Περιγραφική Στατιστική

Οι έννοιες *τυχαία μεταβλητή, τυχαίο δείγμα* και *πληθυσμός* που προσεγγίσαμε και διατυπώσαμε με όρους *Πιθανοτήτων* στο Α΄ Μέρος, αποτελούν βασικές έννοιες και της *Στατιστικής*. Είναι επομένως χρήσιμο να τις δούμε και να τις αποσαφηνίσουμε και με όρους *Στατιστικής*.

Στο *5ο Κεφάλαιο* είδαμε ότι μια ***τυχαία μεταβλητή*** είναι μια πραγματική συνάρτηση που παίρνει τιμές με βάση μια τυχαία διαδικασία και πιο συγκεκριμένα, με βάση το αποτέλεσμα ενός τυχαίου πειράματος. Επίσης, στο *1ο Κεφάλαιο* εξηγήσαμε ότι το αποτέλεσμα ενός τυχαίου πειράματος αναφέρεται/αφορά σε κάποιο κοινό χαρακτηριστικό των υποκειμένων επί των οποίων αυτό εκτελείται. Έτσι, στη Στατιστική, πρακτικά μια τυχαία μεταβλητή εκφράζει ένα κοινό χαρακτηριστικό μιας ομάδας υποκειμένων (ατόμων, αντικειμένων, τόπων, φυτών, κτλ.) το οποίο μεταβάλλεται από υποκείμενο σε υποκείμενο (ή και στο ίδιο υποκείμενο) και παίρνει τιμές με βάση μια τυχαία διαδικασία. Κάθε υποκείμενο επί του οποίου μετράμε/παρατηρούμε την τιμή μιας τυχαίας μεταβλητής ονομάζεται ***απλό στοιχείο*** ή ***πειραματική/δειγματοληπτική μονάδα***[1]. Η κατανομή των τιμών μιας τυχαίας μεταβλητής ονομάζεται ***πληθυσμός*** ή ***στατιστικός πληθυσμός***.

Τυχαίο δείγμα *μεγέθους* ν από έναν πληθυσμό, δηλαδή, από την κατανομή των τιμών μιας τυχαίας μεταβλητής X, ονομάζουμε ν ανεξάρτητες τυχαίες μεταβλητές $X_1, X_2,\ldots,X_\nu$ που παίρνουν τιμές από τον πληθυσμό αυτό, που ακολουθούν δηλαδή την ίδια κατανομή, αυτήν της τ.μ. X. Οι συγκεκριμένες τιμές, $x_1, x_2,\ldots,x_\nu$, της X που έχουμε διαθέσιμες για επεξεργασία μετά τη λήψη του δείγματος αποτελούν μια πραγματοποίηση των $X_1, X_2,\ldots,X_\nu$ και ονομάζονται ***δεδομένα*** ή ***παρατηρήσεις***.

Ας δούμε όμως δύο παράδειγματα που θα μας βοηθήσουν να αποσαφηνίσουμε αυτές τις πολύ βασικές για τη συνέχεια έννοιες.

Παράδειγμα 9.1: *Η πτυχιακή εργασία ενός φοιτητή αφορούσε στα άνθη μιας συγκεκριμένης ποικιλίας ενός φυτού που καλλιεργείται στο νομό Κοζάνης. Στο πλαίσιο αυτής της μελέτης, ο φοιτητής μέτρησε, μεταξύ άλλων, τον αριθμό των πετάλων σε 115 άνθη της συγκεκριμένης ποικιλίας που επέλεξε τυχαία από καλλιέργειες του νομού Κοζάνης. Τα αποτελέσματα αυτών των μετρήσεων φαίνονται στον Πίνακα 9.1.*

7	5	8	7	5	5	6	6	5	7	5	5	5	9	6	8	5
5	5	6	6	5	5	6	5	9	6	5	5	7	6	6	7	5
7	5	5	6	6	5	6	5	6	5	5	5	5	6	6	5	5
8	5	5	5	5	6	5	5	5	6	5	5	6	5	5	5	6
7	5	7	5	5	8	5	5	5	6	5	10	5	6	5	5	6
5	7	5	5	5	9	5	5	7	5	5	5	5	6	7	5	5
6	5	6	5	7	5	10	5	6	5	5	5	8				

Πίνακας 9.1
Οι αριθμοί πετάλων 115 ανθέων συγκεκριμένης ποικιλίας που καλλιεργείται στο νομό Κοζάνης

Προφανώς, η ***τυχαία μεταβλητή*** που μελέτησε ο φοιτητής εκφράζει τον αριθμό των πετάλων του άνθους της συγκεκριμένης ποικιλίας φυτών που καλλιεργείται στο νομό

[1] *Όπως σημειώσαμε και στο εισαγωγικό 1ο Κεφάλαιο, συχνά στη βιβλιογραφία ως δειγματοληπτική μονάδα ορίζεται ένα σύνολο απλών στοιχείων, δηλαδή, στη βιβλιογραφία οι έννοιες απλό στοιχείο και δειγματοληπτική μονάδα γενικά διακρίνονται. Στο πλαίσιο του παρόντος επιλέξαμε με τον όρο δειγματοληπτική μονάδα να εννοούμε όπως και με τον όρο απλό στοιχείο κάθε υποκείμενο επί του οποίου μετράμε/παρατηρούμε την τιμή μιας μεταβλητής.*

Κοζάνης. Ας συμβολίσουμε αυτή την τυχαία μεταβλητή με *X*. Τα 115 άνθη που επέλεξε τυχαία από τις καλλιέργειες του νομού Κοζάνης, αποτελούν τις 115 ***δειγματοληπτικές μονάδες (απλά στοιχεία)*** από τις οποίες αντίστοιχα πήρε τις 115 τιμές $x_1, x_2, \ldots, x_{115}$ της X (του κοινού χαρακτηριστικού τους που μελέτησε) και οι οποίες φαίνονται στον *Πίνακα 9.1*. Αυτές οι 115 τιμές αποτελούν το συγκεκριμένο ***τυχαίο δείγμα*** τιμών της *X* μεγέθους 115 με το οποίο εργάσθηκε. Ο ***πληθυσμός*** που μελέτησε ο φοιτητής, με βάση το *τυχαίο δείγμα* τιμών που πήρε από αυτόν, είναι η κατανομή των τιμών της *X*, δηλαδή, αποτελείται από όλους τους αριθμούς πετάλων που αντιστοιχούν σε όλα τα άνθη όλων των φυτών της συγκεκριμένης ποικιλίας στο νομό Κοζάνης και όχι από όλα τα άνθη όλων των φυτών της συγκεκριμένης ποικιλίας στο νομό Κοζάνης. Ανάλογα, ως *δείγμα* δεν εννοούμε τα άνθη που επέλεξε ο φοιτητής αλλά τις τιμές, $x_1, x_2, \ldots, x_{115}$, της μεταβλητής *X* που παρατήρησε σε αυτά και δίνονται στον *Πίνακα 9.1*. Μπορεί επομένως, στην ίδια ομάδα υποκειμένων *(δειγματοληπτικών μονάδων)* να αναφέρονται διαφορετικοί *πληθυσμοί*. Αν, για παράδειγμα, ο φοιτητής ενδιαφέρεται να μελετήσει και το μήκος, έστω *Y*, του μίσχου του άνθους της συγκεκριμένης ποικιλίας φυτών στο νομό Κοζάνης, τότε πρόκειται για ένα νέο πληθυσμό που αποτελείται από όλα τα *μήκη μίσχων* όλων των ανθέων της συγκεκριμένης ποικιλίας φυτών στο νομό Κοζάνης που είναι ένας διαφορετικός πληθυσμός από αυτόν των *αριθμών των πετάλων* παρότι και οι δύο αναφέρονται στην ίδια ομάδα ανθέων.

■

Παράδειγμα 9.2: *Στο πλαίσιο μιας δημογραφικής έρευνας που αφορούσε στις οικογένειες που κατοικούν μόνιμα στην επαρχία Γορτυνίας του νομού Αρκαδίας, επελέγησαν τυχαία 20 οικογένειες από το σύνολο των οικογενειών που κατοικούν μόνιμα στην επαρχία Γορτυνίας και για κάθε μια από αυτές καταγράφηκαν, μεταξύ άλλων, το επάγγελμα πατέρα, το επίπεδο εκπαίδευσης πατέρα, το μηνιαίο οικογενειακό εισόδημα (σε €) και ο αριθμός παιδιών της οικογένειας. Οι παρατηρήσεις που ελήφθησαν φαίνονται στον Πίνακα 9.2.*

Είναι προφανές ότι στο πλαίσιο της συγκεκριμένης έρευνας, μελετήθηκαν τέσσερις διαφορετικοί *πληθυσμοί* που όμως όλοι αναφέρονται στο ίδιο σύνολο υποκειμένων, στο σύνολο όλων των οικογενειών που κατοικούν μόνιμα στην επαρχία Γορτυνίας. Οι πληθυσμοί αυτοί είναι οι εξής:

- Ο *πληθυσμός* των επαγγελμάτων των πατεράδων των οικογενειών που κατοικούν μόνιμα στη Γορτυνία, δηλαδή, η κατανομή των *επαγγελμάτων* των πατεράδων των οικογενειών που κατοικούν μόνιμα στη Γορτυνία.
- Ο *πληθυσμός* των επιπέδων εκπαίδευσης των πατεράδων των οικογενειών που κατοικούν μόνιμα στη Γορτυνία, δηλαδή, η κατανομή των *επιπέδων εκπαίδευσης* όλων των πατεράδων των οικογενειών που κατοικούν μόνιμα στη Γορτυνία.
- Ο *πληθυσμός* των μηνιαίων οικογενειακών εισοδημάτων των οικογενειών που κατοικούν μόνιμα στη Γορτυνία, δηλαδή, η κατανομή των *μηνιαίων οικογενειακών εισοδημάτων* όλων των οικογενειών που κατοικούν μόνιμα στη Γορτυνία.
- Ο *πληθυσμός* του αριθμού παιδιών ανά οικογένεια όλων των οικογενειών που κατοικούν μόνιμα στη Γορτυνία, δηλαδή, η κατανομή των αριθμών που δηλώνουν το *πλήθος των παιδιών ανά οικογένεια* που κατοικεί μόνιμα στη Γορτυνία.

Οικογένεια	***Επάγγελμα πατέρα*** x_i	***Επίπεδο εκπαίδευσης πατέρα***[2] y_i	***Μηνιαίο οικογενειακό εισόδημα (σε €)*** w_i	***Αριθμός παιδιών οικογένειας*** u_i
1	*Αγρότης*	*1*	*1400*	*0*
2	*Κτηνοτρόφος*	*2*	*1450*	*1*
3	*Εργάτης*	*2*	*1600*	*0*
4	*Δημ. Υπάλληλος*	*4*	*1400*	*2*
5	*Κτηνοτρόφος*	*2*	*1600*	*2*
6	*Αγρότης*	*2*	*1000*	*2*
7	*Κτηνοτρόφος*	*2*	*1800*	*3*
8	*Ιδιωτ. Υπάλληλος*	*4*	*2000*	*2*
9	*Αγρότης*	*2*	*1200*	*4*
10	*Εργάτης*	*2*	*1200*	*1*
11	*Άλλο*	*3*	*1400*	*1*
12	*Αγρότης*	*2*	*1200*	*2*
13	*Δάσκαλος*	*3*	*1600*	*3*
14	*Δημ. Υπάλληλος*	*2*	*1400*	*4*
15	*Ιδιωτ. Υπάλληλος*	*3*	*1800*	*1*
16	*Δάσκαλος*	*3*	*2000*	*2*
17	*Εργάτης*	*1*	*1800*	*2*
18	*Κτηνοτρόφος*	*1*	*1250*	*2*
19	*Άλλο*	*2*	*1450*	*2*
20	*Κτηνοτρόφος*	*2*	*1600*	*2*

Πίνακας 9.2
Δημογραφικά δεδομένα 20 οικογενειών
που κατοικούν μόνιμα στην επαρχία Γορτυνίας

Οι παρατηρήσεις που φαίνονται στον *Πίνακα 9.2*, ελήφθησαν από τις ίδιες, τυχαία επιλεγμένες, δειγματοληπτικές μονάδες, όμως αποτελούν τέσσερα διαφορετικά *τυχαία δείγματα* τιμών, τεσσάρων διαφορετικών τυχαίων μεταβλητών αντίστοιχα. Ας συμβολίσουμε αυτές τις τυχαίες μεταβλητές, που η κάθε μία εκφράζει ένα από τα τέσσερα χαρακτηριστικά που μελετήθηκαν στην έρευνα, με X, Y, W, και U, αντίστοιχα. Έτσι, τη συγκεκριμένη πραγματοποίηση του δείγματος από τη μεταβλητή *επάγγελμα πατέρα* τη συμβολίζουμε με $x_1, x_2, \ldots, x_{20}$, από τη μεταβλητή *επίπεδο εκπαίδευσης πατέρα* με $y_1, y_2, \ldots, y_{20}$, από τη μεταβλητή *μηνιαίο οικογενειακό εισόδημα* με $w_1, w_2, \ldots, w_{20}$ και από τη μεταβλητή *αριθμός παιδιών οικογένειας* με $u_1, u_2, \ldots, u_{20}$.

■

Όπως αναφέραμε και στην εισαγωγή του Β΄ Μέρους *(8ο Κεφάλαιο)*, στη *Στατιστική* «όλα αρχίζουν από τα δεδομένα». Αυτό που κατ' αρχάς απαιτείται είναι κατάλληλη επεξεργασία τους ώστε να μπορέσουμε να τα περιγράψουμε με συνοπτικό και εύληπτο τρόπο για να **κατανοήσουμε την κατανομή τους**. Μάλιστα, αν αυτά έχουν προκύψει από *τυχαία δειγματοληψία*, η περιγραφή της κατανομής τους μας βοηθάει να αποκτήσουμε ***εμπειρική*** γνώση για την άγνωστη κατανομή από την οποία προέρχονται και την οποία ενδιαφερόμαστε να μελετήσουμε[3].

[2] *1=Πρωτοβάθμια εκπαίδευση, 2=Δευτεροβάθμια εκπαίδευση, 3=Τριτοβάθμια εκπαίδευση και 4=Μεταπτυχιακές σπουδές.*

[3] *Αξίζει να επισημάνουμε ότι για την εφαρμογή των μεθόδων περιγραφικής στατιστικής δεν είναι απαραίτητο τα δεδομένα να έχουν προκύψει από πραγματοποίηση τυχαίου δείγματος. Αυτό απαιτείται στη στατιστική συμπερασματολογία. Όμως στη συνέχεια, τόσο στην περιγραφική στατιστική όσο και στη στατιστική συμπερασματολογία, θα αναφερόμαστε σε δεδομένα από τυχαία δείγματα.*

Η ανάγκη επεξεργασίας των δεδομένων για την περιγραφή της κατανομής τους, προκύπτει αβίαστα αν στο *Παράδειγμα 9.1* παρατηρήσουμε τα δεδομένα που χρησιμοποίησε ο φοιτητής στην πτυχιακή του μελέτη. Τα παρουσιάσαμε όπως τα κατέγραψε ο φοιτητής, δηλαδή, χωρίς να έχει προηγηθεί κάποιου είδους επεξεργασία ***(raw data)***. Είναι προφανές ότι με αυτόν τον τρόπο παρουσίασης των δεδομένων δύσκολα μπορούμε να απαντήσουμε ακόμη και σε πολύ απλές ερωτήσεις σχετικές με την κατανομή τους όπως, *ποιος αριθμός πετάλων (δηλαδή, ποια τιμή της μεταβλητής Χ) εμφανίσθηκε πιο συχνά στο δείγμα, ποιο ποσοστό των παρατηρήσεων είναι π.χ., μικρότερες του 7*.

Η *Περιγραφική Στατιστική* αυτή την ανάγκη καλύπτει. Μας προσφέρει μεθόδους επεξεργασίας των δεδομένων για να μπορέσουμε, κατ' αρχάς, και πριν προχωρήσουμε σε επαγωγικά συμπεράσματα για τον πληθυσμό από τον οποίο προέρχονται, **να περιγράψουμε και να κατανοήσουμε την κατανομή τους**. Οι δυνατότητες επεξεργασίας δεδομένων που μας προσφέρει μπορούν να ταξινομηθούν σε τρεις κατηγορίες:

- ***Πινακοποίηση***
- ***Γραφικές αναπαραστάσεις***
- ***Αριθμητικά περιγραφικά μέτρα.***

Όπως θα διαπιστώσουμε, οι δυνατότητες αυτές (σε πολλές περιπτώσεις) διαφοροποιούνται ανάλογα με τον τύπο/είδος της μεταβλητής. Γι΄αυτό, θα τις παρουσιάσουμε ανά τύπο μεταβλητής. Πρώτα για τις ***ποσοτικές*** μεταβλητές, στη συνέχεια για τις ***ποιοτικές*** και τέλος για τις ***διεύθυνσης*** και ***κατεύθυνσης (κυκλικές)*** που αποτελούν ειδική περίπτωση των ποσοτικών.

9.1 Ποσοτικές Μεταβλητές

Ποσοτικές (quantitative) είναι οι μεταβλητές που παίρνουν μόνο αριθμητικές τιμές και διακρίνονται σε ***συνεχείς (continuous)*** και ***διακριτές (discrete)***. *Συνεχείς* είναι οι *ποσοτικές* μεταβλητές που μπορούν να πάρουν ως τιμές τους όλους τους αριθμούς σε ένα διάστημα πιθανών τιμών ενώ *διακριτές* είναι οι *ποσοτικές* μεταβλητές που μπορούν να πάρουν ως τιμές τους μεμονωμένους/διακριτούς αριθμούς όπως 0,1,2,3, … ή αλλιώς, το σύνολο των πιθανών τιμών τους είναι πεπερασμένο ή απείρως αριθμήσιμο. Η τυχαία μεταβλητή *Χ (αριθμός πετάλων άνθους)* του *Παραδείγματος 9.1* και η τυχαία μεταβλητή *U (αριθμός παιδιών οικογένειας)* του *Παραδείγματος 9.2* προφανώς είναι *ποσοτικές διακριτές*. Η τυχαία μεταβλητή *W (μηνιαίο οικογενειακό εισόδημα)* του *Παραδείγματος 9.2* θεωρείται *ποσοτική διακριτή* γιατί παρότι θεωρητικά μπορεί να πάρει ως τιμή της οποιονδήποτε αριθμό στο διάστημα $[0, +\infty)$, εντούτοις πρακτικά παίρνει τιμές το πολύ με ακρίβεια λεπτού (cent).

Για τις *ποσοτικές μεταβλητές,* η *Περιγραφική Στατιστική* προσφέρει τις ακόλουθες δυνατότητες.

9.1.1. Κατασκευή πίνακα (κατανομής) συχνοτήτων

Το πρώτο που κάνουμε μετά τη συγκέντρωση των δεδομένων του δείγματος είναι να δούμε **ποιες τιμές** της μεταβλητής που μελετάμε και **πόσο συχνά** η κάθε μια εμφανίσθηκαν στο δείγμα. Ο *πίνακας συχνοτήτων* κατασκευάζεται για να αναδείξει με εύληπτο τρόπο αυτή την πληροφορία. Ας δούμε πώς.

Έστω $x_1, x_2, \ldots, x_\nu$, ένα τυχαίο δείγμα ν τιμών μιας τυχαίας μεταβλητής X και $y_1, y_2, \ldots, y_k$ $(k \le \nu)$ οι k διαφορετικές, μεταξύ τους, τιμές από τις $x_1, x_2, \ldots, x_\nu$.

Ο ***πίνακας (κατανομής) συχνοτήτων (frequency table)*** ενός τυχαίου δείγματος τιμών, $x_1, x_2, \ldots, x_\nu$, μιας *ποσοτικής* μεταβλητής X, αποτελείται από τρεις στήλες. Στην πρώτη στήλη καταγράφονται σε αύξουσα σειρά οι k διαφορετικές τιμές της X που εμφανίσθηκαν στο δείγμα, δηλαδή οι $y_1, y_2, \ldots, y_k \ (k \le \nu)$ και στις δύο επόμενες στήλες καταγράφονται αντίστοιχα

1. η ***συχνότητα (frequency)*** εμφάνισης, ν_i, κάθε τιμής y_i, $i = 1, 2, \ldots, k$, δηλαδή, πόσες φορές εμφανίσθηκε η αντίστοιχη τιμή, y_i, στο δείγμα και
2. η ***σχετική συχνότητα (relative frequency)*** εμφάνισης, f_i, κάθε τιμής y_i, $i = 1, 2, \ldots, k$ που ορίζεται από το λόγο

$$f_i = \frac{\nu_i}{\nu} \text{ ή } f_i = \frac{\nu_i}{\nu} \cdot 100\%.$$

Τα ζεύγη (y_i, ν_i), $i = 1, 2, \ldots, k$ αποτελούν την ***κατανομή συχνοτήτων*** και τα ζεύγη (y_i, f_i), $i = 1, 2, \ldots, k$ την ***κατανομή σχετικών συχνοτήτων*** των τιμών της X που εμφανίσθηκαν στο δείγμα.

Είναι προφανές ότι

$$\nu_1 + \nu_2 + \ldots + \nu_k = \nu \text{ και } f_1 + f_2 + \ldots + f_k = 1 \ (\text{ή } = 100\%).$$

Ο *πίνακας συχνοτήτων* ενός δείγματος τιμών μιας *ποσοτικής* μεταβλητής, μπορεί να συμπληρωθεί με δύο ακόμη στήλες στις οποίες να καταγράφονται αντίστοιχα

1. η ***αθροιστική συχνότητα (cumulative frequency)***, N_i, κάθε τιμής y_i, $i = 1, 2, \ldots, k$ που ορίζεται ως το άθροισμα των *συχνοτήτων* όλων των τιμών που είναι μικρότερες ή ίσες της y_i και
2. η ***αθροιστική σχετική συχνότητα (cumulative relative frequency)***, F_i, κάθε τιμής y_i, $i = 1, 2, \ldots, k$ που ορίζεται ως το άθροισμα των *σχετικών συχνοτήτων* όλων των τιμών που είναι μικρότερες ή ίσες της y_i.

Στα επόμενα, λέγοντας *πίνακας συχνοτήτων* θα θεωρούμε/εννοούμε ότι περιλαμβάνει και τις *αθροιστικές συχνότητες* και τις *αθροιστικές σχετικές συχνότητες*, δηλαδή, ότι συνολικά αποτελείται από πέντε στήλες. Ας δούμε δύο παραδείγματα.

Παράδειγμα 9.1.1 (συνέχεια του Παραδείγματος 9.1): Ο *Πίνακας 9.1.1* που ακολουθεί είναι ο *πίνακας συχνοτήτων* του *τυχαίου δείγματος* τιμών της *τυχαίας μεταβλητής X* του *Παραδείγματος 9.1 (αριθμός πετάλων του άνθους συγκεκριμένης ποικιλίας φυτών που καλλιεργείται στο νομό Κοζάνης)*.

y_i	ν_i	f_i	N_i	F_i
5	67	0.5826	67	0.5826
6	26	0.2261	93	0.8087
7	12	0.1043	105	0.9130
8	5	0.0435	110	0.9565
9	3	0.0261	113	0.9826
10	2	0.0174	115	1.0000
Σύνολα	**115**	**1.0000**		

Πίνακας 9.1.1
Ο πίνακας συχνοτήτων του δείγματος από την τυχαία μεταβλητή
«αριθμός πετάλων του άνθους συγκεκριμένης ποικιλίας
που καλλιεργείται στο νομό Κοζάνης»
του Παραδείγματος 9.1

Παρατηρώντας τον *πίνακα συχνοτήτων*, άμεσα διαπιστώνουμε ότι οι τιμές της *X* που εμφανίσθηκαν στο *τυχαίο δείγμα* που πήρε ο φοιτητής, είναι οι 5, 6, 7, 8, 9, και 10. Επίσης, πολύ εύκολα μπορούμε να δούμε, πόσο συχνά εμφανίσθηκε κάθε μια από αυτές τις τιμές, ποιος αριθμός πετάλων εμφανίσθηκε πιο συχνά (είναι η τιμή 5 και μάλιστα βλέπουμε ότι αποτελεί το 58.26% του δείγματος, δηλαδή, το 58.26% των τιμών του δείγματος είναι ίσες με 5), ποια από τις τιμές 6 και 7 εμφανίσθηκε πιο συχνά (είναι η τιμή 6), επίσης βλέπουμε ότι οι συχνότητες φθίνουν καθώς ο αριθμός των πετάλων αυξάνει, ότι οι τιμές 5 και 6 αποτελούν το 80.87% του δείγματος, ότι ποσοστό 91.30% των τιμών του δείγματος δεν ξεπερνούν την τιμή 7 (δηλαδή, ότι το 91.30% των ανθέων που εξετάσθηκαν είχαν το πολύ μέχρι και 7 πέταλα και μάλιστα 5, 6 ή 7), κτλ. Ως **γενικό συμπέρασμα** για την κατανομή του τυχαίου δείγματος, δηλαδή για την κατανομή των αριθμών των πετάλων των 115 τυχαία επιλεγμένων ανθέων, μπορούμε να πούμε ότι ένα πολύ μεγάλο ποσοστό των τιμών του δείγματος συγκεντρώνεται στο αριστερό άκρο της κατανομής και ότι οι συχνότητες φθίνουν αυξανομένου του αριθμού των πετάλων. Παρατηρώντας τον *Πίνακα 9.1,* όπου οι τιμές του δείγματος παρουσιάζονται όπως τις κατέγραψε ο φοιτητής, χωρίς να έχει προηγηθεί κάποια επεξεργασία *(raw data),* είναι προφανές ότι τέτοιου είδους πληροφορίες για την κατανομή του *τυχαίου δείγματος* δε μπορούν να προκύψουν με απλή παρατήρηση.

■

Παράδειγμα 9.1.2 (συνέχεια του Παραδείγματος 9.2): Ο *Πίνακας 9.1.2* που ακολουθεί είναι ο *πίνακας συχνοτήτων* του *τυχαίου δείγματος* από τη μεταβλητή *U* (*αριθμός παιδιών οικογένειας)* του *Παραδείγματος 9.2*.

y_i	ν_i	f_i	N_i	F_i
0	2	0.1	2	0.1
1	4	0.2	6	0.3
2	10	0.5	16	0.8
3	2	0.1	18	0.9
4	2	0.1	20	1.0
Σύνολα	**20**	**1.0**		

Πίνακας 9.1.2
Ο πίνακας συχνοτήτων του δείγματος από την τυχαία μεταβλητή «αριθμός παιδιών οικογένειας» του Παραδείγματος 9.2

Όπως και στο προηγούμενο παράδειγμα, από τον *πίνακα συχνοτήτων* μπορούμε πλέον να πάρουμε άμεσα πληροφορίες για την κατανομή του δείγματος (ποιες τιμές εμφανίσθηκαν, πόσο συχνά, κτλ.). Ως **γενικό συμπέρασμα** για την *κατανομή συχνοτήτων* αυτού του τυχαίου δείγματος, μπορούμε να πούμε ότι η τιμή 2 παρουσιάζει τη μεγαλύτερη συχνότητα και ότι αριστερά αυτής της τιμής οι συχνότητες αυξάνουν αυξανομένου του αριθμού των παιδιών, ενώ δεξιά αυτής της τιμής, οι συχνότητες φθίνουν αυξανομένου του αριθμού των παιδιών.

■

Παρατήρηση 9.1.1 (εμπειρική εκτίμηση της συνάρτησης πιθανότητας): *Αν θυμηθούμε τον **στατιστικό ορισμό** της πιθανότητας, είναι προφανές ότι η **κατανομή σχετικών συχνοτήτων** ενός τυχαίου δείγματος τιμών, $x_1, x_2, \ldots, x_\nu$, από μια διακριτή τυχαία μεταβλητή X, μας δίνει μια **εμπειρική εκτίμηση/προσέγγιση της συνάρτησης πιθανότητας,** $f(x) = P(X = x)$, της X.*

Για παράδειγμα, η κατανομή σχετικών συχνοτήτων του τυχαίου δείγματος από τη μεταβλητή U (αριθμός παιδιών οικογένειας) του Παραδείγματος 9.2 (Πίνακας 9.1.2), μας δίνει τις τιμές 0.1, 0.2, 0.5, 0.1 και 0.1, ως προσεγγιστικές τιμές, αντίστοιχα, των πιθανοτήτων $f(0) = P(U = 0)$, $f(1) = P(U = 1)$, $f(2) = P(U = 2)$, $f(3) = P(U = 3)$,

$f(4) = P(U = 4)$ *(και για όλες τις άλλες πιθανές τιμές της U μας δίνει προσεγγιστικές πιθανότητες μηδέν). Βέβαια, αν πάρουμε ένα άλλο τυχαίο δείγμα τιμών της U, ακόμη και αν είναι ιδίου μεγέθους, δεν περιμένουμε να έχει ακριβώς ίδια κατανομή σχετικών συχνοτήτων. Παρόλα αυτά, επειδή τα δείγματα είναι τυχαία, περιμένουμε οι κατανομές σχετικών συχνοτήτων τους να είναι παρόμοιες με την κατανομή της U. Μάλιστα, όσο μεγαλύτερου μεγέθους τυχαίο δείγμα παίρνουμε τόσο η κατανομή σχετικών συχνοτήτων του προσεγγίζει καλύτερα την συνάρτηση πιθανότητας* $f(u) = P(U = u)$ *της U. Αν μάλιστα κατασκευάσουμε τον πίνακα συχνοτήτων των αριθμών παιδιών όλων των οικογενειών που κατοικούν μόνιμα στη Γορτυνία τότε θα έχουμε προσδιορίσει* ***επακριβώς*** *τη συνάρτηση πιθανότητας της U.*

■

Ομαδοποίηση των δεδομένων

Στα δύο προηγούμενα παραδείγματα κατασκευάσαμε τον *πίνακα συχνοτήτων* δεδομένων που προέρχονται από *διακριτές ποσοτικές* μεταβλητές, μάλιστα, και στις δύο περιπτώσεις οι διαφορετικές τιμές, y_i, των αντίστοιχων μεταβλητών είναι λίγες (στο πρώτο παράδειγμα 6 διαφορετικές τιμές και στο δεύτερο 5 διαφορετικές τιμές). Ας δούμε ένα παράδειγμα κατασκευής *πίνακα συχνοτήτων* δεδομένων που προέρχονται από μια *συνεχή ποσοτική* μεταβλητή.

Παράδειγμα 9.1.3: *Στον Πίνακα 9.1.3 φαίνεται για κάθε μια από 50 τυχαία επιλεγμένες γαλακτοπαραγωγές αγελάδες, ο χρόνος X (σε μήνες), από την πρώτη εκδήλωση μιας συγκεκριμένης ασθένειας από την οποία είχαν προσβληθεί, μέχρι την επανεμφάνισή της. (Πρόκειται για μια δύσκολα αντιμετωπίσιμη ασθένεια η οποία ενώ θεραπεύεται, μετά από κάποιο χρονικό διάστημα επανεμφανίζεται).*

2.1	1.7	0.8	0.8	4.1	8.7	1.4	2.9	1.9	2.7
4.4	2.2	5.5	7.0	1.8	0.2	1.0	0.9	4.0	0.7
2.0	6.5	0.7	4.3	0.2	5.6	2.4	1.4	1.3	1.2
0.5	3.9	7.4	3.3	8.8	0.3	2.0	5.7	0.8	2.6
9.9	1.6	2.8	1.0	0.6	1.3	0.8	5.9	0.9	0.4

Πίνακας 9.1.3
Οι χρόνοι επανεμφάνισης (σε μήνες) μιας ασθένειας σε 50 γαλακτοπαραγωγές αγελάδες

Παρατηρούμε ότι οι διαφορετικές τιμές της *X,* δηλαδή τα y_i, που εμφανίσθηκαν σε αυτό το τυχαίο δείγμα είναι πολλές. Η μικρότερη είναι η τιμή 0.2 και η μεγαλύτερη η 9.9. Είναι προφανές, ότι αν οργανώσουμε αυτά τα δεδομένα σε *πίνακα συχνοτήτων* όπως στα δύο προηγούμενα παραδείγματα, δηλαδή, γράφοντας στην πρώτη στήλη, σε αύξουσα σειρά, όλες τις διαφορετικές τιμές y_i, ο πίνακας θα έχει ελάχιστη πρακτική αξία αφού οι διαφορετικές τιμές είναι πολλές και με μικρή συχνότητα η κάθε μια (οι περισσότερες έχουν συχνότητα 1 η κάθε μία, δηλαδή εμφανίζονται 1 φορά η κάθε μία). Σημειώνουμε ότι αυτό είναι λογικό να συμβαίνει σε *συνεχείς* μεταβλητές, όμως πολλές διαφορετικές τιμές μπορεί επίσης να εμφανισθούν και σε δείγματα από *διακριτές* μεταβλητές.

Για αυτές τις περιπτώσεις, όπου στα δεδομένα εμφανίζονται πολλές διαφορετικές τιμές, είτε αυτές προέρχονται από *συνεχείς* τυχαίες μεταβλητές είτε από *διακριτές*, η *Περιγραφική Στατιστική,* προτείνει η κατασκευή του *πίνακα συχνοτήτων* να γίνεται αφού πρώτα γίνει ***ομαδοποίηση*** των δεδομένων. Δηλαδή, να ταξινομούνται τα δεδομένα σε *k* διαφορετικές ***ομάδες/κλάσεις (groups/class intervals)*** και στην πρώτη στήλη του *πίνακα συχνοτήτων* να αναγράφονται όχι οι διαφορετικές τιμές που εμφανίσθηκαν στο δείγμα αλλά οι *k* διαφορετικές κλάσεις τιμών. Έτσι στη δεύτερη στήλη θα αναγράφεται πλέον η *συχνότητα* κάθε κλάσης και στις επόμενες στήλες

αντίστοιχα η *σχετική*, η *αθροιστική* και η *αθροιστική σχετική συχνότητα* κάθε κλάσης. Δηλαδή, ο *πίνακας συχνοτήτων*, σε αυτές τις περιπτώσεις, παρουσιάζει τις συχνότητες κλάσεων τιμών και όχι τιμών. Προφανώς εννοείται, αλλά το επισημαίνουμε, ότι ο καθορισμός των κλάσεων γίνεται έτσι, ώστε κάθε τιμή να ανήκει σε μια μόνο κλάση.

Γεννώνται, βέβαια, τρία βασικά ερωτήματα: σε πόσες κλάσεις ταξινομούμε τα δεδομένα, τι πλάτος πρέπει να έχει κάθε κλάση και αν πρέπει όλες να είναι ίσου πλάτους ή μπορεί να έχουν και άνισα πλάτη.

Στα ερωτήματα αυτά η *Περιγραφική Στατιστική* δε δίνει μονοσήμαντες/αυστηρές απαντήσεις. Κατ' αρχάς, πρέπει να έχουμε υπόψη μας ότι ομαδοποιώντας τα δεδομένα, χάνουμε κάποια από την πληροφορία που περιέχεται στα αρχικά δεδομένα, και επομένως, όσο πιο λίγες και μεγάλου πλάτους κλάσεις κατασκευάσουμε τόσο περισσότερη πληροφορία χάνουμε. Βέβαια, ο αριθμός και το πλάτος των κλάσεων και το αν θα επιλέξουμε αυτές να είναι ίσου ή άνισου πλάτους, εξαρτώνται από την κλίμακα στην οποία θέλουμε να αναδείξουμε διαφορές και από το αν θέλουμε και σε ποια διαστήματα να εστιάσουμε για μεγαλύτερη λεπτομέρεια της κατανομής. Επίσης, ο αριθμός και το πλάτος των κλάσεων εξαρτώνται και από το μέγεθος του δείγματος, ν. Μάλιστα, στη βιβλιογραφία προτείνεται ο τύπος

$$k = 1 + 3.32 \cdot \log_{10}(\nu)$$

γνωστός ως ***τύπος του Sturges***, ο οποίος δίνει τον αριθμό των κλάσεων k ως συνάρτηση του μεγέθους του δείγματος ν και μπορεί να χρησιμοποιηθεί ως ένας οδηγός για την επιλογή κατάλληλου αριθμού κλάσεων. Αναφέρουμε τέλος, ότι συνήθως, οι κλάσεις επιλέγουμε να είναι του ίδιου πλάτους. Όμως για όλες αυτές τις αποφάσεις, ιδιαίτερη αξία και καθοριστική σημασία έχει η εμπειρία του ερευνητή.

Αν αποφασίσουμε οι κλάσεις να έχουν ίδιο πλάτος, έστω r, αυτό υπολογίζεται διαιρώντας το εύρος των δεδομένων, $R = x_{\max} - x_{\min}$, με τον αριθμό των κλάσεων k. Δηλαδή,

$$r = \frac{R}{k} = \frac{x_{\max} - x_{\min}}{k}.$$

Όλες οι κλάσεις ορίζονται ως ημιανοικτά διαστήματα της ίδιας μορφής [α, β) ή (α, β]. Ο καθορισμός τους, δηλαδή η επιλογή του αριστερού άκρου της πρώτης κλάσης, γίνεται έτσι ώστε η πρώτη κλάση να περιέχει τη μικρότερη τιμή που εμφανίσθηκε στο δείγμα, $x_{\min}$, και η τελευταία τη μεγαλύτερη, $x_{\max}$. Φροντίζουμε επίσης, να μη συμπίπτουν τιμές του δείγματος με άκρα κλάσεων (χωρίς όμως να είναι απαραίτητο).

Παράδειγμα 9.1.4 (συνέχεια του Παραδείγματος 9.1.3): *Ο Πίνακας 9.1.4 που ακολουθεί είναι ο πίνακας συχνοτήτων των δεδομένων του Πίνακα 9.1.3 ομαδοποιημένων σε επτά κλάσεις, πλάτους 1.5 μήνες η κάθε μια.*

Χρόνος Επανεμφάνισης	ν_i	f_i	N_i	F_i
[0.0 1.5)	21	0.42	21	0.42
[1.5 3.0)	13	0.26	34	0.68
[3.0 4.5)	6	0.12	40	0.80
[4.5 6.0)	4	0.08	44	0.88
[6.0 7.5)	3	0.06	47	0.94
[7.5 9.0)	2	0.04	49	0.98
[9.0 10.5)	1	0.02	50	1.00
Σύνολα	**50**	**1.00**		

Πίνακας 9.1.4
Ο πίνακας συχνοτήτων των τιμών του δείγματος από τη μεταβλητή «χρόνος επανεμφάνισης μιας ασθένειας σε γαλακτοπαραγωγές αγελάδες» του Παραδείγματος 9.1.3, ομαδοποιημένων σε επτά κλάσεις

Τον αριθμό των κλάσεων τον επιλέξαμε με βάση τον τύπο του *Sturges*

$$k = 1 + 3.32 \cdot \log_{10}(50) = 6.64 \cong 7$$

και το πλάτος, *r*, υπολογίσθηκε από τον τύπο

$$r = \frac{R}{k} = \frac{x_{\max} - x_{\min}}{k} = \frac{9.9 - 0.2}{7} = 1.39 \cong 1.5 \text{ μήνες.}$$

Ως αριστερό άκρο της πρώτης κλάσης πήραμε το μηδέν και τα διαστήματα ημιανοιχτά από δεξιά.

Να σημειώσουμε ότι αν για τον υπολογισμό του *k* ή του *r* κάνουμε στρογγυλοποιήσεις, αυτές πρέπει να γίνονται προς τα πάνω ώστε να διασφαλίζουμε ότι οι κλάσεις που θα δημιουργηθούν θα καλύπτουν όλα τα δεδομένα.

Παρατηρούμε ότι το 68% των τιμών του δείγματος συγκεντρώνεται στις δύο πρώτες κλάσεις, δηλαδή, στο 68% των 50 αγελάδων στις οποίες έγιναν μετρήσεις, ο χρόνος επανεμφάνισης της ασθένειας ήταν μικρότερος από 3 μήνες, μάλιστα, στο 42% η ασθένεια επανεμφανίσθηκε πριν συμπληρωθούν 1.5 μήνες από την πρώτη εμφάνισή της. Επίσης, παρατηρούμε ότι οι συχνότητες φθίνουν καθώς ο χρόνος επανεμφάνισης της νόσου αυξάνει. Σε ένα ποσοστό 6% των αγελάδων, η νόσος επανεμφανίσθηκε μετά από 7.5 ή και περισσότερους μήνες. Ως **γενικό συμπέρασμα** για την κατανομή αυτού του τυχαίου δείγματος, δηλαδή για την κατανομή των χρόνων επανεμφάνισης της νόσου στις 50 τυχαία επιλεγμένες γαλακτοπαραγωγές αγελάδες, μπορούμε να πούμε ότι ένα πολύ μεγάλο ποσοστό, 68%, των τιμών του δείγματος συγκεντρώνεται στο αριστερό άκρο της κατανομής, στο διάστημα [0 3) μήνες, και ότι οι συχνότητες φθίνουν αυξανομένου του χρόνου.

■

Παράδειγμα 9.1.5 (συνέχεια του Παραδείγματος 9.2): *Ο Πίνακας 9.1.5 που ακολουθεί είναι ο πίνακας συχνοτήτων των τιμών του δείγματος από τη μεταβλητή μηνιαίο οικογενειακό εισόδημα του Παραδείγματος 9.2, ομαδοποιημένων σε 6 κλάσεις πλάτους 200€ η κάθε μια.*

Εισόδημα	ν_i	f_i	N_i	F_i
[900 1100)	1	0.05	1	0.05
[1100 1300)	4	0.20	5	0.25
[1300 1500)	6	0.30	11	0.55
[1500 1700)	4	0.20	15	0.75
[1700 1900)	3	0.15	18	0.90
[1900 2100)	2	0.10	20	1.00
Σύνολα	**20**	**1.00**		

Πίνακας 9.1.5
Ο πίνακας συχνοτήτων των τιμών του δείγματος από τη μεταβλητή «μηνιαίο οικογενειακό εισόδημα» του Παραδείγματος 9.2 ομαδοποιημένων σε 6 κλάσεις

Από τον *Πίνακα 9.1.5*, ως **γενικό συμπέρασμα** για την κατανομή του τυχαίου δείγματος των *μηνιαίων οικογενειακών εισοδημάτων* των 20 τυχαία επιλεγμένων οικογενειών, μπορούμε να πούμε ότι το 50% των τιμών του δείγματος συγκεντρώνεται σε ένα κεντρικό διάστημα τιμών, συγκεκριμένα στο διάστημα, [1300 1700) με την κλάση [1300 1500) να έχει τη μεγαλύτερη συχνότητα και τις συχνότητες αριστερά αυτής της κλάσης να αυξάνουν αυξανομένου του εισοδήματος ενώ δεξιά αυτής της κλάσης, αυξανομένου του εισοδήματος, να φθίνουν.

■

Παρατήρηση 9.1.2: *α) Είναι προφανές, όμως το επισημαίνουμε, ότι η κατανομή συχνοτήτων (και σχετικών και αθροιστικών) επηρεάζεται από την επιλογή των κλάσεων. β) Παρατηρείστε ότι από τον πίνακα συχνοτήτων ομαδοποιημένων*

δεδομένων (όπως οι Πίνακες 9.1.4 & 9.1.5) δε μπορούμε να συμπεράνουμε αν και πόσες φόρες μια συγκεκριμένη τιμή εμφανίσθηκε στο δείγμα. γ) Η κατανομή σχετικών συχνοτήτων ενός τυχαίου δείγματος από συνεχή τυχαία μεταβλητή που προκύπτει μετά από ομαδοποίηση, συνδέεται με τη συνάρτηση πυκνότητας της τυχαίας μεταβλητής. Στο θέμα αυτό θα αναφερθούμε στη συνέχεια, αφού πρώτα μιλήσουμε για το ιστόγραμμα.

■

Ας δούμε τώρα ποιες δυνατότητες γραφικής παρουσίασης δεδομένων προσφέρει η *Περιγραφική Στατιστική* για *ποσοτικές μεταβλητές*.

9.1.2 Γραφική παρουσίαση κατανομής συχνοτήτων

Για την περιγραφή της κατανομής δεδομένων από μια *ποσοτική* μεταβλητή, η *Περιγραφική Στατιστική*, μας προσφέρει πολλές δυνατότητες γραφικής παρουσίασής της. Στη συνέχεια θα παρουσιάσουμε αρκετές από αυτές. Κυρίως, θα προσπαθήσουμε να εξηγήσουμε πώς «διαβάζουμε» ένα διάγραμμα γραφικής αναπαράστασης της κατανομής των δεδομένων, δηλαδή, πώς το ερμηνεύουμε και τι μπορούμε να συμπεράνουμε για την κατανομή των δεδομένων από αυτό. Θα αναφερθούμε βέβαια και στην κατασκευή τους, όμως όπως αναφέραμε και στο εισαγωγικό κεφάλαιο του Β΄ Μέρους *($8^{ο}$ Κεφάλαιο)*, αυτό είναι πλέον πολύ εύκολο να γίνει με χρήση κατάλληλου λογισμικού (αρκεί βέβαια να ξέρουμε τι ζητάμε από το λογισμικό) και γι' αυτό θα επιμείνουμε στην ερμηνεία τους.

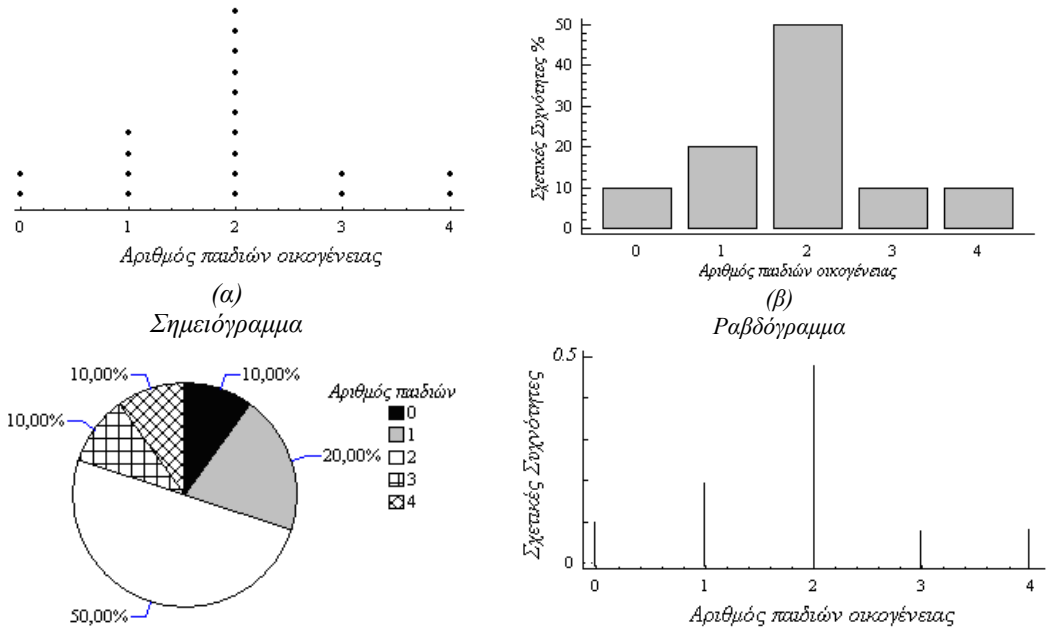
Ειδικότερα, θα αναφερθούμε στα ακόλουθα διαγράμματα.

- ***Σημειόγραμμα***
- ***Ραβδόγραμμα συχνοτήτων*** *και* ***σχετικών συχνοτήτων***
- ***Διάγραμμα συχνοτήτων*** *και* ***σχετικών συχνοτήτων***
- ***Κυκλικό διάγραμμα συχνοτήτων*** *και* ***σχετικών συχνοτήτων***
- ***Ιστόγραμμα συχνοτήτων*** και αντίστοιχα*,* ***σχετικών συχνοτήτων, αθροιστικών συχνοτήτων*** *και* ***αθροιστικών σχετικών συχνοτήτων***
- ***Πολύγωνο συχνοτήτων*** και αντίστοιχα*,* ***σχετικών συχνοτήτων, αθροιστικών συχνοτήτων*** *και* ***αθροιστικών σχετικών συχνοτήτων***
- ***Φυλλογράφημα***
- ***Θηκόγραμμα.***

Τους τρόπους γραφικής παρουσίασης των δεδομένων, επιλέξαμε να τους παρουσιάσουμε μέσω παραδειγμάτων. Να σημειώσουμε, ότι το *Θηκόγραμμα* θα το παρουσιάσουμε αργότερα, αφού πρώτα μιλήσουμε για τα *αριθμητικά περιγραφικά μέτρα*.

Παράδειγμα 9.1.6 (σημειόγραμμα, ραβδόγραμμα, κυκλικό διάγραμμα και διάγραμμα συχνοτήτων): *Στα τέσσερα σχήματα που ακολουθούν φαίνονται τέσσερις διαφορετικές γραφικές αναπαραστάσεις/απεικονίσεις της κατανομής του δείγματος από τη μεταβλητή U (αριθμός παιδιών οικογένειας) του Παραδείγματος 9.2.*

Στο *Σχήμα 9.1.1α* φαίνεται η *κατανομή συχνοτήτων* του δείγματος ως ***σημειόγραμμα (dot diagram)*** ενώ στα *Σχήματα 9.1.1β,γ,δ* φαίνεται η *κατανομή σχετικών συχνοτήτων* του δείγματος ως ***ραβδόγραμμα (barchart)***, ***κυκλικό διάγραμμα (piechart)*** και ***διάγραμμα σχετικών συχνοτήτων (line diagram),*** αντίστοιχα.



(α)
Σημειόγραμμα

(β)
Ραβδόγραμμα

(γ)
Κυκλικό διάγραμμα

(δ)
Διάγραμμα σχετικών συχνοτήτων

Σχήματα 9.1.1
Γραφική αναπαράσταση της κατανομής του δείγματος από τη μεταβλητή «αριθμός παιδιών οικογένειας» του Παραδείγματος 9.2

Και οι τέσσερις γραφικές απεικονίσεις, δίνουν μια πιο παραστατική και πιο ευκρινή εικόνα της κατανομής του δείγματος από αυτήν που δίνει ο *πίνακας συχνοτήτων*. Βέβαια, δε μας δίνουν περισσότερη ή διαφορετική πληροφορία από αυτήν που μας δίνει ο *πίνακας συχνοτήτων*, γιατί απλούστατα κατασκευάσθηκαν με βάση την πληροφορία που παίρνουμε από αυτόν και αυτή την πληροφορία απεικονίζουν γραφικά. Όμως δίνουν αυτήν την πληροφορία πιο παραστατικά και, λογικά, την κάνουν πιο εύκολα κατανοητή. Ιδιαίτερα, συμπεράσματα που αφορούν τη μορφή και τη θέση της κατανομής, προκύπτουν με πιο άμεσο και προφανή τρόπο και χωρίς να απαιτείται ιδιαίτερη εμπειρία. Θυμηθείτε το γενικό συμπέρασμα που διατυπώσαμε όταν σχολιάσαμε τον αντίστοιχο *πίνακα συχνοτήτων (Πίνακας 9.1.2)* και παρατηρείστε πόσο άμεσα και αβίαστα προκύπτει αυτό το συμπέρασμα από το *σημειόγραμμα* ή από το *ραβδόγραμμα* ή από το *διάγραμμα σχετικών συχνοτήτων*. Παρατηρείστε επίσης, ότι είναι πλέον ευδιάκριτη μια εμφανής *συμμετρία* της κατανομής γύρω από την τιμή 2 που παρουσιάζει τη μεγαλύτερη συχνότητα.

Ο τρόπος κατασκευής των διαγραμμάτων αυτών, είναι προφανής και απλός.

- Στο ***σημειόγραμμα***, απεικονίζουμε τα ***δεδομένα*** ως κουκίδες στις αντίστοιχες θέσεις ενός οριζόντιου άξονα.
- Στο ***ραβδόγραμμα***, απεικονίζουμε τις ***συχνότητες*** ή τις ***σχετικές συχνότητες*** των διαφορετικών τιμών y_i, $i=1,2,\ldots,k$, ως ύψη ορθογωνίων που σχεδιάζουμε στις αντίστοιχες θέσεις του οριζόντιου άξονα. Τα ορθογώνια έχουν ίδιο πλάτος βάσης που επιλέγουμε αυθαίρετα. Επίσης, το *ραβδόγραμμα* μπορεί να σχεδιασθεί με οριζόντιο, αντί κατακόρυφο, προσανατολισμό.
- Στο ***διάγραμμα συχνοτήτων (ή σχετικών συχνοτήτων)***, απεικονίζουμε τις ***συχνότητες*** (αντίστοιχα τις ***σχετικές συχνότητες***) των διαφορετικών τιμών y_i, $i=1,2,\ldots,k$ όπως και στο *ραβδόγραμμα*, με τη διαφορά ότι στις θέσεις των y_i χαράσσουμε κάθετα ευθύγραμμα τμήματα αντί ορθογωνίων.

- Στο ***κυκλικό διάγραμμα***, απεικονίζουμε τις ***συχνότητες*** ή τις ***σχετικές συχνότητες*** των διαφορετικών τιμών y_i, $i=1,2,\ldots,k$ με ένα διαφορετικό τρόπο. Πρόκειται για έναν κυκλικό δίσκο χωρισμένο σε k κυκλικούς τομείς, έναν για κάθε y_i, τα τόξα των οποίων, έστω φ_i, είναι ανάλογα με τις αντίστοιχες *συχνότητες* και *σχετικές συχνότητες*. Συγκεκριμένα,

$$\varphi_i = \nu_i \cdot \frac{360^0}{\nu} = 360^0 \cdot f_i, \quad i=1,2,\ldots,k\,.$$

■

Παρατηρήσεις 9.1.3:

α) Το κυκλικό διάγραμμα και το σημειόγραμμα είναι αποτελεσματικά/χρήσιμα όταν οι διαφορετικές τιμές, y_i, της μεταβλητής που εμφανίζονται στο δείγμα είναι λίγες, διαφορετικά γίνονται «δυσανάγνωστα» και επομένως, χωρίς πρακτική αξία.

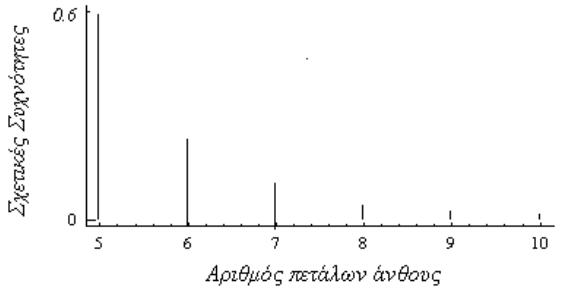
β) Το ραβδόγραμμα (αλλά και το κυκλικό διάγραμμα), χρησιμοποιείται και για δεδομένα από ποιοτικές μεταβλητές. Μάλιστα, στη βιβλιογραφία αλλά και στα στατιστικά πακέτα, σε πολλές περιπτώσεις, το ραβδόγραμμα προτείνεται μόνο για ποιοτικά δεδομένα και όχι για ποσοτικά. Αντίθετα, άλλοι συγγραφείς, για διακριτά ποσοτικά δεδομένα προτείνουν μόνο το ραβδόγραμμα και δεν αναφέρονται στο διάγραμμα συχνοτήτων. Η άποψη μας είναι ότι αν κρίνουμε ότι το ραβδόγραμμα θα βοηθήσει αυτούς στους οποίους απευθύνεται να κατανοήσουν την κατανομή των δεδομένων, τότε μπορεί να χρησιμοποιείται και όταν τα δεδομένα προέρχονται από διακριτή ποσοτική μεταβλητή. Όμως, όχι αντί του διαγράμματος συχνοτήτων αλλά συμπληρωματικά.

γ) Το διάγραμμα συχνοτήτων (ή σχετικών συχνοτήτων) κατασκευάζεται για δεδομένα που προέρχονται από διακριτή ποσοτική μεταβλητή. Αν τα δεδομένα προέρχονται από συνεχή ποσοτική μεταβλητή αντί του διαγράμματος συχνοτήτων (ή σχετικών συχνοτήτων) κατασκευάζεται το αντίστοιχο ιστόγραμμα ή/και το φυλλογράφημα τα οποία όπως θα δούμε κατασκευάζονται και για δεδομένα που προέρχονται από διακριτή ποσοτική μεταβλητή.

*δ) Το **διάγραμμα σχετικών συχνοτήτων** ενός **<u>τυχαίου</u>** δείγματος τιμών, $x_1, x_2, \ldots, x_\nu$, από μια διακριτή τυχαία μεταβλητή X, μας δίνει μια **εμπειρική/προσεγγιστική εικόνα του διαγράμματος πιθανοτήτων** της τυχαίας μεταβλητής X.*

■

Στο *Σχήμα 9.1.2* φαίνεται το *διάγραμμα σχετικών συχνοτήτων* της κατανομής του *τυχαίου δείγματος* του *Παραδείγματος 9.1*. Και στο παράδειγμα αυτό, διαπιστώνουμε ότι από το *διάγραμμα σχετικών συχνοτήτων,* άμεσα προκύπτει το γενικό συμπέρασμα που διατυπώσαμε για την κατανομή αυτού του τυχαίου δείγματος όταν σχολιάσαμε τον αντίστοιχο *πίνακα συχνοτήτων (Πίνακας 9.1.1)*. Μπορούμε επίσης να κατασκευάσουμε και το *κυκλικό διάγραμμα* αλλά και το *σημειόγραμμα* γιατί οι διαφορετικές τιμές που εμφανίσθηκαν στο τυχαίο δείγμα είναι λίγες (δείτε το ως άσκηση).



Σχήμα 9.1.2

Το διάγραμμα σχετικών συχνοτήτων της κατανομής του δείγματος από την τ.μ.

«αριθμός πετάλων των ανθέων συγκεκριμένης ποικιλίας που καλλιεργείται στο νομό Κοζάνης» του Παραδείγματος 9.1

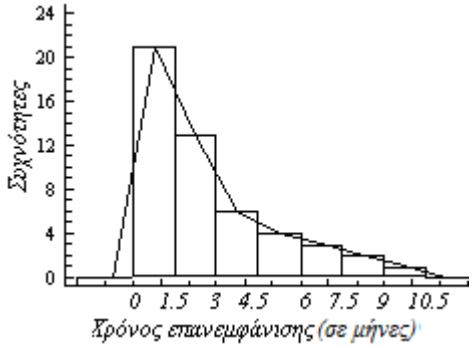
■

Παράδειγμα 9.1.7 (ιστογράμματα): *Ας δούμε πώς μπορούμε να παρουσιάσουμε γραφικά την κατανομή του τυχαίου δείγματος του Παραδείγματος 9.1.3 (χρόνοι επανεμφάνισης της ασθένειας).*

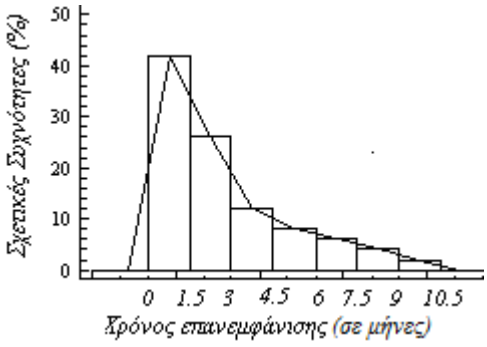
Οι διαφορετικές τιμές, y_i, της μεταβλητής X που εμφανίσθηκαν στο δείγμα είναι πολλές και επομένως δεν έχει πρακτική αξία να κατασκευάσουμε το *σημειόγραμμα* ή το *κυκλικό διάγραμμα* του δείγματος.

Όταν το τυχαίο δείγμα προέρχεται από *συνεχή ποσοτική* μεταβλητή, όπως στο παράδειγμά μας, για τη γραφική παρουσίαση της κατανομής του ενδείκνυνται τα *ιστογράμματα,* τα *πολύγωνα συχνοτήτων* και το *φυλλογράφημα* (είτε οι διαφορετικές τιμές είναι πολλές είτε είναι λίγες).

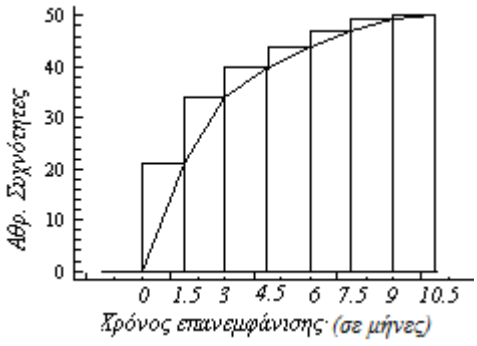
Στα *Σχήματα 9.1.3* φαίνεται η κατανομή του τυχαίου δείγματος του παραδείγματός μας ως ***ιστόγραμμα (histogram)*** και ***πολύγωνο (polygon) συχνοτήτων***, ***σχετικών συχνοτήτων, αθροιστικών συχνοτήτων*** και ***αθροιστικών σχετικών συχνοτήτων***, αντίστοιχα.



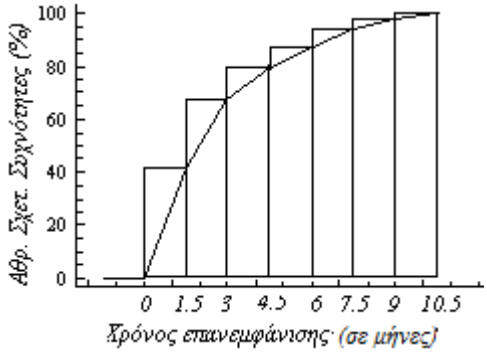
(α)
Ιστόγραμμα και πολύγωνο συχνοτήτων



(β)
Ιστόγραμμα και πολύγωνο σχετικών συχνοτήτων



(γ)
Ιστόγραμμα και πολύγωνο αθροιστικών συχνοτήτων



(δ)
Ιστογραμμα και πολύγωνο αθροιστικών σχετικών συχνοτήτων

Σχήματα 9.1.3
Γραφική αναπαράσταση της κατανομής του δείγματος από τη μεταβλητή «χρόνος επανεμφάνισης μιας ασθένειας σε γαλακτοπαραγωγές αγελάδες» του Παραδείγματος 9.1.3.

Τα ***ιστογράμματα*** κατασκευάζονται για τη γραφική παρουσίαση/αναπαράσταση της πληροφορίας που παίρνουμε από έναν *πίνακα συχνοτήτων* ομαδοποιημένων δεδομένων. Για να κατασκευάσουμε ένα *ιστόγραμμα*, σημειώνουμε στον οριζόντιο

άξονα ενός ορθογωνίου συστήματος αξόνων, με κατάλληλη κλίμακα, τα άκρα των κλάσεων, και στα διαδοχικά διαστήματα που ορίζουν αυτά, υψώνουμε διαδοχικά ορθογώνια. Κάθε ορθογώνιο σχεδιάζεται έτσι ώστε, το **εμβαδόν** του να είναι ίσο με τη ***συχνότητα*** της αντίστοιχης κλάσης αν πρόκειται για το ***ιστόγραμμα συχνοτήτων***, ή με τη ***σχετική συχνότητα*** της αντίστοιχης κλάσης αν πρόκειται για το ***ιστόγραμμα σχετικών συχνοτήτων***, ή με την ***αθροιστική συχνότητα*** της αντίστοιχης κλάσης αν πρόκειται για το ***ιστόγραμμα αθροιστικών συχνοτήτων***, ή με την ***αθροιστική σχετική συχνότητα*** της αντίστοιχης κλάσης αν πρόκειται για το ***ιστόγραμμα αθροιστικών σχετικών συχνοτήτων***.

Αν στο *ιστόγραμμα συχνοτήτων*, αριστερά της πρώτης κλάσης και δεξιά της τελευταίας, θεωρήσουμε από μία υποθετική κλάση με συχνότητα μηδέν και στη συνέχεια ενώσουμε με ευθύγραμμα τμήματα τα μέσα των πάνω βάσεων των ορθογωνίων, δημιουργείται το ***πολύγωνο συχνοτήτων***. Ομοίως δημιουργείται το ***πολύγωνο σχετικών συχνοτήτων*** (από το *ιστόγραμμα σχετικών συχνοτήτων*).

Το ***πολύγωνο αθροιστικών συχνοτήτων*** και το ***πολύγωνο αθροιστικών σχετικών συχνοτήτων*** δημιουργούνται αν στα αντίστοιχα ιστογράμματα, ενώσουμε με ευθύγραμμα τμήματα τα δεξιά άκρα των πάνω βάσεων των ορθογωνίων.

Στο *ιστόγραμμα συχνοτήτων*, το συνολικό εμβαδόν των ορθογωνίων προφανώς είναι ίσο με το μέγεθος του δείγματος ν, και αντίστοιχα, στο *ιστόγραμμα σχετικών συχνοτήτων* είναι ίσο με 1 (ή με 100). Επίσης, το εμβαδόν που περικλείεται μεταξύ του *πολυγώνου συχνοτήτων* (αντίστοιχα, *σχετικών συχνοτήτων)* και του οριζόντιου άξονα είναι ίσο με ν (αντίστοιχα, ίσο με 1 ή με 100).

Σε ότι αφορά στα ύψη των ορθογωνίων, προφανώς πρέπει να είναι τέτοια ώστε τα εμβαδά τους να είναι ίσα με τις αντίστοιχες *συχνότητες/σχετικές συχνότητες*.

Έτσι, στο *ιστόγραμμα σχετικών συχνοτήτων*, το ύψος κάθε ορθογωνίου υπολογίζεται ***αν διαιρέσουμε τη σχετική συχνότητα της αντίστοιχης κλάσης, δηλαδή το ποσοστό των δεδομένων στην αντίστοιχη κλάση, με το πλάτος της κλάσης***. Επομένως, το ύψος κάθε ορθογωνίου εκφράζει το ποσοστό της συγκέντρωσης δεδομένων στην αντίστοιχη κλάση, ***ανά μονάδα*** του οριζόντιου άξονα, δηλαδή εκφράζει ***πυκνότητα***, και γι΄αυτό αυτή η κλίμακα απεικόνισης ονομάζεται ***κλίμακα πυκνότητας (density scale)***.

Στην περίπτωση που όλες οι κλάσεις έχουν το ίδιο πλάτος, όπως στο παράδειγμά μας, συνηθίζεται να κάνουμε το εξής *«trick»*. Ως μονάδα μέτρησης της μεταβλητής στον οριζόντιο άξονα θεωρούμε το πλάτος των κλάσεων. Με αυτό τον τρόπο, και τα εμβαδά αλλά και τα ύψη των ορθογωνίων είναι ίσα με τις αντίστοιχες *σχετικές συχνότητες*. Επομένως, στην περίπτωση αυτή, το ύψος κάθε ορθογωνίου εκφράζει το ποσοστό της συγκέντρωσης δεδομένων στην αντίστοιχη κλάση.

Το *ιστόγραμμα σχετικών συχνοτήτων* του παραδείγματος μας (*Σχήμα 9.1.3β*), με αυτόν τον τρόπο σχεδιάσθηκε. Θεωρήσαμε ως μονάδα στον οριζόντιο άξονα όχι το μήνα αλλά τους 1.5 μήνες και έτσι τα ύψη των ορθογωνίων είναι ίσα με τις σχετικές συχνότητες των αντίστοιχων κλάσεων, όπως τα εμβαδά, δηλαδή είναι ίσα (από αριστερά προς τα δεξιά) με 42%, 26%, 12%, 8%, 6%, 4% και 2% αντίστοιχα, γι'αυτό και στον κατακόρυφο άξονα γράφουμε *«Σχετικές Συχνότητες (%)»*. Αν χρησιμοποιούσαμε *κλίμακα πυκνότητας* τα ύψη θα ήταν, αντίστοιχα, ίσα με 28%, 17.3%, 8%, 5.3%, 4%, 2.7% και 1.3% και τα εμβαδά θα ήταν και πάλι ίσα με 42%, 26%, 12%, 8%, 6%, 4% και 2% όμως δεν θα προέκυπταν με απλή ανάγνωση του κατακόρυφου άξονα αλλά κάνοντας τους αντίστοιχους πολλαπλασιασμούς

($28\cdot 1.5, 17.3\cdot 1.5$ κτλ.). Στην περίπτωση αυτή, στον κατακόρυφο άξονα θα γράφαμε *«Σχετικές Συχνότητες ανά μήνα (%)»*. Ανάλογα, μπορούμε να εργασθούμε για την κατασκευή και των άλλων τύπων ιστογράμματος. Όμως, επαναλαμβάνουμε, μπορούμε να κάνουμε αυτό το *«trick»* που μας διευκολύνει και στην κατασκευή αλλά και στην ανάγνωση του ιστογράμματος μόνο όταν οι κλάσεις είναι του ιδίου πλάτους.

Και τα τέσσερα ιστογράμματα, δίνουν μια πιο παραστατική και πιο ευκρινή εικόνα της κατανομής του δείγματος από αυτήν του *πίνακα συχνοτήτων*. Βέβαια, όπως και οι γραφικές παρουσιάσεις μη ομαδοποιημένων δεδομένων, τα *ιστογράμματα* δε μας δίνουν περισσότερη ή/και διαφορετική πληροφορία από αυτήν που μας δίνει ο *πίνακας συχνοτήτων*, γιατί απλούστατα κατασκευάζονται με βάση την πληροφορία που παίρνουμε από αυτόν και αυτή την πληροφορία απεικονίζουν γραφικά. Όμως δίνουν αυτήν την πληροφορία πιο παραστατικά. Έτσι, από τα *ιστογράμματα*, για παράδειγμα, των *σχετικών* και των *αθροιστικών σχετικών συχνοτήτων*, μπορούμε πολύ εύκολα να πάρουμε το ποσοστό των τιμών του δείγματος που βρίσκονται μεταξύ των άκρων μιας κλάσης ή μιας ομάδας διαδοχικών κλάσεων ή αριστερά/δεξιά ενός άκρου μιας κλάσης. Ιδιαίτερα, συμπεράσματα που αφορούν τη μορφή και τη θέση της κατανομής, προκύπτουν με πιο άμεσο και προφανή τρόπο και χωρίς να απαιτείται ιδιαίτερη εμπειρία. Θυμηθείτε το γενικό συμπέρασμα που διατυπώσαμε όταν σχολιάσαμε τον αντίστοιχο *πίνακα συχνοτήτων (Πίνακας 9.1.4)* και παρατηρείστε πόσο άμεσα και αβίαστα προκύπτει αυτό το συμπέρασμα από τα *ιστογράμματα*. Παρατηρείστε επίσης, ότι είναι πιο ευδιάκριτη μια εμφανής *ασυμμετρία* της κατανομής.

■

Πολύγωνο σχετικών συχνοτήτων και συνάρτηση πυκνότητας

Στα προηγούμενα, είδαμε ότι το διάγραμμα σχετικών συχνοτήτων ενός τυχαίου δείγματος τιμών μιας διακριτής τυχαίας μεταβλητής αποτελεί μια προσέγγιση του διαγράμματος πιθανοτήτων της τυχαίας μεταβλητής. Είναι λογικό, να περιμένουμε να συμβαίνει κάτι ανάλογο, με το ιστόγραμμα και το πολύγωνο σχετικών συχνοτήτων ενός τυχαίου δείγματος τιμών μιας συνεχούς τυχαίας μεταβλητής και τη συνάρτηση πυκνότητάς της.

Πράγματι, το πολύγωνο σχετικών συχνοτήτων αποτελεί μια εκτίμηση/προσέγγιση της γραφικής παράστασης της συνάρτησης πυκνότητας. Θυμηθείτε ότι όταν στο Α΄ Μέρος μιλήσαμε για τη συνάρτηση πυκνότητας, f, μιας συνεχούς τυχαίας μεταβλητής X, είδαμε ότι οι πιθανότητες που αφορούν την X υπολογίζονται ως εμβαδά κατάλληλων περιοχών κάτω από τη γραφική παράσταση της f και ότι για την τιμή της, $f(\alpha)$, στη θέση $x=\alpha$, έχουμε

$$P(\alpha-\frac{\varepsilon}{2}<X<\alpha+\frac{\varepsilon}{2})\cong f(\alpha)\varepsilon$$

ή

$$f(\alpha)\cong\frac{P(\alpha-\frac{\varepsilon}{2}<X<\alpha+\frac{\varepsilon}{2})}{\varepsilon}\ . \qquad (9.1.1)$$

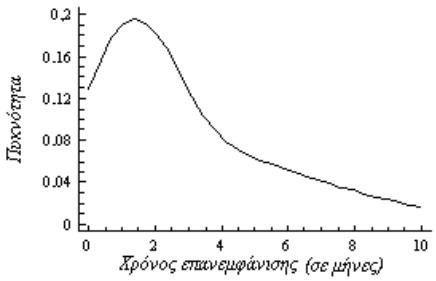
Δηλαδή, η τιμή της συνάρτησης πυκνότητας μιας συνεχούς τυχαίας μεταβλητής στη θέση α είναι ίση με το λόγο της πιθανότητας να πάρει η X τιμές σε ένα διάστημα πλάτους ε γύρω από το α προς το πλάτος αυτό.

Κάτι ανάλογο συμβαίνει, όπως είδαμε, με το ιστόγραμμα σχετικών συχνοτήτων των τιμών του τυχαίου δείγματος από την X (και την προσέγγισή του, το πολύγωνο σχετικών συχνοτήτων). Αν επιλέξουμε τυχαία μια τιμή από το δείγμα του παραδείγματός μας, η πιθανότητα να βρίσκεται αυτή η τιμή, για παράδειγμα, μεταξύ 1.5 και 6 είναι ίση με το

συνολικό εμβαδόν των ορθογωνίων που αντιστοιχούν στις κλάσεις [1.5 3), [3 4.5) και [4.5 6), δηλαδή, είναι ίση με 0.46. Αυτή η πιθανότητα αποτελεί μια προσέγγιση της πιθανότητας $P(1.5 < X < 6)$*, δηλαδή της πιθανότητας, αν επιλέξουμε τυχαία από τον πληθυσμό μια τιμή, αυτή να βρίσκεται μεταξύ 1.5 και 6. Επίσης, το ύψος κάθε ορθογωνίου ορίσθηκε ως λόγος, όπως αυτός στη σχέση (9.1.1). (Θυμηθείτε την κλίμακα πυκνότητας).*

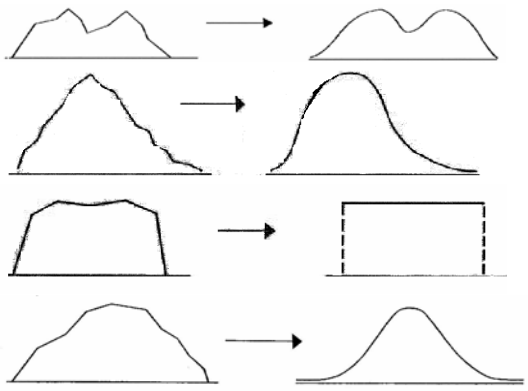
Ομαδοποιώντας τα δεδομένα, όπως ήδη έχουμε αναφέρει, χάνουμε κάποια από την πληροφορία που περιέχεται στο δείγμα και γι'αυτό, είναι λογικό, όσο το μέγεθος του δείγματος αυξάνεται να παίρνουμε όλο και μεγαλύτερο αριθμό κλάσεων με πλάτος όλο και πιο μικρό. Έτσι, η συνάρτηση πυκνότητας της συνεχούς τυχαίας μεταβλητής από την οποία παίρνουμε τα τυχαία δείγματα προσεγγίζεται από τα αντίστοιχα πολύγωνα συχνοτήτων όλο και καλύτερα. Μάλιστα, όσο το μέγεθος του δείγματος αυξάνεται και το πλάτος των κλάσεων μειώνεται, το πολύγωνο σχετικών συχνοτήτων παίρνει όλο και πιο πολύ, μορφή λείας καμπύλης. Και επειδή οι συνεχείς μεταβλητές (τουλάχιστον θεωρητικά) μπορούν να πάρουν άπειρες τιμές, η καμπύλη σχετικών συχνοτήτων ολόκληρου του πληθυσμού προκύπτει ως το όριο των πολυγώνων συχνοτήτων στα οποία το πλάτος των κλάσεων τείνει στο μηδέν και πρόκειται, φυσικά, για τη γραφική παράσταση της συνάρτησης πυκνότητας.

Στο Σχήμα 9.1.4, φαίνεται η λεία καμπύλη σχετικών συχνοτήτων που προσεγγιστικά προκύπτει για τη μεταβλητή «χρόνος επανεμφάνισης της ασθένειας» του Παραδείγματος 9.1.3. Παρατηρείστε την κλίμακα πυκνότητας στον κατακόρυφο άξονα.



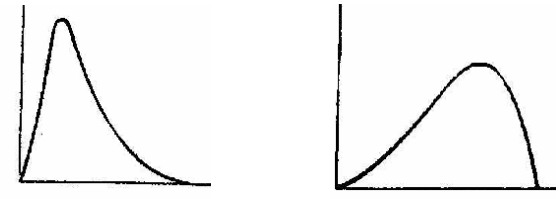
Σχήμα 9.1.4
Καμπύλη σχετικών συχνοτήτων του δείγματος του Παραδείγματος 9.1.3

Τα *πολύγωνα* και αντίστοιχα οι *καμπύλες σχετικών συχνοτήτων* μπορεί να έχουν διάφορες μορφές. Δείτε σχετικά παραδείγματα στο *Σχήματα 9.1.5*.



Σχήματα 9.1.5
Πολύγωνα σχετικών συχνοτήτων
και λείες καμπύλες που προκύπτουν από αυτά

Μπορεί να παρουσιάζουν μια, καμία ή και περισσότερες κορυφές. Μπορεί επίσης να είναι συμμετρικές ή να είναι λοξές/ασσύμμετρες. Μία λοξή/ασύμμετρη καμπύλη σχετικών συχνότητων παρουσιάζει είτε *θετική ασυμμετρία* οπότε εμφανίζει πιο μακριά ουρά προς τα δεξιά είτε *αρνητική ασυμμετρία* οπότε εμφανίζει πιο μακκριά ουρά προς αρστερά *(δες Σχήματα 9.1.6)*. Περισσότερα για το νόημα και την ερμηνεία του είδους της ασυμμετρίας, θα αναφέρουμε στα επόμενα όταν θα μιλήσουμε για τα *αριθμητικά περιγραφικά μέτρα*.



(α) *(β)*
Θετική ασυμμετρία *Αρνητική ασυμμετρία*

Σχήματα 9.1.6
Είδη ασυμμετρίας/λοξότητας

Τέλος, οι *καμπύλες συχνοτήτων,* ανάλογα με το βαθμό συγκέντρωσης των παρατηρήσεων στο μέσο και στα άκρα της κατανομής, διακρίνονται σε ***μεσόκυρτες, λεπτόκυρτες***, και ***πλατύκυρτες*** *(δες Σχήματα 9.1.7).*

(α) *(β)* *(γ)*
Μεσόκυρτη *Λεπτόκυρτη* *Πλατύκυρτη*

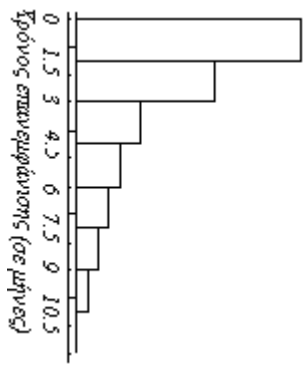
Σχήματα 9.1.7
Είδη κυρτότητας

Όπως ήδη έχουμε διαπιστώσει, με την ομαδοποίηση των τιμών του δείγματος χάνουμε κάποια από την πληροφορία που περιέχεται σε αυτό αφού τόσο ο *πίνακας συχνοτήτων* όσο και τα *ιστογράμματα* δε διατηρούν τις αρχικές τιμές του δείγματος. Αυτό το πρόβλημα μπορεί να αντιμετωπισθεί με την κατασκευή του ***φυλλογραφήματος (steam-leaf plot)*** των τιμών του δείγματος (*είναι μια από τις μεθόδους-τεχνικές της* ***διερευνητικής ανάλυσης δεδομένων)***.

Παράδειγμα 9.1.8 (φυλλογράφημα): *Στο Σχήμα 9.1.8α φαίνεται το φυλλογράφημα του τυχαίου δείγματος τιμών του Παραδείγματος 9.1.3. Ας δούμε πώς κατασκευάσθηκε και πώς «διαβάζεται».*

0	22345677888899
1	00233446789
2	001246789
3	39
4	0134
5	5679
6	5
7	04
8	7□
HI	9,9

(Leaf unit=0.1)



(α) *(β)*

Φυλλογράφημα *Ιστόγραμμα συχνοτήτων*

Σχήμα 9.1.8
Το φυλλογράφημα και το ιστόγραμμα συχνοτήτων του τυχαίου δείγματος τιμών του Παραδείγματος 9.1.3

Η κατασκευή του έγινε σε τέσσερα βήματα:
α) Χωρίσαμε κάθε τιμή του δείγματος σε δύο μέρη. Στο *steam* και το *leaf*. Τα *leaves* επιλέξαμε να εκφράζουν τα δέκατα (και επομένως *τα steams* τις μονάδες). Έτσι, για παράδειγμα, η τιμή 5.7 αναπαρίσταται ως 5|7 όπου το 5 είναι το *steam* και το 7 το *leaf* αυτής της τιμής.
β) Καταγράψαμε όλα τα *steams*, τα διατάξαμε κατακόρυφα σε αύξουσα σειρά και, δεξιά τους, χαράξαμε μια κατακόρυφη γραμμή.
γ) Στη γραμμή κάθε *steam,* καταγράψαμε το *leaf* κάθε τιμής του δείγματος που έχει το συγκεκριμένο *steam*. Έτσι, στη γραμμή που βρίσκεται το *steam* 5, καταγράψαμε δεξιά της κατακόρυφης γραμμής τα *leaves* 5, 6, 7, 9 γιατί οι τιμές του δείγματος που έχουν *steam* 5 είναι οι τιμές: 5.5, 5.6, 5.7 και 5.9. *(Τα leaves κάθε steam, τα καταγράφουμε σε αύξουσα σειρά).*
δ) Σημειώσαμε σε κάποιο σημείο του διαγράμματος (ως υπόμνημα) την αναγκαία πληροφορία για να είναι σαφές τι εκφράζουν τα *steams* και τα *leaves.*

Είναι φανερό ότι από ένα *φυλλογράφημα* μπορεί κανείς, αμέσως, να διαπιστώσει αν μια συγκεκριμένη τιμή ανήκει (και πόσες φορές) στο δείγμα, κάτι το οποίο, δεν είναι δυνατόν να γίνει από ένα *ιστόγραμμα*. Για παράδειγμα, από το παραπάνω *φυλλογράφημα* εύκολα διαπιστώνουμε ότι η τιμή 2.5 δεν εμφανίσθηκε στο δείγμα ενώ η τιμή 2.0 εμφανίσθηκε και μάλιστα δύο φορές.

Το *φυλλογράφημα*, επηρεάζεται από την επιλογή των *steams* όπως και το *ιστόγραμμα* επηρεάζεται από την επιλογή των κλάσεων. Αξίζει επίσης να σημειώσουμε ότι η εικόνα-μορφή ενός *φυλλογραφήματος* είναι ανάλογη με αυτήν του αντίστοιχου *ιστογράμματος συχνοτήτων*. Δείτε τα *Σχήματα 9.1.8* όπου το *ιστόγραμμα συχνοτήτων*, για διευκόλυνση της σύγκρισης, έχει περιστραφεί κατά 90^0.

■

Σημείωση 9.1.1: *Στα φυλλογραφήματα που κατασκευάζουν τα στατιστικά πακέτα, σημειώνονται και άλλες πληροφορίες όπως, σε ποιο steam βρίσκεται η διάμεσος, οι αθροιστικές συχνότητες και οι πιθανές «ακραίες» τιμές (μια τέτοια τιμή, η τιμή 9.9, φαίνεται και στο παράδειγμά μας και συμβολίζεται με HI).*

Παρατήρηση 9.1.4: *Κάποιες φορές, μπορεί τα διαφορετικά steams να είναι πολύ λίγα και κάποια (ή και όλα) να έχουν πολλά leaves. Σε αυτές τις περιπτώσεις, για καλύτερη παρουσίαση των δεδομένων μας, μπορούμε να «απλώσουμε» τα steams που έχουν πολλά leaves, γράφοντας καθένα από αυτά, σε δύο ή περισσότερες γραμμές, ανάλογα με τον αριθμό των leaves που έχουν. Συνήθως ένα steam, ανάλογα με τον αριθμό των leaves του, «απλώνεται» σε περισσότερες γραμμές με δύο τρόπους:*
α) Σε δύο γραμμές, όπου στην πρώτη καταγράφονται τα leaves από 0 μέχρι 4 και στη δεύτερη τα leaves από 5 μέχρι 9.
β) Σε πέντε γραμμές, όπου καταγράφονται, αντίστοιχα, τα leaves, 0 και 1, 2 και 3, 4 και 5, 6 και 7, 8 και 9.
Μια τέτοια περίπτωση, φαίνεται στο φυλλογράφημα του Σχήματος 9.1.9.

```
1 | 99
2 | 011
2 | 23
2 | 4455555
2 | 6677777
2 | 8899
3 | 0011
3 |
3 | 4
```

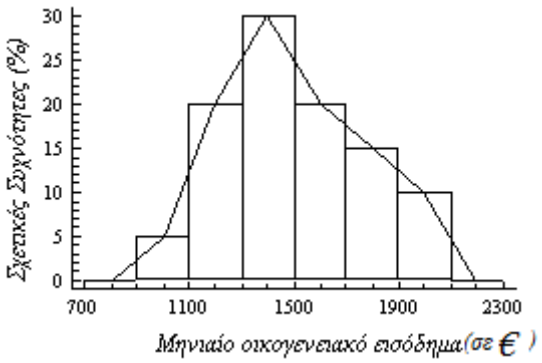
(Leaf unit=0.1)

Σχήμα 9.1.9
Φυλλογράφημα με ίδια steams
σε περισσότερες από μια γραμμές

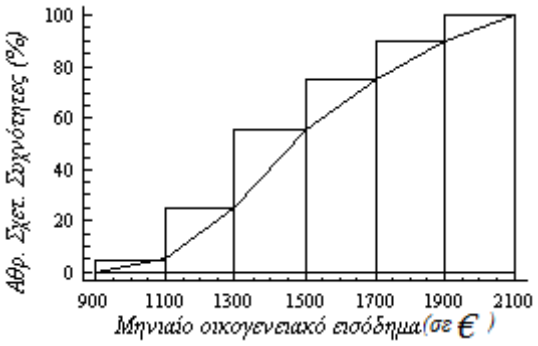
■

Παράδεγμα 9.1.5 (συνέχεια): *Στα Σχήματα 9.1.10 φαίνονται το ιστόγραμμα και το πολύγωνο σχετικών συχνοτήτων, το ιστόγραμμα και το πολύγωνο αθροιστικών σχετικών συχνοτήτων και το φυλλογράφημα της κατανομής του δείγματος από την τυχαία μεταβλητή μηνιαίο οικογενειακό εισόδημα του Παραδείγματος 9.2 και με βάση την ομαδοποίηση των παρατηρήσεων που έγινε στο Παράδειγμα 9.1.5.*

Στο *φυλλογράφημα*, ως *leaves* πήραμε τις δεκάδες (και ως *steams* τις εκατοντάδες). Έτσι, για παράδειγμα, η τιμή 1450 αναπαρίσταται ως 14|5.



(α)
Ιστόγραμμα και πολύγωνο
σχετικών συχνοτήτων



(β)
Ιστόγραμμα και πολύγωνο
αθροιστικών σχετικών συχνοτήτων

(Leaf unit=10.0)

```
10 | 0
11 |
12 | 0005
13 |
14 | 000055
15 |
16 | 0000
17 |
18 | 000
19 |
20 | 00
```

(γ)
Φυλλογράφημα

Σχήματα 9.1.10
Αναπαράσταση της κατανομής του δείγματος από την τυχαία μεταβλητή «μηνιαίο οικογενειακό εισόδημα» του Παραδείγματος 9.2.

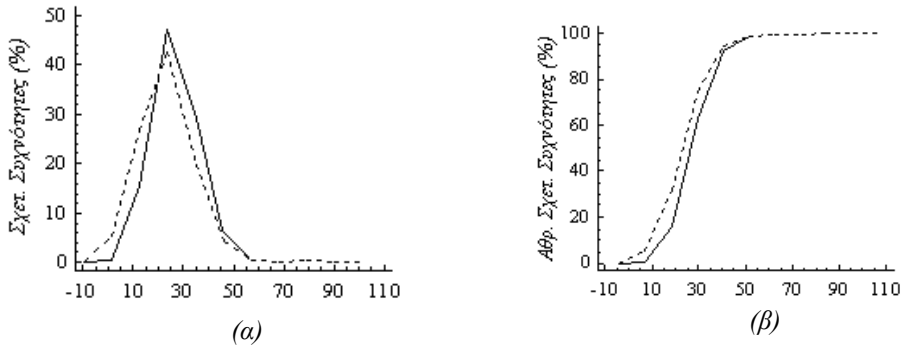
Και στο παράδειγμα αυτό, διαπιστώνουμε ότι από το *ιστόγραμμα σχετικών συχνοτήτων* αλλά και από το *φυλλογράφημα* άμεσα προκύπτει το γενικό συμπέρασμα που διατυπώσαμε για την κατανομή αυτού του τυχαίου δείγματος όταν σχολιάσαμε τον αντίστοιχο *πίνακα συχνοτήτων (Πίνακας 9.1.5)*. Φαίνεται επίσης, ότι η κατανομή του δείγματος είναι περίπου *συμμετρική*.

■

Στο σημείο αυτό αξίζει να κάνουμε μια παρατήρηση/επισήμανση.

Παρατήρηση 9.1.5: *Όπως θα έχετε παρατηρήσει, επιμένουμε στην ανίχνευση πιθανής συμμετρίας στην κατανομή του τυχαίου δείγματος. Αυτό το κάνουμε γιατί, όπως θα διαπιστώσουμε στα επόμενα, πολλές μέθοδοι στατιστικής συμπερασματολογίας προϋποθέτουν ότι το τυχαίο δείγμα προέρχεται από κανονικό πληθυσμό, δηλαδή, ότι οι τιμές του τυχαίου δείγματος είναι τιμές τυχαίας μεταβλητής που ακολουθεί κάποια κανονική κατανομή (η οποία, όπως γνωρίζουμε, είναι συμμετρική, μονοκόρυφη κατανομή και με «όχι παχιές» ουρές/μεσόκυρτη). Όταν επομένως το δείγμα είναι τυχαίο, μπορούμε, από τη μορφή της κατανομής του να κάνουμε μια πρώτη «εκτίμηση» για το αν ο πληθυσμός από τον οποίο προέρχεται μπορεί να είναι κανονικός ή απέχει πολύ από την κανονική κατανομή. Αυτό που παρατηρούμε στην κατανομή του δείγματος είναι αν αυτή είναι (περίπου) συμμετρική ή μήπως παρουσιάζει μεγάλη/εμφανή ασυμμετρία (επίσης, αν είναι μονοκόρυφη ή όχι και αν δεν έχει «παχιές» ουρές, δηλαδή, αν δεν εμφανίσθηκαν ακραίες τιμές, αλλά γι' αυτό, θα πούμε περισσότερα όταν θα μιλήσουμε για το θηκόγραμμα). Βέβαια, για το αν το δείγμα προέρχεται από κανονικό ή όχι πληθυσμό, όπως θα δούμε, δεν (πρέπει να) αποφασίζουμε μόνο από την εικόνα του πολυγώνου και του ιστογράμματος σχετικών συχνοτήτων του δείγματος. Υπάρχουν για το σκοπό αυτό ειδικοί στατιστικοί έλεγχοι. Εξάλλου, από μια άλλη ομαδοποίηση των τιμών του δείγματος, είναι πιθανόν να παίρναμε μια αρκετά διαφορετική εικόνα για τη μορφή της κατανομής του δείγματος. Το ίδιο θα μπορούσε επίσης να συμβεί αν παίρναμε ένα άλλο τυχαίο δείγμα. Όμως, μια πρώτη «εκτίμηση», ιδιαίτερα αν το δείγμα δεν είναι μικρό, είναι πάντοτε χρήσιμη.*

Ερώτηση: *Στα Σχήματα 9.1.11, φαίνονται τα πολύγωνα σχετικών συχνοτήτων και τα πολύγωνα αθροιστικών σχετικών συχνοτήτων δύο κατανομών δεδομένων. Σχολιάστε τη σχετική θέση των αντίστοιχων πολυγώνων στα δύο σχήματα.*



(α) *(β)*

Σχήματα 9.1.11
Σχετική θέση πολυγώνων σχετικών συχνοτήτων

<u>Υπόδειξη</u>: Η κατανομή της οποίας το πολύγωνο σχετικών συχνοτήτων και το πολύγωνο αθροιστικών σχετικών συχνοτήτων βρίσκονται δεξιότερα, είναι «στοχαστικά μεγαλύτερη». Σκεφθείτε τι μπορεί να σημαίνει αυτό. (Θεωρείστε ότι η κατανομή που βρίσκεται δεξιότερα, αφορά, για παράδειγμα, στις τιμές ενός αιματολογικού δείκτη σε μια ομάδα ανδρών και η άλλη κατανομή στις τιμές του ίδιου αιματολογικού δείκτη σε μια ομάδα γυναικών και συγκρίνετε τα ποσοστά στις δύο ομάδες που έχουν τιμή

αιματολογικού δείκτη, π.χ., μεγαλύτερη από 30 μονάδες ή μικρότερη από 10. Επίσης, βρείτε και συγκρίνετε τις τιμές, κάποιου ποσοστημορίου π.χ. του p_{60}, στις δύο κατανομές).

■

9.1.3 Αριθμητικά περιγραφικά μέτρα

Οι γραφικές αναπαραστάσεις της κατανομής του δείγματος (*διαγράμματα*, *ιστογράμματα* και *πολύγωνα συχνοτήτων*, *φυλλογράφημα*, κτλ) μας βοηθούν, όπως είδαμε, να αποκτήσουμε μια παραστατική εικόνα για τη θέση της και τη μορφή της. Όμως, υπάρχουν όρια στην περαιτέρω αξιοποίησή τους και ιδίως στη *στατιστική συμπερασματολογία*. Για παράδειγμα, αν θέλουμε να χρησιμοποιήσουμε το *ιστόγραμμα* ενός *τυχαίου δείγματος* για να οδηγηθούμε σε συμπεράσματα για το *ιστόγραμμα* του *πληθυσμού* από τον οποίο προέρχεται, πώς μπορούμε να μετρήσουμε τις ομοιότητες και τις διαφορές μεταξύ δύο ιστογραμμάτων; Αν ταυτίζονται, μπορούμε να πούμε ότι τα δύο ιστογράμματα «είναι ίδια», αν όμως υπάρχουν διαφορές (που είναι και το πιο πιθανό) πώς θα εκφράσουμε το «πόσο διαφέρουν»; Απαιτείται επομένως, ένα επόμενο βήμα. Πρέπει να μπορούμε να περιγράψουμε την κατανομή του δείγματος (όπως και τον πληθυσμό) και με ποσοτικούς όρους, ώστε η σύγκριση να μπορεί να γίνει με ποσοτικά/μετρήσιμα κριτήρια.

Τα ***αριθμητικά περιγραφικά μέτρα (numerical descriptive measures)*** αυτό ακριβώς μας προσφέρουν. Πρόκειται για ποσοτικά μεγέθη που βοηθούν στην περιγραφή της κατανομής ενός δείγματος ή στην περιγραφή ενός πληθυσμού (της κατανομής μιας τυχαίας μεταβλητής) με όρους ποσοτικούς. Αν αφορούν έναν *πληθυσμό,* ονομάζονται *(όπως είδαμε στο Α΄ Μέρος)* ***παράμετροι (parameters)*** ενώ αν αφορούν ένα *δείγμα* από έναν *πληθυσμό* ονομάζονται ***στατιστικά (statistics)***. Δύο βασικές *παράμετροι* που γνωρίσαμε στο Α΄ Μέρος είναι η *μέση τιμή* μ και η *διακύμανση* σ^2.

Οι *παράμετροι* ενός πληθυσμού είναι συγκεκριμένοι/μοναδικοί αριθμοί, που μπορεί βέβαια, να μας είναι είτε γνωστοί είτε άγνωστοι. Αντίθετα, τα *στατιστικά,* για συγκεκριμένη πραγματοποίηση $x_1, x_2, \ldots, x_\nu$ ενός δείγματος $X_1, X_2, \ldots X_\nu$ μπορούμε να τα υπολογίσουμε και επομένως μας είναι γνωστά, όμως σε μια άλλη πραγματοποίηση του δείγματος η τιμή τους μεταβάλλεται, δηλαδή, τα *στατιστικά* είναι τυχαίες μεταβλητές.

Στη συνέχεια, θα εξηγήσουμε πώς ορίζονται τα βασικά *στατιστικά* ενός δείγματος, πώς τα υπολογίζουμε για συγκεκριμένη πραγματοποίηση, $x_1, x_2, \ldots, x_\nu$, ενός τυχαίου δείγματος και κυρίως θα προσπαθήσουμε να εξηγήσουμε και να αναδείξουμε το νόημά τους, ώστε να μπορούμε να τα ερμηνεύουμε και να τα αξιοποιούμε σωστά.

Τα *στατιστικά* (αλλά και οι *παράμετροι* όπως είδαμε στο Α΄ Μέρος) ταξινομούνται σε τρεις βασικές κατηγορίες:

α) Μέτρα Θέσης/Κεντρικής Τάσης (location measures/ central tendency measures) τα οποία μας δίνουν πληροφορίες για τη θέση της κατανομής του δείγματος.
β) Μέτρα Μεταβλητότητας/Διασποράς (variability measures/dispersion measures) τα οποία μας δίνουν πληροφορίες για τη μεταβλητότητα των τιμών του δείγματος.
γ) Μέτρα Λοξότητας (skewness) και Κύρτωσης (kurtosis) τα οποία μας δίνουν πληροφορίες για τη μορφή της κατανομής του δείγματος.

9.1.3.1 Μέτρα θέσης/Κεντρικής τάσης

Για την περιγραφή της θέσης της κατανομής ενός δείγματος, θα παρουσιάσουμε τέσσερα *μέτρα θέσης*.

- Τον ***δειγματικό μέσο*** ή ***αριθμητικό μέσο*** ή ***μέσο όρο***
- Τη ***δειγματική κορυφή***
- Τη ***δειγματική διάμεσο***
- Τα ***p-ποσοστιαία σημεία του δείγματος.***

(α) Δειγματικός μέσος

Έστω $X_1, X_2, \ldots, X_\nu$ ένα τυχαίο δείγμα από έναν πληθυσμό (από την κατανομή μιας τυχαίας μεταβλητή X) και $x_1, x_2, \ldots, x_\nu$ μια πραγματοποίησή του. Έστω επίσης, $y_1, y_2, ..., y_k$ $(k \le \nu)$ οι k διαφορετικές μεταξύ τους τιμές από τις $x_1, x_2, \ldots, x_\nu$.

Ο ***δειγματικός μέσος (sample mean/arithmetic mean/average)*** $\overline{X}_\nu$ ή $\overline{X}$, όπως και κάθε *στατιστικό*, είναι τυχαία μεταβλητή και γι'αυτό συμβολίζεται με κεφαλαίο γράμμα όπως οι τυχαίες μεταβλητές. Ορίζεται από τον τύπο

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

Η συγκεκριμένη τιμή του $\overline{X}$, που υπολογίζεται για μια πραγματοποίηση $x_1, x_2, \ldots, x_\nu$ του τυχαίου δείγματος $X_1, X_2, \ldots, X_\nu$ συμβολίζεται με $\overline{x}$. Δηλαδή,

$$\overline{x} = \frac{1}{\nu}\sum_{i=1}^{\nu} x_i \ .$$

Θυμηθείτε ότι η *μέση τιμή* του *πληθυσμού* (δηλαδή, η *μέση τιμή* της κατανομής της τυχαίας μεταβλητής X) συμβολίζεται με μ ή με μ_X.

Από τον ορισμό του *δειγματικού μέσου* είναι φανερό ότι αν οι τιμές $x_1, x_2, \ldots, x_\nu$ είναι όλες μεταξύ τους ίσες, θα είναι ίσες και με τον *μέσο* τους. Φαίνεται δηλαδή, ότι με τον *δειγματικό μέσο* ενός δείγματος τιμών, $x_1, x_2, \ldots, x_\nu$, επιδιώκεται να ορισθεί ένας «τυπικός εκπρόσωπος» τους. Το γεγονός όμως, ότι στον υπολογισμό του συμμετέχει το άθροισμα όλων των τιμών, τον καθιστά ευαίσθητο σε *ακραίες-έκτροπες (outlying ή unusual)* τιμές[4]. Κατά συνέπεια, ο *δειγματικός μέσος* αποκρύπτει τις έκτροπες τιμές. Δηλαδή, όταν υπάρχουν έκτροπες τιμές, ο *δειγματικός μέσος* δίνει παραπλανητική εικόνα αν θεωρηθεί «τυπικός εκπρόσωπος» των τιμών του δείγματος. Βέβαια, αν πάρουμε τις διαφορές των $x_1, x_2, \ldots, x_\nu$, από τον *μέσο* τους, οι ακραίες τιμές αποκαλύπτονται.

Παράδειγμα 9.1.9: *Ο ιδιοκτήτης μιας μικρής επιχείρησης που απασχολεί πέντε εργαζομένους ισχυρίσθηκε σε δημοσιογράφο τοπικής εφημερίδας ότι οι εργαζόμενοι στην επιχείρησή του είναι πολύ καλά αμειβόμενοι αφού ο μέσος μισθός τους είναι 2000€. Ο «υποψιασμένος» δημοσιογράφος ερεύνησε λεπτομερέστερα το θέμα και βρήκε ότι οι μισθοί των εργαζομένων ήταν 400, 400, 500, 700 και 8000 € αντίστοιχα! Ο μισθός των 8000 € ήταν του manager και συνιδιοκτήτη! (Ως ένα άλλο αντίστοιχο παράδειγμα σκεφθείτε πόσο θα αλλάξει το μέσο ετήσιο εισόδημα των πελατών ενός συνοικιακού καφενείου αν ξαφνικά μπει στο καφενείο ως πελάτης και ο Bill Gates!!)*

■

Ο υπολογισμός του *δειγματικού μέσου* είναι πολύ απλός. Χρησιμοποιώντας τις *συχνότητες* ή τις *σχετικές συχνότητες* από τον *πίνακα συχνοτήτων* του δείγματος μπορεί εύκολα επίσης να υπολογισθεί από τον τύπο[5]

[4] *Στο πλαίσιο όμως της Θεωρίας Πιθανοτήτων και της Στατιστικής Συμπερασματολογίας, αυτό είναι το μεγάλο του πλεονέκτημα! (βλ. Κεντρικό Οριακό Θεώρημα)*

[5] *Η απόδειξή του είναι πολύ απλή.*

$$\bar{x} = \frac{\sum_{i=1}^{k} \nu_i y_i}{\nu} = \sum_{i=1}^{k} f_i y_i \ .$$

Ας δούμε ένα παράδειγμα.

Παράδειγμα 9.1.10 (συνέχεια του Παραδείγματος 9.2): *Ο μέσος του δείγματος από τη μεταβλητή «αριθμός παιδιών οικογένειας» του Παραδείγματος 9.2, υπολογίζεται εύκολα αν συμπληρώσουμε τον πίνακα συχνοτήτων με μια ακόμη στήλη όπως φαίνεται στον Πίνακα 9.1.6.*

y_i	ν_i	f_i	N_i	F_i	$\nu_i y_i$
0	2	0.1	2	0.1	0
1	4	0.2	6	0.3	4
2	10	0.5	16	0.8	20
3	2	0.1	18	0.9	6
4	2	0.1	20	1.0	8
Σύνολα	**20**	**1.00**			**38**

Πίνακας 9.1.6
Υπολογισμός του μέσου του δείγματος από τη μεταβλητή «αριθμός παιδιών οικογένειας» του Παραδείγματος 9.2

Δημιουργούμε μια στήλη με τα γινόμενα $\nu_i y_i$ και έτσι έχουμε

$$\bar{x} = \frac{\sum_{i=1}^{k} \nu_i y_i}{\nu} = \frac{38}{20} = 1.9 \text{ παιδιά.}$$

■

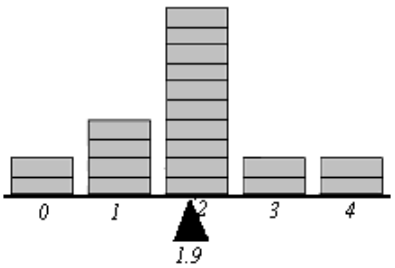
Ιδιότητες του δειγματικού μέσου

Ας δούμε κάποιες ενδιαφέρουσες ιδιότητες του *δειγματικού μέσου* (που παραπέμπουν σε αντίστοιχες ιδιότητες της *πληθυσμιακής μέσης τιμής* μ).

1. Το άθροισμα των αποκλίσεων των τιμών $x_1, x_2, \ldots, x_\nu$ από τον *δειγματικό μέσο* $\bar{x}$ είναι 0. Δηλαδή,

$$\sum_{i=1}^{\nu} (x_i - \bar{x}) = \sum_{i=1}^{k} (y_i - \bar{x}) \nu_i = 0$$

Επομένως, ο *δειγματικός μέσος* μπορεί να ερμηνευθεί και ως ***το σημείο ισορροπίας της κατανομής του δείγματος*** *(θυμηθείτε και την αντίστοιχη ιδιότητα της πληθυσμιακής μέσης τιμής που είδαμε στο Α΄ Μέρος)*. Έτσι, ο *δειγματικός μέσος* $\bar{x} = 1.9$ που υπολογίσαμε στο προηγούμενο παράδειγμα μπορεί να ερμηνευθεί ως εξής: αν στις θέσεις y_i ενός άξονα αμελητέου βάρους τοποθετήσουμε ως βάρη τις αντίστοιχες συχνότητες ν_i τότε ο άξονας θα έχει σημείο ισορροπίας τη θέση 1.9 *(δες Σχήμα 9.1.12)*.



Σχήμα 9.1.12
Ερμηνεία του δειγματικού μέσου ως

το σημείο ισορροπίας της κατανομής του δείγματος

Επίσης, αυτή η ιδιότητα μας λέει ότι αν από τις ν διαφορές $x_i - \bar{x}$ γνωρίζουμε τις $\nu - 1$, τότε μπορούμε να υπολογίσουμε και τη *ν-οστή*. Επομένως, μπορούμε να υπολογίσουμε το άθροισμα $\sum_{i=1}^{\nu}(x_i - \bar{x})^2$ αν γνωρίζουμε τους $\nu - 1$ από τους ν όρους του. Στη συνέχεια θα επανέλθουμε στη σημασία αυτής της ιδιότητας.

2. Το άθροισμα των τετραγώνων των αποκλίσεων των τιμών $x_1, x_2, \ldots, x_\nu$ από τον *δειγματικό μέσο* $\bar{x}$, είναι μικρότερο από το άθροισμα των τετραγώνων των αποκλίσεών τους από οποιαδήποτε άλλη τιμή λ. Δηλαδή,
$$\sum_{i=1}^{\nu}(x_i - \bar{x})^2 < \sum_{i=1}^{\nu}(x_i - \lambda)^2, \quad \forall \lambda .$$
Έτσι, το άθροισμα $\sum_{i=1}^{\nu}(x_i - \lambda)^2$ γίνεται ***ελάχιστο*** αν και μόνο αν $\lambda = \bar{x}$.
3. Αν $t_i = x_i + \beta$ τότε $\bar{t} = \bar{x} + \beta$. Δηλαδή, αν στις τιμές $x_1, x_2, \ldots, x_\nu$ του δείγματος προσθέσουμε μια σταθερή ποσότητα β (θετική ή αρνητική), τότε η ο *δειγματικός μέσος* αυξάνεται (ή μειώνεται) κατά την ίδια ποσότητα.
4. Αν $t_i = \alpha x_i$ τότε $\bar{t} = \alpha\bar{x}$. Δηλαδή, αν οι τιμές $x_1, x_2, \ldots, x_\nu$ του δείγματος πολλαπλασιασθούν με την ίδια ποσότητα α, τότε ο *δειγματικός μέσος* πολλαπλασιάζεται με την ίδια ποσότητα.
5. Αν $t_i = \alpha x_i + \beta$ τότε $\bar{t} = \alpha\bar{x} + \beta$

Επίσης, ο *δειγματικός μέσος* έχει τα ακόλουθα πλεονεκτήματα και μειονεκτήματα.

Πλεονεκτήματα	**Μειονεκτήματα**
• *Για τον υπολογισμό του χρησιμοποιούνται όλες οι τιμές.* • *Είναι μοναδικός.* • *Ο υπολογισμός του είναι απλός.* • *Αξιοποιείται στη στατιστική συμπερασματολογία.*	• *Επηρεάζεται από ακραίες τιμές.* • *Μπορεί να μην αντιστοιχεί σε δυνατή τιμή της μεταβλητής.* • *Δεν υπολογίζεται για ποιοτικά δεδομένα.*

Σταθμικός μέσος

Στις περιπτώσεις που τα δεδομένα $x_1, x_2, \ldots, x_\nu$, έχουν διαφορετική αξία/βάρος $w_1, w_2, \ldots, w_\nu$ αντίστοιχα, υπολογίζεται ο ***σταθμικός μέσος (weighted mean)*** ο οποίος ορίζεται με τον τύπο
$$\bar{x}_w = \frac{\sum_{i=1}^{\nu} w_i x_i}{\sum_{i=1}^{\nu} w_i} .$$
Σημειώνουμε ότι ο *σταθμικός μέσος* διατηρεί τις ιδιότητες του μη σταθμισμένου μέσου. Ας δούμε ένα παράδειγμα.

Παράδειγμα 9.1.11: *Ένας οδηγός φορτηγού διανομής τροφίμων, αγόρασε σε μια ημέρα πετρέλαιο από τρία διαφορετικά πρατήρια. Από το πρώτο αγόρασε 6 λίτρα προς 0.75 € το λίτρο, από το δεύτερο 12 λίτρα προς 0.84 € το λίτρο και από το τρίτο 5 λίτρα προς 0.76 € το λίτρο.*

Προφανώς, για να υπολογισθεί το μέσο ποσό που πλήρωσε ανά λίτρο ο οδηγός πρέπει να χρησιμοποιηθεί ο *σταθμικός μέσος* (θυμηθείτε και το εισαγωγικό για την έννοια της μέσης τιμής παράδειγμα της παραγράφου 5.3.1). Έτσι έχουμε

$$\bar{x}_w = \frac{\sum_{i=1}^{\nu} w_i x_i}{\sum_{i=1}^{\nu} w_i} = \frac{6 \cdot 0.75 + 12 \cdot 0.84 + 5 \cdot 0.76}{6+12+5} = 0.799\,€ \text{ ανά λίτρο.}$$

Παρατήρηση 9.1.6 (μέσος k μέσων): *Ο μέσος των μέσων k δειγμάτων μεγέθους* $\nu_1, \nu_2, \ldots, \nu_k$ *αντίστοιχα, προφανώς είναι*

$$\bar{x} = \frac{\sum_{i=1}^{k} \nu_i \bar{x}_i}{\sum_{i=1}^{k} \nu_i}.$$

Ουσιαστικά πρόκειται για σταθμικό μέσο με βάρη $\nu_1, \nu_2, \ldots, \nu_k$ *αντίστοιχα. Ας δούμε ένα παράδειγμα.*

■

Παράδειγμα 9.1.12: *Αν το μέσο ύψος 10 φοιτητών είναι 170 cm και το μέσο ύψος 5 φοιτητριών είναι 160 cm τότε το μέσο ύψος φοιτητών και φοιτητριών είναι*

$$\bar{x} = \frac{\sum_{i=1}^{2} \nu_i \bar{x}_i}{\sum_{i=1}^{2} \nu_i} = \frac{10 \cdot 170 + 5 \cdot 160}{15} = 166.7\,cm.$$

Ερώτηση: *Στην έκδοση της αμερικανικής κυβέρνησης "Science Indicators" του 1980, αναφέρεται ότι ο μέσος μισθός των γυναικών σε όλους τους επιστημονικούς τομείς είναι μόνο το 77% του μέσου μισθού των ανδρών επιστημόνων. Στην ίδια πηγή όμως, αναφέρεται ότι σε κάθε επιστημονικό τομέα ξεχωριστά, ο μέσος μισθός των γυναικών είναι τουλάχιστον το 92% του μέσου μισθού των ανδρών. Εξηγήστε αυτή τη φαινομενική αντίφαση.*

Απάντηση: Οι γυναίκες είναι συγκεντρωμένες στους τομείς που αμείβονται λιγότερο. Έτσι, για τις γυναίκες, ο μέσος μισθός συνολικά θα είναι χαμηλότερος των ανδρών ακόμη και αν κερδίζουν το ίδιο ποσό με τους άνδρες σε κάθε τομέα ξεχωριστά.

■

Παρατηρήσεις 9.1.7:

α) Αν από τον υπολογισμό του δειγματικού μέσου θέλουμε να παραλείψουμε τις ακραίες τιμές, μπορούμε να δημιουργήσουμε έναν ***ισοσταθμισμένο μέσο (trimmed mean)*** *θέτοντας στον σταθμικό μέσο, βάρος 0 για τις ακραίες τιμές που θέλουμε να παραληφθούν και βάρος 1 για όλες τις υπόλοιπες.*

β) Παρότι ο δειγματικός μέσος ως μέτρο θέσης-κεντρικής τάσης δεν είναι πάντα το καταλληλότερο για την περιγραφή της θέσης της κατανομής των δεδομένων (μάλιστα, μπορεί και να παραπλανήσει), εντούτοις, έχει μεγάλη σημασία και χρησιμοποιείται ευρέως στη στατιστική συμπερασματολογία. Ένας από τους λόγους που συμβαίνει αυτό είναι το γεγονός ότι ελαχιστοποιεί το άθροισμα $\sum_{i=1}^{\nu} (x_i - \lambda)^2$ *. Αυτή η ιδιότητα του δειγματικού μέσου είναι «πολύ καλή» μαθηματική ιδιότητα*[6] *και γι΄ αυτό έχει επηρεάσει τον ορισμό και άλλων στατιστικών μέτρων. Στη συνέχεια θα αναφερθούμε και σε άλλους λόγους που δικαιολογούν τη μεγάλη χρησιμότητα του δειγματικού μέσου στη στατιστική συμπερασματολογία.*

[6] *Ικανοποιεί το* ***κριτήριο των ελαχίστων τετραγώνων.***

(β) Δειγματική Κορυφή ή Επικρατούσα τιμή

Η ***κορυφή (mode)*** της κατανομής του δείγματος είναι η τιμή του δείγματος με τη *μεγαλύτερη συχνότητα.* Συμβολίζεται με M_0 και έχει τα ακόλουθα πλεονεκτήματα και μειονεκτήματα.

Πλεονεκτήματα	**Μειονεκτήματα**
• *Υπολογίζεται εύκολα.* • *Είναι εύκολα κατανοητή.* • *Υπολογίζεται και από ελλιπή δεδομένα.* • *Δεν επηρεάζεται από ακραίες τιμές.* • *Υπολογίζεται και για ποιοτικά δεδομένα.*	• *Δε χρησιμοποιούνται όλες οι τιμές για τον υπολογισμό της.* • *Στη στατιστική συμπερασματολογία έχει περιορισμένη σημασία.* • *Δεν ορίζεται πάντα μονοσήμαντα και επίσης μπορεί να μην υπάρχει.*

Εύκολα αποδεικνύεται ότι αν $t_i = \alpha x_i + \beta$, δηλαδή αν γίνει γραμμικός μετασχηματισμός των $x_1, x_2, \ldots, x_\nu$ τότε και η κορυφή τους, έστω M_{0x}, μετασχηματίζεται αντίστοιχα, δηλαδή, για την κορυφή M_{0t} των $t_1, t_2, \ldots, t_\nu$ έχουμε

$$M_{0t} = \alpha M_{0x} + \beta .$$

Για τον υπολογισμό της σε ομαδοποιημένες παρατηρήσεις, μπορεί να χρησιμοποιηθεί ο τύπος

$$M_0 = L_i + \frac{\Delta_1}{\Delta_1 + \Delta_2} c_i$$

όπου, L_i είναι το κάτω άκρο της *επικρατούσας κλάσης,* δηλαδή, της κλάσης με τη μεγαλύτερη συχνότητα, c_i είναι το πλάτος της *επικρατούσας κλάσης*, $\Delta_1 = \nu_i - \nu_{i-1}$ είναι η διαφορά που προκύπτει αν από τη συχνότητα της *επικρατούσας κλάσης* αφαιρέσουμε τη συχνότητα της προηγούμενης (από την επικρατούσα) κλάσης και αντίστοιχα, $\Delta_2 = \nu_i - \nu_{i+1}$ είναι η διαφορά που προκύπτει αν από τη συχνότητα της *επικρατούσας κλάσης* αφαιρέσουμε τη συχνότητα της επόμενης (από την επικρατούσα) κλάσης.

(γ) Δειγματική διάμεσος

Η ***διάμεσος (median)*** της κατανομής του δείγματος είναι ένας αριθμός για τον οποίο ισχύει ότι το πολύ 50% των τιμών του δείγματος (των παρατηρήσεων) είναι μικρότερες από αυτόν και επίσης το πολύ 50% των τιμών του δείγματος είναι μεγαλύτερες από αυτόν. Εκφράζει την *κεντρική θέση* της κατανομής του δείγματος και γι΄ αυτό στη βιβλιογραφία συναντάται και ως ***μέσος θέσης (position average).*** Συνήθως συμβολίζεται με δ.

Αν το μέγεθος του δείγματος ν, είναι αριθμός περιττός, τότε προφανώς

$$\delta = x_{(\frac{\nu+1}{2})}$$

ενώ αν είναι άρτιος

$$\delta = \frac{x_{(\frac{\nu}{2})} + x_{(\frac{\nu}{2}+1)}}{2}$$

(με $x_{(\nu)}$ συμβολίζουμε τη *ν-οστή* τιμή του δείγματος, σε αύξουσα διάταξή τους).

Δηλαδή, τη θέση της *διαμέσου* δίνει ο αριθμός $0.5(\nu+1)$ εφόσον είναι ακέραιος, ενώ αν δεν είναι ακέραιος, τότε η *διάμεσος* λαμβάνεται ίση (εκτιμάται) με το ημιάθροισμα

των δύο τιμών που οι θέσεις τους είναι οι πλησιέστερες στον αριθμό $0.5(\nu+1)$. Ας δούμε δύο παραδείγματα.

1) Έστω οι παρατηρήσεις 5, 2, 9, 6, 11. Τις διατάσσουμε σε αύξουσα σειρά

2, 5, 6, 9, 11.

Η *διάμεσος* τιμή είναι αυτή που βρίσκεται στη θέση $0.5(5+1)=3$, δηλαδή

$$\delta = 6.$$

2) Έστω οι παρατηρήσεις 2, 5, 6, 27, 11, 9. Τις διατάσσουμε σε αύξουσα σειρά

2, 5, 6, 9, 11, 27.

Επειδή ο αριθμός $0.5(6+1)=3.5$ δεν είναι ακέραιος, η *διάμεσος* τιμή είναι το ημιάθροισμα της $3^{ης}$ και της $4^{ης}$ παρατήρησης, δηλαδή

$$\delta = (6+9)/2 = 7.5.$$

Σε ομαδοποιημένες παρατηρήσεις, για τον υπολογισμό της *διαμέσου* χρησιμοποιείται το *πολύγωνο αθροιστικών σχετικών συχνοτήτων* ή ο τύπος

$$\delta = L_i + \frac{0.5\nu - N_{i-1}}{\nu_i} c_i,$$

όπου, L_i είναι το κάτω άκρο της *μεσαίας κλάσης*, δηλαδή, της κλάσης στην οποία ανήκει η *διάμεσος*, c_i είναι το πλάτος της *μεσαίας κλάσης*, ν_i είναι η *συχνότητα* της *μεσαίας κλάσης* και N_{i-1} είναι η *αθροιστική συχνότητα* της προηγούμενης από τη *μεσαία* κλάσης.

Παραδείγματα υπολογισμού της *διαμέσου* ομαδοποιημένων παρατηρήσεων δίνουμε αφού ολοκληρώσουμε κάποια ακόμη στοιχεία θεωρίας.

Για τη *διάμεσο* ισχύει ότι

$$\sum_{i=1}^{\nu} |x_i - \delta| < \sum_{i=1}^{\nu} |x_i - \lambda|, \quad \forall \lambda.$$

Δηλαδή, το άθροισμα των απόλυτων αποκλίσεων των τιμών $x_1, x_2, \ldots, x_\nu$ από τη *διάμεσό* τους δ, είναι μικρότερο από το άθροισμα των απολύτων αποκλίσεών τους από οποιαδήποτε άλλη τιμή λ. Ή αλλιώς, το άθροισμα $\sum_{i=1}^{\nu} |x_i - \lambda|$ γίνεται ***ελάχιστο*** αν και μόνο αν $\lambda = \delta$.

Εύκολα επίσης αποδεικνύεται ότι αν $t_i = \alpha x_i + \beta$, δηλαδή αν γίνει γραμμικός μετασχηματισμός των $x_1, x_2, \ldots, x_\nu$ τότε και η διάμεσός τους, έστω δ_x, μετασχηματίζεται αντίστοιχα, δηλαδή, για τη διάμεσο δ_t των $t_1, t_2, \ldots, t_\nu$ έχουμε

$$\delta_t = \alpha \delta_x + \beta.$$

Αναφέρουμε τέλος, ότι η διάμεσος έχει τα ακόλουθα πλεονεκτήματα και μειονεκτήματα.

Πλεονεκτήματα	**Μειονεκτήματα**
• *Είναι εύκολα κατανοητή.*	• *Δε χρησιμοποιούνται όλες οι τιμές για τον υπολογισμό της.*
• *Δεν επηρεάζεται από ακραίες τιμές.*	• *Δεν υπολογίζεται για ποιοτικά δεδομένα.*
• *Ο υπολογισμός της είναι απλός.*	• *Για τον υπολογισμό της μπορεί να χρειαστεί παρεμβολή.*
• *Είναι μοναδική.*	

Παρατήρηση 9.1.8: *Επειδή η διάμεσος δεν επηρεάζεται όπως ο μέσος από ακραίες τιμές, για την περιγραφή παρατηρήσεων που εμφανίζουν ακραίες τιμές προτιμάται ως μέτρο θέσης από τον μέσο. Εξηγείται, έτσι, γιατί ο ΟΗΕ ως δείκτη γήρανσης (μεταξύ άλλων) χρησιμοποιεί τη διάμεσο και όχι τον μέσο. Έτσι, μπορούμε, επίσης, να εξηγήσουμε γιατί στις διαπραγματεύσεις των συνδικαλιστών με τους εργοδότες για το ύψος των αποδοχών, συνήθως, οι συνδικαλιστές χρησιμοποιούν τη διάμεσο των αποδοχών ενώ οι εργοδότες τον μέσο.*

■

(δ) p-ποσοστιαία σημεία του δείγματος

Τα ***p-ποσοστιαία σημεία (quantiles)*** της κατανομής του δείγματος συμβολίζονται με $x_p, 0 < p < 1$. Αποτελούν γενίκευση της έννοιας της *διαμέσου* και βοηθούν στην πληρέστερη περιγραφή της θέσης της κατανομής του δείγματος (αλλά και της μεταβλητότητας των τιμών του και της μορφής της κατανομής του, όπως θα δούμε στη συνέχεια).

Το ***p-ποσοστιαίο σημείο*** x_p είναι ένας αριθμός για τον οποίο ισχύει ότι το πολύ $100p\%$ των τιμών του δείγματος είναι μικρότερες από αυτόν και το πολύ $100(1-p)\%$ των τιμών του δείγματος είναι μεγαλύτερες από αυτόν. Τα *p-ποσοστίαια* διακρίνονται σε

εκατοστημόρια (percentiles) $x_{0.01}, x_{0.02}, \ldots, x_{0.99}$

δεκατημόρια (deciles) $x_{0.1}, x_{0.2}, \ldots, x_{0.9}$

τεταρτημόρια (quartiles) $x_{0.25} = Q_1,\ x_{0.5} = Q_2 = \delta,\ x_{0.75} = Q_3$.

Τα *p-ποσοστιαία σημεία* σε ομαδοποιημένες παρατηρήσεις μπορούν να υπολογισθούν από το *πολύγωνο αθροιστικών σχετικών συχνοτήτων* ή από τον τύπο

$$x_p = L_i + \frac{p\nu - N_{i-1}}{\nu_i} c_i$$

όπου, L_i είναι το κάτω άκρο της κλάσης στην οποία βρίσκεται το x_p, c_i είναι το πλάτος της, ν_i είναι η *συχνότητά* της και N_{i-1} είναι η *αθροιστική συχνότητα* της προηγούμενης κλάσης. Το x_p βρίσκεται στην κλάση που βρίσκεται η τιμή με *αθροιστική σχετική συχνότητα* p.

Παράδειγμα 9.1.13 (συνέχεια του Παραδείγμτος 9.2): *Θα υπολογίσουμε τη διάμεσο και το 0.95-ποσοστιαίο σημείο της κατανομής του δείγματος από τη μεταβλητή «μηνιαίο οικογενειακό εισόδημα» του Παραδείγματος 9.2.*

Εισόδημα	ν_i	f_i	N_i	F_i
[900 1100)	1	0.05	1	0.05
[1100 1300)	4	0.20	5	0.25
[1300 1500)	6	0.30	11	0.55
[1500 1700)	4	0.20	15	0.75
[1700 1900)	3	0.15	18	0.90
[1900 2100)	2	0.10	20	1.00
Σύνολα	**20**	**1.00**		

Πίνακας 9.1.7
Ο πίνακας συχνοτήτων των τιμών του δείγματος από τη μεταβλητή «μηνιαίο οικογενειακό εισόδημα» του Παραδείγματος 9.2 ομαδοποιημένων σε 6 κλάσεις

Για να υπολογίσουμε τη *διάμεσο,* $\delta = Q_2 = x_{0.5}$, των παρατηρήσεων εργαζόμαστε ως εξής: η *διάμεσος* βρίσκεται στην κλάση [1300, 1500) γιατί όπως φαίνεται στη στήλη

των *αθροιστικών σχετικών συχνοτήτων* του *πίνακα συχνοτήτων*, σε αυτή την κλάση βρίσκεται η τιμή με *αθροιστική σχετική συχνότητα* 0.5. Επομένως

$$\delta = x_{0.5} = L_i + \frac{0.5\nu - N_{i-1}}{\nu_i} c_i = 1300 + \frac{0.5 \cdot 20 - 5}{6} \cdot 200 = 1466.7 .$$

Το *0.95-ποσοστιαίο σημείο* $x_{0.95}$ υπολογίζεται ανάλογα. Προφανώς, το $x_{0.95}$ ανήκει στην κλάση [1900, 2100), γιατί σε αυτή την κλάση βρίσκεται η τιμή με *αθροιστική σχετική συχνότητα* 0.95, και επομένως

$$x_{0.95} = L_i + \frac{0.95\nu - N_{i-1}}{\nu_i} c_i = 1900 + \frac{0.95 \cdot 20 - 18}{2} \cdot 200 = 2000 .$$

■

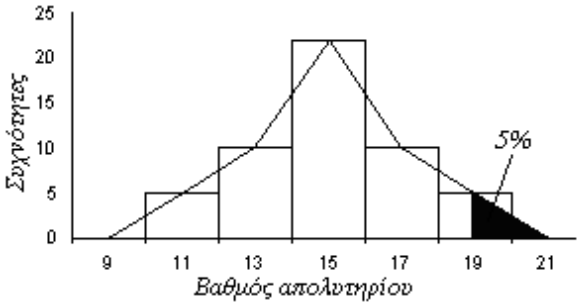
Παράδειγμα 9.1.14: *Στον Πίνακα 9.1.8 δίνεται η κατανομή των βαθμών απολυτηρίου 50 μαθητών Λυκείου. Αν στο 5% των μαθητών με την υψηλότερη βαθμολογία δοθεί υποτροφία, τι βαθμό απολυτηρίου πρέπει να έχει ένας μαθητής για να πάρει υποτροφία;*

Βαθμοί	ν_i	N_i	F_i
[10 12)	5	5	0.1
[12 14)	10	15	0.3
[14 16)	20	35	0.7
[16 18)	10	45	0.9
[18 20)	5	50	1.0

Πίνακας 9.1.8
Η κατανομή των βαθμών απολυτηρίου
50 μαθητών Λυκείου

<u>*Απάντηση:*</u> Προφανώς *(δες και το Σχήμα 9.1.13)* ζητούμενο είναι το *0.95-ποσοστιαίο σημείο*

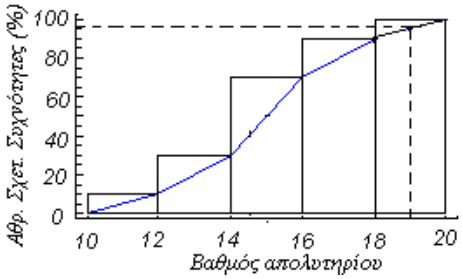
$$x_{0.95} = L_i + \frac{0.95\nu - N_{i-1}}{\nu_i} c_i = 18 + \frac{0.95 \cdot 50 - 45}{5} \cdot 2 = 19 .$$



Σχήμα 9.1.13
Δεξιά του 0.95-ποσοστιαίου σημείου βρίσκεται
5% της κατανομής

■

***Σημείωση 9.1.2*:** Τα *p-ποσοστιαία σημεία, όπως έχουμε αναφέρει, μπορούν να υπολογισθούν και γραφικά από το πολύγωνο αθροιστικών σχετικών συχνοτήτων. Δείτε στο Σχήμα 9.1.14 πώς προκύπτει γραφικά ότι για την κατανομή των βαθμών απολυτηρίου του προηγούμενου ππαραδείγματος είναι* $x_{0.95} = 19$.



Σχήμα 9.1.14
Υπολογισμός του 0.95-ποσοστιαίου σημείου
από το πολύγωνο αθροιστικών σχετικών συχνοτήτων

Αν τα δεδομένα δεν είναι ομαδοποιημένα, τα *p-ποσοστιαία σημεία* υπολογίζονται με τρόπο ανάλογο του τρόπου υπολογισμού της *διαμέσου* για μη ομαδοποιημένα δεδομένα. Δηλαδή, σε αύξουσα διάταξη των παρατηρήσεων (των τιμών του δείγματος) $x_1, x_2, \ldots, x_\nu$, τη θέση του *p-ποσοστιαίου σημείου* δίνει ο αριθμός $p(\nu+1)$ εφόσον είναι ακέραιος, ενώ αν δεν είναι ακέραιος, τότε το *p-ποσοστιαίο σημείο* εκτιμάται με παρεμβολή μεταξύ των δύο τιμών που οι θέσεις τους είναι οι πλησιέστερες στον αριθμό $p(\nu+1)$. Ας δούμε ένα παράδειγμα.

Παράδειγμα 9.1.15: *Θα υπολογίσουμε τα τεταρτημόρια της κατανομής*
α) του δείγματος, 6, 1, 5, 9, 6, 8, 1, 9, 2
β) του δείγματος, 15, 11, 11, 11, 22, 9, 11, 7, 11, 12, 12, 16, 8, 11, 15, 9, 10, 14, 9, 10, 11, 10, 6, 17, 11, 10, 8, 11

α) Διατάσσουμε τις παρατηρήσεις σε αύξουσα σειρά:

1, 1, 2, 5, 6, 6, 8, 9, 9.

Η θέση του Q_1 είναι $0.25(9+1)=2.5$ και επειδή, ο αριθμός αυτός δεν είναι ακέραιος, βρίσκουμε το Q_1 με παρεμβολή μεταξύ της $2^{ης}$ και της $3^{ης}$ θέσης ως εξής: $Q_1 = 1 + 0.5\cdot(2-1) = 1.5$. Αντίστοιχα, η θέση του Q_3 είναι $0.75\cdot(9+1)=7.5$ και επειδή ο αριθμός αυτός δεν είναι ακέραιος, βρίσκουμε το Q_3 με παρεμβολή μεταξύ της $7^{ης}$ και της $8^{ης}$ θέσης ως εξής: $Q_3 = 8 + 0.5\cdot(9-8) = 8.5$.

β) Για διευκόλυνσή μας, κατασκευάζουμε τον *πίνακα συχνοτήτων* του δείγματος.

y_i	ν_i	N_i	f_i	F_i
6	1	1	0.0357	0.0357
7	1	2	0.0357	0.0714
8	2	4	0.0714	0.1428
9	3	7	0.1071	0.2500
10	4	11	0.1428	0.3928
11	9	20	0.3214	0.7143
12	2	22	0.0714	0.7857
14	1	23	0.0357	0.8214
15	2	25	0.0714	0.8928
16	1	26	0.0357	0.9286
17	1	27	0.0357	0.9643
22	1	28	0.0357	1.000
Σύνολα	**28**		**1.000**	

Πίνακας 9.1.9
Ο πίνακας συχνοτήτων του δείγματος
του παραδείγματος 9.1.15β

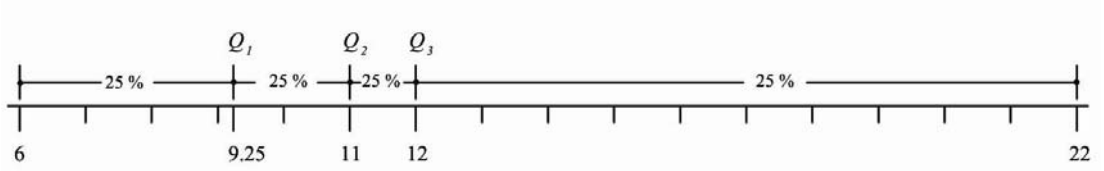
Η θέση του Q_1 είναι $0.25(28+1)=7.25$ και επειδή ο αριθμός αυτός δεν είναι ακέραιος, βρίσκουμε το Q_1 με παρεμβολή μεταξύ της $7^{ης}$ και της $8^{ης}$ θέσης ως εξής: $Q_1 = 9 + 0.25 \cdot (10-9) = 9.25$. Αντίστοιχα, η θέση του Q_3 είναι η $0.75 \cdot (28+1) = 21.75$ και επειδή ο αριθμός αυτός δεν είναι ακέραιος, βρίσκουμε το Q_3 με παρεμβολή μεταξύ της $21^{ης}$ και της $22^{ης}$ θέσης ως εξής: $Q_3 = 12 + 0.75 \cdot (12-12) = 12$. Ομοίως βρίσκουμε ότι $Q_2 = \delta = 11$.

■

Επισημαίνουμε ότι στην περίπτωση που ο αριθμός που δίνει τη θέση του *τεταρτημορίου* δεν είναι ακέραιος, το πρόβλημα αντιμετωπίζεται και με άλλους τρόπους όπως, η συνήθης στρογγυλοποίηση ή το *πρώτο τεταρτημόριο* υπολογίζεται ως *διάμεσος* του πρώτου μισού του συνόλου των παρατηρήσεων και το τρίτο ως η *διάμεσος* του δεύτερου μισού του συνόλου των παρατηρήσεων. Όμως δε θα επεκταθούμε περισσότερο σε θέματα υπολογισμού των *p-ποσοστιαίων σημείων*. Θα επικεντρωθούμε στην ερμηνεία και τη χρησιμότητά τους.

Παρατηρήσεις 9.1.9:

α) Παρατηρείστε ότι τα τεταρτημόρια υποδιαιρούν την κατανομή των παρατηρήσεων σε τέσσερα «ίσα» τμήματα. Όμως προσοχή. Όχι με όρους απόστασης, αλλά με όρους ποσοστών. Δηλαδή, τα τμήματα αυτά είναι «ίσα» με την έννοια ότι περιέχουν ίσα ποσοστά παρατηρήσεων. Έτσι, ίσες αποστάσεις μπορεί να περιέχουν διαφορετικά ποσοστά παρατηρήσεων και αντίστροφα, άνισες αποστάσεις μπορεί να περιέχουν ίδια ποσοστά παρατηρήσεων. Τα τεταρτημόρια (γενικότερα, τα p-ποσοστιαία σημεία) είναι μέτρα σχετικής θέσης και όχι σχετικής απόστασης. Παρατηρείστε το Σχήμα 9.1.15 όπου σημειώνονται τα τεταρτημόρια του δείγματος του Παραδείγματος 9.1.15β. Μεταξύ των άνισων αποστάσεων, 6 έως 9.25, 9.25 έως 11, 11 έως 12 και 12 έως 22, βρίσκονται ίσα ποσοστά παρατηρήσεων (25%).



Σχήμα 9.1.15
Άνισες αποστάσεις περιέχουν ίσα ποσοστά παρατηρήσεων (25%)

<u>Ερώτηση</u>: Αν σε ένα σύνολο παρατηρήσεων η μικρότερη τιμή είναι 20 και η μεγαλύτερη 80, γιατί η διάμεσος δεν είναι, κατ΄ ανάγκη, $50 = (20+80)/2$ *;*

β) Τα p-ποσοστιαία σημεία είναι μέτρα θέσης ιδιαιτέρως χρήσιμα στη μελέτη οικονομικών, κοινωνικών, δημογραφικών κ.ά. φαινομένων γιατί, μεταξύ άλλων, μας επιτρέπουν να απαντήσουμε σε ερωτήσεις που αφορούν ***συγκεκριμένες*** *παρατηρήσεις. Για παράδειγμα, μια συγκεκριμένη παρατήρηση, βρίσκεται κοντά στα άκρα ή κοντά στο κέντρο της κατανομής; ή πόσες παρατηρήσεις είναι μικρότερες από μια συγκεκριμένη παρατήρηση; Έτσι, αν σε μια κατανομή βαθμολογίας φοιτητών, είναι* $x_{0.95} = 7.5$ *τότε, για έναν φοιτητή που έχει βαθμό π.χ. 8 μπορούμε να συμπεράνουμε ότι ανήκει στο 5% των φοιτητών με τη μεγαλύτερη βαθμολογία.*

γ) Τα p-ποσοστιαία σημεία, όπως θα δούμε και στη συνέχεια, είναι χρήσιμα και για την περιγραφή της μορφής της κατανομής των παρατηρήσεων. Αν τα ποσοστιαία σημεία x_p *και* x_{1-p}*, δηλαδή, τα* $x_{0.2}$ *και* $x_{0.8}$*, τα* $x_{0.3}$ *και* $x_{0.7}$*, τα* $x_{0.25}$ *και* $x_{0.75}$ *κ.ο.κ, βρίσκονται σε ίση απόσταση από το κέντρο της κατανομής (τη διάμεσο), τότε η κατανομή είναι συμμετρική.*

δ) Τα p-ποσοστιαία σημεία μπορούν να βοηθήσουν και στην αντιμετώπιση κάποιων πρακτικών προβλημάτων που μπορεί να παρουσιασθούν σε μια έρευνα. Για παράδειγμα, αν ένας ερευνητής θέλει να υπολογίσει το χρόνο ζωής μιας ομάδας πειραματόζωων, πρέπει να περιμένει να πεθάνει και το τελευταίο πειραματόζωο προκειμένου να υπολογίσει το μέσο χρόνο ζωής τους. Για να υπολογίσει όμως τη διάμεσο του χρόνου ζωής ή κάποιο άλλο p-ποσοστιαίο σημείο, δεν απαιτείται να περιμένει μέχρι να πεθάνουν όλα και έτσι κερδίζει χρόνο που μπορεί να είναι σημαντικός για την εξέλιξη της έρευνας.

■

Σύγκριση δειγματικού μέσου, δειγματικής κορυφής και δειγματικής διαμέσου

Αν συγκρίνουμε αυτά τα τρία μέτρα θέσης με *μαθηματικούς όρους,* τότε εύκολα μπορούμε να αποφανθούμε για «το καλύτερο». Δηλαδή, αν για παράδειγμα, θέσουμε ως κριτήριο την ελαχιστοποίηση του αθροίσματος $\sum_{i=1}^{\nu}(x_i-\lambda)^2$ τότε το καλύτερο είναι ο *δειγματικός μέσος* ενώ αν θέσουμε ως κριτήριο την ελαχιστοποίηση του αθροίσματος $\sum_{i=1}^{\nu}|x_i-\lambda|$ τότε το καλύτερο είναι η *διάμεσος*. Αν τα συγκρίνουμε με κριτήριο την καταλληλότητα περιγραφής της *θέσης* της κατανομής, τότε, φαίνεται να υπερέχει η *διάμεσος*. **Όμως, για την περιγραφή της θέσης της κατανομής του δείγματος κάθε μέτρο *θέσης* έχει την ιδιαίτερη αξία του και επομένως πρέπει όλα να μπορούμε να τα ερμηνεύουμε σωστά ώστε αφενός, να τα χρησιμοποιούμε σωστά και αφετέρου, να μην πέφτουμε θύματα πλάνης επιτηδείων ή ημιμαθών.**

Παράδειγμα 9.1.16: *Στον Πίνακα 9.1.10 φαίνονται οι τιμές του ύψους της βροχής (σε mm) στην Αθήνα για τις ημέρες από 1-12-61 έως 31-12-61.*

0	0	0	0	0	0	1.2	28.6	1.2	0	0
0	0	0	0	0	1.1	2.9	1,5	0.4	0	2.8
0	0	0	0	1.2	3	0.1	0	8.5		

Πίνακας 9.1.10
Το ύψος της βροχής (σε mm) στην Αθήνα
από 1-12-61 έως 31-12-61

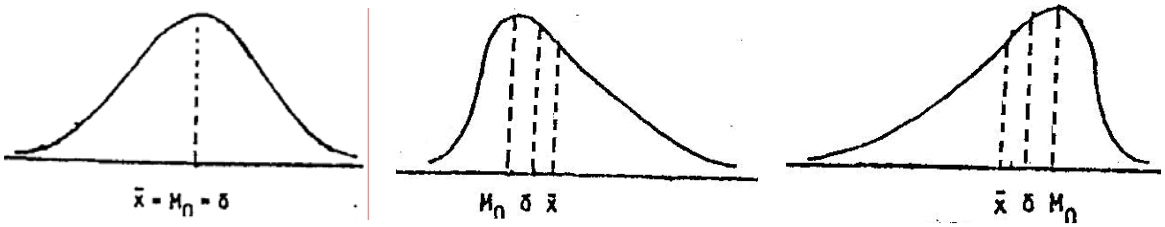
Ο *δειγματικός μέσος* (το μέσο ύψος της βροχής στην Αθήνα από 1-12-61 έως 31-12-61) είναι $\bar{x}=1.7$. Βέβαια, εύκολα διαπιστώνεται, ακόμη και με μια πρόχειρη ματιά, ότι ο *δειγματικός μέσος* δίνει ελάχιστη πληροφορία για τη θέση της κατανομής των υψών της βροχής. Τα *τεταρτημόρια* $Q_1=0,\ Q_2=\delta=0,\ Q_3=1.2$ δίνουν ακριβέστερη εικόνα για τη θέση της κατανομής που είναι η μεγάλη συγκέντρωση τιμών στο 0.

■

Σχετική θέση δειγματικού μέσου, δειγματικής κορυφής και δειγματικής διαμέσου

Για τη σχετική θέση του *δειγματικού μέσου*, της *κορυφής* και της *διαμέσου*, ισχύει, εν γένει, ο ακόλουθος κανόνας *(δείτε και τα Σχήματα 9.1.16)*.

- Όταν η *καμπύλη συχνοτήτων* της κατανομής του δείγματος είναι *συμμετρική* ισχύει
$$\bar{x}=\delta=M_0.$$
- Όταν η *καμπύλη συχνοτήτων* της κατανομής του δείγματος παρουσιάζει *θετική ασυμμετρία* ισχύει
$$\bar{x}>\delta>M_0.$$
- Όταν η *καμπύλη συχνοτήτων* της κατανομής του δείγματος παρουσιάζει *αρνητική ασυμμετρία* ισχύει
$$\bar{x}<\delta<M_0.$$



(α) *Συμμετρική κατανομή*

(β) *Λοξή κατανομή με θετική ασυμμετρία*

(γ) *Λοξή κατανομή με αρνητική ασυμμετρία*

Σχήμα 9.1.16

Σχετική θέση μέσου, κορυφής και διαμέσου

Επισημαίνουμε ότι υπάρχουν περιπτώσεις κατανομών που αυτός ο κανόνας δεν ισχύει. Εξαιρέσεις αυτού του κανόνα μπορεί, για παράδειγμα, να παρουσιασθούν σε κατανομές με πολύ «μακριά» ουρά, π.χ. προς τα αριστερά, αλλά με πολύ «παχιά» ουρά προς τα δεξιά. Σε τέτοιες περιπτώσεις ενώ η κατανομή παρουσιάζει εμφανώς *αρνητική ασυμμετρία*, εντούτοις ο *μέσος* μπορεί να είναι μεγαλύτερος (να βρίσκεται δεξιά) της *διαμέσου*. Ένα τέτοιο παράδειγμα κατανομής δίνεται στην *Άσκηση 9.11*. Επίσης, εξαιρέσεις του κανόνα μπορούν να εμφανισθούν σε κατανομές που δεν είναι *μονοκόρυφες* καθώς και σε περιπτώσεις *διακριτών* μεταβλητών. Ένα τέτοιο παράδειγμα *διακριτής* μεταβλητής, δίνεται στην *Άσκηση 9.17*.

Ερώτηση: *Έστω ότι η κατανομή των (μηνιαίων) μισθών των εργαζομένων μιας επιχείρησης παρουσιάζει θετική ασυμμετρία (όπως η κατανομή στο Σχήμα 9.1.16β) με μέσο 2000€ και διάμεσο 1500€. Σε μια συνάντηση των εκπροσώπων των εργαζομένων με τον εργοδότη, ο εργοδότης αναφέρεται στον υψηλό μέσο μισθό (2000€). Τι αντεπιχειρήματα, που να προκύπτουν από το είδος της ασυμμετρίας της κατανομής, έχουν οι εργαζόμενοι;*
Απάντηση: Ο μέσος μισθός είναι πράγματι 2000€, όμως, ποσοστό εργαζομένων μεγαλύτερο από το 50%, έχει μισθό μικρότερο από το μέσο μισθό. Μάλλιστα, το 50% έχει μισθό μικρότερο από 1500€.

■

Είναι φανερό ότι ακόμη και αν κάποιος μπορεί να ερμηνεύσει σωστά τα *μέτρα θέσης*, απαιτείται αρκετή εμπειρία για να μπορεί να συνοψίζει, να συνδυάζει και να συμπυκνώνει όλες τις πληροφορίες που αυτά δίνουν για την κατανομή του δείγματος. Η ***διερευνητική ανάλυση δεδομένων*** με μια έξυπνη και πολύ απλή τεχνική μας βοηθάει να παρουσιάσουμε τα κυριότερα *μέτρα θέσης* με τέτοιο τρόπο που να διευκολύνεται η εξαγωγή συμπερασμάτων για την κατανομή του δείγματος. Αναφερόμαστε στο *θηκόγραμμα*.

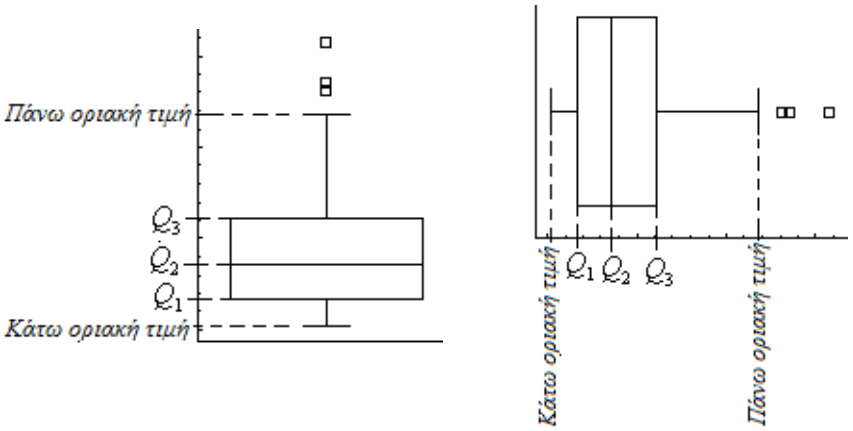
Θηκόγραμμα της κατανομής του δείγματος

Το ***θηκόγραμμα (box plot)*** *(Σχήμα 9.1.17)* είναι γνωστό και ως το ***διάγραμμα των πέντε αριθμών***. Πρόκειται για ένα ορθογώνιο με δύο *κεραίες (whiskers)* το οποίο κατασκευάζεται ως εξής: η κάτω βάση του ορθογωνίου βρίσκεται στο Q_1 και η πάνω στο Q_3. Η *διάμεσος,* $\delta = Q_2$, αναπαριστάνεται με ένα οριζόντιο ευθύγραμμο τμήμα μέσα στο ορθογώνιο και στην κατάλληλη θέση. Το πλάτος των βάσεων του ορθογωνίου καθορίζεται αυθαίρετα. Η πάνω κεραία έχει τη μορφή **T** και εκτείνεται μέχρι την ***πάνω οριακή τιμή***. Ως *πάνω οριακή τιμή* επιλέγεται ή α) η *μέγιστη* τιμή του δείγματος ή β) η μεγαλύτερη τιμή του δείγματος που είναι μικρότερη ή ίση από το *ανώτερο εσωτερικό φράγμα* $Q_3 + 1.5(Q_3 - Q_1)$ ή γ) η μεγαλύτερη τιμή του δείγματος που είναι μικρότερη ή ίση από το *ανώτερο εξωτερικό φράγμα* $Q_3 + 3(Q_3 - Q_1)$. Αντίστοιχα, η κάτω κεραία έχει τη μορφή ανεστραμμένου **T** και εκτείνεται μέχρι την ***κάτω οριακή τιμή***. Ως *κάτω οριακή τιμή* μπορεί να επιλεγεί ή α) η *ελάχιστη* τιμή του

δείγματος ή β) η μικρότερη τιμή του δείγματος που είναι μεγαλύτερη ή ίση από το *κατώτερο εσωτερικό φράγμα* $Q_1 - 1.5(Q_3 - Q_1)$ ή γ) η μικρότερη τιμή του δείγματος που είναι μεγαλύτερη ή ίση από το *κατώτερο εξωτερικό φράγμα* $Q_1 - 3(Q_3 - Q_1)$.

Στην περίπτωση που ως *οριακές τιμές* δε χρησιμοποιηθούν η *μέγιστη* και η *ελάχιστη τιμή* του δείγματος, στο *θηκόγραμμα* σημειώνονται, εφόσον φυσικά υπάρχουν, και οι τιμές του δείγματος που βρίσκονται εκτός των *εσωτερικών* ή/και των *εξωτερικών φραγμάτων*. Οι τιμές αυτές χαρακτηρίζονται ως *ακραίες* ή *εξαιρετικά ακραίες τιμές*, αντίστοιχα. Δηλαδή, μια τιμή του δείγματος χαρακτηρίζεται ***ακραία*** αν η απόστασή της από την πλευρά του ορθογωνίου που αναπαριστά το Q_1 ή από την πλευρά που αναπαριστά το Q_3, είναι μεγαλύτερη από $1.5(Q_3 - Q_1)$ ενώ χαρακτηρίζεται ***εξαιρετικά ακραία***, αν η απόσταση αυτή είναι μεγαλύτερη από $3(Q_3 - Q_1)$.

Σημειώνουμε τέλος, ότι το *θηκόγραμμα* μπορεί, αντί κατακόρυφα, να σχεδιασθεί οριζόντια.



Σχήμα 9.1.17
Θηκόγραμμα κατανομής

Παράδειγμα 9.1.17 (συνέχεια του Παραδείγματος 9.1.15β): *Για την κατανομή του δείγματος του Παραδείγματος 9.1.15β βρήκαμε* $Q_1 = 9.25,\ Q_3 = 12$ *και* $\delta = 11$.

Αν για τον καθορισμό των *οριακών τιμών* χρησιμοποιήσουμε τα *εσωτερικά φράγματα* τότε το *ανώτερο εσωτερικό φράγμα* είναι

$$Q_3 + 1.5(Q_3 - Q_1) = 12 + 1.5(12 - 9.25) = 16.125$$

άρα η *πάνω οριακή τιμή* είναι η παρατήρηση που είναι ίση με 16 *(η μεγαλύτερη παρατήρηση που είναι ίση ή μικρότερη από 16.125)*. Αντίστοιχα, το *κατώτερο εσωτερικό φράγμα* είναι

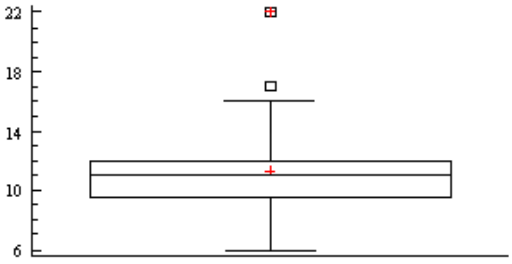
$$Q_1 - 1.5(Q_3 - Q_1) = 9.25 - 1.5(12 - 9.25) = 5.125$$

άρα η *κάτω οριακή τιμή* είναι η παρατήρηση που είναι ίση με 6 *(η μικρότερη παρατήρηση που είναι ίση ή μεγαλύτερη από 5.125)*.

Έτσι, το *θηκόγραμμα* της κατανομής του δείγματος του παραδείγματός μας είναι αυτό του *Σχήματος 9.1.18*.

Ας δούμε τι πληροφορίες μας δίνει για την κατανομή του δείγματος. Η κατανομή παρουσιάζει μια μικρή *αρνητική ασυμμετρία* διότι η *διάμεσος* βρίσκεται πιο κοντά στην πάνω πλευρά του ορθογωνίου. Το 50% των παρατηρήσεων βρίσκεται σε ένα διάστημα ίσο με το ύψος του ορθογωνίου το οποίο είναι αρκετά «συμπιεσμένο» και επιπλέον τοποθετείται περίπου στο μέσο του εύρος των παρατηρήσεων

(εξαιρουμένων των ακραίων). Η κατανομή παρουσιάζει μια *ακραία* τιμή και μια *εξαιρετικά ακραία* τιμή (είναι οι τιμές 17 και 22 αντίστοιχα).



Σχήμα 9.1.18
Το θηκόγραμμα της κατανομής του δείγματος
του Παραδείγματος 9.1.15β

■

Παράδειγμα 9.1.18 (συνέχεια του Παραδείγματος 9.1.16): *Θα κατασκευάσουμε το θηκόγραμμα του δείγματος του Παραδείγματος 9.1.16.*

Το *ανώτερο εσωτερικό φράγμα* είναι

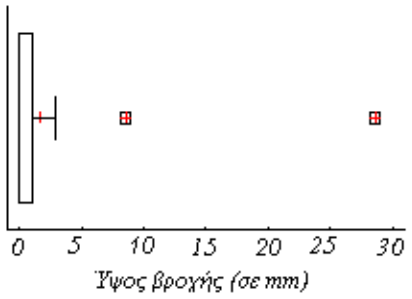
$$Q_3 + 1.5(Q_3 - Q_1) = 1.2 + 1.5(1.2 - 0) = 3$$

άρα η *πάνω οριακή τιμή* είναι η παρατήρηση που είναι ίση με 3 (η μεγαλύτερη παρατήρηση που είναι ίση ή μικρότερη από 3). Το *κατώτερο εσωτερικό φράγμα* είναι

$$Q_1 - 1.5(Q_3 - Q_1) = 0 - 1.5(1.2 - 0) = -1.8$$

άρα η *κάτω οριακή τιμή* είναι η παρατήρηση που είναι ίση με 0 (η μικρότερη παρατήρηση που είναι ίση ή μεγαλύτερη από -1.8).

Έτσι, το *θηκόγραμμα* της κατανομής του δείγματος του παραδείγματός μας είναι αυτό του *Σχήματος 9.1.19*. Είναι φανερό ότι συνοψίζει με παραστατικό τρόπο τα συμπεράσματα που σχολιάσαμε στο *Παράδειγμα 9.1.16* και επιπλέον αναδεικνύει τις ακραίες τιμές του δείγματος.



Σχήμα 9.1.19
Το θηκόγραμμα της κατανομής του δείγματος
του Παραδείγματος 9.1.16

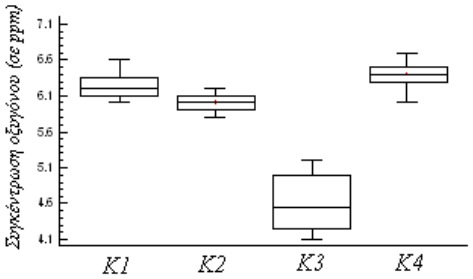
■

Σημείωση 9.1.3 (για τη χρησιμότητα και την ερμηνεία του θηκογράμματος): *Το θηκόγραμμα προσφέρεται ιδιαιτέρως για την ανίχνευση ακραίων τιμών και για την αναγνώριση της συμμετρίας ή του είδους της ασυμμετρίας της κατανομής. Αν οι δύο κεραίες έχουν ίσα ή περίπου ίσα (συγκρίσιμα) μήκη, και το ευθύγραμμο τμήμα που αναπαριστά τη διάμεσο βρίσκεται στο μέσο του ορθογωνίου ή αν τουλάχιστον δεν αποκλίνει σημαντικά προς μια από τις πλευρές του ορθογωνίου που αναπαριστούν τα Q_1 και Q_3, η κατανομή μπορεί να θεωρηθεί συμμετρική. Αν το ευθύγραμμο τμήμα που αναπαριστά τη διάμεσο αποκλίνει σημαντικά προς την πλευρά του ορθογωνίου που αναπαριστά το Q_1, η κατανομή παρουσιάζει θετική ασυμμετρία ενώ αν αποκλίνει*

σημαντικά προς την πλευρά του ορθογωνίου που αναπαριστά το Q_3, η κατανομή παρουσιάζει αρνητική ασυμμετρία (δείτε και το συντελεστή ασυμμετρίας του Bowley στην Παράγραφο 9.1.3.3). Έτσι, το θηκόγραμμα είναι χρήσιμο στην ανίχνευση ενδείξεων για το αν μπορούμε να θεωρήσουμε ότι το δείγμα προέρχεται από κανονική ή όχι κατανομή.

Αν υπάρχουν ενδείξεις ότι η κατανομή του δείγματος είναι συμμετρική και αν δεν υπάρχουν ακραίες τιμές στο δείγμα (δηλαδή, αν οι ουρές της κατανομής του δείγματος δεν είναι «παχιές») τότε το δείγμα μπορεί να προέρχεται από κανονική κατανομή.

Το θηκόγραμμα είναι επίσης πολύ χρήσιμο για τη σύγκριση των κατανομών δύο ή περισσοτέρων δειγμάτων. Δείτε για παράδειγμα τα θηκογράμματα στο Σχήμα 9.1.20. Πρόκειται για τα θηκογράμματα τεσσάρων δειγμάτων από τις κατανομές τεσσάρων τυχαίων μεταβλητών που αντίστοιχα εκφράζουν τη συγκέντρωση διαλυμένου οξυγόνου στις θέσεις Κ1, Κ2, Κ3 και Κ4 της κοίτης ενός ποταμού (σε ppm).



Σχήμα 9.1.20
Σύγκριση τεσσάρων κατανομών
μέσω των αντίστοιχων θηκογραμμάτων

Παρατηρείστε πόσο διευκολύνεται η σύγκριση των κατανομών των τεσσάρων δειγμάτων. Για παράδειγμα, σχετικά με τη θέση τους αβίαστα προκύπτει ότι η κατανομή του δείγματος που αντιστοιχεί στη θέση Κ3 τοποθετείται πιο χαμηλά (σε μικρότερες τιμές) από τις κατανομές των δειγμάτων που αντιστοιχούν στις θέσεις Κ1, Κ2 και Κ4 (κοντά στη θέση Κ3 ρίχνονται βιομηχανικά απόβλητα). Άμεσα επίσης προκύπτει ότι η κατανομή του δείγματος που αντιστοιχεί στη θέση Κ3 παρουσιάζει τη μεγαλύτερη μεταβλητότητα ενώ και οι τέσσερις κατανομές φαίνεται να έχουν την ίδια μορφή (περίπου συμμετρικές).

■

9.1.3.2 Μέτρα μεταβλητότητας/διασποράς

Για την περιγραφή της μεταβλητότητας των τιμών ενός δείγματος θα παρουσιάσουμε τέσσερα ***μέτρα διασποράς***.

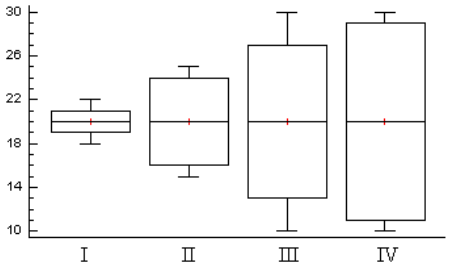
- Το ***εύρος του δείγματος***
- Το ***ενδοτεταρτημοριακό εύρος του δείγματος***
- Τη ***δειγματική διακύμανση***
- Τη ***δειγματική τυπική απόκλιση***.

Στον *Πίνακα 9.1.11* δίνονται τέσσερα (υποθετικά) δείγματα μεγέθους 5, το καθένα.

Δείγμα Ι	**Δείγμα ΙΙ**	**Δείγμα ΙΙΙ**	**Δείγμα ΙV**
18	15	10	10
19	16	13	11
20	20	20	20
21	24	27	29
22	25	30	30

Πίνακας 9.1.11
Τέσσερα δείγματα με ίδιο μέσο και ίδια διάμεσο

Εύκολα διαπιστώνεται ότι καθένα από τα τέσσερα δείγματα έχει *μέσο* 20 και *διάμεσο* επίσης 20. Όμως, αν στο *Σχήμα 9.1.21* παρατηρήσουμε τα αντίστοιχα *θηκογράμματα*, αβίαστα προκύπτει ότι οι κατανομές τους διαφέρουν. Η μεταβλητότητα των τιμών τους γύρω από τα στατιστικά θέσης εμφανώς δεν είναι ίδια. Ας δούμε λοιπόν πώς τα ***μέτρα διασποράς*** ποσοτικοποιούν τη μεταβλητότητα των τιμών ενός δείγματος και βοηθούν έτσι στην ανίχνευση και την ανάδειξη τέτοιων διαφοροποιήσεων.



Σχήμα 9.1.21
Τα θηκογράμματα των δειγμάτων του Πίνακα 9.1.11

(α) Εύρος και Ενδοτεταρτημοριακό εύρος του δείγματος

Το ***εύρος (range)***, *R*, των τιμών του δείγματος ορίζεται ως η διαφορά της μικρότερης από τη μεγαλύτερη τιμή του, δηλαδή

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

Είναι το πιο απλό μέτρο *διασποράς* και έχει τα ακόλουθα πλεονεκτήματα και μειονεκτήματα.

Πλεονεκτήματα

- *Είναι πολύ απλό στον υπολογισμό.*
- *Είναι πολύ χρήσιμο στον έλεγχο ποιότητας.*
- *Μπορεί να χρησιμοποιηθεί για την εκτίμηση της τυπικής απόκλισης.*

Μειονεκτήματα

- *Δε θεωρείται αξιόπιστο μέτρο διασποράς, επειδή βασίζεται μόνο στη μικρότερη και τη μεγαλύτερη παρατήρηση και συνεπώς είναι ευαίσθητο σε ακραίες τιμές.*
- *Δε χρησιμοποιείται για περαιτέρω στατιστική ανάλυση.*

Αν για την περιγραφή με ποσοτικούς όρους της μεταβλητότητας των τεσσάρων δειγμάτων του *Πίνακα 9.1.11*, χρησιμοποιήσουμε το *εύρος,* βλέπουμε ότι ενώ ανιχνεύει τη διαφορά στη μεταβλητότητα μεταξύ, π.χ., των δειγμάτων I και II (το δείγμα I έχει εύρος 22-18 = 4 ενώ το δείγμα II έχει εύρος 25-15 = 10) εντούτοις, δεν ανιχνεύει τη διαφορά που υπάρχει στη μεταβλητότητα μεταξύ των δειγμάτων III και IV (και το δείγμα III και το δείγμα IV έχουν εύρος 30-10 = 20). Δηλαδή, υπάρχουν κατανομές που έχουν ίσους *μέσους*, ίσες *διαμέσους* και ίδιο *εύρος* και εντούτοις, διαφέρουν. Δεν αρκεί επομένως το *εύρος* για να περιγραφεί η μεταβλητότητα των τιμών ενός δείγματος. Είναι φανερό ότι αυτό οφείλεται στο ότι στον υπολογισμό του συμμετέχουν μόνο δυο παρατηρήσεις. Για να αντιμετωπίσουμε αυτό το πρόβλημα μπορούμε να χρησιμοποιήσουμε ως μέτρο της μεταβλητότητας τη διαφορά $Q_3 - Q_1$ για τον υπολογισμό της οποίας συμμετέχουν σαφώς περισσότερες παρατηρήσεις (όσες συμμετέχουν στον υπολογισμό των Q_1 και Q_3). Η διαφορά αυτή ονομάζεται ***ενδοτεταρτημοριακό εύρος (interquartile range)***[7]**.** Το *ενδοτεταρτημοριακό εύρος,* μας δίνει το πλάτος του κεντρικού (γύρω από τη *διάμεσο*) διαστήματος εντός του οποίου βρίσκεται το 50% των τιμών του δείγματος, γι' αυτό θεωρείται η «καρδιά» της

[7] *Ανάλογα ορίζεται το ενδοδεκατημοριακό εύρος* $x_{0.9} - x_{0.1}$.

κατανομής. Όσο μικρότερο είναι αυτό το διάστημα τόσο μικρότερη εν γένει είναι η μεταβλητότητα των τιμών του δείγματος.

Παρατηρείστε ότι αν για την αριθμητική περιγραφή της μεταβλητότητας στα τέσσερα δείγματα του παραδείγματός μας χρησιμοποιήσουμε το *ενδοτεταρτημοριακό εύρος,* ανιχνεύονται πλέον όλες οι υπάρχουσες διαφορές μεταξύ των τεσσάρων δειγμάτων.

Σημειώνουμε ότι το *ενδοτεταρτημοριακό εύρος* δεν επηρεάζεται από ακραίες τιμές. Αξίζει επίσης να επισημάνουμε ότι το *εύρος*, αντίθετα με το *ενδοτεταρτημοριακό εύρος*, είναι πολύ ευαίσθητο σε αλλαγές στο μέγεθος του δείγματος. Δηλαδή, είναι δυνατόν, αύξηση του μεγέθους του δείγματος ακόμη και κατά μια μονάδα, να προκαλέσει δυσανάλογη αύξηση του εύρους. Αν για παράδειγμα, οι παρατηρήσεις

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

συμπληρωθούν με την παρατήρηση 10, το εύρος του δείγματος από 4 γίνεται 9!

(β) Δειγματική διακύμανση και δειγματική τυπική απόκλιση

Η *διακύμανση* ενός δείγματος $X_1, X_2, \ldots, X_\nu$, ορίζεται για να εκφράσει (με κάποιον τρόπο) τον *μέσο* των αποκλίσεων $X_i - \overline{X}$ των τιμών του δείγματος από τον *δειγματικό μέσο* $\overline{X}$ (θυμηθείτε και πώς ορίσαμε την πληθυσμιακή διακύμανση σ^2).

Έστω λοιπόν $X_1, X_2, \ldots, X_\nu$ ένα τυχαίο δείγμα από έναν πληθυσμό και $x_1, x_2, \ldots, x_\nu$ μια πραγματοποίησή του. Έστω επίσης, $y_1, y_2, \ldots, y_k$ $(k \le \nu)$ οι k διαφορετικές μεταξύ τους τιμές από τις $x_1, x_2, \ldots, x_\nu$.

Η ***δειγματική διακύμανση (sample variance)*** συμβολίζεται με S^2 και ορίζεται από τον τύπο

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

Όπως ο *δειγματικός μέσος,* αλλά και κάθε *στατιστικό*, η *δειγματική διακύμανση* είναι τυχαία μεταβλητή γι' αυτό και συμβολίζεται με κεφαλαίο γράμμα. Η τιμή της για συγκεκριμένη πραγματοποίηση $x_1, x_2, \ldots, x_\nu$ του δείγματος συμβολίζεται με s^2. Έτσι,

$$s^2 = \frac{1}{\nu - 1} \sum_{i=1}^{\nu} (x_i - \overline{x})^2 .$$

Εύκολα αποδεικνύεται ότι

$$s^2 = \frac{1}{\nu - 1} \left(\sum_{i=1}^{\nu} x_i^2 - \nu \overline{x}^2 \right) .$$

Αυτή η έκφραση του τύπου ορισμού της *δειγματικής διακύμανσης* χρησιμοποιείται συχνά στην πράξη γιατί διευκολύνει τον υπολογισμό της. Επίσης, πολύ συχνά χρησιμοποιούνται οι τύποι

$$s^2 = \frac{1}{\nu - 1} \sum_{i=1}^{k} (y_i - \overline{x})^2 \nu_i \quad \text{και} \quad s^2 = \frac{1}{\nu - 1} \left(\sum_{i=1}^{k} \nu_i y_i^2 - \nu \overline{x}^2 \right)$$

που επίσης αποδεικνύονται εύκολα. Μέσω αυτών των τύπων μπορούμε να υπολογίζουμε τη *δειγματική διακύμανση* χρησιμοποιώντας τις *συχνότητες* ν_i των διαφορετικών τιμών $y_1, y_2, \ldots, y_k$ $(k \le \nu)$ που εμφανίσθηκαν στο δείγμα.

Η ***δειγματική τυπική απόκλιση (sample standard deviation)*** ορίζεται ως η (θετική) τετραγωνική ρίζα της *δειγματικής διακύμανσης* και συμβολίζεται με κεφαλαίο S. Έτσι,

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

Με πεζό s, συμβολίζουμε τη συγκεκριμένη τιμή της S που υπολογίζεται από μια πραγματοποίηση $x_1, x_2, \ldots, x_\nu$ του τυχαίου δείγματος $X_1, X_2, \ldots, X_\nu$. Δηλαδή,

$$s = \sqrt{\frac{1}{\nu - 1}\sum_{i=1}^{\nu}(x_i - \overline{x})^2} \ .$$

Θυμηθείτε ότι την *τυπική απόκλιση* του *πληθυσμού* (δηλαδή, την *τυπική απόκλιση* της κατανομής της *X*) τη συμβολίζουμε με σ ή με σ_X.

Για την *τυπική απόκλιση του δείγματος*, από τους αντίστοιχους για τη *δειγματική διακύμανση* τύπους, παίρνουμε

$$s = \sqrt{\frac{1}{\nu - 1}\left(\sum_{i=1}^{\nu} x_i^2 - \nu\overline{x}^2\right)} = \sqrt{\frac{1}{\nu - 1}\sum_{i=1}^{k}(y_i - \overline{x})^2 \nu_i} = \sqrt{\frac{1}{\nu - 1}\left(\sum_{i=1}^{k}\nu_i y_i^2 - \nu\overline{x}^2\right)} \ .$$

Ας δούμε ένα παράδειγμα υπολογισμού της *δειγματικής διακύμανσης* και της *δειγματικής τυπικής απόκλισης*.

Παράδειγμα 9.1.19 (συνέχεια του Παραδείγματος 9.2): *Η διακύμανση και η τυπική απόκλιση του δείγματος από τη μεταβλητή «αριθμός παιδιών οικογένειας» του Παραδείγματος 9.2, υπολογίζεται εύκολα αν συμπληρώσουμε τον πίνακα συχνοτήτων με μια ακόμη στήλη όπως φαίνεται στον Πίνακα 9.1.12.*

y_i	ν_i	f_i	N_i	F_i	$\nu_i y_i$	$\nu_i y_i^2$
0	2	0.1	2	0.1	0	0
1	4	0.2	6	0.3	4	4
2	10	0.5	16	0.8	20	40
3	2	0.1	18	0.9	6	18
4	2	0.1	20	1.0	8	32
Σύνολα	**20**	**1.00**			**38**	**94**

Πίνακας 9.1.12
Υπολογισμός της διακύμανσης και της τυπικής απόκλισης του δείγματος από τη μεταβλητή «αριθμός παιδιών οικογένειας» του Παραδείγματος 9.2

Δημιουργούμε μια στήλη με τα γινόμενα $\nu_i y_i^2$ και έτσι έχουμε

$$s^2 = \frac{1}{\nu - 1}\left(\sum_{i=1}^{k}\nu_i y_i^2 - \nu\overline{x}^2\right) = \frac{1}{19}\left(94 - 20 \cdot 1.9^2\right) = 1.147$$

και επομένως

$$s = \sqrt{1.147} = 1.07 \ .$$

■

Παρατήρηση 9.1.10: *Ίσως δημιουργεί απορίες το γεγονός ότι στον τύπο της δειγματικής διακύμανσης το άθροισμα $\sum_{i=1}^{\nu}(x_i - \overline{x})^2$ διαιρείται με $\nu - 1$ αντί με ν. Αυτό γίνεται διότι, όπως θα δούμε στο 11ο Κεφάλαιο, μπορεί να αποδειχθεί ότι αν διαιρέσουμε με $\nu - 1$ η δειγματική διακύμανση S^2 είναι αμερόληπτη εκτιμήτρια της πληθυσμιακής διακύμανσης σ^2. Δηλαδή, αν πάρουμε όλα τα δυνατά δείγματα μεγέθους ν και υπολογίσουμε τις διακυμάνσεις τους, s^2, τότε ο μέσος τους (των s^2) θα είναι ίσος με τη διακύμανση σ^2 του πληθυσμού!! Δηλαδή, κατά μέσο όρο, η S^2 εκτιμά σωστά την σ^2. Ούτε την υποεκτιμά ούτε την υπερεκτιμά. Ενώ αν διαιρούσαμε με ν,*

όπως θα δούμε, η δειγματική διακύμανση κατά μέσο όρο θα υποεκτιμούσε τη διακύμανση σ^2 του πληθυσμού. Γι' αυτή την υποεκτίμηση υπάρχει η εξής λογική εξήγηση: είναι λογικό, οι τιμές x_i να είναι πιο κοντά στον μέσο τους $\bar{x}$, παρά στην πληθυσμιακή μέση τιμή μ (που χρησιμοποιείται για τον υπολογισμό της σ^2). Έτσι, και τα τετράγωνα των αποκλίσεων $(x_i - \bar{x})^2$, των x_i από τον $\bar{x}$, θα τείνουν να είναι μικρότερα από τα τετράγωνα των αποκλίσεων τους $(x_i - \mu)^2$, από τη μ. Διαιρώντας με $\nu - 1$ αντί με ν, αυτή την τάση υποεκτίμησης αντισταθμίζουμε.

■

Ιδιότητες της δειγματικής διακύμανσης και της δειγματικής τυπικής απόκλισης

Εύκολα μπορεί να αποδειχθεί ότι η *δειγματική διακύμανση* και η *δειγματική τυπική απόκλιση* έχουν τις ακόλουθες ιδιότητες.

1. Αν οι τιμές του δείγματος είναι μεταξύ τους ίσες τότε η *διακύμανσή* τους και επομένως και η *τυπική απόκλισή* τους είναι μηδέν.
2. Αν $t_i = x_i + \beta$ τότε $s_t^2 = s_x^2$ και $s_t = s_x$.

   Δηλαδή, αν στις τιμές του δείγματος $x_1, x_2, \ldots, x_\nu$ προσθέσουμε μια σταθερή ποσότητα β (θετική ή αρνητική), τότε οι τιμές t_i που προκύπτουν, έχουν ίδια *διακύμανση* και ίδια *τυπική απόκλιση* με τις αρχικές.
3. Αν $t_i = \alpha x_i$ τότε $s_t^2 = \alpha^2 s_x^2$ και $s_t = |\alpha| s_x$.

   Δηλαδή, αν οι τιμές $x_1, x_2, \ldots, x_\nu$ του δείγματος, πολλαπλασιασθούν με την ίδια ποσότητα α, τότε η *διακύμανσή* τους πολλαπλασιάζεται με α^2 και η *τυπική απόκλισή* τους με $|\alpha|$.
4. Αν $t_i = \alpha x_i + \beta$ τότε $s_t^2 = \alpha^2 s_x^2$ και $s_t = |\alpha| s_x$

Επίσης, η *δειγματική διακύμανση* και η *δειγματική τυπική απόκλιση* έχουν τα ακόλουθα πλεονεκτήματα και μειονεκτήματα.

Πλεονεκτήματα	**Μειονεκτήματα**
• *Για τον υπολογισμό τους, λαμβάνονται υπόψη όλες οι παρατηρήσεις.* • *Έχουν μεγάλη εφαρμογή στη στατιστική συμπερασματολογία*	• *Η διακύμανση δεν εκφράζεται στις ίδιες μονάδες με τη μεταβλητή.*

Το νόημα και η ερμηνεία της δειγματικής τυπικής απόκλισης

Είναι φανερό ότι η *δειγματική τυπική απόκλιση* απαντά στο ερώτημα: *πόσο μακριά από τον μέσο τους βρίσκονται οι τιμές του δείγματος;* Έτσι, όταν οι τιμές του δείγματος δε διαφέρουν πολύ από τον *μέσο* τους, η *δειγματική τυπική απόκλιση* είναι μικρή, ενώ αντίθετα, η *δειγματική τυπική απόκλιση* μεγαλώνει όσο περισσότερο οι τιμές του δείγματος «διασκορπίζονται» γύρω από τον *μέσο* τους. Δηλαδή, η *δειγματική τυπική απόκλιση* μας δίνει ένα μέτρο της μέσης απόστασης-απόκλισης των τιμών του δείγματος από τον *μέσο* τους. Συνεπώς, έχει νόημα να χρησιμοποιείται μόνο σε συνδυασμό με τον *δειγματικό μέσο*.

Πρακτικά όμως τι σημαίνει ***«μεγάλη»*** ή ***«μικρή» τυπική απόκλιση****;* Ας προσπαθήσουμε να απαντήσουμε σε αυτό το ερώτημα μέσα από συγκεκριμένα προβλήματα.

1) Αν για καθένα από τα τέσσερα δείγματα του *Πίνακα 9.1.11*, υπολογίσουμε την *τυπική απόκλισή* του βρίσκουμε 1.6, 4.5, 8.6 και 9.5, αντίστοιχα. Μπορούμε να ισχυρισθούμε ότι η μεταβλητότητα, π.χ. του δείγματος IV είναι μεγαλύτερη από τη

μεταβλητότητα του δείγματος Ι επειδή είναι $9.5 > 1.6$*;* Η απάντηση είναι ναι γιατί τα δείγματα έχουν **τον ίδιο *μέσο***.

Αν όμως, επιχειρήσουμε να συγκρίνουμε τις *μεταβλητότητες* δύο ή περισσοτέρων δειγμάτων που έχουν άνισους *μέσους,* με βάση μόνο τις *τυπικές αποκλίσεις* τους, τότε είναι πιθανό να οδηγηθούμε σε λάθος συμπεράσματα.

Παράδειγμα 9.1.20 (συντελεστής μεταβλητότητας): *Έστω δύο δείγματα με* $\bar{x} = 5,\ s_1 = 1$ *το ένα και* $\bar{w} = 150,\ s_2 = 12$ *το άλλο. Μπορούμε να ισχυρισθούμε ότι το δεύτερο δείγμα παρουσιάζει μεγαλύτερη μεταβλητότητα από το πρώτο επειδή έχει μεγαλύτερη τυπική απόκλιση;*

<u>*Απάντηση:*</u> Ασφαλώς όχι, αφού *«άλλο 1 στα 5 και άλλο 12 στα 150»*. Είναι επομένως λογικό να αναζητήσουμε ένα μέτρο το οποίο να εκφράζει την *τυπική απόκλιση* των τιμών του δείγματος ***ως ποσοστό του μέσου τους***. Δηλαδή, ένα μέτρο *σχετικής μεταβλητότητας*. Ένα τέτοιο μέτρο είναι ο ***συντελεστής μεταβλητότητας (coefficient of variation)*** ο οποίος συμβολίζεται με CV και ορίζεται με τον τύπο

$$CV = \frac{s}{|\bar{x}|}100\% .$$

Έτσι, αν συγκρίνουμε τις *τυπικές αποκλίσεις* των δύο δειγμάτων αφού προηγουμένως κάθε μια τη δούμε ως ***ποσοστό του μέσου με βάση τον οποίο υπολογίσθηκε***, δηλαδή, αν υπολογίσουμε τους ***συντελεστές μεταβλητότητας*** παρατηρούμε ότι για το πρώτο δείγμα είναι

$$CV = \frac{1}{5}100\% = 20\%$$

και για το δεύτερο

$$CV = \frac{12}{150}100\% = 8\% .$$

Δηλαδή, στο πρώτο δείγμα η *τυπική απόκλιση* είναι το 20% του *μέσου* του ενώ στο δεύτερο η *τυπική απόκλιση* είναι το 8% του *μέσου* του. Συνεπώς, μεγαλύτερη μεταβλητότητα παρουσιάζεται στο πρώτο και όχι στο δεύτερο δείγμα (μάλιστα είναι $20/8 = 2.5$ φορές μεγαλύτερη!).

■

Από τα παραπάνω είναι φανερό ότι ο *CV* μπορεί να χρησιμοποιηθεί:

- *Ως μέτρο σύγκρισης της μεταβλητότητας δύο ή περισσοτέρων δειγμάτων που έχουν διαφορετικούς μέσους.*
- *Ως μέτρο ομοιογένειας ενός δείγματος (αν σε ένα δείγμα είναι CV<10% τότε το* δείγμα θεωρείται ομοιογενές).

2) Ας δούμε ένα ακόμη πρακτικό πρόβλημα.

Παράδειγμα 9.1.21 (η τυπική απόκλιση ως μονάδα μέτρησης): *Ένας φοιτητής βαθμολογήθηκε στις εξετάσεις του Ιουνίου 2012 στο μάθημα της Στατιστικής με 8. Ένας άλλος φοιτητής βαθμολογήθηκε στο ίδιο μάθημα στις εξετάσεις του Ιουνίου 2013 με 7. Με κριτήριο το βαθμό στις εξετάσεις, ποιος από τους δύο φοιτητές είναι καλύτερος στη Στατιστική;*

<u>*Απάντηση:*</u> Αν δε βιαστούμε να απαντήσουμε, διαπιστώνουμε ότι ουσιαστικά μας ζητούν να συγκρίνουμε «ανόμοια πράγματα», αφού πρέπει να συγκρίνουμε δυο τιμές η κάθε μια από τις οποίες ανήκει σε διαφορετική κατανομή. Η τιμή 8 ανήκει στην κατανομή των βαθμών στις εξετάσεις του Ιουνίου 2012 ενώ η τιμή 7 ανήκει στην κατανομή των βαθμών στις εξετάσεις του Ιουνίου 2013. Για να συγκριθούν επομένως

οι δύο τιμές, πρέπει να προσδιορισθεί πρώτα η θέση της κάθε μίας μέσα στην κατανομή της.

Έτσι, αν οι βαθμοί των φοιτητών τον Ιούνιο 2012 είχαν *μέσο* 7.5 και *τυπική απόκλιση* 0.6 και τον Ιούνιο του 2013 είχαν *μέσο* 5.5 και *τυπική απόκλιση* 1.1 τότε είναι προφανές ότι το κλάσμα

$$\frac{8-7.5}{0.6}=\frac{0.5}{0.6}=+0.8$$

εκφράζει την απόσταση-απόκλιση της τιμής 8 από τον *μέσο* της κατανομής της **σε μονάδες *τυπικής απόκλισης***. Δηλαδή, δείχνει *«πόσες φορές χωράει η τυπική απόκλιση 0.6 στην απόσταση* $8-7.5$*»*. Ομοίως, το κλάσμα

$$\frac{7-5.5}{1.1}=\frac{1.5}{1.1}=+1.4$$

δείχνει *«πόσες φορές χωράει η τυπική απόκλιση 1.1 στην απόσταση* $7-5.5$*»*.

Είναι πλέον φανερό ότι ο βαθμός 7 είναι καλύτερος από το βαθμό 8 με την έννοια ότι απέχει από τον *μέσο* της κατανομής του +1.4 *τυπικές αποκλίσεις* ενώ ο βαθμός 8 απέχει από τον *μέσο* της δικής του κατανομής +0.8 *τυπικές αποκλίσεις*. Δηλαδή, ***ο βαθμός 7 είναι 1.4 τυπικές αποκλίσεις μεγαλύτερος από τον μέσο της κατανομής του ενώ ο βαθμός 8 είναι 0.8 τυπικές αποκλίσεις μεγαλύτερος από τον μέσο της δικής του κατανομής***.

Η δειγματική τυπική απόκλιση μπορεί, επομένως, να χρησιμοποιηθεί ως μονάδα μέτρησης της απόστασης μιας (οποιασδήποτε) τιμής του δείγματος από τον δειγματικό μέσο.

z-τιμές

Αν από την τιμή x_i, $i=1,2,\ldots,\nu$ αφαιρέσουμε τον *δειγματικό μέσο* $\bar{x}$ και τη διαφορά $x_i-\bar{x}$ που προκύπτει τη διαιρέσουμε με τη *δειγματική τυπική απόκλιση* s, προκύπτει η (μετασχηματισμένη) τιμή

$$\frac{x_i-\bar{x}}{s}$$

η οποία συμβολίζεται με z_i, δηλαδή

$$z_i=\frac{x_i-\bar{x}}{s}$$

και ονομάζεται z_i*-τιμή* της τιμής x_i.

Οι z_i*-τιμές* έχουν τις ακόλουθες, πολύ ενδιαφέρουσες, ιδιότητες.

- Η z_i*-τιμή* μιας τιμής x_i, εκφράζει σε **μονάδες *τυπικής απόκλισης*** την απόσταση της x_i από τον *δειγματικό μέσο* $\bar{x}$.
- Αν μια z_i*-τιμή* είναι *θετική* αυτό σημαίνει ότι η τιμή x_i είναι μεγαλύτερη από τον *δειγματικό μέσο* ενώ αν είναι *αρνητική* σημαίνει ότι η τιμή x_i είναι μικρότερη από τον *δειγματικό μέσο*.
- Ο *μέσος* των z_i*-τιμών* είναι πάντα 0 και *η τυπική τους απόκλιση* είναι πάντα 1. Δηλαδή, $\bar{z}=0$ και $s_z=1$**.** Η απόδειξη είναι προφανής αν παρατηρήσουμε ότι ο μετασχηματισμός

$$z_i=\frac{x_i-\bar{x}}{s}$$

είναι της μορφής $z_i = \alpha x_i + \beta$ με

$$\alpha = \frac{1}{s} \text{ και } \beta = \frac{-\bar{x}}{s}.$$

- Ίσες αποστάσεις z_i*-τιμών* μιας κατανομής, έχουν ταυτόσημο νόημα. Για παράδειγμα, η διαφορά μεταξύ των z_i*-τιμών* 2 και 2.5 είναι ταυτόσημη με τη διαφορά μεταξύ των z_i*-τιμών* 3 και 3.5. Και οι δύο διαφορές δείχνουν μια απόσταση μισής *τυπικής απόκλισης*.
- Στις z_i*-τιμές* το 0 έχει νόημα, δηλαδή, δεν ορίζεται συμβατικά-αυθαίρετα. Η z_i*-τιμή* 0 σημαίνει *«έλλειψη απόστασης»,* δηλαδή, η τιμή x_i συμπίπτει με τον *δειγματικό μέσο* $\bar{x}$.
- Η μορφή της κατανομής των z_i*-τιμών* είναι όμοια με τη μορφή της κατανομής των x_i τιμών (διατηρούνται π.χ. οι ασυμμετρίες ή η συμμετρία). Έτσι, αν η κατανομή των x_i τιμών έχει μορφή *κανονικής* κατανομής τότε και η κατανομή των z_i*-τιμών* θα έχει μορφή *κανονικής* κατανομής.
- Οι z_i*-τιμές* μπορούν να χρησιμοποιηθούν για την ανίχνευση ακραίων τιμών. Στη συνέχεια θα δούμε σχετικά παραδείγματα.
- Οι z_i*-τιμές* μπορούν να χρησιμοποιηθούν για τη σύγκριση τιμών που ανήκουν σε διαφορετικές κατανομές. Ας δούμε ένα ακόμη παράδειγμα.

Παράδειγμα 9.1.22 (σύγκριση τιμών από διαφορετικές κατανομές): *Στην Ελλάδα, όπως είναι γνωστό, η βαθμολογία των αποφοίτων δευτεροβάθμιας εκπαίδευσης δίνεται σε κλίμακα από 1 μέχρι 20. Στις Η.Π.Α., συνήθως δίνεται σε μια κλίμακα από 1 μέχρι 4. Σε πολλές άλλες χώρες δίνεται σε κλίμακα από 1 μέχρι 100. Σε ένα σχολείο των Η.Π.Α. η κατανομή της βαθμολογίας των αποφοίτων έχει μέσο 3.2 και τυπική απόκλιση 0.2, σε ένα ελληνικό σχολείο έχει μέσο 14.2 και τυπική απόκλιση 2.1 και σε ένα ολλανδικό έχει μέσο 76 και τυπική απόκλιση 7. Πώς μπορούμε να συγκρίνουμε το βαθμό 3.6 ενός μαθητή του σχολείου των Η.Π.Α. με το βαθμό 18.4 ενός μαθητή του ελληνικού σχολείου και με το βαθμό 90 ενός μαθητή του ολλανδικού σχολείου;*

<u>*Απάντηση:*</u> Οι αντίστοιχες *z-τιμές* των βαθμών είναι

$$\frac{3.6-3.2}{0.2} = +2\,,\ \frac{18.4-14.2}{2.1} = +2 \ \text{ και } \ \frac{90-76}{7} = +2\,.$$

Συνεπώς, οι τρεις μαθητές πήραν τα απολυτήριά τους με βαθμούς που βρίσκονται σε ίσες αποστάσεις πάνω από τη μέση βαθμολογία του σχολείου τους.

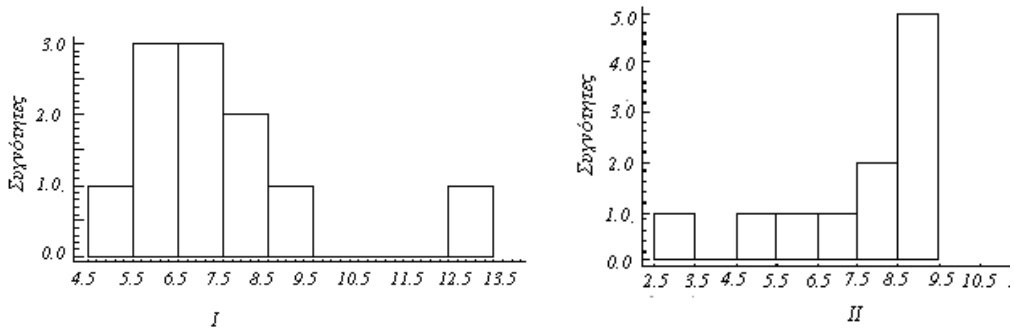
■

Παρατήρηση 9.1.11: *Οι z-τιμές είναι ένα μέτρο σχετικής απόστασης. Επομένως, όταν χρησιμοποιούνται για τη σύγκριση τιμών που ανήκουν σε διαφορετικές κατανομές, θα πρέπει οι κατανομές αυτές να έχουν παραπλήσιες μορφές. Διαφορετικά, η πληροφορία που θα πάρουμε από τη σύγκριση z-τιμών θα είναι διφορούμενη-ασαφής (θυμηθείτε ότι με όρους ποσοστών, ίσες αποστάσεις μπορεί να περιέχουν πολύ διαφορετικά ποσοστά παρατηρήσεων). Ας δούμε ένα παράδειγμα*[8].

Παράδειγμα 9.1.23: *Δίνονται δύο δείγματα και τα αντίστοιχα ιστογράμματα συχνοτήτων.*

Δείγμα Ι	*7.46 6.77 12.74 7.11 7.81 8.84 6.10 5.39 8.15 6.42 5.73*
Δείγμα ΙΙ	*9.14 8.14 8.74 8.77 9.26 8.10 6.10 3.10 9.13 7.26 4.74*

[8] *Tal, J. Reading between the numbers, McGraw-Hill, 2001.*



Από τα δύο *ιστογράμματα* είναι προφανές ότι οι δύο κατανομές διαφέρουν σημαντικά αφού στην I οι τιμές κατανέμονται μεταξύ 4.5 και 9.5 με μια ακραία τιμή προς τα δεξιά, ενώ στη II υπάρχει μεγάλη συγκέντρωση τιμών μεταξύ 8.5 και 9.5 και οι υπόλοιπες κλάσεις έχουν από μία μόνο τιμή (εκτός από μια που έχει δύο τιμές). Παρόλα αυτά, οι δύο κατανομές έχουν ίσους *μέσους* και ίσες *τυπικές αποκλίσεις* (7.5 και 1.93 αντίστοιχα). Η τιμή 6.1 ανήκει και στα δύο δείγματα και επομένως θα έχει και στα δύο δείγματα ίδια *z-τιμή*

$$z = \frac{6.1 - 7.5}{1.93} = -0.73 .$$

Δηλαδή, η τιμή 6.1 και στα δύο δείγματα βρίσκεται 0.73 *τυπικές αποκλίσεις* αριστερά του *μέσου*. Αυτό όμως δεν εμποδίζει να είναι η πραγματικότητα για την τιμή 6.1 πολύ διαφορετική στις δύο κατανομές. Αν παρατηρήσουμε τα αντίστοιχα *ιστογράμματα* των κατανομών βλέπουμε ότι στην κατανομή I η τιμή 6.1 έχει «δεσπόζουσα θέση» (βρίσκεται στο mainstream της κατανομής) ενώ στην II βρίσκεται μόνη της και περιβάλλεται από κλάσεις με μια μόνο τιμή! Η σύγκριση, επομένως, δύο *z-τιμών* από διαφορετικές κατανομές, δεν αποδίδει την πραγματική εικόνα αν οι κατανομές έχουν πολύ διαφορετική μορφή.

■

Ας ολοκληρώσουμε τις επισημάνσεις και τα σχόλια για την ερμηνεία, το νόημα και τη χρησιμότητα της *τυπικής απόκλισης* με ένα ακόμη παράδειγμα αξιοποίησης της.

3) Άραγε μπορούμε, με βάση την *τυπική απόκλιση,* να καθορίσουμε διαστήματα γύρω από τον *δειγματικό μέσο* εντός των οποίων να βρίσκεται συγκεκριμένο ποσοστό τιμών του δείγματος[9]; Η απάντηση είναι ότι μπορούμε.

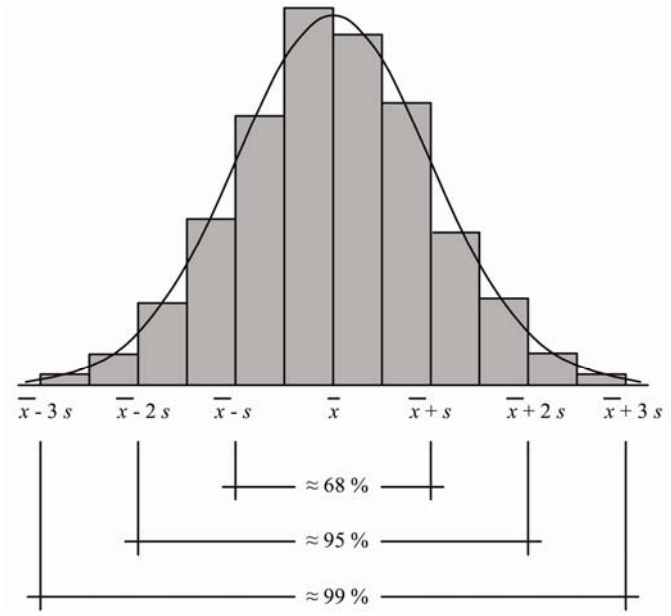
Η *ανισότητα Chebyshev* μας λέει ότι το ποσοστό των τιμών του δείγματος που βρίσκονται στο διάστημα $(\bar{x} - ks,\ \bar{x} + ks)$ είναι τουλάχιστον $1 - (1/k^2)$. Για παράδειγμα, το ποσοστό των τιμών του δείγματος που βρίσκονται στο διάστημα $(\bar{x} - 2s,\ \bar{x} + 2s)$ είναι τουλάχιστον 75%.

Επίσης, αν η κατανομή του δείγματος προσομοιάζει με μια ***κανονική*** *κατανομή* (έχει κωδωνοειδή μορφή), τότε ισχύει ο ακόλουθος κανόνας, γνωστός στη βιβλιογραφία ως ***εμπειρικός κανόνας (empirical rule)*** *(Σχήμα 9.1.22)* γιατί πολύ συχνά επαληθεύεται εμπειρικά σε διάφορα φαινόμενα και πειράματα (και όχι γιατί δε μπορεί να επιβεβαιωθεί θεωρητικά, αφού όπως είδαμε στο Α΄ Μέρος, μπορεί να αποδειχθεί).

- *Στο διάστημα* $(\bar{x} - s,\ \bar{x} + s)$ *βρίσκεται περίπου το 68% των παρατηρήσεων*
- *Στο διάστημα* $(\bar{x} - 2s,\ \bar{x} + 2s)$ *βρίσκεται περίπου το 95% των παρατηρήσεων*

[9] *Δηλαδή κάτι ανάλογο με τα διαστήματα που καθορίζουμε με βάση τα ποσοστιαία σημεία. Π.χ. στο διάστημα που ορίζεται από τα* $x_{0.1}$ *και* $x_{0.9}$ *βρίσκεται το 80% των παρατηρήσεων.*

- *Στο διάστημα* $(\bar{x}-3s,\ \bar{x}+3s)$ *βρίσκονται όλες σχεδόν οι παρατηρήσεις (πάνω από το 99%).*



Σχήμα 9.1.22
Ο εμπειρικός κανόνας

Ας δούμε με ένα παράδειγμα πώς αξιοποιούνται αυτές οι πληροφορίες (για το ποσοστό των παρατηρήσεων που βρίσκονται σε ένα διάστημα πλάτους $2ks$ γύρω από τον *μέσο*).

Παράδειγμα 9.1.24: *Μια αυτόματη μηχανή συσκευασίας τροφίμων έχει προγραμματισθεί να συσκευάζει δημητριακά σε φακελάκια των 13 γραμμαρίων. Ζυγίσαμε 15 τέτοια φακελάκια (ακριβέστερα, το περιεχόμενό τους) τα οποία είχαμε επιλέξει τυχαία από το σύνολο της παραγωγής μιας ημέρας και βρήκαμε μέσο βάρος 12.9gr με τυπική απόκλιση 0.1gr. α) Ποιο ποσοστό συσκευασιών αναμένεται να περιέχει ποσότητα δημητριακών μεταξύ 12.7gr και 13.1gr; β) Μια συσκευασία από τις 15 που επελέγησαν για να ελεγχθούν, έχει βάρος 13.21gr και μια άλλη έχει βάρος 12.75gr. Τι μπορούμε να πούμε για τη θέση αυτών των τιμών στην κατανομή του δείγματος; γ) Αν είναι γνωστό ότι η κατανομή των βαρών των συσκευασιών είναι κανονική, πώς απαντάμε στα ερωτήματα (α) και (β);*

<u>*Απάντηση:*</u> α) Το διάστημα $(12.7, 13.1)$ έχει πλάτος $2ks$ με $k=2$, αφού $13.1-12.7=0.4$ και άρα $0.4=2k0.1 \Rightarrow 0.4=k0.2 \Rightarrow k=2$. Έτσι, με βάση την *ανισότητα Chebyshev,* **τουλάχιστον** το 75% των συσκευασιών αναμένεται να περιέχει ποσότητα δημητριακών στο διάστημα $(12.7, 13.1)$.

β) Η *z-τιμή* της τιμής 12.75 είναι $z=(12.75-12.9)/0.1=-1.5$ άρα η τιμή 12.75 βρίσκεται αριστερά του *μέσου* της κατανομής του δείγματος και σε απόσταση ίση με 1.5 φορά την *τυπική απόκλιση* του δείγματος. Πρόκειται, δηλαδή, για μια όχι σπάνια/ακραία τιμή αλλά για μια συχνά εμφανιζόμενη τιμή αφού ανήκει σε ένα διάστημα γύρω από τον *μέσο* της κατανομής στο οποίο ανήκει **τουλάχιστον** το 75% των παρατηρήσεων. Αντίστοιχα, η *z-τιμή* της τιμής 13.21 είναι $z=(13.21-12.9)/0.1=+3.1$ άρα η τιμή 13.21 βρίσκεται δεξιά του *μέσου* της κατανομής του δείγματος και σε απόσταση ίση με 3.1 φορές την *τυπική απόκλιση* του δείγματος. Δηλαδή, η τιμή 13.21 απέχει από τον *μέσο* του δείγματος περισσότερο από 3 *τυπικές αποκλίσεις* και επομένως, σύμφωνα με την ανισότητα του *Chebyshev,* ανήκει στο 11.11%, **το πολύ**, των παρατηρήσεων που απέχουν από τον *μέσο*

περισσότερο από τρεις *τυπικές αποκλίσεις*. Πρόκειται, δηλαδή, για μια τιμή που βρίσκεται μακριά από τον *μέσο* της κατανομής του δείγματος και δεν εμφανίζεται συχνά, είναι «σπάνια/ακραία». Βέβαια, το «σπάνιο/ακραίο» πρέπει να ορίζεται με σαφήνεια/ακρίβεια. Θα το κάνουμε στα επόμενα, στη *στατιστική συμπερασματολογία*.
γ) Αν γνωρίζουμε ότι η κατανομή των βαρών των συσκευασιών είναι *κανονική,* τότε στο διάστημα $(12.7, 13.1)$ αναμένουμε να βρίσκεται το 95% (περίπου) των τιμών του δείγματος. Σε ό,τι αφορά τις θέσεις των τιμών 12.75 και 13.21, αν γνωρίζουμε ότι η κατανομή των βαρών των συσκευασιών είναι *κανονική,* οι απαντήσεις που δώσαμε στο ερώτημα (β) ισχυροποιούνται. Η τιμή 12.75 είναι μια συχνά εμφανιζόμενη τιμή αφού βρίσκεται σε ένα διάστημα γύρω από τη *μέση τιμή* του δείγματος στο οποίο ανήκει το 95% (περίπου) των παρατηρήσεων του δείγματος (αντί «τουλάχιστον 75%» που αναμέναμε σύμφωνα με την *ανισότητα Chebyshev)*. Αντίστοιχα, η τιμή 13.21, τώρα μπορεί ασφαλέστερα να χαρακτηρισθεί «σπάνια/ακραία» αφού γνωρίζουμε ότι ανήκει μόλις στο 0.3% (περίπου) των τιμών που βρίσκονται πέραν των τριών *τυπικών αποκλίσεων* από τον *μέσο*. Και όχι μόνο αυτό. Λόγω συμμετρίας της *κανονικής κατανομής*, η τιμή 13.21 ανήκει σε ακόμη πιο μικρό ποσοστό, στο 0.15% των τιμών που βρίσκονται πέραν των τριών *τυπικών αποκλίσεων* δεξιότερα (προς μεγαλύτερες τιμές) του *μέσου*.

■

Ολοκληρώνουμε την παρουσίαση των *μέτρων θέσης* και *διασποράς*, με τον υπολογισμό των *μέτρων θέσης* και *διασποράς* του δείγματος από τη μεταβλητή *μηνιαίο οικογενειακό εισόδημα* του *Παραδείγματος 9.2*.

Παράδειγμα 9.1.25 (συνέχεια του Παραδείγματος 9.2): Εφόσον έχουμε στη διάθεσή μας τις αρχικές τιμές δηλαδή τα πρωτογενή δεδομένα, θα εργασθούμε με αυτά. Στη συνέχεια θα δούμε πώς υπολογίζονται τα *μέτρα θέσης* και *διασποράς* αν τα δεδομένα δίνονται ομαδοποιημένα και δε μας είναι γνωστά τα αρχικά. Ο *πίνακας συχνοτήτων* των αρχικών δεδομένων είναι ο *Πίνακας 9.1.13*.

y_i	ν_i	f_i	N_i	F_i
1000	1	0.05	1	0.05
1200	3	0.15	4	0.20
1250	1	0.05	5	0.25
1400	4	0.20	9	0.45
1450	2	0.10	11	0.55
1600	4	0.20	15	0.75
1800	3	0.15	18	0.90
2000	2	0.10	20	1.00
	20	**1.00**		

Πίνακας 9.1.13
Ο πίνακας συχνοτήτων του δείγματος από τη μεταβλητή «μηνιαίο οικογενειακό εισόδημα» του Παραδείγματος 9.2 (χωρίς ομαδοποίηση των αρχικών δεδομένων)

Για ευκολία στον υπολογισμό του *δειγματικού μέσου* και της *δειγματικής διακύμανσης* συμπληρώνουμε τον *πίνακα συχνοτήτων* με δύο ακόμη στήλες όπως φαίνεται στον *Πίνακα 9.1.14* (μια με τα γινόμενα $\nu_i y_i$ και μια με τα γινόμενα $\nu_i y_i^2$).

y_i	ν_i	f_i	N_i	F_i	$\nu_i y_i$	$\nu_i y_i^2$
1000	1	0.05	1	0.05	1000	1000000
1200	3	0.15	4	0.20	3600	4320000
1250	1	0.05	5	0.25	1250	1562500
1400	4	0.20	9	0.45	5600	7840000
1450	2	0.10	11	0.55	2900	4205000
1600	4	0.20	15	0.75	6400	10240000
1800	3	0.15	18	0.90	5400	9720000
2000	2	0.10	20	1.00	4000	8000000
Σύνολα	**20**	**1.00**			**30150**	**46887500**

Πίνακας 9.1.14
Υπολογισμός του μέσου και της διακύμανσης του δείγματος από τη μεταβλητή «μηνιαίο οικογενειακό εισόδημα» του Παραδείγματος 9.2 με βάση τα αρχικά δεδομένα (χωρίς ομαδοποίηση)

Δειγματικός μέσος:

$$\bar{x} = \frac{\sum_{i=1}^{k} \nu_i y_i}{\nu} = \frac{30150}{20} = 1507.5\,€.$$

Δειγματική διακύμανση:

$$s^2 = \frac{1}{\nu - 1}\left(\sum_{i=1}^{k} \nu_i y_i^2 - \nu \bar{x}^2 \right) =$$

$$= \frac{1}{19}\left(46887500 - 20 \cdot 1507.5^2\right) = 75598.7\,€^2.$$

Δειγματική τυπική απόκλιση:

$$s = \sqrt{75598.7} = 274.95\,€.$$

Συντελεστής μεταβλητότητας του δείγματος:

$$CV = \frac{s}{|\bar{x}|} \cdot 100\% = \frac{274.95}{1507.5} \cdot 100\% = 18.24\%\,.$$

Η *κορυφή του δείγματος* δεν ορίζεται μονοσήμαντα (υπάρχουν δύο τιμές με τη μεγαλύτερη συχνότητα, η 1400 και η 1600)

Διάμεσος του δείγματος:

$$\delta = \frac{x_{(10)} + x_{(11)}}{2} = \frac{1450 + 1450}{2} = 1450\,€.$$

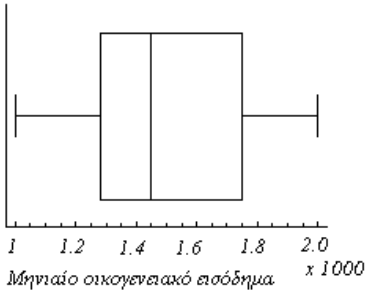
Πρώτο τεταρτημόριο του δείγματος:
Το Q_1 βρίσκεται στη θέση $0.25(\nu + 1) = 0.25 \cdot (20 + 1) = 5.25$, άρα
$Q_1 = x_{(5)} + 0.25(x_{(6)} - x_{(5)}) = 1250 + 0.25 \cdot (1400 - 1250) = 1287.5$.

Τρίτο τεταρτημόριο του δείγματος:
Το Q_3 βρίσκεται στη θέση $0.75(\nu + 1) = 0.75 \cdot (20 + 1) = 15.75$ θέση, άρα
$Q_3 = x_{(15)} + 0.75(x_{(16)} - x_{(15)}) = 1600 + 0.75 \cdot (1800 - 1600) = 1750$.

Ενδοτεταρτημοριακό εύρος του δείγματος:
$Q_3 - Q_1 = 1750 - 1287.5 = 462.5$.

Κατασκευή του θηκογράμματος:

Το *ανώτερο εσωτερικό φράγμα* είναι
$Q_3 + 1.5(Q_3 - Q_1) = 1750 + 1.5 \cdot 462.5 = 2443.75$
άρα το όριο της άνω κεραίας (η *άνω οριακή τιμή*) είναι το 2000 (η μεγαλύτερη παρατήρηση που είναι μικρότερη ή ίση του 2443.75).
Το *κατώτερο εσωτερικό φράγμα* είναι
$Q_1 - 1.5(Q_3 - Q_1) = 1287.5 - 1.5 \cdot 462.5 = 593.75$
άρα το όριο της κάτω κεραίας (η *κάτω οριακή τιμή*) είναι το 1000 (η μικρότερη παρατήρηση που είναι μεγαλύτερη ή ίση του 593.75).
Έτσι το *θηκόγραμμα* του δείγματος είναι αυτό του *Σχήματος 9.1.23.*



Σχήμα 9.1.23
Το θηκόγραμμα της κατανομής του δείγματος από τη μεταβλητή «μηνιαίο οικογενειακό εισόδημα» του Παραδείγματος 9.2

Συμπέρασμα: *Η κατανομή του δείγματος των μηνιαίων οικογενειακών εισοδημάτων φαίνεται να είναι συμμετρική (οι δύο κεραίες έχουν συγκρίσιμα μήκη και η γραμμή που αντιστοιχεί στη διάμεσο δεν αποκλίνει σημαντικά προς κάποια από τις πλευρές του ορθογωνίου). Επίσης, δεν υπάρχουν ακραίες τιμές. Το 50% των μηνιαίων οικογενειακών εισοδημάτων του δείγματος βρίσκεται μεταξύ 1287.5 και 1750€. Ποσοστό 25% των οικογενειών του δείγματος έχει μηνιαίο εισόδημα πάνω από 1750€ και ποσοστό 25%, έχει μηνιαίο εισόδημα κάτω των 1287.5€. Το ποσοστό των οικογενειών του δείγματος που έχουν μηνιαίο εισόδημα μικρότερο από το μέσο μηνιαίο εισόδημα (1507.5€) είναι μεγαλύτερο από 50%.*

Ας δούμε πώς εργαζόμαστε αν δεν έχουμε στη διάθεσή μας τα πρωτογενή δεδομένα αλλά μας έχουν δοθεί ομαδοποιήμενα. Θα χρησιμοποιήσουμε την ομαδοποίηση που είχαμε κάνει για το συγκεκριμένο δείγμα στα προηγούμενα *(Πίνακας 9.1.5)*.

Συμπληρώνουμε τον *πίνακα συχνοτήτων* με μια στήλη στην οποία γράφουμε τις *κεντρικές τιμές* y_i των κλάσεων[10], τις οποίες χρησιμοποιούμε για τον υπολογισμό του *δειγματικού μέσου* και της *δειγματικήςδιακύμανσης*. Θεωρούμε, δηλαδή, την *κεντρική τιμή* κάθε κλάσης ως αντιπροσωπευτική όλων των τιμών που περιέχονται σε αυτή. Επίσης, συμπληρώνουμε τον πίνακα με δύο ακόμη στήλες, μια με τα γινόμενα $\nu_i y_i$ και μια με τα γινόμενα $\nu_i y_i^2$. Προκύπτει έτσι ο *Πίνακας 9.1.15*.

[10] *Ως κεντρική τιμή μιας κλάσης ορίζεται το ημιάθροισμα των άκρων της.*

Εισόδημα	y_i	ν_i	f_i	N_i	F_i	$\nu_i y_i$	$\nu_i y_i^2$
[900 1100)	1000	1	0.05	1	0.05	1000	1000000
[1100 1300)	1200	4	0.20	5	0.25	4800	5760000
[1300 1500)	1400	6	0.30	11	0.55	8400	11760000
[1500 1700)	1600	4	0.20	15	0.75	6400	10240000
[1700 1900)	1800	3	0.15	18	0.90	5400	9720000
[1900 2100)	2000	2	0.10	20	1.00	4000	8000000
Σύνολα		**20**	**1.00**			**30000**	**46480000**

Πίνακας 9.1.15
Ο πίνακας συχνοτήτων των τιμών του δείγματος από τη μεταβλητή «μηνιαίο οικογενειακό εισόδημα» του Παραδείγματος 9.2 ομαδοποιημένων σε 6 κλάσεις

Δειγματικός μέσος:

$$\bar{x} = \frac{\sum_{i=1}^{k} \nu_i y_i}{\nu} = \frac{30000}{20} = 1500 €.$$

Δειγματική διακύμανση:

$$s^2 = \frac{1}{\nu - 1}\left(\sum_{i=1}^{k} \nu_i y_i^2 - \nu \bar{x}^2 \right) = \frac{1}{19}\left(46480000 - 20 \cdot 1500^2\right) = 77894.74 €^2.$$

Δειγματική τυπική απόκλιση:

$$s = \sqrt{77894.74} = 279.1 €.$$

Συντελεστής μεταβλητότητας του δείγματος:

$$CV = \frac{s}{|\bar{x}|} \cdot 100\% = \frac{279.1}{1500} \cdot 100\% = 18.6\% .$$

Η *επικρατούσα κλάση* είναι προφανώς η κλάση [1300, 1500), επομένως για να προσδιορίσουμε την *κορυφή του δείγματος* μπορούμε, σύμφωνα με όσα έχουμε αναφέρει, να κάνουμε παρεμβολή και ως *κορυφή* του δείγματος να θεωρήσουμε την τιμή

$$M_0 = 1300 + \frac{6-4}{(6-4)+(6-4)} \cdot 200 = 1400 €.$$

Διάμεσος του δείγματος:
Έχει υπολογισθεί στο *Παράδειγμα 9.1.13* και βρέθηκε $\delta = 1466.7$.

Πρώτο τεταρτημόριο του δείγματος:
Το Q_1 βρίσκεται στην κλάση [1100, 1300) γιατί όπως φαίνεται στη στήλη των *αθροιστικών σχετικών συχνοτήτων* του *πίνακα συχνοτήτων*, σε αυτή την κλάση βρίσκεται η τιμή με *αθροιστική σχετική συχνότητα* 0.25. Επομένως,

$$Q_1 = x_{0.25} = L_i + \frac{0.25\nu - N_{i-1}}{\nu_i} \cdot c_i = 1100 + \frac{0.25 \cdot 20 - 1}{4} \cdot 200 = 1300 .$$

Τρίτο τεταρτημόριο του δείγματος:
Το Q_3 βρίσκεται στην κλάση [1500, 1700) γιατί όπως φαίνεται στη στήλη των *αθροιστικών σχετικών συχνοτήτων* του *πίνακα συχνοτήτων*, σε αυτή την κλάση βρίσκεται η τιμή με *αθροιστική σχετική συχνότητα* 0.75. Επομένως

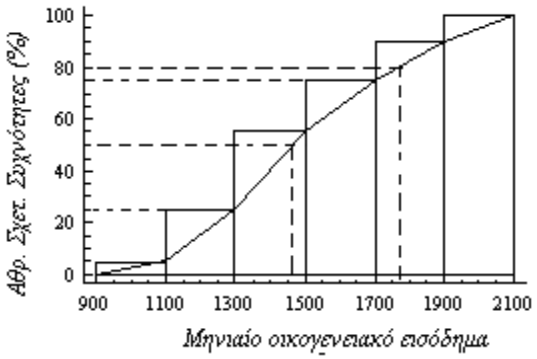
$$Q_3 = x_{0.75} = L_i + \frac{0.75\nu - N_{i-1}}{\nu_i} \cdot c_i = 1500 + \frac{0.75 \cdot 20 - 11}{4} \cdot 200 = 1700 .$$

Ενδοτεταρτημοριακό εύρος του δείγματος:
$Q_3 - Q_1 = 1700 - 1300 = 400$.

Τέλος, ας υπολογίσουμε ένα ακόμη *p-ποσοστιαίο σημείο*, το $x_{0.8}$. Βρίσκεται στην κλάση [1700, 1900) γιατί όπως φαίνεται στη στήλη των *αθροιστικών σχετικών συχνοτήτων* του *πίνακα συχνοτήτων*, σε αυτή την κλάση βρίσκεται η τιμή με *αθροιστική σχετική συχνότητα* 0.8. Επομένως

$$x_{0.8} = L_i + \frac{0.8\nu - N_{i-1}}{\nu_i} \cdot c_i = 1700 + \frac{0.8 \cdot 20 - 15}{3} \cdot 200 = 1766.7$$

Τις τιμές των *p-ποσοστιαίων σημείων,* όπως φαίνεται στο *Σχήμα 9.1.24,* μπορούμε να τις υπολογίσουμε και από το *πολύγωνο αθροιστικών σχετικών συχνοτήτων*.



Σχήμα 9.1.24
Υπολογισμός p-ποσοστιαίων σημείων ($p = 0.25, 0.50, 0.75, 0.80$)
από το πολύγωνο αθροιστικών σχετικών συχνοτήτων

■

Παρατήρηση 9.1.12: *Οι τιμές όλων των στατιστικών που υπολογίσθηκαν από τα ομαδοποιημένα δεδομένα διαφέρουν, προφανώς λόγω της ομαδοποίησης, από τις τιμές των αντίστοιχων στατιστικών που υπολογίσθηκαν από τα πρωτογενή δεδομένα. Γι' αυτό, όταν τα πρωτογενή δεδομένα είναι διαθέσιμα, ο υπολογισμός των στατιστικών πρέπει να γίνεται από αυτά, και ανεξάρτητα από την ομαδοποίηση που θα κάνουμε για την κατασκευή ιστογραμμάτων κτλ.*

■

9.1.3.3 Μέτρα λοξότητας και μέτρα κύρτωσης

Τα *μέτρα λοξότητας ή ασυμμετρίας* και τα *μέτρα κύρτωσης* ορίσθηκαν για να περιγράψουν τη μορφή της κατανομής του δείγματος. Τις έννοιες που συνδέονται με τη μορφή της κατανομής του δείγματος *(συμμετρία, θετική και αρνητική ασυμμετρία* και *κύρτωση)*, τις έχουμε εξηγήσει και μάλιστα είδαμε πώς μπορούμε, με βάση αυτές, να περιγράψουμε τη μορφή της κατανομής του δείγματος ερμηνεύοντας και συνδυάζοντας, με τη βοήθεια του *θηκογράμματος* αλλά και του *ιστογράμματος*, τις πληροφορίες που παίρνουμε από τα *στατιστικά θέσης* και *διασποράς*. Δε θα παρουσιάσουμε αναλυτικά όλα τα μέτρα *λοξότητας* και *κύρτωσης*. Θα σημειώσουμε μόνο ότι στη βιβλιογραφία συναντάμε αρκετά τέτοια μέτρα και ότι όλα, με κάποιο τρόπο, ποσοτικοποιούν την πληροφορία που παίρνουμε αν συνδυάσουμε, κατά περίπτωση, διάφορα *στατιστικά θέσης και διασποράς* που ήδη χρησιμοποιήσαμε για το σκοπό αυτό, όπως τον *μέσο*, τα *τεταρτημόρια*, τη *διάμεσο*, την *κορυφή*, την *τυπική απόκλιση* αλλά και *κεντρικές ροπές*.

Για παράδειγμα, οι ***συντελεστές ασυμμετρίας του Pearson***

$$\gamma_1 = \frac{\bar{x} - M_0}{s} \text{ και } \gamma_2 = \frac{3(\bar{x} - \delta)}{s},$$

ποσοτικοποιούν με ένα συγκεκριμένο τρόπο τη σχέση μεταξύ $\bar{x},\ M_0, s$ και $\bar{x},\ \delta, s,$ αντίστοιχα.

Έτσι, αν $\gamma_1 = \gamma_2 = 0$, η κατανομή είναι *συμμετρική* (γιατί τότε $\bar{x} = \delta = M_0$).

Επίσης, ο ***συντελεστής ασυμμετρίας του Bowley***

$$S_A = \frac{Q_1 + Q_3 - 2Q_2}{Q_3 - Q_1} = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{Q_3 - Q_1},$$

ποσοτικοποιεί τη σχέση των $Q_3 - Q_2$ και $Q_2 - Q_1$ με το *ενδοτεταρτημοριακό εύρος* $Q_3 - Q_1$.

Έτσι, αν $S_A = 0$, η κατανομή είναι *συμμετρική,* ενώ αν ο S_A έχει τιμή μεταξύ 0 και +1, η κατανομή παρουσιάζει *θετική ασυμμετρία* και αν έχει τιμή μεταξύ -1 και 0, η κατανομή παρουσιάζει *αρνητική ασυμμετρία (θυμηθείτε ότι για να ελέγξουμε τη συμμετρία από το θηκόγραμμα, εκτός από τα μήκη των κεραιών, παρατηρούμε και αν η γραμμή που αναπαριστά τη διάμεσο βρίσκεται στη μέση του ορθογωνίου ή αποκλίνει προς τις πλευρές του που αναπαριστούν τα Q_1 και Q_3).*

Από τα *μέτρα κύρτωσης* αναφέρουμε τον ***ποσοστημοριακό συντελεστή κύρτωσης***

$$k = \frac{Q_3 - Q_1}{x_{0.9} - x_{0.1}},$$

ο οποίος χρησιμοποιεί το *ενδοτεταρτημοριακό* και το *ενδοδεκατημοριακό εύρος,* $Q_3 - Q_1$ και $x_{0.9} - x_{0.1}$, αντίστοιχα.

Αναφέρουμε επίσης, το ***συντελεστή κύρτωσης του Pearson***

$$\beta_2 = \frac{m_4}{(m_2)^2},$$

που ορίζεται με βάση τις *δειγματικές κεντρικές ροπές*

$$m_4 = \frac{\sum_{i=1}^{\nu}(x_i - \bar{x})^4}{\nu} \text{ και } m_2 = \frac{\sum_{i=1}^{\nu}(x_i - \bar{x})^2}{\nu}.$$

Για *κανονικές* κατανομές είναι $\beta_2 = 3$ *(μεσόκυρτη)*, ενώ αν $\beta_2 > 3$ η κατανομή χαρακτηρίζεται *πλατύκυρτη* και αν $\beta_2 < 3$, χαρακτηρίζεται *λεπτόκυρτη*.

Με βάση *δειγματικές κεντρικές ροπές* (τις $m_2,\ m_3$) υπολογίζονται και οι ***συντελεστές ασυμμετρίας***

$$\beta_1 = \frac{(m_3)^2}{(m_2)^3} \text{ και } \alpha_3 = \frac{m_3}{(m_2)^{3/2}}.$$

Αν $\beta_1 = 0$ η κατανομή είναι *συμμετρική*, όμως αν $\beta_1 \neq 0$, ο β_1 δε μπορεί να καθορίσει το είδος της ασυμμετρίας, κάτι που συμβαίνει με τον α_3.

9.2 Ποιοτικές Μεταβλητές

Υπενθυμίζουμε ότι ***ποιοτικές (qualitative)*** είναι οι μεταβλητές που δεν είναι *ποσοτικές*, δηλαδή, που δεν παίρνουν αριθμητικές τιμές και διακρίνονται σε ***κατηγορίας (categorical/nominal)*** και ***διάταξης (ordinal)***. Στο *Παράδειγμα 9.2, η* τυχαία μεταβλητή *X (επάγγελμα πατέρα)* προφανώς είναι *ποιοτική κατηγορίας,* ενώ η τυχαία μεταβλητή *Y (επίπεδο εκπαίδευσης πατέρα)* είναι *ποιοτική διάταξης (θυμηθείτε όσα αναφέραμε για τις κλίμακες μέτρησης στο εισαγωγικό 1^{ο} Κεφάλαιο)*.

Για *ποιοτικές μεταβλητές,* η *Περιγραφική Στατιστική* προσφέρει τη δυνατότητα κατασκευής ***πίνακα κατανομής συχνοτήτων***, ***ραβδογράμματος*** και ***κυκλικού διαγράμματος***. Από τα *αριθμητικά περιγραφικά μέτρα*, στις *ποιοτικές μεταβλητές,* ορίζεται (έχει νόημα) μόνο η ***κορυφή/επικρατούσα τιμή*** της κατανομής.

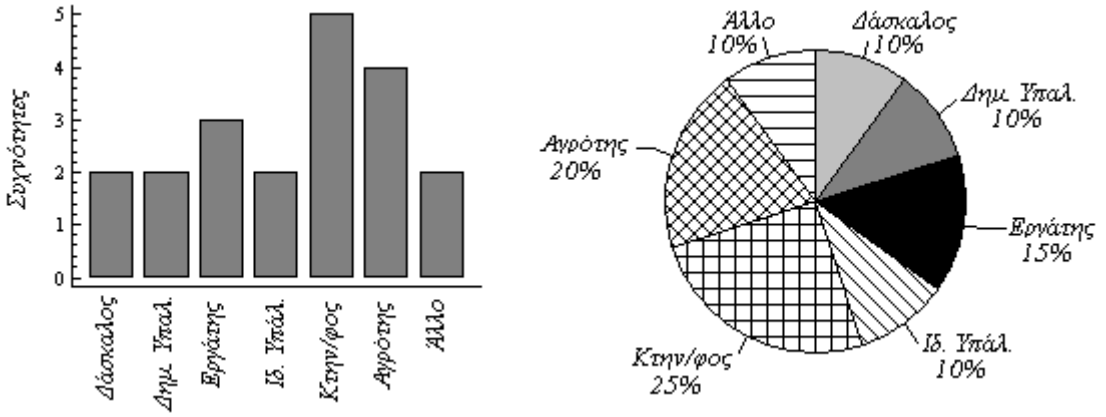
Παράδειγμα 9.2.1 (συνέχεια του Παραδείγματος 9.2): *Ο Πίνακας 9.2.1 είναι ο πίνακας κατανομής συχνοτήτων* του *τυχαίου δείγματος από την ποιοτική μεταβλητή* ***κατηγορίας*** *«επάγγελμα πατέρα» του Παραδείγματος 9.2*

y_i	ν_i	f_i
Δάσκαλος	2	0.10
Δημ. Υπάλληλος	2	0.10
Εργάτης	3	0.15
Ιδιωτ. Υπάλληλος	2	0.10
Κτηνοτρόφος	5	0.25
Αγρότης	4	0.20
Άλλο	2	0.10
Σύνολα	**20**	**1.00**

Πίνακας 9.2.1
Ο πίνακας συχνοτήτων του δείγματος από την τυχαία μεταβλητή «επάγγελμα πατέρα» του Παραδείγματος 9.2

Στην πρώτη στήλη, οι διαφορετικές τιμές y_i της μεταβλητής που εμφανίσθηκαν στο δείγμα, δεν έχουν καταγραφεί με κάποια σειρά γιατί δεν ορίζεται (δεν έχει νόημα) κάποιου είδους διάταξη. Επίσης, δεν υπάρχουν στήλες με *αθροιστικές* και με *αθροιστικές σχετικές συχνότητες* αφού στις *ποιοτικές μεταβλητές κατηγορίας* δεν έχουν νόημα. Τι νόημα έχει και πώς μπορούμε, για παράδειγμα, να προσδιορίσουμε πόσες τιμές του δείγματος είναι μικρότερες ή ίσες από την τιμή δάσκαλος ή μεγαλύτερες από την τιμή κτηνοτρόφος. Όμως, οι *συχνότητες* και οι *σχετικές συχνότητες* των διαφορετικών τιμών y_i, έχουν νόημα και μάλιστα ταυτόσημο με αυτό που έχουν στις *ποσοτικές* μεταβλητές.

Επίσης, τα *ραβδογράμματα* και τα *κυκλικά διάγραμμα* κατασκευάζονται και ερμηνεύονται όπως στις *ποσοτικές* μεταβλητές. Δείτε στα *Σχήματα 9.2.1* το *ραβδόγραμμα συχνοτήτων* και το *κυκλικό διάγραμμα σχετικών συχνοτήτων* της κατανομής του δείγματος. Παρατηρείστε ότι 3 πατεράδες έχουν επάγγελμα *εργάτης* και αποτελούν το 15% του δείγματος, η τιμή αγρότης εμφανίσθηκε 4 φορές, κτλ. Επίσης, η κορυφή της κατανομής του δείγματος είναι η τιμή *κτηνοτρόφος* με *συχνότητα* 5 και σχετική συχνότητα 0.25 ή 25%.



(α)
Ραβδόγραμμα συχνοτήτων

(β)
Κυκλικό διάγραμμα σχετικών συχνοτήτων

Σχήματα 9.2.1
Γραφική αναπαράσταση της κατανομής του δείγματος από τη μεταβλητή «επάγγελμα πατέρα» του Παραδείγματος 9.2

■

Παράδειγμα 9.2.2 (συνέχεια του Παραδείγματος 9.2): *Ο Πίνακας 9.2.2 είναι ο πίνακας συχνοτήτων του δείγματος από την ποιοτική μεταβλητή* ***διάταξης*** *«επίπεδο εκπαίδευσης πατέρα» του Παραδείγματος 9.2.*

y_i	ν_i	f_i	N_i	F_i
1	3	0.15	3	0.15
2	11	0.55	14	0.70
3	4	0.20	18	0.90
4	2	0.10	20	1.00
Σύνολα	**20**	**1.00**		

Πίνακας 9.2.2
Ο πίνακας συχνοτήτων του δείγματος από την τυχαία μεταβλητή «επίπεδο εκπαίδευσης πατέρα» του Παραδείγματος 9.2

Στην πρώτη στήλη του *πίνακα συχνοτήτων* οι διαφορετικές τιμές y_i της μεταβλητής που εμφανίσθηκαν στο δείγμα καταγράφηκαν σε αύξουσα σειρά αφού μπορούν να διαταχθούν. Σε *ποιοτικές μεταβλητές διάταξης* έχουν επίσης νόημα οι *αθροιστικές* και οι *αθροιστικές σχετικές συχνότητες*. Για παράδειγμα, έχει νόημα να πούμε ότι επίπεδο εκπαίδευσης μέχρι και *δευτεροβάθμια εκπαίδευση* έχουν 14 πατεράδες. Η *κορυφή* της κατανομής του δείγματος είναι η τιμή 2 (*δευτεροβάθμια εκπαίδευση*) με *σχετική συχνότητα* 55%. Μπορούμε, τέλος, να κατασκευάσουμε *ραβδογράμματα* και *κυκλικά διαγράμματα* της κατανομής του δείγματος (δείτε το ως άσκηση).

■

Σε μια έρευνα, οι δειγματοληπτικές/πειραματικές μονάδες μπορεί να ταξινομούνται όχι μόνο ως προς ένα χαρακτηριστικό αλλά και ως προς ένα δεύτερο. Σε αυτές τις περιπτώσεις ο πίνακας συχνοτήτων είναι διδιάστατος και ονομάζεται *πίνακας συνάφειας* (θυμηθείτε και όσα αναφέραμε στο *Παράδειγμα 4.5.3*). Η γραφική αναπαράσταση μιας τέτοιας κατανομής συχνοτήτων (διδιάστατης) γίνεται με ***πολλαπλά ραβδογράμματα (multiple barcharts)*** και με ***διαγράμματα mosaic (mosaic plots)***. Ας δούμε ένα παράδειγμα.

Παράδειγμα 9.2.3 (συνέχεια του Παραδείγματος 4.5.3): *Ο Πίνακας 9.2.3 είναι ο πίνακας δεδομένων του Παραδείγματος 4.5.3.*

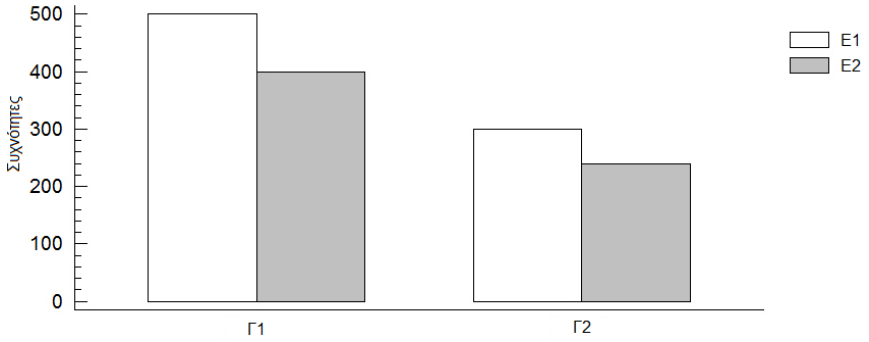
		Γραμμή παραγωγής	
		Γ1	*Γ2*
Σοβαρότητα ελαττώματος	*Ε1*	500 0.3472	300 0.2083
	Ε2	400 0.2778	240 0.1667

Πίνακας 9.2.3

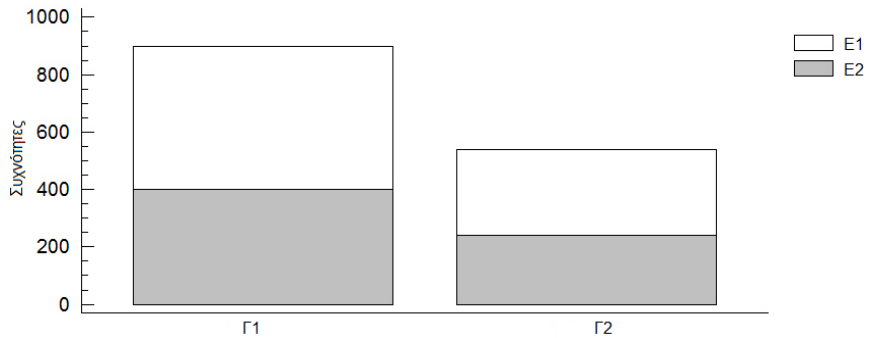
Η κατανομή 1440 ελαττωματικών προϊόντων ως προς δύο χαρακτηριστικά: γραμμή παραγωγής (Γ1 ή Γ2) και σοβαρότητα ελαττώματος (Ε1 ή Ε2).

Πρόκειται για έναν διδιάστατο πίνακα συχνοτήτων αφού δίνει τη συχνότητα που παρατηρήθηκε για κάθε συνδυασμό γραμμής παραγωγής *(Γ1 ή Γ2)* και σοβαρότητας ελαττώματος *(Ε1 ή Ε2)*,. Δίνεται επίσης και η αντίστοιχη σχετική συχνότητα. Παρατηρείστε, για παράδειγμα, ότι τα 500 προϊόντα που προέρχονται από τη γραμμή παραγωγής Γ1 και κατατάσσονται στην κατηγορία ελαττώματος Ε1 αποτελούν το 34.72% των 1440 ελαττωματικών προϊόντων που εξετάσθηκαν. Επίσης, ποσοστό 16.67% των 1440 ελαττωματικών προϊόντων προέρχονται από τη γραμμή παραγωγής Γ2 και κατατάσσονται στην κατηγορία ελαττώματος Ε2.

Στα *Σχήματα 9.2.3&9.2.4* η διδιάστατη αυτή κατανομή παρουσιάζεται γραφικά σε μορφή *πολλαπλού ραβδογράμματος*. Συγκεκριμένα, στο *Σχήμα 9.2.3(α)* παρουσιάζεται ως *πολλαπλό ομαδοποιημένο ραβδόγραμμα* της γραμμής παραγωγής δοθείσης της σοβαρότητας του ελαττώματος και στο *Σχήμα 9.2.4(α)* ως *πολλαπλό ομαδοποιημένο ραβδόγραμμα* της σοβαρότητας του ελαττώματος δοθείσης της γραμμής παραγωγής. Αντίστοιχα, στα *Σχήματα 9.2.3(β)* και *9.2.4(β)* η κατανομή παρουσιάζεται ως *πολλαπλό ραβδόγραμμα στοίβας*. Σημειώνουμε ότι αν για την κατασκευή των ραβδογραμμάτων αντί για τις συχνότητες χρησιμοποιήσουμε τις σχετικές συχνότητες, η εικόνα που παίρνουμε για την κατανομή ασφαλώς δεν αλλάζει.



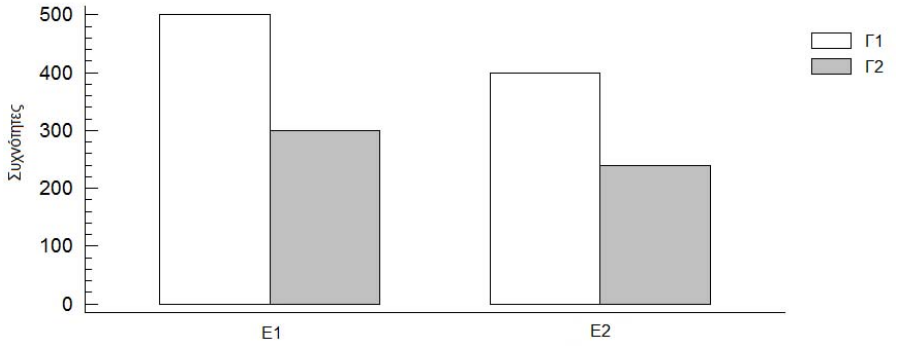
(α)
Ομαδοποιημένο



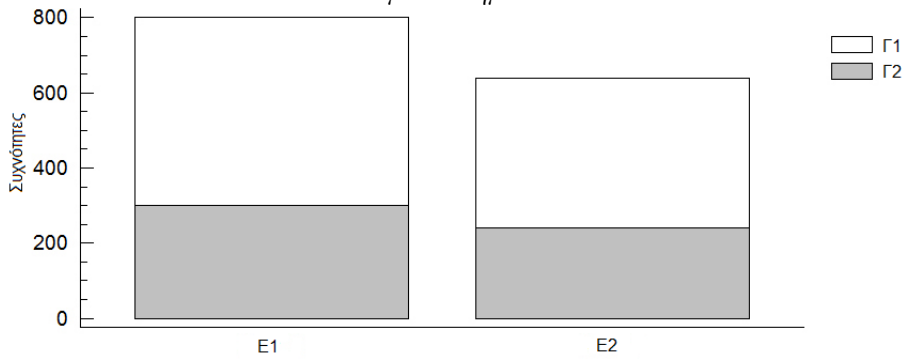
(β)
Στοίβας

Σχήματα 9.2.3
Πολλαπλό ραβδόγραμμα της γραμμής παραγωγής

δοθείσης της σοβαρότητας του ελαττώματος



(α)
Ομαδοποιημένο



(β)
Στοίβας

Σχήματα 9.2.4
Πολλαπλό ραβδόγραμμα της σοβαρότητας του ελαττώματος δοθείσης της γραμμής παραγωγής

Δείτε επίσης τους *Πίνακες 9.2.5&9.2.6.* Στον *Πίνακα 9.2.5* οι συχνότητες εκφράζονται ως ποσοστό της συνολικής συχνότητας ανά γραμμή και στον *Πίνακα 9.2.6* ως ποσοστό της συνολικής συχνότητας ανά στήλη. Δηλαδή, ο *Πίνακας 9.2.5* δίνει τη δεσμευμένη κατανομή της γραμμής παραγωγής δοθείσης της σοβαρότητας του ελαττώματος και ο *Πίνακας 9.2.6* τη δεσμευμένη κατανομή της σοβαρότητας του ελαττώματος δοθείσης της γραμμής παραγωγής (ξαναδείτε και το *Παράδειγμα 4.5.3*). Έτσι, από τον *Πίνακα 9.2.5* άμεσα παίρνουμε ότι από τα ελαττωματικά προϊόντα που κατατάσσονται στην κατηγορία ελαττώματος E1, ποσοστό 62.5% προέρχεται από τη γραμμή παραγωγής Γ1 και ποσοστό 37.5% από την γραμμή παραγωγής Γ2. Παρατηρείστε ότι τα ίδια ποσοστά εμφανίζονται αντίστοιχα και στην κατηγορία E2. Ανάλογα, από τον *Πίνακα 9.2.6* έχουμε ότι από τα ελαττωματικά προϊόντα που προέρχονται από τη γραμμή παραγωγής Γ1, ποσοστό 55.6% κατατάσσεται στην κατηγορία ελαττώματος E1 και ποσοστό 44.44% στην κατηγορία ελαττώματος E2. Τα ίδια αντίστοιχα ποσοστά εμφανίζονται και στα προϊόνται που προέρχονται από τη γραμμή παραγωγής Γ2.

		Γραμμή παραγωγής		
		Γ1	*Γ2*	
Σοβαρότητα ελαττώματος	*E1*	500 0.6250	300 0.3750	800 1.0000
	E2	400 0.6250	240 0.3750	640 1.0000

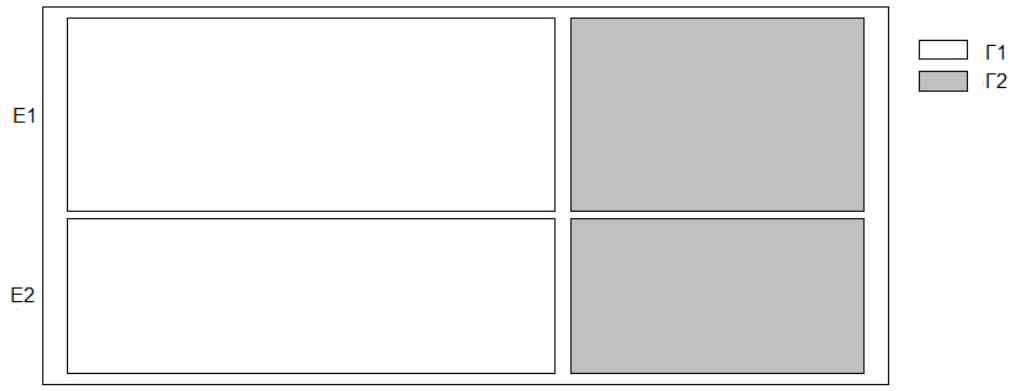
Πίνακας 9.2.5
Η δεσμευμένη κατανομή της γραμμής παραγωγής

δοθείσης της σοβαρότητας του ελαττώματος.

		Γραμμή παραγωγής	
		Γ1	*Γ2*
Σοβαρότητα ελαττώματος	*Ε1*	500 0.5556	300 0.5556
	Ε2	400 0.4444	240 0.4444
		900 1.0000	540 1.000

Πίνακας 9.2.6
Η δεσμευμένη κατανομή της σοβαρότητας του ελαττώματος δοθείσης της γραμμής παραγωγής.

Δείτε στα *διαγράμματα Mosaic* που φαίνονται στα *Σχήματα 9.2.5(α)&(β)* πώς με έναν απλό και έξυπνο τρόπο αναπαρίστανται γραφικά οι πληροφορίες που παίρνουμε από τους *Πίνακες 9.2.5&9.2.6* αντίστοιχα. Παρατηρείστε επίσης πώς (με προφανή τρόπο) φαίνεται γραφικά ότι τα δύο χαρακτηριστικά είναι ανεξάρτητα (δείτε επίσης το *Σχόλιο 14.2.2* και την *Άσκηση 14.24*).



(α)

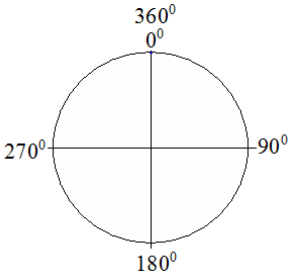
(β)

Σχήματα 9.2.5
Διαγράμματα Mosaic για τη διδιάστατη κατανομή του Παραδείγματος 9.2.3

■

9.3 Μεταβλητές διεύθυνσης και κατεύθυνσης (κυκλικά δεδομένα)

Όπως ήδη έχουμε αναφέρει στο εισαγωγικό κεφάλαιο *(1ο Κεφάλαιο)*, οι μεταβλητές που εκφράζουν χαρακτηριστικά *διεύθυνσης* ή *κατεύθυνσης* μετρώνται σε *κυκλική κλίμακα*. Ένας κύκλος διαιρείται σε 360 ίσα μέρη (μοίρες). Ως μονάδα μέτρησης ορίζεται η μία μοίρα (1^0). Δηλαδή, στις *μεταβλητές κατεύθυνσης* ή *διεύθυνσης* αποδίδονται τιμές γωνιών σε μοίρες[11]. Στη *Γεωλογία*, τη *Μετεωρολογία* και σε άλλες επιστήμες έχει καθιερωθεί οι 0^0 να ορίζονται στο θετικό ημιάξονα Oy, δηλαδή στο Βορρά και οι γωνίες να μετρώνται από το Βορρά και κατά τη φορά των δεικτών του ωρολογίου, δηλαδή η θετική φορά ορίζεται ως η φορά των δεικτών του ωρολογίου. Έτσι, οι 360^0 αντιστοιχίζονται επίσης στο Βορρά, οι 90^0 στην Ανατολή, οι 180^0 στο Νότο και οι 270^0 στη Δύση *(Σχήμα 9.3.1)*.



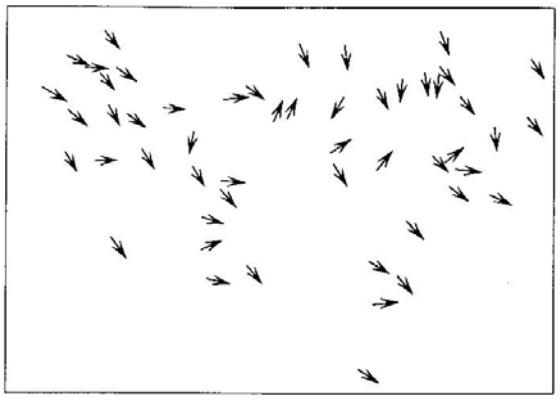
Σχήμα 9.3.1
Οι 0^0 ορίζονται στο Βορρά και οι γωνίες μετρώνται
από το Βορρά και κατά τη φορά της κίνησης *των δεικτών του ωρολογίου*

Όπως δείξαμε με αντιπαραδείγματα *(1ο Κεφάλαιο)*, οι μέθοδοι παρουσίασης, περιγραφής και ανάλυσης *κυκλικών δεδομένων*, διαφέρουν από τις αντίστοιχες που εφαρμόζονται σε δεδομένα *κλίμακας διαστήματος* ή *κλίμακας αναλογίας* (παρότι, ως έννοιες, είναι ανάλογες).

Ας δούμε, μέσω συγκεκριμένων παραδειγμάτων, ποιες μέθοδοι χρησιμοποιούνται για τη γραφική αναπαράσταση *κυκλικών δεδομένων* και πώς ορίζονται και υπολογίζονται τα *αριθμητικά περιγραφικά μέτρα* της κατανομής τους.

9.3.1 Γραφική παρουσίαση κατανομής συχνοτήτων κυκλικών δεδομένων

***Παράδειγμα 9.3.1*:** *Στον Πίνακα 9.3.1 δίνονται 51 τιμές της τυχαίας μεταβλητής που εκφράζει την κατεύθυνση του ίχνους της κίνησης των πάγων (glacial striations) σε μια έκταση 35 Km^2 στη νότια Φινλανδία (Davis, J.C., 2002).*



[11] *Οι γωνίες μετρώνται και σε ακτίνια (rads). Ένα ακτίνιο ορίζεται ως μια επίκεντρη γωνία που βαίνει σε τόξο ίσο με την ακτίνα r του κύκλου. Επειδή ο κύκλος έχει περιφέρεια ίση με $2\pi r$, οι 360^0 αντιστοιχούν σε 2π ακτίνια και επομένως ένα ακτίνιο ισούται με $360^0/2\pi = 180^0/\pi$ μοίρες, δηλαδή, περίπου με 57^0.*

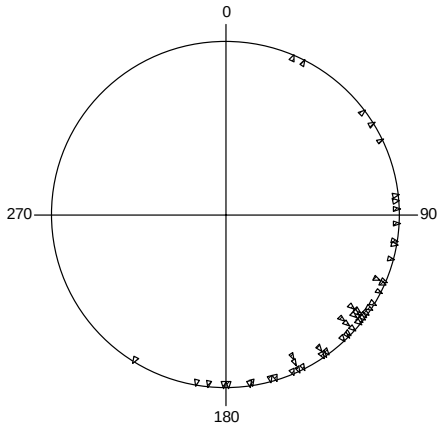
23	93	121	128	137	155	186
27	99	123	128	144	157	190
53	100	125	129	145	163	212
58	105	126	132	145	165	
64	113	126	132	146	171	
83	113	126	132	153	172	
85	114	127	134	155	179	
88	117	127	135	155	181	

Πίνακας 9.3.1
51 τιμές της τυχαίας μεταβλητής που εκφράζει
την κατεύθυνση του ίχνους της κίνησης των πάγων στη νότια Φινλανδία
(σε μοίρες από το Βορρά και κατά τη φορά της κίνησης *των δεικτών του ωρολογίου)*

Πρόκειται για ***δεδομένα κατεύθυνσης***. Δηλαδή, τα δεδομένα αυτά ορίζουν και διεύθυνση και φορά. Για παράδειγμα, οι 53^0 και οι 233^0 ορίζουν την ίδια διεύθυνση 53^0-233^0 αλλά ταυτόχρονα ορίζουν και δύο αντίθετες κατευθύνσεις: την κατεύθυνση 53^0 και την κατεύθυνση 233^0.

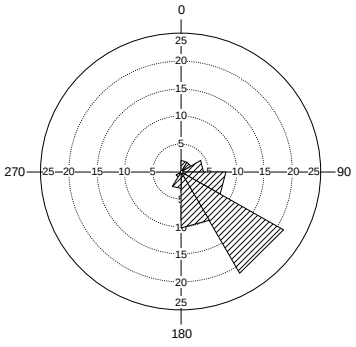
Ας δούμε πώς μπορεί να γίνει η γραφική παρουσίαση αυτών των δεδομένων.

α) *Κυκλικό διάγραμμα διασποράς*
Στο *Σχήμα 9.3.2* φαίνεται το ***κυκλικό διάγραμμα διασποράς*** των δεδομένων του παραδείγματός μας. Πρόκειται για τον πιο απλό και προφανή τρόπο γραφικής παρουσίασης κυκλικών δεδομένων.



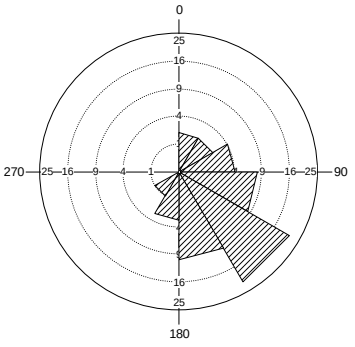
Σχήμα 9.3.2
Το κυκλικό διάγραμμα διασποράς του δείγματος από τη μεταβλητή που εκφράζει
την κατεύθυνση του ίχνους της κίνησης των πάγων στη νότια Φινλανδία

β) *Ροδόγραμμα*
Το ***ροδόγραμμα (rose diagram)*** είναι αντίστοιχο του γραμμικού *ιστογράμματος*. Τα δεδομένα ομαδοποιούνται σε κλάσεις και αντίστοιχα ο κύκλος διαιρείται σε κυκλικούς τομείς. Δηλαδή, αν για παράδειγμα, ως πλάτος της κλάσης επιλεγούν οι 30^0, ο κύκλος διαιρείται σε 12 τομείς των 30^0. Η *συχνότητα* κάθε κλάσης αναπαρίσταται είτε με την ακτίνα *(Σχήμα 9.3.3α)* είτε με το εμβαδόν *(Σχήμα 9.3.3β)* του αντίστοιχου κυκλικού τομέα.



(α)

Η συχνότητα αναπαρίσταται με την ακτίνα του αντίστοιχου κυκλικού τομέα



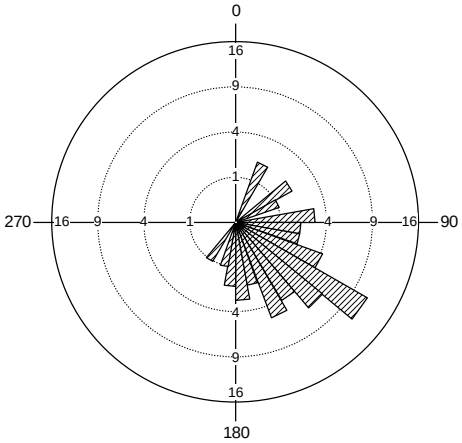
(β)

Η συχνότητα αναπαρίσταται με το εμβαδόν του αντίστοιχου κυκλικού τομέα

Σχήμα 9.3.3

Ροδογράμματα σε τομείς 30^0 του δείγματος από τη μεταβλητή που εκφράζει την κατεύθυνση του ίχνους της κίνησης των πάγων στη νότια Φινλανδία

Επειδή η οπτική εντύπωση που δημιουργεί ένας κυκλικός τομέας διαμορφώνεται πρωτίστως από το εμβαδόν του και δευτερευόντως από την ακτίνα του, το ***ροδόγραμμα*** στο *Σχήμα 9.3.3α* μπορεί να παρασύρει σε λάθος συμπεράσματα αφού υπερτονίζει (οπτικά) τις μεγάλες *συχνότητες* και υποβαθμίζει τις μικρές. Έτσι, μπορεί να δημιουργηθεί η εντύπωση ότι κάποια κατεύθυνση «ξεχωρίζει» ιδιαίτερα έναντι των άλλων, ενώ τα δεδομένα μπορεί να μην υποστηρίζουν ένα τέτοιο συμπέρασμα. Για το λόγο αυτό, στη βιβλιογραφία προτείνεται οι *συχνότητες* (ή οι *σχετικές συχνότητες*) των κλάσεων να αναπαρίστανται με τα εμβαδά και όχι με τις ακτίνες των αντίστοιχων τομέων. Δηλαδή, η ακτίνα κάθε τομέα προτείνεται να είναι ανάλογη με την τετραγωνική ρίζα της αντίστοιχης συχνότητας και όχι με την αντίστοιχη συχνότητα. Είναι προφανές ότι στο ιστόγραμμα *μη κυκλικών δεδομένων* δε δημιουργείται ανάλογο πρόβλημα. Είναι επίσης προφανές ότι το ***ροδόγραμμα***, όπως και το *ιστόγραμμα* μη κυκλικών δεδομένων, επηρεάζεται δραστικά από το πλάτος των κλάσεων *(συγκρίνετε το **ροδόγραμμα** στο Σχήμα 9.3.4 που σχεδιάσθηκε σε τομείς 10^0 με το **ροδόγραμμα** στο Σχήμα 9.3.3β που σχεδιάσθηκε σε τομείς 30^0).*



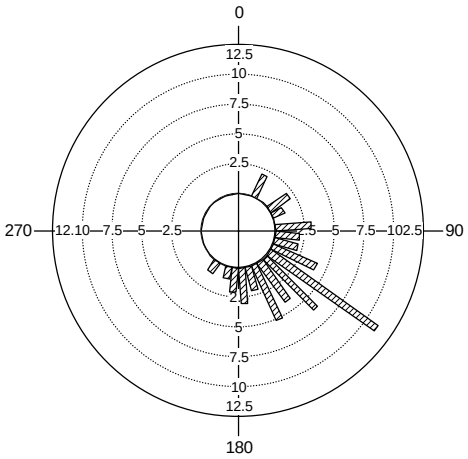
Σχήμα 9.3.4

Το ροδόγραμμα σε τομείς 10^0 του δείγματος από τη μεταβλητή που εκφράζει την κατεύθυνση του ίχνους της κίνησης των πάγων στη νότια Φινλανδία

γ) *Κυκλικό ιστόγραμμα*

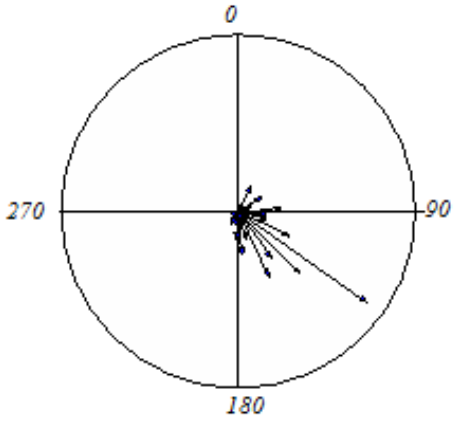
Στο ***κυκλικό ιστόγραμμα (circular histogram)*** οι *συχνότητες* ή οι *σχετικές συχνότητες* αναπαρίστανται με ράβδους αντίστοιχου μήκους που σχεδιάζονται από την περιφέρεια ενός κύκλου. Το *κυκλικό ιστόγραμμα* στο *Σχήμα 9.3.5* είναι το αντίστοιχο

του *ροδογράμματος* του *Σχήματος 9.3.4*. Τα *κυκλικά ιστογράμματα* χρησιμοποιούνται ευρέως για τη γραφική παρουσίαση της κατεύθυνσης ανέμων.



Σχήμα 9.3.5
Κυκλικό ιστόγραμμα του δείγματος από τη μεταβλητή που εκφράζει την κατεύθυνση του ίχνους της κίνησης των πάγων στη νότια Φινλανδία

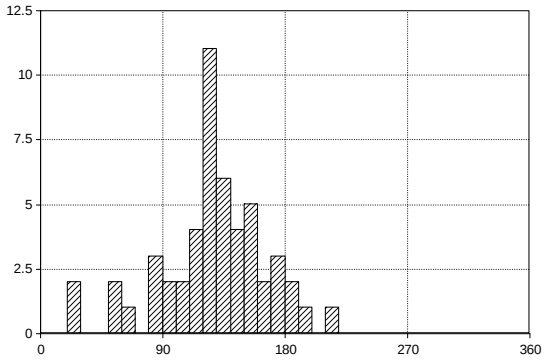
Αν για την αναπαράσταση των *συχνοτήτων* ή των *σχετικών συχνοτήτων* χρησιμοποιηθούν διανύσματα, το *κυκλικό ιστόγραμμα* παίρνει τη μορφή που φαίνεται στο *Σχήμα 9.3.6*.



Σχήμα 9.3.6
Arrow διαγραμμα αντίστοιχο του κυκλικού ιστογράμματος του Σχήματος 9.3.5

δ) *Γραμμικό ιστόγραμμα*

Πρόκειται για το σύνηθες *γραμμικό ιστόγραμμα*. Ένα μειονέκτημά του είναι ότι η οπτική εντύπωση που δημιουργεί διαφοροποιείται σημαντικά ανάλογα με την επιλογή της αρχής των αξόνων. Γι' αυτό, όταν το εύρος των τιμών είναι μεγάλο (πάνω από 180^0) δεν προτείνεται για την αναπαράσταση *κυκλικών δεδομένων*. Το *γραμμικό ιστόγραμμα* στο *Σχήμα 9.3.7* είναι το αντίστοιχο του *ροδογράμματος* του *Σχήματος 9.3.4*.



Σχήμα 9.3.7
Το γ ραμμικό ιστόγραμμα του δείγματος από τη μεταβλητή που εκφράζει την κατεύθυνση του ίχνους της κίνησης των πάγων στη νότια Φινλανδία

■

Ας δούμε ένα ακόμη παράδειγμα.

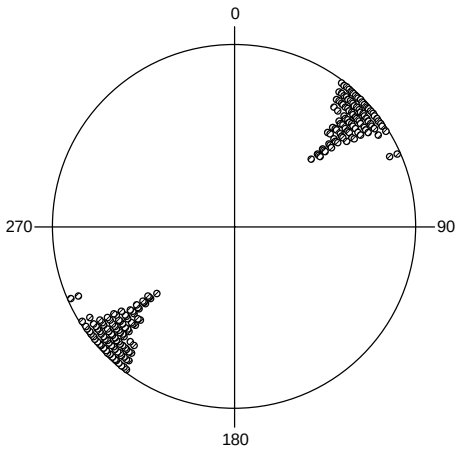
Παράδειγμα 9.3.2: *Στον Πίνακα 9.3.2 δίνονται οι διευθύνσεις των κύριων (μεγαλύτερων) αξόνων 99 ελλειπτικών γεωλογικών σχηματισμών στις νότιες ακτές του Ατλαντικού σε μια περιοχή της North Carolina (Davis, J.C., 2002).*

218	42.8	225.3	50.9	234.8	54	229.6	40.8	231.9	49.4
217.9	42.7	226.8	50.9	234.7	51.8	230	44.9	227.3	46.1
218.8	43.8	227.4	51.4	238.3	51.9	229.3	45.8	221.9	47
219.4	43.9	228.6	51.7	246.8	50.9	228.8	46.6	221.3	44.9
219.8	43.8	228.9	51.9	246.8	50.9	227.9	48	219.4	51.9
220.1	44.8	230	52.4	238.9	51	227	49.9	231.9	54.1
220.8	45.8	229.9	53.7	235.8	50.5	225.8	53	233.7	46.1
220.8	45.8	229.7	53.9	235.8	49.9	225.9	50	235	46
222	46.1	229.9	53.9	233.9	49.8	226	47.9	236	50.8
221.9	45.8	231.3	54.9	232.9	50	222	49.9	229.9	

Πίνακας 9.3.2
99 τιμές της μεταβλητής που εκφράζει τη διεύθυνση του κύριου άξονα ελλειπτικών γεωλογικών σχηματισμών στις νότιες ακτές της North Carolina (σε μοίρες από το Βορρά και κατά τη φορά της κίνησης των δεικτών του ωρολογίου)

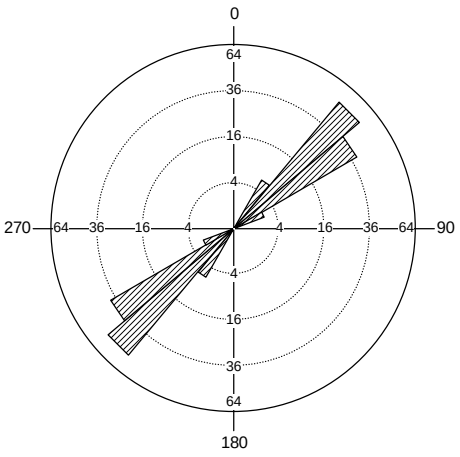
Πρόκειται για ***δεδομένα διεύθυνσης***. Δηλαδή, δεν ενδιαφέρει η φορά αλλά μόνο η διεύθυνση. Πρακτικά, αυτό σημαίνει ότι παρατηρήσεις που διαφέρουν κατά 180^0 προσδιορίζουν μια τιμή της μεταβλητής *«διεύθυνση του κύριου άξονα ελλειπτικών γεωλογικών σχηματισμών στις νότιες ακτές της North Carolina»*. Για παράδειγμα, η παρατήρηση 50^0 προσδιορίζει τη διεύθυνση 50^0-230^0. Την ίδια διεύθυνση 50^0-230^0, προσδιορίζει επίσης και η παρατήρηση 230^0. Δηλαδή, η τιμή 50^0-230^0 της μεταβλητής *«διεύθυνση του κύριου άξονα ελλειπτικών γεωλογικών σχηματισμών στις νότιες ακτές της North Carolina»* μπορεί να αποδοθεί είτε με τις 50^0 είτε με τις 230^0.

Με βάση όσα ήδη έχουμε αναφέρει για τις *μεταβλητές διεύθυνσης* στο *1ο Κεφάλαιο*, η κατανομή δεδομένων διεύθυνσης είναι προφανές ότι αναπαρίσταται σε ημικύκλιο ή σε κύκλο ως δύο συμμετρικά ως προς το κέντρο του κύκλου γραφήματα. Έτσι, για τα δεδομένα του παραδείγματός μας έχουμε τα γραφήματα που φαίνονται στα *Σχήματα 9.3.8*.



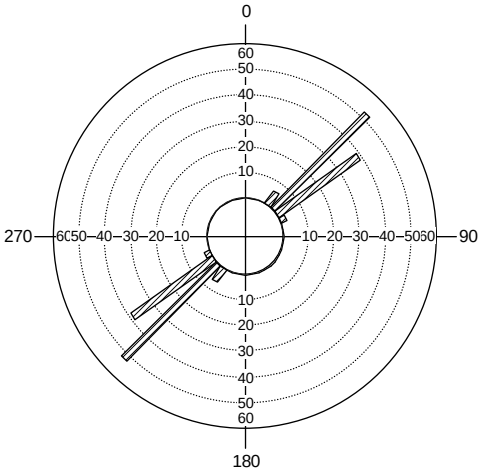
(α)

Κυκλικό διάγραμμα διασποράς



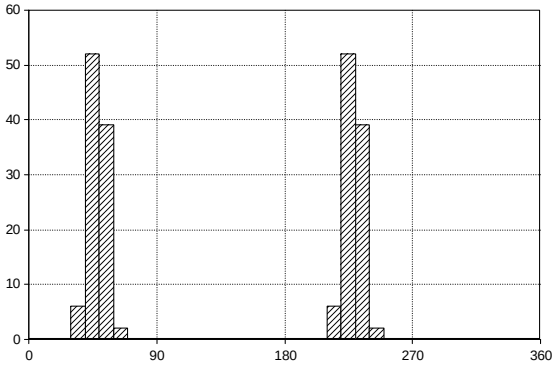
(β)

Ροδόγραμμα (σε τομείς 10^o)



(γ)

Κυκλικό ιστόγραμμα (σε τομείς 10^o)



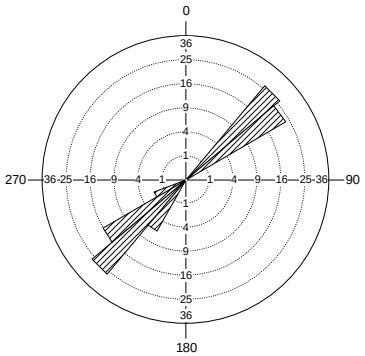
(δ)

Γραμμικό ιστόγραμμα (σε τομείς 10^o)

Σχήματα 9.3.8

Γραφική παρουσίαση της κατανομής του δείγματος από τη μεταβλητή που εκφράζει τη διεύθυνση του κύριου άξονα ελλειπτικών γεωλογικών σχηματισμών στις νότιες ακτές της North Carolina (σε μοίρες από το Βορρά και κατά τη φορά της κίνησης *των δεικτών του ωρολογίου)*

Ερώτηση: *Δείτε το* ***ροδόγραμμα*** *στο Σχήμα 9.3.9. Αφορά στα ίδια δεδομένα και σχεδιάσθηκε σε ίδιους τομείς 10^0, όπως το* ***ροδόγραμμα*** *στο Σχήμα 9.3.8β. Τι μπορεί να συνέβη και έχει αλλάξει.*



Σχήμα 9.3.9
Ροδόγραμμα

■

9.3.2 Αριθμητικά περιγραφικά μέτρα κυκλικών δεδομένων

Όπως ήδη έχουμε αναφέρει, τα *μέτρα θέσης* και *διασποράς* της κατανομής *κυκλικών δεδομένων* ενώ εννοιολογικά είναι ανάλογα με τα αντίστοιχα *μέτρα* της κατανομής *μη κυκλικών δεδομένων*, εντούτοις, τα περισσότερα από αυτά ορίζονται και υπολογίζονται διαφορετικά. Για παράδειγμα, η *διακύμανση* της κατανομής *κυκλικών δεδομένων* εκφράζει, όπως και η *διακύμανση* της κατανομής *μη κυκλικών*, το βαθμό συγκέντρωσης των δεδομένων γύρω από τον *μέσο* τους. Όμως, υπολογίζεται διαφορετικά.

Ας δούμε πώς ορίζονται και πώς υπολογίζονται τα βασικότερα *μέτρα θέσης* και *διασποράς* της κατανομής *κυκλικών δεδομένων*.

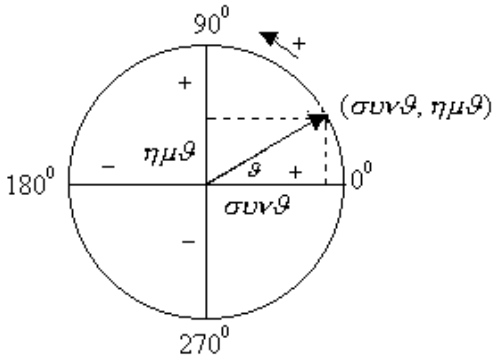
9.3.2.1 Δειγματικός μέσος κυκλικών δεδομένων

(α) Μέση κατεύθυνση

Θα ορίσουμε τη ***μέση κατεύθυνση*** $\overline{\vartheta}$, ν τιμών $\vartheta_1, \vartheta_2, \ldots, \vartheta_\nu$ μιας *μεταβλητής κατεύθυνσης*.

Επειδή τιμές σε μεταβλητές κατεύθυνσης αποδίδονται με γωνίες, είναι προφανές ότι πρέπει να ανατρέξουμε στα αντίστοιχα μαθηματικά εργαλεία. Δηλαδή, στον τριγωνομετρικό κύκλο[12] και τις τριγωνομετρικές συναρτήσεις.

Μια γωνία ϑ, αναπαριστάνεται στην περιφέρεια του τριγωνομετρικού κύκλου *(δες Σχήμα 9.3.10)* με το πέρας ενός μοναδιαίου διανύσματος $(συν\vartheta, ημ\vartheta)$. Είναι επομένως λογικό να ορίσουμε τη μέση τιμή γωνιών μέσω της συνισταμένης μοναδιαίων διανυσμάτων.



Σχήμα 9.3.10
Αναπαράσταση γωνιών σε τριγωνομετρικό κύκλο
όπου οι 0^0 ορίζονται στο θετικό ημιάξονα και ως θετική φορά
ορίζεται η αντίθετη της φοράς της κινησης των δεικτών του ωρολογίου

Έτσι, ως ***μέση κατεύθυνση*** των $\vartheta_1, \vartheta_2, \ldots, \vartheta_\nu$, ορίζεται η κατεύθυνση $\overline{\vartheta}$ της συνισταμένης $\vec{r}$ των μοναδιαίων διανυσμάτων

$$(συν\vartheta_1, ημ\vartheta_1), (συν\vartheta_2, ημ\vartheta_2), \ldots, (συν\vartheta_\nu, ημ\vartheta_\nu)$$[13].

Αν x_r, y_r είναι οι συντεταγμένες της συνισταμένης $\vec{r}$ των μοναδιαίων διανυσμάτων $(συν\vartheta_1, ημ\vartheta_1), (συν\vartheta_2, ημ\vartheta_2), \ldots, (συν\vartheta_\nu, ημ\vartheta_\nu)$, τότε, από τον ορισμό του αθροίσματος διανυσμάτων, έχουμε

---

[12] *Ο τριγωνομετρικός κύκλος είναι ένας προσανατολισμένος κύκλος που έχει ακτίνα ένα. Η αρχή (0^0) ορίζεται στο θετικό ημιάξονα Οχ, και ως θετική φορά ορίζεται η αντίθετη φορά των δεικτών του ωρολογίου.*

[13] *και ασφαλώς, όχι ο αριθμητικός μέσος τους* $(\vartheta_1 + \vartheta_2 + \ldots + \vartheta_\nu)/\nu$.

$$x_r = \sum_{i=1}^{\nu} \sigma\upsilon\nu\vartheta_i \text{ και } y_r = \sum_{i=1}^{\nu} \eta\mu\vartheta_i .$$

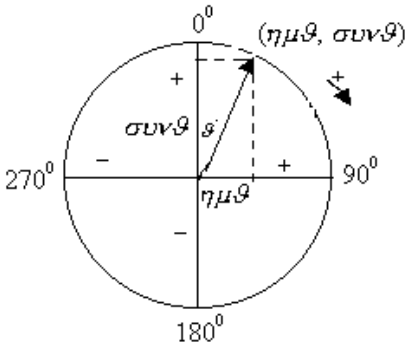
Επομένως, η *μέση κατεύθυνση* $\overline{\vartheta}$, των $\vartheta_1, \vartheta_2, \ldots, \vartheta_\nu$, υπολογίζεται από τη σχέση

$$\overline{\vartheta} = \tau o \xi \varepsilon \phi \frac{\sum_{i=1}^{\nu} \eta\mu\vartheta_i}{\sum_{i=1}^{\nu} \sigma\upsilon\nu\vartheta_i}$$

σε συνδυασμό με το πρόσημο των x_r και y_r (αφού οι τιμές της εφαπτομένης επαναλαμβάνονται ανά 180^0).

Βέβαια, αν οι 0^0 ορισθούν στο θετικό ημιάξονα Oy, δηλαδή στο Βορρά και ως θετική φορά ορισθεί η φορά της κίνησης των δεικτών του ωρολογίου *(Σχήμα 9.3.11),* τότε στην περίπτωση αυτή, το μοναδιαίο διάνυσμα που αντιστοιχεί στη γωνία ϑ έχει συντεταγμένες $(\eta\mu\vartheta, \sigma\upsilon\nu\vartheta)$ (γιατί;) και επομμένως

$$x_r = \sum_{i=1}^{\nu} \eta\mu\vartheta_i \text{ και } y_r = \sum_{i=1}^{\nu} \sigma\upsilon\nu\vartheta_i .$$



Σχήμα 9.3.11
Αναπαράσταση γωνιών όταν οι 0^0 ορίζονται στο Βορρά και ως θετική φορά ορίζεται η φορά της κίνησης των δεικτών του ωρολογίου

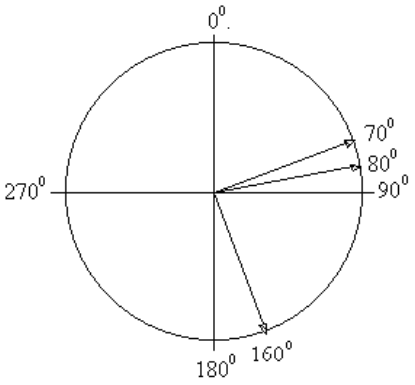
Η *μέση κατεύθυνση* υπολογίζεται και πάλι από τη σχέση

$$\overline{\vartheta} = \tau o \xi \varepsilon \phi \frac{\sum_{i=1}^{\nu} \eta\mu\vartheta_i}{\sum_{i=1}^{\nu} \sigma\upsilon\nu\vartheta_i}$$

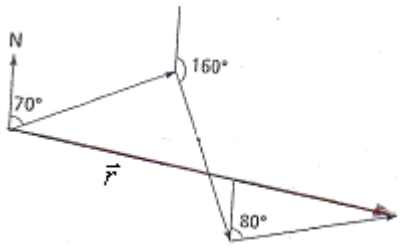
σε συνδυασμό με το πρόσημο των x_r και y_r.

Παράδειγμα 9.3.3: *Ας υπολογίσουμε τον μέσο των κατευθύνσεων τριών ανέμων 70^0, 160^0 και 80^0 αντίστοιχα. Οι κατευθύνσεις αυτές μετρήθηκαν από το βορρά και κατά τη φορά των δεικτών του ωρολογίου (Σχήμα 9.3.12α).*

<u>*Απάντηση:*</u> Γραφικά, η *μέση κατεύθυνση* μπορεί να βρεθεί με το γνωστό *κανόνα του παραλληλογράμμου* ως η συνισταμένη $\vec{r}$ των τριών μοναδιαίων διανυσμάτων $(\eta\mu 70^0, \sigma\upsilon\nu 70^0)$, $(\eta\mu 160^0, \sigma\upsilon\nu 160^0)$ και $(\eta\mu 80^0, \sigma\upsilon\nu 80^0)$ *(δες Σχήμα 9.3.12β).*



(α)
Γραφική αναπαράσταση τριών κατευθύνσεων ανέμων 70^0, 160^0 και 80^0 όταν οι 0^0 ορίζονται στο Βορρά και ως θετική φορά ορίζεται η φορά της κίνησης των δεικτών του ωρολογίου



(β)
Προσδιορισμός του μέσου τριών κατευθύνσεων ανέμων 70^0, 160^0 και 80^0 με τον κανόνα του παραλληλογράμμου

Σχήμα 9.3.12
Αναπαράσταση τριών κατευθύνσεων ανέμων και υπολογισμός του μέσου τους με τον κανόνα του παραλληλογράμμου

Ας υπολογίσουμε τις συντεταγμένες x_r και y_r της $\vec{r}$.

Έχουμε

ϑ	$συν\vartheta$	$ημ\vartheta$
70^0	0.34202	0.939693
160^0	-0.93969	0.34202
80^0	0.173648	0.984808
Αθροίσματα	**-0.42402**	**2.266521**

Δηλαδή

$$x_r = \sum_{i=1}^{3} ημ\vartheta_i = 2.266521 > 0$$

$$y_r = \sum_{i=1}^{3} συν\vartheta_i = -0.42402 < 0$$

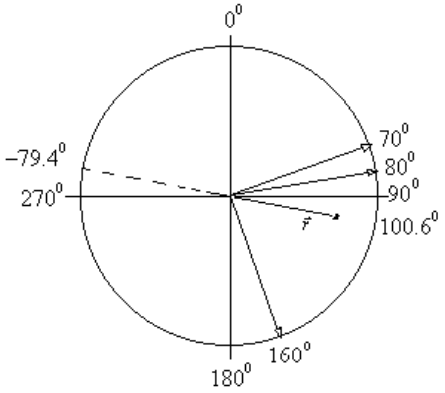
και επομένως

$$\overline{\vartheta} = τοξεφ\frac{2.266521}{-0.42402} = τοξεφ(-5.3453)\,.$$

Άρα

$$\overline{\vartheta} = -79.4^0 \text{ ή } \overline{\vartheta} = 180^0 + (-79.4^0) = 100.6^0$$

και επειδή $x_r > 0$ και $y_r < 0$ η *μέση κατεύθυνση* των τριών ανέμων είναι $\overline{\vartheta} = 100.6^0$ και όχι -79.4^0, δηλαδή, είναι περίπου ανατολική και όχι περίπου δυτική! *(δες και Σχήμα 9.3.13).*



Σχήμα 9.3.13
Μέση κατεύθυνση τριών κατευθύνσεων ανέμων
70^0, 160^0 και 80^0

■

Παρατήρηση 9.3.1: *Στο Σχήμα 9.3.13, οι συντεταγμένες x_r, y_r της συνισταμένης $\vec{r}$, έχουν διαιρεθεί με το μέγεθος του δείγματος ν. Δηλαδή, το πέρας της $\vec{r}$ βρίσκεται στο σημείο*

$$\left(\frac{\sum_{i=1}^{\nu} \eta\mu\vartheta_i}{\nu}, \frac{\sum_{i=1}^{\nu} \sigma\upsilon\nu\vartheta_i}{\nu} \right)$$

δηλαδή, στο σημείο

$$\left(\frac{x_r}{\nu}, \frac{y_r}{\nu} \right)$$

και όχι στο (x_r, y_r). Έτσι, το $\vec{r}$ δεν έχει σχεδιασθεί με μήκος ίσο με το μέτρο του

$$|\vec{r}| = r = \sqrt{x_r^2 + y_r^2}$$

αλλά, με μήκος ίσο με

$$\bar{r} = \sqrt{\left(\frac{x_r}{\nu} \right)^2 + \left(\frac{x_r}{\nu} \right)^2} = \frac{r}{\nu}$$

το οποίο ονομάζεται ***μέσο μέτρο της*** $\vec{r}$.

Ποια σκοπιμότητα εξυπηρετεί ο ορισμός του μέσου μέτρου θα φανεί στη συνέχεια όταν ορίσουμε τα μέτρα διασποράς.

■

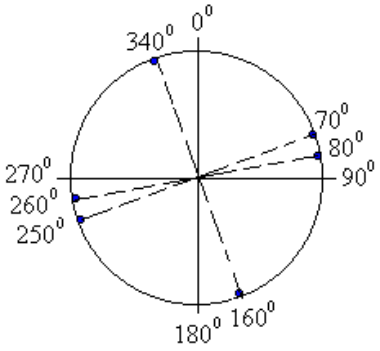
(β) Μέση διεύθυνση

Η ***μέση διεύθυνση*** $\bar{\vartheta}$, ν τιμών $\vartheta_1, \vartheta_2, \ldots, \vartheta_\nu$, μιας μεταβλητής διεύθυνσης ορίζεται όπως και η ***μέση κατεύθυνση*** ν τιμών μιας μεταβλητής κατεύθυνσης, όμως υπολογίζεται αφού προηγουμένως οι τιμές μετασχηματισθούν.

Στο *$1^ο$ Κεφάλαιο* εξηγήσαμε ότι σε μια διεύθυνση αποδίδουμε τιμή επιλέγοντας την τιμή μιας οποιασδήποτε από τις δύο αντίθετες κατευθύνσεις που ορίζει. Εξηγήσαμε επίσης, ότι η ανάλυση δεδομένων που αφορούν μεταβλητές διεύθυνσης γίνεται αφού προηγουμένως αυτά μετασχηματισθούν. Ας δούμε ένα παράδειγμα.

Παράδειγμα 9.3.4: *Τρεις γραμμώσεις σε μια εικόνα Landsat έχουν διευθύνσεις 70^0-250^0, 80^0-260^0 και 160^0-340^0.*

Οι διευθύνσεις των γραμμώσεων μετρήθηκαν από το βορρά και κατά τη φορά της κίνησης των δεικτών του ωρολογίου (Σχήμα 9.3.14). Θα υπολογίσουμε τη μέση διεύθυνση των τριών γραμμώσεων.



Σχήμα 9.3.14
Γραφική αναπαράσταση των διευθύνσεων τριών Landsat γραμμώσεων
70^0-250^0, 80^0-260^0 και 160^0-340^0
όταν οι 0^0 ορίζονται στο Βορρά και ως θετική φορά
ορίζεται η φορά της κίνησης των δεικτών του ωρολογίου

Επιλέγουμε ως τιμές των διευθύνσεων των γραμμώσεων τις 70^0, 80^0 και 160^0 αντίστοιχα. Διπλασιάζουμε τις τιμές αυτές και εργαζόμαστε όπως στο προηγούμενο παράδειγμα *(Παράδειγμα 9.3.3)* που αφορούσε *μεταβλητή κατεύθυνσης*.

Υπολογίζουμε τις συντεταγμένες x_r και y_r της $\vec{r}$.

Έχουμε

ϑ	2ϑ	$συν(2\vartheta)$	$ημ(2\vartheta)$
70^0	140^0	-0.76604	0.64279
80^0	160^0	-0.93969	0.34202
160^0	320^0	0.76604	-0.64279
Αθροίσματα		**-0.93969**	**0.34202**

Δηλαδή

$$x_r = \sum_{i=1}^{3} ημ(2\vartheta_i) = 0.342020 > 0$$

$$y_r = \sum_{i=1}^{3} συν(2\vartheta_i) = -0.93969 < 0$$

και επομένως

$$2\overline{\vartheta} = τοξεφ\frac{0.34202}{-0.93969} = τοξεφ(-0.36397)\,.$$

Άρα

$$2\overline{\vartheta} = -20^0 \text{ ή } 2\overline{\vartheta} = 180 + (-20^0) = 160^0$$

και επειδή $x_r > 0$ και $y_r < 0$ είναι $2\overline{\vartheta} = 160^0$. Έτσι, η *μέση διεύθυνση* των τριών γραμμώσεων είναι $\overline{\vartheta} = 80^0$, δηλαδή, η διεύθυνση 80^0 - 260^0.

Είναι φανερό ότι αν ως τιμές των διευθύνσεων 70^0-250^0, 80^0-260^0 και 160^0-340^0, αντί των 70^0, 80^0 και 160^0 επιλέγαμε π.χ. τις 250^0, 80^0 και 340^0, θα είχαμε το ίδιο αποτέλεσμα αφού για τον υπολογισμό της *μέσης διεύθυνσης* θα χρησιμοποιούσαμε και πάλι τις ίδιες τιμές: $2\cdot(250^0) - 360^0 = 140^0$, $2\cdot(80^0) = 160^0$ και $2\cdot(340^0)-360^0 = 320^0$.

■

Ερωτήσεις:

α) Πότε η μέση κατεύθυνση ή η μέση διεύθυνση δεν ορίζεται[14];

β) Το μέτρο $r = \sqrt{x_r^2 + y_r^2}$ *και το μέσο μέτρο* $\bar{r} = r/\nu$ *του διανύσματος* $\vec{r}$*, τι εκφράζουν άραγε[15];*

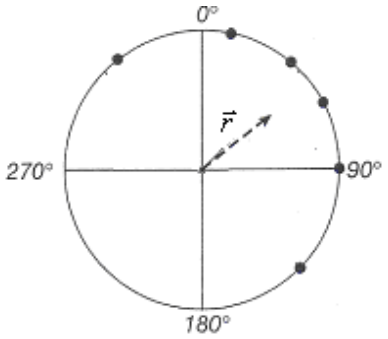
Παρατηρήσεις 9.3.2:

α) Η μέση κατεύθυνση $\bar{\vartheta}$*, όπως ορίσθηκε, έχει το εξής μειονέκτημα. Όταν το μέτρο του* $\vec{r}$ *είναι πολύ μικρό, τότε, μικρή αλλαγή σε κάποιο από τα μοναδιαία διανύσματα (δηλαδή σε κάποια κατεύθυνση) μπορεί να προκαλέσει μεγάλη αλλαγή στη μέση κατεύθυνση* $\bar{\vartheta}$*.*

β) Όπως συμβαίνει και με τον μέσο μη κυκλικών δεδομένων, ο μέσος κυκλικών δεδομένων είναι το κέντρο ισορροπίας της κατανομής τους. Έτσι, αν σε ένα κυκλικό δίσκο αμελητέου βάρους (Σχήμα 9.3.15) θεωρήσουμε ότι στα σημεία $(\eta\mu\vartheta_i, \sigma\upsilon\nu\vartheta_i)$ *βρίσκονται ίσα βάρη, τότε το κέντρο ισορροπίας (το κέντρο βάρους) του δίσκου βρίσκεται στο πέρας του* $\vec{r}$*, δηλαδή, στο σημείο*

$$\left(\frac{x_r}{\nu}, \frac{y_r}{\nu}\right)$$

(το $\vec{r}$ *έχει σχεδιασθεί με μήκος* $\bar{r}$*).*



Σχήμα 9.3.15
Ερμηνεία του δειγματικού μέσου κυκλικών δεδομένων ως το σημείο ισορροπίας της κατανομής του δείγματος

■

9.3.2.2 Διακύμανση κυκλικών δεδομένων

(α) Διακύμανση δεδομένων κατεύθυνσης

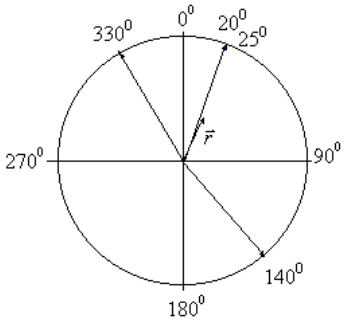
Θα ορίσουμε τη ***διακύμανση*** ν τιμών $\vartheta_1, \vartheta_2, \ldots, \vartheta_\nu$, μιας *μεταβλητής κατεύθυνσης,* δηλαδή ένα μέτρο που να εκφράζει πόσο διασκορπισμένες (ή ισοδύναμα, συγκεντρώμενες) είναι οι $\vartheta_1, \vartheta_2, \ldots, \vartheta_\nu$ γύρω από τον *μέσο* τους, δηλαδή, γύρω από τη *μέση κατεύθυνση*, $\bar{\vartheta}$.

Ας θεωρήσουμε τρεις κατευθύνσεις ανέμων 140^0, 330^0 και 20^0 *(Σχήμα 9.3.16).* Εύκολα βρίσκουμε ότι η *μέση κατεύθυνση* αυτών των ανέμων είναι $\bar{\vartheta} = 25^0$. Άραγε, πώς μπορούμε να εκφράσουμε ποσοτικά το πόσο διασκορπισμένες είναι οι κατευθύνσεις των τριών ανέμων γύρω από την τιμή αυτή;

---

[14] *Σκεφθείτε τι συμβαίνει όταν* $\vec{r} = \vec{0}$

[15] *Η απάντηση στη συνέχεια!*



Σχήμα 9.3.16
Γραφική αναπαράσταση των κατευθύνσεων τριών ανέμων
140^0, 330^0 και 20^0
και του μέσου τους
όταν οι 0^0 ορίζονται στο Βορρά και ως θετική φορά
ορίζεται η φορά της κίνησης των δεικτών του ωρολογίου

Με μια πρώτη ματιά φαίνεται ότι οι τιμές 140^0, 330^0 και 20^0 είναι σαφώς περισσότερο διασκορπισμένες γύρω από τη *μέση κατεύθυνσή* τους $\overline{\vartheta} = 25^0$, από όσο είναι οι τιμές 70^0, 160^0 και 80^0 των κατευθύνσεων των ανέμων του προηγούμενου παραδείγματος από τη μέση κατεύθυνσή τους $\overline{\vartheta} = 100.6^0$ *(δες Σχήμα 9.3.13)*. Επίσης, αν παρατηρήσουμε το *μέσο μέτρο* $\bar{r}$ της συνισταμένης $\vec{r}$ στις δύο ομάδες δεδομένων, διαπιστώνουμε ότι η ομάδα δεδομένων που είναι περισσότερο συγκεντρωμένη γύρω από τη μέση κατεύθυνσή της, που έχει δηλαδή μικρότερη *διακύμανση*, έχει συνισταμένη $\vec{r}$ με μεγαλύτερο *μέσο μέτρο*.

Η ίδια διαπίστωση, αβίαστα προκύπτει και από τα παραδείγματα στο *Σχήμα 9.3.17*, όπου όλες οι ομάδες δεδομένων έχουν την ίδια *μέση κατεύθυνση* $\overline{\vartheta} = 50^0$, όμως οι κατανομές τους είναι διαφορετικές *(Zar, J.H., 2010)*.

Στις περισσότερο διασκορπισμένες (γύρω από τη *μέση κατεύθυνση* $\overline{\vartheta} = 50^0$) ομάδες δεδομένων, αντιστοιχεί μικρότερο $\bar{r}$. Έτσι, για τις περιπτώσεις (α), (β), (γ), (δ), (ε) και (στ), αντίστοιχα έχουμε $\bar{r} = 1$, $\bar{r} = 0.99$, $\bar{r} = 0.90$, $\bar{r} = 0.60$, $\bar{r} = 0.30$ και $\bar{r} = 0.00$.

Φαίνεται δηλαδή, ότι το *μέτρο* r (και φυσικά και το *μέσο μέτρο* $\bar{r}$) της συνισταμένης $\vec{r}$ των μοναδιαίων διανυσμάτων που αντιστοιχούν στις κατευθύνσεις $\vartheta_1, \vartheta_2, \ldots, \vartheta_\nu$, περιέχει πληροφορία για τη *διακύμανση* των $\vartheta_1, \vartheta_2, \ldots, \vartheta_\nu$ γύρω από τον *μέσο* τους $\overline{\vartheta}$.

Έτσι, είναι λογικό, ως ένα μέτρο *διασποράς* των ν κατευθύνσεων $\vartheta_1, \vartheta_2, \ldots, \vartheta_\nu$, να ορισθεί π.χ. η ποσότητα $1-r$ ή η ποσότητα $1-\bar{r}$. Όμως, για να είναι δυνατή η σύγκριση των *διακυμάνσεων* δύο ή περισσοτέρων δειγμάτων διαφορετικού μεγέθους, είναι προφανές ότι πρέπει να χρησιμοποιηθεί η ποσότητα $1-\bar{r}$, αφού, το μέτρο του $\vec{r}$ δεν επηρεάζεται μόνο από τη *διακύμανση* του δείγματος αλλά προφανώς και από το μέγεθός του, ν.

Έτσι, ως ένα *μέτρο διασποράς* των ν κατευθύνσεων $\vartheta_1, \vartheta_2, \ldots, \vartheta_\nu$ γύρω από τη *μέση κατεύθυνσή* τους $\overline{\vartheta}$, ορίζουμε την ποσότητα

$$S^2 = 1 - \bar{r}$$

όπου,

$$\bar{r} = \frac{r}{\nu} = \frac{\sqrt{x_r^2 + y_r^2}}{\nu}$$

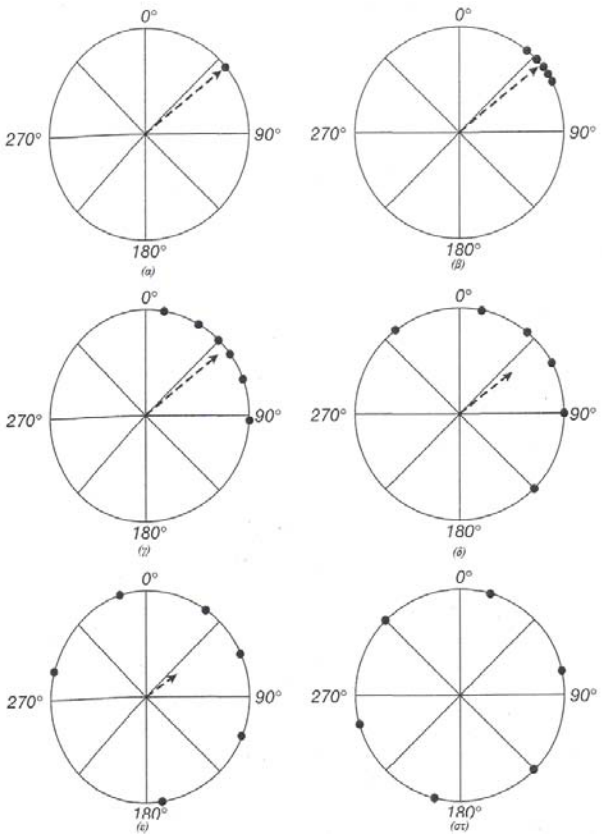
το *μέσο μέτρο* της συνισταμένης $\vec{r}$.

Ως μέτρα διασποράς κυκλικών δεδομένων ορίζονται, με βάση επίσης το *μέσο μέτρο* $\bar{r}$, και τα ακόλουθα.

$$s^2 = 2(1-\bar{r})$$
$$s_0^2 = -2\ln\bar{r}\,.$$

Η ***τυπική απόκλιση*** των $\vartheta_1, \vartheta_2, \ldots, \vartheta_\nu$ ορίζεται ως η (θετική) τετραγωνική ρίζα της *διακύμανσης* s^2 ή της *διακύμανσης* s_0^2. Δηλαδή,

$$s = \sqrt{2(1-\bar{r})} \text{ ή } s_0 = \sqrt{-2\ln\bar{r}}\,.$$



Σχήμα 9.3.17
(α) $\bar{r} = 1$, *β)* $\bar{r} = 0.99$, *(γ)* $\bar{r} = 0.90$,
(δ) $\bar{r} = 0.60$, *(ε)* $\bar{r} = 0.30$, *(στ)* $\bar{r} = 0.00$.

Παρατηρήσεις 9.3.3:

α) Τα τρία μέτρα διακύμανσης S^2, s^2, s_0^2, *όπως ορίσθηκαν παραπάνω, δίνουν τη διακύμανση σε ακτίνια στο τετράγωνο (rad²). Για να πάρουμε την τιμή της διακύμανσης σε μοίρες στο τετράγωνο αρκεί να πολλαπλασιάσουμε με* $(180^0/\pi)^2$. *Έτσι, οι αντίστοιχοι τύποι γίνονται*

$$S^2 = \left(\frac{180^0}{\pi}\right)^2 (1-\bar{r})$$

$$s^2 = 2\left(\frac{180^0}{\pi}\right)^2 (1-\bar{r}) \text{ και } s = \frac{180^0}{\pi}\sqrt{2(1-\bar{r})}$$

$s_0^2 = \left(\frac{180^0}{\pi}\right)^2 (-2\ln\bar{r})$ *και* $s_0 = \frac{180^0}{\pi}\sqrt{-2\ln\bar{r}}$.

β) Η διακύμανση $S^2 = 1-\bar{r}$ *παίρνει τιμές από 0 μέχρι 1. Η τιμή 0 σημαίνει ότι όλες οι κατευθύνσεις είναι συγκεντρωμένες σε μία κατεύθυνση ενώ η τιμή 1 σημαίνει ότι οι παρατηρήσεις έχουν τη μέγιστη διακύμανση. Όμως προσοχή! Η τιμή 1 δε σημαίνει ότι οι παρατηρήσεις (κατευθύνσεις) είναι, κατ' ανάγκη, ομοιόμορφα κατανεμημένες στην περιφέρεια του κύκλου.*

γ) Η διακύμανση $s^2 = 2(1-\bar{r})$ *παίρνει τιμές από 0 μέχρι 2 ενώ η* $s_0^2 = -2\ln\bar{r}$ *παίρνει τιμές από 0 μέχρι* $+\infty$ *. Για μεγάλα* $\bar{r}$ *, οι τιμές των* s, s_0 *δε διαφέρουν πολύ ενώ για μικρές τιμές του* $\bar{r}$ *δε συμβαίνει το ίδιο. Στα παραδείγματα (α), (β), (γ), (δ), (ε) και (στ) του Σχήματος 9.3.17 έχουμε αντίστοιχα*

$\bar{r}=1$ *και* $s=0^0, s_0=0^0$, $\bar{r}=0.99$ *και* $s=8.10^0, s_0=8.12^0$, $\bar{r}=0.90$ *και* $s=25.62^0, s_0=26.30^0$, $\bar{r}=0.60$ *και* $s=51.25^0, s_0=57.91^0$, $\bar{r}=0.30$ *και* $s=67.79^0, s_0=88.91^0$, $\bar{r}=0.00$ *και* $s=81.03^0, s_0=\infty$.

δ) Σε ομαδοποιημένες παρατηρήσεις, για τον υπολογισμό της διακύμανσης, στη βιβλιογραφία προτείνεται να γίνεται «διόρθωση» του $\bar{r}$ *. Έτσι, αντί του* $\bar{r}$ *προτείνεται να χρησιμοποιείται το*

$$\bar{r}_c = \frac{d\pi/360^o}{\eta\mu(d/2)}\cdot\bar{r}$$

όπου, d το πλάτος των κλάσεων σε μοίρες. Για πλάτος κλάσεων μικρότερο των 30^0*, η διόρθωση αυτή είναι αμελητέα.*

■

Παράδειγμα 9.3.5 (συνέχεια του Παραδείγματος 9.3.3): *Στο Παράδειγμα 9.3.3 βρήκαμε ότι ο μέσος των τριών κατευθύνσεων* 70^0*,* 160^0 *και* 80^0 *είναι* $\bar{\vartheta} = 100.6^0$ *και επίσης, ότι οι συντεταγμένες* x_r, y_r *του* $\vec{r}$ *είναι*

$$x_r = \sum_{i=1}^{3} \eta\mu\vartheta_i = 2.266521$$

$$y_r = \sum_{i=1}^{3} \sigma\upsilon\nu\vartheta_i = -0.42402 \cdot$$

Ας υπολογίσουμε τη διακύμανση και την τυπική απόκλιση αυτών των κατευθύνσεων γύρω από τον μέσο τους.

Το *μέσο μέτρο* του $\vec{r}$ είναι

$$\bar{r} = \frac{r}{3} = \frac{\sqrt{2.27^2 + (-0.42)^2}}{3} = \frac{2.3}{3} = 0.77$$

επομένως

$S^2 = 1-\bar{r} = 1-0.77 = 0.23$ *rad*

$s^2 = 2(1-\bar{r}) = 2\cdot(1-0.77) = 0.46$ *rad*^{2} και $s = \sqrt{0.46} = 0.68$ *rad* $= 38.96^0$

$s_0^2 = -2\ln\bar{r} = -2\ln(0.77) = 0.523$ *rad*^{2} και $s_0 = \sqrt{0.523} = 0.723$ *rad* $= 41.42^0$.

■

(β) Διακύμανση δεδομένων διεύθυνσης

Η ***διακύμανση*** και η ***τυπική απόκλισή*** ν τιμών $\vartheta_1, \vartheta_2, \ldots, \vartheta_\nu$, μιας *μεταβλητής διεύθυνσης* γύρω από τη *μέση διεύθυνσή* τους $\bar{\vartheta}$, ορίζονται όπως η *διακύμανση* και η *τυπική απόκλιση δεδομένων κατεύθυνσης*. Όμως, υπολογίζονται αφού προηγουμένως οι τιμές μετασχηματισθούν. Ας δούμε ένα παράδειγμα.

Παράδειγμα 9.3.6 (συνέχεια του Παραδείγματος 9.3.4): *Στο Παράδειγμα 9.3.4 βρήκαμε ότι ο μέσος των διευθύνσεων 70^0-250^0, 80^0-260^0 και 160^0-340^0 είναι $\overline{\vartheta} = 80^0$ και επίσης, ότι οι συντεταγμένες x_r, y_r του $\vec{r}$, για τα μετασχηματισμένα δεδομένα, είναι*

$$x_r = \sum_{i=1}^{3} \eta\mu(2\vartheta_i) = 0.342020$$

$$y_r = \sum_{i=1}^{3} \sigma\upsilon\nu(2\vartheta_i) = -0.93969 .$$

Ας υπολογίσουμε τη διακύμανση και την τυπική απόκλιση των διευθύνσεων αυτών γύρω από τον μέσο τους.

Το *μέσο μέτρο* του $\vec{r}$, για τα μετασχηματισμένα δεδομένα, είναι

$$\bar{r} = \frac{r}{3} = \frac{\sqrt{0.34^2 + (-0.94)^2}}{3} = \frac{0.999}{3} = 0.333 .$$

Άρα, για τα μετασχηματισμένα δεδομένα είναι

$S^2 = 1 - \bar{r} = 1 - 0.333 = 0.666$ *rad*^{2}

$s^2 = 2 \cdot (1 - \bar{r}) = 2 \cdot (1 - 0.333) = 1.33$ *rad*^{2} και $s = \sqrt{1.33} = 1.15$ *rad* $= 66.08^0$

$s_0^2 = -2 \cdot \ln \bar{r} = -2 \cdot \ln(0.333) = 2.199$ *rad*^{2} και $s_0 = \sqrt{2.199} = 1.48$ *rad* $= 84.968^0$

και για τα αρχικά δεδομένα, αντίστοιχα, είναι

$$S^2 = \frac{0.666}{2} = 0.333$$

$$s^2 = \frac{1.33}{2} = 0.665 \text{ rad}^2 \text{ και } s = \frac{1.15}{2} = 0.575 \text{ rad} = 33^0$$

$$s_0^2 = \frac{2.199}{2} = 1.099 \text{ rad}^2 \text{ και } s_0 = \frac{1.48}{2} = 0.74 \text{ rad} = 42.4^0.$$

■

Για τις κατανομές *κυκλικών δεδομένων* ορίζονται και άλλα γνωστά *μέτρα θέσης* και *διασποράς* όπως η ***διάμεσος***, τα ***ποσοστιαία σημεία***, η ***κορυφή***, το ***εύρος***, καθώς και μέτρα ***συμμετρίας*** και ***κύρτωσης***. Ως έννοιες, όλα ορίζονται ανάλογα με τα αντίστοιχα *μη κυκλικών δεδομένων*, όμως υπολογίζονται διαφορετικά. Επίσης, κάποια υπολογίζονται απλά, για αλλά απαιτείται ιδιαίτερη προσοχή γιατί εύκολα μπορεί να γίνουν λάθη. Για παράδειγμα, το *εύρος* ορίζεται ως το μικρότερο τόξο του κύκλου που περιλαμβάνει όλα τα δεδομένα. Έτσι, οι κατανομές που φαίνονται στο *Σχήμα 9.3.17*, έχουν *εύρος*, η (α) 0^0, η (β) 22^0 (μεταξύ 38^0 και 60^0), η (γ) 83^0 (μεταξύ 10^0 και 93^0), η (δ) 173^0 (μεταξύ 322^0 και 135^0) και όχι $322^0 - 135^0 = 187^0$, η (ε) 246^0 (μεταξύ 285^0 και 171^0) και όχι $285^0 - 171^0 = 114^0$ και η (στ) 300^0 (γιατί;).

Άλλα μέτρα απαιτούν πολύπλοκους υπολογισμούς, αλλά πλέον, το πρόβλημα αυτό μπορεί να αντιμετωπισθεί με κατάλληλο λογισμικό. Βέβαια, το λογισμικό μας απαλλάσσει από τους πολύπλοκους και χρονοβόρους αριθμητικούς υπολογισμούς, όμως, όπως έχουμε επισημάνει, για τη σωστή ερμηνεία των αποτελεσμάτων των αριθμητικών υπολογισμών, **απαιτείται να έχουμε κατανοήσει το νόημα και τη σημασία των αντίστοιχων εννοιών**. Από την «υποχρέωση» αυτή, το λογισμικό δε μας απαλλάσσει *(παρότι, αν είναι κατάλληλα σχεδιασμένο, μπορεί επιπλέον να βοηθήσει και στην κατανόηση των εννοιών γιατί διευκολύνει εναλλακτικές προσεγγίσεις, πολλαπλές αναπαραστάσεις, πολλαπλές δοκιμές, διερεύνηση κτλ.).*

■

9.4 Σύντομη ανασκόπηση βασικών εννοιών, προτάσεων και τύπων

Πληθυσμός ή στατιστικός πληθυσμός	Πληθυσμό ή στατιστικό πληθυσμό ονομάζουμε την κατανομή των τιμών μιας τυχαίας μεταβλητής, δηλαδή την κατανομή των τιμών που παίρνει ένα κοινό χαρακτηριστικό μιας ομάδας υποκειμένων. Κάθε υποκείμενο επί του οποίου μετράται/παρατηρείται η τιμή ενός κοινού χαρακτηριστικού ονομάζεται ***απλό στειχείο*** ή ***δειγματοληπτική/πειραματική μονάδα***.
Τυχαίο δείγμα και πραγματοποίηση τυχαίου δείγματος	Τυχαίο δείγμα μεγέθους *ν* από έναν πληθυσμό ονομάζουμε *ν* ανεξάρτητες τυχαίες μεταβλητές $X_1, X_2,\ldots,X_\nu$ που παίρνουν τιμές από τον πληθυσμό αυτό (και έχουν επομένως την ίδια κατανομή). Οι συγκεκριμένες τιμές $x_1, x_2,\ldots,x_\nu$, που έχουμε διαθέσιμες για επεξεργασία μετά τη λήψη του δείγματος αποτελούν μια πραγματοποίηση των $X_1, X_2,\ldots,X_\nu$ και ονομάζονται ***δεδομένα*** ή ***παρατηρήσεις***.
Πίνακας κατανομής συχνοτήτων	**α) Ποσοτικές μεταβλητές** Στην πρώτη στήλη του πίνακα κατανομής συχνοτήτων καταγράφονται σε αύξουσα σειρά οι διαφορετικές τιμές $y_1, y_2,\ldots, y_k$ από τις $x_1, x_2,\ldots,x_\nu$ που εμφανίσθηκαν στο δείγμα. Στις επόμενες στήλες, για κάθε τιμή y_i, $i=1,2,\ldots,k$, καταγράφεται • *η συχνότητά της,* ν_i *(πόσες φορές εμφανίσθηκε)* • *η σχετική συχνότητά της,* $f_i=\nu_i/\nu$ • *η αθροιστική συχνότητά της,* N_i *(το άθροισμα των συχνοτήτων των τιμών που είναι* $\le y_i$*)* • *η αθροιστική σχετική συχνότητά της,* F_i *(το άθροισμα των σχετικών συχνοτήτων των τιμών που είναι* $\le y_i$*)* Αν (έχει) γίνει ομαδοποίηση των τιμών, στην πρώτη στήλη αντί των διαφορετικών τιμών καταγράφονται οι διαφορετικές κλάσεις τιμών. Στις επόμενες στήλες καταγράφεται η συχνότητα, η σχετική συχνότητα, η αθροιστική συχνότητα και η αθροιστική σχετική συχνότητα κάθε κλάσης τιμών. **β) Ποιοτικές μεταβλητές** • Στις ποιοτικές μεταβλητές **κατηγορίας** δεν έχει νόημα η διάταξη των διαφορετικών τιμών $y_1, y_2,\ldots, y_k$ και επομένως δεν έχουν νόημα ούτε οι αθροιστικές ούτε οι αθροιστικές σχετικές συχνότητες αλλά μόνο οι συχνότητες και οι σχετικές συχνότητες. • Στις ποιοτικές μεταβλητές **διάταξης** η διάταξη των διαφορετικών τιμών $y_1, y_2,\ldots, y_k$ έχει νόημα και επομένως έχουν νόημα τόσο οι συχνότητες και οι σχετικές συχνότητες όσο και οι αθροιστικές και οι αθροιστικές σχετικές.
Γραφική παρουσίαση κατανομής συχνοτήτων	**α) Ποσοτικές μεταβλητές** • *Σημειόγραμμα* • *Ραβδόγραμμα συχνοτήτων και σχετικών συχνοτήτων* • *Διάγραμμα συχνοτήτων και σχετικών συχνοτήτων* • *Κυκλικό διάγραμμα συχνοτήτων και σχετικών συχνοτήτων* • *Ιστόγραμμα συχνοτήτων/σχετικών συχνοτήτων/ αθροιστικών συχνοτήτων/αθροιστικών σχετικών συχνοτήτων* • *Πολύγωνο συχνοτήτων/σχετικών συχνοτήτω/αθροιστικών συχνοτήτων/ αθροιστικών σχετικών συχνοτήτων* • *Φυλλογράφημα* • *Θηκόγραμμα* **β) Ποιοτικές μεταβλητές** • *Ραβδόγραμμα συχνοτήτων και σχετικών συχνοτήτων* • *Κυκλικό διάγραμμα συχνοτήτων και σχετικών συχνοτήτων* **γ) Κυκλικές μεταβλητές (διεύθυνσης ή κατεύθυνσης)**

	• *Κυκλικό διάγραμμα διασποράς* • *Ροδόγραμμα* • *Κυκλικό ιστόγραμμα* • *Γραμμικό ιστογραμμα*
Αριθμητικά περιγραφικά μέτρα *(για συγκεκριμένη πραγματοποίηση* $x_1, x_2, \ldots, x_\nu$ *δείγματος με* $y_1, y_2, \ldots, y_k$ *διαφορετικές τιμές)*	**α) Ποσοτικές μεταβλητές** **Μέτρα θέσης** • ***Δειγματικός μέσος,*** $\bar{x}$ $\bar{x} = \frac{1}{\nu}\sum_{i=1}^{\nu} x_i = \frac{1}{\nu}\sum_{i=1}^{k} \nu_i y_i = \sum_{i=1}^{k} f_i y_i$ • ***Κορυφή του δείγματος,*** M_0 Η τιμή με τη μεγαλύτερη συχνότητα • ***Διάμεσος του δείγματος,*** δ ***ή*** Q_2 Το πολύ 50% των τιμών του δείγματος είναι μικρότερες από τη διάμεσο και επίσης το πολύ 50% των τιμών του δείγματος είναι μεγαλύτερες από τη διάμεσο. Σε αύξουσα διάταξη των $x_1, x_2, \ldots, x_\nu$, τη θέση της διαμέσου δίνει ο αριθμός $0.5(\nu+1)$ εφόσον είναι ακέραιος, ενώ αν δεν είναι ακέραιος, τότε η διάμεσος είναι ίση με το ημιάθροισμα των δύο τιμών που οι θέσεις τους είναι οι πλησιέστερες στον αριθμό $0.5(\nu+1)$. • ***p-ποσοστιαία σημεία του δείγματος,*** x_p **,** $0 < p < 1$ Το πολύ $100p\%$ των τιμών του δείγματος είναι μικρότερες από το p-ποσοστιαίο σημείο και το πολύ $100(1-p)\%$ των τιμών του δείγματος είναι μεγαλύτερες από το p-ποσοστιαίο σημείο. Σε αύξουσα διάταξη των $x_1, x_2, \ldots, x_\nu$, τη θέση του p-ποσοστιαίου σημείου δίνει ο αριθμός $p(\nu+1)$ εφόσον είναι ακέραιος, ενώ αν δεν είναι ακέραιος, τότε το p-ποσοστιαίο σημείο εκτιμάται με παρεμβολή μεταξύ των δύο τιμών που οι θέσεις τους είναι οι πλησιέστερες στον αριθμό $p(\nu+1)$. • ***Τεταρτημόρια,*** Q_1, Q_2, Q_3 $Q_1 = x_{0.25}$, $Q_2 = x_{0.5} = \delta$, $Q_3 = x_{0.75}$ Αν (έχει) γίνει ομαδοποίηση των τιμών του δείγματος σε *k* κλάσεις: • Η ***κορυφή*** υπολογίζεται από τον τύπο $M_0 = L_i + \frac{\Delta_1}{\Delta_1 + \Delta_2} c_i$ όπου, L_i το κάτω άκρο της επικρατούσας κλάσης, δηλαδή της κλάσης με τη μεγαλύτερη συχνότητα, c_i το πλάτος της επικρατούσας κλάσης, $\Delta_1 = \nu_i - \nu_{i-1}$ και $\Delta_2 = \nu_i - \nu_{i+1}$ όπου ν_i η συχνότητα της επικρατούσας κλάσης. • Στον τύπο υπολογισμού του ***δειγματικού μέσου*** $\bar{x} = \frac{1}{\nu}\sum_{i=1}^{k} \nu_i y_i = \sum_{i=1}^{k} f_i y_i$ τα y_i, $i = 1, 2, \ldots, k$ είναι οι κεντρικές τιμές των κλάσεων. • Η ***διάμεσος*** υπολογίζεται από τον τύπο $\delta = L_i + \frac{0.5\nu - N_{i-1}}{\nu_i} c_i$ όπου, L_i το κάτω άκρο της μεσαίας κλάσης, δηλαδή της κλάσης στην οποία ανήκει η διάμεσος, c_i το πλάτος της

μεσαίας κλάσης, ν_i η συχνότητα της μεσαίας κλάσης και N_{i-1} η αθροιστική συχνότητα της προηγούμενης κλάσης από τη μεσαία.

- Τα ***p-ποσοστιαία σημεία*** υπολογίζονται από τον τύπο

$$x_p = L_i + \frac{p\nu - N_{i-1}}{\nu_i} c_i$$

όπου, L_i το κάτω άκρο της κλάσης στην οποία βρίσκεται το x_p, c_i το πλάτος της, ν_i η συχνότητά της και N_{i-1} η αθροιστική συχνότητα της προηγούμενης κλάσης.

Μέτρα μεταβλητότητας/διασποράς

- ***Εύρος***

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

- ***Ενδοτεταρτημοριακό εύρος***

$$Q_3 - Q_1$$

- ***Δειγματική διακύμανση***

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

$$= \frac{1}{\nu-1}\sum_{i=1}^{k}(y_i - \bar{x})^2\nu_i = \frac{1}{\nu-1}\left(\sum_{i=1}^{k}\nu_i y_i^2 - \nu\bar{x}^2\right)$$

- ***Δειγματική τυπική απόκλιση***

$$s = \sqrt{\frac{1}{\nu-1}\sum_{i=1}^{\nu}(x_i - \bar{x})^2} = \sqrt{\frac{1}{\nu-1}\left(\sum_{i=1}^{\nu} x_i^2 - \nu\bar{x}^2\right)} =$$

$$= \sqrt{\frac{1}{\nu-1}\sum_{i=1}^{k}(y_i - \bar{x})^2\nu_i} = \sqrt{\frac{1}{\nu-1}\left(\sum_{i=1}^{k}\nu_i y_i^2 - \nu\bar{x}^2\right)}$$

- ***Συντελεστής μεταβλητότητας***

$$CV = \frac{s}{|\bar{x}|}100\%$$

Αν (έχει) γίνει ομαδοποίηση των τιμών του δείγματος σε k κλάσεις τα y_i, $i = 1, 2, \ldots, k$ είναι οι κεντρικές τιμές των κλάσεων.

Μέτρα λοξότητας και μέτρα κύρτωσης

- ***Συντελεστές ασυμμετρίας του Pearson***

$$\gamma_1 = \frac{\bar{x} - M_0}{s} \ , \ \gamma_2 = \frac{3(\bar{x} - \delta)}{s}$$

Αν $\gamma_1 = \gamma_2 = 0$ η κατανομή είναι συμμετρική

- ***Συντελεστής ασυμμετρίας του Bowley***

$$S_A = \frac{Q_1 + Q_3 - 2Q_2}{Q_3 - Q_1}$$

Αν $S_A = 0$ η κατανομή είναι συμμετρική

- ***Συντελεστής κύρτωσης του Pearson***

$$\beta_2 = \frac{m_4}{(m_2)^2}$$

Αν $\beta_2 = 3$ η κατανομή είναι μεσόκυρτη, αν $\beta_2 > 3$ είναι πλατύκυρτη και αν $\beta_2 < 3$ είναι λεπτόκυρτη

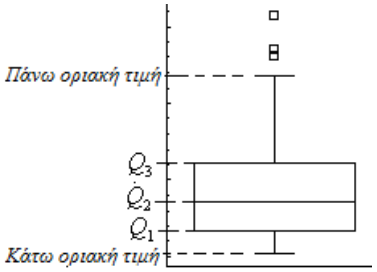
β) Ποιοτικές μεταβλητές

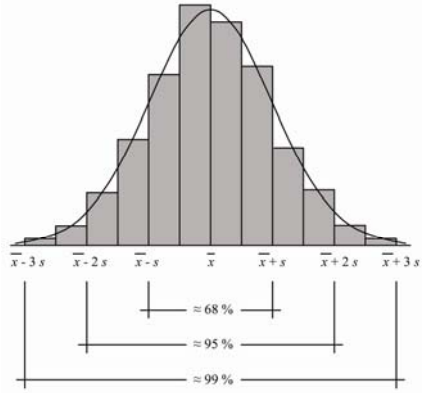
Ορίζεται (και έχει νόημα) μόνο η ***κορυφή*** της κατανομής

γ) Μεταβλητές διεύθυνσης και κατεύθυνσης

Μέτρα θέσης

- ***Μέση κατεύθυνση*** $\bar{\vartheta}$

	ν κατευθύνσεων $\vartheta_1, \vartheta_2, \ldots, \vartheta_\nu$. $$\overline{\vartheta} = \tau o \xi \varepsilon \phi \frac{\sum_{i=1}^{\nu} \eta\mu\vartheta_i}{\sum_{i=1}^{\nu} \sigma\upsilon\nu\vartheta_i}$$ σε συνδυασμό με το πρόσημο των $$x_r = \sum_{i=1}^{\nu} \eta\mu\vartheta_i \text{ και } y_r = \sum_{i=1}^{\nu} \sigma\upsilon\nu\vartheta_i .$$ **Μέτρα μεταβλητότητας** • ***Διακύμανση*** ν κατευθύνσεων $\vartheta_1, \vartheta_2, \ldots, \vartheta_\nu$. $S^2 = 1 - \bar{r}$ $s^2 = 2(1 - \bar{r})$ $s_0^2 = -2\ln\bar{r}$ Όπου, $\bar{r} = \frac{r}{\nu} = \frac{\sqrt{x_r^2 + y_r^2}}{\nu}$ • ***Τυπική απόκλιση*** ν κατευθύνσεων $\vartheta_1, \vartheta_2, \ldots, \vartheta_\nu$. $s = \sqrt{2(1-\bar{r})}$ $s_0 = \sqrt{-2\ln\bar{r}}$ Η μέση διεύθυνση, η διακύμανση και η τυπική απόκλιση ν **διευθύνσεων** $\vartheta_1, \vartheta_2, \ldots, \vartheta_\nu$, υπολογίζεται όπως τα αντίστοιχα μέτρα ν **κατευθύνσεων** αφού προηγουμένως οι τιμές $\vartheta_1, \vartheta_2, \ldots, \vartheta_\nu$ μετασχηματισθούν κατάλληλλα
Μέτρα θέσης και μεταβλητότητας γραμμικού μετασχηματισμού των παρατηρήσεων/δεδομένων	Αν $t_i = \alpha x_i + \beta$ τότε • $\bar{t} = \alpha\bar{x} + \beta$ • $s_t^2 = \alpha^2 s_x^2$ • $s_t = \lvert\alpha\rvert s_x$ • $\delta_t = \alpha\delta_x + \beta$ • $M_{0t} = \alpha M_{0x} + \beta$
Θηκόγραμμα	 Πάνω οριακή τιμή: η μεγαλύτερη τιμή του δείγματος που είναι $\leq Q_3 + 1.5(Q_3 - Q_1)$ ή $\leq Q_3 + 3(Q_3 - Q_1)$ Κάτω οριακή τιμή: η μικρότερη τιμή του δείγματος που είναι $\geq Q_1 - 1.5(Q_3 - Q_1)$ ή $\geq Q_1 - 3(Q_3 - Q_1)$

Ο εμπειρικός κανόνας	Αν η κατανομή του δείγματος προσομοιάζει με μια κανονική κατανομή (έχει κωδωνοειδή μορφή), τότε • στο διάστημα $(\bar{x}-s,\ \bar{x}+s)$ βρίσκεται περίπου το 68% των παρατηρήσεων • στο διάστημα $(\bar{x}-2s,\ \bar{x}+2s)$ βρίσκεται περίπου το 95% των παρατηρήσεων • στο διάστημα $(\bar{x}-3s,\ \bar{x}+3s)$ βρίσκονται περίπου όλες οι παρατηρήσεις (πάνω από το 99%). 
Η ανισότητα *Chebyshev*	Το ποσοστό των τιμών του δείγματος που βρίσκονται στο διάστημα $(\bar{x}-ks,\ \bar{x}+ks)$ είναι **τουλάχιστον** $1-(1/k^2)$,

9.5 Προβλήματα και Ασκήσεις

1. Στο κατάστημα ενός αγροτοτουριστικού συνεταιρισμού, πωλείται μέλι τεσσάρων ειδών (ανθέων, ελάτης, θυμαρίσιο και πεύκου), παραγωγής τριών ντόπιων μελισσοκόμων (*Α, Β* και *Γ*), σε γυάλινα βάζα τριών μεγεθών (μικρού, μεσαίου και μεγάλου). Επιλέξαμε τυχαία από τα ράφια του καταστήματος 25 βάζα μελιού και για κάθε ένα καταγράψαμε τον παραγωγό. Τα δεδομένα που προέκυψαν φαίνονται στον πίνακα που ακολουθεί.

Α	*Β*	*Α*	*Γ*	*Α*	*Α*	*Α*	*Γ*	*Γ*	*Γ*	*Α*	*Β*	*Α*
Β	*Β*	*Α*	*Β*	*Γ*	*Α*	*Β*	*Γ*	*Γ*	*Γ*	*Β*	*Α*	

α) Ποια είναι η *δειγματοληπτική μονάδα*; β) Ποιας *μεταβλητής* καταγράψαμε τιμές; Είναι *ποσοτική* ή *ποιοτική*; γ) Από ποιον *πληθυσμό* πήραμε το τυχαίο δείγμα; δ) Να κατασκευάσετε το *ραβδόγραμμα* και το *κυκλικό διάγραμμα* της κατανομής του δείγματος. ε) Ποιο ποσοστό των βάζων μελιού που επελέγησαν, παράγεται από τον μελισσοκόμο *Γ*; στ) Ποια είναι η *κορυφή* του δείγματος;

2. *Συνέχεια της Άσκησης 9.1*: Για καθένα από τα 25 βάζα μελιού που επιλέξαμε τυχαία από τα ράφια του καταστήματος, καταγράψαμε επίσης α) το είδος του μελιού (ανθέων, ελάτης, θυμαρίσιο, πεύκου) β) το μέγεθος της συσκευασίας (μικρό, μεσαίο, μεγάλο) γ) την περιεκτικότητα του μελιού σε σάκχαρα δ) την ποσότητα μελιού που περιέχεται σε κάθε βάζο. Για κάθε μια από αυτές τις περιπτώσεις, να απαντήσετε στα ερωτήματα (α), (β) και (γ) της *Άσκησης 9.1*.

3. Είναι γνωστό ότι η πετρελαϊκή ρύπανση των θαλασσών προκαλεί, μεταξύ άλλων, την ανάπτυξη ενός συγκεκριμένου τύπου βακτηρίων. Μια ομάδα ερευνητών, προκειμένου να μελετήσει αυτό το φαινόμενο σε μια θαλάσσια περιοχή που έχει πληγεί από πετρελαϊκή ρύπανση, πήρε νερό από 10 διαφορετικά σημεία αυτής της περιοχής και έκανε σχετικές μετρήσεις. Συγκεκριμένα, μέτρησε τον αριθμό, έστω *X*, αυτών των βακτηρίων ανά 100 *milliliters* νερού. Οι τιμές, $x_1, x_2, ..., x_{10}$, της μεταβλητής X στα δέκα σημεία ήταν

   49, 70, 54, 67, 59, 40, 61, 69, 71, 52.

   α) Ποιον *πληθυσμό* μελετούν οι ερευνητές; β) Να υπολογίσετε και να ερμηνεύσετε τα *μέτρα θέσης* και *διασποράς* της κατανομής του δείγματος. γ) Να κατασκευάσετε το *θηκόγραμμα* του δείγματος και με βάση αυτό να περιγράψετε την κατανομή του.

4. Ένας φοιτητής, στο πλαίσιο της πτυχιακής του εργασίας, μελέτησε μεταξύ άλλων, την ποσότητα νατρίου, έστω *X*, που περιέχεται στο κασέρι συνήθους τύπου (όχι *light*) που παράγει μια γνωστή γαλακτοβιομηχανία. Τα αποτελέσματα εννέα σχετικών μετρήσεων που έκανε ο φοιτητής σε κασέρι που επέλεξε τυχαία από εννέα παρτίδες παραγωγής της γαλακτοβιομηχανίας, ήταν *(σε mg/100gr)*

   340, 300, 340, 320, 320, 290, 330, 320, 310.

   α) Να υπολογίσετε και να ερμηνεύσετε τα *μέτρα θέσης* και *διασποράς* της κατανομής του δείγματος. β) Να κατασκευάσετε το *θηκόγραμμα* της κατανομής του δείγματος.

5. *Συνέχεια της Άσκησης 9.4:* Ο φοιτητής μελέτησε επίσης, την ποσότητα νατρίου στο κασέρι τύπου *light* της ίδιας γαλακτοβιομηχανίας. Τα αποτελέσματα οκτώ σχετικών μετρήσεων ήταν *(σε mg/100gr)*

   300, 300, 310, 290, 280, 280, 285, 275.

Να συγκρίνετε την κατανομή αυτού του δείγματος με την κατανομή του προηγούμενου δείγματος (ως προς τη συμμετρία, τη θέση και τη μεταβλητότητά τους).

6. Η πτυχιακή μελέτη ενός φοιτητή αφορούσε, μεταξύ άλλων, στον αριθμό των πετάλων, έστω *X*, των ανθέων μιας συγκεκριμένης ποικιλίας ενός φυτού που καλλιεργείται στο νομό Κοζάνης. Στο πλαίσιο αυτής της μελέτης, ο φοιτητής μέτρησε τον αριθμό των πετάλων σε 115 άνθη της συγκεκριμένης ποικιλίας που επέλεξε τυχαία από καλλιέργειες του νομού Κοζάνης. Τα αποτελέσματα αυτών των μετρήσεων φαίνονται στον πίνακα που ακολουθεί.

7	5	8	7	5	5	6	6	5	7	5	5	5	9	6	8	5
5	5	6	6	5	5	6	5	9	6	5	5	7	6	6	7	5
7	5	5	6	6	5	6	5	6	5	5	5	5	6	6	5	5
8	5	5	5	5	6	5	5	5	6	5	5	6	5	5	5	6
7	5	7	5	5	8	5	5	5	6	5	10	5	6	5	5	6
5	7	5	5	5	9	5	5	7	5	5	5	5	6	7	5	5
6	5	6	5	7	5	10	5	6	5	5	5	8				

α) Να υπολογίσετε και να ερμηνεύσετε τα *μέτρα θέσης* και *διασποράς* της κατανομής του δείγματος. β) Να κατασκευάσετε το *θηκόγραμμα* του δείγματος. Τι συμπεραίνετε για την κατανομή του δείγματος; γ) Για κάποιο άνθος βρέθηκε $x = 7$. Τι μπορούμε να πούμε για τη θέση αυτής της τιμής στην κατανομή του δείγματος; δ) Αν $x_{0.98} = 9.68$, τι μπορούμε να πούμε για τη θέση της τιμής $x = 10$ στην κατανομή του δείγματος; ε) Να κατασκευάσετε το *θηκόγραμμα* των *z-τιμών,* $z_1, z_2, ..., z_{115}$, των τιμών $x_1, x_2, ..., x_{115}$ της *X*. Τι συμπεραίνετε για την κατανομή των *z-τιμών;*

7. Σε μια περιοχή του Μαινάλου αιχμαλωτίσθηκαν από μια ομάδα ερευνητών, με βάση ένα σχέδιο τυχαίας δειγματοληψίας, 100 αλεπούδες για να ελεγχθούν ως προς το αν έχουν προσβληθεί από παράσιτα (ενός συγκεκριμένου τύπου). Στη συνέχεια οι ερευνητές κατέγραψαν τον αριθμό, έστω *X,* των παράσιτων που βρέθηκαν ανά αλεπού. Στον πίνακα που ακολουθεί φαίνονται οι συχνότητες όλων των τιμών *x* της μεταβλητής *X* που εμφανίσθηκαν στο δείγμα (μηδέν παράσιτα σε κάθε μία από 69 αλεπούδες, ένα παράσιτο σε κάθε μια από 17 αλεπούδες, 2 παράσιτα σε κάθε μία από 6 αλεπούδες, κτλ).

Αριθμός παράσιτων	0	1	2	3	4	5	6	7	8
Αριθμός αλεπούδων	69	17	6	3	1	2	1	0	1

α) Να υπολογίσετε και να ερμηνεύσετε τα *μέτρα θέσης* και *διασποράς* της κατανομής του δείγματος. β) Να κατασκευάσετε το *θηκόγραμμα* του δείγματος. Τι συμπεραίνετε για την κατανομή του δείγματος; γ) Να υπολογίσετε τα *ποσοστιαία σημεία* $x_{0.95}$ και $x_{0.98}$. Τι μπορούμε να πούμε για τη θέση των τιμών, $x = 4$ και $x = 6$ στην κατανομή του δείγματος;

8. Σε 50 φύλλα πορτοκαλιάς, τυχαία επιλεγμένα, από έναν πορτοκαλεώνα στον κάμπο της Αργολίδας, μετρήθηκε ο αριθμός, έστω *X*, ζωυφίων ανά φύλλο. Στον πίνακα που ακολουθεί φαίνονται οι συχνότητες όλων των τιμών *x* της μεταβλητής *X* που εμφανίσθηκαν στο δείγμα.

Αριθμός ζωυφίων	0	1	2	3	4	5	6	7
Αριθμός φύλλων	2	5	9	11	10	7	4	2

α) Να υπολογίσετε και να ερμηνεύσετε τα *μέτρα θέσης* και *διασποράς* της κατανομής του δείγματος. β) Να κατασκευάσετε το *θηκόγραμμα* του δείγματος. Τι συμπεραίνετε για την κατανομή του δείγματος;

9. Στον πίνακα που ακολουθεί φαίνεται ο αριθμός σταφίδων που περιέχονται σε καθένα από 14, τυχαία επιλεγμένα, μικρά φακελάκια (των 30gr) παραγωγής μιας μεγάλης εταιρείας συσκευασίας τροφίμων και αντίστοιχα, σε 14 τυχαία επιλεγμένα φακελάκια (επίσης των 30gr) οικοτεχνικής παραγωγής. Να συγκρίνετε τις κατανομές των δύο δειγμάτων ως προς τη θέση τους και τη μεταβλητότητά τους.

Αριθμός σταφίδων σε φακελάκια (των 30 gr)

Εταιρείας τροφίμων					**Οικοτεχνικής παραγωγής**				
25	26	25	28	26	25	29	24	24	28
28	28	27	26	27	24	28	22	25	28
24	25	26	26		30	27	28	24	

10. Ένας ερευνητής σχεδίασε και εκτέλεσε ένα πείραμα για να μελετήσει το χρόνο, έστω *X (σε ημέρες)*, που απαιτείται για την αποδόμηση μιας συγκεκριμένης χημικής ουσίας από το μέλι (η ουσία αυτή χρησιμοποιείται για την καταπολέμηση των ακάρεων). Στον πίνακα που ακολουθεί φαίνονται 50 σχετικές παρατηρήσεις.

38	47	32	55	42	40	36	35	45	45	40	35	34
39	50	48	41	40	42	38	30	34	41	33	37	36
43	30	41	46	35	43	30	32	39	31	48	46	36
36	39	41	46	32	33	36	40	37	50	31		

α) Να υπολογίσετε τον *μέσο,* την *τυπική απόκλιση,* την *κορυφή* και τη *διάμεσο* του δείγματος. β) Να ομαδοποιήσετε τις παρατηρήσεις σε 6 κλάσεις με πλάτος 5 ημέρες η κάθε μια και αριστερό άκρο της πρώτης κλάσης τις 30 ημέρες. Να υπολογίσετε και πάλι τον *μέσο,* την *τυπική απόκλιση,* την *κορυφή* και τη *διάμεσο* του δείγματος χρησιμοποιώντας τώρα τις ομαδοποιημένες παρατηρήσεις και να συγκρίνετε τα αποτελέσματα με αυτά του ερωτήματος (α). γ) Να κατασκευάσετε το *ιστόγραμμα συχνοτήτων* της κατανομής του δείγματος με βάση την ομαδοποίηση που κάνατε στο (β). Τι συμπεραίνετε για τη μορφή της; δ) Να σχολιάσετε τη θέση της *κορυφής*, της *διαμέσου* και του *μέσου* του δείγματος σε σχέση με τη μορφή της κατανομής που προκύπτει από το (γ). ε) Να υπολογίσετε τα ποσοστά των παρατηρήσεων που βρίσκονται εντός των διαστημάτων $(\bar{x}-s, \bar{x}+s)$, $(\bar{x}-2s, \bar{x}+2s)$, $(\bar{x}-3s, \bar{x}+3s)$ και να τα συγκρίνετε με τα αντίστοιχα ποσοστά που αναμένονται από την *ανισότητα Chebyshev* και από τον *εμπειρικό κανόνα*.

11. Προκειμένου μια βιομηχανία παραγωγής χάλυβα να εκτιμήσει τη *μέση περιεκτικότητα* του χάλυβα που παράγει σε μαγγάνιο, έκανε με βάση ένα σχέδιο τυχαίας δειγματοληψίας, 40 μετρήσεις. Τα αποτελέσματα των μετρήσεων αυτών *(ποσοστά, %)* φαίνονται στον πίνακα που ακολουθεί.

1.50	1.28	1.54	1.50	1.58	1.40	1.34	1.46	1.52	1.70
1.54	1.46	1.62	1.72	1.38	1.58	1.46	1.44	1.36	1.08
1.60	1.34	1.18	1.44	1.46	1.52	1.58	1.62	1.42	1.34
1.58	1.12	1.56	1.42	1.36	1.44	1.38	1.52	1.58	1.64

α) Να ομαδοποιήσετε τις παρατηρήσεις σε 8 κλάσεις με πλάτος 0.1 η κάθε μια και αριστερό άκρο της πρώτης κλάσης το 1. β) Να κατασκευάσετε το *ιστόγραμμα συχνοτήτων* της κατανομής του δείγματος με βάση την ομαδοποίηση που κάνατε

στο (α). Τι συμπεραίνετε για τη συμμετρία της κατανομής; γ) Να υπολογίσετε τον *μέσο*, τη *διάμεσο* και την *κορυφή* του δείγματος και να σχολιάσετε τη θέση τους σε σχέση με τη μορφή της κατανομής που προκύπτει από το (β). δ) Να κατασκευάσετε το *θηκόγραμμα* της κατανομής του δείγματος.

12. Στον πίνακα που ακολουθεί φαίνεται το ποσοστό (%) οξειδίου του αργιλίου *(aluminum oxide)* σε καθένα από 24 κεραμικά αγγεία που βρέθηκαν σε αρχαιολογικές ανασκαφές που έγιναν σε δύο διαφορετικές περιοχές (*A* και *B*).

Περιοχή Α					**Περιοχή Β**			
14.4	11.6	13.8	11.1	14.6	18.3	18.0	17.7	14.8
11.5	12.4	13.8	13.1	10.9	15.8	18.0	18.3	
10.1	12.5	13.4	12.7		20.8	19.1	16.7	

α) Να κατασκευάσετε το *ιστόγραμμα* και το *πολύγωνο σχετικών συχνοτήτων* των 24 παρατηρήσεων. Παρατηρείτε κάτι αξιοσημείωτο; β) Να κατασκευάσετε το πολύγωνο *σχετικών συχνοτήτων* και το *θηκόγραμμα* των παρατηρήσεων από την περιοχή *A* και αντίστοιχα από την περιοχή *B*. Μπορείτε τώρα να εξηγήσετε γιατί συμβαίνει αυτό που παρατηρήσατε στο (α);

13. *Συνέχεια του Παραδείγματος 9.1.3:* Να υπολογίσετε τα μέτρα θέσης και διασποράς του δείγματος. Ο υπολογισμός να γίνει α) με βάση την ομαδοποίηση που έχουμε κάνει β) από τα πρωτογενή δεδομένα χωρίς κάποια ομαδοποίηση. Επίσης να κατασκευάσετε το θηκόγραμμα της κατανομής του δείγματος.

14. Να εφαρμόσετε κατάλληλες (κατά περίπτωση) μεθόδους περιγραφικής στατιστικής για να περιγράψετε την κατανομή καθενός από τα παρακάτω δείγματα. Να συνοψίσετε τα συμπεράσματά σας σε μια σύντομη παράγραφο.

(α) Στον πίνακα που ακολουθεί φαίνεται η ποσότητα DNA που βρέθηκε στο συκώτι καθενός από 52 ποντίκια.

3.4	13.2	6.7	1.4	1.3	3.8	3.9	2.9	13.2	3.9	2.7
4.4	3.6	1.4	2.4	3.6	3.1	7.5	2.9	7.8	2.7	3.9
3.3	1.7	2.0	4.4	3.3	0.7	3.9	1.6	5.6	3.0	3.4
1.4	3.5	2.8	1.4	1.9	2.3	2.9	2.8	1.5	4.1	5.9
3.1	8.7	2.8	3.8	13.0	3.0	3.0	4.1			

(β) Μετρήθηκε ο χρόνος ζωής πενήντα εξαρτημάτων, τυχαία επιλεγμένων από την αποθήκη του εργοστασίου παραγωγής τους. Οι μετρήσεις αυτές έδωσαν τα ακόλουθα αποτελέσματα *(σε ώρες)*.

46	104	94	114	35	214	15	272	118	193	48	97	37
126	64	5	27	26	57	56	236	72	46	73	38	184
23	85	122	43	159	102	14	73	17	314	143	9	171
120	8	146	117	35	14	263	4	64	113	25		

(γ) Μια ομάδα ερευνητών επέλεξε τυχαία 200 άτομα από έναν πληθυσμό και για καθένα από αυτά κατέγραψε την ομάδα αίματός του. Στον πίνακα που ακολουθεί φαίνεται η συχνότητα κάθε ομάδας αίματος που παρατηρήθηκε στο δείγμα.

	Ομάδα αίματος			
	A	B	AB	O
Παρατηρηθείσα συχνότητα	89	18	12	81

(δ) Τα τελευταία χρόνια παρατηρείται συνεχώς αυξανόμενο ενδιαφέρον για τη μελέτη της συγκέντρωσης τοξικών στοιχείων στον οργανισμό των θαλάσσιων θηλαστικών. Στο πλαίσιο μιας σχετικής μελέτης για τη συγκέντρωση, έστω *X*,

υδραργύρου στο συκώτι ενός είδους αρσενικών δελφινιών, έγιναν σχετικές μετρήσεις σε ένα τυχαίο δείγμα 28 αρσενικών δελφινιών αυτού του είδους με τα ακόλουθα αποτελέσματα *(σε microgr/gr)*.

1.70	101	168	481	252	278	397
1.72	85.40	218	485	329	286	209
8.80	118	180	221	316	315	314
5.90	183	264	406	445	241	318

(ε) Ενδιαφερόμαστε να μελετήσουμε την τυχαία μεταβλητή X που εκφράζει το χρόνο που μεσολαβεί μεταξύ διαδοχικών εκρήξεων του ηφαιστείου *Mauna Loa* της Χαβάης. Στον πίνακα που ακολουθεί φαίνονται 36 τιμές (σε μήνες) της X (αφορούν 37 διαδοχικές εκρήξεις του ηφαιστείου που έγιναν από το 1832 μέχρι το 1950).

126	26	11	6	68	94	23	12
73	73	3	12	41	16	51	
26	23	3	38	38	40	20	
6	21	2	6	50	77	18	
41	18	6	65	37	91	61	

(στ) *Συνέχεια της Άσκησης 9.1*: Στον πίνακα που ακολουθεί φαίνεται για καθένα από τα 25 βάζα μελιού που επιλέξαμε τυχαία από τα ράφια του καταστήματος i) ο παραγωγός *(Α, Β, Γ)* ii) το είδος του μελιού (ανθέων, ελάτης, θυμαρίσιο, πεύκου) iii) το μέγεθος της συσκευασίας (μικρό, μεσαίο, μεγάλο) iv) η περιεκτικότητα του μελιού σε σάκχαρα και v) η ποσότητα μελιού που περιέχεται σε κάθε βάζο.

Παραγωγός	**Είδος**	**Μέγεθος συσκευασίας**	**Περιεκτικότητα σε σάκχαρα (%)**	**Ποσότητα (σε gr)**
Α	Ανθέων	Μικρό	75	250
Β	Ανθέων	Μεσαίο	77	500
Α	Ανθέων	Μεσαίο	70	490
Γ	Ανθέων	Μικρό	78	240
Α	Θυμαρίσιο	Μεγάλο	77	1000
Α	Ανθέων	Μεγάλο	75	950
Α	Πεύκου	Μεγάλο	52	1100
Γ	Πεύκου	Μεσαίο	55	550
Γ	Ανθέων	Μεσαίο	77	450
Γ	Ελάτης	Μεσαίο	60	500
Α	Θυμαρίσιο	Μικρό	77	250
Β	Θυμαρίσιο	Μικρό	75	270
Α	Θυμαρίσιο	Μεγάλο	78	1000
Β	Θυμαρίσιο	Μεγάλο	78	1050
Β	Πεύκου	Μεγάλο	60	1000
Α	Πεύκου	Μεσαίο	50	500
Β	Ελάτης	Μεσαίο	55	550
Γ	Ελάτης	Μικρό	59	250
Α	Ελάτης	Μικρό	60	250
Β	Θυμαρίσιο	Μεσαίο	75	560
Γ	Ανθέων	Μεσαίο	77	500
Γ	Πεύκου	Μεσαίο	55	500
Γ	Ανθέων	Μικρό	77	240
Β	Ανθέων	Μεγάλο	72	990
Α	Ανθέων	Μικρό	75	250

(ζ) Μια ομάδα ερευνητών, στο πλαίσιο ενός πειράματος, ράντισε μια καλλιέργεια σέλινου με παραθείο με σκοπό να εκτιμήσει το υπόλοιπο παραθείου στο σέλινο μετά ορισμένο χρονικό διάστημα από το ράντισμα. Στον πίνακα που ακολουθεί φαίνονται ομαδοποιημένες σε πέντε κλάσεις οι μετρήσεις *(σε milligrams)* που έκανε η ερευνητική ομάδα σε 100 τυχαία επιλεγμένα φυτά.

Ποσότητα παραθείου *(σε mgr)*	**Αριθμός φυτών**
[0, 20)	10
[20, 40)	10
[40, 60)	20
[60, 80)	40
[80, 100)	20

(η) Ένας φοιτητής, στο πλαίσιο της πτυχιακής του εργασίας, επέλεξε τυχαία 100 φύλλα από φυτά της ίδιας οικογένειας που καλλιεργούνται στο θερμοκήπιο του Πανεπιστημίου του και μέτρησε το μήκος τους, έστω X (*σε cm*). Τα αποτελέσματα των μετρήσεων αυτών φαίνονται, ομαδοποιημένα σε πέντε κλάσεις, στον πίνακα που ακολουθεί.

Μήκος φύλλου (*σε cm*)	**Αριθμός φύλλων**
[0, 4)	51
[4, 8)	20
[8, 12)	16
[12, 16)	4
[16, 20)	9

(θ) Στον πίνακα που ακολουθεί, δίνονται ομαδοποιημένες σε οκτώ κλάσεις, 200 παρατηρήσεις, $x_1, x_2, ..., x_{200}$, για το ύψος της ετήσιας βροχόπτωσης, έστω X *(σε cm),* που ελήφθησαν από 200 μετεωρολογικούς σταθμούς μιας χώρας.

Ύψος βροχόπτωσης *(σε cm)*	**Αριθμός σταθμών**
[20, 30)	11
[30, 40)	14
[40, 50)	31
[50, 60)	48
[60, 70)	41
[70, 80)	30
[80, 90)	15
[90, 100)	10

(ι) Ένας ερευνητής σχεδίασε και εκτέλεσε το εξής πείραμα. Σε έναν κλειστό διάδρομο στο τέλος του οποίου υπήρχαν τρεις έξοδοι διαφορετικού χρώματος (πράσινη, κόκκινη και μπλε), απελευθέρωσε ένα ποντίκι 90 φορές και κατέγραψε πόσες φορές αυτό διέφυγε από την πράσινη έξοδο, πόσες από την κόκκινη και πόσες από τη μπλε. Η συχνότητα που παρατηρήθηκε για κάθε έξοδο διαφυγής φαίνεται στον πίνακα που ακολουθεί.

	Έξοδος διαφυγής		
	Πράσινη	Κόκκινη	Μπλε
Παρατηρηθείσα συχνότητα	20	39	31

15. Μια βιομηχανία τροφίμων παράγει από τρεις γραμμές παραγωγής (Γ_1, Γ_2 και Γ_3) ελαφρά συμπυκνωμένο χυμό τομάτας σε συσκευασίες των 200gr. Το τμήμα ποιοτικού ελέγχου της βιομηχανίας, όταν διαπιστώνει ότι κάποιο προϊόν είναι ελαττωματικό το κατατάσσει σε μία από τέσσερις κατηγορίες (A_1, A_2, A_3 και A_4)

ανάλογα με το είδος και τη σοβαρότητα των ελαττωμάτων που παρουσιάζει. Επίσης, καταγράφει από ποια γραμμή παραγωγής παρήχθη. Στον πίνακα που ακολουθεί φαίνεται πώς κατανέμονται στις τέσσερις κατηγορίες A_1, A_2, A_3, A_4 και τις τρεις γραμμές παραγωγής Γ_1, Γ_2, Γ_3, 309 προϊόντα που βρέθηκαν ελαττωματικά.

		Γραμμή Παραγωγής		
		Γ_1	Γ_2	Γ_3
Κατηγορία κατάταξης	A_1	15	26	33
	A_2	21	31	17
	A_3	45	34	49
	A_4	13	5	20

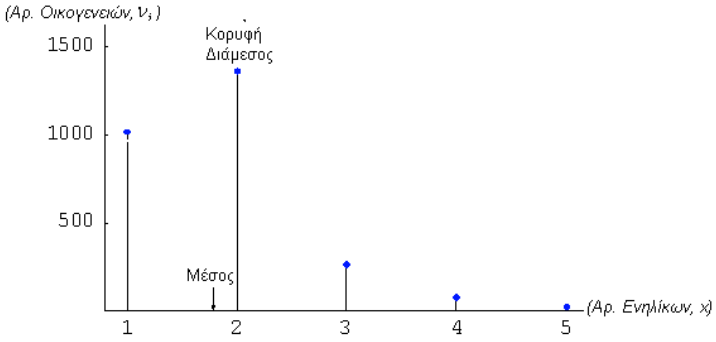
Να γίνει γραφική αναπαράσταση της κατανομής αυτών των δεδομένων.

16. Στο πλαίσιο μιας έρευνας αγοράς, ρωτήθηκε καθένας από 670 δυνητικούς αγοραστές ενός νέου προϊόντος να δηλώσει ποιο από τρία μοντέλα (M1, M2 και M3) του νέου προϊόντος προτιμά καθώς και να κατατάξει τον εαυτό του σε έναν από τέσσερις τύπους καταναλωτή που του προτάθηκαν (T1, T2, T3 και T4). Τα δεδομένα που προέκυψαν φαίνονται στον πίνακα που ακολουθεί.

		Τύπος καταναλωτή			
		T1	T2	T3	T4
Μοντέλο	M1	20	22	56	38
	M2	40	44	68	42
	M3	80	90	75	95

Να γίνει γραφική αναπαράσταση της κατανομής αυτών των δεδομένων

17. Έστω X τυχαία μεταβλητή που εκφράζει τον αριθμό των ενηλίκων μελών *(18 ετών και άνω),* των οικογενειών μιας πολιτείας των Η.Π.Α το έτος 2002. Στο διάγραμμα που ακολουθεί φαίνεται η κατανομή ενός αντιπροσωπευτικού δείγματος τιμών της X, μεγέθους $\nu = 2660$.

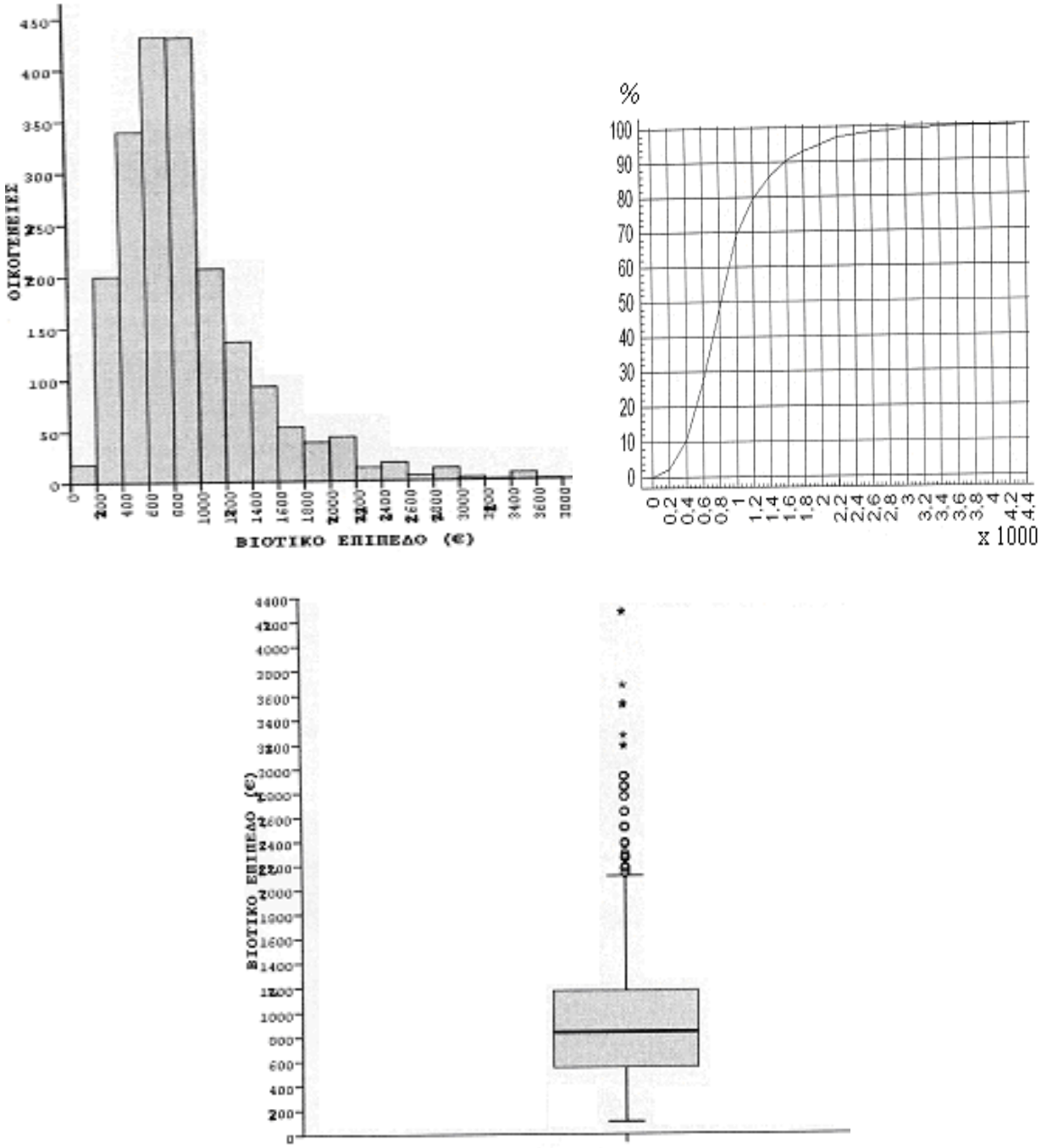


Δίνεται επίσης, ο *πίνακας συχνοτήτων* του δείγματος.

Αριθμός ενηλίκων	1	2	3	4	5
Αριθμός οικογενειών	1020	1300	250	70	20

α) Να επαληθεύσετε ότι οι θέσεις που έχουν σημειωθεί στο *διάγραμμα συχνοτήτων*, αντιστοιχούν πράγματι στον *μέσο*, τη *διάμεσο* και την *κορυφή* της κατανομής του δείγματος. β) Μπορείτε να εξηγήσετε γιατί, ενώ η κατανομή εμφανώς παρουσιάζει *θετική ασυμμετρία*, ο *μέσος* βρίσκεται αριστερά της *διαμέσου* και της *κορυφής*;

18. Έστω X τυχαία μεταβλητή που εκφράζει το μηνιαίο βιοτικό επίπεδο των μελών των οικογενειών στο Νομό Αττικής το έτος 2008[16,17]. Στα σχήματα που ακολουθούν παρουσιάζεται η κατανομή ενός αντιπροσωπευτικού δείγματος τιμών της X μεγέθους $\nu = 2051$. Για την κατανομή αυτή, δίνεται επίσης, ο *δειγματικός μέσος*, $\bar{x} = 923.12\,€$ και η *δειγματική τυπική απόκλιση*, $s = 538.01€$.



[16] *Το μηνιαίο βιοτικό επίπεδο μιας οικογένειας είναι ίδιο για όλα τα μέλη της οικογένειας και προκύπτει από τη διαίρεση του συνολικού καθαρού μηνιαίου εισοδήματος της οικογένειας με ένα σταθμικό άθροισμα των μελών της. Το σταθμικό άθροισμα προκύπτει ως εξής: για τον πρώτο ενήλικα βάρος 1, για κάθε άλλο ενήλικα και κάθε παιδί άνω των 14 ετών βάρος 0.5 και για κάθε παιδί κάτω των 14 ετών βάρος 0.3. Για παράδειγμα, το μηνιαίο βιοτικό επίπεδο μιας οικογένειας με συνολικό καθαρό μηνιαίο εισόδημα 2800€ που αποτελείται από τον πατέρα, τη μητέρα, τη γιαγιά, ένα παιδί 8 ετών και ένα παιδί 16 ετών είναι,* $2800/(1+0.5+0.5+0.3+0.5) = 1000\,€$.

[17] *Από θέμα στις εισαγωγικές εξετάσεις για την Εθνική Σχολή Δημόσιας Διοίκησης.*

α) Ποιον *πληθυσμό* μελετάμε και ποια είναι η *δειγματοληπτική μονάδα.* β) Να υπολογίσετε (κατά προσέγγιση) και να ερμηνεύσετε τη *διάμεσο* και το *1ο* και *3ο τεταρτημόριο* της κατανομής του δείγματος. γ) Τι ποσοστό (περίπου) των οικογενειών του δείγματος έχει μηνιαίο βιοτικό επίπεδο πάνω από 2000*€;* δ) Αν είστε εκπρόσωπος των εργαζομένων, ποιες πληροφορίες από την κατανομή του δείγματος θα χρησιμοποιούσατε ως επιχειρήματα σε μια συνάντηση με τον υπουργό οικονομικών; ε) Τι ποσοστό (περίπου) των οικογενειών του δείγματος βρίσκεται κάτω από το όριο της φτώχιας (το όριο της φτώχιας ορίζεται ως το 60% του διάμεσου μηνιαίου βιοτικού επιπέδου). στ) Αν η *z-τιμή* μιας τιμής του δείγματος είναι -1.3, ποια είναι η θέση αυτής της τιμής στην κατανομή του δείγματος; ζ) Αν μια τιμή του δείγματος είναι 1500*€*, ποια είναι η θέση της στην κατανομή του δείγματος; η) Τι ποσοστό (περίπου) των τιμών του δείγματος βρίσκεται στο διάστημα $(\bar{x}-2s,\ \bar{x}+2s)$; Συμφωνεί αυτό το ποσοστό με αυτό που αναμένουμε από την *ανισότητα Chebyshev*;

19. Με τη βοήθεια κατάλληλου λογισμικού να εφαρμόσετε κατάλληλες στατιστικές μεθόδους για να περιγράψετε την κατανομή καθενός από τα παρακάτω δείγματα. Να συνοψίσετε τα συμπεράσματά σας σε μια σύντομη παράγραφο.

(α) Στον πίνακα που ακολουθεί δίνονται (σε μοίρες από το βορρά και κατά τη φορά της κίνησης των δεικτών του ωρολογίου) οι κατευθύνσεις των σταυρωτών στρωματώσεων σε ένα σχηματισμό ψαμμίτη.

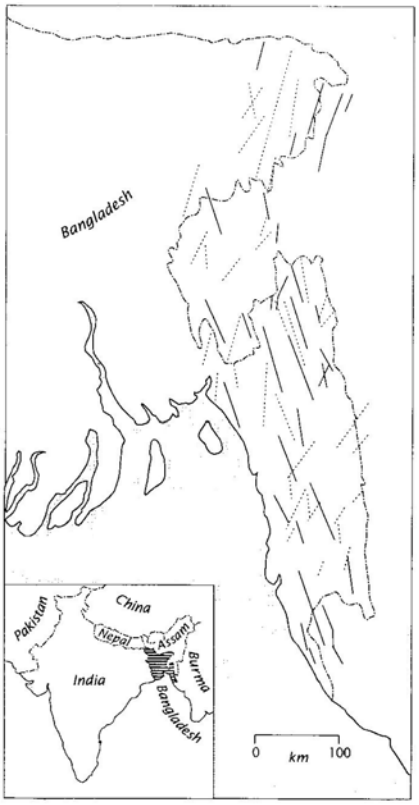
121	113	97	113	100	118	354	256	220	192
283	128	145	335	333	6	342	45	54	169
172	160	146	177	179	169	33	14	25	4
338	321	335	22	338	128	44	59	199	208
28	30	24	58	199	208	175	197	328	339
215	176	85	295	299	1	16	334		

(β) Στον πίνακα που ακολουθεί δίνονται (σε μοίρες από το βορρά και κατά τη φορά της κίνησης των δεικτών του ωρολογίου) οι κατευθύνσεις των αμμορυτίδων δύο σχηματισμών ψαμμίτη.

Σχηματισμός Α΄									
216	118	223	305	242	198	172	222	155	233
269	238	189	219	111	217	141	201	260	276
182	212	245	221	177	248	192	210	222	251
214	228	217	262	280	234	244	218	208	191
201									

Σχηματισμός Β΄									
72	216	255	217	341	222	257	278	242	148
171	251	190	196	290	288	313	300	275	135
255	227	163	96	35	93	106	320	349	215
333	237	15	105	118	179	205	180	271	223
123									

(γ) Στον πίνακα που ακολουθεί δίνονται (σε μοίρες από το βορρά και κατά τη φορά της κίνησης των δεικτών του ωρολογίου) οι διευθύνσεις των αξονικών επιπέδων των αντικλίνων καθώς και των Landsat γραμμώσεων στο ανατολικό Μπανγκλαντές *(Davis, J.C., 2002)*.



Διεύθυνση των αξονικών επιπέδων των αντικλίνων				**Διεύθυνση των Landsat γραμμώσεων**			
12	16	14	5	350	32	15	8
192	202	169	163	214	192	16	26
186	186	24	344	356	218	198	221
343	346	161	341	350	18	221	342
339	150	169	336	160	205	35	337
351	156	159	352	2	171	196	14
152	150	341	181	184	246	175	25
348	156	156		354	213	26	212
330	162	20		42	354	13	202