AL-FARABI KAZAKH NATIONAL UNIVERSITY
FACULTY OF BIOLOGY AND BIOTECHNOLOGY
DEPARTMENT OF BIOTECHNOLOGY

Lecture 7

STATISTICAL METHODS IN BIOLOGY

PhD, Acting
Associate Professor,
Doktyrbay Gulina

Lecture Aim

To develop students' understanding of statistical methods used in biological research and their application in data analysis, interpretation, and scientific decision-making.

Expected Learning Outcomes

By the end of this lecture, students will be able to:

1. Explain the basic concepts of biostatistics.
2. Distinguish descriptive and inferential statistics.
3. Apply appropriate statistical tests to biological data.
4. Interpret statistical results, including p-values and confidence intervals.
5. Present biological data using tables, graphs, and statistical summaries

Agenda & Structure

1. Introduction & Data Types · Slides 3-5
2. Sampling & Preprocessing · Slides 6-8
3. Descriptive Statistics & Visualization · Slides 9-12
4. Probability & Hypothesis Testing · Slides 13-20
5. Advanced Statistical Methods & Practical Applications · Slides 21-30

Why Biology Needs Biostatistics

Natural Variation

Living systems exhibit inherent variation – genetic, environmental, and physiological – making every biological measurement unique.

Signal Detection

Biostatistics provides rigorous tools to separate true biological signals from random variation and sampling er ror.

Biological Noise

Unlike physics, biological data always contains noise from measurement error, individual differences, and environmental factors.

Understanding Data Structures

Quantitative Continuous

Measured values thatcan takeany value within a range. Examples: enzyme reaction rates, blood pressure, body temperature, and oxygen saturation levels in biological specimens.

Quantitative Discrete

Counted values that take only whole numbers. Examples: cell counts per field of view, number of offspring, mutation events, or colonies on an agar plate.

Qualitative Nominal

Categories with noinherent order or ranking. Examples: blood types (A, B, AB, O), species names, genotype labels, and tissue types in histological studies.

Qualitative Ordinal

Ranked categorieswithout equal intervals between levels. Examples: disease severity stages (mild, moderate, severe), pain scales, and ecological abundance classes.

Measurement Scales

Concept / test	Meaning and application
Nominal	Categories without order or ranking. No mathematical operations apply. Examples: blood types (A, B, AB, O), plant species, sex (male/female). Only equality or difference can be determined.
Ordinal	Ranked categories with meaningful order, but intervals between ranks are unequal. Examples: disease severity (mild, moderate, severe), pain scales. Direction of difference is known; magnitude is not.
Interval	Ordered scale with equal intervals between values, but no true zero point. Ratios are meaningless. Example: temperature in °Celsius. Addition and subtraction are valid; multiplication is not.
Ratio	Interval scale with an absolute zero, enabling ratio comparisons. All mathematical operations are valid. Examples: glucose concentration (mg/dL), body mass, enzyme activity. Most powerful scale.

Inferring Truth From Subsets

Biological Populations

Entire populations (e.g., all individuals of a species) are too large to measure completely; full census is rarely feasible.

Parameter Estimation

Statistical inference uses sample statistics ($\bar{x}$, s) to estimate true population parameters (μ, σ) with quantified uncertainty.

Representative Samples

A sample is a carefully selected subset drawn from the population to represent its diversity and characteristics.

Avoiding Sampling Bias

Simple Random Sampling

Every individualin the population has an equal and independent probability of being selected. Minimises selection bias and supports valid statistical inference.

Stratified Sampling

Populationis divided into distinct subgroups (strata) based on a relevant characteristic. Random samples are drawn from each stratum to ensure proportional representation.

Systematic Sampling

Individuals are selected at fixed intervals (every k-th unit) from an ordered list or spatial arrangement. Efficient and practical for large or spatially distributed populations.

Transect Sampling

Aline or belt transectis laid across a habitat at defined intervals. Organisms or features intersecting the transect are recorded – widely used in ecological field surveys.

Data Preprocessing

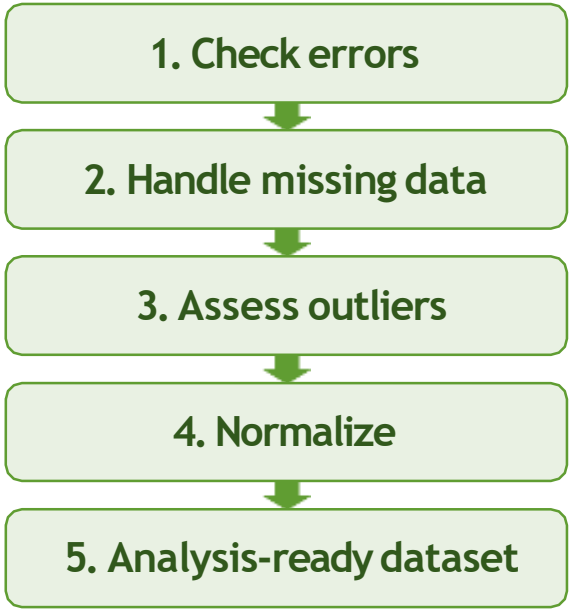
Error Checking
Identify data entry errors, impossible values, and format inconsistencies in raw biological data

Analysis-Ready Dataset
Clean, validated data meeting assumptions for downstream statistical tests

Missing Data Handling
Apply imputation, exclusion, or flagging strategies for incomplete observations Use IQR rules or Z-scores to identify and assess extreme biological measurements

Normalization
Rescale variables to comparable ranges; log- transform skewed distributions

Outlier Detection



Measures Of Central Tendency

Mean	Sum of all values divided by n; most common measure but sensitive to extreme outliers.
Mode	Most frequently occurring value; useful for nominal and bimodal biological data.
Median	Middle value when data are ranked; robust to skewed distributions and outliers.

Quantifying Biological Variation

Variance (s^2)	Average squared deviation from the mean; emphasises large deviations due to squaring
IQR	Range between 25th and 75th percentiles; robust to outliers and skewed distributions
Standard Deviation (s)	Square root of variance; expressed in the same units as the original data

Estimating Precision Of Population Metrics

Standard Error (SEM)	SEM = s / √n. Quantifies how precisely the sample mean estimates the true population mean; smaller n yields larger SEM.
SD vs. SEM Error Bars	SD describes data spread around the mean. SEM describes precision of the mean estimate. Use SEM for inference; SD for variability.
95% Confidence Interval	95% CI: $\bar{x} \pm 1.96 \times SEM$. The interval expected to contain the true population parameter in 95% of repeated samples.

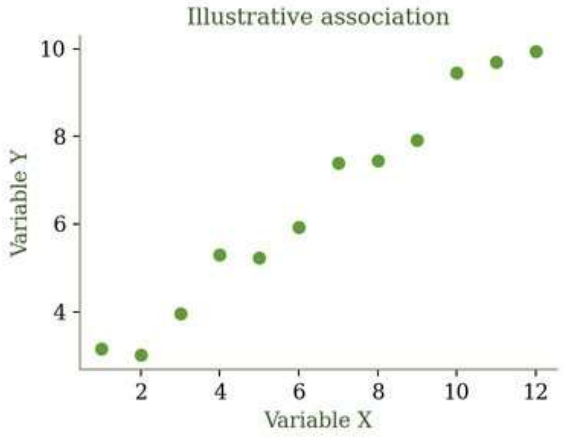
Visualizing Complex Data Distributions

Histogram
Displays frequency distribution of continuous variables. Used to assess shape, spread, and normality of biological data (e.g., body mass, enzyme activity levels).

Scatter Plot
Visualizes relationships between two continuous variables. Reveals direction, strength, and linearity of association (e.g., dose vs. response).

Box Plot
Summarizes distribution via median, IQR, and outliers. Ideal for comparing groups across categorical variables (e.g., treatment vs. control).

Bar Chart
Compares mean values across discrete groups with error bars (SD or SEM). Effective for summarizing categorical group differences in biological experiments.



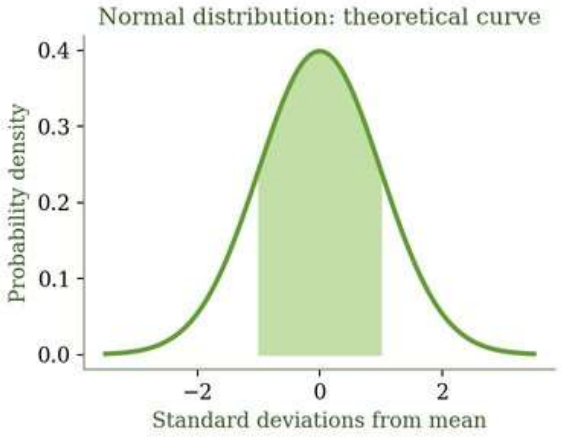
Educational illustration

The Normal Distribution

Bell-Shaped Curve
Many biological traits — height, enzyme activity, blood pressure — follow a symmetric bell-shaped distribution centred on the mean μ.

Empirical Rule
±1 SD ≈ 68% of data; ±2 SD ≈ 95%; ±3 SD ≈ 99.7%. These zones define expected biological variation around μ.

Central Limit Theorem
Many small additive genetic and environmental factors combine to produce normality. Sample means converge to normal as n increases.



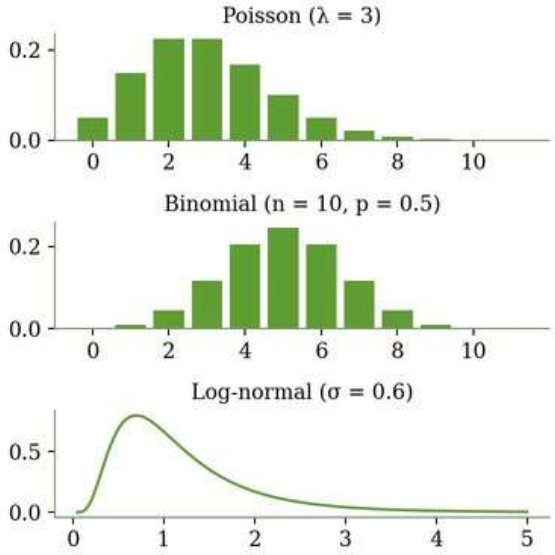
Educational illustration

Non-gaussian Distributions

Poisson Distribution
Models rare count events per unit time or space (e.g., DNA mutations per cell, radioactive decay events).

Log-Normal Distribution
Models right-skewed continuous data where log(X) is normally distributed (e.g., incubation periods, enzyme concentrations).

Binomial Distribution
Models the number of successes in a fixed number of independent binary trials (e.g., allele presence, survival outcomes).



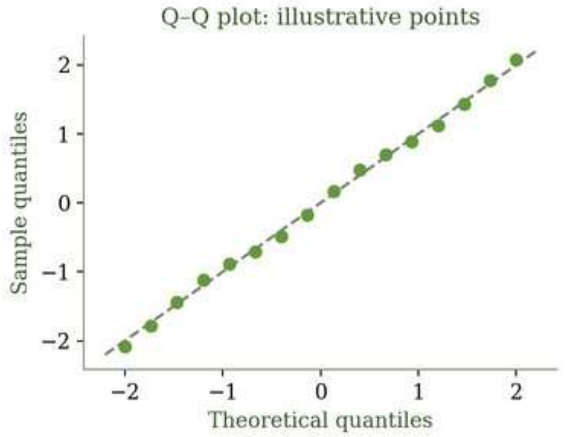
Educational illustration

Validating Parametric Assumptions

Q-Q Plots
Plot sample quantiles against theoretical normal quantiles; points on the diagonal line indicate normality.

Shapiro-Wilk Test
Formal significance test for normality; preferred for small samples (n < 50). Also: Kolmogorov-Smirnov test for larger samples.

Histograms
Visualise the frequency distribution; a symmetric bell shape suggests normality of the data.



Educational illustration

Making Decisions Under Uncertainty

Null Hypothesis (H_0)

States that there is noeffect or difference between groups. Example: $\mu_1 = \mu_2$. It is the default assumption that the researcher seeks to challenge with evidence.

Alternative Hypothesis (H_1)

States that a real effector difference exists. Example: $\mu_1 \neq \mu_2$. Accepted when evidence against H_0 is sufficiently strong ($p < \alpha$).

Test Statistic

A value calculated from sample data (e.g., t, F, χ^2) that quantifies how far the observed result deviates from what H_0 predicts, relative to sampling variability.

Decision: Reject / Retain H_0

Compare p-value to significance level α (typically 0.05). If $p < \alpha$, reject H_0 and support H_1. If $p \geq \alpha$, retain H_0. Never "accept" H_0 – only fail to reject it.

Quantifying Evidence Against The Null

p-value	Probability of observing data as extreme as (or more extreme than) the sample result, assuming H_0 is true. A small p-value indicates evidence against H_0.
Rejection Region	Tail area(s) of the null distribution where test statistics lead to rejection of H_0. For two-tailed tests, α is split equally between both tails ($\alpha/2$ each).
Significance Level α	Threshold α (commonly 0.05) set before testing. If $p \leq \alpha$, we reject H_0. Choosing α balances Type I error risk against detection sensitivity.

Type I & Type II Errors

Concept / test	Meaning and application
Type I Error (α)	False Positive. Rejecting a true null hypothesis (H_0). Concluding an effect exists when it does not. Probability equals the significance level α (typically 0.05).
Type II Error (β)	False Negative. Failing to reject a false null hypothesis (H_0). Missing a real biological effect. Probability is β; statistical power = 1 − β.
Correct Rejection	True Positive. Rejecting a false H_0. Correctly detecting a real effect or difference. Probability equals statistical power (1 − β).
Correct Retention	True Negative. Retaining a true H_0. Correctly concluding no effect is present. Probability equals 1 − α, the specificity of the test.

Comparing Means Across Biological Treatments

Concept / test	Meaning and application
Two-Sample t-Test	Compares means of two independent groups (e.g., control vs. treatment). Assumes normality and equal or unequal variance.
One-Way ANOVA	Compares means across 3 or more groups on one factor (e.g., control, low-dose, high-dose). Post-hoc tests identify which groups differ.
Paired t-Test	Used for matched or repeated-measures designs where each subject serves as its own control, reducing variability.

Non-parametric Tests For Group Comparisons

Mann-Whitney U Test	Compares two independent groups on ranked data; non- parametric alternative to the independent-samples t-test
Kruskal-Wallis Test	Compares 3 or more independent groups on ranked data; non-parametric alternative to one-way ANOVA
Wilcoxon Signed-Rank Test	Compares two related/paired samples using signed ranks; non-parametric alternative to the paired t-test

Chi-square (χ^2) Tests

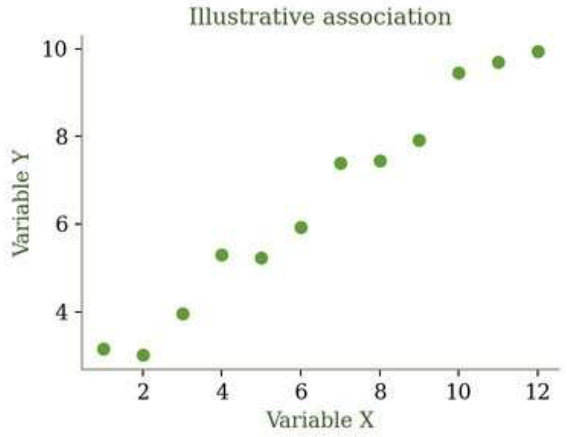
Observed vs. Expected	Compares observed counts to expected counts under a null model; $\chi^2 = \Sigma\ (O - E)^2 / E$
Mendelian Genetics	Tests inheritance ratios (e.g., 3:1 dominant:recessive) against observed phenotype frequencies.
Contingency Tables	Analyses associations between two categorical variables in a cross-tabulated frequency table.

Measuring Association Between Variables

Pearson's r
Measures linear association between two continuous, normally distributed variables. r ranges from −1 to +1.

Correlation Strength
|r| ≥ 0.7: strong | 0.3–0.7: moderate | < 0.3: weak. Sign indicates direction of association.

Spearman's ρ
Measures monotonic association for ordinal or non-normal data. Rank-based; robust to outliers and skew.



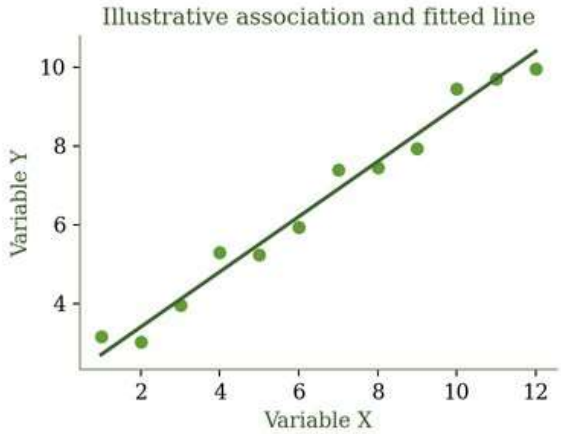
Educational illustration

Simple Linear Regression

Regression Model
$Y = \beta_0 + \beta_1 X$ predicts dependent variable Y from independent variable X; β_0 is the intercept constant

Goodness-of-Fit (R^2)
R^2 measures the proportion of variance in Y explained by X; ranges from 0 (no fit) to 1 (perfect fit)

Effect Size (Slope β_1)
Slope β_1 quantifies the change in Y per unit increase in X; represents the magnitude and direction of the effect



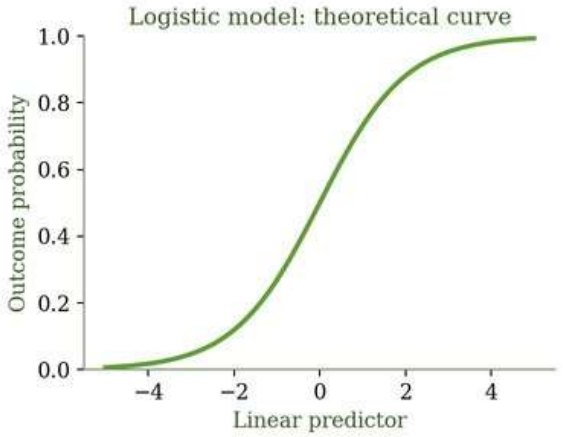
Educational illustration

Complex Multi-variable Biological Models

Multiple Regression
Models continuous outcome Y from several predictors: $Y = \beta_0 + \beta_1X_1 + \beta_2X_2 + \ldots + \beta\square X\square$; each β reflects independent effect size

Odds Ratios (OR)
OR quantifies association strength; OR > 1 indicates increased risk; derived directly from logistic regression coefficients

Logistic Regression
Models binary outcomes (e.g., diseased/healthy); estimates probability via sigmoid curve; suited for clinical and genetic data



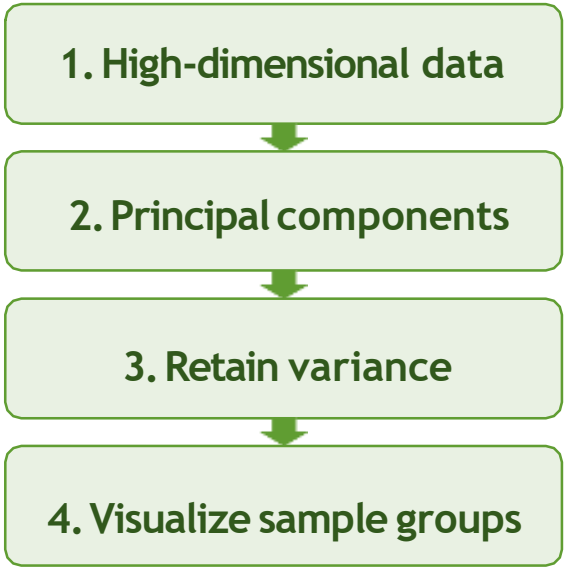
Educational illustration

Multivariate Analysis: PCA & Clustering

Dimensionality Reduction
PCA compresses high-dimensional omics data into principal components, capturing maximum variance with fewer variables

Clustering & Grouping
2D PCA score plots reveal natural clustering of treated vs. control samples in complex biological datasets

Variance Retention
Each principal component retains the greatest possible proportion of remaining variance in the dataset

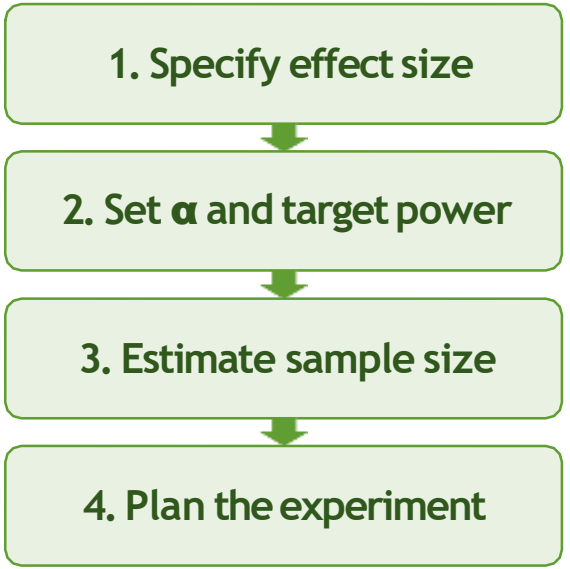


Statistical Power & Sample Size

Power Analysis
Calculates minimum n needed to detect a true effect at significance level α with adequate probability (typically 80–90%).

Power Curves
Power curves plot sample size vs. statistical power across effect sizes — larger effects require fewer subjects to achieve adequate power.

Avoiding Type II Errors
Insufficient sample size raises Type II error risk (β), causing real biological effects to go undetected — a critical flaw in study design.



Computational Tools In Modern Biology

R / RStudio

Open-source environment with extensive biostatistics packages (e.g., ggplot2, vegan, limma, survival). Ideal for reproducible research, custom analysis, and publication- quality graphics.

P ython

Versatile programming ecosystem with SciPy, Statsmodels, Pandas, and Seabom. Supports statistical modelling, data wrangling, and integration with machine learning pipelines.

GraphPad Prism

User-friendly GUI software widely used in laboratory and biomedical research. Provides t-tests, ANOVA, regression, and nonlinear curve fitting with clear visual output.

SPSS

Industry-standard platform common in clinical and epidemiological studies. Offers point-and-click access to descriptive statistics, GLMs, survival analysis, and survey data tools.

Ethical Biostatistics

Avoid P-Hacking & HARKing

Donot selectively report results aftertesting multiple hypotheses (p-hacking) or reframe hypotheses after seeing data (HARKing). Report all tests conducted and planned analyses transparently.

Pre-Registration

Register hypotheses,methods, and analysis plans before data collection. Pre-registration prevents post-hoc rationalization and strengthens the credibility of confirmatory research findings.

Open Data Sharing

Share rawdatasetsin public repositories (e.g., Dryad, OSF, Zenodo). Open data enables independent verification, reanalysis, and meta-analysis, reinforcing scientific reproducibility.

Reproducible Code & Reporting

Publishanalysisscripts (R, Python) alongside manuscripts. Report effect sizes, confidence intervals, and exact p- values. Full transparent reporting allows others to replicate and validate results.

Workshop Assignment

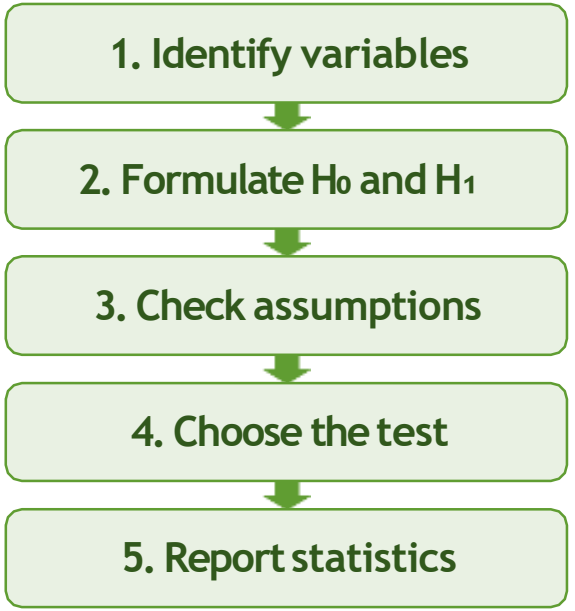
Formulate Hypotheses
State the Null (H_0) and Alternative (H_1) hypotheses for this experiment

Descriptive Statistics
Report two descriptive statistics for the treatment group (e.g., median, IQR)

Scenario
Plant extract effect on blood glucose in diabetic mice (n=12 treatment, n=12 control)

Choose the Test
Select parametric or non-parametric test; justify based on normality assumptions

Identify Variables
Name the dependent variable and classify the data type for each group



References & Suggested Reading

Zar, J. H. (2010)
Biostatistical Analysis (5th ed.). Pearson Prentice Hall.

Field, Miles & Field (2012)
Discovering Statistics Using R. SAGE Publications.

Gotelli & Ellison (2013)
A Primer of Ecological Statistics (2nd ed.). Sinauer Associates.

Sokal & Rohlf (2012)
Biometry: Principles and Practice of Statistics in Biological Research (4th ed.). W. H. Freeman.

Whitlock & Schluter (2020)
The Analysis of Biological Data (3rd ed.). W. H. Freeman and Company.

