδομένα. Γι' αυτό συνήθως χρησιμοποιούμε την τυπική απόκλιση, που εκφράζεται στην ίδια μονάδα μέτρησης.

ii) Τυπική απόκλιση $s = \sqrt{s^2}$

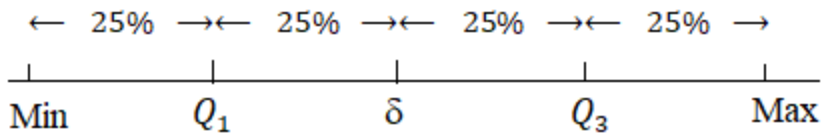
$$s = \sqrt{s^2} = \sqrt{26.5} = 5.15$$

iii) Συντελεστής μεταβλητότητας CV

$$CV = \frac{s}{\bar{x}} \cdot 100\% = \frac{5.15}{6} \cdot 100\% = 85.83\%$$

Καθώς ο $CV >> 10\%$ το δείγμα έχει πολλή μεγάλη μεταβλητότητα.

iv) Ενδοτεταρτημοριακό εύρος $Q = Q_3 - Q_1$



Για τον προσδιορισμό των Q_1 και Q_3 σε ομαδοποιημένα δεδομένα δουλεύουμε όπως στη διάμεσο. Δηλαδή προσδιορίζουμε καταρχάς τις κλάσεις μέσα στις οποίες βρίσκονται τα Q_1 και Q_3 και στη συνέχεια με τους παρακάτω τύπους υπολογίζουμε τις τιμές τους.

$$Q_1 = L_i + (0.25\nu - N_{i-1})\frac{c}{\nu_i} \qquad Q_3 = L_i + (0.75\nu - N_{i-1})\frac{c}{\nu_i}$$

Στο συγκεκριμένο δείγμα των 100 παρατηρήσεων (αρτίου πλήθους δείγμα) εύκολα καταλαβαίνουμε ότι:

$$\delta = \frac{x_{(50)} + x_{(51)}}{2}, \qquad Q_1 = \frac{x_{(25)} + x_{(26)}}{2}, \qquad Q_3 = \frac{x_{(75)} + x_{(76)}}{2}$$

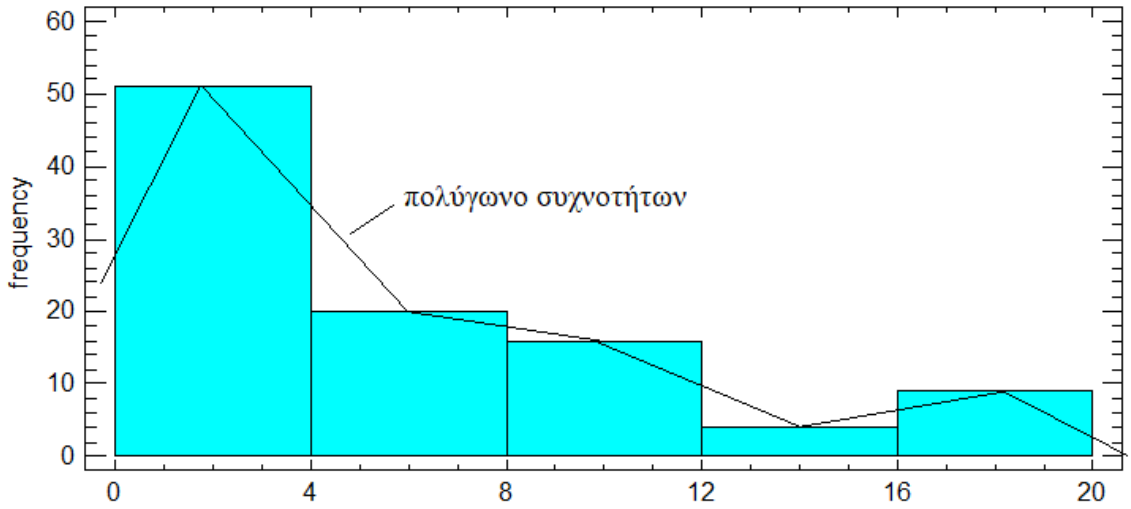
Με τη βοήθεια της αθροιστικής συχνότητας N_i παρατηρούμε ότι το Q_1 βρίσκεται στην 1η κλάση (0 – 4], ενώ το Q_3 βρίσκεται στην 3η κλάση (8 – 12]. Στη συνέχεια εφαρμόζοντας τους αντίστοιχους τύπους υπολογίζουμε τις τιμές τους των Q_1 και Q_3:

$$Q_1 = L_i + (0.25\nu - N_{i-1})\frac{c}{\nu_i} = 0 + (0.25 \cdot 100 - 0)\frac{4}{51} = 0 + \frac{100}{51} = 1.96$$

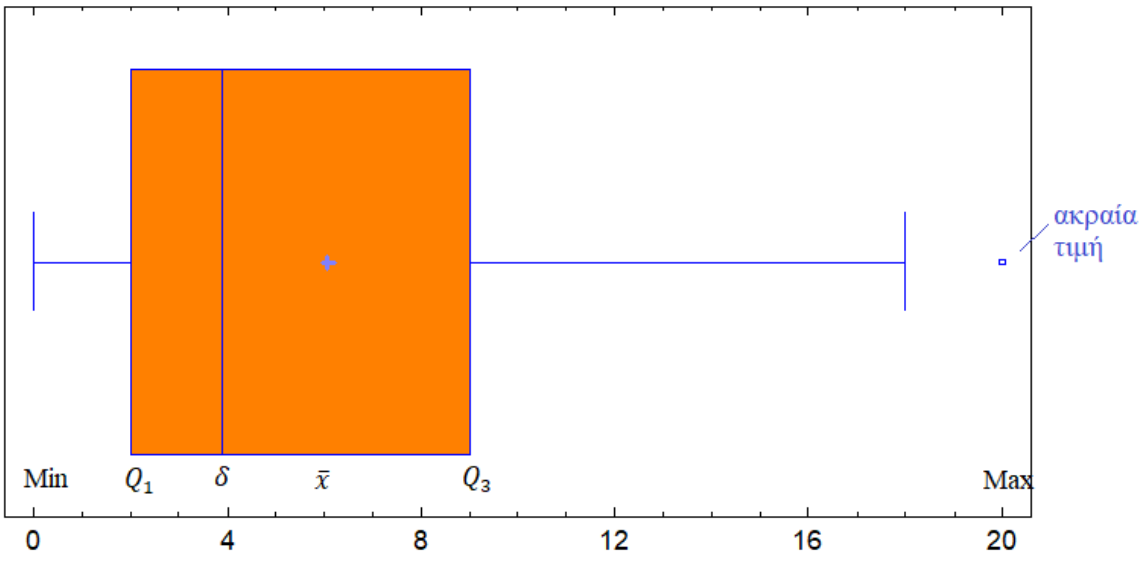
$$Q_3 = L_i + (0.75\nu - N_{i-1})\frac{c}{\nu_i} = 8 + (0.75 \cdot 100 - 71)\frac{4}{16} = 8 + 1 = 9$$

Επομένως $Q = Q_3 - Q_1 = 9 - 1.96 = 7.04$

Ιστόγραμμα και πολύγωνο συχνοτήτων



Στη συνέχεια παραθέτουμε ένα *ενδεικτικό* θηκόγραμμα παρόμοιου δείγματος με αυτό της άσκησης, καθώς δεν μας δίνονται οι 100 μετρήσεις, αλλά μόνο ο πίνακας συχνοτήτων. Δεν γνωρίζουμε επακριβώς την ελάχιστη και την μέγιστη μέτρηση, έτσι ώστε να μπορούμε να κατασκευάσουμε το ακριβές θηκόγραμμα.



Και στα δύο διαγράμματα (ιστόγραμμα & θηκόγραμμα) είναι ολοφάνερη η έντονη θετική ασυμμετρία της κατανομής του δείγματος.

β) $\bar{y} = \alpha\bar{x} + \beta = 0.2\,\bar{x} + 6 = 0.2 \cdot 6 + 6 = 7.2$

$s_y^2 = \alpha^2 s_x^2 = 0.2^2 \cdot 26.5 = 1.06$ και $s_y = |\alpha| s_x = |0.2| \cdot 5.15 = 1.03$

Τότε:

$$CV_y = \frac{s_y}{\bar{y}} 100\% = \frac{1.03}{7.2} 100\% = 14.31\%$$

γ) Για τον μετασχηματισμό:

$$z_i = \frac{x_i - \bar{x}}{s} = \frac{1}{s} x_i - \frac{\bar{x}}{s}$$

όπου $\bar{x}$ και s η μέση τιμή και η τυπική απόκλιση των αρχικών μετρήσεων $x_1, x_2, \ldots, x_{100}$ έχουμε:

$\alpha = \frac{1}{s}$ και $\beta = -\frac{\bar{x}}{s}$ και τότε η μέση τιμή $\bar{z}$ και η τυπική απόκλιση s_z των μετασχηματισμένων δεδομένων είναι:

$\bar{z} = \alpha\bar{x} + \beta = \frac{1}{s}\,\bar{x} - \frac{\bar{x}}{s} = 0$

$s_z^2 = \alpha^2 s_x^2 = (\frac{1}{s})^2 s_x^2 = 1, \quad$ καθώς $\quad s^2 = s_x^2$

και $\quad s_z = |\alpha| s_x = \left|\frac{1}{s}\right| s_x = 1\,, \quad$ καθώς $\quad s = s_x$