Κεφάλαιο 1

ΠΕΡΙΓΡΑΦΙΚΗ ΣΤΑΤΙΣΤΙΚΗ

Σ' αυτό το κεφάλαιο θα δούμε πρώτα τρόπους να παρουσιάσουμε τα δεδομένα με στατιστικούς πίνακες και διαγράμματα και μετά να συνοψίσουμε τα δεδομένα υπολογίζοντας συνοπτικά μέτρα.

1.1 Περιγραφή Στατιστικών Δεδομένων

Οι στατιστικοί πίνακες και γραφικές παραστάσεις αποτελούν χρήσιμα μέσα για να παρουσιάσουμε τα δεδομένα καθαρά, σύντομα και με σαφήνεια. Επίσης μπορούν να αποκαλύψουν σημαντικά χαρακτηριστικά των δεδομένων, όπως το εύρος τους, τη συμμετρικότητα τους ή την ύπαρξη ακραίων τιμών.

Πίνακας συχνοτήτων Τα δεδομένα ενός δείγματος για μια τ.μ. X που παίρνει τιμές σ' ένα σχετικά μικρό σύνολο διακεκριμένων τιμών (κατηγορίες ή αριθμητικές τιμές) μπορούν εύκολα να παρουσιαστούν σ' ένα **πίνακα συχνοτήτων** (frequency table). Ο πίνακας συχνοτήτων παρουσιάζει για κάθε τιμή x_i της X τη συχνότητα εμφάνισής της f_i, δηλαδή πόσες φορές εμφανίζεται η κάθε διακεκριμένη τιμή στο δείγμα. Εύκολα μπορούμε επίσης να υπολογίσουμε και τη **σχετική συχνότητα** (relative frequency) εμφάνισης ή αλλιώς το **ποσοστό** (percent) p_i που ορίζεται από το λόγο της συχνότητας εμφάνισης f_i μιας τιμής x_i προς το σύνολο των παρατηρήσεων n του δείγματος

$$p_i = \frac{f_i}{n}. \tag{1.1}$$

Ορίζεται επίσης η **αθροιστική συχνότητα** (cumulative frequency) F_i μιας τιμής x_i ως το άθροισμα των συχνοτήτων όλων των τιμών που είναι μικρότερες ή ίσες της x_i,

$$F_i = \sum_{j=1}^{i} f_j \quad \text{όπου} \quad x_j \leq x_i \quad \text{για} \quad j \leq i.$$

Με τον ίδιο τρόπο ορίζεται και η αθροιστική σχετική συχνότητα P_i

$$P_i = \sum_{j=1}^{i} p_j \quad \text{όπου} \quad x_j \leq x_i \quad \text{για} \quad j \leq i.$$

Ο πίνακας της σχετικής συχνότητας αντιστοιχεί στην εμπειρική ή δειγματική συνάρτηση μάζας πιθανότητας της διακριτής τ.μ. X, δηλαδή στην εκτίμηση της $f_X(x)$ βασισμένη στο δείγμα. Όμοια ο πίνακας της αθροιστικής σχετικής συχνότητας μας δίνει μια εκτίμηση της αθροιστικής συνάρτησης κατανομής $F_X(x)$ από το δείγμα.

Ραβδόγραμμα Τα δεδομένα του πίνακα συχνοτήτων εύκολα μπορούν να παρασταθούν γραφικά σ' ένα **ραβδόγραμμα** (bar chart), όπου η κάθε ράβδος παρουσιάζει τη συχνότητα (ή τη σχετική συχνότητα ή την αθροιστική συχνότητα, ή ακόμα την αθροιστική σχετική συχνότητα) για κάθε τιμή x_i. Το ραβδόγραμμα σχετικής συχνότητας είναι το γράφημα της εμπειρικής ή δειγματικής συνάρτησης μάζας πιθανότητας (δηλαδή της εκτίμησης της $f_X(x)$ από το δείγμα). Αντίστοιχα το ραβδόγραμμα της αθροιστικής σχετικής συχνότητας απεικονίζει τη δειγματική αθροιστική συνάρτηση κατανομής. Η ίδια πληροφορία μπορεί να δοθεί και με άλλου είδους γραφήματα, όπως μ' ένα **κυκλικό διάγραμμα** ή **διάγραμμα πίτας** (pie chart) όπου το κάθε κομμάτι της επιφάνειας του κύκλου ('πίτα') παρουσιάζει τη συχνότητα της αντίστοιχης τιμής.

Παράδειγμα 1.1. *Μια εταιρεία τροφοδοσίας χημικών προϊόντων πουλάει ένα χημικό προϊόν στους πελάτες της σε παρτίδες των 5 λίτρων. Η εταιρεία θέλει να μελετήσει την ποσότητα του προϊόντος σε κάθε παραγγελία. Από τα αρχεία της εταιρείας βρέθηκαν 120 πρόσφατες παραγγελίες αυτού του χημικού προϊόντος και ο αριθμός των παρτίδων σε κάθε παραγγελία δίνεται στον Πίνακα 1.1.*

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

Πίνακας 1.1: Αριθμός παρτίδων σε δείγμα 120 παραγγελιών ενός χημικού προϊόντος.

Η τ.μ. X είναι ο αριθμός των παρτίδων του χημικού προϊόντος ανά παραγγελία. Τα δεδομένα του Πίνακα 1.1 είναι διακριτά αριθμητικά (ως αριθμοί από το 1 ως το ανώτατο αριθμό παρτίδων). Με κατάλληλη επεξεργασία θα μπορούσαμε να οργανώσουμε τα δεδομένα σε διατακτικά κατηγορικά, δηλαδή ως κατηγορίες μεγέθους παραγγελίας με βάση τον αριθμό των παρτίδων, μικρού μεγέθους για 1 ή 2 παρτίδες, μεσαίου μεγέθους για 3 ή 4 παρτίδες και μεγάλου μεγέθους για περισσότερες από 4 παρτίδες.

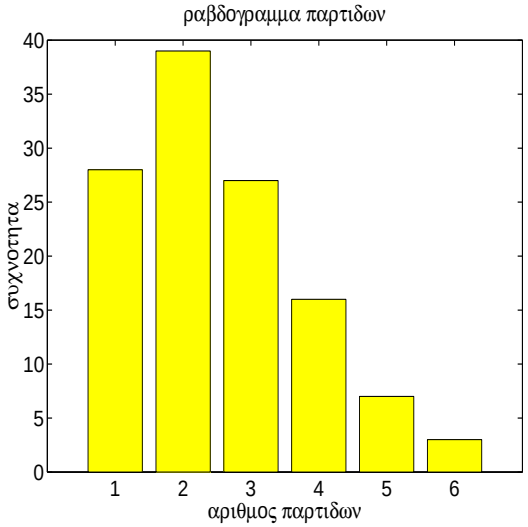
Με βάση τον Πίνακα 1.1 δεν είναι εύκολο να μελετήσουμε την συχνότητα εμφάνισης των διαφόρων αριθμών παρτίδων. Ποιος αριθμός παρτίδων εμφανίζεται συχνότερα; Είναι περισσότερες παραγγελίες των 2 παρτίδων ή των 3 παρτίδων; Για να απαντήσουμε σε τέτοια ερωτήματα μπορούμε να μετρήσουμε πόσες φορές εμφανίζεται ο κάθε αριθμός παρτίδων και να φτιάξουμε έτσι του πίνακα συχνοτήτων. Αυτούς τους απλούς υπολογισμούς μπορούμε εύκολα να τους κάνουμε μόνοι μας αλλά όταν το δείγμα είναι μεγάλο θα χρειαστεί να χρησιμοποιήσουμε κάποιο υπολογιστικό πρόγραμμα (όπως το στατιστικό πακέτο SPSS).

x_i	f_i	p_i	F_i	P_i
1	28	0.23	28	0.23
2	39	0.33	67	0.56
3	27	0.23	94	0.78
4	16	0.13	110	0.92
5	7	0.06	117	0.97
6	3	0.03	120	1.00
Άθροισμα	120	1.00		

Πίνακας 1.2: Πίνακας συχνοτήτων για τον αριθμό παρτίδων σε 120 παραγγελίες του Πίνακα 1.1 που περιλαμβάνει τη συχνότητα f_i, τη σχετική συχνότητα p_i, τη αθροιστική συχνότητα F_i και τη σχετική αθροιστική συχνότητα P_i).

Ο Πίνακας 1.2 παρουσιάζει τις τιμές της X (αριθμός παρτίδων x_i για $i = 1, \ldots, 6$) στην πρώτη στήλη, τη συχνότητα f_i της κάθε τιμής x_i στη δεύτερη στήλη, τη σχετική συχνότητα (ποσοστό) p_i στην τρίτη στήλη, την αθροιστική συχνότητα F_i στην τέταρτη στήλη και τη σχετική αθροιστική συχνότητα P_i στην πέμπτη στήλη.

Στο Σχήμα 1.1 παρουσιάζεται το ραβδόγραμμα που προκύπτει από τις συχνότητες εμφάνισης του κάθε αριθμού παρτίδων. Από του πίνακα συχνοτήτων και το ραβδόγραμμα είναι φανερό πως



Σχήμα 1.1: Ραβδόγραμμα του αριθμού παρτίδων σε 120 παραγγελίες του Πίνακα 1.1.

οι περισσότερες παραγγελίες στο δείγμα μας είναι 2 παρτίδων, λιγότερες είναι 1 παρτίδας ή 3 παρτίδων και η συχνότητα φθείνει καθώς αυξάνει ο αριθμός παρτίδων, δηλαδή για παραγγελίες 4, 5 και 6 παρτίδων.

Ομαδοποίηση Όταν τα δεδομένα είναι αριθμητικά και είτε ο αριθμός των διακεκριμένων τιμών που παίρνει η τ.μ. X είναι μεγάλος, ή η X είναι συνεχής (δηλαδή παίρνει τιμές σ' ένα διάστημα τιμών), είναι προφανές ότι οι πίνακες και τα γραφήματα συχνοτήτων των τιμών δεν προσφέρονται για την απεικόνιση των δεδομένων. Σε τέτοιες περιπτώσεις πρώτα χωρίζουμε τα

δεδομένα σε k **ομάδες** (groups), ή κλάσεις διαστημάτων, και μετά παρουσιάζουμε σε πίνακα ή σε γράφημα τη συχνότητα της κάθε ομάδας, δηλαδή τον αριθμό των παρατηρήσεων που ανήκουν σε κάθε ομάδα.

Το εύρος τιμών r_i της κάθε i ομάδας είναι συνήθως το ίδιο ($r_i = r$). Δεν υπάρχει συγκεκριμένος τρόπος να το καθορίσουμε και το διαλέγουμε ανάλογα με την κλίμακα τιμών για την οποία μας ενδιαφέρει να δούμε διαφορές. Γενικά φροντίζουμε να είναι τέτοιο ώστε να μην προκύπτουν πολλές ομάδες με αποτέλεσμα να έχουμε μικρές συχνότητες γιατί τότε δε μπορούμε να διακρίνουμε κάποιο σχηματισμό στα δεδομένα. Από την άλλη δε θα πρέπει οι ομάδες να είναι πολύ λίγες γιατί τότε δε θα μπορούμε να διακρίνουμε διαφορές παρά μόνο για μεγάλες κλίμακες τιμών.

Για το χωρισμό των δεδομένων σε ομάδες βρίσκουμε πρώτα τη μικρότερη (ελάχιστη) τιμή $x_{\min}$ και μεγαλύτερη (μέγιστη) τιμή $x_{\max}$ και υπολογίζουμε το εύρος των δεδομένων

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

Διαιρώντας το R με τον αριθμό των ομάδων k που επιλέγουμε έχουμε το εύρος τιμών r της κάθε ομάδας το οποίο συνήθως στρογγυλοποιούμε για να έχουμε εύχρηστα νούμερα. Η πρώτη ομάδα έχει σαν κάτω άκρο του διαστήματος κάποιον κατάλληλα στρογγυλοποιημένο αριθμό μικρότερου ή ίσου του $x_{\min}$, τα διαστήματα των ομάδων είναι ισομήκη και το διάστημα της τελευταίας ομάδας περιλαμβάνει το $x_{\max}$. Για τα διαστήματα διαλέγουμε συμβατικά να είναι κλειστά από αριστερά (να περιέχουν την ακραία μικρότερη τιμή) κι ανοιχτά από δεξιά (να μην περιέχουν την ακραία μεγαλύτερη τιμή).

Ιστόγραμμα Έχοντας ομαδοποιήσει τα αριθμητικά δεδομένα μπορούμε να κάνουμε τον πίνακα και τα γραφήματα συχνοτήτων όπως και πριν. Ειδικότερα για το ραβδόγραμμα για την κάθε ράβδο παίρνουμε το κέντρο του διαστήματος που αντιστοιχεί σε κάθε ομάδα. Επίσης δεν υπάρχει κενό διάστημα μεταξύ των ράβδων και το γράφημα αυτό λέγεται **ιστόγραμμα** (histogram). Στον κάθετο άξονα του ιστογράμματος μπορεί να είναι η συχνότητα f_i, η σχετική συχνότητα (ποσοστό) p_i, η αθροιστική συχνότητα F_i, ή ακόμα η σχετική αθροιστική συχνότητα P_i για την κάθε i ομάδα. Σε πλήρη αντιστοιχία με το ραβδόγραμμα σχετικής συχνότητας το ιστόγραμμα σχετικής συχνότητας (ή απλά το ιστόγραμμα συχνότητας) δίνει μια εκτίμηση του γραφήματος της συνάρτησης πυκνότητας πιθανότητας $f_X(x)$ της συνεχούς τ.μ. X, βασισμένη στο δείγμα και στην επιλεγμένη ομαδοποίηση των παρατηρήσεων.

Το ιστόγραμμα είναι χρήσιμο για να κρίνουμε αν μπορούμε να δεχθούμε ότι η τ.μ. X, όπως την παρατηρήσαμε από το δείγμα, ακολουθεί κάποια γνωστή κατανομή. Εδώ κυρίως ενδιαφερόμαστε για την **κανονική κατανομή** γιατί τότε μπορούμε να χρησιμοποιήσουμε μια ολόκληρη μεθοδολογία εκτιμητικής που βασίζεται στην κανονική κατανομή της X. Για μικρά δείγματα (με λιγότερες από 30 παρατηρήσεις) δεν περιμένουμε το ιστόγραμμα να δίνει το σχήμα καμπάνας της κανονικής κατανομής (ακόμα και για δεδομένα που προέρχονται από κανονική κατανομή είναι δυνατόν να παρατηρήσουμε σημαντικές αποκλίσεις). Συνήθως δεχόμαστε ότι η τ.μ. X ακολουθεί κανονική κατανομή όταν το ιστόγραμμα φαίνεται να διατηρεί κάποια σχετική συμμετρία με τις υψηλότερες συχνότητες να παρουσιάζονται στα κεντρικά διαστήματα τιμών.

Υπάρχουν κι άλλα γραφήματα που απεικονίζουν την εμπειρική κατανομή της τ.μ. X με διαφορετικό τρόπο. Αναφέρουμε ενδεικτικά χωρίς να τα περιγράψουμε το *φυλλογράφημα* (stem and leaf plot) και το *σημειογράφημα* (dotplot).

Παράδειγμα 1.2. *Στον Πίνακα 1.3 δίνονται οι μετρήσεις του πορώδους ηλίου (σε ποσοστά) από 25 δοκίμια γαιάνθρακα που διαλέχτηκαν τυχαία από ένα κοίτασμα Α (και αντίστοιχα για 20 δοκίμια γαιάνθρακα από ένα άλλο κοίτασμα Β που θα μελετήσουμε αργότερα).*

Α/Α	κοίτασμα Α	κοίτασμα Β
1	5.3	5.0
2	4.5	4.2
3	5.7	5.4
4	5.8	5.5
5	4.8	4.6
6	6.4	6.1
7	6.4	6.1
8	5.6	5.3
9	5.8	5.5
10	5.7	5.4
11	5.5	5.2
12	6.1	5.8
13	5.2	4.9
14	7.0	6.7
15	5.5	5.2
16	5.7	5.4
17	6.3	6.0
18	5.6	5.3
19	5.5	5.2
20	5.0	4.8
21	5.8	
22	4.7	
23	6.1	
24	6.7	
25	5.1	
Σύνολο	141.8	107.6

Πίνακας 1.3: Δεδομένα πορώδους ηλίου (σε ποσοστά) από δοκίμια γαιάνθρακα δύο κοιτασμάτων Α και Β.

Ονομάζουμε X την τυχαία μεταβλητή του πορώδους ηλίου σε γαιάνθρακα του κοιτάσματος Α. Το μικρότερο πορώδες ηλίου είναι $x_{min} = 4.5$, το μεγαλύτερο πορώδες είναι $x_{max} = 7.0$ και το εύρος των δεδομένων του δείγματος είναι

$$R = x_{max} - x_{min} = 7.0 - 4.5 = 2.5.$$

Διαλέγουμε να χωρίσουμε τα δεδομένα σε 10 ομάδες ($k = 10$) κι άρα η κάθε ομάδα καλύπτει εύρος τιμών

$$r = \frac{R}{k} = \frac{2.5}{10} = 0.25,$$

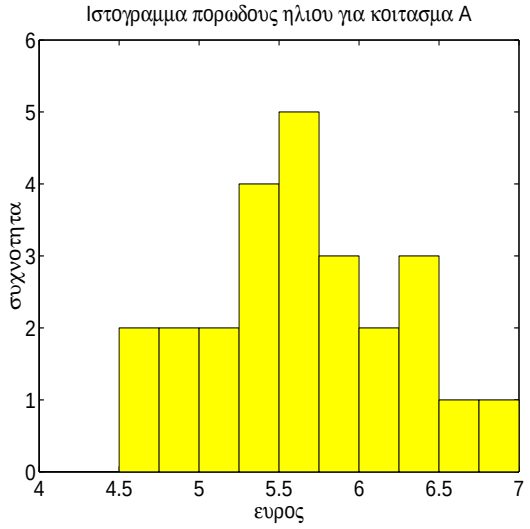
αρχίζοντας από την τιμή 4.5. Στη συνέχεια μπορούμε να υπολογίσουμε του πίνακα συχνοτήτων μετρώντας τον αριθμό των δεδομένων σε κάθε ομάδα.

Στον Πίνακα 1.4 δίνεται ο πίνακας συχνοτήτων, που περιλαμβάνει τη συχνότητα, τη σχετική συχνότητα, την αθροιστική συχνότητα και την σχετική αθροιστική συχνότητα. Παρατηρούμε ότι

Διάστημα τιμών	f_i	p_i	F_i	P_i
4.50 – 4.75	2	0.08	2	0.08
4.75 – 5.00	2	0.08	4	0.16
5.00 – 5.25	2	0.08	6	0.24
5.25 – 5.50	4	0.16	10	0.40
5.50 – 5.75	5	0.20	15	0.60
5.75 – 6.00	3	0.12	18	0.72
6.00 – 6.25	2	0.08	20	0.80
6.25 – 6.50	3	0.12	23	0.92
6.50 – 6.75	1	0.04	24	0.96
6.75 – 7.00	1	0.04	25	1.00
Άθροισμα	25	1.00		

Πίνακας 1.4: Πίνακας συχνοτήτων για τα δεδομένα του πορώδες ηλίου σε γαιάνθρακα κοιτάσματος Α που περιλαμβάνει τη συχνότητα f_i, τη σχετική συχνότητα p_i, την αθροιστική συχνότητα F_i και τη σχετική αθροιστική συχνότητα P_i).

οι τιμές του ποσοστού πορώδες ηλίου στο δείγμα των 25 δοκιμίων γαιάνθρακα συγκεντρώνονται σε ένα κεντρικό διάστημα τιμών (περίπου οι μισές παρατηρήσεις εμφανίζονται στις τρεις ομάδες που καλύπτουν το διάστημα $[5.25, 6.0]$*). Μπορούμε να σχηματίσουμε καλύτερη εντύπωση για την κατανομή της τυχαίας μεταβλητής X του πορώδες ηλίου από το ιστόγραμμα συχνοτήτων, που παρουσιάζεται στο Σχήμα 1.2. Πράγματι από το ιστόγραμμα φαίνεται ότι οι παρατηρήσεις*



Σχήμα 1.2: Ιστόγραμμα των δεδομένων πορώδες ηλίου σε γαιάνθρακα του κοιτάσματος Α του Πίνακα 1.3.

συγκεντρώνονται σ΄ ένα κεντρικό διάστημα τιμών κι απλώνονται με κάποια συμμετρία γύρω από αυτό. Η εικόνα που δίνει το ιστόγραμμα είναι συνεπής με την γραφική παράσταση καυουικής

κατανομής (σχήμα καμπάνας) και μας επιτρέπει να δεχτούμε ότι η κατανομή του πορώδες ηλίου σε γαιάνθρακα του κοιτάσματος Α είναι κανονική.

1.2 Περιγραφικά Μέτρα Στατιστικών Δεδομένων

Ο πίνακας συχνοτήτων και το ραβδόγραμμα ή ιστόγραμμα δίνουν μια συνοπτική παρουσίαση των δεδομένων και μας επιτρέπουν να μελετήσουμε ποιοτικά την κατανομή της τυχαίας μεταβλητής που παρατηρήσαμε. Στη συνέχεια θα ορίσουμε ποσοτικά μεγέθη που περιγράφουν περιληπτικά τα βασικά χαρακτηριστικά της κατανομής της τ.μ. X και λέγονται **συνοπτικά ή περιγραφικά μέτρα** (summarizing or descriptive statistics). Κάθε τέτοιο μέτρο υπολογίζεται από τις παρατηρήσεις του δείγματος κι όπως θα δούμε στα επόμενα κεφάλαια αποτελεί εκτίμηση κάποιας παραμέτρου της κατανομής της τ.μ. που μελετάμε.

Θα ασχοληθούμε με δύο τύπους περιγραφικών μέτρων:

- τα **μέτρα θέσης** (measures of location) που προσδιορίζουν χαρακτηριστικές θέσεις μέσα στο εύρος των δεδομένων και
- τα **μέτρα μεταβλητότητας** (variability measures) που δίνουν περιληπτικά τη διασκόρπιση και μεταβλητότητα των δεδομένων.

1.2.1 Μέτρα θέσης

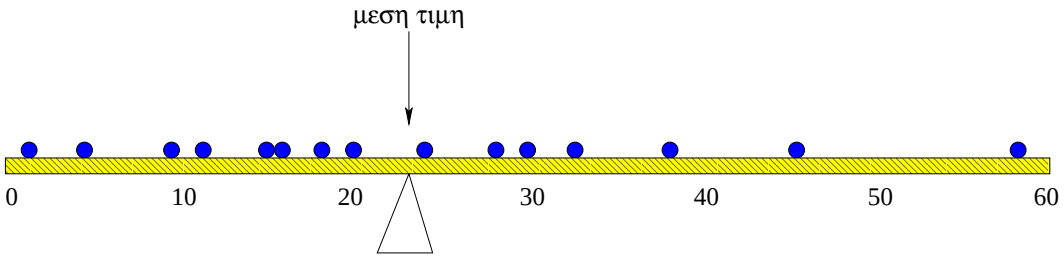
Ως μέτρα θέσης εννοούμε κυρίως τα μέτρα κεντρικής τάσης που προσδιορίζουν ένα κεντρικό σημείο γύρω από το οποίο τείνουν να συγκεντρώνονται τα δεδομένα. Τα κυριότερα μέτρα κεντρικής τάσης είναι:

- η **δειγματική μέση τιμή** (sample mean value) ή **αριθμητικός μέσος** (arithmetic mean), ή **μέσος όρος** (average),
- η **δειγματική διάμεσος** (sample median),
- η **δειγματική επικρατούσα τιμή** (sample mode).

Μέση τιμή Η δειγματική μέση τιμή είναι το πιο γνωστό και χρήσιμο μέτρο του κέντρου των δεδομένων. Έστω $x_1, x_2, \ldots, x_n$, οι τιμές των παρατηρήσεων του δείγματος για μια τ.μ. X που μελετάμε. Η δειγματική μέση τιμή συμβολίζεται $\bar{x}$ κι ορίζεται ως

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i. \tag{1.2}$$

Η μέση τιμή είναι το 'κέντρο ισορροπίας' των δεδομένων. Για να καταλάβουμε τη φυσική της σημασία ας φανταστούμε μία σανίδα πάνω στην οποία σκορπίζουμε ένα αριθμό n ίδιων βαριδιών. Το σημείο στήριξης της σανίδας (ώστε να ισορροπεί σε οριζόντια θέση) είναι η μέση τιμή της θέσης των βαριδίων πάνω στη σανίδα, όπως φαίνεται και στο Σχήμα 1.3.



Σχήμα 1.3: Σχηματική παρουσίαση της μέσης τιμής.

Διάμεσος Η δειγματική διάμεσος είναι ένα άλλο μέτρο του κέντρου των δεδομένων και ορίζεται ως η κεντρική τιμή όταν διατάξουμε τα δεδομένα σε αύξουσα σειρά. Θα τη συμβολίζουμε ως $\tilde{x}$. Αν ο αριθμός n των δεδομένων είναι περιττός τότε η διάμεσος είναι η τιμή στη θέση $(n+1)/2$, ενώ αν το n είναι άρτιος τότε είναι το ημιάθροισμα των τιμών στις θέσεις $n/2$ και $n/2+1$, δηλαδή

$$\tilde{x} = \begin{cases} x_{(n+1)/2} & n = 2k+1 \\ \dfrac{x_{n/2} + x_{n/2+1}}{2} & n = 2k. \end{cases} \tag{1.3}$$

Για παράδειγμα σε δείγμα τριών τιμών η διάμεσος είναι η δεύτερη μικρότερη τιμή και σε δείγμα τεσσάρων τιμών η διάμεσος είναι ο μέσος όρος της δεύτερης και τρίτης μικρότερης τιμής.

Επικρατούσα τιμή Η δειγματική επικρατούσα τιμή χρησιμοποιείται επίσης για να δηλώσει την κεντρική τάση των δεδομένων κι ορίζεται ως η τιμή που εμφανίζεται με τη μεγαλύτερη συχνότητα. Αν υπάρχουν πάνω από μία τέτοιες τιμές, τότε όλες αυτές θεωρούνται επικρατούσες τιμές. Είναι φανερό πως η επικρατούσα τιμή δεν έχει νόημα όταν το δείγμα δεν αποτελείται από διακεκριμένες επαναλαμβονόμενες τιμές.

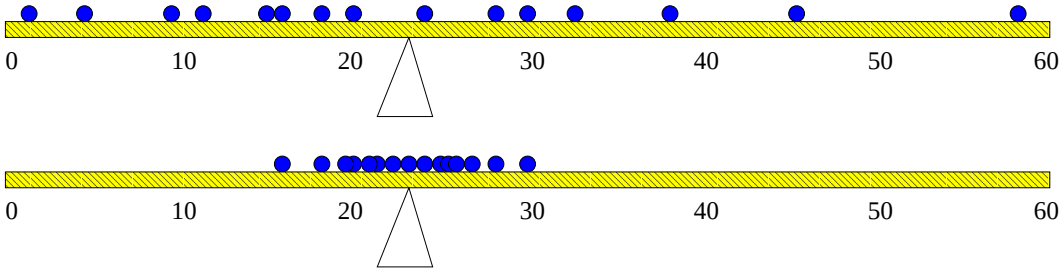
Η δειγματική μέση τιμή είναι το πιο σημαντικό από τα τρία μέτρα κεντρικής τάσης και θα μας απασχολήσει ιδιαίτερα καθώς θα τη χρησιμοποιήσουμε στη στατιστική συμπερασματολογία (στα επόμενα κεφάλαια) για να βγάλουμε συμπεράσματα για τη μέση τιμή μ του πληθυσμού. Για τον υπολογισμό της μέσης τιμής χρησιμοποιούνται όλες οι τιμές του δείγματος, ενώ για τη διάμεσο μόνο η τάξη τους. Γι αυτό και η μέση τιμή επηρεάζεται από μακρινές τιμές αλλά η διάμεσος όχι. Όταν η κατανομή των αριθμητικών δεδομένων είναι μονοκόρυφη και συμμετρική, τότε και τα τρία μέτρα κεντρικής τάσης συμπίπτουν.

Η ύπαρξη μακρινών παρατηρήσεων στο δείγμα δυσκολεύει τη στατιστική περιγραφή κι ανάλυση. Γι αυτό πριν προχωρήσουμε θα πρέπει να αποφασίσουμε αν θα συμπεριλάβουμε τη μακρινή παρατήρηση (αν πιστεύουμε ότι είναι σωστή) ή αν θα την αγνοήσουμε (αν έχουμε λόγους να πιστεύουμε ότι δεν είναι ακριβής).

1.2.2 Μέτρα μεταβλητότητας

Εκτός από την κεντρική τάση μας ενδιαφέρει επίσης και η μεταβλητότητα ή διασπορά των παρατηρήσεων. Όταν τα δεδομένα είναι συγκεντρωμένα γύρω από μια κεντρική τιμή, δηλαδή η διασπορά των δεδομένων είναι μικρή, τότε η κεντρική τιμή αντιπροσωπεύει ικανοποιητικά τα δεδομένα. Από την άλλη, όταν τα δεδομένα είναι πολύ σκορπισμένα τα μέτρα κεντρικής τιμής δε δίνουν καλή περιληπτική περιγραφή των δεδομένων. Επίσης, διαφορετικά δείγματα

από τον ίδιο πληθυσμό μπορεί να έχουν το ίδιο μέτρο κεντρικής τάσης αλλά να διαφέρουν κατά κάποιο σημαντικό τρόπο ως προς τη διασπορά των παρατηρήσεων. Χρησιμοποιώντας το παράδειγμα με τα βαρίδια και τη σανίδα βλέπουμε στο Σχήμα 1.4 πώς δύο δείγματα που έχουν την ίδια μέση τιμή μπορεί να διαφέρουν σημαντικά και κατά χαρακτηριστικό τρόπο ως προς τη διασπορά τους.



Σχήμα 1.4: Σχηματική παρουσίαση δύο δειγμάτων ίσου πλήθους με ίδια μέση τιμή και διαφορετική μεταβλητότητα.

Τα κυριότερα μέτρα διασποράς είναι:

- το **δειγματικό εύρος** (sample range) R,
- η **δειγματική διακύμανση** ή **δειγματική διασπορά** (sample variance) s^2 και η **δειγματική τυπική απόκλιση** (standard deviation) s.
- τα **εκατοστιαία σημεία** (percentiles) και το **ενδοτεταρτομοριακό εύρος** (interquartile range).

Εύρος Όπως αναφέρθηκε παραπάνω το εύρος των δεδομένων $R = x_{\max} - x_{\min}$ είναι η διαφορά της ελάχιστης από τη μέγιστη τιμή του δείγματος. Το εύρος υπολογίζεται εύκολα αλλά δεν είναι ανθεκτικό μέτρο μεταβλητότητας. Εξαρτάται μόνο από τις δύο ακραίες παρατηρήσεις $x_{\min}$ και $x_{\max}$ και αγνοεί τις υπόλοιπες παρατηρήσεις. Γι αυτό μπορεί να αλλάζει σημαντικά από δείγμα σε δείγμα (ίδιου πλήθους κι από τον ίδιο πληθυσμό). Γενικά το εύρος αυξάνει όταν μεγαλώνει το δείγμα καθώς αναμένεται να συμπεριληφθούν πιο ακραίες τιμές.

Διασπορά Η διασπορά ή διακύμανση μετράει τη μεταβλητότητα των παρατηρήσεων γύρω από τη μέση τιμή. Αν ορίσουμε την απόκλιση μιας παρατήρησης x_i από τη μέση τιμή ως $x_i - \bar{x}$, είναι φανερό πως το άθροισμα όλων αυτών των αποκλίσεων είναι 0 γιατί χρησιμοποιώντας τον ορισμό της δειγματικής μέσης τιμής (1.2) έχουμε

$$\sum_{i=1}^{n}(x_i - \bar{x}) = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \bar{x} = n\bar{x} - n\bar{x} = 0.$$

Η δειγματική μέση τιμή $\bar{x}$ έχει οριστεί έτσι ώστε οι θετικές αποκλίσεις για τιμές μεγαλύτερες του $\bar{x}$ να είναι αθροιστικά ίδιες με τις αρνητικές αποκλίσεις για τιμές μικρότερες του $\bar{x}$. Για να μετρήσουμε λοιπόν τη μεταβλητότητα των παρατηρήσεων γύρω από τη μέση τιμή διαλέγουμε να αθροίσουμε όχι τις ίδιες τις αποκλίσεις αλλά τα τετράγωνα των αποκλίσεων. Επίσης για να πάρουμε ένα μέτρο της μέσης απόκλισης που δεν εξαρτάται από το πλήθος των παρατηρήσεων

θα πρέπει να διαιρέσουμε με το πλήθος n των παρατηρήσεων. Όμως για τεχνικούς λόγους που θα εξηγήσουμε παρακάτω διαιρούμε με $n-1$ αντί για n και η δειγματική διασπορά s^2 ορίζεται ως

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2. \tag{1.4}$$

Αναπτύσσοντας τα τετράγωνα της (1.4) έχουμε τον ισοδύναμο τύπο

$$s^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right), \tag{1.5}$$

που είναι πιό εύχρηστος και χρησιμοποιείται στους υπολογισμούς.

Η διασπορά s^2 προκύπτει από τα τετράγωνα των παρατηρήσεων και συχνά είναι δύσκολο να την ερμηνεύσουμε ως πραγματικό φυσικό μέγεθος. Γι αυτό ορίζουμε τη δειγματική τυπική απόκλιση s, που είναι απλά η θετική ρίζα της δειγματικής διασποράς s^2. Η τυπική απόκλιση s μετριέται με τη μονάδα μέτρησης της τ.μ. X κι εκφράζει (όπως δηλώνει η ονομασία της) την τυπική απόκλιση των δεδομένων από τη δειγματική μέση τιμή, δηλαδή μέχρι πόσο περίπου περιμένουμε μια τυπική τιμή της X να απέχει από τη μέση τιμή.

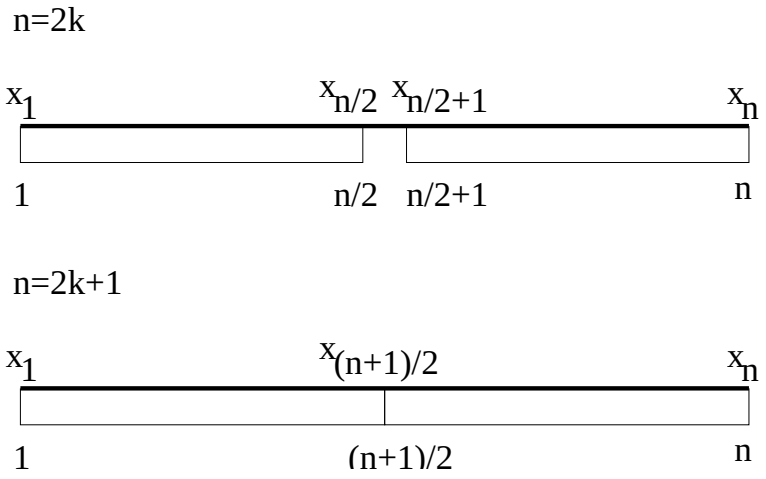
Σημείωση: Χρήση του $n-1$ αντί του n

Όπως η δειγματική μέση τιμή εκτιμά τη μέση τιμή του πληθυσμού μ, έτσι και η δειγματική διασπορά s^2 εκτιμά τη διασπορά του πληθυσμού σ^2. Αν γνωρίζαμε τη μ τότε θα τη χρησιμοποιούσαμε στον τύπο για τον υπολογισμό του s^2, αλλά συνήθως η μ είναι άγνωστη. Οι παρατηρήσεις x_i, τείνουν να είναι πιο κοντά στη $\bar{x}$ παρά στη μ κι άρα οι υπολογισμοί με βάση τις αποκλίσεις $x_i - \bar{x}$ δίνουν μικρότερες τιμές απ' ότι αν χρησιμοποιούσαμε τις αποκλίσεις $x_i - \mu$. Για να αντισταθμίσουμε αυτήν την τάση για υποεκτίμηση της διασποράς του πληθυσμού σ^2 διαιρούμε με $n-1$ αντί με n.

Μία άλλη πιο τεχνική εξήγηση βασίζεται στους *βαθμούς ελευθερίας* (degrees of freedom). Οι n 'ελεύθερες' παρατηρήσεις αποτελούν τους n βαθμούς ελευθερίας. Για τον υπολογισμό της $\bar{x}$ σχηματίζουμε το μέσο όρο διαιρώντας το άθροισμα των παρατηρήσεων με τους βαθμούς ελευθερίας n αφού δεν έχουμε καμιά συνθήκη για τις n παρατηρήσεις που χρησιμοποιούμε. Για τον υπολογισμό της s^2 όμως έχουμε της συνθήκη $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$, δηλαδή αν ξέρουμε $n-1$ από τις αποκλίσεις μπορούμε να βρούμε αυτήν που απομένει. Άρα για τον υπολογισμό της s^2 οι βαθμοί ελευθερίας είναι $n-1$ και γι αυτό διαιρούμε με $n-1$.

Εκατοστιαία σημεία – ενδοτεταρτομοριακό εύρος – θηκόγραμμα Η διάμεσος χωρίζει τα δεδομένα στα δύο. Μπορούμε να ορίσουμε άλλα σημεία χωρισμού του διατεταγμένου συνόλου τιμών που παίρνουμε από το δείγμα. Τέτοια σημεία είναι τα εκατοστιαία σημεία. Μια παρατήρηση καλείται το **p-εκατοστιαίο σημείο** (p–percentile) όταν ποσοστό παρατηρήσεων το πολύ $p\%$ είναι μικρότερες απ' αυτήν την παρατήρηση ($0 \le p < 1$). Η διάμεσος είναι το 50-εκατοστιαίο σημείο. Αλλά χαρακτηριστικά εκατοστιαία σημεία είναι αυτά που ορίζουν τέταρτα ή *τεταρτομόρια* (quartiles). Το 25-εκατοστιαίο σημείο είναι το **πρώτο** ή **κατώτερο τεταρτομόριο** (first or lower quartile) και το συμβολίζουμε Q_1, ενώ το 75-εκατοστιαίο σημείο είναι το **τρίτο** ή **ανώτερο τεταρτομόριο** (third or upper quartile) και το συμβολίζουμε Q_3. Το πρώτο και τρίτο τεταρτομόριο ορίζονται όπως η διάμεσος αλλά περιορίζοντας το σύνολο των δεδομένων στα αντίστοιχα υποσύνολα (κατώτερο ή ανώτερο μισό). Ειδικότερα, έχοντας πρώτα διατάξει τις παρατηρήσεις σε αύξουσα σειρά, αν το σύνολο των παρατηρήσεων n είναι άρτιος αριθμός τότε το κατώτερο υποσύνολο περιέχει τις παρατηρήσεις από 1 ως $n/2$ και το ανώτερο από $n/2+1$

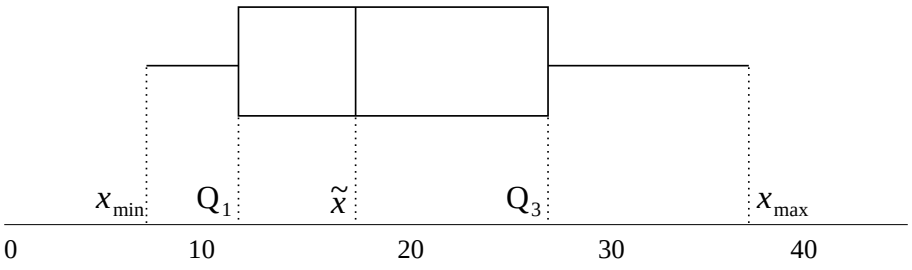
ως n, ενώ αν το n είναι περιττός το κατώτερο υποσύνολο περιέχει τις παρατηρήσεις από 1 ως $(n+1)/2$ και το ανώτερο από $(n+1)/2$ ως n (δες Σχήμα 1.5).



Σχήμα 1.5: Σχηματική παρουσίαση των δύο υποσυνόλων του δείγματος που σχηματίζονται για άρτιο και περιττό πλήθος παρατηρήσεων. Το πρώτο και τρίτο τεταρτομόριο υπολογίζονται ως οι διάμεσοι του πρώτου και δεύτερου υποσυνόλου αντίστοιχα.

Η διαφορά $I = Q_3 - Q_1$ λέγεται **ενδοτεταρτομοριακό εύρος** (interquartile range) και δίνει το εύρος που καλύπτουν τα μισά από τα δεδομένα που είναι πιο κοντά στην κεντρική τιμή (διάμεσο). Το I είναι ένα άλλο μέτρο διασποράς των δεδομένων (πιο ανθεκτικό από το εύρος R) που δεν ορίζεται ως προς τη δειγματική μέση τιμή κι άρα δεν επηρεάζεται απ' αυτήν.

Η διάμεσος $\tilde{x}$, το πρώτο και τρίτο τεταρτομόριο Q_1 και Q_3 αντίστοιχα, καθώς και η ελάχιστη και μέγιστη τιμή των δεδομένων $x_{\min}$ και $x_{\max}$, αποτελούν τη **σύνοψη των 5 αριθμών** (five number summary). Γραφικά η παρουσίαση της σύνοψης των 5 αριθμών γίνεται με το **θηκόγραμμα** (box plot) σε οριζόντια ή κάθετη θέση όπως δείχνει το Σχήμα 1.6.



Σχήμα 1.6: Σχηματική παρουσίαση οριζόντιου θηκογράμματος.

Σημείωση: Θηκόγραμμα στον υπολογιστή

Το θηκόγραμμα που παράγει κάποιο στατιστικό πρόγραμμα, όπως το SPSS, μπορεί να διακόπτει τις γραμμές που ενώνουν τα άκρα του κουτιού με την ελάχιστη και μέγιστη τιμή (που λέγονται *μύστακες* (whiskers)) σε κάποια άλλα σημεία και να παρουσιάζει με ειδικά σύμβολα τις υπόλοιπες μακρινές τιμές. Για το σκοπό αυτό χρησιμοποιείται κάποιο κριτήριο για να χαρακτηριστεί μια μακρινή τιμή x_i σαν *ύποπτη ακραία τιμή* (outlier), ή *ακραία τιμή* (extreme). Το κριτήριο είναι η απόσταση της τιμής από τα άκρα του κουτιού (δηλαδή από το Q_1 ή Q_3). Τυπικά αν η απόσταση αυτή είναι μεγαλύτερη από $1.5I$

(όπου τελειώνει ο μύστακας) τότε η τιμή χαρακτηρίζεται *ύποπτη ακραία*, ενώ αν είναι μεγαλύτερη από $3I$ χαρακτηρίζεται *ακραία*.

Θηκόγραμμα και κανονική κατανομή Το θηκόγραμμα, όπως και το ιστόγραμμα, μας επιτρέπει να κρίνουμε αν μπορούμε να δεχτούμε ότι η κατανομή της συνεχούς τυχαίας μεταβλητής που παρατηρήσαμε είναι κανονική. Για να κάνουμε αυτήν την παραδοχή θα πρέπει:

- η διάμεσος να μην αποκλίνει σημαντικά προς το πρώτο ή το τρίτο τεταρτομόριο, δηλαδή η γραμμή που αντιστοιχεί στη διάμεσο να μην πλησιάζει σε κάποιο από τα δύο άκρα του κουτιού (γιατί αλλιώς αυτό θα σήμαινε πως η κατανομή δεν είναι συμμετρική και δείχνει λοξότητα),

- το εύρος των τιμών στα δύο ακραία τεταρτομόρια να μη διαφέρει σημαντικά, δηλαδή τα μήκη των δύο μυστάκων να είναι συγκρίσιμα (για τη διατήρηση της συμμετρίας).

- να μην υπάρχουν ακραίες τιμές, δηλαδή να μην υπάρχουν σημεία μακριά από τους δύο μύστακες (η ύπαρξη ακραίων σημείων δηλώνει πως οι ουρές της κατανομής είναι 'παχιές ' που δε συμφωνεί με την κανονική κατανομή).

Είναι σημαντικό να αναλογιστούμε ότι με λίγα δεδομένα δεν είναι δυνατόν να τηρούνται αυστηρά οι παραπάνω προϋποθέσεις για το θηκόγραμμα ακόμα κι αν τα δεδομένα προέρχονται πράγματι από κανονική κατανομή. Αν όμως το θηκόγραμμα (όπως και το ιστόγραμμα, ή οποιοδήποτε άλλο γράφημα της κατανομής των δεδομένων) δίνει ενδείξεις σημαντικής απόκλισης από συμμετρική κατανομή τότε δεν θα πρέπει να θεωρήσουμε πως η κατανομή είναι κανονική στην στατιστική ανάλυση που θα κάνουμε στη συνέχεια.

Παράδειγμα 1.3. *Θέλουμε να εκτιμήσουμε την περιεκτικότητα σε ραδιενέργεια του χάλυβα που παράγεται από ένα εργοστάσιο Α. Γι αυτό έγιναν μετρήσεις της ραδιενέργειας (σε Bq/g) σε 10 δοκίμια από το εργοστάσιο Α. Τα αποτελέσματα δίνονται στον Πίνακα 1.5.*

Α/Α	εργοστάσιο Α	εργοστάσιο Β
1	0.40	0.11
2	0.51	0.13
3	0.51	0.26
4	0.54	0.27
5	0.55	0.33
6	0.59	0.37
7	0.63	0.52
8	0.67	0.65
9	0.75	
10	2.10	
Σύνολο	7.25	2.64

Πίνακας 1.5: Δεδομένα περιεκτικότητας ραδιενέργειας (σε μονάδα μέτρησης Bq/g) σε δοκίμια από χάλυβα κατασκευασμένα από δύο εργοστάσια Α και Β.

Θα ασχοληθούμε με τις παρατηρήσεις της περιεκτικότητας ραδιενέργειας στο χάλυβα από το εργοστάσιο Α κι ας ονομάσουμε αυτήν την τυχαία μεταβλητή X. Η δειγματική μέση τιμή υπολογίζεται από το άθροισμα των 10 παρατηρήσεων που δίνονται στη δεύτερη στήλη του Πίνακα 1.5 ως

$$\bar{x} = \frac{1}{10}\sum_{i=1}^{10} x_i = \frac{7.25}{10} = 0.725.$$

Η δειγματική διάμεσος εύκολα μπορεί να βρεθεί μια κι οι παρατηρήσεις δίνονται σε αύξουσα σειρά. Αφού το n είναι άρτιο η διάμεσος δίνεται ως

$$\tilde{x} = \frac{x_{n/2} + x_{n/2+1}}{2} = \frac{x_5 + x_6}{2} = \frac{0.55 + 0.59}{2} = 0.57.$$

Παρατηρούμε ότι η δειγματική μέση τιμή δίνει μεγαλύτερη τιμή στην εκτίμηση της κεντρικής τάσης των δεδομένων από τη δειγματική διάμεσο. Αυτό συμβαίνει γιατί η μέση τιμή επηρεάζεται από την ακραία τιμή της δεκάτης παρατήρησης που είναι πολύ μεγαλύτερη από όλες τις άλλες. Θα πρέπει να γνωρίζουμε αν αυτή η ακραία τιμή είναι πραγματική ή οφείλεται σε κάποιο σφάλμα της παρατήρησης (σφάλμα του μηχανήματος μέτρησης, σφάλμα στην καταγραφή κτλ).

Η μικρότερη περιεκτικότητα ραδιενέργειας στο δείγμα είναι $x_{min} = 0.40\,Bq/g$ κι η μεγαλύτερη είναι $x_{max} = 2.10\,Bq/g$. Άρα το εύρος των δεδομένων είναι $R = 1.70\,Bq/g$, που είναι μεγάλο εξαιτίας της ακραίας τιμής που συμπεριλάβαμε στο δείγμα.

Για να βρούμε τη διασπορά s^2 της περιεκτικότητας της ραδιενέργειας στο χάλυβα στο δείγμα μας υπολογίζουμε πρώτα το άθροισμα των τετραγώνων των παρατηρήσεων

$$\sum_{i=1}^{10} x_i^2 = 0.40^2 + 0.51^2 + \cdots + 0.75^2 + 2.10^2 = 7.44$$

και αντικαθιστώντας το στον τύπο της δειγματικής διασποράς (1.5) βρίσκουμε

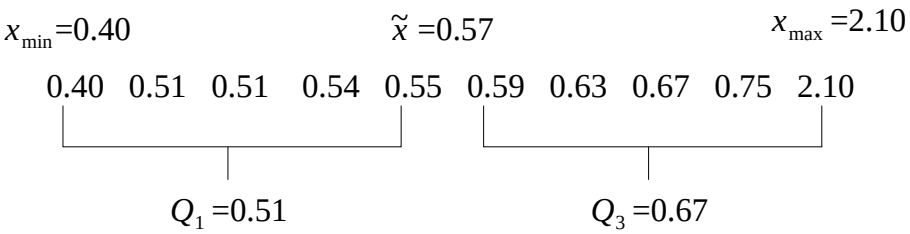
$$s^2 = \frac{1}{9}\left(\sum_{i=1}^{10} x_i^2 - 10\bar{x}^2\right) = \frac{1}{9}\left(7.44 - 10 \cdot 0.725^2\right) = 0.243.$$

Η δειγματική τυπική απόκλιση είναι

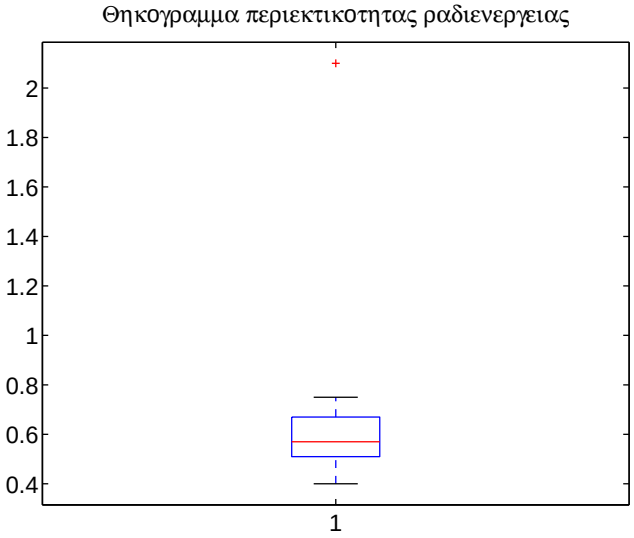
$$s = \sqrt{s^2} = \sqrt{0.243} = 0.493,$$

δηλαδή η αντιπροσωπευτική τυπική απόκλιση από τη μέση περιεκτικότητα ραδιενέργειας είναι περίπου $0.5\,Bq/g$, που είναι πολύ μεγάλη (όση περίπου και η διάμεσος).

Η σύνοψη των 5 αριθμών δίνεται στο Σχήμα 1.7 όπου παρουσιάζεται σχηματικά και η εύρεση του πρώτου και τρίτου τεταρτομορίου. Από το πρώτο και τρίτο τεταρτομόριο υπολογίζουμε το ενδοτεταρτομοριακό εύρος, $I = Q_3 - Q_1 = 0.67 - 0.51 = 0.16\,Bq/g$. Έχοντας βρεί τη σύνοψη των 5 αριθμών μπορούμε εύκολα να παραστήσουμε το θηκόγραμμα. Στο Σχήμα 1.8 δίνεται το θηκόγραμμα σε κατακόρυφη θέση. Παρατηρούμε ότι ο άνω μύστακας δεν προεκτείνεται ως τη μέγιστη τιμή των δεδομένων που είναι 2.10. Αυτή η τιμή δηλώνεται ως ακραία τιμή με ιδιαίτερο σύμβολο, αφού η απόσταση του αντίστοιχου σημείου από το πάνω μέρος του κουτιού, $Q_3 = 0.67$, είναι μεγαλύτερη από $3I = 3 \cdot 0.16 = 0.48$.



Σχήμα 1.7: Σχηματική παρουσίαση της σύνοψης των 5 αριθμών μαζί με τις 10 παρατηρήσεις της περιεκτικότητας ραδιενέργειας.



Σχήμα 1.8: Θηκόγραμμα των 10 παρατηρήσεων περιεκτικότητας ραδιενέργειας σε χάλυβα του εργοστασίου Α.

Παράδειγμα 1.4. *Θεωρούμε πως η ακραία τιμή στο προηγούμενο παράδειγμα οφείλεται πράγματι σε κάποιο σφάλμα της μέτρησης και την απαλείφουμε. Για το νέο δείγμα των 9 παρατηρήσεων υπολογίζουμε τα μέτρα θέσης και μεταβλητότητας.*

Η μέση τιμή είναι

$$\bar{x} = \frac{1}{9}\sum_{i=1}^{9} x_i = \frac{5.15}{9} = 0.572$$

και η διάμεσος (το n είναι περιττό)

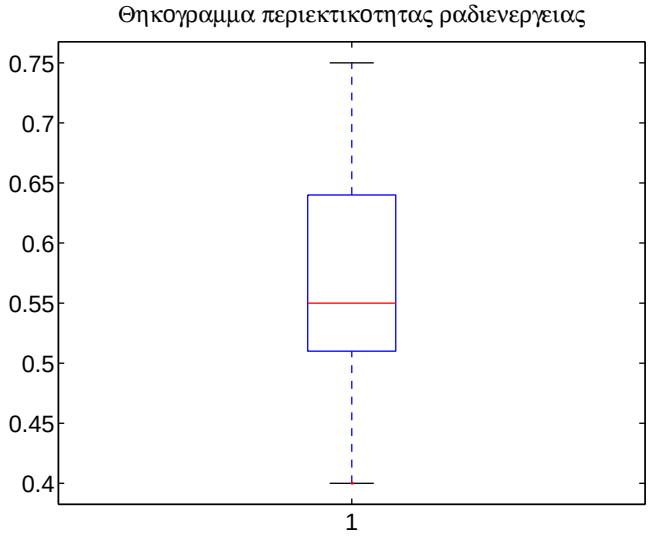
$$\tilde{x} = x_{(n+1)/2} = x_5 = 0.55.$$

Βλέπουμε λοιπόν ότι μετά την απαλοιφή της ακραίας τιμής η δειγματική μέση τιμή και η διάμεσος δίνουν περίπου την ίδια εκτίμηση της κεντρικής τάσης των δεδομένων, που είναι ένδειξη συμμετρίας της κατανομής.

Για τα άλλα μέτρα έχουμε:

Διασπορά	$s^2 = \frac{1}{8}(5.15 - 9 \cdot 0.572^2) = 0.010.$
Τυπική απόκλιση	$s = \sqrt{0.010} = 0.10$
Ελάχιστη τιμή	$x_{min} = 0.40$
Μέγιστη τιμή	$x_{max} = 0.75$
Εύρος	$R = 0.75 - 0.40 = 0.35$
Πρώτο τεταρτομόριο	*(διάμεσος των* $\{x_1, \ldots, x_5\}$*)* $Q_1 = x_3 = 0.51$
Τρίτο τεταρτομόριο	*(διάμεσος των* $\{x_5, \ldots, x_9\}$*)* $Q_3 = x_7 = 0.63$
Ενδοτεταρτομοριακό εύρος	$I = 0.63 - 0.51 = 0.12$

Το αντίστοιχο θηκόγραμμα δίνεται στο Σχήμα 1.9. Από τα μέτρα θέσης και μεταβλητότητας



Σχήμα 1.9: Θηκόγραμμα των 9 παρατηρήσεων περιεκτικότητας ραδιενέργειας σε χάλυβα του εργοστασίου Α.

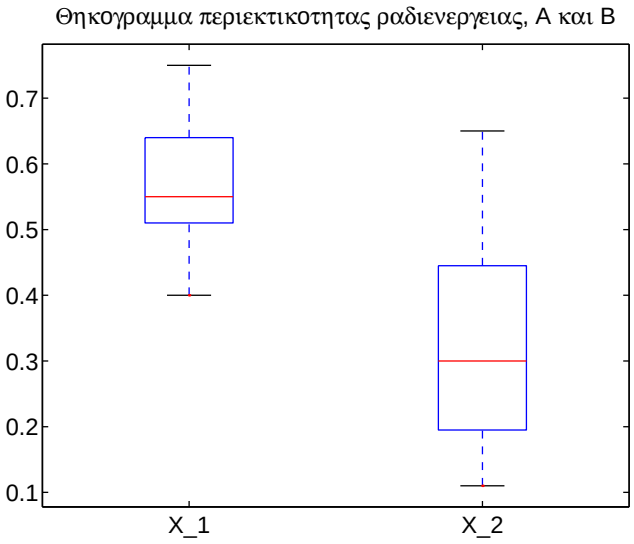
καθώς κι από το θηκόγραμμα παρατηρούμε ότι με την απαλοιφή της ακραίας τιμής η κατανομή της περιεκτικότητας ραδιενέργειας στο χάλυβα του εργοστασίου Α φαίνεται να είναι συμμετρική και μπορούμε να κάνουμε τώρα την παραδοχή ότι η κατανομή είναι κανονική με βάση το δείγμα.

Παράδειγμα 1.5. *Στον Πίνακα 1.5 παρουσιάζονται επίσης οι μετρήσεις της περιεκτικότητας ραδιενέργειας σε 8 δοκίμια χάλυβα που κατασκευάστηκαν από ένα άλλο εργοστάσιο Β. Θέλουμε να συγκρίνουμε την περιεκτικότητα ραδιενέργειας στο χάλυβα από τις δύο μονάδες παραγωγής με βάση τα δείγματα των 9 και 8 παρατηρήσεων που έχουμε από το εργοστάσιο Α και από το εργοστάσιο Β αντίστοιχα. Έστω* X_1 *η τ.μ. της περιεκτικότητας ραδιενέργειας στο χάλυβα για το εργοστάσιο Α και* X_2 *η αντίστοιχη τ.μ. για το εργοστάσιο Β. Στο προηγούμενο παράδειγμα υπολογίσαμε τα μέτρα θέσης και μεταβλητότητας για τη* X_1*. Κάνουμε το ίδιο για τη* X_2*. Τα συγκεντρωτικά αποτελέσματα δίνονται στον Πίνακα 1.6. Επίσης στο Σχήμα 1.10 δίνεται το συνδυασμένο θηκόγραμμα για το δείγμα περιεκτικότητας ραδιενέργειας σε χάλυβα κι από τα δύο εργοστάσια.*

Από τα μέτρα θέσης (μέση τιμή και διάμεσο) φαίνεται ότι η κεντρική τάση της περιεκτικότητας ραδιενέργειας είναι μεγαλύτερη για το χάλυβα του εργοστασίου Α. Από τα μέτρα μεταβλητότητας (τυπική απόκλιση, εύρος δεδομένων και ενδοτεταρτομοριακό εύρος) φαίνεται πως η περιεκτικότητα ραδιενέργειας μεταβάλλεται λιγότερο στο χάλυβα του εργοστασίου Α.

Μέτρο	X_1	X_2
Μέση τιμή	$\bar{x}_1 = 0.572$	$\bar{x}_2 = 0.33$
Διάμεσος	$\tilde{x}_1 = 0.55$	$\tilde{x}_2 = 0.30$
Διασπορά	$s_1^2 = 0.010$	$s_2^2 = 0.034$
Τυπική απόκλιση	$s_1 = 0.10$	$s_2 = 0.18$
Ελάχιστη τιμή	$x_{1,\min} = 0.40$	$x_{2,\min} = 0.11$
Μέγιστη τιμή	$x_{1,\max} = 0.75$	$x_{1,\max} = 0.65$
Εύρος	$R_1 = 0.35$	$R_2 = 0.54$
Πρώτο τεταρτομόριο	$Q_{1,1} = 0.51$	$Q_{2,1} = 0.195$
Τρίτο τεταρτομόριο	$Q_{1,3} = 0.63$	$Q_{2,3} = 0.445$
Ενδοτεταρτομοριακό εύρος	$I_1 = 0.12$	$I_2 = 0.250$

Πίνακας 1.6: Μέτρα θέσης και μεταβλητότητας για τα δεδομένα περιεκτικότητας ραδιενέργειας σε 9 και 8 δοκίμια χάλυβα κατασκευασμένα από τα εργοστάσια Α και Β στη στήλη 2 και 3 αντίστοιχα.



Σχήμα 1.10: Θηκόγραμμα των 9 παρατηρήσεων περιεκτικότητας ραδιενέργειας σε χάλυβα του εργοστασίου Α και των 8 παρατηρήσεων περιεκτικότητας ραδιενέργειας σε χάλυβα του εργοστασίου Β.