

Introduction to Biostatistics

Introduction to Biostatistics

Key concepts, terminology, and tools for applying statistical methods to biological and health-related data.

1. What Is Biostatistics?

Aspect	Description
Definition	The branch of statistics that deals with the collection, analysis, interpretation, and presentation of data arising from biological, medical, and public health research.
Purpose	- Summarize complex biological data - Quantify uncertainty and variability - Test scientific hypotheses - Inform clinical and policy decisions
Typical Applications	Clinical trials, epidemiologic studies, genetics, environmental health, health services research, bioinformatics.

2. Types of Data in Biostatistics

Data Type	Scale	Examples	Typical Analyses
Nominal (Categorical)	No intrinsic order	Sex, blood type, disease presence (yes/no)	Chi-square test, logistic regression
Ordinal	Ordered categories	Pain rating (none, mild, moderate, severe)	Mann-Whitney U, ordinal logistic regression
Discrete Count	Integer values, often non-negative	Number of infections, gene mutation counts	Poisson/negative binomial regression
Continuous	Measured on a real number scale	Blood pressure, cholesterol, gene expression levels	t-test, ANOVA, linear regression, survival analysis

Note: Many biological variables are **right-skewed** (e.g., serum triglycerides). Transformations (log, square-root) are often used to meet model assumptions.

Introduction to Biostatistics

3. Study Designs & Their Statistical Implications

Design	Key Features	Primary Statistical Considerations
Cross-sectional	Snapshot of a population at one time point	Prevalence estimation, logistic regression
Case-control	Subjects selected by outcome status (cases vs. controls)	Odds ratios, conditional logistic regression
Cohort (prospective/retrospective)	Follow subjects over time to observe incidence	Relative risk, Cox proportional hazards model
Randomized Controlled Trial (RCT)	Random allocation to intervention vs. control	Intention-to-treat analysis, ANCOVA, survival analysis
Longitudinal / Repeated Measures	Multiple observations per subject	Mixed-effects models, generalized estimating equations (GEE)
Clustered / Hierarchical	Groups (e.g., clinics, schools) as sampling units	Multilevel models, intraclass correlation coefficient (ICC)

4. Core Statistical Concepts

Concept	Definition	Why It Matters in Biostatistics
Population vs. Sample	Population = all possible units; Sample = subset actually observed	Inference relies on sample representing the population
Parameter vs. Statistic	Parameter = true value in population; Statistic = estimate from sample	Estimation (e.g., mean, proportion) aims to approximate parameters
Bias	Systematic deviation of an estimate from the true value	Study design and measurement methods must minimize bias

Introduction to Biostatistics

Concept	Definition	Why It Matters in Biostatistics
Variance & Standard Error	Variability of an estimator; SE = SD of sampling distribution	Determines precision of estimates and width of confidence intervals
Confidence Interval (CI)	Range of values that, with a specified probability (e.g., 95%), contains the true parameter	Provides a measure of uncertainty beyond a single point estimate
p-value	Probability of observing data as extreme as (or more extreme than) those observed, assuming the null hypothesis is true	Used