ΚΕΦΑΛΑΙΟ 3
ΑΡΙΘΜΗΤΙΚΗ ΠΕΡΙΓΡΑΦΗ ΔΕΔΟΜΕΝΩΝ

Οι γραφικές μέθοδοι είναι εξαιρετικά χρήσιμες στην παρουσίαση δεδομένων και στο να προσφέρουν μια γρήγορη γενική περιγραφή των δεδομένων. Αυτό επιβεβαιώνει με πολλούς τρόπους την γνωστή ρήση ότι "μια εικόνα αξίζει όσο χίλιες λέξεις". Υπάρχουν όμως περιορισμοί στη χρησιμοποίηση των γραφικών τεχνικών για παρουσίαση και ανάλυση δεδομένων. Για παράδειγμα, υπάρχει το ενδεχόμενο να θέλουμε να παρουσιάσουμε τα δεδομένα μας σε μια ομάδα ανθρώπων χωρίς να έχουμε άλλα μέσα περιγραφής τους εκτός από προφορική περιγραφή. Σε τέτοιες περιπτώσεις θα πρέπει να χρησιμοποιήσουμε άλλα περιγραφικά μέτρα που θα μεταφέρουν τα ίδια χαρακτηριστικά στο ακροατήριο. Ενας δεύτερος και όχι τόσο προφανής περιορισμός της τεχνικής των ιστογραμμάτων και των άλλων γραφικών τεχνικών είναι ότι δεν είναι εύκολο να χρησιμοποιηθούν προκειμένου να γίνει με αυτές στατιστική συμπερασματολογία. Κατ'αρχήν βέβαια χρησιμοποιούμε το ιστόγραμμα με τα δεδομένα του δείγματος για να βγάλουμε συμπεράσματα γύρω από τη μορφή και τη θέση του αντίστοιχου ιστογράμματος του πληθυσμού, από τον οποίο έχει προέλθει το δείγμα το οποίο περιγράφει τον άγνωστο σε εμάς πληθυσμό. Η συμπερασματολογία βασίζεται στην υπόθεση ότι, σε κάποιο βαθμό, υπάρχει ομοιότητα μεταξύ των δύο ιστογραμμάτων. Εχουμε όμως τη δυσκολία με τα ιστογράμματα να μετρήσουμε το βαθμό αυτής της ομοιότητας. Βέβαια ξέρουμε πότε δύο σχήματα ταυτίζονται αλλά κάτι τέτοιο σπάνια θα συμβεί στην πράξη. Επομένως, εάν το δειγματικό ιστόγραμμα και το ιστόγραμμα του πληθυσμού διαφέρουν, γεννάται το ερώτημα με τι τρόπο θα μπορέσουμε να μετρήσουμε το βαθμό της διαφοράς τους, ή, αν θέλουμε να το δούμε από μια πιο θετική σκοπιά, το βαθμό της ομοιότητάς τους;

Οι περιορισμοί των γραφικών μεθόδων περιγραφής δεδομένων μπορούν να αρθούν με τη χρησιμοποίηση **αριθμητικών περιγραφικών μέτρων (numerical descriptive measures).** Υπενθυμίζουμε δύο ορισμούς που έχουμε δει και προηγουμένως. Αριθμητικά περιγραφικά μέτρα που αναφέρονται στον πληθυσμό ονομάζονται **'παράμετροι (parameters).** Τα αντίστοιχα αριθμητικά περιγραφικά μέτρα που αναφέρονται σε ένα δείγμα υπολογίζονται ως τιμές **στατιστικών συναρτήσεων (statistics).** Επομένως, ενδιαφερόμαστε να χρησιμοποιήσουμε τα δειγματικά δεδομένα για να υπολογίσουμε ένα σύνολο τιμών στατιστικών συναρτήσεων που θα μας δώσουν μια καλή θεωρητική εικόνα της δειγματικής κατανομής σχετικής συχνότητας και τα οποία θα είναι χρήσιμα για να βοηθηθούμε στη στατιστική συμπερασματολογία που αναφέρεται στην κατανομή της σχετικής συχνότητας του πληθυσμού.

Δύο είναι, κυρίως, τα χαρακτηριστικά ενός δείγματος που μπορούν να μας δώσουν μια καλή συνοπτική εικόνα για το δείγμα. Το ένα είναι κάποια τιμή γύρω από την οποία τα δεδομένα τείνουν να συσσωρεύονται. Μέτρα που αναφέρονται στον καθορισμό μιας τέτοιας τιμής ονομάζονται **μέτρα θέσης (measures of location)** ή **μέτρα κεντρικής τάσης (measures of central tendency).**

Μια άλλη έννοια αναφέρεται στη μεταβλητότητα των δεδομένων. Στον καθορισμό δηλαδή της διασποράς των δεδομένων γύρω από κάποιο μέτρο αριθμητικής θέσης. Η περιγραφή των εννοιών αυτών θα γίνει στη συνέχεια.

3.1 Μέτρα Θέσης ή Κεντρικής Τάσης (Measures of Location or Central Tendency)

Τα κυριότερα μέτρα κεντρικής τάσης είναι ο **αριθμητικός μέσος (arithmetic mean),** η **διάμεσος (median)** και η **επικρατούσα τιμή (mode).**

Αριθμητικός Μέσος (arithmetic mean)

Ο **αριθμητικός μέσος (arithmetic mean)** που συνήθως αναφέρεται απλά ως μέσος, είναι ο μέσος των δεδομένων. Αναφερόμαστε στον δειγματικό μέσο όταν υπολογίζουμε το μέσο ενός δείγματος και στην μέση τιμή του πληθυσμού όταν αναφερόμαστε στο γενικό μέσο του πληθυσμού. Συνήθως συμβολίζουμε με n τον αριθμό των παρατηρήσεων ενός δείγματος και με N τον αριθμό των μονάδων ενός πληθυσμού. Με $\overline{X}$ συμβολίζουμε τον μέσο των παρατηρήσεων $X_1, X_2, ..., X_n$ ενώ με μ συμβολίζουμε τον μέσο των τιμών $x_1, x_2, ..., x_N$ των N μονάδων του πληθυσμού.

Προφανώς ισχύει ότι

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

Ο αριθμητικός μέσος υπολογίζεται εύκολα στο Minitab με την εντολή MEAN.

Για παράδειγμα, έστω ότι με C1, C2 και C3 είχαμε συμβολίσει τις στήλες των δεδομένων που αναφέρονται στις βαθμολογίες Στατιστικής των τριών ομάδων της ενότητας 2.2.

Ομάδα 1 : 66, 74, 78, 78, 80, 81, 85, 88, 89, 97

Ομάδα 2 : 65, 66, 68, 71, 77, 78, 80, 82, 88

Ομάδα 3 : 48, 61, 62, 71, 75, 82, 84, 91, 99

Τότε με την εντολή

MEAN C1

θα παίρναμε από το Minitab την απάντηση

MEAN = 81.600

Επίσης, με τις εντολές MEAN C2 και MEAN C3, αντίστοιχα, θά

παίρναμε τις απαντήσεις

MEAN = 75.000
MEAN = 74.778

Παραλλαγές του Αριθμητικού Μέσου

Ως **α-περικομμένο μέσο (α-trimmed mean)** ορίζουμε τον μέσο των παρατηρήσεων που απομένουν αφού παραλειφθεί το ανώτερο 100α% και το κατώτερο 100α% των διατεταγμένων παρατηρήσεων. (Το Minitab δίνει την τιμή του 0.05 - περικομμένου μέσου (που ονομάζει απλά **περικομμένο μέσο (trimmed mean)**). Αν αντί να περικόψουμε το άνω και κάτω 25% των διατεταγμένων παρατηρήσεων αντικαταστήσουμε κάθε μια από τις παρατηρήσεις του άνω τετάρτου με το άνω τεταρτημόριο και κάθε μια από τις παρατηρήσεις του κάτω τετάρτου με το κάτω τεταρτημόριο, ο μέσος των παρατηρήσεων που προκύπτουν λέγεται **Winsorised μέσος (Winsorised mean)**. Τόσο ο περικομμένος μέσος όσο και ο Winsorised μέσος χρησιμοποιούνται όταν θέλουμε να αποφύγουμε την επίδραση των ακραίων τιμών.

Διάμεσος (median)

Η **διάμεσος (median)** είναι η τιμή που χωρίζει ένα σύνολο δεδομένων περίπου στη μέση όταν τα δεδομένα αυτά τοποθετηθούν με σειρά τάξης μεγέθους. Συγκεκριμένα η διάμεσος ορίζεται ως εξής:

Ορισμός: Η διάμεσος (median) ενός συνόλου μετρήσεων είναι η τιμή εκείνη με την ιδιότητα ότι το πολύ 50% των μετρήσεων είναι μικρότερες από την τιμή αυτή και το πολύ 50% των μετρήσεων είναι μεγαλύτερες από την τιμή αυτή.

Για παράδειγμα, αν παρατηρήσουμε τους βαθμούς της δεύτερης

ομάδας, οι οποίοι έχουν ήδη διαταχθεί κατά σειρά μεγέθους, βλέπουμε ότι

Ομάδα 2 : 65, 66, 68, 71, 77, 78, 80, 82, 88

Η διάμεσος για τα δεδομένα αυτά είναι ο βαθμός 77, δοθέντος ότι υπάρχουν τέσσερις βαθμοί μικρότεροι από αυτόν και τέσσερις βαθμοί μεγαλύτεροι από αυτόν.

Για δεδομένα με άρτιο αριθμό παρατηρήσεων, ως διάμεσο θεωρούμε το μέσο των δύο τιμών που είναι πλησιέστερα στο κέντρο των διατεταγμένων δεδομένων. Ετσι, για παράδειγμα, για τους βαθμούς της πρώτης ομάδας

Ομάδα 1 : 66, 74, 78, 78, 80, 81, 85, 88, 89, 97

διάμεσος είναι η τιμή 80.5.

Στο Minitab η σχετική εντολή είναι MEDIAN.

Για παράδειγμα, για την τρίτη ομάδα δεδομένων (C3) η σχετική εντολή είναι

MEDIAN C3

και η απάντηση που παίρνουμε από το Minitab είναι

MEDIAN = 75.000

Η διάμεσος είναι το πιο κατάλληλο μέτρο κεντρικής τάσης όταν τα δεδομένα είναι διατεταγμένα ποσοτικά δεδομένα και όχι ποιοτικά.

Επικρατούσα τιμή (mode)

Ως επικρατούσα τιμή (mode) χαρακτηρίζουμε την τιμή εκείνη των δεδομένων που έχει τη μεγαλύτερη συχνότητα εμφάνισης.

Για παράδειγμα, για τη σειρά δεδομένων

1, 3, 4, 3, 4, 5, 4, 2

η επικρατούσα τιμή είναι το 4 (εμφανίζεται τρεις φορές).

Οταν υπάρχουν δύο τιμές οι οποίες έχουν την ίδια συχνότητα εμφάνισης, τότε λέμε ότι τα δεδομένα αυτά έχουν δύο επικρατούσες τιμές . Η κατανομή των δεδομένων που έχουν μια μόνο επικρατούσα τιμή λέγεται **μονοκόρυφη (unimodal)** ενώ εάν έχει δύο επικρατούσες

τιμές λέγεται **δικόρυφη (bimodal)**. Η επικρατούσα τιμή δεν είναι υποχρεωτικό να ταυτίζεται με τον μέσο των διατεταγμένων παρατηρήσεων ενός δείγματος παρ'ότι αυτό συμβαίνει πολύ συχνά.

Η επικρατούσα τιμή χρησιμοποιείται όταν τα δεδομένα που έχουμε είναι ποιοτικά. (Στην περίπτωση αυτή ούτε ο μέσος ούτε η διάμεσος έχουν έννοια).

Γεωμετρικός Μέσος (Geometric Mean)

Ενα άλλο μέτρο θέσης (κεντρικής τάσης) που χρησιμοποιείται πολύ λιγότερο από αυτά που ήδη αναφέρθηκαν είναι ο *γεωμετρικός μέσος (geometric mean).*

Ο γεωμετρικός μέσος ορίζεται από τον τύπο:

$$\text{Γεωμετρικός μέσος} = \sqrt[n]{X_1 X_2 \cdots X_n}$$

όπου n είναι ο αριθμός των παρατηρήσεων που έχουν ληφθεί για τη μεταβλητή X και $X_1, X_2, \ldots, X_n$ είναι οι τιμές των παρατηρήσεων αυτών. Για παράδειγμα, αν έχουμε τις παρατηρήσεις 3, 25 και 45, ο γεωμετρικός μέσος είναι:

$$\text{Γεωμετρικός μέσος} = \sqrt[3]{3 \times 25 \times 45} = \sqrt[3]{3375} = 15.$$

Είναι προφανές ότι ο γεωμετρικός μέσος δεν ορίζεται αν στο δείγμα περιλαμβάνονται παρατηρήσεις με αρνητικές τιμές ή παρατηρήσεις που έχουν την τιμή μηδέν. Μπορεί επίσης να αποδειχθεί ότι ο γεωμετρικός μέσος μιας σειράς παρατηρήσεων είναι πάντοτε μικρότερος από τον αριθμητικό μέσο των παρατηρήσεων αυτών (εκτός και αν οι παρατηρήσεις ταυτίζονται).

Ενα σημαντικό χαρακτηριστικό του γεωμετρικού μέσου σε σύγκριση με τον αριθμητικό μέσο είναι ότι ο γεωμετρικός μέσος επηρεάζεται λιγότερο από παρατηρήσεις με πολύ μεγάλες τιμές. Αν

θεωρήσουμε, για παράδειγμα, το ημερομίσθιο που λαμβάνει μια ομάδα επτά εργαζομένων το οποίο είναι (σε χιλιάδες δραχμές)

6 8 10 10 10 12 16

βρίσκουμε εύκολα ότι ο στρογγυλεμένος αριθμητικός μέσος είναι 10.286 δρχ.

Ο γεωμετρικός μέσος, αντίστοιχα, είναι

$$\sqrt[7]{6\times8\times10\times10\times10\times12\times16} = \sqrt[7]{9216000} = 9884 .$$

Ας υποθέσουμε ότι ο ιδιοκτήτης της επιχείρησης, θεωρώντας ότι ο υψηλότερα αμειβόμενος εργαζόμενος έχει συνεισφέρει σημαντικά στην αύξηση των εργασιών της επιχείρησης, αποφασίζει να αυξήσει το ημερομίσθιό του στο ποσό των 48 χιλ. δρχ.. Τότε θα έχει το εξής σύνολο παρατηρήσεων:

6 8 10 10 10 12 48.

Ο αριθμητικός μέσος για τις τιμές αυτές είναι 14.857 ενώ, αντίστοιχα, ο γεωμετρικός μέσος είναι 11.564.

Είναι προφανές ότι ο τριπλασιασμός του ημερομισθίου του υψηλότερα αμειβόμενου εργαζόμενου επηρέασε σε πολύ μικρότερο βαθμό τον γεωμετρικό μέσο από ότι τον αριθμητικό μέσο. Αυτό γιατί ο γεωμετρικός μέσος μεταβλήθηκε μόνο κατά ένα παράγοντα της τάξεως της $\sqrt[7]{3}$.

Ο γεωμετρικός μέσος είναι χρήσιμος όταν μεταβάλλονται μερικές μόνο από τις παρατηρήσεις μιας ακολουθίας παρατηρήσεων. Στην περίπτωση αυτή, όπως είδαμε και στο παράδειγμά μας, ο γεωμετρικός μέσος είναι πολύ περισσότερο σταθερός από ότι ο αριθμητικός μέσος. Για το λόγο αυτό χρησιμοποιείται και στον υπολογισμό του δείκτη συνήθων βιομηχανικών μετοχών των Financial Times. (Financial Times Industrial Ordinary Share Index).

Ο γεωμετρικός μέσος είναι επίσης χρήσιμος για την κατασκευή

εκτιμήσεων από δεδομένα που αυξάνονται ή ελαττώνονται σύμφωνα με μια γεωμετρική πρόοδο. Οι πληθυσμοί, για παράδειγμα, αυξάνονται με τέτοιο τρόπο. Η αύξηση, δηλαδή, σε ένα πληθυσμό είναι ανάλογη του μεγέθους του πληθυσμού σε οποιαδήποτε χρονική στιγμή και όχι ανάλογη του μεγέθους του πληθυσμού στην αρχή της χρονικής περιόδου μελέτης (όπως θα συνέβαινε εάν η αύξηση ήταν αριθμητική και όχι γεωμετρική). Για παράδειγμα, αν ξέραμε ότι ο πληθυσμός μιας πόλης το 1960 ήταν 270000 και το 1970 ήταν 510000 και θέλαμε να εκτιμήσουμε το μέγεθος του πληθυσμού το 1965 θα μπορούσαμε, χρησιμοποιώντας τον αριθμητικό μέσο, να ισχυρισθούμε ότι το μέγεθος αυτό ήταν

$$\frac{270000 + 510000}{2} = 390000.$$

Κάτι τέτοιο όμως θα ήταν λογικό αν ο πληθυσμός αυξανόταν κατά σταθερό αριθμό κάθε χρόνο. Παρ'όλα αυτά, είναι γνωστό ότι όσο περισσότεροι άνθρωποι κατοικούν σε μια πόλη, τόσο μεγαλύτερος είναι ο αριθμός των ατόμων που προστίθεται στον πληθυσμό αυτό κάθε χρόνο. Είναι, επομένως, περισσότερο ρεαλιστικό να υποθέσουμε ότι υπάρχει μία γεωμετρική αύξηση στον πληθυσμό αυτό. Να υποθέσουμε, δηλαδή, ότι ο πληθυσμός αυξάνει με σταθερό ρυθμό κάθε χρόνο. Ο γεωμετρικός μέσος, για το παράδειγμά μας, θα είναι

$$\sqrt{270000 \times 510000} = 363730.$$

Από την τελευταία παρατήρηση προκύπτει ότι η χρήση του γεωμετρικού μέσου αποτελεί μία κατάλληλη τεχνική όταν κανείς θέλει να κάνει εκτιμήσεις για οποιαδήποτε σύνολα στοιχείων που αυξάνονται με τον τρόπο αυτό, όπως, για παράδειγμα, το συνολικό ποσόν χρημάτων που επενδύεται με σύνθετο τόκο (compound interest).

Αρμονικός μέσος (Harmonic Mean)

Ο *αρμονικός μέσος (harmonic mean)* για ένα δείγμα n

παρατηρήσεων $X_1, X_2, ..., X_n$ ορίζεται από τη σχέση:

$$\text{Αρμονικός Μέσος} = \frac{n}{\sum_{i=1}^{n} \left(\frac{1}{X_i}\right)}$$

Ο αρμονικός μέσος χρησιμοποιείται συνήθως ως ένας μέσος για ρυθμούς.

Αν π.χ. ένα αυτοκίνητο κινείται στην διαδρομή προς μία ορισμένη πόλη με ταχύτητα 60km/h και στο ταξίδι της επιστροφής κινείται με ταχύτητα 40km/h, η μέση ταχύτητα δεν μπορεί να θεωρηθεί ότι είναι 50km/h όπως, ενδεχομένως, θα μπορούσε να θεωρήσει κανείς με μια πρώτη αντιμετώπιση. Αυτή θα ήταν η μέση ταχύτητα αν το αυτοκίνητο είχε κινηθεί για μια ώρα με ταχύτητα 60km/h και στη συνέχεια για μια ακόμα ώρα με ταχύτητα 40km/h. Προκειμένου να καθορίσουμε τη μέση ταχύτητα όταν έχουν χρησιμοποιηθεί διαφορετικές ταχύτητες πάνω στην ίδια απόσταση χρησιμοποιείται ο αρμονικός μέσος.

Για το συγκεκριμένο παράδειγμα ταχυτήτων 60km/h και 40km/h η μέση ταχύτητα, όπως αυτή εκφράζεται από τον αρμονικό μέσο, είναι

$$\frac{2}{\left(\frac{1}{60} + \frac{1}{40}\right)} = \frac{2}{\frac{5}{120}} = 48\text{km/h}$$

Η αρχή αυτή μπορεί να εφαρμοσθεί και σε άλλα παράμοια προβλήματα. Εστω, για παράδειγμα, ότι μια δακτυλογράφος σε ένα γραφείο μπορεί να δακτυλογραφήσει τιμολόγια με ρυθμό 30 την ώρα, καταστάσεις με ρυθμό 40 την ώρα και σημειώματα υπενθύμισης με ρυθμό 80 την ώρα. Εστω επίσης ότι θα πρέπει να δακτυλογραφηθούν ίσοι αριθμοί από κάθε κατηγορία των εγγράφων αυτών. Ο διευθυντής του γραφείου θα πρέπει να αναφέρει τη μέση παραγωγικότητα ανά ώρα της δακτυλογράφου. Ο μέσος αριθμητικός θα είναι (30+40+80)/3 = 50 κείμενα την ώρα.

Ο μέσος αριθμητικός όμως δεν λαμβάνει υπόψη του το γεγονός

ότι τα τιμολόγια χρειάζονται περισσότερο χρόνο να δακτυλογραφηθούν από ότι οι καταστάσεις και οι καταστάσεις χρειάζονται περισσότερο χρόνο από ότι τα σημειώματα υπενθύμισης. Ο πραγματικός μέσος ρυθμός μπορεί να βρεθεί με τη θεώρηση του αρμονικού μέσου:

$$\frac{3}{\left(\frac{1}{30} + \frac{1}{40} + \frac{1}{80}\right)} = 42 \text{ (σε προσέγγιση ακεραίου).}$$

3.2 Σύγκριση των Ιδιοτήτων Μέσου, Διαμέσου και Επικρατούσας τιμής

Δοθέντος ότι και τα τρία κύρια μέτρα κεντρικής τάσης που ορίσαμε (αριθμητικός μέσος, διάμεσος και επικρατούσα τιμή) έχουν τον ίδιο στόχο, αναρωτιέται κανείς ποιο από τα τρία πρέπει να χρησιμοποιεί και για ποιο λόγο. Προκειμένου να αποφασίσει ο χρήστης ποιο μέτρο να χρησιμοποιήσει θα πρέπει να λάβει υπόψη του τέσσερις παράγοντες.

α) Την ευαισθησία των μέτρων αυτών στην παρουσία ακραίων τιμών (outliers) στο δείγμα.

β) Την ευαισθησία τους στο σχήμα της κατανομής.

γ) Τη χρησιμότητά τους για συμπερασματολογία σε σχέση με τα αντίστοιχα μέτρα του πληθυσμού.

δ) Την αντίστοιχη θεωρητική ανάπτυξη.

Για τις περισσότερες περιπτώσεις η σύγκριση γίνεται μεταξύ του μέσου και της διαμέσου γιατί η επικρατούσα τιμή χρησιμοποιείται σχεδόν αποκλειστικά για τις περιπτώσεις ποιοτικών δεδομένων.

α) Η παρουσία ακραίων τιμών σε ένα δείγμα είναι δυνατόν να επηρεάσει σημαντικά τη συνοπτική παρουσίασή του και να οδηγήσει σε λανθασμένα συμπεράσματα εάν δεν επιλεγεί το κατάλληλο μέτρο.

Ας θεωρήσουμε, για παράδειγμα τέσσερις οικογένειες των οποίων

το μηνιαίο εισόδημα έχει τις τιμές (σε χιλιάδες δραχμές)

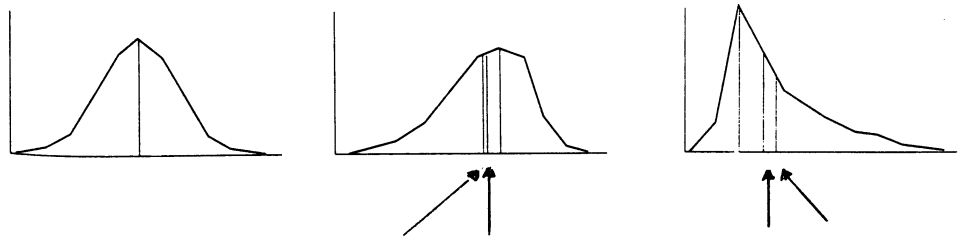
120 , 140 , 150 , 110 , 500 .

Οπως παρατηρούμε, τα δεδομένα αυτά έχουν μια ακραία τιμή, την τιμή 500. Ας θεωρήσουμε τον αριθμητικό μέσο.Παρατηρούμε ότι $\overline{X}$ = 200. Αν, επομένως θεωρήσουμε τη μέση τιμή ως ένα μέτρο θέσης των τιμών του πληθυσμού και αντίστοιχα τον μέσο του δείγματος ως ένα μέτρο θέσης των παρατηρήσεων, θα οδηγηθούμε στο συμπέρασμα ότι το μέσο εισόδημα των οικογενειών αυτών είναι 200 χιλ. δραχμές. Κάτι τέτοιο όμως, προφανώς, δίνει μια λανθασμένη εικόνα υψηλού εισοδήματος για τις οικογένειες αυτές. Είναι προφανές ότι καταλήξαμε στη μέση αυτή τιμή γιατί στο δείγμα υπάρχει μια ακραία τιμή η οποία και "παρέσυρε" το μέσο σε υψηλότερη τιμή.Αντίθετα, η διάμεσος (που έχει την τιμή 140) χαρακτηρίζει πολύ καλύτερα τα δεδομένα αυτά. Εξ ίσου κατάλληλα μέτρα θα ήταν ο περικομμένος μέσος και ο Winsorised μέσος.

β) Οσο αφορά την εικόνα της κατανομής, είναι προφανές ότι και εδώ οι ακραίες τιμές παίζουν σημαντικό ρόλο. Αν δεν υπάρχουν τέτοιες ακραίες τιμές, η κατανομή τους είναι περίπου συμμετρική.

Αν υπάρχουν μερικές ακραίες τιμές στα δεξιά του κύριου όγκου των τιμών ενός δείγματος, τότε λέμε ότι η κατανομή των δεδομένων έχει **δεξιά ασυμμετρία (στρεβλότητα** ή **λοξότητα)** ή ότι έχει **δεξιά κλίση (skewed to the right** ή **positively skewed)**. Αντίθετα, εάν υπάρχουν μερικές ακραίες παρατηρήσεις στα αριστερά του κύριου όγκου των παρατηρήσεων θα λέμε ότι η κατανομή είναι **ασύμμετρη (στρεβλή** ή **λοξή) προς τα αριστερά** ή ότι έχει **αριστερή κλίση (skewed to the left** ή **negatively skewed)**.

Οταν μια κατανομή είναι απόλυτα συμμετρική, η μέση τιμή και η διάμεσος συμπίπτουν. Οταν η κατανομή έχει δεξιά ασυμμετρία (δεξιά στρεβλότητα) η μέση τιμή είναι δεξιότερα από τη διάμεσο, ενώ όταν η κατανομή έχει αριστερή στρεβλότητα η μέση τιμή είναι αριστερότερα από τη διάμεσο.



Σχήμα 3.2.1

Εν γένει η διάμεσος είναι προτιμότερη ως μέτρο θέσης από την μέση τιμή για κατανομές δεδομένων που είναι στρεβλές με μία μόνο κορυφή (επικρατούσα τιμή).

γ) Οταν θέλουμε να χρησιμοποιήσουμε το μέτρο θέσης για να βγάλουμε κάποιο συμπέρασμα για το σύνολο των παρατηρήσεων ενός πληθυσμού, η μέση τιμή είναι το μόνο κατάλληλο τέτοιο μέτρο. Αυτό γιατί όταν ξέρουμε την μέση τιμή και τον αριθμό των μονάδων του πληθυσμού μπορούμε, πολλαπλασιάζοντας τους δύο αυτούς αριθμούς, να εκτιμήσουμε το άθροισμα των παρατηρήσεων του πληθυσμού. Κάτι τέτοιο, προφανώς, δεν είναι δυνατόν να γίνει με τη διάμεσο.

δ) Η θεωρητική ανάπτυξη της στατιστικής συμπερασματολογίας που αναφέρεται και χρησιμοποιεί τον μέσο ενός δείγματος είναι πολύ μεγαλύτερη από την αντίστοιχη που χρησιμοποιεί τη διάμεσο. Αυτό οφείλεται στο ότι τα μαθηματικά που μπορεί και πρέπει να χρησιμοποιήσει κανείς για κάτι τέτοιο είναι πολύ πιο εύχρηστα για

τον μέσο από ότι για τη διάμεσο. Παρ'όλα αυτά θα πρέπει να επισημάνει κανείς ότι με τη μεγάλη πρόοδο των υπολογιστών και τις μεθόδους που χρησιμοποιούνται με τη χρήση των υπολογιστών έχει εμφανισθεί τελευταία μια σημαντική εξέλιξη και ενδιαφέρον σε συμπερασματολογία που βασίζεται στην έννοια της διαμέσου.

Συνοπτικά, όσον αφορά τη σύγκριση του μέσου και της διαμέσου, μπορούμε να πούμε τα εξής:

Σύγκριση Μέσου και Διαμέσου

Μέσος	**Διάμεσος**
- Ευαίσθητος στην επίδραση ακραίων τιμών, ειδικά σε μικρά σύνολα δεδομένων	- Οχι ευαίσθητη στην επίδραση ακραίων τιμών
- Λιγότερο αντιπροσωπευτικός ως "τυπική τιμή" για στρεβλές κατανομές με μία μόνο επικρατούσα τιμή	- Περισσότερο αντιπροσωπευτική ως "τυπική τιμή" για στρεβλές κατανομές με μία μόνο επικρατούσα τιμή
- Χρήσιμος για συμπερασματολογία που αναφέρεται στο άθροισμα των τιμών του πληθυσμού	- Οχι χρήσιμη για συμπερασματολογία που αναφέρεται στο άθροισμα των τιμών του πληθυσμού
- Ευκολότερος για να εργασθούμε με αυτόν θεωρητικά	- Δύσκολο να εργασθούμε με αυτήν θεωρητικά.

3.3 Μέτρα Μεταβλητότητας ή Διασποράς (Measures of Variation or Dispersion)

Εχουμε ήδη τονίσει στα προηγούμενα κεφάλαια την σημασία της μελέτης της μεταβλητότητας σε μια σειρά δεδομένων. Οπως είναι φυσικό, μεταβλητότητα υπάρχει σε κάθε σύνολο δεδομένων αφού, αντίστοιχα, υπάρχει μεταβλητότητα σε κάθε πληθυσμό ή διαδικασία. Με απλά λόγια, θα μπορούσε να πει κανείς ότι *μεταβλητότητα είναι το "άπλωμα" ή η διασπορά των τιμών σε ένα σύνολο δεδομένων.* Είναι επομένως σημαντικό να μετρήσουμε τη διασπορά αυτή, να την κατανοήσουμε και να καθορίσουμε τις αιτίες που την προκαλούν έτσι ώστε να έχουμε τη δυνατότητα να πάρουμε τις κατάλληλες αποφάσεις.

Εχουμε ήδη μελετήσει κάποια μέτρα θέσης των οποίων οι τιμές συνοψίζουν μία σειρά δεδομένων. Οπως όμως είναι φυσικό από τη στιγμή που ορίζουμε ένα μέτρο θέσης για μία σειρά μετρήσεων, τίθεται στη συνέχεια το ερώτημα: πόσο αντιπροσωπευτικό είναι αυτό το μέτρο θέσης σε σχέση με όλες τις μετρήσεις του συνόλου των δεδομένων; Με άλλα λόγια, πώς "απλώνονται" οι μετρήσεις (οι παρατηρήσεις) γύρω από αυτό το μέτρο θέσης; Είναι οι μετρήσεις πολύ απλωμένες γύρω από το μέτρο θέσης ή συγκεντρώνονται κοντά σε αυτό;

Για να αντιληφθούμε καλύτερα την ιδέα της μεταβλητότητας, ας θεωρήσουμε την καθολική έλλειψη μεταβλητότητας. Για παράδειγμα, έστω ότι κάθε φοιτητής που παρακολούθησε ένα μάθημα Στατιστικής πέτυχε τον ίδιο ακριβώς βαθμό στο τελικό διαγώνισμα. Αν μας δοθεί η πληροφορία αυτή, χρειαζόμαστε οποιαδήποτε άλλη πληροφορία για να συνοψίσουμε το σύνολο των βαθμών της τάξης; Η απάντηση είναι, προφανώς, όχι, δεδομένου ότι το να συνοψίσει κανείς ένα σύνολο δεδομένων χρησιμοποιώντας την έννοια της κατανομής τους και κάποιου μέτρου θέσης δεν έχει καμιά έννοια εάν δεν υπάρχει μεταβλητότητα στο συγκεκριμένο σύνολο δεδομένων. Επομένως, για να κατανοήσει κανείς τη μεταβλητότητα και να την αντιμετωπίσει με σωστό τρόπο θα πρέπει πρώτα να την μετρήσει.

Στην ενότητα αυτή εξετάζουμε μία σειρά από τα ποιο σημαντικά μέτρα μεταβλητότητας που χρησιμοποιούνται στην πράξη. Συγκεκριμένα, θα παρουσιάσουμε τις έννοιες της *έκτασης (εύρους)*, της *μέσης απόλυτης απόκλισης*, της *διακύμανσης (διασποράς)* και της *τυπικής απόκλισης*.

Εκταση ή Εύρος (Range)

Ως **έκταση** ή **εύρος (range)** ορίζουμε τη διαφορά μεταξύ της μεγαλύτερης και της μικρότερης τιμής ενός συνόλου δεδομένων.

Δηλαδή

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

όπου X_{max} η μέγιστη τιμή του συνόλου των δεδομένων και X_{min} η ελάχιστη τιμή του συνόλου αυτού.

Το σημαντικότερο πλεονέκτημα της έννοιας της έκτασης είναι η απλότητά της και η ευκολία στον υπολογισμό της. Το μεγαλύτερο μειονέκτημά της είναι ότι εξαρτάται από δύο μόνο τιμές του συνόλου των παρατηρήσεων. Είναι δηλαδή απόλυτα εξαρτημένη (και επομένως ευαίσθητη) από τις δύο ακραίες τιμές των παρατηρήσεων χωρίς να λαβαίνει καθόλου υπόψη της τις άλλες τιμές.

Για παράδειγμα, η έκταση του συνόλου των τιμών 1,2,3,7,12 είναι

$$12 - 1 = 11.$$

Η έκταση του συνόλου των δεδομένων

$$1,\ 1,\ 1,\ 12,\ 12$$

είναι επίσης 11.

Είναι όμως προφανές ότι το δεύτερο σύνολο δεδομένων εμφανίζει πολύ μεγαλύτερη συνολική μεταβλητότητα από ότι το πρώτο.

Το χαρακτηριστικό αυτό είναι ένα μεγάλο μειονέκτημα όταν χρησιμοποιείται η έκταση για τη μέτρηση της μεταβλητότητας σε μεγάλα σύνολα δεδομένων. Η έκταση ήταν ιδιαίτερα χρήσιμη τις παλιότερες εποχές όταν δεν γινόταν μεγάλη χρήση υπολογιστών, οπότε

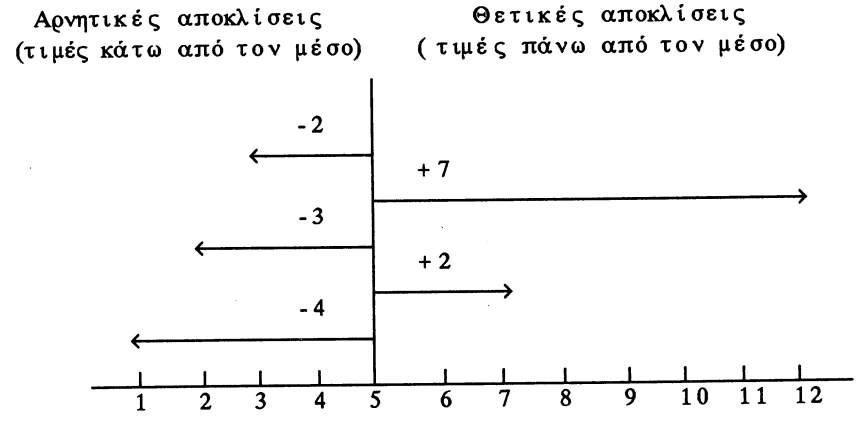
η απλότητά της ως μέτρου μεταβλητότητας την έκανε εξαιρετικά χρήσιμη. Σήμερα χρησιμοποιείται πολύ λιγότερο σε στατιστικές εφαρμογές.

Μέση Απόλυτη Απόκλιση (Mean Absolute Deviation)

Στην ενότητα αυτή θα ορίσουμε ένα μέτρο απόκλισης που είναι ευαίσθητο σε όλες τις τιμές ενός συνόλου δεδομένων, λόγω του ότι χρησιμοποιεί όλα τα δεδομένα. Το μέτρο αυτό, όπως και άλλα παρόμοια μέτρα που θα ορίσουμε στη συνέχεια, *θεωρούν τη μεταβλητότητα ως το βαθμό στον οποίο οι τιμές ενός συνόλου δεδομένων αποκλίνουν από τον μέσο τους.*

Μια προφανής αρχική σκέψη είναι να παρατηρήσει κανείς την απόκλιση κάθε σημείου των δεδομένων από τον μέσο τους και να προσδιορίσει τη μέση απόκλιση. Ας θεωρήσουμε, για παράδειγμα, το σύνολο των δεδομένων 1, 2, 3, 7, 12 των οποίων ο μέσος είναι 5. Οι αποκλίσεις από τον μέσο των τιμών είναι, αντίστοιχα, -4, -3, -2, +2 και +7.

Οι αποκλίσεις αυτές εμφανίζονται στο σχήμα 3.3.1 που ακολουθεί.



Σχήμα 3.3.1
Οι αποκλίσεις από τον μέσο πέντε παρατηρήσεων

Ο μέσος αυτών των αποκλίσεων είναι 0, δοθέντος ότι το άθροισμα των αποκλίσεων είναι 0. Το τελευταίο βέβαια ισχύει πάντα. Δηλαδή **το άθροισμα των αποκλίσεων των τιμών ενός συνόλου παρατηρήσεων από τον μέσο τους είναι πάντοτε 0**, για οποιοδήποτε σύνολο δεδομένων, δοθέντος ότι οι θετικές αποκλίσεις αντισταθμίζονται πάντοτε από τις αρνητικές αποκλίσεις (αφού ο μέσος είναι το κέντρο βάρους των παρατηρήσεων). Είναι επομένως προφανές ότι, αφού η μέση απόκλιση από τον μέσο είναι πάντοτε 0, η χρησιμοποίησή της ως μέτρου μεταβλητότητας δεν έχει έννοια.

Προκειμένου να ξεπεράσουμε το πρόβλημα αυτό, θεωρούμε την *απόσταση κάθε παρατήρησης από τον μέσο*, δηλαδή την απόκλισή της από τον μέσο αγνοώντας το κατά πόσον η παρατήρηση είναι μεγαλύτερη ή μικρότερη από αυτόν. Με τον τρόπο αυτό έχουμε πάντοτε μη αρνητικές τιμές, αφού οι αποστάσεις μιας παρατήρησης από τον μέσο είναι πάντοτε μη αρνητικές. Εν γένει, υπάρχουν δύο θεωρήσεις της έννοιας της απόστασης μεταξύ δύο τιμών. Η πρώτη βασίζεται στην απόλυτη τιμή της διαφοράς των τιμών και η δεύτερη στο τετράγωνο της διαφοράς τους. Η υιοθέτηση της πρώτης από αυτές οδηγεί στον ορισμό της έννοιας της *μέσης απόλυτης απόκλισης* ενώ η υιοθέτηση της δεύτερης από αυτές οδηγεί στον ορισμό της έννοιας της *διακύμανσης (διασποράς)* που θα δούμε στη συνέχεια.

Ορισμός: Ως **μέση απόλυτη απόκλιση (mean absolute deviation)** ορίζουμε τον μέσο των απολύτων τιμών των αποκλίσεων των παρατηρήσεων από τον μέσο των παρατηρήσεων αυτών.

Σημείωση: Από τα αρχικά των αγγλικών λέξεων του όρου χρησιμοποιείται, για συντομογραφία, ο συμβολισμός MAD για την μέση απόλυτη απόκλιση.

Επομένως, έχουμε

$$MAD = \frac{\sum_{i=1}^{n} |X_i - \overline{X}|}{n}$$

όπου $|X_i - \overline{X}|$ συμβολίζει την απόλυτη τιμή της διαφοράς μεταξύ της τιμής X_i των δεδομένων και του μέσου των δεδομένων αυτών $\overline{X}$.

Για παράδειγμα, η μέση απόλυτη απόκλιση του συνόλου των δεδομένων 1, 2, 3, 7, 12 που έχει μέσο 5 είναι

$$MAD = \frac{|1-5| + |2-5| + |3-5| + |7-5| + |12-5|}{5} = 3.6 .$$

Η μέση απόλυτη απόκλιση είναι ένα χρήσιμο μέτρο μέτρησης της μεταβλητότητας. Οπως είπαμε ήδη, εξαρτάται από όλες τις τιμές των δεδομένων και είναι εύκολο να ερμηνευθεί. Ενα μειονέκτημα είναι ότι υπάρχει κάποια δυσκολία ως προς τη δυνατότητα ανάπτυξης στατιστικών μεθόδων με βάση το μέτρο αυτό δεδομένου ότι είναι μία συνάρτηση απολύτων τιμών. Αυτό έχει σαν αποτέλεσμα να μη χρησιμοποιείται συχνά σε πρακτικές εφαρμογές. Αυτό οδήγησε στην αξιοποίηση της άλλης έννοιας απόστασης, δηλαδή της τετραγωνικής απόστασης και στον ορισμό της διακύμανσης.

Διακύμανση ή Διασπορά (Variance)

Ορισμός: Ορίζουμε ως **διακύμανση ή διασπορά (variance)** ενός πληθυσμού N τιμών $x_1, x_2, ..., x_N$ με μέση τιμή μ την μέση τετραγωνική απόκλιση των n μετρήσεων από τη μέση τιμή μ του πληθυσμού. Για τη διακύμανση χρησιμοποιείται ο συμβολισμός σ^2.

Επομένως έχουμε

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N} .$$

Για παράδειγμα, έστω ότι έχουμε τον πληθυσμό 11, 12, 13, 17, 22 με μέση τιμή 15. Τότε θα έχουμε ότι

$$\sigma^2 = \frac{(11-15)^2+(12-15)^2+(13-15)^2+(17-15)^2+(22-15)^2}{5} = 16.4 .$$

Η διακύμανση ως μέτρο μεταβλητότητας έχει ένα μεγάλο πλεονέκτημα σε σχέση με τη μέση απόλυτη απόκλιση: είναι περισσότερο εύχρηστη όσο αφορά τις μαθηματικές πράξεις. Για το λόγο αυτό χρησιμοποιείται ως το κυριότερο μέτρο της μεταβλητότητας στην ανάπτυξη της στατιστικής θεωρίας και θεωρείται ως μια παράμετρος μεγάλου ενδιαφέροντος.

Δεδομένου ότι στη Στατιστική σπανίως γνωρίζουμε τις τιμές για το σύνολο των στοιχείων ενός πληθυσμού, χρησιμοποιούμε δειγματικά δεδομένα με βάση τα οποία ορίζουμε δειγματικές στατιστικές συναρτήσεις για να εκτιμήσουμε τις τιμές των παραμέτρων του πληθυσμού. Προκειμένου, επομένως, να εκτιμήσουμε την τιμή της διακύμανσης ενός πληθυσμού χρειάζεται να ορίσουμε ένα αντίστοιχο μέτρο δειγματικής διακύμανσης.

Οσο αφορά την θεωρητική κατασκευή ενός μέτρου της δειγματικής διακύμανσης, αυτή γίνεται με τρόπο αντίστοιχο αυτού που ακολουθήθηκε για την κατασκευή του μέτρου της διακύμανσης του πληθυσμού (στην οποία αναφερθήκαμε προηγουμένως) με αντικατάσταση της άγνωστης μέσης τιμής μ του πληθυσμού με τον δειγματικό μέσο $\overline{X}$ και του μεγέθους N του πληθυσμού με το μέγεθος n του δείγματος. Η δειγματική διακύμανση συμβολίζεται με S^2.

Επομένως, για τη δειγματική διακύμανση θα έχουμε ότι

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n} .$$

Εχει παρατηρηθεί, όμως, ότι η χρησιμοποίηση του παραπάνω τύπου, ο οποίος είναι το ακριβές ανάλογο του αντίστοιχου τύπου για

τη διακύμανση του πληθυσμού, τείνει να οδηγήσει σε μια εκτιμήτρια του σ^2 η οποία είναι "χαμηλότερη" από την πραγματική του τιμή. Με στατιστική ορολογία δηλαδή μπορούμε να πούμε ότι αν κανείς θεωρήσει πολλά τυχαία δείγματα μεγέθους n, η εκτιμήτρια S^2 του σ^2, όπως αυτή ορίσθηκε παραπάνω, οδηγεί τις περισσότερες φορές σε τιμές μικρότερες από την πραγματική τιμή του σ^2 από ότι σε τιμές μεγαλύτερες του σ^2. Η κατάσταση αυτή ονομάζεται *μεροληπτικότητα (bias).*

Η θεωρία, αλλά και η πράξη, έχουν αποδείξει ότι αν χρησιμοποιηθεί ως διαιρέτης στον τύπο του S^2 το n-1 αντί του n η εκτιμήτρια στην οποία καταλήγουμε, και την οποία συμβολίζουμε με S^{*2} είναι *αμερόληπτη* εκτιμήτρια του σ^2. Δηλαδή, αν θέλουμε να κατασκευάσουμε μια αμερόληπτη εκτιμήτρια του σ^2 θα πρέπει να θεωρήσουμε τη στατιστική συνάρτηση

$$S^{*2} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}.$$

Τυπική Απόκλιση (Standard Devation)

Η αριθμητική τιμή της διακύμανσης ενός πληθυσμού, ή ενός δείγματος, είναι δύσκολο να ερμηνευθεί δεδομένου ότι εκφράζεται σε τετράγωνα των μονάδων των παρατηρήσεων. Για να αποφευχθεί το πρόβλημα αυτό ορίστηκε ένα μέτρο μεταβλητότητας που να εκφράζεται στις ίδιες μονάδες όπως τα αρχικά δεδομένα.

Ορισμός: Ορίζουμε ως **τυπική απόκλιση (standard deviation)** την θετική τετραγωνική ρίζα της διακύμανσης.

Θα έχουμε, δηλαδή, όσο αφορά τον πληθυσμό

$$\sigma = +\sqrt{\sigma^2}$$

και, οσο αφορά το δείγμα

$$S = +\sqrt{S^2} \quad \text{και} \quad S^* = +\sqrt{S^{*2}}.$$

Είναι προφανές ότι ούτε η διακύμανση ούτε η τυπική απόκλιση μπορούν να έχουν αρνητικές τιμές. Η διακύμανση δεν είναι δυνατόν να πάρει αρνητική τιμή γιατί είναι ένας μέσος τετραγωνικών ποσοτήτων. Αντίστοιχα, η τυπική απόκλιση δεν μπορεί να πάρει αρνητική τιμή γιατί από τον ορισμό της είναι η θετική τετραγωνική ρίζα της διακύμανσης.

Υπολογισμός της δειγματικής διακύμανσης και της δειγματικής τυπικής απόκλισης

Ο υπολογισμός της δειγματικής διακύμανσης και της δειγματικής τυπικής απόκλισης γίνεται είτε με υπολογιστή είτε με ηλεκτρονική μηχανή (calculator). Για τις περιπτώσεις όμως που είναι κανείς υποχρεωμένος να κάνει κάποιους υπολογισμούς με το χέρι, θα ήταν χρήσιμο να δώσουμε έναν εναλλακτικό τύπο που διευκολύνει τους υπολογισμούς αυτούς.

Αν συμβολίσουμε με

$$SST = \sum_{i=1}^{n}(X_i - \bar{X})^2$$

το συνολικό άθροισμα τετραγώνων των αποκλίσεων των παρατηρήσεων από τον μέσο (total sum of squares), θα έχουμε

$$S^2 = \frac{SST}{n} \quad \text{και} \quad S^{*2} = \frac{SST}{n-1}.$$

Είναι εύκολο να αποδείξουμε ότι

$$SST = \sum_{i=1}^{n} X_i^2 - \frac{\left(\sum_{i=1}^{n} X_i\right)^2}{n}.$$

Επομένως, οι εναλλακτικοί ισοδύναμοι τύποι για τον υπολογισμό της μεροληπτικής και της αμερόληπτης δειγματικής διακύμανσης είναι

$$S^2 = \frac{1}{n}\left[\sum_{i=1}^{n} X_i^2 - \frac{\left(\sum_{i=1}^{n} X_i\right)^2}{n}\right]$$

και

$$S^{*2} = \frac{1}{n-1}\left[\sum_{i=1}^{n} X_i^2 - \frac{\left(\sum_{i=1}^{n} X_i\right)^2}{n}\right]$$

αντίστοιχα.

Σημείωση: Η έννοια του συνολικού αθροίσματος τετραγώνων των αποκλίσεων είναι σημαντική γιατί χρησιμοποιείται σε πολλές στατιστικές μεθόδους όπως π.χ. στην Ανάλυση Παλινδρόμησης και στην Ανάλυση Διακύμανσης. Το συνολικό άθροισμα τετραγώνων αναφέρεται στην *συνολική* μεταβλητότητα μεταξύ των τιμών ενός συνόλου δεδομένων (ενώ η διακύμανση μετρά την *μέση* απόκλιση). Για παράδειγμα, αν όλες οι τιμές σε ένα δείγμα ήταν ακριβώς οι ίδιες δεν θα υπήρχε μεταβλητότητα και το SST θα ήταν ίσο με το 0. Οσο μεγαλύτερη είναι η τιμή του SST τόσο μεγαλύτερη είναι η μεταβλητότητα μεταξύ των τιμών σε ένα σύνολο δεδομένων.

Παράδειγμα: Τα αμοιβαία κεφάλαια (mutual funds) έχουν γίνει τελευταία μια δημοφιλής μορφή επένδυσης για μικρούς επενδυτές. Προκειμένου να βοηθηθούν οι επενδυτές για να αποφασίσουν σε ποιο συγκεκριμένο από τα αμοιβαία κεφάλαια θα επενδύσουν, πολλές οικονομικές εφημερίδες στον κόσμο δίνουν πληροφορίες για τον μέσο ετήσιο ρυθμό απόδοσης (average annual rate of return) για μία σειρά αμοιβαίων κεφαλαίων και για μια σειρά ετών στο παρελθόν. (Σημείωση: Ο ετήσιος ρυθμός απόδοσης ενός αμοιβαίου κεφαλαίου δίνεται από την σχέση $(P_1-P_0)/P_0$ όπου P_0 και P_1 είναι οι τιμές των μετοχών του κεφαλαίου στην αρχή και στο τέλος του χρόνου, αντίστοιχα. Ο ορισμός αυτός ισχύει κάτω από την υπόθεση ότι δεν πληρώνονται μερίσματα (dividents) από το κεφάλαιο κατά τη διάρκεια του χρόνου.)

Εστω ότι ο ετήσιος (ποσοστιαίος) ρυθμός απόδοσης για τα τελευταία δέκα χρόνια δύο αμοιβαίων κεφαλαίων δίνεται από τα στοιχεία που ακολουθούν.

Κεφ. Α: 8.3, -6.2, 20.9, -2.7, 33.6, 42.9, 24.4, 5.2, 3.1, 30.5
Κεφ. Β: 12.1, -2.8, 6.4, 12.2, 27.8, 25.3, 18.2, 10.7, -1.3, 11.4

Ποιο από τα δύο αμοιβαία κεφάλαια θα θεωρούσατε ότι περικλείει τον μεγαλύτερο βαθμό επενδυτικού κινδύνου;

Λύση :

Για κάθε αμοιβαίο κεφάλαιο θα πρέπει να υπολογίσουμε τη διακύμανση ή την τυπική απόκλιση του δείγματος των ρυθμών απόδοσης.

Για το κεφάλαιο Α έχουμε

$$\text{A:} \quad \sum_{i=1}^{10} x_i = 160 \quad , \quad \sum_{i=1}^{n} x_i^2 = 5083.06$$

$$s_A^{*2} = \frac{1}{9}\left[5083.06 - \frac{160^2}{10} \right] = 280.34$$

$$s_A^{*} = 16.74.$$

Για το κεφάλαιο Β έχουμε

$$\text{B:} \quad \sum_{i=1}^{10} x_i = 120 \quad , \quad \sum_{i=1}^{n} x_i^2 = 2334.36$$

$$s_B^{*2} = \frac{1}{9}\left[2334.36 - \frac{120^2}{10} \right] = 99.37$$

$$s_B^{*} = 9.97.$$

Επομένως, οι τιμές που υπολογίσθηκαν παρέχουν μια ένδειξη ότι το αμοιβαίο κεφάλαιο Α έχει το μεγαλύτερο επίπεδο επικινδυνότητας, όπως αυτό μετριέται με τη διακύμανση, δοθέντος ότι η διακύμανση

των ρυθμών απόδοσης του κεφαλαίου αυτού είναι μεγαλύτερη από τη διακύμανση των αντίστοιχων ρυθμών του κεφαλαίου Β.

Μπορούμε επίσης να παρατηρήσουμε ότι το κεφάλαιο Α έχει υψηλότερο μέσο ρυθμό απόδοσης για τα τελευταία χρόνια από ότι το κεφάλαιο Β δοθέντος ότι

$$\bar{x}_A = \frac{160}{10} = 16 \qquad \text{και} \qquad \bar{x}_B = \frac{120}{10} = 12.$$

Επομένως ο μέσος ετήσιος ρυθμός απόδοσης είναι 16% για το κεφάλαιο Α και 12% για το κεφάλαιο Β.

Το αποτέλεσμα αυτό συμβαδίζει με τη διαισθητική αντίληψη ότι μία επένδυση που έχει υψηλότερο κίνδυνο αποδίδει ένα μεγαλύτερο ρυθμό απόδοσης. (Οι φοιτητές των Οικονομικών θα αντιλαμβάνονται ότι ο επενδυτής θα πρέπει να επενδύει σε μια ποικιλία αμοιβαίων κεφαλαίων προκειμένου να ισχύει η προηγούμενη παρατήρηση όταν ως μέτρο επικινδυνότητας χρησιμοποιείται η διακύμανση).

Βαθμοί Ελευθερίας (Degrees of Freedom)

Οπως είδαμε, για τον υπολογισμό της τυπικής απόκλισης που προέρχεται από την αμερόληπτη εκτιμήτρια S^{*2} της διακύμανσης χρησιμοποιείται ο διαιρέτης n-1. Ο αριθμός αυτός αναφέρεται συνήθως στην βιβλιογραφία ως **βαθμοί ελευθερίας (degrees of freedom)**. Η ονομασία οφείλεται στο ότι εάν πρόκειται να διαλέξουμε n τιμές που θα πρέπει να έχουν ένα δεδομένο μέσο, αριθμός των τιμών που μπορούμε να διαλέξουμε ελεύθερα και αυθαίρετα είναι n-1. Αυτό γιατί όταν ο μέσος n τιμών είναι καθορισμένος, και το άθροισμά τους είναι καθορισμένο. Τότε όμως, μόνο n-1 από αυτές τις τιμές μπορούν να είναι αυθαίρετες ενώ η τελευταία τιμή δεν μπορεί να είναι αυθαίρετη. (Αυτή θα προκύπτει ως η διαφορά του αθροίσματος των n τιμών μείον το άθροισμα των n-1 τιμών που έχουν ήδη επιλεγεί αυθαίρετα). Δοθέντος ότι η δειγματική διακύμανση υπολογίζεται μέσω των αποκλίσεων των n τιμών των δεδομένων από τον μέσο τους, λέμε ότι έχει n-1 βαθμούς ελευθερίας.

Σχέση μέσης απόλυτης απόκλισης και τυπικής απόκλισης

Για οποιοδήποτε σύνολο δεδομένων η μέση απόλυτη απόκλιση είναι πάντοτε μικρότερη από την τυπική απόκλιση διότι είναι λιγότερο ευαίσθητη στην επίδραση ακραίων παρατηρήσεων. (Αυτό οφείλεται στο γεγονός ότι για τον καθορισμό της διακύμανσης θεωρούμε το τετράγωνο των αποκλίσεων). Επομένως, όταν ένα σύνολο δεδομένων περιέχει μερικές ακραίες παρατηρήσεις, η μέση απόλυτη απόκλιση ίσως αποτελεί ένα πιο ρεαλιστικό μέτρο της μεταβλητότητας από ότι η τυπική απόκλιση. Οπως έχουμε όμως ήδη επισημάνει, η τυπική απόκλιση χρησιμοποιείται πολύ περισσότερο στις στατιστικές εφαρμογές λόγω των μαθηματικών ιδιοτήτων που έχει οι οποίες την κάνουν πιο χρήσιμη για τη θεωρητική στατιστική ανάπτυξη.

Ερμηνεία και χρήση της τυπικής απόκλισης

Οπως ήδη παρατηρήσαμε, η τυπική απόκλιση προτιμάται, εν γένει, της διακύμανσης ως ένα περιγραφικό μέτρο της μεταβλητότητας κυρίως διότι εκφράζεται στις ίδιες μονάδες μέτρησης όπως τα δεδομένα από τα οποία έχει προέλθει.

Ας εξετάσουμε τώρα πώς μπορούμε να ερμηνεύσουμε την τυπική απόκλιση ως ένα περιγραφικό μέτρο της μεταβλητότητας.

Ας υποθέσουμε ότι για τις αμοιβές μιας ομάδας εργαζομένων έχει υπολογισθεί ότι το μέσο ημερομίσθιο είναι 10000δρχ. και η τυπική απόκλιση 3000δρχ.. Το ερώτημα που τίθεται είναι πώς μπορεί κανείς να εξηγήσει σε κάποιο εργοδότη που δεν έχει γνώσεις Στατιστικής ποιες ενδείξεις προκύπτουν από τα στοιχεία αυτά όσο αφορά τη μεταβλητότητα μεταξύ των απολαβών των εργαζομένων. Μία πιθανή απάντηση είναι δυνατόν να στηρίζεται στον καθορισμό του ποσοστού των τιμών των δεδομένων που μπορεί κανείς να περιμένει ότι θα περιλαμβάνονται σε ένα διάστημα που έχει ως άκρα την απόσταση μιας τυπικής απόκλισης από τον μέσο (ή δύο τυπικών

αποκλίσεων ή τριων τυπικών αποκλίσεων κ.λ.π.). Στο παράδειγμά μας δηλαδή, ποιό είναι το ποσοστό των εργαζομένων που μπορεί κανείς να περιμένει ότι θα έχουν απολαβές μεταξύ 7000δρχ. και 13000δρχ. (όπου 7000δρχ. είναι ο μέσος μείον μία τυπική απόκλιση και 13000δρχ. είναι ο μέσος συν μία τυπική απόκλιση). Η απάντηση στο ερώτημα αυτό δόθηκε από ένα Ρώσσο μαθηματικό, τον Pavroly Chebyshev. (Σε πολλά αγγλικά βιβλία η γραφή του ονόματος του μαθηματικού αυτού με λατινικούς χαρακτήρες είναι Tchebysheff. Πολλοί επίσης ονομάζουν το αποτέλεσμα αυτό *κανόνα των Bienaymè-Tchebysheff*).

Θεώρημα Chebyshev: Δοθέντος ενός συνόλου τιμών (μετρήσεων) και ενός αριθμού $k \geq 1$, το ποσοστό των τιμών (μετρήσεων) αυτών που περιλαμβάνεται σε απόσταση k τυπικών αποκλίσεων από τον αντίστοιχο μέσο τους είναι **τουλάχιστον** $1-(1/k^2)$.

Από το αποτέλεσμα αυτό έχουμε, για τις διάφορες τιμές του k, τα εξής ποσοστά παρατηρήσεων σε αντίστοιχες αποστάσεις από τον μέσο:

Θεώρημα του Chebyshev για διάφορες τιμές του k

k	Διάστημα	Ποσοστό μετρήσεων στο διάστημα αυτό	
1	$(\mu-\sigma, \mu+\sigma)$	τουλάχιστον 0	(≥ 0%)
2	$(\mu-2\sigma, \mu+2\sigma)$	τουλάχιστον 3/4	(≥ 75%)
2.5	$(\mu-2.5\sigma, \mu+2.5\sigma)$	τουλάχιστον 21/25	(≥ 84%)
3	$(\mu-3\sigma, \mu+3\sigma)$	τουλάχιστον 8/9	(≥ 89%)

Το σημαντικό χαρακτηριστικό του θεωρήματος του Chebyshev είναι ότι μπορεί να εφαρμοσθεί για οποιοδήποτε σύνολο παρατηρήσεων ανεξάρτητα από την κατανομή που οι παρατηρήσεις αυτές έχουν. Γιαυτό όμως το λόγο οδηγεί και σε εξαιρετικά συντηρητικά όρια.

Αυτό βέβαια αποτελεί και ένα σημαντικό περιορισμό για τη χρησιμότητά του. Ετσι, για τα περισσότερα σύνολα δεδομένων μπορούμε να περιμένουμε ένα πολύ μεγαλύτερο ποσοστό παρατηρήσεων σε απόσταση μιας, δύο ή τριών τυπικών αποκλίσεων από τον μέσο από αυτό που δίνει η εφαρμογή του θεωρήματος.

Σημείωση: Οπως προκύπτει από τη διατύπωση του θεωρήματος του Chebyshev, το θεώρημα αυτό έχει διατυπωθεί για το σύνολο των τιμών ενός πληθυσμού. Για το λόγο αυτό άλλωστε χρησιμοποιούνται η μέση τιμή μ και η τυπική απόκλιση σ του πληθυσμού. Για την απόδειξη του θεωρήματος μπορεί να ανατρέξει κανείς σε ένα βιβλίο πιθανοτήτων (βλέπε π.χ. *Πιθανότητες και Στοιχεία Στοχαστικών Ανελίξεων* των συγγραφέων). Από την απόδειξη του θεωρήματος προκύπτει ότι το συμπέρασμά του ισχύει και για τις μετρήσεις, ή παρατηρήσεις, ενός τυχαίου δείγματος. Στην περίπτωση αυτή, η μέση τιμή μ και η τυπική απόκλιση σ του πληθυσμού θα αντικατασταθούν από τη μέση τιμή $\overline{X}$ και την τυπική απόκλιση S του δείγματος αντίστοιχα. Το θεώρημα επίσης ισχύει και όταν χρησιμοποιηθεί ως τυπική απόκλιση το S^*, η τετραγωνική, δηλαδή, ρίζα της αμερόληπτης εκτιμήτρια της διακύμανσης. Αυτό γιατί το S^{*2} είναι πάντοτε μεγαλύτερο από το αντίστοιχο S^2. (Δοθέντος ότι $n>n-1$). Σε κάθε περίπτωση εξάλλου υπάρχει μικρή διαφορά μεταξύ των δύο διαφορετικών υπολογισμών της απόκλισης, ιδίως όταν πρόκειται για μεγάλο δείγμα.

Παράδειγμα: Τα δεδομένα για το δείγμα της διάρκειας τηλεφωνημάτων από το τηλεφωνικό κέντρο ενός Πανεπιστημίου που δώσαμε νωρίτερα έχουν μέσο $\bar{x} = 10.26$ και τυπική απόκλιση $s^* = 4.29$. Εάν δεν είχαμε καμιά άλλη πληροφορία σε σχέση με την κατανομή της διάρκειας των τηλεφωνημάτων, από το θεώρημα του Chebyshev προκύπτει ότι τουλάχιστον τα 3/4, δηλαδή τουλάχιστον το 75% των τηλεφωνημάτων, έχουν διάρκεια που βρίσκεται στο διάστημα $(\bar{x}-2s^*, \bar{x}+2s^*) = (1.68, 18.84)$.

Παρ'όλα αυτά, παρατηρώντας τα δεδομένα που είναι διαθέσιμα βλέπουμε ότι όλες, εκτός από την μεγαλύτερη από τις 30 παρατηρήσεις, βρίσκονται μέσα στο διάστημα αυτό. Δηλαδή το διάστημα αυτό στην πραγματικότητα περιέχει το 96.7% της διάρκειας των τηλεφωνημάτων, ένα ποσοστό που είναι πολύ μεγαλύτερο από αυτό που δίνει το θεώρημα του Chebyshev. Αυτό οφείλεται στην μορφή της κατανομής των δεδομένων.

Οταν έχουμε κάποια καλύτερη γνώση της μορφής της κατανομής των παρατηρήσεων (ή του πληθυσμού) τότε, όπως προκύπτει και από το προηγούμενο παράδειγμα, τα αντίστοιχα ποσοστά των παρατηρήσεων που περιλαμβάνονται στα διαστήματα που κατασκευάζονται με βάση την απόσταση σε τυπικές αποκλίσεις από τον μέσο είναι πολύ μεγαλύτερα από αυτά που δίνει το θεώρημα του Chebyshev. Ειδικότερα, όταν τα δεδομένα έχουν μια συμμετρική κατανομή με μία μόνο κορυφή, που συνήθως ταιριάζει με μια πάρα πολύ γνωστή κατανομή στη Στατιστική που ονομάζεται **κανονική κατανομή (normal distribution)**, είναι δυνατόν να προσδιορίσουμε ακριβέστερα τα σχετικά ποσοστά των παρατηρήσεων που περιλαμβάνονται σε απόσταση μιας, δυο ή τριων τυπικών αποκλίσεων από τον μέσο των παρατηρήσεων. (Με τη μελέτη και τις ιδιότητες της κανονικής κατανομής θα ασχοληθούμε αργότερα).

Η εμπειρία έχει δείξει ότι τα ποσοστά αυτά των παρατηρήσεων (που θα δώσουμε στον πίνακα που ακολουθεί) είναι κατά προσέγγιση ακριβή όχι μόνο για δεδομένα που ακολουθούν την κανονική κατανομή αλλά ακόμα και για δεδομένα που έχουν κατανομή που προσεγγίζει την κανονική. Ισχύουν δηλαδή και για δεδομένα των οποίων η κατανομή έχει μία μόνο κορυφή και είναι συμμετρική ή έχει μέτρια ασυμμετρία. Δοθέντος ότι τα συμπεράσματα αυτά βασίζονται σε εμπειρικές διαπιστώσεις που έχουν προκύψει από μελέτες πολλών συνόλων δεδομένων, ο κανόνας που ακολουθεί ονομάζεται **εμπειρικός κανόνας (empirical rule)**. Ο εμπειρικός αυτός κανόνας για την κανονική κατανομή περιγράφεται στον πίνακα που ακολουθεί.

Ο εμπειρικός κανόνας

Για σύνολα δεδομένων με κατανομή συχνότητας που προσεγγίζει την κανονική κατανομή ή περιέχουν πολλά στοιχεία, έχουμε ότι:

- *το 68% των παρατηρήσεων βρίσκονται στο διάστημα που τα άκρα του απέχουν μια τυπική απόκλιση από τον μέσο.*
- *το 95% των παρατηρήσεων βρίσκονται στο διάστημα που τα άκρα του απέχουν δύο τυπικές αποκλίσεις από τον μέσο.*
- *το 99% των παρατηρήσεων βρίσκονται στο διάστημα που τα άκρα του απέχουν τρεις τυπικές αποκλίσεις από τον μέσο.*

Οταν τα δεδομένα ακολουθούν μια κατανομή που είναι έντονα μη συμμετρική ή έχει περισσότερες από δύο κορυφές, ο εμπειρικός κανόνας δεν θα πρέπει να εφαρμόζεται. Στο σημείο αυτό πρέπει να κάνουμε την εξής παρατήρηση. Εάν καταλήξουμε σε μια τυπική απόκλιση που είναι εξαιρετικά μεγάλη σε σχέση με τον μέσο των δεδομένων από τα οποία τα στοιχεία έχουν προέλθει, αυτό μπορεί να θεωρηθεί ως μία ένδειξη σημαντικής ασυμμετρίας ή παρουσίας ακραίων τιμών. Παρ'όλα αυτά μπορεί να λεχθεί εν γένει ότι ο εμπειρικός κανόνας είναι σωστός για πολύ μεγάλο αριθμό περιπτώσεων συνόλου δεδομένων.

Ως μια τελευταία παρατήρηση σε σχέση με τον εμπειρικό κανόνα, μπορούμε να πούμε ότι αυτός μπορεί να χρησιμοποιηθεί για μια πρώτη προσέγγιση της τυπικής απόκλισης ενός συνόλου παρατηρήσεων που έχουν κατανομή συχνότητας σχεδόν συμμετρική με μια μόνο κορυφή. Δοθέντος ότι οι περισσότερες από τις μετρήσεις του δείγματος (περίπου το 95%) βρίσκεται στο διάστημα που έχει άκρα δύο τυπικές αποκλίσεις από τον μέσο, η έκταση (το εύρος) των μετρήσεων είναι περίπου ίσο με 4S (ή, κατά προσέγγιση, $4S^*$). Επομένως, έχοντας καθορίσει την έκταση των μετρήσεων μπορούμε να προσεγγίσουμε τη δειγματική τυπική απόκλιση με τον τύπο

$$S \cong \text{έκταση}/4.$$

Η προσέγγιση της τυπικής απόκλισης μέσω της έννοιας της έκτασης των παρατηρήσεων είναι ένας χρήσιμος και γρήγορος έλεγχος για επιβεβαίωση του ότι η υπολογισθείσα τιμή της τυπικής απόκλισης είναι λογική.

Για παράδειγμα, η έκταση της διάρκειας των τηλεφωνημάτων στο σχετικό παράδειγμα είναι 17.2. Επομένως, η τιμή 17.2/4=4.3 είναι μια προσέγγιση του S. Βλέπουμε δηλαδή, ότι στην περίπτωση αυτή η προσέγγιση με την χρήση της τιμής της έκτασης είναι πολύ κοντά στην τιμή 4.29 που αποτελεί την υπολογισθείσα τιμή για την τυπική απόκλιση. Τέτοια βέβαια ακρίβεια στην προσέγγιση δεν επιτυγχάνεται πολύ συχνά.

Ενδοτεταρτημοριακό Εύρος (Interquartile range)

Ενα άλλο μέτρο απόκλισης που στηρίζεται στα εκατοστιαία σημεία είναι το **ενδοτεταρτημοριακό εύρος (interquartile range)**. Αυτό ορίζεται ως εξής:

$$\text{Ενδοτεταρτημοριακό εύρος} = Q_3 - Q_1$$

όπου Q_3 είναι το τρίτο τεταρτημόριο και Q_1 το πρώτο τεταρτημόριο. (Τα τεταρτημόρια είναι αριθμοί που χωρίζουν τα διατεταγμένα δεδομένα σε τέσσερα ίσα μέρη· Αποτελούν μέτρα σχετικής θέσης και ως τέτοια θα μελετηθούν σε επόμενη ενότητα).

Τα πλεονεκτήμα του ενδοτεταρτημοριακού εύρους ως μέτρου απόκλισης, σε σχέση με την έκταση, έγκειται στο ότι δεν επηρεάζεται από ακραίες τιμές.

3.4 Χρήση του MINITAB και του SAS για τον υπολογισμό των μέτρων θέσης και μεταβλητότητας

MINITAB

Οταν τα δεδομένα του δείγματος έχουν τοποθετηθεί στη μεταβλητή C1, η εντολή η οποία χρειάζεται για να δώσει το MINITAB τις τιμές των στατιστικών συναρτήσεων μέσω των οποίων ορίζονται τα διάφορα μέτρα θέσης και μεταβλητότητας είναι η

DESCRIBE C1

Για παράδειγμα, εάν στο προηγούμενό μας παράδειγμα με τα αμοιβαία κεφάλαια τοποθετήσουμε τις τιμές του αμοιβαίου κεφαλαίου Α στη μεταβλητή C1 και χρησιμοποιήσουμε τις εντολές

```
SET C1
8.3  -6.2  20.9  -2.7  33.6  42.9  24.4  5.2  3.1  30.5
END
DESCRIBE C1
```

η απάντηση του υπολογιστή θα είναι η εξής

N	MEAN	MEDIAN	TRMEAN	STDEV	SEMEAN
10	16.00	14.60	15.41	16.74	5.29

MIN	MAX	Q1	Q3
-6.20	42.90	1.65	31.27

Τοποθετώντας, αντίστοιχα, τις τιμές του αμοιβαίου κεφαλαίου Β στην μεταβλητή C2 και ακολουθώντας την ίδια διαδικασία θα πάρουμε το αποτέλεσμα

N	MEAN	MEDIAN	TRMEAN	STDEV	SEMEAN
10	12.00	11.75	11.87	9.97	3.15

MIN	MAX	Q1	Q3
-2.80	27.80	4.47	19.98

Τα αποτελέσματα αυτά περιλαμβάνουν τις ακόλουθες στατιστικές συναρτήσεις:

N	το μέγεθος του δείγματος
MEAN	ο μέσος του δείγματος
MEDIAN	η διάμεσος του δείγματος
TRMEAN	(trimmed mean) ένα άλλο μέτρο θέσης, ο περικομμένος μέσος, ο οποίος προκύπτει αφού παραλείψουμε το κατώτερο 5% και το ανώτερο 5% των διατεταγμένων παρατηρήσεων
STDEV	(standard deviation) η τυπική απόκλιση
SEMEAN	(standard error of the mean) το τυπικό σφάλμα του μέσου που ισούται με $S/\sqrt{N}$
MIN	(MINIMUM) η ελάχιστη παρατήρηση
MAX	(MAXIMUM) η μέγιστη παρατήρηση
Q1	(first or lower quartile) πρώτο ή κάτω τεταρτημόριο
Q3	(third or upper quartile) τρίτο ή άνω τεταρτημόριο

SAS

Προκειμένου να πάρουμε τις τιμές των περιγραφικών στατιστικών συνάρτήσεων με το στατιστικό πακέτο SAS, απαιτείται η εντολή

PROC UNIVARIATE;

Η εντολή αυτή θα δώσει ως αποτέλεσμα τον μέσο, τη διάμεσο, την επικρατούσα τιμή, την τυπική απόκλιση, τη διακύμανση και τα τεταρτημόρια για ένα σύνολο δεδομένων. Εάν θέλουμε να αποκτήσουμε τις ίδιες πληροφορίες για συγκεκριμένες μεταβλητές μπορούμε να περιλάβουμε μια VAR εντολή, όπως για παράδειγμα, για τη διαδικασία PRINT.

Εστω ότι θέλουμε να υπολογίσουμε τις περιγραφικές στατιστικές συναρτήσεις για το αμοιβαίο κεφάλαιο A του παραδείγματός μας. Θα προχωρήσουμε ως εξής:

```
DATA;
INPUT A;
CARDS;
8.3
-6.2
20.9
-2.7
33.6
42.9
24.4
5.2
3.1
30.5
PROC UNIVARIATE;
VAR A;
```

Ενα μέρος της απάντησης του προγράμματος θα είναι το εξής:

Variable = A

Moments

N	10		
Mean	16	Sum	160
Std Dev	16.7433	Variance	280.34

Σημείωση: Συνήθως, όταν αναφερόμαστε σε δειγματικά δεδομένα χρησιμοποιούμε το σύμβολο n για τον αριθμό των παρατηρήσεων. Ο λόγος που δεν αλλάξαμε το συμβολισμό και εξακολουθήσαμε να χρησιμοποιούμε το N είναι ότι έτσι εμφανίζεται συνήθως στο MINITAB.

3.5 Προσέγγιση περιγραφικών μέτρων για ομαδοποιημένα δεδομένα

Τα δύο πιο σημαντικά περιγραφικά μέτρα από αυτά που εξετάσαμε μέχρι τώρα είναι ο μέσος και η διακύμανση (ή, εναλλακτικά, η τυπική απόκλιση). Στην ενότητα αυτή θα εξετάσουμε πώς μπορούμε να προσεγγίσουμε τα δύο αυτά μέτρα για δεδομένα τα οποία έχουν ομαδοποιηθεί σε κατανομές συχνότητας.

Οπως θα δούμε, οι προσεγγίσεις αυτές είναι χρήσιμες σε δύο περιπτώσεις.

Η πρώτη περίπτωση είναι όταν θέλουμε να εξετάσουμε ένα πολύ μεγάλο σύνολο μη ομαδοποιημένων δεδομένων. Παρ'ότι στην περίπτωση αυτή μπορούμε να υπολογίσουμε το μέσο και τη διακύμανση με ακρίβεια με τη βοήθεια ενός υπολογιστή, ίσως θεωρήσουμε ότι μία προσέγγιση των μέτρων αυτών είναι αρκετή για τις ανάγκες του προβλήματός μας. Σε μια τέτοια περίπτωση ίσως θεωρήσουμε ταχύτερο και περισσότερο οικονομικό να ομαδοποιήσουμε τα δεδομένα και να κατασκευάσουμε μια κατανομή συχνότητας και στη συνέχεια να χρησιμοποιήσουμε τις μεθόδους που θα περιγράψουμε παρακάτω.

Η δεύτερη περίπτωση που συχνά συναντάται στην πράξη είναι όταν αντλούμε τα στοιχεία που θέλουμε να μελετήσουμε από κυβερνητικές πηγές ή οργανισμούς (π.χ. στοιχεία της Εθνικής Στατιστικής Υπηρεσίας, της ΕΟΚ, του ΟΗΕ κ.λ.π.). Σε τέτοιες περιπτώσεις δεδομένα που έχουν συλλεγεί από άλλους παρουσιάζονται συνήθως με τη μορφή μιας κατανομής συχνότητας και ο ερευνητής δεν έχει πρόσβαση στα μη ομαδοποιημένα δεδομένα. Στην περίπτωση αυτή δεν υπάρχει καν επιλογή και το μόνο που μπορούμε να κάνουμε είναι

να χρησιμοποιήσουμε μία προσεγγιστική μέθοδο για να καταλήξουμε στον υπολογισμό των περιγραφικών μέτρων.

Ας θεωρήσουμε ότι έχουμε ένα δείγμα n μετρήσεων (παρατηρήσεων) που έχουν ομαδοποιηθεί σε k κλάσεις (τάξεις). Αν f_i συμβολίζει τη συχνότητα των παρατηρήσεων στην κλάση i (i = 1,2,...,k), τότε

$$n = f_1 + f_2 + ... + f_k.$$

Μία καλή προσέγγιση για το δειγματικό μέσο $\bar{x}$ μπορεί να επιτευχθεί κάνοντας την παραδοχή ότι το ενδιάμεσο σημείο m_i κάθε κλάσης i προσεγγίζει ικανοποιητικά τον μέσο των μετρήσεων που ανήκουν στην κλάση i. Μια τέτοια παραδοχή δεν απέχει πολύ από την πραγματικότητα όταν οι μετρήσεις σε μια κλάση κατανέμονται σχετικά συμμετρικά γύρω από το μέσο σημείο του διαστήματος. Το άθροισμα των μετρήσεων (παρατηρήσεων) στην κλάση i είναι τότε, κατά προσέγγιση, ίσο με το γινόμενο $f_i m_i$.

Η προσέγγιση για την δειγματική διακύμανση επιτυγχάνεται με τη χρησιμοποίηση μιας προσεγγιστικής μεθόδου υπολογισμού της δειγματικής διακύμανσης για μη ομαδοποιημένα δεδομένα που εξετάσαμε προηγουμένως. Η προσέγγιση της δειγματικής διακύμανσης στην πραγματικότητα απαιτεί μια πιο αυστηρή υπόθεση από την υπόθεση της συμμετρίας στην οποία αναφερθήκαμε προηγουμένως. Στην περίπτωση της διακύμανσης, η υπόθεση που χρειάζεται να κάνουμε είναι ότι κάθε μέτρηση (παρατήρηση) μιας κλάσης είναι ίση με το μέσο του διαστήματος της κλάσης αυτής. Οσο περισσότερο η υπόθεση αυτή ανταποκρίνεται στην πραγματικότητα τόσο καλύτερη θα είναι η προσέγγιση της δειγματικής διακύμανσης.

Σύμφωνα με τα όσα είπαμε παραπάνω, οι τύποι που θα πρέπει να χρησιμοποιούμε για προσέγγιση του μέσου και της διακύμανσης ομαδοποιημένων δεδομένων είναι οι ακόλουθοι.

$$\bar{x} \cong \frac{\sum_{i=1}^{k} f_i m_i}{n}$$

$$s^2 \cong \frac{1}{n}\left[\sum_{i=1}^{k} f_i m_i^2 + \frac{\left(\sum_{i=1}^{k} f_i m_i\right)^2}{n}\right]$$

$$s^{*2} \cong \frac{1}{n-1}\left[\sum_{i=1}^{k} f_i m_i^2 + \frac{\left(\sum_{i=1}^{k} f_i m_i\right)^2}{n}\right].$$

Σημείωση: Οταν οι μετρήσεις από τις οποίες θέλουμε να υπολογίσουμε την διακύμανση αναφέρονται σε ολόκληρο τον πληθυσμό τότε, όπως είναι φυσικό, θα πρέπει να χρησιμοποιούμε τον τύπο για το S^2 και όχι για το S^{*2}.

Παράδειγμα: Ο πίνακας που ακολουθεί αναφέρεται στην κατανομή συχνότητας των στοιχείων για τη διάρκεια των τηλεφωνημάτων που εξετάσαμε νωρίτερα.

Στον πίνακα αυτόν έχουν προστεθεί τρεις στήλες που αναφέρονται σε πληροφορίες απαραίτητες για τον υπολογισμό του μέσου και της δειγματικής διακύμανσης μέσω των τύπων που ήδη δώσαμε.

Πίνακας 3.5.1

Κατανομή συχνότητας της διάρκειας των τηλεφωνημάτων

Κλάση	Ορια κλάσης	Συχνότητα f_i	Μέσο σημείο m_i	$f_i m_i$	$f_i m_i^2$
1	2 έως 5	3	3.5	10.5	36.75
2	5 έως 8	6	6.5	39.0	253.50
3	8 έως 11	8	9.5	76.0	722.00
4	11 έως 14	7	12.5	87.5	1093.75
5	14 έως 17	4	15.5	62.0	961.00
6	17 έως 20	2	18.5	37.0	684.50
Αθροισμα		n=30		312.0	3751.50

Τότε, υπολογίζοντας την αμερόληπτη εκτιμήτρια της διακύμανσης της διάρκειας όλων των τηλεφωνημάτων θα έχουμε τα εξής αποτελέσματα:

$$\bar{x} \cong \frac{\sum_{i=1}^{6} f_i m_i}{30} = \frac{312}{30} = 10.4$$

$$s^{*2} \cong \frac{1}{29} \left[\sum_{i=1}^{6} f_i m_i^2 + \frac{\left(\sum_{i=1}^{6} f_i m_i \right)^2}{30} \right] =$$

$$= \frac{1}{29} \left[3751.5 + \frac{312^2}{30} \right] =$$

$$= 17.47.$$

Οι προσεγγίσεις αυτές, όπως βλέπουμε, είναι πολύ κοντά στις πραγματικές τιμές $\bar{x} = 10.26$ και $s^{*2} = 18.40$ που έχουμε ήδη υπολογίσει.

Ανάλογοι τύποι ισχύουν για τον υπολογισμό των ροπών του δείγματος τάξης $r>2$, όπως επίσης και για την διάμεσο. (Η έννοια της ροπής αναπτύσσεται σε επόμενη ενότητα).

3.6 Μέτρα Σχετικής Θέσης

Σε πολλά πρακτικά προβλήματα η γνώση της κεντρικής τάσης των δεδομένων δεν εξασφαλίζει την ικανοποιητική περιγραφή τους. Τέτοια προβλήματα απαιτούν τον καθορισμό και άλλων μέτρων που να αναφέρονται στην σχετική θέση των δεδομένων μεταξύ τους. Στην συνέχεια θα εξετάσουμε μερικά τέτοια μέτρα.

Ποσοστιαία ή Εκατοστιαία Σημεία (Percentiles)

Η έννοια της διαμέσου μπορεί να επεκταθεί έτσι ώστε να προσδιορίζεται μία κατανομή συχνότητας με μεγαλύτερη ακρίβεια μέσω

της χρήσης διαφόρων μέτρων. Θα μπορούσαμε, για παράδειγμα, να προσδιορίσουμε τρεις τιμές οι οποίες να διαιρούν τη συνολική συχνότητα σε τέσσερα ίσα μέρη. Η μεσαία από αυτές τις τρεις τιμές θα είναι, βεβαίως, η διάμεσος. Οι άλλες δύο θα ονομάζονται *κάτω* και *άνω τεταρτημόριο*, αντίστοιχα (*lower* and *upper quartiles*). Με όμοιο τρόπο μπορούμε να προσδιορίσουμε 9 τιμές οι οποίες να διαιρούν τη συνολική συχνότητα σε δέκα ίσα μέρη - *τα δεκατημόρια* - ή 99 τιμές οι οποίες να διαιρούν την συνολική συχνότητα σε εκατό ίσα μέρη - *τα ποσοστιαία σημεία (percentiles)* ή *εκατοστιαία σημεία (percentage points).*

Γενικότερα, θα μπορούσαμε να προσδιορίσουμε ένα σύνολο n-1 τιμών οι οποίες διαιρούν τη συνολική συχνότητα σε n ίσα μέρη, τα *ποσοτικά σημεία (quantiles).* Προφανώς η γνώση των ποσοτικών σημείων για κάποια αρκετά μεγάλη τιμή του n, όπως για παράδειγμα 10, δίνει μια πολύ καλή εικόνα της γενικής μορφής της κατανομής συχνότητας. Ακόμη και μόνο τα τεταρτημόρια και η διάμεσος αποτελούν πολύτιμους γενικούς οδηγούς.

Τα παραπάνω μπορούν να συνοψισθούν στους εξής ορισμούς:

Ορισμός: Ορίζουμε ως **p-ποσοστιαίο** ή **p-εκατοστιαίο σημείο (p-percintile** ή **p-percentage point)** ενός συνόλου μετρήσεων την τιμή εκείνη που έχει την ιδιότητα ότι το πολύ p% των μετρήσεων είναι μικρότερες από την τιμή αυτή και το πολύ (100-p)% των μετρήσεων είναι μεγαλύτερες από την τιμή αυτή.

Ορισμός: Ορίζουμε ως **k-ποσοτικό σημείο (kth quantile)** ενός συνόλου τιμών την τιμή εκείνη που έχει την ιδιότητα ότι το πολύ 100k/n% των τιμών είναι μικρότερες από την τιμή αυτή και το πολύ 100(n-k)/n% των τιμών είναι μεγαλύτερες από την τιμή αυτή.

Οπως φαίνεται από τα προηγούμενα, το p-εκατοστιαίο σημείο ορίζεται με τον ίδιο ακριβώς τρόπο όπως η διάμεσος η οποία διαιρεί

μια σειρά μετρήσεων ώστε το πολύ το 50% των μετρήσεων να είναι μικρότερες από τη διάμεσο και το πολύ το 50% των μετρήσεων να είναι μεγαλύτερες από τη διάμεσο. Ετσι, στην πραγματικότητα, η διάμεσος είναι το 50ό εκατοστιαίο σημείο.

Με την ίδια λογική που έχουμε το ειδικό όνομα *διάμεσος* για το εκατοστιαίο σημείο που διαιρεί μια σειρά διατεταγμένων μετρήσεων σε δύο περίπου ίσα μέρη, έχουμε είδικά ονόματα για ποσοστιαία σημεία που χωρίζουν ένα σύνολο διατεταγμένων μετρήσεων σε τέταρτα και σε δέκατα. Οι αντίστοιχες ονομασίες είναι **τεταρτημόρια (quartiles)** και **δεκατημόρια (deciles)**.

Ετσι θα έχουμε

Πρώτο (κάτω) δεκατημόριο = 10ο εκατοστιαίο σημείο
Q_1 = πρώτο (κάτω) τεταρτημόριο = 25ο εκατοστιαίο σημείο
Q_2 = δεύτερο (μεσαίο) τεταρτημόριο = διάμεσος (50ο εκατοστιαίο σημείο)
Q_3 = τρίτο (άνω) τεταρτημόριο = 75ο εκατοστιαίο σημείο
Ενατο (άνω) δεκατημόριο = 90ο εκατοστιαίο σημείο.

Η γενική μέθοδος για τον προσδιορισμό των εκατοστιαίων σημείων ενός συνόλου μετρήσεων ή παρατηρήσεων είναι απλή. Πρώτα κατατάσσουμε τις N τιμές των δεδομένων κατά αύξουσα σειρά. Στη συνέχεια τις χωρίζουμε σε τέσσερα περίπου ίσα μέση με τον καθορισμό των τιμών με τάξεις $\frac{N+1}{4}$ (για το κάτω τεταρτημόριο), $\frac{N+1}{2}$ (για το δεύτερο τεταρτημόριο, δηλαδή τη διάμεσο) και $\frac{3(N+1)}{4}$ (για το πάνω τεταρτημόριο). Οι τιμές αυτές αντιστοιχούν στο πρώτο τεταρτημόριο Q_1, στο δεύτερο τεταρτημόριο Q_2 (διάμεσο) και στο άνω τεταρτημόριο Q_3 όπως αυτά ορίσθηκαν παραπάνω. Πράγματι, το πολύ N/4 παρατηρήσεις είναι μικρότερες από την $\frac{N+1}{4}$ διατεταγμένη παρατήρηση (δηλαδή το Q_1) και το πολύ 3N/4 είναι μεγαλύτερές της. Το πολύ N/2 παρατηρήσεις είναι μικρότερες από την $\frac{N+1}{2}$

διατεταγμένη παρατήρηση (δηλαδή τη διάμεσο) και το πολύ N/2 είναι μεγαλύτερές της. Τέλος, το πολύ 3N/4 παρατηρήσεις είναι μικρότερες από την $\frac{3(N+1)}{4}$ διατεταγμένη παρατήρηση (δηλαδή το Q_3) και το πολύ N/4 είναι μεγαλύτερές της.

Στις περισσότερες εφαρμογές έχουμε μεγάλους αριθμούς δεδομένων και οι ακριβείς τιμές των τεταρτημορίων προσδιορίζονται από υπολογιστές. Ο ερευνητής ασχολείται με την ερμηνεία των τιμών αυτών.

Ενώ τα εκατοστιαία σημεία χρησιμοποιούνται κυρίως ως ένδειξη της σχετικής θέσης μιας συγκεκριμένης τιμής των δεδομένων, τα τεταρτημόρια είναι περισσότερο χρήσιμα για να συνοψίζουν τη συνολική κατανομή όλων των δεδομένων.

Τα εκατοστιαία σημεία, παρ'ότι κατά βάση είναι μέτρα θέσης, χρησιμοποιούνται συχνά ως μέτρα σχετικής θέσης επειδή είναι πολύ εύκολο να ερμηνευθούν (Σε μέτρα σχετικής θέσης θα αναφερθούμε αργότερα). Για αυτόν το λόγο χρησιμοποιούνται συχνά ως ένδειξη της σχετικής απόδοσης ενός υποψηφίου που συναγωνίζεται σε κάποια εξέταση (όπως π.χ. οι γενικές εξετάσεις). Τα εκατοστιαία σημεία χρησιμοποιούνται επίσης πολύ συχνά για να χαρακτηρίσουν τη σχετική απόδοση συνταξιοδοτικών προγραμμάτων και επενδύσεων.

Παράδειγμα: Για να βρούμε τα τεταρτημόρια των παρατηρήσεων

7, 18, 12, 17, 29, 18, 4, 27, 30, 2, 4, 10, 21, 5, 8

διατάσσουμε πρώτα τις παρατηρήσεις αυτές κατά αύξουσα σειρά:

2, 4, 4, 5, 7, 8, 10, 12, 17, 18, 18, 21, 27, 29, 30

κάτω τεταρτημόριο ↓ 5 — διάμεσος ↓ 12 — άνω τεταρτημόριο ↓ 21

Σύμφωνα με τον ορισμό, το κάτω τεταρτημόριο είναι εκείνη η τιμή από την οποία το πολύ 0.25×15 = 3.75 των μετρήσεων είναι μικρότερες και το πολύ 0.75×15 = 11.25 των μετρήσεων είναι μεγαλύτερες. Η μόνη μέτρηση που ικανοποιεί τα κριτήρια αυτά είναι το 5. Επομένως το 5 είναι το κάτω τεταρτημόριο. Το δεύτερο

τεταρτημόριο, δηλαδή η διάμεσος είναι η τιμή 12, (η μεσαία τιμή). Το άνω τεταρτημόριο είναι εκείνη η τιμή από την οποία το πολύ $0.75 \times 15 = 11.25$ των παρατηρήσεων είναι μικρότερες και το πολύ $0.25 \times 15 = 3.75$ των παρατηρήσεων είναι μεγαλύτερες. Η μόνη μέτρηση που ικανοποιεί αυτά τα κριτήρια είναι το 21. Επομένως το 21 είναι το τρίτο τεταρτημόριο.

Στα ίδια αποτελέσματα θα καταλήγαμε αν αντί του ορισμού των τεταρτημορίων χρησιμοποιούσαμε την "θέση" που αυτά έχουν στην ακολουθία των διατεταγμένων τιμών όπως περιγράψαμε προηγουμένως. Πράγματι, η θέση του πρώτου τεταρτημορίου είναι η $(N+1)/4 = (15+1)/4 = 4$. Είναι, επομένως, η τέταρτη κατά σειρά μεγέθους παρατήρηση, δηλαδή η τιμή 5. Με όμοιο τρόπο η θέση της διαμέσου είναι η $(N+1)/2 = (15+1)/2 = 8$, δηλαδή η διάμεσος είναι η όγδοη κατά σειρά μεγέθους παρατήρηση που είναι η τιμή 12. Τέλος, η θέση του άνω τεταρτημορίου είναι η $3(N+1)/4 = 3(15+1)/4 = 12$, δηλαδή το άνω τεταρτημόριο είναι η δωδεκάτη κατά σειρά μεγέθους παρατήρηση που είναι η τιμή 21.

Σε πολλές περιπτώσεις (όταν το πλήθος των δεδομένων είναι άρτιος αριθμός) το ποσοστιαίο ή το εκατοστιαίο σημείο που αναζητούμε τοποθετείται μεταξύ δύο διαδοχικών τιμών από το σύνολο των διατεταγμένων παρατηρήσεων ή μετρήσεων. Στις περιπτώσεις αυτές θεωρούμε ως αντίστοιχο ποσοστιαίο ή εκατοστιαίο σημείο το ενδιάμεσο σημείο μεταξύ των δύο αυτών παρατηρήσεων που προκύπτει μέσω γραμμικής παρεμβολής χρησιμοποιώντας τις τιμές των αντίστοιχων τάξεών τους.

Παράδειγμα: Να προσδιορισθούν τα κάτω και άνω τεταρτημόρια του συνόλου παρατηρήσεων:

16, 25, 4, 18, 11, 13, 20, 8, 11, 9.

Λύση:

Διατάσουμε τις παρατηρήσεις κατά αύξουσα σειρά μεγέθους

4, 8, 9, 11, 11, 13, 16, 18, 20, 25.

Τότε έχουμε για τη θέση του κάτω τεταρτημορίου ότι $(N+1)/4 = (10+1)/4 = 2.75$. Επομένως, το πρώτο τεταρτημόριο είναι η τιμή που βρίσκεται στο 75% της απόστασης μεταξύ της δεύτερης και της τρίτης διατεταγμένης παρατήρησης, δηλαδή, στο 75% της απόστασης μεταξύ 8 και 9. Αρα το πρώτο τεταρτημόριο είναι ο αριθμός

$$Q_1 = 8 + 0.75(9-8) = 8.75.$$

Με όμοιο τρόπο, η θέση του άνω τεταρτημορίου είναι η $3(N+1)/4 = 3(10+1)/4 = 8.25$. Είναι, επομένως, το άνω τεταρτημόριο η τιμή που βρίσκεται στο 25% της απόστασης μεταξύ της όγδοης και της ένατης διατεταγμένης παρατήρησης, δηλαδή, στο 25% της απόστασης μεταξύ 18 και 20. Κατά συνέπεια, το άνω τεταρτημόριο είναι ο αριθμός

$$Q_3 = 18 + 0.25(20-18) = 18.5.$$

Οσο αφορά το δεύτερο τεταρτημόριο (διάμεσος), αυτό θα βρίσκεται στο 50% της απόστασης μεταξύ της πέμπτης και της έκτης διατεταγμένης παρατήρησης αφού $(N+1)/2 = (10+1)/2 = 5.5$, δηλαδή, $Q_2 = (11+13)/2 = 12$. Το αποτέλεσμα αυτό είναι σε πλήρη αρμονία με τα όσα έχουμε ήδη αναφέρει σχετικά με τον προσδιορισμό της διαμέσου.

Σημείωση: Μερικά εγχειρίδια δεν χρησιμοποιούν γραμμική παρεμβολή για τον προσδιορισμό των τεταρτημορίων. Απλώς στρογγυλοποιούν την τάξη στον πλησιέστερο ακέραιο και θεωρούν ως τεταρτημόριο την παρατήρηση με τάξη την προκύπτουσα στρογγυλοποιημένη τιμή. Θεωρούν δε ως τεταρτημόριο τον μέσο των τιμών των παρατηρήσεων που αντιστοιχούν στις δύο γειτονικές τάξεις μόνο όταν οι τιμές $(N+1)/4$ και $3(N+1)/4$ λήγουν σε 0.5.

Για παράδειγμα, αν στα παραπάνω δεδομένα ακολουθήσουμε τη

μέθοδο της στρογγυλοποίησης των τάξεων 2.75 του κάτω τεταρτημορίου και 8.25 του άνω τεταρτημορίου σε 3 και 8 αντίστοιχα, τότε ως κάτω και άνω τεταρτημόρια προκύπτουν οι τιμές $Q_1 = 9$ και $Q_3 = 18$, αντίστοιχα. Στην περίπτωση αυτή δηλαδή τα Q_1 και Q_3 είναι οι διάμεσοι του πρώτου και του δεύτερου μισού των διατεταγμένων παρατηρήσεων αντίστοιχα. (Οπως θα δούμε αργότερα, οι "διάμεσοι" αυτές ονομάζονται *άξονες* ή *κεντρικά σημεία (hinges)*. Στα περισσότερα εγχειρίδια όμως δε γίνεται διάκριση μεταξύ τεταρτημορίων και αξόνων, αλλά οι έννοιες αυτές θεωρούνται ταυτόσημες).

Παρατήρηση: Αν για την κατασκευή του διαγράμματος μίσχου-φύλλου χρησιμοποιήσουμε τις παρατηρήσεις του δείγματος αφού πρώτα τις διατάξουμε κατά αύξουσα σειρά μπορούμε, χρησιμοποιώντας το διάγραμμα αυτό, να εντοπίσουμε τιμές πολλών στατιστικών συναρτήσεων που μας ενδιαφέρουν, όπως την διάμεσο, το πρώτο και το τρίτο τεταρτημόριο και, φυσικά, τη μέγιστη και την ελάχιστη των παρατηρήσεων.

Ετσι, στο παράδειγμά μας, αφού έχουμε διατάξει τις παρατηρήσεις κατά αύξουσα σειρά κατασκευάζουμε το διάγραμμα μίσχου-φύλλου.

Stem and Leaf για διάμεσο, Q_1,Q_3,X_{min},X_{max}

$X_{min} = 2$, $Q_1 = 5$	0	[2] 4 4 [5] 7 8
Διαμ.= 12	1	0 [2] 7 8 8
$Q_3 = 21$	2	[1] 7 9
$X_{max} = 30$	3	[0]

Από τους ορισμούς των αντιστοίχων εννοιών, όπως αυτές δόθηκαν προηγουμένως, είναι προφανές ότι

- η διάμεσος θα βρίσκεται στη θέση $d(M)=\frac{N+1}{2}$, δηλαδή στη θέση της $\frac{N+1}{2}$ διατεταγμένης παρατήρησης
- το πρώτο τεταρτημόριο θα βρίσκεται στη θέση $d(Q_1)=\frac{N+1}{4}$, δηλαδή στη θέση της $\frac{N+1}{4}$ διατεταγμένης παρατήρησης
- το τρίτο τεταρτημόριο θα βρίσκεται στη θέση $d(Q_3)=\frac{N+1}{4}$, δηλαδή στη θέση της $\frac{3(N+1)}{4}$ διατεταγμένης παρατήρησης.

Ετσι, για το παράδειγμά μας, όπου έχουμε $N=15$ προκύπτει ότι:

- η διάμεσος βρίσκεται στη θέση της $\frac{15+1}{2}$ = 8ης διατεταγμένης παρατήρησης.

Δηλαδή

$$\text{Διάμεσος} = 12.$$

Επίσης

- το Q_1 θα βρίσκεται στη θέση της $\frac{15+1}{4}$ = 4ης διατεταγμένης παρατήρησης
- το Q_3 θα βρίσκεται στη θέση της $\frac{3(15+1)}{4}$ = 12ης διατεταγμένης παρατήρησης.

Αρα

$$Q_1 = 5 \quad , \quad Q_3 = 21.$$

Επίσης είναι προφανές ότι η ελάχιστη παρατήρηση είναι η $X_{min} = 2$ και η μέγιστη παρατήρηση είναι η $X_{max} = 30$.

Παρατηρούμε, δηλαδή, ότι οι τιμές αυτών των στατιστικών συναρτήσεων μπορούν να προσδιορισθούν εύκολα, αλλά και να σημειωθούν, στο διάγραμμα μίσχου-φύλλου.

Σημείωση: Οπως προαναφέραμε, αν κάποιο από τα κλάσματα που προσδιορίζει τη θέση ενός από τα εκατοστιαία σημεία δεν είναι ακέραιος, το αντίστοιχο εκατοστιαίο σημείο βρίσκεται μεταξύ των θέσεων που αντιστοιχούν στον ακέραιο που προηγείται και στον

ακέραιο που έπεται της τιμής που βρέθηκε και είναι η τιμή που παρεμβάλλεται γραμμικά μεταξύ των παρατηρήσεων που βρίσκονται σε αυτές τις δύο θέσεις. Στο διάγραμμα μίσχου-φύλλου τότε σημειώνουμε και τις δύο αυτές τιμές.

Τυποποιημένες τιμές (Standardized values)

Ενα άλλο μέτρο σχετικής θέσης (relative standing) των τιμών των παρατηρήσεων (μετρήσεων) σε ένα σύνολο παρατηρήσεων (μετρήσεων) είναι οι λεγόμενες **τυποποιημένες τιμές (standardized values)** ή αλλιώς **Z-τιμές (Z-values)**.

Ορισμός: Ως **Z-τιμή** ή **τυποποιημένη τιμή** μιας παρατήρησης (μέτρησης) ορίζεται η απόσταση της παρατήρησης αυτής από τον μέσο του συνόλου των παρατηρήσεων εκφρασμένη σε μονάδες τυπικής απόκλισης.

Εναλλακτικά, η Z-τιμή ορίζεται ως ο αριθμός των τυπικών αποκλίσεων κατά τις οποίες μια παρατήρηση βρίσκεται πάνω ή κάτω από τον μέσο.

Επομένως, ο γενικός τρόπος υπολογισμού της Z-τιμής ενός συνόλου παρατηρήσεων είναι

$$Z = \frac{X - \mu}{\sigma}$$

όπου X η συγκεκριμένη παρατήρηση που μας ενδιαφέρει, μ ο μέσος και σ η τυπική απόκλιση.

Με παρόμοιο τρόπο ορίζεται και η τυποποιημένη τιμή μιας παρατήρησης σε ένα δείγμα. Στην τελευταία αυτή περίπτωση ο μέσος μ αντικαθίσταται από τον δειγματικό μέσο $\overline{X}$ και η τυπική απόκλιση σ από την τυπική απόκλιση του δείγματος S ή S^*.

Είναι σημαντικό να παρατηρήσουμε ότι οι Z-τιμές είναι καθαροί αριθμοί (δεν εκφράζονται σε κάποια μονάδα μέτρησης). Για αυτόν το λόγο, κυρίως, **οι Z-τιμές είναι ιδιαίτερα χρήσιμες όταν θέλουμε να συγκρίνουμε αποδόσεις που έχουν μετρηθεί σε διαφορετικές κλίμακες.** Ο μετασχηματισμός μιας τιμής των δεδομένων στην αντίστοιχη Z-τιμή

ονομάζεται *μετασχηματισμός τυποποίησης (standardizing transformation).*

Χαρακτηριστική περίπτωση για την εφαρμογή της έννοιας της τυποποιημένης τιμής για σύγκριση στοιχείων που βρίσκονται σε διαφορετικές κλίμακες μέτρησης είναι η περίπτωση των βαθμολογιών σε μαθήματα των Γενικών Εξετάσεων. Οπως είναι γνωστό, στις Γενικές Εξετάσεις ένας υποψήφιος έχει το δικαίωμα να "διατηρήσει" την βαθμολογία του σε ένα μάθημα και στις επόμενες Γενικές Εξετάσεις. Η δυνατότητα αυτή έχει δημιουργήσει επανειλημμένα προβλήματα λόγω της διαφοράς στο βαθμό δυσκολίας των θεμάτων των Γενικών Εξετάσεων στα συγκεκριμένα μαθήματα από έτος σε έτος. Χαρακτηριστικότερη είναι η περίπτωση της εξέτασης του μαθήματος της Φυσικής της Α' δέσμης στις Γενικές Εξετάσεις του 1993 και η άνιση, όπως χαρακτηρίστηκε, μεταχείρηση αυτών που εξετάστηκαν για πρώτη φορά στη Φυσική στις εξετάσεις του 1993 σε σχέση με αυτούς που είχαν εξετασθεί στη Φυσική στις Γενικές Εξετάσεις του 1992 και είχαν "διατηρήσει" το βαθμό τους.

Ο καλύτερος - και επιστημονικά ενδεδειγμένος - τρόπος αντιμετώπισης του προβλήματος αυτού θα ήταν η χρησιμοποίηση για την αξιολόγηση των υποψηφίων των τυποποιημένων βαθμολογιών τους για κάθε μάθημα και έτος και όχι των απολύτων βαθμολογιών τους.

Για να εξηγήσουμε καλύτερα τη μέθοδο και τη σημασία της για τη δικαιότερη διεξαγωγή των γενικών εξετάσεων, ας θεωρήσουμε το εξής παράδειγμα: (Για ευκολία θεωρούμε τις βαθμολογίες των υποψηφίων σε κλίμακα από 0-20. Είναι προφανές ότι τα συμπεράσματα θα ισχύουν για οποιαδήποτε κλίμακα).

Εστω ότι η βαθμολογία όλων των υποψηφίων στο μάθημα της Φυσικής της Α' δέσμης στις γενικές εξετάσεις του 1992 είχε μέση τιμή $\mu_1 = 11$ και τυπική απόκλιση $\sigma_1 = 2$.

Εστω ότι η αντίστοιχη βαθμολογία στο ίδιο μάθημα στις εξετάσεις του 1993 είχε μέση τιμή $\mu_2 = 6$ και τυπική απόκλιση $\sigma_2 = 3$.

Εστω ότι έχουμε τις βαθμολογίες στο μάθημα της Φυσικής της Α'

δέσμης 10 υποψηφίων που πήραν μέρος στις εξετάσεις του 1992 και "διατήρησαν" το βαθμό τους για τις εξετάσεις του 1993. Εστω επίσης ότι έχουμε τις βαθμολογίες 10 άλλων υποψηφίων που πήραν μέρος στις εξετάσεις Φυσικής της Α' δέσμης για πρώτη φορά το 1993.

Οι βαθμολογίες των 20 αυτών υποψηφίων στη Φυσική όπως επίσης και οι αντίστοιχες τυποποιημένες τιμές τους εμφανίζονται στον πίνακα 3.6.1. (Οι τυποποιημένες τιμές προκύπτουν από τον τύπο της τυποποίησης π.χ. η τιμή -0.46 προκύπτει από τον τύπο $\frac{10.08-11}{2}$).

Πίνακας 3.6.1

Οι βαθμολογίες (πραγματικές και τυποποιημένες) των 20 υποψηφίων

Υποψήφιος	Βαθμός Φυσικής	Τυποποιημένος βαθμός z	Ετος εξέτασης
Α	10.08	-0.46	92
Β	12.72	0.86	92
Γ	14.06	1.53	92
Δ	12.14	0.57	92
Ε	10.53	-0.24	92
Ζ	12.05	0.52	92
Η	11.14	0.07	92
Θ	12.81	0.91	92
Ι	11.19	0.09	92
Κ	9.35	-0.83	92
Λ	6.99	0.33	93
Μ	8.13	0.71	93
Ν	6.32	0.11	93
Ξ	12.71	2.24	93
Ο	6.52	0.17	93
Π	13.31	2.44	93
Ρ	3.08	-0.97	93
Σ	5.38	-0.21	93
Τ	4.12	-0.63	93
Υ	7.60	0.53	93

Ας υποθέσουμε ότι και οι 20 αυτοί υποψήφιοι έχουν τον ίδιο ακριβώς μέσο όρο σε όλα τα άλλα μαθήματα (οπότε η εισαγωγή τους σε κάποιο συγκεκριμένο τμήμα εξαρτάται αποκλειστικά από την απόδοσή τους στο μάθημα της Φυσικής). Ο πίνακας 3.6.2 συνοψίζει την

απόδοση των 20 υποψηφίων με τον συνήθη τρόπο "μέτρησής" της μέσω των πραγματικών βαθμών τους και κατατάσσει τους υποψηφίους κατά σειρά "επιτυχίας" με βάση αυτούς τους βαθμούς.

Πίνακας 3.6.2

Οι βαθμολογίες (πραγματικές και τυποποιημένες) των 20 υποψηφίων με φθίνουσα σειρά των πραγματικών βαθμών

Υποψήφιος	Σειρά επιτυχίας	Βαθμός Φυσικής	Τυποποιημένος βαθμός z	Ετος εξέτασης
Γ	1	14.06	1.53	92
Π	2	13.31	2.44	93
Θ	3	12.81	0.91	92
Β	4	12.72	0.86	92
Ξ	5	12.71	2.24	93
Δ	6	12.14	0.57	92
Ζ	7	12.05	0.52	92
Ι	8	11.19	0.09	92
Η	9	11.14	0.07	92
Ε	10	10.53	-0.24	92
Α	11	10.08	-0.46	92
Κ	12	9.35	-0.83	92
Μ	13	8.13	0.71	93
Υ	14	7.60	0.53	93
Λ	15	6.99	0.33	93
Ο	16	6.52	0.17	93
Ν	17	6.32	0.11	93
Σ	18	5.38	-0.21	93
Τ	19	4.12	-0.63	93
Ρ	20	3.08	-0.97	93

Αν λοιπόν υποθέσουμε ότι από τους 20 αυτούς υποψήφιους υπάρχει δυνατότητα να εισαχθούν σε κάποιο συγκεκριμένο τμήμα μόνο 5, από τα δεδομένα του πίνακα 3.6.2, παρατηρούμε ότι θα εισαχθούν 3 από αυτούς που εξετάσθηκαν το 1992 (αυτοί με τους βαθμούς 14.06 (1ος), 12.81 (3ος) και 12.72 (4ος)), και 2 από τους εξετασθέντες το 1993 (αυτοί με τους βαθμούς 13.31 (2ος) και 12.71 (5ος)).

Αν όμως η απόδοση των υποψηφίων "μετρηθεί" μέσω των τυποποιημένων βαθμολογιών τους (πίνακας 3.6.3), τα αποτελέσματα

δεν θα είναι ακριβώς τα ίδια.

Πίνακας 3.6.3

Οι βαθμολογίες (πραγματικές και τυποποιημένες) των 20 υποψηφίων με φθίνουσα σειρά των τυποποιημένων βαθμών

Υποψήφιος	Σειρά επιτυχίας	Βαθμός Φυσικής	Τυποποιημένος βαθμός z	Ετος εξέτασης
Π	1	13.31	2.44	93
Ξ	2	12.71	2.24	93
Γ	3	14.06	1.53	92
Θ	4	12.81	0.91	92
Β	5	12.72	0.86	92
Μ	6	8.13	0.71	93
Δ	7	12.14	0.57	92
Υ	8	7.60	0.53	93
Ζ	9	12.05	0.52	92
Λ	10	6.99	0.33	93
Ο	11	6.52	0.17	93
Ν	12	6.32	0.11	93
Ι	13	11.19	0.09	92
Η	14	11.14	0.07	92
Σ	15	5.38	-0.21	93
Ε	16	10.53	-0.24	92
Α	17	10.08	-0.46	93
Τ	18	4.12	-0.63	93
Κ	19	9.35	-0.83	92
Ρ	20	3.08	-0.97	93

Πράγματι, κατατάσσοντας τους υποψήφιους με βάση τις τυποποιημένες βαθμολογίες τους παρατηρούμε, από τον πίνακα 3.6.3, ότι εισάγονται και πάλι 3 από αυτούς που εξετάσθηκαν το 1992, αυτοί με τους τυποποιημένους βαθμούς:

1.53 (3ος) (με τον πραγματικό του βαθμό 14.06 ήταν 1ος),

0.91 (4ος) (με τον πραγματικό του βαθμό 12.81 ήταν 3ος)

0.86 (5ος) (με τον πραγματικό του βαθμό 12.72 ήταν 4ος)).

Από αυτούς που διαγωνίσθηκαν το 1993 εισάγονται 2, αυτοί που έχουν τυποποιημένους βαθμούς:

2.48 (1ος) (με τον πραγματικό του βαθμό 13.31 ήταν 2ος)
2.24 (2ος) (με τον πραγματικό του βαθμό 12.71 ήταν 5ος)).

Βλέπουμε δηλαδή ότι στην περίπτωση αυτή οι εισακτέοι παραμένουν οι ίδιοι είτε χρησιμοποιήσουμε τις τυποποιημένες βαθμολογίες είτε χρησιμοποιήσουμε τις πραγματικές βαθμολογίες. Αλλάζει όμως ουσιαστικά η σειρά επιτυχίας.

Αν όμως εξετάσουμε την περίπτωση που από τους 20 αυτούς υποψηφίους θα εισαχθούν στο συγκεκριμένο τμήμα 9, βλέπουμε ότι:

Με βάση τις πραγματικές βαθμολογίες από τους υποψηφίους του 1992, σύμφωνα με τον πίνακα 3.6.2, εισάγονται 7, αυτοί με τους βαθμούς

14.06 (1ος), 12.81 (3ος), 12.72 (4ος), 12.14 (6ος),
12.05 (7ος), 11.19 (8ος) και 11.14 (9ος).

Από τους υποψήφιους του 1993, σύμφωνα με τον πίνακα 3.6.2, εισάγονται 2, αυτοί με τους βαθμούς 13.31 (2ος) και 12.71 (5ος).

Αν χρησιμοποιηθούν οι τυποποιημένες βαθμολογίες από τους υποψηφίους του 1992, σύμφωνα με τον πίνακα 3.6.3, εισάγονται μόνο 4, αυτοί που είχαν τους τυποποιημένους βαθμούς:

1.53 (3ος) (με τον πραγματικό του βαθμό 14.06 ήταν 1ος),
0.91 (4ος) (με τον πραγματικό του βαθμό 12.81 ήταν 3ος),
0.86 (5ος) (με τον πραγματικό του βαθμό 12.72 ήταν 4ος),
0.57 (7ος) (με τον πραγματικό του βαθμό 12.14 ήταν 6ος).

Από τους υποψήφιους του 1993, σύμφωνα με τον πίνακα 3.6.3, εισάγονται 5, αυτοί με τους τυποποιημένους βαθμούς:

2.48 (1ος) (με τον πραγματικό του βαθμό 13.31 ήταν 2ος),
2.24 (2ος) (με τον πραγματικό του βαθμό 12.71 ήταν 5ος),
0.71 (6ος) (με τον πραγματικό του βαθμό 8.13 δεν εισαγόταν),
0.53 (8ος) (με τον πραγματικό του βαθμό 7.6 δεν εισαγόταν),
1.33 (9ος) (με τον πραγματικό του βαθμό 6.99 δεν εισαγόταν).

Βλέπουμε δηλαδή ότι στην περίπτωση των 9 εισακτέων στο τμήμα δεν αλλάζει μόνο η σειρά των επιτυχόντων αλλά, και το σημαντικότερο, και οι επιτυχόντες. (Τρεις από τους υποψήφιους του 1992 που θα πετύχαιναν με τους πραγματικούς βαθμούς, οι Ξ, Δ και Ζ θα ήταν αποτυχόντες με τους τυποποιημένους βαθμούς, ενώ τρεις από τους αποτυχόντες με τους πραγματικούς βαθμούς υποψήφιους του 1993, οι Γ, Θ και Β θα πετύχαιναν με τους τυποποιημένους βαθμούς).

Είναι εξ άλλου προφανές ότι λόγω της διαφοράς δυσκολίας των θεμάτων, η αξιολόγηση με βάση την τυποποιημένη βαθμολογία είναι δικαιότερη. Αυτό εξηγείται από το γεγονός ότι η τυποποιημένη βαθμολογία στηρίζεται στη σχετική απόδοση των υποψηφίων και όχι στην απόλυτη απόδοσή τους. (Ολες οι τυποποιημένες τιμές έχουν μέση τιμή 0 και τυπική απόκλιση 1). Επομένως, η τυποποίηση των βαθμών των γενικών εξετάσεων θα εξασφάλιζε ένα αντικειμενικό και δίκαιο τρόπο επιλογής με υπέρβαση των αδικιών που δημιουργούνται από τη διατήρηση της βαθμολογίας και τη διαφορά δυσκολίας των θεμάτων από έτος σε έτος.

Παρατήρηση: Η ερμηνεία των Z-τιμών εξαρτάται από την κατανομή των δεδομένων. Εάν η κατανομή είναι τέτοια ώστε να μπορεί να εφαρμοσθεί ο εμπειρικός κανόνας, η ερμηνεία μιας Z-τιμής είναι απλή. Μια Z-τιμή με απόλυτη τιμή 1 ή λιγότερο από 1 δεν είναι καθόλου ασυνήθιστη. Αυτό γιατί, για τις περιπτώσεις που ο εμπειρικός κανόνας μπορεί να εφαρμοσθεί, γνωρίζουμε ότι το 68% όλων των τυποποιημένων τιμών των δεδομένων θα πρέπει να βρίσκεται μεταξύ +1 και -1. Αντίθετα μια Z-τιμή με απόλυτη τιμή μεγαλύτερη από 3 είναι ασυνήθιστη δοθέντος ότι λιγότερο από το 1% όλων των παρατηρήσεων θα έχουν Z-τιμή μεγαλύτερη από 3 ή μικρότερη από -3. Μια αρνητική Z-τιμή αποτελεί ένδειξη ότι μια παρατήρηση βρίσκεται κάτω από τον μέσο. Μια θετική Z-τιμή, αντίστοιχα, αποτελεί ένδειξη ότι η αντίστοιχη παρατήρηση είναι μεγαλύτερη από τον μέσο.

Για παράδειγμα, ας υποθέσουμε ότι σε μία εξέταση με βαθμολογία σε κλίμακα από 0 έως 100 ο μέσος βαθμός των εξετασθέντων φοιτητών είναι 74 και η τυπική απόκλιση είναι 8. Τότε, για ένα φοιτητή με βαθμό 92 θα έχουμε ότι

$$Z = \frac{92-74}{8} = 2.25.$$

Επομένως, η απόδοση του συγκεκριμένου φοιτητή βρίσκεται 2.25 τυπικές αποκλίσεις πάνω από το μέσο.

Εξαιρετικά μεγάλες ή εξαιρετικά μικρές τυποποιημένες τιμές (μικρότερες του -3 ή μεγαλύτερες του 3) δημιουργούν ερωτηματικά για την ακρίβεια της αντίστοιχης παρατήρησης. Η συγκεκριμένη παρατήρηση ίσως έχει καταγραφεί λάθος για κάποιο λόγο ή ίσως δεν ανήκει στον πληθυσμό από τον οποίο έχουμε επιλέξει το δείγμα. Παρατηρήσεις με εξαιρετικά μικρές ή μεγάλες τυποποιημένες τιμές ονομάζονται *ακραίες παρατηρήσεις (outliers)* ακριβώς γιατί βρίσκονται μακριά από το "κέντρο" των δεδομένων.

3.7 Μέτρα Σχετικής Μεταβλητότητας

Εστω ότι ένας επενδυτής αναρωτιέται κατά πόσο μια απώλεια 5 χιλ. δρχ. είναι περισσότερο επώδυνη αν προέλθει από μια επένδυση 100 χιλ. δρχ. από ότι αν προέλθει από μια επένδυση 500 χιλ. δρχ.. Η διαισθητική απάντηση είναι ότι, ίσως, ο επενδυτής θα έπρεπε να ανησυχεί περισσότερο αν η απώλεια αυτή προέρχεται από επένδυση 100 χιλ. δρχ., δοθέντος ότι η απώλεια 5 χιλ. δρχ. στις 100 χιλ. δρχ. αντιπροσωπεύει ποσοστιαία μεταβολή στην αξία της επένδυσης μεγαλύτερη από την ποσοστιαία μεταβολή που προκαλεί η απώλεια 5 χιλ. δρχ. από επένδυση 500 χιλ. δρχ..

Προβλήματα αυτής της μορφής οδηγούν στην ανάγκη ορισμού μέτρων που συνδυάζουν μέτρα θέσης με μέτρα απόκλισης.

Συντελεστής Μεταβλητότητας (Coefficient of Variation)

Το πιο γνωστό μέτρο σχετικής απόκλισης είναι ο **συντελεστής μεταβλητότητας (coefficient of variation).**

Σε αντίθεση με τα μέτρα θέσης και απόκλισης που εξετάσαμε μέχρι τώρα, ο συντελεστής μεταβλητότητας (coefficient of variation) είναι ένα μέτρο σχετικής μεταβλητότητας. Εκφράζεται δε συνήθως ως ποσοστό κάι όχι μέσω των μονάδων των δεδομένων στα οποία αναφέρεται.

Ο συντελεστής μεταβλητότητας που συμβολίζεται με CV μετρά το "άπλωμα" των δεδομένων σε σχέση με το μέσο, δίνεται δε από τον τύπο

$$CV = \frac{S^*}{\overline{X}}$$

όπου S^* η τυπική απόκλιση του δείγματος που προέρχεται από την αμερόληπτη εκτιμήτρια της διακύμανσης και $\overline{X}$ ο μέσος του δείγματος.

Ο συντελεστής μεταβλητότητας είναι επίσης χρήσιμος για τη σύγκριση δύο ή περισσοτέρων συνόλων δεδομένων που έχουν μετρηθεί με τις ίδιες μονάδες αλλά διαφέρουν σε τέτοιο βαθμό ώστε μία

απευθείας σύγκριση των αντιστοίχων τυπικών αποκλίσεων να μην είναι πολύ χρήσιμη. Επειδή δε είναι ένα καθαρός αριθμός (δηλαδή ανεξάρτητος των μονάδων στις οποίες έχουν μετρηθεί τα δεδομένα), χρησιμοποιείται επίσης για τη σύγκριση της μεταβλητότητας δύο ή περισσοτέρων συνόλων δεδομένων που έχουν μετρηθεί σε διαφορετικές μονάδες (π.χ. χιλιάδες δραχμές και εκατομμύρια δραχμές).

Για να γίνει καλύτερα αντιληπτή η χρησιμότητα του συντελεστή αυτού, ας γυρίσουμε στο παράδειγμα των αμοιβαίων κεφαλαίων. Στο παράδειγμα αυτό οι συντελεστές μεταβλητότητας για τους δειγματικούς ρυθμούς απόδοσης των κεφαλαίων Α και Β είναι

$$CV_A = \frac{s^*_A}{\bar{x}_A} = \frac{16.74}{16} = 1.05$$

και

$$CV_B = \frac{s^*_B}{\bar{x}_B} = \frac{9.97}{12} = 0.83.$$

αντίστοιχα.

Βλέπουμε, επομένως, ότι η σύγκριση των συντελεστών μεταβλητότητας οδηγεί στο ίδιο συμπέρασμα που είχε οδηγήσει και η σύγκριση των αντιστοίχων τυπικών αποκλίσεων. Οτι, δηλαδή, οι μετρήσεις στο δείγμα Α έχουν μεγαλύτερη μεταβλητότητα. Αν όμως η μέση ετήσια απόδοση για το κεφάλαιο Α, στο παράδειγμά μας, ήταν 21% με την ίδια τυπική απόκλιση $s^*_A = 16.74\%$, ο συντελεστής μεταβλητότητας της απόδοσης για το κεφάλαιο Α θα ήταν

$$CV_A = \frac{16.74}{21} = 0.80.$$

Στην περίπτωση αυτή, δηλαδή, το αμοιβαίο κεφάλαιο Α θα είχε σχετικά μικρότερη μεταβλητότητα από ότι το κεφάλαιο Β (του οποίου ο αντίστοιχος συντελεστής μεταβλητότητας θα ήταν 0.83), συμπέρασμα αντίθετο από αυτό στο οποίο θα καταλήγαμε αν χρησιμοποιούσαμε τις τυπικές αποκλίσεις.

Σημείωση: Ο συντελεστής μεταβλητότητας πολλαπλασιάζεται μερικές

φορές με το 100 και δίνεται ως ποσοστό, πράγμα που σημαίνει ότι εκφράζει την τυπική απόκλιση σαν ένα ποσοστό του μέσου. Με τον τρόπο αυτό έκφρασης για το κεφάλαιο Α του παραδείγματός μας, ο συντελεστής μεταβλητότητας θα ήταν 105%.

Μέση Διαφορά του Gini (Gini's Mean Difference)

Ενα άλλο μέτρο σχετικής μεταβλητότητας που αναφέρεται στην "συγκέντρωση" μιας σειράς μετρήσεων είναι η μέση διαφορά του Gini.

Ορισμός: Για μια σειρά μετρήσεων $x_1, x_2, \ldots, x_n$ ορίζουμε ως **μέση διαφορά του Gini (Gini's mean difference)** την ποσότητα

$$g = \frac{1}{n(n-1)} \sum_{i,j=1}^{n} |x_i - x_j|.$$

Επειδή το άθροισμα αναφέρεται σε όλα τα πιθανά ζεύγη των i και j (διπλό άθροισμα) ο δείκτης g εκφράζει τον μέσο των μεταξύ των μετρήσεων αποκλίσεων. (Κάθε διαφορά εμφανίζεται δύο φορές στο άθροισμα, ως $|x_i - x_j|$ και ως $|x_j - x_i|$. Για το λόγο αυτό, προκειμένου να υπολογισθεί ο μέσος όρος των απολύτων διαφορών, διαιρούμε το άθροισμα με n(n-1), δηλαδή με το συνολικό αριθμό των ζευγών που είναι δυνατόν να σχηματίζουν οι n μετρήσεις).

Παράδειγμα: Το μηνιαίο εισόδημα των 4 μελών μιας οικογένειας είναι (σε χιλιάδες δραχμές):

300 250 280 100

Η τιμή του g είναι

$$\begin{aligned} g = \frac{1}{4\times 3} = \Big[& |300-250| + |250-300| + |300-280| + |280-300| + \\ & + |300-100| + |100-300| + |250-280| + |280-250| + \\ & + |250-100| + |100-250| + |280-100| + |100-280| \Big] = \\ & = 83.3. \end{aligned}$$

Επομένως, ο μέσος όρος των απολύτων διαφορών του μισθού κάθε

μέλους της οικογένειας αυτής από του μισθούς των υπολοίπων μελών της οικογένειας είναι 83.3 χιλ. δρχ.

Σημείωση: Επειδή στον τύπο του υπολογισμού του g κάθε διαφορά εμφανίζεται δύο φορές, για απλοποίηση των υπολογισμών μπορούμε να γράφουμε, ισοδύναμα

$$g = \frac{2}{n(n-1)} \sum_{\substack{i,j=1 \\ i<j}}^{n} |x_i - x_j|$$

Παρατήρηση: Η διαφορά του δείκτη g από τα άλλα μέτρα σχετικής μεταβλητότητας που εξετάσαμε είναι ότι ο g αναφέρεται σε μέση απόκλιση των μετρήσεων μεταξύ τους ενώ τα άλλα μέτρα κάνουν χρήση των αποκλίσεων των μετρήσεων από το μέσο τους.

3.8 Διάγραμμα πλαισίου και απολήξεων (Box and Whisker Plot)

Εχουμε ήδη αναφερθεί σε πολλά μέτρα θέσης, μεταβλητότητας και σχετικής θέσης. Εχουμε επίσης αναφερθεί σε περιπτώσεις σημείων από ένα σύνολο δεδομένων που βρίσκονται πολύ μακριά από τον κύριο όγκο των δεδομένων. (Ακραία σημεία (outliers)).

Ο Αμερικανός στατιστικός John Tukey παρουσίασε, σχετικά πρόσφατα, μια μέθοδο παρουσίασης των δεδομένων που όχι μόνο προσδιορίζει τιμές διαφόρων μέτρων θέσης και σχετικής θέσης (διαμέσου, ποσοστιαίων σημείων), τιμές διαφόρων μέτρων μεταβλητότητας και σχετικής μεταβλητότητας (εύρους, ενδοτεταρτημοριακού εύρους) και ακραίες τιμές αλλά και τις παρουσιάζει γραφικά. Το διάγραμμα αυτό που προτάθηκε από τον Tukey το 1977 λέγεται *διάγραμμα πλαισίου και απολήξεων (box and whisker plot)*. Μερικοί στατιστικοί ονομάζουν το διάγραμμα αυτό *θηκόγραμμα* ενώ στη διεθνή βιβλιογραφία χρησιμοποιείται συχνά μόνο ο όρος *box plot*.

Το διάγραμμα πλαισίου και απολήξεων, όπως και το διάγραμμα

μίσχου-φύλλου (που επίσης εισήγαγε ο John Tukey), αποτελούν τα κύρια στοιχεία μιας νέας μεθοδολογίας απλής αλλά ισχυρής παρουσίασης δεδομένων που είναι γνωστή με την ονομασία *ανιχνευτική ανάλυση δεδομένων (exploratory data analysis* ή, για συντομογραφία, *EDA).*

Το όνομα του διαγράμματος αυτού είναι πλήρως περιγραφικό της αποστολής του. Υποδεικνύει τις τιμές των μέτρων κεντρικής θέσης και απόκλισης που έχουν την ιδιότητα της ανθεκτικότητας (robustness), δηλαδή τις τιμές στατιστικών συναρτήσεων που δεν επηρεάζονται εν γένει από ακραίες τιμές ή από αλλαγές σε κάποια από τα δεδομένα. Τέτοιες στατιστικές συναρτήσεις είναι η διάμεσος και τα τεταρτημόρια. Το διάγραμμα πλαισίου και απολήξεων συνδυάζει τις προαναφερθείσες ανθεκτικές στατιστικές συναρτήσεις με πληροφορίες που αναφέρονται στις ακραίες τιμές.

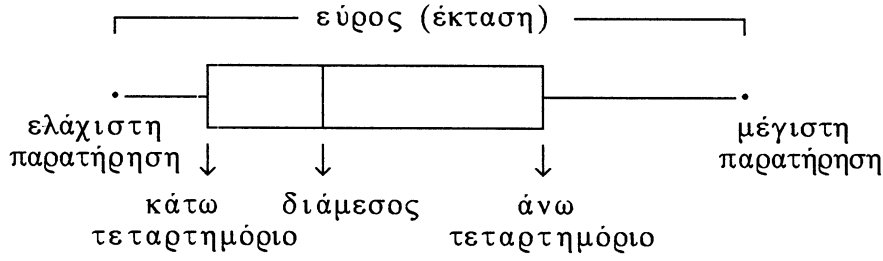
Συγκεκριμένα τα στοιχεία τα οποία δίνει το διάγραμμα πλαισίου-απολήξεων είναι οι τιμές

$$X_{min}\ ,\ Q_1\ ,\ \text{διάμεσος}\ ,\ Q_3\ ,\ X_{max}$$

όπου, όπως είδαμε, Q_1 και Q_3 είναι το πρώτο και το τρίτο τεταρτημόριο, αντίστοιχα και X_{min}, X_{max} είναι η ελάχιστη και η μέγιστη τιμή των δεδομένων.

Επειδή η μέθοδος αυτή παρουσίασης των δεδομένων χρησιμοποιεί πέντε αριθμούς λέγεται και **σύνοψη των πέντε αριθμών (five-number summary).**

Το σχήμα 3.8.1 απεικονίζει ένα πρότυπο αυτού του διαγράμματος πλαισίου και απολήξεων. Το διάγραμμα εδώ απεικονίζεται οριζόντια, αλλά θα μπορούσε να απεικονισθεί και κατακόρυφα.



Σχήμα 3.8.1
Κύρια στοιχεία ενός διαγράμματος πλαισίου και απολήξεων

Οι απολήξεις του διαγράμματος αντιστοιχούν στην ελάχιστη και μέγιστη παρατήρηση του συνόλου των δεδομένων. Η απόσταση μεταξύ αυτών των σημείων παριστάνει την έκταση. Τα άκρα του πλαισίου τοποθετούνται στο πρώτο και τρίτο τεταρτημόριο. Επομένως το μήκος του πλαισίου παριστάνει το ενδοτεταρτημοριακό εύρος. Το διαχωριστικό ευθύγραμμο τμήμα στο εσωτερικό του πλαισίου διέρχεται από τη διάμεσο των δεδομένων. Το 50% των τιμών των δεδομένων βρίσκονται στο εσωτερικό του πλαισίου ενώ 25% βρίσκεται στην αριστερή απόληξη και το υπόλοιπο 25% στη δεξιά απόληξη.

Το παράδειγμα που ακολουθεί δείχνει πόσο αποτελεσματικά είναι τα διαγράμματα πλαισίου και απολήξεων στο να δίνουν μια συνοπτική γραφική παρουσίαση της κατανομής συχνότητας των δεδομένων και στο να διευκολύνουν συγκρίσεις αρκετών συνόλων δεδομένων.

Παράδειγμα: Μια περιφερειακή οικονομική έκθεση παρουσίασε τον πίνακα που ακολουθεί στον οποίο συνοψίζονται δεδομένα για τον αριθμό των εργαζόμενων σε εταιρείες σε τέσσερις συγγενεύουσες βιομηχανίες: βιομηχανία δέρματος, υφαντουργία, πλεκτοβιομηχανία, βιομηχανία ένδυσης.

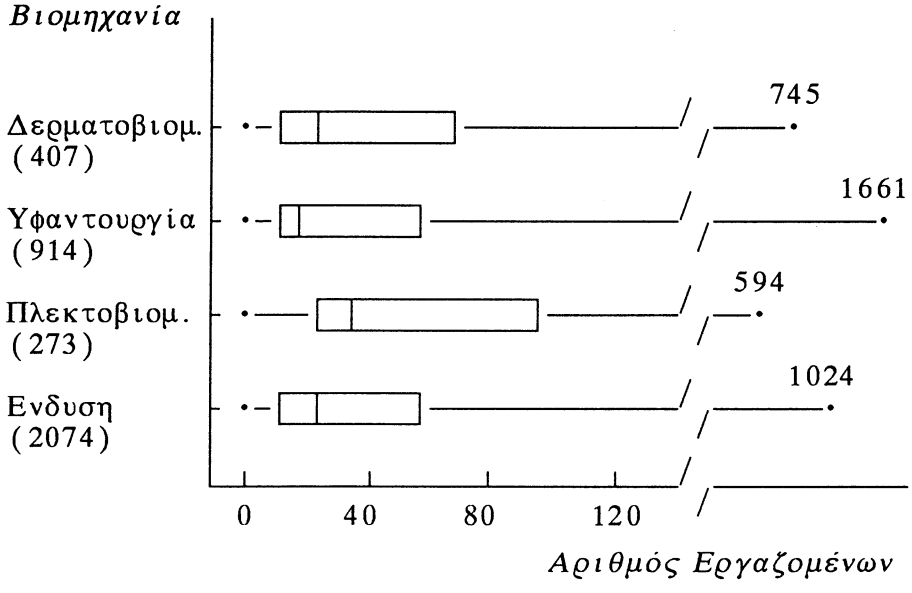
Πίνακας 3.8.1

Περιληπτικά Μέτρα

Αριθμός Εργαζομένων

Βιομηχανικός Κλάδος	X_{min}	Q_1	Q_2	Q_3	X_{max}	Αριθμός Εταιρειών
Δερματο-βιομηχανία	1	9	30	74	745	407
Υφαντουργία	1	7	18	59	1661	914
Πλεκτο-βιομηχανία	2	6	39	107	594	273
Βιομηχανία ένδυσης	1	4	26	59	1024	2074

Η έκθεση παρουσίασε επίσης τα διαγράμματα πλαισίου και απολήξεων που απεικονίζονται στο σχήμα 3.8.1.



Σχήμα 3.8.2

Διαγράμματα πλαισίου - απολήξεων

(Οι μέγιστες τιμές δεν σημειώνονται σύμφωνα με την κλίμακα)

Ας σημειωθεί οτι για τις δεξιές απολήξεις έχει χρησιμοποιηθεί η θλάση για να είναι δυνατό να σημειωθούν οι μέγιστες τιμές (οι οποίες είναι πολύ μεγάλες). Για κάθε έναν από τους βιομηχανικούς κλάδους, το διάγραμμα πλαισίου και απολήξεων παρέχει μια συνεκτική οπτική σύνοψη του εργατικού δυναμικού στη συγκεκριμένη βιομηχανία. Για τη βιομηχανία δέρματος, για παράδειγμα, βλέπουμε από το άνω πέρας του πλαισίου (το άνω τεταρτημόριο) ότι περίπου 75% των εταιρειών αυτού του βιομηχανικού κλάδου απασχολεί λιγότερους από 74 εργαζόμενους. Βλέπουμε, επίσης, ότι η κατανομή του αριθμού των εργαζόμενων σε αυτή τη βιομηχανία παρουσιάζει μεγάλη θετική ασυμμετρία. Αυτό το χαρακτηριστικό προκύπτει από μήκος της απόληξης (προς τη θετική κατεύθυνση) και από το ότι η διάμεσος βρίσκεται πολύ πλησιέστερα στο κάτω τεταρτημόριο παρά στο άνω.

Τα τέσσερα διαγράμματα πλαισίου και απολήξεων παρέχουν ένα αποτελεσματικό τρόπο σύγκρισης του εργατικού δυναμικού στους τέσσερις βιομηχανικούς κλάδους. Οπως βλέπουμε, ο διάμεσος αριθμός εργαζομένων είναι μέγιστος για την πλεκτοβιομηχανία. Βλέπουμε επίσης ότι τα μεγέθη των εταιρειών έχουν μεγάλη μεταβλητότητα στην πλεκτοβιομηχανία όπως προκύπτει από το γεγονός ότι το ενδοτεταρτημοριακό εύρος για τη βιομηχανία αυτή είναι το μεγαλύτερο (παρ'ότι η συνολική έκταση των τιμών της είναι η μικρότερη). Η υφαντουργία και η βιομηχανία ένδυσης τείνουν να έχουν εταιρείες με μικρότερους αριθμούς εργαζομένων και με παρόμοιες κατανομές συχνότητας όσο αφορά τη θέση και τη μεταβλητότητα.

Τέλος οι κατανομές του μεγέθους των εταιρειών έχουν θετική ασυμμετρία και στους τέσσερις βιομηχανικούς κλάδους.

Η συνοπτική περιγραφή με πέντε αριθμούς που μας δίνει το διάγραμμα πλαισίου και απολήξεων μας βοηθά να μελετήσουμε το σχήμα της κατανομής των δεδομενων. Εάν τα δεδομένα ήταν απόλυτα συμμετρικά, είναι προφανές ότι θα ίσχυαν τα παρακάτω:

1. Η απόσταση του Q_1 από τη διάμεσο θα ήταν ίση με την απόσταση της διαμέσου από το Q_3.
2. Η απόσταση του X_{min} από το Q_1 θα ήταν ίση με την απόσταση του Q_3 από το X_{max}.
3. Η διάμεσος και η μέση έκταση θα ήταν ίσες (και προφανώς ίσες με το μέσο των δεδομένων).

Οταν η κατανομή δεν είναι συμμετρική θα έχουμε τα εξής:
Για μια κατανομή με δεξιά ασυμμετρία:
1. Η απόσταση του Q_3 από το X_{max} θα είναι σαφώς μεγαλύτερη από την απόσταση του X_{min} από το Q_1.
2. Διάμεσος < μέση έκταση.
3. Η διάμεσος είναι πιο κοντά στο κάτω τεταρτημόριο παρά στο άνω τεταρτημόριο.

Για κατανομές με αριστερή ασυμμετρία:
1. Η απόσταση του X_{min} από το Q_1 θα είναι σαφώς μεγαλύτερη από την απόσταση του Q_3 από το X_{max}.
2. Μέση έκταση < διάμεσος.
3. Η διάμεσος είναι πιο κοντά στο άνω τεταρτημόριο παρά στο κάτω τεταρτημόριο.

Το διάγραμμα πλαισίου και απολήξεων κατασκευάζεται μερικές φορές με την χρήση της διαμέσου και δύο άλλων μέτρων που ονομάζονται *άξονες* ή *κεντρικά σημεία (hinges)*. Οι άξονες είναι οι τιμές στο μέσο κάθε μισού στα οποία έχει χωρίσει τα δεδομένα η διάμεσος (είναι, δηλαδή, οι "διάμεσοι" κάθε "μισού" των παρατηρήσεων). Οι άξονες μοιάζουν πολύ με τα τεταρτημόρια και στην πραγματικότητα εξυπηρετούν τους ίδιους σκοπούς. Η διαφορά μεταξύ των τιμών ενός άξονα και ενός τεταρτημορίου είναι πολύ μικρή και ελαττώνεται όσο ο αριθμός των μετρήσεων (παρατηρήσεων) αυξάνει.

Αυτός εξ άλλου είναι και ο λόγος για τον οποίο, όπως αναφέρθηκε παραπάνω, σε πολλά εγχειρίδια οι άξονες και τα τεταρτημόρια θεωρούνται ταυτόσημες έννοιες. Εστω ότι d(M) είναι η θέση της διαμέσου. Τότε το d(M) δίνει το "βάθος" (depth) της διαμέσου όπως αυτό μετριέται από κάθε άκρο των διατεταγμένων παρατηρήσεων. Οπως παρατηρήθηκε παραπάνω, οι άξονες είναι οι διάμεσοι των δύο "μισών" των διατεταγμένων παρατηρήσεων. Το "βάθος" επομένως, των αξόνων, όπως αυτό μετριέται από το αντίστοιχο άκρο των δεδομένων δίνεται από την σχέση

$$d(H) = \frac{[d(M)]+1}{2}$$

όπου [d(M)] συμβολίζει το ακέραιο μέρος του d(M).

Οι άξονες δηλαδή των δεδομένων είναι οι τιμές που κατέχουν τη θέση d(H) όπως αυτή μετριέται από κάθε άκρο των διατεταγμένων παρατηρήσεων (μετρήσεων). Για παράδειγμα, αν έχουμε δέκα παρατηρήσεις (N=10), η θέση της διαμέσου είναι d(M) = (10+1)/2 = 5.5. Επομένως, [d(M)] = 5 και d(H) = (5+1)/2 = 3. Αρα οι άξονες είναι οι τιμές των παρατηρήσεων στην τρίτη κατά σειρά θέση, όπως η θέση αυτή καθορίζεται από τα δύο άκρα των 10 διατεταγμένων παρατηρήσεων. Αν η τιμή d(H) περιέχει 1/2, ο κάθε άξονας θα είναι ο μέσος των δύο γειτονικών τιμών στο διατεταγμένο δείγμα.

Η διάχυση (το "άπλωμα") των δεδομένων μετριέται μέσω της διαφοράς των δύο αξόνων που ονομάζεται *Η-άπλωμα (H-spread)*, το οποίο είναι περίπου ίσο με την διαφορά $Q_3 - Q_1$, το ενδοτεταρτημοριακό εύρος, όπως το είχαμε ορίσει.

Θα θεωρούμε ότι μία τιμή των δεδομένων είναι ακραία τιμή (outlier) ανάλογα με τη σχετική θέση που έχει αυτή η τιμή σε σχέση με δύο *συνοριακά σημεία (boundary points)* που ονομάζονται *εσωτερικός και εξωτερικός φράκτης (inner and outer fences).*

Οι εσωτερικοί φράκτες ορίζονται ως εξής:

κάτω εσωτερικός φράκτης =
= κάτω άξονας - 1.5 (Η-άπλωμα),

άνω εσωτερικός φράκτης =
= άνω άξονας + 1.5 (Η-άπλωμα).

Οι εξωτερικοί φράκτες ορίζονται ως εξής:
κάτω εξωτερικός φράκτης =
= κάτω άξονας - 3 (Η-άπλωμα),

άνω εξωτερικός φράκτης =
= άνω άξονας + 3 (Η-άπλωμα).

Οι τιμές των δεδομένων στα άκρα των διατεταγμένων παρατηρήσεων που είναι πλησίον αλλά εξακολουθούν να βρίσκονται μέσα στους εσωτερικούς φράκτες ονομάζονται *γειτονικές τιμές (adjacent values).* Τιμές οι οποίες βρίσκονται μεταξύ ενός εσωτερικού φράκτη και του γειτονικού εξωτερικού φράκτη ονομάζονται "εξωτερικές" ("outside") και θεωρούνται ήπιες ακραίες τιμές (mild outliers). Τιμές που βρίσκονται έξω από τους εξωτερικούς φράκτες ονομάζονται "πολύ απομακρυσμένες" ("far outside") και θεωρούνται εξαιρετικά ακραίες τιμές. Ενα διάγραμμα πλαισίου και απολήξεων συνδυάζει όλες αυτές τις πληροφορίες σε μια γραφική παράσταση με την επισήμανση των ακραίων τιμών και τη γραφική τους παρουσίαση σε σχέση με το κέντρο του συνόλου των δεδομένων.

Σημείωση: Το πλεονέκτημα αυτής της μεθόδου για τον καθορισμό των ακραίων τιμών, σε σύγκριση με τη χρήση της μεθόδου των τυποποιημένων τιμών είναι ότι η διάμεσος και οι άξονες δεν εξαρτώνται από τις τιμές των ακραίων παρατηρήσεων.

Σημείωση: Στις περισσότερες πρακτικές εφαρμογές και στα περισσότερα εγχειρίδια το Η-άπλωμα αντικαθίσταται από την απόσταση Q_3- Q_1 του πρώτου από το τρίτο τεταρτημόριο (ενδοτεταρτημοριακό εύρος) που όπως είπαμε δεν διαφέρει σημαντικά. Γιαυτό και στην

σύνοψη των πέντε αριθμών χρησιμοποιούνται τα Q_1 και Q_3 και όχι οι άξονες.

Παράδειγμα: Εστω ότι έχουμε N=10 μετρήσεις διατεταγμένες κατά αύξουσα σειρά, οι οποίες είναι οι εξής:

0, 0, 1, 2, 2, 3, 3, 3, 4, 15.

Η θέση της διαμέσου, σύμφωνα με όσα έχουμε πει είναι

$$d(M) = \frac{10+1}{2} = 5.5.$$

Επομένως, η διάμεσος είναι ο μέσος της πέμπτης και της έκτης διατεταγμένης παρατήρησης. Επομένως η διάμεσος είναι

$$m = \frac{2+3}{2} = 2.5.$$

Το βάθος των αξόνων είναι

$$d(H) = \frac{[d(M)]+1}{2} = \frac{5+1}{2} = 3.$$

Επομένως, ο κάτω άξονας είναι η παρατήρηση με την τιμή 1 και ο άνω άξονας είναι η παρατήρηση με την τιμή 3.

Το H-άπλωμα είναι 3-1=2. Ο κάτω και ο άνω εσωτερικός φράκτης (lower and upper inner fences) είναι

$$1 - 1.5\ (2) = -2 \quad \text{και} \quad 3 + 1.5 = 6.$$

αντίστοιχα. Ο κάτω και ο άνω εξωτερικός φράκτης (lower and upper outer fences) είναι

$$1 - 3\ (2) = -5 \quad \text{και} \quad 3 + 3 = 6.$$

αντίστοιχα. Οι γειτονικές τιμές που είναι πλησιέστερα, αλλά μεταξύ του κάτω και του άνω εσωτερικού φράκτη είναι το 0 και το 4, αντίστοιχα. Η τιμή 15, η οποία βρίσκεται έξω από τον άνω εξωτερικό φράκτη είναι μια εξαιρετικά ακραία τιμή (extreme outlier).

Είναι δυνατόν να κατασκευάσουμε, όπως είπαμε και στην αρχή, ένα πλαίσιο με τις απολήξεις του που να αποτυπώνει τα όσα αναφέραμε. Τα άκρα του πλαισίου θα αντιστοιχούν στις τιμές των

αξόνων. (Οπως ήδη αναφέραμε, στα περισσότερα εγχειρίδια ως άκρα του πλαισίου λαμβάνονται οι τιμές του πρώτου και του τρίτου τεταρτημορίου). Κατασκευάζεται επίσης μια γραμμή που διαπερνά το πλαίσιο στην τιμή της διαμέσου. Στη συνέχεια κατασκευάζεται μια διακεκομμένη γραμμή που ξεκινά από κάθε ένα από τα άκρα του πλαισίου και φθάνει μέχρι την αντίστοιχη γειτονική τιμή (adjacent value). Οι απολήξεις του διαγράμματος (whiskers) δεν πρέπει να ξεπερνούν σε μήκος μιάμιση φορά το μήκος του πλαισίου από τα δεξιά και αριστερά των κεντρικών σημείων (αντίστοιχα των Q_1 και Q_3).

Οι τιμές των παρατηρήσεων που βρίσκονται μέσα στο διάστημα αυτό, μέσα στο διάστημα δηλαδή που περιέχεται μεταξύ των εσωτερικών φρακτών:

$$[H_1 - 1.5(H_3 - H_1),\ H_3 - 1.5(H_3 - H_1)],$$

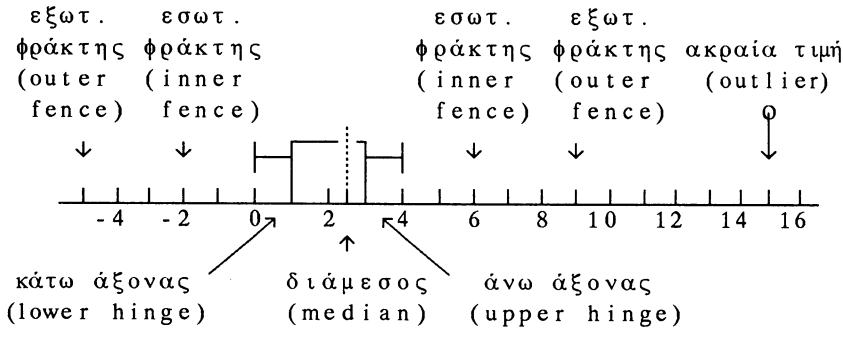
θεωρούνται ακραίες τιμές και σημειώνονται στο διάγραμμα πλαισίου και απολήξεων είτε με έναν αστερίσκο (*) είτε με έναν μικρό κύκλο (ο). Με αστερίσκο σημειώνονται συνήθως οι ακραίες εκείνες τιμές οι οποίες βρίσκονται έξω από τα όρια των εσωτερικών φρακτών αλλά μέσα στα όρια των εξωτερικών φρακτών, δηλαδή δεν βρίσκονται έξω από το διάστημα:

$$[H_1 - 3(H_3 - H_1),\ H_3 - 3(H_3 - H_1)].$$

Οι παρατηρήσεις αυτές λέγονται *ήπιες ακραίες τιμές (mild outliers)* και σημειώνονται με το (*). Οι τιμές που βρίσκονται εκτός των εξωτερικών φρακτών συμβολίζονται με το (ο) και θεωρούνται *εξαιρετικά ακραίες τιμές (extreme outliers).*

Στο διάγραμμα επίσης εμφανίζεται και η κλίμακα του διαγράμματος έτσι ώστε τιμές της διαμέσου, των κεντρικών σημείων και των ακραίων τιμών να είναι δυνατόν να προκύψουν από το διάγραμμα αυτό.

Το σχεδιάγραμμα πλαισίου - απολήξεων για τα δεδομένα του παραδείγματός μας είναι αυτό που δίνεται στο σχήμα 3.8.3.



Σχήμα 3.8.3

Διάγραμμα πλαισίου και απολήξεων

Το διάγραμμα πλαισίου-απολήξεων δίνει έμφαση στο γεγονός ότι οι ακραίες τιμές βρίσκονται μακρυά από τον κεντρικό όγκο του 50% των παρατηρήσεων (μετρήσεων) οι οποίες βρίσκονται μεταξύ των κεντρικών σημείων (hinges). Το διάγραμμα αυτό δίνει επίσης την ένδειξη ότι τα συγκεκριμένα δεδομένα έχουν θετική ασυμμετρία, δεδομένου ότι η διάμεσος δεν απέχει εξίσου από τα κεντρικά σημεία αλλά βρίσκεται πλησιέστερα προς το άνω κεντρικό σημείο (upper hinge).

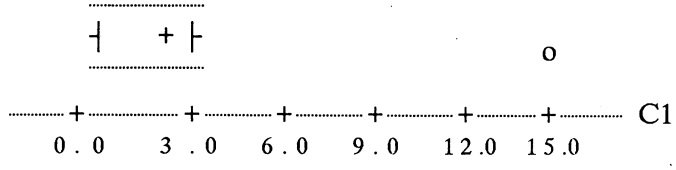
Διάγραμμα πλαισίου και απολήξεων με το MINITAB και το SAS

MINITAB

Με δεδομένο ότι τα στοιχεία του δείγματος έχουν εισαχθεί στη μεταβλητή C1, η εντολή για την κατασκευή του διαγράμματος πλαισίου-απολήξεων στο MINITAB είναι

MTB > BOXPLOT C1

Το αποτέλεσμα της εντολής αυτής δίνεται στο σχήμα 3.8.4 που ακολουθεί.



Σχήμα 3.8.4

Διάγραμμα πλαισίου - απολήξεων με το MINITAB

Παρατηρούμε ότι το διάγραμμα πλαισίου απολήξεων που κατασκευάζεται με το MINITAB είναι παρόμοιο με αυτό που κατασκευάσαμε προηγουμένως.

SAS

Οσο αφορά το στατιστικό πακέτο SAS, η εντολή για την κατασκευή του διαγράμματος πλαισίου-απολήξεων είναι η ίδια με αυτή που χρειάζεται για την κατασκευή του διαγράμματος μίσχου-φύλλου. Πράγματι το SAS παρουσιάζει και τα δύο αυτά διαγράμματα ταυτόχρονα με την ίδια εντολή. Εκτός από τα δύο αυτά διαγράμματα η συγκεκριμένη εντολή έχει ως αποτέλεσμα την παρουσίαση από το πρόγραμμα διαφόρων περιγραφικών μέτρων, εκατοστιαίων σημείων και ακραίων τιμών. Η σχετική εντολή, όπως έχουμε ήδη δει, είναι η

```
PROC UNIVARIATE PLOT;
VAR X;
```

Και στο πακέτο SAS το διάγραμμα πλαισίου απολήξεων εμφανίζεται οριζόντια όπως και στο MINITAB.

3.9 Μέτρα ασυμμετρίας και κύρτωσης

Οπως ήδη αναφέρθηκε στην παράγραφο 3.2, μια μη συμμετρική κατανομή συχνότητας ενός συνόλου δεδομένων ονομάζεται **ασύμμετρη, λοξή** ή **στρεβλή**. Ενας τρόπος για να μελετήσουμε την ασυμμετρία των

δεδομένων είναι, όπως νωρίτερα υποδείχθηκε, να συγκρίνουμε τις τιμές της επικρατούσας τιμής, της διαμέσου και του μέσου: Η επικρατούσα τιμή είναι η θέση της κατακόρυφης κλίμακας με την μεγαλύτερη συγκέντρωση παρατηρήσεων. Η διάμεσος είναι η τιμή κάτω από την οποία βρίσκονται οι μισές περίπου παρατηρήσεις με τις υπόλοιπες μισές πάνω από αυτήν. Αντίθετα, ο μέσος τείνει να βρίσκεται προς την κατεύθυνση των απομακρυσμένων (ακραίων) παρατηρήσεων, δηλαδή πλησιέστερα στην ουρά της κατανομής, όπως εξ άλλου φαίνεται στο σχήμα 3.2.1. Η διάμεσος βρίσκεται πάντα μεταξύ μέσου και επικρατούσας τιμής. Ετσι, όταν ο μέσος διαφέρει ουσιωδώς από τη διάμεσο και την επικρατούσα τιμή, έχουμε ενδείξεις ασυμμετρίας στο σύνολο των δεδομένων.

Πέρα από αυτή την άτυπη ανάλυση της συμμετρίας ή ασυμμετρίας ενός συνόλου δεδομένων μπορούμε να κάνουμε χρήση κάποιων άμεσων αριθμητικών μέτρων. Τέτοια μέτρα σχετίζονται με τη διασπορά των παρατηρήσεων. Ο Karl Pearson πρότεινε ως μέτρο ασυμμετρίας τη συνάρτηση

$$\frac{\text{μέσος - επικρατούσα τιμή}}{\text{τυπική απόκλιση}}$$

που βέβαια έχει το μειονέκτημα ότι προυποθέτει γνώση της επικρατούσας τιμής η οποία δεν είναι πάντα εύκολο να προσδιορισθεί, κυρίως στην περίπτωση ομαδοποιημένων δεδομένων. Λόγω αυτής της δυσκολίας και δεδομένου ότι για μια ελαφρά ασύμμετρη κατανομή ισχύει κατά προσέγγιση η σχέση

$$\text{μέσος - επικρατούσα τιμή} = 3(\text{μέσος - διάμεσος}),$$

συχνά χρησιμοποιείται ο συντελεστής

$$\frac{3(\text{μέσος - διάμεσος})}{\text{τυπική απόκλιση}}.$$

Τα παραπάνω δύο μέτρα μπορούν εναλλακτικά να υπολογισθούν μέσω των εννοιών των *ροπών (moments).*

Η έννοια της ροπής προέρχεται από τη Μηχανική και αναφέρεται στη μέτρηση της τάσης μιας δύναμης να παράγει περιστροφή. Η τάση αυτή εξαρτάται από το μέγεθος της δύναμης και από την απόσταση του σημείου στο οποίο ασκείται η δύναμη από κάποια αρχή.

Ορισμός: Ως *ροπή r τάξεως γύρω από τον μέσο* μιας σειράς n παρατηρήσεων ορίζεται η στατιστική συνάρτηση

$$m_r = \sum_{i=1}^{n} (X_i - \overline{X})^r / n.$$

Δεδομένου ότι

$$m_2 = \sum_{i=1}^{n} (X_i - \overline{X})^2 / n$$

συνεπάγεται ότι η ροπή β' τάξεως (ή δεύτερη ροπή) γύρω από τον μέσο είναι η διασπορά του δείγματος.

Εξ άλλου ορίζουμε ως *ροπή r τάξεως* (ή *ροπή r τάξεως γύρω από το 0)* τη στατιστική συνάρτηση

$$m'_r = \sum_{i=1}^{n} X_i^r / n.$$

Είναι προφανές ότι ο μέσος ενός δείγματος είναι η ροπή πρώτης τάξης m'_1 του δείγματος.

Οι ροπές ενός δείγματος χρησιμοποιούνται ως εκτιμήτριες των αντιστοίχων ροπών του πληθυσμού από τον οποίο προέρχονται. Για τον ορισμό και τις ιδιότητες των ροπών του πληθυσμού ο αναγνώστης παραπέμπεται σε σχετικό εγχειρίδιο Πιθανοτήτων (π.χ. *Πιθανότητες και Στοιχεία Στοχαστικών Ανελίξεων* των συγγραφέων).

Χρησιμοποιώντας λοιπόν την έννοια της ροπής, οι δύο

συντελεστές ασυμμετρίας που ορίσθηκαν παραπάνω μπορούν να εκφρασθούν μέσω των στατιστικών συναρτήσεων

$$\frac{m_3}{S^3} \text{ και } \frac{m_4}{S^4}$$

όπου προφανώς

$$m_3 = \sum_{i=1}^{n}(X_i - \bar{X})^3/n,$$

$$m_4 = \sum_{i=1}^{n}(X_i - \bar{X})^4/n.$$

Η στατιστική συνάρτηση m_3/S_3 συχνά χρησιμοποιείται ως μέτρο ασυμμετρίας και ονομάζεται **τυποποιημένος συντελεστής ασυμμετρίας (standardized coefficient of skewness).**

Είναι προφανές από τον ορισμό αυτού του συντελεστή μέσω της m_3 ότι μεγάλες τιμές των αποκλίσεων $(X_i - \bar{X})$ είναι καθοριστικές της τιμής του αριθμητή (αφού εξ άλλου αυτές υψώνονται στην τρίτη δύναμη). Αν λοιπόν οι μεγάλες θετικές αποκλίσεις είναι επικρατέστερες, τότε η m_3 θα είναι θετική. Αντίστοιχα, αν επικρατούν οι μεγάλες αρνητικές αποκλίσεις, η m_3 θα είναι αρνητική. Επειδή οι μεγάλες αποκλίσεις συνδέονται με τη μακρυά ουρά μιας κατανομής, βλέπουμε και από το σχήμα 3.2.1 ότι η m_3, και κατά συνέπεια και ο συντελεστής ασυμμετρίας, θα έχει θετική ή αρνητική τιμή ανάλογα με το εάν η κατεύθυνση της ασυμμετρίας είναι θετική (δεξιά) ή αρνητική (αριστερή). Αν οι παρατηρήσεις στο σύνολο δεδομένων κατανέμονται συμμετρικά γύρω από τον μέσο τους, η m_3 - και επομένως και ο συντελεστής ασυμμετρίας - θα είναι 0. Τα παραπάνω συνοψίζονται από τη σχέση που ακολουθεί.

$$\frac{m_3}{S^3} \begin{cases} > 0 & \text{θετική ή δεξιά ασυμμετρία} \\ = 0 & \text{συμμετρία} \\ < 0 & \text{αρνητική ή αριστερή ασυμμετρία} \end{cases}$$

Θα πρέπει να σημειωθεί ότι ο συντελεστής ασυμμετρίας που θεωρήσαμε δεν είναι κατάλληλος για σύνολα δεδομένων με ακραίες τιμές.

Σημείωση: Συχνά ο παραπάνω συντελεστής υπολογίζεται με βάση τις τιμές των

$$m_3^* = \frac{n}{n-1} m_3 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^3$$

και

$$S^* = \sqrt{\frac{n}{n-1} S^2} = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2}.$$

Η στατιστική συνάρτηση $\frac{m_4}{S^4} - 3$ ονομάζεται **συντελεστής κύρτωσης (coefficient of kurtosis)** και χρησιμοποιείται ως μέτρο της οξύτητας της κορυφής της κατανομής των δεδομένων.

Στην περίπτωση που ο συντελεστής αυτός είναι 0 η κατανομή των δεδομένων έχει μέτρια κύρτωση και ονομάζεται *μεσόκυρτη (mesokurtic)*. Αν είναι θετικός ή αρνητικός αριθμός η κατανομή έχει οξεία ή ελαφρά κύρτωση και ονομάζεται *λεπτόκυρτη (leptokurtic)* ή *πλατύκυρτη (platykurtic)* αντίστοιχα.

Και στην περίπτωση του συντελεστή αυτού χρησιμοποιείται πολλές φορές η

$$m_4^* = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^4$$

αντί της m_4 και η S^* αντί της S.

Παράδειγμα: Εστω ότι έχουμε το εξής σύνολο δεδομένων:

13 16 14 7 13 7 10 9 6 7 10 6 5 15 7

11 16 17 13 12 9 11 6 15 9 5 8 6 12 9

Εύκολα μπορεί να διαπιστωθεί ότι ο μέσος των παραπάνω μετρήσεων είναι 10.13, η επικρατούσα τιμή είναι 7, η διάμεσος είναι 9.5 και η τυπική απόκλιση, όπως δίνεται από την S^*, είναι 3.63. Τότε η τιμή του μέτρου ασυμμετρίας που ορίσθηκε από τον Pearson είναι

$$\frac{\text{μέσος - επικρατούσα τιμή}}{\text{τυπική απόκλιση}} = \frac{10.13 - 7}{3.63} = 0.86.$$

Η τιμή αυτή υποδηλώνει ελαφρά θετική ασυμμετρία αφού αντανακλά το γεγονός ότι ο μέσος βρίσκεται 0.86 τυπικές αποκλίσεις δεξιότερα από την επικρατούσα τιμή.

Στο ίδιο συμπέρασμα μας οδηγεί η τιμή του ισοδύναμου συντελεστή

$$\frac{3(\text{μέσος - διάμεσος})}{\text{τυπική απόκλιση}} = \frac{10.13 - 9.5}{3.63} = 0.52$$

αφού, σύμφωνα με αυτή, ο μέσος βρίσκεται 0.52 τυπικές αποκλίσεις δεξιότερα από τη διάμεσο.

Επίσης, η τιμή της τρίτης ροπής των δεδομένων γύρω από τον μέσο τους είναι

$$m_3^* = \frac{1}{29}\left[(13-10.13)^3+(16-10.13)^3+\ldots+(9-10.13)^3\right] = 14.176.$$

Τότε, ο τυποποιημένος συντελεστής ασυμμετρίας έχει την τιμή

$$\frac{m_3^*}{s^{*3}} = \frac{33.81}{(3.63)^3} = 0.296$$

που οδηγεί και πάλι στο συμπέρασμα ότι η κατανομή συχνότητας των δεδομένων είναι ασύμμετρη προς τα δεξιά.

Τέλος, η τιμή της τέταρτης ροπής των δεδομένων γύρω από τον μέσο τους είναι

$$m_4^* = \frac{1}{29}\left[(13\text{-}10.13)^4 + (16\text{-}10.13)^4 + \ldots + (9\text{-}10.13)^4\right] = 312.001.$$

Επομένως, ο συντελεστής κύρτωσης έχει την τιμή

$$\frac{m_4}{S^4} = \frac{306.72}{(3.63)^3} - 3 = -1.179$$

η οποία συνεπάγεται ότι η κατανομή συχνότητας των δεδομένων είναι πλατύκυρτη.

ΑΣΚΗΣΕΙΣ

3.1 Τα δεδομένα που ακολουθούν αναφέρονται στον αριθμό των ημερών στις οποίες απουσίαζε κάθε ένας από 15 υπαλλήλους ενός γραφείου για μια χρονική περίοδο ενός μηνός.

0, 1, 1, 3, 0, 0, 2, 5, 0, 1, 1, 2, 0, 1, 1.

Να υπολογισθεί ο μέσος, η διάμεσος και η επικρατούσα τιμή του αριθμού των ημερών απουσίας.

3.2 Τα δεδομένα που ακολουθούν αναφέρονται στον αριθμό των ελαττωματικών προιόντων που βρέθηκαν σε κάθε μια από τις οκτάωρες βάρδιες τις εργάσιμες μέρες μιας εβδομάδας.

	ΔΕ	ΤΡ	ΤΕ	ΠΕ	ΠΑ
Βάρδια 1	24	17	35	15	19
Βάρδια 2	21	13	15	20	18

Να βρεθεί ο μέσος αριθμός ελαττωματικών αντικειμένων ανά βάρδια.

3.3 Ποιά είναι τα χαρακτηριστικά ενός συνόλου δεδομένων για τα οποία ο μέσος, η διάμεσος και η επικρατούσα τιμή ταυτίζονται;

3.4 Οι ηλικίες των υπαλλήλων σε ένα υποκατάστημα γρήγορου φαγητού είναι οι εξής:

19, 19, 65, 20, 21, 18, 20.

α) Να υπολογισθεί ο μέσος, η διάμεσος και η επικρατούσα τιμή των ηλικιών.

β) Πώς θα επηρεάζονταν τα τρία αυτά μέτρα κεντρικής θέσης εάν ο μεγαλύτερος από τους τρεις υπαλλήλους έπαυε πια να δουλεύει στο κατάστημα αυτό;

3.5 Σε πολλούς αθλητικούς αγώνες όπου χρησιμοποιείται η γνώμη κάποιων κριτών προκειμένου να καθορισθεί η απόδοση των αθλητών, συχνά απαλείφεται η μικρότερη και η μεγαλύτερη βαθμολογία πριν να υπολογισθεί η μέση βαθμολογία του αθλητή, έτσι ώστε να ελαχιστοποιηθεί η επίδραση των ακραίων τιμών στον υπολογισμό του μέσου. Εστω ότι δύο αθλητές Α και Β έλαβαν τις ακόλουθες βαθμολογίες:

Α : 6.0, 7.0, 7.25, 7.25, 7.5, 7.5, 7.5
Β : 7.0, 7.0, 7.0, 7.25, 7.5, 7.5, 8.5

α) Να συγκριθεί η απόδοση των αθλητών Α και Β με βάση τη μέση βαθμολογία τους πριν και μετά την απαλοιφή των ακραίων βαθμολογιών.
β) Να επαναληφθεί η προηγούμενη διαδικασία για τη διάμεσο στη θέση του μέσου.

3.6 Ενας προπονητής στίβου πρόκειται να διαλέξει έναν από δύο αθλητές ο οποίος θα πάρει μέρος σε αγώνα ταχύτητας 100 μέτρων. Ο προπονητής θα πάρει την απόφασή του με βάση τα αποτελέσματα που θα πετύχουν οι δύο αθλητές σε πέντε αγώνες που θα γίνουν μεταξύ τους σε διάστημα μιας ώρας με διάλειμμα 15 λεπτών μεταξύ κάθε αγώνα. Οι ώρες που πέτυχαν οι δύο αθλητές (σε δευτερόλεπτα) δίνονται στον πίνακα που ακολουθεί.

Αθλητής	1	2	3	4	5
Α	12.1	12.0	12.0	16.8	12.1
Β	12.3	12.4	12.4	12.5	12.4

α) Με βάση τα δεδομένα αυτά ποιον από τους δύο αθλητές θα πρέπει να διαλέξει ο προπονητής και γιατί;
β) Θα έπρεπε, κατά τη γνώμη σας, η επιλογή να είναι

διαφορετική αν ο προπονητής γνώριζε ότι ο αθλητής Α έπεσε στο ξεκίνημα του τέταρτου αγώνα; Γιατί;

γ) Να συζητήσετε τις διαφορές στις έννοιες του μέσου και της διαμέσου ως μέτρων κεντρικής τάσης και πώς οι διαφορές αυτές σχετίζονται με τα ερωτήματα (α) και (β).

3.7 Ενας κατασκευαστής μπαταριών εξέτασε ένα δείγμα από 13 μπαταρίες από την παραγωγή μιας συγκεκριμένης μέρας και τις χρησιμοποίησε συνεχώς μέχρις ότου εξαντλήθηκαν. Ο αριθμός των ωρών που οι μπαταρίες αυτές διήρκεσαν δίνεται στη συνέχεια:

342, 426, 317, 545, 264, 451, 1049, 631, 512, 266, 492, 562, 298.

α) Να υπολογισθεί ο μέσος και η διάμεσος των δεδομένων αυτών. Παρατηρώντας την κατανομή των χρόνων της διάρκειας ζωής, ποιό από τα δύο μέτρα θεωρείτε ότι είναι το πιο κατάλληλο και γιατί;

β) Συζητήστε τους τρόπους με τους οποίους οι παραπάνω πληροφορίες θα είναι χρήσιμες στον κατασκευαστή.

γ) Δώστε την σύνοψη των πέντε αριθμών για δεδομένα αυτά.

δ) Να κατασκευάσετε το διάγραμμα πλαισίου και απολήξεων και να περιγράψετε το σχήμα του.

3.8 Μια εταιρεία που πουλά βιβλία ταχυδρομικώς, επιλέγει τυχαία ένα δείγμα 20 λογαριασμών πελατών της. Τα ποσά που χρωστούν οι πελάτες αυτοί είναι τα εξής: (σε χιλ. δρχ.)

4, 18, 11, 7, 7, 10, 5, 33, 9, 12,
3, 11, 10, 6, 26, 37, 15, 18, 10, 21.

α) Να υπολογισθεί ο μέσος, η διάμεσος, η επικρατούσα τιμή και μέση η έκταση.

β) Αν συνολικά στην εταιρεία αυτή υπήρχαν 350 λογαριασμοί με χρεωστικό, να χρησιμοποιήσετε τον μέσο που υπολογίσατε προηγουμένως για να εκτιμήσετε το συνολικό ποσό που οι πελάτες χρωστάνε στην εταιρεία.

γ) Να γράψετε ένα σημείωμα που θα στέλνατε στο διευθυντή της εταιρείας αν εργαζόσασταν σε αυτή με βάση τις διαπιστώσεις σας από το προηγούμενο ερώτημα.

δ) Διατυπώστε την άποψή σας για τη χρησιμότητα που θα είχαν οι πληροφορίες που στείλατε στο διευθυντή της εταιρείας.

ε) Να κατασκευάσετε το διάγραμμα πλαισίου και απολήξεων και να περιγράψετε το σχήμα του.

3.9 Υποθέστε ότι έχετε μια στήλη σε ένα περιοδικό για φορητούς υπολογιστές. Εστω ότι οι τιμές για τα διάφορα μοντέλα τέτοιων υπολογιστών που αναφέρονται στο άρθρο σας είναι οι εξής: (σε χιλ. δρχ.)

575, 259, 1049, 340, 499, 675, 599, 450, 649,
475, 520, 550, 398, 490, 560, 625, 875, 749.

α) Να προσδιορίσετε τον μέσο, τη διάμεσο και την έκταση των παραπάνω δεδομένων.

β) Αν στην αγορά εμφανίζονταν τρία νέα μοντέλα με τιμές (σε χιλ. δρχ.) 345, 375 και 355, ποιές θα ήταν οι νέες τιμές του μέσου και της διαμέσου; Ποιό από τα δύο μέτρα μεταβάλλεται περισσότερο και γιατί;

γ) Να γράψετε το κομμάτι για το περιοδικό με βάση τα παραπάνω δεδομένα.

δ) Να κατασκευάσετε το διάγραμμα πλαισίου και απολήξεων για τα δεδομένα αυτά και να περιγράψετε το σχήμα του.

3.10 Τις τελευταίες δέκα μέρες ενός συγκεκριμένου μήνα, το τραίνο που εκτελούσε μια συγκεκριμένη διαδρομή από την πόλη Α στην πόλη Β έφθανε καθυστερημένα στον προορισμό του κατά τους ακόλουθους χρόνους (σε λεπτά). (Ενας αρνητικός αριθμός σημαίνει ότι το τραίνο έφθασε στην ώρα του τόσα λεπτά νωρίτερα όσο ο αριθμός).

-3, 6, 4, 10, -4, 124, 2, -1, 4, 1.

α) Αν δουλεύατε στον ΟΣΕ ως στατιστικός, ποιο από τα συνοπτικά στατιστικά μέτρα θα χρησιμοποιούσατε για να αποδείξετε ότι ο οργανισμός παρέχει ικανοποιητικές υπηρεσίες;
β) Αν εργαζόσαστε σε ένα τηλεοπτικό κανάλι το οποίο προετοίμαζε μια εκπομπή με σκοπό να δείξει ότι ο οργανισμός δεν παρέχει ικανοποιητικές υπηρεσίες, ποιό συνοπτικό μέτρο θα χρησιμοποιούσατε και γιατί;
γ) Αν θέλατε να είστε αντικειμενικός και αμερόληπτος στην αξιολόγηση της απόδοσης του ΟΣΕ, ποιό μέτρο θα χρησιμοποιούσατε και γιατί; (Αυτό είναι το πιο δύσκολο μέρος, διότι δεν μπορείτε να απαντήσετε χωρίς να κάνετε πρόσθετες υποθέσεις για το σχετικό κόστος που συνεπάγεται η καθυστέρηση σε διαφορετικούς χρόνους καθυστέρησης. Η επιστημονική περιοχή της Θεωρίας Αποφάσεων βοηθά στην αντιμετώπιση τέτοιων προβλημάτων όταν το κόστος αυτό είναι σαφώς προσδιορισμένο).

3.11 Ενας καταναλωτής της ΕΥΔΑΠ ενδιαφέρεται να μελετήσει την κατανάλωση νερού της οικογένειάς του με βάση τα στοιχεία της τελευταίας εξαετίας. Οπως προκύπτει από τους λογαριασμούς της ΕΥΔΑΠ, η κατανάλωση της οικογένειας αυτής ανά τρίμηνο, αρχίζοντας από το πρώτο τρίμηνο του 1988 (λογαριασμός από 19/12/87 μέχρι 18/3/88) δίνεται (σε m^3) στον πίνακα που ακολουθεί.

	Τρίμηνο			
	1	2	3	4
1988:	48,	45,	51,	46
1989:	41,	52,	39,	51
1990:	43,	52,	30,	34
1991:	32,	31,	28,	37
1992:	40,	50,	25,	48
1993:	26,	31,	30	

α) Να υπολογισθούν ο μέσος, η διάμεσος και η επικρατούσα τιμή για τα συνολικά δεδομένα, για κάθε έτος κατανάλωσης καθώς και για τα δεδομένα που αφορούν την κατανάλωση των πρώτων τριμήνων των παραπάνω ετών, των δεύτερων τριμήνων, των τρίτων τριμήνων, και των τέταρτων τριμήνων.

β) Να υπολογισθούν, για τις αντίστοιχες περιόδους, η μέση απόλυτη απόκλιση, η διακύμανση, η έκταση και ο συντελεστής μεταβλητότητας, όπου αυτός έχει έννοια.

γ) Διατυπώσατε τις απόψεις σας όσο αφορά την κατανάλωση νερού της συγκεκριμένης οικογένειας και για το κατά πόσο από την εποχή που ελήφθησαν τα μέτρα για τη λειψυδρία επηρεάσθηκε η κατανάλωση νερού της οικογένειας αυτής.

3.12 Προκειμένου να εκτιμήσει πόσο νερό θα χρειασθεί για την ύδρευση μιας πόλης για την επόμενη δεκαετία, το Δημοτικό Συμβούλιο της πόλης ζήτησε να πληροφορηθεί την ποσότητα του νερού που χρησιμοποιούν οι οικογένειες της πόλης. Ενα δείγμα 15 οικογενειών έδωσε τα εξής στοιχεία για την ετήσια κατανάλωση νερού (σε m^3) για το χρόνο που πέρασε.

112, 215, 164, 197, 146, 169, 322, 182,
131, 238, 183, 155, 188, 227, 140.

α) Να υπολογισθεί η μέση ποσότητα νερού και η διάμεσος της ποσότητας του νερού που χρησιμοποιείται από κάθε οικογένεια.

β) Εστω ότι το Δημοτικό Συμβούλιο υπολογίζει πως σε δέκα χρόνια στην πολή αυτή θα ζουν 45000 οικογένειες. Ποιά είναι η ποσότητα νερού ανά χρόνο που θα απαιτείται αν ο αριθμός των κυβικών κατανάλωσης ανά οικογένεια παραμένει σταθερός;

γ) Με ποιούς τρόπους μπορούν να χρησιμοποιηθούν από το Δημοτικό Συμβούλιο οι πληροφορίες που παρέχουν οι απαντήσεις στα ερωτήματα (α) και (β); Εξηγείστε.

δ) Να κατασκευασθεί το διάγραμμα πλαισίου-απολήξεων και να περιγραφεί το σχήμα του.

ε) Γιατί θα έπρεπε το Δημοτικό Συμβούλιο να χρησιμοποιήσει τα δεδομένα μιας δειγματοληψίας μάλλον παρά να μετρήσει τη συνολική κατανάλωση της πόλης; (Σκεφθείτε τί είδους καταναλωτές δεν έχουν περιληφθεί στη διαδικασία της εκτίμησης).

3.13 Ο ημερήσιος αριθμός εισαγωγών στην πτέρυγα επειγόντων περιστατικών ενός νοσοκομείου στη βραδυνή βάρδια τα τελευταία έντεκα βράδυα ήταν ο εξής:

2, 3, 2, 9, 0, 3, 5, 3, 1, 4, 2.

α) Να υπολογισθεί ο μέσος αριθμός και η διάμεσος των εισαγωγών ανά βράδυ.

β) Επιλέξτε ένα κατάλληλο μέτρο μεταβλητότητας για τα δεδομένα αυτά και υπολογίστε το.

γ) Κατασκευάστε το διάγραμμα πλαισίου - απολήξεων και υπολογίστε τα πέντε μέτρα θέσης και απόκλισης που αναφέρονται σε αυτό.

3.14 Σε μια στρατιωτική μονάδα υπηρετούν 125 αξιωματικοί, των οποίων ο μέσος χρόνος υπηρεσίας είναι 8.5 χρόνια και 23 πολίτες, των οποίων ο μέσος χρόνος υπηρεσίας είναι 6.1 χρόνια. Να υπολογισθεί ο μέσος χρόνος υπηρεσίας ολόκληρου του προσωπικού (και των 148 ατόμων) της μονάδας αυτής.

3.15 Σε μια εκπομπή για το δημογραφικό πρόβλημα, ο συντονιστής της συζήτησης, ακούγοντας ότι η επικρατούσα τιμή του αριθμού των παιδιών των οικογενειών είναι μηδέν είπε " ο ρυθμός των γεννήσεων έχει πέσει τόσο χαμηλά ώστε η πλειοψηφία των οικογενειών στη χώρα δεν έχει καθόλου παιδιά". Σχολιάστε.

3.16 Σε ένα άρθρο που αναφερόταν στα αυτοκίνητα, μια εφημερίδα έγραφε το εξής : "το μέσο νοικοκυριό στην πόλη μας έχει στην κατοχή του 1.79 αυτοκίνητα". Ποιός είναι ο μέσος που χρησιμοποιήθηκε για να βγεί το συμπέρασμα αυτό; Υπάρχει κάποιο άλλο μέτρο θέσης το οποίο θα είχε περισσότερο νόημα στην περίπτωση αυτή; Συζητήστε τις απαντήσεις σας.

3.17 Ενας αναλυτής στοιχείων σε μια παρουσίαση έκανε την εξής παρατήρηση: "όταν ένα σύνολο δεδομένων έχει μια ή περισσότερες ακραίες τιμές που προέρχονται ίσως από κάποιο λάθος στη συλλογή των δεδομένων, προτιμώ να χρησιμοποιώ τη διάμεσο ως μέτρο θέσης αντί του μέσου". Εξηγείστε γιατί ο αναλυτής παίρνει αυτή τη θέση.

3.18 Μια καθημερινή εφημερίδα δίνει στις οικονομικές σελίδες της τις τιμές συναλλάγματος για τα κυριότερα ξένα συναλλάγματα σε σχέση με τη δραχμή. Σε ένα φύλλο της έδωσε τις παρακάτω ποσοστιαίες μεταβολές για τις τιμές των ξένων συναλλαγμάτων σε σχέση με τη δραχμή για τις τελευταίες έξι εβδομάδες:

-3.5 1.8 2.1 2.1 5.1 6.6 9.2

α) Να κατασκευάσετε το διάγραμμα πλαισίου - απολήξεων για τα δεδομένα αυτά.

β) Να καθορίσετε τη σύνοψη των πέντε σημείων που αναφέρονται σε μέτρα θέσης και απόκλισης.

3.19 Μια αθλητική εφημερίδα κάνει σύγκριση των τμημάτων κατάδυσης δύο αθλητικών ομάδων Α και Β. Η ομάδα Α έχει 9 καταδύτες των οποίων οι βαθμολογίες σε μια πρόσφατη αθλητική συνάντηση είχαν έκταση 2.3. Η ομάδα Β έχει 9 καταδύτες των οποίων οι βαθμολογίες στους ίδιους αγώνες είχαν έκταση 1.6. Ο δημοσιογράφος της εφημερίδας ισχυρίζεται ότι η ομάδα Β έχει περισσότερο ομοιόμορφη απόδοση όσο αφορά τους καταδύτες δεδομένου ότι οι βαθμολογίες τους έχουν μικρότερη έκταση. Συμφωνείτε με την άποψή αυτή του

δημοσιογράφου; Εξηγείστε την άποψή σας.

3.20 Οι μηνιαίοι ρυθμοί απόδοσης (monthly rates of return) μιας μετοχής για τους τέσσερις τελευταίους μήνες, εμφανίζονται στον πίνακα που ακολουθεί. Οι ρυθμοί απόδοσης είναι βασισμένοι αποκλειστικά σε μεταβολές των τιμών των μετοχών (και αγνοούν οποιεσδήποτε πληρωμές μερισμάτων).

Μήνας	1	2	3	4
Ρυθμός απόδοσης (%)	1.8	-0.6	1.3	2.6

α) Ο ρυθμός απόδοσης για τον πρώτο μήνα, 1.8%, αποτελεί ένδειξη ότι ο λόγος της τιμής της μετοχής στο κλείσιμο για τον μήνα 1 προς της τιμή κλεισίματος της μετοχής για τον προηγούμενο μήνα (μήνας 0) ήταν 1.018. Να καθορισθούν οι αντίστοιχοι λόγοι τιμών για τους υπόλοιπους μήνες (δηλαδή ο λόγος της τιμής κλεισίματος για κάθε μήνα προς την τιμή κλεισίματος του προηγούμενου μήνα).

β) Να υπολογισθεί ο γεωμετρικός μέσος των τεσσάρων λόγων τιμών που υπολογίσθηκε στο ερώτημα (α). Με βάση το γεωμετρικό μέσο, ποιός ήταν ο μέσος ρυθμός απόδοσης της μετοχής αυτής για τους τελευταίους τέσσερις μήνες;

γ) Η τιμή κλεισίματος της μετοχής για το μήνα 0 ήταν 5 (σε χιλ. δρχ.). Να χρησιμοποιηθεί η τιμή αυτή και ο γεωμετρικός μέσος που καθορίσθηκε στην ερώτηση (β) για να υπολογισθεί η τιμή κλεισίματος της μετοχής για τον μήνα 4.

3.21 Τα στοιχεία για ατυχήματα σε μια βιομηχανική μονάδα εμφανίζονται στον πίνακα που ακολουθεί. Ο πίνακας αυτός δείχνει τον ετήσιο αριθμό ωρών εργασίας ανά ατύχημα κατά τη διάρκεια μιας περιόδου τεσσάρων ετών στην οποία πραγματοποιήθηκαν τέσσερα εκατομμύρια ώρες εργασίας ανά έτος.

Ετος	1	2	3	4
Ωρες ανά ατύχημα	12535	10810	11691	14735

α) Να υπολογισθεί ο αρμονικός μέσος αριθμός των ωρών ανά ατύχημα για την περίοδο των τεσσάρων ετών.

β) Είναι ο αρμονικός μέσος ένα κατάλληλο μέτρο θέσης για την περίπτωση αυτή; (Υπόδειξη: Πόσα ατυχήματα συνέβησαν κατά τη διάρκεια της περιόδου των τεσσάρων ετών στην οποία καταγράφηκε εργασία 16 εκατομμυρίων ωρών;)

γ) Θα επηρεαζόταν η απάντησή σας στο ερώτημα (β) αν ο ετήσιος αριθμός των ωρών εργασίας δεν ήταν σταθερός κατά τη διάρκεια της περιόδου των τεσσάρων ετών; Εξηγείστε.

3.22 Μια πολτοποιητική μηχανή έχει πέντε αντλίες του ίδιου τύπου, κάθε μια από τις οποίες λειτούργησε για τον ίδιο περίπου χρόνο το τελευταίο έτος. Ο αριθμός των ωρών λειτουργίας ανά επισκευή για τις πέντε αντλίες κατά τη διάρκειά του έτους δίνεται στον πίνακα που ακολουθεί.

	Αντλία 1	Αντλία 2	Αντλία 3	Αντλία 4	Αντλία 5
Ωρες ανά επισκευή	617	866	703	452	620

α) Να υπολογισθεί ο αρμονικός μέσος αριθμός ωρών λειτουργίας ανά επισκευή για τις πέντε αντλίες. Ποιά είναι η μονάδα μέτρησης στην οποία εκφράζεται ο αρμονικός αυτός μέσος;

β) Εάν η αντλία 1 λειτούργησε για 8640 ώρες το τελευταίο έτος, πόσες επισκευές θα πρέπει να είχε κατά τη διάρκεια του έτους αυτού;

γ) Εάν ο συνολικός αριθμός ωρών λειτουργίας για το τελευταίο έτος, για όλες τις πέντε αντλίες διαιρείτο με το συνολικό αριθμό επισκευών και των πέντε αντλιών για το έτος αυτό, θα

μπορούσε να υπολογισθεί ο αρμονικός μέσος του ερωτήματος (α); Εξηγείστε.

3.23 Ο κατασκευαστής ενός είδους μπαταριών ισχυρίζεται ότι το βάρος τους είναι 11gr. Ενα δείγμα τριών μπαταριών έχει τα εξής βάρη: $x_1 = 10.98$, $x_2 = 11.01$, $x_3 = 10.97$.

α) Να υπολογισθεί ο μέσος και η τυπική απόκλιση των τριών αυτών παρατηρήσεων. Στη συνέχεια, να εκφρασθεί κάθε μια παρατήρηση ως η απόκλιση από το επιθυμητό βάρος των 11gr (να γίνει, δηλαδή, ο μετασχηματισμός $Y_i = X_i - 11.0$) και να υπολογισθεί ο μέσος και η τυπική απόκλιση των παρατηρήσεων που θα προκύψουν από τον μετασχηματισμό. Να εξετασθεί η σχέση του μέσου και της τυπικής απόκλισης των μετασχηματισθέντων παρατηρήσεων με τον μέσο και την τυπική απόκλιση των αρχικών παρατηρήσεων.

β) Να εκφρασθούν τα βάρη των τριών μπαταριών σε κιλά. (Δηλαδή να γίνει ο μετασχηματισμός $Y_i = X_i/1000$). Να υπολογισθεί ο μέσος και η τυπική απόκλιση των μετασχηματισθέντων παρατηρήσεων. Να εξετασθεί η σχέση του μέσου και της τυπικής απόκλισης των μετασχηματισθέντων παρατηρήσεων με τον μέσο και την τυπική απόκλιση των αρχικών παρατηρήσεων.

3.24 Θεωρούμε τρεις μετρήσεις θερμοκρασίας (σε βαθμούς Κελσίου) $x_1 = 15.3$, $x_2 = 21.3$, $x_3 = 17.4$. Ο μέσος και η τυπική απόκλιση των τριών αυτών παρατηρήσεων είναι, αντίστοιχα, $\bar{x} = 18.0$ και $s = 3.04$. Με βάση τα συμπεράσματα στο (α) και (β) ερώτημα της προηγούμενης άσκησης, να βρεθεί ο μέσος και η τυπική απόκλιση των τριών μετρήσεων θερμοκρασίας εκφρασμένων σε βαθμούς Φαρενάιτ. (Να χρησιμοποιηθεί ο μετασχηματισμος $Y_i = 32 + (9/5) X_i$ που μεταχηματίζει τα δεδομένα από βαθμούς Κελσίου σε βαθμούς Φαρενάιτ).

3.25 Να αποδειχθεί ότι η τρίτη ροπή των τυποποιημένων παρατηρήσεων ενός συνόλου στοιχείων είναι ίση με το τυποποιημένο μέτρο ασυμμετρίας των δεδομένων αυτών.

3.26 Να αποδειχθεί ότι ο μέσος και η διακύμανση των τυποποιημένων παρατηρήσεων ενός συνόλου στοιχείων είναι, αντίστοιχα, 0 και 1.

3.27 Η κίνηση των τιμών των μετοχών στα χρηματιστήρια ενδιαφέρει ιδιαίτερα την κυβέρνηση, τους επιχειρηματίες αλλά και τους συναλλασσόμενους στο χρηματιστήριο. Η σωστή και αποτελεσματική σύνοψη του τεράστιου όγκου των δεδομένων που παράγεται κάθε εργάσιμη μέρα στο χρηματιστήριο χρειάζεται προσεκτική επιλογή των κατάλληλων στατιστικών μέτρων.

α) Να περιγράψετε τα κύρια στατιστικά μέτρα που χρησιμοποιεί η εφημερίδα που διαβάζετε για να παραθέσει τη γενική τάση των κοινών μετοχών σε καθημερινή βάση και τις κινήσεις της τιμής οποιασδήποτε συγκεκριμένης μετοχής σε ημερήσια, εβδομαδιαία, μηνιαία και ετήσια βάση.

β) Διατυπώστε την άποψή σας για το πόσο κατάλληλα είναι τα μέτρα αυτά και ποιά είναι τα πλεονεκτήματα και τα μειονεκτήματά τους σε σύγκριση με άλλα μέτρα τα οποία θα μπορούσε να χρησιμοποιήσει κανείς για να συνοψίσει τέτοια δεδομένα τιμών μετοχών.

3.28 Οι μετοχές 15 εταιρειών στις οποίες είχε επενδύσει ένας επενδυτής είχαν τις εξής εκατοστιαίες μεταβολές στην αξία τους κατά τη διάρκεια του περασμένου έτους:

3, 0, 6, -5, -2, 5, -18, 20, 14, 18, -10, 10, 50, -20, 14.

α) Να υπολογισθεί ο μέσος και η διακύμανση του πληθυσμού αυτού των δεδομένων. (Οι απαντήσεις να εκφρασθούν στις κατάλληλες μονάδες).

β) Να υπολογισθεί η έκταση, η διάμεσος, το 20ο εκατοστιαίο σημείο και το 60ο εκατοστιαίο σημείο για τα δεδομένα αυτά.
γ) Να κατασκευασθεί το διάγραμμα πλαισίου - απολήξεων.

3.29 Ο ιδιοκτήτης ενός καταστήματος ηλεκτρικών ειδών πωλεί το ηλεκτρικό καλώδιο με το μέτρο. Προκειμένου να μειώσει το εργατικό κόστος, ο ιδιοκτήτης σκέπτεται να πουλά το καλώδιο σε κομμάτια κομμένα εκ των προτέρων σε καθορισμένα μήκη. Ενα δείγμα μηκών (σε μέτρα) καλωδίου που πουλήθηκε κατά τη διάρκεια μιας ημέρας έδωσε τα εξής αποτελέσματα:

3, 7, 4, 2.5, 3, 20, 5, 5, 15, 3.5, 3.

α) Να βρεθεί ο μέσος, η διάμεσος και η επικρατούσα τιμή των μηκών του καλωδίου που πωλήθηκε τη συγκεκριμένη μέρα.
β) Σχολιάστε τα μειονεκτήματα που κάθε ένα από τα τρία αυτά μέτρα κεντρικής τάσης έχει ως προς τις πληροφορίες που παρέχει στον ιδιοκτήτη του καταστήματος.
γ) Να βρεθεί το πρώτο και το τρίτο τεταρτημόριο των δεδομένων αυτών.
δ) Με ποιό τρόπο θα μπορούσε ο ιδιοκτήτης του καταστήματος να αποφασίσει για τα μήκη στα οποία θα κόβει εκ των προτέρων το καλώδιο αυτό;

3.30 Θεωρείστε τα δύο αμοιβαία κεφάλαια Α και Β του παραδείγματος της ενότητας 3.3. Οπως είχαμε δει, η ετήσια ποσοστιαία απόδοση των κεφαλαίων αυτών για τα δέκα τελευταία χρόνια ήταν:

Κεφ. Α: 8.3, -6.2, 20.9, -2.7, 33.6, 42.9, 24.4, 5.2, 3.1, 30.5
Κεφ. Β: 12.1, -2.8, 6.4, 12.2, 27.8, 25.3, 18.2, 10.7, -1.3, 11.4

Εστω ότι ένας επενδυτής πριν από 10 χρόνια δημιούργησε ένα χαρτοφυλάκιο επενδύοντας ίσα ποσά χρημάτων σε κάθε ενα από τα

δύο αμοιβαία κεφάλαια. Η απόδοση που θα είχε ο επενδυτής κατά τον πρώτο χρόνο θα ήταν $0.5(8.3)+0.5(12.1)=10.2\%$.

α) Να υπολογισθεί η απόδοση της επένδυσης για τον κάτοχο του χαρτοφυλακίου για κάθε ένα από τα δέκα έτη.

β) Να βρεθεί η μέση απόδοση του χαρτοφυλακίου για τα δέκα τελευταία έτη.

γ) Να βρεθεί η τυπική απόκλιση της απόδοσης του χαρτοφυλακίου για τα δέκα τελευταία έτη.

δ) Ιεραρχήσατε τις τρεις πιθανές επενδύσεις (κεφάλαιο Α, κεφάλαιο Β και το χαρτοφυλάκιο) σύμφωνα με τη μέση απόδοσή τους και σύμφωνα με την επικινδυνότητά τους (όπως αυτή μετριέται με την τυπική απόκλιση) για τα δέκα τελευταία έτη.

3.31 Ο ετήσιος συνολικός αριθμός απόδοσης των κοινών μετοχών και των μακροχρονίων κρατικών ομολόγων στο χρηματιστήριο του Καναδά για τα τελευταία 25 χρόνια δίνονται στον πίνακα που ακολουθεί.

α) Να βρεθεί ο μέσος, η διάμεσος, η έκταση και η τυπική απόκλιση του δείγματος αυτού των κοινών μετοχών.

β) Να βρεθεί το κάτω και το άνω τεταρτημόριο της απόδοσης των κοινών μετοχών.

Ετήσια συνολική απόδοση

Ετος	Κοινές μετοχές				
1960-1964	1.66	32.54	-7.52	15.56	25.30
1965-1969	6.54	-7.10	18.00	22.36	-0.96
1970-1974	-3.60	8.07	27.31	-0.42	-26.61
1975-1979	19.70	10.94	9.93	29.22	44.38
1980-1984	29.93	-10.29	5.51	34.84	-2.44

Ετος	Μακροχρόνια κρατικά ομόλογα				
1960-1964	7.10	9.78	3.05	4.60	6.59
1965-1969	0.96	1.55	-2.20	-0.52	-2.31
1970-1974	-21.98	11.55	1.11	1.71	-1.69
1975-1979	2.82	19.02	5.97	1.29	-2.62
1980-1984	2.06	-3.02	42.98	9.60	15.09

(Πηγή: Report on Canadian Economic Statistics: 1924-1984 (Canadian Institute of Actuaries, May 1985) pp 22,32).

γ) Να επαναληφθούν τα ερωτήματα (α) και (β) για την απόδοση των ομολόγων.

δ) Ποιό είδος επένδυσης (κοινές μετοχές ή ομόλογα) φαίνεται να περικλείει το μεγαλύτερο κίνδυνο; την υψηλότερη απόδοση;

ε) Να υπολογισθεί η μέση απόδοση των μετοχών για κάθε μια από τις πέντε πενταετείς περιόδους.

στ) Να υπολογισθεί ο μέσος και η τυπική απόκλιση των πέντε μέσων αποδόσεων του προηγουμένου ερωτήματος. Να συγκριθούν οι τιμές αυτές με το μέσο και την τυπική απόκλιση που υπολογίσθηκε στο ερώτημα (α) και να εξηγηθούν οι διαφορές που εμφανίζονται.

3.32 Ενας επιθεωρητής της αγορανομίας θέλει να ελέγξει το πραγματικό βάρος του περιεχομένου όλων των συσκευασιών καφέ που έχουν την ένδειξη "160gr". Ενα δείγμα 30 συσκευασιών του συγκεκριμένου καφέ επελέγη τυχαία και το περιεχόμενο καθεμιάς από αυτές ζυγίσθηκε. Ο μέσος του βάρους του περιεχομένου των 30 αυτών συσκευασιών βρέθηκε να είναι 159.2gr. Η τυπική απόκλιση των παρατηρήσεων αυτών βρέθηκε να είναι 0.4gr. Να περιγράψετε την κατανομή των μετρήσεων των βαρών για τις 30 συσκευασίες, χρησιμοποιώντας

α) το θεώρημα του Chebyshev

β) τον εμπειρικό κανόνα (θα περιμένατε να είναι ο εμπειρικός

κανόνας κατάλληλος για την περιγραφή των δεδομένων αυτών;)

γ) Εστω ότι ο υπεύθυνος της αγορανομίας είχε ελέγξει το περιεχόμενο μόνο τεσσάρων συσκευασιών καφέ και είχε βρει τα αντίστοιχα βάρη να είναι 158.4, 160.0, 159.2 και 158.4. Θα ήταν ο εμπειρικός κανόνας κατάλληλος για να περιγράψει τις 4 μετρήσεις; Εξηγείστε την απάντησή σας.

3.33 Ο χρόνος που χρειάζεται μία ταμίας να "χτυπήσει" τις τιμές τροφίμων στη μηχανή του ταμείου σε κάποια αγορά τροφίμων που χρησιμοποιεί αυτόματες ταμειακές μηχανές έχει καταγραφεί για 10 πελάτες. Οι χρόνοι αυτοί σε δευτερόλεπτα, ήταν:

15, 62, 53, 11, 38, 75, 112, 40, 22, 57.

α) Ελέγξτε τα δεδομένα και χρησιμοποιήστε την έκτασή τους για να προσεγγίσετε την τυπική απόκλιση. Χρησιμοποιήστε την τιμή αυτή για να ελέγξετε τους υπολογισμούς σας στο δεύτερο ερώτημα.

β) Να υπολογισθεί ο δειγματικός μέσος και η τυπική απόκλιση για τα δεδομένα αυτά. Να συγκρίνετε την απάντησή σας με την απάντηση στο ερώτημα (α).

3.34 Ο αριθμός των ωρών παρακολούθησης τηλεόρασης ανά νοικοκυριό και η ώρα της ημέρας κατά την οποία παρακολουθούν τηλεόραση οι περισσότεροι θεατές είναι δύο παράγοντες που επηρεάζουν τα εισοδήματα των τηλεοπτικών σταθμών από διαφημίσεις. Ενα τυχαίο δείγμα από 25 νοικοκυριά σε κάποια συγκεκριμένη περιοχή έδωσε τα ακόλουθα στοιχεία για τον αριθμό των ωρών παρακολούθησης τηλεόρασης ανά νοικοκυριό.

3.0, 6.0, 7.5, 15.0, 12.0, 6.5, 8.0, 4.0, 5.5, 6.0, 5.0, 12.0,
1.0, 3.5, 3.0, 7.5, 5.0, 10.0, 8.0, 3.5, 9.0, 2.0, 6.5, 1.0, 5.0.

α) Να ελέγξετε τα δεδομένα αυτά και να χρησιμοποιήσετε την έκταση για να βρείτε μια προσεγγιστική τιμή για την τυπική

απόκλιση.

β) Να υπολογίσετε το δειγματικό μέσο και τη δειγματική τυπική απόκλιση. Να συγκρίνετε την τιμή της δειγματικής τυπικής απόκλισης με την προσεγγιστική τιμή που βρήκατε στο ερώτημα (α).

γ) Να βρείτε το ποσοστό των ωρών παρακολούθησης τηλεόρασης ανά νοικοκυριό που περιέχεται στο διάστημα $(\bar{x}\pm 2s)$. Να συγκρίνετε της απάντησή σας με το αντίστοιχο ποσοστό που δίνει ο εμπειρικός κανόνας.

3.35 Στην υλοτομία, μία μέθοδος που χρησιμοποιείται για τον καθορισμό του αριθμού των υλοτομήσιμων δέντρων (αυτών δηλαδή με διάμετρο μεγαλύτερη από 40cm) είναι η εξής: Ο ειδικός, προκειμένου να εκτιμήσει τον αριθμό αυτών των δέντρων της υπό εξέταση έκτασης, διασχίζει με τυχαίο τρόπο τυχαία επιλεγμένες επιφάνειες διαστάσεων περίπου 15m×15m και μετρά τον αριθμό των δέντρων που έχουν διάμετρο που υπερβαίνει τα 40cm. Εστω ότι 70 τέτοια τετράγωνα επελέγησαν τυχαία και εξετάσθηκαν. Εστω ότι οι αριθμοί των δέντρων στα τετράγωνα αυτά με διάμετρο μεγαλύτερη από 40cm είναι οι εξής:

7, 9, 3, 10, 9, 6, 10, 8, 8, 5, 8, 6, 9, 2, 6, 11, 8,
9, 7, 8, 7, 4, 5, 7, 8, 9, 8, 10, 9, 8, 10, 9, 9, 4, 8,
11, 5, 7, 9, 7, 4, 10, 9, 8, 8, 7, 9, 7, 6, 9, 8, 9, 8,
5, 7, 7, 9, 7, 8, 13, 6, 8, 7, 10, 8, 11, 8, 5, 9, 8.

α) Να κατασκευασθεί ένα ιστόγραμμα σχετικής συχνότητας για να περιγράψει τα δεδομένα αυτά.

β) Να υπολογισθεί ο δειγματικός μέσος $\bar{x}$ ως μια εκτίμηση του μέσου αριθμού των υλοτομήσιμων δέντρων με διάμετρο μεγαλύτερη από 40 cm για όλα τα τετράγωνα στην έκταση αυτή.

γ) Να υπολογισθεί η τυπική απόκλιση των δεδομένων. Να κατασκευασθούν τα διαστήματα $(\bar{x}\pm s)$, $(\bar{x}\pm 2s)$ και $(\bar{x}\pm 3s)$. Να

βρεθεί το ποσοστό των τετραγώνων δάσους διαστάσεων 15m×15m που περιέχονται σε κάθε ένα από τα τρία αυτά διαστήματα. Να συγκρίνετε τις απαντήσεις σας με τα αντίστοιχα ποσοστά που δίνονται με τη χρήση του εμπειρικού κανόνα και του θεωρήματος του Chebyshev

3.36 Διαφορετικές γλώσσες έχουν διαφορετικό αριθμό συλλαβών στις λέξεις τους. Ο πίνακας που ακολουθεί δίνει μερικές σχετικές συχνότητες για τέσσερις διαφορετικές γλώσσες.

Αριθμός συλλαβών	Σχετικές συχνότητες			
	Αραβικά	Αγγλικά	Γερμανικά	Ιαπωνικά
1	.23	.71	.56	.36
2	.50	.19	.31	.34
3	.22	.07	.09	.18
4	.05	.02	.03	.09
5	.00	.01	.01	.02
6	.00	.00	.00	.01
Σύνολο	1.00	1.00	1.00	1.00

α) Να κατασκευασθεί ένα διάγραμμα σχετικής συχνότητας για κάθε μια από τις τέσσερις αυτές γλώσσες.

β) Να υπολογισθεί ο μέσος, η διάμεσος και η επικρατούσα τιμή για κάθε μια από τις γλώσσες αυτές.

γ) Να υπολογισθεί η μέση απόλυτη απόκλιση, η τυπική απόκλιση και ο συντελεστής μεταβλητότητας για κάθε μια από τις γλώσσες.

δ) Με βάση τις παραπάνω περιγραφικές στατιστικές συναρτήσεις να περιγράψετε τη διαφορά μεταξύ των τεσσάρων αυτών γλωσσών.

3.37 Ο μέσος και η τυπική απόκλιση των βαθμών 500 φοιτητών που έδωσαν εξετάσεις στη Στατιστική στο Τμήμα Διοίκησης Επιχειρήσεων ήταν 69 και 7 αντίστοιχα.

α) Ποιός είναι ο ελάχιστος αριθμός από τους φοιτητές αυτούς που πήραν βαθμό στα διαστήματα ($\bar{x}$±2s), ($\bar{x}$±2.5s) και ($\bar{x}$±3s);

β) Πόσοι, το πολύ, από τους φοιτητές αυτούς δεν έλαβαν βαθμό στο διάστημα ($\bar{x}$±3.5s) και ($\bar{x}$±4s);

γ) Εάν η κατανομή των βαθμών είχε περίπου το σχήμα κανονικής κατανομής πόσοι φοιτητές πιστεύετε ότι πήραν βαθμούς στα διαστήματα ($\bar{x}$±s), ($\bar{x}$±2s) και ($\bar{x}$±3s);

3.38 Οι ακόλουθες 20 τιμές αντιπροσωπεύουν, σε δευτερόλεπτα, την ώρα που χρειάζεται κάθε ένας από 20 εργαζόμενους σε ένα εργαστάσιο για να κάνει μία συγκεκριμένη εργασία.

2.1, 2.7, 2.6, 2.8, 2.3, 2.5, 2.6, 2.4, 2.6, 2.7,
2.4, 2.6, 2.8, 2.5, 2.6, 2.4, 2.9, 2.4, 2.7, 2.3.

α) Να υπολογισθεί η διασπορά και η τυπική απόκλιση για το δείγμα αυτό των 20 παρατηρήσεων.

β) Να χρησιμοποιηθεί η προσέγγιση της έκτασης για την τυπική απόκλιση για να ελεγχθούν οι υπολογισμοί στο ερώτημα (α). Ποιές είναι οι υποθέσεις που θα πρέπει να κάνουμε για να ισχύει η προσέγγιση αυτή;

γ) Χρησιμοποιώντας το θεώρημα του Chebyshev τί θα μπορούσατε να πείτε για το ποσοστό των μετρήσεων πού βρίσκονται σε ένα διάστημα 1.5 τυπικών απολήξεων από το μέσο για τις 20 αυτές μετρήσεις;

δ) Να συγκρίνετε την απάντησή σας στο ερώτημα (γ) με το πραγματικό ποσοστό μετρήσεων που βρίσκονται 1.5 τυπικές απολήξεις από τον μέσο.

3.39 Ενα βιβλιοπωλείο έχει υπολογίσει ότι οι εβδομαδιαίες πωλήσεις

ενός περιοδικού έχουν μια περίπου συμμετρική κατανομή με μία επικρατούσα τιμή, με μέσο αριθμό πωλήσεων 85 και τυπική απόκλιση 6.

α) Ποιό είναι το ποσοστό των φορών που ο ιδιοκτήτης του βιβλιοπωλείου θα πρέπει να περιμένει τις εβδομαδιαίες πωλήσεις του περιοδικού να βρίσκονται στα διαστήματα $(\bar{x}\pm s)$ και $(\bar{x}\pm 3s)$;

β) Ποιό είναι το ποσοστό των φορών που ο ιδιοκτήτης θα πρέπει να περιμένει να έχουν οι εβδομαδιαίες πωλήσεις για το περιοδικό αυτό μια τιμή που να είναι μεγαλύτερη από το μέσο περισσότερο από δύο τυπικές αποκλίσεις;

γ) Εάν το βιβλιοπωλείο παραλαμβάνει 97 αντίτυπα του περιοδικού αυτού κάθε εβδομάδα, ποιό είναι το ποσοστό των εβδομάδων που θα βρεθεί να μην έχει αρκετό αριθμό αντιτύπων ώστε να αντιμετωπίσει τη ζήτηση για το περιοδικό αυτό;

3.40 Τον τελευταίο χρόνο οι ρυθμοί απόδοσης των κοινών μετοχών σε κάποιο μεγάλο χαρτοφυλάκιο είχαν μια συμμετρική κατανομή με μία κορυφή, μέσο 20% και τυπική απόκλιση 50%.

α) Ποιό είναι το ποσοστό των μετοχών που είχαν απόδοση μεταξύ 10% και 30%; μεταξύ -10% και 50%;

β) Ποιό είναι το ποσοστό των μετοχών που είχαν απόδοση είτε μικρότερη από 10% είτε μεγαλύτερη από 30%;

γ) Ποιό είναι το ποσοστό των μετοχών που έχουν θετική απόδοση;

3.41 α) Να υπολογισθεί, κατά προσέγγιση, ο μέσος και η διακύμανση των δειγματικών δεδομένων που δίνονται στην κατανομή συχνότητας που ακολουθεί.

β) Να χρησιμοποιηθεί η προσέγγιση του s μέσω της έκτασης των δεδομένων για να ελεγχθεί η προσέγγιση της διακύμανσης του

ερωτήματος (α).

Κλάση	Συχνότητα
0 έως 16	50
16 έως 32	160
32 έως 48	110
48 έως 64	80

3.42 Η ωριαία αμοιβή (σε εκατοντάδες δραχμές) μιας ομάδας εργαζομένων που επιλέχθηκε τυχαία από την κατάσταση πληρωμής μιας βιομηχανίας εμφανίζεται στην κατανομή συχνότητας που ακολουθεί.

Ωριαία αμοιβή (σε εκατοντ. δρχ.)	Αριθμός εργαζομένων
8 έως 10	11
10 έως 12	17
12 έως 14	32
14 έως 16	27
16 έως 18	13

Να υπολογισθεί, κατά προσέγγιση, ο μέσος και η τυπική απόκλιση των ωριαίων αμοιβών των εργαζομένων για το συγκεκριμένο δείγμα.

3.43 Οι ηλικίες ενός δείγματος 25 χρηματιστών είναι οι εξής:

50, 64, 32, 55, 41, 44, 24, 46, 58, 47, 36, 52, 54,
44, 66, 47, 59, 51, 61, 57, 49, 28, 42, 38, 45.

α) Να κατασκευασθεί ένα διάγραμμα μίσχου-φύλλου για τις ηλικίες αυτές.

β) Να βρεθεί η διάμεσος των ηλικιών, το κάτω και το άνω τεταρτημόριο και το 80ο εκατοστιαίο σημείο των ηλικιών.

γ) Υπάρχει κάποιο συμπέρασμα που προκύπτει από την κατανομή των ηλικιών των χρηματιστών αυτών;

3.44 Ο λόγος της τιμής προς την απόδοση μιας μετοχής ενδιαφέρει ιδιαίτερα τους αναλυτές επενδύσεων γιατί δίνει πληροφορίες τόσο για το ρίσκο όσο και για τις ευκαιρίες ανάπτυξης της μετοχής. Ο λόγος των τιμών προς την απόδοση για 30 μετοχές τραπεζών δίνεται στα στοιχεία που ακολουθούν.

6.3, 9.6, 10.6, 8.0, 8.6, 9.2, 6.1, 4.8, 9.9, 8.0,
6.4, 9.7, 7.7, 5.3, 7.6, 6.9, 8.4, 8.1, 6.9, 6.2,
9.4, 8.6, 11.0, 8.9, 8.4, 7.9, 9.0, 8.1, 10.0, 7.0.

α) Να υπολογισθεί ο μέσος και η τυπική απόκλιση του δείγματος των 30 αυτών παρατηρήσεων.

β) Να κατασκευασθεί ένα ιστόγραμμα σχετικής συχνότητας για τις παραπάνω τιμές.

γ) Να καθορισθεί η διάμεση τιμή του λόγου τιμών-απόδοσης και να καθορισθεί ο μέσος και η διάμεσος στο ιστόγραμμα αυτό.

δ) Να χρησιμοποιηθεί η κατανομή σχετικής συχνότητας για να εκτιμηθεί ο μέσος και η τυπική απόκλιση των 30 παρατηρήσεων. Στη συνέχεια να συγκριθούν οι εκτιμήσεις αυτές με τις τιμές που προέκυψαν στο ερώτημα (α).

ε) Να κατασκευασθεί το διάγραμμα πλαισίου-απολήξεων για τα δεδομένα αυτά.

3.45 Η απόδοση μιας κοινής μετοχής είναι ο λόγος του ετήσιου μερίσματος της μετοχής δια της τιμής της μετοχής. Η απόδοση για το έτος 1985 εκτιμήθηκε για 30 μετοχές τραπεζών στο χρηματιστήριο της Νέας Υόρκης. Η εκτίμηση της απόδοσης των 30 αυτών μετοχών με τη μορφή ποσοστών δίνεται στα στοιχεία που ακολουθούν.

4.0, 2.9, 4.3, 3.1, 3.5, 3.4, 5.2, 6.8, 4.5, 3.3,
4.9, 3.4, 4.3, 6.4, 3.0, 4.7, 4.4, 4.1, 5.0, 4.9,
3.6, 3.7, 1.9, 3.1, 4.0, 4.0, 3.7, 4.4, 6.3, 4.4.

α) Να υπολογισθεί ο μέσος και η τυπική απόκλιση του δείγματος των 30 αποδόσεων των μετοχών.

β) Να κατασκευασθεί το ιστόγραμμα της σχετικής συχνότητας για τα στοιχεία αυτά.

γ) Να βρεθεί η διάμεσος και να εντοπισθούν ο μέσος και η διάμεσος στο ιστόγραμμα.

δ) Να κατασκευασθεί το διάγραμμα πλαισίου-απολήξεων με τα στοιχεία αυτά.

3.46 Τα στοιχεία που ακολουθούν αναφέρονται στο ύψος των ενοικίων (σε χιλιάδες δραχμές) για ένα δείγμα από 10 πολυτελή γραφεία σε δύο διαφορετικά σημεία της Αθήνας, έστω Α και Β.

Α: 955, 1000, 985, 980, 940, 975, 965, 999, 1247, 1119
Β: 750, 775, 725, 705, 694, 725, 690, 745, 575, 800

α) Για κάθε ομάδα δεδομένων να υπολογισθούν ο μέσος, η διάμεσος, η έκταση, η τυπική απόκλιση και ο συντελεστής μεταβλητότητας.

β) Τί σχόλια θα μπορούσαν να γίνουν για το ύψος των ενοικίων γραφείων στην περιοχή Α σε σχέση με το αντίστοιχο ύψος των ενοικίων στην περιοχή Β με βάση τα στοιχεία του δείγματος;

γ) Πώς θα μορούσε κανείς να χρησιμοποιήσει τις πληροφορίες αυτές αν ενδιαφερόταν να νοικιάσει γραφεία στην περιοχή Α ή στην περιοχή Β;

3.47 Τα στοιχεία που ακολουθούν αναφέρονται στις τιμές ενός δείγματος από 29 διαφορετικές μάρκες υπολογιστών υψηλής ποιότητας (σε χιλιάδες δραχμές).

899, 1199, 850, 1000, 600, 1195, 750, 1595, 1050,
1200, 799, 700, 1500, 629, 899, 1150, 889, 629, 999,

650, 799, 900, 580, 850, 700, 799, 1200, 729, 899.

α) Να υπολογισθούν ο μέσος, η διάμεσος, η επικρατούσα τιμή, η έκταση, το ενδοτεταρτημοριακό εύρος, η διακύμανση, η τυπική απόκλιση και ο συντελεστής μεταβλητότητας.

β) Να κατασκευασθεί το διάγραμμα πλαισίου-απολήξεων και να καθορισθεί αν τα δεδομένα είναι συμμετρικά ή αν έχουν δεξιά ή αριστερή ασυμμετρία. Να δικαιολογηθεί το συμπέρασμα αυτό.

γ) Πώς θα μπορούσε κανείς να αξιοποιήσει τα στοιχεία αυτά;

3.48 Τα δεδομένα που ακολουθούν αναφέρονται στο χρόνο (σε δευτερόλεπτα) που χρειάζονται 22 γερμανικά αυτοκίνητα και 30 ιαπωνικά σε ένα δοκιμαστικό έλεγχο για να αναπτύξουν ταχύτητα από 0 σε 90km/h.

Γερμανικά αυτοκίνητα				**Ιαπωνικά αυτοκίνητα**				
10.0	7.9	7.1	8.6	9.4	7.7	5.7	8.2	9.3
6.4	6.9	8.7	8.3	8.9	9.3	8.3	9.7	8.6
8.5	6.4	7.5	6.7	6.7	9.1	9.5	11.7	10.0
5.5	6.0	5.4	6.9	7.2	6.8	8.0	6.3	8.8
5.1	4.9	8.5	8.8	8.5	7.1	6.5	12.0	9.2
10.9	8.9			9.5	10.5	12.5	6.2	6.6

(Στοιχεία Οκτωβρίου 1990)

α) Να γίνει σύγκριση των χρόνων επιτάχυνσης για τα δύο αυτά δείγματα.

β) Γράψτε τις απόψεις που θα διατυπώνατε σε ένα φίλο σας σε σχέση με την απάντηση του ερωτήματος (α) για να τον βοηθήσετε στην επιλογή αυτοκινήτου με δεδομένο ότι ο φίλος σας ενδιαφέρεται για το συγκεκριμένο χαρακτηριστικό.

3.49 Εκτελέστε το εξής πείραμα: στρίψτε 10 νομίσματα μία φορά το κάθε ένα και σημειώστε τον αριθμό X των φορών που το αποτέλεσμα

ήταν "γράμματα". Επαναλάβατε τη διαδικασία αυτή 50 φορές, ώστε να έχετε 50 τιμές για το Χ.

α) Κατασκευάστε τη σχετική συχνότητα για τις μετρήσεις αυτές.

β) Υπολογίστε το μέσο και την τυπική απόκλιση των μετρήσεων αυτών.

γ) Υπολογίστε τα διαστήματα $(\bar{x}\pm s)$, $(\bar{x}\pm 2s)$ και $(\bar{x}\pm 3s)$.

δ) Βρείτε το ποσοστό των παρατηρήσεων που ανήκουν σε κάθε ένα από τα διαστήματα αυτά. Συμφωνεί το ποσοστό αυτό με το θεώρημα του Chebyshev; Συμφωνεί με τον εμπειρικό κανόνα;

3.50 Τα βιομηχανικά ρομπότ εμφανίσθηκαν για πρώτη φορά στις Η.Π.Α. το 1960 και από τότε χρησιμοποιούνται όλο και περισσότερο στη βιομηχανία. Αρχικά ένα ρομπότ που εχρησιμοποιείτο στη γραμμή παραγωγής κόστιζε, κατά μέσο όρο, 4.20 δολλάρια ανά ώρα, λίγο περισσότερο από τη μέση ωριαία αμοιβή ενός βιομηχανικού εργάτη. Σήμερα οι βιομηχανικοί εργάτες πληρώνονται μεταξύ 15 και 20 δολλαρίων την ώρα ενώ τα ρομπότ εξακολουθούν και εργάζονται με κόστος λιγότερο από 5 δολλάρια την ώρα. Εστω ότι έχει καταγραφεί το κόστος λειτουργίας για 100 εταιρείες που χρησιμοποιούν βιομηχανικά ρομπότ. Εστω ότι ο μέσος και η διασπορά του δείγματος βρέθηκαν να είναι, αντίστοιχα, 4.86δολ. και $2.50(\text{δολ.})^2$. Να υπολογισθούν τα διαστήματα $(\bar{x}\pm s)$, $(\bar{x}\pm 2s)$ και $(\bar{x}\pm 3s)$ και να δηλωθεί το κατά προσέγγιση ποσοστό των μετρήσεων που περιμένει κανείς να ανήκουν σε κάθε ένα από τα διαστήματα αυτά σύμφωνα με τον εμπειρικό κανόνα.

3.51 Η μέση διάρκεια των τηλεοπτικών διαφημίσεων σε κάποιο κανάλι είναι 75 δευτερόλεπτα και η τυπική απόκλιση είναι 20 δευτερόλεπτα. Υποθέτοντας ότι η κατανομή της συχνότητας της διάρκειας των διαφημίσεων είναι περίπου κανονική, να υπολογισθούν:

α) το κατά προσέγγιση ποσοστό των διαφημίσεων που διαρκούν λιγότερο από 35 δευτερόλεπτα,

β) το κατά προσέγγιση ποσοστό των διαφημίσεων που διαρκούν περισσότερο από 55 δευτερόλεπτα.

3.52 Ας θεωρήσουμε το σύνολο που αποτελείται από τα μέλη του Διδακτικού Ερευνητικού Προσωπικού (ΔΕΠ) που απασχολούνται σε τμήματα τεσσάρων τυχαία επιλεγμένων Πανεπιστημίων. Εστω ότι ο αριθμός των μελών ΔΕΠ ανά τμήμα έχει μέσο $\bar{x} = 25$ και τυπική απόκλιση $s = 25$.

α) Να χρησιμοποιηθεί το θεώρημα του Chebyshev για να εξαχθεί κάποιο συμπέρασμα που να αναφέρεται στο ποσοστό των τμημάτων που απασχολούν από 15 έως 35 μέλη ΔΕΠ.

β) Υποθέτοντας ότι ο πληθυσμός για το συγκεκριμένο πρόβλημα είναι κανανικός, να καθορισθεί το ποσοστό των τμημάτων που απασχολούν περισσότερα από 30 μέλη ΔΕΠ.

3.53 Μια μέθοδος για να αυξηθούν οι εισπράξεις στα περιοδικά είναι να χρησιμοποιήσουν περισσότερες ολοσέλιδες διαφημίσεις σε κάθε τεύχος. Τα δεδομένα που ακολουθούν αναφέρονται στον αριθμό των ολοσέλιδων διαφημίσεων σε ένα τυχαία επιλεγμένο δείγμα εβδομαδιαίων περιοδικών.

12, 10, 16, 7, 18, 13, 14, 20, 9, 23,
8, 13, 14, 6, 19, 6, 11, 15, 10, 16.

Να υπολογισθεί η τιμή της τυπικής απόκλισης για τα δεδομένα αυτά και να χρησιμοποιηθεί η μέθοδος της προσέγγισης της τυπικής απόκλισης μέσω της έκτασης ως ένας έλεγχος για τους υπολογισμούς. Πιστεύετε ότι η διαφορά της υπολογισθείσας από την εκτιμηθείσα τιμή είναι εξαιρετικά μεγάλη;

3.54 Η καταγραφή του βαθμού συννεφιάς σε κλίμακα από 0 έως 10 στο

Greenwich κάθε μέρα του Ιουλίου για την περίοδο 1890 έως 1904 έδωσε τα εξής αποτελέσματα:

Βαθμός συννεφιάς	Αριθμός ημερών
0	320
1	129
2	74
3	68
4	45
5	45
6	55
7	65
8	90
9	148
10	676

Να κατασκευασθεί ένα διάγραμμα σχετικής συχνότητας και να υπολογισθεί ο μέσος βαθμός συννεφιάς. Αποτελεί ο μέσος μια ακριβή περιγραφή μιας "τυπικής" μέρας στο Greenwich;

3.55 Μια μελέτη των ηλικιών στις οποίες παντρεύτηκαν οι Αυστραλοί άνδρες κατά την περίοδο 1907 έως 1914 έδωσε τα εξής αποτελέσματα:

Ηλικία	Αριθμός ανδρών
15 έως 18	294
18 έως 21	10995
21 έως 24	61101
24 έως 27	73054
27 έως 30	56501
30 έως 33	33478
33 έως 36	20569
36 έως 39	14281
39 έως 42	9320
42 έως 48	11006
48 έως 54	5810
54 έως 60	2755
60 έως 66	1459
66 έως 78	1143
78 έως 90	119

Χρησιμοποιήστε ένα ιστόγραμμα για να παρουσιάσετε την κατανομή των ηλικιών γάμου. Ποιά είναι η επικρατούσα κλάση; Να εκτιμηθεί ο μέσος και η διάμεσος των ηλικιών στις οποίες παντρεύονταν οι Αυστραλοί άνδρες σύμφωνα με τα δεδομένα αυτά.

3.56 Η Εθνική Υπηρεσία Οικονομικής Ερευνας (National Bureau of Economic Research) των Η.Π.Α. προσδιορίζει την αρχή μιας περιόδου ύφεσης και το τέλος της. Ο πίνακας που ακολουθεί δίνει τα στοιχεία για την αρχή και το τέλος των περιόδων ύφεσης από το έτος 1920 έως το 1982.

Κύκλοι της οικονομίας των ΗΠΑ	Υφεση (σε μήνες)	Προηγηθείσα ανάπτυξη (σε μήνες)
Ιαν.1920 - Ιουλ.1921	18	10
Μάιος 1923 - Ιουλ.1924	14	22
Οκτ.1926 - Νοεμ.1927	13	27
Αυγ.1929 - Μαρ.1933	43	21
Μάιος 1937 - Ιουν.1938	13	50
Φεβ.1945 - Οκτ.1945	8	80
Νοεμ.1948 - Οκτ.1949	11	37
Ιουλ.1953 - Μάιος 1954	10	45
Αυγ.1957 - Απρ.1958	8	39
Απρ.1960 - Φεβ.1961	10	24
Δεκ.1969 - Νοεμ.1970	11	106
Νοεμ.1973 - Μαρ.1975	16	36
Ιαν.1980 - Ιουλ.1980	6	58
Ιουλ.1981 - Δεκ.1982	18	12

Να υπολογισθεί ο μέσος και η διάμεσος των χρονικών περιόδων για τις 14 περιόδους υφέσεων και τις 14 προηγηθείσες περιόδους ανάπτυξης. Ποιός είναι ο λόγος για τον οποίο κάθε ένα από αυτά τα δύο μέτρα θέσης διαφέρει από την αντίστοιχη διάμεσο;

3.57 Εχει λεχθεί ότι "ο μέσος επενδυτής είναι μία λευκή γυναίκα 53 ετών". Τί σημαίνει μέσος στην περίπτωση αυτή;

3.58 α) Εάν κάποιος οδηγεί με ταχύτητα 20km/h για μία ώρα και 40km/h για μία ακόμα ώρα ποιά είναι η μέση ταχύτητά του;
β) Ποιά είναι η μέση ταχύτητα εάν οδηγεί κάποιος με ταχύτητα 20km/h για 20 χιλιόμετρα και 40km/h για 40 χιλιόμετρα;

3.59 Ενα ερώτημα που απασχολεί τους κοινωνιολόγους είναι το κατά πόσον οι μεγάλες ανακαλύψεις έγιναν από ανθρώπους που ήταν νέοι

στην ηλικία και δραστήριοι ή από ανθρώπους που ήταν μεγάλοι σε ηλικία και ώριμοι. Τα δεδομένα που ακολουθούν αναφέρονται στις ηλικίες στις οποίες 12 γνωστοί επιστήμονες έκαναν κάποιες μεγάλες ανακαλύψεις. Να χρησιμοποιηθούν τα δεδομένα αυτά για να υπολογισθεί ο μέσος και η διάμεσος ηλικία στην οποία οι ανακαλύψεις αυτές έγιναν.

Επιστήμονας	Ανακάλυψη	Ηλικία
Κοπέρνικος	Περιστροφή της γης γύρω από τον ήλιο	40
Γαλιλαίος	Νόμοι της Αστρονομίας	34
Νεύτων	Κίνηση, Βαρύτητα, Απειροστικός Λογισμός	23
Φραγκλίνος	Φύση του ηλεκτρισμού	40
Lavoisier	Η καύση ως οξείδωση	31
Lyell	Σταδιακή εξέλιξη της γης	33
Δαρβίνος	Φυσική επιλογή στην εξέλιξη	49
Maxwell	Εξισώσεις για το φως	33
Curie	Ραδιενέργεια	34
Planck	Κβαντική Θεωρία	43
Einstein	Θεωρία σχετικότητας	26
Schroedinger	Εξισώσεις για την κβαντική Θεωρία	39

Να χρησιμοποιηθούν τα δεδομένα αυτά για να υπολογισθούν η μέση και η διάμεσος ηλικία στην οποία έγιναν οι ανακαλύψεις αυτές.

3.60 Ενα ερώτημα που συχνά θέτουν περιοδικά που απευθύνοται σε επιχειρηματίες είναι το κατά πόσον οι επιχειρηματίες που ξεκινούν νέες επιχειρήσεις το κάνουν αυτό όταν είναι νέοι ή όταν είναι

μεσήλικες και έχουν κουρασθεί πια να δουλεύουν για κάποιον άλλο. Προκειμένου να απαντήσετε στο ερώτημα αυτό, χρησιμοποιήστε τα δεδομένα που ακολουθούν.

Ηλικία επιχειρηματία	Ποσοστό επιχειρηματιών
20 έως 25	9%
25 έως 30	17%
30 έως 35	21%
35 έως 40	18%
40 έως 45	15%
45 έως 49	9%
50 έως 60	11%

α) Να χρησιμοποιηθούν τα δεδομένα αυτά για να κατασκευασθεί ένα ραβδόγραμμα και ένα ιστόγραμμα. Υπάρχει ουσιαστική διαφορά στην εμφάνισή τους;

β) Να εκτιμηθεί ο μέσος και η διάμεσος ηλικία στην οποία οι άνθρωποι ξεκινούν νέες επιχειρήσεις με χρησιμοποίηση των δεδομένων αυτών. Να εξηγηθεί γιατί ο μέσος είναι μεγαλύτερος ή μικρότερος από τη διάμεσο.

3.61 Ο πίνακας που ακολουθεί δίνει τα ποσοστά ανεργίας για επτά χώρες τα έτη 1960, 1970, 1980 και 1990.

Χώρα	Ποσοστό ανεργίας			
	1960	1970	1980	1990
Καναδάς	6.5%	5.7%	7.5%	7.5%
Γαλλία	1.5%	2.5%	6.4%	10.1%
Ιταλία	3.7%	3.2%	4.4%	7.8%
Ιαπωνία	1.7%	1.2%	2.0%	2.3%
Μ.Βρετανία	2.2%	3.1%	7.0%	5.9%
Η.Π.Α	5.5%	4.9%	7.1%	5.3%
Δ.Γερμανία	1.1%	0.5%	2.9%	5.5%

α) Με βάση τα στοιχεία αυτά, ποιά από τις επτά αυτές χώρες είχε το μικρότερο ποσοστό ανεργίας; Ποιά είχε το μεγαλύτερο; β) Να υπολογισθεί το μέσο ποσοστό ανεργίας στα έτη 1960, 1970, 1980 και 1990. Υπάρχει κάποια από τις επτά χώρες για τις οποίες διαθέτουμε στοιχεία η οποία να έχει ποσοστό ανεργίας σταθερά μικρότερο από το μέσο όρο; Υπάρχει κάποια άλλη που έχει ποσοστό ανεργίας σταθερά μεγαλύτερο από το μέσο όρο;

3.62 Ο Ronald Reagan στην προεκλογική εκστρατεία του 1980 για την εκλογή προέδρου των Η.Π.Α. ισχυρίσθηκε επανειλημμένα οτι το 1980 οι ψηφοφόροι βρίσκονταν σε καλύτερη οικονομική κατάσταση από ότι τέσσερα χρόνια πριν, όταν ο απερχόμενος πρόεδρος Carter (που ήταν αντίπαλος του Reagan) είχε εκλεγεί πρόεδρος. Ο πίνακας που ακολουθεί δίνει κάποια στοιχεία για την ανεργία και τον πληθωρισμό κατά τη διάρκεια των τεσσάρων ετών προεδρίας του Carter (1977-1980) και των οκτώ χρόνων προεδρίας του Reagan (1981-1988) στις ΗΠΑ.

Ετος	Ποσοστό ανεργίας	Πληθωρισμός
1977	7.1%	6.7%
1978	6.1%	9.0%
1979	5.8%	13.3%
1980	7.1%	12.5%
1981	7.6%	8.9%
1982	9.7%	3.8%
1983	9.6%	3.8%
1984	7.5%	3.9%
1985	7.2%	3.8%
1986	7.0%	1.1%
1987	6.2%	4.4%
1988	5.5%	4.4%

Να υπολογισθεί το μέσο ποσοστό ανεργίας και ο μέσος πληθωρισμός κατά τη διάρκεια των ετών διακυβέρνησης του Carter και του Reagan αντίστοιχα. Υπάρχει κάποια σημαντική διαφορά στην ανεργία ή στον πληθωρισμό μεταξύ των δύο διαφορετικών κυβερνήσεων;

3.63 Κάποιος που οδηγεί με 90km/h στην εθνική οδό παρατηρεί ότι ο αροθμός των αυτοκινήτων που τον προσπερνούν είναι ίσος με τον αριθμό των αυτοκινήτων που ο ίδιος προσπερνά. Σημαίνει αυτό ότι η ταχύτητα των 90km/h είναι ο μέσος, η διάμεσος ή η επικρατούσα τιμή της ταχύτητας;