Living Document - Biostatistical Analysis

ALLIANCE FOR DISEASE DETECTION AND RESPONSE

September 2025, Dr. Atwau Pius

1 INTRODUCTION

- Analysis begins from research idea, concept, proposal.
  - Sample size and Sampling method
  - Hypothesis formulation and testing - **Inferential statistics**
  - Choice of variables - **type of variable determines which analysis can be done on it; this guides your data analysis plan (DAP)**
    a) Actual analysis
    b) Data management - **Version control (Git and GitHub)**
  - Study design

Note:

- Every stage before data collection affects the quality of analysis.

2 STUDY DESIGN

Non-Experimental (Observational)		**Experimental**
Descriptive	**Analytical**	
Case reports	Case-control studies	Randomized Controlled Trials (RCTs)
Case series	Cohort studies	Quasi-experimental studies
Cross-sectional studies		Field trials
Ecological studies		Community trials

2.0.1 For each of these study designs, one needs to know:

1. Out come
2. Measure of effect

2.1 Non-Experimental (Observational) Studies

2.1.1 A. Descriptive Studies

Study Design	**Typical Outcome**	**Measure of Effect**
Case Report	Description of an unusual occurrence or novel finding in a single patient (no comparison group).	None (purely descriptive).
Case Series	Description of a group of patients with a similar condition or exposure.	None (no comparison or measure of association).
Cross-Sectional Study	Prevalence of disease or exposure at a single point in time.	- Prevalence - Prevalence ratio or Prevalence odds ratio when comparing groups.
Ecological Study	Association between exposure and outcome at the population or group level.	- Correlation coefficient (r) - Regression coefficient - Rate ratios (ecological associations).

Table 1: Descriptive Observational Study Designs, Their Outcomes, and Measures of Effect

2.1.2 B. Analytical Studies

Study Design	**Typical Outcome**	**Measure of Effect**
Case-Control Study	Presence or absence of disease (starting with outcome, looking backward for exposure).	- Odds Ratio (OR): estimates the strength of association between exposure and outcome.
Cohort Study	Incidence of disease among exposed vs unexposed over time.	- Risk Ratio (RR) or Relative Risk - Rate Ratio (if person-time used) - Attributable Risk (AR) or Risk Difference

Table 2: Analytical Observational Study Designs, Their Outcomes, and Measures of Effect

2.2 Experimental Studies

Study Design	Typical Outcome	Measure of Effect
Randomized Controlled Trial (RCT)	Incidence of outcome (disease, improvement, death, etc.) between intervention and control groups.	- Risk Ratio (RR) - Risk Difference (RD) - Number Needed to Treat (NNT) - Odds Ratio (OR)
Quasi-Experimental Study	Outcome after an intervention, but without randomization.	- Risk Ratio (RR) - Risk Difference (RD) - Pre-post difference (e.g., before-and-after comparison)
Field Trial	Incidence of outcome among individuals receiving a preventive intervention in a community (e.g., vaccine trials).	- Risk Ratio (RR) - Vaccine Efficacy (%) = (1 RR) $\times$ 100
Community Trial	Outcome measured at the population level following community-level interventions.	- Rate Ratio - Difference in rates or proportions between communities

Table 3: Experimental Study Designs, Their Outcomes, and Measures of Effect

2.3 Summary Table

Type	Study Design	Outcome	Measure of Effect
Descriptive	Case Report	Unusual single case	None
Descriptive	Case Series	Group of similar cases	None
Descriptive	Cross-Sectional	Prevalence	Prevalence ratio / odds ratio
Descriptive	Ecological	Group-level associations	Correlation coefficient / rate ratio
Analytical	Case-Control	Odds of exposure among cases vs controls	Odds Ratio (OR)
Analytical	Cohort	Incidence of disease	Risk Ratio (RR), Rate Ratio, Attributable Risk
Experimental	RCT	Incidence of outcome	RR, RD, OR, NNT
Experimental	Quasi-Experimental	Outcome before vs after intervention	RR, RD, pre-post difference
Experimental	Field Trial	Disease incidence (preventive)	RR, Vaccine efficacy
Experimental	Community Trial	Population-level outcomes	Rate ratio / Difference in rates

Table 4: Summary of Study Designs, Outcomes, and Measures of Effect

3 SAMPLING

3.1 Sampling Procedure

3.1.1 Sampling techniques

a) Lottery method

b) Table of random digits

c) Computer packages

3.1.2 Sampling methods

a) Random (probability) sampling methods - **Sampling techniques MUST be applied to each of the random sampling methods**

- Simple random sampling
- Systematic random sampling
- Stratified random sampling
- Cluster random sampling
- Multi-stage random sampling

b) Non probability Sampling methods

- Purposive sampling
- Quota sampling
- Convenience sampling
- Snowball sampling

3.2 Sample Size Calculation

3.2.1 Methods for sample size calculation

a) Statistical methods for the calculation of sample size

- Sample size estimation using population mean (μ)
  - i) Infinite population (N unknown)
  - ii) Finite population (N known)
- Estimation of the sample size using population proportion (p)
  - i) Infinite population (N unknown)
  - ii) Finite population (N known)

b) Non-statistical methods for sample size calculation

3.2.2 Population parameters versus sample statistics

Sample Statistics	Population Parameters
$\bar{x}$ (sample mean)	μ (population mean)
$\hat{p}$ (sample proportion)	p (population proportion)
s (sample std. deviation)	σ (population std. deviation)
s^2 (sample variance)	σ^2 (population variance)

Note:

- We have to know how to calculate each of these manually from the first principles.
- The sample size is calculated using either mean or proportion

4 VARIABLES

4.1 Types of Variables

a) Qualitative variables (text) Exist in categories.

- Nominal
- Ordinal (can be ordered in ranks)

b) Quantitative (numerical) variables

- Discrete (no successive values between values)
- Continuous (infinite values exist between values)

4.2 Groups of variables

a) Independent variables (x variables).

b) Dependent variables (y variables).

c) Confounding variables (important in causal effect analysis - regression)

5 DESCRIPTIVE STATISTICS

6 INFERENTIAL STATISTICS

6.1 Testing for normality and deviation from normality

a) Anderson Darling Test

b) Shapiro-Wilk Test - suitable for data n ¡ 2000.

c) Kolmogorov-Smirnov (K-S) Test

d) Chi-Square Goodness-of-Fit Test

6.2 Choosing tests - Reference table

Statistical Objective	**Parametric Test (mean)**	**Non-Parametric Test (median)**	**Application Context**
Compare single group to known value	One Sample t-test	Wilcoxon Signed-Rank Test	Testing if sample differs from hypothesized value
Compare 2 independent groups (same parameter measured in two different groups)	Independent t-test	Mann-Whitney U Test	Two unrelated groups comparison
Compare 2 related samples (same parameter measured in the same group at different time intervals)	Paired t-test	Wilcoxon Signed-Rank Test	Pre-test/post-test designs
Compare 3+ independent groups	One-Way ANOVA	Kruskal-Wallis Test	Multiple group comparison

Continued on next page

Table 5 – *Continued from previous page*

Statistical Objective	**Parametric Test (mean)**	**Non-Parametric Test (median)**	**Application Context**
Compare 3+ related groups	Related Samples ANOVA	Friedman Test	Longitudinal measurements
Assess relationships	Pearson's Correlation	Spearman's Correlation	Variable association
Cause effect analysis (Predictive modeling)	Regression analysis (Various methods)	Regression analysis (Various methods)	Type of regression analysis determined by type of data
2 variables (simple regression analysis)	Regression analysis	Regression analysis	Modeling relationship between 2 variables
More than 2 variables (multiple regression analysis)	Multiple regression analysis	Multiple regression analysis	Modeling relationship with multiple predictors
Difference in frequencies of categorical variables	Chi-square test for independence	Chi-square test for independence	Association between categorical variables
Association between categorical variables	Chi-square test for association	Chi-square test for association	Association between categorical variables