

ΚΕΦΑΛΑΙΟ 3

Αριθμητικά Δεδομένα

3.1 Γενικά

Τα αριθμητικά δεδομένα προκύπτουν από την παρατήρηση ή καταγραφή μιας ποσοτικής μεταβλητής. Διακρίνουμε δύο περιπτώσεις. Πρώτον, η μεταβλητή μπορεί να είναι είτε συνεχής είτε διακριτή (δηλαδή μπορεί να πάρει λίγες αριθμητικές τιμές) οπότε στα δεδομένα θα υπάρχουν επαναλαμβανόμενες τιμές και δεύτερον, η μεταβλητή είναι συνεχής.

Η περιγραφική στατιστική στις δύο περιπτώσεις διαφοροποιείται κυρίως ως προς τον τρόπο που οργανώνονται τα δεδομένα σε μία κατανομή και ως προς τις γραφικές παραστάσεις που είναι κατάλληλες σε κάθε περίπτωση. Υπενθυμίζεται ότι στη στατιστική πρακτική, οι διακριτές μεταβλητές οι οποίες μπορούν να πάρουν μεγάλο πλήθος τιμών αντιμετωπίζονται ως συνεχείς (βλέπε και § 1.5).

3.2 Πίνακες και Γραφήματα για Επαναλαμβανόμενες Τιμές

Έστω ότι τα δεδομένα αποτελούνται από n μετρήσεις μιας ποσοτικής μεταβλητής η οποία είναι διακριτή και οι δυνατές τιμές της είναι λιγότερες από n . Είναι προφανές ότι κάποιες ή όλες οι τιμές της επαναλαμβάνονται. Ο πιο φυσικός τρόπος να οργανωθούν οι n παρατηρήσεις είναι ένας **πίνακας συχνοτήτων** (frequency table) στον οποίον κάθε τιμή επαναλαμβάνεται f_i φορές. Ο πίνακας ονομάζεται εναλλακτικά και **κατανομή συχνοτήτων** (frequency distribution).

Είναι προφανές ότι ισχύει: $\sum_i f_i = n$.

Εναλλακτικά, μπορούμε να εκφράσουμε τις συχνότητες ως ποσοστά επί του πλήθους των παρατηρήσεων οπότε προκύπτει η **κατανομή σχετικών συχνοτήτων** (relative frequency distribution) ή **επί τοις εκατό ποσοστιαία κατανομή** (percent distribution) η οποία προσφέρεται για συγκρίσεις.

Είναι προφανές ότι ισχύει: $\sum_i f_i/n = 1$.

παράδειγμα 3.1

Για τους 20 πρώτους εισαχθέντες στο Τμήμα Οικονομικών Επιστημών του ΑΠΘ, καταγράψαμε τις συμμετοχές τους στις Πανελλήνιες εξετάσεις και πήραμε τις εξής τιμές:

1	2	1	1	2	3	1	1	1	2
1	2	2	1	1	1	1	1	2	4

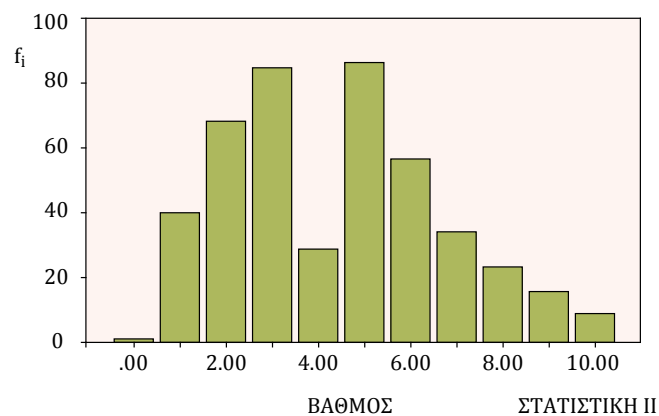
Τα δεδομένα οργανώνονται στην ακόλουθη κατανομή συχνοτήτων και σχετικών συχνοτήτων, αντίστοιχα.

Συμμετοχές x_i	Συχνότητα f_i	Σχετική συχνότητα f_i/n
1	12	0.60
2	6	0.30
3	1	0.05
4	1	0.05
Άθροισμα	20	1.00

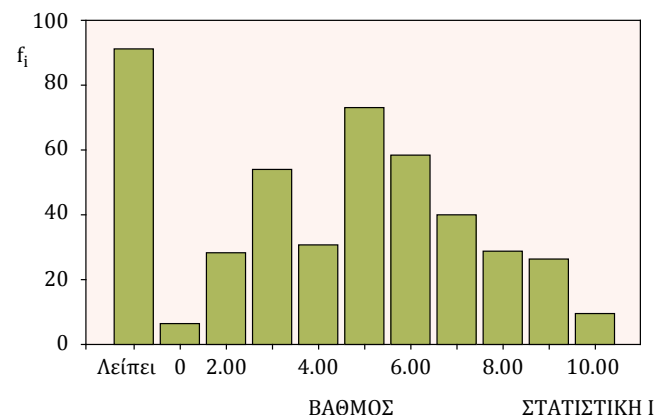
Είναι εύκολο να διαπιστωθεί ότι στο παράδειγμα αυτό, η μεταβολή μιας συχνότητας κατά 1 θα έχει ως αποτέλεσμα τη μεταβολή της σχετικής συχνότητας κατά 0.05.

παράδειγμα 3.2

Η βαθμολογία των φοιτητών του Οικονομικού Τμήματος του ΑΠΘ στο μάθημα Στατιστική ΙΙ, τον Ιούνιο 2006 δίνεται στο ακόλουθο γράφημα



Στις ίδιες εξετάσεις ζητήθηκε από τους εξεταζόμενους να καταγράψουν τον βαθμό τους στη Στατιστική Ι και τα αποτελέσματα δίνονται στο ακόλουθο γράφημα



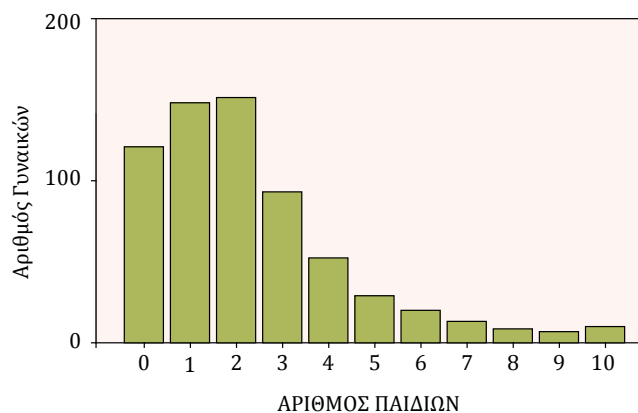
Επισημαίνεται ότι περίπου 90 άτομα δεν έδωσαν τον βαθμό τους στη Στατιστική Ι (είτε δεν τον θυμόταν είτε προσήλθαν στις εξετάσεις στη Στατιστική ΙΙ χωρίς να έχουν εξεταστεί στην Στατιστική Ι). ◀

παράδειγμα 3.3

Δίνεται η κατανομή συχνοτήτων του αριθμού των παιδιών που γέννησαν 653 γυναίκες οι οποίες παντρεύτηκαν μετά τα 30 τους χρόνια και στη συνέχεια το αντίστοιχο ραβδόγραμμα συχνοτήτων.

Αριθμός παιδιών x_i	Συχνότητα f_i	Σχετική συχνότητα f_i/n
0	123	0.18836
1	152	0.23277
2	154	0.23584
3	96	0.14701
4	53	0.08116
5	30	0.04594
6	17	0.02603
7	10	0.01531
8	6	0.00919
9	5	0.00766
10	7	0.01072
Άθροισμα	653	$\simeq 1.00000$

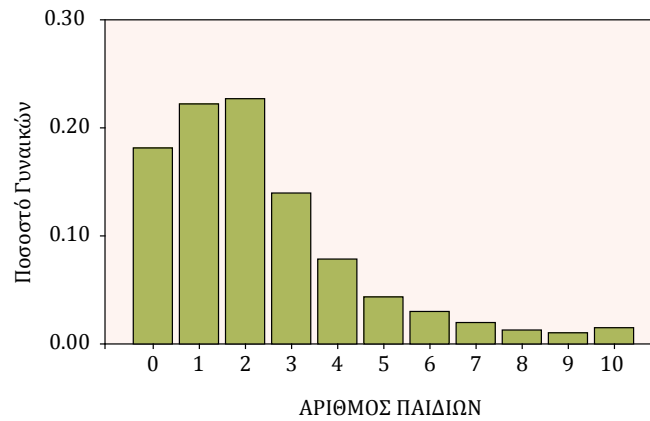
Μια τέτοια κατανομή μπορεί να δοθεί γραφικά με το ραβδόγραμμα.¹



Την ίδια ακριβώς εικόνα θα πάρουμε αν στον κάθετο άξονα σημειώσουμε τις σχετικές συχνότητες (όπως φαίνεται στο σχήμα που δίνεται στην επόμενη σελίδα).

ΣΗΜΕΙΩΣΗ Η κατανομή της βαθμολογίας, γενικά, θεωρείται εντάξει όταν είναι συμμετρική. Δεξιά λοξότητα είναι ένδειξη ότι τα θέματα ήταν δύσκολα ενώ η αριστερή λοξότητα ότι τα θέματα ήταν πολύ εύκολα. Το συγκεκριμένο γράφημα δίνει ενδείξεις για θέματα μάλλον δύσκολα και επιπλέον αντανάκλα την τάση του διδάσκοντα να “σπρώχνει” το μη δημοφιλές 4 προς το 3 ή στο 5!

¹Να σημειωθεί ότι για το SPSS το ραβδόγραμμα δεν διαφοροποιείται είτε οι παρατηρήσεις είναι ποσοτικές αλλά διακριτές και επαναλαμβανόμενες είτε είναι κατηγορικές.



3.2.1 Η αθροιστική και η αφαιρετική κατανομή συχνοτήτων

Η αθροιστική συχνότητα που αντιστοιχεί στην τιμή x_i συμβολίζεται με F_i και είναι το πλήθος των παρατηρήσεων που έχουν τιμή μέχρι και x_i . Ο τρόπος με τον οποίο προσδιορίζεται είναι ανάλογος με αυτόν που είδαμε στα σταδιακά δεδομένα (βλέπε ενότητα 2.3). Ειδικότερα, θέτοντας ως αρχική συνθήκη $F_0 = 0$, και επειδή ισχύει η αναδρομική σχέση

$$F_i = F_{i-1} + f_i,$$

έχουμε

$$F_1 = F_0 + f_1 = f_1$$

$$F_2 = F_1 + f_2 = f_1 + f_2$$

$$F_3 = F_2 + f_3 = f_1 + f_2 + f_3$$

κ.ο.κ.

Ο πίνακας που σε κάθε τιμή x_i αντιστοιχεί την αθροιστική της συχνότητα ονομάζεται **πίνακας ή κατανομή αθροιστικών συχνοτήτων** (cumulative or “less than” distribution). Αν μας ενδιαφέρει το πλήθος των παρατηρήσεων οι οποίες έχουν τιμή μεγαλύτερη ή ίση με x_i τότε αρκεί από το άθροισμα των παρατηρήσεων να αφαιρέσουμε τις συχνότητες που αντιστοιχούν στις μικρότερες της x_i τιμές. Έτσι προκύπτει η αφαιρετική συχνότητα και, αντίστοιχα, η **κατανομή αφαιρετικών συχνοτήτων** (“more than” distribution). Οι αφαιρετικές συχνότητες συμβολίζονται με F'_i .

Εκφράζοντας τις αθροιστικές ή τις αφαιρετικές συχνότητες ως ποσοστά επί του συνόλου των παρατηρήσεων προκύπτει η κατανομή των σχετικών αθροιστικών συχνοτήτων και, αντίστοιχα, η κατανομή των σχετικών αφαιρετικών συχνοτήτων.

παράδειγμα 3.4

Για την κατανομή συχνοτήτων του παραδείγματος 3.1 οι αθροιστικές και οι αφαιρετικές συχνότητες υπολογίζονται στον ακόλουθο πίνακα.

x_i	f_i	F_i	F'_i	F_i/n	F'_i/n
1	12	12	20	0.60	1.00
2	6	18	8	0.90	0.40
3	1	19	2	0.95	0.10
4	1	20	1	1.00	0.05

Έτσι π.χ. μέχρι και 2 φορές έδωσαν 18 άτομα ή ποσοστό 90% ενώ 2 και περισσότερες φορές έδωσαν 8 άτομα δηλαδή ποσοστό 40%.

3.3 Ομαδοποίηση

Όταν στις παρατηρήσεις υπάρχουν πολλές διαφορετικές τιμές, τα δεδομένα ομαδοποιούνται σε **τάξεις** (classes or bins) τιμών και παρουσιάζονται με τη μορφή μιας **κατανομής συχνοτήτων** (frequency distribution). Ειδικότερα, έστω θέλουμε να ομαδοποιήσουμε τις παρατηρήσεις $x_1, x_2, \dots, x_n$ μιας συνεχούς μεταβλητής X (η X μπορεί ακόμη να είναι διακριτή με μεγάλο πλήθος τιμών οπότε, όπως ειπώθηκε στο προηγούμενο κεφάλαιο, αντιμετωπίζεται ως συνεχής). Για το σκοπό αυτό χωρίζουμε το **εύρος των δεδομένων** (data range), δηλαδή τη διαφορά

$$R = x_{\text{μέγιστη}} - x_{\text{ελάχιστη}}$$

σε ίσα διαδοχικά διαστήματα ή τάξεις και καταγράφουμε τη συχνότητα f_i των παρατηρήσεων που ανήκουν σε κάθε τάξη. Πρέπει επομένως να επιλέξουμε το πλήθος, το εύρος και τα όρια των τάξεων.

ΠΛΗΘΟΣ ΤΑΞΕΩΝ. Επιλέγεται παίρνοντας υπόψη το πλήθος των παρατηρήσεων καθώς και τον σκοπό για τον οποίο προορίζεται η κατανομή. Οι ακόλουθες πρακτικές παρατηρήσεις μπορεί να φανούν χρήσιμες:

- Όταν τα δεδομένα είναι λιγότερα από 100 επιλέγουμε συνήθως μέχρι 10 τάξεις. Από 100 ως 1000 συνήθως αρκούν 10-12 τάξεις. Από 1000 ως 10000 αρκούν 13-16 τάξεις.
- Αν υπάρχουν πολλές μηδενικές συχνότητες συνιστάται η μείωση του αριθμού των τάξεων.
- Αν η κατανομή είναι πολύ ανώμαλη δοκιμάζουμε να αυξήσουμε τον αριθμό των τάξεων.
- Όταν σκοπός της ομαδοποίησης των δεδομένων είναι απλώς η περιληπτική τους παρουσίαση, το πλήθος των τάξεων δεν πρέπει να ξεπερνά το 20.

ΤΑΞΙΚΟ ΕΥΡΟΣ δηλαδή η διαφορά της κατώτερης από την ανώτερη τιμή. Συμβολίζεται με c και προσδιορίζεται ως εξής: διαιρούμε το εύρος R με τον αριθμό των τάξεων k και στρογγυλοποιούμε το αποτέλεσμα έτσι ώστε να προκύπτει ένας αριθμός εύχρηστος και οικείος στον μέσο αναγνώστη. Τέτοιοι είναι π.χ. οι αριθμοί

0.5, 0.10, 2, 4, 5, 10, 15, 20, 25 κλπ.

ΟΡΙΑ ΤΑΞΕΩΝ δηλαδή την κατώτερη και την ανώτερη τιμή για κάθε τάξη. Επιλέγοντας τα όρια αυτά πρέπει να έχουμε υπόψη ότι η κεντρική τιμή κάθε τάξης πρόκειται να χρησιμοποιηθεί σε περαιτέρω υπολογισμούς, οπότε κεντρικές τιμές όπως 5, 10, 15, ... είναι προτιμότερες από τις 7.2, 14.4, 21.6.

ΚΑΤΩΤΕΡΗ ΤΙΜΗ ή **ΑΓΚΥΡΑ** δηλαδή το κάτω όριο της πρώτης τάξης, το οποίο, όπως θα δούμε στα παραδείγματα που ακολουθούν, μπορεί να επηρεάσει σημαντικά την εικόνα που δίνει η κατανομή και η γραφική της παράσταση.

Περισσότερα για την ομαδοποίηση

❖ Ανεξάρτητα από την ακρίβεια των μετρήσεων από τις οποίες προέκυψαν τα δεδομένα, συνηθίζεται να ορίζουμε την ίδια τιμή για το ανώτερο όριο μιας τάξης και το κατώτερο της επόμενης. Αν έχουμε μια παρατήρηση η οποία έχει αυτή την κοινή τιμή, συμβατικά θεωρούμε ότι ανήκει στην επόμενη τάξη –σύμβαση πάντως που δεν ακολουθείται από όλους. Έτσι π.χ. αν έχουμε τις τάξεις 200-300 και 300-400 και στα δεδομένα υπάρχει η τιμή $x = 300$ θα την θεωρήσουμε ότι ανήκει στο ταξικό διάστημα 300-400. Εξάλλου, επειδή σε περαιτέρω υπολογισμούς κάνουμε την υπόθεση ότι οι παρατηρήσεις κατανέμονται ομοιόμορφα σε κάθε τάξη, φροντίζουμε ώστε η κατώτερη τιμή της πρώτης τάξης να είναι μικρότερη από την ελάχιστη τιμή των δεδομένων και η ανώτερη τιμή της τελευταίας μεγαλύτερη

από την μέγιστη τιμή των δεδομένων. Αν χρειαστεί, αυξάνουμε το πλήθος των τάξεων ή (και) τα όριά τους.

❖ Όταν σκοπός της ομαδοποίησης των δεδομένων είναι η περαιτέρω στατιστική τους ανάλυση είναι χρήσιμο να δοκιμάζονται διαφορετικές ομαδοποιήσεις μεταβάλλοντας το πλήθος και συνακόλουθα και το εύρος των τάξεων. Αυτό ισχύει ιδιαίτερα αν το γράφημα της κατανομής δίνει εικόνα για κάτι ιδιαίτερο στα δεδομένα όπως π.χ. δύο κορυφές ή απρόσμενη λοξότητα. Όπως ειδικότερα θα δούμε λίγο πιο κάτω, είναι δυνατόν σε μια ομαδοποίηση με περισσότερες τάξεις να υπάρξει μόνο μία κορυφή που επιμένει ή αντίστροφα, να αποκαλύψει μια δεύτερη κορυφή.

παράδειγμα 3.5

Δίνονται τα αποτελέσματα της μέτρησης του βάρους γεννήσεως 60 βρεφών σε γραμμάρια

1950	2090	2700	3350	4200	3720	4400	2980	3850	4550
3050	2350	1850	2820	3670	2950	3750	1850	2420	3150
3000	3470	3920	3100	2400	2900	2650	3450	3650	4020
4450	3120	3660	3070	3550	2020	3500	2500	3780	3940
3540	2800	2850	4450	1950	3020	2800	3500	1480	4495
2850	3100	2250	3300	4100	3220	3600	2130	4020	4075

Στα δεδομένα αυτά η μέγιστη τιμή είναι η 4550 και η ελάχιστη η 1480, οπότε το εύρος τους είναι το $R = 4550 - 1480 = 3070$. Επειδή το πλήθος των παρατηρήσεων είναι μικρότερο από 100 θα προτιμήσουμε να τις ομαδοποιήσουμε σε 6, 7 ή 8 τάξεις. Διαιρώντας το εύρος R δια τον καθέναν από τους αριθμούς αυτούς παίρνουμε:

$$3070 : 6 = 511.7$$

$$3070 : 7 = 438.6$$

$$3070 : 8 = 383.8$$

Στρογγυλοποιώντας σε αριθμούς εύχρηστους και οικείους επιλέγουμε ως ταξικό εύρος το 400 ή το 500. Για ταξικό εύρος 400 θα πάρουμε μια από τις ακόλουθες κατανομές ανάλογα με την άγκυρα, η οποία επιλέγεται.

Αν επιλέξουμε ως άγκυρα τον αριθμό 1400 προκύπτει η ακόλουθη κατανομή με 8 τάξεις.

Τάξεις βάρους	Αριθμός βρεφών f_i
1400-1800	2
1800-2200	6
2200-2600	5
2600-3000	9
3000-3400	11
3400-3800	13
3800-4200	8
4200-4600	6
Άθροισμα	60

Αν επιλέξουμε ως άγκυρα τον αριθμό 1200 προκύπτει η ακόλουθη κατανομή η οποία όμως έχει 9 τάξεις.

Τάξεις βάρους	Αριθμός βρεφών f_i
1200-1600	1
1600-2000	4
2000-2400	5
2400-2800	5
2800-3200	16
3200-3600	9
3600-4000	10
4000-4400	5
4400-4800	6
Άθροισμα	60

Μπορούμε ακόμη να επιλέξουμε ως ταξικό εύρος το 500. Στην περίπτωση αυτή όμως είναι λογικό να ορίσουμε ως όρια των τάξεων αριθμούς πολλαπλασίους του 500 οπότε επιλέγουμε ως άγκυρα το 1000 και προκύπτει η ακόλουθη κατανομή:

Τάξεις βάρους	Αριθμός βρεφών f_i
1000-1500	1
1500-2000	4
2000-2500	7
2500-3000	11
3000-3500	14
3500-4000	13
4000-4500	9
4500-5000	1
Άθροισμα	60

Οι συχνότητες συχνά εκφράζονται ως ποσοστά επί του συνόλου των παρατηρήσεων και η αντίστοιχη κατανομή ονομάζεται η κατανομή σχετικών συχνοτήτων ή ποσοστιαία.

3.4 Κατανομή Συχνοτήτων με Ανοιχτές τις Ακραίες Τάξεις

Όταν οι περισσότερες παρατηρήσεις είναι συγκεντρωμένες σε ένα διάστημα τιμών και υπάρχουν πολύ λίγες ακραίες τιμές έξω από το διάστημα αυτό, μπορούμε να χρησιμοποιήσουμε διαστήματα με ανοιχτή τη μια άκρη για την πρώτη ή (και) την τελευταία τάξη. Ένα τέτοιο διάστημα θα ήταν της μορφής “κάτω από x ” για την πρώτη τάξη και “πάνω από y ” για την τελευταία τάξη.

Τα ανοιχτά διαστήματα πολλές φορές χρησιμοποιούνται για να εξασφαλίσουν την εμπιστευτικότητα της πληροφορίας. Έτσι, όταν είναι γνωστή η ταυτότητα των λίγων επιχειρήσεων ή ατόμων με υψηλό εισόδημα, ο προσδιορισμός του ανώτερου ορίου της τελευταίας τάξης εισοδήματος μπορεί να θεωρηθεί πολύ αποκαλυπτικός. Ο υπολογισμός ορισμένων συναρτήσεων των παρατηρήσεων πάντως δεν μπορεί να γίνει από μια κατανομή η οποία έχει ανοιχτές τάξεις. Θα πρέπει για τον σκοπό αυτό να διαθέτουμε τα αρχικά δεδομένα.

παράδειγμα 3.6

Κατανομή συχνοτήτων με ανοιχτές και τις δύο ακραίες τάξεις είναι η ακόλουθη:

Ηλικία της μητέρας σε έτη	Αριθμός γεννήσεων f_i
κάτω των 15	187
15-20	18.446
20-25	54.118
25-30	42.981
30-35	22.033
35-40	7.697
40-45	2.287
45-50	270
50 και άνω	41
Άθροισμα	148.060

Πηγή: ΕΣΥΕ, Στατιστική Επετηρίς της Ελλάδος, 1981

Πίνακας 3.1 Κατανομή των γεννήσεων της Ελλάδας το 1980 κατά ηλικία της μητέρας

3.5 Το Ιστόγραμμα και το Πολύγωνο Συχνοτήτων

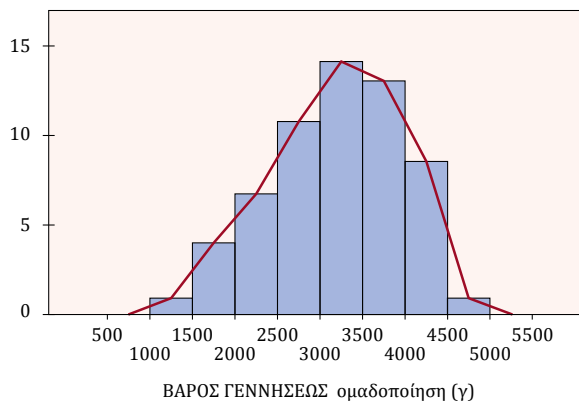
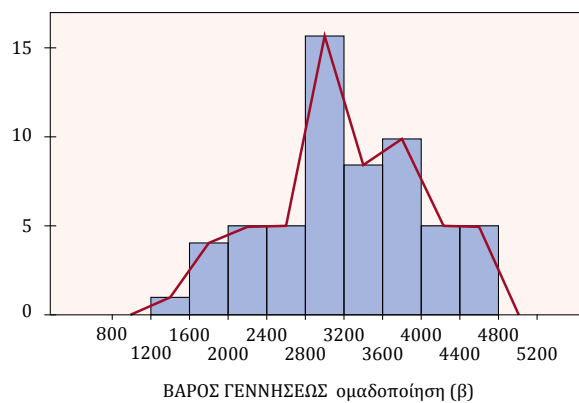
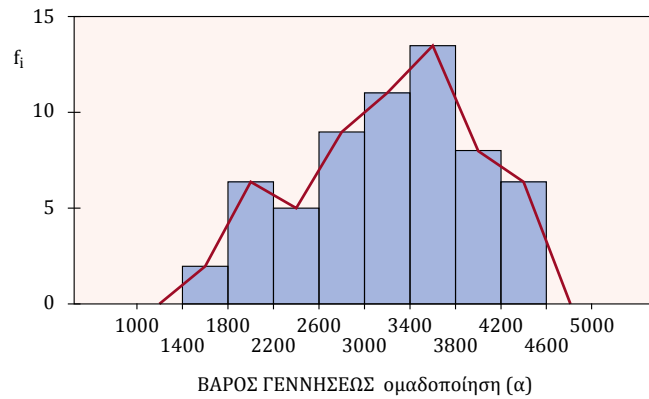
Η γραφική παρουσίαση δεδομένων τα οποία είναι ομαδοποιημένα σε μια κατανομή συχνοτήτων γίνεται με το ιστόγραμμα συχνοτήτων (frequency histogramm) ή απλώς ιστόγραμμα ή (και) την πολυγωνική γραμμή ή πολύγωνο συχνοτήτων (frequency polygon).

ΙΣΤΟΓΡΑΜΜΑ. Είναι ίσως η πιο σημαντική γραφική παράσταση των δεδομένων και όπως θα δούμε σε επόμενα μέρη μπορεί να παίξει σημαντικό ρόλο στη στατιστική συμπερασματολογία. Αποτελεί μια ακολουθία ορθογώνιων παραλληλογράμμων, όσα και οι τάξεις της κατανομής. Κάθε παραλληλόγραμμο έχει πλάτος ίσο με το εύρος και ύψος ίσο με τη συχνότητα της. Αν οι συχνότητες είναι σχετικές, τότε το ιστόγραμμα αναφέρεται ως ιστόγραμμα σχετικών συχνοτήτων.

ΠΟΛΥΓΩΝΙΚΗ ΓΡΑΜΜΗ. Αποτελείται από ένα σύνολο ευθύγραμμων τμημάτων, κάθε ένα από τα οποία ενώνει τα μέσα των επάνω πλευρών δύο διαδοχικών παραλληλογράμμων του ιστογράμματος. Συμβατικά, η πολυγωνική γραμμή επεκτείνεται από τα δύο άκρα στον οριζόντιο άξονα και σε απόσταση μισού ταξικού εύρους. Έτσι, η επιφάνεια που περικλείει, ισούται με την επιφάνεια του ιστογράμματος.

παράδειγμα 3.7

Στα σχήματα (α)-(γ) που ακολουθούν δίνεται το ιστόγραμμα και η πολυγωνική γραμμή για τις 3 παραλλαγές της κατανομής του βάρους γεννήσεως 60 βρεφών (παράδειγμα 3.5). Αν στον κάθετο άξονα μετρήσουμε τις σχετικές συχνότητες τότε έχουμε, αντίστοιχα, το ιστόγραμμα και το πολύγωνο σχετικών συχνοτήτων. Το ιστόγραμμα σχετικών συχνοτήτων δίνει την ίδια εικόνα με το ιστόγραμμα συχνοτήτων, επιπλέον όμως επιτρέπει τις συγκρίσεις κατανομών με διαφορετικό πλήθος παρατηρήσεων. Παρατηρούμε ότι παρόλο που στην τελευταία ομαδοποίηση αντιστοιχεί ένα ομαλό και μάλλον συμμετρικό πολύγωνο συχνοτήτων, οι τρεις γραφικές παραστάσεις δίνουν χονδρικά παρόμοια εικόνα για τα δεδομένα.



3.5.1 Το ιστόγραμμα μιας κατανομής με άνισα ταξικά διαστήματα

Είδαμε ότι το ιστόγραμμα συχνοτήτων μας δίνει σημαντικές οπτικές πληροφορίες για την κατανομή των δεδομένων. Όταν όλες οι τάξεις έχουν το ίδιο εύρος, το ύψος των παραλληλογράμμων επιλέγεται ίσο με τις αντίστοιχες συχνότητες οπότε οι διαφορές στις συχνότητες απεικονίζονται με ανάλογες διαφορές στα εμβαδά.

Όταν η κατανομή έχει άνισες τάξεις, το ύψος των παραλληλογράμμων θα πρέπει να τροποποιηθεί έτσι ώστε και πάλι τα εμβαδά των επιφανειών που περικλείουν να είναι ανάλογα με τις συχνότητες. Έτσι π.χ. αν μια τάξη είναι διπλάσια από τις υπόλοιπες, τότε το ύψος του παραλληλογράμμου της θα πρέπει να ισούται με το μισό της συχνότητάς της. Γενικά, για να κατασκευάσουμε το ιστόγραμμα μιας

κατανομής με άνισες τάξεις υπολογίζουμε τις διορθωμένες συχνότητες ως εξής:

- ο διαλέγουμε την τάξη της οποίας το εύρος θέλουμε να πάρουμε ως μονάδα,
- ο διαιρούμε το εύρος της κάθε τάξης δια του εύρους-μονάδα,
- ο θέτουμε ως διορθωμένη συχνότητα το πηλίκο της διαίρεσης της αρχικής συχνότητας δια του αντίστοιχου αριθμού που προέκυψε από το βήμα 2.

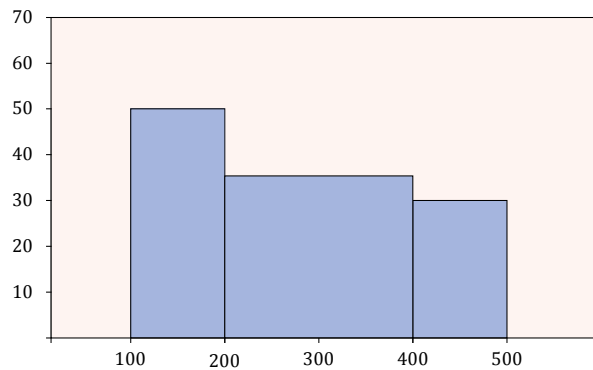
Είναι προφανές ότι για τη γραφική παράσταση των αθροιστικών ή αφαιρετικών συχνοτήτων δε χρειάζονται διορθώσεις.

παράδειγμα 3.8

Δίνεται μια υποθετική κατανομή συχνοτήτων με τρία ταξικά διαστήματα. Επιλέγουμε ως μονάδα το ταξικό εύρος 100. Τότε, επειδή η δεύτερη τάξη έχει διπλάσιο εύρος από το εύρος-μονάδα, θα διαιρέσουμε τη συχνότητά της με 2.

Τάξεις	Εύρος τάξεως	Συχνότητα	Διορθωμένη συχνότητα
100-200	100	50	$\left(\frac{100}{100} = 1\right) 50 : 1 = 50$
200-400	200	70	$\left(\frac{200}{100} = 2\right) 70 : 2 = 35$
400-500	100	32	$\left(\frac{100}{100} = 1\right) 32 : 1 = 32$

Το ιστόγραμμα της κατανομής αυτής δίνεται στο σχήμα που ακολουθεί:



3.5.2 Ιστόγραμμα και πολύγωνο συχνοτήτων που περικλείουν επιφάνεια ίση με τη μονάδα

Η επιφάνεια E_i που περικλείεται από το παραλληλόγραμμο i ενός ιστογράμματος συχνοτήτων ισούται με

$$E_i = c_i f_i \quad i = 1, \dots, \kappa$$

όπου c_i το ταξικό εύρος και f_i η αντίστοιχη συχνότητα. Όταν όλες οι τάξεις είναι ίσες δηλαδή $c_i = c$ τότε, τα κ παραλληλόγραμμα του ιστογράμματος περικλείουν επιφάνεια ίση με

$$E = \sum_{i=1}^{\kappa} c f_i = c \sum_{i=1}^{\kappa} f_i = c \cdot n$$

Σε επόμενα κεφάλαια θα δούμε ότι σε ορισμένες περιπτώσεις το ιστόγραμμα συχνοτήτων χρησιμοποιείται για να προσεγγίσει το γράφημα μιας κατανομής πιθανοτήτων και επομένως θα πρέπει να περικλείει συνολική επιφάνεια ίση με 1. Για τον σκοπό αυτό αρκεί το ύψος του κάθε παραλληλόγραμμου i να ισούται με την σχετική συχνότητα ανά ταξικό εύρος δηλαδή

$$d_i = \frac{f_i}{nc_i}, \quad i = 1, \dots, \kappa$$

Η d_i ($i = 1, \dots, \kappa$) ονομάζεται **πυκνότητα** (density) και το αντίστοιχο ιστόγραμμα περικλείει επιφάνεια ίση με 1. Πράγματι έχουμε

$$E_i = c \frac{f_i}{nc} = \frac{f_i}{n}, \quad i = 1, \dots, \kappa$$

και

$$E' = \sum_{i=1}^{\kappa} E_i = \sum_{i=1}^{\kappa} \frac{f_i}{n} = \frac{1}{n} \sum_{i=1}^{\kappa} f_i = \frac{n}{n} = 1$$

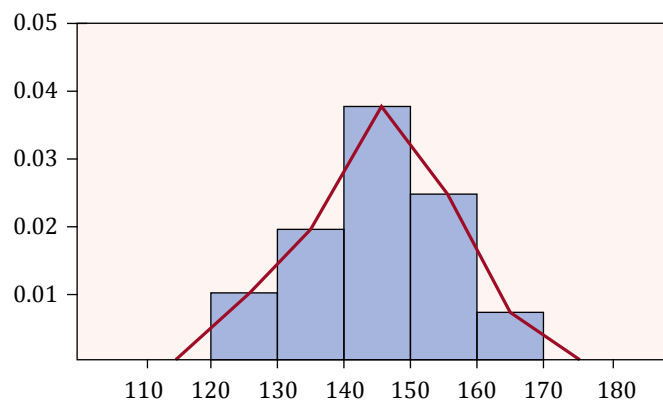
Το ιστόγραμμα των πυκνοτήτων προσφέρεται για συγκρίσεις στις κατανομές με άνισα ταξικά διαστήματα ή/και διαφορετικό πλήθος παρατηρήσεων.

παράδειγμα 3.9

Στον πίνακα που ακολουθεί δίνεται η κατανομή συχνοτήτων του αριθμού των ωρών εργασίας που απαιτήθηκαν σε κάθε μία από τις 100 μονάδες του X προϊόντος. Στην τελευταία στήλη δίνονται οι σχετικές συχνότητες ανά μονάδα ταξικού εύρους.

X	f_i	f_i/n	f_i/nc
100-130	10	0.10	0.010
130-140	20	0.20	0.020
140-150	38	0.38	0.038
150-160	25	0.25	0.025
160-170	7	0.07	0.007
Άθροισμα	100	1.00	0.100

Στο σχήμα που ακολουθεί δίνεται το ιστόγραμμα πυκνότητας και η αντίστοιχη πολυγωνική γραμμή.



3.6 Ερμηνεύοντας το Ιστογράμμα

Σχεδόν πάντα το ιστογράμμα αναφέρεται σε ένα σύνολο δειγματικών δεδομένων τα οποία μελετούμε με σκοπό την εξαγωγή συμπερασμάτων για τον πληθυσμό. Έτσι, από την κατ' αρχήν οπτική εξέταση του ιστογράμματος των δειγματικών τιμών, μπορούμε να ανιχνεύσουμε τα χαρακτηριστικά της άγνωστης κατανομής πληθυσμού. Όπως ειδικότερα θα δούμε και σε επόμενα μέρη, σε ένα τέτοιο πλαίσιο, η κατανομή συχνοτήτων των δειγματικών παρατηρήσεων ονομάζεται **εμπειρική** κατανομή σε αντιδιαστολή με την κατανομή πληθυσμού η οποία ονομάζεται **θεωρητική**. Από το ιστογράμμα μπορούμε να αντλήσουμε σημαντικές πληροφορίες για την κατανομή των δεδομένων και ειδικότερα αν

- είναι συμμετρική ή λοξή,
- έχει μια ή περισσότερες κορυφές,
- υπάρχουν ασυνήθιστα σημεία.

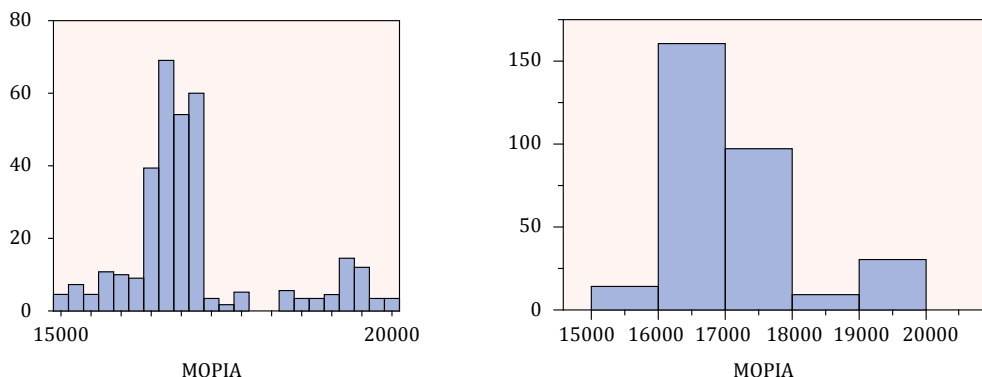
Η ερμηνεία του ιστογράμματος θα πρέπει να γίνεται με προσοχή και σε συνδυασμό με το φυλλογράφημα (που παρουσιάζεται αμέσως στη συνέχεια) αλλά και το θηκόγραμμα (που παρουσιάζεται στην ενότητα 4.3). Στη συνέχεια επισημαίνονται ειδικότερα ζητήματα τα οποία μπορεί να ανακύψουν κατά την "ανάγνωση" του ιστογράμματος

3.6.1 Ιστογράμμα και ομαδοποίηση

Η εικόνα που δίνει το ιστογράμμα μπορεί να μεταβάλλεται σημαντικά με το πλήθος των τάξεων, άρα και με το εύρος τους, καθώς και την αρχική τιμή ή άγκυρα. Ήδη στα τρία ιστογράμματα των δεδομένων του παραδείγματος του βάρους γεννήσεως βρεφών παρατηρούμε κάποιες διαφορές παρόλο που το πλήθος των τάξεων δε μεταβάλλεται σημαντικά. Οι διαφορές αυτές μπορεί να γίνουν σημαντικές, όταν π.χ. διπλασιάζεται το πλήθος των τάξεων.

παράδειγμα 3.10

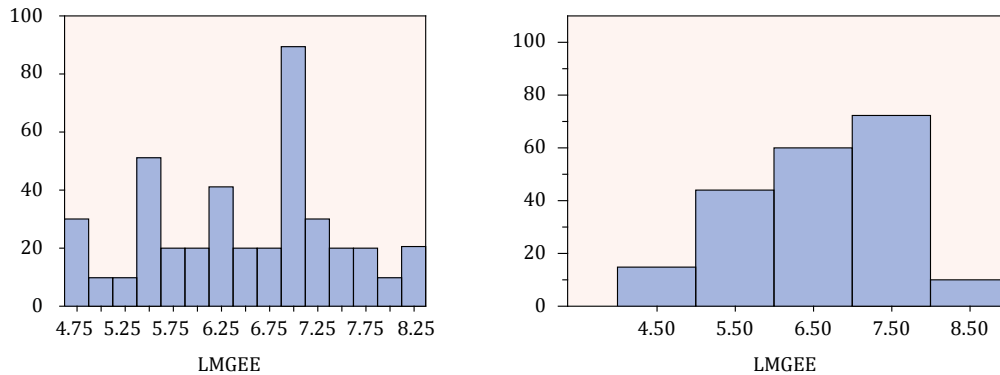
Τα ακόλουθα δύο ιστογράμματα αποτελούν γραφική παράσταση του ίδιου συνόλου δεδομένων και έγιναν από τα προγράμματα SPSS και S-plus, αντίστοιχα, με αυτόματη επιλογή (by default).



Ομοίως, τα ακόλουθα δύο ιστογράμματα (επόμενη σελίδα) αναφέρονται σε ίδιο σύνολο δεδομένων και έγιναν με τα προγράμματα SPSS και S-plus, αντίστοιχα.



Πολλά προγράμματα Η/Υ δίνουν τη δυνατότητα στο χρήστη να ζητήσει πολλαπλά ιστογράμματα μεταβάλλοντας το πλήθος των τάξεων ή (και) την άγκυρα. Αν η εμφάνιση του ιστογράμματος δε μεταβάλλεται σημαντικά όταν μεταβάλλεται το πλήθος των τάξεων σε ένα λογικά μεγάλο εύρος τότε ο



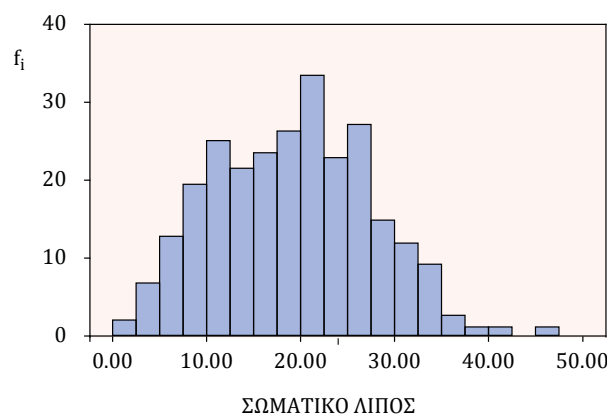
αναλυτής μπορεί να εμπιστευτεί την εικόνα που παίρνει από τη γραφική παράσταση των δεδομένων του (βλ. και Chatterjee-Handcock-Simonoff, 1995).

3.6.2 Συμμετρία

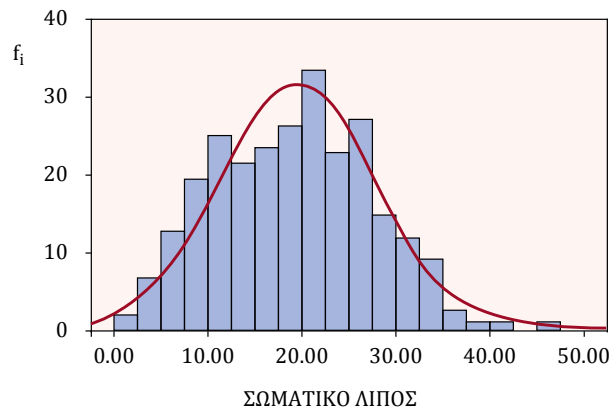
Στις περισσότερες περιπτώσεις το ιστόγραμμα το οποίο μελετούμε αναφέρεται σε εμπειρική κατανομή. Η τυχαιότητα των παρατηρήσεων και το πεπερασμένο και συνήθως μικρό, μέγεθος του δείγματος έχει ως αποτέλεσμα το ιστόγραμμα και η πολυγωνική γραμμή να μην είναι ομαλή. Έτσι, ακόμα και αν η κατανομή στον πληθυσμό είναι π.χ. ομαλή και απολύτως συμμετρική, στο δείγμα θα παρατηρήσουμε αποκλίσεις από μια τέτοια συμπεριφορά. Με την εμπειρία ο ερευνητής μπορεί να κρίνει αν ένα ιστόγραμμα αποκτά την ικανότητα να θεωρηθεί (περίπου) συμμετρικό ή όχι.

παράδειγμα 3.11

Μπορούμε με σχετική ασφάλεια να θεωρήσουμε ότι είναι συμμετρικό το ακόλουθο ιστόγραμμα των μετρήσεων του λίπους σε 252 ενήλικες άντρες.



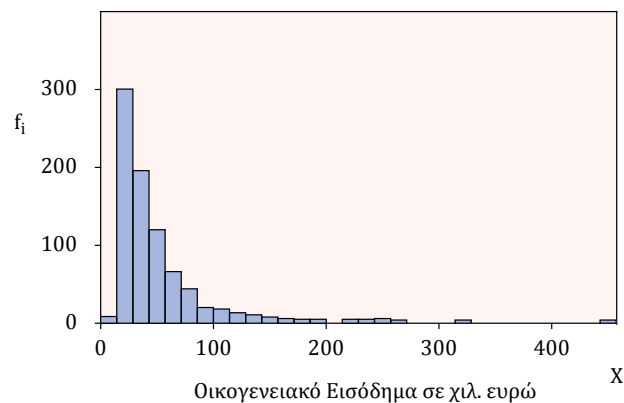
Στο επόμενο σχήμα δίνεται το ίδιο ιστόγραμμα μαζί με την κανονική καμπύλη η οποία ορίζει μια θεωρητική κατανομή που είναι συμμετρική γύρω από τη μέση τιμή. Μπορούμε λοιπόν κατ' αρχήν να υποθέσουμε ότι οι παρατηρήσεις αυτού του δείγματος έχουν προέλθει από μια τέτοια κατανομή. Πιο αυστηρούς στατιστικούς ελέγχους για κανονικότητα θα δούμε σε επόμενα κεφάλαια.



3.6.3 Λοξότητα και μετασχηματισμοί

Όπως ειδικότερα θα δούμε στο μέρος ΙΙΙ αυτού του Βιβλίου, στη στατιστική συμπερασματολογία ορισμένες τεχνικές (π.χ. ο έλεγχος με το κριτήριο t) μπορούν να εφαρμοστούν μόνο σε δεδομένα τα οποία προέρχονται από κατανομή συμμετρική. Όταν η εμπειρική κατανομή είναι εμφανώς λοξή, τότε οι αντίστοιχες τεχνικές δεν είναι πια αξιόπιστες.

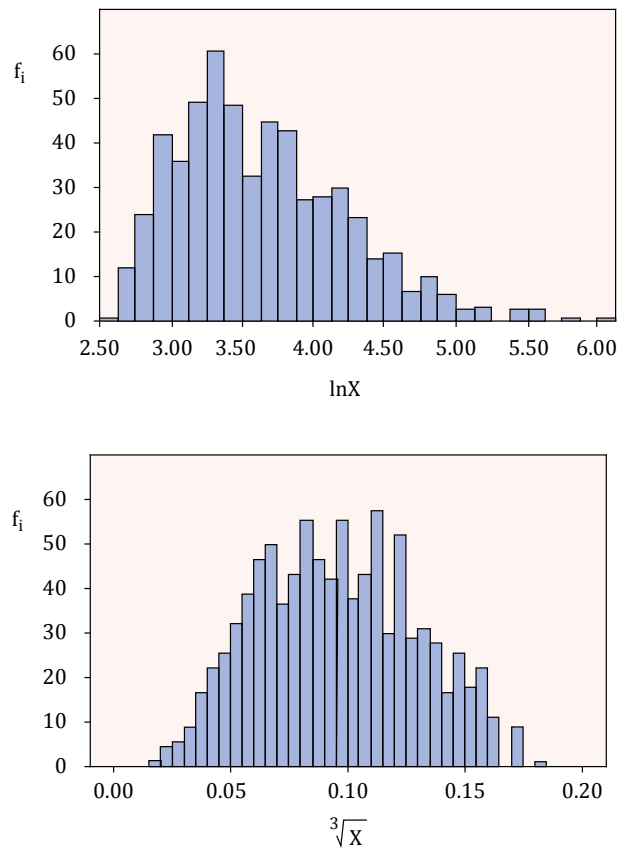
Λοξές κατανομές παρατηρούμε συχνά σε οικονομικές εφαρμογές όπου π.χ. στα σύνολα δεδομένων που αναφέρονται σε εισοδήματα παρατηρείται το εξής: το μεγαλύτερο ποσοστό των ατόμων έχει εισοδήματα κοντά στο βασικό και όσο απομακρυνόμαστε από αυτό τόσο το ποσοστό των δικαιούχων μικραίνει. Αυτό έχει ως συνέπεια η κατανομή συχνοτήτων σε εισοδηματικά δεδομένα να έχει δεξιά λοξότητα όπως στο ακόλουθο σχήμα που αποτελεί το ιστόγραμμα του οικογενειακού εισοδήματος σε δείγμα 850 νοικοκυριών.



Σε μια τέτοια περίπτωση, για να κάνουμε την εμπειρική κατανομή (περίπου) συμμετρική, δοκιμάζουμε κάποιον μετασχηματισμό στα δεδομένα αρχίζοντας από τον λογαριθμικό. Στα επόμενα σχήματα δίνεται το ιστόγραμμα του νεπέριου λογαρίθμου του εισοδήματος και της τρίτης ρίζας του εισοδήματος αντίστοιχα. Ο δεύτερος φαίνεται να είναι ένας μετασχηματισμός που φέρνει την κατανομή πιο κοντά στη συμμετρία (για τον λογαριθμικό μετασχηματισμό γίνεται ξανά αναφορά στην ενότητα 4.7).

3.6.4 Η δίκورφη κατανομή

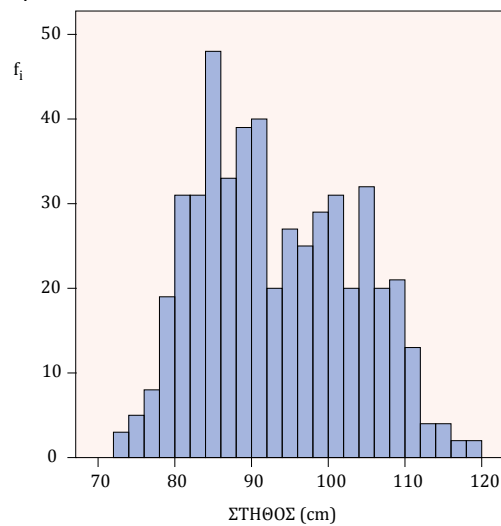
Είναι δυνατόν στο ιστόγραμμα να εμφανίζονται δύο κορυφές οι οποίες επιμένουν ακόμα και αν δοκιμάσουμε διαφορετική ομαδοποίηση. Μια τέτοια εικόνα μας οδηγεί στο βασίμο συμπέρασμα ότι τα



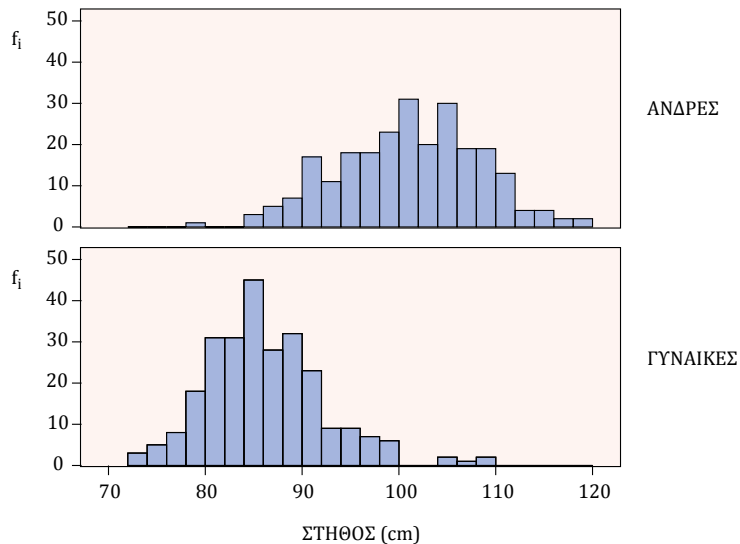
δεδομένα μας μπορούν να καταταγούν σε δύο κατηγορίες ως προς κάποιο χαρακτηριστικό (ή, ισοδύναμα, ότι προέρχονται από δύο διαφορετικούς πληθυσμούς). Το αν θα πρέπει να εξεταστούν χωριστά είναι κάτι που θα κριθεί κατά περίπτωση.

παράδειγμα 3.12

Το ιστόγραμμα που ακολουθεί είναι η γραφική παράσταση της περιμέτρου του στήθους σε τυχαίο δείγμα 507 αθλουμένων. Υποπευόμαστε ότι οι δύο κορυφές οφείλονται στο ότι τα δεδομένα αναφέρονται σε άνδρες και γυναίκες.

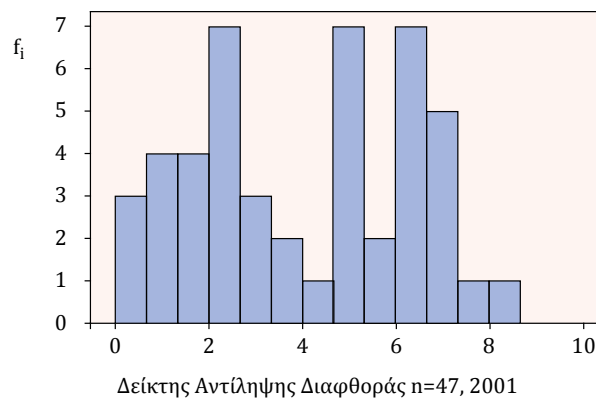


Αυτό επιβεβαιώνεται στο ακόλουθο γράφημα όπου δίνεται το ιστόγραμμα για τους άνδρες και τις γυναίκες του δείγματος, χωριστά.



παράδειγμα 3.13

Οι δύο κορυφές στο ιστόγραμμα του Δείκτη Αντίληψης Διαφθοράς (Corruption Perception Index ή CPI) για 47 χώρες αντιστοιχεί στις δύο κατηγορίες χωρών: αναπτυσσόμενες και υπανάπτυκτες. (Η τιμή 10 του δείκτη αντιστοιχεί σε μηδενική διαφθορά.)



3.7 Το Φυλλογράφημα

Η ομαδοποίηση των δεδομένων σε μια κατανομή με k τάξεις και η γραφική τους παρουσίαση με το ιστόγραμμα συχνοτήτων έχει δύο μειονεκτήματα: πρώτον εξαρτάται από τις επιλογές του ερευνητή και δεύτερο και πιο σημαντικό, δεν διατηρεί τις αρχικές τιμές.

Το φυλλογράφημα ονομαζέται και διάγραμμα μίσχου-και-φύλλο (stemplot or stem-and-leaf plot) και είναι ένας έξυπνος¹ και απλός τρόπος παρουσίασης της κατανομής ο οποίος διατηρεί τις πραγματικές

¹ Παρουσιάστηκε για πρώτη φορά από τον Tukey J. (1977), θεμελιωτή της λεγόμενης Διερευνητικής Ανάλυσης των Δεδομέ-

τιμές των δεδομένων. Από το φυλλογράφημα μπορούμε να πάρουμε αμέσως σημαντικές πληροφορίες για την κατανομή δεδομένων και ειδικότερα:

- αν είναι συμμετρική ή λοξή,
- αν έχει μία ή περισσότερες κορυφές,
- αν έχει ασυνήθιστα σημεία.

Επίσης, από το φυλλογράφημα μπορούμε να προσδιορίσουμε, σχετικά εύκολα, τη θέση της κατανομής είτε με το μάτι είτε προσδιορίζοντας τη διάμεσο ή μεσαία τιμή μετρώντας οποιοδήποτε άκρο του φυλλογραφήματος μέχρι να βρούμε τις μισές παρατηρήσεις.

Το φυλλογράφημα λειτουργεί κυρίως για λίγα δεδομένα με μικρούς αριθμούς, που είναι όλοι μεγαλύτεροι από το μηδέν. Έτσι, όταν χρειάζεται θα πολλαπλασιάσουμε τα δεδομένα με την κατάλληλη δύναμη του 10, ώστε το ακέραιο μέρος τους να είναι μονοψήφιο ή διψήφιο. Στη συνέχεια εργαζόμαστε ως εξής:

- Χωρίζουμε κάθε παρατήρηση σε ένα μίσχο και ένα φύλλο. Ο μίσχος μπορεί να έχει όσα ψηφία χρειάζεται αλλά το φύλλο μόνο ένα ψηφίο.
- Τοποθετούμε τους μίσχους (δηλαδή τα πρώτα ψηφία κάθε παρατήρησης) σε μία στήλη σε αύξουσα σειρά από πάνω προς τα κάτω, τραβούμε μια κάθετη γραμμή προς τα δεξιά των μίσχων και παραθέτουμε τα φύλλα (δηλαδή το τελευταίο ψηφίο όλων των παρατηρήσεων) από τη δεξιά πλευρά της γραμμής.
- Ταξινομούμε τα φύλλα σε αύξουσα σειρά από αριστερά προς τα δεξιά.

παράδειγμα 3.14

Σε τυχαίο δείγμα 2 παιδικών σταθμών μετρήθηκε με ειδικό τεστ το IQ και τα αποτελέσματα των μετρήσεων σε αύξουσα τάξη μεγέθους είναι τα εξής:

61	72	75	77	79	80	80	81	81	82	83	84	84	85	85	85
86	88	89	90	90	91	92	93	94	96	96	97	98	98	100	100
101	101	102	102	102	103	104	105	106	107	107	108	109	109	109	110
110	111	112	113	113	114	115	115	115	116	117	118	119	120	121	121
121	122	123	124	125	125	126	127	129	130	132	136	142	146	151	

Το φυλλογράφημα για τα δεδομένα αυτά είναι το εξής:

Μίσχοι	Φύλλα
6	1
7	2 5 7 9
8	0 0 1 1 2 3 4 4 5 5 5 6 8 9
9	0 0 1 2 3 4 6 6 7 8 8
10	0 0 1 1 2 2 2 3 4 5 6 7 7 8 9 9 9
11	0 0 1 2 3 3 4 5 5 5 6 7 8 9
12	0 1 1 1 2 3 4 5 5 6 7 9
13	0 2 6
14	2 6
15	1

νων (Exploratory Data Analysis or EDA). Αυτή είναι μια πρακτική αντίληψη που ανάγει την οπτική ανάλυση των δεδομένων σε στάδιο ουσιαστικό, που καθορίζει τα επόμενα στάδια της στατιστικής επεξεργασίας

παράδειγμα 3.15

Σε ορισμένο ορυχείο βωξίτη καταγράψαμε την ημερήσια παραγωγή μεταλλεύματος σε τόνους στη διάρκεια 28 εργάσιμων ημερών και πήραμε τα ακόλουθα αποτελέσματα.

9.0	9.5	10.0	10.5	10.8	11.0	11.2	11.4	11.7	12.0	12.4
12.6	13.0	13.0	13.2	13.4	13.5	13.5	13.7	13.8	14.3	14.6
14.9	15.0	15.4	16.0	16.2	16.6	17.1				

Το φυλλογράφημα αυτών των παρατηρήσεων πολλαπλασιασμένων επί 10 είναι το εξής:

Μίσχοι	Φύλλα
9	0 5
10	0 5 8
11	0 2 4 7
12	0 4 6
13	0 0 2 4 5 5 7 8
14	3 6 9
15	0 4
16	0 2 6
17	1

παράδειγμα 3.16

Σε τυχαίο δείγμα 38 βροχοπτώσεων μετρήσαμε το ύψος του νερού και πήραμε τα ακόλουθα αποτελέσματα:

0.03	0.03	0.04	0.06	0.07	0.07	0.08	0.09	0.15	0.16
0.17	0.18	0.21	0.22	0.25	0.25	0.26	0.28	0.28	0.31
0.35	0.44	0.45	0.45	0.48	0.52	0.52	0.53	0.54	0.54
0.55	0.55	0.56	0.58	0.59	0.60	0.60	0.62		

Πολλαπλασιάζουμε τις παρατηρήσεις επί 100 και κατασκευάζουμε το ακόλουθο φυλλογράφημα:

Μίσχοι	Φύλλα
0	3 3 4 6 7 7 8 9
1	5 6 7 8
2	1 2 5 5 6 8 8
3	1 5
4	4 5 5 8
5	2 2 3 4 4 5 5 6 8 9
6	0 0 2

ΑΠΛΩΜΑ. Αν τα ψηφία-μίσχος έχουν μικρό εύρος, οπότε το φυλλογράφημα δεν δίνει αρκετές πληροφορίες για τη μεταβλητότητα των παρατηρήσεων, μπορούμε να σπάσουμε τα φύλλα σε δύο ομάδες με τον ίδιο μίσχο: στην πρώτη θα αντιστοιχήσουμε τα φύλλα από 0-4 και στη δεύτερη τα φύλλα 5-9. Έτσι προκύπτει το λεγόμενο **απλωμένο φυλλογράφημα** (spread-out stem-and-leaf).

Το φυλλογράφημα είναι κατάλληλο για μικρά σύνολα δεδομένων –το πολύ 50 παρατηρήσεις.

παράδειγμα 3.17

Δίνεται η μέση ετήσια δαπάνη για ιατροφαρμακευτική περίθαλψη ανά ασφαλισμένο συνταξιούχο ορισμένου ταμείου, στη διάρκεια ενός έτους, σε χιλιάδες ευρώ.

3.0	3.3	3.3	3.9	4.2	2.8	2.1	3.2	4.0	3.1	4.4	3.0	3.4
3.5	3.7	3.2	1.8	3.2	2.9	2.4	4.0	3.7	3.9	5.3	3.8	3.5
3.3	3.7	2.8	2.9	2.6	4.7	3.1	3.5	3.7	3.8	4.0	2.3	3.4
4.6	3.6	3.2	2.5	3.4	3.1	3.3	3.6	3.5	2.6	4.3	2.7	4.3

Το φυλλογράφημα που προκύπτει αν απλώσουμε τους μίσχους διπλασιάζοντάς τους είναι το εξής:

Μίσχοι	Φύλλα
1	8
2	1 3 4
2	5 6 6 7 8 8 9 9
3	0 0 1 1 1 2 2 2 2 3 3 3 3 4 4 4
3	5 5 5 5 6 6 7 7 7 7 8 8 9 9
4	0 0 0 2 3 3 4
4	6 7
5	3

Αν και πάλι η στήλη των μίσχων είναι μικρή και οι παρατηρήσεις πολλές, μπορούμε να απλώσουμε το φυλλογράφημα ακόμη περισσότερο πενταπλασιάζοντας τις ομάδες με τον ίδιο μίσχο: στην πρώτη αντιστοιχούμε τα φύλλα 0 και 1 στη δεύτερη τα φύλλα 2 και 3, κ.ο.κ., όπως στο ακόλουθο παράδειγμα.

παράδειγμα 3.18

Σε 36 κεσεδάκια γιαούρτης μετρήσαμε το pH (μονάδα μέτρησης της οξύτητας) και πήραμε τα ακόλουθα αποτελέσματα:

3.8	4.4	4.5	4.6	4.6	4.7	4.7	4.7	4.8	4.9
4.9	5.0	5.0	5.0	5.1	5.2	5.2	5.2	5.3	5.3
5.3	5.4	5.4	5.5	5.5	5.5	5.5	5.5	5.6	5.6
5.7	5.7	5.7	5.8	5.9	6.6				

Πολυπλασιάζουμε τις παρατηρήσεις με 10 και επειδή έχουμε μόνον 4 δυνατές τιμές για το πρώτο ψηφίο πενταπλασιάζουμε τις ομάδες με τον ίδιο μίσχο, οπότε προκύπτει το ακόλουθο φυλλογράφημα.

Μίσχοι	Φύλλα
3	8
4	
4	
4	4 5
4	6 6 7 7 7
4	8 9 9
5	0 0 0 1
5	2 2 2 3 3 3
5	4 4 5 5 5 5 5
5	6 6 7 7 7
5	8 9
6	
6	
6	
6	
6	6

ΑΠΟΚΟΠΗ. Σε πολλές περιπτώσεις, όταν οι παρατηρούμενες τιμές έχουν πολλά ψηφία, είναι χρήσιμο να αποκόψουμε ένα ή περισσότερα από τα τελικά ψηφία των παρατηρήσεων.

Το αν θα προτιμήσουμε αποκοπή ή διπλασιασμό των μίσχων επαφίεται στην κρίση του ερευνητή. Πρέπει να θυμόμαστε ότι ο σκοπός ενός φυλλογράφηματος είναι να παρουσιάσει το σχήμα της κατανομής. Έτσι, αν έχει λιγότερους από 5 μίσχους είναι καλό να απλώσουμε το φυλλογράφημα. Αν πολλοί μίσχοι δεν έχουν φύλλα ή έχουν μόνον ένα, βλέπουμε μήπως η αποκοπή βοηθήσει.

παράδειγμα 3.19

Δίνονται οι ημερήσιες πωλήσεις, σε ευρώ, ενός καταστήματος αλυσίδας σούπερ μάρκετ.

9760	17221	13697	14568	13766	12541
10265	8377	12431	17592	7366	12417
14102	10401	12323	10634	11670	9170
18468	10244	11894	12829	13679	14959
9206	15642	13267	12527	12990	12772

Αποκόβουμε τα δύο τελικά ψηφία και κατασκευάζουμε το ακόλουθο φυλλογράφημα

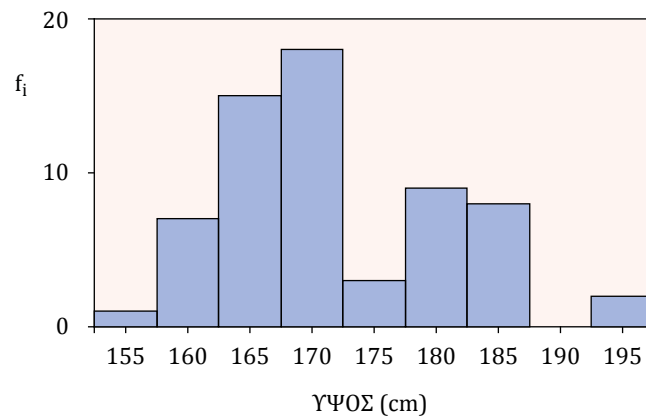
Μίσχοι	Φύλλα
7	3
8	3
9	1 2 7
10	2 2 4 6
11	6 8
12	3 4 4 5 5 7 8 9
13	2 6 6 7
14	1 5 9
15	6
16	
17	2 5
18	4

3.7.1 Η πρακτική χρησιμότητα του φυλλογραφήματος

Το φυλλογράφημα είναι ένας έξυπνος τρόπος ομαδοποίησης των δεδομένων που δεν χάνει τις αρχικές τιμές. Δυστυχώς δεν κατάφερε να κερδίσει σε δημοτικότητα τη θέση του ιστογράμματος, παρόλο που σε πολλές πρακτικές εφαρμογές δίνει καλύτερη εικόνα όπως στο παράδειγμα που ακολουθεί

παράδειγμα 3.20

Σε ένα τμήμα 68 πρωτοετών καταγράψαμε το ύψος τους (όπως το δήλωσαν) και στα δύο σχήματα που ακολουθούν δίνεται το ιστόγραμμα και το φυλλογράφημα των αντίστοιχων 68 μετρήσεων.



Μίσχοι	Φύλλα
15	7 8 8 8
16	0 0 2 2 3 3 4
16	5 5 5 5 5 6 6 6 7 7 7 7 8 8 8 8 8 9
17	0 0 0 0 0 0 0 0 0 1 2 2 3 4
17	5 8 8 8 9 9
18	0 0 0 2 3 3 4 5 5 5 5 7
19	3 7

Παρατηρούμε ότι ενώ από το ιστόγραμμα θα υπέθετε κάποιος ότι υπάρχουν δύο τιμές που είναι πολύ μακριά από τις υπόλοιπες, στο φυλλογράφημα βλέπει ότι οι μεγαλύτερες τιμές είναι το 193 και 195 και η αμέσως προηγούμενη είναι η 187 άρα δεν έχουμε ασυνήθιστα ακραίες τιμές.