Usage of Statistical Methods in Data Handling: from Collection to Analysis

Dr. Md Saef Ullah Miah

Associate Professor, Department of Computer Science

Faculty of Science and Technology

Additional Director, Institutional Quality Assurance Cell (IQAC)

American International University-Bangladesh

Laboni Point, Cox's Bazar, Bangladesh
24th August 2025, 10:00 am.

Understanding Data: Primary and Secondary Sources

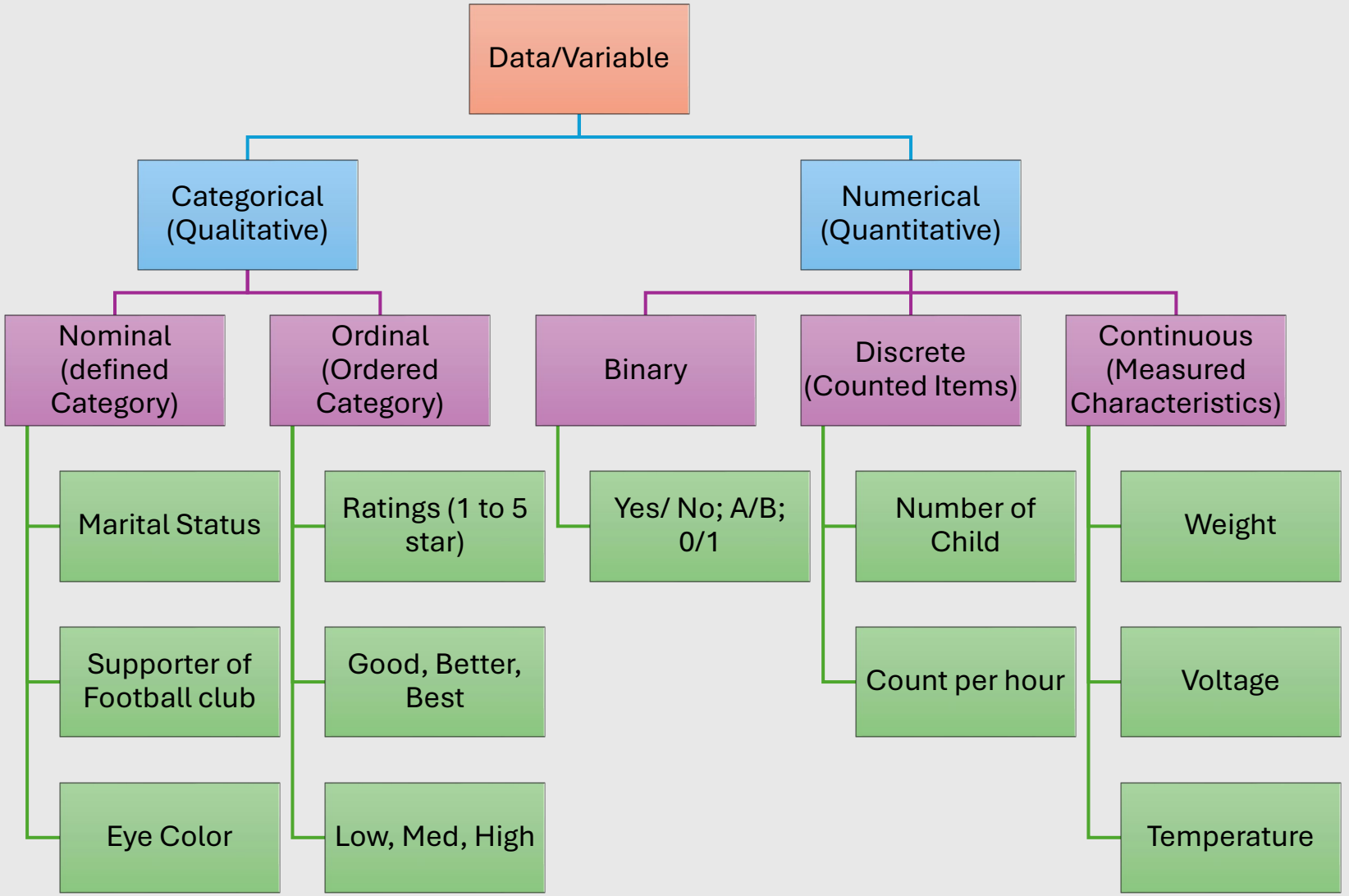
Primary Data

- Methods: Surveys, Experiments, Observations, Interviews
- Advantages: Relevant, up-to-date, tailored to objectives
- Limitations: Time-consuming, costly

Secondary Data

- Sources: Databases, Research Reports, Journals, Government Records
- Advantages: Quick, inexpensive, wide coverage
- Limitations: May be outdated or less specific

Types of Data / Variable



Data Collection

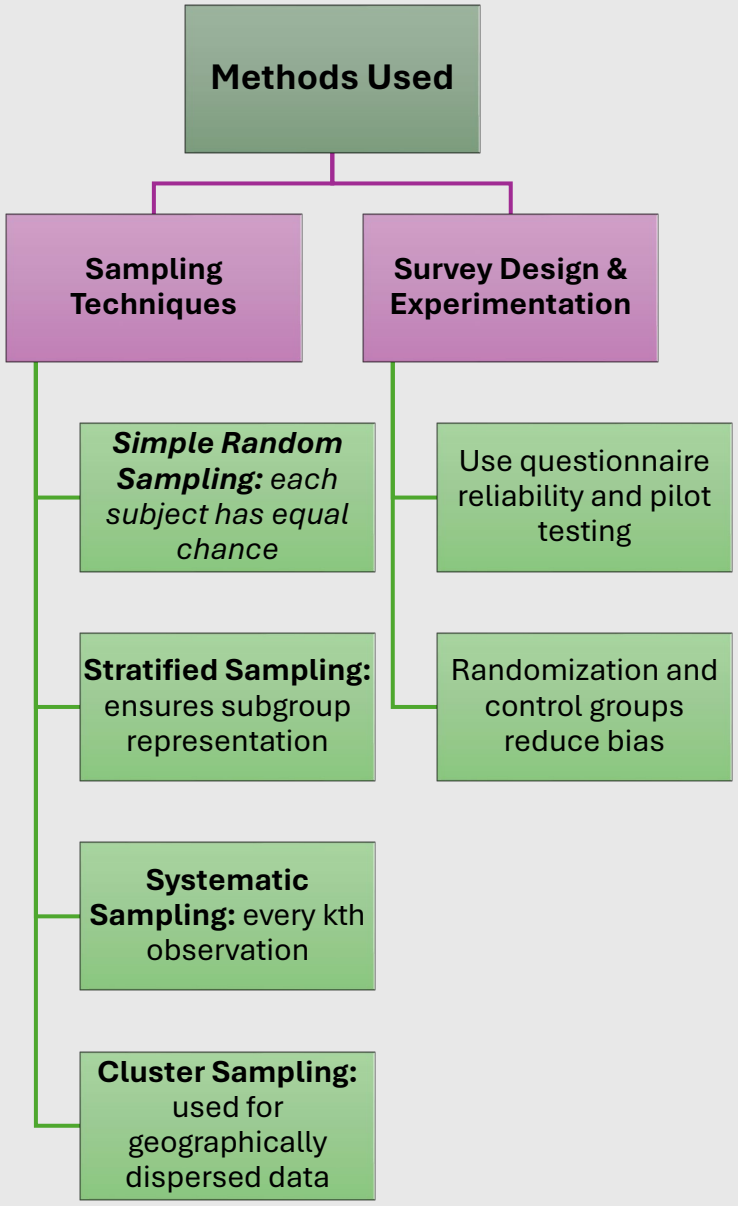
Purpose: Ensure data is representative, unbiased, and sufficient for analysis.

Guidelines:

✓ Define population clearly

✓ Use appropriate sample size (power analysis)

✓ Minimize sampling bias



Randomization and Control Group

Randomization

Definition:
Randomization is the process of assigning subjects or experimental units to different groups (e.g., treatment or control) purely by chance rather than by choice or pattern.

Purpose:
To eliminate selection bias and ensure that every participant has an equal chance of being assigned to any group.

Common Techniques:

Simple randomization: Flip a coin or use a random number generator.

Block randomization: Ensures equal sample sizes in each group.

Stratified randomization: Randomize within subgroups (e.g., age, gender) to ensure balance.

Control Group

Definition:
A control group is a baseline group that does not receive the experimental treatment, used for comparison with the treated group.

Purpose:
To isolate the effect of the independent variable (the intervention or treatment).

Types of Control Groups:

Placebo Control: Receives an inactive treatment (e.g., sugar pill).

Active Control: Receives a standard or existing treatment.

No-treatment Control: Receives nothing.

Waitlist Control: Receives the treatment later (common in social studies).

Data Validation

Purpose: Confirm accuracy, completeness, and consistency of collected data.

Guidelines:

✓ Detect anomalies early
✓ Standardize units and formats
✓ Document validation criteria

Statistical Methods Used

- **Descriptive Statistics:** Mean, SD, Interquartile Range (IQR) for range checking
- **Outlier Detection:**
  - *Z-score* (> ±3)
    - If **Z > +3** or **Z < –3**, the data point is considered an **outlier**
  - *IQR Rule*
    - Data Point < Lower Bound:Q1 – 1.5×IQR
    - Data Point > Upper Bound Q3 + 1.5×IQR
- **Missing Data Analysis:**
  - *Little's MCAR test* to check randomness of missingness
- **Cross-validation Checks:** Compare data from multiple sources

Data Validation (Summary)

Method	Basis	Works Best For	Outlier Threshold	Pros	Cons
Z-Score	Mean & SD	Normal data	Z > ±3	Simple, standardized	Affected by skewed data
IQR Rule	Quartiles (Q1–Q3)	Non-normal data	Below Q1–1.5×IQR or above Q3+1.5×IQR	Robust, nonparametric	Less effective on small samples

In short:

- **Z-score** measures how far a value is from the mean in SD units.
- **IQR rule** measures how far a value is from the middle 50% of data.

Both identify data points that don't "fit" the expected pattern — crucial for ensuring data quality and valid analysis.

Data Verification

Purpose: Ensure data integrity, reliability, and authenticity.

Guidelines:

✓ Recheck data entry or coding errors

✓ Use control totals or benchmarks

✓ Validate through replication

Statistical Methods Used

1. **Reproducibility Tests:** Used to ensure that repeated measurements or studies yield consistent results
2. **Correlation Analysis:** Used to check internal consistency and relationships between variables.
3. **Inter-rater Reliability:** Used to measure the agreement between observers, coders, or instruments.
4. **Data Audits:** Used to verify data integrity and identify inconsistencies or errors.

1. Reproducibility Tests

- **Test–retest reliability** (correlation between repeated measurements)
- **Bland–Altman analysis** (agreement between two measurements)
- **Intraclass Correlation Coefficient (ICC)** (consistency of quantitative measures)
- **Coefficient of Variation (CV)** (relative variability across trials)
- **Paired t-test** (compare repeated measures for mean differences)

2. Correlation Analysis

- **Pearson correlation coefficient (r)** – for linear relationships between continuous variables
- **Spearman's rank correlation (ρ)** – for nonparametric or ordinal data
- **Kendall's Tau (τ)** – for ordinal or small sample data
- **Partial correlation** – controlling for other variables
- **Canonical correlation** – for multiple variable relationships

3. Inter-rater Reliability

- **Cohen's Kappa (κ)** – agreement between
- two raters (categorical data)
- **Fleiss' Kappa** – agreement among more than two raters
- **Cronbach's Alpha (α)** – internal consistency for scale items
- **Intraclass Correlation Coefficient (ICC)** – reliability for continuous ratings
- **Krippendorff's Alpha** – general-purpose reliability across data types

4. Data Audits

- **Random record sampling and cross-verification**
- **Benford's Law analysis** – detect anomalies in numeric data
- **Descriptive summary comparison** (mean, median, totals across datasets)
- **Error rate analysis** – proportion of mismatched entries
- **Chi-square goodness-of-fit test** – compare expected vs observed distributions

Statistical Methods for Data Verification and Reliability

Category	Purpose	Common Statistical Methods / Tests	Typical Data Type
Reproducibility Tests	Assess consistency when measurements or samples are repeated	• Test–retest reliability • Bland–Altman analysis • Intraclass Correlation Coefficient (ICC) • Coefficient of Variation (CV) • Paired t-test	Continuous / Interval
Correlation Analysis	Measure strength and direction of association between variables	• Pearson correlation (r) • Spearman's rank correlation (ρ) • Kendall's Tau (τ) • Partial correlation • Canonical correlation	Continuous / Ordinal
Inter-rater Reliability	Evaluate agreement between raters or instruments	• Cohen's Kappa (κ) • Fleiss' Kappa • Cronbach's Alpha (α) • Intraclass Correlation Coefficient (ICC) • Krippendorff's Alpha	Categorical / Ordinal / Continuous
Data Audits	Verify accuracy, completeness, and authenticity of data	• Random record sampling and cross-verification • Benford's Law analysis • Descriptive summary comparison • Error rate analysis • Chi-square goodness-of-fit test	Numeric / Categorical

Data Analysis

Purpose: Extract patterns, relationships, and insights.

Guidelines:

✓ Choose test based on data type (categorical vs continuous)

✓ Check assumptions (normality, independence)

✓ Report effect sizes and confidence intervals

Statistical Methods Used

- **Descriptive Analysis:** Mean, Median, Variance, Frequency
- **Inferential Analysis:**
  - *t-tests, ANOVA, Chi-square tests*
  - *Regression, Correlation, Factor Analysis*
- **Predictive/Exploratory Methods:**
  - *Machine Learning (e.g., Logistic Regression, Decision Trees)*
  - *Time Series & Trend Analysis*

Inferential Analysis (Draw conclusions about a population based on a sample)

Hypothesis Testing Methods

Used to determine if observed differences or relationships are statistically significant.

Method	Purpose	Example Use Case
t-test	Compares means between **two groups**	Compare average blood pressure between men and women
Paired t-test	Compares means from **same group at two times**	Before-and-after treatment analysis
ANOVA (Analysis of Variance)	Compares means among **three or more groups**	Compare student performance across multiple schools
Chi-square test (χ^2)	Tests **association between categorical variables**	Relationship between gender and smoking habits

** **Key Output:** p-value (probability of observing results by chance).
If $p < 0.05$, the result is considered statistically significant.

Inferential Analysis (Contd...)

Relationship and Dependence Methods

Used to analyze how variables relate or influence each other.

Method	Purpose	Example Use Case
Correlation analysis	Measures strength & direction of relationship between variables	Relationship between income and education
Regression analysis	Predicts value of one variable based on another	Predict sales from advertising spend
Multiple Regression	Examines influence of multiple predictors	Predict house prices using area, location, and number of rooms
Factor Analysis	Reduces many correlated variables into fewer underlying “factors”	Identify psychological traits from questionnaire items

Predictive / Exploratory Methods (Go beyond inference: to predict future outcomes or uncover hidden structures)

Machine Learning Methods (Predictive Analytics)

Used when the focus is on prediction and pattern discovery rather than classical inference.

Algorithm	Type	Typical Use Case
Logistic Regression	Classification	Predict whether a patient has a disease (yes/no)
Decision Trees	Classification / Regression	Predict customer churn based on demographics
Random Forests / Gradient Boosting	Ensemble Models	Improve accuracy over single models
K-Means Clustering	Unsupervised	Group customers by purchasing behavior
Support Vector Machines (SVM)	Classification	Image recognition or anomaly detection

** **Key Evaluation Metrics:** Accuracy, Precision, Recall, F1-score, AUC-ROC.

Predictive / Exploratory Methods (Contd..)

Time Series & Trend Analysis
Used for data collected over time to identify trends, patterns, or seasonality.

Method	Purpose	Example Use Case
Moving Average / Exponential Smoothing	Smooth short-term fluctuations	Sales forecasting
ARIMA (AutoRegressive Integrated Moving Average)	Model and forecast time-dependent data	Predict stock prices or demand
Seasonal Decomposition (STL)	Separate trend, seasonality, and noise	Monthly temperature pattern analysis
Trend Analysis / Linear Regression	Detect long-term upward or downward trend	GDP growth over years

** **Key Outputs:** Trend line, forecast intervals, autocorrelation (ACF/PACF) plots.

Summary Map: Analysis tasks

Category	Main Focus	Common Techniques	Outcome
Inferential Analysis	Testing hypotheses	t-test, ANOVA, Chi-square	Statistical significance
Dependence Analysis	Exploring relationships	Regression, Correlation, Factor Analysis	Quantified relationships
Predictive Analytics	Making predictions	Machine Learning models	Future outcomes
Time Series Analysis	Understanding trends over time	ARIMA, Exponential Smoothing	Forecasts and seasonality insights

Key Takeaways

01

Statistical methods are vital from data collection to insight generation.

02

Focus on validity, reliability, and transparency at each step.

03

Proper method selection ensures accuracy, reproducibility, and actionable results.

Download the resource

https://ping543f.github.io/downloads_/stat-method.pdf