q— title: “Statistical Methods: Comprehensive Teaching Guide” author: “Dr. Siddalingaiah H S, Professor, Community Medicine, SIMSRH, Tumkur” date: “2025” geometry: margin=1in fontsize: 11pt colorlinks: true linkcolor: blue urlcolor: blue

citecolor: green toc: true toc-depth: 2 numbersections: true —

1 Statistical Methods: Comprehensive Teaching Guide

Created by: Dr. Siddalingaiah H S, Professor, Community Medicine, SIMSRH, Tumkur

1.1 1. Introduction to Statistical Methods

2 Module 1: Introduction to Statistical Methods

2.1 Learning Objectives

- Define statistics and its role in research
- Understand the difference between descriptive and inferential statistics
- Identify types of data and measurement scales
- Understand the research process and study designs
- Recognize ethical considerations in statistical analysis

2.2 What is Statistics?

Statistics is the science of collecting, organizing, analyzing, interpreting, and presenting data. It provides methods for making inferences about populations from samples and for quantifying uncertainty.

Key terms: - Population: Complete set of individuals or objects of interest **- Sample**: Subset of the population used for study **- Parameter**: Numerical characteristic of a population **- Statistic**: Numerical characteristic of a sample **- Variable**: Characteristic that varies among individuals

2.3 Branches of Statistics

2.3.1 1. Descriptive Statistics

- Organize and summarize data
- Describe main features of a dataset
- Include measures of central tendency, dispersion, and distribution

2.3.2 2. Inferential Statistics

- Make inferences about populations from samples
- Test hypotheses and draw conclusions
- Include estimation, hypothesis testing, and prediction

2.4 Types of Data

2.4.1 By Nature

- **Quantitative**: Numerical measurements
  - **Discrete**: Countable values (e.g., number of children)
  - **Continuous**: Measurable values (e.g., height, weight)
- **Qualitative**: Categorical measurements
  - **Nominal**: Categories without order (e.g., gender, blood type)
  - **Ordinal**: Categories with order (e.g., education level, pain scale)

2.4.2 By Measurement Scale

- **Nominal**: Categories, no quantitative meaning
- **Ordinal**: Ordered categories, unequal intervals
- **Interval**: Equal intervals, no true zero
- **Ratio**: Equal intervals, true zero point

2.5 The Research Process

2.5.1 1. Problem Identification

- Define research question
- Review existing literature
- Formulate hypotheses

2.5.2 2. Study Design

- Choose appropriate design
- Determine sample size
- Select measurement methods

2.5.3 3. Data Collection

- Implement data collection procedures
- Ensure data quality
- Maintain ethical standards

2.5.4 4. Data Analysis

- Clean and prepare data

- Apply statistical methods
- Interpret results

2.5.5 5. Reporting Results

- Present findings clearly
- Draw appropriate conclusions
- Suggest future research

2.6 Study Designs in Research

2.6.1 Experimental Designs

- **Randomized Controlled Trials (RCTs)**: Gold standard for establishing causality
- **Quasi-experimental**: Intervention without randomization
- **Field Experiments**: Natural setting interventions

2.6.2 Observational Designs

- **Cohort Studies**: Follow groups over time
- **Case-Control Studies**: Compare cases and controls
- **Cross-Sectional Studies**: Snapshot at single point
- **Ecological Studies**: Population-level data

2.7 Ethical Considerations

2.7.1 Key Principles

- **Respect for persons**: Informed consent, voluntary participation
- **Beneficence**: Maximize benefits, minimize harms
- **Justice**: Fair distribution of benefits and burdens
- **Confidentiality**: Protect participant privacy

2.7.2 Statistical Ethics

- **Data integrity**: Accurate data collection and analysis
- **Transparency**: Clear reporting of methods and results
- **Objectivity**: Avoid bias in analysis and interpretation
- **Reproducibility**: Enable verification of results

2.8 Software Tools for Statistics

2.8.1 Statistical Software

- **R**: Free, powerful statistical programming language
- **Python**: General programming with statistical libraries

- **SPSS**: User-friendly interface for statistical analysis
- **SAS**: Enterprise statistical software
- **Stata**: Comprehensive statistical package

2.8.2 Data Visualization

- **Tableau**: Interactive data visualization
- **Power BI**: Business intelligence and analytics
- **ggplot2 (R)**: Advanced plotting capabilities
- **matplotlib (Python)**: Flexible plotting library

2.9 Summary

Statistical methods provide the tools for extracting meaningful insights from data. Understanding the fundamentals of statistics is essential for conducting valid research and making evidence-based decisions in health sciences and beyond.

2.10 2. Descriptive Statistics

3 Module 2: Descriptive Statistics

3.1 Learning Objectives

- Understand measures of central tendency
- Calculate and interpret measures of dispersion
- Describe data distributions
- Create and interpret data visualizations
- Understand sampling distributions

3.2 Measures of Central Tendency

3.2.1 Mean (Arithmetic Average)

Formula: $\bar{x} = \frac{\sum x_i}{n}$

Example: Test scores: 85, 90, 78, 92, 88 Mean = (85 + 90 + 78 + 92 + 88) / 5 = 86.6

Advantages: - Uses all data points - Mathematically useful - Foundation for many statistical tests

Disadvantages: - Affected by extreme values (outliers) - Not appropriate for ordinal data

3.2.2 Median (Middle Value)

Calculation: - Sort data in ascending order - For odd n: Middle value - For even n: Average of two middle values

Example: Ages: 23, 25, 28, 30, 35 Median = 28

Advantages: - Not affected by outliers - Appropriate for ordinal data - Easy to understand

3.2.3 Mode (Most Frequent Value)

Definition: Value that appears most frequently

Example: Blood types: A, B, AB, A, O, A Mode = A

Uses: - Categorical data - Identifying most common category - Bimodal/multimodal distributions

3.3 Measures of Dispersion

3.3.1 Range

Formula: Range = Maximum - Minimum

Example: Heights: 160, 165, 170, 175, 180 cm Range = 180 - 160 = 20 cm

Limitations: - Only uses extreme values - Affected by outliers - No information about data distribution

3.3.2 Variance

Population Variance: $\sigma^2 = \frac{\sum (x_i - \mu)^2}{N}$

Sample Variance: $s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$

Interpretation: Average squared deviation from mean

3.3.3 Standard Deviation

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

Example: Data: 10, 12, 14, 16, 18 Mean = 14, Variance = 8, SD = $\sqrt{8}$ ≈ 2.83

Coefficient of Variation: CV = (s/mean) × 100%

3.4 Data Distributions

3.4.1 Normal Distribution

- Bell-shaped, symmetric
- Mean = Median = Mode
- 68% within 1 SD, 95% within 2 SD, 99.7% within 3 SD
- Many statistical tests assume normality

3.4.2 Skewness

- **Positive skew**: Tail extends to right
- **Negative skew**: Tail extends to left
- **Symmetric**: No skew

3.4.3 Kurtosis

- **Mesokurtic**: Normal peakedness
- **Leptokurtic**: More peaked than normal
- **Platykurtic**: Less peaked than normal

3.5 Data Visualization

3.5.1 Histograms

- Show frequency distribution
- Bar chart for continuous data
- Height represents frequency

3.5.2 Box Plots (Box-and-Whisker Plots)

- Show median, quartiles, outliers
- Box: IQR (Q3-Q1)
- Whiskers: 1.5 × IQR
- Points: Outliers

3.5.3 Scatter Plots

- Show relationship between two variables
- X-axis: Independent variable
- Y-axis: Dependent variable
- Pattern indicates correlation

3.5.4 Bar Charts and Pie Charts

- Categorical data
- Bar charts: Compare categories

- Pie charts: Show proportions

3.6 Sampling Distributions

3.6.1 Central Limit Theorem

- Sample means follow normal distribution
- Regardless of population distribution
- For sufficiently large samples ($n \geq 30$)

3.6.2 Standard Error

Formula: $SE = \frac{s}{\sqrt{n}}$

Interpretation: Standard deviation of sampling distribution

Uses: - Confidence interval calculation - Hypothesis testing - Sample size determination

3.7 Summary

Descriptive statistics provide the foundation for understanding data. Measures of central tendency describe typical values, while measures of dispersion describe variability. Data visualization helps identify patterns and outliers, and sampling distributions form the basis for inferential statistics.

3.8 3. Probability Theory

4 Module 3: Probability Theory

4.1 Learning Objectives

- Understand basic probability concepts
- Calculate probabilities using different rules
- Work with common probability distributions
- Understand conditional probability and independence
- Apply Bayes’ theorem

4.2 Basic Probability Concepts

4.2.1 Probability

Definition: Measure of likelihood that an event will occur

Range: $0 \leq P(A) \leq 1$ - $P(A) = 0$: Impossible event - $P(A) = 1$: Certain event

4.2.2 Sample Space

Definition: All possible outcomes of an experiment

Examples: - Coin flip: {Heads, Tails} - Die roll: {1, 2, 3, 4, 5, 6}

4.2.3 Events

- **Simple event**: Single outcome
- **Compound event**: Combination of outcomes
- **Mutually exclusive**: Cannot occur together
- **Independent**: Occurrence doesn't affect other events

4.3 Probability Rules

4.3.1 Addition Rule

For mutually exclusive events: $P(A \cup B) = P(A) + P(B)$

For non-mutually exclusive events: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

4.3.2 Multiplication Rule

For independent events: $P(A \cap B) = P(A) \times P(B)$

For dependent events: $P(A \cap B) = P(A) \times P(B|A)$

4.3.3 Complement Rule

$P(A') = 1 - P(A)$

4.3.4 Conditional Probability

Formula: $P(A|B) = P(A \cap B) / P(B)$

Example: Probability of having disease given positive test

4.4 Bayes' Theorem

4.4.1 Formula

$P(A|B) = [P(B|A) \times P(A)] / P(B)$

4.4.2 Extended Form

$P(A|B) = [P(B|A) \times P(A)] / [P(B|A) \times P(A) + P(B|A') \times P(A')]$

4.4.3 Applications

- Medical diagnosis
- Spam filtering
- Risk assessment
- Quality control

4.5 Common Probability Distributions

4.5.1 Discrete Distributions

4.5.1.1 Binomial Distribution

- Fixed number of trials (n)
- Each trial: success/failure
- Constant probability of success (p)
- Independent trials

Mean: $\mu = n \times p$ **Variance**: $\sigma^2 = n \times p \times (1-p)$

Example: Number of heads in 10 coin flips

4.5.1.2 Poisson Distribution

- Counts rare events
- Events occur randomly over time/space
- Constant average rate (λ)

Mean = Variance = λ

Example: Number of accidents per day

4.5.2 Continuous Distributions

4.5.2.1 Normal Distribution

- Bell-shaped, symmetric
- Defined by mean (μ) and standard deviation (σ)

- 68-95-99.7 rule

Standard Normal: $\mu = 0$, $\sigma = 1$ **Z-score**: $z = (x - \mu) / \sigma$

4.5.2.2 t-Distribution

- Similar to normal but with heavier tails
- Used for small samples
- Degrees of freedom = n - 1

4.5.2.3 Chi-Square Distribution

- Sum of squared standard normal variables
- Used for goodness of fit and independence tests
- Degrees of freedom vary

4.6 Expected Value and Variance

4.6.1 Expected Value (Mean)

Discrete: $E[X] = \Sigma\, x \times P(x)$

Continuous: $E[X] = \int x \times f(x)\, dx$

4.6.2 Variance

$Var(X) = E[X^2] - (E[X])^2$

4.6.3 Properties

- $E[aX + b] = aE[X] + b$
- $Var(aX + b) = a^2Var(X)$
- $Var(X + Y) = Var(X) + Var(Y)$ (if independent)

4.7 Summary

Probability theory provides the foundation for statistical inference. Understanding probability rules, distributions, and concepts like conditional probability and Bayes' theorem is essential for advanced statistical analysis and decision-making under uncertainty.

4.8 4. Inferential Statistics

5 Module 4: Inferential Statistics

5.1 Learning Objectives

- Understand sampling and sampling distributions
- Calculate and interpret confidence intervals
- Perform hypothesis testing
- Understand Type I and Type II errors
- Interpret p-values and statistical significance

5.2 Sampling Theory

5.2.1 Populations and Samples

- **Population**: Complete set of interest
- **Sample**: Subset used for inference
- **Sampling frame**: List of population elements
- **Sampling methods**: Random, stratified, cluster, systematic

5.2.2 Sampling Distributions

- **Sampling distribution**: Distribution of a statistic over many samples
- **Standard error**: Standard deviation of sampling distribution
- **Central Limit Theorem**: Sample means approach normal distribution

5.3 Confidence Intervals

5.3.1 Concept

- Range of values likely to contain true population parameter
- Based on sample data and desired confidence level

5.3.2 Formula for Mean (σ known)

CI = $\bar{x} \pm z \times \frac{\sigma}{\sqrt{n}}$

5.3.3 Formula for Mean (σ unknown)

CI = $\bar{x} \pm t \times \frac{s}{\sqrt{n}}$

5.3.4 Formula for Proportion

CI = $p \pm z \sqrt{\frac{p(1-p)}{n}}$

5.3.5 Interpretation

- 95% CI: 95% confidence that interval contains true parameter
- Wider intervals: More uncertainty
- Factors affecting width: Sample size, variability, confidence level

5.4 Hypothesis Testing

5.4.1 Steps in Hypothesis Testing

1. **State hypotheses**
    - H_0: Null hypothesis (no effect/difference)
    - H_1: Alternative hypothesis (effect/difference exists)
2. **Choose significance level (α)**
    - Common values: 0.05, 0.01, 0.10
3. **Select test statistic**
4. **Determine critical value or p-value**
5. **Make decision**
    - Reject H_0 if $p < \alpha$
    - Fail to reject H_0 if p α

5.4.2 Types of Errors

- **Type I Error (α)**: Reject true H_0 (false positive)
- **Type II Error (β)**: Fail to reject false H_0 (false negative)
- **Power (1-β)**: Probability of correctly rejecting false H_0

5.4.3 Power Analysis

Factors affecting power: - Sample size (primary factor) - Effect size - Significance level (α) - Variability

Sample size formula: $n = [(z\alpha + z\beta)/\delta]^2 \times \sigma^2$

5.5 Common Statistical Tests

5.5.1 Parametric Tests (Normal data, equal variances)

5.5.1.1 One-sample t-test

- Compare sample mean to known value
- H_0: $\mu = \mu_0$

5.5.1.2 Two-sample t-test

- Compare means of two groups
- Independent samples or paired

5.5.1.3 One-way ANOVA

- Compare means of three or more groups
- Tests if at least one group differs

5.5.1.4 Pearson Correlation

- Measure linear relationship between variables
- Range: -1 to +1

5.5.2 Non-parametric Tests (No normality assumption)

5.5.2.1 Chi-square test

- Test independence of categorical variables
- Goodness of fit

5.5.2.2 Mann-Whitney U test

- Compare two independent groups
- Alternative to t-test

5.5.2.3 Kruskal-Wallis test

- Compare three or more independent groups
- Alternative to ANOVA

5.5.2.4 Spearman correlation

- Measure monotonic relationship
- Alternative to Pearson correlation

5.6 p-Values and Significance

5.6.1 p-Value Definition

- Probability of observing data as extreme as sample data, assuming H_0 true
- Smaller p-values: Stronger evidence against H_0

5.6.2 Significance Levels

- **$p < 0.05$**: Statistically significant
- **$p < 0.01$**: Highly significant
- **$p < 0.001$**: Very highly significant

5.6.3 Misconceptions

- p-value ≠ probability that H_0 is true
- p-value ≠ importance of result
- Statistical significance ≠ clinical significance

5.7 Effect Size

5.7.1 Importance

- Magnitude of relationship or difference
- Independent of sample size
- More meaningful than p-values

5.7.2 Common Measures

- **Cohen's d**: Standardized mean difference
- **Odds ratio**: For categorical outcomes
- **Relative risk**: For incidence data
- **R^2**: Proportion of variance explained

5.8 Summary

Inferential statistics allow us to make conclusions about populations from sample data. Confidence intervals provide estimates with uncertainty, while hypothesis testing helps determine if observed effects are real. Understanding Type I/II errors, power, and effect sizes is crucial for proper interpretation of statistical results.

5.9 5. Parametric Tests

6 Module 5: Parametric Statistical Tests

6.1 Learning Objectives

- Understand assumptions of parametric tests
- Perform and interpret t-tests
- Conduct ANOVA and post-hoc tests
- Calculate and interpret correlation coefficients
- Understand linear regression

6.2 Assumptions of Parametric Tests

6.2.1 Normality

- Data follows normal distribution
- Check with histograms, Q-Q plots, Shapiro-Wilk test
- Central Limit Theorem helps with large samples

6.2.2 Homoscedasticity (Equal Variances)

- Variances equal across groups
- Test with Levene's test or Bartlett's test
- Important for t-tests and ANOVA

6.2.3 Independence

- Observations independent of each other
- Violated in repeated measures designs
- Important for all parametric tests

6.2.4 Linearity

- Relationship between variables is linear
- Check with scatter plots
- Important for correlation and regression

6.3 t-Tests

6.3.1 One-Sample t-Test

Purpose: Test if sample mean differs from known population mean

Formula: $t = (\bar{x} - \mu_0) / (s /)$

Example: Test if average height differs from national average

6.3.2 Independent Samples t-Test

Purpose: Compare means of two independent groups

Formula: $t = (\bar{x}_1 - \bar{x}_2) /$

Assumptions: - Independent observations - Normal distribution in each group - Equal variances (or use Welch's correction)

6.3.3 Paired t-Test

Purpose: Compare means of two related groups

Formula: $t = (\bar{x}\_d) / (s\_d /)$

Where $\bar{x}\_d$ is mean of differences

Uses: - Before-after studies - Matched pairs - Repeated measures

6.4 Analysis of Variance (ANOVA)

6.4.1 One-Way ANOVA

Purpose: Compare means of three or more groups

Logic: Partition total variation into between-group and within-group

F-statistic: $F = MS\_between / MS\_within$

Post-hoc tests: - Tukey's HSD: Compare all pairs - Bonferroni: Control family-wise error - Dunnett's: Compare to control group

6.4.2 Two-Way ANOVA

Purpose: Examine effects of two factors and their interaction

Model: $Y = \mu + A + B + AB + \varepsilon$

Main effects: Effect of each factor alone **Interaction**: Combined effect of factors

6.4.3 Repeated Measures ANOVA

Purpose: Compare means across multiple time points or conditions

Advantages: - Controls for individual differences - Requires fewer subjects - More powerful than independent groups

6.5 Correlation Analysis

6.5.1 Pearson Correlation

Purpose: Measure strength and direction of linear relationship

Formula: $r = \Sigma[(x\_i - \bar{x})(y\_i - \bar{y})] /$

Interpretation: - +1: Perfect positive correlation - 0: No linear relationship - -1: Perfect negative correlation

Hypothesis testing: $t = r$

6.5.2 Correlation vs. Causation

- Correlation does not imply causation
- Third variable may explain relationship
- Experimental design needed for causality

6.6 Linear Regression

6.6.1 Simple Linear Regression

Model: $Y = \beta_0 + \beta_1 X + \varepsilon$

Parameters: - β_0: Intercept (Y when X=0) - β_1: Slope (change in Y per unit X) - ε: Error term

Estimation: Least squares method

6.6.2 Multiple Linear Regression

Model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon$

Assumptions: - Linearity - Independence - Homoscedasticity - Normality of residuals

6.6.3 Model Evaluation

- **R^2**: Proportion of variance explained
- **Adjusted R^2**: Penalizes for additional variables
- **F-test**: Overall model significance
- **t-tests**: Individual coefficient significance

6.6.4 Regression Diagnostics

- **Residual plots**: Check assumptions
- **Influential points**: Cook’s distance
- **Multicollinearity**: VIF > 10

- **Outliers**: Standardized residuals > 3

6.7 Summary

Parametric tests provide powerful tools for comparing groups and examining relationships when assumptions are met. Understanding test assumptions, proper interpretation, and appropriate use of post-hoc tests is essential for valid statistical analysis.

6.8 6. Non-Parametric Tests

7 Module 6: Non-Parametric Statistical Tests

7.1 Learning Objectives

- Understand when to use non-parametric tests
- Perform and interpret chi-square tests
- Apply Mann-Whitney and Kruskal-Wallis tests
- Use Spearman’s correlation
- Understand non-parametric alternatives

7.2 When to Use Non-Parametric Tests

7.2.1 Advantages

- **No normality assumption**: Work with any distribution
- **Ordinal data**: Appropriate for ranked data
- **Robust**: Less affected by outliers
- **Small samples**: Often more powerful with small n

7.2.2 Disadvantages

- **Less powerful**: May miss real effects with normal data
- **Less precise**: Often only test for differences, not magnitude
- **Ordinal results**: May lose information from continuous data

7.2.3 Decision Criteria

- Data not normally distributed
- Ordinal or nominal data
- Small sample sizes

- Presence of outliers

7.3 Chi-Square Tests

7.3.1 Chi-Square Goodness of Fit

Purpose: Test if observed frequencies match expected distribution

Formula: $\chi^2 = \Sigma [(O_i - E_i)^2 / E_i]$

Example: Test if die is fair - Expected: Each face 1/6 = 16.67 - Observed: 20, 15, 18, 16, 17, 14

7.3.2 Chi-Square Test of Independence

Purpose: Test if two categorical variables are independent

Contingency table:

```
         Variable B
Variable A | B1 | B2 | Total
-----------|----|----|------
A1         |    |    |
A2         |    |    |
-----------|----|----|------
Total      |    |    |
```

Formula: Same as goodness of fit, but for all cells

Expected frequency: E_{ij} = (Row total × Column total) / Grand total

7.3.3 Fisher's Exact Test

Purpose: Alternative to chi-square for small samples

Uses: 2×2 tables with small expected frequencies (<5)

7.4 Mann-Whitney U Test

7.4.1 Purpose

- Compare two independent groups
- Alternative to independent samples t-test
- Works with ordinal or continuous data

7.4.2 Logic

- Rank all observations combined

- Compare sum of ranks between groups
- Test if groups come from same distribution

7.4.3 Formula

$U = n_1n_2 + n_1(n_1+1)/2 - R_1$

Where R_1 is sum of ranks in group 1

7.4.4 Effect Size

$r = |z| / \sqrt{(n_1 + n_2)}$

7.5 Kruskal-Wallis Test

7.5.1 Purpose

- Compare three or more independent groups
- Extension of Mann-Whitney test
- Alternative to one-way ANOVA

7.5.2 Logic

- Rank all observations
- Compare mean ranks between groups
- Chi-square approximation for large samples

7.5.3 Post-hoc Tests

- Dunn’s test for pairwise comparisons
- Bonferroni correction for multiple tests

7.6 Wilcoxon Signed-Rank Test

7.6.1 Purpose

- Compare two related samples
- Alternative to paired t-test
- Works with ordinal data

7.6.2 Procedure

1. Calculate differences between pairs
2. Rank absolute differences
3. Assign signs based on direction
4. Test if positive and negative ranks balanced

7.7 Spearman’s Rank Correlation

7.7.1 Purpose

- Measure monotonic relationship between variables
- Alternative to Pearson correlation
- Works with ordinal data

7.7.2 Calculation

1. Rank both variables
2. Calculate Pearson correlation on ranks

Formula: $r_s = 1 - (6\Sigma d_i^2) / (n(n^2-1))$

7.8 Summary

Non-parametric tests provide robust alternatives when parametric assumptions are not met. Understanding when to use each test and how to interpret results is essential for appropriate statistical analysis of various data types.