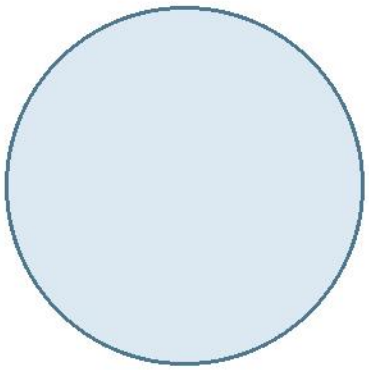
# BASIC BIOSTATISTICS

**Author: Manjunath Kulkarni, Nephrologist, Mangaluru, India**

An Evidence Integration Labs initiative



This downloadable handbook is designed to accompany the calculator and provide a quick conceptual foundation for descriptive statistics, analytic statistics, and sample size calculation.

It is intentionally practical and written for medical learners, clinicians, researchers, and postgraduate trainees who want a clean starting point for interpreting biomedical data.

# Index



Author: Manjunath Kulkarni
Mangaluru, India

# 1. What biostatistics does in medical research

Biostatistics is the language used to summarize uncertainty in medical data. It helps clinicians and researchers move from observation to evidence by organizing measurements, describing patterns, and comparing groups in a transparent way.

In practice, biostatistics supports study design, data collection, analysis, interpretation, and communication. A good statistical plan helps avoid avoidable bias, underpowered studies, and overstated conclusions.

The goal is not only to produce a p value. The larger goal is to quantify effects, precision, variation, and credibility so that clinical interpretation becomes more disciplined.

Key point: biostatistics is most useful when it improves clarity, not when it adds unnecessary complexity.

## 2. Types of medical data

Medical variables are commonly grouped as nominal, ordinal, and continuous. Nominal variables are categories without order, such as blood group or treatment arm. Ordinal variables carry rank, such as disease stage or symptom grade. Continuous variables are numerical measurements, such as age, blood pressure, and serum creatinine.

Choosing the correct statistical summary starts with understanding the data type. Counts and percentages fit nominal variables, while means or medians fit continuous variables depending on the distribution.

Key point: most statistical mistakes begin with misclassifying the type of data being analyzed.

# 3. Descriptive statistics overview

Descriptive statistics summarize what the sample looks like before any formal comparison is made. They answer questions such as: what is the average age, how variable are the values, and how many people had a positive outcome?

Descriptive analysis is not a minor preliminary task. It is often where data-quality problems, outliers, impossible values, and inconsistent coding first become visible.

Key point: strong description is the foundation on which all later analytic interpretation depends.

# 4. Frequency tables and percentages

Categorical data are usually summarized in frequency tables. A frequency is the number of observations in a category. Percentages make the same information easier to compare across groups of different sizes.

When percentages are presented, the denominator should always be clear. A percentage without a denominator can be misleading, especially when there are missing values.

Key point: percentages should always be linked to their denominator and never shown in isolation.

# 5. Measures of central tendency

The mean is the arithmetic average and is most useful when data are approximately symmetric. The median is the middle value and is more robust to skewness and outliers. The mode is the most frequent value but is used less often in clinical reports.

For normal-looking continuous data, mean and standard deviation are usually appropriate. For skewed data, median and interquartile range are generally more informative.

Key point: choose mean or median based on the data shape rather than habit.

# 6. Measures of variability

Variation matters because averages alone can hide important differences. Standard deviation describes spread around the mean. Interquartile range describes the spread of the middle 50% of observations around the median.

Range is easy to understand but unstable because it depends heavily on the smallest and largest observations. Variance is mathematically useful but less intuitive than standard deviation.

Key point: averages without measures of spread are incomplete and often misleading.

# 7. Normal distribution and skewness

Many familiar parametric methods assume data are approximately normally distributed within groups. A normal distribution is symmetric and bell-shaped. Skewed data have a long tail to one side and often require different summaries or tests.

Normality should not be judged by a formal test alone. Visual review using histograms, box plots, or QQ plots is often more informative when used alongside subject-matter knowledge.

Key point: distributional assumptions should be checked visually as well as statistically.

# 8. Presenting data in tables and graphs

Clear display is part of good statistics. Tables should use explicit labels, units, and denominators. Graphs should match the data type: bar charts for counts, box plots for skewed continuous data, scatter plots for paired numeric variables, and line plots for trends over time.

Presentation should reduce ambiguity rather than decorate the page. The best tables and figures make the main result obvious without oversimplifying it.

Key point: the best figure is the one that makes the clinical message easier to understand.

# 9. Analytic statistics overview

Analytic statistics are used when the goal is comparison, association, or prediction rather than description alone. Examples include comparing proportions between treatment groups, testing whether average blood pressure differs between groups, or modelling the relation between BMI and systolic blood pressure.

The analytic method depends on the study question, the outcome type, the number of groups, and assumptions such as normality and independence.

Key point: every analytic test should answer a clearly stated research question.

# 10. Hypothesis testing and p values

A hypothesis test compares the observed data with what would be expected under a null hypothesis. The p value is the probability of seeing data at least this extreme if the null hypothesis were true.

A small p value suggests incompatibility with the null model, but it is not the probability that the null hypothesis is false. It does not measure effect size or clinical importance.

Key point: a p value does not measure the size, importance, or truth of an effect.

# 11. Confidence intervals and effect estimates

A point estimate summarizes the observed effect in the sample. A confidence interval gives a range of plausible values for the corresponding population effect under the model used.

Confidence intervals are often more informative than p values because they display both direction and precision. Wide intervals suggest uncertainty; narrow intervals suggest greater precision.

Key point: confidence intervals tell you how precise your estimate is, not just whether it is significant.

# 12. Chi-square and Fisher exact tests

Chi-square testing is commonly used for nominal data arranged in contingency tables. It evaluates whether the observed distribution differs from the distribution expected if the variables were independent.

When expected counts are small in a 2x2 table, Fisher exact testing is preferred because the usual large-sample chi-square approximation may become unreliable.

Key point: small expected counts should prompt a check for whether Fisher exact testing is more appropriate.

# 13. t tests and Mann-Whitney tests

The Student t test compares average values between two groups when the outcome is continuous and reasonably normal. The main effect estimate is usually the mean difference with a 95% confidence interval.

The Mann-Whitney U test is a non-parametric alternative for two groups when a rank-based approach is more appropriate, especially with skewness or influential outliers.

Key point: the test should match the distribution and scale of the outcome, not just the number of groups.

# 14. ANOVA and Kruskal-Wallis tests

ANOVA extends mean comparison to three or more groups when normality assumptions are acceptable. The main test statistic is an F statistic that compares between-group to within-group variability.

Kruskal-Wallis is the non-parametric alternative for three or more groups. It works on ranks rather than raw values and is useful when distributions are clearly non-normal.

Key point: comparing more groups does not automatically mean ANOVA is appropriate; assumptions still matter.

# 15. Correlation and regression

Correlation quantifies the strength and direction of association between two variables. Pearson correlation measures linear association, while Spearman correlation measures monotonic rank-based association.

Simple linear regression goes further by modelling the expected change in one variable for a one-unit change in another. Regression is particularly useful when prediction or effect estimation is the aim.

Key point: correlation describes association, whereas regression adds an interpretable predictive equation.

# 16. Diagnostic accuracy statistics

Diagnostic studies often summarize test performance using sensitivity, specificity, positive predictive value, negative predictive value, prevalence, and accuracy. Each metric answers a different question, so they should not be treated as interchangeable.

Sensitivity and specificity are intrinsic to the test under a given context, while predictive values depend strongly on disease prevalence in the studied population.

Key point: predictive values are prevalence-dependent, so they change when the tested population changes.

# 17. Statistical significance versus clinical importance

A statistically significant result may still be clinically trivial if the effect size is too small to matter in practice. Conversely, an effect can be clinically important but fail to reach conventional statistical significance in a small or noisy study.

Interpretation should therefore combine effect size, confidence interval, plausibility, study quality, and clinical context.

Key point: clinical importance and statistical significance should always be interpreted together.

# 18. Sample size principles

Sample size planning aims to recruit enough participants to answer the question with acceptable precision or power. Too few participants lead to unstable estimates and false-negative findings. Too many participants may waste time, effort, and resources.

Every sample size calculation depends on assumptions. These usually include the expected effect or prevalence, acceptable error, confidence level, statistical power, and design type.

Key point: every sample size number is only as good as the assumptions used to calculate it.

# 19. Sample size for prevalence studies

For a single proportion or prevalence study, the basic planning formula uses the expected prevalence, desired margin of error, and confidence level. The sample size grows when the desired precision becomes tighter.

If the expected prevalence is uncertain, researchers often use prior literature or a pilot estimate. The planning assumptions should always be recorded in the protocol.

Key point: better precision requires a larger sample, especially when prevalence is near 50 percent.

# 20. Sample size for estimating a mean

When the aim is to estimate a mean, the main inputs are the anticipated standard deviation and the desired absolute precision. More variable outcomes require larger samples to achieve the same precision.

This design is often used in descriptive physiological or laboratory studies where the primary aim is estimation rather than comparison.

Key point: outcomes with greater variability need more observations to estimate precisely.

# 21. Sample size for two proportions

Two-group comparison of proportions is common in cohort studies, randomized trials, and service evaluations with binary outcomes. Planning usually depends on the expected event rates in each group, type I error, and desired power.

Smaller anticipated differences require larger samples, while larger expected differences require fewer participants.

Key point: detecting a small absolute difference between proportions usually requires a large sample.

## 22. Sample size for two means

Two-group comparison of means depends on the anticipated common standard deviation, the target mean difference considered important, the significance level, and the desired power.

This is a standard planning approach for two-arm trials or comparative observational studies with continuous outcomes.

Key point: mean-difference studies are highly sensitive to the assumed standard deviation.

# 23. Sample size for case-control studies

In an unmatched case-control study, planning often relies on the expected exposure proportion in cases and controls. The same general logic used for comparing two proportions can be adapted when groups are selected by disease status.

Case-control planning should also consider feasibility, source population, and the validity of exposure assessment.

Key point: in case-control work, exposure assumptions should come from the most credible prior evidence available.

# 24. Sample size for diagnostic accuracy studies

Diagnostic accuracy planning can focus on estimating sensitivity or specificity with a desired precision. In that setting, the target parameter, acceptable margin of error, and anticipated disease prevalence are key inputs.

Because prevalence affects how many participants fall into the diseased and non-diseased groups, it directly influences the total number of people who need to be sampled.

Key point: prevalence determines how efficiently a diagnostic study accumulates diseased and non-diseased participants.

# 25. Practical checklist and references

Before starting analysis, confirm the research question, outcome type, grouping structure, distribution assumptions, missing-data approach, and sample size logic. Record the chosen test, the main estimate, confidence interval, and any important limitations.

Formula references used in the companion calculator include the WHO practical manual by Lwanga and Lemeshow, the overview by Charan and Biswas, and the diagnostic precision paper by Buderer.

Key point: a short pre-analysis checklist can prevent many avoidable interpretation errors.

Author: Manjunath Kulkarni
Mangaluru, India