

ΔΗΜΗΤΡΗ Α. ΙΩΑΝΝΙΔΗ

Στατιστικές Μέθοδοι

ΠΕΡΙΓΡΑΦΙΚΗ ΣΤΑΤΙΣΤΙΚΗ

ΘΕΩΡΙΑ ΠΙΘΑΝΟΤΗΤΩΝ

ΣΤΑΤΙΣΤΙΚΗ ΣΥΜΠΕΡΑΣΜΑΤΟΛΟΓΙΑ

ΑΠΛΗ ΚΑΙ ΠΟΛΛΑΠΛΗ ΓΡΑΜΜΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ

ΧΡΟΝΟΛΟΓΙΚΕΣ ΣΕΙΡΕΣ

ΑΝΑΛΥΣΗ ΔΙΑΚΥΜΑΝΣΗΣ

3^η έκδοση



Κάθε γνήσιο αντίτυπο υπογράφεται από το συγγραφέα

ISBN 960-431-970-1

© Copyright: Δ. Ιωαννίδη, Εκδόσεις Ζήτη, Μάρτιος 1999, Μάρτιος 2001,
Σεπτέμβριος 2005 Θεσσαλονίκη

Το παρόν έργο πνευματικής ιδιοκτησίας προστατεύεται κατά τις διατάξεις του Ελληνικού νόμου (Ν.2121/1993 όπως έχει τροποποιηθεί και ισχύει σήμερα) και τις διεθνείς συμβάσεις περί πνευματικής ιδιοκτησίας. Απαγορεύεται απολύτως η άνευ γραπτής άδειας του εκδότη και συγγραφέα κατά οποιοδήποτε τρόπο ή μέσο αντιγραφή, φωτοανατύπωση και εν γένει αναπαραγωγή, εκμίσθωση ή δανεισμός, μετάφραση, διασκευή, αναμετάδοση στο κοινό σε οποιαδήποτε μορφή (ηλεκτρονική, μηχανική ή άλλη) και η εν γένει εκμετάλλευση του συνόλου ή μέρους του έργου.



*Φωτοστοιχειοθεσία
Εκτύπωση*

Βιβλιοπωλείο

www.ziti.gr

Π. ΖΗΤΗ & ΣΙΑ ΟΕ

180 χλμ Θεσ/νίκης-Περαιάς
Τ.Θ. 4171 • Περαιά Θεσσαλονίκης • Τ.Κ. 570 19
Τηλ.: 23920 72.222 (5 γραμ.) - Fax: 23920 72.229
e-mail: info@ziti.gr

ΕΚΔΟΣΕΙΣ ΖΗΤΗ

Αρμενοπούλου 27 • 546 35 Θεσσαλονίκη
Τηλ. 2310 203.720, Fax 2310 211.305
e-mail: sales@ziti.gr

Πρόλογος

Η χρησιμοποίηση των στατιστικών μεθόδων μπορεί να συμβάλλει σημαντικά σε μία αποτελεσματικότερη διοίκηση στο Δημόσιο και στον Ιδιωτικό τομέα. Η σημασία της Στατιστικής είναι εμφανής, αφ' ενός από το γεγονός ότι στα προγράμματα των περισσότερων πανεπιστημιακών τμημάτων, ιδιαίτερα της αλλοδαπής, περιλαμβάνονται ποικίλα μαθήματα Στατιστικής, και αφ' ετέρου πολλοί δημόσιοι και ιδιωτικοί οργανισμοί διεθνώς διαθέτουν στατιστικά γραφεία που ασχολούνται με την καταγραφή και συλλογή δεδομένων που τους αφορούν, αλλά και συχνά με τη λήψη αποφάσεων ή προβλέψεων.

Σκοπός αυτού του εγχειριδίου είναι να παρουσιασθούν οι πρώτες βασικές αρχές της Στατιστικής μ' ένα μεθοδικό και επεξηγηματικό τρόπο, αποφεύγοντας τις ιδιαίτερες μαθηματικές δυσκολίες. Τα δύο πρώτα κεφάλαια ασχολούνται με εισαγωγικές έννοιες της στατιστικής, και συγκεκριμένα της περιγραφικής. Η κατανόηση αυτών είναι βοηθητική για την ευκολότερη εκμάθηση των εννοιών των επόμενων κεφαλαίων. Στα Κεφάλαια 3, 4 και 5 γίνεται μια σύντομη εισαγωγή στις πιθανότητες, και μελετώνται οι βασικές θεωρητικές κατανομές. Μία εισαγωγή στις πολυδιάστατες κατανομές δίνεται στο Κεφάλαιο 6, με έμφαση τη μελέτη των ανεξάρτητων τυχαίων μεταβλητών. Στο Κεφάλαιο 7 δίνονται οι γενικές συνθήκες κάτω από τις οποίες η κατανομή του αθροίσματος τυχαίων μεταβλητών προσεγγίζεται από μία κανονική κατανομή. Στα υπόλοιπα κεφάλαια διατυπώνονται και αναπτύσσονται οι βασικότερες μέθοδοι για την εξαγωγή στατιστικών συμπερασμάτων σχετικά με τις παραμέτρους που χαρακτηρίζουν έναν πληθυσμό, με βάση πάντα ένα τυχαίο δείγμα που λαμβάνεται απ' αυτόν. Έτσι στο Κεφάλαιο 8 παρουσιάζονται τα απαραίτητα τεχνικά μέσα που θα μας βοηθήσουν στην εξαγωγή αυτών των συμπερασμάτων. Στο Κεφάλαιο 9 αναπτύσσονται διάφοροι Μέθοδοι Σημειοεκτίμησης των παραμέτρων ενός πληθυσμού, ενώ στο Κεφάλαιο 10 η εκτίμηση αυτών γίνεται με τη βοήθεια των Διαστημάτων Εμπιστοσύνης. Άλλες μέθοδοι εξαγωγής στατιστικών συμπερασμάτων είναι αυτές των Ελέγχων Υποθέσεων και δίνονται στο Κεφάλαιο 11. Στα Κεφάλαια 9, 10 και 11 πάντοτε δεχόμαστε ότι ο πληθυσμός που μελετάμε ακολουθεί μία συγκεκριμένη κατανομή κάτι όμως που μπορούμε να το ελέγξουμε στατιστικά. Το τελευταίο επιτυγχάνεται στο Κεφάλαιο 12, όπου παράλληλα αναπτύσσονται και κάποια άλλα στατιστικά προ-

βλήματα. Η απλή γραμμική παλινδρόμηση στην οποία πολύ συχνά υπακούουν δύο μεταβλητές, καθώς και ένας εμπειρικός τρόπος υπολογισμού της γραμμικής συσχέτισης αυτών αναπτύσσονται στο Κεφάλαιο 13. Βασικές αρχές της ανάλυσης διακύμανσης δίνονται στο Κεφάλαιο 14. Τέλος όλα τα παραδείγματα και οι ασκήσεις αποτελούν εφαρμογές της Στατιστικής, χρήσιμες για οικονομολόγους και στελέχη επιχειρήσεων. Όλα τα ανωτέρω αποτελούν μέρος των βασικότερων μεθόδων της Στατιστικής, που όμως δεν εξαντλώνται με το παρόν εγχειρίδιο. Έτσι οι μη-παραμετρικοί μέθοδοι, οι μέθοδοι κατά Bayes, στοιχεία από τις χρονολογικές σειρές κ.λπ. θα συμπλήρωναν ικανοποιητικά αυτό το εγχειρίδιο. Πιστεύουμε ότι θα ακολουθήσει στο μέλλον.

Θεσσαλονίκη, Μάρτιος 1999

Δημήτρης Ιωαννίδης

Πρόλογος για τη 2η έκδοση

Κατά τη δεύτερη έκδοση προβήκαμε σε διορθώσεις και βελτιώσεις του κειμένου της πρώτης. Προσθέσαμε δύο νέα κεφάλαια, που αφορούν την πολλαπλή παλινδρόμηση και τις χρονολογικές σειρές, καθώς και άλλες έννοιες που συμπληρώνουν πιο ικανοποιητικά τα αρχικά κεφάλαια αυτού του συγγράμματος.

Θεσσαλονίκη, Μάρτιος 2001

Δημήτρης Ιωαννίδης

Πρόλογος για τη 3η έκδοση

Στην τρίτη έκδοση εμπλουτίζουμε την ύλη της προηγούμενης έκδοσης, εισάγοντας νέες μεθόδους περισσότερα παραδείγματα και σχήματα, αναλύοντας με τον πιο διαισθητικό τρόπο τις «Στατιστικές Μεθόδους» στις πρακτικές εφαρμογές. Κρίναμε απαραίτητο να αναπτύξουμε περισσότερο ορισμένες έννοιες και να προσθέσουμε στο τέλος κάθε κεφαλαίου ορισμένες ερωτήσεις που θα διευκολύνουν τον αναγνώστη στην κατανόηση της ύλης.

Τέλος προβήκαμε σε διόρθωση τυπογραφικών λαθών της προηγούμενης έκδοσης.

Θεσσαλονίκη, Μάρτιος 2005

Δημήτρης Ιωαννίδης

*Στους λατρευτούς μου
Λέσποινα & Σταύρο*

Περιεχόμενα

σελ.

1. Γενικά περί Στατιστικής	13
2. Περιγραφική Στατιστική	
2.1 Στατιστικές μονάδες. Στατιστικά γνωρίσματα και Μεταβλητές	17
2.2 Κατανομές Συχνοτήτων	20
2.3 Απόλυτη και Σχετική Συχνότητα	21
2.4 Ιστογράμματα	23
2.5 Αθροιστικές Κατανομές Συχνοτήτων	27
2.6 Μέτρα Θέσης (ή Κεντρικής Τάσης)	30
2.7 Μέτρα Απόκλισης	39
2.8 Άλλες Παράμετροι Κατανομών Συχνοτήτων	47
Ερωτήσεις	53
3. Βασικές Αρχές της Θεωρίας Πιθανοτήτων	
3.1 Βασικές Έννοιες: Τυχαία πειράματα - Δειγματοχώροι - Ενδεχόμενα	55
3.2 Ορισμός της Πιθανότητας	60
3.3 Βασικές Αρχές Απαρίθμησης	67
3.4 Δεσμευμένες Πιθανότητες	71
3.5 Πολλαπλασιαστικός κανόνας πιθανοτήτων	72
3.6 Ανεξάρτητα Ενδεχόμενα	73
3.7 Νόμος του Bayes - Υποκειμενικές Πιθανότητες	74
3.8 Γενικές Ασκήσεις	77
Ερωτήσεις	78
4. Τυχαίες Μεταβλητές και Κατανομές	
4.1 Τυχαίες Μεταβλητές	79
4.2 Διακριτές κατανομές	84
4.3 Συνεχείς κατανομές	93
Ερωτήσεις	101
5. Παράμετροι Κατανομών	
5.1. Παράμετροι θέσεων: Μέση τιμή, Διάμεσος, Επικρατούσα τιμή	103
5.2. Παράμετροι Απόκλισης: Διακύμανση ή Διασπορά. Τυπική απόκλιση	107
5.3. Λυμένες Ασκήσεις	110

5.4. Η Λοξότητα σαν μέτρο ασυμμετρίας μιας κατανομής.....	113
5.5. Κυρτότητα μιας κατανομής.....	114
Ερωτήσεις	115

6. Κοινές κατανομές και διάφορες προτάσεις

γύρω από τη μέση τιμή και διακύμανση

6.1. Από κοινού συνάρτηση κατανομής και πυκνότητα πιθανότητας. Περιθωριακή συνάρτηση κατανομής και πυκνότητα πιθανότητας.....	117
6.2. Μέση τιμή και Διακύμανση μιας συνάρτησης δύο τ.μ.....	121
6.3. Δεσμευμένη πυκνότητα πιθανότητας.....	122
6.4. Δεσμευμένη μέση τιμή.....	123
6.5. Ανεξάρτητες τ.μ.	125
6.6. Συνδιακύμανση και Συσχέτιση.....	126
6.7. Διάφορες προτάσεις.....	131
Ερωτήσεις	134

7. Νόμος των Μεγάλων Αριθμών και Κεντρικό Οριακό Θεώρημα

7.1. Νόμος των Μεγάλων αριθμών.....	135
7.2. Κεντρικό Οριακό Θεώρημα.....	136
7.3. Η Poisson κατανομή σαν προσέγγιση της Διωνυμικής κατανομής.....	140
Ερωτήσεις	141

8. Στατιστική Συμπερασματολογία - Βασικές Έννοιες

8.1. Γενικότητες.....	143
8.2. Σημαντικές Στατιστικές Συναρτήσεις	146
8.3. Δειγματοληπτικές Κατανομές.....	147
8.4. Κατανομές Μερικών Σπουδαίων Τυχαίων Μεταβλητών	151
8.5. Γενικές Ασκήσεις.....	154

9. Σημειοεκτιμητική

9.1. Γενικότητες.....	155
9.2. Αμερόληπτοι Εκτιμητές - Αποδοτικοί Εκτιμητές.....	156
9.3. Συνεπείς Εκτιμητές.....	161
9.4. Μέθοδος των Ελαχίστων Τετραγώνων.....	163
9.5. Εκτιμητές Μέγιστης Πιθανοφάνειας.....	164
Ερωτήσεις	168

10. Εκτίμηση με Διαστήματα Εμπιστοσύνης

10.1. Γενικότητες.....	169
10.2. Συμμετρικά Διαστήματα Εμπιστοσύνης για τη μέση τιμή μ	170
10.3. Διαστήματα Πρόβλεψης	178
10.4. Γενικές Ασκήσεις	179
10.5. Συμμετρικό Διάστημα Εμπιστοσύνης για τη διακύμανση σ^2 ενός κανονικού πληθυσμού	181

10.6. Σύμμετρα Διαστήματα Εμπιστοσύνης για τη διαφορά των μέσων μ_1, μ_2 δύο πληθυσμών.....	183
Ερωτήσεις	191

11. Έλεγχοι Υποθέσεων

11.1. Γενικότητες	193
11.2. Έλεγχοι για τη μέση τιμή μ	197
11.3. Έλεγχοι για τη διακύμανση σ^2	204
11.4. Έλεγχοι για τη μέση τιμή μη κανονικού πληθυσμού.....	206
11.5. Έλεγχοι για τη διαφορά των μέσων δύο πληθυσμών	207
11.6. Έλεγχοι σύγκρισης των διακυμάνσεων δύο πληθυσμών.....	212
11.7. Ισοδυναμία Ελέγχων Υποθέσεων και Διαστημάτων Εμπιστοσύνης.....	216
11.8. Ισχύς Στατιστικών Ελέγχων - p - τιμή	218
11.9. Ποιοτικός Έλεγχος	224
Ερωτήσεις	226

12. Ανάλυση Κατηγοροποιημένων Δεδομένων

12.1. Πολυωνυμική κατανομή	227
12.2. χ^2 -Έλεγχοι.....	228
12.3. Έλεγχοι Καλής Προσαρμογής.....	231
12.4. Πίνακες Συνάφειας (Έλεγχοι Ανεξαρτησίας).....	234
Ερωτήσεις	238

13. Παλινδρόμηση και Συσχέτιση

13.1. Γενικότητες	239
13.2. Γραμμική παλινδρόμηση (απλή)	239
13.3. Μέθοδος Ελαχίστων Τετραγώνων.....	244
13.4. Εκτίμηση διακύμανσης της ευθείας παλινδρόμησης.....	247
13.5. Στατιστικός έλεγχος και διάστημα εμπιστοσύνης για την κλίση της ευθείας παλινδρόμησης.....	250
13.6. Μέση εκτίμηση της (εξααρτημένης μεταβλητής) Y για δοθέν x της ανεξάρ- τητης X	251
13.7. Πρόβλεψη τιμών της Y για δοθέν x	253
13.8. Δειγματοληπτική Συσχέτιση (Pearson συντελεστής συσχέτισης).....	254
Ερωτήσεις	258

14. Ανάλυση Διακύμανσης

14.1. Γενικότητες	259
14.2. Μονοπαραγοντική Ανάλυση Διακύμανσης.....	260
14.3. Διπαραγοντική Ανάλυση Διακύμανσης – με μια παρατήρηση ανά γραμμή και στήλη	264
14.4. Διπαραγοντική Ανάλυση Διακύμανσης – με περισσότερες από μια παρατηρήσεις ανά γραμμή και στήλη.....	268
14.5. Σπουδαίες παρατηρήσεις	272
Ερωτήσεις	274

15. Πολυδιάστατη Παλινδρόμηση

15.1. Γενικότητες	275
15.2. Πολυδιάστατο Γραμμικό Μοντέλο	275
15.3. Μέθοδος Ελαχίστων Τετραγώνων.....	276
15.4. Εκτίμηση Διακύμανσης της Πολλαπλής Παλινδρόμησης και Πολυδια- στατος Συντελεστής Συσχέτισης	281
15.5. Μερικά Στατιστικά Συμπεράσματα στο Πολυδιάστατο Γραμμικό Μοντέ- λο.....	284
15.6. Πρόβλεψη στο Πολυδιάστατο Μοντέλο Παλινδρόμησης	286
15.7. Ανάλυση της Διακύμανσης και Έλεγχοι Υποθέσεων.....	286
15.8. F-έλεγχος: Μερικές από τις β παραμέτρους είναι 0.....	288
15.9. Γενικές Παρατηρήσεις στην Πολυδιάστατη Παλινδρόμηση	289
<i>Ερωτήσεις</i>	290

16. Χρονολογικές σειρές

16.1. Γενικότητες	291
16.2. Παραδείγματα χρονολογικών σειρών.....	291
16.3. Διαμερισμός χρονολογικών σειρών σε συνιστώσες.....	302
16.4. Μοντέλα των Box και Jenkins	311
<i>Ερωτήσεις</i>	314

Παράρτημα

A. Πίνακες κατανομών.....	317
B. Ειδικοί πίνακες κατανομών.....	348

<i>Βιβλιογραφία</i>	355
---------------------------	-----

<i>Ευρετήριο Όρων</i>	357
-----------------------------	-----

Συντομογραφίες

β.ε.	βαθμοί ελευθερίας
βλ.	βλέπε
δηλ.	δηλαδή
ε.σ.	επίπεδο σημαντικότητας
κ.λπ.	και λοιπά
π.π.	πυκνότητα πιθανότητας
σ.ε.	συντελεστής εμπιστοσύνης
σ.κ.	συνάρτηση κατανομής
σ.σ.	στατιστική συνάρτηση
τ.μ.	τυχαία μεταβλητή
τ.π.	τυχαίο πείραμα
υποδ.	υπόδειξη

1

Γενικά περί Στατιστικής

Οι περισσότεροι άνθρωποι όταν αναφέρονται στη λέξη “**Στατιστική**” συνήθως σκέπτονται πίνακες από αριθμούς (αριθμητικά δεδομένα). Ως επί το πλείστον είναι εξοικειωμένοι με αριθμητικά δεδομένα που αναφέρονται σε απογραφές πληθυσμών, σε αναλύσεις αποτελεσμάτων αθλημάτων, στην οικονομία κ.λπ.. Το περιεχόμενο της στατιστικής όμως δεν ανταποκρίνεται μόνο σε κάποιους πίνακες αριθμητικών δεδομένων. Η στατιστική σήμερα αποτελεί έναν ανεξάρτητο επιστημονικό κλάδο, με δικές της μεθόδους ανάλυσης. Το αντικείμενό της είναι πολύ ευρύ, ασχολείται με σχεδιασμούς πειραμάτων, μεθόδους συλλογής δεδομένων, καθώς και με την ανάλυση, παρουσίαση, επεξήγηση των δεδομένων με απώτερο σκοπό τη λήψη αποφάσεων και τη δημιουργία προβλέψεων. Τα αριθμητικά δεδομένα αποτελούν ένα πολύ μικρό μέρος αυτής.

Οι πλέον απλές στατιστικές μέθοδοι είναι αυτές που ασχολούνται με τους μέσους των δεδομένων καθώς και άλλων μέτρων που μπορούν να περιγράψουν τα δεδομένα και να μας δώσουν πληροφορίες (**Περιγραφική Στατιστική**). Για παράδειγμα, έστω ότι ενδιαφερόμαστε για να αποκτήσουμε γνώση γύρω από το ετήσιο οικογενειακό εισόδημα 1000 νοικοκυριών μιας περιοχής, η επιλογή των οποίων γίνεται κατά έναν αντιπροσωπευτικό τρόπο. Μερικά χαρακτηριστικά αριθμητικά μεγέθη που μπορούν να μας δώσουν πληροφορίες γύρω απ’ αυτά είναι ο μέσος όρος αυτών, καθώς και ο τρόπος απόκλισης τους από το μέσο αυτών (Διακύμανση ή Διασπορά).

Η σύγχρονη στατιστική διαθέτει μεθόδους για εξαγωγή συμπερασμάτων και πέρα από ένα συγκεκριμένο αριθμό δεδομένων που γνωρίζουμε, καταλήγοντας σε συμπεράσματα που στηρίζονται στην ανάλυση αυτών.

Έτσι στο προηγούμενο παράδειγμα ένα εύλογο ερώτημα είναι ποιο μπορεί να είναι το μέσο ετήσιο εισόδημα των νοικοκυριών της περιοχής αυτής. Μπορούμε να διατυπώσουμε την (στατιστική) υπόθεση ότι το μέσο εισόδημα είναι λιγότερο από 14.000 €, οπότε με βάση τη γνώση των 1000 εισοδημάτων να δεχθούμε ή να απορρίψουμε την υπόθεση αυτή (**Έλεγχος Υποθέσεων**). Επίσης μπορούμε να έχουμε

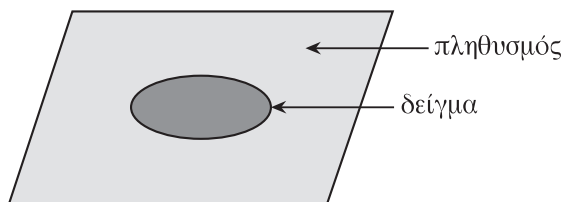
μια εκτίμηση, δηλ. μια μόνο αριθμητική τιμή για το μέσο εισόδημα όλων των οικογενειών με βάση πάλι τα 1000 γνωστά εισοδήματα (**Σημειοεκτίμηση**). Τέλος αντί να έχουμε μια εκτίμηση του μέσου εισοδήματος, να κατασκευάσουμε ένα διάστημα τιμών που μέσα σ' αυτές να εμπεριέχεται το αληθινό μέσο εισόδημα (**Διαστήματα Εμπιστοσύνης**). Οι τρεις τελευταίες έννοιες που αναφέρθηκαν αποτελούν αντικείμενο της "**Στατιστικής Συμπερασματολογίας**" ή της "**Επαγωγικής Στατιστικής**".

Στηριζόμενοι στα παραπάνω μπορούμε να χαρακτηρίσουμε την περιοχή σαν υψηλών ή μεσαίων ή χαμηλών εισοδημάτων.

Το βοηθητικό επιστημονικό εργαλείο που θα μας δίνει τη δυνατότητα εξαγωγής συμπερασμάτων για όλο το πλήθος των οικογενειών με βάση τα γνωστά σε μας 1000 εισοδήματα είναι η "**Θεωρία Πιθανοτήτων**".

Να συζητήσουμε μερικές ακόμη πτυχές αυτού του παραδείγματος. Κάποιος θα ρωτούσε γιατί δεν μαθαίνουμε όλα τα εισοδήματα των κατοίκων αυτής της περιοχής. Οπωσδήποτε για να γίνει αυτό απαιτεί αρκετό χρόνο και σημαντικά οικονομικά έξοδα. Επειδή όμως λογικό είναι να ελαχιστοποιήσουμε το χρόνο εργασίας μας, και το κόστος που απαιτεί αυτή, περιοριζόμαστε σ' ένα μικρότερο αριθμό εισοδημάτων. Το τελευταίο πρέπει να γίνει μ' ένα τρόπο μεθοδικό. Τα 1000 εισοδήματα που επιλέξαμε κατά ένα τυχαίο τρόπο πρέπει να είναι αντιπροσωπευτικά γι' όλα τα υπόλοιπα εισοδήματα. Εν ολίγοις, κάθε εισόδημα πρέπει να έχει την ίδια δυνατότητα ή τον ίδιο βαθμό βεβαιότητας για να είναι ένα μεταξύ των 1000 εισοδημάτων. Υπάρχει ξεχωριστός κλάδος της στατιστικής που ασχολείται μ' αυτό το πρόβλημα και καλείται "**Θεωρία Δειγματοληψιών**". Θα ασχοληθούμε μ' αυτήν εν συντομία σ' ένα από τα επόμενα κεφάλαια.

Το πλήθος των οικογενειών που θα αποτελέσουν αντικείμενο δυνατών παρατηρήσεων είναι ο πληθυσμός μας. Γενικά το πλήθος των αντικειμένων ή ατόμων που συνδέονται με τη (στατιστική) έρευνά μας ονομάζεται "**πληθυσμός**". Τα 1000 νοικοκυριά που τα εισοδήματά τους παρατηρήθηκαν αποτελούν το "δείγμα μας". Κάθε υποσύνολο ενός πληθυσμού ονομάζεται **δείγμα**.



Στη Στατιστική θα δημιουργούμε δείγμα απ' ένα πληθυσμό με βάση το οποίο θα εξάγουμε συμπεράσματα γύρω απ' αυτόν.

Ας συζητήσουμε μερικά ακόμα παραδείγματα, όπου συναντάται η Στατιστική.

Δημοσκοπήσεις: Έστω ότι ενδιαφερόμαστε να γνωρίσουμε αν το κόμμα Α ή το Β κερδίζει μία εκλογική αναμέτρηση. Εδώ ο πληθυσμός μας είναι το σύνολο όλων των ψηφοφόρων. Οι διάφορες εταιρίες δημοσκοπήσεων προβλέπουν το ποιο κόμμα θα κερδίσει με βάση ένα πολύ μικρότερο αριθμό ερωτηθέντων ατόμων (Δείγμα).

Ποιοτικός Έλεγχος:

α) Σε ένα εργοστάσιο παραγωγής τσιμέντου, οι σάκκοι των τσιμέντων οφείλουν να έχουν βάρος 50 kg, και η πακετοποίηση γίνεται με μια αυτόματη μηχανή. Για να ελέγξει κανείς την αξιοπιστία αυτής της μηχανής, θα έπρεπε να ζυγίζει κάθε σάκκο τσιμέντου. Επειδή αυτή η διαδικασία προδικάζει πολύ χρόνο και αρκετά έξοδα, ζυγίζει κανείς κάθε 20^ο ή 50^ο σάκκο και προσδιορίζει το μέσο βάρος των ζυγισθέντων σάκκων. Αν το μέσο βάρος απέχει σημαντικά από τα 50 kg τότε αυτό θα σημαίνει ότι η μηχανή χρειάζεται μια επιδιόρθωση. Οι σάκκοι που παρήχθησαν κατά τη διάρκεια μιας ημέρας αποτελούν τον πληθυσμό μας, ενώ το σύνολο των ζυγισθέντων σάκκων το δείγμα μας.

β) Ορισμένα αντικείμενα κατασκευάζονται σύμφωνα με μια βιομηχανική μέθοδο. Θέλουμε να ελέγξουμε το πλήθος των ελαττωματικών αντικειμένων που παράγονται κατά τη διάρκεια μιας χρονικής μονάδας παραγωγής. Τα αντικείμενα που παρήχθησαν κατ' αυτήν τη χρονική μονάδα αποτελούν τον πληθυσμό μας, ενώ τα αντικείμενα που εξετάσαμε το δείγμα μας.

Στα επόμενα κεφάλαια θα ασχοληθούμε κατ' αρχήν με την Περιγραφική Στατιστική, που κύρια ασχολείται με τρόπους περιγραφής των δειγμάτων μέσω αριθμητικών ποσοτήτων ή γραφικών παραστάσεων. Έπειτα θα αναφερθούμε σε βασικές αρχές της Θεωρίας Πιθανοτήτων, όπου γίνεται η υποδειγματοποίηση (Μοντελοποίηση) των διαφόρων πληθυσμών που συναντώνται. Τέλος με τη Στατιστική Συμπερασματολογία, εξάγουμε συμπεράσματα γύρω από διάφορα χαρακτηριστικά αριθμητικά μεγέθη των πληθυσμών (Παράμετροι πληθυσμών).

2

Περιγραφική Στατιστική

2.1.. Στατιστικές μονάδες. Στατιστικά γνωρίσματα και Μεταβλητές

Τα μέλη ενός πληθυσμού θα τα καλούμε **στατιστικές μονάδες**, όπου η κάθε μια έχει ένα κοινό χαρακτηριστικό με τις υπόλοιπες, το ονομαζόμενο **στατιστικό γνώρισμα** ή απλά **γνώρισμα**. Ένα στατιστικό γνώρισμα θα συμβολίζεται στο εξής μ' ένα κεφαλαίο γράμμα X ή Y κ.λπ., που θα καλείται **μεταβλητή**.

Έτσι στο παράδειγμα με τα οικογενειακά εισοδήματα οι στατιστικές μονάδες είναι τα νοικοκυριά και το στατιστικό γνώρισμα το εισόδημα (X). Στο παράδειγμα των δημοσκοπήσεων η στατιστική μονάδα είναι ο ψηφοφόρος, ενώ το στατιστικό γνώρισμα είναι η ψήφος (X) για το A ή το B κόμμα. Στο πρώτο παράδειγμα ποιοτικού ελέγχου οι στατιστικές μονάδες είναι οι σάκκοι των τσιμεντών που παράγονται, ενώ το στατιστικό γνώρισμα είναι το βάρος (X). Σαν στατιστικό γνώρισμα μπορεί να είναι η ηλικία, το βάρος, το φύλο, η θρησκεία κ.λπ.. Γενικά ένα στατιστικό γνώρισμα μπορεί να είναι **ποσοτικό**, δηλ. να επιδέχεται μετρήσεις, ή **ποιοτικό**, δηλ. να επιδέχεται κατηγοροποιήσεις. Παραδείγματα ποσοτικών στατιστικών γνωρισμάτων είναι το εισόδημα, η ηλικία, το βάρος κ.λπ., ενώ ποιοτικών είναι το φύλλο, η θρησκεία κ.λπ.. Επιπλέον τα ποσοτικά γνωρίσματα διακρίνονται σε **συνεχή** και **διακριτά**. Ένα ποσοτικό γνώρισμα καλείται διακριτό, αν οι αριθμοί των μετρήσεων του είναι διακριτοί πραγματικοί αριθμοί ή στην πλέον συνήθη περίπτωση οι θετικοί ακέραιοι, $0, 1, 2, \dots$. Έτσι το πλήθος των εργαζομένων σ' ένα εργοστάσιο, το πλήθος των αντικειμένων που παράγονται από μια βιομηχανική μονάδα, είναι διακριτά γνωρίσματα. Συνεχές χαρακτηριστικό είναι εκείνο όπου οι αριθμοί των μετρήσεων του είναι πραγματικοί αριθμοί (το χαρακτηριστικό επιδέχεται μετρήσεις και μεταξύ ακεραίων). Τυπικά παραδείγματα γι' αυτήν την περίπτωση είναι το εισόδημα, ο χρόνος λειτουργίας μιας μηχανής κ.λπ.. Στην περίπτωση που έχουμε ποιοτικά χαρακτηριστικά, μπορούμε να έχουμε κατά συμβιβασμό ένα είδος μετρήσεων ανάλογο της διακριτής περίπτωσης. Για παράδειγμα εξετάζοντας το θρήσκευμα των κατοίκων σε μια χώρα, μπο-

ρούμε να παραστήσουμε κατά συμβιβασμό

Άθεος = 0, Ορθόδοξος Χριστ. = 1, Καθολικός Χριστ. = 2, Μουσουλμάνος = 3

ή εξετάζοντας τα αντικείμενα μιας βιομηχανικής παραγωγής
έχουμε ελαττωματικό = 0, μη ελαττωματικό = 1.

Βέβαια η εδώ ποσοτικοποίηση που επιτυγχάνεται είναι εντελώς εξωτερική. Μπορούμε να συμβολίσουμε τα θρησκευματα με άλλους αριθμούς, χωρίς να έχουμε απώλεια σε πληροφορίες. Έτσι μπορούσαμε να έχουμε Άθεος = 3, Μουσουλμάνος = 2 κ.λπ. Αν όμως ρωτούσαμε μια νοικοκυρά να κατατάξει 4 απορρυπαντικά Α, Β, Γ και Δ σαν πρώτο καλύτερο με 1, δεύτερο καλύτερο με 2, τρίτο καλύτερο με 3, και τελευταίο με 4, τότε αυτοί οι αριθμοί που αντιστοιχούν σ' αυτές τις κατατάξεις έχουν ιδιαίτερη ιεραρχική σημασία. Στην περίπτωση με τις θρησκείες θα λέμε ότι είμαστε στην **ονομαστική κλίμακα**, ενώ στη δεύτερη με τις επιδόσεις στην **τακτική κλίμακα**.

Οι **μετρήσεις** λαμβάνονται με **ερωτήσεις** (μέσω συνεντεύξεων ή γραπτών εξετάσεων), **παρατηρήσεις** (παρατηρούμε το πλήθος των αυτοκινήτων που περνούν από κάποιο σημείο κατά μία χρονική περίοδο ή παρατηρούμε το χρόνο ανamonής των πελατών στη θυρίδα μιας τράπεζας), και με **πειράματα** (μετρούμε τη διάρκεια ζωής ορισμένων ηλεκτρικών μηχανημάτων). Όλες οι μετρήσεις γίνονται σύμφωνα με κάποια κλίμακα. Εκτός από τις δύο που αναφέραμε, έχουμε την **κλίμακα διαστήματος** και την **κλίμακα πηλίκου**. Έτσι αν αναφερόμαστε σε επιδόσεις βαθμολογιών μαθητών, και χαρακτηρίζαμε αυτές σε κακή, μέτρια, καλή, πολύ καλή, άριστη, αντιστοιχίζοντας σ' αυτές τους αριθμούς 1, 2, 3, 4, 5, θα λέμε ότι βρισκόμαστε στην κλίμακα διαστήματος. Βέβαια μεταξύ της κακής επίδοσης 1 και της μέτριας 2 υπάρχει διαφορά μια μονάδα, που είναι όπως αυτής μεταξύ της πολύ καλής 4 και της άριστης 5. Φυσικά δεν μπορούμε να ισχυρισθούμε ότι αυτές με την άριστη βαθμολογία 5 είναι πέντε φορές καλύτερες απ' αυτές με τη βαθμολογία 1. Για παράδειγμα αν πολλαπλασιάσουμε κάθε επίδοση με 2 και κατόπιν προσθέσουμε τον αριθμό 2 τότε έχουμε την εξής αντιστοιχία

12	κακή
14	μέτρια
16	καλή
18	πολύ καλή
20	άριστη

Οι διαφορές μεταξύ δύο διαδοχικών επιδόσεων παραμένουν ίδιες, αλλά ο λόγος μεταξύ αυτών διαφέρουν. Έτσι ο λόγος 20 προς 12 είναι διαφορετικός απ' αυτόν του 5 προς 1. Η κλίμακα διαστήματος δεν περιέχει κάποια φυσική βάση και μπορούμε να δεχθούμε σαν τέτοια μια οποιαδήποτε τιμή. Ανάλογο παράδειγμα

μπορούμε να δώσουμε με την μέτρηση της θερμοκρασίας. Ως γνωστόν η θερμοκρασία εκφράζεται σε βαθμούς Κελσίου ή σε βαθμούς Φάρανταιϋ. Έτσι 0°C αντιστοιχούν σε 32°F που είναι διαφορετική από 0°F . Σε κάθε περίπτωση όμως το 0° δεν σημαίνει ότι δεν υπάρχει θερμοκρασία, και λέμε ότι εδώ το 0 δεν έχει φυσική σημασία. Η κλίμακα που περιέχει το φυσικό 0 ονομάζεται κλίμακα πηλίκου. Έτσι μετρώντας την ηλικία των ατόμων μπορούμε να εκφράσουμε με αριθμούς ότι το Α άτομο έχει διπλάσια ηλικία από το Β. Η τελευταία είναι η ισχυρότερη όλων των κλιμάκων και η πλέον εύχρηστη. Κάθε μέτρηση της κλίμακας πηλίκου, αποτελεί μέτρηση της κλίμακας διαστήματος, επίσης της τακτικής κλίμακας, καθώς και της ονομαστικής κλίμακας.

Σύμφωνα με τα παραπάνω εφ' όσον με μία μεταβλητή X παριστάνουμε ένα γνώρισμα, τότε αυτή μπορεί να είναι του διακριτού ή του συνεχούς τύπου, όταν και το αντίστοιχο γνώρισμα είναι διακριτό ή συνεχές.

Το πλήθος των δυνατών τιμών μιας μεταβλητής αν αυτή είναι του διακριτού τύπου μπορεί να είναι πεπερασμένο, δηλ. να παίρνει τιμές $\{a_1, \dots, a_m\}$ ή μη (αλλά αριθμήσιμο), δηλ. να παίρνει τιμές $\{a_1, \dots, a_m, a_{m+1} \dots\}$. Στην περίπτωση που η μεταβλητή μας είναι του συνεχούς τύπου, οπωσδήποτε το πλήθος των δυνατών τιμών είναι μη πεπερασμένο. Πολλές φορές δεν μας ενδιαφέρει η ακριβής τιμή μιας παρατήρησης, αλλά περισσότερο σε πιο αριθμητικό διάστημα βρίσκεται. Αυτό γίνεται συνήθως, όταν η μεταβλητή μας είναι διακριτή και με πάρα πολύ μεγάλο πλήθος τιμών ή όταν αυτή είναι του συνεχούς τύπου. Έτσι στην περίπτωση που είχαμε τη μεταβλητή X του εισοδήματος, μπορούσαμε να διακρίνουμε εισοδήματα σύμφωνα με τις κλάσεις $[0, 6.000)$, $[6.000, 12.000)$, $[12.000$ και άνω), και να μας ικανοποιούσε μόνο η γνώση του διαστήματος όπου η μεταβλητή X θα έπαιρνε τιμή. Επίσης αν εξετάζαμε την ηλικία των ατόμων μιας χώρας σε έτη, πιθανώς, θα είχαμε ηλικίες από 1 έτους έως 100 ετών. Μπορεί στην έρευνά μας να μας ενδιαφέρει περισσότερο αν η ηλικία του ατόμου είναι στο $[0, 5)$ διάστημα ή στο $[5, 10)$... ή $[95, 100)$. Σ' αυτές τις περιπτώσεις θα μιλάμε για **ομαδοποιημένες παρατηρήσεις ή μετρήσεις**, ενώ στην αντίθετη για μη ομαδοποιημένες.

Έστω ένα ποσοτικό ή ποιοτικό γνώρισμα X που παρατηρείται σε n στατιστικές μονάδες πληθυσμού. Οι αριθμοί x_1, \dots, x_n που προκύπτουν ονομάζονται **παρατηρήσεις ή δεδομένα ή μετρήσεις**. Το x_i θα σημαίνει την παρατήρηση του γνωρίσματος X στην $i^{\text{η}}$ στατιστική μονάδα. Τα δεδομένα μας μπορεί να είναι: **Διαστρωματικά**, αυτά που λαμβάνονται από μία μεταβλητή την ίδια χρονική στιγμή και πάνω σε διαφορετικές μονάδες (άτομα, χώρες, εταιρίες, κ.λπ.) τα **χρονολογικά**, αυτά που λαμβάνονται από μία μεταβλητή σε διαδοχικούς χρόνους. Όταν παρατηρούμε τα εισοδήματα 1000 νοικοκυριών σε μια χρονική στιγμή από πέντε διαφορετικές χώρες, τότε αυτά τα δεδομένα είναι διαστρωματικά. Όταν παρατηρούμε το εισόδημα ενός νοικοκυριού σε διαδοχικούς χρόνους, τότε αυτά τα δεδομένα είναι χρονολογικά. Επίσης μπορεί να υπάρξει συνδυασμός των

παραπάνω. Παρατηρούμε τα εισοδήματα των 1000 νοικοκυριών από πέντε χώρες σε διαδοχικούς χρόνους.

Άσκηση 2.1.1: Ποια από τα επόμενα στατιστικά γνωρίσματα είναι διακριτά, και ποια συνεχή;

- α) Ο αριθμός των παιδιών σ’ ένα νοικοκυριό.
- β) Η μηνιαία κατανάλωση σε ηλεκτρική ενέργεια απ’ ένα νοικοκυριό.
- γ) Ο αριθμός των πλοίων που καταφθάνουν σ’ ένα λιμάνι.
- δ) Η τιμή του χρυσού.

Άσκηση 2.1.2: Ο επόμενος πίνακας μας δίνει πληροφορίες γύρω από πέντε άτομα.

Φύλο	Μισθός	Εκπαίδευση	Έτη προϋπηρεσίας
A (άνδρας)	1.200 €	Πανεπιστημιακή	8
Γ (γυναίκα)	900 €	Μέση	4
A	1.100 €	Μέση	5
Γ	1.600 €	Πανεπιστημιακή	10
Γ	950 €	Μέση	6

- α) Ποια από τα παραπάνω γνωρίσματα είναι ποιοτικά, και ποια ποσοτικά.
- β) Σε ποια στατιστικά γνωρίσματα θα χρησιμοποιήσετε την τακτική κλίμακα και σε ποια την ονομαστική.

Άσκηση 2.1.3: Ποια από τα επόμενα στατιστικά γνωρίσματα μας δίνουν χρονολογικά δεδομένα ή διαστρωματικά.

- α) Η τιμή μιας μετοχής ανά μέρα τα δύο τελευταία χρόνια.
- β) Ο αριθμός των τηλεφωνημάτων που έγινε από κάθε νοικοκυριό χθές.
- γ) Η τιμή των σιτηρών τα τελευταία 50 χρόνια.
- δ) Οι καταθέσεις σε μια τράπεζα που κάνουν 60 άτομα.

2.2. Κατανομές Συχνοτήτων

Η μεμονωμένη γνώση των δεδομένων μας δίνουν μια συγκεχυμένη εικόνα του δείγματός μας.

Παράδειγμα 2.2.1: Παρατηρήθηκε ότι ο κύριος προμηθευτής μιας επιχείρησης για 50 παραγγελίες χρειάστηκε τους εξής χρόνους (σε ημέρες) διεκπεραίωσης των.

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

Για το γνώρισμα X [= χρόνος παράδοσης σε ημέρες] παρατηρήθηκαν τα εξής:
 $x_1 = 4$, $x_2 = 5$, $x_3 = 3$, $x_4 = 1$, ..., $x_{49} = 7$, $x_{50} = 5$.

Το ενδιαφέρον μας εδώ εστιάζεται στο να βγάλουμε κάποιο συμπέρασμα, στο κατά πόσο γρήγορα διεκπεραιώνονται οι παραγγελίες από τον εν λόγω προμηθευτή. Έτσι η διάσπαρτη γνώση των δεδομένων μας δεν μας βοηθά αποτελεσματικά σ' αυτό το σκοπό.

Για να έχουμε μια καλύτερη εικόνα των δεδομένων είναι απαραίτητο να γνωρίζουμε με τι τρόπο κατανέμονται αυτές οι τελευταίες τιμές. Έτσι προχωρούμε στην ανάπτυξη των επομένων εννοιών.

2.3. Απόλυτη και Σχετική Συχνότητα

Σε μια λίστα n δεδομένων θεωρούμε τις $\alpha_1, \dots, \alpha_k$ δυνατές τιμές, που τις θεωρούμε εξ αρχής διατεταγμένες

$$\alpha_1 < \alpha_2 < \dots < \alpha_k$$

Το πλήθος εμφάνισης του α_j καλείται **απόλυτη συχνότητα** και συμβολίζεται με $h(\alpha_j)$. Ενώ το πηλίκο $f(\alpha_j) = \frac{1}{n} h(\alpha_j)$ θα καλείται **σχετική συχνότητα**. Ισχύουν τα ακόλουθα:

$$\sum_{j=1}^k h(\alpha_j) = n, \quad \text{και} \quad \sum_{j=1}^k f(\alpha_j) = 1. \tag{2.3.1}$$

Σύμφωνα με τις τελευταίες έννοιες τα δεδομένα του παραδείγματος 2.2 μπορούν να παρασταθούν υπό μορφή **πίνακα συχνοτήτων**. Έτσι έχουμε το ακόλουθο:

Παράδειγμα 2.3.1: Να βρεθούν οι απόλυτες και σχετικές συχνότητες των τιμών της μεταβλητής X του Παραδείγματος 2.2.1.

Δυνατές τιμές α_j	$\alpha_1=1$	$\alpha_2=2$	$\alpha_3=3$	$\alpha_4=4$	$\alpha_5=5$	$\alpha_6=6$	$\alpha_7=7$
Απόλυτη Συχνότητα $h(\alpha_j)$	1	1	2	12	17	9	8
Σχετική Συχνότητα $f(\alpha_j)$	$\frac{1}{50}$	$\frac{1}{50}$	$\frac{2}{50}$	$\frac{12}{50}$	$\frac{17}{50}$	$\frac{9}{50}$	$\frac{8}{50}$

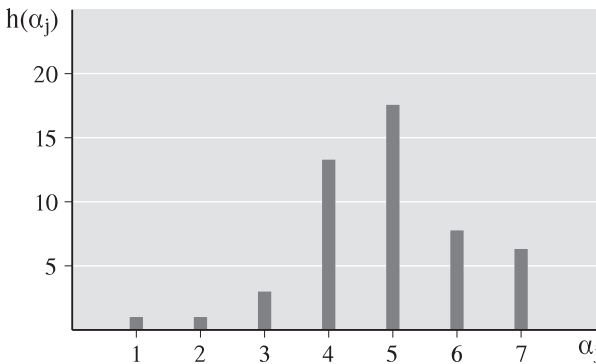
Σαν **κατανομή συχνότητας** θα ονομάζουμε τη διάταξη των συχνοτήτων (απόλυτων ή σχετικών) των διαφόρων τιμών της μεταβλητής X .

Οι συνηθισμένες γραφικές μέθοδοι των κατανομών συχνοτήτων είναι οι εξής:

- **Πίνακας Συχνοτήτων**
- **Ακιδωτά Διαγράμματα**
- **Κυκλικά Διαγράμματα**
- ή στην περίπτωση ομαδοποιημένων δεδομένων τα **ιστογράμματα**.

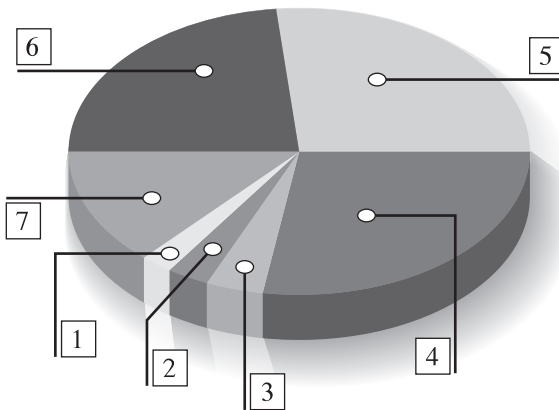
Ο τρόπος παρουσίασης των δεδομένων του Παραδείγματος 2.3.1 έγινε με τον πίνακα συχνοτήτων.

Ένα ακιδωτό διάγραμμα λαμβάνεται σ' ένα ορθογώνιο σύστημα συντεταγμένων, τοποθετώντας τα k σημεία $\{\alpha_j, h(\alpha_j) \text{ ή } f(\alpha_j)\}$ σ' αυτό. Έτσι το ακιδωτό διάγραμμα του Παραδείγματος 2.3.1, είναι το σχήμα 2.3.1.



Σχήμα 2.3.1

Σ' ένα κυκλικό διάγραμμα θα θεωρούμε τις συχνότητες σαν εμβαδά κυκλικών τομέων, όπου το κάθε εμβαδόν θα είναι ανάλογο της συχνότητας που θα παριστάνει. Το κυκλικό διάγραμμα του Παραδείγματος 2.3.1 δίνεται από το σχήμα 2.3.2.



Σχήμα 2.3.2

Άσκηση 2.3.1: Σ' ένα εργοστάσιο ρωτήθηκαν 40 εργαζόμενοι με τι τρόπο πηγαινούν στη δουλειά τους. Είχαμε τις ακόλουθες παρατηρήσεις

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

όπου

1= Δημόσιο μέσο, 2 = Ιδιωτικό, 3 = Μηχανή,
4 = ποδήλατο, 5 = οδοιπορικώς.

Να κατασκευασθούν: το ακιδωτό διάγραμμα, το κυκλικό διάγραμμα, καθώς και ο πίνακας συχνοτήτων.

Άσκηση 2.3.2: Να σχεδιασθεί το κυκλικό διάγραμμα των επόμενων δεδομένων:

Μέσα Μηνιαία Έξοδα σε ευρώ.

Είδος	1988	1990
Τρόφιμα	640	840
Ρούχα	200	280
Ενοίκιο	420	520
Διάφορα	170	300

2.4. Ιστογράμματα

Σε πολλές στατιστικές αναλύσεις η απαρίθμηση των δυνατών τιμών μιας μεταβλητής δεν είναι δυνατή ή και αν είναι δεν έχει νόημα, επειδή

- α) Το πλήθος των δυνατών τιμών είναι πολύ μεγάλο.
- β) Η μεταβλητή είναι συνεχής.
- γ) Και αν ακόμη χρησιμοποιούσαμε τον πίνακα συχνοτήτων ή το ακιδωτό διάγραμμα δεν θα είχαμε καμία σαφή εικόνα των δεδομένων μας.

Τα προηγούμενα διαφωτίζονται με το επόμενο παράδειγμα.

Παράδειγμα 2.4.1: Οι μηνιαίες αποδοχές 100 υπαλλήλων μιας εταιρείας το 1987, είναι

1.600,30	1.819,20	1.920,30	1.970,30	2.091,20
1.632,40	1.819,80	1.926,90	1.979,40	2.091,20
1.673,90	1.821,70	1.928,30	1.982,80	2.096,20
1.685,00	1.825,50	1.929,90	1.996,00	2.097,40
1.711,00	1.826,80	1.939,90	2.000,30	2.100,10
1.715,00	1.839,10	1.939,40	2.006,50	2.109,20
1.722,70	1.840,30	1.940,70	2.017,40	2.111,20
1.738,70	1.848,80	1.942,30	2.025,80	2.119,70
1.742,00	1.852,20	1.942,30	2.039,50	2.140,80
1.751,00	1.857,30	1.943,50	2.043,30	2.149,20
1.751,60	1.859,60	1.948,80	2.049,10	2.149,60
1.768,20	1.872,20	1.950,10	2.049,70	2.158,90
1.771,80	1.883,10	1.952,60	2.050,40	2.159,70
1.782,20	1.886,30	1.952,60	2.050,90	2.162,20
1.784,30	1.892,10	1.952,60	2.058,20	2.186,80
1.789,20	1.893,60	1.958,70	2.069,40	2.198,70
1.790,40	1.901,20	1.959,20	2.078,60	2.223,70
1.791,70	1.913,50	1.960,80	2.082,70	2.248,40
1.800,50	1.913,70	1.969,20	2.083,50	2.269,80
1.817,90	1.914,90	1.969,99	2.083,90	2.296,50

Τότε το ακιδωτό διάγραμμα στο σχήμα 2.4.1 δεν μας δίνει καμιά πληροφορία για τις αποδοχές των υπαλλήλων. Απλά συμπεραίνουμε ότι κάθε τιμή εμφανίζεται μια φορά (βλ. Παράδειγμα 2.4.2).



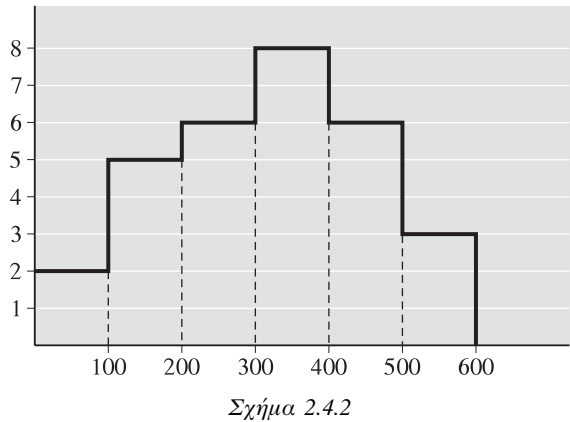
Σχήμα 2.4.1

Σ’ αυτές τις περιπτώσεις ενδείκνυται να χωρίσουμε τον άξονα της μεταβλητής X σε k διαδοχικά διαστήματα, όπου το j -οστό διάστημα είναι το $[α_{j-1}, α_j)$, και η απόλυτη συχνότητα $h(α_j)$ του διαστήματος $[α_{j-1}, α_j)$ ορίζεται σαν το πλήθος των παρατηρήσεων που περιλαμβάνονται σ’ αυτό. Κατόπιν δημιουργούμε ορθογώνια παραλληλόγραμμα πάνω σε κάθε διάστημα εμβαδού ανάλογου της απόλυ-

της συχνότητας του διαστήματος. Η γραφική παράσταση που προκύπτει ονομάζεται **Ιστόγραμμα**. Έτσι αν είχαμε τον ακόλουθο πίνακα συχνοτήτων

Διαστήματα	Απόλυτες Συχνότητες
[0, 100)	2
[100, 200)	5
[200, 300)	6
[300, 400)	8
[400, 500)	6
[500, 600)	3

λαμβάνουμε το παρακάτω Ιστόγραμμα (σχ. 2.4.2)



Τα ερωτήματα που δημιουργούνται κατά τη κατασκευή ενός Ιστογράμματος είναι τα εξής: Πρώτον, ποιο είναι το κατάλληλο μήκος των διαστημάτων, και αν αυτά πρέπει να είναι ίσα. Πόσα διαστήματα πρέπει να λάβουμε; Από που πρέπει να ξεκινήσουμε τη σκιαγράφησή του;

Στην περίπτωση που δεχτούμε ότι έχουμε διαστήματα ίσου μήκους, τότε το πλήθος k αυτών πρέπει να είναι ένας αριθμός μεταξύ 5 και 20 ή να ικανοποιεί τη σχέση $k = 1 + 3,3 \ln n$, όπου n το πλήθος των δεδομένων μας. Το μήκος ℓ του διαστήματος δίνεται από τη σχέση

$$\ell = \frac{\text{μέγιστο}\{x_i\} - \text{ελάχιστο}\{x_i\}}{k}.$$

Σε αρκετές περιπτώσεις λαμβάνουμε διαστήματα ανίσου μήκους, επειδή υπάρχει ο κίνδυνος στην περίπτωση ίσων διαστημάτων να υπάρξουν τέτοια που να μην συμπεριλαμβάνουν καμία μέτρηση.

Παράδειγμα 2.4.2: Να κατασκευασθεί το Ιστόγραμμα του Παραδείγματος 2.4.1.

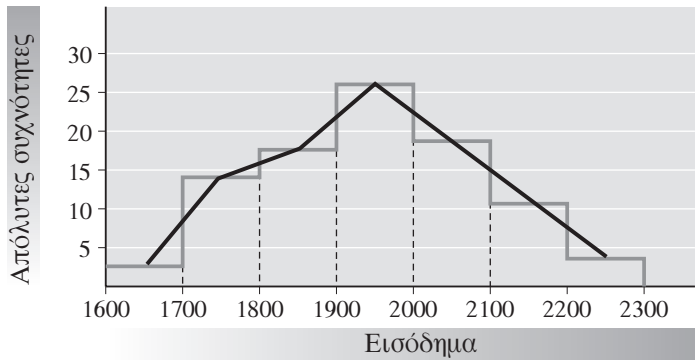
Λύση: Από τη σχέση $k = 1+3,3 \ln n$, έχουμε για $n=100$, ότι $k \approx 7$, ενώ

$$\ell = \frac{229.650 - 160.030}{7} = 9,95 \approx 10.000.$$

Αν θεωρήσουμε σαν αρχή το 1.600, τότε λαμβάνουμε τον εξής πίνακα συχνοτήτων

Διαστήματα	Απόλυτες Συχνότητες
[1.600, 1.700)	4
[1.700, 1.800)	14
[1.800, 1.900)	18
[1.900, 2.000)	28
[2.000, 2.100)	20
[2.100, 2.200)	12
[2.200, 2.300)	4

Με βάση τον πίνακα συχνοτήτων έχουμε την εξής γραφική παράσταση του Ιστογράμματος (σχ. 2.4.3).



Σχήμα 2.4.3

Αν ενώσουμε τα μέσα των πλευρών των ορθογώνιων που είναι παράλληλες προς τον άξονα των εισοδημάτων σχηματίζουμε το λεγόμενο **Πολύγωνο Συχνοτήτων**.

Αν αντί για τις απόλυτες συχνότητες θεωρούσαμε τις σχετικές συχνότητες, τότε το Ιστόγραμμα που θα προκύψει θα έχει εμβαδόν ίσον με μονάδα, όταν το μήκος κάθε διαστήματος θεωρείται μοναδιαίο.

Άσκηση 2.4.1: Είκοσι φοιτητές ρωτήθηκαν για τα μηνιαία έξοδά τους και είχαμε

1000, 580, 520, 350, 620, 800, 120, 600, 550, 420, 470, 200, 560, 480,
1000, 600, 1150, 800, 250, 650.

Να σχηματίσετε το Ιστόγραμμα, καθώς και το Πολύγωνο Συχνοτήτων.

2.5. Αθροιστικές Κατανομές Συχνοτήτων

Πολλές φορές μας ενδιαφέρει να γνωρίζουμε πόσες παρατηρήσεις είναι ίσες, ή μικρότερες ή μεγαλύτερες από κάποιον αριθμό. Επίσης, ποιο ποσοστό παρατηρήσεων είναι ίσο ή μικρότερο ή μεγαλύτερο από κάποιον αριθμό. Οι απαντήσεις σ’ αυτά τα ερωτήματα δίδονται με τη βοήθεια της αθροιστικής κατανομής συχνοτήτων. Η συνάρτηση $H(x) = \sum_{j: \alpha_j \leq x} h(x)$ ονομάζεται **απόλυτη αθροιστική κατανομή**

συχνοτήτων, ενώ η $F(x) = \frac{H(x)}{n}$ **σχετική αθροιστική κατανομή συχνοτήτων**. Για την περίπτωση μη ομαδοποιημένων παρατηρήσεων δημιουργούμε τον ακόλουθο πίνακα.

Τιμή Μεταβλητής	Απόλυτη Συχνότητα	Σχετική Συχνότητα	Απόλυτη Αθροιστική Κατανομή	Σχετική Αθροιστική Κατανομή
α_1	$h_1 = h(\alpha_1)$	$f_1 = f(\alpha_1)$	$H_1 = H(\alpha_1) = h_1$	$F_1 = f_1$
α_2	$h_2 = h(\alpha_2)$	$f_2 = f(\alpha_2)$	$H_2 = H(\alpha_2) = h_1 + h_2$	$F_2 = f_1 + f_2$
\vdots	\vdots	\vdots	\vdots	\vdots
α_k	$h_k = h(\alpha_k)$	$f_k = f(\alpha_k)$	$H_k = H(\alpha_k) = h_1 + h_2 + \dots + h_k$	$F_k = f_1 + \dots + f_k$

Για την περίπτωση των ομαδοποιημένων παρατηρήσεων έχουμε τον ίδιο πίνα-

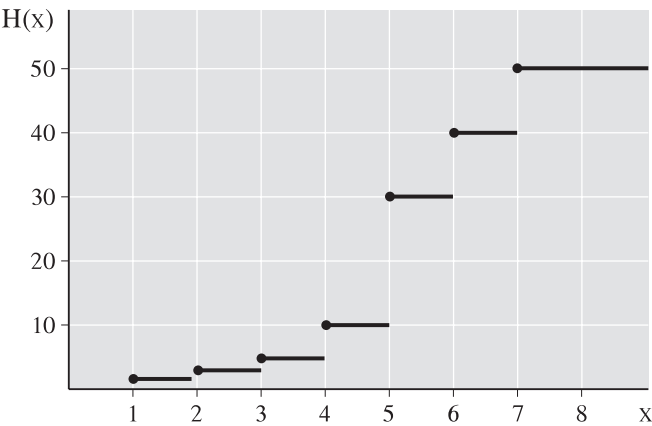
κα, βάζοντας όπου α_{j-1} το αριστερό άκρο του $[\alpha_{j-1}, \alpha_j)$, και h_j, f_j τις απόλυτες και σχετικές συχνότητες των διαστημάτων, αντίστοιχα.

Παράδειγμα 2.5.1: Να βρεθεί η αθροιστική κατανομή συχνοτήτων (απόλυτη και σχετική) του Παραδείγματος 2.2.1.

Λύση: Η απόλυτη κατανομή συχνότητας $H(x)$ συμβολίζει το πλήθος των παραγγελιών που χρειάστηκαν χρόνο λιγότερο ή ίσο με x . Έτσι οι τιμές της θα εξαρτώνται ανάλογα σε ποιο διάστημα θα κυμαίνεται το x .

$$H(x) = \begin{cases} 0 & x < 1 & \text{(καμία παραγγελία δεν έγινε σε χρόνο μικρότερο του 1)} \\ 1 & 1 \leq x < 2 & \text{(μία παραγγελία έγινε σε χρόνο μεγαλύτερο ή ίσο του 1)} \\ 2 & 2 \leq x < 3 & \text{και μικρότερο του 2)} \\ 4 & 3 \leq x < 4 \\ 16 & 4 \leq x < 5 \\ 33 & 5 \leq x < 6 \\ 42 & 6 \leq x < 7 \\ 50 & 7 \leq x \end{cases}$$

Η γραφική παράσταση της $H(x)$ δίνεται στο σχήμα 2.5.1.



Σχήμα 2.5.1

Άσκηση 2.5.1: Σε μια ασφαλιστική εταιρεία 100 άτομα κάνουν ασφάλεια ζωής με τον εξής τρόπο.

Ηλικία	Ποσά	Απόλυτες συχνότητες	Σχετικές συχνότητες
[20, 30)	70.000	26	$\frac{26}{100}$
[30, 40)	60.000	33	$\frac{33}{100}$
[40, 50)	40.000	21	$\frac{21}{100}$
[50, 60)	25.000	14	$\frac{14}{100}$
[60, 70)	10.000	6	$\frac{6}{100}$

- 1) Να γίνουν τα ιστογράμματα και τα ακιδωτά διαγράμματα.
- 2) Ποιό ποσοστό ασφαλισμένων είναι 50 ετών και γηραιότερο;
- 3) Ποιό ποσοστό είναι ασφαλισμένο για 40.000 και λιγότερο;
- 4) Να βρεθούν οι αθροιστικές κατανομές.

Άσκηση 2.5.2: Ποια η αθροιστική κατανομή απόλυτης ή (σχετικής) συχνότητας του Παραδείγματος 2.4.1.

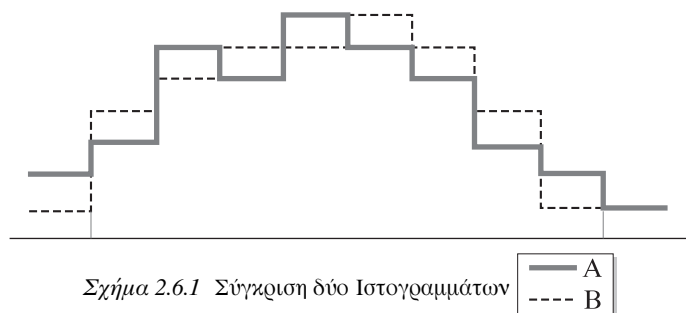
Άσκηση 2.5.3: Οι 150 επιβάτες των λεωφορείων μιας γραμμής ρωτήθηκαν για το χρόνο αναμονής τους στη στάση. Λάβαμε τα εξής δεδομένα:

Χρόνος αναμονής: (σε λεπτά)	0–1	1–2	2–3	3–4	4–5	5–6	6–7	7–8	8–9	9–10	10–11	11–12
Απόλυτη συχνότητα:	3	9	18	30	24	21	12	9	3	6	9	6

- 1) Να υπολογισθεί η αθροιστική κατανομή συχνοτήτων.
- 2) Η συνάρτηση $A(x) = \sum_{j: a_j > x} h(x) = n - H(x)$ ονομάζεται αφαιρετική κατανομή
συχνοτήτων. Στην άσκηση αυτή παριστάνει το πλήθος των επιβατών που ανέ-
μειναν περισσότερο από x λεπτά.

2.6. Μέτρα Θέσης (ή Κεντρικής Τάσης)

Έστω ότι ενδιαφερόμαστε να συγκρίνουμε τους μισθούς των εργαζομένων μεταξύ δύο επιχειρήσεων Α και Β. Μπορούμε να δημιουργήσουμε τα ιστογράμματα ή τα πολύγωνα συχνотήτων των μισθών και να συγκρίνουμε αυτά. Αυτού του είδους η σύγκριση είναι αρκετά δύσκολη, π.χ. ενδέχεται να λαμβάναμε τα ιστογράμματα του σχήματος 2.6.1. Από τη σύγκριση αυτών δεν έχουμε σαφή ένδειξη κατά πόσο οι μισθοί των εργαζομένων στην Α επιχείρηση είναι μικρότεροι ή μεγαλύτεροι των αντίστοιχων εργαζομένων στην Β επιχείρηση. Στο συγκεκριμένο παράδειγμα το ευκολότερο που μπορούσαμε να κάνουμε είναι να συγκρίνουμε τους μέσους μισθούς.



Στην Περιγραφική Στατιστική ενδιαφέρων παρουσιάζει στο να υπολογίζουμε κάποιες αριθμητικές χαρακτηριστικές τιμές των κατανομών συχνотήτων, που θα μας δίνουν μια “συνολτική” πληροφορία για τα δεδομένα μας. Θα εξετάσουμε με τη σειρά που παρατίθενται τον Αριθμητικό Μέσο, τη Διάμεσο, το Γεωμετρικό Μέσο, και τέλος την Επικρατούσα Τιμή.

α. Αριθμητικός Μέσος

Αν διαθέτουμε τα δεδομένα x_1, \dots, x_n , τότε ο (απλός) αριθμητικός μέσος \bar{x} , ορίζεται από τη σχέση

$$\bar{x} = \frac{x_1 + \dots + x_n}{n} .$$

Αν επιπλέον παρατηρήσουμε ότι h_1 από τα δεδομένα μας έχουν την τιμή a_1 , h_2 την τιμή a_2 , ..., h_k την τιμή a_k , τότε ο (σταθμικός) αριθμητικός μέσος ορίζεται από τη σχέση

$$\bar{x} = \frac{\sum_{j=1}^k h_j \alpha_j}{\sum_{i=1}^k h_i} = \frac{\sum_{j=1}^k h_j \alpha_j}{n} \quad (\text{Άθροισμα γινομένων απολύτων συχνοτήτων με τις αντίστοιχες δυνατές τιμές του γνωρίσματος, διαιρούμενου με το πλήθος των δεδομένων})$$

ή

$$\bar{x} = \sum_{j=1}^k f_j \alpha_j \quad (\text{Άθροισμα γινομένων σχετικών συχνοτήτων με τις αντίστοιχες δυνατές τιμές του γνωρίσματος})$$

Ο τελευταίος ορισμός εφαρμόζεται όταν δεν έχουμε ομαδοποιημένες παρατηρήσεις, και παρατηρούμε τη συχνότητα κάθε τιμής. Στην περίπτωση που έχουμε ομαδοποιημένες παρατηρήσεις, εφαρμόζουμε ακριβώς τους ίδιους τύπους με τη διαφορά ότι αντί για α_j θέτουμε την κεντρική τιμή του διαστήματος $[\alpha_{j-1}, \alpha_j]$, δηλαδή την $\frac{\alpha_{j-1} + \alpha_j}{2}$. Σ' αυτή την περίπτωση ο αριθμητικός μέσος υπολογίζεται προσεγγιστικά.

Παρατήρηση 2.6.1:

- α) Δεν πρέπει να ξεχνάμε ότι ο αριθμητικός μέσος που ορίσθηκε εδώ είναι αυτός του δείγματος, και όχι όλου του πληθυσμού. Συνήθως το μέσο του πληθυσμού θα τον συμβολίζουμε με μ , και παραμένει άγνωστος στον ενδιαφερόμενο ερευνητή.
- β) Αν είχαμε δύο διαφορετικά σύνολα δεδομένων, ώστε στο πρώτο να είχαμε n_1 παρατηρήσεις και με μέσο \bar{x}_1 , ενώ στο δεύτερο να είχαμε n_2 παρατηρήσεις και με μέσο \bar{x}_2 , τότε ο ολικός αριθμητικός μέσος και των δυο δεδομένων \bar{x} ορίζεται από τη σχέση,

$$\bar{x} = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2}{n_1 + n_2}.$$

Παράδειγμα 2.6.1: Ερωτήθηκαν 30 υπάλληλοι μιας επιχείρησης για την ηλικία τους και είχαμε τις εξής ηλικίες

24, 24, 40, 22, 32, 51, 41, 22, 42, 43, 44, 51, 23, 30, 20, 32, 34, 64, 19, 23, 22, 50, 50, 33, 60, 40, 20, 50, 42, 41

- α) Να βρεθεί ο μέσος χρόνος ηλικίας των.
- β) Ποια είναι η μέση ηλικία, αν ομαδοποιήσουμε τις ηλικίες ανά 10 χρόνια. Για τα ανοικτά διαστήματα λάβατε σαν κεντρικές τιμές $x_1=18$ και $x_6=63$.

Λύση: α) Εδώ το γνώρισμα είναι η ηλικία, και έχουμε τις εξής δυνατές τιμές α_j

και απόλυτες συχνότητες h_j

$\alpha_1 = 19,$	$h_1 = 1$	$\alpha_{10} = 40,$	$h_{10} = 2$
$\alpha_2 = 20,$	$h_2 = 2$	$\alpha_{11} = 41,$	$h_{11} = 2$
$\alpha_3 = 22,$	$h_3 = 3$	$\alpha_{12} = 42,$	$h_{12} = 2$
$\alpha_4 = 23,$	$h_4 = 2$	$\alpha_{13} = 43,$	$h_{13} = 1$
$\alpha_5 = 24,$	$h_5 = 2$	$\alpha_{14} = 44,$	$h_{14} = 1$
$\alpha_6 = 30,$	$h_6 = 1$	$\alpha_{15} = 50,$	$h_{15} = 3$
$\alpha_7 = 32,$	$h_7 = 2$	$\alpha_{16} = 51,$	$h_{16} = 2$
$\alpha_8 = 33,$	$h_8 = 1$	$\alpha_{17} = 60,$	$h_{17} = 1$
$\alpha_9 = 34,$	$h_9 = 1$	$\alpha_{18} = 64,$	$h_{18} = 1$

Άρα ο αριθμητικός μέσος υπολογίζεται ως εξής:

$$\bar{x} = \frac{h_1\alpha_1 + \dots + h_{18}\alpha_{18}}{\sum_{i=1}^{18} h_i} = 36,3 \text{ .}$$

β) Εύκολα σχηματίζουμε τον επόμενο πίνακα

Ηλικία	Κεντρική Τιμή	Απόλυτες Συχνότητες
Κάτω από 20	18	1
[20, 30)	25	9
[30, 40)	35	5
[40, 50)	45	8
[50, 60)	55	5
πάνω από 60	63	2

Ο αριθμητικός μέσος εδώ είναι $\bar{x} = 39,4$, που αποτελεί μια προσεγγιστική τιμή του πραγματικού αριθμητικού μέσου.

β. Διάμεσος

Κάθε αριθμός $\tilde{x}_{n/2}$ που χωρίζει τα δεδομένα μας σε δύο ίσα στο πλήθος μέρη δεδομένων, όταν αυτά διαταχθούν σ' αύξουσα σειρά μεγέθους, ονομάζεται **διάμεσος**. Αν είχαμε τις μεμονωμένες τιμές x_1, \dots, x_n , και τις διατάσσαμε ως εξής, $x_{(1)} \leq x_{(2)} \leq \dots x_{(n)}$, όπου $x_{(1)}$ είναι μικρότερη τιμή, $x_{(2)}$ η αμέσως μεγαλύτερη τιμή, ... και $x_{(n)}$ η μέγιστη τιμή, τότε η διάμεσος $\tilde{x}_{n/2}$ ορίζεται από την ακόλουθη σχέση.

$$\tilde{x}_{n/2} = \begin{cases} x_{\left(\frac{n+1}{2}\right)} & , \text{ αν } n \text{ περιττός} \\ \frac{1}{2} \left[x_{\left(\frac{n}{2}\right)} + x_{\left[\left(\frac{n}{2}\right)+1\right]} \right] & , \text{ αν } n \text{ άρτιος} \end{cases}$$

Έτσι η διάμεσος των δεδομένων 3, 2, 4, 10 είναι $\tilde{x}_{n/2} = 3,5$, επειδή αν διατάξουμε αυτές έχουμε 2, 3, 4, 10, οπότε το 3,5 χωρίζει τα δεδομένα μας στο αυτό το πλήθος μέρη.

Στη συνέχεια αν τα δεδομένα μας είναι ομαδοποιημένα, τότε εντοπίζουμε το διάστημα που πιθανόν να βρίσκεται η διάμεσός μας. Αν αυτό ήταν το $[\alpha_{j-1}, \alpha_j]$, θα πρέπει να ισχύει η σχέση

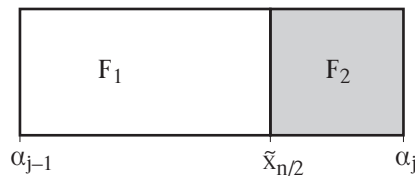
$$H_{j-1} \leq \frac{n}{2} < H_j \quad \text{ή} \quad F_{j-1} \leq \frac{1}{2} < F_j,$$

όπου H_j και F_j είναι οι τιμές των αλθιοιστικών κατανομών συχνοτήτων, απόλυτων ή σχετικών. Αν υποθέσουμε ότι κάθε διάστημα $[\alpha_{j-1}, \alpha_j]$ περιέχει ομοιόμορφα τις μετρήσεις μας, τότε μπορεί να δειχθεί ότι

$$\tilde{x}_{n/2} = \alpha_{j-1} + \frac{\alpha_j - \alpha_{j-1}}{h_j} \left(\frac{n}{2} - H_{j-1} \right)$$

ή

$$\tilde{x}_{n/2} = \alpha_{j-1} + \frac{\alpha_j - \alpha_{j-1}}{f_j} \left(\frac{1}{2} - F_{j-1} \right)$$



Σχήμα 2.6.2

Μια απόδειξη της τελευταίας σχέσης αποτελεί το παρακάτω:

Έστω ότι F_1 παρατηρήσεις ότι βρίσκονται ομοιόμορφα στο διάστημα $[\alpha_{j-1}, m_n]$, και ανάλογα F_2 παρατηρήσεις στο $[\tilde{x}_{n/2}, \alpha_j]$. Τότε

$$H_{j-1} + F_1 = H(\alpha_{j-1}) + F_1 = \frac{n}{2} \quad \text{και} \quad H_j - F_2 = H(\alpha_j) - F_2 = \frac{n}{2}.$$

Έτσι αν h_j είναι το πλήθος των παρατηρήσεων που βρίσκονται στο διάστημα

$[a_{j-1}, a_j]$, ισχύει η $\frac{F_1}{\tilde{x}_{n/2}-a_{j-1}} = \frac{h_j}{a_j-a_{j-1}}$, επειδή οι παρατηρήσεις κατανέμονται

ομοιόμορφα. Από την τελευταία έχουμε $F_1 = \frac{\tilde{x}_{n/2}-a_{j-1}}{a_j-a_{j-1}} h_j$ που σε συνδυασμό με τη σχέση $H_{j-1}+F_1 = \frac{n}{2}$, μας δίνει τη σχέση που επρόκειτο να δειχθεί.

Στην περίπτωση που τα δεδομένα μας είναι από ποιοτική μεταβλητή και χρησιμοποιούμε την τακτική κλίμακα ενδείκνυται η χρήση της διαμέσου.

Παράδειγμα 2.6.2: α) Σε ένα διαγωνισμό είχαμε τις ακόλουθες επιδόσεις,

8, 8, 9, 10, 10, 11, 11, 12, 12, 12, 13, 13, 13, 14, 18.

Ποιά η διάμεσος;

β) Δίδονται τα ακόλουθα 3000 εισοδήματα

Εισοδήματα	Απόλυτες Συχνότητες	Αθροιστική Κατανομή Απόλυτων Συχνοτήτων
[0, 500)	400	400
[500, 1000)	800	1200
[1000, 2000)	600	1800
[2000, 4000)	1200	3000

Ποιά η διάμεσος;

Λύση: α) Εδώ $\tilde{x}_{n/2}=12$ και μάλιστα το δεύτερο στη σειρά 12 από τα δεδομένα μας.

β) Από τον τύπο της διαμέσου,

$$\tilde{x}_{n/2} = 1000 + \frac{2000-1000}{600}(1500-1200) = 1500 ,$$

Επειδή το μισό του $n=3000$ είναι το 1500 και βρίσκεται στο διάστημα $[1000, 2000]$. Άρα εδώ $a_{j-1}=1000$, $a_j=2000$, $h_j=600$, $H_{j-1}=1200$ με $n=3000$.

γ. Γεωμετρικός μέσος

Σε πολλές περιπτώσεις δεν μπορούμε να χρησιμοποιήσουμε τον αριθμητικό μέσο ή τη διάμεσο. Αν για παράδειγμα η τιμή μιας μετοχής είναι 100 €, τον επόμενο χρόνο γίνει 120 €, το μεθεπόμενο 150, και καταλήξει στον τρέχοντα χρόνο 100 €, τότε τίθεται το ερώτημα ποια είναι η μέση ποσοστιαία μεταβολή της μετοχής. Προφανώς είναι 0. Όμως αν χρησιμοποιούσαμε έναν από τους αναφερόμενους μέσους δεν θα λαμβάναμε αυτήν την τιμή. Γενικά σε περιπτώσεις μετα-

βολής πληθυσμών, κεφαλαίων κ.λπ. ενδείκνυται η χρησιμοποίηση του Γεωμετρικού μέσου για τον υπολογισμό της μέσης ποσοστιαίας μεταβολής. Ο τελευταίος ορίζεται ως εξής:

Στην περίπτωση μεμονωμένων δεδομένων, x_1, \dots, x_n , σαν τον αριθμό G , που δίνεται από τη σχέση

$$G = \sqrt[n]{x_1 \dots x_n}.$$

Ενώ αν h_1 από τα δεδομένα μας έχουν την τιμή a_1 , h_2 από τα δεδομένα μας έχουν την τιμή a_2, \dots, h_k από τα δεδομένα μας έχουν την τιμή a_k , σαν τον αριθμό G , που δίνεται από τη σχέση

$$G = \sqrt[h]{a_1^{h_1} \dots a_k^{h_k}}, \quad \text{όπου } h_1 + h_2 + \dots + h_k = n.$$

Αν τα δεδομένα μας είναι ομαδοποιημένα χρησιμοποιούμε την τελευταία σχέση θέτοντας όπου a_j τις κεντρικές τιμές των διαστημάτων $[a_{j-1}, a_j]$.

ΕΦΑΡΜΟΓΗ 2.6.1

Αν σε n διαδοχικές χρονικές μονάδες έχουμε τις αντίστοιχες ποσοστιαίες μεταβολές r_1, \dots, r_n ενός κεφαλαίου k_0 , τότε η μέση ποσοστιαία μεταβολή r δίνεται από τη σχέση

$$1+r = \sqrt[n]{(1+r_1)(1+r_2)\dots(1+r_n)}$$

Απόδειξη: Η r_j , $j=1, \dots, n$, ποσοστιαία μεταβολή ορίζεται από τη σχέση

$$r_j = \frac{k_j - k_{j-1}}{k_{j-1}} = \frac{k_j}{k_{j-1}} - 1$$

όπου k_j συμβολίζει το κεφάλαιο στο τέλος της $j^{\text{ης}}$ χρονικής μονάδας. Από την τελευταία έχουμε

$$k_j = k_{j-1} (1+r_j), \quad j = 1, \dots, n$$

και συνεπώς

$$k_n = k_0 (1+r_1)(1+r_2) \dots (1+r_n).$$

Επίσης αν r η μέση ποσοστιαία μεταβολή στις n διαδοχικές χρονικές μονάδες, τότε το κεφάλαιο k_n θα δίνεται και από τη σχέση

$$k_n = k_0 (1+r)^n = (k_0 (1+r) \dots (1+r)).$$

Εξισώνοντας τις δύο τελευταίες σχέσεις έχουμε

$$1+r = \sqrt[n]{(1+r_1)(1+r_2)\dots(1+r_n)}.$$

Παράδειγμα 2.6.3: Ένα κεφάλαιο αυξάνει κατά τον εξής τρόπο: Από 100.000, γίνεται 180.000 τον επόμενο χρόνο, και στο τέλος του μεθεπόμενου 198.000. Ποια είναι η μέση ποσοστιαία μεταβολή του κεφαλαίου;

Λύση: Εδώ $r_1=0,8$, $r_2=0,1$, και συνεπώς από την

$$1+r = \sqrt{1,8 \cdot 1,1} = 1,40,$$

έχουμε $r=0,40$.

Παρατήρηση 2.6.2: Ο γεωμετρικός μέσος είναι μικρότερος του αριθμητικού μέσου.

δ. Επικρατούσα Τιμή

Μεταξύ των μεμονωμένων μετρήσεων x_1, \dots, x_n , η τιμή εκείνη που εμφανίζεται πιο συχνά ονομάζεται επικρατούσα τιμή και τη συμβολίζουμε με το γράμμα d . Συμβολικά $h(d) = \max_j h(x_j)$. Υπάρχει το ενδεχόμενο να έχουμε περισσότερες

από μία επικρατούσα τιμή. Η επικρατούσα τιμή ενδείκνυται να χρησιμοποιείται, όταν έχουμε δεδομένα από ποιοτικά χαρακτηριστικά, και εφαρμόζουμε την ονομαστική κλίμακα.

Όταν τα δεδομένα είναι ομαδοποιημένα εμφανίζεται στο διάστημα με τις περισσότερες παρατηρήσεις.

ε. Ποσοστιαία σημεία

Αν έχουμε n παρατηρήσεις x_1, \dots, x_n και τις διατάξουμε σε αύξουσα σειρά μεγέθους, τότε το p -οστό ποσοστιαίο σημείο είναι μία τιμή \tilde{x}_p , $0 < p < 1$, κάτω από την οποία υπάρχουν np των παρατηρήσεων, ενώ $(1-p)n$ παρατηρήσεις βρίσκονται άνω.

Το 25° ποσοστιαίο σημείο $\left(p = \frac{1}{4}\right)$ ονομάζεται **πρώτο τεταρτημόριο** και θα συμβολίζεται με $\tilde{x}_{1/4}$ ή και με Q_1 .

Το 50° ποσοστιαίο σημείο $\left(p = \frac{1}{2}\right)$ ονομάζεται **δεύτερο τεταρτημόριο** και θα συμβολίζεται με $\tilde{x}_{1/2}$ ή Q_2 . Το τελευταίο αποτελεί ένα διαφορετικό ορισμό της διαμέσου.

Το 75° ποσοστιαίο σημείο $\left(p = \frac{3}{4}\right)$ ονομάζεται **τρίτο τεταρτημόριο** και θα συμβολίζεται με $\tilde{x}_{3/4}$ ή και με Q_3 .

Υπάρχει μία άμεση σχέση μεταξύ ποσοσטיών σημείων και σχετικής αθροιστικής κατανομής συχνοτήτων. Την τελευταία τη συναντήσαμε στην παράγραφο 2.5. Οι τιμές μιας (σχετικής) αθροιστικής κατανομής συχνοτήτων $F(x)$ είναι μεταξύ 0 και 1, και μπορούν να εκφραστούν σαν ποσοστά $p\%$ από 0% έως 100%. Έτσι αν η τιμή της $F(x)$ στη θέση x είναι p , δηλαδή, $F(x) = p$, το τελευταίο σημείναι ότι $p\%$ από τα δεδομένα μας είναι μικρότερα ή ίσα από την τιμή x . Παριστάνοντας στον κατακόρυφο άξονα τις τιμές της σχετικής αθροιστικής κατανομής $F(x)$ και στον οριζόντιο τα δεδομένα μας μπορούμε να υπολογίσουμε:

- 1) Το ποσοστιαίο σημείο \tilde{x}_p γνωρίζοντας το p και
- 2) Γνωρίζοντας μια τιμή (\tilde{x}_p) της μεταβλητής μας σε ποιο ποσοστό (p) αντιστοιχεί (βλ. άσκηση 2.6.4.).

Παράδειγμα 2.6.4: Να υπολογισθούν τα ποσοστιαία σημεία, των παρακάτω δεδομένων:

11, 24, 41, 45, 46, 48, 49, 51, 51, 51, 52, 56, 57, 61, 62, 64, 68, 69, 71, 75, 83, 87, 94.

Λύση: Επειδή $n=24$, $n\ 25\% = 24 \cdot \frac{25}{100} = 6$, και επομένως 6 παρατηρήσεις πρέπει να βρίσκονται κάτω του $Q_1 = \tilde{x}_{1/4}$ και 18 άνω απ' αυτό. Συνεπώς, εδώ $\tilde{x}_{1/4} = 48$.

Η διάμεσος, δηλαδή το 50^ο ποσοστιαίο σημείο είναι, μία οποιοδήποτε τιμή μεταξύ του 52 και 56. Μπορούμε να θεωρήσουμε $\tilde{x}_{1/2} = Q_2 = 54$.

Ανάλογα το 75^ο ποσοστιαίο σημείο είναι, $\tilde{x}_{3/4} = Q_3 = 68,5$.

Παρατήρηση 2.6.3:

- α) Συχνά οι διάφοροι μέσοι δεν συμπίπτουν με καμία από τις μετρήσεις μας. Ο αριθμητικός μέσος επηρεάζεται από το μέγεθος των μετρήσεων, ενώ η διάμεσος από το πλήθος.
- β) Αν τα δεδομένα x_i πολλαπλασιαθούν με ένα αριθμό β , και μετά προσθέσουμε ένα αριθμό α , δηλ. έχουμε τα νέα δεδομένα $y_i = \alpha + \beta x_i$, τότε $\bar{y} = \alpha + \beta \bar{x}$, ενώ η νέα διάμεσος είναι $\alpha + \beta m_n$.
- γ) Το άθροισμα, $\sum_{i=1}^n (x_i - \lambda)^2$, ελαχιστοποιείται για $\lambda = \bar{x}$.
- δ) Το άθροισμα, $\sum_{i=1}^n |x_i - \lambda|$, ελαχιστοποιείται για $\lambda = \tilde{x}_{1/2}$.
- ε) Το άθροισμα, $\sum_{i=1}^n S(x_i, \lambda)$, με $S(x_i, \lambda) = \begin{cases} 0 & x_i = \lambda \\ 1 & x_i \neq \lambda \end{cases}$ ελαχιστοποιείται αν λ γίνει ίσο με την επικρατούσα τιμή.

στ) Ο αριθμητικός μέσος k διαφορετικών δειγμάτων υπολογίζεται, όχι όμως και η διάμεσος.

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

Άσκηση 2.6.1: Ενδιαφερόμαστε να γνωρίσουμε τις τιμές βενζίνης σε μία περιοχή. Ρωτώντας 20 βενζινάδικα λάβαμε τις ακόλουθες τιμές σε λεπτά ευρώ

73	73	71	71	71	71	71	76	80	80
83	83	83	87	87	87	87	89	89	89

α) Να βρεθεί η επικρατούσα τιμή, και η διάμεσος. Επίσης να υπολογιστεί ο αριθμητικός μέσος, χρησιμοποιώντας κατάλληλο γραμμικό μετασχηματισμό.

β) Για να έχουμε μια καλύτερη εικόνα γύρω από την αγορά βενζίνης ρωτάμε 12 άλλα βενζινάδικα και λαμβάνουμε μια μέση τιμή 86 λεπτά. Να υπολογισθεί ο αριθμητικός μέσος των τιμών στα 32 βενζινάδικα.

Άσκηση 2.6.2: Το 1972 μία χώρα χρειάστηκε 354,3 Mio t λιγνίτες για ενέργεια. Για τα χρόνια έως το 1976 αυξήθηκαν οι ανάγκες της ως εξής

Χρονικές Μονάδες	Σχετική Μεταβολή
1973-72	1,07
1974-73	0,97
1975-74	0,95
1976-75	1,06

Να βρεθεί η μέση σχετική μεταβολή με τη βοήθεια του γεωμετρικού μέσου, καθώς και του αριθμητικού.

Άσκηση 2.6.3: Ο επόμενος πίνακας τιμών δείχνει το μέγεθος 300 ζευγαριών υποδημάτων που πουλήθηκαν από ένα κατάστημα:

Μέγεθος υποδημάτων:	34	35	36	37	38	39	40	41	42
Αριθμός υποδημάτων:	2	3	68	72	85	40	16	11	3

Να βρεθεί ο αριθμητικός μέσος και η διάμεσος.

Άσκηση 2.6.4: Πολλές χώρες έχουν το φόρο προστιθέμενης αξίας (ΦΠΑ) σε διάφορα προϊόντα. Ο παρακάτω πίνακας δείχνει το φόρο προστιθέμενης αξίας σε ορισμένες χώρες:

Χώρα	ΦΠΑ%
Αγγλία	17,5
Βέλγιο	19,5
Γαλλία	18,6
Γερμανία	15,0
Δανία	15,0
Ελλάδα	18,0
Ελβετία	6,5
Ιαπωνία	3,0
Ιταλία	19,0
Ισπανία	15,0
Καναδάς	7,0
Λουξεμβούργο	15,0
Νέα Ζηλανδία	12,5
Νορβηγία	22,0
Ολλανδία	18,5
Πορτογαλία	16,0
Τουρκία	12,0

- α) Να κατασκευάσετε το ιστόγραμμα και το πολύγωνο συχνοτήτων.
- β) Να βρεθεί η διάμεσος και ο αριθμητικός μέσος.
- γ) Να κατασκευάσετε και να παραστήσετε γραφική τη σχετική αθροιστική κατανομή συχνοτήτων.
- δ) Να βρεθούν τα 20° και το 80° ποσοστιαίο σημείο.
- ε) Οι τιμές 3,0 , 22,0 και 15,0 σε ποίο ποσοστό p αντιστοιχεί. Τί σημαίνει το τελευταίο;

2.7. Μέτρα Απόκλισης

Πολλές φορές οι μέσοι δεν είναι αρκετό από μόνοι τους να χαρακτηρίσουν τα δεδομένα μας. Θα ήταν ενδιαφέρον να γνωρίζουμε κατά πόσο οι παρατηρήσεις μας απέχουν από το μέσο αυτών. Γι’ αυτό το σκοπό θεωρούμε τα δεδομένα $x_1, ..., x_n$, και ορίζουμε τις επόμενες έννοιες:

α. Εύρος

Η διαφορά που προκύπτει μεταξύ της μεγίστης και ελαχίστης παρατήρησης ονομάζεται, **εύρος**. Συμβολικά $\Delta = \max\{x_i\} - \min\{x_i\}$. Όταν έχουμε κάποιες τιμές στα δεδομένα μας που είναι πολύ χαμηλές ή πολύ υψηλές σε σχέση με τις υπό-

λοιπες τιμές, δηλαδή ακραίες, τότε το εύρος δεν είναι αντιπροσωπευτική παράμετρος απόκλισης.

β. Διακύμανση (ή Διασπορά)

Το μέσο άθροισμα των τετραγωνικών αποστάσεων των παρατηρήσεών μας από τον αριθμητικό μέσο ονομάζεται **διακύμανση ή διασπορά**.

Στην περίπτωση μεμονωμένων δεδομένων, x_1, \dots, x_n , ορίζεται από τον αριθμό S^2 , που δίνεται από τη σχέση

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

Ενώ αν h_1 από τα δεδομένα μας έχουν την τιμή α_1 , h_2 την τιμή α_2 , ..., h_k την τιμή α_k , ο S^2 δίνεται από την

$$S^2 = \frac{1}{n} \sum_{j=1}^k h_j (\alpha_j - \bar{x})^2.$$

Στην περίπτωση των ομαδοποιημένων δεδομένων στη θέση των α_j θέτουμε τις κεντρικές τιμές των διαστημάτων $[\alpha_j, \alpha_{j+1})$.

Πολλές φορές ο S^2 ορίζεται από τη σχέση

$$S^2 = \frac{1}{n-1} \sum_{j=1}^n h_j (\alpha_j - \bar{x})^2,$$

δηλ. ο παράγοντας $\frac{1}{n}$ αντικαθίσταται από τον $\frac{1}{n-1}$, και εντελώς ανάλογους τύπους έχουμε και για τις λοιπές περιπτώσεις. Η αντικατάσταση του $\frac{1}{n}$ με το $\frac{1}{n-1}$ δεν μας δίνει το “μέσο των τετραγωνικών αποστάσεων”, και για την Περι-

γραφική Στατιστική δεν έχει σημασία. Εκεί που παρουσιάζει ενδιαφέρον είναι όταν επιχειρούμε να εκτιμήσουμε ή να προσδιορίσουμε τη διακύμανση όλου του πληθυσμού μας. Επ’ αυτού θα επανέλθουμε στη Στατιστική Συμπερασματολογία.

Είναι λογικό να δεχθούμε ότι η μονάδα μέτρησης για τη διακύμανση είναι η αυτή με του γνωρίσματος που εξετάζουμε υψούμενη στο τετράγωνο. Έτσι αν X εξέφραζε ηλικία σε έτη, η διακύμανση S^2 των δεδομένων θα ήταν (έτη)², οπότε για να αποφύγουμε αυτή την ανεπιθύμητη ιδιότητα εισάγουμε την έννοια της τυπικής απόκλισης. Διακύμανση 0 ($S^2=0$) θα σημαίνει ότι όλες μου οι μετρήσεις είναι σταθερές.

γ. Τυπική Απόκλιση

Η θετική τετραγωνική ρίζα της διακύμανσης ονομάζεται **τυπική απόκλιση**, και ορίζεται από τη σχέση

$$S = \sqrt{S^2} = \sqrt{\frac{1}{n} \sum_{j=1}^n (x_j - \bar{x})^2}$$

ή

$$= \sqrt{\frac{1}{n} \sum_{j=1}^k h_j (\alpha_j - \bar{x})^2}$$

δ. Απόλυτη Διακύμανση

Ο μέσος των απολύτων αποστάσεων των δεδομένων μας από μια παράμετρο θέσης λ , ονομάζεται **απόλυτη διακύμανση**. Ορίζεται στην περίπτωση μεμονωμένων δεδομένων x_1, \dots, x_n από τη σχέση

$$\bar{S} = \frac{1}{n} \sum_{j=1}^n |x_j - \lambda|.$$

Ενώ αν h_1 από τα δεδομένα έχουν την τιμή α_1 , h_2 έχουν την τιμή α_2, \dots, h_k έχουν την τιμή α_k , τότε ορίζεται από τη σχέση

$$\bar{S} = \frac{1}{n} \sum_{j=1}^k h_j |\alpha_j - \lambda|.$$

Και στις δύο τελευταίες περιπτώσεις λ θεωρείται ότι είναι μία παράμετρος θέσης. Σύμφωνα με την Παρατήρηση 2.6.3(δ), \bar{S} γίνεται ελάχιστη αν $\lambda = m_n$ (διάμεσος). Άλλωστε είναι και αυτός ο ορισμός της απόλυτης διακύμανσης από πολylούς άλλους συγγραφείς.

ε. Συντελεστής Μεταβλητότητας

Το πηλίκο τυπικής απόκλισης προς τον αριθμητικό μέσο το ονομάζουμε **συντελεστή μεταβλητότητας**, και είναι ανεξάρτητο από τις μονάδες μετρήσεων των δεδομένων μας.

Ο συντελεστής μεταβλητότητας συμβολίζεται με V , ορίζεται από τη σχέση

$$V = \frac{S}{\bar{X}},$$

και εκφράζει τη μεταβλητότητα των δεδομένα μας σαν ποσοστό του μέσου.

στ. Ενδοτεταρτημοριακό Εύρος

Το εύρος εξαρτάται από το μέγεθος της υψηλότερης (μέγιστης), και της χαμηλότερης (ελάχιστης) παρατήρησής μας.

Για να αποφύγουμε το τελευταίο ορίζουμε την έννοια του Ενδοτεταρτημοριακού Εύρους σαν την διαφορά $Q_3 - Q_1 = \tilde{x}_{3/4} - \tilde{x}_{1/4}$, δηλαδή της διαφοράς του 75^{ου} και 25^{ου} ποσοστιαίου σημείου.

ζ. z-τιμή

Πολλές φορές παρουσιάζει ενδιαφέρον να γνωρίσουμε την θέση μιας παρατήρησης σε σχέση με τις υπόλοιπες. Το τελευταίο δίνεται από τη σχέση:

$$z\text{-τιμή} = \frac{x - \bar{x}}{S},$$

όπου \bar{x} , S οι τιμές του αριθμητικού μέσου και της τυπικής απόκλισης αντίστοιχα. Η z-τιμή είναι ένα μέτρο που εκφράζει την απόσταση μιας παρατήρησης x από τον μέσο \bar{x} σε μονάδες τυπικής απόκλισης, και είναι καθαρός αριθμός.

Παρατήρηση 2.7.1:

- α) Αν τα δεδομένα x_1, \dots, x_n , πολλαπλασιασθούν μ' ένα αριθμό β , και προστεθεί μετά ο αριθμός α , τότε για τα νέα δεδομένα, $y_1 = \alpha + \beta x_1, \dots, y_n = \alpha + \beta x_n$, έχουμε $S_y^2 = \beta^2 \cdot S^2$, όπου S^2 η διακύμανση των αρχικών δεδομένων, ενώ S_y^2 η διακύμανση των νέων.
- β) Ο τύπος της διακύμανσης μπορεί να αναπτυχθεί, ώστε να χρησιμοποιείται ευκολότερα σε αριθμητικούς υπολογισμούς ως εξής. Στην περίπτωση μεμονωμένων δεδομένων

$$S^2 = \frac{1}{n} \sum_{j=1}^n (x_j - \bar{x})^2 = \frac{1}{n} \left[\sum_{j=1}^n x_j^2 - 2x \sum_{j=1}^n x_j + \bar{x}^2 \right] = \frac{1}{n} \sum_{j=1}^n x_j^2 - \bar{x}^2$$

Ενώ στην περίπτωση όπου h_1 από τα δεδομένα μας λαμβάνουν την τιμή α_1 , h_2 την τιμή α_2, \dots, h_k την τιμή α_k

$$S^2 = \frac{1}{n} \sum_{j=1}^k h_j \alpha_j^2 - \bar{x}^2 \quad \left(\begin{array}{l} h_1, \dots, h_k \text{ απόλυτες συχνότητες} \\ h_1 + h_2 + \dots + h_k = n \end{array} \right)$$

ή

$$S^2 = \sum_{j=1}^k f_j \alpha_j^2 - \bar{x}^2 \quad \left(\begin{array}{l} f_1, \dots, f_k \text{ σχετικές συχνότητες} \\ f_1 + f_2 + \dots + f_k = 1 \end{array} \right)$$

γ) Πολλές φορές θέλουμε να απαντήσουμε στο ερώτημα: Πόσες μετρήσεις βρίσκονται εντός μιας ή και περισσότερων μονάδων τυπικής απόκλισης από τον μέσο \bar{x} . Αυτό μπορεί να απαντηθεί, αν τα δεδομένα μας σχηματίζουν ένα ιστόγραμμα που το πολύγωνο συχνοτήτων πλησιάζει τη μορφή “σχήματος καμπάνας” (ή όπως θα εξετάσουμε αργότερα την κανονική καμπύλη κατανομής). Έτσι υπολογίζεται ότι:

Στο διάστημα $[\bar{x} - S, \bar{x} + S]$ βρίσκονται το 68% των δεδομένων

Στο διάστημα $[\bar{x} - 2S, \bar{x} + 2S]$ βρίσκονται το 95% των δεδομένων

Στο διάστημα $[\bar{x} - 3S, \bar{x} + 3S]$ βρίσκονται σχεδόν όλα τα δεδομένα

Αν τα δεδομένα μας δεν προέρχονται από μια τέτοια κατανομή, τότε υπάρχει πάλι σχετικός κανόνας, βλέπε κεφάλαιο 6, παράγραφο 6.7.

Από την τελευταία παρατήρηση προκύπτει μια σχέση μεταξύ του εύρους Δ και της τυπικής απόκλισης S .

Έτσι στη δεύτερη περίπτωση έχουμε: $\Delta \approx 4S$.

δ) Αν είχαμε δύο διαφορετικά δείγματα, όπου το πρώτο είναι μεγέθους n_1 και έχει διακύμανση S_1^2 ενώ αντίστοιχα το δεύτερο είναι μεγέθους n_2 με διακύμανση S_2^2 , τότε η **ολική διακύμανση** S^2 ορίζεται από την

$$S^2 = \frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2}$$

Παράδειγμα 2.7.1: α) Δίνονται οι ακόλουθες παρατηρήσεις:

27, 4, 8, 3, 12, 10, 26, 6, 19, 16.

Ποιό το εύρος αυτών;

β) 18 εργαζόμενοι μιας εταιρείας ρωτήθηκαν για το μηνιαίο μισθό τους και είχαμε:

2670, 2549, 2738, 2629, 2582, 2518, 2784, 2763, 2628, 2684, 2654, 2648, 2537, 2723, 2733, 2616, 2572, 2772.

Να υπολογισθεί η διακύμανση αυτών όταν τα δεδομένα θεωρούνται μεμονωμένα, και όταν ομαδοποιηθούν σε διαστήματα μήκους 100.

Λύση: α) $\Delta = 27 - 3 = 24$

β) Για μεμονωμένα δεδομένα είναι

$$\bar{X} = \frac{2670 + \dots + 2772}{18} = 2655,56 (\times 100),$$

και

$$S^2 = \frac{1}{18} \left[(2670 - 2655,56)^2 + \dots + (2772 - 2655,56)^2 \right] = 6654 (\times 100^2).$$

Για ομαδοποιημένα δεδομένα έχουμε τον ακόλουθο πίνακα

Διαστήματα	Κεντρικές Τιμές	Απόλυτες Συχνότητες
[2500, 2600)	2550	5
[2600, 2700)	2650	7
[2700, 2800)	2750	6

και λαμβάνουμε

$$\bar{x} = \frac{5 \cdot 2550 + 7 \cdot 2650 + 6 \cdot 2750}{18} = 2655,56 (\times 100)$$

ενώ

$$\begin{aligned} S^2 &= \frac{\sum_{j=1}^3 h_j (\alpha_j - \bar{x})^2}{18} = \\ &= \frac{5(2550 - 2655,56)^2 + 7(2650 - 2655,56)^2 + 6(2750 - 2655,56)^2}{18} = \\ &= 6056,64 (\times 100^2). \end{aligned}$$

Παράδειγμα 2.7.2: Δίνεται η ακόλουθη κατανομή συχνοτήτων

Διαστήματα	Κεντρικές Τιμές	Απόλυτες Συχνότητες
[100, 150)	125	30
[150, 200)	175	25
[200, 250)	225	40
[250, 300)	275	5

Να βρεθεί η απόλυτη απόκλιση.

Λύση: Η διάμεσος $\tilde{x}_{1/2}$ είναι

$$\tilde{x}_{1/2} = 150 + \frac{200 - 150}{25} [50 - 30] = 190.$$

Άρα,

$$\bar{S} = \frac{\sum_{j=1}^k h_j |a_j - m_n|}{\sum_{j=1}^k h_j} = \frac{|125 - 190|30 + |175 - 190|25 + |225 - 190|40 + |275 - 190|5}{100} = 41,5.$$

Μπορούσαμε αντί της διαμέσου m_n να χρησιμοποιήσουμε τον αριθμητικό μέσο \bar{x} . Σ' αυτήν την περίπτωση σύμφωνα με την Παρατήρηση 2.6.3(δ), η αριθμητική τιμή που θα βρούμε πρέπει να είναι μεγαλύτερη από το 41,5.

Παράδειγμα 2.7.3: Αναφερόμενοι στο Παράδειγμα 2.5.1 θεωρούμε ότι κάποιο διαφορετικό προϊόν διατίθεται απ' έναν άλλο προμηθευτή, όπου οι παραδόσεις των εμπορευμάτων διαρκούν κατά 20 ημέρες περισσότερο. Δηλ. έχουμε τα δεδομένα 24, 25, 24, 21, κ.λπ.. Να υπολογισθούν οι συντελεστές μεταβλητότητας και στις δύο περιπτώσεις.

Λύση: Εδώ

$$V_1 = \frac{S}{5,04}, \quad \text{και} \quad V_2 = \frac{S}{25,04},$$

όπου στους παρονομαστές έχουμε τους αριθμητικούς μέσους για τις δύο περιπτώσεις, ενώ ο αριθμητής παραμένει ίδιος (σύμφωνα με την Παρατήρηση 2.6.4(α)). Είναι φανερό ότι στη δεύτερη περίπτωση εμφανίζονται μικρότερες “μεταβολές” των τιμών.

Παράδειγμα 2.7.4: Αναφερόμενοι στο Παράδειγμα 2.6.2, ποιο το ενδοτεταρτημοριακό εύρος;

Λύση: Εύκολα, $\tilde{x}_{3/4} - \tilde{x}_{1/4} = 68,5 - 18 = 50,5$.

Παράδειγμα 2.7.5: Έστω ότι έχουμε 10 παρατηρήσεις 1, 3, 15, 0, 1, 2, 3, 4, 1, 0. Να βρεθεί η σχετική θέση της παρατήρησης 20.

Λύση: Εδώ, $\bar{x}=3$, 0, $S^2=17$, 6, ενώ $S \approx 4,2$. Άρα, z-τιμή $= \frac{20-3}{4,2} \approx 4,06$, δηλαδή

η τιμή 20 είναι περίπου 4 τυπικές πάνω από το 3. Η τελευταία χαρακτηρίζεται σαν μία «ακραία» τιμή, σύμφωνα με την Παρατήρηση 2.7.1 (γ).

Παράδειγμα 2.7.6: Να υπολογιστούν τα $Q_1 = \tilde{x}_{1/4}$, $Q_2 = \tilde{x}_{1/2}$ και $Q_3 = \tilde{x}_{3/4}$ πο-

σοσταία σημεία, καθώς και το ενδοτεταρτημορικό εύρος, των παρακάτω δεδομένων:

30, 35, 45, 40, 95, 70, 101, 36, 99, 58.

Λύση: Επειδή $n=10$, τότε $n\ 25\% = 10 \frac{25}{100} = 2,5$ παρατηρήσεις πρέπει να είναι κάτω από το $\tilde{x}_{1/4}$ ενώ 7,5 παρατηρήσεις άνω από το $\tilde{x}_{1/4}$. Αν διατάξουμε τις παρατηρήσεις μας, έχουμε:

30, 35, 36, 40, 45, 58, 70, 95, 99, 101.

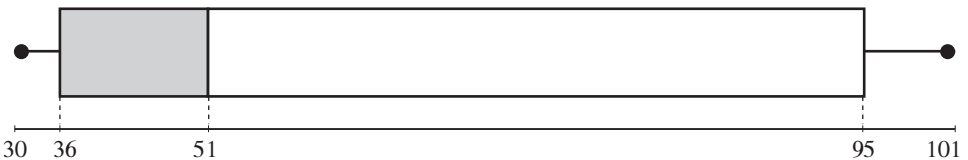
Έτσι η τρίτη παρατήρηση στην παραπάνω διάταξη, 36, ικανοποιεί τα παραπάνω, και επομένως $\tilde{x}_{1/4} = 36$.

Όμοια, ή λόγω συμμετρίας $\tilde{x}_{3/4} = 95$, οπότε

$$\text{Ενδοτεταρτημορικό Εύρος} = \tilde{x}_{3/4} - \tilde{x}_{1/4} = 95 - 36 = 59.$$

Η διάμεσος $\tilde{x}_{1/2}$ είναι κάθε αριθμός μεταξύ του 45 και 58. Λαμβάνουμε $Q_2 = \tilde{x}_{1/2} = 51$.

Ένας γραφικός τρόπος για να διαπιστώσουμε ότι μία παρατήρησή μας είναι «ακραία» είναι το λεγόμενο **Boxplot διάγραμμα** ή **διάγραμμα του κουτιού**. Ένα τέτοιο διάγραμμα περιέχει την ελάχιστη παρατήρηση (χαμηλότερη), τη μέγιστη (υψηλότερη) παρατήρηση καθώς και τα τρία τεταρτημόρια $\tilde{x}_{1/4}$, $\tilde{x}_{1/2}$, και $\tilde{x}_{3/4}$. Το Boxplot κατασκευάζεται ως εξής: Απεικονίζουμε τα 5, παραπάνω σημεία στον οριζόντιο άξονα, και από κάποιο μικρό ύψος του νοερού κατακόρυφου άξονα που τέμνει τον οριζόντιο στην ελάχιστη τιμή, φέρνουμε μία ευθεία παράλληλη προς τον οριζόντιο άξονα, μέχρι τον κατακόρυφο άξονα του $\tilde{x}_{1/4}$. Από το $\tilde{x}_{1/4}$ έως το $\tilde{x}_{3/4}$ σχηματίζουμε ένα ορθογώνιο παραλληλόγραμμο με πλευρές δύο τμήματα από τις κατακόρυφες στα σημεία $\tilde{x}_{1/4}$ και $\tilde{x}_{3/4}$. Φέρνουμε την ενδιάμεση κατακόρυφο στο $\tilde{x}_{1/2}$ και στη συνέχεια προεκτείνουμε την πρώτη παράλληλο που φέραμε έως το σημείο που συναντά νοερά την κατακόρυφο στη μέγιστη τιμή. Έτσι το Boxplot του παραδείγματος 2.7.5 δίνεται ως εξής:



Οι τιμές 35 και 99 μπορούν να θεωρηθούν σαν ακραίες.

Άσκηση 2.7.1: Για 5 διαδοχικές μέρες μετρήθηκαν οι μέσες ταχύτητες 1106 φορτηγών και είχαμε τα ακόλουθα.

Ημέρα	Απόλυτη Συχνότητα	Αριθμητικός Μέσος	Διακύμανση
1	$h_1 = 180$	$\bar{X}_1 = 48,2$	$S_1^2 = 36$
2	$h_2 = 270$	$\bar{X}_2 = 46,5$	$S_2^2 = 22$
3	$h_3 = 215$	$\bar{X}_3 = 47,1$	$S_3^2 = 48$
4	$h_4 = 248$	$\bar{X}_4 = 49,1$	$S_4^2 = 29$
5	$h_5 = 193$	$\bar{X}_5 = 47,6$	$S_5^2 = 41$

Να υπολογισθεί η μέση ταχύτητα, καθώς και η διακύμανση των ταχυτήτων όλων των φορτηγών.

Άσκηση 2.7.2: Οι τιμές εγχρώμων τηλεοράσεων ενός τύπου έχουν μέση τιμή $\bar{x} = 2348$, και διακύμανση $S^2 = 510^2$, ενώ των μαυρόασπρων $\bar{x} = 530$ και $S^2 = 120^2$. Να υπολογισθούν οι συντελεστές μεταβλητότητας.

Άσκηση 2.7.3: Να υπολογισθούν τα τεταρτημόρια και να σχεδιασθεί το Boxplot των εξής δεδομένων:

- α) 3, 7, 7, 8, 5.
- β) 30, 35, 36, 45, 70, 40, 58, 99, 101, 95.

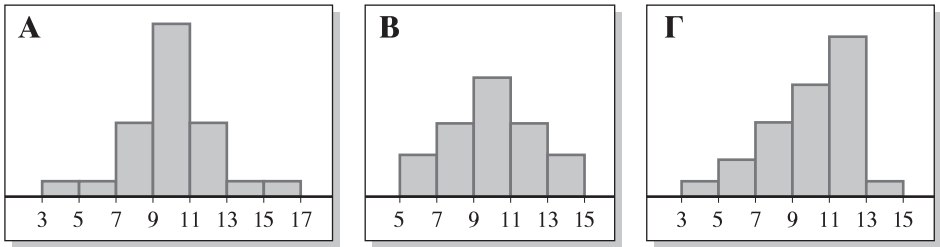
2.8. Άλλες Παράμετροι Κατανομών Συχνοτήτων

Μια κατανομή συχνοτήτων γενικά μπορεί να χαρακτηριστεί αρκετά καλά από τα μέτρα θέσεων και αποκλίσεων. Υπάρχουν όμως περιπτώσεις, όπου οι κατανομές συχνοτήτων ενώ διαφέρουν, να έχουν τα ίδια μέτρα θέσεων και αποκλίσεων. Το επόμενο παράδειγμα διευκρινίζει τα παραπάνω.

Παράδειγμα 2.8.1: Δίνονται οι ακόλουθες κατανομές συχνοτήτων

Διάστημα	Κεντρικές τιμές	Απόλ. Συχ. κατ. Α	Απόλ. Συχ. κατ. Β	Απόλ. Συχ. κατ. Γ
[3, 5)	4	5		5
[5, 7)	6	5	15	10
[7, 9)	8	15	20	15
[9, 11)	10	50	30	25
[11, 13)	12	15	20	40
[13, 15)	14	5	15	5
[15, 17)	16	5		
		100	100	100

Αν υπολογίσουμε τους αριθμητικούς μέσους καθώς και τις διακυμάνσεις των παραπάνω κατανομών έχουμε $\bar{x}=10$ και $S^2=6,4$. Αν δε σκιαγραφήσουμε τα ιστογράμματα αυτών, έχουμε τα επόμενα γραφήματα.



Οι κατανομές Α και Β είναι συμμετρικές, ενώ η Γ ασύμμετρη. Το παράδειγμα αυτό δείχνει την ανάγκη εισαγωγής και άλλων μέτρων.

Ροπές

Όταν έχουμε ένα στατιστικό γνώρισμα X και τις x_1, \dots, x_n παρατηρήσεις του τότε η ν -οστή ροπή αυτού σε σχέση με την παράμετρο θέσης λ , ορίζεται

$$M_\nu^\lambda = \frac{1}{n} \sum_{j=1}^n (x_j - \lambda)^\nu, \quad \nu \text{ φυσικός αριθμός, } \lambda \in \mathbb{R}, \quad \text{όταν } x_1, \dots, x_n \text{ θεωρούνται σαν μεμονωμένα δεδομένα}$$

$$= \frac{1}{n} \sum_{j=1}^k h_j (\alpha_j - \lambda)^\nu, \quad \gg \gg \quad \text{όταν η τιμή } \alpha_j \text{ εμφανίζεται } h_j \text{ φορές,}$$

$$= \sum_{j=1}^k f_j (\alpha_j - \lambda)^\nu, \quad \gg \gg, \quad \text{όπου } f_j \text{ η σχετική συχνότητα του } \alpha_j.$$

Συνήθως λαμβάνουμε $\lambda = \bar{x}$, οπότε σ' αυτήν την περίπτωση

$$\text{για } \nu=1 \quad M_1^{\bar{x}} = \frac{1}{n} \sum_{j=1}^n (x_j - \bar{x}) = 0, \quad \text{ενώ για } \nu=2 \quad M_2^{\bar{x}} = \frac{1}{n} \sum_{j=1}^n (x_j - \bar{x})^2 = S^2.$$

Όταν $\lambda = \bar{x}$ οι ροπές που προκύπτουν θα ονομάζονται **κεντρικές ροπές**. Αν αντί για $(x_j - \lambda)$ θέσουμε $|x_j - \lambda|$, οι ροπές που προκύπτουν ονομάζονται **απόλυτες**.

Συμμετρία και Λοξότητα (Ασυμμετρία)

Μια κατανομή συχνοτήτων θα είναι **συμμετρική αναφορικά με \bar{x}** αν $\bar{x} - c$

και $\bar{x} + c$ έχουν την ίδια απόλυτη (ή σχετική) συχνότητα.

Αν έχουμε μία συμμετρική κατανομή συχνοτήτων τότε

αριθμητικός μέσος = διάμεσος = επικρατούσα τιμή, και αντιστρόφως.

Μια κατανομή συχνοτήτων που δεν είναι συμμετρική θα ονομάζεται **λοξή ή ασύμμετρη**.

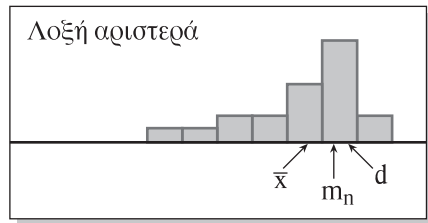
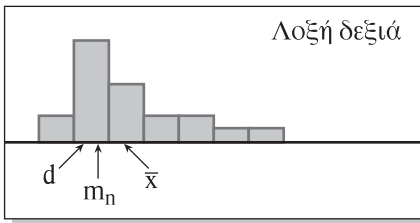
Αν $\bar{x} > m_n > d$ (αριθμητικός μέσος > διάμεσος > επικρατούσα τιμή), τότε θα λέμε ότι είναι

Λοξή προς τα δεξιά (αφήνει δεξιά ουρά)

Αν $\bar{x} < m_n < d$ (αριθμητικός μέσος < διάμεσος < επικρατούσα τιμή), τότε θα λέμε ότι είναι

Λοξή προς τα αριστερά (αφήνει ουρά αριστερά)

Επίσης ένας άλλος τρόπος για να διαπιστώσουμε την συμμετρία είναι να εξετάσουμε κατά πόσο η διάμεσος στο Boxplot το χωρίζει σε δύο ίσα μέρη. Στο Παράδειγμα 2.7.5 η διάμεσος 51 δεν το χωρίζει σε δύο ίσα μέρη και κατά συνέπεια τα δεδομένα μας δεν προέρχονται από συμμετρική κατανομή.



Μέτρα Λοξότητας

Υπάρχουν διάφοροι υπολογιστικοί τρόποι για να δούμε αν μια κατανομή συχνοτήτων είναι λοξή. Αναφέρουμε ορισμένους τέτοιους τρόπους.

Αν \bar{x} (αριθμητικός μέσος), m_n (διάμεσος), d (επικρατούσα τιμή), S (τυπική απόκλιση), $M_3^{\bar{x}}$ (τρίτη κεντρική ροπή), τότε έχουμε τα εξής μέτρα:

1. **Pearson,**

$$g_1 = \frac{\bar{x} - d}{S}$$

2. **Yule-Pearson,**

$$g_2 = \frac{3(\bar{x} - m_n)}{S}$$

3. **Μέτρο λοξότητας τρίτης ροπής,**

$$g_3 = \frac{M_3^{\bar{x}}}{S^3}$$

Έτσι, αν $g_j=0$, $j=1, 2, 3$ τότε η κατανομή είναι συμμετρική.

Αν $g_j<0$, $j=1, 2, 3$ τότε η κατανομή είναι λοξή προς τα αριστερά
(Θετική λοξότητα).

Αν $g_j>0$, $j=1, 2, 3$ τότε η κατανομή είναι λοξή προς τα δεξιά
(Αρνητική λοξότητα).

4. Ένας άλλος πιο εποπτικός τρόπος για την λοξότητα είναι αυτός με το **διάγραμμα του Boxplot**. Έτσι αν η διάμεσος πλησιάζει προς το τρίτο ποσοστιαίο σημείο Q_3 τότε μπορούμε να συμπεράνουμε ότι η κατανομή μας είναι λοξή προς τα αριστερά. Αν η διάμεσος πλησιάζει το Q_1 η κατανομή μας θα είναι λοξή προς τα δεξιά.

Παράδειγμα 2.8.2: Για τις κατανομές B, και Γ έχουμε

$$B: \bar{x} = 10, \quad m = 10, \quad d = 10, \quad S = 2,53, \quad M_3^{\bar{x}} = 0,$$

δηλ. $g_1 = g_2 = g_3 = 0$, δηλ. συμμετρική

$$Γ: \bar{x} = 10, \quad m = 10,6 \quad d = 12, \quad S = 2,53, \quad M_3^{\bar{x}} = -12.$$

$$\text{Εδώ } g_1 = \frac{10-12}{2,53} = 0,79, \quad g_2 = \frac{3(10-10,6)}{2,53} = -0,71, \text{ και } g_3 = \frac{-12}{2,53^3} = -0,74,$$

δηλ. λοξή προς τα αριστερά.

Κυρτότητα

Η διαφορά των κατανομών συχνοτήτων A και B εξετάζεται μέσω της κυρτότητας.

Η ακόλουθη ποσότητα μας δίνει το συντελεστή **κυρτότητας** μιας κατανομής συχνοτήτων.

$$W = \frac{M_4^{\bar{x}}}{S^4} - 3$$

Αν $W>0$ η κατανομή θα είναι λεπτόκυρτη.

Αν $W=0$ η κατανομή θα είναι μεσόκυρτη.

Αν $W<0$ η κατανομή θα είναι πλατόκυρτη.

Παράδειγμα 2.8.3: Για τις κατανομές A και B έχουμε

$$A: M_4^{\bar{x}}=160 \text{ και } W = \frac{160,0}{2,53^4} - 3 = 0,91,$$

δηλ. η κατανομή συχνοτήτων είναι λεπτόκυρτη και τα εμβαδά των ορθογωνίων παραλληλογράμμων στα άκρα είναι λεπτότερα από τα υπόλοιπα κάτι που διαπιστώνεται, επίσης εύκολα από το σχήμα της σελίδας 44.

Ενώ για την

$$B: M_4^{\bar{x}}=83,2 \text{ και } W = \frac{83,2}{40,97} - 3 = -0,97 ,$$

δηλαδή η κατανομή συχνοτήτων είναι πλατόκυρτη και έχουμε «πλατύτερα» εμβαδά των ορθογωνίων παραλληλογράμμων στα άκρα.

Παρατήρηση 2.8.1:

- α) Τα μέτρα που αναφέρθηκαν σ' αυτό το εδάφιο είναι αναλοίωτα κάτω από γραμμικούς μετασχηματισμούς, δηλ. αν τα δεδομένα μας πολλαπλασιασθούν μ' έναν αριθμό β, και μετά προσθέσουμε έναν αριθμό α, τα νέα δεδομένα, θα μας δώσουν τα ίδια μέτρα.
- β) Πολλοί συγγραφείς αντί της ορολογίας **μέτρα θέσης** ή απόκλισης κ.λπ., χρησιμοποιούν τον όρο **παράμετροι θέσης** ή **απόκλισης** αντίστοιχα, κάτι που θα ήταν απόλυτα ακριβές, αν το δείγμα μας ήταν ολόκληρος ο πληθυσμός μας.

Άσκηση 2.8.1: Ο παρακάτω πίνακας δείχνει τους μηνιαίους μισθούς των υπαλλήλων μιας εταιρίας.

Μηνιαίοι μισθοί σε ευρώ	Αριθμός των εργαζομένων
1.000 - 1.500	15
1.500 - 2.000	24
2.000 - 2.500	29
2.500 - 3.000	36
3.000 - 3.500	34
3.500 - 4.000	21
4.000 - 4.500	11

- α) Να υπολογισθεί ο αριθμητικός μέσος, καθώς και η διάμεσος.
- β) Να εξετασθεί από την άποψη λοξότητας και κυρτότητας η εν λόγω κατανομή συχνότητας.

Άσκηση 2.8.2: Ο παρακάτω πίνακας δείχνει τους φόρους για αγαθά και παροχή υπηρεσιών σαν ποσοστά του εθνικού εισοδήματος σε ορισμένες χώρες

Χώρα	Εθνικό εισόδημα	Φόροι
Αγγλία	6,7	18,5
Βέλγιο	7,2	16,0
Γαλλία	7,9	17,8
Δανία	9,9	20,6
Γερμανία	6,4	16,4
Ελλάδα	9,9	25,9
Ελβετία	3,0	9,7
Ιαπωνία	1,4	4,4
Ιταλία	5,7	14,3
Ισπανία	5,5	15,9
Καναδάς	5,3	14,1
Λουξεμβούργο	7,2	14,9
Νέα Ζηλανδία	8,6	23,8
Νορβηγία	8,2	17,4
Ολλανδία	7,3	15,6
Πορτογαλία	6,8	19,0
Τουρκία	6,5	9,7

- α) Να υπολογισθεί η τυπική απόκλιση κάθε μεταβλητής.
- β) Να υπολογισθεί ο συντελεστής μεταβλητότητας κάθε μεταβλητής.
- γ) Να υπολογισθεί το εύρος κάθε μεταβλητής.
- δ) Να συγκριθούν τα δύο παραπάνω μέτρα απόκλισης.
- ε) Να σχεδιασθεί το boxplot διάγραμμα κάθε μεταβλητής.
- στ) Να σχεδιασθούν οι σχετικές αθροιστικές κατανομές κάθε μεταβλητής.
- ζ) Να εξετασθούν από την άποψη λοξότητας και κυρτότητας οι τελευταίες.

Ερωτήσεις

1. Τι είναι πληθυσμός;
2. Τι είναι δείγμα;
3. Τι είναι η στατιστική μονάδα, τι το στατιστικό γνώρισμα και τι η μεταβλητή;
4. Πόσων ειδών μεταβλητές έχουμε;
5. Πώς λαμβάνονται οι μετρήσεις;
6. Πόσων ειδών στατιστικές κλίμακες έχουμε;
7. Ποιούς όρους χρησιμοποιούμε αντί του όρου μετρήσεις;
8. Ταξινομήσατε τα είδη μετρήσεων.
9. Η διάσπαρτη γνώση των μετρήσεων μας απ' έναν πληθυσμό μας βοηθά στο να αντλήσουμε πληροφορίες γύρω απ' αυτόν. Σωστό ή λάθος;
10. Τι είναι η σχετική συχνότητα, τι η απόλυτη; Τι είναι η κατανομή συχνότητας;
11. Ποιους τρόπους παρουσίασης γραφικών δεδομένων γνωρίζετε;
12. Τι είναι η αθροιστική κατανομή συχνοτήτων; Πόσα είδη γνωρίζετε;
13. Τι είναι τα μέτρα θέσης; Αναφέρατε τα μέτρα θέσεων και ερμηνεύσατε αυτά.
14. Τι είναι τα μέτρα απόκλισης; Αναφέρατε τα μέτρα απόκλισης και ερμηνεύσατε αυτά.
15. Τι είναι λοξότητα; Πόσα είδη λοξότητας γνωρίζετε; Πως επιβεβαιώνεται η λοξότητα;
16. Τι είναι κυρτότητα;
17. Τι συμπεράσματα μπορείτε να έχετε μ' ένα σχεδιάγραμμα Box-plot;