

Εισαγωγή

Το παρόν βιβλίο προέκυψε έπειτα από πολύχρονη εμπειρία μας στη διδασκαλία των μαθηματικών και της στατιστικής σε φοιτητές των γεωπονικών επιστημών. Κατά τη διάρκεια αυτής της πορείας, διαμορφώθηκε μια συγκεκριμένη φιλοσοφία για το κατάλληλο περιεχόμενο του αντικειμένου σε προπτυχιακό επίπεδο. Η φιλοσοφία αυτή στηρίζεται στην πεποίθησή μας ότι η διδασκαλία των μαθηματικών και της στατιστικής οφείλει να είναι λειτουργική, προσαρμοσμένη στις ανάγκες του φοιτητή-γεωπόνου και εστιασμένη στην ερμηνεία και στην εφαρμογή των εννοιών σε πραγματικά προβλήματα του αγροδιατροφικού τομέα.

Η προσέγγιση αυτή αποκρυσταλλώνεται στο παρόν εγχειρίδιο, το οποίο είναι εκτενές και πρακτικά προσανατολισμένο, με ειδική έμφαση στις ανάγκες της γεωπονικής και των βιολογικών επιστημών. Ο στόχος του είναι να γεφυρώσει τη θεωρητική κατανόηση με τις πραγματικές εφαρμογές στο πεδίο, δίνοντας έμφαση όχι μόνο στην παρουσίαση των βασικών εννοιών αλλά και στη δυνατότητα αξιοποίησής τους σε προβλήματα που προκύπτουν από τη μελέτη του φυσικού κόσμου.

Η επιλογή της ύλης έγινε με γνώμονα τη σαφήνεια, τη συνάφεια και τη δυνατότητα σύνδεσης με εφαρμογές στη γεωπονία, στην περιβαλλοντική επιστήμη και στην αγροτική οικονομία. Ιδιαίτερη βαρύτητα δόθηκε σε παραδείγματα και προβλήματα που προέρχονται από το ευρύτερο γεωργικό και αγροβιολογικό πεδίο, ώστε η θεωρία να γίνεται εύκολα κατανοητή και η χρήση της άμεσα αξιοποιήσιμη.

Αναγνωρίζουμε με ειλικρίνεια τη συμβολή των φοιτητών μας, των οποίων η ενεργή συμμετοχή, οι ερωτήσεις, οι απορίες και οι προτάσεις διαμόρφωσαν καθοριστικά τη διδακτική μας προσέγγιση και, τελικά, τη δομή και το περιεχόμενο του βιβλίου αυτού. Το ενδιαφέρον και η συνέπειά τους υπήρξαν για εμάς πηγή έμπνευσης και διαρκούς ανανέωσης. Σε αυτούς αφιερώνεται αυτό το εγχείρημα.

Επιπλέον, αναγνωρίζουμε την ουσιαστική συνδρομή των πλατφορμών ChatGPT της OpenAI και Gemini της Google, οι οποίες συνέβαλαν σημαντικά στη συγγραφή και στον γλωσσικό έλεγχο του βιβλίου, προσφέροντας υποστήριξη στην οργάνωση της ύλης, στη διατύπωση και στη διασύνδεση εννοιών με εφαρμογές. Η ενσωμάτωση αυτών των εργαλείων αποτέλεσε για εμάς έναν πολύτιμο σύμμαχο σε κάθε στάδιο της συγγραφικής διαδικασίας.

Η δομή του βιβλίου έχει σχεδιαστεί ώστε να ισορροπεί αρμονικά μεταξύ περιγραφής, σύντομης θεωρητικής ανάλυσης, επίλυσης ασκήσεων και πρακτικής εφαρμογής στον υπολογιστή με χρήση της γλώσσας R. Υπάρχουν πλέον άφθονες πηγές για εισαγωγή στην R και στο περιβάλλον RStudio, που διευκολύνουν σημαντικά την εκπαιδευτική της χρήση. Ιδιαίτερα χρήσιμο για φοιτητές των επιστημών της ζωής είναι το υλικό του RStudio Education Team, διαθέσιμο στη διεύθυνση: <https://education.rstudio.com/>

learn/, όπου παρουσιάζονται διαδραστικά μαθήματα για αρχάριους και πιο προχωρημένους χρήστες, με παιδαγωγικά άρτια δομή και πρακτική στόχευση, ιδανικά για χρήση σε πανεπιστημιακά μαθήματα.

Η προσέγγιση που ακολουθήσαμε αποσκοπεί στην ενίσχυση της κατανόησης των μαθηματικών και στατιστικών εννοιών, καθιστώντας τις πιο προσίτες και λειτουργικές για τον φοιτητή-γεωπόνο. Μέσω της σταδιακής μετάβασης από τη θεωρία στην πράξη, επιδιώκουμε την ανάπτυξη μιας στέρεης αλλά και διαισθητικής γνώσης, ικανής να αξιοποιηθεί σε ποικίλα επιστημονικά και επαγγελματικά περιβάλλοντα. Η ενσωμάτωση της R ως εργαλείου ανάλυσης και διερεύνησης επιτρέπει την άμεση εφαρμογή των εννοιών σε πραγματικά δεδομένα και προβλήματα της γεωπονικής επιστήμης. Κάθε ενότητα συνοδεύεται από παραδείγματα και ασκήσεις που έχουν επιλεγεί ώστε να ενισχύουν τη σύνδεση θεωρίας και πράξης, καλλιεργώντας παράλληλα την αναλυτική σκέψη και την ικανότητα λήψης αποφάσεων με βάση δεδομένα.

Η παρουσίαση των επιμέρους κεφαλαίων ακολουθεί μια λογική προοδευτικότητας, ξεκινώντας από τις βασικές στατιστικές έννοιες και καταλήγοντας σε πιο σύνθετες μαθηματικές δομές και μεθόδους, πάντα με γνώμονα την εφαρμοσιμότητα στην επιστήμη της γεωπονίας.

Στο **Κεφάλαιο 1** ο αναγνώστης εισάγεται στην περιγραφική στατιστική, δηλαδή στην επιστήμη της οργάνωσης, παρουσίασης και αρχικής ανάλυσης δεδομένων. Δίνεται έμφαση στους τύπους βιολογικών και γεωπονικών δεδομένων που προκύπτουν από πειράματα, μετρήσεις πεδίου ή εργαστηρίου. Παρουσιάζονται θεμελιώδη περιγραφικά μέτρα, όπως ο μέσος όρος, η διάμεσος, η διασπορά και η τυπική απόκλιση, που επιτρέπουν τη σύντομη αλλά ουσιαστική αποτύπωση της πληροφορίας. Η ενότητα εμπλουτίζεται με τεχνικές οπτικοποίησης, όπως ιστογράμματα, γραφήματα κατανομής και διαγράμματα κουτιών (box plots), τα οποία ενισχύουν τη διαισθητική κατανόηση των δεδομένων.

Το **Κεφάλαιο 2** πραγματεύεται τη συνδυαστική, έναν κλάδο των μαθηματικών που μελετά τρόπους απαρίθμησης και διάταξης αντικειμένων. Αναλύονται οι έννοιες των μεταθέσεων, διατάξεων και συνδυασμών, καθώς και ο τρόπος με τον οποίο αυτές εφαρμόζονται σε προβλήματα επιλογής, ομαδοποίησης και κατανομής πόρων. Ιδιαίτερη σημασία δίνεται στη χρήση της συνδυαστικής στον σχεδιασμό πειραμάτων – για παράδειγμα, στην επιλογή δειγμάτων, στην ομαδοποίηση μεταχειρίσεων ή στην κατασκευή πειραματικών πινάκων. Μέσω τέτοιων εφαρμογών, ο φοιτητής αντιλαμβάνεται τη σύνδεση της αφηρημένης μαθηματικής λογικής με πρακτικά ερωτήματα της έρευνας και παραγωγής.

Στο **Κεφάλαιο 3** το ενδιαφέρον μετατοπίζεται στις συναρτήσεις, οι οποίες αποτελούν βασικό εργαλείο για την περιγραφή της σχέσης μεταξύ μεταβλητών. Ξεκινώντας από την έννοια της εξάρτησης μιας ποσότητας από μια άλλη, αναλύονται γραμμικές, εκθετικές, λογαριθμικές και άλλες μορφές συναρτήσεων, με στόχο τη σύνδεσή τους με πραγματικές εφαρμογές, όπως η ανάπτυξη καλλιεργειών, η αποσύνθεση οργανικής ύλης ή η απόδοση λιπασμάτων. Η έννοια της αναπαράστασης μέσω γραφημάτων βοηθά στην οπτική κατανόηση της συμπεριφοράς των συναρτήσεων, ενώ πρακτικά παραδείγματα δείχνουν πώς αυτές χρησιμοποιούνται στη μοντελοποίηση και πρόβλεψη γεωπονικών παραμέτρων.

Το **Κεφάλαιο 4** είναι αφιερωμένο στους πίνακες και στις ιδιότητές τους. Οι πίνακες παρουσιάζονται τόσο ως μαθηματικές δομές όσο και ως πρακτικά εργαλεία για την αποθήκευση και επεξεργασία δεδομένων. Αναλύονται οι βασικές πράξεις μεταξύ πινάκων, οι ρόλοι των διανυσμάτων, η έννοια του αντιστρέψιμου πίνακα και οι ορίζουσες, με σαφείς γεωμετρικές και αλγεβρικές ερμηνείες. Εφαρμογές περιλαμβάνουν τη λύση συστημάτων εξισώσεων που σχετίζονται με γεωπονικά μοντέλα, π.χ. σε προβλήματα ισορροπίας θρεπτικών στοιχείων ή διαχείρισης πόρων.

Με το **Κεφάλαιο 5** εισερχόμαστε στον διαφορικό λογισμό και ειδικότερα στην έννοια της παραγώγου. Η παράγωγος αναλύεται ως μέτρο της στιγμιαίας μεταβολής μιας ποσότητας, με εφαρμογές που εκτείνονται από την κατανόηση του ρυθμού ανάπτυξης ενός οργανισμού μέχρι τη βελτιστοποίηση της απόδοσης μιας καλλιέργειας. Ιδιαίτερη έμφαση δίνεται στην εύρεση ακροτάτων (μέγιστα-ελάχιστα), σε μεθόδους βελτιστοποίησης (όπως η αριθμητική μέθοδος Newton-Raphson), καθώς και στην προσέγγιση συναρτήσεων μέσω του αναπτύγματος Taylor, που επιτρέπει τη γραμμικοποίηση μη γραμμικών συστημάτων για πιο εύκολη ανάλυση.

Το **Κεφάλαιο 6** πραγματεύεται τον ολοκληρωτικό λογισμό. Η παρουσίαση ξεκινά από την έννοια της αθροιστικής ποσότητας και οδηγεί στα ορισμένα και αόριστα ολοκληρώματα, παρέχοντας το μαθηματικό υπόβαθρο για την εκτίμηση εκτάσεων, όγκων ή σωρευτικών μεγεθών σε ένα πεδίο. Εξετάζονται εφαρμογές όπως η εκτίμηση της απορρόφησης νερού από το έδαφος, η κατανομή ουσιών κατά μήκος ενός αγρού ή η μεταβολή της βιομάζας στον χρόνο. Περιλαμβάνονται και μέθοδοι αριθμητικής ολοκλήρωσης, πολύτιμες όταν η αναλυτική λύση δεν είναι διαθέσιμη.

Στο **Κεφάλαιο 7** ο αναγνώστης εισάγεται στην έννοια της πιθανότητας, η οποία αποτελεί τη μαθηματική γλώσσα της αβεβαιότητας. Μελετώνται συναρτήσεις πυκνότητας και πιθανότητας, κατανομές όπως η διωνυμική, η Poisson και η κανονική, καθώς και η έννοια των τυχαίων μεταβλητών και των αναμενόμενων τιμών. Ιδιαίτερη έμφαση δίνεται στην εφαρμογή των πιθανοτήτων στη γεωπονική πειραματική ανάλυση, σε προβλέψεις αποδόσεων, σε εκτιμήσεις κινδύνου και στη διαχείριση φυσικής ή καλλιεργητικής μεταβλητότητας.

Το **Κεφάλαιο 8** εστιάζει στις διαφορικές εξισώσεις, μέσω των οποίων μοντελοποιούνται φυσικά φαινόμενα που εξελίσσονται δυναμικά στον χρόνο. Αναλύονται απλές μορφές εξισώσεων με χωριζόμενες μεταβλητές και γραμμικές εξισώσεις πρώτης τάξης, με εφαρμογές σε πληθυσμιακά μοντέλα, εξάπλωση παθογόνων, κατανάλωση θρεπτικών ή αποσύνθεση ουσιών. Η έμφαση δίνεται στην κατανόηση της μαθηματικής δομής, χωρίς να απαιτείται προηγούμενη γνώση σύνθετων τεχνικών.

Τέλος, στο **Κεφάλαιο 9** η μελέτη επεκτείνεται σε συναρτήσεις πολλών μεταβλητών. Αναλύονται έννοιες όπως η μερική παράγωγος, τα ολικά διαφορικά, η συνθήκη για ακρότατα σε περισσότερες από δύο διαστάσεις, καθώς και τα διπλά ολοκληρώματα. Μέσω πρακτικών παραδειγμάτων, ο φοιτητής μπορεί να αντιληφθεί πώς η ανάλυση σε περισσότερες διαστάσεις επιτρέπει την περιγραφή πολύπλοκων γεωργικών φαινομένων, όπως η κατανομή ενός ρύπου σε ένα αγρόκτημα ή η βελτιστοποίηση μιας πολυμεταβλητής γεωργικής απόφασης.

Με ασκήσεις στο τέλος κάθε κεφαλαίου και πληθώρα παραδειγμάτων εμπνευσμένων από την πράξη, το βιβλίο αυτό φιλοδοξεί να λειτουργήσει όχι μόνο ως εισαγωγικό εργαλείο αλλά και ως σημείο αναφοράς για τη μελέτη και εφαρμογή των μαθηματικών και της στατιστικής στις γεωπονικές επιστήμες. Απευθύνεται πρωτίστως σε φοιτητές γεωπονικών σχολών σε προπτυχιακό επίπεδο, αλλά και σε κάθε αναγνώστη που επιθυμεί να κατανοήσει πώς τα μαθηματικά εργαλεία μπορούν να στηρίξουν τη σύγχρονη επιστημονική σκέψη και πρακτική στον αγροδιατροφικό τομέα.

Η σύνδεση των μαθηματικών με τη γεωπονία δεν περιορίζεται στη χρηστική διάσταση των υπολογισμών ή των μετρήσεων· αντιθέτως, αναδεικνύει μια βαθύτερη συγγένεια ανάμεσα στην αναλυτική σκέψη και στη λειτουργία της φύσης. Πίσω από την ποικιλομορφία των βιολογικών φαινομένων, τις διακυμάνσεις του καιρού ή τις διαφορές στις αποδόσεις των καλλιεργειών, υπάρχει μια λογική δομή την οποία τα μαθηματικά βοηθούν να αποκαλυφθεί. Η χρήση εννοιών όπως η συνάρτηση, η μεταβλητότητα, η βελτιστοποίηση ή η πιθανότητα επιτρέπει στον γεωπόνο να προσεγγίζει κάθε πρόβλημα με ακρίβεια, πληρότητα και τεκμηρίωση, ακόμα και όταν τα δεδομένα είναι ατελή ή οι συνθήκες μεταβαλλόμενες.

Η αναλυτική προσέγγιση, όταν εφαρμόζεται με ευαισθησία στον γεωπονικό χώρο, δεν απομακρύνει τον επιστήμονα από την εμπειρική πραγματικότητα, αλλά αντίθετα την ερμηνεύει με τρόπο συνεπή και κατανοητό. Φέρνει στο φως πρότυπα, τάσεις και αλληλεξαρτήσεις που αλλιώς θα παρέμεναν αόρατα, προσφέροντας ταυτόχρονα εργαλεία πρόβλεψης και λήψης αποφάσεων. Στον πυρήνα αυτής της σχέσης βρίσκεται μια μορφή αρμονίας: η ισορροπία ανάμεσα στην πολυπλοκότητα του φυσικού συστήματος και στην απλότητα των μαθηματικών εργαλείων που το περιγράφουν. Έτσι, η γεωπονία, μέσα από τα μαθηματικά, αποκτά ένα μέσο έκφρασης που είναι ταυτόχρονα ακριβές, ευέλικτο και βαθιά συνδεδεμένο με τη φύση που υπηρετεί.

1

Περιγραφική στατιστική

Οι επιστήμες της ζωής είναι εγγενώς πειραματικές επιστήμες, κατά συνέπεια, για την κατανόηση ή την ερμηνεία ενός φαινομένου βρισκόμαστε αντιμέτωποι με δεδομένα τα οποία καλούμαστε να επεξεργαστούμε με ποσοτικές μεθόδους. Τα δεδομένα μπορεί να προέρχονται από πειραματικές διαδικασίες ή από την έρευνα σε βάσεις δεδομένων.

Η αφετηρία κάθε έρευνας είναι η διατύπωση ενός **ερευνητικού ερωτήματος** με τη μορφή υπόθεσης που πρέπει να ελεγχθεί. Αυτό μπορεί να γίνει είτε μέσω παρατήρησης (π.χ. της φυσικής εξέλιξης ενός φαινομένου) είτε πειραματικά (αλλάζοντας παραμέτρους για την αξιολόγηση των μεταβολών) ή θεωρητικά (αναλύοντας τις επιπτώσεις διαφόρων υποθέσεων). Καθεμιά από αυτές τις προσεγγίσεις περιλαμβάνει ποσοτικές μεθόδους, τις βάσεις για τις οποίες θα προσπαθήσουμε να θέσουμε στη συνέχεια.

Για να απαντήσουμε στο ερευνητικό ερώτημα, στη στατιστική είναι κρίσιμο να διακρίνουμε μεταξύ του **πληθυσμού** (το σύνολο των ατόμων ή παρατηρήσεων που μας ενδιαφέρει, π.χ. όλες οι ελιές σε έναν ελαιώνα) και του **δείγματος** (το υποσύνολο του πληθυσμού που πραγματικά μελετάμε, π.χ. 50 ελιές από έναν συγκεκριμένο ελαιώνα). Οι υπολογισμοί που κάνουμε στο δείγμα, όπως η δειγματική μέση τιμή (\bar{x}) και η δειγματική διακύμανση (s^2), αποτελούν εκτιμήσεις για τις πραγματικές παραμέτρους του πληθυσμού, όπως η πληθυσμιακή μέση τιμή (μ) και η διακύμανση (σ^2).

Η συλλογή των δεδομένων γίνεται μέσα από κατάλληλα σχεδιασμένες διαδικασίες, ανάλογα με τους στόχους της έρευνας. Αυτή μπορεί να περιλαμβάνει την κατάστρωση απογραφής ή, πολύ συνθηθέστερα, δειγματοληψίας, με στόχο τη λήψη αντιπροσωπευτικού του πληθυσμού δείγματος με βάση το οποίο θα προσπαθήσει ο ερευνητής να εξαγάγει συμπεράσματα για το ερευνητικό του ερώτημα.

Συνολικά, η κατανόηση των δεδομένων σε μια έρευνα είναι μια διαδικασία που αποτελείται από τα εξής βήματα:

1. Διατύπωση του ερευνητικού ερωτήματος και της στρατηγικής συλλογής δεδομένων.
2. Συλλογή των δεδομένων (δειγματοληψία) και κατασκευή της κατάλληλης βάσης δεδομένων με στόχο την εξαγωγή συμπερασμάτων σε σχέση με τους στόχους της έρευνας.
3. Σύνοψη των δεδομένων (**περιγραφική στατιστική**).
4. Ανάλυση των δεδομένων (στατιστική συμπερασματολογία και μοντελοποίηση).
5. Ερμηνεία των αποτελεσμάτων και ορθή επικοινωνία τους.

Σε αυτό το κεφάλαιο θα ασχοληθούμε με τη σύνοψη των δεδομένων, που ακολουθεί τη δειγματοληψία (συλλογή των δεδομένων), ως το πρώτο βήμα της στατιστικής ανά-

λυσος. Η δειγματοληψία αφορά τη συλλογή δεδομένων για καθορισμένο πλήθος *πειραματικών μονάδων*, ως μέρος του συνόλου για το οποίο θέλουμε να επικεντρώσουμε τη μελέτη μας.

1.1 Αδρή κατηγοριοποίηση των βιολογικών δεδομένων

Τα βιολογικά δεδομένα συλλέγονται σε *βάσεις δεδομένων* (στην απλούστερη και συνηθέστερη περίπτωση σε λογιστικά φύλλα) με τη μορφή *τυχαίων μεταβλητών*. Οι τελευταίες αφορούν χαρακτηριστικά που αποδίδονται από αριθμητικές ή αλφαριθμητικές τιμές που μπορούν να μεταβάλλονται (δεν είναι σταθερές στη διάρκεια ενός πειράματος ή διαδικασίας συλλογής δεδομένων). Οι τυχαίες μεταβλητές (τ.μ.) μπορούν να ταξινομηθούν σε τρεις μεγάλες κατηγορίες:

- Κατηγορικές τ.μ.: Αφορούν δεδομένα σε λογικές κατηγορίες *χαρακτηρισμών*, όπως χρώμα, φύλο, όνομα ποικιλίας/είδους, κατηγορία μεταχείρισης κ.λπ. Από την πλευρά των μαθηματικών, για τις τιμές των διακριτών τ.μ. δεν έχει νόημα η έννοια της απόστασης μεταξύ των τιμών αλλά ούτε και η έννοια της διάταξης αντίστοιχα. Για παράδειγμα, αν το αποτέλεσμα μιας καταγραφής πειραματικής μονάδας διακριτής τ.μ. είναι «κόκκινο», ενώ για μια άλλη καταγραφή της ίδιας τ.μ. για μια άλλη πειραματική μονάδα είναι «κίτρινο», δεν έχει νόημα να ορίσουμε την έννοια της απόστασης μεταξύ αυτών των τιμών (π.χ. η φράση «το κόκκινο απέχει 5 μονάδες από το κίτρινο» δεν έχει νόημα) αλλά ούτε και την έννοια της διάταξης (π.χ. η φράση «το κόκκινο είναι μεγαλύτερο από το κίτρινο» δεν έχει νόημα ως χαρακτηρισμός δύο πειραματικών μονάδων αποκλειστικά με βάση το συγκεκριμένο χαρακτηριστικό), τουλάχιστον για τους σκοπούς της συγκεκριμένης έρευνας¹
- Ταξινομημένες κατηγορικές τ.μ.: Αφορούν δεδομένα σε ταξινομημένες κατηγορίες *χαρακτηρισμών*, για παράδειγμα, απαντήσεις σε ερωτηματολόγιο της μορφής «καθόλου», «λίγο», «αρκετά», «πολύ». Σε τέτοιες περιπτώσεις, η έννοια της διάταξης προφανώς και έχει νόημα αλλά η έννοια της απόστασης δεν ορίζεται τυπικά (π.χ. το «καθόλου» από το «λίγο» δεν απέχει όσο το «αρκετά» από το «πολύ», αλλά σίγουρα «καθόλου» < «λίγο»).
- Μετρήσιμες τ.μ.: Οι τιμές αφορούν *μετρήσεις* σε συγκεκριμένη ποσοτική κλίμακα (συνεχή δεδομένης ακρίβειας ή διακριτή). Για παράδειγμα, για δύο μετρήσεις απόδοσης πειραματικών μονάδων (σε αυθαίρετη μονάδα μέτρησης), 35 και 42 αντίστοιχα, μπορούμε να γράψουμε $35 < 42$ και $d = 7$ (ως αποτέλεσμα της πράξης $42 - 35$).

1. Γενικά, κάποιος μπορεί να ποσοτικοποιήσει χρώματα με βάση την ένταση της απόχρωσης στη σύνθεση από τα βασικά...

Πίνακας 1.1 Ιδιότητες τυχαίων μεταβλητών

	Απόσταση	Διάταξη
Γενικές κατηγορικές	✗	✗
Ταξινομημένες κατηγορικές	✗	✓
Μετρήσιμες	✓	✓

Η παραπάνω κατηγοριοποίηση καθορίζει και τον τρόπο σύνοψης/παρουσίασης των δεδομένων μας, είτε μέσω δεικτών/μέτρων που τα χαρακτηρίζουν είτε μέσω αντιπροσωπευτικών γραφημάτων, οπότε είναι στοχευμένη σε σχέση με τις βασικές μεθόδους στατιστικής ανάλυσης και τον τύπο των δεδομένων στα οποία είναι εφαρμόσιμες. Υπάρχουν εναλλακτικές δυνατότητες κατηγοριοποίησης της φύσης των δεδομένων, αλλά αυτή που παρουσιάζεται εδώ αρκεί για τους σκοπούς επικοινωνίας με τα αντίστοιχα λογισμικά στατιστικής ανάλυσης.

Στη συνέχεια παρουσιάζονται μέτρα παρουσίασης/σύνοψης δεδομένων ανάλογα με τη φύση τους.

1.2 Σύνοψη δεδομένων με περιγραφικά στατιστικά μέτρα

Η σύνοψη δεδομένων με περιγραφικά στατιστικά μέτρα βοηθά στην κατανόηση της κατανομής και των βασικών χαρακτηριστικών ενός συνόλου δεδομένων (τυχαίου δείγματος). Τα κύρια περιγραφικά στατιστικά μέτρα χωρίζονται σε τρεις κατηγορίες:

1.2.1 Μέτρα θέσης και κεντρικής τάσης

Τα **μέτρα θέσης** είναι στατιστικοί δείκτες που περιγράφουν πού «βρίσκονται» οι τιμές μιας κατανομής. Για συνεχείς μεταβλητές, βασικά μέτρα θέσης είναι το **ελάχιστο**, το **μέγιστο** και τα **τεταρτημόρια**.

- Το **ελάχιστο** είναι η μικρότερη παρατηρούμενη τιμή στο σύνολο των δεδομένων δείχνει την κάτω ακραία τιμή της κατανομής.
- Το **μέγιστο** είναι η μεγαλύτερη παρατηρούμενη τιμή και υποδεικνύει το άνω άκρο της κατανομής.
- Τα **τεταρτημόρια** χωρίζουν τα δεδομένα σε τέσσερα ίσα μέρη:
 - Το **πρώτο τεταρτημόριο (Q_1)** είναι η τιμή κάτω από την οποία βρίσκεται το 25% των δεδομένων.
 - Το **δεύτερο τεταρτημόριο (Q_2)** είναι η διάμεσος (median), που χωρίζει το σύνολο στα δύο.
 - Το **τρίτο τεταρτημόριο (Q_3)** είναι η τιμή κάτω από την οποία βρίσκεται το 75% των δεδομένων.

Ο **τυπικός μαθηματικός ορισμός των τεταρτημορίων** βασίζεται στην κατάταξη των τιμών μιας μεταβλητής από τη μικρότερη στη μεγαλύτερη (δηλαδή στην **ταξινόμηση του δείγματος**).

Έστω ένα δείγμα δεδομένων μεγέθους n , ταξινομημένο ως:

$$x_1 \leq x_2 \leq \dots \leq x_n$$

Τα **τεταρτημόρια** είναι τιμές που χωρίζουν το δείγμα σε τέσσερα ίσα μέρη:

Πρώτο τεταρτημόριο Q_1 : Η τιμή κάτω από την οποία βρίσκεται το **25%** των παρατηρήσεων. Μαθηματικά, είναι η διάμεσος του **κάτω μισού** του δείγματος (χωρίς να περιλαμβάνεται η κεντρική τιμή αν το n είναι περιττό).

Δεύτερο τεταρτημόριο Q_2 : Είναι η **διάμεσος** του συνόλου των δεδομένων, που θα δούμε και παρακάτω. Εάν το n είναι:

περιττό: $Q_2 = x_{(n+1)/2}$

άρτιο: $Q_2 = \frac{1}{2}(x_{n/2} + x_{n/2+1})$

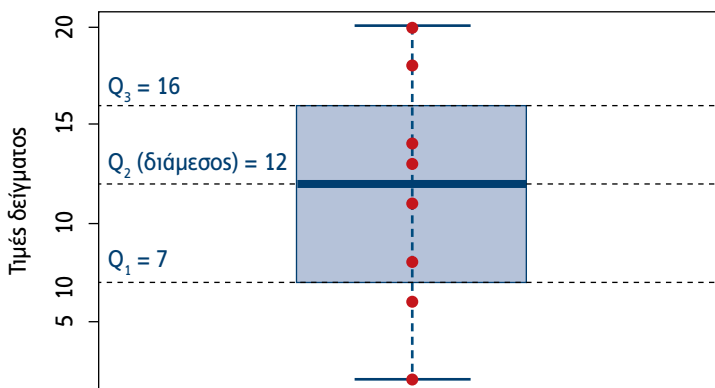
Τρίτο τεταρτημόριο Q_3 : Η τιμή κάτω από την οποία βρίσκεται το **75%** των παρατηρήσεων. Είναι η διάμεσος του **πάνω μισού** του δείγματος.

Παράδειγμα:

Για το δείγμα δεδομένων $\{8, 11, 2, 13, 6, 18, 14, 20\}$:

Ταξινομούμε τα δεδομένα σε σειρά από τη μικρότερη τιμή προς τη μεγαλύτερη $\{2, 6, 8, 11, 13, 14, 18, 20\}$. Το $n = 8$, άρα άρτιος αριθμός.

- $Q_2 = (11 + 13)/2 = 12$
- $Q_1 = (6 + 8)/2 = 7$
- $Q_3 = (14 + 18)/2 = 16$



Διάγραμμα 1.1 Θηκόγραμμα που δείχνει τα τεταρτημόρια και την κατανομή του δείγματος δεδομένων

Αυτά τα μέτρα βοηθούν στην κατανόηση της **κατανομής**, της **διασποράς** και της **συμμετρίας** των τιμών, και είναι ιδιαίτερα χρήσιμα για τον εντοπισμό ακραίων παρατηρήσεων (π.χ. μέσω boxplot).

Τα **μέτρα κεντρικής τάσης** περιγράφουν ένα «αντιπροσωπευτικό σημείο» γύρω από το οποίο συγκεντρώνονται οι τιμές μιας κατανομής. Τα βασικότερα είναι η **μέση τιμή**, η **διάμεσος** και η **επικρατούσα τιμή**.

- Η **μέση τιμή** (ή αριθμητικός μέσος) είναι το άθροισμα όλων των τιμών διαιρούμενο με το πλήθος τους· αποτελεί τον πιο διαδεδομένο δείκτη κεντρικής τάσης, αλλά είναι ευαίσθητη σε ακραίες τιμές. Η **μέση τιμή ενός πληθυσμού** συμβολίζεται συνήθως με το ελληνικό γράμμα μ , ενώ η **μέση τιμή δείγματος** συμβολίζεται με x .
- Η **διάμεσος** είναι η τιμή που βρίσκεται στο μέσο της κατανομής όταν τα δεδομένα ταξινομηθούν· χωρίζει το σύνολο σε δύο ίσα μέρη και είναι πιο ανθεκτική στην επίδραση ακραίων παρατηρήσεων.
- Η **επικρατούσα τιμή** είναι η τιμή που εμφανίζεται πιο συχνά στο δείγμα· χρησιμοποιείται κυρίως για **κατηγορικές** ή **ποσοτικές μεταβλητές με επαναλαμβανόμενες τιμές** και μπορεί να υπάρχει μία, καμία ή περισσότερες.

Η επιλογή του κατάλληλου μέτρου εξαρτάται από τον τύπο της μεταβλητής, την κατανομή των τιμών και την παρουσία ακραίων παρατηρήσεων.

- **Μέση τιμή (Mean, \bar{x}):** Το άθροισμα όλων των τιμών διαιρούμενο με το πλήθος τους. Η μέση τιμή υπολογίζεται ως:

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$

- **Διάμεσος (median, $x_{0,5}$):** Η μεσαία τιμή ενός διατεταγμένου συνόλου δεδομένων. Η διάμεσος είναι η **μεσαία τιμή** των διατεταγμένων δεδομένων, όπως αναφέρθηκε και παραπάνω ως Q_2 .

- Αν το n είναι **περιττός αριθμός**, τότε:

$$x_{0,5} = \frac{x_{n+1}}{2}$$

- Αν το n είναι **άρτιος αριθμός**, τότε:

$$x_{0,5} = \frac{\frac{x_n}{2} + \frac{x_{n+1}}{2}}{2}$$

- **Επικρατούσα τιμή (mode):** Η τιμή που εμφανίζεται συχνότερα στα δεδομένα, μπορεί να είναι περισσότερες από μία.

1.2.2 Μέτρα διασποράς (μεταβλητότητας)

Τα **μέτρα διασποράς** εκφράζουν το πόσο «απλωμένα» είναι τα δεδομένα γύρω από ένα κεντρικό σημείο (συνήθως τη μέση τιμή). Μας βοηθούν να καταλάβουμε την ομοιογένεια και τη μεταβλητότητα των τιμών μιας μεταβλητής. Τα βασικότερα μέτρα απόλυτης διασποράς είναι:

- **Εύρος (range):** Η διαφορά μεταξύ της μέγιστης και της ελάχιστης τιμής. Είναι το πιο απλό μέτρο αλλά και το πιο ευαίσθητο σε ακραίες τιμές.

$$\text{Εύρος} = X_{\max} - X_{\min}$$

- **Διακύμανση (variance, s^2):** Η μέση τιμή των τετραγωνικών αποκλίσεων των παρατηρήσεων από τη μέση τιμή τους. Η διακύμανση δείχνει τη διασπορά των τιμών γύρω από τη μέση τιμή. Μια μεγάλη διακύμανση υποδηλώνει μεγάλη διασπορά.

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

Στον τύπο της δειγματικής διακύμανσης διαιρούμε με τον όρο $n - 1$, ο οποίος ονομάζεται «βαθμοί ελευθερίας». Διαισθητικά, αυτό γίνεται γιατί, εφόσον έχουμε ήδη χρησιμοποιήσει τη μέση τιμή του δείγματος στους υπολογισμούς μας, μία τιμή θεωρείται «δεσμευμένη». Αυτή η διόρθωση με το $n - 1$ μας δίνει μια πιο ακριβή εκτίμηση της πραγματικής διακύμανσης του πληθυσμού.

- **Τυπική απόκλιση (standard deviation, s):** Η τετραγωνική ρίζα της διακύμανσης. Είναι το πιο συνηθισμένο μέτρο διασποράς, καθώς εκφράζεται στις ίδιες μονάδες με τα αρχικά δεδομένα.

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

- **Ενδοτεταρτημοριακό εύρος (interquartile range, IQR):** Το IQR είναι η διαφορά μεταξύ του 3ου τεταρτημορίου (Q_3) και του 1ου τεταρτημορίου (Q_1), η οποία μετρά το εύρος των «μεσαίων» 50% των τιμών. Δεν επηρεάζεται από ακραίες τιμές:

$$IQR = Q_3 - Q_1$$

Τα παραπάνω μέτρα είναι μέτρα **απόλυτης διασποράς**. Αυτό καθιστά δύσκολη τη σύγκριση της μεταβλητότητας μεταξύ δύο διαφορετικών συνόλων δεδομένων (π.χ. η σύγκριση της διασποράς στην απόδοση μιας καλλιέργειας (σε κιλά/στρέμμα) με τη διασπορά στην περιεκτικότητα σε πρωτεΐνη (%) είναι προβληματική, καθώς οι μονάδες και οι μέσες τιμές τους είναι εντελώς διαφορετικές).

Για να επιλύσουμε αυτό το πρόβλημα, χρησιμοποιούμε τον **συντελεστή μεταβλητότητας (CV)**, ο οποίος είναι ένα **σχετικό μέτρο διασποράς**. Εκφράζει την τυπική απόκλιση ως ποσοστό της μέσης τιμής, καθιστώντας τον έναν «καθαρό» αριθμό χωρίς μονάδες, ιδανικό για συγκρίσεις.

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

Γενικά, τιμές $CV < 15\%$ θεωρούνται χαμηλές (μικρή μεταβλητότητα), ενώ τιμές $CV > 15\%$ θεωρούνται υψηλές (μεγάλη μεταβλητότητα).

Παράδειγμα:

Θα εξετάσουμε τις αποδόσεις μιας καλλιέργειας σιταριού σε 15 διαφορετικά αγροτεμάχια, μετρημένες σε κιλά ανά στρέμμα.

Δεδομένα (απόδοση σε κιλά/στρέμμα):

[420, 500, 550, 480, 600, 520, 490, 510, 530, 580, 495, 515, 505, 560, 540]

Θα υπολογίσουμε τα περιγραφικά στατιστικά μέτρα για αυτά τα δεδομένα.

Ακολουθούν αναλυτικά **οι πράξεις για τον υπολογισμό των περιγραφικών στατιστικών μέτρων** που δόθηκαν, με βάση τα δεδομένα απόδοσης σε κιλά/στρέμμα:

1. Μέσος όρος (mean):

$$\bar{x} = \frac{420 + 500 + 550 + 480 + 600 + 520 + 490 + 510 + 530 + 580 + 495 + 515 + 505 + 560 + 540}{15}$$

$$\bar{x} = \frac{7795}{15} = 519.67 \text{ κιλά}$$

2. Διάμεσος (median):

Ταξινομούμε τα δεδομένα:

[420, 480, 490, 495, 500, 505, 510, 515, 520, 530, 540, 550, 560, 580, 600]

Το $n = 15$, άρα περιττός αριθμός. Η διάμεσος είναι η μεσαία τιμή (θέση 8 για 15 αριθμούς):

$$\text{Διάμεσος} = 515$$

3. Επικρατούσα τιμή (mode):

Ψάχνουμε τη συχνότερη τιμή. Όλες εμφανίζονται **μία φορά**, άρα **τεχνικά δεν υπάρχει επικρατούσα τιμή**.

4. Εύρος (range):

$$\text{Εύρος} = \max(x) - \min(x) = 600 - 420 = 180 \text{ κιλά}$$

5. Διακύμανση (variance, s^2):

Υπολογισμός διαφορών από τον μέσο όρο:

$$(420 - 519.67)^2 = 9913.79$$

$$(500 - 519.67)^2 = 387.11$$

$$(550 - 519.67)^2 = 912.11$$

$$(480 - 519.67)^2 = 1583.11$$

$$(600 - 519.67)^2 = 6453.44$$

$$(520 - 519.67)^2 = 0.11$$

$$(490 - 519.67)^2 = 887.11$$

$$(510 - 519.67)^2 = 93.44$$

$$(530 - 519.67)^2 = 106.78$$

$$(580 - 519.67)^2 = 3627.11$$

$$(495 - 519.67)^2 = 610.78$$

$$(515 - 519.67)^2 = 21.78$$

$$(505 - 519.67)^2 = 216.78$$

$$(560 - 519.67)^2 = 1620.11$$

$$(540 - 519.67)^2 = 412.11$$

Άθροισμα όλων των παραπάνω:

$$\sum_{i=1}^{15} (x_i - \bar{x})^2 \approx 26,873.67$$

Συνεπώς:

$$s^2 = \frac{26873.67}{15 - 1} \approx 1919.52$$

6. Τυπική απόκλιση (standard deviation):

$$s = \sqrt{s^2} = \sqrt{1919.52} \approx 43.81 \text{ κιλά}$$

7. Ενδοτεταρτημοριακό εύρος (IQR):

Χρειάζεται να υπολογιστούν το 1ο (Q1) και 3ο (Q3) τεταρτημόριο:

Ταξινομημένα δεδομένα (ξανά):

$$[420, 480, 490, 495, 500, 505, 510, 515, 520, 530, 540, 550, 560, 580, 600]$$

- Q_1 = διάμεσος του κάτω μισού (πρώτα 7): 420, 480, 490, 495, 500, 505, 510 → $Q_1 = 495$
- Q_3 = διάμεσος του άνω μισού (τελευταία 7): 520, 530, 540, 550, 560, 580, 600 → $Q_3 = 550$

$$IQR = Q_3 - Q_1 = 550 - 495 = 55 \text{ κιλά}$$

8. Συντελεστής μεταβλητότητας (CV):

$$CV = \frac{43.81}{519.67} \cdot 100\% \approx 8.43$$

Η τιμή του CV (8.43%) είναι πολύ χαμηλή (<15%), επομένως συμπεραίνουμε ότι οι αποδόσεις των χωραφιών στην περιοχή είναι αρκετά ομοιογενείς και παρουσιάζουν χαμηλή μεταβλητότητα.

Τελική σύνοψη:

Μέτρο	Τιμή
Μέσος όρος	519.67 κιλά
Διάμεσος	515.00 κιλά
Επικρατούσα τιμή	(καμία)
Εύρος	180 κιλά
Διακύμανση	1919.52
Τυπική απόκλιση	43.81 κιλά
Ενδοτεταρτημοριακό εύρος	47.5 κιλά
Συντελεστής μεταβλητότητας	8.43%

Κώδικας σε R:

```
# Δημιουργία δεδομένων (Απόδοση σε κιλά/στρέμμα)
data <- c(420, 500, 550, 480, 600, 520, 490, 510, 530, 580,
495, 515, 505, 560, 540)

# Μέτρα κεντρικής τάσης
mean_value <- mean(data) # Μέσος όρος
median_value <- median(data) # Διάμεσος
mode_value <- as.numeric(names(sort(table(data),
decreasing=TRUE)[1])) # Επικρατούσα τιμή (αν δεν υπάρχει επι-
κρατούσα τιμή θα εμφανιστεί η πρώτη τιμή της σειράς)

# Μέτρα διασποράς
range_value <- max(data) - min(data) # Εύρος
variance_value <- var(data) # Διακύμανση
std_dev_value <- sd(data) # Τυπική απόκλιση
iqr_value <- IQR(data) # Ενδοτεταρτημοριακό εύρος
cv_value <- (std_dev_value / mean_value) * 100 # Υπολογισμός CV

# Εμφάνιση αποτελεσμάτων
stats_list <- list(
  "Μέσος όρος" = mean_value,
  "Διάμεσος" = median_value,
  "Επικρατούσα τιμή" = mode_value,
  "Εύρος" = range_value,
  "Διακύμανση" = variance_value,
  "Τυπική απόκλιση" = std_dev_value,
  "Ενδοτεταρτημοριακό εύρος" = iqr_value
  "Συντελεστής μεταβλητότητας (%)" = cv_value
)

print(stats_list)
```

1.2.3 Περιγραφικά στατιστικά μέτρα για κατηγορικές μεταβλητές

Τα **περιγραφικά στατιστικά μέτρα για κατηγορικές μεταβλητές** (δηλαδή μεταβλητές που παίρνουν διακριτές κατηγορίες ως τιμές, όπως το φύλο, η εκπαίδευση, η προτίμηση προϊόντος κ.λπ.) εστιάζουν σε διαφορετικά στατιστικά μέτρα από αυτά που χρησιμοποιούμε για αριθμητικά δεδομένα. Η ανάλυση βασίζεται σε δύο πυλώνες: την καταμέτρηση των συχνοτήτων και τη μέτρηση της ποικιλότητας.

Βασικά μέτρα: Συχνότητες (frequencies) και επικρατούσα τιμή

Το πρώτο βήμα είναι πάντα η καταμέτρηση της εμφάνισης κάθε κατηγορίας.

- **Απόλυτες συχνότητες (absolute frequencies):** Ο αριθμός των παρατηρήσεων σε κάθε κατηγορία.
- **Σχετικές συχνότητες (relative frequencies):** Το ποσοστό των παρατηρήσεων σε κάθε κατηγορία, που μας επιτρέπει να κάνουμε συγκρίσεις..
- **Αθροιστικές συχνότητες (cumulative frequencies):** Το άθροισμα των συχνοτήτων έως μια συγκεκριμένη κατηγορία (χρησιμοποιείται κυρίως για διατεταγμένες κατηγορίες, π.χ. «καλό», «μέτριο», «κακό»).

Το αντίστοιχο μέτρο κεντρικής τάσης είναι:

- **Επικρατούσα τιμή (mode):** Είναι η κατηγορία με τη μεγαλύτερη συχνότητα εμφάνισης. Ένα σύνολο δεδομένων μπορεί να έχει μία (unimodal), δύο (bimodal) ή περισσότερες επικρατούσες τιμές.

Προχωρημένα μέτρα: Ποικιλότητα, ομοιομορφία και αβεβαιότητα

Πέρα από την απλή καταμέτρηση, συχνά θέλουμε να απαντήσουμε σε πιο σύνθετες ερωτήσεις: Πόσο «πλούσια» σε κατηγορίες είναι τα δεδομένα μας; Μία κατηγορία κυριαρχεί ή η κατανομή είναι ισορροπημένη; Για να ποσοτικοποιήσουμε αυτές τις ιδιότητες, χρησιμοποιούμε τους παρακάτω δείκτες, οι οποίοι είναι ιδιαίτερα διαδεδομένοι σε επιστήμες όπως η οικολογία, η αγρονομία και η ανάλυση δεδομένων.

- **Δείκτης ποικιλότητας (diversity index):** Οι δείκτες ποικιλότητας ποσοτικοποιούν το **πλήθος των κατηγοριών (π.χ. είδη)** και την **κατανομή τους**. Δύο βασικοί δείκτες είναι ο **δείκτης του Shannon** και ο **δείκτης του Simpson**, και δείχνουν πόσο ποικίλες είναι οι κατηγορίες.
- **Δείκτης Shannon (H):** Ο δείκτης αυτός μετρά τη συνολική **ποικιλότητα**, λαμβάνοντας υπόψη τόσο το πλήθος των κατηγοριών όσο και την ισορροπία στην κατανομή τους.

$$H = -\sum_{i=1}^S p_i \ln(p_i)$$

όπου S είναι το πλήθος των κατηγοριών και p_i το σχετικό ποσοστό της κατηγορίας i .

Ερμηνεία: Μια υψηλή τιμή δείχνει μεγάλη ποικιλότητα και αβεβαιότητα (π.χ. πολλά είδη φυτών σε ισορροπημένες ποσότητες).

- **Δείκτης Simpson (D):** Ο δείκτης αυτός εστιάζει περισσότερο στην κυριαρχία. Συγκεκριμένα, ο τύπος D υπολογίζει την πιθανότητα δύο στοιχεία που επιλέγονται τυχαία να ανήκουν στην ίδια κατηγορία. Για τον λόγο αυτόν, ως δείκτης ποικιλότητας χρησιμοποιείται συνήθως η μορφή $1 - D$.

$$D = \sum_{i=1}^S p_i^2$$

όπου S είναι το πλήθος των κατηγοριών και p_i το σχετικό ποσοστό της κατηγορίας i . Ο δείκτης ποικιλότητας είναι $1 - D$.

Ερμηνεία: Μια τιμή $1 - D$ κοντά στο 1 σημαίνει ότι είναι πολύ πιθανό να επιλέξουμε δύο διαφορετικά στοιχεία, άρα έχουμε μεγάλη ποικιλότητα.

Χρησιμότητα: Οι δείκτες ποικιλότητας μετρούν τη **βιολογική ποικιλότητα** (π.χ. φυτικά είδη σε αγρό), την **πολυμορφία παραγόμενων προϊόντων** ή την **ποικιλία καταναλωτικών επιλογών**.

- **Συντελεστής ομοιομορφίας (evenness coefficient, E):** Ο δείκτης αυτός απαντά στην ερώτηση: «Δεδομένης της ποικιλότητας που έχουμε, πόσο ισορροπημένη είναι η κατανομή;». Συγκρίνει την πραγματική ποικιλότητα (Shannon) με τη μέγιστη δυνατή που θα μπορούσε να υπάρξει με αυτό το πλήθος κατηγοριών.

$$E = \frac{H}{\ln(S)}$$

όπου H είναι ο δείκτης Shannon και S το πλήθος κατηγοριών.

Ερμηνεία: Μια τιμή κοντά στο 1 σημαίνει ότι οι κατηγορίες είναι σχεδόν ισοκατανεμημένες. Μια χαμηλή τιμή σημαίνει ότι, παρότι μπορεί να υπάρχουν πολλές κατηγορίες, μία ή μερικές εξ αυτών κυριαρχούν αριθμητικά.

Χρησιμότητα: Δείχνει αν η παρουσία ποικιλίας συνοδεύεται και από **ισορροπημένη κατανομή** (π.χ. αν όλα τα είδη είναι εξίσου συχνά ή κυριαρχεί κάποιο).

- **Εντροπία (entropy):** Δείχνει την αβεβαιότητα ή την ανομοιογένεια της κατανομής. Ο μαθηματικός της τύπος είναι πρακτικά ταυτόσημος με αυτόν του δείκτη ποικιλότητας Shannon.

$$H = - \sum_{i=1}^S p_i \log(p_i)$$

Ερμηνεία: Όσο μεγαλύτερη η εντροπία, τόσο πιο **τυχαία** ή **μη προβλέψιμη** η κατανομή των κατηγοριών.

Η σχέση της εντροπίας με τον δείκτη ποικιλότητας Shannon είναι άμεση, με τη διαφορά τους να έγκειται στη βάση του λογαρίθμου που χρησιμοποιείται. Στην πράξη, αυτό σημαίνει ότι η εντροπία, που στη Θεωρία της Πληροφορίας μετρείται σε bits (\log_2), μπορεί να υπολογιστεί απευθείας από τον δείκτη Shannon, ο οποίος στην οικολογία χρησιμοποιεί τον φυσικό λογάριθμο (\ln), απλώς διαιρώντας την τιμή του Shannon με τη σταθερά $\ln(2)$.

Χρησιμότητα: Χρησιμοποιείται για **ταξινομήσεις, λήψη απόφασης και εκτίμηση διασποράς πληθυσμών**, καθώς και σε **μηχανική μάθηση** για τον υπολογισμό πληροφορίας σε decision trees.

Συγκριτική επισκόπηση και ερμηνεία

Δείκτης	Εύρος τιμών	Τι εκφράζει	Ερμηνεία τιμών
Shannon (H)	0 έως $\ln(S)$	Συνολική ποικιλότητα (ποσότητα + ισορροπία)	– Κοντά στο 0: κυριαρχεί 1 κατηγορία – Μέγιστο: ισοκατανομή
Simpson ($1 - D$)	0 έως $1 - \frac{1}{S}$	Πιθανότητα δύο τυχαίες τιμές να διαφέρουν	– Κοντά στο 0: χαμηλή ποικιλότητα – Κοντά στο 1: μεγάλη ποικιλία
Evenness (E)	0 έως 1	Πόσο ισοκατανομημένες είναι οι παρατηρήσεις	– Κοντά στο 1: ισοκατανομή – Χαμηλό: κυριαρχία κάποιου είδους
Entropy (\log_2)	0 έως $\log_2(S)$	Πληροφορία/αβεβαιότητα για την κατηγορία	– 0: απόλυτη βεβαιότητα (μία μόνο τιμή) – Υψηλό: μέγιστη αβεβαιότητα

Σύνοψη:

- **Οι δείκτες Shannon και Entropy** μετρούν την **ποικιλία και την αβεβαιότητα**: όσο πιο ισομερώς κατανομημένες οι τιμές, τόσο μεγαλύτερες οι τιμές.
- **Ο δείκτης Simpson** εστιάζει περισσότερο στην **κυριαρχία** μίας ή περισσότερων κατηγοριών.
- **Ο δείκτης Evenness** δείχνει **πόσο «ισόρροπη»** είναι η κατανομή.

Για να κατανοήσουμε τη δομή των κατηγορικών δεδομένων μας, ξεκινάμε με τις συχνότητες. Στη συνέχεια, οι δείκτες **Shannon** και **Simpson** μας δίνουν μια εικόνα της συνολικής ποικιλότητας, ενώ ο δείκτης **Evenness** εξειδικεύει στο πόσο ισορροπημένη είναι αυτή η ποικιλότητα.

Παράδειγμα: Είδη φυτών και αριθμός παρατηρήσεων

Για να κατανοήσουμε πώς εφαρμόζονται οι παραπάνω δείκτες στην πράξη, ας εξετάσουμε τα δεδομένα ενός γεωπόνου που κατέγραψε την κατανομή 100 φυτών σε ένα αγροτεμάχιο.

Αρχικά δεδομένα και σχετικές συχνότητες

Το πρώτο βήμα είναι να μετατρέψουμε τις απόλυτες παρατηρήσεις σε σχετικές συχνότητες (p_i), καθώς όλοι οι τύποι των δεικτών βασίζονται σε αυτές.

Για τον υπολογισμό των σχετικών συχνοτήτων χρησιμοποιούμε τον τύπο:

$$p_i = \frac{n_i}{\sum n_i}$$

Είδος	Παρατηρήσεις n_i	Σχετική συχνότητα p_i
Στάρι	40	0.40
Βρώμη	30	0.30
Κριθάρι	20	0.20
Τριφύλλι	10	0.10
Σύνολο	100	1.00

Υπολογισμός των δεικτών

Τώρα θα υπολογίσουμε τους δείκτες ποικιλότητας και ομοιομορφίας χρησιμοποιώντας τις σχετικές συχνότητες (p_i).

Δείκτης Shannon (H')

$$H = -\sum_{i=1}^4 p_i \ln(p_i)$$

$$-(0.40 \cdot \ln 0.40) = -(0.40 \cdot -0.91629) = 0.36652$$

$$-(0.30 \cdot \ln 0.30) = -(0.30 \cdot -1.20397) = 0.36119$$

$$-(0.20 \cdot \ln 0.20) = -(0.20 \cdot -1.60944) = 0.32189$$

$$-(0.10 \cdot \ln 0.10) = -(0.10 \cdot -2.30259) = 0.23026$$

$$H' = 0.36652 + 0.36119 + 0.32189 + 0.23026 = \boxed{1.27986}$$

Δείκτης Simpson (D και 1 - D)

$$D = \sum_{i=1}^4 p_i^2$$

$$D = (0.40)^2 + (0.30)^2 + (0.20)^2 + (0.10)^2 = 0.16 + 0.09 + 0.04 + 0.01 = \boxed{0.30}$$

$$1 - D = 1 - 0.30 = \boxed{0.70}$$

Συντελεστής ομοιομορφίας (evenness coefficient, E):

$$E = \frac{H}{\ln(4)}$$

$$\ln(4) = 1.38629 \Rightarrow E = \frac{1.27986}{1.38629} \approx \boxed{1.846}$$

Εντροπία (H):

Εφόσον χρησιμοποιούμε τα **ίδια** p_i , η **εντροπία σε \log_2** είναι **ίση με τον δείκτη Shannon διαιρεμένο με $\ln(2)$** :

$$H = \frac{H'}{\ln(2)} \approx \frac{1.27986}{0.6931} \approx \boxed{1.846}$$

Αφού κάναμε τους υπολογισμούς, το σημαντικότερο βήμα είναι να ερμηνεύσουμε τι σημαίνουν αυτοί οι αριθμοί για το αγροτεμάχιό μας.

Δείκτης	Τιμή
Shannon (H')	1.280
Simpson Diversity (1 - D)	0.700
Evenness (E)	0.923
Εντροπία	1.846

Κώδικας σε R:

```
# Δεδομένα
species <- c("Σιτάρι", "Βρώμη", "Κριθάρι", "Τριφύλλι")
counts <- c(40, 30, 20, 10)

# Αναλογίες (p_i)
p <- counts / sum(counts)

# Shannon Index
H <- -sum(p * log(p)) # ln βάση
H_log2 <- -sum(p * log2(p)) # log2 βάση (εντροπία)

# Simpson Index
D_simpson <- sum(p^2)
Simpson_diversity <- 1 - D_simpson

# Evenness
S <- length(species)
Evenness <- H / log(S)

# Εμφάνιση αποτελεσμάτων
cat("Δείκτης Shannon (H):", round(H, 3), "\n")
cat("Δείκτης Simpson (1 - D):", round(Simpson_diversity, 3), "\n")
cat("Συντελεστής Ομοιομορφίας (Evenness):", round(Evenness, 3), "\n")
cat("Εντροπία (log2 βάση):", round(H_log2, 3), "\n")
```

Ερμηνεία αποτελεσμάτων

Τα αποτελέσματα δείχνουν μια υγιή ποικιλότητα στο αγροτεμάχιο.

- **Shannon H** ≈ 1.28 \rightarrow Υψηλή ποικιλότητα. Η τιμή αυτή υποδεικνύει ότι στο αγροτεμάχιο δεν κυριαρχεί ένα μόνο είδος, αλλά συνυπάρχουν πολλαπλά είδη σε υγιείς, σημαντικές ποσότητες.
- **Simpson (1 - D)** ≈ 0.70 \rightarrow Υψηλή πιθανότητα διαφορετικότητας. Υπάρχει 70% πιθανότητα δύο φυτά που θα επιλέξουμε τυχαία από το χωράφι να είναι διαφορετικού είδους. Η υψηλή αυτή πιθανότητα επιβεβαιώνει την ύπαρξη μεγάλης ποικιλίας.
- **Evenness** ≈ 0.92 \rightarrow Καλή σχετική ισορροπία μεταξύ των ειδών. Μια τιμή τόσο κοντά στο 1 (που είναι η τέλεια ισοκατανομή) αποδεικνύει ότι οι ποσότητες των διαφορετικών ειδών είναι πολύ ισορροπημένες.
- **Εντροπία** ≈ 1.85 \rightarrow Υπάρχει αβεβαιότητα στην τυχαία επιλογή είδους (δεν είναι μονοκαλλιέργεια). Το αποτέλεσμα είναι αρκετά μακριά από το 0, δείχνοντας ότι η επιλογή ενός φυτού στην τύχη δεν είναι εύκολα προβλέψιμη. Αυτή η υψηλή αβεβαιότητα είναι άμεση συνέπεια της μεγάλης ποικιλότητας και της καλής ισορροπίας που είδαμε παραπάνω.

1.3 Απεικόνιση δεδομένων μετρήσεων σε συνεχή κλίμακα με γραφήματα

Η απεικόνιση δεδομένων αποτελεί κρίσιμο βήμα στην αρχική διερεύνηση κάθε συνόλου δεδομένων. Μέσω κατάλληλων γραφημάτων μπορούμε να αποτυπώσουμε με οπτικό τρόπο βασικά χαρακτηριστικά, όπως η κατανομή, οι ακραίες τιμές, η παρουσία ασυμμετρίας, αλλά και η πιθανή συσχέτιση μεταξύ μεταβλητών. Τα γραφήματα δεν είναι απλώς βοηθήματα παρουσίασης· αποτελούν εργαλεία ερμηνείας, επιτρέποντας στον ερευνητή να διατυπώσει υποθέσεις, να εντοπίσει μοτίβα ή σφάλματα και να αξιολογήσει την ποιότητα των δεδομένων του.

Για την οπτική ανάλυση συνεχών μεταβλητών, υπάρχουν τρία βασικά γραφήματα.

Κατανομή μίας μεταβλητής

Όταν θέλουμε να εξετάσουμε μία μόνο μεταβλητή, τα κύρια εργαλεία μας είναι:

Το **ιστόγραμμα (histogram)**, το οποίο είναι ένα από τα βασικότερα εργαλεία. Παρουσιάζει τη συχνότητα εμφάνισης των τιμών, χωρίζοντας το εύρος τους σε διαδοχικά διαστήματα (κλάσεις). Μας βοηθά να δούμε άμεσα τη μορφή της κατανομής (π.χ. αν είναι συμμετρική, κωδωνοειδής ή ασύμμετρη).

Για να φτιάξουμε τις κλάσεις ενός ιστογράμματος, ομαδοποιούμε τις τιμές σε διαδοχικά διαστήματα ίσου μεγέθους. Η λογική είναι απλή:

1. Βρίσκουμε το εύρος των δεδομένων (δηλαδή τη διαφορά μεταξύ της μεγαλύτερης και της μικρότερης τιμής).

2. Αποφασίζουμε σε πόσες κλάσεις (διαστήματα) θέλουμε να χωρίσουμε αυτό το εύρος. Ένας καλός γενικός κανόνας είναι να επιλέγουμε έναν αριθμό μεταξύ 5 και 15. Στόχος είναι να έχουμε αρκετές κλάσεις για να δούμε τη μορφή της κατανομής, αλλά όχι τόσο πολλές ώστε το γράφημα να γίνεται περίπλοκο.
3. Το πλάτος κάθε κλάσης (και άρα κάθε ράβδου στο γράφημα) προκύπτει από τη διαίρεση του εύρους με τον αριθμό των κλάσεων που επιλέξαμε.

Σκοπός είναι το τελικό γράφημα να αποκαλύπτει καθαρά τη βασική μορφή της κατανομής, χωρίς να είναι ούτε υπερβολικά απλό ούτε υπερβολικά πολύπλοκο. Τα στατιστικά λογισμικά συνήθως προτείνουν μια αυτόματη ομαδοποίηση, αλλά η κατανόηση αυτής της λογικής είναι το κλειδί για τη σωστή ερμηνεία και προσαρμογή των ιστογραμμάτων.

Το **θηκόγραμμα (boxplot)** είναι ιδιαίτερα χρήσιμο για την απεικόνιση της θέσης, της διασποράς και της ύπαρξης ακραίων τιμών. Παρουσιάζει γραφικά τα τεταρτημόρια, τη διάμεσο και τις πιθανές ακραίες/έκτροπες παρατηρήσεις (outliers), προσφέροντας μια συνοπτική και άμεση απεικόνιση της κατανομής.

Διερεύνηση σχέσης μεταξύ δύο μεταβλητών

Όταν θέλουμε να διερευνήσουμε την πιθανή σχέση μεταξύ δύο ποσοτικών μεταβλητών, χρησιμοποιούμε το **σημειόγραμμα (scatter plot)**. Κάθε σημείο στο γράφημα αναπαριστά ένα ζεύγος τιμών (x, y) , επιτρέποντάς μας να ανιχνεύσουμε οπτικά γραμμικές ή άλλου είδους συσχετίσεις.

Η οπτική διερεύνηση είναι το πρώτο βήμα για κάθε στατιστική ανάλυση· λειτουργεί ως γέφυρα μεταξύ των δεδομένων και της ερμηνείας τους.

Πρακτική εφαρμογή

Τυπικά καταφεύγουμε σε λογισμικά για την κατασκευή των γραφημάτων. Για να δούμε πώς λειτουργούν αυτά τα εργαλεία, ας υποθέσουμε ότι ένας γεωπόνος καταγράφει για 20 αγροτεμάχια την ποσότητα λίπανσης που εφάρμοσε και την τελική απόδοση της καλλιέργειας, με σκοπό να διερευνήσει οπτικά τόσο την κατανομή της απόδοσης όσο και την πιθανή σχέση μεταξύ των δύο μεταβλητών.

Κώδικας σε R:

Παρακάτω παρατίθεται ο κώδικας σε R για την εισαγωγή των δεδομένων, με τον οποίο μπορούμε να κατασκευάσουμε το ιστόγραμμα και το θηκόγραμμα που αναφέρθηκαν παραπάνω.

```
# Φόρτωση βιβλιοθηκών
library(ggplot2)

# Δεδομένα
# Απόδοση καλλιέργειας σε κιλά/στρέμμα
Yield <- c(450, 460, 470, 480, 490, 495, 500, 505, 510, 515,
520, 525, 530, 540, 550, 560, 570, 580, 590, 600)
```