Κεφάλαιο 2

ΕΡΕΥΝΗΤΙΚΗ ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ

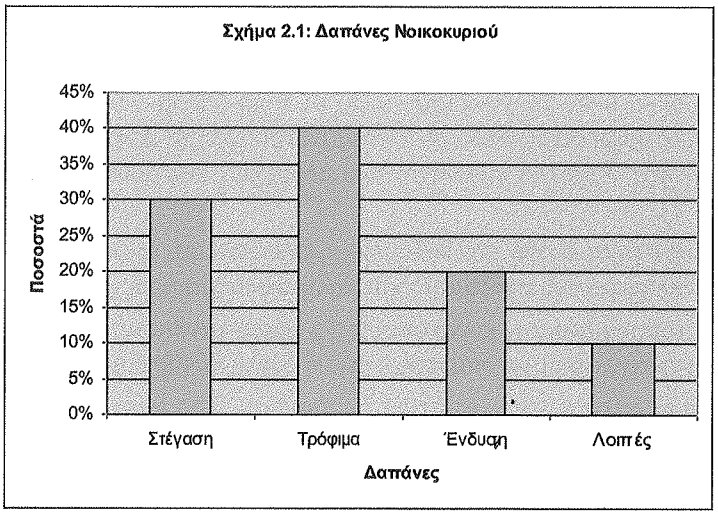
Η πραγματοποίηση ενός ερευνητικού έργου απαιτεί την εύρεση και σωστή χρήση δεδομένων σχετικών με το υπό εξέταση φαινόμενο. Ένα *σύνολο δεδομένων* (data set) περιλαμβάνει όλα τα στοιχεία που έχουν συλλεχθεί σε μια συγκεκριμένη περιοχή μελέτης. Ένα *στοιχείο* του συνόλου δεδομένων είναι στην ουσία ένα δεδομένο που έχει επιλεγεί. Μετά τη συλλογή πρέπει να αναλύσουμε το σύνολο δεδομένων, έτσι ώστε να συλλάβουμε κάθε υπάρχουσα σχέση που συνεπάγονται τα δεδομένα αυτά. Η εφαρμογή συγκεκριμένων στατιστικών μεθόδων, όπως οι γραφικές παραστάσεις και οι διάφορες στατιστικές μετρήσεις, είναι γνωστές ως *ερευνητική ανάλυση δεδομένων*. Αυτό είναι το πρώτο βήμα προκειμένου να γνωρίσουμε το σύνολο δεδομένων που μας ενδιαφέρει. Αξίζει να σημειώσουμε ότι τα δεδομένα αυτά μπορεί να μην αντιπροσωπεύουν πλήρως την πραγματικότητα. Υπάρχει πιθανότητα να περιέχουν ένα αριθμό λαθών που οφείλονται σε:

1. Προβλήματα στη συλλογή των δεδομένων, όπως ελλιπή ερωτηματολόγια, μη τυχαιότητα στον πειραματικό σχεδιασμό και πολυπλοκότητα στις μεθόδους επιλογής.
2. Σφάλματα τυπογραφικά ή κωδικοποίησης, όπως λανθασμένη δακτυλογράφηση ή εσφαλμένη αντιγραφή μιας παρατήρησης ή διπλή καταχώρηση ενός δεδομένου.
3. Παράλειψη παρατηρήσεων, εξαιτίας της άρνησης του ανταποκρινόμενου να απαντήσει σε μια ερώτηση.
4. Παράτυπες ή ακραίες τιμές (outliers), όπου μία παρατήρηση εμφανίζεται ασυνεπής με τις υπόλοιπες παρατηρήσεις. Μια ακραία τιμή μπορεί να είναι σφάλμα κωδικοποίησης ή αληθινή αξία, και είναι απαραίτητο να ελέγξουμε το σύνολο των δεδομένων προκειμένου να δούμε τι πραγματικά συμβαίνει στην περίπτωσή μας.

Τα βασικά εργαλεία για την ερευνητική ανάλυση δεδομένων (Ε.Α.Δ.) αποτελούνται από γραφικές παραστάσεις (ιστογράμματα, ραβδογράμματα, κυκλικά διαγράμματα κ.ά.) και περιγραφικές τεχνικές με μετρήσεις θέσης και κεντρικής τάσης (μέσος, διάμεσος, επικρατούσα τιμή), καθώς και μετρήσεις διασποράς και μεταβλητότητας (εύρος, τεταρτημόρια, τυπική απόκλιση, διακύμανση, συντελεστής μεταβλητότητας).

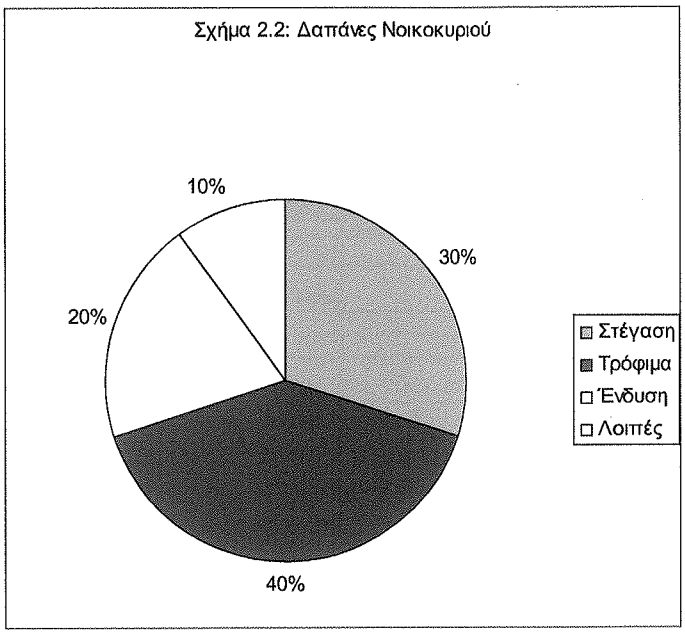
2.1 Γραφική παράσταση των δεδομένων

Η γραφική παράσταση του συνόλου των δεδομένων μας βοηθά να αντιληφθούμε πώς κατανέμονται τα δεδομένα, να αναγνωρίσουμε την κατανομή των επιλεγμένων τιμών και συγχρόνως να ελέγξουμε αν υπάρχουν ακραίες τιμές που μπορούν να επηρεάσουν τη στατιστική ανάλυση. Τα ποιοτικά δεδομένα (ονομαστικής κλίμακας) μπορούν να παρασταθούν γραφικά με ραβδογράμματα και κυκλικά διαγράμματα ή πίτες. Η πρώτη μέθοδος γραφικής παράστασης περιγράφει ποσότητες για διαφορετικές κατηγορίες δεδομένων με μια σειρά από ράβδους. Οι ράβδοι αυτές έχουν μορφή ορθογωνίου παραλληλογράμμου με ύψος ίσο με τη συχνότητα της αντίστοιχης κατηγορίας. Είναι δηλαδή ένα γράφημα στο οποίο η συχνότητα ή η αρίθμηση των παρατηρήσεων κάθε κατηγορίας παρουσιάζεται με το ύψος ή το μήκος κάθε ράβδου. Οι ράβδοι μπορούν να απεικονιστούν είτε οριζόντια είτε κάθετα. Το σχήμα 2.1 δείχνει ένα παράδειγμα ραβδογράμματος, με τον κάθετο άξονα να παρουσιάζει κάποια ποσοστά και τον οριζόντιο άξονα να παρουσιάζει υποθετικές δαπάνες κάποιων νοικοκυριών.



Το γράφημα σε μορφή κυκλικού διαγράμματος (πίτας) είναι ένα γράφημα στο οποίο ολόκληρος ο αριθμός των παρατηρήσεων παρουσιάζεται με ένα κύκλο και η αναλογία κάθε κατηγορίας παριστάνεται ως κομμάτι του κύκλου. Το κυκλικό διάγραμμα παρουσι-

άζει ποσοστά σε σχέση με το σύνολο. Αυτό το είδος γραφήματος είναι πολύ εύχρηστο σε μελέτες οικονομικής ανάλυσης και σε αναφορές προϋπολογισμού. Το σχήμα 2.2 δείχνει ένα παράδειγμα γραφήματος σε μορφή κυκλικού διαγράμματος.[1] Συγκεκριμένα παρουσιάζονται οι δαπάνες για στέγαση, τρόφιμα, ενδύματα και λοιπές δαπάνες.



Για τη σχεδίαση ποσοτικών δεδομένων (διαστημικής και αναλογικής κλίμακας) μπορούμε να χρησιμοποιήσουμε ένα *ιστόγραμμα*.[2] Αυτό μοιάζει με ένα ραβδόγραμμα, οι τάξεις του οποίου είναι τοποθετημένες με αριθμητική σειρά στον οριζόντιο άξονα και το εμβαδόν κάθε ράβδου καθορίζεται από τον αριθμό των παρατηρήσεων της αντίστοιχης τάξης. Δηλαδή, το ιστόγραμμα είναι ένα σχήμα που κατασκευάζεται από ράβδους διαφορετικού ύψους, με το ύψος κάθε ράβδου να παρουσιάζει τη συχνότητα των τιμών στην

[1] Αν θέλουμε να χρησιμοποιήσουμε εικόνες, προκειμένου να παρουσιάσουμε τις συχνότητες ενός αντικειμένου μελέτης, τότε μπορούμε να χρησιμοποιήσουμε μία άλλη γραφική παράσταση, το γράφημα μέσω εικόνας (pictograph). Για παράδειγμα οι πωλήσεις τηλεοράσεων μπορούν να παρουσιαστούν με τη χρήση του σχήματος της τηλεόρασης και με αντιστοιχία, για παράδειγμα, 10.000 πωλήσεων ανά τηλεόραση. Μισή τηλεόραση αντιστοιχεί σε 5.000 πωλήσεις τηλεοράσεων και ομοίως για τα υπόλοιπα.

[2] Άλλοι τρόποι γραφικής παρουσίασης τέτοιων δεδομένων είναι τα φυλλογραφήματα (stem and leaf) και τα θηκογράμματα (boxplot). Τα θηκογράμματα θα εξετασθούν παρακάτω, ενώ τα φυλλογραφήματα παραλείπονται, καθώς τα συμπεράσματα από την ανάλυσή τους είναι όμοια με αυτά των ιστογραμμάτων.

τάξη που αντιπροσωπεύεται από τη ράβδο. Μια διαφορά μεταξύ του ιστογράμματος και ενός ραβδογράμματος είναι ότι στο ιστόγραμμα ο οριζόντιος άξονας είναι ταξινομημένος σε ομάδες που αντιπροσωπεύουν τα δεδομένα διαστημικής ή αναλογικής κλίμακας. Άλλη διαφορά είναι ότι ο κάθετος άξονας είναι ταξινομημένος με τέτοιον τρόπο, ώστε το εμβαδόν κάθε ράβδου να ισούται με τη συχνότητα της αντίστοιχης τάξης.

Παράδειγμα 2.1

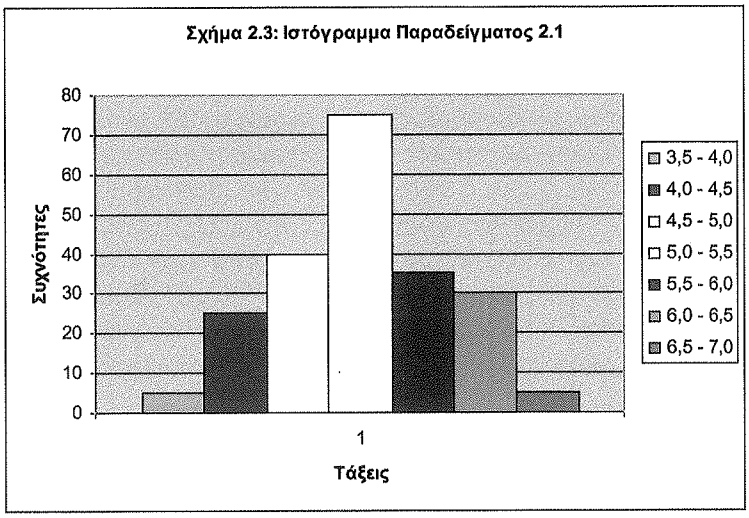
Υποθέστε ότι έχουμε τα ακόλουθα ετήσια επίπεδα εισοδήματος για 215 οικογένειες σε μια συνοικία.

Εισόδημα σε χιλιάδες €	**Αριθμός οικογενειών**
3,5 - 4,0	5
4,0 - 4,5	25
4,5 - 5,0	40
5,0 - 5,5	75
5,5 - 6,0	35
6,0 - 6,5	30
6,5 - 7,0	5
	Σ= 215

Ο αριθμός των ατόμων που ανήκουν σε κάθε τάξη αποτελεί τη συχνότητα της τάξης. Έτσι, για παράδειγμα η συχνότητα της τάξης 3,5 - 4,0 είναι 5. Η λίστα των τάξεων και των συχνοτήτων ονομάζεται *κατανομή συχνοτήτων*. Οι εγγραφές 3,5 - 4,0, 4,0 - 4,5, ..., 6,5 - 7,0 ονομάζονται *τάξεις διαστημάτων*. Το 3,5 είναι το χαμηλότερο όριο και το 4,0 είναι το υψηλότερο όριο της πρώτης τάξης. Το μέσο σημείο της τάξης (midpoint) είναι ο μέσος όρος της υψηλότερης και της χαμηλότερης τιμής κάθε τάξης, δηλαδή

$$\text{Αριθμητικός μέσος τάξης} = \frac{(\text{χαμηλότερο όριο τάξης} + \text{υψηλότερο όριο τάξης})}{2} = \frac{(3{,}5+4{,}0)}{2} = 3{,}75$$

Το Σχήμα 2.3 παρουσιάζει ένα ιστόγραμμα για το σύνολο των δεδομένων του παραδείγματος 2.1. Τα ιστογράμματα μάς δίνουν μια ιδέα για τη μορφή της κατανομής των δεδομένων. Έτσι, στην περίπτωση του παραδείγματος 2.1 φαίνεται ότι η πιο κοινή σειρά οικογενειακού εισοδήματος είναι ανάμεσα σε 5 και 5,5 χιλιάδες €, ενώ οι συχνότητες των τάξεων του πρώτου και του τελευταίου διαστήματος είναι πολύ χαμηλές.



Όταν τα διαστήματα των τάξεων είναι ίσα, τότε στον κάθετο άξονα χρησιμοποιούμε τις αντίστοιχες συχνότητες. Αν τα διαστήματα δεν είναι ίσα, τότε είναι προτιμότερο να χρησιμοποιούμε για τον κάθετο άξονα τις «πυκνότητες». *Πυκνότητα* είναι ο αριθμός των παρατηρήσεων σε ένα διάστημα (f_i), διαιρούμενο με το εύρος (πλάτος) του διαστήματος αυτού (δ_i). Μια κατακόρυφη γραμμή δείχνει τις τιμές των παρατηρήσεων στον άξονα των X και τις συχνότητες στον άξονα των Y. Το ύψος κάθε κατακόρυφης γραμμής δείχνει τον αριθμό των παρατηρήσεων σε κάθε τιμή.

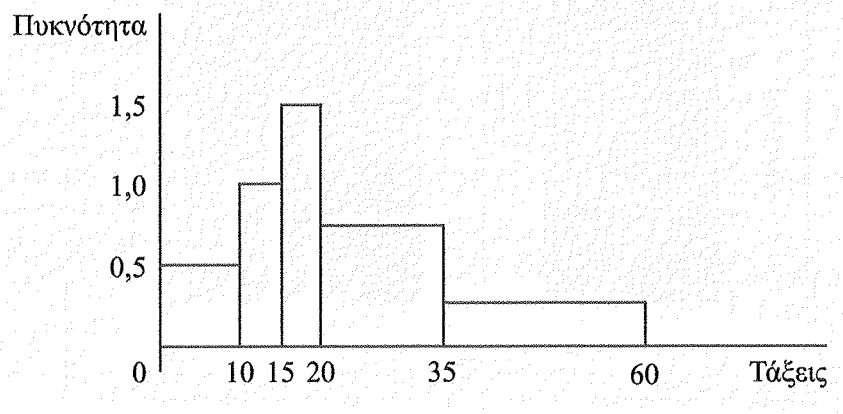
Παράδειγμα 2.2

Υποθέστε ότι έχουμε τα ακόλουθα ομαδοποιημένα δεδομένα με τάξεις άνισων μεγεθών.

Τάξεις	Συχνότητες (f_i)	Πυκνότητα (f_i / δ_i)
0-10	5	0,5
10-15	5	1,0
15-20	8	1,6
20-35	10	0,7
35-60	6	0,24
	$\Sigma = 34$	

Το σχήμα 2.4 δείχνει ένα ιστόγραμμα στην περίπτωση άνισων τάξεων και σύμφωνα με τα δεδομένα του παραδείγματος 2.2. Όπως έχει αναφερθεί, στην περίπτωση άνισων τάξεων χρησιμοποιούμε πυκνότητες αντί για συχνότητες. Αξίζει να σημειωθεί ότι είναι δυνατόν να κατασκευάσουμε ένα πολύγωνο συχνοτήτων από ένα ιστόγραμμα συνδέοντας τα μεσαία σημεία στην κορυφή της κάθε ράβδου.

Σχήμα 2.4: Ιστόγραμμα με άνισες τάξεις



2.2 Δημιουργία ομαδοποιημένων δεδομένων

Πολλές φορές κρίνεται απαραίτητο να παρουσιάσουμε τα συλλεχθέντα στοιχεία με τέτοιον τρόπο, ώστε να είναι δυνατή η γραφική τους παρουσίαση και η ποσοτική τους ανάλυση. Αυτό απαιτεί τακτοποίηση των δεδομένων με την έννοια της ομαδοποίησης. Ας δούμε τα βασικά βήματα για ομαδοποίηση δεδομένων με τη χρήση ενός παραδείγματος.

Παράδειγμα 2.3

Ας υποθέσουμε ότι έχουμε συλλέξει το ακόλουθο σύνολο δεδομένων και πρέπει να χωρίσουμε τα δεδομένα αυτά τουλάχιστον σε πέντε (5) τάξεις.

55	90	81	70	64	39	52	75	83	99	60	47	81	67	58	75	90	89	54	58	58	69	52
70	81	52	83	62	87	62	69	91	45	66	51	73	60	86	52	71	55	41	89	56	71	45
78	34	55	51	62	52	84	50	53	79	51	60	93	58	53	64	55	71	83	69	54	59	50

α) Αρχικά οι αριθμοί πρέπει να καταταχθούν με σειρά (π.χ. από το μικρότερο στο μεγαλύτερο). Με τη στατιστική ορολογία, η διαδικασία αυτή ονομάζεται *διάταξη* και είναι η λίστα των παρατηρήσεων σε ένα καθορισμένο σύνολο δεδομένων με αύξουσα ή φθίνουσα σειρά.

β) Βρίσκουμε το εύρος των μετρήσεων, που είναι η διαφορά ανάμεσα στη μεγαλύτερη και τη μικρότερη μέτρηση. Στην περίπτωσή μας έχουμε 69 μετρήσεις τής υπό εξέταση τυχαίας μεταβλητής. Ο μικρότερος αριθμός είναι το 34 και ο μεγαλύτερος

το 99. Η απόσταση ανάμεσα στη μεγαλύτερη και τη μικρότερη μέτρηση είναι γνωστή ως *εύρος* και στη συγκεκριμένη περίπτωση ισούται με 99-34=65.

γ) Διαιρούμε το εύρος των μετρήσεων με τον αριθμό των τάξεων που έχουμε σκοπό να δημιουργήσουμε. Στη συνέχεια στρογγυλοποιούμε το αποτέλεσμα στην πιο κοντινή ακέραια μονάδα, προκειμένου να γίνουν πιο εύκολα οι υπολογισμοί. Στο παράδειγμα αυτό, χρειαζόμαστε τουλάχιστον 5 τάξεις. Αν διαιρέσουμε 65/7=9,29 για επτά (7) τάξεις. Η επιλογή του αριθμού των τάξεων γίνεται εμπειρικά. Εάν έχουμε πάρα πολύ λίγες τάξεις, πιθανόν να αποκρύψουμε σημαντικά χαρακτηριστικά των συλλεχθέντων στοιχείων, ενώ αν έχουμε πάρα πολλές τάξεις, μπορεί να δημιουργηθεί σύγχυση και οι υπολογισμοί να γίνουν επίπονοι. Είναι λογικό να στρογγυλοποιήσουμε το 9,29 σε 10 για να διευκολυνθούμε.

δ) Η πρώτη τάξη διαστήματος πρέπει να περιλαμβάνει τη μικρότερη μέτρηση και η τελευταία τάξη διαστήματος τη μεγαλύτερη μέτρηση. Κατόπιν, βρίσκουμε τη συχνότητα κάθε τάξης μετρώντας κάθε παρατήρηση που ανήκει σε μία συγκεκριμένη τάξη. Αυτό είναι αρκετά επίπονη εργασία, αλλά μπορεί να μας δείξει ξεκάθαρα πως κατανέμονται τα δεδομένα.

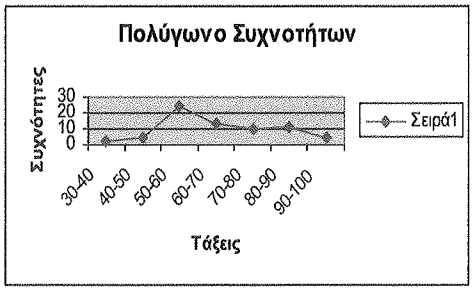
Αφού μετρήσουμε τις παρατηρήσεις και τις κατατάξουμε στις αντίστοιχες τάξεις, τότε μπορούμε να βρούμε κάποιες άλλες ενδιαφέρουσες συχνότητες. Οι σχετικές συχνότητες μπορούν να παραχθούν από τη διαίρεση τής κάθε συχνότητας με το συνολικό αριθμό των παρατηρήσεων σε ολόκληρη την κατανομή. Ομοίως, οι *αθροιστικές* (σωρευτικές) *συχνότητες* βρίσκονται αθροίζοντας τις συχνότητες καθώς μετακινούμαστε από τάξη σε τάξη. Για παράδειγμα, για την τρίτη τάξη η σωρευτική συχνότητα θα είναι οι συχνότητες της πρώτης, δεύτερης και τρίτης τάξης μαζί. Τέλος, υπάρχουν και οι *αθροιστικές σχετικές συχνότητες*, που όπως υποδηλώνει η ονομασία τους, δημιουργούνται από την άθροιση των σχετικών συχνοτήτων καθώς κινούμαστε από τάξη σε τάξη. Στην δική μας περίπτωση, έχουμε:

Τάξεις	Συχνότητες (f_i)	Σχετικές Συχνότητες ($f_i / \Sigma_i f_i$)	Αθροιστικές Συχνότητες	Αθροιστικές Σχετικές Συχνότητες
30 και κάτω από 40	2	0,03	2	0,03
40 και κάτω από 50	4	0,06	6	0,09
50 και κάτω από 60	24	0,35	30	0,44
60 και κάτω από 70	13	0,19	43	0,63
70 και κάτω από 80	10	0,14	53	0,77
80 και κάτω από 90	11	0,16	64	0,93
90 και κάτω από 100	5	0,07	69	1,00

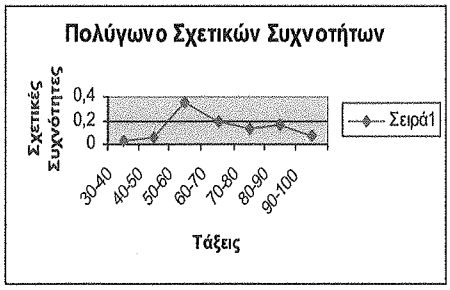
Το διάγραμμα που σε γενικές γραμμές χρησιμοποιείται για συνεχή δεδομένα διαστημικής και αναλογικής κλίμακας είναι το *πολύγωνο συχνοτήτων*, που ενώνει μια σειρά από σημεία, καθένα από τα οποία είναι τοποθετημένο στο μέσο τής κάθε τάξης κατά μήκος του κάθετου άξονα. Το πολύγωνο συχνοτήτων είναι μια τεθλασμένη γραμμή από διανεμημένες συχνότητες. Μία παρέκκλιση του πολυγώνου συχνοτήτων είναι το *πολύγωνο σχετικών συχνοτήτων*, το οποίο είναι όμοιο με το πολύγωνο συχνοτήτων, με τη μόνη διαφορά ότι στον κάθετο άξονα τοποθετούμε τις σχετικές συχνότητες. Έτσι, κάθε σχετική συχνότητα είναι μια αναλογία.

Μια άλλη παρέκκλιση είναι το *πολύγωνο των αθροιστικών συχνοτήτων* ή ogive. Το ogive είναι μία καμπύλη που εξάγεται από την ίδια συχνότητα δεδομένων, όπως το ιστόγραμμα. Σχεδιάζουμε τις αθροιστικές συχνότητες από τα υψηλότερα άκρα τής κάθε τάξης διαστημάτων με τέτοιον τρόπο, ώστε κάθε σημείο να δίνει ένα ποσοστό των δεδομένων σε όρους των μεταβλητών που μελετάμε. Αυτή η καμπύλη μπορεί να χρησιμοποιηθεί για να εκτιμήσουμε εκατοστημόρια (percentiles) της κατανομής των δεδομένων. Οι γραφικές παραστάσεις για το δικό μας παράδειγμα παρουσιάζονται στα παρακάτω σχήματα.

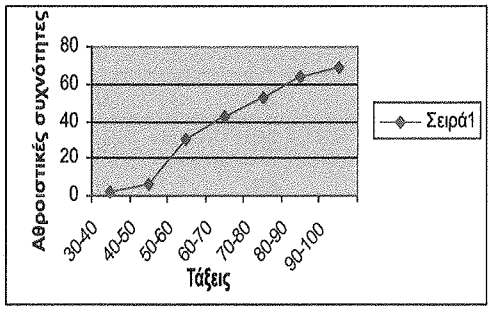
Σχήμα 2.5:
Πολύγωνο συχνοτήτων



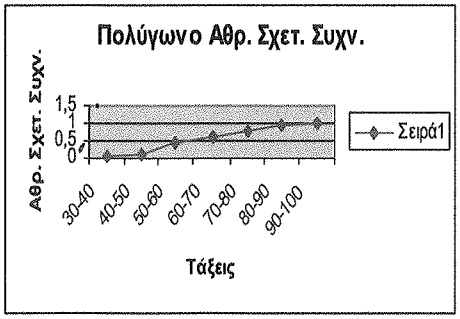
Σχήμα 2.6:
Πολύγωνο σχετικών συχνοτήτων



Σχήμα 2.7:
Πολύγωνο αθροιστικών συχνοτήτων



Σχήμα 2.8: Πολύγωνο
αθροιστικών σχετικών συχνοτήτων



2.3 Αριθμητική περιγραφή των δεδομένων

Μετά τη γραφική παράσταση των δεδομένων απαιτείται η αριθμητική τους επεξεργασία, με σκοπό την εύρεση ποσοτικών πληροφοριών σχετικά με τη θέση, την κεντρική τάση και τη διασπορά των διαφόρων μετρήσεων.

2.3.1 *Μέτρα θέσης και κεντρικής τάσης (measures of location and central tendency)*

Το πιο σημαντικό αριθμητικό μέτρο θέσης για μία μεταβλητή είναι ο *αριθμητικός μέσος* ή η *μέση τιμή*. Είναι εύκολο να υπολογίσουμε το μέσο, αφού το μόνο που πρέπει να κάνουμε είναι να προσθέσουμε όλες τις παρατηρήσεις και το άθροισμά τους να το διαιρέσουμε με το συνολικό αριθμό των παρατηρήσεων αυτών. Για να παραστήσουμε το αλγεβρικό άθροισμα χρησιμοποιούμε το ελληνικό γράμμα Σ που εκφράζει μαθηματικά την άθροιση. Τα όρια του αθροίσματος αυτού αναφέρονται ως δείκτες των υπό άθροιση μετρήσεων της μεταβλητής και παίρνουν τιμές από i = 1 ... n (όπου n είναι ο συνολικός αριθμός των παρατηρήσεων). Έτσι, σύμφωνα με τις σχέσεις που ακολουθούν, ο μέσος ισούται με το άθροισμα όλων των παρατηρήσεων από την πρώτη τιμή της μεταβλητής μέχρι την n-οστή τιμή, διαιρεμένο με το συνολικό αριθμό των παρατηρήσεων. Έχοντας υπόψη τη διάκριση ανάμεσα στο δείγμα και στον πληθυσμό και θυμίζοντας ότι ελληνικά γράμματα αναφέρονται στον πληθυσμό, ενώ τα αγγλικά γράμματα στο δείγμα, τότε οι τύποι υπολογισμού του μέσου σε περιπτώσεις χρήσης δειγμάτων ή πληθυσμών είναι:[3]

Σε περίπτωση ανάλυσης ενός δείγματος

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} \qquad (2.1)$$

Σε περίπτωση ανάλυσης ενός πληθυσμού

$$\mu = \frac{\sum_{i=1}^{\nu} X_i}{\nu} \qquad (2.2)$$

[3] Για απλούστευση πολλών αλγεβρικών εκφράσεων θα χρησιμοποιήσουμε το σύμβολο της άθροισης (Σ). Για μια τυχαία μεταβλητή X η άθροιση των τιμών $x_1, x_2, ..., x_n$ της μεταβλητής αυτής μπορεί να εκφραστεί ως $x_1+x_2+...+x_n$ και να απλουστευτεί ως $\sum_{i=1}^{n} X_i$. Αυτό σημαίνει ότι $\sum_{i=1}^{n} X_i = x_1 + x_2 + ... + x_n$. Το σύμβολο Σ σημαίνει "άθροιση" και ο δείκτης *i* είναι ο δείκτης άθροισης από την πρώτη ως τη n-οστή παρατήρηση. Η έκφραση $\sum_{i=1}^{n}$ μπορεί να διαβαστεί ως "άθροιση των τιμών της μεταβλητής X από 1 έως n". Ο αριθμός 1 αποτελεί το χαμηλό και η τιμή n το υψηλό όριο της άθροισης. Συχνά γράφουμε $\Sigma_i X_i$ ή ΣX_i αντί $\Sigma_{i=1}^{n} X_i$.

Ο μέσος γίνεται εύκολα κατανοητός. Το μειονέκτημά του είναι ότι είναι πολύ ευαίσθητος, καθώς επηρεάζεται εύκολα από τις ακραίες τιμές που μπορεί να έχουν οι μετρήσεις μιας μεταβλητής. Αυτές είναι τιμές της μεταβλητής που είναι κατά πολύ μικρότερες ή/και κατά πολύ μεγαλύτερες από την πλειοψηφία των υπολοίπων τιμών της μεταβλητής. Αν πιστεύουμε πως αυτές οι τιμές δεν είναι αντιπροσωπευτικές και μπορούν να μας οδηγήσουν σε λανθασμένα συμπεράσματα, μπορούμε να μην τις λάβουμε υπόψη μας στον υπολογισμό του μέσου. Με αυτό τον τρόπο υπολογίζουμε τον *τετριμμένο μέσο* (trimmed mean). Αυτός υπολογίζεται αν παραλείψουμε ένα καθορισμένο ποσοστό (συνήθως 5% ή 10%) από τις μικρότερες και τις μεγαλύτερες παρατηρήσεις.

Παράδειγμα 2.4

Υποθέστε ότι έχουμε συλλέξει τους παρακάτω αριθμούς

1 5 9 12 15 28

Ο μέσος αυτών των αριθμών είναι 11,66. Αυτός βρίσκεται χρησιμοποιώντας τον τύπο (2.1) που παρουσιάστηκε παραπάνω. Δηλαδή:

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{1+5+9+12+15+28}{6} = \frac{70}{6} = 11{,}66$$

Προφανώς όμως ο πρώτος, και κυρίως, ο τελευταίος αριθμός επηρεάζουν σημαντικά το μέσο. Αν παραλείψουμε αυτές τις δύο τιμές (την πρώτη και την τελευταία), τότε ο τετριμμένος μέσος θα είναι 10,25 σύμφωνα με τη σχέση:

$$\overline{X}_T = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{5+9+12+15}{4} = \frac{41}{4} = 10{,}25$$

όπου ο δείκτης Τ υποδηλώνει τον τετριμμένο μέσο.

Αν έχουμε τώρα δύο ομάδες δεδομένων και γνωρίζουμε τους μέσους των δεδομένων αυτών και τον αριθμό των παρατηρήσεων της κάθε ομάδας δεδομένων, μπορούμε να υπολογίσουμε το μέσο αριθμητικό της ομάδας των δεδομένων που θα περιλαμβάνει όλα τα δεδομένα των δύο διαφορετικών ομάδων που είχαμε αρχικά. Αυτός είναι ο *σταθμικός μέσος* και υπολογίζεται από τον παρακάτω τύπο:

$$\overline{X} = \frac{n_1 \overline{X}_1 + n_2 \overline{X}_2}{n_1 + n_2}$$

Αυτή η σχέση μπορεί να χρησιμοποιηθεί και για n ομάδες δεδομένων σύμφωνα με τον τύπο:

$$\overline{X} = \frac{\sum_{i=1}^{n} n_i \overline{X}_i}{\sum_{i=1}^{n} n_i} = \frac{n_1 \overline{X}_1 + n_2 \overline{X}_2 + \ldots + n_n \overline{X}_n}{n_1 + n_2 + \ldots + n_n} \qquad (2.3)$$

Αν οι ομάδες δεδομένων έχουν τον ίδιο αριθμό παρατηρήσεων, ο μέσος της συνδυασμένης ομάδας δεδομένων είναι απλά ο μέσος των μέσων αριθμητικών όλων αυτών των ομάδων δεδομένων.

Παράδειγμα 2.5

Μια βιομηχανική μονάδα χρησιμοποίησε τα ακόλουθα αποθέματα πετρελαίου στις παραγωγικές της διαδικασίες: 5.000 βαρέλια ελαφρύ ανεπεξέργαστου πετρελαίου, 20.000 βαρέλια πετρελαίου υψηλής συγκέντρωσης σε θείο, 1.000 βαρέλια πετρελαίου χαμηλής συγκέντρωσης σε θείο και 6.000 βαρέλια μεσαίου βάρους ακατέργαστου πετρελαίου. Αν το κόστος για τα διάφορα είδη του πετρελαίου είναι 29, 24, 38 και 26 νομισματικές μονάδες (ν.μ.) αντιστοίχως, ποιο είναι το μέσο κόστος ανά βαρέλι πετρελαίου που χρησιμοποιήθηκε συνολικά στην παραγωγική διαδικασία;

$$\overline{X} = \frac{\sum_{i=1}^{4} n_i \overline{X}_i}{\sum_{i=1}^{4} n_i} = \frac{n_1 \overline{X}_1 + n_2 \overline{X}_2 + n_3 \overline{X}_3 + n_4 \overline{X}_4}{n_1 + n_2 + n_3 + n_4} =$$

$$= \frac{5.000(29) + 20.000(24) + 1.000(38) + 6.000(26)}{5.000 + 20.000 + 1.000 + 6.000} = \frac{819.000}{32.000} = 25{,}59 \text{ ν.μ.}$$

Το μέσο κόστος είναι 25,59 νομισματικές μονάδες ανά βαρέλι. Αυτό το υπολογίσαμε από το πηλίκο του αθροίσματος των γινομένων των τιμών του κάθε είδους πετρελαϊκού αποθέματος, πολλαπλασιασμένων με τον αριθμό των βαρελιών, δια του συνολικού αριθμού των βαρελιών.

Στην περίπτωση που έχουμε ομαδοποιημένα δεδομένα, ο μέσος δίνεται από τη σχέση:

$$\overline{X} = \frac{\sum_{j=1}^{k} f_j X_{mj}}{\sum_{j} f_j} \qquad (2.4)$$

όπου X_{mj} είναι το μέσο σημείο της j ομάδας (midpoint), το k ισούται με τον αριθμό των ομάδων και f_j είναι η συχνότητα της ομάδας j. Το παράδειγμα που ακολουθεί δείχνει τη χρήση του μέσου σε περίπτωση ομαδοποιημένων δεδομένων.

Παράδειγμα 2.6

Υποθέστε πως έχουμε τον παρακάτω πίνακα, όπου στις δύο πρώτες στήλες έχουμε τις διάφορες κατηγορίες εισοδήματος (σε εβδομαδιαία βάση) και τον αριθμό των εργατών σε μια γειτονιά. Αν θέλουμε να βρούμε το μέσο εισόδημα των εργατών αυτών, τότε θα πρέπει να υπολογίσουμε το μέσο, πολλαπλασιάζοντας αρχικά το μέσο σημείο της κάθε τάξης με την αντίστοιχη συχνότητα (αριθμό εργατών) της κάθε τάξης. Αυτοί οι υπολογισμοί φαίνονται στον παρακάτω πίνακα:

Τάξεις εισοδημάτων (σε χιλιάδες ν.μ.)	**Αριθμός εργατών (f_j)**	**Μέσο σημείο τάξης (X_{mj})**	**($f_j X_{mj}$)**
70-80	4	75	300
80-90	8	85	680
90-100	18	95	1.710
100-110	35	105	3.675
110-120	15	115	1.725
	$\Sigma = 80$		$\Sigma = 8.090$

Κατόπιν, το άθροισμα αυτών των πολλαπλασιασμών αποτελεί τον αριθμητή στη σχέση που χρησιμοποιούμε για τον υπολογισμό του μέσου. Ο μέσος υπολογίζεται διαιρώντας το άθροισμα της τελευταίας στήλης με το συνολικό αριθμό των εργατών.

$$\overline{X} = \frac{\sum_{j=1}^{5} f_j X_{mj}}{\sum_{j} f_j} = \frac{8.090}{80} = 101{,}125$$

Αξίζει να αναφέρουμε ότι εκτός από τον αριθμητικό μέσο έχουμε και το γεωμετρικό μέσο (geometric mean) που είναι αρκετά χρήσιμος στον υπολογισμό του ρυθμού της μεταβολής. Υπολογίζεται από την ακόλουθη σχέση:

$$\overline{X}_g = \sqrt[n]{x_1 x_2 x_3 \ldots x_n} \qquad (2.5)$$

Παράδειγμα 2.7

Υποθέστε ότι οι συνολικές πωλήσεις μιας επιχείρησης (σε χιλιάδες €) για τα χρόνια μεταξύ 2016 και 2019 ήταν :

Έτος	2016	2017	2018	2019
Πωλήσεις	150	155	180	210

Για να βρούμε τον ετήσιο μέσο ρυθμό αύξησης των πωλήσεων αυτή την περίοδο των 4 ετών, θα πρέπει να υπολογίσουμε την πραγματική ποσοστιαία μεταβολή από έτος σε έτος. Στο παράδειγμά μας έχουμε:

2016-2017 3,33%
2017-2018 16,13%
2018-2019 16,70%

Ο αριθμητικός μέσος αυτών των τιμών δεν είναι η κατάλληλη μέτρηση, αφού αυτά τα ποσοστά υπολογίζονται το καθένα χρησιμοποιώντας διαφορετική βάση. Παίρνοντας το γεωμετρικό μέσο έχουμε:

$$\bar{X}_g = \sqrt[n]{x_1 x_2 x_3 ... x_n} = \sqrt[3]{3,33x16,13x16,7} = 9,64$$

ανά έτος μεταξύ του 2016 και 2019. Για να ορίσουμε το μέσο ετήσιο ρυθμό αύξησης αυτής της περιόδου, μπορούμε να χρησιμοποιήσουμε την ακόλουθη έκφραση:

$$r = \left(\sqrt[n-1]{\frac{x_n}{x_1}} - 1 \right) * 100 \qquad (2.6)$$

όπου r είναι ο ποσοστιαίος ρυθμός αύξησης και x_1 και x_n είναι η πρώτη και η τελευταία τιμή στην αρχική σειρά των n τιμών. Έτσι, στην περίπτωση του παραδείγματος που είδαμε, ο μέσος ρυθμός αύξησης είναι:

$$r = \left(\sqrt[n-1]{\frac{x_n}{x_1}} - 1 \right) x100 = \left(\sqrt[4-1]{\frac{210}{150}} - 1 \right) x100 = 11,87 \text{ ετησίως}$$

Τα δύο βασικά μεγέθη μέτρησης της κεντρικής τάσης είναι η διάμεσος και η επικρατούσα τιμή. Η **διάμεσος** (median) είναι εκείνο το σημείο σε μια σειρά δεδομένων που έχει ίσο αριθμό παρατηρήσεων πριν και μετά από αυτό, βρίσκεται δηλαδή στο

μέσο των παρατηρήσεων. Αναφέρουμε πως έχουμε μια σειρά δεδομένων όταν τα δεδομένα είναι καταταγμένα σε αύξουσα ή φθίνουσα κλίμακα. Αντίθετα με το μέσο αριθμητικό, η διάμεσος δεν επηρεάζεται από τις ακραίες παρατηρήσεις.

Η διάμεσος είναι η [(n+1)/2] παρατήρηση από τα καταταγμένα δεδομένα, όπου n είναι ο συνολικός αριθμός των παρατηρήσεων. Όταν τα δεδομένα μας αποτελούνται από περιττό αριθμό παρατηρήσεων η διάμεσος ισούται με μια από τις παρατηρήσεις, ενώ όταν τα δεδομένα αποτελούνται από άρτιο αριθμό παρατηρήσεων, η διάμεσος βρίσκεται μεταξύ δύο παρατηρήσεων. Στην περίπτωση αυτή χωρίζουμε την απόσταση μεταξύ των δύο αυτών παρατηρήσεων παίρνοντας τη μέση τιμή των δύο αυτών τιμών.

Παράδειγμα 2.8

Υποθέστε πως έχουμε τους ακόλουθους αριθμούς

5 6 8 11 15 17 20

Καθώς ο αριθμός των παρατηρήσεων είναι περιττός (n=7), η διάμεσος θα είναι η παρατήρηση που βρίσκεται στη μέση σε αυτή την κατάταξη των δεδομένων. Η παρατήρηση που αντιστοιχεί στη διάμεσο είναι η 4η, αφού (n+1)/2= (7+1)/2= 4. Έτσι η τέταρτη παρατήρηση είναι η διάμεσος και στο συγκεκριμένο παράδειγμα η τιμή αυτή είναι το 11.

Υποθέστε τώρα ότι έχουμε την ίδια λίστα αριθμών με αυτή που είδαμε παραπάνω συν έναν ακόμα αριθμό, έστω το 25. Τώρα ο αριθμός των παρατηρήσεων είναι άρτιος (n=8) και θα πρέπει να προσδιορίσουμε το μέσο σημείο μεταξύ δύο αριθμών. Η διάμεσος θα είναι ο μέσος όρος δύο τιμών που βρίσκονται από τον τύπο (n+1)/2. Στο συγκεκριμένο παράδειγμα (n+1)/2=(8+1)/2= 4,5. Άρα θα πρέπει να βρούμε την 4η και την 5η παρατήρηση και να πάρουμε το μέσο όρο. Δηλαδή (11+15)/2=13, αφού η 4η παρατήρηση είναι ο αριθμός 11 και η 5η παρατήρηση ο αριθμός 15. Έτσι η διάμεσος στην περίπτωση αυτή είναι ο αριθμός 13.

Στα ομαδοποιημένα δεδομένα η διάμεσος δεν μπορεί να προσδιοριστεί με πλήρη ακρίβεια. Αρχικά πρέπει να βρούμε την τάξη που περιλαμβάνει τη διάμεσο και κατόπιν να υπολογίσουμε τη διάμεσο με τη χρήση του παρακάτω τύπου:

$$M = W_M \left[\frac{\frac{n+1}{2} - f_b}{f_M} \right] + L_M \qquad (2.7)$$

όπου M είναι η διάμεσος, W_M είναι το πλάτος της τάξης της διαμέσου, η οποία τάξη περιλαμβάνει την (n+1)/2 παρατήρηση υποθέτοντας ότι οι τιμές είναι διαταγμένες, f_M είναι η συχνότητα της τάξης της διαμέσου, f_b είναι η αθροιστική συχνότητα των τάξεων

που βρίσκονται πριν από την τάξη της διαμέσου και L_M είναι το κατώτερο όριο της τάξης που περιλαμβάνει τη διάμεσο.

Παράδειγμα 2.9

Στο παράδειγμα 2.6 υπολογίσαμε το μέσο των ομαδοποιημένων δεδομένων. Για να βρούμε τη διάμεσο, αρχικά βρίσκουμε την τάξη που περιλαμβάνει τη διάμεσο. Αυτή είναι η τάξη που περιλαμβάνει την 40,5η παρατήρηση, αφού (80+1)/2=40,5. Η διάμεσος είναι μια τιμή μεταξύ της 40ής και 41ης παρατήρησης. Κοιτώντας τον πίνακα του παραδείγματος 2.6 βλέπουμε ότι η τάξη που περιλαμβάνει την 41η παρατήρηση είναι η τέταρτη τάξη, όπου η αθροιστική συχνότητα είναι 65 (=4+8+18+35). Οι τρεις πρώτες τάξεις περιλαμβάνουν αθροιστικά τις πρώτες 30 παρατηρήσεις. Καθώς ψάχνουμε την 40ή και την 41η παρατήρηση αυτές βρίσκονται στην τέταρτη τάξη, όπου έχουμε την αθροιστική συχνότητα από το 31 μέχρι το 65. Έπειτα, αντικαθιστώντας στην έκφραση υπολογισμού της διαμέσου έχουμε:

$$M = W_M\left[\frac{\frac{n+1}{2} - f_b}{f_M}\right] + L_M = 10[(40{,}5-30)/35]+100 \cong 103$$

όπου $W_M = 10$ (το πλάτος της τάξης της διαμέσου), L_M=100 (το κατώτερο όριο της τάξης της διαμέσου), f_M =35 (η συχνότητα της τάξης της διαμέσου) και f_b =30 (το σύνολο των συχνοτήτων των τάξεων πριν την τάξη της διαμέσου, δηλαδή 4+8+18).

Η **επικρατούσα τιμή** (mode) είναι η παρατήρηση που εμφανίζεται πιο συχνά στα δεδομένα. Αν έχουμε ένα σύνολο μετρήσεων, όπως τους αριθμούς 3, 4, 6, 6, 7 και 9 τότε ο αριθμός 6 είναι η επικρατούσα τιμή, καθώς εμφανίζεται δύο φορές σε σύγκριση με τους υπόλοιπους αριθμούς που εμφανίζονται μόνο μια φορά. Αντίθετα οι αριθμοί του παραδείγματος 2.4 δεν παρουσιάζουν επικρατούσα τιμή. Αξίζει να αναφερθεί ότι υπάρχουν περιπτώσεις που μπορεί να έχουμε περισσότερες από μία επικρατούσες τιμές. Οι αριθμοί 1, 3, 3, 4, 5, 5, 7 και 8 έχουν δυο επικρατούσες τιμές, τους αριθμούς 3 και 5 (bimodal).

Αν τα δεδομένα είναι ομαδοποιημένα, η επικρατούσα τάξη είναι αυτή που έχει τις περισσότερες παρατηρήσεις. Για ομαδοποιημένα δεδομένα έχουμε:

$$M_0 \cong W_m\left[\frac{f_m - f_{m-1}}{(f_m - f_{m-1}) + (f_m - f_{m+1})}\right] + L_m \qquad (2.8)$$

όπου f_m είναι η συχνότητα της επικρατούσας τάξης, f_{m-1} είναι η συχνότητα της τάξης πριν την επικρατούσα τάξη και f_{m+1} είναι η συχνότητα της τάξης που βρίσκεται μετά την επικρατούσα τάξη. Η επικρατούσα τιμή συμβολίζεται με M_0 (Mode).

Παράδειγμα 2.10

Ας υπολογίσουμε την επικρατούσα τιμή των δεδομένων του παραδείγματος 2.6. Η επικρατούσα τάξη είναι η τάξη με την υψηλότερη συχνότητα και αυτή είναι η τάξη 100-110 με συχνότητα 35. Οπότε, αντικαθιστώντας στην προηγούμενη σχέση έχουμε:

$$M_0 \cong W_m\left[\frac{f_m - f_{m-1}}{(f_m - f_{m-1}) + (f_m - f_{m+1})}\right] + L_m = 10\left[\frac{35-18}{(35-18)+(35-15)}\right] + 100 = 104{,}32$$

Τέλος ας αναφερθούμε και στα **εκατοστημόρια** (percentiles) και τα **τεταρτημόρια** (quartiles). Το P-οστό εκατοστημόριο των καταταγμένων δεδομένων ανάλογα με το μέγεθος είναι η τιμή που έχει το P% των παρατηρήσεων πριν από αυτήν και το (100-P)% των παρατηρήσεων μετά από αυτήν. Η θέση του P-οστού εκατοστημορίου είναι [(n+1)P%/100]. Πιο συχνά χρησιμοποιούμε τα τεταρτημόρια και συγκεκριμένα το πρώτο τεταρτημόριο (25ο εκατοστημόριο), το δεύτερο τεταρτημόριο (50ό εκατοστημόριο) και το τρίτο τεταρτημόριο (75ο εκατοστημόριο). Το πρώτο τεταρτημόριο είναι η τιμή της μεταβλητής που έχει το 25% του συνόλου των παρατηρήσεων πριν από αυτήν και το 75% των παρατηρήσεων μετά από αυτήν. Ομοίως το τρίτο τεταρτημόριο είναι η τιμή της μεταβλητής που έχει το 75% των παρατηρήσεων πριν από αυτή και το υπόλοιπο 25% μετά από αυτήν. Το δεύτερο τεταρτημόριο βρίσκεται στο κέντρο των δεδομένων και είναι η διάμεσος, αφού έχει το 50% των παρατηρήσεων από τη μια πλευρά του και το υπόλοιπο 50% των παρατηρήσεων από την άλλη πλευρά του.

Οι μαθηματικές σχέσεις με τις οποίες υπολογίζουμε τα τεταρτημόρια είναι οι εξής:

Πρώτο τεταρτημόριο (first quartile) $$Q_1 = \frac{n+1}{4} \qquad (2.9)$$

Δεύτερο τεταρτημόριο (second quartile, median) $$Q_2 = \frac{2(n+1)}{4} = \frac{n+1}{2} \qquad (2.10)$$

Τρίτο τεταρτημόριο (third quartile) $$Q_3 = \frac{3(n+1)}{4} \qquad (2.11)$$

Ενδοτεταρτημοριακό εύρος[4] (interquartile range) $$IQR = Q_3 - Q_1 \qquad (2.12)$$

[4] Το ενδοτεταρτημοριακό εύρος θεωρείται μέτρο διασποράς. Σημειώνεται εδώ, καθώς γίνεται αναφορά και γνωριμία με την έννοια των τεταρτημορίων.

Για ομαδοποιημένα δεδομένα οι τύποι υπολογισμού είναι πιο πολύπλοκοι, όπως φαίνεται παρακάτω:

$$Q_1 = W_{Q1}\left[\frac{\frac{n}{4} - f_b}{f_{Q1}}\right] + L_{Q1} \tag{2.13}$$

$$Q_3 = W_{Q3}\left[\frac{\frac{3n}{4} - f_b}{f_{Q3}}\right] + L_{Q3} \tag{2.14}$$

όπου W_{Q1} και W_{Q3} είναι το πλάτος των τάξεων που περιλαμβάνουν το πρώτο και το τρίτο τεταρτημόριο αντίστοιχα και υπολογίζεται χρησιμοποιώντας την αντίστοιχη έκφραση σε κάθε περίπτωση. Για παράδειγμα, η τάξη που περιλαμβάνει το πρώτο τεταρτημόριο είναι αυτή που έχει την [(n+1)/4] ή πιο απλά την (n/4) παρατήρηση, ενώ η τάξη που περιλαμβάνει το τρίτο τεταρτημόριο είναι αυτή που περιλαμβάνει την [3(n+1)/4] ή (3n/4) παρατήρηση υποθέτοντας ότι οι παρατηρήσεις είναι διαταγμένες. Τα f_{Q1} και f_{Q3} είναι οι συχνότητες των τάξεων που περιλαμβάνουν το πρώτο και το τρίτο τεταρτημόριο αντίστοιχα, f_b είναι η αθροιστική συχνότητα των τάξεων που βρίσκονται πριν από την τάξη του πρώτου ή του τρίτου τεταρτημορίου. Τέλος, L_{Q1} και L_{Q3} είναι τα κατώτερα όρια της αντίστοιχης τάξης που περιλαμβάνει το πρώτο και το τρίτο τεταρτημόριο αντίστοιχα.

Παράδειγμα 2.11

Βρείτε το μέσο, τη διάμεσο (Q_2), την επικρατούσα τιμή, το πρώτο και το τρίτο τεταρτημόριο (Q_1 και Q_3) και το ενδοτεταρτημοριακό εύρος (IQR) για τους ακόλουθους αριθμούς:

5, 6, 6, 7, 8, 9, 10, 12, 13, 16, 18

Ο μέσος θα είναι:

$$\overline{X} = \frac{\sum_{i=1}^{11} X_i}{n} = \frac{5+6+6+7+8+9+10+12+13+16+18}{11} = \frac{110}{11} = 10$$

Η διάμεσος θα είναι ο αριθμός που αντιστοιχεί στην [(n+1)/2] παρατήρηση. Αυτή είναι η (11+1)/2=6η παρατήρηση. Έτσι στην παραπάνω σειρά των δεδομένων η διάμεσος είναι ο αριθμός 9, δηλαδή η 6η παρατήρηση. Η επικρατούσα τιμή είναι ο αριθμός 6.

Το πρώτο τεταρτημόριο είναι η [(n+1)/4] παρατήρηση. Αυτή είναι η (11+1)/4=3η παρατήρηση. Έτσι το πρώτο τεταρτημόριο (Q_1) είναι η τρίτη παρατήρηση στη σειρά των δεδομένων.[5] Αυτή είναι ο αριθμός 6. Ομοίως, το τρίτο τεταρτημόριο (Q_3) είναι η [3(n+1)/4] παρατήρηση. Αυτή είναι η [3(11+1)/4]=9, οπότε το τρίτο τεταρτημόριο είναι η 9η παρατήρηση στη σειρά των δεδομένων μας. Αυτή είναι ο αριθμός 13. Τέλος, το ενδοτεταρτημοριακό εύρος (IQR) είναι η διαφορά μεταξύ του τρίτου και του πρώτου τεταρτημορίου. Δηλαδή:

$$IQR = Q_3 - Q_1 = 13-6 = 7$$

Σύγκριση των μέτρων θέσης και κεντρικής τάσης

Συγκρίνοντας τα στατιστικά μέτρα που έχουμε δει μέχρι τώρα μπορούμε να πούμε ότι ο μέσος είναι πολύ εύχρηστος, χρησιμοποιεί όλα τα δεδομένα (όλη την πληροφορία που μας παρέχουν τα δεδομένα), όμως είναι ευαίσθητος στις ακραίες τιμές και δεν μπορεί να υπολογιστεί με ακρίβεια για ομαδοποιημένα δεδομένα. Η διάμεσος έχει το πλεονέκτημα να μην επηρεάζεται από ακραίες τιμές, αλλά δεν έχει αριθμητικές ιδιότητες, με την έννοια ότι δεν μπορούμε να προσθέσουμε ή να αφαιρέσουμε διαμέσους διαφορετικών δεδομένων. Η επικρατούσα τιμή δε χρησιμοποιεί πλήρως την πληροφορία που μας παρέχουν όλα τα δεδομένα, δεν επηρεάζεται από τις ακραίες τιμές των παρατηρήσεων και είναι πιθανό να μην υπάρχει για κάποια δεδομένα. Θεωρούμε ότι ο μέσος είναι η καλύτερη από αυτές τις μετρήσεις για να τον χρησιμοποιήσουμε όταν εξετάζουμε κάποιο δείγμα για να γενικεύσουμε για τον πληθυσμό.[6] Μερικές φορές μπορούμε να χρησιμοποιήσουμε τον τετριμμένο μέσο βγάζοντας έξω τις ακραίες τιμές όταν υπολογίζουμε το μέσο. Μια παρόμοια σύγκριση μπορεί να γίνει όταν χρησιμοποιούμε καμπύλες συχνότητας, όπως θα δούμε παρακάτω στην παράγραφο 2.4.

---

[5] Τα τεταρτημόρια μπορούν να βρεθούν γραφικά από τη γραμμή των αθροιστικών σχετικών συχνοτήτων, όπου στον κάθετο άξονα έχουμε τις αθροιστικές σχετικές συχνότητες και στον οριζόντιο άξονα τις τάξεις. Τότε στον κάθετο βρίσκουμε τις σχετικές αθροιστικές συχνότητες 0,25, 0,5 και 0,75. Από τα σημεία αυτά φέρνουμε παράλληλες ευθείες προς τον οριζόντιο άξονα και όπου συναντήσουν τη γραμμή της αθροιστικής συχνότητας φέρνουμε κάθετες στον οριζόντιο άξονα. Τα σημεία επαφής των κάθετων ευθειών με τον οριζόντιο άξονα μάς δίνουν το Q_1, τη διάμεσο (Q_2) και το Q_3.

[6] Σε περίπτωση που χρησιμοποιούμε δεδομένα κατηγορικής κλίμακας, δεν υπάρχει νόημα στον υπολογισμό περιγραφικών στατιστικών. Το ίδιο συμβαίνει και με τα δεδομένα τακτικής κλίμακας. Δεν μας λένε πόση είναι η διαφορά ανάμεσα σε αυτό που κατατάχθηκε πρώτο και σε αυτό που κατατάχθηκε δεύτερο αλλά απλώς μας πληροφορούν για την κατάταξή τους.

2.3.2 Μέτρα διασποράς και μεταβλητότητας των δεδομένων (measures of dispersion)

Η περιγραφή της μεταβλητότητας ορισμένων τιμών από μια μέση τιμή ονομάζεται *διασπορά* αυτών των τιμών. Το πιο απλό μέτρο μέτρησης της μεταβλητότητας που χρησιμοποιείται συχνά είναι το *εύρος* (*range*), δηλαδή η διαφορά μεταξύ της μεγαλύτερης και της μικρότερης τιμής των δεδομένων. Ένα από τα προβλήματα που προκύπτει όταν χρησιμοποιούμε το εύρος ως μέτρο της διασποράς, είναι ότι επηρεάζεται από τις μεταβολές των ακραίων τιμών. Ένας τρόπος για να ξεπεράσουμε το πρόβλημα αυτό είναι να χρησιμοποιήσουμε το *ημιενδοτεταρτημοριακό εύρος* (semi-interquartile range), το οποίο δείχνει τη μέση διαφορά μεταξύ του τρίτου και του πρώτου τεταρτημορίου. Το μειονέκτημα στη χρήση του είναι ότι δεν χρησιμοποιεί όλη την πληροφορία που παρέχεται από τα δεδομένα.

Οι μετρήσεις που χρησιμοποιούμε πιο συχνά για τη μέτρηση της διασποράς είναι η *διακύμανση* (variance) και η *τυπική απόκλιση* (standard deviation).[7] Αυτές οι μετρήσεις δείχνουν το κατά πόσο κοντά μια κατανομή βρίσκεται γύρω από το μέσο ή κατά πόσο απέχει από το μέσο. Οι αποκλίσεις του μέσου μετράνε την απόσταση (τη διαφορά) τής κάθε τιμής των δεδομένων από το συνολικό μέσο. Η τυπική απόκλιση είναι η αριθμητική μέτρηση του μέσου όρου της απόκλισης των δεδομένων γύρω από το μέσο.

Αν μία κατανομή είναι στενά συγκεντρωμένη γύρω από το μέσο, οι αποκλίσεις θα είναι μικρές. Θα πρέπει λοιπόν να υπολογίσουμε όλες τις αποκλίσεις και να βρούμε το μέσο όρο αυτών των αποκλίσεων. Αλλά, καθώς το άθροισμα των διαφορών κάθε τιμής από το μέσο θα είναι μηδέν,[8] χρησιμοποιούμε το τετράγωνο της απόκλισης αυτής. Αυτό είναι:

Για πληθυσμό

$$\sigma = \sqrt{\frac{\sum_{i=1}^{\nu}(X_i - \mu)^2}{\nu}} \qquad (2.15)$$

[7] Σε περίπτωση που μας ενδιαφέρει ο συσχετισμός δύο μεταβλητών, τότε πέρα από το συντελεστή συσχέτισης που θα δούμε αργότερα, υπάρχει και η *συνδιακύμανση* (covariance) η οποία υπολογίζεται με τους τύπους:

$$S_{XY} = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{n-1} \text{ και } \sigma_{XY} = \frac{\sum(X_i - \mu_X)(Y_i - \mu_Y)}{\nu}$$

όπου ο πρώτος τύπος χρησιμοποιείται σε περίπτωση δειγμάτων και ο δεύτερος σε περίπτωση πληθυσμών. Μια μεγάλη θετική (αρνητική) τιμή της συνδιακύμανσης δείχνει μία δυνατή θετική (αρνητική) γραμμική σχέση. Το πρόβλημα του μέτρου αυτού βρίσκεται στις μονάδες μέτρησης των μεταβλητών. Αυτό το πρόβλημα αποφεύγεται με τη χρήση του συντελεστή συσχέτισης.

[8] Με απλά λόγια μπορούμε να βρούμε τη μέση απόκλιση από τον τύπο $\frac{\sum(X_i - \mu)}{n}$ για απλά δεδομένα ή από τον τύπο $\frac{\sum(X_i - \mu)f_i}{\sum f_i}$ για ομαδοποιημένα δεδομένα.

Για δείγμα $$s = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}} \qquad (2.16)$$

Παρατηρούμε ότι στην περίπτωση που έχουμε δείγμα, η τυπική απόκλιση διαιρείται με (n-1) και όχι με (n). Αυτό οφείλεται στο γεγονός ότι χρησιμοποιούμε το μέσο του δείγματος και όχι το μέσο του πληθυσμού στους τύπους του υπολογισμού της δειγματικής τυπικής απόκλισης. Καθώς δεν γνωρίζουμε τον πληθυσμιακό μέσο (μ) χρειάζεται να τον αντικαταστήσουμε με την όσο το δυνατόν καλύτερη προσέγγισή του, που στην περίπτωσή μας είναι ο δειγματικός μέσος ($\bar{X}$). Σίγουρα όμως η απόκλιση με τη χρήση ενός δείγματος δεν αναμένεται να είναι η ίδια με αυτή του πληθυσμού, αλλά μεγαλύτερη. Αυτό θα γίνει σαφές στο κεφάλαιο της Επαγωγικής Στατιστικής.[9] Γι' αυτό και διαιρούμε με (n-1) και όχι με το (n) για να λάβουμε υπόψη μας τη χρήση του δειγματικού μέσου και όχι του άγνωστου πληθυσμιακού μέσου. Είναι αξιοσημείωτο ότι κατά μέσο όρο η δειγματική τυπική απόκλιση (s) δεν θα είναι χαμηλότερη ή υψηλότερη από την πραγματική πληθυσμιακή (σ), καθώς η δειγματοληψία επαναλαμβάνεται.

Στην περίπτωση των ομαδοποιημένων δεδομένων η τυπική απόκλιση δίνεται από τη σχέση:

$$s = \sqrt{\frac{\sum_{i=1}^{k} f_i(X_{mi} - \bar{X})^2}{(\sum f_i) - 1}} \qquad (2.17)$$

Για υπολογιστικούς λόγους μπορούμε να χρησιμοποιήσουμε και την ακόλουθη σχέση:

$$s = \sqrt{\frac{\sum_{i=1}^{k} f_i X_{mi}^2 - n\bar{X}^2}{(\sum f_i) - 1}} \qquad (2.18)$$

όπου f_i είναι η συχνότητα της i τάξης και X_{mi} είναι το μέσο σημείο της i τάξης.

Αντίστοιχα, η διακύμανση των δεδομένων είναι το τετράγωνο της τυπικής απόκλισης, οπότε και έχουμε:

Για πληθυσμό $$\sigma^2 = \frac{\sum_{i=1}^{\nu}(X_i - \mu)^2}{\nu} \qquad (2.19)$$

[9] Εκεί θα συζητηθούν και οι βαθμοί ελευθερίας που αποτελούν περαιτέρω εξήγηση για τη χρήση του n-1 και όχι του n.

Για δείγμα
$$s^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1} \tag{2.20}$$

Για ομαδοποιημένα δεδομένα έχουμε:

Για πληθυσμό
$$\sigma^2 = \frac{\sum_{i=1}^{k} f_i(X_{mi} - \mu)^2}{\sum f_i} \tag{2.21}$$

Για δείγμα
$$s^2 = \frac{\sum_{i=1}^{k} f_i(X_{mj} - \bar{X})^2}{(\sum f_i) - 1} \tag{2.22}$$

Για λόγους απλότητας μπορούμε να χρησιμοποιήσουμε και την ακόλουθη σχέση για να υπολογίσουμε τη διακύμανση του δείγματος.

$$s^2 = \frac{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2}{n-1} \tag{2.23}$$

Παράδειγμα 2.12

Χρησιμοποιώντας τα δεδομένα του παραδείγματος 2.11 υπολογίστε τη διακύμανση και την τυπική απόκλιση των αριθμών:

5 6 6 7 8 9 10 12 13 16 18

Φτιάχνουμε έναν πίνακα που θα παρουσιάζει τα βήματα που κάνουμε για τον υπολογισμό της διακύμανσης και κατ' επέκταση της τυπικής απόκλισης. Ήδη γνωρίζουμε ότι ο μέσος αυτών των αριθμών είναι 10.

Αριθμοί	$(X_I - \bar{X})$	$(X_I - \bar{X})^2$	X_I^2
5	5-10=-5	25	25
6	6-10=-4	16	36
6	6-10=-4	16	36
7	7-10=-3	9	49
8	8-10=-2	4	64
9	9-10=-1	1	81
10	10-10=0	0	100
12	12-10=2	4	144
13	13-10=3	9	169
16	16-10=6	36	256
18	18-10=8	64	324
		$\Sigma = 184$	$\Sigma = 1.284$

Η δεύτερη και η τρίτη στήλη του πίνακα αυτού δείχνει τους όρους που χρησιμοποιούμε για τον υπολογισμό της διακύμανσης σύμφωνα με τον τύπο (2.20). Δηλαδή:

$$s^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1} = \frac{184}{11-1} = 18{,}4$$

Η τελευταία στήλη μπορεί να χρησιμοποιηθεί για τον υπολογισμό της διακύμανσης σύμφωνα με την έκφραση (2.23). Έχουμε:

$$s^2 = \frac{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2}{n-1} = \frac{1.284 - 11(10)^2}{11-1} = \frac{184}{10} = 18{,}4$$

Μπορούμε να πούμε ότι οι υπολογισμοί με βάση τη δεύτερη σχέση είναι ευκολότεροι. Η τυπική απόκλιση, αφού έχουμε υπολογίσει τη διακύμανση, είναι ένας απλός αριθμητικός υπολογισμός. Αυτό που κάνουμε είναι να πάρουμε απλώς την τετραγωνική ρίζα της διακύμανσης και αυτή είναι η τυπική απόκλιση. Στη συγκεκριμένη περίπτωση, η τυπική απόκλιση ισούται με $\sqrt{18{,}4} = 4{,}29$. Αυτό σημαίνει ότι οι τιμές της μεταβλητής αποκλίνουν από τη μέση τιμή κατά 4,29 κατά μέσο όρο.

Παράδειγμα 2.13

Υποθέστε πως έχουμε τα παρακάτω ομαδοποιημένα δεδομένα.

Τάξεις	Συχνότητες
0-19	5
20-39	7
40-59	4
60-79	2
80-99	2
	Σ = 20

Για να βρούμε τη διακύμανση και την τυπική απόκλιση, δημιουργούμε τον πίνακα που ακολουθεί και δείχνει όλους τους υπολογισμούς μας.

Τάξεις	f_j	X_{mj}	f_jX_{mj}	$X_j-\mu$	$(X_j-\mu)^2$	$f_j(X_j-\mu)^2$
0-19	5	9,5	47,5	-29	841	4.205
20-39	7	29,5	206,5	-9	81	567
40-59	4	49,5	198	11	121	484
60-79	2	69,5	139	31	961	1.922
80-99	2	89,5	179	51	2.601	5.202
	Σ=20		Σ = 770			Σ = 12.380

Υποθέτουμε πως οι αριθμοί του πίνακα αυτού αναφέρονται στον πληθυσμό που μας ενδιαφέρει. Πρώτα υπολογίζουμε το μέσο:

$$\mu = \frac{\sum_{j=1}^{k} f_j X_{mj}}{\sum_j f_j} = \frac{770}{20} = 38,5.$$

Κατόπιν βλέπουμε ότι η διακύμανση ισούται με 619 σύμφωνα με τη σχέση:

$$\sigma^2 = \frac{\sum_{j=1} f_j \left(X_{mj} - \mu\right)^2}{\sum f_j} = \frac{12.380}{20} = 619$$

και η τυπική απόκλιση θα είναι η τετραγωνική ρίζα αυτού του αριθμού, δηλαδή $\sigma = \sqrt{619} = 24{,}88$.

Ιδιότητες της διακύμανσης

Ορισμένες από τις ιδιότητες της διακύμανσης είναι οι ακόλουθες:

1. Η διακύμανση μιας σταθεράς είναι μηδέν. Δηλαδή αν

$$x_1 = x_2 = x_3 = \ldots = x_v = \alpha = \text{σταθερό} \qquad \text{τότε } \sigma^2 = 0.$$

2. Αν όλες οι τιμές της μεταβλητής X πολλαπλασιαστούν με τον ίδιο αριθμό, α, τότε η διακύμανση πολλαπλασιάζεται με το τετράγωνο αυτού του αριθμού.

$$\text{Var}(\alpha x) = \alpha^2 \text{Var}(x)$$

3. Αν όλες οι τιμές της μεταβλητής X μεταβληθούν (αυξηθούν ή μειωθούν) κατά μια σταθερά α, τότε η διακύμανση δεν μεταβάλλεται. Δηλαδή:

$$\text{Var}(x \pm \alpha) = \text{Var}(x)$$

Συντελεστής Μεταβλητότητας

Ένα άλλο χρήσιμο μέτρο της μεταβλητότητας είναι ο *συντελεστής μεταβλητότητας* (Coefficient of Variation), ο οποίος χρησιμοποιείται για να συγκρίνουμε σχετικές αποκλίσεις μεταξύ των πληθυσμών. Ισούται με την τυπική απόκλιση της κατανομής διαιρεμένη με το μέσο της κατανομής. Ο συντελεστής αυτός είναι ένα αδιάστατο μέγεθος και δεν είναι απαραίτητο να μετατραπεί σε κοινές μονάδες για να συγκρίνουμε τη *μεταβλητότητα* δυο κατανομών.

Ένας περιορισμός παρουσιάζεται στη χρήση του συντελεστή. Όταν οι κατανομές που συγκρίνονται έχουν αρνητικές παρατηρήσεις, ο συντελεστής *μεταβλητότητας* παρέχει έναν μη αξιόπιστο τρόπο σύγκρισης της μεταβλητότητας μεταξύ των δεδομένων. Με αρνητικές τιμές στα δεδομένα, οι μέσοι των δεδομένων μπορεί να είναι μηδέν ή αρνητικοί και η ερμηνεία του συντελεστή ως σχετική μεταβλητότητα χάνεται.

Η μεταβλητότητα για έναν πληθυσμό είναι: $\omega = (\sigma/\mu)*100$ (2.24)

Η μεταβλητότητα για ένα δείγμα είναι: $w = (s/\overline{X})*100$ (2.25)

Χαμηλές τιμές του ω (ή του w) φανερώνουν ομοιογένεια, δηλαδή μικρή μεταβλητότητα των δεδομένων. Ο συντελεστής της μεταβλητότητας χάνει τη σημασία του όταν ο πληθυσμός ή το μέγεθος του δείγματος είναι πολύ μικρό, κοντά στο μηδέν.

Παράδειγμα 2.14

Θεωρήστε το ακόλουθο παράδειγμα, στο οποίο έχουμε δύο μεταβλητές (X και Y) και θέλουμε να υπολογίσουμε ποια από τις δύο εμφανίζει τη μικρότερη μεταβλητότητα.

X_i: 12 18 35 44 61
Y_i: 24 59 71 82 99

Η σχετική μεταβλητότητα μπορεί να υπολογιστεί χρησιμοποιώντας το συντελεστή μεταβλητότητας. Υποθέστε πως ασχολούμαστε με τον πληθυσμό. Τότε έχουμε:

X_i	Y_I	$X_i - \mu_X$	$Y_i - \mu_Y$	$(X_i - \mu_X)^2$	$(Y_i - \mu_Y)^2$
12	24	12-34= -22	24-67= -43	484	1.849
18	59	18-34= -16	59-67= -8	256	64
35	71	35-34= 1	71-67= 4	1	16
44	82	44-34=10	82-67= 15	100	225
61	99	61-34=27	99-67= 32	729	1.024
Σ = 170	Σ = 335			Σ = 1.570	Σ = 3.178

Οι μέσοι αριθμητικοί των δεδομένων αυτών είναι οι εξής:

$\mu_X=(170/5)=34$ και $\mu_Y=(335/5)=67$.

Ακολουθώντας τους υπολογισμούς που φαίνονται στον πίνακα βρίσκουμε τη διακύμανση. Αν χρησιμοποιήσουμε την έκφραση της διακύμανσης του πληθυσμού, υποθέτοντας πως έχουμε όλα τα στοιχεία που μας ενδιαφέρουν, τότε έχουμε:

$$\sigma_X^2 = \frac{\sum_{i=1}^{v}(X_i - \mu_X)^2}{v} = \frac{1.570}{5} = 314$$

$$\sigma_Y^2 = \frac{\sum_{i=1}^{v}(Y_i - \mu_Y)^2}{v} = \frac{3.178}{5} = 635{,}6$$

Η τυπική απόκλιση θα είναι $\sigma_X = \sqrt{314} = 17,72$ και $\sigma_Y = \sqrt{635,6} = 25,21$ και ο συντελεστής μεταβλητότητας για την πρώτη ομάδα των δεδομένων θα είναι:

$$\omega_x = \frac{\sigma_x}{\mu_x} = \frac{17,72}{34} \times 100 = 52,1\%$$

και για τη δεύτερη ομάδα των δεδομένων:

$$\omega_Y = \frac{\sigma_Y}{\mu_Y} = \frac{25,21}{67} \times 100 = 37,63\%$$

Αυτό σημαίνει ότι η μεταβλητή Y_i παρουσιάζει μικρότερη μεταβλητότητα.

Σύγκριση των μέτρων διασποράς

Τόσο το εύρος όσο και το ενδοτεταρτημοριακό εύρος είναι εύκολο να υπολογιστούν, αλλά δίνουν ένα περιορισμένο μέτρο της διασποράς καθώς δεν λαμβάνουν υπόψη το σύνολο των δεδομένων. Ωστόσο περιορίζουν την ικανότητα να απεικονίσουν την πραγματική διακύμανση όλων των τιμών. Αυτό το μειονέκτημα είναι πιο σοβαρό για το εύρος, καθώς βασίζεται στις μεγαλύτερες και στις μικρότερες τιμές.

Η διακύμανση είναι ένα άλλο πολύ χρήσιμο μέτρο διασποράς, το οποίο δείχνει πόσο συγκεντρωμένα είναι τα δεδομένα γύρω από το μέσο. Αλλά λόγω του γεγονότος ότι ο εκτιμημένος αριθμός για τη διακύμανση μπορεί να μην είναι πάντα της ίδιας τάξης μεγέθους, όπως προκύπτει από το σύνολο των δεδομένων, μπορούμε να πούμε ότι η διακύμανση δεν είναι ιδιαίτερα χρήσιμη στην περιγραφή της διασποράς των δεδομένων. Αξίζει να σημειωθεί ότι η διακύμανση χρησιμοποιείται σε πολλές προχωρημένες στατιστικές τεχνικές, όπως στην Ανάλυση Διακύμανσης (ANOVA) κ.ά. Η τυπική απόκλιση είναι προφανώς το καλύτερο στατιστικό μέγεθος για την περιγραφή της διασποράς των δεδομένων. Λαμβάνει υπόψη όλες τις διαθέσιμες πληροφορίες και παράγει έναν υπολογισμό για τη μέση απόκλιση από το μέσο. Προτιμάται γιατί είναι εύκολο να ερμηνευτεί, καθώς, σε αντίθεση με τη διακύμανση, οι μονάδες της τυπικής απόκλισης είναι της ίδιας τάξης μεγέθους με αυτές των αρχικών μας δεδομένων.

Εμπειρική Εφαρμογή

Ανάλυση Κινδύνου (Risk Analysis)

Αν δεν γνωρίζουμε την πιθανή έκβαση κάποιας ενέργειας (π.χ. επενδυτικής), τότε αντιμετωπίζουμε την κατάσταση με συνθήκες κινδύνου ή αβεβαιότητας. Ο **κίνδυνος (risk)** αναφέρεται στην περίπτωση κατά την οποία υπάρχουν περισσότερες από μία εκβάσεις

(outputs) σε μια απόφαση (decision) και η *πιθανότητα* κάθε έκβασης είναι γνωστή ή μπορεί να υπολογιστεί. Αν οι πιθανότητες αυτές δεν μπορούν να υπολογιστούν τότε μιλάμε για **αβεβαιότητα** (**uncertainty**). Σε περίπτωση σύγκρισης επενδυτικών σχεδίων με κίνδυνο μπορούμε να χρησιμοποιήσουμε τα αναμενόμενα κέρδη (Expected Profits), την τυπική απόκλιση (Standard deviation) και το συντελεστή μεταβλητότητας (Coefficient of variation). Τα αναμενόμενα κέρδη θα δίνονται ως:

$$E(\Pi) - \bar{\Pi} = \sum_{i=1}^{n} \Pi_i P_i$$

Όπου Π_i το επίπεδο κέρδους σε σχέση με κάποιο αποτέλεσμα i και P_i η πιθανότητα εμφάνισης του αποτελέσματος. Για επενδυτικά σχέδια με τον ίδιο κίνδυνο η επιχείρηση θα επιλέξει αυτό με τα μεγαλύτερα κέρδη.

Ο απόλυτος κίνδυνος μια επένδυσης μετράται με την τυπική απόκλιση (Standard Deviation) των πιθανών κερδών από το επενδυτικό σχέδιο και ορίζεται ως:

$$\sigma_i = \sqrt{\sum_{i=1}^{n} \left(\Pi_i - \bar{\Pi}\right)^2 P_i}$$

Όσο μεγαλύτερη η πιθανή απόκλιση των κερδών για ένα επενδυτικό σχέδιο τόσο μεγαλύτερη η τυπική απόκλιση, άρα και ο κίνδυνος. Για σύγκριση σχετικών αποκλίσεων των πιθανών κερδών δύο ή περισσότερων επενδυτικών σχεδίων (projects) χρησιμοποιούμε το *συντελεστή μεταβλητότητας* (*coefficient of variation*, w) που ορίζεται ως:

$$w_i = \frac{\sigma_i}{\mathrm{E}(\Pi_i)}$$

Ένας επενδύτης συνήθως προτιμά ένα πιο επικίνδυνο επενδυτικό σχέδιο αν τα αναμενόμενα κέρδη είναι επαρκώς υψηλότερα από αυτά ενός λιγότερου επικίνδυνου σχεδίου (project). Ως παράδειγμα, ας θεωρήσουμε την κατάσταση μιας οικονομίας με προβλεπόμενες πιθανότητες εμφάνισης άνθισης, κανονικής ανάπτυξης και ύφεσης όπως δίνονται στον παρακάτω πίνακα και το δίλημμα επιλογής ενός από τα δύο πιθανά επενδυτικά σχέδια. Τα κέρδη από κάθε επενδυτικό σχέδιο δίνονται στην 3η στήλη, ενώ οι αναμενόμενες τιμές (από τον πολλαπλασιασμό των πιθανοτήτων επί τα κέρδη) εμφανίζονται στην τελευταία στήλη.

Κατάσταση της Οικονομίας	Πιθανότητα Κατάστασης	Κέρδη	Αναμενόμενη Τιμή
Επενδυτικό σχέδιο Α			
Άνθιση	0,30	1.000	300
Κανονική Ανάπτυξη	0,50	800	400
Ύφεση	0,20	500	100
Αναμενόμενο κέρδος επενδυτικού σχεδίου Α = €800			
Επενδυτικό σχέδιο Β			
Άνθιση	0,30	900	270
Κανονική Ανάπτυξη	0,50	900	450
Ύφεση	0,20	400	80
Αναμενόμενο Κέρδος επενδυτικού σχεδίου Β = €800			

Καθώς και τα δύο επενδυτικά σχέδια αποφέρουν το ίδιο αναμενόμενο κέρδος ύψους € 800 θα πρέπει να κάνουμε την επιλογή με γνώμονα την επενδυτική απόδοση και την επικινδυνότητα κάθε σχεδίου. Δηλαδή, να επιλέξουμε το επενδυτικό σχέδιο με την ίδια απόδοση αλλά με το χαμηλότερο κίνδυνο. Για το σκοπό αυτό μπορούμε να υπολογίσουμε αρχικά τις αποκλίσεις από τα αναμενόμενα μέσα κέρδη ανά περίπτωση προβλεπόμενης οικονομικής κατάστασης και με τη βοήθεια του τύπου εκτίμησης της τυπικής απόκλισης να υπολογίσουμε τον απόλυτο κίνδυνο των επενδυτικών σχεδίων. Έτσι έχουμε:

Απόκλιση $(\pi_{ij} - \bar{\pi})$	$(\pi_{ij} - \bar{\pi})^2$	Πιθανότητα (P_{ij})	$(\pi_{ij} - \bar{\pi})^2 (P_{ij})$
Σχέδιο Α			
€1.000-€800= €200	40.000	0,30	12.000
€800 - €800= €0	0	0,50	0
€500 - €800= -€300	90.000	0,20	18.000
Διακύμανση σ^2 = € 30.000 και τυπική απόκλιση $\sigma = \sqrt{\sigma^2} = \sqrt{30.000}$ = €173,205			
Σχέδιο Β			
€900-€800= €100	10.000	0,30	3.000
€900- €800= €100	10.000	0,50	5.000
€400- €800=-€400	160.000	0,20	32.000
Διακύμανση σ^2 = € 40.000 και τυπική απόκλιση $\sigma = \sqrt{\sigma^2} = \sqrt{40.000}$ =€ 200			

Άρα επιλέγεται το επενδυτικό σχέδιο Α το οποίο έχει την ίδια απόδοση με το επενδυτικό σχέδιο Β, αλλά χαμηλότερο κίνδυνο (173,205< 200).

Αν δυο επενδυτικά σχέδια έχουν διαφορετικά αναμενόμενα κέρδη π.χ. $\bar{\Pi}_A > \bar{\Pi}_B$ αλλά και διαφορετικό κίνδυνο π.χ. $\bar{\sigma}_A > \bar{\sigma}_B$ τότε χρησιμοποιώ το συντελεστή μεταβλητότητας ως:

$$w_A = \frac{\sigma_A}{\bar{\Pi}_A} \qquad \text{ή} \qquad w_B = \frac{\sigma_B}{\bar{\Pi}_B}$$

Στην περίπτωση αυτή επιλέγουμε το επενδυτικό σχέδιο με το χαμηλότερο συντελεστή μεταβλητότητας.

2.4 Ασυμμετρία και κύρτωση

Είδαμε ότι η γραφική παράσταση των μετρήσεων μιας μεταβλητής μας δείχνει τον τρόπο με τον οποίο κατανέμεται η μεταβλητή αυτή. Η κατανομή συχνοτήτων μπορεί να είναι συμμετρική ή ασύμμετρη. Συμμετρική είναι μια κατανομή, όταν οι τιμές της τοποθετούνται συμμετρικά γύρω από τη μέση αριθμητική τιμή (π.χ. η κανονική κατανομή είναι συμμετρική).

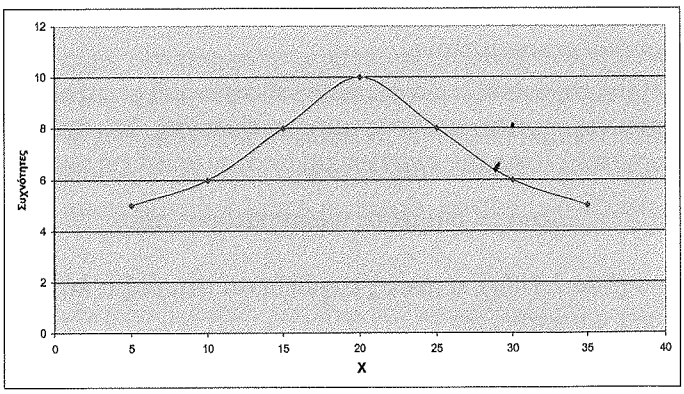
Παράδειγμα 2.15

Υποθέστε ότι έχουμε την παρακάτω κατανομή συχνοτήτων της μεταβλητής Χ:

X_i	5	10	15	20	25	30	35
f_i	5	6	8	10	8	6	5

Ο μέσος αριθμητικός των μετρήσεων αυτών είναι μ=20. Οι τιμές αυτές κατανέμονται συμμετρικά γύρω από το μέσο τους. Αν παραστήσουμε γραφικά τα δεδομένα αυτά θα σχηματιστεί μια συμμετρική καμπύλη γύρω από την τιμή μ=20.

Σχήμα 2.9: Κατανομή τιμών Παραδείγματος 2.15



Όποια κατανομή δεν είναι συμμετρική γύρω από το μέσο της παρουσιάζει ασυμμετρία. Μία κατανομή ασύμμετρη προς τα δεξιά έχει *θετική ασυμμετρία*. Η ασυμμετρία σύμφωνα με τον Pearson υπολογίζεται από τη διαφορά ανάμεσα στο μέσο και την επικρατούσα τιμή μιας κατανομής. Για να γίνει αυτό το μέγεθος αδιάστατο, διαιρούμε τη διαφορά αυτή με ένα μέτρο διασποράς, όπως είναι η τυπική απόκλιση.

Συντελεστής ασυμμετρίας (Pearson)

$$S_k = \frac{\bar{X} - M_0}{s} \tag{2.26}$$

όπου M_0 είναι η επικρατούσα τιμή, $\bar{X}$ ο δειγματικός αριθμητικός μέσος και s η δειγματική τυπική απόκλιση.

Από την άλλη πλευρά υπάρχει ο συντελεστής ασυμμετρίας του Bowley που κάνει χρήση των τεταρτημορίων. Συγκεκριμένα, ο συντελεστής αυτός ισούται με:

Συντελεστής ασυμμετρίας (Bowley)

$$S_k = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{(Q_3 - Q_2) + (Q_2 - Q_1)} \tag{2.27}$$

όπου το δεύτερο τεταρτημόριο είναι η διάμεσος. Η τιμή του συντελεστή ασυμμετρίας είναι μεταξύ ±1. Όταν ο συντελεστής είναι μηδέν, τότε η κατανομή είναι συμμετρική, ενώ όσο απομακρύνεται από το μηδέν και τείνει προς το ±1, τόσο πιο έντονη γίνεται η ασυμμετρία. Για τιμές γύρω στο ±0,5 η ασυμμετρία είναι μέτρια. Αν $S_k>0$ έχουμε θετική ασυμμετρία, ενώ αν $S_k<0$ έχουμε αρνητική ασυμμετρία.

Σε οποιοδήποτε σύνολο δεδομένων που είναι τέλεια συμμετρικό και έχει μόνο μία επικρατούσα τιμή, ο μέσος, η διάμεσος και η επικρατούσα τιμή έχουν την ίδια τιμή. Αν, όμως το σύνολο των δεδομένων έχει μία επικρατούσα τιμή και δεν είναι συμμετρικό, τότε οι τιμές του μέσου, της διαμέσου και της επικρατούσας τιμής διαφέρουν. Αν η κατανομή παρουσιάζει θετική ασυμμετρία, τότε έχει μια μεγάλη δεξιά ουρά και σε αυτή την περίπτωση ο μέσος είναι μεγαλύτερος από τη διάμεσο και την επικρατούσα τιμή. Αντίθετα, αν η καμπύλη παρουσιάζει αρνητική ασυμμετρία, τότε η ουρά εμφανίζεται στα αριστερά της κατανομής και ο μέσος είναι μικρότερος από τη διάμεσο και την επικρατούσα τιμή. Έτσι, αν γνωρίζουμε τις τιμές του μέσου, της διαμέσου και της επικρατούσας τιμής για μια κατανομή, τότε μπορούμε να πούμε αν και προς ποια κατεύθυνση η κατανομή παρουσιάζει ασυμμετρία.

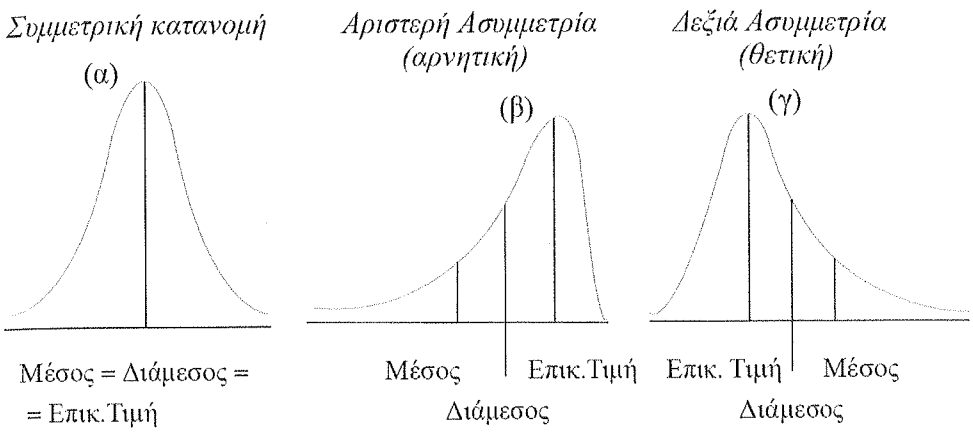
Το σχήμα 2.10(α) παρουσιάζει μια συμμετρική κατανομή με τις αριθμητικές τιμές του μέσου, της διαμέσου και της επικρατούσας τιμής να είναι ίσες. Στο σχήμα 2.10(β), ο μέσος παγιδεύεται στα αριστερά λόγω των ακραίων τιμών στην αριστερή ουρά και ισχύει ότι

μέσος < διάμεσος < επικρατούσα τιμή

Ομοίως στο σχήμα 2.10(γ), ο μέσος παγιδεύεται στα δεξιά λόγω των ακραίων τιμών στη δεξιά ουρά και ισχύει ότι

μέσος > διάμεσος > επικρατούσα τιμή

Σχήμα 2.10



Παράδειγμα 2.16

Χρησιμοποιώντας τα αποτελέσματα των παραδειγμάτων 2.6, 2.9 και 2.10, όπου είχαμε υπολογίσει το μέσο, τη διάμεσο και την επικρατούσα τιμή ενός συγκεκριμένου συνόλου δεδομένων, μπορούμε να πούμε ότι τα δεδομένα παρουσιάζουν αρνητική ασυμμετρία, αφού

$$\text{μέσος} < \text{διάμεσος} < \text{επικρατούσα τιμή}$$
$$101{,}125 < 103 < 104{,}32.$$

Η περίπτωση της θετικής ασυμμετρίας εμφανίζεται στη λυμένη επαναληπτική άσκηση παρακάτω.

Σε κατανομές δεδομένων με ελαφρά ασυμμετρία προτείνεται ο συντελεστής ασυμμετρίας κατά Pearson ως:

$$\text{Ασυμμετρία} \qquad S_k = \frac{3(\text{Μέσος} - \text{Διάμεσος})}{\text{Τυπική απόκλιση}}$$

Για να απαλείψουμε την ασυμμετρία μπορούμε, για παράδειγμα, να εργαστούμε με τη λογαριθμική κλίμακα, με την προϋπόθεση όμως, ότι δεν έχουμε αρνητικές τιμές.

Η κύρτωση είναι ένα μέτρο της κορύφωσης ή επιπεδοποίησης της κατανομής συχνοτήτων. Συνήθως αυτό μετράται σε σχέση με μία κανονική κατανομή. Με απλά λόγια η κύρτωση μετράει πόσο πλατιά ή λεπτή ή επίπεδη είναι η υπό εξέταση κατανομή[10].

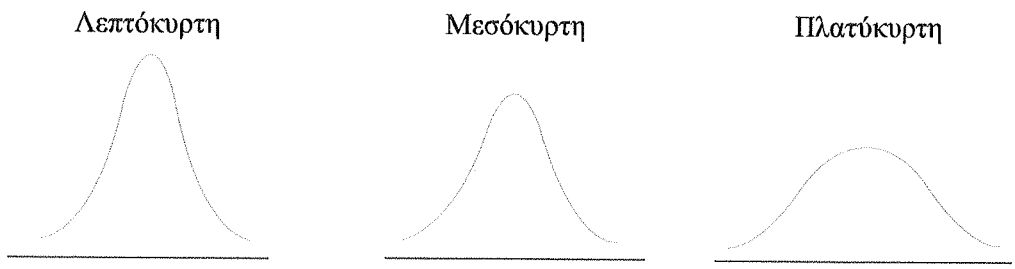
[10] Η κύρτωση, όπως και η ασυμμετρία, μετράται και με τις ροπές της κατανομής συχνοτήτων. Η ροπή μετριέται με τον πολλαπλασιασμό της συχνότητας μιας τάξης επί την απόσταση από το σημείο της μεταβλητής το οποίο θεωρούμε ως αρχή (συνήθως παίρνουμε X=μ ή X=0).

Μια κατανομή μπορεί να έχει μία από τις τρεις μορφές κορύφωσης:

1. *Λεπτόκυρτη*, όταν η κατανομή έχει σχετικά υψηλή κορυφή.
2. *Μεσόκυρτη*, όταν η κορυφή είναι ούτε πολύ υψηλή ούτε πολύ χαμηλή. Η κορυφή της κανονικής κατανομής είναι τέτοιο παράδειγμα.
3. *Πλατύκυρτη*, όταν η κατανομή είναι επίπεδη και η κορυφή χαμηλή.

Το Σχήμα 2.11 παρουσιάζει τις τρεις περιπτώσεις.

Σχήμα 2.11: Μορφές κορύφωσης κατανομών



2.4.1 Συνέπειες της ασυμμετρίας

Μερικές από τις βασικές συνέπειες της ασυμμετρίας είναι ότι:

- ✓ Επηρεάζει την προτίμηση της μέτρησης. Για παράδειγμα, οι περισσότερες κατανομές εισοδημάτων ή κερδών παρουσιάζουν θετική ασυμμετρία. Έτσι, η μέση τιμή των κερδών θα είναι μεγαλύτερη από αυτή της διαμέσου. Σε μια διαπραγμάτευση μισθών μεταξύ εργαζομένων και εργοδοτών, οι εκπρόσωποι των σωματείων των εργαζομένων, θα ήθελαν να χρησιμοποιήσουν τη διάμεσο ως μέτρηση του μέσου όρου των κερδών των εργατών, ενώ η διοίκηση θα προτιμούσε να χρησιμοποιήσει το μέσο.
- ✓ Η άθροιση γίνεται περισσότερο δύσκολη. Για παράδειγμα, παρακάτω παρουσιάζεται η περίπτωση των ημερών απουσίας 12 εργαζομένων λόγω ασθένειας.

0 0 1 1 1 1 1 1 1 2 3 36

(ημέρες απουσίας ανά εργαζόμενο)

Ο μέσος όρος απουσιών για τους 12 εργαζομένους σε αυτή την περίπτωση είναι 4 και η διάμεσος 1. Ποιο είναι το πιο αντιπροσωπευτικό μέτρο; Η απάντηση στην ερώτηση αυτή υποδεικνύει ότι η άθροιση σε τέτοιες περιπτώσεις γίνεται περισσότερο δύσκολη λόγω της μη κανονικότητας των δεδομένων. Ο μέσος και η διάμεσος μπορεί να διαφέρουν κατά πολύ.

2.4.2 Ροπές (Moments)

Ο όρος ροπή πηγάζει από τη Μηχανική και από τη μέτρηση της τάσης μιας δύναμης να παράγει περιστροφή. Συγκεκριμένα η δύναμη της τάσης εξαρτάται από το μέγεθος της δύναμης και από την απόσταση από την αρχή του σημείου πάνω στο οποίο ασκείται η δύναμη.

$$\text{Ροπή} = r * F$$

όπου r η απόσταση από το σημείο εφαρμογής της δύναμης και F η ασκούμενη δύναμη.

Στη Στατιστική και σε ένα ιστόγραμμα μιας κατανομής κάθε στήλη ασκεί πίεση στον άξονα των τετμημένων ίση με την αντίστοιχη της στήλης των συχνοτήτων. Η ροπή κάθε στήλης μετριέται με το γινόμενο της συχνότητας της τάξης και της απόστασης από το σημείο της. Για μια ποσοτική μεταβλητή η *ροπή k τάξης* είναι η μέση τιμή της συνάρτησης $\varphi(X) = X^k$ και συμβολίζεται ως m_k. Δηλαδή

$$m_k = E(X^k) = \frac{\sum_{i=1}^{P} f_i X_i^k}{\sum f_i} \tag{2.28}$$

Βάσει του ορισμού η πρώτη ροπή ισούται με τη μέση αριθμητική τιμής της. Καθώς η τιμή των ροπών βασίζεται στον καθορισμό της αρχής είτε ως x=0 ή ως x=μ έχουμε ροπές περί την αρχή ή περί το μέσο αριθμητικό αντίστοιχα. Ο ακόλουθος πίνακας μας δίνει τους τύπους υπολογισμού των ροπών περί την αρχή και περί το μέσο αριθμητικό.

	Ροπές περί την αρχή (x=0)		**Ροπές περί το μέσο αριθμητικό (x=μ)**	
1η ροπή	$V_1 = \frac{\sum X_i}{N}$	$V_1 = \frac{\sum f_i X_i}{\sum f_i}$	$\mu_1 = \frac{\sum (X_i - \mu)}{N}$	$\mu_1 = \frac{\sum f_i (X_i - \mu)}{\sum f_i}$
2η ροπή	$V_2 = \frac{\sum X_i^2}{N}$	$V_2 = \frac{\sum f_i X_i^2}{\sum f_i}$	$\mu_2 = \frac{\sum (X_i - \mu)^2}{N}$	$\mu_2 = \frac{\sum f_i (X_i - \mu)^2}{\sum f_i}$
3η ροπή	$V_3 = \frac{\sum X_i^3}{N}$	$V_3 = \frac{\sum f_i X_i^3}{\sum f_i}$	$\mu_3 = \frac{\sum (X_i - \mu)^3}{N}$	$\mu_3 = \frac{\sum f_i (X_i - \mu)^3}{\sum f_i}$
Ροπή n τάξης	$V_n = \frac{\sum X_i^n}{N}$	$V_n = \frac{\sum f_i X_i^n}{\sum f_i}$	$\mu_n = \frac{\sum (X_i - \mu)^n}{N}$	$\mu_n = \frac{\sum f_i (X_i - \mu)^n}{\sum f_i}$

Βάσει των ροπών ο συντελεστής ασυμμετρίας του Pearson δίνεται ως $\beta_1 = \frac{\mu_3^2}{\mu_2^3}$. Ο συντελεστής είναι θετικός ή μηδέν ($\beta_1 \geq 0$). Αν $\beta_1 = 0$ έχουμε συμμετρική κατανομή, ενώ αν $\beta_1 > 0$ έχουμε θετική ή αρνητική ασυμμετρία.

Ομοίως ο συντελεστής ασυμμετρίας του Fisher δίνεται ως $\gamma_1 = \frac{\mu_3}{\sqrt{\mu_2^3}} = \frac{\mu_3}{\sigma^3}$. Αν γ_1=0 έχουμε συμμετρική κατανομή ενώ αν $\gamma_1 > (<)$ θετική (αρνητική) ασυμμετρία.

Οι κεντρικές ροπές μπορούν να υπολογιστούν ως συνάρτηση των ροπών περί την αρχή ως:

$$\mu_n = (V - V_1)^n$$
$$\mu_2 = (V - V_1)^2 = V_2 - V_1^2$$
$$\mu_3 = (V - V_1)^3 = V_3 - 3V_2V_1 + 2V_1^3$$
$$\mu_4 = (V - V_1)^4 = V_4 - 4V_3V_1 + 6V_1^2V_2 - 3V_1^4$$

2.4.2.1 Παράμετροι κυρτότητας ή αιχμηρότητας

Ο συντελεστής κύρτωσης του Pearson υπολογίζεται ως:

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{\mu_4}{\sigma^4} \tag{2.29}$$

Όταν β_2= 3 η κατανομή είναι μεσόκυρτη (κανονική), ενώ όταν β_2> 3 έχουμε λεπτόκυρτη και όταν β_2< 3 έχουμε πλατύκυρτη κατανομή.

Ομοίως ο συντελεστής κύρτωσης του Fisher δίνεται ως:

$$\gamma_2 = \frac{\mu_4}{\mu_2^2} - 3 = \beta_2 - 3 \tag{2.30}$$

Όταν γ_2=0 η κατανομή είναι μεσόκυρτη (κανονική) ενώ όταν γ_2>0 έχουμε λεπτόκυρτη και όταν γ_2<0 έχουμε πλατύκυρτη κατανομή.

Παράδειγμα 2.17[11]

Έστω οι θερμοκρασίες σε 100 πόλεις μιας χώρας μία φθινοπωρινή ημέρα του 2019. Ο παρακάτω πίνακας παρουσιάζει τις τάξεις των θερμοκρασιών και τις συχνότητες εμφάνισης.

[11] Από Χάλκος (2019, σελ. 120-121).

Τάξεις	f_i	x_i	$f_i x_i$	$f_i x_i^2$	$f_i x_i^3$	$f_i x_i^4$
2-4	5	3	15	45	135	405
4-6	10	5	50	250	1.250	6.250
6-8	15	7	105	735	5.145	36.015
8-10	40	9	360	3.240	29.160	262.440
10-12	15	11	165	1.815	19.965	219.615
12-14	10	13	130	1.690	21.970	285.610
14-16	5	15	75	1.125	16.875	253.125
Σύνολο	100		900	8.900	94.500	1.063.460

Η μέση θερμοκρασία των 100 πόλεων είναι

$$\mu = \frac{\sum f_i X_i}{\sum f_i} = \frac{900}{100} = 9$$

Η διάμεση θερμοκρασία ισούται με

$$M \approx w_M \left[\frac{\frac{n}{2} - f_b}{f_M} \right] + L_M = 2\left(\frac{\frac{100}{2} - 30}{40} \right) + 8 = 9$$

με 50% των πόλεων να έχουν θερμοκρασίες μικτότερες των 9° και το άλλο 50% θερμοκρασίες υψηλότερες των 9° Κελσίου.

Η επικρατούσα τιμή βρίσκεται ως

$$M_0 \approx w_{M_0} \left[\frac{f_{M_0} - f_{M_{0-1}}}{\left(f_{M_0} - f_{M_0-1}\right) + \left(f_{M_0} - f_{M_0+1}\right)} \right] + L_{M_0} = 2\left(\frac{40-15}{(40-15)+(40-15)} \right) + 8 = 9$$

Δηλαδή, βλέπουμε ότι υπάρχει πλήρης συμμετρία κάθως

$$\mu = M = M_0$$

Πριν τον υπολογισμό των ροπών ας βρούμε τη διακύμανση και την τυπική απόκλιση.

$$\sigma^2 = \frac{\sum f_i x_i^2}{\sum f_i} - \mu^2 = \frac{8.900}{100} - 9^2 = 8 \quad \text{και} \quad \sigma = \sqrt{\sigma^2} = \sqrt{8} = 2,82$$

Για το συντελεστή κυρτότητας υπολογίζουμε πρώτα τις ροπές ως προς την αρχή (x = 0)

$$V_1 = \frac{\sum f_i X_i}{\sum f_i} = \frac{900}{100} = 9 \qquad V_2 = \frac{\sum f_i X_i^2}{\sum f_i} = \frac{8.900}{100} = 89$$

$$V_3 = \frac{\sum f_i X_i^3}{\sum f_i} = \frac{94.500}{100} = 945 \qquad V_4 = \frac{\sum f_i X_i^4}{\sum f_i} = \frac{1.063.460}{100} = 10.634,6$$

$$\mu_2 = (V - V_1)^2 = V_2 - V_1^2 = 89 - 9^2 = 8$$

$$\mu_3 = (V - V_1)^3 = V_3 - 3V_2V_1 + 2V_1^3 = 945 - 3*89*9 + 2*9^3 = 945 - 2403 + 1458 = 0$$

$$\mu_4 = (V - V_1)^4 = V_4 - 4V_3V_1 + 6V_1^2V_2 - 3V_1^4 = 10634,6 - 34020 + 43254 - 19683 = 185,6$$

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{\mu_4}{\sigma^4} = \frac{185,6}{2,82^4} = 2,936 \approx 3$$

Καθώς ο συντελεστής ισούται με 3 η καμπύλη είναι μεσόκυρτη.

Ομοίως ο συντελεστής ασυμμετρίας κατά Fisher υπολογίζεται ως

$$\gamma_1 = \frac{\mu_3}{\sqrt{\mu_2^3}} = \frac{\mu_3}{\sigma^3} = \frac{0}{2,82^3} = 0$$

Επειδή η κατανομή είναι συμμετρική $\beta_1 = \gamma_1 = 0$

Επαναληπτική λυμένη άσκηση

Ας υποθέσουμε ότι έχουμε συγκεντρώσει τις παρακάτω παρατηρήσεις σχετικά με τα μηνιαία ενοίκια που πληρώνουν οι καταστηματάρχες σε μία περιοχή κάποιου Νομού. Να υπολογιστούν τα βασικά περιγραφικά στατιστικά μεγέθη αυτών των παρατηρήσεων.

325 330 330 335 335 335 335 335 340 340 340 340 345 345 345 345 345 345 345
350 350 350 350 350 350 360 360 360 365 365 365 370 370 372 375 375 375 380
380 380 380 385 390 390 390 400 400 400 400 410 410 415 425 425 425 435 449
450 470 470 475 475 480 490 500 500 500 500 515 515

Λύση

Ο δειγματικός μέσος των παρατηρήσεων είναι η πρόσθεση όλων των τιμών, διαιρεμένη με το 70 (το συνολικό αριθμό των παρατηρήσεων). Αυτό μας δίνει κατά μέσο όρο μια τιμή ίση με 390,8. Η διάμεσος είναι η [(n+1)/2] παρατήρηση, η οποία είναι η (70+1)/2=35,5 παρατήρηση. Αυτή είναι ο μέσος όρος της 35ης και 36ης παρατήρησης, δηλαδή [(375+375)/2]=375. Η επικρατούσα τιμή είναι η παρατήρηση που επαναλαμβάνεται τις περισσότερες φορές. Αυτή είναι η τιμή 345.

Το πρώτο τεταρτημόριο είναι η [(n+1)/4]= [(70+1)/4]=17,75 παρατήρηση, που είναι, για απλούστευση, ο μέσος όρος της 17ης και 18ης παρατήρησης. Αύτη είναι η τιμή 345. Ομοίως, το τρίτο τεταρτημόριο είναι μεταξύ της 53ης και 54ης παρατήρησης ως [3(n+1)/4]=53,25. Αυτή είναι η τιμή 425.

Παρόμοια, η δειγματική διακύμανση είναι ίση με 2.997,52 και η δειγματική τυπική απόκλιση είναι ίση με 54,74. Αυτοί οι αριθμοί υπολογίζονται χρησιμοποιώντας τον τύπο της δειγματικής διακύμανσης και μετά παίρνοντας την τετραγωνική ρίζα. Δηλαδή:

$$s^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1} = \frac{(325-390,8)^2 + (330-390,8)^2 + ... + (515-390,8)^2}{70-1} =$$

$$= \frac{206.828,56}{69} = 2.997,52 \qquad \text{και} \qquad s = \sqrt{2.997,52} = 54,74$$

Θα μπορούσαμε, όπως ξέρουμε, να παρουσιάσουμε αυτά τα δεδομένα με ένα πιο "τακτικό" τρόπο ομαδοποίησης σχηματίζοντας ένα πίνακα σύμφωνα με τη διαδικασία που περιγράψαμε στην Παράγραφο 2.2 και στο Παράδειγμα 2.2.

Ενοίκια	Συχνότητες f_i	Μέσο σημείο τάξης M_i	$f_i * M_i$	$M_i - \bar{X}$	$(M_i - \bar{X})^2$	$(M_i - \bar{X})^2 f_i$
320-339	8	329,5	2.636,0	-63.71	4058.96	32.471,71
340-359	17	349,5	5.941,5	-43.71	1910.56	32.479,59
360-379	12	369,5	4.434,0	-23.71	562.16	6.745,97
380-399	8	389,5	3.116,0	-3.71	13.76	110,11
400-419	7	409,5	2.866,5	16.29	265.36	1.857,55
420-439	4	429,5	1.718,0	36.29	1316.96	5.267,86
440-459	2	449,5	899,0	56.29	3168.56	6.337,13
460-479	4	469,5	1.878,0	76.29	5820.16	23.280,66
480-499	2	489,5	979,0	96.29	9271.76	18.543,53
500-519	6	509,5	3.057,0	116.29	13523.36	81.140,19
	$\Sigma = 70$		Σ=27.525			Σ=208.234,3

Τότε ο μέσος των ομαδοποιημένων δεδομένων θα είναι:

$$\bar{X} = \frac{\sum_{j=1}^{k} f_j X_{mj}}{\sum_j f_j} = \frac{27.525}{70} = 393,21$$

Αυτή η προσέγγιση διαφέρει κατά 2,41 από τον πραγματικό δειγματικό μέσο. Τώρα αν θέλουμε να υπολογίσουμε τη διάμεσο για τα ομαδοποιημένα δεδομένα πρέπει πρώτα να βρούμε την τάξη της διαμέσου. Αυτό γίνεται υπολογίζοντας την n/2 παρατήρηση, δηλαδή την 70/2 παρατήρηση. Αυτή είναι η 35η παρατήρηση, η οποία ανήκει στην 3η τάξη (8 + 17 + 12), δηλαδή στην τάξη 360 – 379. Χρησιμοποιώντας τη σχέση για τη διάμεσο έχουμε:

$$M \cong W_M \left[\frac{\frac{n}{2} - f_b}{f_M} \right] + L_M = 20 \left[\frac{\frac{70}{2} - 25}{12} \right] + 360 = 376,67$$

Αυτή είναι μια προσέγγιση της διαμέσου για τα ομαδοποιημένα δεδομένα.

Για να βρούμε την επικρατούσα τιμή χρησιμοποιούμε την τάξη με τη μεγαλύτερη συχνότητα, που είναι η 340–349. Στη συνέχεια αντικαθιστώντας στη σχέση για την επικρατούσα τιμή έχουμε:

$$M_0 \cong W_m \left[\frac{f_m - f_{m-1}}{(f_m - f_{m-1}) + (f_m - f_{m+1})} \right] + L_m - 20 \left[\frac{17-8}{(17-8)+(17-12)} \right] + 340 =$$

$$= 20 \left[\frac{9}{9+5} \right] + 340 = 352,86$$

Η δειγματική διακύμανση των ομαδοποιημένων δεδομένων είναι ίση με 3.017,89 και η δειγματική τυπική απόκλιση είναι ίση με 54,94. Αυτοί οι αριθμοί υπολογίζονται χρησιμοποιώντας τον τύπο της δειγματικής διακύμανσης και μετά παίρνοντας την τετραγωνική ρίζα. Δηλαδή:

$$s^2 = \frac{\sum_{i=1}^{k} f_i (X_{mi} - \bar{X})^2}{(\sum f_i) - 1} = \frac{208.234,3}{69} = 3.017,89$$

και $$s = \sqrt{3.017,89} = 54,94$$

Το πρώτο τεταρτημόριο μπορεί να προσεγγιστεί βρίσκοντας πρώτα την τάξη που το περιέχει. Αυτή είναι η δεύτερη τάξη όπου ανήκει η (n/4] παρατήρηση. Στην περίπτωσή μας έχουμε 70/4=17,5.

$$Q_1 \cong W_{Q_1}[\frac{\frac{n}{4} - f_b}{f_{Q_1}}] + L_{Q_1} = 20\left[\frac{17,5-8}{17}\right] + 340 =$$

$$= 20(0,559) + 340 = 351,18$$

Στη συνέχεια, ομοίως για το τρίτο τεταρτημόριο, βρίσκουμε την τάξη που το περιέχει. Και αυτή η τάξη είναι η 420–439 ως (3n)/4=52,5. Κοιτώντας για την 52η παρατήρηση, αυτή ανήκει στην έκτη τάξη, όπου η αθροιστική συχνότητα εμπεριέχει αυτή την παρατήρηση. Στη συνέχεια αντικαθιστούμε στην αντίστοιχη σχέση και έχουμε:

$$Q_3 \cong W_{Q_3}\left[\frac{\frac{3(n)}{4} - f_b}{f_{Q_3}}\right] + L_{Q_3} = 20\left[\frac{52,5-52}{4}\right] + 420 = 20(0.125) + 420 = 422,5$$

Αξίζει να σημειωθεί το σύνολο των παρατηρήσεων παρουσιάζει θετική ασυμμετρία, όπως αναφέρεται και εξηγείται στην παράγραφο 2.4

2.5 Χρησιμοποιώντας τον Η/Υ

Το MINITAB και το SPSS είναι δύο πολύ διαδεδομένα στατιστικά πακέτα, τα οποία μπορούν να διεκπεραιώσουν σειρά στατιστικών αναλύσεων και τεχνικών. Ομοίως το πακέτο EXCEL είναι ένα πρόγραμμα φύλλου εργασίας (spreadsheet) με ποικίλες χρησιμότητες, το οποίο μπορεί να πραγματοποιήσει στατιστική ανάλυση σε πολύ αξιόλογο επίπεδο. Εδώ επιχειρείται η εισαγωγή στην πραγματοποίηση των βασικών βημάτων του Κεφαλαίου 2. Για μεγαλύτερη εξοικείωση συνιστάται η ανάγνωση των οδηγιών (manuals) για τα προγράμματα.

Μπορούμε να εισάγουμε τη βάση των δεδομένων μας από το πληκτρολόγιο. Τα προγράμματα αυτά διαθέτουν φύλλα εργασίας που χωρίζονται σε γραμμές και στήλες. Οι γραμμές συμβολίζονται με 1,2,3,...και οι στήλες εμφανίζονται διαδοχικά, και στο MINITAB έχουν τα ονόματα C1, C2, C3,..., στο SPSS VAR00001, VAR00002, VAR00003,...και στο EXCEL A, B, C,...

Η εισαγωγή των δεδομένων απαιτεί τη διαδοχική πληκτρολόγηση των στοιχείων της βάσης δεδομένων, γραμμή γραμμή. Η πληροφορία εισάγεται μια γραμμή τη φορά.

Πρέπει μόνο να κινούμαστε στην αρχή κάθε στήλης, είτε με τον κέρσορα είτε με το ποντίκι, να πληκτρολογούμε την τιμή και να πατάμε ENTER. Στη συνέχεια πληκτρολογούμε την επόμενη τιμή και πατάμε ENTER κ.ο.κ. Όταν τελειώσουμε με όλες τις στήλες μπορούμε να αρχίσουμε την ανάλυσή μας.

Αν θέλουμε μπορούμε να ονομάσουμε τις μεταβλητές στο MINITAB στο ακριβώς επάνω κελί που είναι διαθέσιμο για αυτόν το σκοπό, στο EXCEL αφήνοντας ένα κενό κελί πάνω από την πρώτη παρατήρηση και στο SPSS πηγαίνοντας κάτω αριστερά και επιλέγοντας με το ποντίκι Variable View. Αμέσως ανοίγουν μία σειρά επιλογών, η πρώτη από τις οποίες είναι NAME. Γράφουμε το όνομα που χαρακτηρίζει τη μεταβλητή και επιλέγουμε DATA VIEW για να επιστρέψουμε στα δεδομένα.

Σημειώνουμε εδώ ότι μπορούμε να εισάγουμε δεδομένα από ένα ήδη υπάρχον αρχείο πληροφοριών. Μπορούμε να διαβάσουμε τα δεδομένα από ένα αρχείο αριθμών ASCII. Επιπλέον, μπορούμε να διαβάσουμε πληροφορίες από ένα αρχείο ASCII μέσα σε ένα φύλλο εργασίας SPSS/MINITAB υπό την προϋπόθεση ότι γνωρίζουμε τη μορφή του αρχείου ASCII (δηλαδή ποιες μεταβλητές είναι αποθηκευμένες σε ποιες στήλες) ή να ανοίξουμε ένα αρχείο EXCEL ή να αντιγράψουμε (COPY) και να επικολλήσουμε (PASTE) τα δεδομένα στα φύλλα εργασίας των προγραμμάτων.

Από τη στιγμή που έχουμε εισάγει όλη την πληροφορία, το SPSS, το MINITAB και το EXCEL, αλλά και οποιοδήποτε στατιστικό πακέτο, μας παρέχει τα βασικά περιγραφικά στατιστικά μεγέθη. Ας πάρουμε τα προγράμματα με τη σειρά.

MINITAB

Ας εξετάσουμε τα δεδομένα του παραδείγματος με τα ενοίκια. Αφού εισάγουμε τα δεδομένα στη συνέχεια πηγαίνουμε στην επιλογή

- **STAT**,
- επιλέγουμε **BASIC STATISTICS**
- και από αυτή την επιλογή επιλέγουμε **DESCRIPTIVE STATISTICS**.

Τα αποτελέσματα για το παράδειγμα των ενοικίων έχουν ως εξής:

Descriptive Statistics

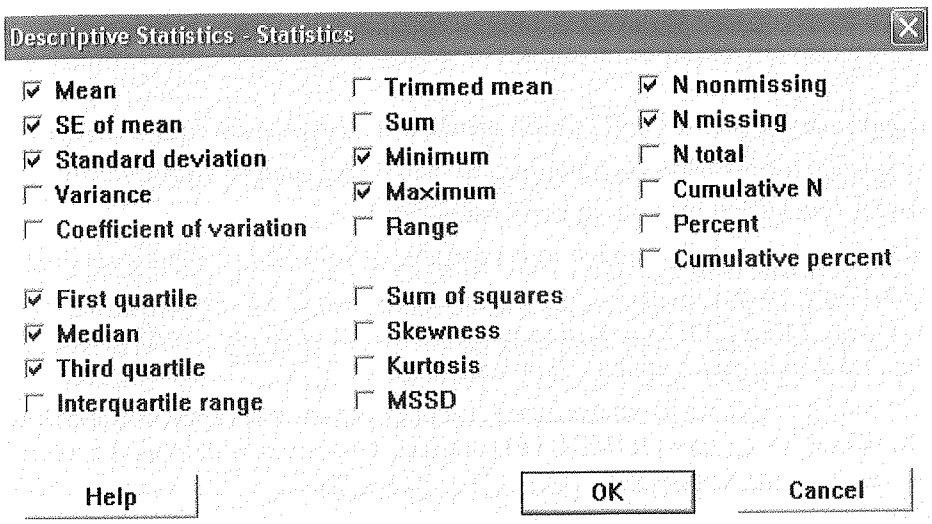
Variable	N	Mean	Median	Tr Mean	StDev	SE Mean
Ενοίκια	70	390.80	375.00	387.19	54.74	6.54

Variable	Min	Max	Q1	Q3
Ενοίκια	325.00	515.00	345.00	425.00

Όπως παρατηρούμε το MINITAB μας παρέχει μεταξύ άλλων, τον αριθμό του συνόλου των παρατηρήσεων (N), το μέσο (Mean), τη διάμεσο (Median), τον 5% τετριμμένο μέσο (Tr Mean), την τυπική απόκλιση (St Dev), την ελάχιστη (Min) και τη μέγιστη (Max) τιμή των παρατηρήσεων και το 1ο (Q1) και 3ο (Q3) τεταρτημόριο. Το αποτέλεσμα με τον τίτλο SE Mean είναι το τυπικό σφάλμα του μέσου (ένα μέγεθος που υπολογίζεται διαιρώντας την τυπική απόκλιση με την τετραγωνική ρίζα του N). Ο σκοπός αυτής της διαίρεσης θα επεξηγηθεί περισσότερο στην ενότητα της Επαγωγικής Στατιστικής.

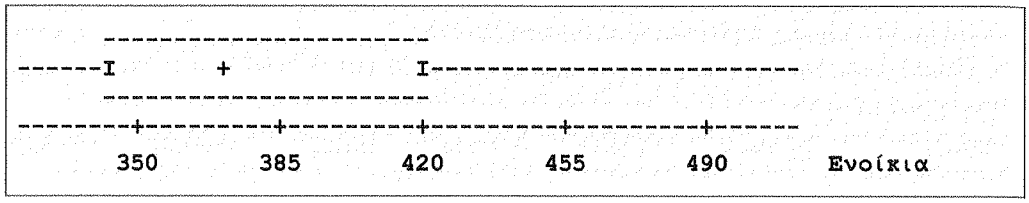
Παρόλο που οι τιμές του εύρους (Range) των παρατηρήσεων, του ενδοτεταρτημοριακού εύρους (Interquartile range), της διακύμανσης (Variance) και του συντελεστή μεταβλητότητας (Coefficient of variation) δεν εμφανίζονται στα αρχικά αποτελέσματα του MINITAB, μπορούμε να ζητήσουμε τις τιμές αυτές, χρησιμοποιώντας την επιλογή **STATISTICS**, επιλέγοντας τα επιπρόσθετα στατιστικά μεγέθη (όπως φαίνεται στην Εικόνα 2.1) και πατώντας **OK**.

Εικόνα 2.1: Επιλογές μέτρων περιγραφικής Στατιστικής



Η χρήση του MINITAB μπορεί να συνεχιστεί με τη γραφική παρουσίαση αυτών των δεδομένων. Αυτό επιτυγχάνεται με την επιλογή **GRAPH**. Ενδιαφέρον παρουσιάζει η επιλογή **BOXPLOT**, δηλαδή το γραφικό κουτί παρουσίασης των δεδομένων. Αυτό το κουτί χρησιμοποιεί το σύμβολο "+" για να εντοπίσει τη διάμεσο. Το σύμβολο "I" εντοπίζει το 1ο και 3ο τεταρτημόριο και αναγνωρίζει το τέλος του κουτιού που περιέχει το 50% της πληροφορίας. Αν ένας "*" παρουσιάζεται τότε έχουμε μια ακραία τιμή. Αν το MINITAB χρησιμοποιεί το " ° " σύμβολο, τότε έχουμε μια εξαιρετικά ακραία τιμή.

Σχήμα 2.12: Γραφική παράσταση θηκογράμματος (Boxplot)



Παράδειγμα 2.18

Έστω ότι έχουμε συγκεντρώσει τους παρακάτω αριθμούς:

28	14	19	55	28	22	7	28
41	52	21	81	40	32	90	11
3	50	48	56	30	71	24	51
34	44	28	69	37	42	23	79
11	53	24	39	27	45	39	28

Χρησιμοποιώντας το MINITAB ας πάρουμε τα περιγραφικά στατιστικά μεγέθη. Για να διατάξουμε αυτούς τους αριθμούς, το MINITAB απαιτεί τη χρήση του

- **DATA** (για να χειριστούμε τη βάση δεδομένων)→
- **SORT** (για να ταξινομήσουμε τους αριθμούς σε αύξουσα (ή φθίνουσα) σειρά →
- επιλέγουμε τη μεταβλητή για να διαταχθεί, πατώντας τη μεταβλητή δύο φορές ή πατώντας SELECT). Στην περίπτωση μας υπάρχει μόνο μία μεταβλητή, η C1, την οποία βάζουμε στην υποδοχή "Sort column(s)".
- Στη συνέχεια γράφουμε το όνομα της μεταβλητής στην πρώτη γραμμή **SORT BY COLUMN (C1)**, όπως το MINITAB απαιτεί, ώστε να καθορίσουμε τη μεταβλητή που διατάχτηκε δύο φορές→
- **OK**.

Τα αποτελέσματα θα είναι τα εξής:

3	7	11	11	14	19	21	22
23	24	24	27	28	28	28	28
28	30	32	34	37	39	39	40
41	42	44	45	48	50	51	52
53	55	56	69	71	79	81	90

Σημειώνουμε ότι το MINITAB δεν επηρεάζεται από το εάν διατάξουμε τα δεδομένα ή όχι. Διατάσσουμε τα στοιχεία προς όφελός μας και για να κατανοήσουμε τους υπολογισμούς που γίνονται από το πρόγραμμα. Στην συνέχεια μπορούμε να ζητήσουμε τα περιγραφικά στατιστικά μεγέθη ως ακολούθως:

- **STAT**→
- **BASIC STATISTICS**→
- **DESCRIPTIVE STATISTICS**→ μπορούμε να επιλέξουμε όλες τις διαθέσιμες μεταβλητές, αλλά επειδή στο συγκεκριμένο παράδειγμα έχουμε μόνο μία επιλέγουμε τη C1 και στη συνέχεια **SELECT**, ώστε να βάλουμε τη μεταβλητή στις υπό εξέταση μεταβλητές→
- **OK**.

Τα αποτελέσματα θα έχουν ως εξής:

Descriptive Statistics

Variable	N	Mean	Median	Tr Mean	StDev	SE Mean
C1	40	38.10	35.50	37.31	20.46	3.24

Variable	Min	Max	Q1	Q3
C1	3.00	90.00	24.00	50.75

Όπως αναφέραμε ήδη, το MINITAB μας δίνει τη δυνατότητα επιλογής εκτός των παραπάνω στατιστικών μεγεθών να πάρουμε στα αποτελέσματά μας μεταξύ των άλλων και τη διακύμανση (variance), το ενδοτεταρτημοριακό εύρος (interquartile range) και τους συντελεστές κύρτωσης και ασυμμετρίας (skewness and kurtosis). Αν ακολουθήσουμε τα βήματα αυτά, τότε τα αρχικά αποτελέσματα γίνονται:

Descriptive Statistics C1

Variable	Total Count	N	N*	Mean	SE MEan	Tr Mean	StDev	Variance	CoefVar
C1	40	40	0	38.10	3.24	37.31	20.46	418.66	53.70

Variable	Sum	Minimum	Q1	Median	Q3	Maximum	Range	IQR
C1	1524.00	3.00	24.00	35.50	50.75	90.00	87.00	26.75

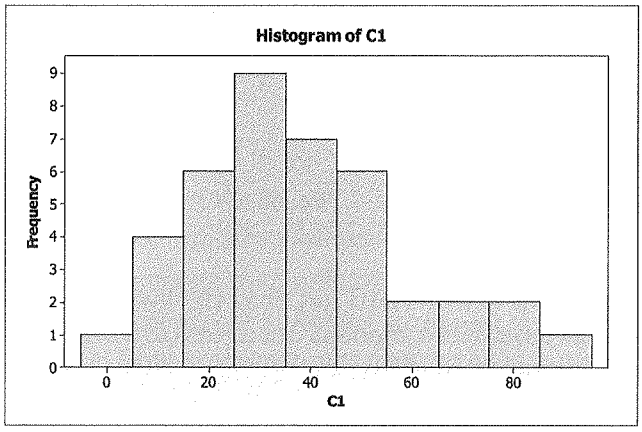
Variable	Skewness	Kurtosis
C1	0.68	0.20

Επίσης, στην επιλογή GRAPHS μπορούμε να επιλέξουμε:

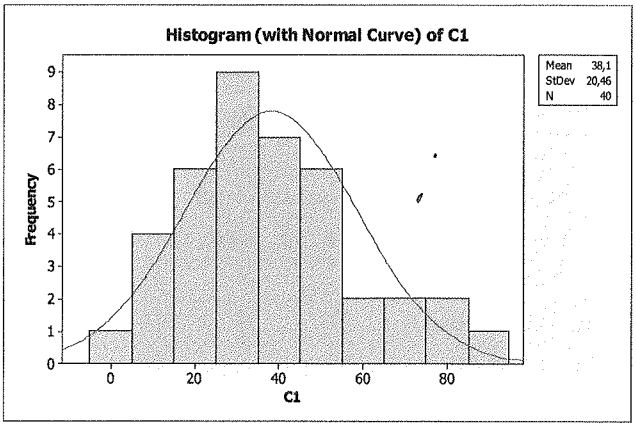
- **Histogram of data** (Σχήμα 2.13)
- **Histogram of data, with normal curve** (Σχήμα 2.14)
- **Individual value plot**
- **Boxplot of data**

Δύο παραδείγματα γραφικών παραστάσεων παρουσιάζονται στα Σχήματα 2.13 και 2.14. Το πρώτο παρουσιάζει το ιστόγραμμα των τιμών της μεταβλητής C1 και το δεύτερο το ιστόγραμμα της C1, αυτή τη φορά με την εμφάνιση της κανονικής κατανομής, ώστε να είναι ευκολότερη η αναγνώριση/εντοπισμός της κανονικότητας των τιμών της C1. Ως πρακτική εξάσκηση αφήνεται στον αναγνώστη η εξαγωγή των σχημάτων για το παράδειγμα με τα ενοίκια.

Σχήμα 2.13: Ιστόγραμμα



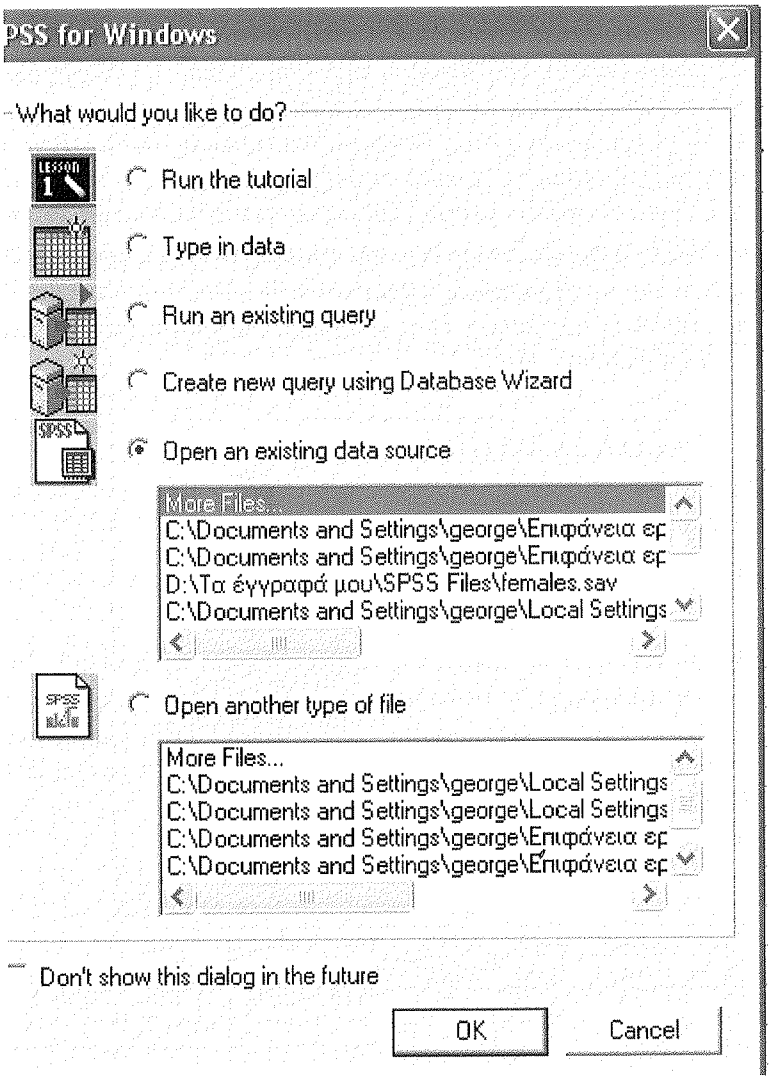
Σχήμα 2.14: Ιστόγραμμα με εμφάνιση καμπύλης κανονικής κατανομής



SPSS

Ανοίγοντας το πρόγραμμα βλέπουμε την ακόλουθη εικόνα (Εικόνα 2.2) στην οποία ερωτόμαστε τι επιθυμούμε να κάνουμε. Ανάμεσα στις επιλογές έχουμε την εισαγωγή δεδομένων (Type in data), ή το άνοιγμα ενός υπάρχοντος SPSS τύπου αρχείο (Open an existing data source) ή κάποιας άλλης μορφής αρχείο (π.χ. EXCEL). Πατάμε είτε OK, αφού επιλέξουμε κάποια από τις δυνατότητες αυτές, είτε κλείνοντας το παράθυρο με CANCEL ή X (πάνω δεξιά στο συγκεκριμένο παράθυρο). Επίσης μπορούμε να διαλέξουμε να μην εμφανίζεται το παράθυρο αυτό επιλέγοντας Don't show this dialog in the future.

Εικόνα 2.2: Επιλογές αρχικών εργασιών με SPSS



Μόλις βρεθούμε στο φύλλο εργασίας εισάγουμε τα δεδομένα. Το SPSS διαθέτει διάφορες διαδικασίες για την πραγματοποίηση Περιγραφικής Στατιστικής στις τιμές μιας μεταβλητής. Για ποιοτικά χαρακτηριστικά μπορούμε να χρησιμοποιήσουμε πίνακες συχνοτήτων και γραφικές παραστάσεις. Από τη στιγμή που είμαστε στο πρόγραμμα πρέπει να πληκτρολογήσουμε τις ακόλουθες εντολές ώστε να εμφανιστούν στην οθόνη τα περιγραφικά στατιστικά μεγέθη.

- **ANALYZE→**
- **SUMMARIZE**
- **FREQUENCIES**
- Επιλέγουμε τη μεταβλητή (-ές) που θέλουμε να αναλύσουμε και την (τις) τοποθετούμε στο παράθυρο Variable(s).
- Αν και το πρόγραμμα στη συγκεκριμένη επιλογή δεν θα υπολογίσει κανένα από τα στατιστικά μέτρα (μόνο η επικρατούσα τιμή έχει νόημα), η επιλογή **Statistics** δίνει τη δυνατότητα εξαγωγής περιγραφικών στατιστικών μεγεθών, όπως μέτρα θέσης και κεντρικής τάσης (μέσος, διάμεσος, επικρατούσα τιμή, άθροιση), κατανομής (κύρτωση, ασυμμετρία), διασποράς (τυπική απόκλιση, διακύμανση, εύρος, ελάχιστη και μέγιστη τιμή, τυπικό σφάλμα μέσου), καθώς και τεταρτημόρια και εκατοστημόρια.
- Η επιλογή **Chart Type** μας επιτρέπει την κατασκευή ραβδογραμμάτων, κυκλικών διαγραμμάτων και ιστογραμμάτων (για ποσοτικές μεταβλητές). Αν επιλέξουμε την εμφάνιση κανονικής καμπύλης, τότε μπορούμε διαγραμματικά να αποφασίσουμε για την ύπαρξη κανονικότητας.

Σε ποσοτικά χαρακτηριστικά επιλέγουμε:

- **ANALYZE→**
- **DESCRIPTIVE STATISTICS →**
- **DESCRIPTIVES**

Στο παράθυρο που ανοίγει επιλέγουμε τη μεταβλητή (-ές) που θέλουμε να αναλύσουμε. Αν θέλουμε μπορούμε να επιλέξουμε **OPTIONS** (και με αυτόν τον τρόπο να συμπεριλάβουμε έναν αριθμό αποτελεσμάτων στην έξοδο), όπου διαλέγουμε όποια επιλογή επιθυμούμε (διακύμανση, εύρος, κύρτωση κ.λπ.). Όταν είμαστε έτοιμοι πατάμε **CONTINUE→OK**.

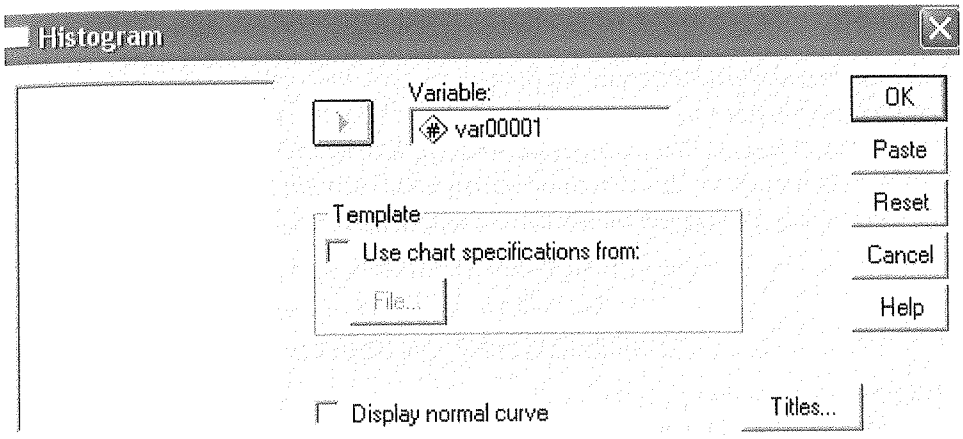
Τα αποτελέσματα έχουν τη μορφή που ακολουθεί:

Descriptive Statistics

	N	Range	Minimum	Maximum	Mean	Std.Deviation	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
VAR00001	40	87,00	3,00	90,00	38,1000	20,46110	,676	,374	,204	,733
Valid N (listwise)	40									

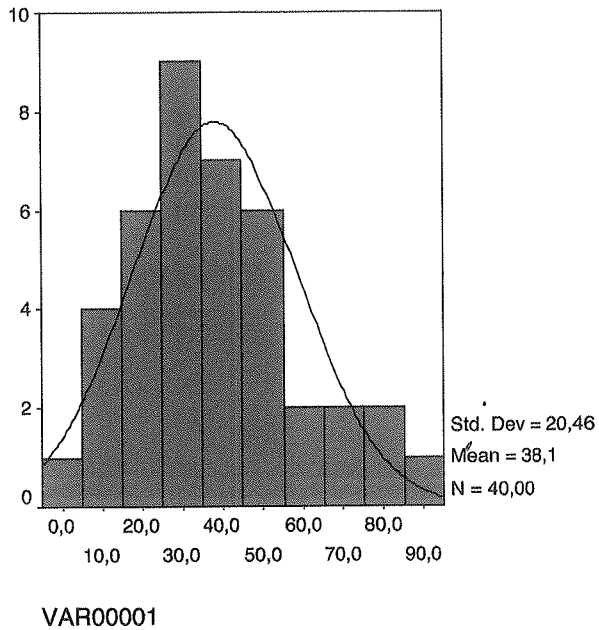
Μπορούμε να ζητήσουμε γραφική παράσταση με αρχική επιλογή **GRAPHS**, κατόπιν **HISTOGRAM** και στο πλαίσιο που μας δίνεται βάζουμε στην υποδοχή Variable την προς εξέταση μεταβλητή (Εικόνα 2.3).

Εικόνα 2.3: Επιλογή κατασκευής ιστογράμματος



Επιλέγοντας **Display normal curve**, παίρνουμε το ιστόγραμμα μαζί με την καμπύλη της κανονικής κατανομής, όπως φαίνεται στο Σχήμα 2.15.

Σχήμα 2.15: Ιστόγραμμα με εμφάνιση καμπύλης κανονικής κατανομής



EXCEL

Προγράμματα όπως το EXCEL μπορούν να μας παράσχουν τα βασικά στατιστικά μεγέθη. Αυτό μπορεί να επιτευχθεί είτε με τις επιλογές ΕΙΣΑΓΩΓΗ (Insert), f_x ΣΥΝΑΡΤΗΣΕΙΣ (f_x Function...) και αμέσως μετά ΣΤΑΤΙΣΤΙΚΕΣ (Statistical) είτε με την ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ (Data Analysis) στην επιλογή ΕΡΓΑΛΕΙΑ (Tools).

Στην πρώτη περίπτωση εμφανίζεται μια λίστα με διαθέσιμα στατιστικά μεγέθη. Για παράδειγμα, αν επιλέξουμε Average, θα βρούμε το μέσο των αριθμών που ενδιαφερόμαστε και που έχουμε εισαγάγει σε συγκεκριμένα κελιά του EXCEL. Είναι σημαντικό να ορίσουμε τα κελιά, για να μπορέσει το πρόγραμμα να μας δώσει το αποτέλεσμα. Ομοίως μπορούμε να βρούμε τη διακύμανση, την τυπική απόκλιση κ.λπ. Ας βρούμε το γεωμετρικό μέσο του Παραδείγματος 2.7. Εισάγουμε τα δεδομένα στις στήλες Α και Β (ή όπου επιθυμούμε) ως

2016-2017	3,33%
2017-2018	16,13%
2018-2019	16,70%

Κατόπιν επιλέγουμε

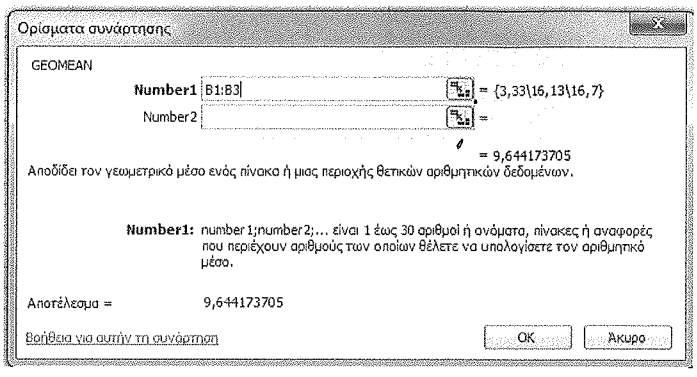
- **ΕΙΣΑΓΩΓΗ (Insert),**
- **f_x ΣΥΝΑΡΤΗΣΕΙΣ (f_x Function...)** και αμέσως μετά
- **ΣΤΑΤΙΣΤΙΚΕΣ (Statistical)**
- **GEOMEAN**

Στο παράθυρο που ανοίγει (Εικόνα 2.4) εισάγουμε τη στήλη με τις τιμές, που θέλουμε να βρούμε το γεωμετρικό μέσο (b1:b3) και αυτομάτως στο κάτω τμήμα εμφανίζεται το αποτέλεσμα

Formula result= 9,644174.

Αν πατήσουμε OK το αποτέλεσμα εμφανίζεται στο κελί που έχουμε επιλέξει αφήνοντας τον κέρσορα πριν τις επιλογές πραγματοποίησης της ανάλυσης.

Εικόνα 2.4: Υπολογισμός γεωμετρικού μέσου

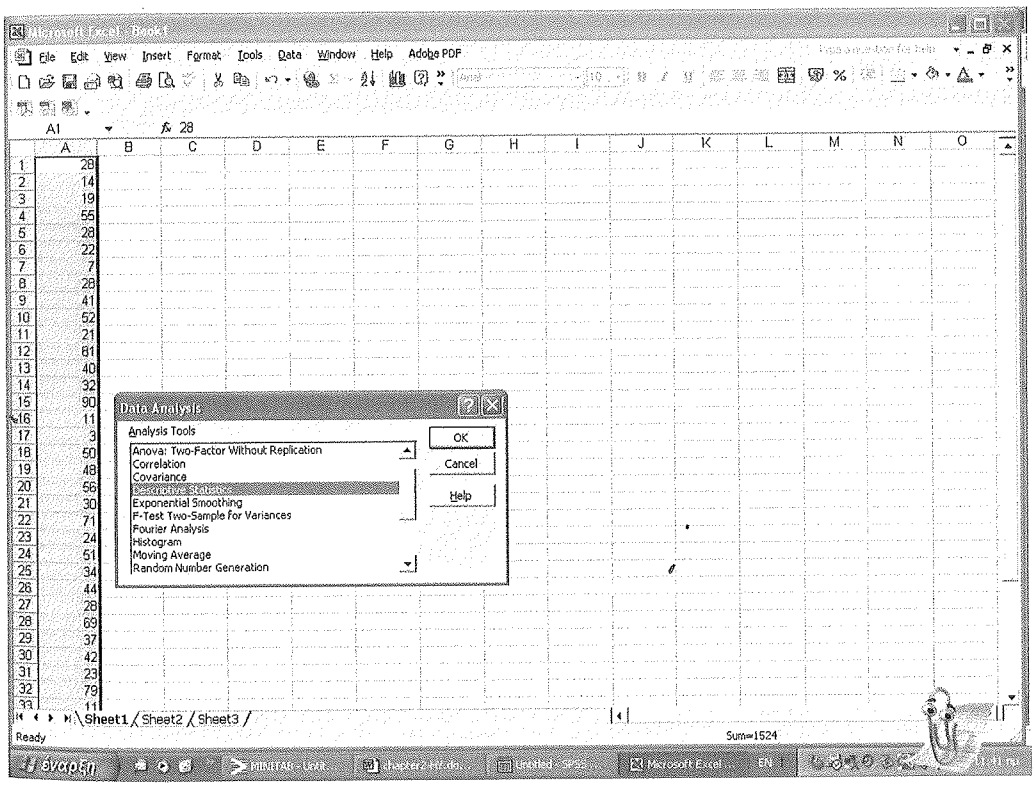


Στη δεύτερη περίπτωση, τα αποτελέσματα είναι συγκεντρωμένα και η επιλογή ΠΕΡΙΓΡΑΦΙΚΑ ΣΤΑΤΙΣΤΙΚΑ (DESCRIPTIVE STATISTICS) μας δίνει μια σειρά περιγραφικών στατιστικών, αφού επιλέξουμε Summary Statistics (μέσος, διάμεσος, τυπικό σφάλμα, επικρατούσα τιμή, τυπική απόκλιση, διακύμανση, συντελεστή κύρτωσης, συντελεστή ασυμμετρίας, εύρος, ελάχιστη και μέγιστη τιμή, άθροιση και αριθμό παρατηρήσεων). Αν δεν υπάρχει η επιλογή Data Analysis, τότε στο Tools επιλέγουμε Add-Ins και, αμέσως, Analysis ToolPack.

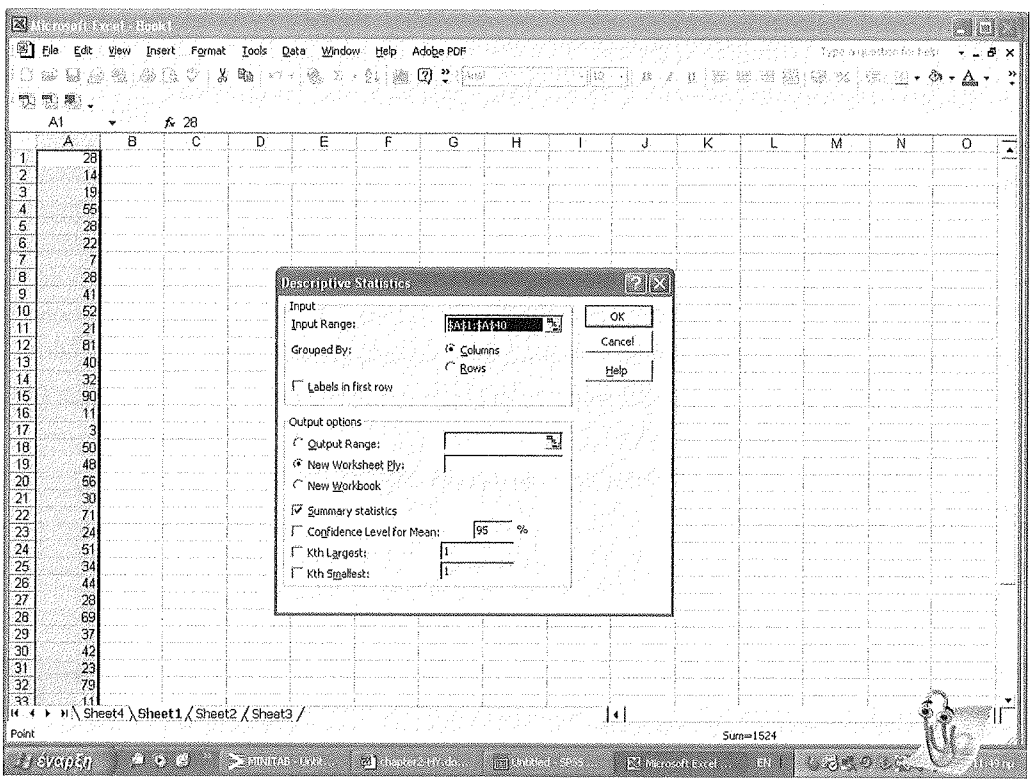
Αν θέλουμε να πραγματοποιήσουμε το παράδειγμα 2.17 τότε επιλέγουμε

- **Tools** (για όσους έχουν Ελληνική έκδοση Εργαλεία)
- **Data Analysis** (Ανάλυση Δεδομένων)
- **Descriptive Statistics** (Περιγραφικά Στατιστικά) (Εικόνα 2.5)
- Στο κουτί **Input Range** βάζουμε τα δεδομένα που θέλουμε να αναλύσουμε (πχ. Στη περίπτωσή μας a1:a40)
- Απαραίτητη είναι η επιλογή του **Summary Statistics**
- **OK** (Εικόνα 2.6)

Εικόνα 2.5: Πραγματοποίηση περιγραφικής στατιστικής



Εικόνα 2.6: Επιλογές για εύρεση μέτρων Περιγραφικής Στατιστικής



Τα αποτελέσματα είναι τα ακόλουθα:

Column1	
Mean	38,1
Standard Error	3,235183187
Median	35,5
Mode	28
Standard Deviation	20,46109504
Sample Variance	418,6564103
Kurtosis	0,204196401
Skewness	0,676180021
Range	87
Minimum	3
Maximum	90
Sum	1524
Count	40

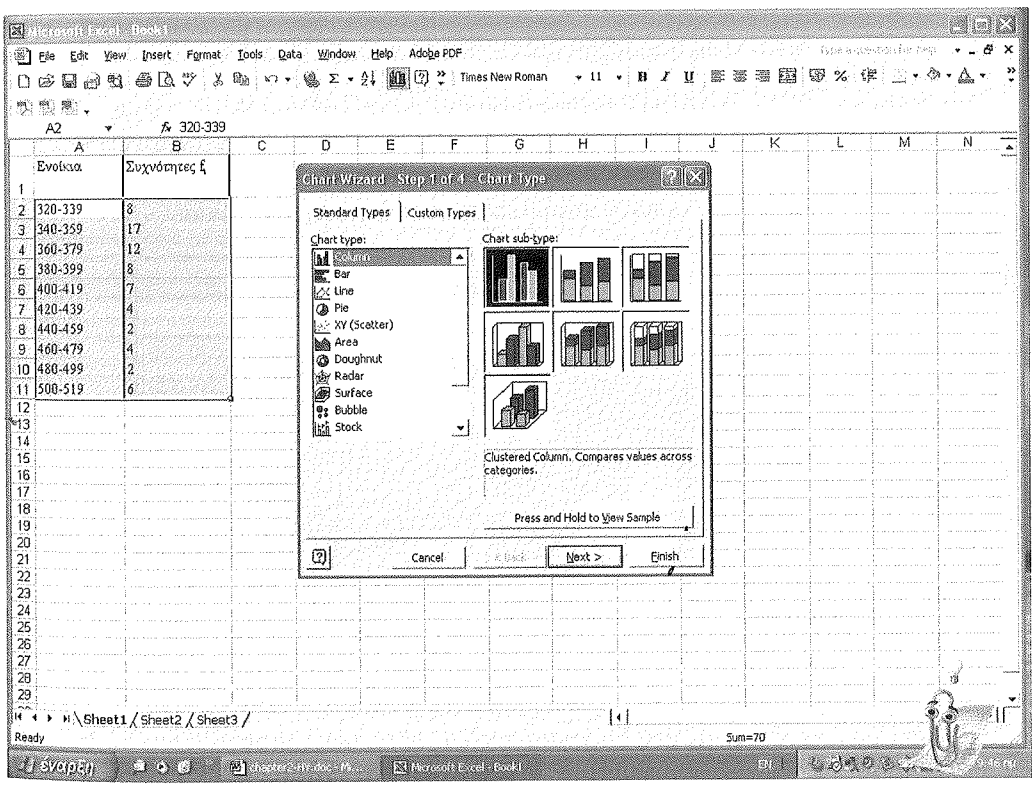
Οι αντίστοιχοι ελληνικοί στατιστικοί όροι των αποτελεσμάτων του EXCEL είναι μέσος (mean), τυπικό σφάλμα (standard error), διάμεσος (median), επικρατούσα τιμή (mode), τυπική απόκλιση (standard deviation), διακύμανση (variance), συντελεστής κύρτωσης (kurtosis), συντελεστής ασυμμετρίας (skewness), εύρος (range), ελάχιστη (minimum) και μέγιστη (maximum) τιμή, άθροιση (sum) και αριθμός παρατηρήσεων (count).

Αν θέλουμε να προχωρήσουμε σε γραφικές παραστάσεις τότε με τη βοήθεια του εικονιδίου "οδηγός γραφημάτων" (το ίδιο επιτυγχάνεται και με τις επιλογές INSERT, Chart), επιλέγουμε το γράφημα που επιθυμούμε.

Μπορούμε να επιλέξουμε ανάμεσα σε μια σειρά επιλογών για ράβδους, γραμμές, πίτες, διαγράμματα διασποράς, κ.λπ. Με την επιλογή Next (Επόμενο), βλέπουμε μια προεπισκόπηση του διαγράμματος. Με τις επιλογές Next (Επόμενο) και Titles (Τίτλοι) δίνουμε (αν επιθυμούμε) ονόματα στους άξονες των διαγραμμάτων καθώς και τίτλο στο κατασκευασμένο διάγραμμα.

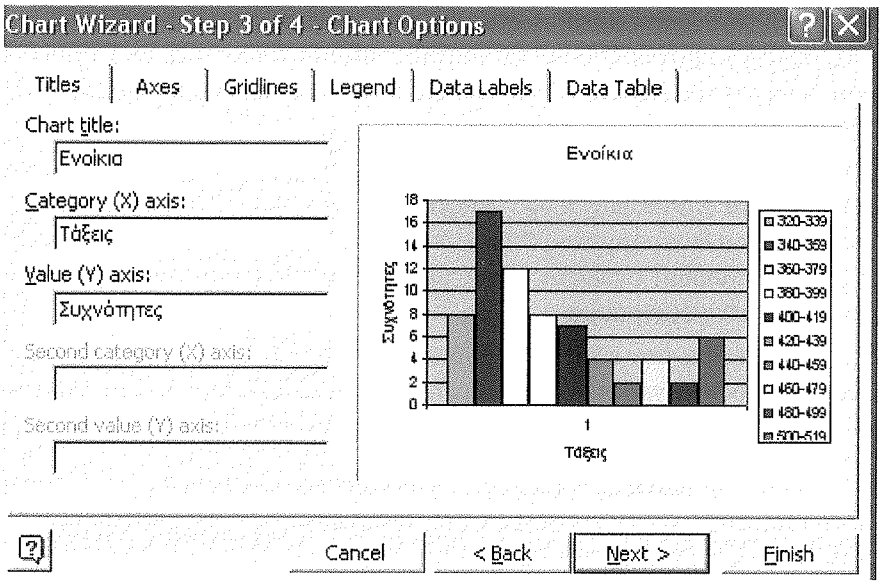
Για το παράδειγμα των ενοικίων αφού εισάγουμε τα δεδομένα στις στήλες A και B επιλέγουμε τον "οδηγό γραφημάτων" και αμέσως Column, την πρώτη επιλογή πάνω αριστερά και επιλογή Rows και Next. Τα βήματα αυτά φαίνονται στην εικόνα 2.7.

Εικόνα 2.7: Επιλογές γραφικών παραστάσεων



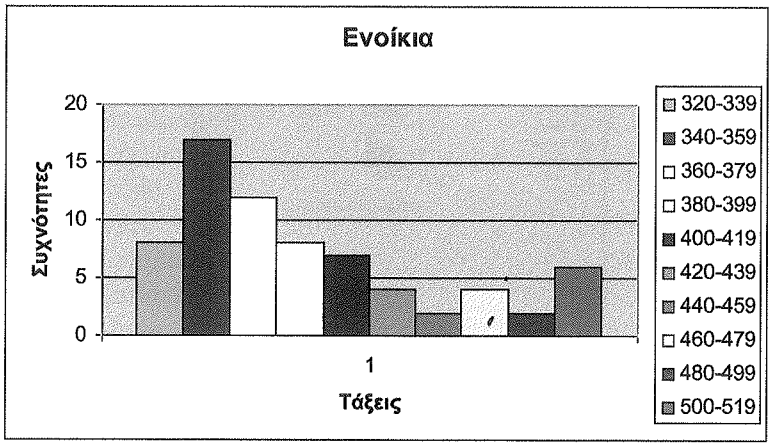
Εξάγουμε το ακόλουθο γράφημα (Εικόνα 2.8), όπου στο αριστερό τμήμα εισάγουμε αρχικά τον τίτλο του γραφήματος, το χαρακτηρισμό του οριζόντιου άξονα (Τάξεις) και κατόπιν του κάθετου άξονα (Συχνότητες).

Εικόνα 2.8: Τοποθέτηση λεπτομερειών στο γράφημα



Επιλέγουμε NEXT, κατόπιν Finish και έχουμε το τελικό γράφημα (Σχήμα 2.16).

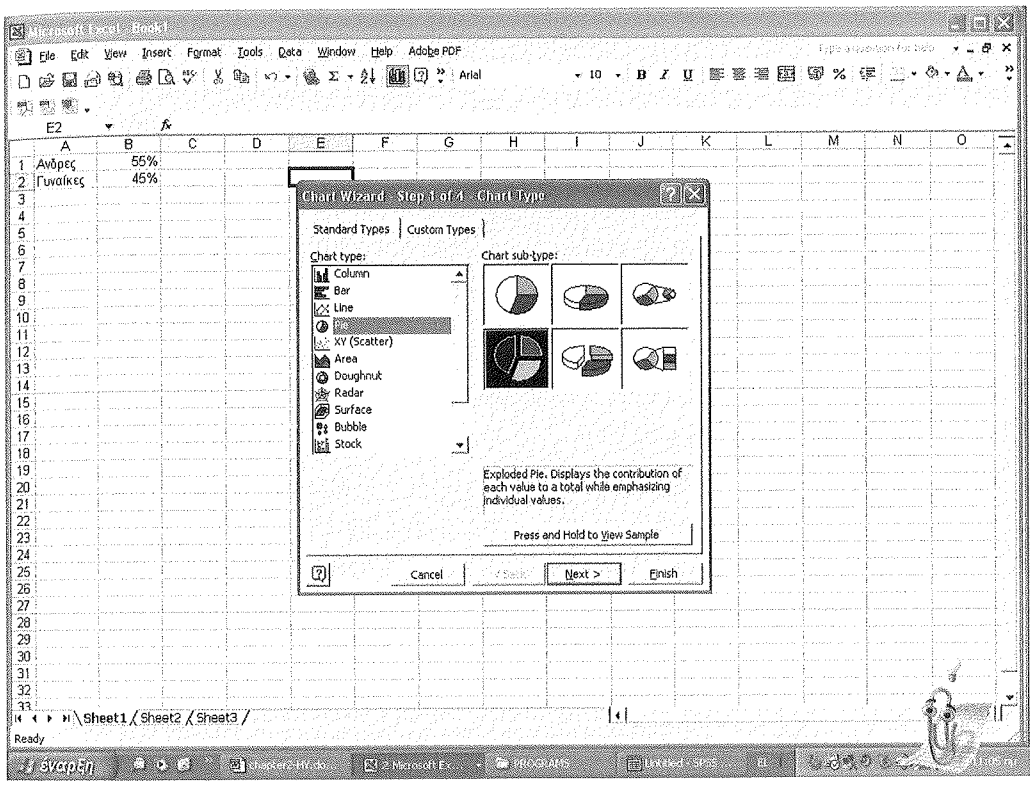
Σχήμα 2.16: Γραφική παράσταση παραδείγματος ενοικίων



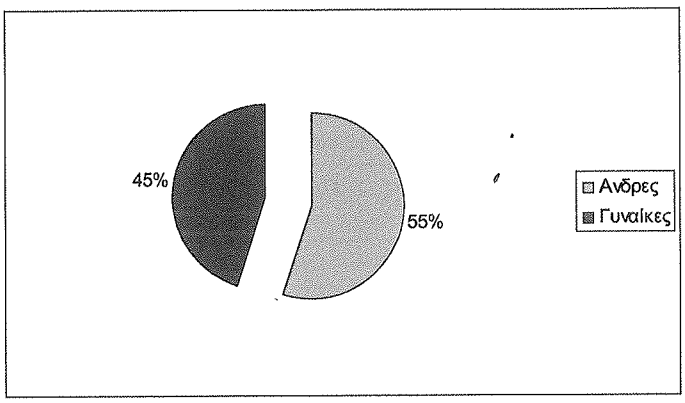
Αν θέλουμε να εξάγουμε ένα διάγραμμα πίτας τότε, αφού εισαγάγουμε τα δεδομένα επιλέγουμε τον "οδηγό γραφημάτων", πίτα (Pie) και πατάμε NEXT (Εικόνα 2.9).

Μετά το πάτημα του NEXT στο επόμενο παράθυρο επιλέγουμε εκ νέου NEXT και στην υποδοχή Chart Title μπορούμε να γράψουμε (αν επιθυμούμε) τον τίτλο του διαγράμματος. Πατώντας NEXT παίρνουμε την τελική μορφή του γραφήματος (Σχήμα 2.17) που αφορά την ποσοστιαία κατανομή ανδρών και γυναικών σε μια ερευνητική προσπάθεια.

Εικόνα 2.9: Εξαγωγή κυκλικών διαγραμμάτων (πίτες)



Σχήμα 2.17: Παράδειγμα κυκλικού διαγράμματος



Ασκήσεις

Λυμένες στο Κεφάλαιο 15

1. Για τις ακόλουθες μετρήσεις υπολογίστε τα βασικά στατιστικά μεγέθη και σχολιάστε τα αποτελέσματα.

5 7 9 11 13 17 22 72

2. α) Σχεδιάστε ένα ιστόγραμμα που να δείχνει το χρόνο αναμονής σε λεπτά που χρειάζεται ένας επιβάτης στο σιδηροδρομικό σταθμό της Ομόνοιας με προορισμό την Κηφισιά.

β) Υπολογίστε την τυπική απόκλιση και σχολιάστε το αποτέλεσμα, με έμφαση στο τι υποδηλώνει ο υπολογισμός της τυπικής απόκλισης.

Διάστημα λεπτών αναμονής	**Συχνότητα**
0 και κάτω από 5	5
5 και κάτω από 10	12
10 και κάτω από 15	5
15 και κάτω από 20	5
20 και κάτω από 25	4
25 και κάτω από 30	5

3. Μία χημική βιομηχανία πρέπει να επιλέξει τον τρόπο διαφήμισης του προϊόντος της ανάμεσα σε τρεις επιλογές: διαφήμιση στην τηλεόραση, σε εφημερίδα ή σε περιοδικό. Το τμήμα προώθησης του προϊόντος (marketing) έχει εκτιμήσει τις πωλήσεις και τις πιθανότητες σύμφωνα με τα εναλλακτικά σχέδια διαφήμισης ως ακολούθως:

Στρατηγική Α Τηλεοπτική διαφήμιση		**Στρατηγική Β Διαφήμιση σε εφημερίδα**		**Στρατηγική Γ Διαφήμιση σε Περιοδικό**	
Πωλήσεις	**Πιθανότητες**	**Πωλήσεις**	**Πιθανότητες**	**Πωλήσεις**	**Πιθανότητες**
4.000	0,2	4.000	0,25	3.000	0,3
6.000	0,6	5.000	0,5	4.000	0,4
8.000	0,2	6.000	0,25	6.000	0,3

Το περιθώριο κέρδους της επιχείρησης είναι το 50% των πωλήσεων.

(α) Υπολογίσατε το αναμενόμενο κέρδος κάθε στρατηγικής.

(β) Υπολογίστε την τυπική απόκλιση της κατανομής των κερδών για κάθε στρατηγική προώθησης του προϊόντος

(γ) Ποια από τις στρατηγικές αυτές είναι πιο επικίνδυνη και ποια θα επιλέξει η χημική βιομηχανία;

Προς επίλυση

4. Ποιο από τα παρακάτω σύνολα δεδομένων παρουσιάζει τη μικρότερη μεταβλητότητα;

A={3, 5, 12, 19, 26} ή B={2.110, 3.111, 5.923, 6.311, 11.771}

5. Αν μια πόλη το 1990 είχε 90.000 κατοίκους και το 2019 130.000, βρείτε τη μέση ετήσια αύξηση του πληθυσμού.

6. Ένας φοιτητής στο μάθημα της Στατιστικής πρέπει να συμπληρώσει τρεις εργασίες. Η δεύτερη εργασία έχει συντελεστή βαρύτητας τρεις φορές μεγαλύτερο από την πρώτη και η τρίτη τέσσερεις φορές μεγαλύτερο από την πρώτη. Ο φοιτητής βαθμολογήθηκε με 5,5 στην 1η, 6,0 στη 2η και 7,5 στην 3η. Υπολογίστε το σταθμικό μέσο.

7. Για τα παρακάτω ομαδοποιημένα δεδομένα, υπολογίστε το μέσο, την επικρατούσα τιμή, τον διάμεσο, την τυπική απόκλιση, τη διακύμανση, το 1ο και 3ο τεταρτημόριο, το εύρος και το ενδοτεταρτημοριακό εύρος.

Τάξεις	Συχνότητες
0-9	3
10-19	7
20-29	9
30-39	21
40-49	6
50-59	4

8. Ο παρακάτω πίνακας αφορά τον αριθμό τερμάτων που πέτυχαν οι ομάδες της Superleague την περίοδο 2018-2019.

Αρ. Τερμάτων	f_i	**Αθροιστικές Συχνότητες**	$f_i/\Sigma_i f_i$	**Αθροιστικές Σχετικές Συχνότητες**	$f_i X_{mi}$
20-29	4				
30-39		12			
40-49					135
50-59					
Total	16				

α) Συμπληρώστε τον πίνακα χρησιμοποιώντας τα στοιχεία που είναι συμπληρωμένα, καθώς και το γεγονός ότι μόλις το 25% των ομάδων σημείωσαν περισσότερα από 40 τέρματα.

β) Σχεδιάστε ένα ιστόγραμμα που να δείχνει τον αριθμό τερμάτων των ομάδων.

γ) Υπολογίστε το μέσο αριθμό τερμάτων ανά ομάδα, την τυπική απόκλιση και σχολιάστε τα αποτελέσματα.

9. Οι οκτώ εργαζόμενοι μιας επιχείρησης είχαν μέσο μισθό €1.000 το 2018. Το 2019 εξαιτίας μιας νέας πολιτικής της επιχείρησης αυξήθηκε ο μισθός τεσσάρων εργαζομένων κατά 50% με αποτέλεσμα όλοι οι υπάλληλοι να λαμβάνουν τους ίδιους μισθούς.

α) Υπολογίστε τους αρχικούς μισθούς των υπαλλήλων για το 2018.

β) Ποια ήταν η διακύμανση των μισθών στην επιχείρηση το 2018 και το 2019;

γ) Το 2020 η επιχείρηση προτίθεται να προσλάβει έναν διευθυντή και επιθυμεί ο νέος μέσος μισθός που θα διαμορφωθεί να ισούται με €1.250. Ποιος θα είναι ο μισθός του γενικού διευθυντή, δεδομένου ότι οι μισθοί των υπολοίπων εργαζομένων δεν θα μεταβληθούν σε σχέση με το 2019.

10. Από ένα τυχαίο δείγμα 25 φοιτητών παρουσιάζεται παρακάτω ο συνολικός αριθμός ωρών μελέτης που αφιέρωσε κάθε φοιτητής για μια προγραμματισμένη πρόοδο.

40	44	12	13	6
3	45	3	60	24
13	22	17	31	2
14	43	40	23	15

α) Υπολογίστε το δειγματικό μέσο και την τυπική απόκλιση.

β) Βρείτε το ποσοστό των παρατηρήσεων που περιέχονται στα διαστήματα $\overline{X} \pm s$ και $\overline{X} \pm 2s$.

11. Οι δυο πίνακες παρακάτω παρουσιάζουν τις πιθανότητες πιθανών κερδών από δυο επενδυτικά σχέδια Α και Β

Επενδυτικό Σχέδιο Α	
Κέρδος ($)	**Πιθανότητα**
2.000	0,3
3.000	0,5
5.000	0,2

Επενδυτικό Σχέδιο Β	
Κέρδος ($)	**Πιθανότητα**
2.000	0,2
2.500	0,3
4.000	0,3
8.000	0,1
12.000	0,1

Να υπολογισθούν για κάθε επενδυτικό σχέδιο τα αναμενόμενα κέρδη και η τυπική απόκλιση των κερδών. Ποιο από τα δυο επενδυτικά σχέδια θα επιλέγατε; Αιτιολογείστε την απάντησή σας.