ΜΑΘΗΜΑΤΙΚΑ
ΚΑΙ
ΣΤΟΙΧΕΙΑ ΣΤΑΤΙΣΤΙΚΗΣ

Γ΄ ΓΕΝΙΚΟΥ ΛΥΚΕΙΟΥ

Τόμος 2ος

ΣΤΟΙΧΕΙΑ ΑΡΧΙΚΗΣ ΕΚΔΟΣΗΣ

ΣΥΓΓΡΑΦΕΙΣ:

Αδαμόπουλος Λεωνίδας
Επ. Σύμβουλος Παιδαγωγικού Ινστιτούτου

Δαμιανού Χαράλαμπος
Αναπλ. Καθηγητής Παν/μίου Αθηνών

Σβέρκος Ανδρέας
Σχολικός Σύμβουλος

ΚΡΙΤΕΣ:

Κουνιάς Στρατής
Καθηγητής Παν/μίου Αθηνών

Μακρής Κωνσταντίνος
Σχολικός Σύμβουλος

Τσικαλουδάκης Γεώργιος
Καθηγητής Β/θμιας Εκπαίδευσης

Γλωσσική Επιμέλεια:

Μπουσούνη Λία
Καθηγήτρια Β/θμιας Εκπαίδευσης

Δακτυλογράφηση:

Μπολιώτη Πόπη

Σχήματα:

Μπούτσικας Μιχάλης

ΣΤΟΙΧΕΙΑ ΕΠΑΝΕΚΔΟΣΗΣ

Η επανέκδοση του παρόντος βιβλίου πραγματοποιήθηκε από το Ινστιτούτο Τεχνολογίας Υπολογιστών & Εκδόσεων «Διόφαντος» μέσω ψηφιακής μακέτας, η οποία δημιουργήθηκε με χρηματοδότηση από το ΕΣΠΑ / ΕΠ «Εκπαίδευση & Διά Βίου Μάθηση» / Πράξη «ΣΤΗΡΙΖΩ».

Ευρωπαϊκή Ένωση
Ευρωπαϊκό Κοινωνικό Ταμείο
ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΕΚΠΑΙΔΕΥΣΗ ΚΑΙ ΔΙΑ ΒΙΟΥ ΜΑΘΗΣΗ
επένδυση στην κοινωνία της γνώσης
ΥΠΟΥΡΓΕΙΟ ΠΑΙΔΕΙΑΣ & ΘΡΗΣΚΕΥΜΑΤΩΝ, ΠΟΛΙΤΙΣΜΟΥ & ΑΘΛΗΤΙΣΜΟΥ
ΕΙΔΙΚΗ ΥΠΗΡΕΣΙΑ ΔΙΑΧΕΙΡΙΣΗΣ
ΕΣΠΑ 2007-2013
ΕΥΡΩΠΑΪΚΟ ΚΟΙΝΩΝΙΚΟ ΤΑΜΕΙΟ
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Οι διορθώσεις πραγματοποιήθηκαν κατόπιν έγκρισης του Δ.Σ. του Ινστιτούτου Εκπαιδευτικής Πολιτικής

Η αξιολόγηση, η κρίση των προσαρμογών και η επιστημονική επιμέλεια του προσαρμοσμένου βιβλίου πραγματοποιείται από τη Μονάδα Ειδικής Αγωγής του Ινστιτούτου Εκπαιδευτικής Πολιτικής.

Η προσαρμογή του βιβλίου για μαθητές με μειωμένη όραση από το ΙΤΥΕ – ΔΙΟΦΑΝΤΟΣ πραγματοποιείται με βάση τις προδιαγραφές που έχουν αναπτυχθεί από ειδικούς εμπειρογνώμονες για το ΙΕΠ.

ΠΡΟΣΑΡΜΟΓΗ ΤΟΥ ΒΙΒΛΙΟΥ
ΓΙΑ ΜΑΘΗΤΕΣ ΜΕ ΜΕΙΩΜΕΝΗ ΟΡΑΣΗ

ΙΤΥΕ - ΔΙΟΦΑΝΤΟΣ

ΥΠΟΥΡΓΕΙΟ ΠΑΙΔΕΙΑΣ, ΕΡΕΥΝΑΣ ΚΑΙ ΘΡΗΣΚΕΥΜΑΤΩΝ
ΙΝΣΤΙΤΟΥΤΟ ΕΚΠΑΙΔΕΥΤΙΚΗΣ ΠΟΛΙΤΙΚΗΣ

ΑΔΑΜΟΠΟΥΛΟΣ ΛΕΩΝΙΔΑΣ
ΔΑΜΙΑΝΟΥ ΧΑΡΑΛΑΜΠΟΣ
ΣΒΕΡΚΟΣ ΑΝΔΡΕΑΣ

Η συγγραφή και η επιστημονική επιμέλεια του βιβλίου πραγματοποιήθηκε υπό την αιγίδα του Παιδαγωγικού Ινστιτούτου

ΜΑΘΗΜΑΤΙΚΑ ΚΑΙ ΣΤΟΙΧΕΙΑ ΣΤΑΤΙΣΤΙΚΗΣ

Γ΄ ΓΕΝΙΚΟΥ ΛΥΚΕΙΟΥ

Τόμος 2ος

ΙΝΣΤΙΤΟΥΤΟ ΤΕΧΝΟΛΟΓΙΑΣ ΥΠΟΛΟΓΙΣΤΩΝ
ΚΑΙ ΕΚΔΟΣΕΩΝ «ΔΙΟΦΑΝΤΟΣ»

2 ΣΤΑΤΙΣΤΙΚΗ

Εισαγωγή

Ο όρος "Στατιστική" ενδεχομένως να προέρχεται από τη λατινική λέξη "status" (πολιτεία, κράτος) η οποία, χρησιμοποιήθηκε αρχικά για το χαρακτηρισμό αριθμητικών δεδομένων που αναφέρονται κυρίως στον πληθυσμό μιας χώρας. Μπορεί όμως να προέρχεται από την αρχαία ελληνική λέξη στατίζω (τοποθετώ, ταξινομώ, συμπεραίνω). Με την εμφάνιση της Στατιστικής και στα πρώτα στάδια της ανάπτυξής της οι άνθρωποι την ταύτισαν με την παράθεση τεράστιων πινάκων με δεδομένα σχετικά με τους θανάτους, τις γεννήσεις, τους φόρους, τα προϊόντα, τους άνδρες σε στρατεύσιμη ηλικία κτλ., προσπαθώντας έτσι να περιγράψουν διάφορα δημογραφικά, οικονομικά και πολιτικά φαινόμενα. Η αρχαιότερη ίσως συλλογή στατιστικών στοιχείων θεωρείται η απογραφή πληθυσμού που έγινε το 2238 π.Χ. στην Κίνα από τον αυτοκράτορα Yao. Επίσης, στοιχειώδεις απογραφές φαίνεται να έχουν πραγματοποιηθεί από τους Σίνες, τους Αιγυπτίους και τους Πέρσες.

Ο όρος Στατιστική αναφέρεται επίσης και από το Σωκράτη (Ξενοφώντος "Απομνημονεύματα") και από τον Αριστοτέλη ("Πολιτεία"). Όπως γνωρίζουμε απογραφή πληθυσμού είχε επίσης διαταχθεί και από τον καίσαρα Αύγουστο στην περίοδο της γέννησης του Χριστού. Στην αρχαιότητα, η συγκέντρωση στατιστικών στοιχείων είχε στόχο τον εντοπισμό των πολιτών που είχαν υποχρέωση να υπηρετήσουν ως πολεμιστές ή να πληρώσουν φόρο. Συστηματική συλλογή δεδομένων για τον πληθυσμό και την οικονομία άρχισε κατά τη διάρκεια της Αναγέννησης στις πόλεις Βενετία και Φλωρεντία στην Ιταλία, και γρήγορα επεκτάθηκε και σε άλλες χώρες της Δυτικής Ευρώπης. Ο μεγάλος ρυθμός θνησιμότητας στην Ευρώπη οφειλόταν στις επιδημικές ασθένειες, στους πολέμους και στις λιμοκτονίες. Στις αρχικές καταγραφές των θανάτων από την πανώλη, τη φοβερή ασθένεια που εμφανίστηκε το 1348 και κράτησε πάνω από 400 χρόνια, προστέθηκαν στη συνέχεια και οι θάνατοι από άλλες αιτίες. Στα 1620 ο Άγγλος εμπορευόμενος Graunt από δειγματοληπτική έρευνα που έκανε σε οικογένειες του Λονδίνου βρήκε ότι σε κάθε 88 άτομα υπήρχαν 3 θάνατοι. Χρησιμοποιώντας τους καταλόγους του Λονδίνου, που έδιναν 13.200 θανάτους το 1620, εκτίμησε τον πληθυσμό του Λονδίνου το έτος αυτό στα 387.200 άτομα.

Μια πραγματικά σπουδαία στατιστική απογραφή στην εποχή του Γουλιέλμου του Κατακτητή, στο τέλος του 11ου αιώνα, αναφέρεται σε διάφορες μονάδες παραγωγής της Αγγλίας όπως μεταλλεία, ιχθυοτροφεία κ.ά. Από το 16ο έως το 19ο αιώνα, η ραγδαία ανάπτυξη του εμπορίου ώθησε τις πολιτειακές αρχές στη μελέτη οικονομικών δεδομένων, όπως είναι το εξαγωγικό εμπόριο, το πλήθος και η δυναμικότητα των βιομηχανιών κτλ. Ενώ παλαιότερα η Στατιστική ασχολείτο μόνο με την παράθεση τεράστιων πινάκων με δεδομένα και αναρίθμητων διαγραμμάτων, σήμερα μπορούμε να διακρίνουμε σε μια στατιστική έρευνα τρία στάδια: Τη συλλογή του στατιστικού υλικού, την επεξεργασία και παρουσίασή του και τέλος την ανάλυση αυτού του υλικού και την εξαγωγή χρήσιμων συμπερασμάτων. Τα τρία αυτά στάδια επιτυγχάνονται με την εφαρμογή καταλλήλων για κάθε περίπτωση στατιστικών μεθόδων, όπως και με τη βοήθεια των Υπολογιστών, οι οποίοι σημείωσαν τεράστια ανάπτυξη στις μέρες μας. Συμπερασματικά λοιπόν μπορούμε να δώσουμε ως ορισμό της "Στατιστικής" το συνηθέστερο και πλέον γνωστό ορισμό του R.A. Fisher (1890-1962), πατέρα της σύγχρονης Στατιστικής:

Στατιστική είναι ένα σύνολο αρχών και μεθοδολογιών για:

- **το σχεδιασμό της διαδικασίας συλλογής δεδομένων**
- **τη συνοπτική και αποτελεσματική παρουσίασή τους**
- **την ανάλυση και εξαγωγή αντίστοιχων συμπερασμάτων.**

Ο κλάδος της Στατιστικής που ασχολείται με τον πρώτο στόχο λέγεται σχεδιασμός πειραμάτων (experimental design) ενώ, με τον δεύτερο ασχολείται η περιγραφική στατιστική (descriptive statistics), που αποτελεί και το αντικείμενο μελέτης μας στη συνέχεια. Τέλος, η επαγωγική στατιστική ή στατιστική συμπερασματολογία (inferential statistics) περιλαμβάνει τις μεθόδους με τις οποίες γίνεται η προσέγγιση των χαρακτηριστικών ενός μεγάλου συνόλου δεδομένων, με τη μελέτη των χαρακτηριστικών ενός μικρού υποσυνόλου των δεδομένων. Έτσι αν, για παράδειγμα, ο Διευθυντής ενός σχολείου εξετάζοντας ένα δείγμα 100 απουσιών των μαθητών από το σύνολο των απουσιών ενός τριμήνου αναφέρει στο σύλλογο των καθηγητών ότι 20 από τις 100 απουσίες είναι αδικαιολόγητες, τότε απλώς περιγράφει αυτό που παρατήρησε. Αν όμως αναφέρει ότι το 20% των απουσιών είναι αδικαιολόγητες, τότε συμπεραίνει ότι το ποσοστό των απουσιών όλων των μαθητών του σχολείου

είναι (περίπου) το ίδιο με αυτό του δείγματος. Προβαίνει δηλαδή σε μια **επαγωγή** από το δείγμα στον πληθυσμό.
Η Στατιστική σήμερα χρησιμοποιείται ευρύτατα σε όλους σχεδόν τους τομείς της ανθρώπινης δραστηριότητας. Βασικές έννοιες της Στατιστικής έχουν εισχωρήσει και ενσωματωθεί σε όλες τις επιστήμες. Από τις Ανθρωπιστικές, Νομικές και Κοινωνικές Επιστήμες (Αρχαιολογία, Λαογραφία, Κοινωνιολογία, Δημογραφία, …), τις Φυσικές Επιστήμες (Φυσική, Χημεία, Αστρονομία, …), τις Επιστήμες Υγείας (Ιατρική, Φαρμακευτική, Βιολογία, …), τις Τεχνολογικές Επιστήμες (Μηχανολογία, Τοπογραφία, Ναυπηγική, …) μέχρι τις Επιστήμες Οικονομίας και Διοίκησης (Οικονομικά, Χρηματιστηριακά, Διαφήμιση, Marketing, …), βλέπουμε να υπεισέρχεται η Στατιστική είτε με την αρχική περιγραφική μορφή της είτε με τις προηγμένες αναλυτικές τεχνικές της.
Η ανάλυση στατιστικών ερευνών είναι το κυριότερο εργαλείο έρευνας σε ένα μεγάλο φάσμα εφαρμογών των παραπάνω επιστημών. Οι έρευνες των ανθρώπινων πληθυσμών (συχνά αναφερόμενες και ως **δημοσκοπήσεις**) αποτελούν σπουδαίες πηγές βασικής γνώσης των κοινωνικών επιστημών. Οικονομολόγοι, ψυχολόγοι, κοινωνιολόγοι και πολιτικοί επιστήμονες μελετούν ποικίλα θέματα όπως πρότυπα εσόδων-εξόδων των οικογενειών και των επιχειρήσεων, την επίδραση της

επαγγελματικής απασχόλησης των γυναικών στην οικογενειακή ζωή, τις συγκοινωνιακές και ταξιδιωτικές συνήθειες των κατοίκων μιας πόλης, τις προτιμήσεις των ψηφοφόρων για τους υποψηφίους και τις θέσεις τους. Πολλά προβλήματα που αντιμετωπίζουν σήμερα οι επιχειρήσεις αφορούν τη διατήρηση, αντικατάσταση ή το κρίσιμο σημείο αντοχής συσκευών ή προσωπικού. Ο διευθυντής μιας βιομηχανίας πρέπει να είναι σε θέση να κατανοεί στατιστικές έρευνες που αφορούν την ποιότητα του προϊόντος και την αποδοτικότητα της παραγωγικής διαδικασίας. Πρέπει επίσης να αντιλαμβάνεται την αποτελεσματικότητα της διαφήμισης και τις προτιμήσεις του καταναλωτή σε μια έρευνα αγοράς. Συμβουλευόμενος και τον στατιστικό μπορεί να πάρει σωστές αποφάσεις αναφορικά με την επέκταση ή μη της επιχείρησης. Σήμερα κάθε γιατρός πρέπει να έχει βασικές γνώσεις Στατιστικής που θα τον βοηθήσουν τόσο στην έρευνα όσο και στην καθημερινή άσκηση του κάθε μορφής και είδους ιατρικού ή βιοϊατρικού, γενικότερα, επαγγέλματος. Η Εθνική Στατιστική Υπηρεσία κάθε χώρας διενεργεί σε τακτά χρονικά διαστήματα δειγματοληπτικές έρευνες, για να πάρει πληροφορίες για τον πληθωρισμό, την απασχόληση και την ανεργία στη χώρα. Ανάλογα με τα αποτελέσματα διαμορφώνεται και η κυβερνητική πολιτική στα θέματα αυτά.

Πέρα από όλα αυτά, διαπιστώνουμε ολοένα και περισσότερο να γίνεται χρήση μεθόδων της Στατιστικής για την υποστήριξη διάφορων θέσεων. Ακόμα και σε τηλεοπτικές αντιπαραθέσεις (κυρίως σε προεκλογικές περιόδους) βλέπουμε τους συνομιλητές να κάνουν χρήση αριθμών, στατιστικών στοιχείων, γραφημάτων και διαγραμμάτων, για να δώσουν εγκυρότητα στις απόψεις τους και να πείσουν για τα λεγόμενά τους.

Παραπάνω έχουν αναφερθεί ελάχιστα από τα πεδία εφαρμογών της Στατιστικής. Προφανώς μια λεπτομερής περιγραφή όλων των εφαρμογών δεν είναι δυνατή. Η μελέτη όμως και η γνώση της Στατιστικής βοηθά όχι μόνο στη σωστή χρήση των γνωστών μεθόδων αλλά και στην ανάπτυξη νέων τεχνικών για την αποτελεσματικότερη εξαγωγή χρήσιμων συμπερασμάτων.

2.1 ΒΑΣΙΚΕΣ ΕΝΝΟΙΕΣ

Πληθυσμός - Μεταβλητές

Όπως αναφέρθηκε και προηγουμένως, αυτό που μας ενδιαφέρει είναι να εξετάσουμε τα στοιχεία ενός συνόλου ως προς ένα ή περισσότερα χαρακτηριστικά τους. Αυτό συμβαίνει, για παράδειγμα, όταν ενδιαφερόμαστε για:

α) τις προτιμήσεις των ψηφοφόρων εν όψει των προσεχών εκλογών
β) τον αριθμό των υπαλλήλων μιας επιχείρησης
γ) το ύψος, το βάρος, την ομάδα αίματος και το φύλο των μαθητών της Γ΄ τάξης Λυκείου
δ) τις συνέπειες του καπνίσματος στην υγεία των καπνιστών κτλ.

Σε καθένα από τα παραδείγματα αυτά έχουμε ένα σύνολο και θέλουμε να εξετάσουμε τα στοιχεία του ως προς ένα ή περισσότερα χαρακτηριστικά τους. Ένα τέτοιο σύνολο λέγεται πληθυσμός (population). Τα στοιχεία του πληθυσμού συχνά αναφέρονται και ως μονάδες ή άτομα του πληθυσμού. Στο πρώτο παράδειγμα έχουμε το σύνολο των ψηφοφόρων και μας ενδιαφέρει η προτίμησή

τους, ποιο "κόμμα" π.χ. υποστηρίζουν. Στο τρίτο παράδειγμα έχουμε το σύνολο των μαθητών της Γ΄ Λυκείου και μας ενδιαφέρουν τα τέσσερα χαρακτηριστικά τους: ύψος, βάρος, ομάδα αίματος και φύλο.

Τα χαρακτηριστικά ως προς τα οποία εξετάζουμε έναν πληθυσμό λέγονται μεταβλητές (variables) και τις συμβολίζουμε συνήθως με τα κεφαλαία γράμματα Χ, Υ, Ζ, Β,… Οι δυνατές τιμές που μπορεί να πάρει μια μεταβλητή λέγονται τιμές της μεταβλητής. Από τη διαδοχική εξέταση των ατόμων του πληθυσμού ως προς ένα χαρακτηριστικό τους προκύπτει μια σειρά από δεδομένα, που λέγονται στατιστικά δεδομένα ή παρατηρήσεις. Τα στατιστικά δεδομένα δεν είναι κατ' ανάγκη διαφορετικά. Για παράδειγμα, αν εξετάζουμε την ομάδα αίματος δέκα ατόμων, τα στατιστικά δεδομένα ή παρατηρήσεις που θα προκύψουν μπορεί να είναι: Α, Β, Α, ΑΒ, Ο, ΑΒ, ΑΒ, ΑΒ, Ο, Β. Οι δυνατές όμως τιμές που μπορεί να πάρει η μεταβλητή "ομάδα αίματος" είναι οι εξής τέσσερις: Α, Β, ΑΒ και Ο.

Τις μεταβλητές τις διακρίνουμε:

1. **Σε ποιοτικές ή κατηγορικές μεταβλητές, των οποίων οι τιμές τους δεν είναι αριθμοί. Τέτοιες είναι, για παράδειγμα, η ομάδα αίματος (με τιμές Α, Β, ΑΒ, Ο), το φύλο (με τιμές αγόρι, κορίτσι), οι συνέπειες του καπνίσματος (με τιμές καρδιακά νοσήματα, καρκίνος κτλ), όπως επίσης και η οικονομική κατάσταση και**

η υγεία των ανθρώπων (που μπορεί να χαρακτηριστεί ως κακή, μέτρια, καλή ή πολύ καλή), καθώς και το ενδιαφέρον των μαθητών για τη Στατιστική, που μπορεί να χαρακτηριστεί ως υψηλό, μέτριο, χαμηλό ή μηδαμινό.

2. Σε **ποσοτικές** μεταβλητές, των οποίων οι τιμές είναι αριθμοί και διακρίνονται:
   i) Σε **διακριτές** μεταβλητές, που παίρνουν μόνο "μεμονωμένες" τιμές. Τέτοιες μεταβλητές είναι, για παράδειγμα, ο αριθμός των υπαλλήλων μιας επιχείρησης (με τιμές 1, 2, …), το αποτέλεσμα της ρίψης ενός ζαριού (με τιμές 1, 2, …, 6) κτλ.
   ii) Σε **συνεχείς** μεταβλητές, που μπορούν να πάρουν οποιαδήποτε τιμή ενός διαστήματος πραγματικών αριθμών (α, β). Τέτοιες μεταβλητές είναι το ύψος και το βάρος των μαθητών της Γ΄ Λυκείου, ο χρόνος που χρειάζονται οι μαθητές να απαντήσουν στα θέματα μιας εξέτασης, η διάρκεια μιας τηλεφωνικής συνδιάλεξης κτλ.

Συλλογή Στατιστικών Δεδομένων

Ένας τρόπος για να πάρουμε τις απαραίτητες πληροφορίες που χρειαζόμαστε για κάποιο πληθυσμό είναι να

εξετάσουμε όλα τα άτομα (στοιχεία) του πληθυσμού ως προς το χαρακτηριστικό που μας ενδιαφέρει. Η μέθοδος αυτή συλλογής των δεδομένων καλείται απογραφή (census). Για παράδειγμα, η Στατιστική Υπηρεσία της χώρας μας (ΕΣΥΕ) κάνει κάθε 10 χρόνια απογραφή του πληθυσμού, η οποία αποτελεί κύρια πηγή δεδομένων δημογραφικού, οικονομικού, εμπορικού και βιομηχανικού χαρακτήρα. Η τελευταία απογραφή έγινε το 1991. Σε πολλές όμως περιπτώσεις η εξέταση όλων των μονάδων του πληθυσμού είναι δύσκολη ή ακόμα και αδύνατη. Ένας υποψήφιος βουλευτής, για παράδειγμα, πριν από τις εκλογές είναι δύσκολο να εξετάσει όλους τους ψηφοφόρους, για να προσδιορίσει τι αντίληψη έχουν για τις θέσεις του. Επίσης ο κόπος, ο χρόνος και τα έξοδα που χρειάζονται για τη διεξαγωγή μιας απογραφής είναι πολλές φορές αρκετά μεγάλα, ιδίως όταν ο πληθυσμός που εξετάζεται είναι αρκετά μεγάλος. Εξάλλου ένας κατασκευαστής εκρηκτικών μηχανισμών ή ηλεκτρικών λυχνιών είναι αδύνατο να δοκιμάζει όλους τους παραγόμενους μηχανισμούς, για να ελέγχει την αποτελεσματικότητά τους, ή όλες τις παραγόμενες λυχνίες για να ελέγχει το χρόνο ζωής τους. Ομοίως ο γιατρός για να υπολογίσει την αποτελεσματικότητα ενός νέου φαρμάκου στην καταπολέμηση μιας ασθένειας είναι αδύνατο να περιμένει να δοκιμαστεί το φάρμακο σε όλα τα άτομα που πάσχουν από τη συγκεκριμένη

ασθένεια. Όπου λοιπόν η απογραφή είναι δύσκολη, αδύνατη ή οικονομικά και χρονικά ασύμφορη, ο ερευνητής μαζεύει πληροφορίες από κάποια μικρή ομάδα ή υποσύνολο του πληθυσμού, το οποίο καλείται δείγμα. Κάνει τις παρατηρήσεις του στο δείγμα αυτό και μετά γενικεύει τα συμπεράσματά του για ολόκληρο τον πληθυσμό. Τα συμπεράσματα όμως που θα προκύψουν από τη μελέτη του δείγματος θα είναι αξιόπιστα, θα ισχύουν δηλαδή με ικανοποιητική ακρίβεια για ολόκληρο τον πληθυσμό, αν η επιλογή του δείγματος γίνει με σωστό τρόπο, ώστε το δείγμα να είναι, όπως λέμε, αντιπροσωπευτικό του πληθυσμού. Στην πράξη, ένα δείγμα θεωρείται αντιπροσωπευτικό ενός πληθυσμού, εάν έχει επιλεγεί κατά τέτοιο τρόπο, ώστε κάθε μονάδα του πληθυσμού να έχει την ίδια δυνατότητα να επιλεγεί. Η επιλογή του αντιπροσωπευτικού δείγματος είναι "εκ των ων ουκ άνευ". Αποτελεί πολύ σοβαρή και δύσκολη διαδικασία. Ο κακός σχεδιασμός και η εκτέλεση της στατιστικής έρευνας, η μη αντιπροσωπευτικότητα του δείγματος, ο μη σωστός καθορισμός του μεγέθους του δείγματος αποτελούν μερικά βασικά μειονεκτήματα στη διαδικασία επιλογής ενός δείγματος. Από την άλλη πλευρά, στις απογραφές απαιτείται συνήθως μεγάλος αριθμός απογραφέων. Παρουσιάζεται έτσι η ανάγκη πρόσληψης και εκπαίδευσης μεγάλου αριθμού υπαλλήλων. Λόγω του μεγάλου χρόνου και κυρίως των σημαντικών εξόδων που απαιτούνται, πολλές φορές χρησιμοποιούνται ανεπαρκώς εκπαιδευμένοι απογραφείς με

κίνδυνο να σημειώνονται λάθη οφειλόμενα σ' αυτούς.

Αξίζει να σημειωθεί ότι μία "προσεκτική" επιλογή μικρότερου δείγματος είναι δυνατόν να δώσει καλύτερα αποτελέσματα από ένα μεγαλύτερο δείγμα που δεν έχει εκλεγεί κατάλληλα. Ενδεικτικό είναι το παράδειγμα των προεδρικών εκλογών των ΗΠΑ το 1936. Το περιοδικό Literary Digest χρησιμοποιώντας δείγμα 2.400.000 ατόμων πρόβλεψε νίκη του Landon με ποσοστό 57%. Αντίθετα, το δημοσκοπικό γραφείο του G. Gallup χρησιμοποιώντας δείγμα 50.000 ατόμων πρόβλεψε το σωστό αποτέλεσμα που ήταν νίκη του Roosvelt με ποσοστό 62%! Η παταγώδης αποτυχία της δημοσκόπησης του περιοδικού οφειλόταν στο γεγονός ότι το δείγμα που επελέγη δεν ήταν αντιπροσωπευτικό του πληθυσμού. Οι αρχές και οι μέθοδοι για τη συλλογή και ανάλυση δεδομένων από πεπερασμένους πληθυσμούς είναι το αντικείμενο της Δειγματοληψίας (Sampling), που αποτελεί τη βάση της Στατιστικής. Γενικά, μπορούμε να πούμε ότι η οργάνωση της συλλογής και επεξεργασίας των σχετικών δεδομένων και πληροφοριών γίνεται κατά τρόπο που για δεδομένη ακρίβεια να επιτυγχάνεται το χαμηλότερο δυνατό κόστος ή, αντιστρόφως, να εξασφαλίζεται η μέγιστη δυνατή ακρίβεια την οποίαν επιτρέπουν τα μέσα που διαθέτουμε.

ΑΣΚΗΣΕΙΣ

1. **Ποιες από τις παρακάτω μεταβλητές είναι ποιοτικές και ποιες ποσοτικές;**
Από τις ποσοτικές ποιες είναι διακριτές και ποιες συνεχείς;

α) Βάρος
β) Αριθμός τροχαίων ατυχημάτων
γ) Φύλο
δ) Οικογενειακή κατάσταση
ε) Στάθμη της λίμνης του Μαραθώνα
στ) Τόπος καταγωγής
ζ) Επάγγελμα
η) Αριθμός παιδιών στην οικογένεια
θ) Βαθμολογία στο σκάκι
ι) Νούμερο γυναικείων παπουτσιών.

2. **Στις παρακάτω περιπτώσεις ποιες μπορεί να είναι οι μεταβλητές που μας ενδιαφέρουν; Να γίνει η διάκρισή τους σε ποιοτικές ή ποσοτικές και να αναφερθούν μερικές δυνατές τιμές τους:**
α) Εξετάζουμε ένα δείγμα υπαλλήλων μιας εταιρείας.

β) Εξετάζουμε ένα δείγμα προϊόντων από μια παραγωγή.

γ) Εξετάζουμε ένα δείγμα τηλεθεατών.

δ) Εξετάζουμε τους καλαθοσφαιριστές μιας ομάδας σε έναν αγώνα.

3. Για να βρούμε ποιες εκπομπές στην τηλεόραση έχουν τη μεγαλύτερη ακροαματικότητα αποφασίσαμε να πάρουμε δείγμα 500 τηλεθεατών. Ποιος είναι, κατά τη γνώμη σας, ο καλύτερος από τους παρακάτω τρόπους, για να πάρουμε το δείγμα; Είναι καλύτερο να πάρουμε:

α) μόνο άνδρες, β) μόνο γυναίκες, γ) άτομα από τις μεγάλες πόλεις, δ) άτομα μόνο από την επαρχία, ε) άτομα από διάφορες περιοχές.

4. Τι έχετε να παρατηρήσετε για τα παρακάτω επιλεγόμενα δείγματα;

α) Για να βρούμε τα ποσοστά των ανδρών και των γυναικών στην Ελλάδα, πηγαίνουμε σε μια μεγάλη στρατιωτική μονάδα και ρωτάμε όλους τους στρατιώτες, πόσοι άνδρες και

πόσες γυναίκες υπάρχουν στην οικογένειά τους.

β) Κάποιος θέλει να σχηματίσει μια ιδέα για το αποτέλεσμα των επερχόμενων βουλευτικών εκλογών. Τηλεφωνεί λοιπόν σε συγγενείς και φίλους του και τους ρωτάει σχετικά.

γ) Για να εκτιμήσουμε το κατά κεφαλή εισόδημα των Ελλήνων παίρνουμε ένα δείγμα από το Κολωνάκι των Αθηνών.

δ) Για να δούμε πώς διασκεδάζουν οι νέοι της χώρας μας επιλέγουμε κάποιους μαθητές από διάφορα Λύκεια της Αττικής.

ε) Ο διευθυντής ενός Λυκείου αποφάσισε να καταγράψει τους λόγους της απουσίας των μαθητών από το Λύκειο κατά τη διάρκεια της ακαδημαϊκής χρονιάς. Γι' αυτό τον λόγο πήρε ως δείγμα όσους απουσίασαν το Νοέμβριο.

2.2 ΠΑΡΟΥΣΙΑΣΗ ΣΤΑΤΙΣΤΙΚΩΝ ΔΕΔΟΜΕΝΩΝ

Στατιστικοί Πίνακες

Μετά τη συλλογή των στατιστικών δεδομένων είναι αναγκαία η κατασκευή συνοπτικών πινάκων ή γραφικών παραστάσεων, ώστε να είναι εύκολη η κατανόησή τους και η εξαγωγή σωστών συμπερασμάτων. Η παρουσίαση των στατιστικών δεδομένων σε πίνακες γίνεται με την κατάλληλη τοποθέτηση των πληροφοριών σε γραμμές και στήλες, με τρόπο που να διευκολύνεται η σύγκριση των στοιχείων και η καλύτερη ενημέρωση του αναγνώστη σχετικά με τη δομή του πληθυσμού που ερευνάμε.

Οι πίνακες διακρίνονται στους:
α) γενικούς πίνακες, οι οποίοι περιέχουν όλες τις πληροφορίες που προκύπτουν από μία στατιστική έρευνα (συνήθως με αρκετά λεπτομερειακά στοιχεία) και αποτελούν πηγές στατιστικών πληροφοριών στη διάθεση των επιστημόνων-ερευνητών για παραπέρα ανάλυση και εξαγωγή συμπερασμάτων,
β) ειδικούς πίνακες, οι οποίοι είναι συνοπτικοί και σαφείς. Τα στοιχεία τους συνήθως έχουν ληφθεί από τους γενικούς πίνακες.

Κάθε πίνακας που έχει κατασκευαστεί σωστά πρέπει να περιέχει:

α) τον τίτλο, που γράφεται στο επάνω μέρος του πίνακα και δηλώνει με σαφήνεια και συνοπτικά το περιεχόμενο του πίνακα,

β) τις επικεφαλίδες των γραμμών και στηλών, που δείχνουν συνοπτικά τη φύση και τις μονάδες μέτρησης των δεδομένων,

γ) το κύριο σώμα (κορμό), που περιέχει διαχωρισμένα μέσα στις γραμμές και στις στήλες τα στατιστικά δεδομένα,

δ) την πηγή, που γράφεται στο κάτω μέρος του πίνακα και δείχνει την προέλευση των στατιστικών στοιχείων, έτσι ώστε ο αναγνώστης να ανατρέχει σ' αυτήν, όταν επιθυμεί, για επαλήθευση στοιχείων ή για λήψη περισσότερων πληροφοριών.

Παρακάτω δίνονται μερικοί στατιστικοί πίνακες, που διευκρινίζουν την εφαρμογή των προηγούμενων εννοιών.

Πίνακας 1

Πληθυσμός της Ελλάδος (σε εκατομμύρια) κατά μεγάλες ομάδες ηλικιών

Ηλικία (σε έτη)	Απο-γραφή 1971	Απο-γραφή 1981	Απο-γραφή 1991	Εκτί-μηση 1993	Εκτί-μηση 1994
0 - 14	2,22	2,31	1,97	1,85	1,81
15 - 64	5,58	6,19	6,88	6,99	7,04
≥ 65	0,96	1,24	1,40	1,54	1,58

Πηγή: ΕΣΥΕ, 1996

Πίνακας 2

Επιφάνεια και πληθυσμός των κατοικημένων νησιών της Ελλάδας με πληθυσμό, κατά την απογραφή του 1991, άνω των 10.000 κατοίκων.

Κατοικημένες νήσοι	Επιφάνεια σε τ.χμ.	Πληθυσμός κατά τις απογραφές		
		1971	1981	1991
Κρήτη	8.261,183	456.471	502.082	539.938
Εύβοια	3.661,637	162.986	185.626	205.502
Λέσβος	1.635,998	97.008	88.601	87.151
Ρόδος	1.401,459	66.606	87.831	98.181
Χίος	842,796	52.487	48.700	51.060
Κεφαλληνία	734,014	31.787	27.649	29.392
Κέρκυρα	585,312	89.578	96.533	104.781
Σάμος	477,942	32.664	31.629	33.032
Λήμνος	476,288	17.367	15.721	17.645
Ζάκυνθος	406,612	30.180	30.011	32.556
Νάξος	389,434	14.201	14.037	14.838
Θάσος	383,672	13.316	13.111	13.527
Λευκάδα	301,106	22.917	19.947	19.350
Κως	287,611	16.650	20.350	26.379
Κάλυμνος	110,581	13.097	14.295	15.706
Σαλαμίνα	91,503	23.065	28.574	34.272
Σύρος	84,069	18.642	19.668	19.870
Αίγινα	77,014	9.553	11.127	11.639

Πηγή: ΕΣΥΕ, Απογραφή 1991

Πίνακας 3

Εργατικά ατυχήματα κατά ομάδες ηλικιών

Έτη 1990-94

Ηλικία	1990	1991	1992	1993	1994
Κάτω των 15	18	9	10	16	5
15 - 19	731	564	437	735	442
20 - 24	3323	2785	2755	2981	2696
25 - 29	4277	3921	4246	3881	3717
30 - 34	3952	3700	3388	3348	3282
35 - 39	3589	3146	3233	3230	3000
40 - 44	3237	2803	2911	2880	2903
45 - 49	2839	2593	2784	2608	2403
50 - 54	2727	2564	2286	2095	1877
55 - 59	2304	2230	2185	1699	1664
60 - 64	728	720	688	420	523
65 - 69	121	150	140	66	96
Σύνολο	27846	25185	25063	23959	22608

Πηγή: ΙΚΑ, Ελληνικό Ινστιτούτο Υγιεινής και Ασφάλειας της Εργασίας

Πίνακας 4

Χαρακτηριστικά 40 μαθητών Γ΄ τάξης ενός Λυκείου.

α.α	Φύλο	Ασχολία *	Αριθμός αδελφών	Βαθμός μαθημα-τικών Β' λυκείου	Ύψος (cm)	Βάρος (kg)	Ύψος πατέρα (cm)	Ύψος μητέρας (cm)
1	Κ	4	1	15	170	60	172	168
2	Α	1	0	17	180	68	185	165
3	Κ	4	2	12	178	62	181	160
4	Κ	5	1	18	165	47	180	162
5	Κ	5	0	15	170	54	180	168
6	Κ	4	3	16	168	56	185	168
7	Κ	4	2	15	175	58	193	162
8	Α	4	1	15	175	72	174	174
9	Α	2	3	13	173	67	182	160
10	Κ	3	1	15	162	50	176	170
11	Κ	4	1	16	160	51	176	164
12	Α	2	1	11	170	58	182	165
13	Κ	7	3	20	167	50	174	170

14	A	1	1	18	177	81	177	169
15	A	1	0	17	180	70	170	165
16	K	2	2	19	170	63	165	174
17	A	2	0	14	182	71	176	173
18	A	7	2	17	178	73	182	170
19	K	4	1	14	165	58	180	161
20	A	5	1	16	178	74	173	168
21	K	5	1	12	156	44	170	158
22	K	5	1	13	175	53	170	165
23	A	5	2	18	172	60	178	165
24	K	6	1	16	173	64	182	162
25	K	6	2	14	167	57	172	157
26	A	5	0	14	187	85	185	170
27	K	6	1	17	170	62	180	165
28	A	3	1	12	180	80	180	167
29	A	3	0	15	178	73	173	170
30	A	2	1	10	191	86	180	170
31	A	2	0	16	176	65	180	172
32	K	4	1	12	169	57	170	167

α.α	Φύλο	Ασχολία *	Αριθμός αδελφών	Βαθμός μαθημα-τικών Β' λυκείου	Ύψος (cm)	Βάρος (kg)	Ύψος πατέρα (cm)	Ύψος μητέρας (cm)
33	Κ	4	2	14	167	61	179	158
34	Κ	4	1	19	166	62	178	165
35	Α	3	1	19	179	76	178	160
36	Α	3	1	16	178	68	180	160
37	Α	5	1	19	180	85	170	163
38	Κ	5	1	19	164	64	184	170
39	Κ	3	0	15	170	63	165	167
40	Κ	4	1	15	173	63	186	162

*1: Υπολογιστές, 2:Αθλητισμός, 3:Διασκέδαση - Ντίσκο, 4:Μουσική, 5:Τηλεόραση -Κινηματογράφος, 6:Διάβασμα εξωσχολικών βιβλίων, 7:Άλλο

Πηγή: Δειγματοληπτική έρευνα μεταξύ μαθητών 1ου Λυκείου Αμαρουσίου (Σεπτ. ’98).

Πίνακες Κατανομής Συχνοτήτων

Ας υποθέσουμε ότι $x_1, x_2, ..., x_κ$ είναι οι τιμές μιας μεταβλητής Χ, που αφορά τα άτομα ενός δείγματος μεγέθους ν, $κ \leq ν$. Στην τιμή x_i αντιστοιχίζεται η (απόλυτη) **συχνότητα** (frequency) $ν_i$, δηλαδή ο φυσικός αριθμός που δείχνει πόσες φορές εμφανίζεται η τιμή x_i της εξεταζόμενης μεταβλητής Χ στο σύνολο των παρατηρήσεων. Είναι φανερό ότι το άθροισμα όλων των συχνοτήτων είναι ίσο με το μέγεθος ν του δείγματος, δηλαδή:

$$ν_1 + ν_2 + ... + ν_κ = ν \qquad (1)$$

Για παράδειγμα, για τη μεταβλητή Χ: "αριθμός αδελφών" του πίνακα 4 οι συχνότητες για τις τιμές $x_1 = 0$, $x_2 = 0$, $x_3 = 2$, $x_4 = 3$ είναι, αντίστοιχα, $ν_1 = 8$, $ν_2 = 22$, $ν_3 = 7$, $ν_4 = 3$ με $ν_1 + ν_2 + ν_3 + ν_4 = 40$. Ο υπολογισμός των συχνοτήτων γίνεται με τη **διαλογή** των παρατηρήσεων, όπως φαίνεται στον παρακάτω πίνακα 5. Διατρέχοντας με τη σειρά τη λίστα των δεδομένων καταγράφουμε κάθε παρατήρηση με συμβολικό τρόπο σαν μια γραμμή " | " στην αντίστοιχη τιμή της μεταβλητής.

Πίνακας 5

Κατανομή συχνοτήτων της μεταβλητής X: “αριθμός αδελφών” των μαθητών του πίνακα 4.

Αριθμός αδελφών x_i	Διαλογή	Συχνότητα v_i	Σχετική Συχνότητα f_i	Σχετική Συχνότητα f_i %
0	𝍸 |||	8	0,200	20,0
1	𝍸 𝍸 𝍸 𝍸 ||	22	0,550	55,0
2	𝍸 ||	7	0,175	17,5
3	|||	3	0,075	7,5
Σύνολο:		40	1,000	100,0

Αν διαιρέσουμε τη συχνότητα v_i με το μέγεθος ν του δείγματος, προκύπτει η **σχετική συχνότητα** (relative frequency) f_i της τιμής x_i, δηλαδή

$$f_i = \frac{v_i}{v},\ i = 1,2,...,\kappa. \qquad (2)$$

Για τη σχετική συχνότητα ισχύουν οι ιδιότητες:

i) $0 \le f_i \le 1$ για $i = 1, 2, ..., \kappa$ αφού $0 \le v_i \le v$.

ii) $f_1 + f_2 + ... + f_\kappa = 1$, αφού

$$f_1 + f_2 + ... + f_\kappa = \frac{v_1}{v} + \frac{v_2}{v} + ... + \frac{v_\kappa}{v} = \frac{v_1 + v_2 + ... + v_\kappa}{v} = \frac{v}{v} = 1.$$

Συνήθως, τις σχετικές συχνότητες f_i τις εκφράζουμε επί τοις εκατό, οπότε συμβολίζονται με $f_i\%$, δηλαδή $f_i\% = 100f_i$. Για παράδειγμα, οι σχετικές συχνότητες για τις τιμές $x_1 = 0$, $x_2 = 1$, $x_3 = 2$, $x_4 = 3$ της μεταβλητής Χ: "αριθμός αδελφών" είναι αντιστοίχως:

$$f_1 = \frac{8}{40} = 0,20,\quad f_2 = \frac{22}{40} = 0,55,\quad f_3 = \frac{7}{40} = 0,175 \text{ και}$$

$$f_4 = \frac{3}{40} = 0,075 \text{ με}$$

$$f_1 + f_2 + f_3 + f_4 = 0,20 + 0,55 + 0,175 + 0,075 = 1.$$

Συνεπώς $f_1\% = 20$, $f_2\% = 55$, $f_3\% = 17,5$ και $f_4\% = 7,5$

με $f_1\% + f_2\% + f_3\% + f_4\% = 100$.
Οι ποσότητες x_i, v_i, f_i για ένα δείγμα συγκεντρώνονται σε ένα συνοπτικό πίνακα, που ονομάζεται **πίνακας κατανομής συχνοτήτων** ή απλά **πίνακας συχνοτήτων**. Για μια μεταβλητή, το σύνολο των ζευγών (x_i, v_i) λέμε ότι αποτελεί την **κατανομή συχνοτήτων** και το σύνολο των ζευγών (x_i, f_i), ή των ζευγών (x_i, $f_i\%$), την **κατανομή των σχετικών συχνοτήτων**. Στον πίνακα 5 παρουσιάζονται οι κατανομές συχνοτήτων και σχετικών συχνοτήτων της μεταβλητής Χ: "αριθμός αδελφών" των μαθητών του πίνακα 4.

Αθροιστικές Συχνότητες

Στην περίπτωση των **ποσοτικών μεταβλητών** εκτός από τις συχνότητες v_i και f_i χρησιμοποιούνται συνήθως και οι λεγόμενες **αθροιστικές συχνότητες** (cumulative frequencies) N_i και οι **αθροιστικές σχετικές συχνότητες** (cumulative relative frequencies) F_i, οι οποίες εκφράζουν το πλήθος και το ποσοστό αντίστοιχα των παρατηρήσεων που είναι μικρότερες ή ίσες της τιμής x_i. Συχνά οι F_i πολλαπλασιάζονται επί 100 εκφραζόμενες έτσι επί τοις εκατό, δηλαδή $F_i\% = 100F_i$, βλέπε πίνακα 6. Αν οι τιμές $x_1, x_2, ..., x_\kappa$ μιας ποσοτικής μεταβλητής Χ είναι σε αύξουσα διάταξη, τότε η αθροιστική συχνότητα

της τιμής x_i είναι $N_i = v_1 + v_2 + ... + v_i$. Όμοια, η αθροιστική σχετική συχνότητα είναι $F_i = f_1 + f_2 + ... + f_i$, για $i = 1, 2, ..., κ$. Για παράδειγμα, για τη μεταβλητή Χ: "αριθμός αδελφών" του πίνακα 4 είναι $N_1 = v_1 = 8$, $N_2 = v_1 + v_2 = 30$, $N_3 = v_1 + v_2 + v_3 = 37$ και $N_4 = v_1 + v_2 + v_3 + v_4 = v = 40$, οπότε

$F_1 = f_1 = 0,20$, $F_2 = f_1 + f_2 = 0,75$, $F_3 = f_1 + f_2 + f_3 = 0,925$ και $F_4 = f_1 + f_2 + f_3 + f_4 = 1$, οπότε $F_1\% = 20$, $F_2\% = 75$, $F_3\% = 92,5$ και $F_4\% = 100$. Είναι φανερό ότι ισχύουν οι σχέσεις:

$v_1 = N_1$, $v_2 = N_2 - N_1$, ... , $v_κ = N_κ - N_{κ-1}$

και

$f_1 = F_1$, $f_2 = F_2 - F_1$, ... , $f_κ = F_κ - F_{κ-1}$.

Πίνακας 6

Κατανομή συχνοτήτων και αθροιστικών συχνοτήτων της μεταβλητής “αριθμός αδελφών” των μαθητών του πίνακα 4.

Αριθμός αδελφών x_i	Συχνότητα v_i	Σχετ. Συχν. f_i	Σχετ. Συχν. f_i%	Αθροισ. Συχν. N_i	Αθροισ. Σχετ. Συχν. F_i	Αθροισ. Σχετ. Συχν. F_i %
0	8	0,200	20,0	8	0,200	20,0
1	22	0,550	55,0	30	0,750	75,0
2	7	0,175	17,5	37	0,925	92,5
3	3	0,075	7,5	40	1,000	100,0
Σύνολο:	40	1,000	100,0	–	–	–

Γραφική Παράσταση Κατανομής Συχνοτήτων

Τα στατιστικά δεδομένα παρουσιάζονται πολλές φορές και υπό μορφή γραφικών παραστάσεων ή διαγραμμάτων. Οι γραφικές παραστάσεις παρέχουν πιο σαφή εικόνα του χαρακτηριστικού σε σχέση με τους πίνακες, είναι πολύ πιο ενδιαφέρουσες και ελκυστικές, χωρίς βέβαια να προσφέρουν περισσότερη πληροφορία από εκείνη που περιέχεται στους αντίστοιχους πίνακες συχνοτήτων. Επί πλέον με τα διαγράμματα διευκολύνεται η σύγκριση μεταξύ ομοειδών στοιχείων για το ίδιο ή για διαφορετικά χαρακτηριστικά.
Υπάρχουν διάφοροι τρόποι γραφικής παρουσίασης, ανάλογα με το είδος των δεδομένων που έχουμε. Όπως όμως οι στατιστικοί πίνακες έτσι και τα στατιστικά διαγράμματα πρέπει να συνοδεύονται από α) τον τίτλο, β) την κλίμακα με τις τιμές των μεγεθών που απεικονίζονται, γ) το υπόμνημα που επεξηγεί συνήθως τις τιμές της μεταβλητής και δ) την πηγή των δεδομένων.

α) Ραβδόγραμμα

Το ραβδόγραμμα (barchart) χρησιμοποιείται για τη γραφική παράσταση των τιμών μιας ποιοτικής μεταβλητής. Το ραβδόγραμμα αποτελείται από ορθογώνιες

στήλες που οι βάσεις τους βρίσκονται πάνω στον οριζόντιο ή τον κατακόρυφο άξονα. Σε κάθε τιμή της μεταβλητής Χ αντιστοιχεί μια ορθογώνια στήλη της οποίας το ύψος είναι ίσο με την αντίστοιχη συχνότητα ή σχετική συχνότητα. Έτσι έχουμε αντίστοιχα το **ραβδόγραμμα συχνοτήτων** και το **ραβδόγραμμα σχετικών συχνοτήτων**. Τόσο η απόσταση μεταξύ των στηλών όσο και το μήκος των βάσεών τους καθορίζονται αυθαίρετα. Στον πίνακα 7 έχουμε την κατανομή συχνοτήτων της μεταβλητής Χ: "απασχόληση στον ελεύθερο χρόνο" και στα σχήματα 1(α), (β) τα αντίστοιχα ραβδογράμματα συχνοτήτων και σχετικών συχνοτήτων.

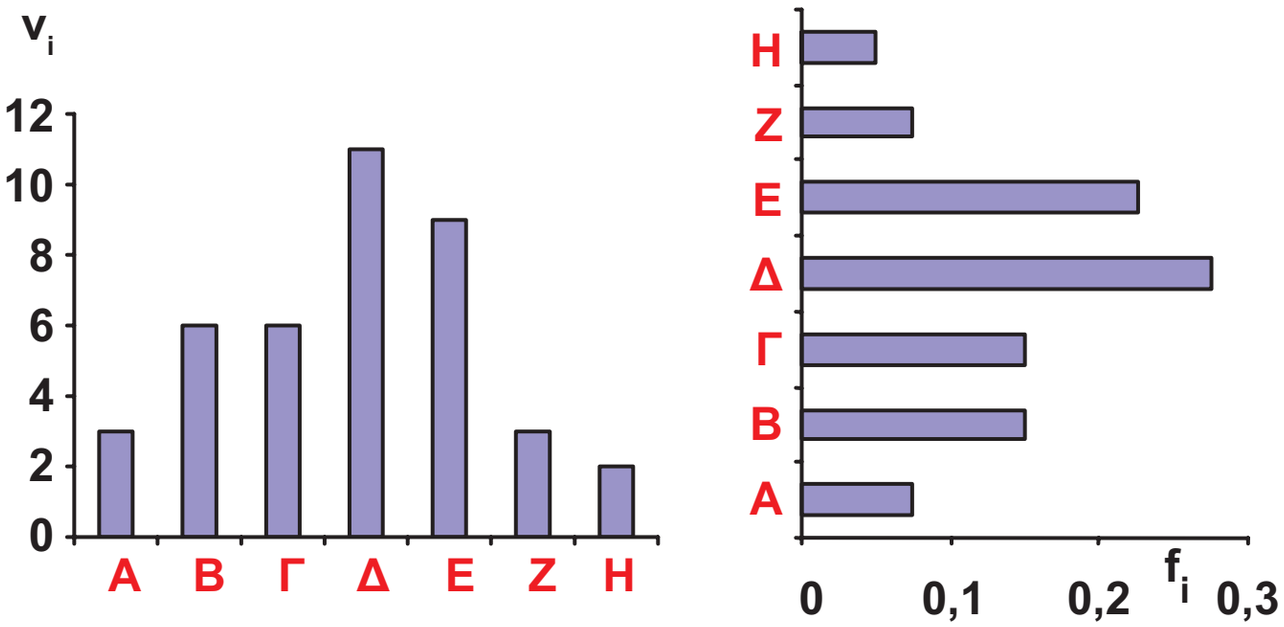
Πίνακας 7

Κατανομή συχνοτήτων για την απασχόληση στον ελεύθερο χρόνο τους των μαθητών του πίνακα 4.

i	Απασχόληση x_i	Συχνότητα v_i	Σχετική συχνότητα f_i	Σχετική συχνότητα f_i%
1	Υπολογιστές	3	0,075	7,5
2	Αθλητισμός	6	0,150	15,0
3	Διασκέδαση-ντίσκο	6	0,150	15,0
4	Μουσική	11	0,275	27,5
5	Τηλεόραση-Κινηματογράφος.	9	0,225	22,5
6	Διάβασμα εξωσχ. Βιβλίων	3	0,075	7,5
7	Άλλο	2	0,050	5,0
Σύνολο:		40	1,000	100,0

Μερικές φορές σε ένα ραβδόγραμμα συχνοτήτων ο ρόλος των δύο αξόνων είναι δυνατόν να αντιστραφεί, όπως φαίνεται στο σχήμα 1(β), που παριστάνεται το ραβδόγραμμα σχετικών συχνοτήτων της ίδιας μεταβλητής.

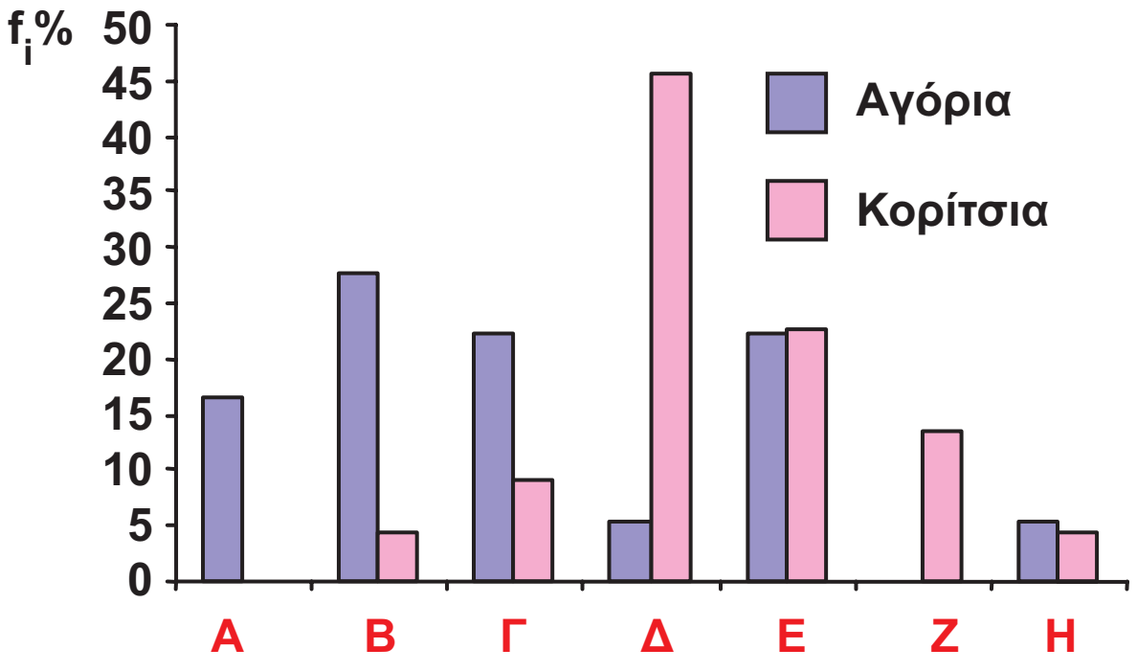
Αν θέλουμε να συγκρίνουμε τον τρόπο που περνούν τον ελεύθερο χρόνο τους τα αγόρια και τα κορίτσια, τότε κατασκευάζουμε το ραβδόγραμμα σχετικών συχνοτήτων του σχήματος 1(γ), όπως προκύπτει από τον πίνακα 4.



(α) (β)

Ραβδόγραμμα συχνοτήτων (α) και σχετικών συχνοτήτων (β) για την απασχόληση των μαθητών του πίνακα 7.

1



(γ)

Ραβδόγραμμα σχετικών συχνοτήτων για την απασχόληση των μαθητών του πίνακα 4 ανάλογα με το φύλο.

Α. Η/Υ **Β.** Αθλητισμός **Γ.** Διασκέδαση-Ντίσκο

Δ. Μουσική **Ε.** Τηλεόραση-Κινηματογράφος.

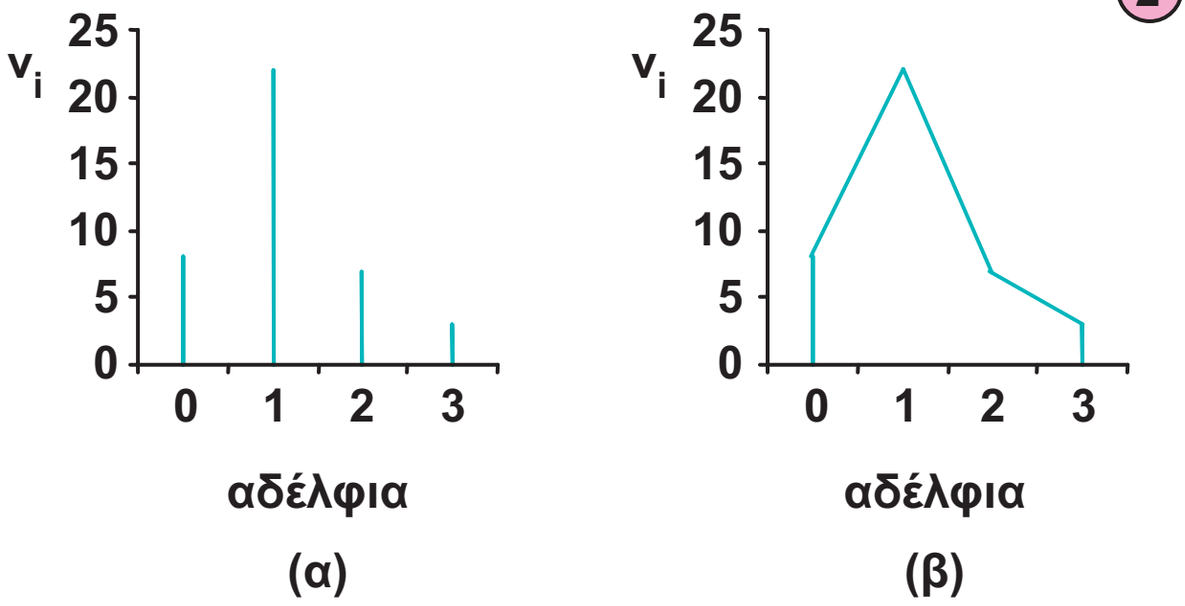
Ζ. Διάβασμα εξωσχ. Βιβλίων **Η.** Άλλο

β) Διάγραμμα Συχνοτήτων

Στην περίπτωση που έχουμε μια ποσοτική μεταβλητή αντί του ραβδογράμματος χρησιμοποιείται το διάγραμμα συχνοτήτων (line diagram). Αυτό μοιάζει με το ραβδόγραμμα με μόνη διαφορά ότι αντί να χρησιμοποιούμε συμπαγή ορθογώνια υψώνουμε σε κάθε x_i (υποθέτοντας ότι $x_1 < x_2 < ... < x_κ$) μία κάθετη γραμμή με μήκος ίσο προς την αντίστοιχη συχνότητα, όπως φαίνεται στο σχήμα 2(α). Μπορούμε επίσης αντί των συχνοτήτων v_i στον κάθετο άξονα να βάλουμε τις σχετικές συχνότητες f_i, οπότε έχουμε το διάγραμμα σχετικών συχνοτήτων.

Ενώνοντας τα σημεία (x_i, v_i) ή (x_i, f_i) έχουμε το λεγόμενο πολύγωνο συχνοτήτων ή πολύγωνο σχετικών συχνοτήτων, αντίστοιχα, που μας δίνουν μια γενική ιδέα για τη μεταβολή της συχνότητας ή της σχετικής συχνότητας όσο μεγαλώνει η τιμή της μεταβλητής που εξετάζουμε, βλέπε σχήμα 2(β).

2



Διάγραμμα συχνοτήτων (α) και πολύγωνο συχνοτήτων (β) για τη μεταβλητή"αριθμός αδελφών" του πίνακα 4.

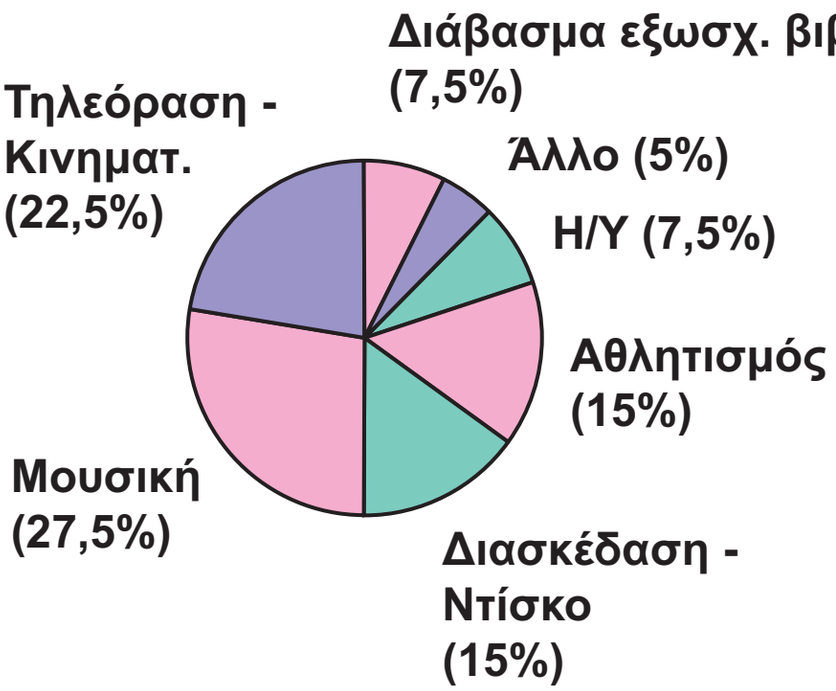
γ) Κυκλικό Διάγραμμα

Το κυκλικό διάγραμμα (piechart) **χρησιμοποιείται για τη γραφική παράσταση τόσο των ποιοτικών όσο και των ποσοτικών δεδομένων, όταν οι διαφορετικές τιμές της μεταβλητής είναι σχετικά λίγες. Το κυκλικό διάγραμμα είναι ένας κυκλικός δίσκος χωρισμένος σε κυκλικούς τομείς, τα εμβαδά ή, ισοδύναμα, τα τόξα των οποίων είναι ανάλογα προς τις αντίστοιχες συχνότητες v_i ή τις σχετικές συχνότητες f_i των τιμών x_i της μεταβλητής. Αν συμβολίσουμε με α_i το αντίστοιχο τόξο ενός κυκλικού τμήματος στο κυκλικό διάγραμμα συχνοτήτων, τότε**

$$\alpha_i = v_i \frac{360^o}{v} = 360^o f_i \quad \text{για } i = 1, 2, \ldots, \kappa.$$

Στο σχήμα 3 παριστάνεται το αντίστοιχο κυκλικό διάγραμμα σχετικών συχνοτήτων της "απασχόλησης των μαθητών" για τα δεδομένα του πίνακα 4.

3

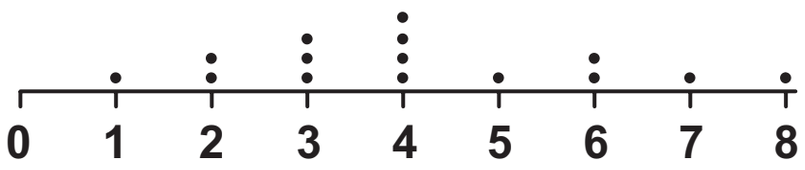


Κυκλικό διάγραμμα σχετικών συχνοτήτων της απασχόλησης των μαθητών για τα δεδομένα του πίνακα 4.

δ) Σημειόγραμμα

Όταν έχουμε λίγες παρατηρήσεις, η κατανομή τους μπορεί να περιγραφεί με το σημειόγραμμα (dot diagram), στο οποίο οι τιμές παριστάνονται γραφικά σαν σημεία υπεράνω ενός οριζόντιου άξονα. Στο σχήμα 4 έχουμε το σημειόγραμμα των χρόνων (σε λεπτά) 4, 2, 3, 1, 5, 6, 4, 2, 3, 4, 7, 4, 8, 6, 3 που χρειάστηκαν δεκαπέντε μαθητές, για να λύσουν ένα πρόβλημα.

4



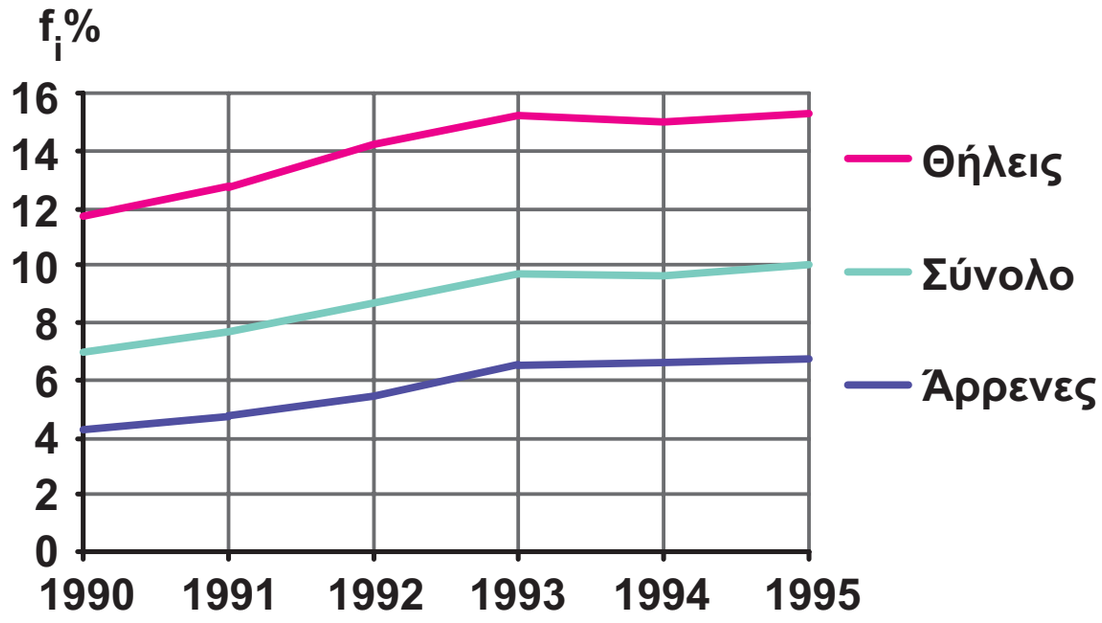
ε) Χρονόγραμμα

Το χρονόγραμμα ή χρονολογικό διάγραμμα χρησιμοποιείται για τη γραφική απεικόνιση της διαχρονικής εξέλιξης ενός οικονομικού, δημογραφικού ή άλλου μεγέθους. Ο οριζόντιος άξονας χρησιμοποιείται συνήθως ως άξονας μέτρησης του χρόνου και ο κάθετος ως άξονας μέτρησης της εξεταζόμενης μεταβλητής.

Στο σχήμα 5 έχουμε το χρονόγραμμα του ποσοστού ανεργίας στη χώρα μας από το 1990 έως το 1995. (Πηγή ΕΣΥΕ).

Παρατηρούμε ότι στο γυναικείο πληθυσμό υπάρχει συστηματικά μεγαλύτερο ποσοστό ανεργίας, γύρω στις 8 εκατοστιαίες μονάδες. Στο διάστημα 1993-95 το ποσοστό ανεργίας έχει σταθεροποιηθεί γύρω στο 6,5% για τους άνδρες και γύρω στο 15% για τις γυναίκες.

5



Ποσοστά ανεργίας στην Ελλάδα

Ομαδοποίηση των Παρατηρήσεων

Οι πίνακες συχνοτήτων και κατ' αναλογίαν τα αντίστοιχα διαγράμματα είναι δύσκολο να κατασκευαστούν, όταν το πλήθος των τιμών μιας μεταβλητής είναι αρκετά μεγάλο. Αυτό μπορεί να συμβεί είτε στην περίπτωση μιας διακριτής μεταβλητής είτε, πολύ περισσότερο, στην περίπτωση μιας συνεχούς μεταβλητής, όπου αυτή μπορεί να πάρει οποιαδήποτε τιμή στο διάστημα ορισμού της. Σ' αυτές τις περιπτώσεις είναι απαραίτητο να ταξινομηθούν (ομαδοποιηθούν) τα δεδομένα σε μικρό πλήθος ομάδων, που ονομάζονται και κλάσεις (class intervals), έτσι ώστε κάθε τιμή να ανήκει μόνο σε μία κλάση. Τα άκρα των κλάσεων καλούνται όρια των κλάσεων (class boundaries). Συνήθως υιοθετούμε την περίπτωση που μια κλάση περιέχει το κάτω άκρο της (κλειστή αριστερά) αλλά όχι το άνω άκρο της (ανοικτή δεξιά), δηλαδή που οι κλάσεις είναι της μορφής [,). Οι παρατηρήσεις κάθε κλάσης θεωρούνται όμοιες, οπότε μπορούν να "αντιπροσωπευθούν" από τις κεντρικές τιμές, τα κέντρα δηλαδή κάθε κλάσης.

- **Το πρώτο βήμα στην ομαδοποίηση των δεδομένων είναι η εκλογή του αριθμού κ των ομάδων ή κλάσεων. Ο αριθμός αυτός συνήθως ορίζεται αυθαίρετα από τον ερευνητή σύμφωνα με την πείρα του. Γενικά όμως μπορεί να χρησιμοποιηθεί ως οδηγός ο παρακάτω πίνακας:**

Μέγεθος δείγματος ν	Αριθμός κλάσεων κ	Μέγεθος δείγματος ν	Αριθμός κλάσεων κ
< 20	5	200 - 400	9
20 - 50	6	400 - 700	10
50 - 100	7	700 - 1000	11
100 - 200	8	≥ 1000	12

• Το δεύτερο βήμα είναι ο προσδιορισμός του πλάτους των κλάσεων. **Πλάτος μιας κλάσης** ονομάζεται η διαφορά του κατωτέρου από το ανώτερο όριο της κλάσης. Στην πλειονότητα των πρακτικών εφαρμογών οι κλάσεις έχουν το ίδιο πλάτος. Φυσικά υπάρχουν και περιπτώσεις όπου επιβάλλεται οι κλάσεις να έχουν άνισο πλάτος, όπως, για παράδειγμα, στις κατανομές εισοδήματος, ημερών απεργίας κτλ. Για να κατασκευάσουμε ισοπλατείς κλάσεις, χρησιμοποιούμε το **εύρος** (range) R του δείγματος, δηλαδή τη διαφορά της μικρότερης παρατήρησης από τη μεγαλύτερη παρατήρηση του συνολικού δείγματος. Τότε υπολογίζουμε το πλάτος c των κλάσεων διαιρώντας το εύρος R διά του αριθμού των κλάσεων κ, στρογγυλεύοντας, αν χρειαστεί για λόγους διευκόλυνσης, πάντα προς τα πάνω.

• Το επόμενο βήμα είναι η κατασκευή των κλάσεων. Ξεκινώντας από την μικρότερη παρατήρηση, ή για

πρακτικούς λόγους λίγο πιο κάτω από την μικρότερη παρατήρηση, και προσθέτοντας κάθε φορά το πλάτος c δημιουργούμε τις κ κλάσεις. Αυτονόητο είναι ότι η μεγαλύτερη τιμή του δείγματος θα (πρέπει να) ανήκει οπωσδήποτε στην τελευταία κλάση.

- Τέλος, γίνεται η **διαλογή** των παρατηρήσεων. Το πλήθος των παρατηρήσεων v_i που προκύπτουν από τη διαλογή για την κλάση i καλείται **συχνότητα της κλάσης** αυτής ή **συχνότητα της κεντρικής τιμής** x_i, $i = 1, 2, ..., κ$.

Έστω, για παράδειγμα, ότι από τα δεδομένα του πίνακα 4 εξετάζουμε το ύψος των μαθητών. Το ύψος των μαθητών, όπως έχει καταγραφεί με τη σειρά, δίνεται στον παρακάτω πίνακα 8.

Πίνακας 8

Το ύψος (σε cm) των μαθητών της Γ΄ Λυκείου, όπως έχει καταγραφεί στον πίνακα 4. Σε αγκύλες έχουμε τη μικρότερη και τη μεγαλύτερη τιμή.

170	180	178	165	170	168	175	175	173	162
160	170	167	177	180	170	182	178	165	178
[156]	175	172	173	167	187	170	180	178	[191]
176	169	167	166	179	178	180	164	170	173

Παρατηρούμε ότι το εύρος του δείγματος είναι $R = 191 - 156 = 35$. Επειδή έχουμε $\nu = 40$ παρατηρήσεις, χρησιμοποιούμε $\kappa = 6$ κλάσεις. Το πλάτος των κλάσεων είναι $c = \frac{R}{\kappa} = \frac{35}{6} = 5{,}83 \approx 6$. Αν θεωρήσουμε ως αρχή της πρώτης κλάσης το 156, θα έχουμε τον επόμενο πίνακα 9.

Πρέπει να προσεχτεί ότι:

- Καμία παρατήρηση δεν μπορεί να μείνει έξω από κάποια κλάση.
- Οι κεντρικές τιμές διαφέρουν μεταξύ τους όσο και το πλάτος των κλάσεων, που εδώ είναι ίσο με 6.
- Μία παρατήρηση που συμπίπτει με το άνω άκρο μιας κλάσης θα τοποθετηθεί κατά τη διαλογή στην αμέσως επόμενη κλάση. Για παράδειγμα, ο μαθητής με ύψος 180 θα τοποθετηθεί στην πέμπτη κλάση [180, 186).

Πίνακας 9

Κατανομές συχνοτήτων (απόλυτων, σχετικών, αθροιστικών) για τα δεδομένα του πίνακα 8.

Κλάσεις [-)	Κεντρικές τιμές x_i	Διαλογή	Συχν. v_i	Σχετική Συχνότητα f_i%	Αθρ.συχν. N_i	Αθρ. Σχετ. Συχν. F_i%
156-162	159	||	2	5,0	2	5,0
162-168	165	𝍸 |||	8	20,0	10	25,0
168-174	171	𝍸 𝍸 ||	12	30,0	22	55,0
174-180	177	𝍸 𝍸 |	11	27,5	33	82,5
180-186	183	𝍸	5	12,5	38	95,0
186-192	189	||	2	5,0	40	100,0
	Σύνολο	—	40	100	—	—

Ιστόγραμμα Συχνοτήτων

Η αντίστοιχη γραφική παράσταση ενός πίνακα συχνοτήτων με ομαδοποιημένα δεδομένα γίνεται με το λεγόμενο ιστόγραμμα (histogram) συχνοτήτων. Στον οριζόντιο άξονα ενός συστήματος ορθογωνίων αξόνων σημειώνουμε, με κατάλληλη κλίμακα, τα όρια των κλάσεων. Στη συνέχεια, κατασκευάζουμε διαδοχικά ορθογώνια (ιστούς), από καθένα από τα οποία έχει βάση ίση με το πλάτος της κλάσης και ύψος τέτοιο, ώστε το εμβαδόν του ορθογωνίου να ισούται με τη συχνότητα της κλάσης αυτής.

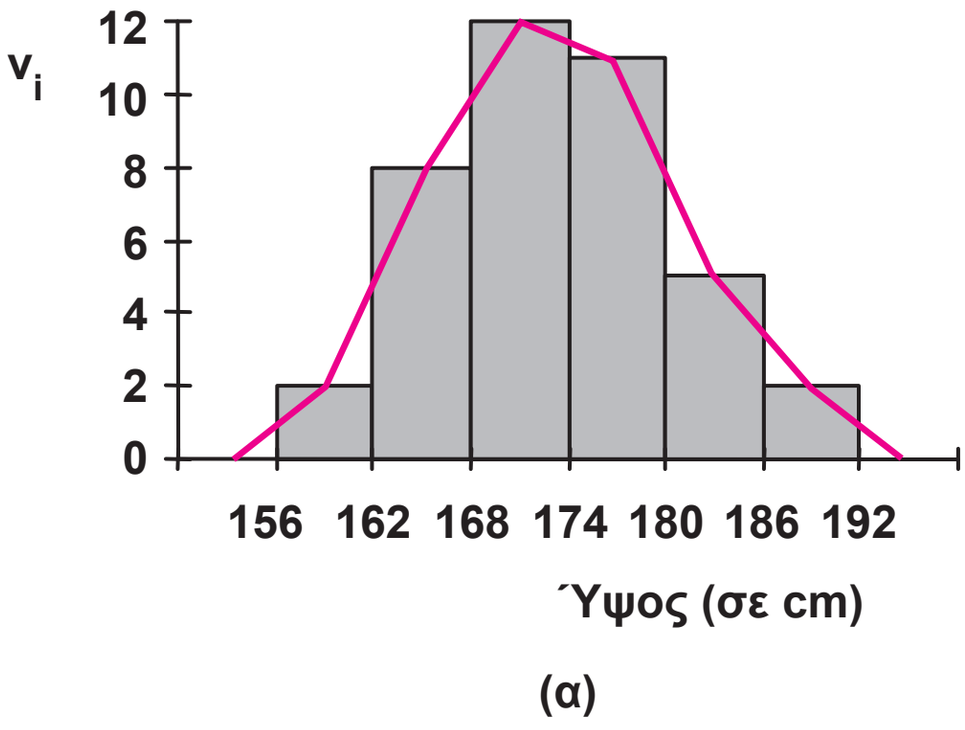
α) Κλάσεις Ίσου Πλάτους

Θεωρώντας το πλάτος c ως μονάδα μέτρησης του χαρακτηριστικού στον οριζόντιο άξονα, το ύψος κάθε ορθογωνίου είναι ίσο προς τη συχνότητα της αντίστοιχης κλάσης, έτσι ώστε να ισχύει πάλι ότι το εμβαδόν των ορθογωνίων είναι ίσο με τις αντίστοιχες συχνότητες. Επομένως, στον κατακόρυφο άξονα σε ένα ιστόγραμμα συχνοτήτων βάζουμε τις συχνότητες. Με ανάλογο τρόπο κατασκευάζεται και το ιστόγραμμα σχετικών συχνοτήτων, οπότε στον κάθετο άξονα βάζουμε τις σχετικές συχνότητες.

Αν στα ιστογράμματα συχνοτήτων θεωρήσουμε δύο ακόμη υποθετικές κλάσεις, στην αρχή και στο τέλος, με

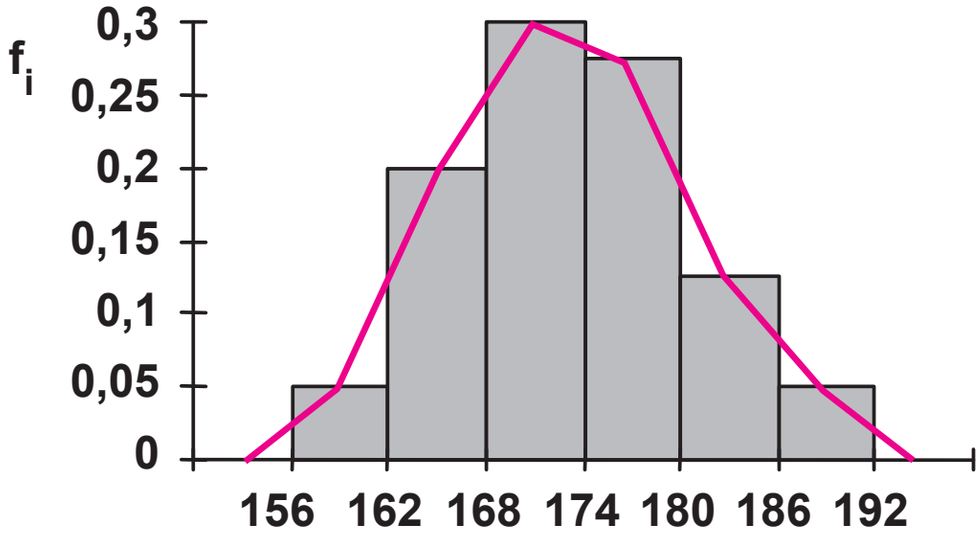
συχνότητα μηδέν και στη συνέχεια ενώσουμε τα μέσα των άνω βάσεων των ορθογωνίων με ευθύγραμμα τμήματα, σχηματίζεται το λεγόμενο πολύγωνο συχνοτήτων (frequency polygon). Το εμβαδόν του χωρίου που ορίζεται από το πολύγωνο συχνοτήτων και τον οριζόντιο άξονα είναι ίσο με το άθροισμα των συχνοτήτων, δηλαδή με το μέγεθος του δείγματος ν. Όμοια κατασκευάζεται από το ιστόγραμμα σχετικών συχνοτήτων και το πολύγωνο σχετικών συχνοτήτων με εμβαδόν ίσο με 1, (βλέπε σχήμα 6).

6α



(α)

6β

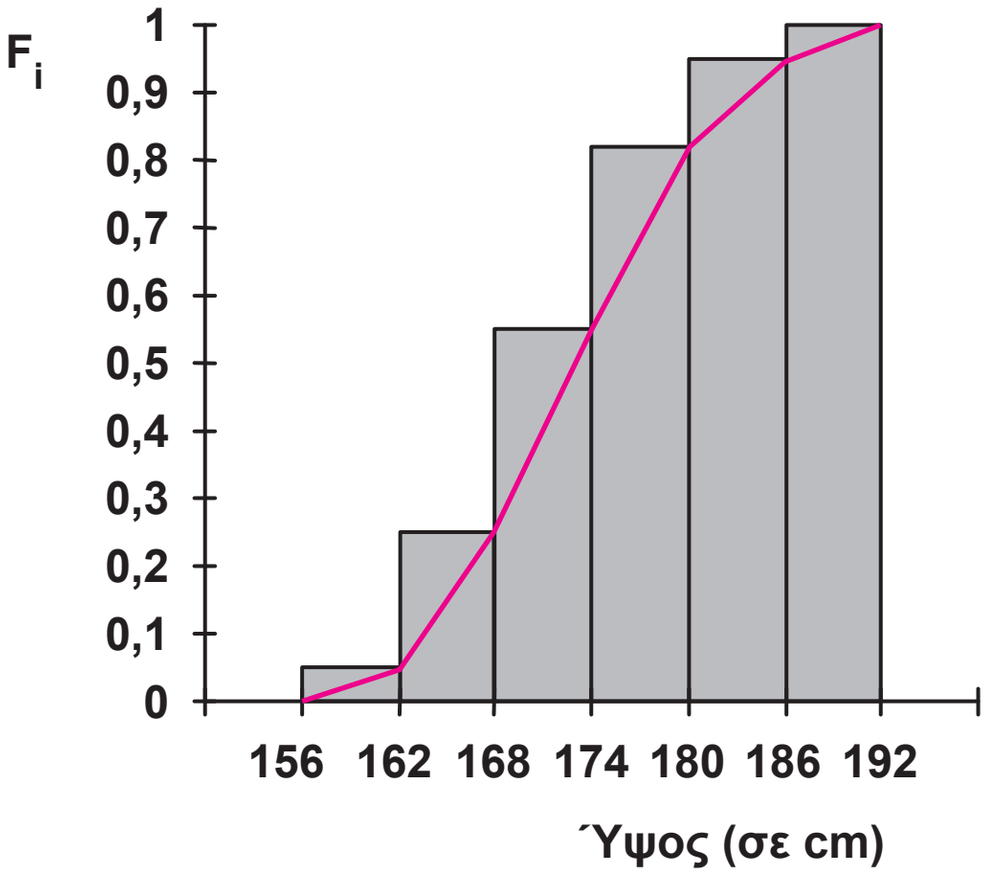


(β)

Ιστόγραμμα και πολύγωνο (α) συχνοτήτων και (β) σχετικών συχνοτήτων για τα δεδομένα του πίνακα 9.

Με τον ίδιο τρόπο κατασκευάζονται και τα **ιστογράμματα αθροιστικών συχνοτήτων** και **αθροιστικών σχετικών συχνοτήτων**. Αν ενώσουμε σε ένα ιστόγραμμα αθροιστικών συχνοτήτων τα **δεξιά άκρα** (όχι μέσα) των άνω βάσεων των ορθογωνίων με ευθύγραμμα τμήματα βρίσκουμε το **πολύγωνο αθροιστικών συχνοτήτων** (ogive) της κατανομής. Στο σχήμα 7 παριστάνεται το ιστόγραμμα και το πολύγωνο αθροιστικών σχετικών συχνοτήτων για το ύψος των μαθητών του πίνακα 9.

7



β) Κλάσεις Άνισου Πλάτους

Όπως προαναφέραμε, συνήθως επιλέγουμε κλάσεις ίσου πλάτους. Υπάρχουν όμως και περιπτώσεις που είναι απαραίτητο να έχουμε κλάσεις διαφορετικού πλάτους όπως, για παράδειγμα, στην κατανάλωση νερού και ηλεκτρικού ρεύματος ή ακόμα και περιπτώσεις όπου οι συχνότητες σε κάποιες κλάσεις να είναι πολύ μικρές οπότε γίνεται συγχώνευση κλάσεων.

Έστω, για παράδειγμα, η διάρκεια (σε sec) v = 80 τηλεφωνημάτων που έγιναν τυχαία από ένα κινητό τηλέφωνο, η οποία δίνεται στον παρακάτω πίνακα συχνοτήτων.

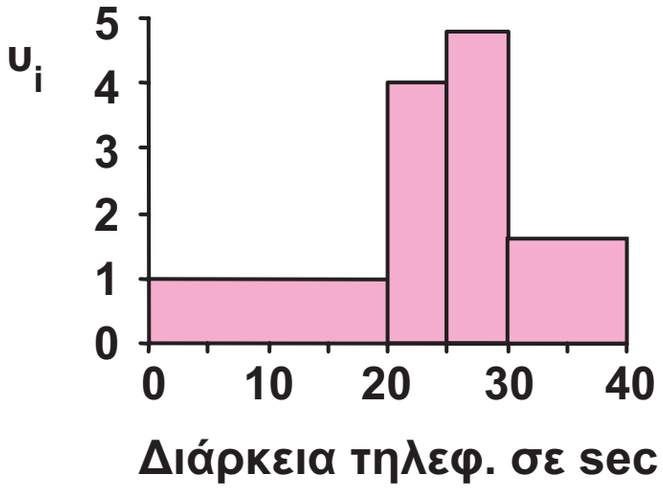
Διάρκεια τηλεφ. σε sec	Συχνότητα v_i
0-20	20
20-25	20
25-30	24
30-40	16
Σύνολο	v = 80

Το αντίστοιχο ιστόγραμμα συχνοτήτων κατασκευάζεται πάλι, έτσι ώστε το εμβαδόν κάθε ορθογωνίου να ισούται με τη συχνότητα της αντίστοιχης κλάσης. Άρα, αν c_i είναι το πλάτος της κλάσης i με συχνότητα v_i, το ύψος του ορθογωνίου θα είναι $u_i = \frac{v_i}{c_i}$, i = 1, 2, ..., κ. Επομένως, για την κατασκευή του ιστογράμματος συχνοτήτων χρειαζόμαστε τα πλάτη των κλάσεων και τα ύψη των ορθογωνίων. Αυτά δίνονται στον επόμενο πίνακα.

Διάρκεια τηλεφ. σε sec	Πλάτος κλάσης c_i	Συχνότητα v_i	Ύψος $u_i = \frac{v_i}{c_i}$	Ύψος $u_i^* = \frac{f_i\%}{c_i}$
0-20	20	20	1,0	1,25
20-25	5	20	4,0	5,00
25-30	5	24	4,8	6,00
30-40	10	16	1,6	2,00

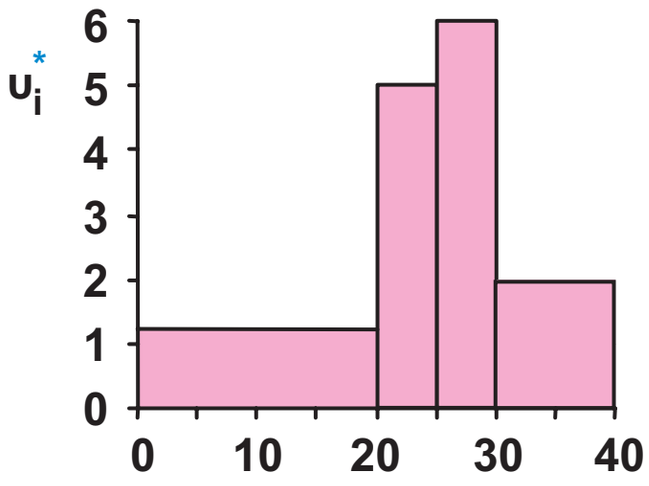
Τότε το ιστόγραμμα συχνοτήτων δίνεται στο σχήμα 8(α). Παρατηρούμε ότι το άθροισμα των εμβαδών όλων των ορθογωνίων είναι ίσο με το συνολικό μέγεθος δείγματος ν, όπως δηλαδή συμβαίνει και στο ιστόγραμμα με κλάσεις ίσου πλάτους.

8α



(α)

8β



(β)

Ιστόγραμμα συχνοτήτων (α) και σχετικών συχνοτήτων (β) της διάρκειας τηλεφωνημάτων.

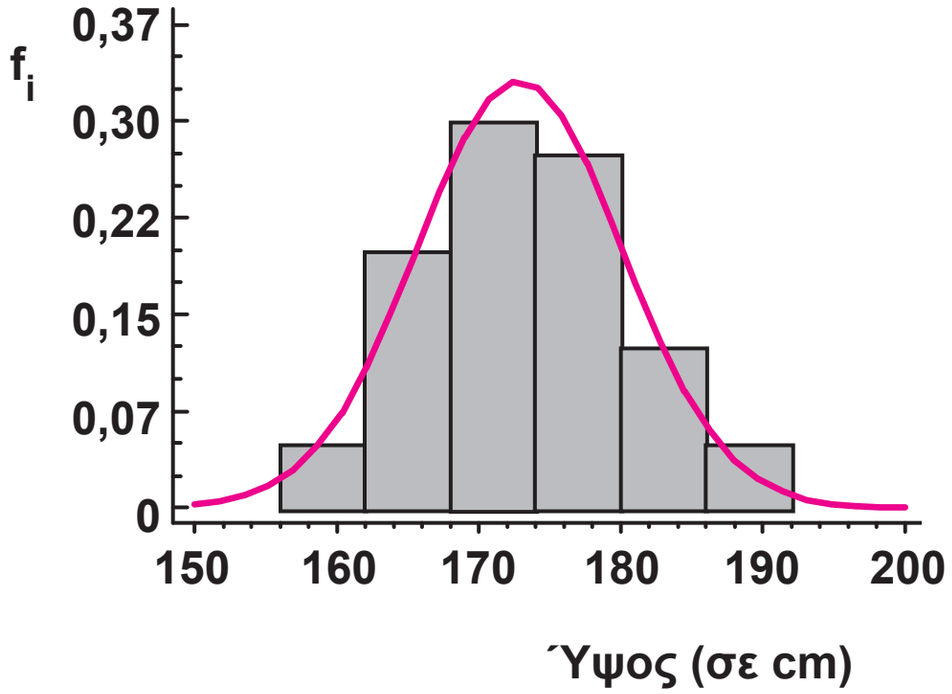
Με ανάλογο τρόπο κατασκευάζεται και το ιστόγραμμα σχετικών συχνοτήτων, (σχήμα 8(β)) αρκεί να χρησιμοποιήσουμε ως ύψος των ορθογωνίων το λόγο των σχετικών συχνοτήτων προς το πλάτος των κλάσεων, δηλαδή $u_i^* = \frac{f_i\%}{c_i}$.

Καμπύλες Συχνοτήτων

Εάν υποθέσουμε ότι ο αριθμός των κλάσεων για μια συνεχή μεταβλητή είναι αρκετά μεγάλος (τείνει στο άπειρο) και ότι το πλάτος των κλάσεων είναι αρκετά μικρό (τείνει στο μηδέν), τότε η πολυγωνική γραμμή

συχνοτήτων τείνει να πάρει τη μορφή μιας ομαλής καμπύλης, η οποία ονομάζεται **καμπύλη συχνοτήτων** (frequency curve), όπως δείχνει το σχήμα 9. Οι καμπύλες συχνοτήτων έχουν μεγάλη εφαρμογή στη Στατιστική, όπου οι ιδιότητες τους μπορούν να χρησιμοποιηθούν για την εξαγωγή χρήσιμων συμπερασμάτων.

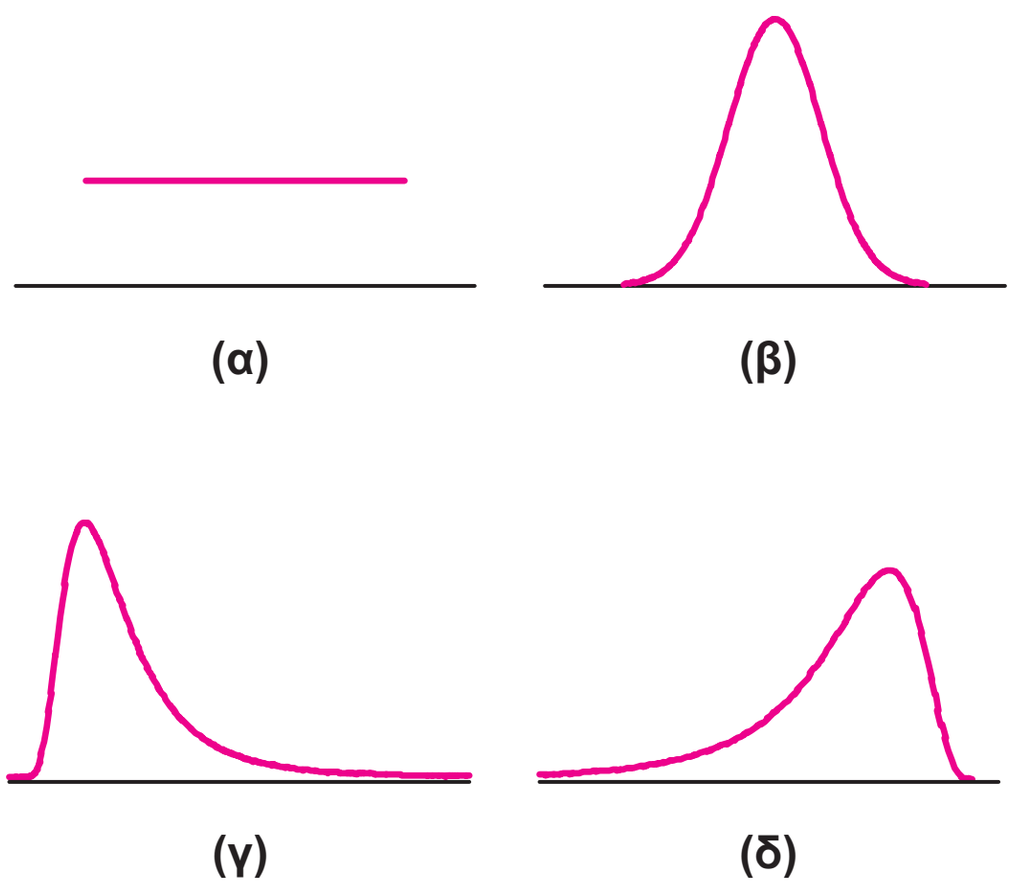
9



Καμπύλη συχνοτήτων για το ύψος των μαθητών του πίνακα 4

Η μορφή μιας κατανομής συχνοτήτων εξαρτάται από το πώς είναι κατανεμημένες οι παρατηρήσεις σε όλη την έκταση του εύρους τους. Μερικές χαρακτηριστικές καμπύλες συχνοτήτων που συναντάμε συχνά στις

εφαρμογές δίνονται στο σχήμα 10. Η κατανομή (β), με "κωδωνοειδή" μορφή λέγεται **κανονική κατανομή** (normal distribution) και παίζει σπουδαίο ρόλο στη Στατιστική. Όταν οι παρατηρήσεις "κατανέμονται" ομοιόμορφα σε ένα διάστημα [α, β], όπως στην κατανομή (α), η κατανομή λέγεται **ομοιόμορφη**. Όταν οι παρατηρήσεις δεν είναι συμμετρικά κατανεμημένες, η κατανομή λέγεται ασύμμετρη με θετική ασυμμετρία όπως στην κατανομή (γ) ή αρνητική ασυμμετρία όπως στην κατανομή (δ).



Μερικές χαρακτηριστικές κατανομές συχνοτήτων

ΕΦΑΡΜΟΓΕΣ

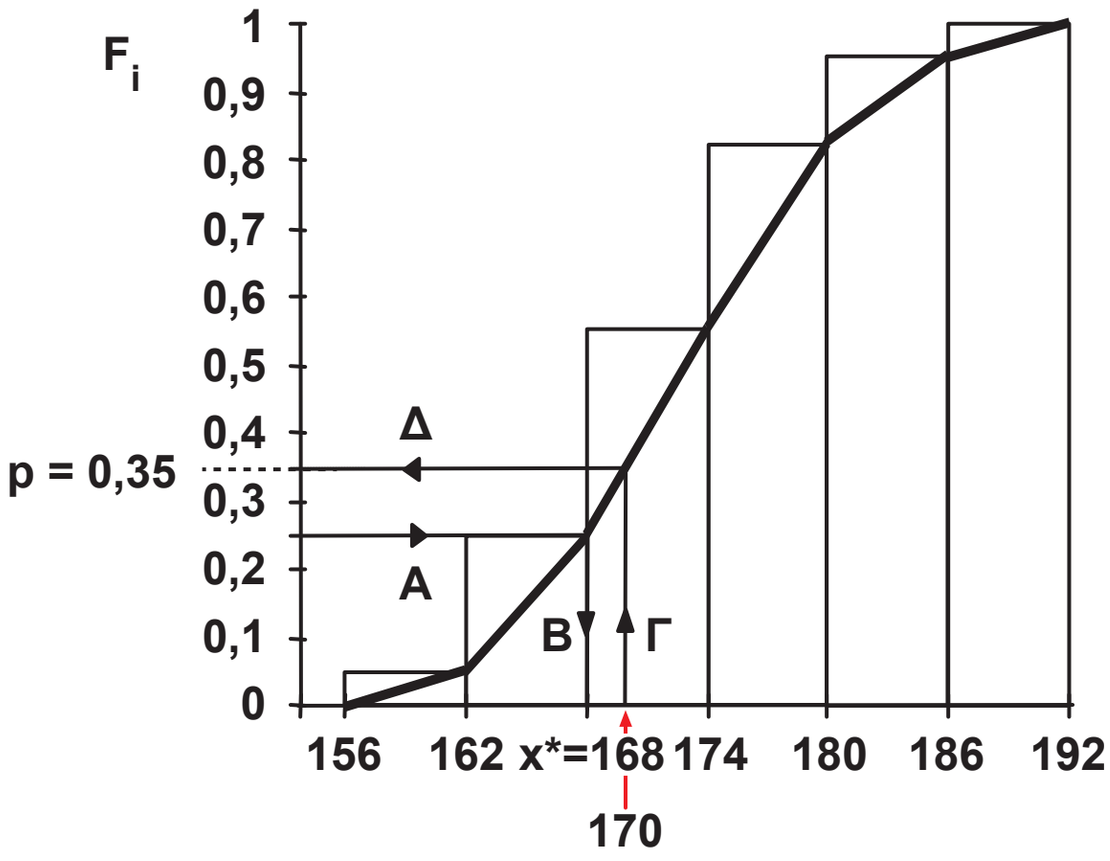
1. Από το πολύγωνο σχετικών αθροιστικών συχνοτήτων του παρακάτω διαγράμματος να βρεθεί

α) το ύψος x*, κάτω από το οποίο ανήκει το 25% των μαθητών

β) το ποσοστό p των μαθητών που έχουν ύψος μέχρι και 170 cm.

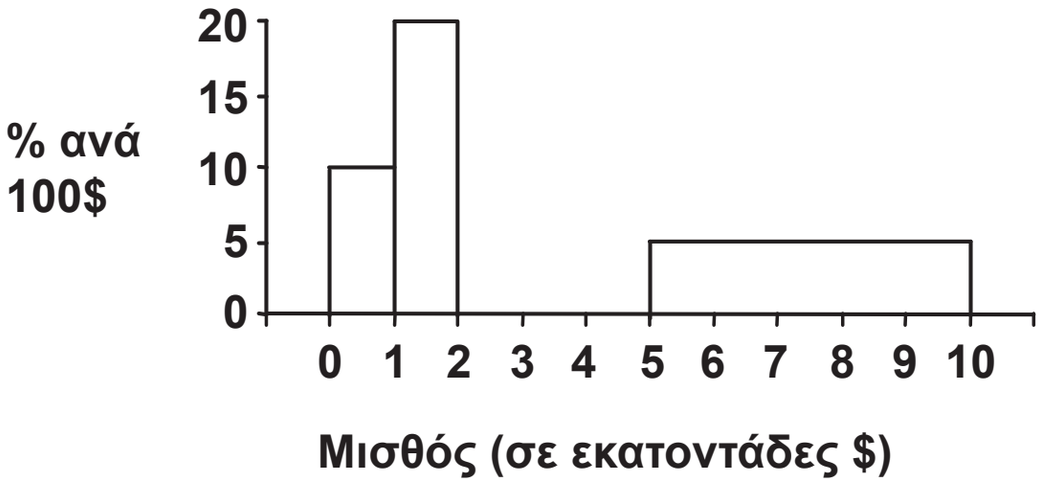
ΛΥΣΗ

α) Ακολουθούμε τη διαδρομή ΑΒ, όπως φαίνεται στο διάγραμμα, και ξεκινώντας από το σημείο (0, 0,25) πηγαίνουμε παράλληλα προς τον άξονα 0x μέχρι το αθροιστικό διάγραμμα και μετά κάθετα στον άξονα 0x μέχρι το σημείο (x*, 0). Το x* = 168 είναι το ζητούμενο ύψος.



β) Όμοια, ακολουθώντας τη διαδρομή ΓΔ από το σημείο (170, 0) καταλήγουμε, όπως φαίνεται στο σχήμα, στο σημείο (0, p). Το p = 0,35 = 35% είναι το ζητούμενο ποσοστό.

2. Στο παρακάτω ιστόγραμμα σχετικών συχνοτήτων σβήστηκε κατά λάθος το ορθογώνιο της κλάσης [2-5). Εάν είναι γνωστό ότι δεν υπάρχει μισθός άνω των $1000, να κατασκευάσετε το ορθογώνιο αυτό.



ΛΥΣΗ

Επειδή έχουμε ένα ιστόγραμμα σχετικών συχνοτήτων (f_i%), το άθροισμα των εμβαδών όλων των ορθογωνίων θα πρέπει να ισούται με 100. Το εμβαδόν του πρώτου ορθογωνίου είναι $E_1 = (1 - 0) \cdot 10 = 10$, του δεύτερου ορθογωνίου $E_2 = (2 - 1) \cdot 20 = 20$, και του τέταρτου $E_4 = (10 - 5) \cdot 5 = 25$. Άρα, το εμβαδόν του τρίτου ορθογωνίου θα είναι $E_3 = 100 - (10 + 20 + 25) = 45$.

Επειδή το πλάτος του ορθογωνίου είναι 5 − 2 = 3, το ύψος του θα είναι $\frac{45}{3}$ = 15, όπως φαίνεται στο παρακάτω σχήμα.

ΑΣΚΗΣΕΙΣ

Α΄ ΟΜΑΔΑΣ

1. Η βαθμολογία 50 φοιτητών στις εξετάσεις ενός μαθήματος είναι:

3	4	5	8	9	7	6	8	7	10
8	7	6	5	9	3	8	5	6	6
6	3	5	6	4	2	9	8	7	7
1	6	3	1	5	8	1	2	3	4
5	6	7	9	10	9	8	7	6	5

α) Να κατασκευάσετε τον πίνακα κατανομής συχνοτήτων και σχετικών συχνοτήτων (απολύτων και αθροιστικών).

β) Από τον πίνακα αυτό να εκτιμήσετε το ποσοστό των φοιτητών που πήραν βαθμό

i) κάτω από τη βάση (μικρότερο του 5)

ii) άριστα (9 ή 10) iii) τουλάχιστον 7 αλλά το πολύ 9.

2. Οι παραπάνω φοιτητές ήταν αντίστοιχα αγόρια (Α) ή κορίτσια (Κ):

Α	Α	Κ	Α	Κ	Α	Α	Α	Κ	Κ
Κ	Κ	Α	Α	Α	Κ	Α	Κ	Α	Α
Α	Α	Α	Α	Κ	Κ	Α	Κ	Α	Κ
Κ	Κ	Κ	Α	Κ	Κ	Α	Α	Α	Α
Α	Α	Κ	Α	Κ	Κ	Α	Α	Α	Κ

Να συμπληρώσετε τον επόμενο πίνακα χρησιμοποιώντας απόλυτες συχνότητες.

Φύλο	Βαθμολογία		Σύνολο	
	≤ 5	> 5		
Α				
Κ				
Σύνολο				

3. **Να μετατρέψετε τον προηγούμενο πίνακα συχνοτήτων της άσκησης 2 σε πίνακα σχετικών συχνοτήτων επί τοις εκατό:**
α) ως προς το σύνολο των φοιτητών
β) ως προς το φύλο (γραμμές)
γ) ως προς τη βαθμολογία (στήλες)
και να ερμηνεύσετε τα αποτελέσματα.

4. **Χρησιμοποιώντας τον παρακάτω πίνακα συχνοτήτων, που δίνει την κατανομή του αριθμού των ημερών απουσίας από την εργασία τους λόγω ασθένειας 50 εργατών, να βρεθεί ο αριθμός και το ποσοστό των εργατών που απουσίασαν:**

α) τουλάχιστον 1 ημέρα

β) πάνω από 5 ημέρες

γ) από 3 έως 5 ημέρες

δ) το πολύ 5 ημέρες

ε) ακριβώς 5 ημέρες.

Αριθμός ημερών	Συχνότητα	Αριθμός ημερών	Συχνότητα
0	12	5	8
1	8	6	0
2	5	7	5
3	4	8	2
4	5	9	1

5. Να συμπληρώσετε τον παρακάτω πίνακα.

x_i	v_i	f_i	N_i	F_i	f_i%	F_i%
1						10
2	4	0,20	6			
3				0,60		
4					25	
5	2					
6						
Σύνολο						

6. Να κατασκευάσετε το διάγραμμα συχνοτήτων του βαθμού Μαθηματικών για τα αγόρια και κορίτσια (χωριστά) του πίνακα 4.

7. Τα δημοφιλέστερα ξένα μουσικά συγκροτήματα των 18 αγοριών του πίνακα 4 ήσαν:
Metallica, Iron Maiden, Άλλο, Scorpions, Oasis, Άλλο, Άλλο, Rolling Stones, Metallica, Metallica, Rolling Stones, Metallica, Iron Maiden, Iron Maiden, Scorpions, Scorpions, Scorpions, Metallica.
Να κατασκευάσετε α) το ραβδόγραμμα και β) το κυκλικό διάγραμμα σχετικών συχνοτήτων.

8. Σε ένα κυκλικό διάγραμμα παριστάνεται η βαθμολογία των 450 μαθητών ενός Γυμνασίου σε τέσσερις κατηγορίες “Άριστα”, “Λίαν Καλώς”, “Καλώς” και “Σχεδόν Καλώς”. Το 30% των μαθητών έχουν επίδοση “Λίαν Καλώς”. Η γωνία του κυκλικού τομέα για την επίδοση “Καλώς” είναι 144°. Οι μαθητές με βαθμό “Σχεδόν Καλώς” είναι διπλάσιοι των μαθητών με “Άριστα”. Να μετατρέψετε το κυκλικό διάγραμμα σε ραβδόγραμμα σχετικών συχνοτήτων. Πόσοι μαθητές έχουν επίδοση τουλάχιστον λίαν καλώς;

9. Από το 1960 έως το 1998 (Πρωταθλήματα Α΄ Εθνικής) ο Παναθηναϊκός έχει κατακτήσει 15 τίτλους, ο Ολυμπιακός 12, η ΑΕΚ 9, ο ΠΑΟΚ 2 και η Λάρισα 1. Να κατασκευάσετε το ραβδόγραμμα και το κυκλικό διάγραμμα σχετικών συχνοτήτων.

10. Παρακάτω δίνονται τα μετάλλια που πήραν μερικές χώρες στο 17ο Ευρωπαϊκό Πρωτάθλημα Στίβου, το 1998. Να παρασταθούν τα δεδομένα αυτά σε ένα ραβδόγραμμα.

Χώρα	Χρυσά	Ασημένια	Χάλκινα
Μ. Βρετανία	9	4	3
Γερμανία	8	7	8
Ρωσία	6	9	7
Πολωνία	3	4	1
Ρουμανία	3	2	2
Ουκρανία	3	2	1
Ιταλία	2	4	3
Πορτογαλία	2	3	1
Ισπανία	2	1	4
Γαλλία	2	1	1
Ελλάδα	1	0	2

11. Τα κρούσματα δύο λοιμωδών νόσων από το 1987 έως το 1997 δίνονται στον παρακάτω πίνακα. (Πηγή: ΕΚΕΠΑΠ.)
Να κατασκευάσετε τα αντίστοιχα χρονογράμματα και να τα σχολιάσετε.

Έτος	Έρπης ζωστήρ	Ηπατίτιδα Α
1987	85	351
1988	58	254
1989	123	273
1990	178	172
1991	134	213
1992	201	127
1993	241	123
1994	252	259
1995	338	295
1996	296	107
1997	256	131

12. Τα παρακάτω δεδομένα αντιπροσωπεύουν την επίδοση 50 υποψηφίων για την πρόσληψή τους σε μια ιδιωτική σχολή (κλίμακα 0-10).

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

α) Να παραστήσετε τα δεδομένα σε έναν πίνακα συχνοτήτων.

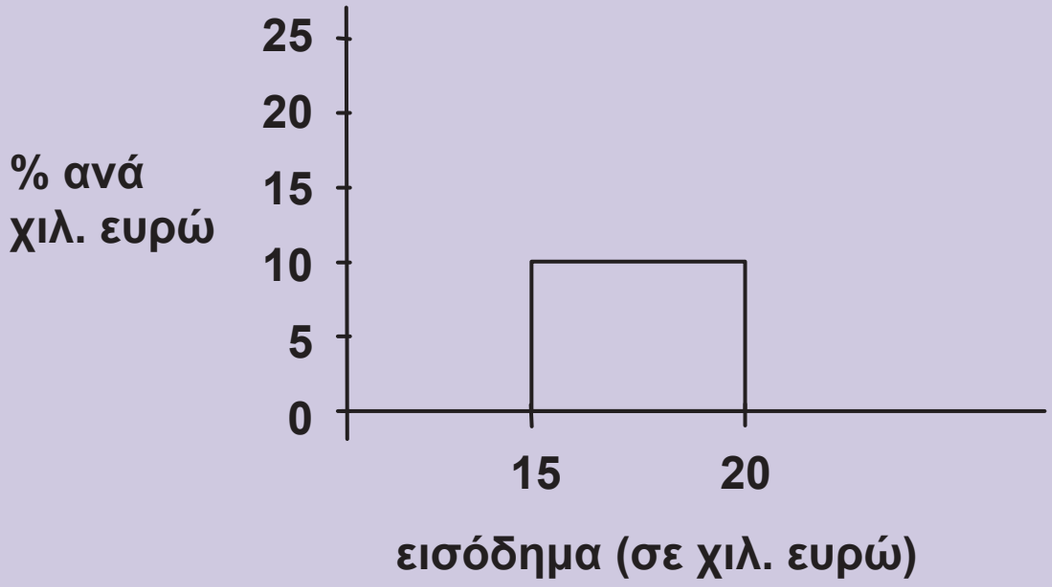
β) Να κατασκευάσετε το διάγραμμα σχετικών

και αθροιστικών σχετικών συχνοτήτων.

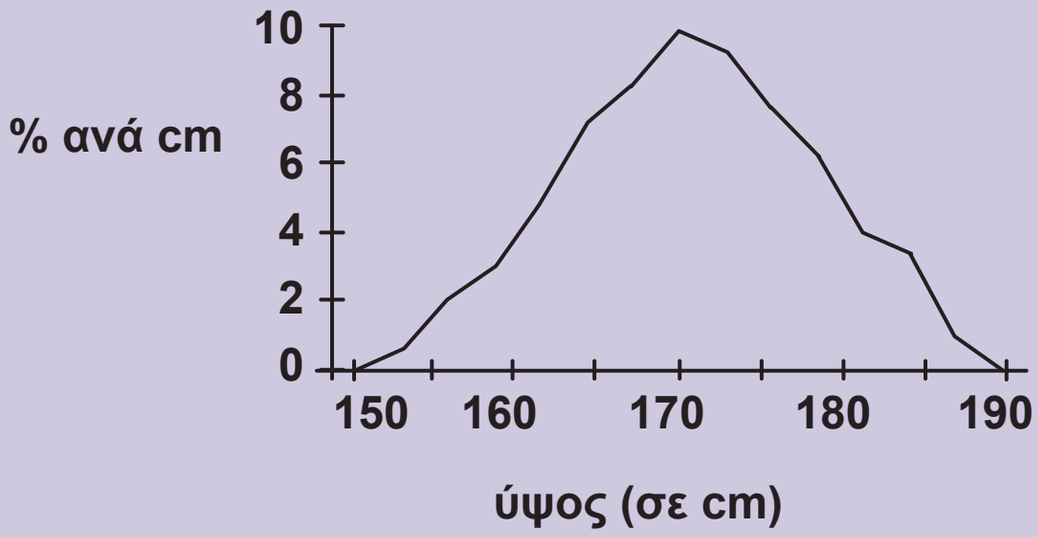
γ) Αν η σχολή θελήσει να πάρει όσους είχαν επίδοση μεγαλύτερη ή ίση του 8, πόσους θα πάρει;

δ) Αν η σχολή πάρει μόνο το 36% των υποψηφίων, τι επίδοση πρέπει να έχει κάποιος για να επιλεγεί;

13. Παρακάτω δίνεται μόνο ένα ορθογώνιο από το ιστόγραμμα του ετήσιου εισοδήματος των οικογενειών μιας περιοχής. Τι ποσοστό οικογενειών της περιοχής είχαν εισόδημα 15.000 ευρώ έως 20.000 ευρώ;



14. Ένας μαθητής έκανε το παρακάτω πολύγωνο σχετικών συχνοτήτων για το ύψος των αγοριών της τάξης του και ο καθηγητής το διέγραψε σαν λάθος. Είχε δίκιο ο καθηγητής;



Β΄ ΟΜΑΔΑΣ

1. Να κατασκευάσετε τα αντίστοιχα χρονογράμματα για τον πληθυσμό των νησιών α) Λέσβου, β) Θάσου, γ) Σαλαμίνας με βάση τα δεδομένα του πίνακα 2. Τι συμπέρασμα συνάγετε;

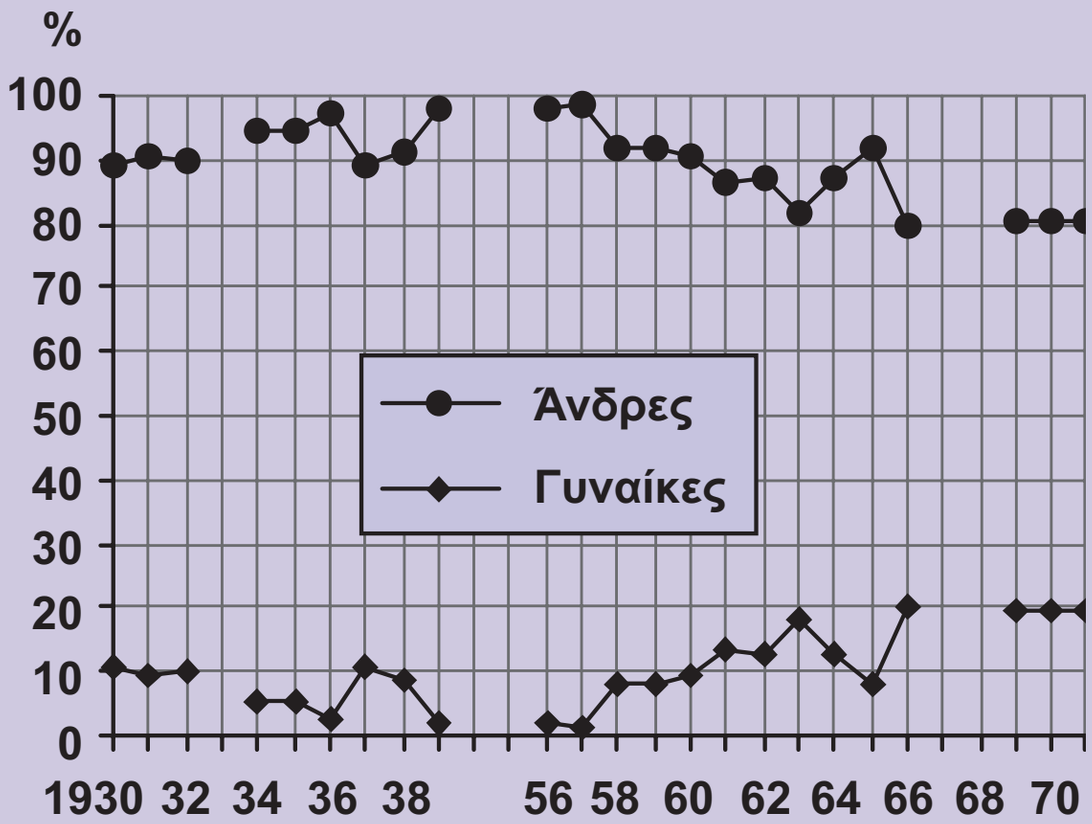
2. Οι βεβαιωθέντες θάνατοι από χρήση ναρκωτικών κατά τα έτη 1988-1998 (για το 1998 έως 8 Απριλίου) σύμφωνα με τον Οργανισμό κατά

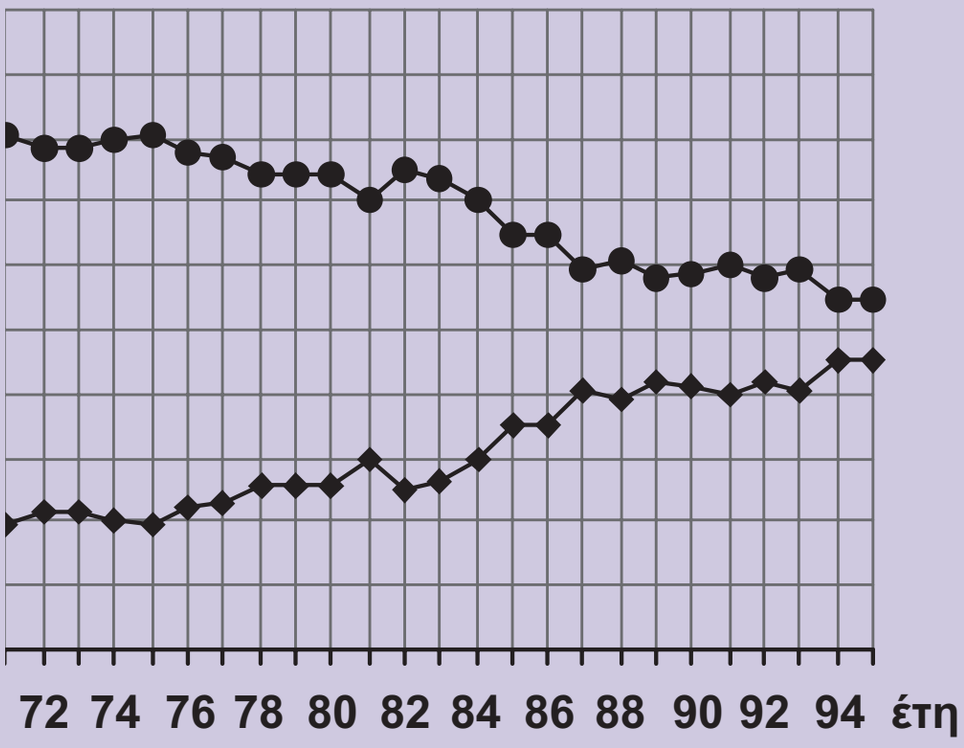
των Ναρκωτικών (ΟΚΑΝΑ) ήταν 62, 72, 66, 79, 79, 78, 146, 176, 222, 222 και 65 αντίστοιχα. Από αυτούς είχαμε 7, 4, 2, 2, 1, 4, 8, 7, 14, 22 και 6 μέχρι και 20 ετών, 43, 51, 34, 44, 47, 49, 71, 90, 98, 99 και 33 από 21-30 ετών και οι υπόλοιποι ήσαν άνω των 30 ετών.
Να παρασταθούν τα δεδομένα αυτά σε μορφή πίνακα.

3. Να παρασταθούν τα παραπάνω δεδομένα της άσκησης 2 σε μορφή πίνακα αναφορικά με το έτος και το φύλο των ατόμων, αν γνωρίζουμε ότι από τους βεβαιωθέντες θανάτους από χρήση ναρκωτικών κατά τα έτη 1988-1998 οι 8, 10, 7, 5, 9, 8, 11, 14, 20, 20 και 9 αντίστοιχα ήταν γυναίκες.

4. Το παρακάτω χρονόγραμμα δίνει τη σχετική συχνότητα των νέων πτυχιούχων Μαθηματικών σε όλη την Ελλάδα από το 1930 έως το 1995 ανάλογα με το φύλο. α) Μελετώντας προσεκτικά το χρονόγραμμα αυτό ποια συμπεράσματα εξάγονται; β) Ο συνολικός αριθμός νέων πτυχιούχων Μαθηματικών το έτος 1995 ήταν 789.

Πόσες ήσαν οι γυναίκες και πόσοι οι άνδρες; γ) Ο αριθμός των γυναικών που έγιναν πτυχιούχοι Μαθηματικών το έτος 1974 ήσαν 173. Πόσοι ήσαν οι άνδρες που έγιναν πτυχιούχοι Μαθηματικοί το ίδιο έτος; δ) Πόσοι άνδρες και πόσες γυναίκες πήραν πτυχίο Μαθηματικών στην Ελλάδα το 1985;





5. Να δοθεί και να ερμηνευτεί το χρονόγραμμα των δεδομένων του πίνακα 1 για κάθε ομάδα ηλικιών.

6. Στον παρακάτω πίνακα δίνεται η κατανομή συχνοτήτων της συστολικής πίεσης 150 γυναικών ηλικίας 17-24 ετών που χρησιμοποιούν το φάρμακο Α για κάποια πάθηση και 200 γυναικών, ανάλογης ηλικίας, που χρησιμοποιούν το φάρμακο Β.

α) Να συγκρίνετε τα ποσοστά γυναικών που παίρνουν τα φάρμακα Α και Β και έχουν συστολική πίεση μεγαλύτερη ή ίση των 130 mm Hg

β) Να κατασκευάσετε τα πολύγωνα αθροιστικών σχετικών συχνοτήτων, χρησιμοποιώντας τους ίδιους άξονες συντεταγμένων.

Συστολική πίεση (σε mm Hg)	Φάρμακο Α	Φάρμακο Β
	v_i	v_i
95-99	6	4
100-104	15	14
105-109	16	18
110-114	22	24
115-119	30	32
120-124	20	28
125-129	15	28
130-134	12	26
135-139	6	12
140-144	5	8
145-149	3	6
Σύνολο	150	200

Πηγή: Υποθετικά δεδομένα

7. Οι χρόνοι (σε λεπτά) που χρειάστηκαν 55 μαθητές να λύσουν ένα πρόβλημα δίνονται παρακάτω:

3,4	13,2	6,7	1,4	1,3	3,8	3,9	2,9	13,8	3,9	2,7
4,4	3,6	1,4	2,4	3,6	3,1	7,5	6,9	7,8	12,7	3,9
3,3	9,7	2,0	4,4	3,3	8,7	3,9	11,6	5,6	9,0	3,4
1,4	3,5	2,8	10,4	11,9	12,3	2,9	2,8	1,5	4,1	5,9
3,1	8,7	2,8	3,8	13,0	3,0	6,4	3,2	5,9	7,0	8,2

α) Να ομαδοποιήσετε τα δεδομένα σε κατάλληλο αριθμό κλάσεων.

β) Να κατασκευάσετε τον πίνακα με τις συχνότητες v_i, $f_i\%$, N_i, $F_i\%$.

γ) Να κατασκευάσετε το πολύγωνο σχετικών συχνοτήτων και αθροιστικών σχετικών συχνοτήτων.

2.3 ΜΕΤΡΑ ΘΕΣΗΣ ΚΑΙ ΔΙΑΣΠΟΡΑΣ

Εισαγωγή

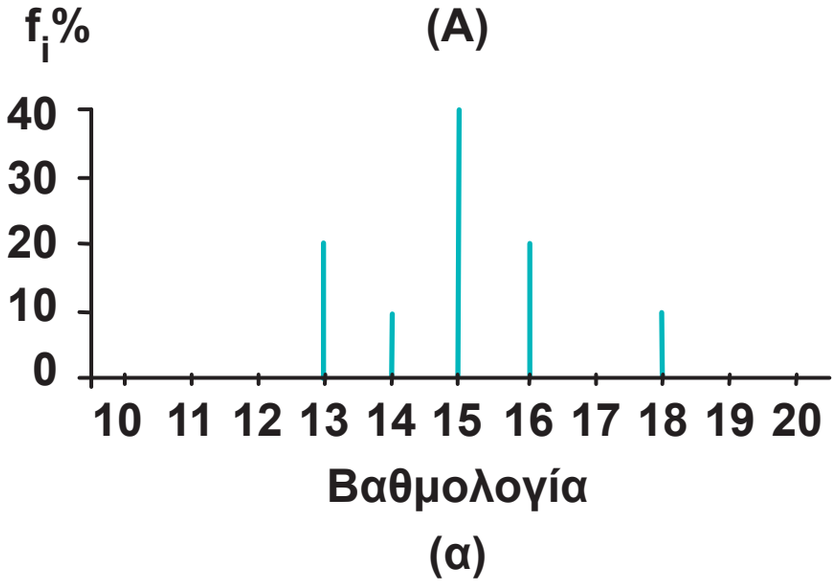
Εκτός από τους στατιστικούς πίνακες και τα διαγράμματα υπάρχουν και αριθμητικά μέτρα με τα οποία μπορούμε να περιγράψουμε με συντομία μια κατανομή συχνοτήτων. Η γνώση των μέτρων αυτών διευκολύνει και την παραπέρα στατιστική επεξεργασία των δεδομένων. Έστω, για παράδειγμα, ένας καθηγητής ο οποίος, για να συγκρίνει δύο διαφορετικά τμήματα Α και Β της ίδιας τάξης ως προς την επίδοσή τους σε ένα μάθημα, πήρε τυχαία 10 μαθητές από κάθε τμήμα. Η βαθμολογία τους στο μάθημα αυτό ήταν:

Τμήμα Α: 13 13 14 15 15 15 15 16 16 18

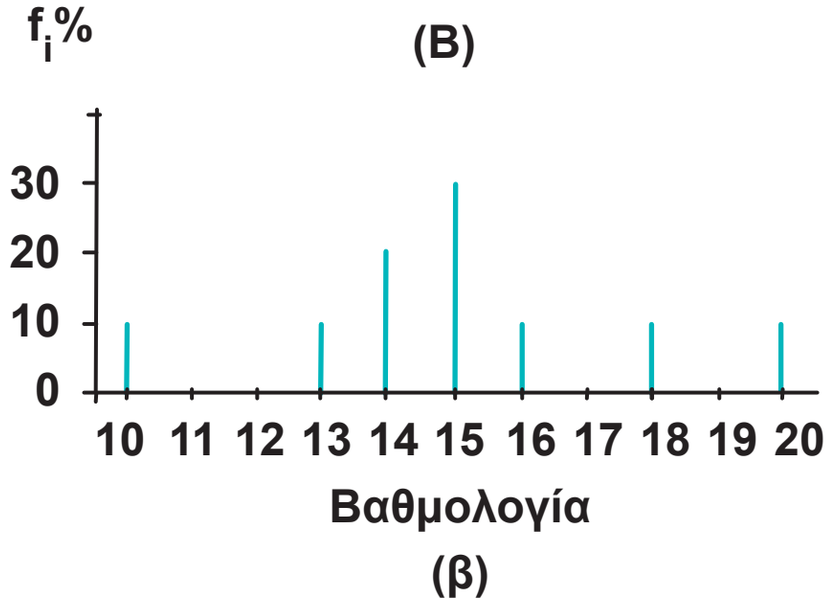
Τμήμα Β: 10 13 14 14 15 15 15 16 18 20.

Τα διαγράμματα σχετικών συχνοτήτων δίνονται στα σχήματα 11(α), (β).

11



(α)



(β)

Παρατηρούμε ότι η βαθμολογία και των δύο τμημάτων είναι συγκεντρωμένη γύρω στο 15, αλλά το δεύτερο τμήμα παρουσιάζει μεγαλύτερη διασπορά βαθμών από το πρώτο. Δηλαδή, οι βαθμοί του Β΄ τμήματος είναι περισσότερο διασκορπισμένοι γύρω από μια "κεντρική"

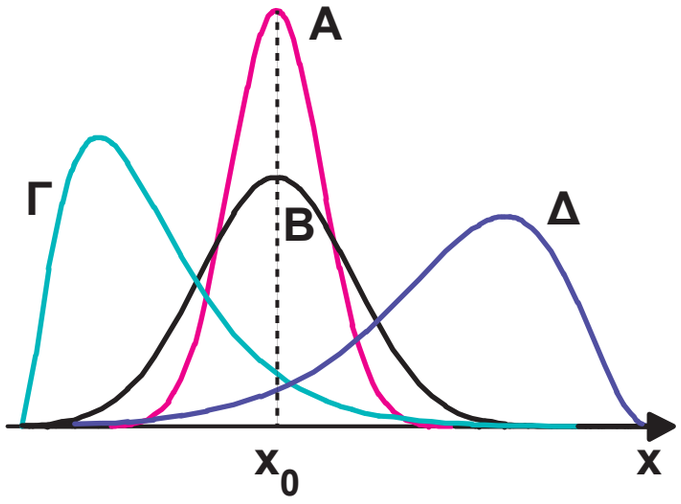
τιμή. Οι έννοιες “κεντρική τιμή” και “διασπορά των παρατηρήσεων” μας δίνουν το ερέθισμα για έναν ακόμα πιο σύντομο τρόπο περιγραφής της κατανομής ενός συνόλου δεδομένων. Για να ορίσουμε δηλαδή κάποια **μέτρα** (αριθμητικά μεγέθη), που να μας δίνουν α) τη θέση του “κέντρου” των παρατηρήσεων στον οριζόντιο άξονα και β) τη διασπορά των παρατηρήσεων, δηλαδή πόσο αυτές εκτείνονται γύρω από το “κέντρο” τους. Τα πρώτα τα καλούμε **μέτρα θέσης** της κατανομής (location measures), ενώ τα δεύτερα **μέτρα διασποράς** ή **μέτρα μεταβλητότητας** (measures of variability).

Εκτός από τα μέτρα θέσης και διασποράς μιας κατανομής πολλές φορές είναι απαραίτητος και ο προσδιορισμός κάποιων άλλων μέτρων, που καθορίζουν τη **μορφή** της κατανομής. Κατά πόσο δηλαδή η αντίστοιχη καμπύλη συχνοτήτων είναι συμμετρική ή όχι ως προς την ευθεία $x = x_0$, για δεδομένο σημείο x_0 του άξονα 0x. Τα μέτρα αυτά, που συνήθως εκφράζονται σε συνάρτηση με τα μέτρα θέσης και διασποράς, καλούνται **μέτρα ασυμμετρίας** (measures of skewness).

Υπολογίζοντας από ένα σύνολο δεδομένων κάποια από τα ανωτέρω μέτρα, μπορούμε να έχουμε μια σύντομη περιγραφή της μορφής της καμπύλης συχνοτήτων. Στο σχήμα 12 οι καμπύλες συχνοτήτων Α και Β είναι συμμετρικές με το ίδιο “κέντρο” x_0, αλλά η Β έχει μεγαλύτερη μεταβλητότητα από την Α. Οι καμπύλες Γ και Δ

είναι ασύμμετρες, με τη Γ όπως λέμε να παρουσιάζει θετική ασυμμετρία και τη Δ αρνητική ασυμμετρία. Το "κέντρο" της Γ είναι αριστερότερα του x_0, ενώ της Δ είναι δεξιότερα του x_0. Η Δ παρουσιάζει μεγαλύτερη μεταβλητότητα από τη Γ.

12



Μέτρα Θέσης

Τα πιο συνηθισμένα μέτρα που χρησιμοποιούνται για την περιγραφή της θέσης ενός συνόλου δεδομένων πάνω στον οριζόντιο άξονα 0x, εκφράζοντας την "κατά μέσο όρο" απόστασή τους από την αρχή των αξόνων, είναι ο αριθμητικός μέσος ή μέση τιμή (arithmetic mean or average), η διάμεσος (median) και η κορυφή ή επικρατούσα τιμή (mode).

α) Μέση Τιμή ($\bar{x}$)

Η μέση τιμή ενός συνόλου ν παρατηρήσεων αποτελεί το σπουδαιότερο και χρησιμότερο μέτρο της Στατιστικής και ορίζεται ως το άθροισμα των παρατηρήσεων διά του πλήθους των παρατηρήσεων.

Όταν σε ένα δείγμα μεγέθους ν οι παρατηρήσεις μιας μεταβλητής Χ είναι t_1, t_2, ..., t_v τότε η μέση τιμή συμβολίζεται με $\bar{x}$ και δίνεται από τη σχέση:

$$\bar{x} = \frac{t_1 + t_2 + \ldots + t_v}{v} = \frac{\sum_{i=1}^{v} t_i}{v} = \frac{1}{v}\sum_{i=1}^{v} t_i \qquad (1)$$

όπου το σύμβολο $\sum_{i=1}^{v} t_i$ παριστάνει μια συντομογραφία του αθροίσματος $t_1 + t_2 + \ldots + t_v$ και διαβάζεται "άθροισμα των t_i από i = 1 έως ν". Συχνά, όταν δεν υπάρχει πρόβλημα σύγχυσης, συμβολίζεται και ως $\sum t_i$ ή ακόμα πιο απλά με $\sum t$.

Σε μια κατανομή συχνοτήτων, αν x_1, x_2, ..., $x_κ$ είναι οι τιμές της μεταβλητής Χ με συχνότητες v_1,v_2, ..., $v_κ$ αντίστοιχα, η μέση τιμή ορίζεται ισοδύναμα από τη σχέση:

$$\bar{x} = \frac{x_1 v_1 + x_2 v_2 + \ldots + x_\kappa v_\kappa}{v_1 + v_2 + \ldots + v_\kappa} = \frac{\sum_{i=1}^{\kappa} x_i v_i}{\sum_{i=1}^{\kappa} v_i} = \frac{1}{v} \sum_{i=1}^{\kappa} x_i v_i \qquad (2)$$

Η παραπάνω σχέση ισοδύναμα γράφεται:

$$\bar{x} = \sum_{i=1}^{\kappa} x_i \frac{v_i}{v} = \sum_{i=1}^{\kappa} x_i f_i$$

όπου f_i οι σχετικές συχνότητες.

Για παράδειγμα, η μέση επίδοση των μαθητών στο τμήμα Α θα είναι σύμφωνα με την (1)

$$\bar{x}_A = \frac{13 + 13 + 14 + \ldots + 18}{10} = \frac{150}{10} = 15$$

ή ισοδύναμα από τον αντίστοιχο πίνακα συχνοτήτων

σύμφωνα με την (2).

Βαθμός x_i	Συχνότητα v_i	$x_i v_i$
13	2	26
14	1	14
15	4	60
16	2	32
18	1	18
Σύνολο	$v_A = 10$	$\sum x_i v_i = 150$

$$\bar{x}_A = \frac{\sum x_i v_i}{v_A} = \frac{150}{10} = 15.$$

Ομοίως, υπολογίζεται και η μέση επίδοση για το τμήμα Β, η οποία είναι πάλι

$$\bar{x}_B = 15.$$

Επίσης, το μέσο ύψος των 40 μαθητών της Γ΄ Λυκείου του πίνακα 8, σύμφωνα με τη σχέση (1) είναι

$\bar{x} = \frac{6918}{40} = 172,95$ cm.

Για ευκολότερο όμως υπολογισμό χρησιμοποιούμε τον πίνακα συχνοτήτων, όπως αυτός δίνεται παρακάτω, ομαδοποιώντας τα δεδομένα σε κ = 6 κλάσεις.

Αν x_i είναι το κέντρο της i κλάσης και v_i η αντίστοιχη

συχνότητα, τότε σύμφωνα με τη σχέση (2) η μέση τιμή θα είναι:

$$\bar{x} = \frac{\sum x_i v_i}{v} = \frac{6930}{40} = 173{,}25 \text{ cm.}$$

Παρατηρούμε ότι οι δύο μέσες τιμές του ίδιου συνόλου δεδομένων δεν είναι ακριβώς οι ίδιες. Πού οφείλεται αυτή η, έστω και μικρή, διαφορά;

Η διαφορά αυτή οφείλεται στο γεγονός ότι κατά την ομαδοποίηση υποθέσαμε ότι οι παρατηρήσεις κάθε κλάσης είναι ομοιόμορφα κατανεμημένες και ότι οι τιμές της μεταβλητής σε κάθε κλάση εκπροσωπούνται από την αντίστοιχη κεντρική τιμή x_i. Η υπόθεση αυτή σημαίνει απώλεια πληροφοριών για τις αρχικές τιμές. Χάνουμε λοιπόν λίγο ως προς την ακρίβεια κερδίζουμε όμως χρόνο!

Ύψος σε cm	Κεντρικές τιμές x_i	Συχνότητα v_i	$x_i\, v_i$
156-162	159	2	318
162-168	165	8	1320
168-174	171	12	2025
174-180	177	11	1947
180-186	183	5	915
186-192	189	2	378
	Σύνολο	$\sum v_i = 40$	$\sum x_i v_i = 6930$

β) Σταθμικός Μέσος

Στις περιπτώσεις που δίνεται διαφορετική βαρύτητα (έμφαση) στις τιμές $x_1, x_2, ..., x_v$ ενός συνόλου δεδομένων, τότε αντί του αριθμητικού μέσου χρησιμοποιούμε τον **σταθμισμένο αριθμητικό μέσο** ή **σταθμικό μέσο** (weighted mean). Εάν σε κάθε τιμή $x_1, x_2, ..., x_v$ δώσουμε διαφορετική βαρύτητα, που εκφράζεται με τους λεγόμενους συντελεστές στάθμισης (βαρύτητας) $w_1, w_2, ..., w_v$, τότε ο σταθμικός μέσος βρίσκεται από τον τύπο:

$$\bar{x} = \frac{x_1 w_1 + x_2 w_2 + \ldots + x_v w_v}{w_1 + w_2 + \ldots + w_v} = \frac{\sum_{i=1}^{v} x_i w_i}{\sum_{i=1}^{v} w_i}.$$

Για παράδειγμα, με το νέο σύστημα, για την εισαγωγή ενός μαθητή στην τριτοβάθμια εκπαίδευση θα συνυπολογίζονται ο βαθμός x_1 του απολυτηρίου του Ενιαίου Λυκείου με συντελεστή (βάρος) $w_1 = 7,5$, ο βαθμός x_2 στο τεστ δεξιοτήτων με συντελεστή $w_2 = 1$, ο βαθμός x_3 στο 1ο βασικό μάθημα με συντελεστή $w_3 = 1$ και ο βαθμός x_4 στο 2ο βασικό μάθημα με συντελεστή $w_4 = 0,5$. Εάν ένας μαθητής πάρει τους βαθμούς $x_1 = 16,5$, $x_2 = 18$, $x_3 = 17$ και $x_4 = 16,6$, τότε ο σταθμικός μέσος της επίδοσης του θα είναι:

$$\bar{x} = \frac{16,5 \times 7,5 + 18 \times 1 + 17 \times 1 + 16,6 \times 0,5}{7,5 + 1 + 1 + 0,5} = \frac{167}{10} = 16,7.$$

γ) Διάμεσος (δ)

Οι χρόνοι (σε λεπτά) που χρειάστηκαν 9 μαθητές, για να λύσουν ένα πρόβλημα είναι: 3, 5, 5, 36, 6, 7, 4, 7, 8 με μέση τιμή $\bar{x} = 9$. Παρατηρούμε όμως ότι οι οκτώ από τις εννέα παρατηρήσεις είναι μικρότερες του 9 και μία

(ακραία τιμή), η οποία επηρεάζει και τη μέση τιμή είναι, αρκετά μεγαλύτερη του 9. Αυτό σημαίνει ότι η μέση τιμή δεν ενδείκνυται ως μέτρο θέσης (“κέντρο”) των παρατηρήσεων αυτών. Αντίθετα, ένα άλλο μέτρο θέσης που δεν επηρεάζεται από ακραίες παρατηρήσεις είναι η διάμεσος (median), η οποία ορίζεται ως εξής:

Διάμεσος (δ) ενός δείγματος ν παρατηρήσεων οι οποίες έχουν διαταχθεί σε αύξουσα σειρά ορίζεται ως η μεσαία παρατήρηση, όταν το ν είναι περιττός αριθμός, ή ο μέσος όρος (ημιάθροισμα) των δύο μεσαίων παρατηρήσεων όταν το ν είναι άρτιος αριθμός.

Για παράδειγμα, για να βρούμε τη διάμεσο των δεδομένων:

α) 3, 4, 0, 6, 5, 8, 1, 1, 6, 1, 2, 8, 9

β) 3, 4, 0, 6, 5, 8, 1, 1, 6, 1, 2, 8, 9, 9

εργαζόμαστε ως εξής:

α) Έχουμε ν = 13 παρατηρήσεις, οι οποίες σε αύξουσα σειρά είναι:

0 1 1 1 2 3 4 5 6 6 8 8 9 .

Άρα, η διάμεσος είναι η μεσαία παρατήρηση (έβδομη στη σειρά), δ = 4.

β) Έχουμε ν = 14 παρατηρήσεις οι οποίες σε αύξουσα σειρά είναι:

0 1 1 1 2 3 4 5 6 6 8 8 9 9 .

Άρα, η διάμεσος είναι το ημιάθροισμα των δύο μεσαίων παρατηρήσεων (της έβδομης και όγδοης στη σειρά), δηλαδή $\delta = \frac{4+5}{2} = 4,5$.

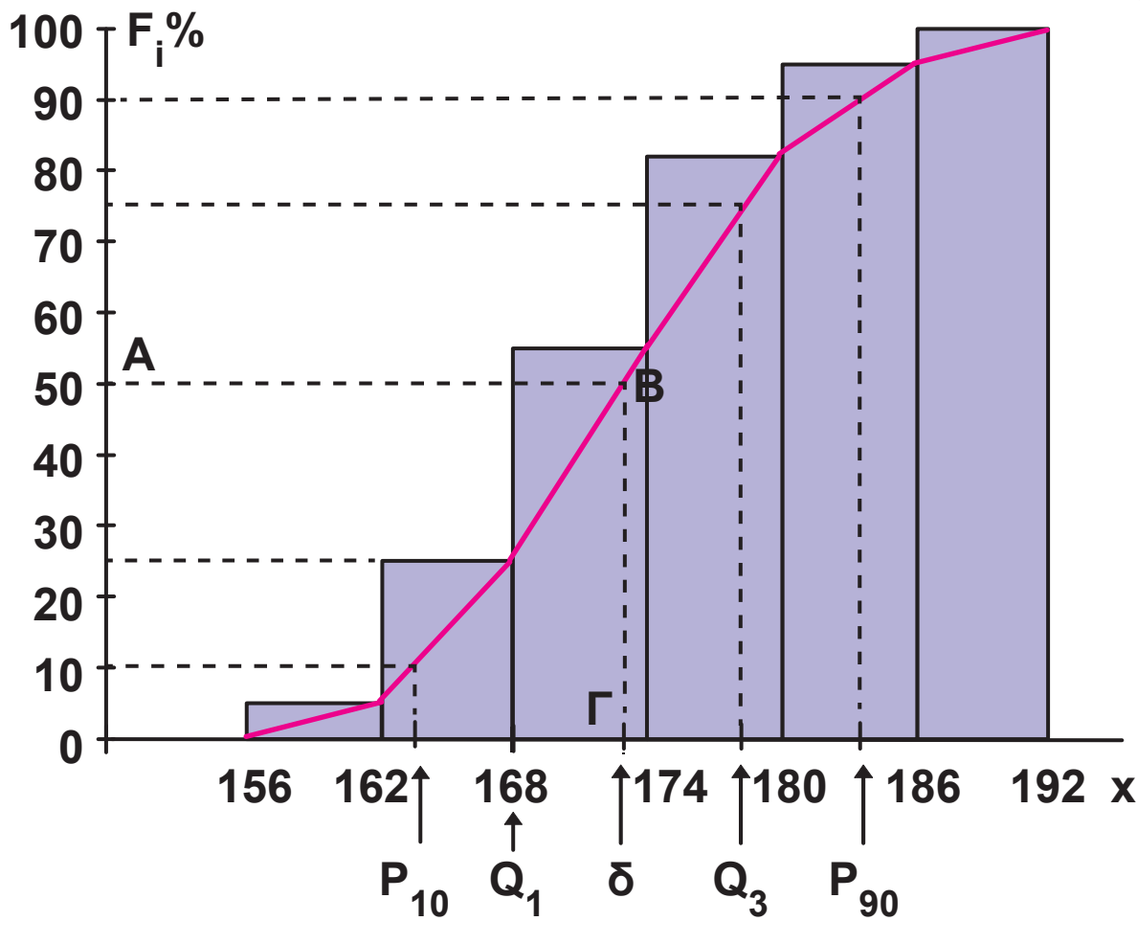
Παρατηρούμε ότι, η διάμεσος είναι η τιμή που χωρίζει ένα σύνολο παρατηρήσεων σε δύο ίσα μέρη όταν οι παρατηρήσεις αυτές τοποθετηθούν με σειρά τάξης μεγέθους. Ακριβέστερα, η διάμεσος είναι η τιμή για την οποία το πολύ 50% των παρατηρήσεων είναι μικρότερες από αυτήν και το πολύ 50% των παρατηρήσεων είναι μεγαλύτερες από την τιμή αυτήν.

Διάμεσος σε Ομαδοποιημένα Δεδομένα

Θεωρούμε τα δεδομένα του ύψους των μαθητών στον πίνακα 9 και το αντίστοιχο ιστόγραμμα αθροιστικών σχετικών συχνοτήτων με την πολυγωνική γραμμή, σχήμα 13. Η διάμεσος, όπως ορίστηκε, αντιστοιχεί στην τιμή x = δ της μεταβλητής X (στον οριζόντιο άξονα), έτσι ώστε το 50% των παρατηρήσεων να είναι μικρότερες ή ίσες του δ. Δηλαδή, η διάμεσος θα έχει αθροιστική σχετική συχνότητα F_i = 50%. Εφόσον στον κάθετο άξονα έχουμε τις αθροιστικές σχετικές συχνότητες, από το

σημείο Α (50% των παρατηρήσεων) φέρουμε την ΑΒ//0x και στη συνέχεια τη ΒΓ⊥0x. Τότε, στο σημείο Γ αντιστοιχεί η διάμεσος δ των παρατηρήσεων. Δηλαδή, δ ≈ 173.

13



δ) Εκατοστημόρια ($P_κ$)

Όπως ορίσαμε τη διάμεσο δ, έτσι ώστε το πολύ 50% των παρατηρήσεων να είναι μικρότερες του δ και το πολύ 50% των παρατηρήσεων να είναι μεγαλύτερες του

δ, μπορούμε ανάλογα να ορίσουμε και τα **εκατοστημόρια** (percentiles) $P_κ$, κ = 1, 2, ..., 99. Οι τιμές P_1, P_2, ..., P_{99} χωρίζουν τη συνολική συχνότητα σε 100 ίσα μέρη. Επομένως, αναλόγως και με τον ορισμό της διαμέσου, ορίζουμε ως κ-εκατοστιαίο σημείο ή $P_κ$ εκατοστημόριο ενός συνόλου παρατηρήσεων την τιμή εκείνη για την οποία το πολύ κ% των παρατηρήσεων είναι μικρότερες του $P_κ$ και το πολύ (100 − κ)% των παρατηρήσεων είναι μεγαλύτερες από την τιμή αυτήν.

Ειδική περίπτωση εκατοστημορίων είναι τα P_{25}, P_{50}, P_{75}, τα οποία καλούνται **τεταρτημόρια** (quartiles) και συμβολίζονται με Q_1, Q_2 και Q_3, αντίστοιχα.

Για το Q_1 έχουμε αριστερά το πολύ 25% των παρατηρήσεων και δεξιά το πολύ 75%. Όμοια για το Q_3 έχουμε αριστερά το πολύ 75% των παρατηρήσεων και δεξιά το πολύ 25% των παρατηρήσεων. Προφανώς το $Q_2 = P_{50}$ συμπίπτει και με τη διάμεσο, δηλαδή Q_2 = δ. Τα μέτρα αυτά χρησιμοποιούνται αρκετά συχνά για τη μελέτη ενός συνόλου δεδομένων.

Συχνά για ευκολία ο υπολογισμός των τεταρτημορίων Q_1 και Q_3 ενός συνόλου δεδομένων γίνεται **κατά προσέγγιση** υπολογίζοντας τις διαμέσους του πρώτου και του δεύτερου μισού των διατεταγμένων παρατηρήσεων, αντίστοιχα. Για παράδειγμα, προκειμένου να υπολογίσουμε τα τεταρτημόρια των δεδομένων 3, 4, 0, 6, 5,

8, 1, 1, 6, 1, 2, 8, 9, εργαζόμαστε ως εξής:

- **Διατάσσουμε τις παρατηρήσεις σε αύξουσα σειρά μεγέθους:**

Έχουμε ν = 13 παρατηρήσεις, οι οποίες σε αύξουσα σειρά είναι:

0 1 1 1 2 3 4 5 6 6 8 8 9 .

- **Υπολογίζουμε τη διάμεσο, όπως προαναφέραμε:**

Η διάμεσος είναι η έβδομη στη σειρά παρατήρηση, δηλαδή δ = 4.

- **Υπολογίζουμε τη διάμεσο του πρώτου μισού των διατεταγμένων παρατηρήσεων, δηλαδή των παρατηρήσεων που είναι αριστερά του δ. Η τιμή αυτή είναι το Q_1:**

Η διάμεσος των παρατηρήσεων που είναι αριστερά του δ, δηλαδή των

0 1 1 1 2 3, είναι το $Q_1 = \frac{1+1}{2} = 1$.

• Υπολογίζουμε τη διάμεσο του δεύτερου μισού των διατεταγμένων παρατηρήσεων, δηλαδή των παρατηρήσεων που είναι δεξιά του δ. Η τιμή αυτή είναι το Q_3.

Η διάμεσος των παρατηρήσεων που είναι δεξιά του δ, δηλαδή των 5 6 6 8 8 9, είναι το $Q_3 = \frac{6+8}{2} = 7$. (Όμως το ακριβές, σύμφωνα με τον ορισμό είναι $Q_3 = 6$).

Εκατοστημόρια σε Ομαδοποιημένα Δεδομένα

Ο υπολογισμός των εκατοστημορίων (ή τεταρτημορίων) σε ομαδοποιημένα δεδομένα γίνεται όπως και στη διάμεσο από το πολύγωνο αθροιστικών σχετικών συχνοτήτων. Στο σχήμα 13 δίνονται τα Q_1, $Q_2 = δ$, Q_3 και P_{10}, P_{90} για τα δεδομένα του πίνακα 9, από το οποίο βρίσκουμε κατά προσέγγιση:

$$P_{10} = 162{,}5, \quad Q_1 = 168, \quad δ = 173, \quad Q_3 = 178$$

$$\text{και} \quad P_{90} = 184.$$

ε) Επικρατούσα Τιμή (M_0)

Στην περίπτωση μη ομαδοποιημένων δεδομένων **επικρατούσα τιμή** ή **κορυφή** (mode) M_0 ορίζεται ως η παρατήρηση με τη μεγαλύτερη συχνότητα. Είναι προφανές ότι η επικρατούσα τιμή μπορεί να οριστεί και στην περίπτωση ποιοτικών δεδομένων, ενώ τα άλλα μέτρα που είδαμε ορίζονται μόνο για ποσοτικά δεδομένα. Για παράδειγμα:

α) Η επικρατούσα τιμή (επικρατούσα απασχόληση) για την απασχόληση των μαθητών του πίνακα 7 στον ελεύθερο χρόνο τους είναι M_0 = “Μουσική”.

β) Η επικρατούσα τιμή του αριθμού των αδελφών των μαθητών στον πίνακα 6 είναι $M_0 = 1$, δηλαδή οι περισσότερες οικογένειες (55%) έχουν δύο παιδιά.

γ) Για να βρούμε την επικρατούσα τιμή των παρατηρήσεων 0 1 1 2 2 2 3 4 4 4 5 5 7 8, κατασκευάζουμε τον παρακάτω πίνακα συχνοτήτων. Οι τιμές 2 και 4 είναι και οι δύο επικρατούσες τιμές, γιατί καθεμιά έχει συχνότητα

3. Βλέπουμε εδώ ότι η επικρατούσα τιμή μπορεί να μην είναι μοναδική. Όταν έχουμε δύο κορυφές, η αντίστοιχη κατανομή συχνοτήτων λέγεται δικόρυφη (bimodal), ενώ όταν έχουμε πολλές κορυφές λέγεται πολυκόρυφη (multimodal).

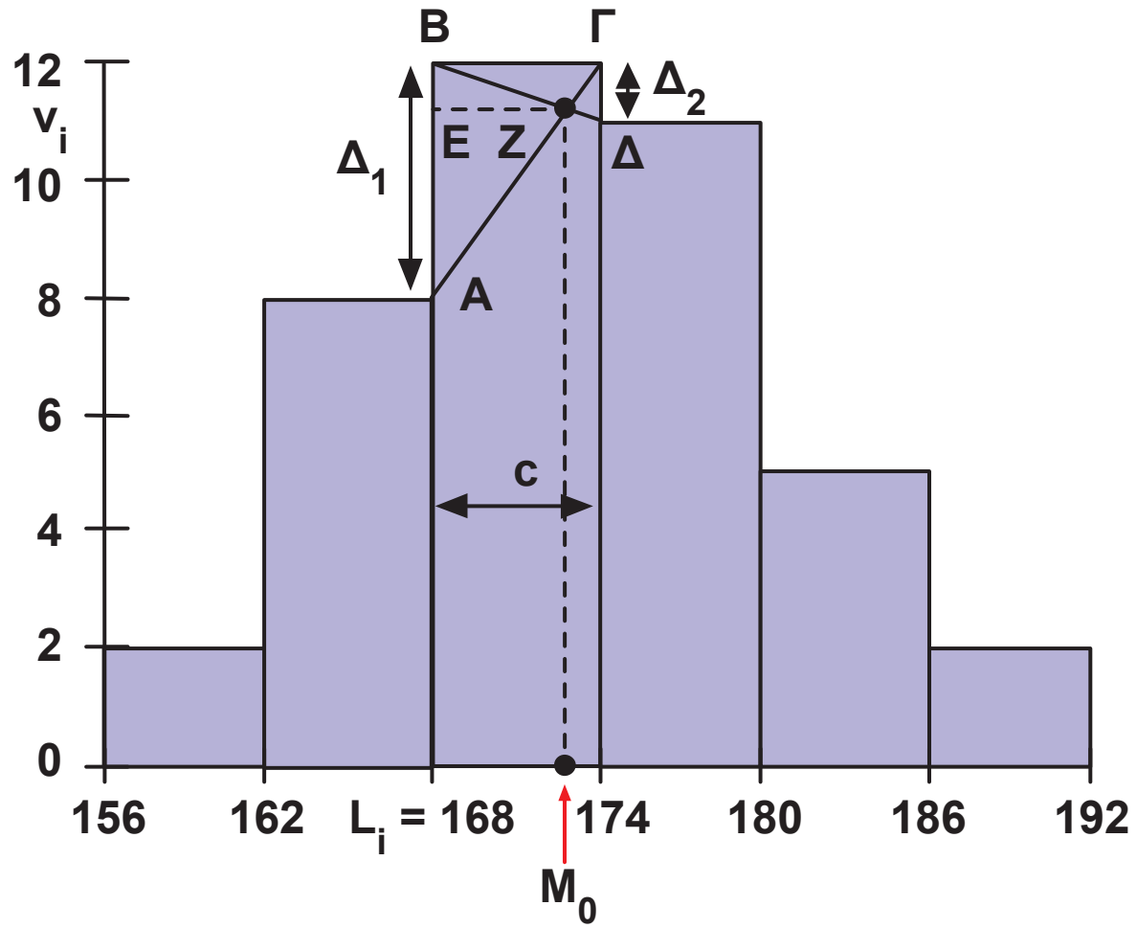
x_i	v_i
0	1
1	2
2	3
3	1
4	3
5	2
7	1
8	1

δ) Όταν όλες οι παρατηρήσεις είναι διαφορετικές, τότε λέμε ότι δεν υπάρχει επικρατούσα τιμή. Έτσι, για τις παρατηρήσεις 0, 1, 2, 7, 8, 9 δεν έχουμε επικρατούσα τιμή.

Επικρατούσα Τιμή σε Ομαδοποιημένα Δεδομένα

Όταν έχουμε ομαδοποιημένα (ποσοτικά) δεδομένα σε ισοπλατείς κλάσεις, τότε βρίσκουμε πρώτα την επικρατούσα κλάση i, δηλαδή την κλάση με τη μεγαλύτερη συχνότητα.
Υποθέτοντας, όπως έχουμε ήδη αναφέρει και προηγουμένως, ότι οι παρατηρήσεις στις κλάσεις κατανέμονται ομοιόμορφα, η επικρατούσα τιμή προσδιορίζεται, όπως φαίνεται στο παρακάτω σχήμα 14, ως η τετμημένη του σημείου τομής Z των ευθύγραμμων τμημάτων ΑΓ και ΒΔ. Στο σχήμα αυτό δίνεται η κορυφή για το ύψος των μαθητών του πίνακα 9. Κατά προσέγγιση η κορυφή (επικρατέστερο ύψος) είναι $M_0 \approx 173$ cm.

14



Μέτρα Διασποράς

Στα προηγούμενα είδαμε ότι τα μέτρα θέσης παρέχουν κάποια πληροφορία για την κατανομή ενός πληθυσμού. Αυτά όμως δεν επαρκούν, για να περιγράψουν πλήρως την κατανομή, όπως διαπιστώσαμε στην αρχή της § 2.3 συγκρίνοντας τις βαθμολογίες των μαθητών δύο τμημάτων Α και Β στα σχήματα 11(α), (β).

Ενώ οι βαθμολογίες των δύο τμημάτων Α και Β έχουν ίσες μέσες τιμές $\bar{x}_A = \bar{x}_B = 15$ και ίσες διαμέσους $\delta_A = \delta_B = 15$, είναι φανερό ότι οι κατανομές τους διαφέρουν σημαντικά ως προς τη μεταβλητότητά τους. Οι βαθμοί του τμήματος Α είναι περισσότερο "συγκεντρωμένοι" γύρω από τη μέση τιμή, ενώ, αντίθετα, οι βαθμοί του τμήματος Β διασπείρονται περισσότερο, έχουν δηλαδή μεγάλες αποκλίσεις γύρω από τη μέση τιμή τους. Παράλληλα λοιπόν με τα μέτρα θέσης κρίνεται απαραίτητη και η εξέταση κάποιων μέτρων διασποράς ή μεταβλητότητας, δηλαδή μέτρων που εκφράζουν τις αποκλίσεις των τιμών μιας μεταβλητής γύρω από τα μέτρα κεντρικής τάσης.

Τέτοια μέτρα λέγονται **μέτρα διασποράς** (measures of variation, dispersion measures). Τα σπουδαιότερα μέτρα διασποράς είναι το εύρος, η ενδοτεταρτημοριακή απόκλιση, η διακύμανση και η τυπική απόκλιση.

α) Εύρος (R)

Το απλούστερο από τα μέτρα διασποράς είναι το **εύρος** ή **κύμανση** (range) (R), που ορίζεται ως η διαφορά της ελάχιστης παρατήρησης από τη μέγιστη παρατήρηση, δηλαδή:

Εύρος R = Μεγαλύτερη παρατήρηση − Μικρότερη παρατήρηση

Έτσι, για τη βαθμολογία του τμήματος Α το εύρος είναι $R_A = 18 - 13 = 5$, ενώ για το τμήμα $R_B = 20 - 10 = 10$, τιμές που επιβεβαιώνουν ότι πράγματι στο τμήμα Β έχουμε μεγαλύτερη διασπορά βαθμολογίας παρά στο τμήμα Α.

Όταν έχουμε ομαδοποιημένα δεδομένα, το εύρος δίνεται από τη διαφορά του κατώτερου ορίου της πρώτης κλάσης από το ανώτερο όριο της τελευταίας κλάσης. Το εύρος των υψών των μαθητών του δείγματος στον πίνακα 9 είναι R = 192 − 156 = 36. Προφανώς, το εύρος σε ομαδοποιημένα δεδομένα μπορεί να διαφέρει ελαφρώς από τα αντίστοιχα δεδομένα πριν αυτά ομαδοποιηθούν. Για παράδειγμα, το εύρος των υψών στον πίνακα 8, πριν αυτά ομαδοποιηθούν, βρήκαμε ότι είναι R = 191 − 156 = 35.

Το εύρος είναι ένα αρκετά απλό μέτρο, που υπολογίζεται εύκολα δε θεωρείται όμως αξιόπιστο μέτρο διασποράς, γιατί βασίζεται μόνο στις δυο ακραίες παρατηρήσεις.

β) Ενδοτεταρτημοριακό Εύρος (Q)

Το ενδοτεταρτημοριακό εύρος (interquartile range) **είναι η διαφορά του πρώτου τεταρτημορίου Q_1 από το τρίτο τεταρτημόριο Q_3, δηλαδή:**

$$Q = Q_3 - Q_1$$

Στο μεταξύ τους διάστημα περιλαμβάνεται το 50% των παρατηρήσεων.

Επομένως όσο μικρότερο είναι αυτό το διάστημα, τόσο μεγαλύτερη θα είναι η συγκέντρωση των τιμών και άρα μικρότερη η διασπορά των τιμών της μεταβλητής.

Από τα δεδομένα του σχήματος 13 βρήκαμε κατά προσέγγιση $Q_1 = 168$, $Q_3 = 178$ επομένως το ενδοτεταρτημοριακό εύρος είναι $Q = 10$. Δηλαδή το 50% των μαθητών έχουν ύψος μεταξύ 168 και 178 cm.

γ) Διακύμανση (s^2)

Ένας άλλος τρόπος για να υπολογίσουμε τη διασπορά των παρατηρήσεων $t_1, t_2, ..., t_v$ μιας μεταβλητής X θα ήταν να αφαιρέσουμε τη μέση τιμή $\overline{x}$ από κάθε παρατήρηση και να βρούμε τον αριθμητικό μέσο των διαφορών αυτών, δηλαδή τον αριθμό:

$$\frac{(t_1-\bar{x})+(t_2-\bar{x})+\ldots+(t_v-\bar{x})}{v}=\frac{\sum_{i=1}^{v}(t_i-\bar{x})}{v}.$$

Ο αριθμός όμως αυτός είναι ίσος με μηδέν, αφού

$$\frac{(t_1-\bar{x})+(t_2-\bar{x})+\ldots+(t_v-\bar{x})}{v}=\frac{t_1+t_2+\ldots+t_v}{v}-\frac{v\bar{x}}{v}=\bar{x}-\bar{x}=0.$$

Γι' αυτό, ως ένα μέτρο διασποράς παίρνουμε τον μέσο όρο των τετραγώνων των αποκλίσεων των t_i από τη μέση τιμή τους $\bar{x}$. Το μέτρο αυτό καλείται διακύμανση ή διασπορά (variance) και ορίζεται από τη σχέση

$$s^2=\frac{1}{v}\sum_{i=1}^{v}(t_i-\bar{x})^2 \qquad (1)$$

Ο τύπος αυτός αποδεικνύεται ότι μπορεί να πάρει την ισοδύναμη μορφή:

$$s^2=\frac{1}{v}\left\{\sum_{i=1}^{v}t_i^2-\frac{\left(\sum_{i=1}^{v}t_i\right)^2}{v}\right\} \qquad (2)$$

η οποία διευκολύνει σημαντικά τους υπολογισμούς κυρίως όταν η μέση τιμή $\bar{x}$ δεν είναι ακέραιος αριθμός. Όταν έχουμε πίνακα συχνοτήτων ή ομαδοποιημένα δεδομένα, η διακύμανση ορίζεται από τη σχέση:

$$s^2 = \frac{1}{v}\sum_{i=1}^{\kappa}(x_i - \bar{x})^2 v_i \qquad (3)$$

ή την ισοδύναμη μορφή:

$$s^2 = \frac{1}{v}\left\{\sum_{i=1}^{\kappa} x_i^2 v_i - \frac{\left(\sum_{i=1}^{\kappa} x_i v_i\right)^2}{v}\right\}. \qquad (4)$$

όπου $x_1, x_2, \ldots, x_\kappa$ οι τιμές της μεταβλητής (ή τα κέντρα των κλάσεων) με αντίστοιχες συχνότητες $v_1, v_2, \ldots, v_\kappa$.

Για παράδειγμα, η διακύμανση της βαθμολογίας των μαθητών του τμήματος Α είναι σύμφωνα με την (1)

$$s_A^2 = \frac{(13-15)^2 + (13-15)^2 + (14-15)^2 + \ldots + (18-15)^2}{10} = \frac{20}{10} = 2,$$

ενώ για τους μαθητές του τμήματος Β βρίσκουμε

$s_B^2 = 6,6$, που επιβεβαιώνει τη διαπίστωσή μας ότι η βαθμολογία των μαθητών του τμήματος Β παρουσιάζει μεγαλύτερη μεταβλητότητα από τη βαθμολογία των μαθητών του τμήματος Α.

Ομοίως, η διακύμανση του ύψους των μαθητών για τα ομαδοποιημένα δεδομένα του πίνακα 9, υπολογίζεται σύμφωνα με τον τύπο (3), όπως φαίνεται στον επόμενο πίνακα:

Κλάσεις [-)	Κεντρικές τιμές x_i	Συχνότητα v_i	x_i^2	$x_i v_i$	$x_i^2 v_i$
156-162	159	2	25281	318	50562
162-168	165	8	27225	1320	217800
168-174	171	12	29241	2052	350892
174-180	177	11	31329	1942	344619
180-186	183	5	33489	915	167445
186-192	189	2	35721	378	71442
Σύνολο:		$v = 40$	—	$\Sigma x_i v_i =$ 6930	$\Sigma x_i^2 v_i =$ 1202760

Επομένως:

$$s^2 = \frac{1}{v}\left\{\sum_{i=1}^{\kappa} x_i^2 v_i - \frac{\left(\sum_{i=1}^{\kappa} x_i v_i\right)^2}{v}\right\} = \frac{1}{40}\left\{1202776 - \frac{6930^2}{40}\right\} = 53{,}4$$

Εάν υπολογίσουμε τη διακύμανση από τα μη ομαδοποιημένα δεδομένα του πίνακα 8, βρίσκουμε $s^2 = 50{,}9$. Η διαφορά αυτή οφείλεται στην απώλεια πληροφορίας λόγω ομαδοποίησης των παρατηρήσεων.

δ) Τυπική Απόκλιση (s)

Η διακύμανση είναι μια αξιόπιστη παράμετρος διασποράς, αλλά έχει ένα μειονέκτημα. Δεν εκφράζεται με τις μονάδες με τις οποίες εκφράζονται οι παρατηρήσεις. Για παράδειγμα, αν οι παρατηρήσεις εκφράζονται σε cm, η διακύμανση εκφράζεται σε cm^2. Αν όμως πάρουμε τη θετική τετραγωνική ρίζα της διακύμανσης, θα έχουμε ένα μέτρο διασποράς που θα εκφράζεται με την ίδια μονάδα μέτρησης του χαρακτηριστικού, όπως ακριβώς είναι και όλα τα άλλα μέτρα θέσης, που εξετάσαμε έως τώρα. Η ποσότητα αυτή λέγεται τυπική απόκλιση (standard deviation), συμβολίζεται με s και δίνεται από

τη σχέση:

$$s = \sqrt{s^2}$$

Η τυπική απόκλιση για το ύψος των μαθητών του πίνακα 4 είναι από το προηγούμενο παράδειγμα $s = \sqrt{53,4} = 7,3$ cm, αν αυτή υπολογιστεί από τα ομαδοποιημένα δεδομένα του πίνακα 9, ή $s = \sqrt{50,9} = 7,13$ cm, αν υπολογιστεί από τα μη ομαδοποιημένα δεδομένα του πίνακα 8.

Αξίζει να σημειωθεί ότι αν η καμπύλη συχνοτήτων για το χαρακτηριστικό που εξετάζουμε είναι κανονική ή περίπου κανονική, τότε η τυπική απόκλιση s έχει τις παρακάτω ιδιότητες:

i) το 68% περίπου των παρατηρήσεων βρίσκεται στο διάστημα

$$(\bar{x} - s, \bar{x} + s)$$

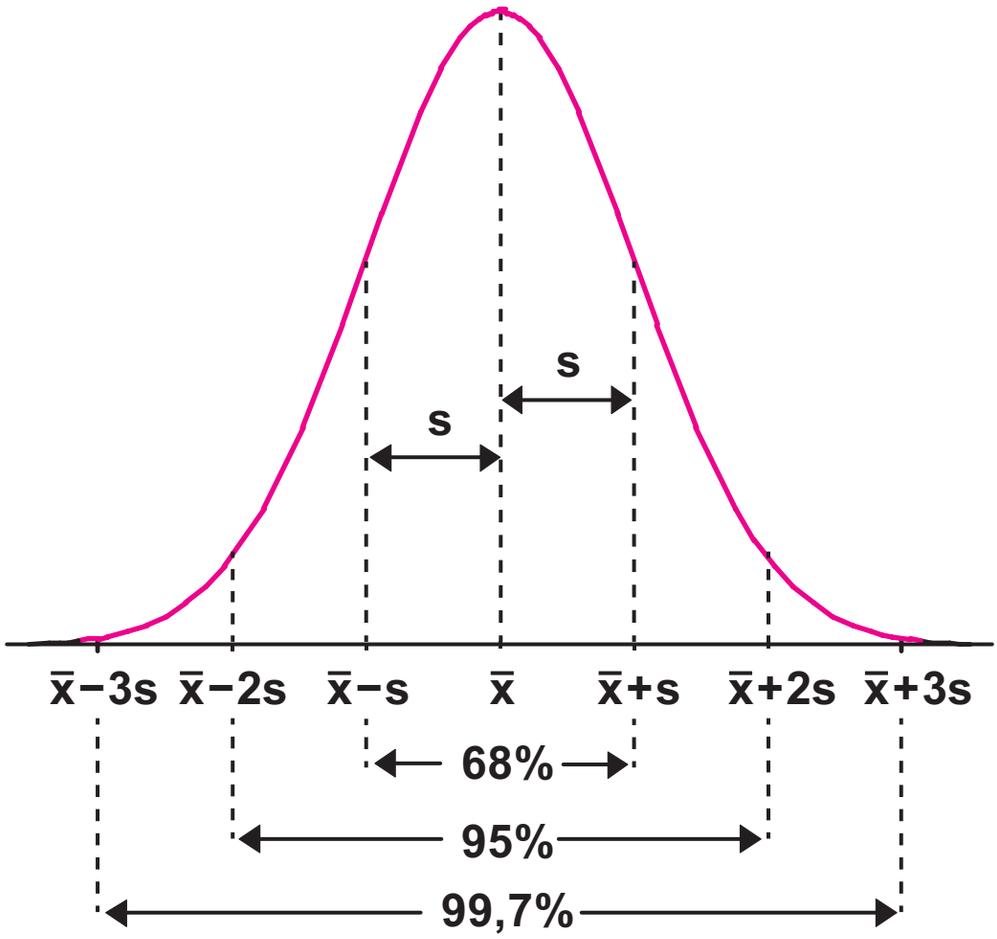
ii) το 95% περίπου των παρατηρήσεων βρίσκεται στο διάστημα

$$(\bar{x} - 2s, \bar{x} + 2s)$$

iii) το 99,7% περίπου των παρατηρήσεων βρίσκεται στο διάστημα

$$(\bar{x} - 3s, \bar{x} + 3s)$$

iv) το εύρος ισούται περίπου με έξι τυπικές αποκλίσεις, δηλαδή $R \approx 6s$.



Συντελεστής Μεταβολής (CV)

Έστω ότι από ένα δείγμα είκοσι μαθητών της Α΄ Γυμνασίου βρήκαμε μέσο βάρος $\bar{x}_A = 40$ kgr και τυπική απόκλιση $s_A = 6$ kgr, ενώ από ένα δεύτερο δείγμα τριάντα μαθητών της Γ΄ Λυκείου βρήκαμε μέσο βάρος $\bar{x}_B = 75$ kgr και τυπική απόκλιση $s_B = 6$ kgr. Όπως αντιλαμβανόμαστε, είναι λάθος να πούμε ότι το βάρος των μαθητών του Λυκείου έχει τον ίδιο βαθμό μεταβλητότητας με το

βάρος των μαθητών του Γυμνασίου, καθόσον η βαρύτητα που έχουν τα 6 kgr στο μέσο βάρος των 40 kgr είναι διαφορετική από αυτήν που έχουν στο μέσο βάρος των 75 kgr.

Ακόμη, ας υποθέσουμε ότι ο μέσος μισθός των υψηλόβαθμων υπαλλήλων μιας εταιρείας Α είναι $\overline{x}_A = 2.500$ € με τυπική απόκλιση $s_A = 420$ €, ενώ για τους υπαλλήλους μιας εταιρείας Β είναι $\overline{x}_B = 1.400$ $ με τυπική απόκλιση $s_B = 350$ $. Στην περίπτωση αυτή έχουμε διαφορετικές μονάδες μέτρησης του μισθού, επομένως οι διασπορές των παρατηρήσεων δεν είναι άμεσα συγκρίσιμες.

Ένα μέτρο με το οποίο μπορούμε να ξεπεράσουμε τις παραπάνω δυσκολίες και το οποίο μας βοηθά στη σύγκριση ομάδων τιμών, που είτε εκφράζονται σε διαφορετικές μονάδες μέτρησης είτε εκφράζονται στην ίδια μονάδα μέτρησης, αλλά έχουν σημαντικά διαφορετικές μέσες τιμές, είναι ο **συντελεστής μεταβολής** ή **συντελεστής μεταβλητότητας** (coefficient of variation), ο οποίος για $\overline{x} \neq 0$ ορίζεται από το λόγο:

$$CV = \frac{\text{τυπική απόκλιση}}{\text{μέση τιμή}} = \frac{s}{\overline{x}}$$

Αν $\overline{x} < 0$, τότε αντί της $\overline{x}$ χρησιμοποιούμε την $|\overline{x}|$.

Ο συντελεστής μεταβολής είναι ανεξάρτητος από τις μονάδες μέτρησης, εκφράζεται επί τοις εκατό και παριστάνει ένα μέτρο **σχετικής διασποράς** των τιμών και όχι της απόλυτης διασποράς, όπως έχουμε δει έως τώρα.
Για το πρώτο παράδειγμα του βάρους έχουμε συντελεστή μεταβολής για τις δύο ομάδες μαθητών:

$$CV_A = \frac{s_A}{\bar{x}_A} = \frac{6}{40} = 0{,}15 = 15\% \text{ και}$$

$$CV_B = \frac{s_B}{\bar{x}_B} = \frac{6}{75} = 0{,}08 = 8\%$$

δηλαδή, ο βαθμός διασποράς του βάρους των μαθητών Γυμνασίου είναι μεγαλύτερος από το βαθμό διασποράς του βάρους των μαθητών Λυκείου (για τα συγκεκριμένα δείγματα).
Ανάλογα συμπεράσματα βγάζουμε και για το δεύτερο παράδειγμα, όπου βρίσκουμε $CV_A = 16{,}8\%$ και $CV_B = 25\%$. Παρ' όλο που η τυπική απόκλιση των μισθών στην εταιρεία Α είναι μεγαλύτερη από την τυπική απόκλιση στην εταιρεία Β, ο συντελεστής μεταβολής δίνει μεγαλύτερη σχετική διασπορά στην εταιρεία Β. Αυτό μεταφράζεται στο να λέμε ότι έχουμε μεγαλύτερη **ομοιογένεια** μισθών στην εταιρεία Α παρά στη Β.
Γενικά δεχόμαστε ότι ένα δείγμα τιμών μιας μεταβλητής θα είναι ομοιογενές, εάν ο συντελεστής μεταβολής δεν ξεπερνά το 10%.

ΕΦΑΡΜΟΓΕΣ

1. **Ο παρακάτω πίνακας συχνοτήτων δίνει την κατανομή του χρόνου X (σε sec) 60 μαθητών που χρειάστηκαν, για να τρέξουν μια δεδομένη απόσταση. Να υπολογιστούν:**

α) **ο μέσος, ο διάμεσος και ο επικρατέστερος χρόνος για την κάλυψη της συγκεκριμένης απόστασης,**

β) **η τυπική απόκλιση,**

γ) **σε πόσο χρόνο από της στιγμή της εκκίνησης κάλυψε την απόσταση το 25% των μαθητών.**

x_i	v_i
50	4
55	6
60	8
65	12
70	14
75	10
80	6

ΛΥΣΗ

α) • Για τον υπολογισμό της μέσης τιμής συμπληρώνουμε τις τρεις πρώτες στήλες του παρακάτω πίνακα:

x_i	v_i	$x_i v_i$	$x_i^2 v_i$
50	4	200	10000
55	6	330	18150
60	8	480	28800
65	12	780	50700
70	14	980	68600
75	10	750	56250
80	6	480	38400
Σύνολο	$v = 60$	$\sum x_i v_i = 4000$	$\sum x_i^2 v_i = 270900$

Επομένως, ο μέσος χρόνος για την κάλυψη της συγκεκριμένης απόστασης είναι:

$$\bar{x} = \frac{\sum x_i v_i}{v} = \frac{4000}{60} \approx 66,67 \text{ sec.}$$

• Έχουμε $v = 60$ παρατηρήσεις σε αύξουσα σειρά, άρα η διάμεσος είναι ο μέσος όρος της 30ής και 31ης παρατήρησης, δηλαδή ο μέσος όρος των παρατηρήσεων 65

και 70, άρα $\delta = \frac{65+70}{2} = 67,5$ sec.

• Η επικρατούσα τιμή είναι η τιμή με τη μεγαλύτερη συχνότητα, άρα $M_0 = 70$ sec.

β) Για τον υπολογισμό της τυπικής απόκλισης είναι προτιμότερο να εφαρμόσουμε τη σχέση (4), μιας και η μέση τιμή δεν είναι ακέραιος αριθμός.
Με βάση τον παραπάνω πίνακα η διακύμανση της μεταβλητής X είναι:

$$s^2 = \frac{1}{v}\left\{\sum x_i^2 v_i - \frac{\left(\sum x_i v_i\right)^2}{v}\right\} = \frac{1}{60}\left\{270900 - \frac{4000^2}{60}\right\} =$$

$$= 70,56 \text{ sec}^2.$$

και η τυπική απόκλιση $s = \sqrt{70,56} = 8,4$ sec.
γ) Θέλουμε να υπολογίσουμε το πρώτο τεταρτημόριο, Q_1. Αριστερά της διαμέσου $\delta = 67,5$ έχουμε 30 παρατηρήσεις. Η διάμεσος αυτών των 30 πρώτων παρατηρήσεων είναι το ημιάθροισμα της 15ης και 16ης παρατήρησης, δηλαδή $Q_1 = \frac{(60+60)}{2} = 60$ sec. Δηλαδή, ύστερα από μία ώρα από τη στιγμή της εκκίνησης το 25% των μαθητών κάλυψαν τη συγκεκριμένη απόσταση.

2. Να αποδειχτεί ότι η συνάρτηση

$$f(\lambda) = \sum_{i=1}^{v}(x_i - \lambda)^2 = (x_1 - \lambda)^2 + (x_2 - \lambda)^2 + \ldots + (x_v - \lambda)^2$$

γίνεται ελάχιστη, όταν $\lambda = \bar{x}$.

ΛΥΣΗ

Λαμβάνοντας την πρώτη παράγωγο της f(λ), βρίσκουμε

$$f'(\lambda) = -2(x_1 - \lambda) - 2(x_2 - \lambda) - \ldots - 2(x_v - \lambda).$$

Έχουμε διαδοχικά:

$$f'(\lambda) = 0$$

$$x_1 - \lambda + x_2 - \lambda + \ldots + x_v - \lambda = 0$$

$$x_1 + x_2 + \ldots + x_v - v\lambda = 0$$

$$\lambda = \frac{x_1 + x_2 + \ldots + x_v}{v} = \bar{x}.$$

Η δεύτερη παράγωγος της f(λ) είναι:

$$f''(\lambda) = \underbrace{2 + 2 + \ldots + 2}_{v} = 2v$$

και επειδή $f''(\bar{x}) = 2v > 0$, συνεπάγεται ότι για $\lambda = \bar{x}$ η f(λ) γίνεται ελάχιστη.

3. Έστω $x_1, x_2, \ldots, x_v$ v παρατηρήσεις με μέση τιμή $\bar{x}$ και τυπική απόκλιση s_x.

α) Αν $y_1, y_2, \ldots, y_v$ είναι οι παρατηρήσεις που προκύπτουν αν προσθέσουμε σε καθεμιά από τις $x_1, x_2,$

..., x_v μια σταθερά c, να δειχτεί ότι:

i) $\bar{y} = \bar{x} + c$ ii) $s_y = s_x$

β) Αν $y_1, y_2, ..., y_v$ είναι οι παρατηρήσεις που προκύπτουν αν πολλαπλασιάσουμε τις $x_1, x_2, ..., x_v$ επί μια σταθερά c, να αποδειχτεί ότι:

i) $\bar{y} = c\bar{x}$, ii) $s_y = |c| s_x$

ΑΠΟΔΕΙΞΗ

α) Έχουμε $y_i = x_i + c$, $i = 1, 2, ..., v$ επομένως:

$$\text{i) } \bar{y} = \frac{y_1 + y_2 + ... + y_v}{v} = \frac{x_1 + c + x_2 + c + ... + x_v + c}{v} =$$

$$= \frac{x_1 + x_2 + ... + x_v}{v} + \frac{vc}{v} = \bar{x} + c$$

$$\text{ii) } s_y^2 = \frac{(y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + ... + (y_v - \bar{y})^2}{v} =$$

$$= \frac{(x_1 + c - \bar{x} - c)^2 + (x_2 + c - \bar{x} - c)^2 + ... + (x_v + c - \bar{x} - c)^2}{v} =$$

$$= \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + ... + (x_v - \bar{x})^2}{v} = s_x^2.$$

Άρα και $s_y = s_x$.

β) Έχουμε $y_i = cx_i$, $i = 1, 2, ..., v$, επομένως:

$$\text{i) } \bar{y} = \frac{y_1 + y_2 + ... + y_v}{v} = \frac{cx_1 + cx_2 + ... + cx_v}{v} =$$

$$= c\frac{x_1 + x_2 + ... + x_v}{v} = c\bar{x}$$

$$\text{ii) } s_y^2 = \frac{(y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + ... + (y_v - \bar{y})^2}{v} =$$

$$= \frac{(cx_1 - c\bar{x})^2 + (cx_2 - c\bar{x})^2 + ... + (cx_v - c\bar{x})^2}{v} =$$

$$= \frac{c^2[(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + ... + (x_v - \bar{x})^2]}{v} = c^2 s_x^2.$$

Άρα και $s_y = |c|s_x$.

ΑΣΚΗΣΕΙΣ

Α΄ ΟΜΑΔΑΣ

1. Έξι διαδοχικοί άρτιοι αριθμοί έχουν μέση τιμή 15. Να βρείτε τους αριθμούς και τη διάμεσό τους.

2. Έχουμε ένα δείγμα ν = 10 παρατηρήσεων, όπου κάθε παρατήρηση μπορεί να είναι 1, 2 ή 3. Είναι δυνατό η μέση τιμή να είναι α) 1 β) 4 γ) 1,8;

3. Ένας επενδυτής επένδυσε το ίδιο ποσό χρημάτων σε 8 διαφορετικές μετοχές στο χρηματιστήριο. Κατά τη διάρκεια του περασμένου έτους οι μετοχές είχαν τις παρακάτω εκατοστιαίες μεταβολές στην αξία τους:
5, 16, −10, 0, 27, 14, −20, 34. Να βρεθεί η μέση εκατοστιαία απόδοση της επένδυσης.

4. Το μέσο ύψος 9 καλαθοσφαιριστών μιας ομάδας είναι 205 cm.
α) Για να "ψηλώσει" την ομάδα ο προπονητής πήρε έναν ακόμη παίκτη με ύψος 216 cm. Ποιο είναι το μέσο ύψος της ομάδας τώρα;

β) Εάν ήθελε να “ψηλώσει” την ομάδα στα 208 cm, πόσο ύψος έπρεπε να έχει ο καλαθοσφαιριστής που πήρε;

5. Η μέση ηλικία 18 αγοριών και 12 κοριτσιών μιας τάξης είναι 15,4 χρόνια. Εάν η μέση ηλικία των αγοριών είναι 15,8 χρόνια, να βρείτε τη μέση ηλικία των κοριτσιών.

6. Σε μια κάλπη υπάρχουν άσπρες, μαύρες, κόκκινες και πράσινες μπάλες σε αναλογία 10%, 20%, 30% και 40% αντίστοιχα. Μια άσπρη μπάλα έχει βάρος 10 gr, μια μαύρη 11 gr, μια κόκκινη 12 gr και μια πράσινη 13 gr. Να βρείτε τη μέση τιμή, τη διάμεσο και την επικρατούσα τιμή του βάρους για όλες τις μπάλες, αν ξέρουμε ότι στην κάλπη υπάρχουν α) 10 μπάλες, β) 20 μπάλες, γ) δε γνωρίζουμε πόσες μπάλες υπάρχουν στην κάλπη.

7. Η επίδοση ενός μαθητή σε πέντε μαθήματα είναι 12, 10, 16, 18, 14.

α) Να βρείτε τη μέση επίδοση.

β) Αν τα μαθήματα είχαν συντελεστές στάθμισης 2, 3, 1, 1 και 3, ποια θα ήταν η μέση επίδοση; Σε ποια μαθήματα έπρεπε να δώσει ιδιαίτερη προσοχή ο μαθητής;

8. Μία εταιρεία απασχολεί 5 υπαλλήλους στο Τμήμα Α με μέσο (μηνιαίο) μισθό 1249 ευρώ, 6 υπαλλήλους στο Τμήμα Β με μέσο μισθό 1280 ευρώ, και 4 υπαλλήλους στο Τμήμα Γ με μέσο μισθό 1360 ευρώ. Ποιος είναι ο μέσος μισθός όλων των υπαλλήλων;

9. Η μέση τιμή και η διάμεσος πέντε αριθμών είναι 6. Οι τρεις από αυτούς είναι οι 5, 8, 9. Να βρείτε τους άλλους δύο.

10. Στον παρακάτω πίνακα δίνονται οι τιμές μιας μεταβλητής Χ με τις αντίστοιχες συχνότητές τους. Η πέμπτη συχνότητα χάθηκε! Μπορείτε να την "ανακαλύψετε", εάν γνωρίζετε ότι
α) η μέση τιμή είναι 4,4,
β) η διάμεσος είναι το 4,5,
γ) υπάρχουν δύο επικρατούσες τιμές;

x_i	v_i
2	1
3	3
4	1
5	2
6	;
7	1

11. Για την κατανομή του βαθμού των Μαθηματικών της Β΄ τάξης των 40 μαθητών και μαθητριών της Γ΄ Λυκείου του πίνακα 4 να βρείτε:

α) τη μέση τιμή,

β) τη διάμεσο,

γ) την επικρατούσα τιμή,

δ) το πρώτο και το τρίτο τεταρτημόριο και να ερμηνεύσετε τα αποτελέσματα.

12. Ο παρακάτω πίνακας δίνει τον αριθμό των επισκέψεων 40 μαθητών σε διάφορα μουσεία της χώρας κατά τη διάρκεια ενός έτους.

Επισκέψεις	Συχνότητα
[0-2)	8
2-4	12
4-6	10
6-8	6
8-10	4

Να υπολογιστούν:

α) η μέση τιμή,

β) η επικρατούσα τιμή,

γ) η διάμεσος,

δ) το πρώτο και τρίτο τεταρτημόριο.

13. Το μέσο ύψος των 30 μαθητών και μαθητριών μιας τάξης είναι 170 cm. Ποιο θα είναι το μέσο ύψος της τάξης:

α) αν φύγει ένας μαθητής με ύψος 180 cm,

β) αν έρθει μια νέα μαθήτρια με ύψος 170 cm,

γ) αν φύγει ένας μαθητής με ύψος 180 cm και έλθει μια μαθήτρια με ύψος 170 cm;

14. Καθεμία από τις παρακάτω λίστες δεδομένων έχουν μέση τιμή 50.

α) Σε ποια λίστα υπάρχει (i) μεγαλύτερη (ii) μικρότερη διασπορά παρατηρήσεων; (Να μη γίνουν πράξεις).

0, 20, 40, 50, 60, 80, 100

0, 48, 49, 50, 51, 52, 100

0, 1, 2, 50, 98, 99, 100.

β) Μπορεί να χρησιμοποιηθεί για σύγκριση των δεδομένων αυτών το εύρος;

15. Η βαθμολογία δέκα μαθητών σε ένα διαγώνισμα ήταν: 7, 11, 10, 13, 15, 3, 12, 11, 4, 14. Να υπολογίσετε:

α) τη μέση τιμή, την επικρατούσα τιμή και τη διάμεσο,

β) τα Q_1 και Q_3,

γ) το εύρος, την τυπική απόκλιση και το συντελεστή μεταβολής.

16. Να υπολογιστεί η τυπική απόκλιση των δεδομένων της άσκησης 12.

17. Ο μέσος χρόνος που χρειάζονται οι μαθητές ενός σχολείου να πάνε το πρωί από το σπίτι τους μέχρι το σχολείο είναι 10 λεπτά με τυπική απόκλιση 2 λεπτά. Υποθέτοντας ότι έχουμε περίπου κανονική κατανομή, να βρείτε κατά προσέγγιση το ποσοστό των μαθητών που χρειάζονται:

α) κάτω από 8 λεπτά

β) πάνω από 14 λεπτά

γ) το πολύ 10 λεπτά

δ) από 6 έως 12 λεπτά

για να πάνε στο σχολείο τους.

18. Να υπολογίσετε τη μέση τιμή και τη διάμεσο για τα παρακάτω δείγματα δεδομένων και να σχολιάσετε τα αποτελέσματα:

α) 1 2 6

β) 2 4 12

γ) 11 12 16

δ) 12 14 22.

19. Να υπολογίσετε τη διακύμανση και την τυπική απόκλιση για καθεμιά από τις παρακάτω λίστες δεδομένων. Συγκρίνοντας τα δεδομένα και τα αποτελέσματα τι συμπέρασμα βγάζετε;

α) 1, 3, 4, 5, 7

β) 3, 9, 12, 15, 21

γ) 6, 8, 9, 10, 12

δ) −1, −3, −4, −5, −7.

20. Οι μαθητές του Γ2 ξόδεψαν ετησίως κατά μέσο όρο 625 ευρώ αγοράζοντας διάφορα τρόφιμα από το κυλικείο του σχολείου. Εάν ο συντελεστής μεταβολής είναι 27,2%, να βρείτε την τυπική απόκλιση. Εάν επιπλέον γνωρίζετε ότι το $\sum x_i^2 = 11.746.700$, πόσοι είναι οι μαθητές του Γ2;

Β΄ ΟΜΑΔΑΣ

1. Η βαθμολογία 50 μαθητών στην Ιστορία κυμαίνεται από 10 μέχρι 20 (κανένας δεν είναι κάτω από τη βάση). Γνωρίζουμε επίσης ότι πέντε μαθητές έχουν βαθμό κάτω από 12, δεκαπέντε κάτω από 14, πέντε μεγαλύτερο ή ίσο του 18

και δεκαπέντε μεγαλύτερο ή ίσο του 16.

α) Να παρασταθούν τα δεδομένα σε έναν πίνακα συχνοτήτων.

β) Να υπολογίσετε: i) τη μέση τιμή, ii) τη διάμεσο.

γ) Εάν στο 5% των μαθητών με την καλύτερη επίδοση δοθεί έπαινος, πόσο βαθμό πρέπει να έχει κάποιος μαθητής για να πάρει έπαινο;

2. Η μέση τιμή και η διακύμανση των 5 τιμών ενός δείγματος είναι $\bar{x} = 4$ και $s^2 = 10$, αντίστοιχα. Εάν, για τις τέσσερις τιμές ισχύει $\sum_{i=1}^{4}(x_i - \bar{x})^2 = 14$, να βρεθεί η πέμπτη τιμή.

3. Ένας μαθητής αγόρασε 10 βιβλία που κόστιζαν χωρίς Φ.Π.Α. 15, 9, 6, 18, 21, 6, 18, 27, 9, 12 ευρώ αντίστοιχα.

α) Ποια είναι η μέση, η διάμεση και η επικρατούσα αξία (τιμή) των βιβλίων;

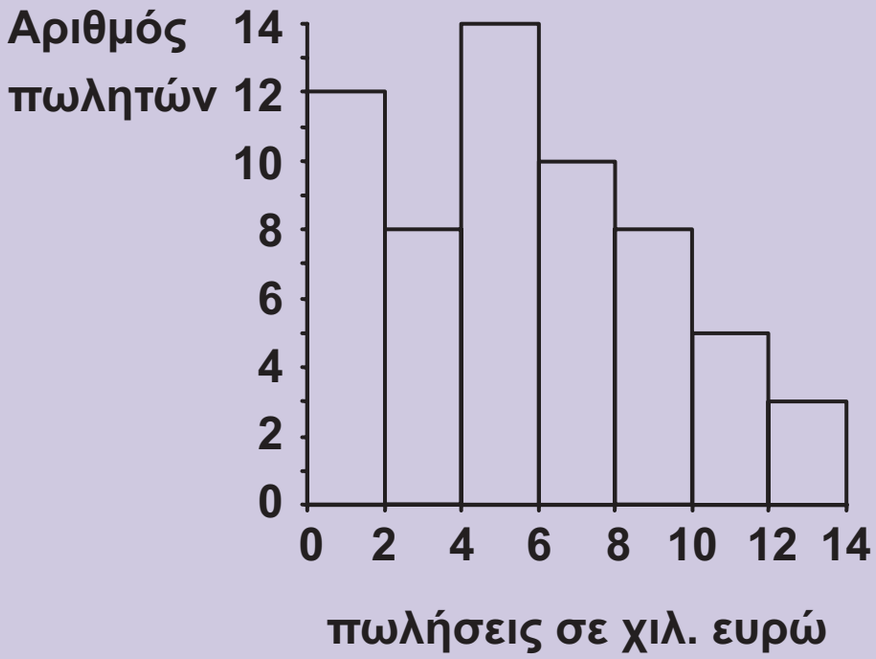
β) Πώς μεταβάλλονται οι απαντήσεις του ερωτήματος (α), αν προσθέσουμε και το Φ.Π.Α., που είναι 18%;

γ) Αν ο μαθητής πληρώσει επί πλέον 0,3 ευρώ

(χωρίς Φ.Π.Α.) για το ντύσιμο κάθε βιβλίου, πώς διαμορφώνονται τώρα οι απαντήσεις στο ερώτημα (β);

4. Να δείξετε ότι εάν από όλες τις τιμές 0, 2, 4, 6, 8, 10 και 12 ενός δείγματος αφαιρέσουμε τη μέση τιμή τους και διαιρέσουμε με την τυπική τους απόκλιση, τότε οι τιμές που θα προκύψουν θα έχουν μέση τιμή 0 και διασπορά 1.

5. Στο παρακάτω σχήμα φαίνονται τα ύψη των πωλήσεων σε χιλιάδες ευρώ που έγιναν από τους πωλητές μιας εταιρείας κατά τη διάρκεια ενός έτους.



α) Πόσοι είναι οι πωλητές;

β) Πόσοι πωλητές έκαναν πωλήσεις πάνω από 5 χιλιάδες ευρώ;

γ) Να βρεθεί η επικρατούσα τιμή των πωλήσεων.

δ) Να κατασκευάσετε τον πίνακα συχνοτήτων και να υπολογίσετε τη μέση τιμή και τη διακύμανση.

ε) Από το πολύγωνο αθροιστικών σχετικών συχνοτήτων να εκτιμήσετε τα τρία τεταρτημόρια Q_1, Q_2, Q_3.

6. Στον παρακάτω πίνακα δίνεται η κατανομή της ηλικίας των ατόμων μιας πόλης. Να υπολογίσετε:

α) Την τυπική απόκλιση και το συντελεστή μεταβολής.

β) Το ενδοτεταρτημοριακό εύρος.

Ηλικία (σε έτη)	Συχνότητα (σε χιλιάδες)
0-20	12
20-40	14
40-60	20
60-80	10
80-100	4

7. **Από τη Στατιστική της Φυσικής Κίνησης Πληθυσμού της Ελλάδας οι θάνατοι λόγω υπερτασικής νόσου το 1995 δίνονται στον παρακάτω πίνακα (ΕΣΥΕ):**

Να κατασκευάσετε στο ίδιο σχήμα τα πολύγωνα αθροιστικών σχετικών συχνοτήτων για την ηλικία ανδρών και γυναικών, αντίστοιχα, που πέθαναν από υπερτασική νόσο το 1995, και στη συνέχεια να τα συγκρίνετε.

Ηλικία	Θάνατοι	
	Άνδρες	Γυναίκες
50-54	10	7
55-59	10	4
60-64	17	21
65-69	36	57
70-74	44	61
75-79	73	109
80-84	117	162
85-89	123	195

2.4 ΓΡΑΜΜΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ

Εισαγωγή

Στα διάφορα προβλήματα που εξετάσαμε έως τώρα στη Στατιστική ασχοληθήκαμε κάθε φορά με μία μόνο μεταβλητή (ξεχωριστά), π.χ. με το βάρος, με το ύψος, με την επίδοση μαθητών κτλ. Για καθεμιά μεμονωμένη μεταβλητή αρκεστήκαμε στη μελέτη της κατανομής συχνοτήτων, στον υπολογισμό διάφορων μέτρων όπως μέση τιμή, διάμεσος, διακύμανση κτλ. Σε αρκετές όμως περιπτώσεις εξίσου ενδιαφέρουσα είναι και η ταυτόχρονη μελέτη δύο ή περισσότερων μεταβλητών, για να προσδιορίσουμε με ποιο τρόπο οι μεταβλητές αυτές σχετίζονται μεταξύ τους. Για παράδειγμα:

α) Η ηλικία και το βάρος ενός παιδιού έχουν κάποια θετική εξάρτηση (συσχέτιση) μεταξύ τους με την έννοια ότι όσο πιο μεγάλο είναι το παιδί τόσο μεγαλύτερο βάρος θα έχει.

β) Η διάρκεια ζωής των ζώντων οργανισμών σε μια περιοχή και το ποσοστό μόλυνσης της περιοχής έχουν αρνητική εξάρτηση μεταξύ τους, με την έννοια ότι όσο πιο μεγάλο είναι το ποσοστό μόλυνσης της περιοχής τόσο μικρότερη είναι η διάρκεια ζωής των

οργανισμών που ζουν στην περιοχή.

γ) Όσο μεγαλύτερη (μέχρι ένα ανώτερο όριο) είναι η περιεκτικότητα σε φθόριο στο πόσιμο νερό, τόσο μικρότερες είναι οι περιπτώσεις στη φθορά των δοντιών των μικρών παιδιών.

δ) Η συνολική παραγωγή καλαμποκιού εξαρτάται από τη θέση του χωραφιού, από την ποσότητα λιπάσματος, από την επίδραση της θερμοκρασίας, της υγρασίας κτλ.

ε) Το ύψος των αποδοχών των υπαλλήλων μιας εταιρείας δεν εξαρτάται από το βάρος τους.

Έτσι λοιπόν είναι ενδιαφέρον να εξεταστούν οι επιδράσεις που κάποιες μεταβλητές ασκούν σε κάποιες άλλες μεταβλητές. Η ύπαρξη μιας συναρτησιακής σχέσης (εξίσωσης) μεταξύ των μεταβλητών μπορεί να είναι εξαιρετικά πολύτιμη για την πρόβλεψη των τιμών μιας μεταβλητής από τις γνώσεις που διαθέτουμε για τις άλλες μεταβλητές, όταν ισχύουν κάποιες συγκεκριμένες συνθήκες.

Ο κλάδος της Στατιστικής που εξετάζει τη σχέση μεταξύ δύο ή περισσότερων μεταβλητών με απώτερο σκοπό την πρόβλεψη μιας απ' αυτές μέσω των άλλων χαρακτηρίζεται με την ονομασία **ανάλυση παλινδρόμησης** (regression analysis). Ιστορικά, ο όρος "regression" χρησιμοποιήθηκε για πρώτη φορά από τον Άγγλο ανθρωπολόγο Galton (1822-1911) το 1885. Με τη μελέτη

του ύψους των παιδιών σε σχέση με το ύψος των γονέων διαπιστώθηκε ότι παιδιά ψηλών γονέων τείνουν, κατά μέσο όρο, να είναι κοντύτερα των γονιών τους, ενώ παιδιά κοντών γονέων τείνουν, κατά μέσο όρο, να γίνονται ψηλότερα των γονιών τους.

Απλή Γραμμική Παλινδρόμηση

Η απλούστερη περίπτωση παλινδρόμησης είναι η απλή γραμμική παλινδρόμηση (simple linear regression), κατά την οποία υπάρχει μόνο μια **ανεξάρτητη μεταβλητή** X (independent or input variable), και η **εξαρτημένη μεταβλητή** Y (dependent or response variable), η οποία μπορεί να προσεγγιστεί ικανοποιητικά από μία γραμμική συνάρτηση του X. Η περίπτωση αυτή εμφανίζεται τόσο σε πειραματικές όσο και σε μη πειραματικές μελέτες. Στις πειραματικές μελέτες ο ερευνητής καθορίζει, για παράδειγμα, από πριν τις δόσεις ενός φαρμάκου (ανεξάρτητη μεταβλητή) που δίνει στα πειραματόζωα και μετρά τις αντιδράσεις τους (εξαρτημένη μεταβλητή). Με την παλινδρόμηση ενδιαφέρεται να προσδιορίσει μία σχέση δόσης-αντίδρασης για το συγκεκριμένο φάρμακο. Στις μη πειραματικές μελέτες ή δειγματοληψίες, γίνονται μετρήσεις σε δύο χαρακτηριστικά (μεταβλητές) για κάθε άτομο (μονάδα) του δείγματος. Σε ένα δείγμα 10 μαθητών μετράμε, για παράδειγμα, το βάρος και το ύψος

τους. Η διάκριση εδώ μεταξύ ανεξάρτητης και εξαρτημένης μεταβλητής είναι δύσκολη. Αν αυτό που μας ενδιαφέρει είναι το "τι συμβαίνει με το βάρος των παιδιών όταν αλλάζει το ύψος τους", τότε θεωρούμε ως ανεξάρτητη μεταβλητή Χ το ύψος και ως εξαρτημένη μεταβλητή Υ το βάρος. Οπότε, ενδιαφερόμαστε για την **παλινδρόμηση του βάρους (Υ) πάνω στο ύψος (Χ)**. Αντίθετα, αν μας ενδιαφέρει το "τι συμβαίνει με το ύψος των παιδιών όταν αλλάζει το βάρος τους", τότε θεωρούμε ως ανεξάρτητη μεταβλητή Χ το βάρος και ως εξαρτημένη μεταβλητή Υ το ύψος. Τότε έχουμε παλινδρόμηση του ύψους (Υ) πάνω στο βάρος (Χ).

Διάγραμμα Διασποράς

Ο παρακάτω πίνακας 10 δίνει τα ύψη Χ (σε cm) και τα βάρη Υ (σε kg) των 18 αγοριών της Γ΄ Λυκείου του πίνακα 4. Οι τιμές του ύψους δίνονται σε αύξουσα σειρά.

Πίνακας 10

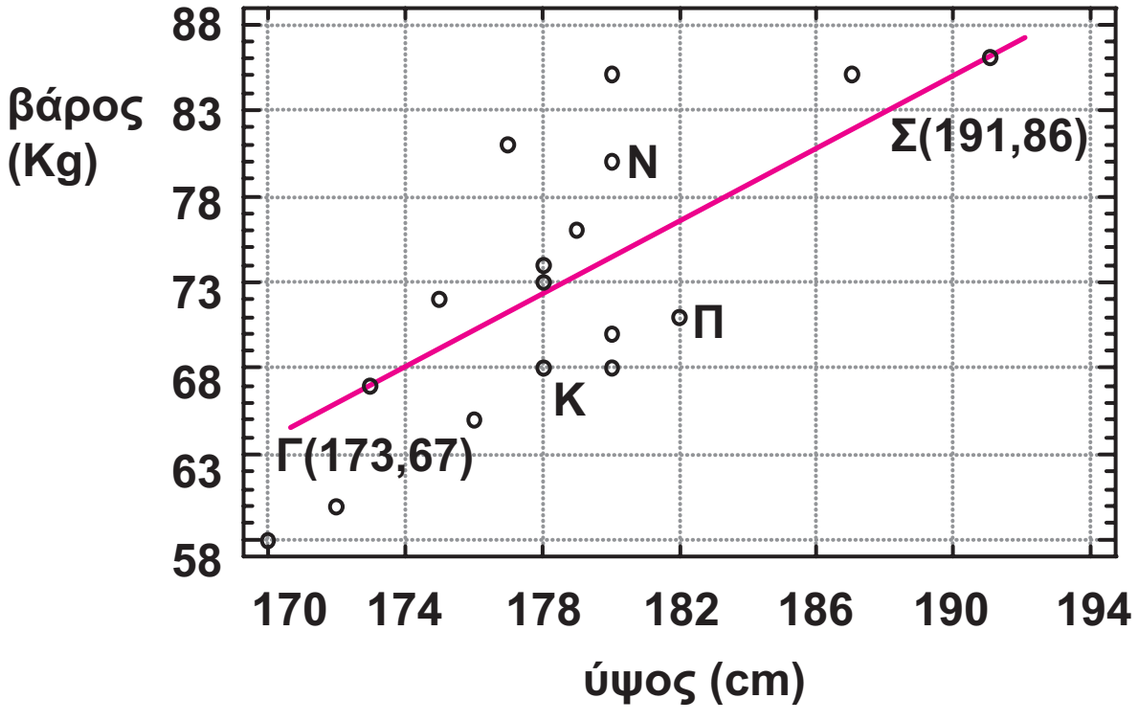
Λίστα υψών (σε cm) και βαρών (σε kg) των 18 αγοριών του πίνακα 4.

Μαθητής	Ύψος X	Βάρος Y	Μαθητής	Ύψος X	Βάρος Y
Α	170	58	Κ	178	68
Β	172	60	Λ	179	76
Γ	173	67	Μ	180	68
Δ	175	72	Ν	180	80
Ε	176	65	Ξ	180	70
Ζ	177	81	Ο	180	85
Η	178	73	Π	182	71
Θ	178	74	Ρ	187	85
Ι	178	73	Σ	191	86

Στο παράδειγμα αυτό έχουμε την περίπτωση όπου σε κάθε άτομο (μαθητή) γίνονται δύο μετρήσεις. Δηλαδή το δείγμα αποτελείται από τα ζεύγη τιμών των συνεχών μεταβλητών X (ύψος) και Y (βάρος).

Αν παραστήσουμε τα ζεύγη (x, y) των παρατηρήσεων σε ένα σύστημα ορθογώνιων αξόνων, παρατηρούμε ότι προκύπτει μία "διασπορά" των σημείων που αντιστοιχούν στους μαθητές που εξετάζουμε. Η παράσταση αυτή των σημείων καλείται διάγραμμα διασποράς (scatter diagram), βλέπε σχήμα 16.

16



Διάγραμμα διασποράς και ευθεία προσαρμοσμένη "με το μάτι" για τα δεδομένα του πίνακα 10.

Η προσεκτική παρατήρηση ενός διαγράμματος διασποράς μπορεί να μας δώσει σημαντικές πληροφορίες για τη σχέση εξάρτησης που ενδεχομένως υπάρχει μεταξύ των μεταβλητών τις οποίες εξετάζουμε. Η πείρα μας λέει ότι υπάρχει κάποια σχέση μεταξύ του ύψους και του βάρους κάθε μαθητή. Στο παράδειγμα αυτό το διάγραμμα διασποράς δείχνει, γενικά, ότι οι ψηλοί μαθητές είναι συνήθως και πιο βαρείς. Για παράδειγμα, ο Ν είναι ψηλότερος και βαρύτερος από τον Κ, ο Π είναι ψηλότερος

και βαρύτερος από τον Κ, αλλά υπάρχουν και εξαιρέσεις, όπως ο Π είναι ψηλότερος από τον Ν αλλά ο Ν είναι βαρύτερος από τον Π.

Ευθεία Παλινδρόμησης

Από το διάγραμμα διασποράς του προηγούμενου παραδείγματος φαίνεται καθαρά ότι υπάρχει μία σχέση ανάμεσα στο ύψος Χ και το βάρος Υ των 18 αγοριών της Γ΄ Λυκείου. Τα σημεία (x, y) είναι συγκεντρωμένα περίπου γύρω από μία ευθεία, δηλαδή η σχέση μεταξύ των Χ και Υ είναι κατά προσέγγιση γραμμική. Όπως έχουμε ήδη αναφέρει, μπορούμε να θεωρήσουμε τη μία μεταβλητή ως ανεξάρτητη μεταβλητή και την άλλη ως εξαρτημένη. Εδώ θεωρούμε ως ανεξάρτητη μεταβλητή το ύψος Χ και ως εξαρτημένη μεταβλητή το βάρος Υ, οπότε η ευθεία που θα προσαρμόζεται καλύτερα στα σημεία αυτά καλείται **ευθεία παλινδρόμησης της Υ πάνω στη Χ**.

Όπως γνωρίζουμε, η εξίσωση μιας ευθείας δίνεται από τη σχέση:

$$y = \alpha + \beta x \qquad (1)$$

όπου α και β είναι παράμετροι τις οποίες θέλουμε να υπολογίσουμε ή, όπως λέμε, να “εκτιμήσουμε”, έτσι

ώστε η ευθεία που θα προκύψει να μας δίνει όσο το δυνατόν την καλύτερη περιγραφή της σχέσης (εξάρτησης) που υπάρχει μεταξύ των μεταβλητών Χ και Υ. Η παράμετρος α μας δίνει το σημείο, (0, α) όπου η ευθεία αυτή τέμνει τον άξονα y΄y, ενώ η παράμετρος β παριστάνει το συντελεστή διεύθυνσης της ευθείας.
Ο πιο εύκολος τρόπος χάραξης της ευθείας είναι αυτός που γίνεται "με το μάτι". Μια τέτοια ευθεία έχουμε φέρει και στο διάγραμμα διασποράς του σχήματος 16. Για να βρούμε τα α και β, εργαζόμαστε ως εξής:

- **Επιλέγουμε δύο σημεία, έστω τα Γ(173,67) και Σ(191,86) πάνω στην ευθεία που φέραμε "με το μάτι".**

- **Αντικαθιστούμε τις συντεταγμένες (x, y) των σημείων αυτών στην (1), οπότε προκύπτει το σύστημα:**

$$\begin{cases} y_1 = \alpha + \beta x_1 \\ y_2 = \alpha + \beta x_2 \end{cases} \Leftrightarrow \begin{cases} 67 = \alpha + 173\beta \\ 86 = \alpha + 191\beta \end{cases}$$

• Επιλύοντας το σύστημα αυτό βρίσκουμε α = −115,6 και β = 1,06 οπότε η εξίσωση της ευθείας (1) γίνεται:

$$y = -115{,}6 + 1{,}06x. \qquad (2)$$

Επομένως, η ευθεία που κατά τη γνώμη μας προσαρμόζεται καλύτερα στα σημεία του διαγράμματος διασποράς διέρχεται από το σημείο (0, −115,6) και έχει συντελεστή διεύθυνσης 1,06.

Μέθοδος των Ελαχίστων Τετραγώνων

Είδαμε ότι η πιο απλή διαδικασία προσαρμογής μιας ευθείας γραμμής σε ένα διάγραμμα διασποράς είναι "με το μάτι". Αυτή όμως έχει πολλά μειονεκτήματα παρά την απλότητά της. Το κυριότερο είναι η έλλειψη αντικειμενικότητας, αφού διάφορα άτομα μπορούν να χαράξουν διαφορετικές μεταξύ τους ευθείες. Ακόμα και το ίδιο άτομο μπορεί να χαράζει διαφορετικές ευθείες κάθε φορά. Χρειαζόμαστε λοιπόν μια ακριβέστερη μέθοδο για την προσαρμογή μιας ευθείας γραμμής σε τέτοιου είδους δεδομένα.

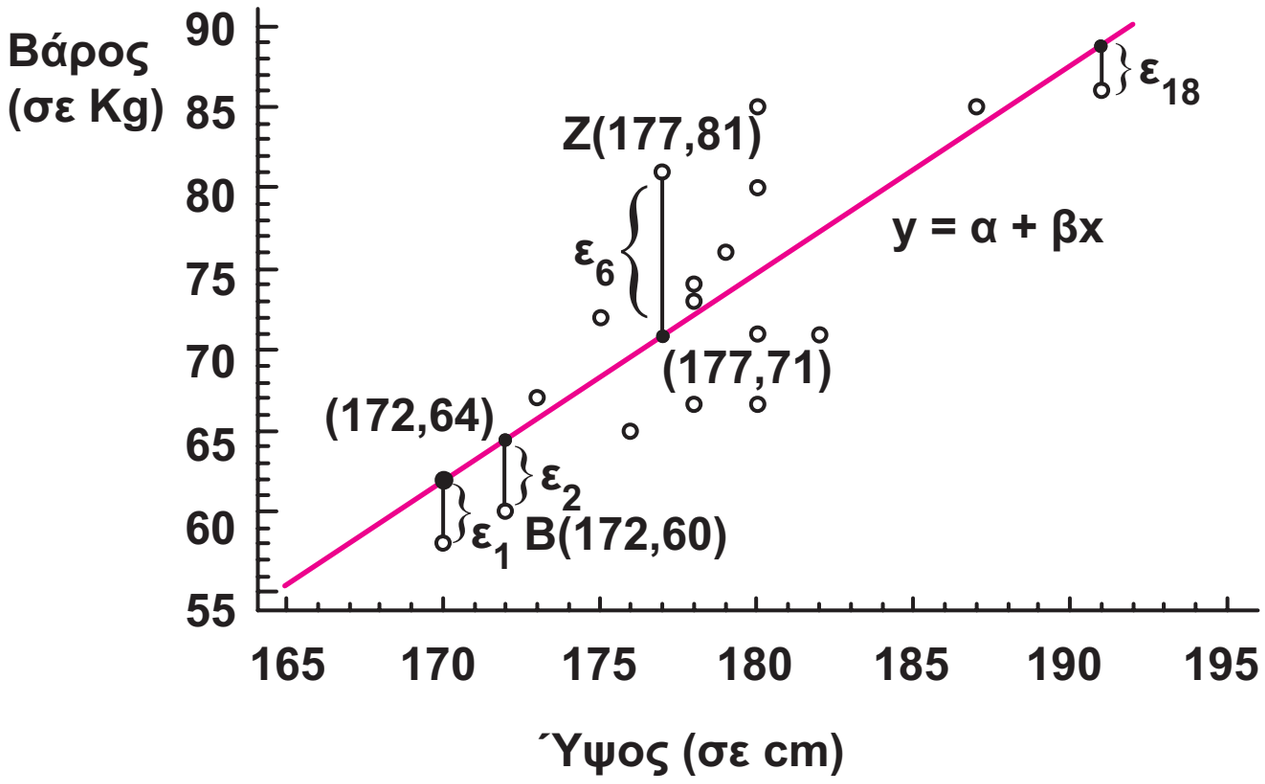
Μια μέθοδος που χρησιμοποιείται για την εκτίμηση των παραμέτρων α και β, άρα και για την εύρεση της εξίσωσης της καλύτερης ευθείας που προσαρμόζεται στα δεδομένα, είναι η "μέθοδος ελαχίστων τετραγώνων".

Η πρώτη αναφορά με ολοκληρωμένη ανάπτυξη της μεθόδου των ελαχίστων τετραγώνων εμφανίζεται το 1805 σε μια εργασία του Γάλλου μαθηματικού Legendre, (1752-1833) και αμέσως μετά από το Γερμανό μαθηματικό Gauss, (1777-1855) στην αστρονομική του πραγματεία "Theoria Motus" για τον προσδιορισμό της τροχιάς του μικρού πλανήτη Δήμητρα. Μάλιστα εδώ ο Gauss αναφέρει ότι χρησιμοποίησε την αρχή των ελαχίστων τετραγώνων πριν από το 1794 (σε ηλικία μόλις 17 ετών), έτσι ώστε να προηγείται του Legendre ως προς

την ανακάλυψη αυτής της μεθόδου.

Ας δούμε ξανά το διάγραμμα διασποράς στο σχήμα 17 του προηγούμενου παραδείγματος για τα ύψη X και τα βάρη Y των 18 μαθητών του πίνακα 10. Στο διάγραμμα αυτό έχουμε φέρει και μία ευθεία $y = \alpha + \beta x$, που πιστεύουμε ότι προσαρμόζεται καλύτερα στα σημεία (x_i, y_i) για τις $v = 18$ συνολικά μετρήσεις των μεταβλητών X και Y.

17



Προσαρμογή ευθείας ελαχίστων τετραγώνων στο διάγραμμα διασποράς του δεδομένων του πίνακα 10.

Έτσι, για παράδειγμα, για το μαθητή Β, σημείο Β(172, 60), με ύψος $x_2 = 172$ cm έχουμε βρει, όπως φαίνεται στον πίνακα 10, βάρος $y_2 = 60$ kg, ενώ, σύμφωνα με την ευθεία που φέραμε, το βάρος του αναμένεται να είναι (περίπου) 64 kg, έχουμε δηλαδή ένα σφάλμα $\varepsilon_2 = 60 - 64 = -4$, δηλαδή βάρος 4 kg λιγότερο από το αναμενόμενο. Ομοίως για το μαθητή Ζ, σημείο Ζ(177, 81), το βάρος του που μετρήθηκε ήταν $y_6 = 81$ kg, ενώ το αναμενόμενο βάρος του σύμφωνα με την ευθεία που φέραμε είναι 71 kg, έχουμε δηλαδή ένα σφάλμα $\varepsilon_6 = 81 - 71 = 10$, δηλαδή βάρος 10 kg περισσότερο από το αναμενόμενο.

Ανάλογα σφάλματα υπολογίζονται και για τους άλλους μαθητές. Θα θέλαμε λοιπόν να βρούμε με κάποια μέθοδο εκείνη την ευθεία $y = \alpha + \beta x$, έτσι ώστε τα σφάλματα που προκύπτουν να είναι όσο το δυνατόν μικρότερα.

Η μέθοδος των ελαχίστων τετραγώνων συνίσταται στον προσδιορισμό των παραμέτρων α, β, έτσι ώστε να ελαχιστοποιείται το άθροισμα των τετραγώνων των κατακόρυφων αποστάσεων των σημείων (x_i, y_i) από την ευθεία $y = \alpha + \beta x$, δηλαδή το

$$\sum_{i=1}^{v} \varepsilon_i^2 = \sum_{i=1}^{v} (y_i - \alpha - \beta x_i)^2 \qquad (4)$$

να γίνεται ελάχιστο.

Οι τιμές των παραμέτρων α και β, που ελαχιστοποιούν

την (4), καλούνται **εκτιμήτριες ελαχίστων τετραγώνων** (least square estimators), συμβολίζονται με $\hat{\alpha}$ ("α καπέλο") και $\hat{\beta}$ ("β καπέλο"), αντιστοίχως, και αποδεικνύεται (η απόδειξη εδώ παραλείπεται) ότι δίνονται από τις σχέσεις:

$$\hat{\beta} = \frac{v\sum_{i=1}^{v} x_i y_i - \left(\sum_{i=1}^{v} x_i\right)\left(\sum_{i=1}^{v} y_i\right)}{v\sum_{i=1}^{v} x_i^2 - \left(\sum_{i=1}^{v} x_i\right)^2} \qquad (5)$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

όπου $\bar{y} = \frac{1}{v}\sum_{i=1}^{v} y_i$, $\bar{x} = \frac{1}{v}\sum_{i=1}^{v} x_i$.

Η ευθεία

$$\hat{y} = \hat{\alpha} + \hat{\beta}x \qquad (6)$$

καλείται **ευθεία ελαχίστων τετραγώνων ή ευθεία παλινδρόμησης** της Υ (πάνω) στη Χ. Αντικαθιστώντας το $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$ στη σχέση (6) βρίσκουμε την

$$\hat{y} - \bar{y} = \hat{\beta}(x - \bar{x}),$$

η οποία φανερώνει ότι η ευθεία ελαχίστων τετραγώνων $\hat{y} = \hat{\alpha} + \hat{\beta}x$ διέρχεται από το σημείο με συντεταγμένες

($\overline{x}$, $\overline{y}$) και έχει συντελεστή διεύθυνσης το $\hat{\beta}$.
Αντικαθιστώντας τις τιμές x_i και y_i από τον πίνακα 10 στις σχέσεις (5) βρίσκουμε:

$$\hat{\beta} = 1{,}28 \text{ και } \hat{\alpha} = -156{,}1$$

οπότε η ευθεία ελαχίστων τετραγώνων που προσαρμόζεται καλύτερα στα δεδομένα είναι από τη σχέση (6), η

$$\hat{y} = -156{,}1 + 1{,}28x.$$

Παρατηρούμε ότι υπάρχει σημαντική διαφορά από την ευθεία y = −115, 6 + 1,06x που προσαρμόσαμε "με το μάτι" στο σχήμα 16.

Ερμηνεία των $\hat{\alpha}$ και $\hat{\beta}$

Στην εξίσωση ελαχίστων τετραγώνων $\hat{y} = \hat{\alpha} + \hat{\beta}x$ η τιμή της εκτιμήτριας $\hat{\alpha}$ της παραμέτρου α παριστάνει την τεταγμένη του σημείου στο οποίο η ευθεία τέμνει τον άξονα y΄y, δηλαδή την τιμή της εξαρτημένης μεταβλητής Y όταν x = 0. Όταν το $\hat{\alpha} = 0$ τότε η ευθεία διέρχεται από την αρχή των αξόνων.
Έστω τώρα δυο τιμές x_1 και $x_2 = x_1 + 1$ της ανεξάρτητης μεταβλητής. Τότε λαμβάνοντας τη διαφορά των αντίστοιχων προβλεπόμενων τιμών της εξαρτημένης μεταβλητής βρίσκουμε

$\hat{y}_2 - \hat{y}_1 = (\hat{\alpha} + \hat{\beta}x_2) - (\hat{\alpha} + \hat{\beta}x_1) = \hat{\alpha} + \hat{\beta}(x_1 + 1) - (\hat{\alpha} + \hat{\beta}x_1) = \hat{\beta}$

δηλαδή $\hat{y}_2 = \hat{y}_1 + \hat{\beta}$. Συνεπώς ο συντελεστής διεύθυνσης $\hat{\beta}$ της ευθείας $\hat{y} = \hat{\alpha} + \hat{\beta}x$ παριστά τη μεταβολή της εξαρτημένης μεταβλητής Υ όταν το Χ μεταβληθεί κατά μια μονάδα. Έτσι, όταν το x αυξηθεί κατά μια μονάδα τότε το $\hat{y}$ αυξάνεται κατά $\hat{\beta}$ μονάδες όταν $\hat{\beta} > 0$ ή ελαττώνεται κατά $\hat{\beta}$ μονάδες όταν $\hat{\beta} < 0$.

ΕΦΑΡΜΟΓΕΣ

1. Ένας ερευνητής, για να εξετάσει την επίδραση ενός αναισθητικού, εμβολίασε 10 ποντίκια με διαφορετική δόση κάθε φορά. Οι χρόνοι που μεσολάβησαν ώσπου τα ποντίκια να χάσουν τις αισθήσεις τους (λιποθυμήσουν) καταγράφονται στον παρακάτω πίνακα.

Δόση (σε mgr)	0,30	0,35	0,40	0,45	0,55	0,60	0,65	0,70	0,75	0,80
Χρόνος λιποθυμίας (sec)	12,5	11,5	11	8,5	7	6	5	4	2,5	2

α) Να γίνει το διάγραμμα διασποράς

β) Να επιλεγούν δύο πιθανά σημεία από τα οποία διέρχεται η προσαρμοσμένη ευθεία παλινδρόμησης και με βάση αυτά να βρεθεί η εξίσωση της ευθείας.

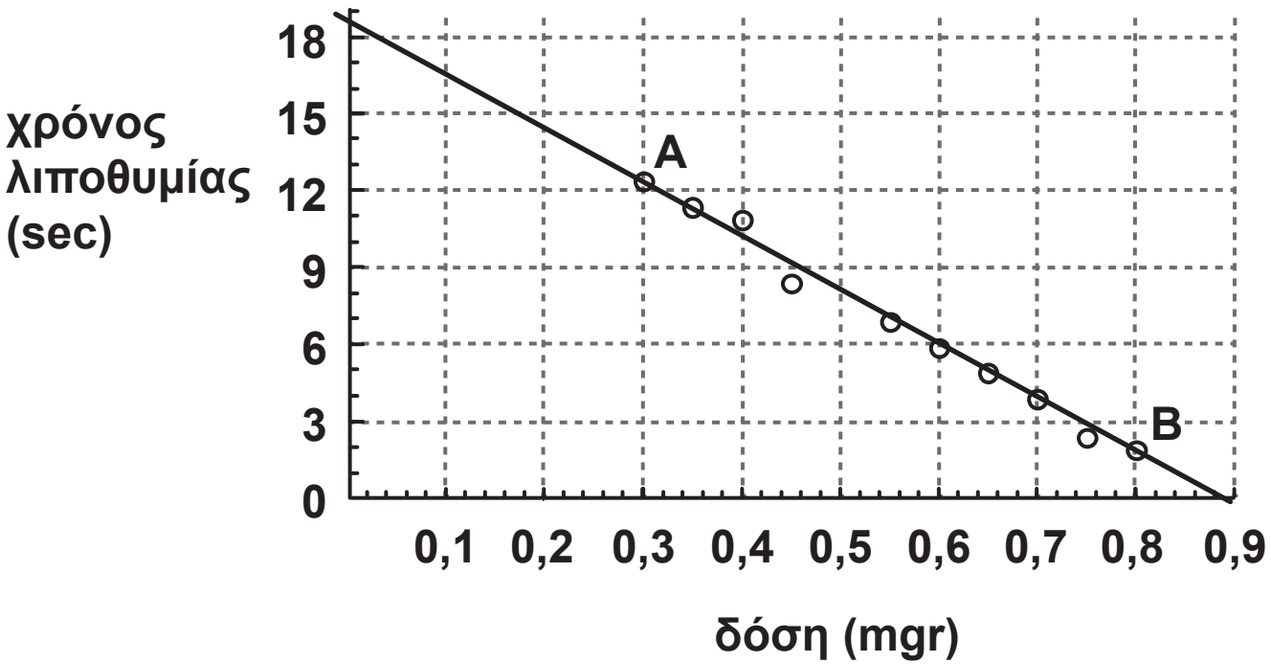
γ) Ύστερα από πόσο χρόνο αναμένεται να λιποθυμήσει ένα ποντίκι, εάν του γίνει ένεση (εμβολιαστεί) με 0, 0,5, 1 mgr αναισθητικού, αντίστοιχα; Να σχολιαστούν τα αποτελέσματα.

ΛΥΣΗ

α) Η περίπτωση που εξετάζεται εδώ αναφέρεται σε μια πειραματική κατάσταση, κατά την οποία ο ερευνητής καθορίζει εκ των προτέρων τη δόση του αναισθητικού που θα δώσει στα πειραματόζωα και μετρά τις αντιδράσεις τους. Ενδιαφέρεται δηλαδή να προσδιορίσει μια σχέση δόσης αναισθητικού και χρόνου λιποθυμίας. Έτσι, η δόση αναισθητικού παριστάνει την ανεξάρτητη

μεταβλητή (X) και ο χρόνος λιποθυμίας την εξαρτημένη μεταβλητή (Y).

Επομένως, το διάγραμμα διασποράς παριστάνεται παρακάτω:



β) Δύο σημεία από τα οποία εκτιμάται “με το μάτι” ότι θα διέρχεται η καλύτερη ευθεία παλινδρόμησης είναι τα A(0,3 , 12,3) και B(0,8 , 1,9). Αντικαθιστώντας στην εξίσωση της ευθείας $y = \alpha + \beta x$ έχουμε το σύστημα:

$$\begin{cases} 12{,}2 = \alpha + 0{,}3\beta \\ 1{,}9 = \alpha + 0{,}8\beta \end{cases}$$

από την επίλυση του οποίου βρίσκουμε $\alpha = 18{,}5$ και $\beta = -20{,}8$, οπότε η ευθεία παλινδρόμησης έχει εξίσωση $y = 18{,}5 - 20{,}8x$.

Παρατηρούμε, όπως άλλωστε αναμενόταν και από το διάγραμμα διασποράς, ότι ο συντελεστής διεύθυνσης είναι αρνητικός, δηλαδή έχουμε αρνητική εξάρτηση του χρόνου λιποθυμίας ως προς τη δόση αναισθητικού. Μεγαλύτερη δόση επιφέρει γρηγορότερη (σε μικρότερο χρόνο) λιποθυμία.

γ) Αντικαθιστώντας στην παραπάνω εξίσωση τις τιμές της δόσης x = 0, 0,5 και 1 mgr, βρίσκουμε τον προβλεπόμενο χρόνο λιποθυμίας του ποντικιού y = 18,5 , 8,1 και −2,3 sec, αντίστοιχα.

ΣΧΟΛΙΟ

Παρατηρούμε ότι εμφανίζονται εδώ δυο παράδοξα:

- **Μηδενική δόση του αναισθητικού προκαλεί λιποθυμία σε 18,5 sec.**

Υποθέτουμε όμως εδώ ότι τα ποντίκια χάνουν τις αισθήσεις τους μόνο από τη δόση του αναισθητικού και όχι από το φόβο τους!

- **Δόση αναισθητικού 1 mgr προκαλεί λιποθυμία σε −2,3 sec. Όμως και πάλι εδώ υποθέτουμε ότι η ένεση γίνεται σε ποντίκι που έχει τις αισθήσεις του και όχι να είναι πριν από 2,3 sec λιπόθυμο!**

Οι δύο αυτές παρατηρήσεις οδηγούν στο βασικό συμπέρασμα ότι οι προβλέψεις της εξαρτημένης μεταβλητής

έχουν νόημα, είναι δηλαδή δυνατές, μόνο για τις τιμές της ανεξάρτητης μεταβλητής οι οποίες βρίσκονται στο διάστημα που έχουμε εξετάσει ή τουλάχιστον πολύ κοντά στα άκρα του διαστήματος αυτού.

2. **Ο παρακάτω πίνακας δίνει τις πωλήσεις (ζήτηση) ενός προϊόντος (π.χ. των κερασιών) Υ (σε κιλά) από το ψαράδικο μιας περιοχής και τις αντίστοιχες τιμές Χ του προϊόντος σε ευρώ ανά κιλό για μια ορισμένη χρονική περίοδο**

Τιμή ψαριών ανά κιλό (σε ευρώ), Χ	15	13	11	9	9	6	5	4
Πωλήσεις σε κιλά, Υ	5	6	8	10	9	12	15	11

α) Να βρεθεί η ευθεία ελαχίστων τετραγώνων $\hat{y} = \hat{\alpha} + \hat{\beta}x$ (των πωλήσεων σε συνάρτηση με την τιμή) και να χαραχτεί στο αντίστοιχο διάγραμμα διασποράς.

β) Να ερμηνευθεί η έννοια των $\hat{\alpha}$ και $\hat{\beta}$

γ) Ποια είναι η αναμενόμενη ζήτηση (πωλήσεις), όταν η τιμή είναι 8 ευρώ/κιλό;

δ) Με βάση την ευθεία αυτή μπορούμε να προβλέψουμε την τιμή του προϊόντος, όταν η ζήτηση είναι 10 κιλά;

ΛΥΣΗ

α) Για τον προσδιορισμό της εξίσωσης της ευθείας ελάχιστων τετραγώνων διευκολύνει ο παρακάτω πίνακας με τις απαραίτητες πράξεις.
Επομένως, έχουμε:

x	y	x^2	xy
15	5	225	75
13	6	169	78
11	8	121	88
9	10	81	90
9	9	81	81
6	12	36	72
5	15	25	75
4	11	16	44
$\Sigma x = 72$	$\Sigma y = 76$	$\Sigma x^2 = 754$	$\Sigma xy = 603$

• $v = 8$

• $\bar{x} = \frac{\sum x}{v} = \frac{72}{8} = 9$

• $\bar{y} = \frac{\sum y}{v} = \frac{76}{8} = 9,5$

• $\hat{\beta} = \frac{v\sum xy - (\sum x)(\sum y)}{v\sum x^2 - \left(\sum x\right)^2} =$

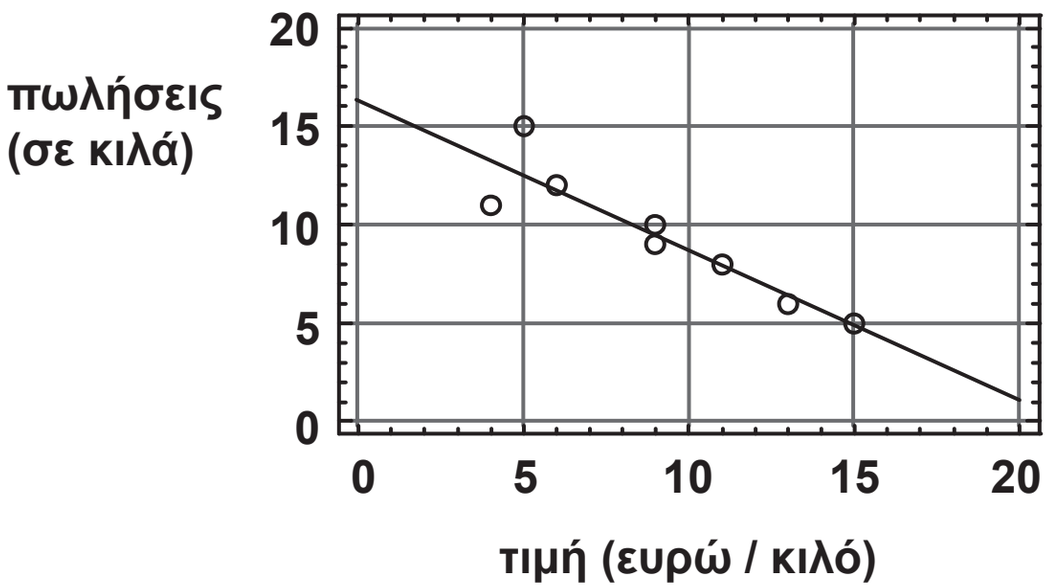
$= \frac{8(603) - (72)(76)}{8(754) - (72)^2} = \frac{-648}{848} = -0,76$

• $\hat{\alpha} = \bar{y} - \hat{\beta}\,\bar{x} = 9,5 + (0,76)(9) = 16,34.$

Άρα, η εξίσωση της ζήτησης (των ψαριών) σε συνάρτηση με την τιμή τους είναι $\hat{y} = 16,34 - 0,76x$.
Επειδή γνωρίζουμε ότι η ευθεία διέρχεται από τα σημεία $(0, \hat{\alpha})$ και $(\bar{x}, \bar{y})$, είναι εύκολο να χαραχτεί στο διάγραμμα διασποράς, όπως φαίνεται παρακάτω.
β) Το $\hat{\alpha} = 16,34$ προσδιορίζει την προβλεπόμενη ζήτηση του προϊόντος, όταν η τιμή του είναι 0 ευρώ/κιλό. Προφανώς εδώ τέτοια περίπτωση δεν είναι ρεαλιστική.
Το $\hat{\beta}$ προσδιορίζει τη μεταβολή που επέρχεται στην εξαρτημένη μεταβλητή Υ, όταν η Χ μεταβληθεί κατά μία μονάδα. Δηλαδή όταν η τιμή των ψαριών αυξηθεί κατά 1 ευρώ/κιλό (μία μονάδα), οι πωλήσεις θα ελαττωθούν κατά 0,76 κιλά.
γ) Όταν η αξία του προϊόντος είναι 8ευρώ/κιλό, σημαίνει ότι x = 8. Συνεπώς, για x = 8 η αναμενόμενη ζήτηση $\hat{y}$ είναι

$$\hat{y} = 16{,}34 - 0{,}76 \cdot 8 = 16{,}34 - 6{,}08 = 10{,}26 \text{ κιλά.}$$

δ) Προφανώς, στην περίπτωση που χρησιμοποιείται η μέθοδος των ελαχίστων τετραγώνων για την εκτίμηση της ευθείας παλινδρόμησης της εξαρτημένης μεταβλητής Υ πάνω στην ανεξάρτητη μεταβλητή Χ, δεν μπορεί να χρησιμοποιηθεί η προκύπτουσα ευθεία για πρόβλεψη της Χ, όταν δίνεται το Υ. Για να γίνει κάτι τέτοιο, πρέπει εξαρχής να εκτιμηθεί η ευθεία παλινδρόμησης της Χ πάνω στην Υ.



ΑΣΚΗΣΕΙΣ

Α΄ ΟΜΑΔΑΣ

1. Να κατασκευάσετε το διάγραμμα διασποράς και να βρείτε την εξίσωση της ευθείας που προσαρμόζεται (με το μάτι) καλύτερα στα δεδομένα.

α)

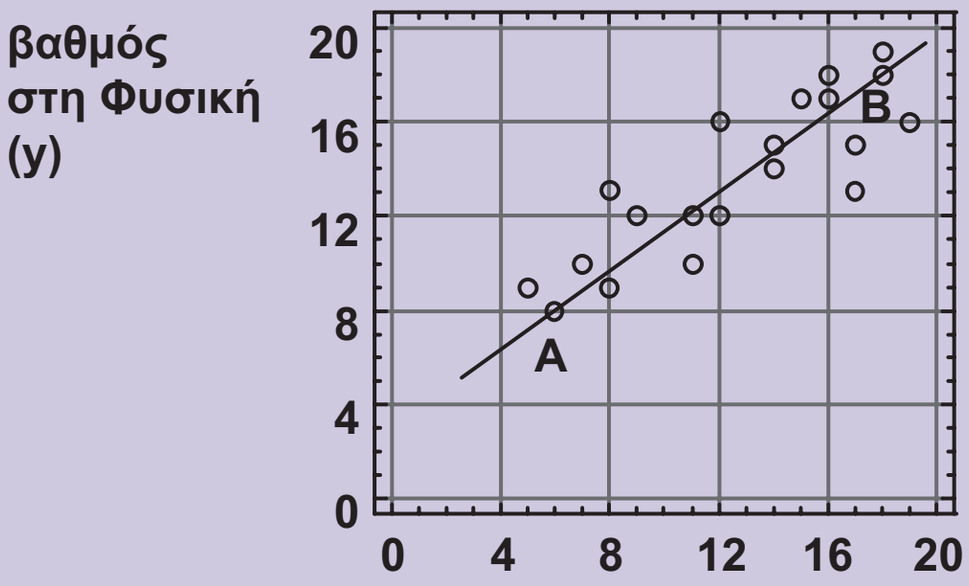
x	12	15	16	18	18
y	13	14	18	18	20

β)

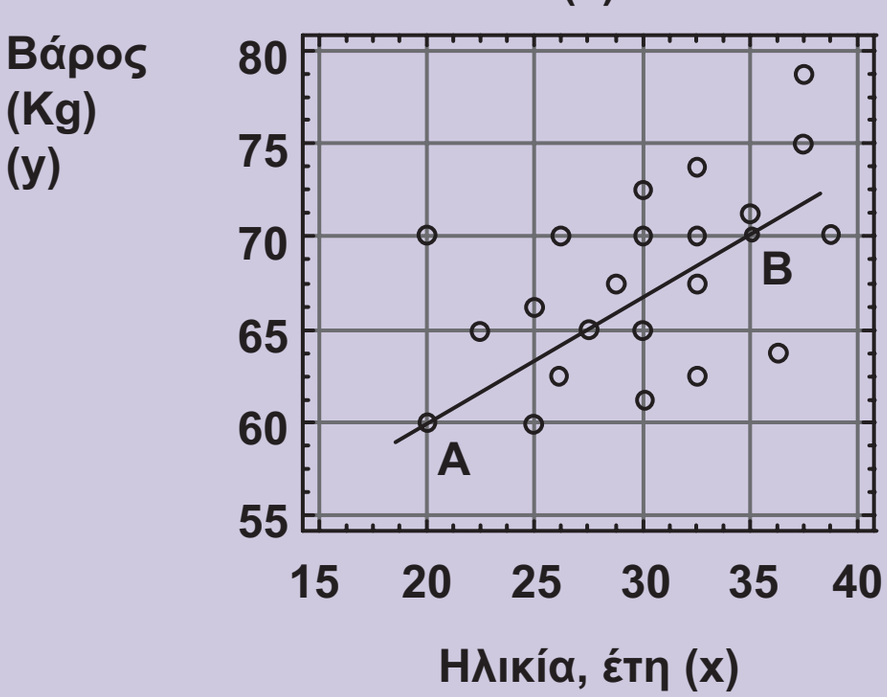
x	1	2	3	4	5	6	7	8	9
y	12	10	10	6	6	3	5	2	4

2. Παρακάτω δίνονται δύο διαγράμματα διασποράς με τις προσαρμοσμένες ευθείες, όπως τις χάραξε “με το μάτι” ένας μαθητής. Χρησιμοποιώντας τα σημεία Α και Β να βρείτε τις εξισώσεις $y = \alpha + \beta x$ των αντίστοιχων ευθειών.

(i)



(ii)



3. Για καθένα από τα παρακάτω ζεύγη τιμών να βρείτε την εξίσωση ελαχίστων τετραγώνων $\hat{y} = \hat{\alpha} + \hat{\beta}x$ και να τη χαράξετε στο αντίστοιχο διάγραμμα διασποράς. Στη συνέχεια να προβλέψετε την τιμή της y, όταν x = 6.

α)

x	1	2	3	4	5
y	1	2	3	4	5

β)

x	1	2	3	4	5
y	5	4	3	2	1

γ)

x	1	2	3	4	5
y	4	2	5	1	3

δ)

x	1	2	3	4	5
y	3	1	5	2	4

4. Να εφαρμόσετε τη μέθοδο ελαχίστων τετραγώνων για τα δεδομένα της εφαρμογής 1, για να εκτιμήσετε την ευθεία γραμμικής παλινδρόμησης του χρόνου λιποθυμίας (Y) των ποντικιών στη δόση αναισθητικού (X). Να συγκρίνετε τα αποτελέσματα που προκύπτουν με τη μέθοδο αυτή με τα αντίστοιχα αποτελέσματα που προέκυψαν με την προσαρμογή της ευθείας "με το μάτι".

5. α) Με τη μέθοδο ελαχίστων τετραγώνων να βρείτε την ευθεία γραμμικής παλινδρόμησης της Y πάνω στη X για τα παρακάτω δεδομένα 5 μαθητών.

Επίδοση στα Μαθηματικά, x	12	15	16	18	18
Επίδοση στη Φυσική, y	13	14	18	18	20

β) Αν υποθέσουμε ότι χάθηκε ο βαθμός της Φυσικής για το μαθητή που πήρε 15 στα Μαθηματικά και για να μην υποχρεωθεί ο μαθητής να ξαναδώσει εξετάσεις στη Φυσική, ποιο βαθμό, κατά τη γνώμη σας, πρέπει να πάρει;

Β΄ ΟΜΑΔΑΣ

1. Ο παρακάτω πίνακας δίνει τα αποτελέσματα των μετρήσεων της συστολικής πίεσης και της ηλικίας 10 γυναικών:

Ηλικία (έτη)	25	30	35	40	45	50	55	60	65	70
Πίεση (mm Hg)	116	117	121	147	111	133	105	153	155	176

α) Ποια από τις δύο μεταβλητές ηλικία και συστολική πίεση μπορεί να θεωρηθεί ως ανεξάρτητη μεταβλητή;

β) Να γίνει το αντίστοιχο διάγραμμα διασποράς.

γ) Να χαράξετε “με το μάτι” την ευθεία που προσαρμόζεται καλύτερα στα δεδομένα.

δ) Τι συστολική πίεση προβλέπετε για μια γυναίκα 75 ετών;

2. Να χρησιμοποιήσετε τα δεδομένα του Πίνακα 4 μόνο για τα αγόρια, για να προβλέψετε το ύψος ενός μαθητή όταν:

α) ο πατέρας του έχει ύψος 180 cm
β) ο μέσος όρος του ύψους των γονιών του είναι 170 cm.

3. Να χρησιμοποιήσετε τα δεδομένα του Πίνακα 4 μόνο για τα κορίτσια, για να προβλέψετε το ύψος μιας μαθήτριας όταν:
α) η μητέρα της έχει ύψος 167 cm
β) ο μέσος όρος του ύψους των γονιών της είναι 170 cm.

4. Από 8 γάμους που έγιναν σε μια εκκλησία κατά τη διάρκεια ενός μηνός, οι ηλικίες των ανδρογύνων ήσαν:

Ηλικία γαμπρού, y	20	22	24	25	28	30	33	38
Ηλικία νύφης , x	20	20	22	27	24	25	28	34

α) Να υπολογίσετε με τη μέθοδο ελαχίστων τετραγώνων την ευθεία γραμμικής παλινδρόμησης της Υ πάνω στη Χ και να τη χαράξετε στο αντίστοιχο διάγραμμα διασποράς.
β) Να βρείτε την αναμενόμενη ηλικία του γαμπρού για μια υποψήφια νύφη 25 ετών.
γ) Για κάθε έτος που μια γυναίκα καθυστερεί να παντρευτεί πόσο αυξάνεται η ηλικία του υποψήφιου γαμπρού.

5. Για τα ίδια δεδομένα της προηγούμενης άσκησης (4) να βρείτε:

α) την ευθεία γραμμικής παλινδρόμησης της Χ πάνω στη Υ,

β) την αναμενόμενη ηλικία της νύφης για έναν υποψήφιο γαμπρό 28 ετών,

γ) για κάθε έτος που ένας άνδρας καθυστερεί να παντρευτεί πόσο αυξάνεται η ηλικία της υποψήφιας νύφης;

2.5 ΓΡΑΜΜΙΚΗ ΣΥΣΧΕΤΙΣΗ

Εισαγωγή

Έχουμε δει μέχρι τώρα ότι ένα σύνολο παρατηρήσεων μιας μεταβλητής περιγράφεται με τα μέτρα θέσης και διασποράς, όπως για παράδειγμα, η μέση τιμή και η τυπική απόκλιση, αντιστοίχως. Επιπλέον, με τη γραμμική παλινδρόμηση που εξετάσαμε στην προηγούμενη παράγραφο είδαμε πώς βρίσκουμε την ευθεία γραμμικής παλινδρόμησης η οποία προσαρμόζεται καλύτερα στο "σμήνος" των σημείων όπως αυτά παριστάνονται σε ένα διάγραμμα διασποράς.

Ας δούμε, για παράδειγμα, τα παρακάτω ζεύγη τιμών (x_i, y_i) για τις μεταβλητές Χ, Υ και (x_i', y_i') για τις μεταβλητές Χ′, Υ′:

x_i	y_i
1	2,5
1	4,0
2	3,0
3	4,5
3	4,0
4	3,5
5	5,5
5	5,0

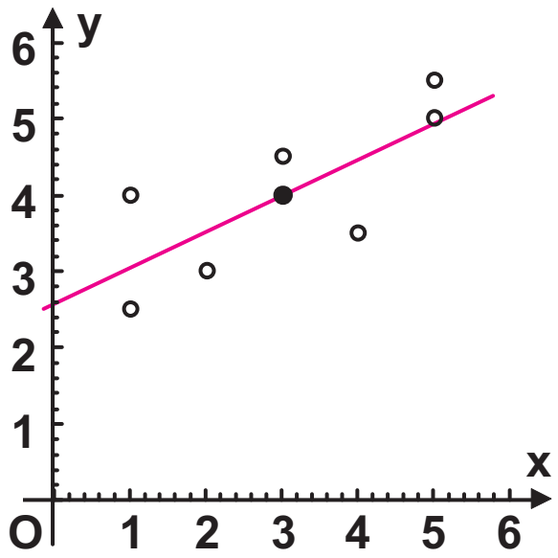
x_i'	y_i'
1,0	2,5
1,5	1,0
2,0	4,0
2,5	5,5
3,0	6,0
4,0	3,5
4,5	6,0
5,5	3,5

από τα οποία βρίσκουμε:

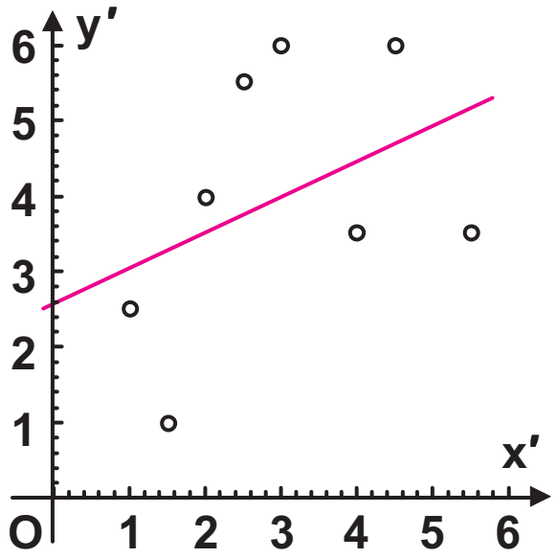
- $\overline{x} = 3$, $\overline{y} = 5$, $s_x = 1,5$, $s_y = 0,94$
- $\overline{x}' = 3$, $\overline{y}' = 5$, $s_{x'} = 1,46$, $s_{y'} = 1,66$

- το διάγραμμα διασποράς στο σχήμα 18(α) των σημείων (x_i, y_i) και την αντίστοιχη ευθεία ελαχίστων τετραγώνων $\hat{y} = 2,58 + 0,47x$

- το διάγραμμα διασποράς στο σχήμα 18(β) των σημείων (x_i', y_i') και την αντίστοιχη ευθεία ελαχίστων τετραγώνων $\hat{y}' = 2,58 + 0,47x'$.



(α)



(β)

Στα δύο αυτά διαγράμματα διασποράς βλέπουμε ότι προσαρμόζεται η ίδια ευθεία γραμμικής παλινδρόμησης. Όμως τα σημεία του σμήνους στο διάγραμμα (α) είναι περισσότερο συγκεντρωμένα γύρω από την ευθεία ενώ στο διάγραμμα (β) έχουμε ένα πιο χαλαρό σμήνος σημείων γύρω από την αντίστοιχη ευθεία παλινδρόμησης. Δηλαδή στην πρώτη περίπτωση η γραμμική σχέση μεταξύ των μεταβλητών είναι ισχυρότερη παρά στη δεύτερη περίπτωση. Ένα μέτρο που μας δίνει το μέγεθος της γραμμικής σχέσης ή το βαθμό συγκέντρωσης των σημείων του διαγράμματος διασποράς γύρω από την ευθεία παλινδρόμησης είναι ο λεγόμενος **συντελεστής γραμμικής συσχέτισης** (linear correlation coefficient).

Συντελεστής Γραμμικής Συσχέτισης

Ο συντελεστής γραμμικής συσχέτισης δύο μεταβλητών X και Y ορίζεται με βάση ένα δείγμα ν ζευγών παρατηρήσεων (x_i, y_i), i = 1, 2, ..., ν συμβολίζεται με r(X,Y) ή απλά με r και δίνεται από τον τύπο:

$$r = \frac{\sum_{i=1}^{\nu}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{\nu}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{\nu}(y_i - \bar{y})^2}} \qquad (1)$$

154 / 118 - 119

αναφέρεται δε και ως συντελεστής συσχέτισης του Pearson.

Από τον ορισμό του r παρατηρούμε ότι για μεγάλες τιμές x_i της Χ και y_i της Υ (μεγαλύτερες από τη μέση τιμή τους) οι διαφορές $(x_i - \overline{x})$ και $(y_i - \overline{y})$ είναι θετικές, οπότε το γινόμενό τους είναι θετικό. Όμοια για μικρές τιμές x_i και y_i, οι διαφορές $(x_i - \overline{x})$ και $(y_i - \overline{y})$ είναι αρνητικές, οπότε το γινόμενό τους είναι πάλι θετικό. Επομένως, όταν σε μεγάλες τιμές της μεταβλητής Χ αντιστοιχούν και μεγάλες τιμές της Υ, ή σε μικρές τιμές της Χ αντιστοιχούν μικρές τιμές της Υ τότε ο συντελεστής συσχέτισης είναι θετικός και λέμε ότι οι Χ, Υ είναι θετικά συσχετισμένες. Ανάλογα μπορούμε να δούμε ότι ο r παίρνει αρνητικές τιμές όταν σε μεγάλες τιμές της μιας μεταβλητής αντιστοιχούν μικρές τιμές της άλλης, οπότε λέμε ότι οι μεταβλητές αυτές είναι αρνητικά συσχετισμένες.

Με βάση τον παρακάτω πίνακα και τον τύπο (1) υπολογίζουμε το συντελεστή

x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
1	2,5	−2	−1,5	4	2,25	3
1	4,0	−2	0	4	0	0
2	3,0	−1	−1	1	1	1
3	4,5	0	0,5	0	0,25	0
3	4,0	0	0	0	0	0
4	3,5	1	−0,5	1	0,25	−0,5
5	5,5	2	1,5	4	0,25	3
5	5,0	2	1	4	1	2
24	32	0	0	18	7	8,5

γραμμικής συσχέτισης για τα δεδομένα του πρώτου παραδείγματος

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2}\sqrt{\sum(y_i - \bar{y})^2}} = \frac{8,5}{\sqrt{18}\sqrt{7}} \approx 0,76.$$

Με ανάλογο τρόπο υπολογίζουμε και το συντελεστή γραμμικής συσχέτισης για τα δεδομένα του δεύτερου παραδείγματος όπου βρίσκουμε $r' \approx 0,41$.

Συγκρίνοντας τους δύο συντελεστές συσχέτισης βλέπουμε ότι $r > r'$. Αυτό δηλώνει ότι οι μεταβλητές X, Y του

πρώτου παραδείγματος είναι περισσότερο γραμμικά συσχετισμένες παρά οι μεταβλητές X′, Y′ του δεύτερου παραδείγματος.

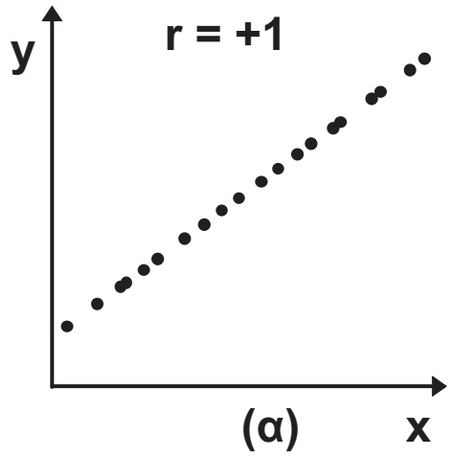
Ο συντελεστής συσχέτισης είναι καθαρός αριθμός, δηλαδή δεν εκφράζεται σε συγκεκριμένες μονάδες μέτρησης, επομένως είναι ανεξάρτητος των χρησιμοποιούμενων μονάδων μέτρησης των μεταβλητών X και Y. Επί πλέον ισχύει πάντοτε ότι:

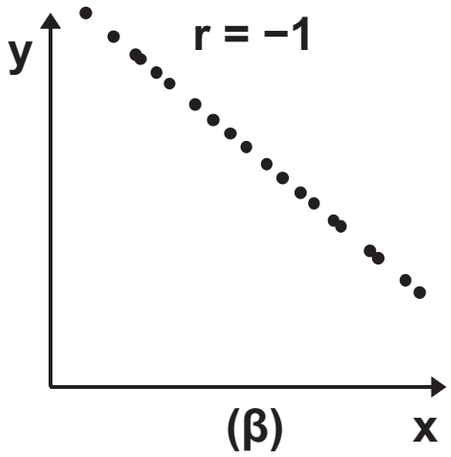
$$-1 \leq r \leq +1.$$

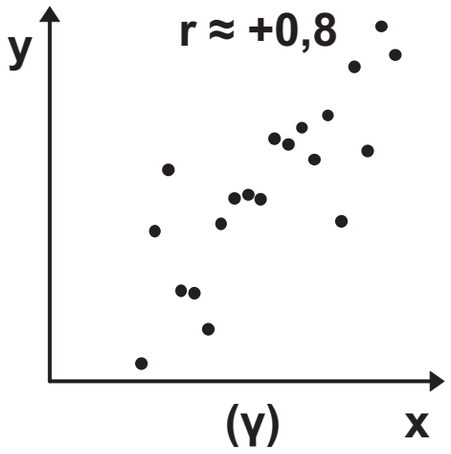
Πιο συγκεκριμένα όταν:

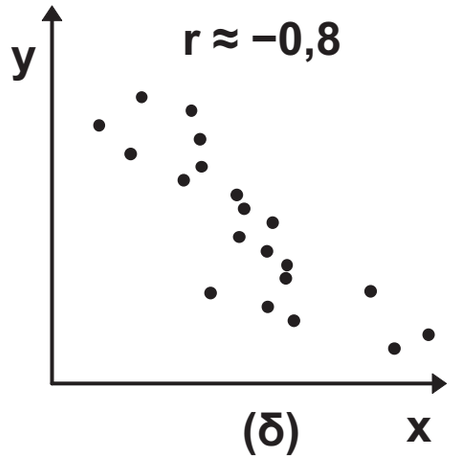
- $0 < r < +1$, τότε οι X, Y είναι **θετικά** γραμμικά συσχετισμένες (σχήμα 19(γ), (ε))
- $-1 < r < 0$, τότε οι X, Y είναι **αρνητικά** γραμμικά συσχετισμένες (σχήμα 19(δ), (στ))
- $r = +1$, τότε έχουμε **τέλεια θετική γραμμική συσχέτιση** και όλα τα σημεία βρίσκονται πάνω σε μια ευθεία με θετική κλίση (σχήμα 19(α)), δηλαδή $y = \alpha + \beta x$, $\beta > 0$
- $r = -1$, τότε έχουμε **τέλεια αρνητική γραμμική συσχέτιση** και όλα τα σημεία βρίσκονται πάνω σε μια ευθεία με αρνητική κλίση (σχήμα 19(β)), δηλαδή $y = \alpha + \beta x$, $\beta < 0$

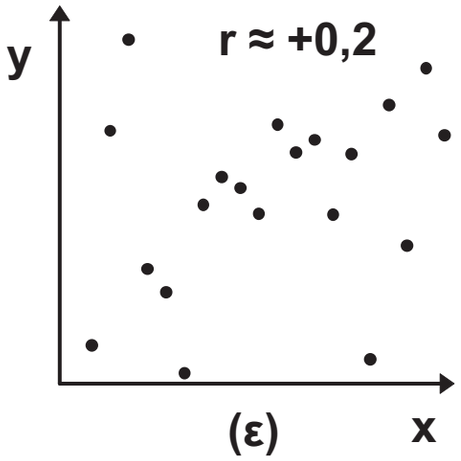
- **r = 0,** **τότε δεν υπάρχει γραμμική συσχέτιση μεταξύ των μεταβλητών. Οι μεταβλητές δηλαδή Χ, Υ είναι γραμμικά ασυσχέτιστες (σχήμα 19(ζ)).**

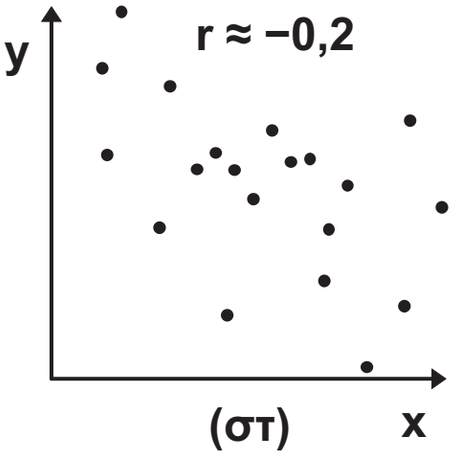


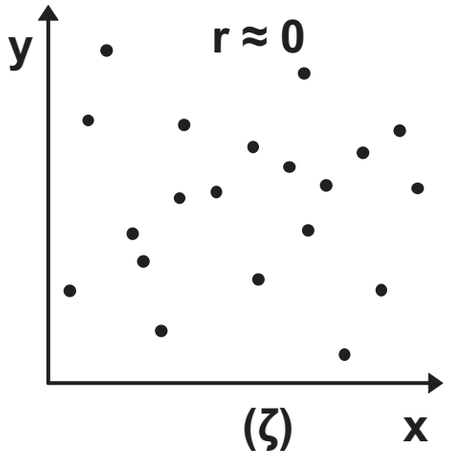












Διαγράμματα διασποράς και συντελεστές συσχέτισης για διάφορα ζεύγη παρατηρήσεων (x_i, y_i).

Αποδεικνύεται ότι ο συντελεστής γραμμικής συσχέτισης r δίνεται ισοδύναμα και από τον παρακάτω τύπο, η χρήση του οποίου διευκολύνει συχνά τους υπολογισμούς κυρίως στην περίπτωση που οι $\overline{x}$, $\overline{y}$ δεν είναι ακέραιοι:

$$r = \frac{v\sum_{i=1}^{v} x_i y_i - \left(\sum_{i=1}^{v} x_i\right)\left(\sum_{i=1}^{v} y_i\right)}{\sqrt{v\sum_{i=1}^{v} x_i^2 - \left(\sum_{i=1}^{v} x_i\right)^2}\sqrt{v\sum_{i=1}^{v} y_i^2 - \left(\sum_{i=1}^{v} y_i\right)^2}} \qquad (2)$$

Συσχέτιση και Παλινδρόμηση

Η παλινδρόμηση και η συσχέτιση, όπως τις εξετάσαμε έως τώρα, είναι δύο διαδικασίες μελέτης διμεταβλητών πληθυσμών. Η παλινδρόμηση προσδιορίζει τη σχέση

εξάρτησης μεταξύ δύο μεταβλητών, ενώ ο συντελεστής γραμμικής συσχέτισης δίνει ένα μέτρο του μεγέθους της γραμμικής συσχέτισης μεταξύ δύο μεταβλητών. Επομένως, οι δύο διαδικασίες δεν είναι άσχετες μεταξύ τους. Όταν δεν έχουμε πειραματικά δεδομένα, να προκαθορίζονται δηλαδή οι τιμές της μιας μεταβλητής, τότε μπορεί να μελετηθεί είτε η εξάρτηση της Y από τη X είτε η εξάρτηση της X από την Y. Το πόσο έντονη είναι η σχέση εξάρτησης μεταξύ των δύο μεταβλητών μας το δίνει ο συντελεστής συσχέτισης. Όσο το r πλησιάζει στο +1 τόσο τα σημεία του διαγράμματος διασποράς τείνουν να βρίσκονται σε μια ευθεία με συντελεστή διεύθυνσης $\hat{\beta} > 0$. Όσο το r πλησιάζει στο −1 τόσο τα σημεία τείνουν να βρίσκονται σε μια ευθεία με $\hat{\beta} < 0$. Όταν $r \approx 0$, τότε $\hat{\beta} \approx 0$. Συνήθως στις εφαρμογές εξετάζεται η συσχέτιση και η παλινδρόμηση μαζί, οπότε έχουμε πληρέστερη και πιο ολοκληρωμένη εξέταση των δύο μεταβλητών.

ΕΦΑΡΜΟΓΗ

Να υπολογιστεί και να ερμηνευτεί ο συντελεστής συσχέτισης r μεταξύ των μεταβλητών X και Y με βάση τις παρακάτω τιμές:

x	10	13	17	21	25	28	30
y	21	24	29	25	36	33	40

ΛΥΣΗ

Για τον υπολογισμό του συντελεστή συσχέτισης μεταξύ των Χ και Υ διευκολύνει ο παρακάτω πίνακας:

x	y	x^2	y^2	xy
10	21	100	441	210
13	24	169	576	312
17	29	289	841	493
21	25	441	625	525
25	36	625	1296	900
28	33	784	1089	924
30	40	900	1600	1200
$\Sigma x = 144$	$\Sigma y = 208$	$\Sigma x^2 = 3308$	$\Sigma y^2 = 6468$	$\Sigma xy = 4564$

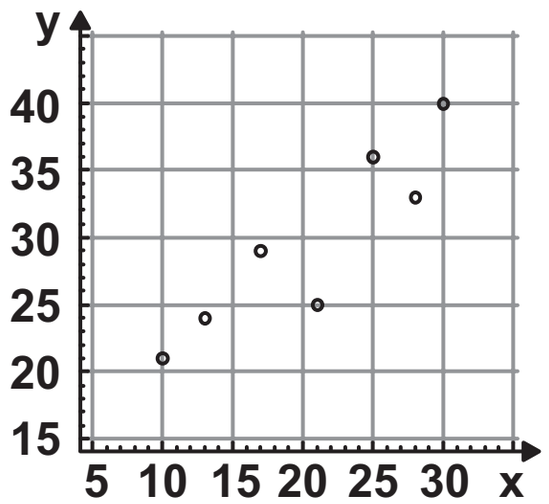
v = 7

Ο συντελεστής συσχέτισης υπολογίζεται από τη σχέση:

$$r = \frac{v\sum xy - \left(\sum x\right)\left(\sum y\right)}{\sqrt{v\sum x^2 - \left(\sum x\right)^2}\sqrt{v\sum y^2 - \left(\sum y\right)^2}} =$$

$$= \frac{7(4564) - (144)(208)}{\sqrt{7(3308) - (144)^2}\sqrt{7(6468) - (208)^2}} \approx 0,9.$$

Η υψηλή τιμή του r μας δείχνει ότι υπάρχει πολύ έντονη θετική γραμμική συσχέτιση μεταξύ των μεταβλητών Χ και Υ, όπως εξάλλου μπορούμε να το διαπιστώσουμε και από το αντίστοιχο διάγραμμα διασποράς.



ΑΣΚΗΣΕΙΣ

Α΄ ΟΜΑΔΑΣ

1. Να διατάξετε τις παρακάτω τιμές του r σε αύξουσα τάξη του βαθμού γραμμικής συσχέτισης δύο μεταβλητών: −0,6, 0,9, −0,7, 0,2, 0, −1.

2. Από τα διαγράμματα διασποράς των παρακάτω ζευγών (x_i, y_i) να εκτιμήσετε (χωρίς πράξεις) εάν η γραμμική συσχέτιση μεταξύ των μεταβλητών Χ και Υ είναι θετική ή αρνητική και επιπλέον αν είναι μικρή, μέτρια ή μεγάλη:

α)

x	1	3	5	7	9	10	12	13
y	−2	0	1	3	5	6	8	10

β)

x	1	3	5	7	9	10	12	13
y	4	5	1	6	4	3	8	10

γ)

x	1	3	5	7	9	10	12	13
y	10	8	3	4	6	3	2	5

δ)

x	1	3	5	7	9	10	12	13
y	10	8	3	4	1	6	5	12

ε)

x	1	3	5	7	9	10	12	13
y	6	8	6	3	5	7	8	9

3. Να υπολογίσετε τους συντελεστές γραμμικής συσχέτισης για τα ζεύγη τιμών (x_i, y_i) της προηγούμενης άσκησης και να τους συγκρίνετε με τις αντίστοιχες εκτιμήσεις σας.

4. Να βρείτε τους συντελεστές γραμμικής συσχέτισης για τα παρακάτω ζεύγη τιμών και να σχολιάσετε τα αποτελέσματα.

α)

x	1	2	3	4	5
y	4	2	0	−2	−4

β)

x	1	2	3	4	5
y	−4	−2	0	2	4

5. Για τέσσερα ζεύγη παρατηρήσεων (x_i, y_i) έχουμε:

$\bar{x} = 7$, $\bar{y} = 4{,}5$, $\sum x_i^2 = 210$, $\sum y_i^2 = 92$, $\sum x_i y_i = 138$.

Να υπολογίσετε το συντελεστή συσχέτισης.

6. Να συμπληρώσετε τον παρακάτω πίνακα και να υπολογίσετε το συντελεστή συσχέτισης ($x, y > 0$):

x	y	$(x - \bar{x})^2$	$(y - \bar{y})^2$	$(x - \bar{x})(y - \bar{y})$
1	3			
2	6	9	1	3
-	1			
6	-			
8	8			
9	13			

Β΄ ΟΜΑΔΑΣ

1. Η κατά άτομο κατανάλωση (σε γαλόνια) άπαχου (Y) και πλήρους (X) γάλακτος για τα έτη 1982-87 στις ΗΠΑ δίνεται στον παρακάτω πίνακα:

	Έτος					
	1982	1983	1984	1985	1986	1987
Πλήρες γάλα, x	15,6	15,2	14,7	14,3	13,4	12,8
Άπαχο γάλα, y	10,8	11,1	11,5	12,1	12,8	13,2

α) Να υπολογίσετε και να ερμηνεύσετε το συντελεστή συσχέτισης.

β) Αν μετατραπούν οι ποσότητες γάλακτος σε λίτρα, ποια θα είναι η τιμή του συντελεστή συσχέτισης; (1 γαλόνι ≈ 3,8 λίτρα)

2. Τα παρακάτω δεδομένα παριστάνουν τους δείκτες ευφυΐας (I.Q.) 10 μητέρων (X) και των θυγατέρων τους (Y):

I.Q. μητέρας	I.Q. θυγατέρας	I.Q. μητέρας	I.Q. θυγατέρας
85	90	115	110
90	100	120	125
95	90	120	110
100	105	130	130
110	120	135	120

α) Να κατασκευάσετε το διάγραμμα διασποράς

β) Από το διάγραμμα διασποράς να εκτιμήσετε το συντελεστή συσχέτισης

γ) Να υπολογίσετε και να ερμηνεύσετε το συντελεστή συσχέτισης

3. α) Να δείξετε ότι $\sum_{i=1}^{v}(x_i-\bar{x})(y_i-\bar{y}) = \sum_{i=1}^{v}x_iy_i - v\bar{x}\bar{y}$.

β) Για εφτά ζεύγη παρατηρήσεων (x_i, y_i) έχουμε $\sum(x_i-\bar{x})^2 = 28$, $\sum(y_i-\bar{y})^2 = 112$, $\sum x_iy_i = 308$, $\bar{x} = 4$, $\bar{y} = 9$. Να υπολογίσετε το συντελεστή συσχέτισης.

4. Σε μια εξέταση στα Μαθηματικά οκτώ μαθητών η βαθμολογία δύο εξεταστών Α, Β ήταν ως ακολούθως:

	Μαθητής							
	1	2	3	4	5	6	7	8
Εξεταστής Α	55	62	71	66	63	56	72	51
Εξεταστής Β	54	56	61	66	63	61	73	54

Να εξετάσετε εάν υπάρχει γραμμική συσχέτιση μεταξύ της βαθμολογίας των δύο εξεταστών.

ΓΕΝΙΚΕΣ ΑΣΚΗΣΕΙΣ

1. Ο αριθμός των παιδιών σε ένα δείγμα 80 οικογενειών μιας πόλης δίνεται στον παρακάτω πίνακα:

Αριθμός Παιδιών	0	1	2	3	4	5	6
Οικογένειες	10	25	20	12	6	5	2

α) Να βρείτε τη μέση τιμή, τη διάμεση τιμή, την επικρατούσα τιμή και την τυπική απόκλιση του αριθμού των παιδιών.

β) Να κατασκευάσετε το διάγραμμα σχετικών συχνοτήτων και το πολύγωνο αθροιστικών σχετικών συχνοτήτων.

γ) Από το πολύγωνο αθροιστικών σχετικών συχνοτήτων να εκτιμήσετε τα τρία τεταρτημόρια.

2. Ο αριθμός των τυπογραφικών λαθών που βρέθηκαν στις 60 σελίδες ενός κειμένου στην πρώτη του διόρθωση ήταν:

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

α) Να ομαδοποιήσετε τα δεδομένα σε πέντε ισοπλατείς κλάσεις πλάτους δύο και να κατασκευάσετε τον πίνακα συχνοτήτων.

β) Να κατασκευάσετε τα ιστογράμματα συχνοτήτων και αθροιστικών συχνοτήτων και τα αντίστοιχα πολύγωνα συχνοτήτων.

γ) Να υπολογίσετε τη μέση τιμή, τη διάμεσο, την κορυφή και την τυπική απόκλιση.

3. Σε μια εταιρεία συνολικά εργάζονται 200 άτομα. Όπως προέκυψε από ένα τυχαίο δείγμα υπαλλήλων, ο συνολικός χρόνος υπηρεσίας τους δίνεται στο παρακάτω κυκλικό διάγραμμα.

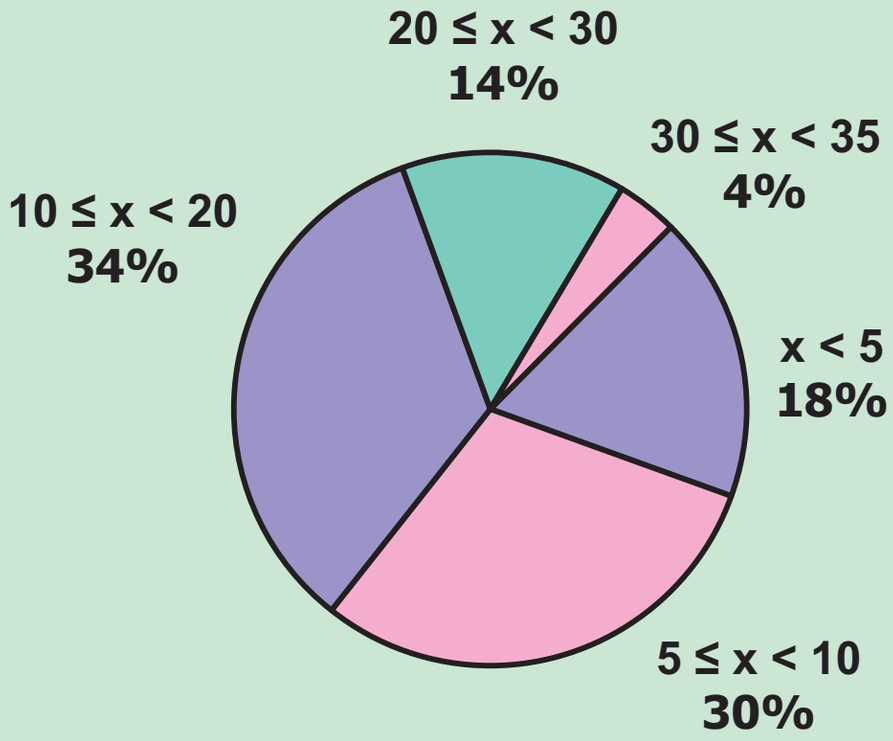
α) Να κατασκευάσετε τον πίνακα συχνοτήτων.

β) Να υπολογίσετε τη μέση τιμή και την τυπική απόκλιση.

γ) Πόσοι συνολικά υπάλληλοι αναμένονται να

συνταξιοδοτηθούν (συμπληρώνοντας 35-ετία) μέσα στα επόμενα i) 5 χρόνια, ii) 10 χρόνια;

δ) Να κατασκευάσετε το ιστόγραμμα σχετικών συχνοτήτων.



4. Τα εργατικά ατυχήματα που συνέβησαν το 1990 και το 1994 δίνονται στον παρακάτω πίνακα (Στοιχεία Υπουργείου Εργασίας).

α) Να απεικονίσετε τα δεδομένα σε ένα ραβδόγραμμα συχνοτήτων.

β) Πόσα ατυχήματα συνέβησαν κατά μέσο όρο για τα έτη 1990 και 1994;

γ) Το 1,4% των ατυχημάτων του 1990 και το 2,1%

του 1994 ήταν θανατηφόρα. Πόσα ατυχήματα ήταν θανατηφόρα για τα αντίστοιχα έτη; Ποιο είναι το συμπέρασμά σας;

Μήνες	1990	1994
Ιαν.-Φεβρ.	1057	692
Μαρ.-Απρ.	927	716
Μάιος-Ιούν.	1114	829
Ιουλ.-Αυγ.	1020	783
Σεπ.-Οκτ.	941	809
Νοεμ.-Δεκ.	775	636
Σύνολο	5834	4465

5. Ο παρακάτω πίνακας δίνει τη διάρκεια ζωής δύο τύπων ηλεκτρικών συσκευών Α και Β σε χιλιάδες ώρες. Μια ηλεκτρική συσκευή τύπου Α στοιχίζει 230 ευρώ.

A	B
12	12
14	13
23	16
30	22
36	32

α) Ποιου τύπου ηλεκτρική συσκευή θα προτιμήσετε, αν η μία ηλεκτρική συσκευή τύπου Β στοιχίζει:
i) 180 ευρώ ii) 190 ευρώ iii) 200 ευρώ.
Να αιτιολογήσετε σε κάθε περίπτωση την απάντησή σας.

β) Ποιου τύπου οι ηλεκτρικές συσκευές παρουσιάζουν μεγαλύτερη ομοιογένεια ως προς τη διάρκεια λειτουργίας τους;

6. Σε δειγματοληπτική έρευνα που έγινε στις 15 χώρες της Ευρωπαϊκής Ένωσης (Ε.Ε.) μία βδομάδα πριν και μία βδομάδα μετά το Συμβούλιο Κορυφής, (Σ.Κ.) που έγινε το Μάιο του 1998, τα ποσοστά των ατόμων που αισθάνονταν πολύ καλά πληροφορημένα για το ενιαίο νόμισμα (ευρώ) δίνονται στον παρακάτω πίνακα:

Χώρα	Πριν το Σ.Κ.	Μετά το Σ. Κ.
Αυστρία	50	47
Βέλγιο	55	55
Βρετανία	40	35
Γαλλία	61	72
Γερμανία	44	48
Δανία	51	53
Ελλάδα	26	22
Ιρλανδία	41	29
Ισπανία	30	39
Ιταλία	49	39
Λουξεμβούργο	56	62
Ολλανδία	56	55
Πορτογαλία	18	20
Σουηδία	40	38
Φινλανδία	45	45

α) Να παραστήσετε τα δεδομένα σε μορφή ραβδογράμματος.

β) Να βρεθεί το μέσο ποσοστό των πολύ καλά ενημερωμένων για τις 15 χώρες της Ε.Ε. πριν και μετά το Σ.Κ., υπολογίζοντας i) τον αριθμητικό μέσο και ii) το σταθμικό μέσο ποσοστό με βάρη τους πληθυσμούς των 15 χωρών μελών

της Ε.Ε. Ποιος από τους δύο μέσους είναι ο αντιπροσωπευτικότερος;

7. Στον παρακάτω πίνακα παριστάνονται οι χρόνοι (σε λεπτά και δευτερόλεπτα) των νικητών των Ολυμπιακών αγώνων στην κολύμβηση στα 400 μέτρα ελευθέρας (freestyle) ανδρών και γυναικών. Να δώσετε (για κάθε φύλο χωριστά) το χρονόγραμμα των δεδομένων αυτών. Τι συμπεράσματα βγάζετε;

Έτος	Χρόνος Ανδρών	Χρόνος Γυναικών
1904	6:16.2	—
1908	5:36.8	—
1912	5:24.4	—
1920	5:26.8	—
1924	5:04.2	6:02.2
1928	5:01.6	5:42.8
1932	4:48.4	5:28.5
1936	4:44.5	5:26.4
1948	4:41.0	5:17.8
1952	4:30.7	5:12.1
1956	4:27.3	4:54.6
1960	4:18.3	4:50.6

Έτος	Χρόνος Ανδρών	Χρόνος Γυναικών
1964	4:12.2	4:43.3
1968	4:09.0	4:31.8
1972	4:00.3	4:19.4
1976	3:51.9	4:09.9
1980	3:51.3	4:08.8
1984	3:51.2	4:07.1
1988	3:47.0	4:03.9
1992	3:45.0	4:07.2

Πηγή: The World Almanac and Book of Facts, 1994.

8. **Οι κάτοικοι ανά km^2 από το 1960 έως και το 1974 στην Ελλάδα ήταν:**

Κάτοικοι, Υ	63	64	64	64	66	65	65	66
Έτος, Χ	1960	61	62	63	64	65	66	67

Κάτοικοι, Υ	66	67	67	67	67	68	68
Έτος, Χ	68	69	1970	71	72	73	74

α) Να εκτιμήσετε την ευθεία γραμμικής παλινδρόμησης της Υ στη Χ, και να την παραστήσετε στο διάγραμμα διασποράς.

β) Το 1976 είχαμε 69,5 κατοίκους/km^2. Είναι αυτό αναμενόμενο;

(Υπόδειξη: Θεωρούμε ως έτος αναφοράς το 1960 με τιμή x = 1).

9. Ο συντελεστής γενικής θνησιμότητας (Σ.Γ.Θ.) της Ελλάδας για τα χρόνια 1931-1964 παρουσίασε την παρακάτω πορεία.

Έτος, Χ	1931	1936	1940	1950	1956	1961	1964
Σ.Γ.Θ.%, Υ	17,7	15,1	12,8	7,9	7,4	7,6	8,2

α) Να χαράξετε "με το μάτι" την ευθεία γραμμικής παλινδρόμησης y = α + βx στο διάγραμμα διασποράς και από την ευθεία αυτή να εκτιμήσετε το Σ.Γ.Θ. για το έτος 1965.

β) Χρησιμοποιώντας τη μέθοδο των δύο σημείων να υπολογίσετε την εξίσωση της ευθείας παλινδρόμησης και στη συνέχεια να εκτιμήσετε πάλι το Σ.Γ.Θ. για το έτος 1965. Συγκρίνετε με το προηγούμενο αποτέλεσμα.

γ) Να επαναλάβετε το ίδιο χρησιμοποιώντας τη μέθοδο ελαχίστων τετραγώνων.

(Υπόδειξη: Για την ανεξάρτητη μεταβλητή Χ να

θέσετε για το έτος 1931 ως τιμή το 1, οπότε για το 1936 το x = 6 και για το 1965 το 35).

10. Ο παρακάτω πίνακας δίνει το ημερήσιο εισόδημα X και τις αντίστοιχες δαπάνες διατροφής Y πέντε υπαλλήλων, που πάρθηκαν τυχαία από μια εταιρεία.

Εισόδημα, (δεκάδες ευρώ) ,	X	3,5	3,7	4,2	4,3	6,9
Δαπάνες διατροφής (δεκάδες ευρώ),	Y	1,1	1,5	1,8	1,5	2,5

α) Με τη μέθοδο των "ελαχίστων τετραγώνων" να βρείτε την εξίσωση της ευθείας γραμμικής παλινδρόμησης των εξόδων διατροφής (πάνω) στο εισόδημα.

β) Μια υπάλληλος της εταιρείας έχει ημερήσιο εισόδημα 50 ευρώ.
Πόσο εκτιμάτε εσείς ότι θα ξοδεύει για διατροφή την ημέρα;

γ) Αν γνωρίζετε ότι μια υπάλληλος ξοδεύει 30 ευρώ για διατροφή μπορείτε, με βάση τα παραπάνω, να προβλέψετε το ημερήσιο εισόδημά της;

11. Η ποσότητα $s_{xy} = \frac{1}{v}\sum_{i=1}^{v}(x_i - \bar{x})(y_i - \bar{y})$ καλείται συνδιακύμανση των μεταβλητών X και Y. Αν καλέσουμε με s_x^2, s_y^2 τις διακυμάνσεις των X και Y αντίστοιχα, να δείξετε ότι ισχύουν οι σχέσεις:

α) $\hat{\beta} = \frac{s_{xy}}{s_x^2}$ β) $r = \hat{\beta}\frac{s_x}{s_y}$.

12. Ένας μαθητής γνώριζε ότι η σχέση που συνδέει τους βαθμούς Φαρενάιτ (°F) με τους βαθμούς Κελσίου (°C) είναι γραμμική, δηλαδή $F = \alpha + \beta C$. Επειδή όμως δε θυμόταν τις σταθερές α, β, μέτρησε τη θερμοκρασία του δωματίου του σε πέντε διαφορετικές ώρες με δύο θερμόμετρα με κλίμακα σε °F και °C, αντίστοιχα, και πήρε τα παρακάτω ζεύγη τιμών:

°C	15	20	25	30	35
°F	59	68	77	86	95

Να βρείτε τη σχέση $\hat{F} = \hat{\alpha} + \hat{\beta}C$ που συνδέει τις δύο κλίμακες θερμοκρασίας.

13. Δίνεται δείγμα ν ζευγών (x_1, y_1), (x_2, y_2), …, (x_v, y_v) δύο μεταβλητών X και Y και έστω r(X, Y) ο συντελεστής συσχέτισης. Εάν $Z = \lambda Y$ όπου λ θετική σταθερά, να δείξετε ότι ισχύει:

$$r(X, Z) = r(X, Y).$$

Τι γίνεται, εάν $\lambda < 0$;

14. Ο αριθμός των διαζυγίων που εκδόθηκαν στην Κύπρο από το 1974 έως το 1994 δίνεται παρακάτω (Πηγή: Τμήμα Στατιστικής και Ερευνών Κύπρου).

Έτος	x	Αριθμός Διαζυγίων y
1974	1	140
1975	2	121
1976	3	110
1977	4	136
1978	5	158
1979	6	161

Έτος	x	Αριθμός Διαζυγίων y
1980	7	164
1981	8	175
1982	9	216
1983	10	262
1984	11	250
1985	12	258
1986	13	276
1987	14	326
1988	15	312
1989	16	335
1990	17	348
1991	18	304
1992	19	433
1993	20	504
1994	21	555
	Δίνονται: $v = 21$ $\Sigma x^2 = 3.311$ $\Sigma x = 231$ $\Sigma y^2 = 17.726.800$ $\Sigma y = 5.544$ $\Sigma xy = 75.512$	

Να βρείτε την ευθεία "ελαχίστων τετραγώνων" και να εκτιμήσετε τον αριθμό των διαζυγίων για τα έτη 1995, 2000.

ΕΡΩΤΗΣΕΙΣ ΚΑΤΑΝΟΗΣΗΣ

Στις ερωτήσεις 1-10 να βάλετε σε κύκλο το Σ (Σωστό) ή το Λ (Λάθος).

1. **Πάντοτε ένα μεγαλύτερο δείγμα δίνει πιο αξιόπιστα αποτελέσματα από ένα μικρότερο δείγμα.**

Σ Λ

2. **Όταν έχουμε συμμετρική κατανομή, η μέση τιμή συμπίπτει με τη διάμεσο.**

Σ Λ

3. **Όταν έχουμε ακραίες παρατηρήσεις, είναι προτιμότερο να χρησιμοποιούμε τη μέση τιμή αντί της διαμέσου.**

Σ Λ

4. **Ο λόγος της μέσης τιμής προς την τυπική απόκλιση καλείται συντελεστής μεταβολής και είναι καθαρός αριθμός.**

Σ Λ

5. Όταν προσθέσουμε μια σταθερά στις παρατηρήσεις μιας μεταβλητής τότε η μέση τιμή και η τυπική απόκλιση αυξάνουν κατά τη σταθερά αυτή.

Σ Λ

6. Όταν πολλαπλασιάσουμε τις τιμές μιας μεταβλητής επί μια σταθερά, τότε η μέση τιμή πολλαπλασιάζεται επί την ίδια σταθερά.

Σ Λ

7. Όταν πολλαπλασιάσουμε τις τιμές μιας μεταβλητής επί μια σταθερά, τότε η τυπική απόκλιση πολλαπλασιάζεται επί την ίδια σταθερά.

Σ Λ

8. Η διάμεσος και το δεύτερο τεταρτημόριο έχουν πάντα την ίδια τιμή.

Σ Λ

9. Το βάρος της ζάχαρης που βάζουμε στους καφέδες είναι ποιοτική μεταβλητή, γιατί χαρακτηρίζει τον καφέ σκέτο, μέτριο ή γλυκύ.

Σ Λ

10. Η σχετική συχνότητα μπορεί να πάρει και αρνητικές τιμές.

Σ Λ

11. Για την ανεξάρτητη μεταβλητή οι παρατηρήσεις είτε προκαθορίζονται είτε λαμβάνονται χωρίς να υπεισέρχεται σφάλμα μέτρησης.

Σ Λ

12. Η $\hat{\beta}$ παριστάνει την αύξηση της εξαρτημένης μεταβλητής, όταν η ανεξάρτητη μεταβλητή αυξηθεί κατά μία μονάδα.

Σ Λ

13. Ένας συντελεστής συσχέτισης $r = +0,6$ δείχνει μεγαλύτερη γραμμική συσχέτιση μεταξύ δύο μεταβλητών παρά ο $r = -0,9$.

Σ Λ

14. Όταν $r(X, Y) > 0$, τότε συνεπάγεται ότι οι μεταβλητές X, Y είναι θετικά συσχετισμένες.

Σ Λ

Στις ερωτήσεις 15-24 να βάλετε σε κύκλο τη σωστή απάντηση.

15. Ένα μέτρο που χρησιμοποιείται τόσο για ποιοτικά όσο και για ποσοτικά δεδομένα είναι:

Α. η μέση τιμή

Β. η επικρατούσα τιμή

Γ. η τυπική απόκλιση

Δ. κανένα από τα παραπάνω.

16. Η διακύμανση των παρατηρήσεων $x_1, x_2, \ldots, x_v$ δίνεται από τον τύπο:

Α. $s^2 = \frac{1}{v}\sum(x_i - \bar{x})$

Β. $s^2 = v\sum x_i^2 - \left(\sum x_i\right)^2$

Γ. $s^2 = \frac{1}{v}\left\{\sum x_i^2 - \left(\sum x_i\right)^2\right\}$

Δ. $s^2 = \frac{v\sum x_i^2 - \left(\sum x_i\right)^2}{v^2}$

17. Εάν οι συντελεστές μεταβολής δύο συνόλων δεδομένων Α και Β είναι 15% και 20% αντιστοίχως, τότε:

Α: τα δεδομένα Α παρουσιάζουν μεγαλύτερη ομοιογένεια από τα Β

Β: τα δεδομένα Α παρουσιάζουν μικρότερη ομοιογένεια από τα Β

Γ: τα δεδομένα Α παρουσιάζουν μεγαλύτερη διασπορά από τα Β

Δ: τα δεδομένα Α παρουσιάζουν μικρότερη διασπορά από τα Β

18. Με βάση την ευθεία παλινδρόμησης $\hat{y} = -10 + 3{,}25x$ η προβλεπόμενη τιμή $\hat{y}$ για $x = 10$ είναι:

Α. 3,25

Β. −10

Γ. 22,5

Δ. Δεν μπορούμε να ξέρουμε.

19. Με βάση την ευθεία παλινδρόμησης $\hat{y} = 2 - 3x$ ο συντελεστής γραμμικής συσχέτισης των Χ, Υ είναι:

Α. −3

Β. Θετικός

Γ. Αρνητικός

Δ. $-\frac{3}{2}$.

20. Εάν ο συντελεστής γραμμικής συσχέτισης δύο μεταβλητών X, Y είναι r = +1, τότε η ευθεία γραμμικής παλινδρόμησης της Y στη X μπορεί να διέρχεται από τα σημεία:

Α. (0, 0) και (1, −1)

Β. (1, −1) και (1, 0)

Γ. (−1, −1) και (1, 1)

Δ. (0, 1) και (1, 0).

21. Ο συντελεστής γραμμικής συσχέτισης r και ο συντελεστής $\hat{\beta}$ στην ευθεία γραμμικής παλινδρόμησης $\hat{y} = \hat{\alpha} + \hat{\beta}x$ έχουν:

Α. πάντα το ίδιο πρόσημο

Β. πάντα διαφορετικό πρόσημο

Γ. άλλοτε το ίδιο πρόσημο και άλλοτε διαφορετικό

Δ. δεν έχουν καμιά σχέση ως προς το πρόσημό τους.

22. Εάν r(X, Y) = 0, τότε οι X, Y είναι:

Α. ασυσχέτιστες

Β. γραμμικά ασυσχέτιστες

Γ. τέλεια θετικά συσχετισμένες

Δ. τέλεια αρνητικά συσχετισμένες.

23. Στην παλινδρόμηση με $\hat{y}$ συμβολίζουμε:

Α. τις πραγματικές τιμές της εξαρτημένης μεταβλητής

Β. τις τιμές της ανεξάρτητης μεταβλητής

Γ. τις προβλεπόμενες τιμές της εξαρτημένης μεταβλητής, που προκύπτουν από την εξίσωση γραμμικής παλινδρόμησης

Δ. κανένα από τα παραπάνω.

24. Οι μεταβλητές X, Y έχουν συντελεστή συσχέτισης $r_1 = +0,9$, ενώ οι Z, W έχουν συντελεστή συσχέτισης $r_2 = +0,3$.

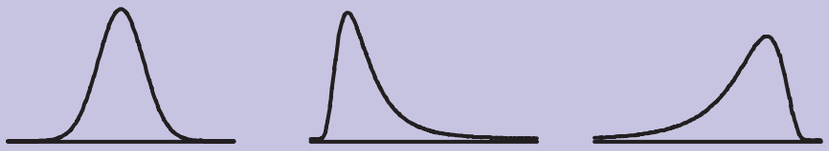
Α. Οι X, Y είναι τριπλάσια συσχετισμένες από τις Z, W

Β. Οι X, Y είναι περισσότερο (σε μεγαλύτερο βαθμό) συσχετισμένες από τις Z, W

Γ. Δεν μπορούμε να συγκρίνουμε διαφορετικές μεταβλητές.

Στις ερωτήσεις 25-35 να γίνει αντιστοίχιση των (α), (β)... με τα (i), (ii), ..., όπου αυτή είναι δυνατή.

25.



(α) (β) (γ)

• • •

• • •

i) $\bar{x} = \delta$ ii) $\bar{x} < \delta$ iii) $\bar{x} > \delta$

26. α) 1 2 10 18 19 • • i) $\bar{x} = 10$, $s \approx 7,5$

• ii) $\bar{x} = 20$, $s \approx 7,5$

β) 18 19 20 21 22 • • iii) $\bar{x} = 10$, $s \approx 1,4$

• iv) $\bar{x} = 20$, $s \approx 1,4$

γ) 8 9 10 11 12 • • v) $\bar{x} = 15$, $s \approx \sqrt{2}$.

27. α) διάμεσος •

β) επικρατούσα τιμή • • i) μέτρο θέσης

γ) τυπική απόκλιση •

δ) εύρος • • ii) μέτρο διασπο-

ε) διακύμανση • ράς

στ) μέση τιμή •

28. α) 5 7 8 10 13 24 • • i) $\overline{x} = 9$

β) 1 2 8 9 9 25 • • ii) $\delta = 9$

γ) 1 2 9 12 12 18 • • iii) $M_0 = 9$

29. α) 10 11 12 13 14 • • i) $\overline{x} < \delta$

β) 10 11 12 13 24 •

γ) 1 11 12 13 14 • • ii) $\overline{x} = \delta$

δ) 20 21 22 23 24 •

ε) 30 33 36 39 42 • • iii) $\overline{x} > \delta$

30. Παρακάτω δίνονται οι καμπύλες συχνοτήτων (α) έως (δ) τεσσάρων μεταβλητών (i) έως (iv) από μια μελέτη που έγινε σε κάποια πόλη.

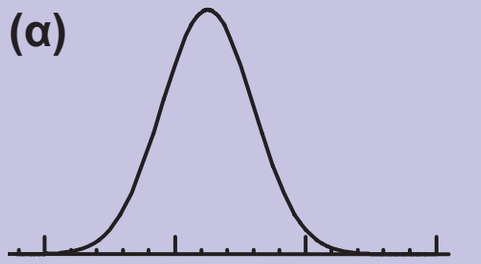
i) Ύψος των μελών των νοικοκυριών στα οποία οι γονείς είναι και οι δύο κάτω των 24 ετών.

ii) Ύψος των παντρεμένων ζευγαριών.

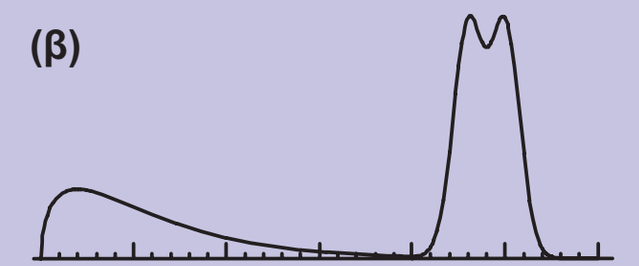
iii) Ύψος όλων των ατόμων.

iv) Ύψος όλων των αυτοκινήτων.

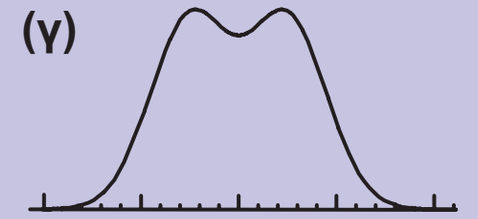
(α)



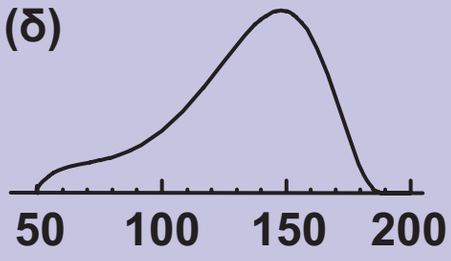
(β)



(γ)



(δ)



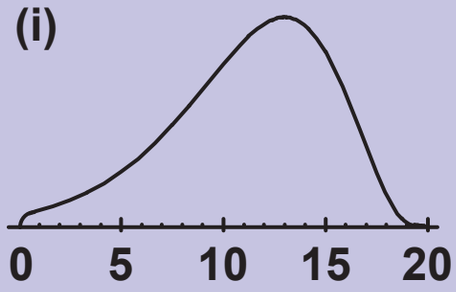
31. Παρακάτω δίνονται οι καμπύλες σχετικών συχνοτήτων ((i) έως (iii)) της βαθμολογίας τριών τμημάτων σε ένα διαγώνισμα, κατά το οποίο

α) στο πρώτο τμήμα πέρασε το 50%

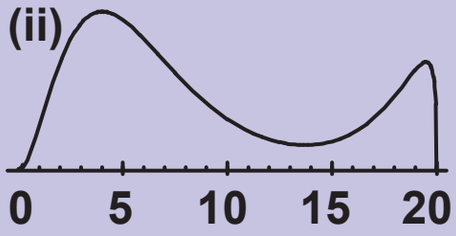
β) Στο δεύτερο τμήμα πέρασε ποσοστό άνω του 50%

γ) Στο τρίτο τμήμα πέρασε ποσοστό κάτω του 50%.

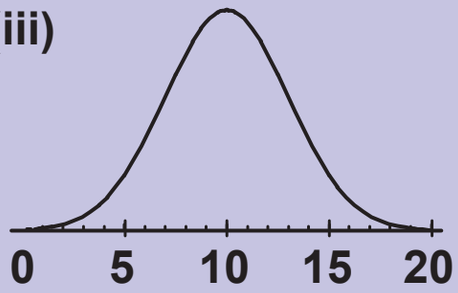
(i)



(ii)

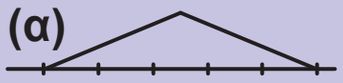


(iii)



32. Παρακάτω δίνονται κατά προσέγγιση οι καμπύλες συχνοτήτων (α) έως (γ) τριών διαφορετικών συνόλων δεδομένων και διάφορες τιμές (i) έως (iv) της μέσης τιμής και της τυπικής απόκλισης:

(α)



(β)

(γ)

(i) $\bar{x} \approx 3{,}5$, $s \approx 1$

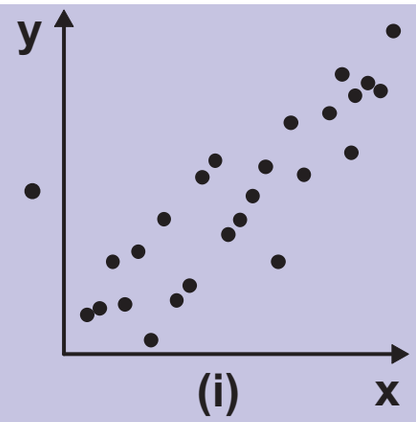
(ii) $\bar{x} \approx 3{,}5$, $s \approx 2$

(iii) $\bar{x} \approx 2{,}5$, $s \approx 1$

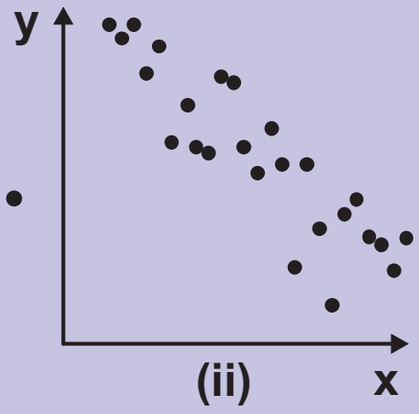
(iv) $\bar{x} \approx 2{,}5$, $s \approx 2$.

33.

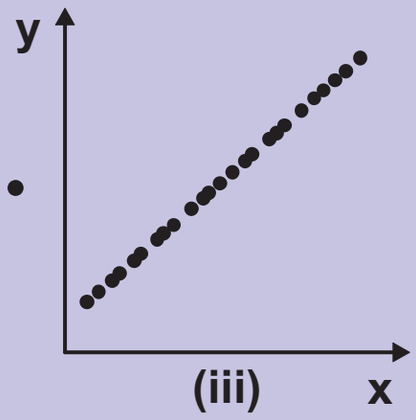
(α) $r \approx 0$



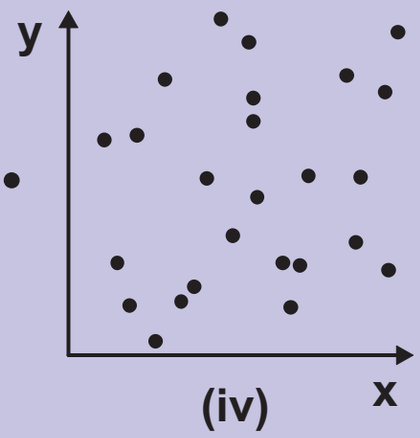
(β) $r \approx 0{,}8$



(γ) $r \approx +1$



(δ) $r \approx -0{,}8$



34.

$r = 0$	•	•	$\hat{\beta} = 0$
$r > 0$	•	•	$\hat{\beta} < 0$
$r < 0$	•	•	$\hat{\beta} > 0$.

35. Για την ευθεία γραμμικής παλινδρόμησης $\hat{y} = 2x$ ισχύει:

$\hat{\alpha} = 0$	•		
$r > 0$	•	•	Σωστό
$\hat{\beta} = 2$	•	•	Λάθος
$r = \hat{\beta}$	•	•	Δεν μπορούμε να ξέρουμε.

ΥΠΟΔΕΙΞΕΙΣ – ΑΠΑΝΤΗΣΕΙΣ ΑΣΚΗΣΕΩΝ

2 ΣΤΑΤΙΣΤΙΚΗ

§ 2.1

1. Ποιοτικές: γ, δ, στ, ζ
   Ποσοτικές-διακριτές: β, η, θ, ι
   Ποσοτικές-συνεχείς: α, ε
2. Είναι δυνατόν να έχουμε διάφορες μεταβλητές για κάθε περίπτωση. Για παράδειγμα:
   α) μισθός (ποσοτική-συνεχής), ηλικία (ποσοτική-διακριτή), φύλο (ποιοτική) κτλ.
3. (ε).
4. Είναι δυνατό να έχουμε διάφορους λόγους ακαταλληλότητας των επιλεγομένων δειγμάτων. Για παράδειγμα:
   α) Θα έχουμε υπερεκτίμηση των ανδρών κτλ.

§ 2.2 Α΄ Ομάδας

1. β) i) 26% ii) 14% iii) 38%.
2. Έχουμε 11 αγόρια με βαθμό < 5, κτλ.
3. α) Το 22% των φοιτητών είναι αγόρια με βαθμό < 5, κτλ.
4. α) 76% β) 16% γ) 34% δ) 84% ε) 16%.
5. Από τη σχέση $f_2 = \frac{v_2}{v}$ βρίσκουμε πρώτα το μέγεθος $v = 20$. Να βρεθεί μετά το N_1 κτλ.
6. Να εργαστείτε όπως στο σχήμα 1(γ).
7. Να κατασκευάσετε πρώτα τον πίνακα συχνοτήτων και μετά να εργαστείτε όπως στα σχήματα 1(β) και 3.
8. Για τους $v = 450$ μαθητές έχουμε 30% με τιμή x_2 = "Λίαν καλώς". Δηλαδή
   $f_2\% = 30 \Leftrightarrow v_2 = 135 \Leftrightarrow \alpha_2 = 108°$, κτλ.
9. Να εργαστείτε όπως στα σχήματα 1(β) και 3.
10. Να εργαστείτε όπως στο σχήμα 1(γ).
11. Να εργαστείτε όπως στο σχήμα 5.
12. γ) 11 άτομα
    δ) επίδοση ≥ 7.

13. 50%.

14. ΝΑΙ. Το εμβαδόν πρέπει να είναι 100.

---

§ 2.2 Β΄ Ομάδας

---

1. Για τη Λέσβο υπάρχει πτωτική τάση ενώ για τη Σαλαμίνα υπάρχει ανοδική τάση.
Για τη Θάσο υπάρχει περίπου σταθερή κατάσταση.

2. Να συμπληρώσετε τον πίνακα:

Έτος	Ηλικία			Σύνολο
	≤ 20	21 - 30	≥31	

3. Ανάλογα με την άσκηση 2.

4. β) 355 γυναίκες, 434 άνδρες
γ) 692
δ) Δεν μπορούμε να ξέρουμε με τα στοιχεία που μας δίνονται.

5. Να εργαστείτε όπως στο σχήμα 5.

6. α) Φάρμακο Α 17,3% και Φάρμακο Β 26%.

7. Επειδή ν = 55 να χρησιμοποιήσετε κ = 7 κλάσεις με πλάτος c = 1,8.

§ 2.3 Α΄ Ομάδας

1. 10, 12, 14, 16, 18, 20 με διάμεσο δ = 15.
2. α) ΝΑΙ, β) ΟΧΙ, γ) ΝΑΙ.
3. 8,25%.
4. α) 206,1 cm β) 235 cm.
5. 14,8.
6. Και στις 3 περιπτώσεις έχουμε:
   $\overline{x}$ = 12gr, δ = 12gr, M_0 = 13gr.
7. α) 14, β) 13.
8. 1291 ευρώ.
9. Οι 2 και 6.
10. α) 2, β) 2, γ) 3.
11. α) $\overline{x}$ = 15,45, β) δ = 15, γ) M_0 = 15,
    δ) Q_1 = 14, Q_3 = 17.
12. α) $\overline{x}$ = 4,3 β) M_0 = 3,3, γ) $Q_1 \approx 2{,}33$, δ = 4, $Q_3 \approx 6$.
13. α) 169,66 cm, β) 170 cm, γ) 169,67 cm.
14. α) Μικρότερη διασπορά έχουμε στη δεύτερη λίστα και μεγαλύτερη διασπορά στην τρίτη.
    β) ΟΧΙ.

15. α) $\overline{x} = 10$, $M_0 = 11$, $\delta = 11$

β) $Q_1 = 7$, $Q_3 = 13$

γ) $R = 12$, $s = 3{,}87$, $cv = 38{,}7\%$.

16. 2,47.

17. α) 16%, β) 2,5%, γ) 50%, δ) 81,5%.

18. α) $\overline{x} = 3$, $\delta = 2$

β) $\overline{x} = 6$, $\delta = 4$

γ) $\overline{x} = 13$, $\delta = 12$

δ) $\overline{x} = 16$, $\delta = 14$.

19. α) $s^2 = 4$, β) $s^2 = 36$, γ) $s^2 = 4$, $s^2 = 4$.

20. $v = 28$.

§ 2.3 Β΄ Ομάδας

1. β) $\overline{x} = 15$, $\delta \approx 15$ γ) $P_{95} \approx 19$.

2. 10 ή −2.

3. α) $\overline{x} = 13{,}20$ ευρώ, $\delta = 10{,}50$ ευρώ,
$M_0 = 9$ ευρώ.

β) Αυξάνουν κατά 18%.

γ) Αυξάνουν κατά 0,30 ευρώ.

4. Να γίνουν οι πράξεις.

5. α) 60, β) 33, γ) 5,2 χιλιάδες ευρώ,
δ) $\overline{x}$ = 5,7 χιλιάδες ευρώ, s^2 = 12,24
ε) $Q_1 \approx 2,8$, $\delta \approx 5,4$, $Q_3 \approx 8,3$.

6. α) s = 23,29, cv = 53,75%
β) $Q = Q_3 - Q_1 \approx 59 - 24 = 35$.

7. Να εργαστείτε όπως στο σχήμα 7.

§ 2.4 Α΄ Ομάδας

1. Να εργαστείτε όπως στο σχήμα 16.

2. i) $y = 3 + \frac{5}{6}x$ ii) y = 46,7 + 0,67x.

3. α) $\hat{y} = x$, β) $\hat{y} = 6 - x$, γ) $\hat{y} = 3,9 - 0,3x$
δ) $\hat{y} = 2,1 + 0,3x$.

4. $\hat{y} = 18,98 - 21,6x$.

5. α) $\hat{y} = -0,35 + 1,07x$, β) 16.

§ 2.4 Β΄ Ομάδας

1. α) Η ηλικία.
   γ,δ) Καθένας από τους μαθητές μπορεί να φέρει τη "δική του" ευθεία, οπότε θα έχει και διαφορετική πρόβλεψη συστολικής πίεσης.
2. α) 178,7 cm β) 177,5 cm.
3. α) 168,3 cm β) 168,1 cm.
4. α) $\hat{y} = -1{,}88 + 1{,}18x$.
   β) Περίπου 27 έτη και 7 μήνες.
   γ) Περίπου 1 έτος και 2 μήνες.
5. α) $\hat{x} = 5{,}25 + 0{,}72y$.
   β) Περίπου 25 έτη και 5 μήνες.
   γ) Περίπου 9 μήνες.

§ 2.5 Α΄ Ομάδας

1. 0, 0,2, −0,6, −0,7, 0,9, −1.
2. α) Θετική -μεγάλη, κτλ.

3. α) $r = 0,99$ β) $r = 0,59$ γ) $r = -0,70$

   δ) $r = -0,05$ ε) $r = 0,33$.

4. α) $r_1 = -1$ β) $r_2 = 1$.

5. $r = 0,97$.

6. $r = 0,74$.

§ 2.5 Β΄ Ομάδας

1. α) $r \approx -1$ β) $r \approx -1$.

2. $r = 0,87$.

3. $r = 1$.

4. $r = 0,77$.

ΓΕΝΙΚΕΣ ΑΣΚΗΣΕΙΣ

1. $\overline{x} = 2,025$, $\delta = 2$, $M_0 = 1$, $s = 1,49$.

2. α) $v = 60$, $\kappa = 5$, $c = 2$ γ) $\overline{x} = 4,4$, $\delta = 4,3$, $M_0 \approx 4,25$, $s = 2,29$.

3. $v_1 = \frac{18}{100} \cdot 200 = 36$ κτλ.

β) $\bar{x} = 14{,}1$, $s = 7{,}4$ γ) i) 8, ii) 22

4. β) $\bar{x}_{1990} \approx 486$, $\bar{x}_{1994} \approx 372$ γ) 82, 94.

5. α) i) τις ηλ. συσκευές τύπου Β
ii) δεν έχουμε προτίμηση
iii) τις ηλ. συσκευές τύπου Α.
β) $cv_A = 39{,}8\%$, $cv_B = 38{,}8\%$.

6. α) Να εργαστείτε όπως στο σχήμα 1(γ).
β) Ο σταθμικός μέσος.

7. Να εργαστείτε όπως στο σχήμα 5.

8. α) $\hat{y} = 63{,}1 + 0{,}34x$ β) 68,9 κάτοικοι/km^2

9. γ) $\hat{y} = 16{,}53 - 0{,}3x$.

10. α) $\hat{y} = 0{,}06 + 0{,}36x$ β) 18,6 ευρώ γ) Πρέπει να βρείτε την ευθεία ελαχίστων τετραγώνων της Χ πάνω στην Υ.

11. Δείτε και την άσκηση 7(α) Α΄ Ομάδας της §2.5.

12. $\hat{F} = 32 + 1{,}8C$.

13. $r(X,Y) = \begin{cases} r(X,Y), & \text{εάν } \lambda > 0 \\ -r(X,Y), & \text{εάν } \lambda < 0 \end{cases}$.

14. $\hat{y} = 56,4 + 18,87x$

$\hat{y}_{1995} \approx 472$ διαζύγια

$\hat{y}_{2000} \approx 547$ διαζύγια.

Ευρετήριο Όρων

Στο Ευρετήριο όρων τα γράμματα Α, Β και Γ δηλώνουν τον 1ο, 2ο και 3ο τόμο αντίστοιχα, ενώ οι αριθμοί αναφέρονται στην πρώτη από τις δύο ενδείξεις που αναγράφονται σε κάθε σελίδα.

αδύνατο ενδεχόμενο	Γ' 11
αθροιστικές συχνότητες	Β' 32
αθροιστικές σχετικές συχνότητες	Β' 32
ανεξάρτητα ενδεχόμενα	Γ' 71
ανεξάρτητη μεταβλητή	Α' 9, Β' 126
αξιωματικός ορισμός πιθανότητας	Γ' 32
απλός προσθετικός νόμος	Γ' 34
απογραφή	Β' 15
ασυμβίβαστα ενδεχόμενα	Γ'15, Γ' 34
βασική αρχή απαρίθμησης	Γ' 50
βέβαιο ενδεχόμενο	Γ' 11
γραμμική συσχέτιση	Β' 151, Β' 156
γραφική παράσταση συνάρτησης	Α' 11
δείγμα	Β' 16
δειγματικός χώρος	Γ' 10
δειγματοληψία	Β' 17
δειγματοληψία με επανατοποθέτηση	Γ' 20
δεντροδιάγραμμα	Γ' 49

δεσμευμένη πιθανότητα Γ' 67
δεύτερη παράγωγος Α' 45
δημοσκόπηση Β' 9
διάγραμμα διασποράς Β' 128
διάγραμμα συχνοτήτων Β' 40
διακριτή μεταβλητή Β' 14
διακύμανση Β' 99
διαλογή Β' 29, Β' 47
διάμεσος Β' 87
διάμεσος ομαδοποιημένης κατανομής Β' 88
διασπορά Β' 99
διατάξεις Γ' 51
εκατοστημόριο Β' 89
εκθετική συνάρτηση Α' 14
εκτιμήτριες ελαχίστων τετραγώνων Β' 135
ενδεχόμενο Γ' 10
ενδοτεταρτημοριακό εύρος Β' 98
εξαρτημένα ενδεχόμενα Γ' 70
εξαρτημένη μεταβλητή Α' 9, Β' 126
εξίσωση γραφικής παράστασης Α' 12
επαγωγή Β' 8
επικρατούσα τιμή Β' 92
ευθεία παλινδρόμησης Β' 130
ευνοϊκές περιπτώσεις Γ' 11, Γ' 31
εύρος Β' 48, Β' 96
εφαπτομένη καμπύλης Β' 28

ισοπίθανα απλά ενδεχόμενα Γ' 30
ιστόγραμμα Β' 50
ιστόγραμμα αθροιστικών συχνοτήτων Β' 52
καμπύλη συχνοτήτων Β' 57
κανόνες παραγώγισης Α' 50
κανονική κατανομή Β' 58
καμπύλη συνάρτησης Α' 11
κατανομή συχνοτήτων Β' 32
κατηγορική μεταβλητή Β' 13
κεντρική τιμή κλάσης Β' 45
κλάσεις Β' 45
κλάσεις ανίσου πλάτους Β' 53
κλάσεις ίσου πλάτους Β' 50
κλασικός ορισμός πιθανότητας Γ' 31
κορυφή Β' 92
κριτήριο δεύτερης παραγώγου Α' 74
κριτήριο πρώτης παραγώγου Α' 67
κυκλικό διάγραμμα Β' 41
κύμανση Β' 96
λογαριθμική συνάρτηση Α' 14
μέθοδος ελαχίστων τετραγώνων Β' 132
μέση τιμή Β' 81
μεταβλητή Β' 13
μεταθέσεις Γ' 53
μέτρα ασυμμετρίας Β' 79
μέτρα διασποράς Β' 96

μέτρα θέσης Β' 79
μονοτονία Α' 17
ολικό ελάχιστο Α' 17
ολικό μέγιστο Α' 17
ομαδοποίηση παρατηρήσεων Β' 45
ομοιογένεια Β' 106
ομοιόμορφη κατανομή Β' 58
όρια κλάσης Β' 45
όριο συνάρτησης Α' 19
παλινδρόμηση Β' 125
παραβολή Α' 13
παραγοντικό Γ' 53
παράγωγος συνάρτησης Α' 44
παράγωγος σύνθετης συνάρτησης Α' 53
παράγωγος της f στο x_0 Α' 35
πείραμα τύχης Γ' 8
περιγραφική στατιστική Β' 8
πίνακας συχνοτήτων Β' 32
πλάτος κλάσης Β' 46
πληθυσμός Β' 12
ποιοτική μεταβλητή Β' 13
πολλαπλασιαστικός νόμος Γ' 69
πολύγωνο συχνοτήτων Β' 40, Β' 51
ποσοτική μεταβλητή Β' 14
πράξεις με ενδεχόμενα Γ' 12
πράξεις με συναρτήσεις Α' 11

προσθετικός νόμος Γ' 36
ραβδόγραμμα Β' 35
ρυθμός μεταβολής Α' 36
σημειόγραμμα Β' 43
σταθμικός μέσος Β' 85
στατιστική ομαλότητα Γ' 29
στατιστικοί πίνακες Β' 21
στιγμιαία ταχύτητα Α' 30
συμπληρωματικά ενδεχόμενα Γ' 35
συνάρτηση αύξουσα Α' 17
συνάρτηση γνησίως μονότονη Α' 17
συνάρτηση ημίτονο Α' 15
συνάρτηση πραγματική Α' 9
συνάρτηση συνεχής Α' 23
συνάρτηση συνημίτονο Α' 15
συνάρτηση φθίνουσα Α' 17
συνδυασμοί Γ' 55
συνεχής μεταβλητή Β' 14
συντελεστής γραμμικής συσχέτισης Β' 154
συντελεστής μεταβολής Β' 105
συχνότητα Β' 29
συχνότητα κλάσης Β' 47
σχεδιασμός πειραμάτων Β' 8
σχετική συχνότητα Β' 31
τεταρτημόριο Β' 90
τοπικό ελάχιστο Α' 18

τοπικό μέγιστο **Α' 18**
τυπική απόκλιση **Β' 102**
υπερβολή **Α' 13**
χρονόγραμμα **Β' 43**

ΠΕΡΙΕΧΟΜΕΝΑ 2ου ΤΟΜΟΥ

Σελίδα

ΚΕΦΑΛΑΙΟ 2ο: Στατιστική

2.1 Βασικές Έννοιες 12
2.2 Παρουσίαση Στατιστικών Δεδομένων 21
2.3 Μέτρα Θέσης και Διασποράς 77
2.4 Γραμμική Παλινδρόμηση 124
2.5 Γραμμική Συσχέτιση 151

ΥΠΟΔΕΙΞΕΙΣ - ΑΠΑΝΤΗΣΕΙΣ ΑΣΚΗΣΕΩΝ 195