For more details please visit:
https://cintarch.com/books/

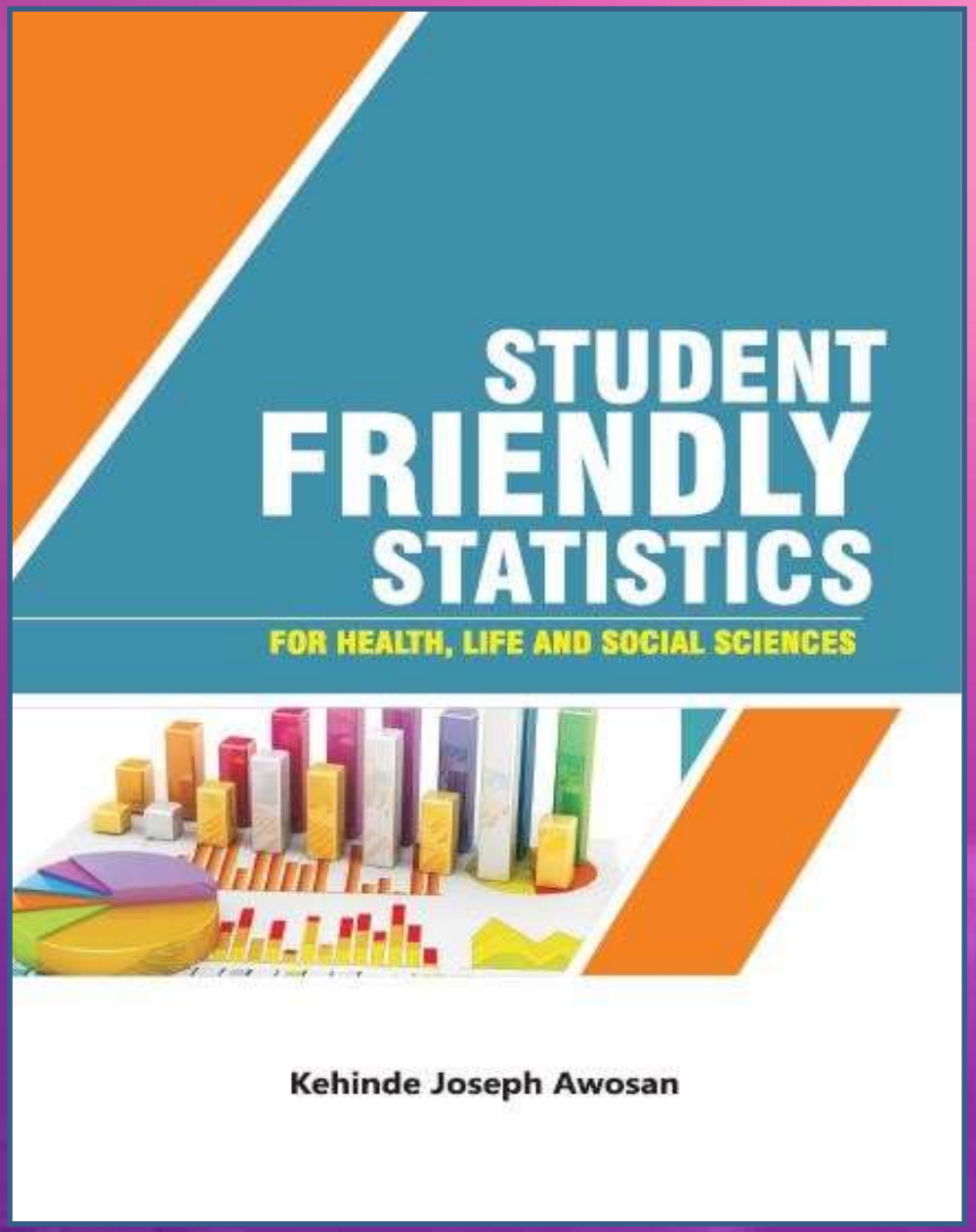
Excerpts from
Student Friendly Statistics
for Health, Life and Social Sciences

This book is an attempt by the author to present the subject in a simple, concise, and lucid approach. The book adequately covers all the necessary topics for its targeted audience, and provides a detailed step-by-step description of how to perform the various statistical procedures and tests.

I am therefore strongly constrained, on the basis of its merit, to recommend this book as an invaluable companion to both undergraduate and postgraduate students, and as a handy guide for self-tutoring by students in higher education. Practicing professionals would also find it as a quick reference.

I am sure that readers would find it a pleasurable and exciting exposition to Statistics simplified at its best.

Professor Taofeek M.T.O. Ibrahim
MBBS, Cert. Epid, FWACP, FIMC
Former Vice Chancellor
Al-Hikmah University, Ilorin, Nigeria



ORDER NOW

Please click on the link below to access the 'ORDER FORM'
https://forms.gle/3J8dWn6Ng6oMKd2e8

DISTRIBUTION CENTERS:

ABUJA ADO-EKITI ENUGU LAGOS SOKOTO

Student Friendly Statistics for Health, Life and Social Sciences

Kehinde Joseph Awosan
Associate Professor of Public Health and Community Medicine,
Department of Community Health,
College of Health Sciences,
Usmanu Danfodiyo University, Sokoto, Nigeria

Contents

Chapter 1

Basic Concepts in Statistics

Definition of Terms

Mathematics: This is the science that deals with the logic of shape, quantity and arrangement. It is all around us, and in everything we do. It is the building block for everything in our daily lives, including mobile devices, architecture (ancient and modern), art, money, engineering, sports, medicine, etc.

Statistics: This is a branch of mathematics that deals with the collection, arrangement, summarization, analysis, interpretation and presentation of data. Statistics is relevant to all disciplines.

Biostatistics: This is a subset of statistics that is concerned with data on living things (from bios – life). The disciplines involved here include Medicine (both human and veterinary), Food and Agricultural Sciences, Botany, Zoology, etc.

Health or Medical statistics: This is a subset of biostatistics that is concerned with data on health related issues.

Vital statistics refer to records of vital events (i.e., births, deaths, marriages and divorces) obtained through a civil registration process. The data obtained are used for generating birth and mortality rates for the whole population or subgroup.

An **observation** is an event that is seen to occur. **Data** refer to the records of two or more observations, while the record of a single observation is called **datum**. Data are collected on variables.

A **data set** is a collection of the data of the individual cases or objects, and each of the respective observations in a data set constitutes a **data point**.

Primary data refers to the first hand data gathered by the researcher. The sources include interviews, focus group discussions, observations, questionnaire surveys, experiments, etc. Although, it is expensive and takes a longer time to collect as compared to secondary data, it is more accurate and reliable than secondary data.

Secondary data refers to the data collected by someone else earlier (either published or unpublished) but now being accessed and used by the researcher. The sources include institutional records, government records and publications, books, journals, magazines, websites, etc.

Variables: These are characteristics of some events, objects or persons that can take on different values or amounts; e.g., age (25yrs, 45yrs, 70yrs, etc.), Height (1.5m, 1.65m, 1.7m, etc.), sex (male or female), marital status (single, married, separated, divorced or widowed).

There are 2 types of variables, i.e., qualitative and quantitative variables.

1. **Qualitative variables**: These are variables that are classified by attributes or categories, e.g., sex (male or female), marital status (single, married, separated, divorced or widowed.

   Qualitative variables can be measured on:
   - ***Nominal scale***: In this case only names are assigned, e.g. sex (male or female).
   - ***Ordinal or ranking scale***: In this case the variables are listed in a specified order or rank, e.g. the ranks of university lecturers (Graduate Assistant < Assistant Lecturer < Lecturer II < Lecturer I < Senior Lecturer < Reader < Professor).

2. **Quantitative variables**: These are variables that result from counting or measurement. They can be:
   - **Discrete:** These assume whole numbers only, e.g. the number of students in a class can be 1, 2, 3, 4, etc., it is not possible to have 3.5 students.
   - **Continuous**: These can assume fractions e.g., the weights of the students in a class can be 45kg, 50.5kg, 64.7kg, etc.

   Quantitative variables can be measured on;
   - ***Interval scale***: The starting point for this scale is arbitrary, it does not have a true zero point (i.e., zero on this scale does not indicate absence of the quantity measured (e.g. a temperature of 0^0C represents a temperature reading, as there can be temperatures below 0^0C such as: -2^0C, -10^0C, -20^0C, etc.).

Chapter 2

Measures of Central Tendency and Dispersion

Measures of Central Tendency

A measure of central tendency is a single value that describes a set of data by identifying the central position within that set of data. Measures of central tendency are also called measures of central location. They include:

a. Mean – Arithmetic mean
- Weighted arithmetic mean
- Geometric mean

b. Mode

c. Median (this is also one of the quantiles)

The appropriate measure of central tendency to be used depends on the type of data (i.e., mean for normally distributed data, and median for data that are not normally distributed).

Arithmetic mean

This refers to the sum of all the observations divided by the number of observations. The mean for a population is denoted by "mu" (μ), while the mean for a sample is denoted by "x bar" ($\bar{x}$), and the formula for computing it is:

$$\bar{x} = \frac{x_1 + x_2 + x_3 \ldots\ldots x_n}{n}$$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i^n$$

Example 2.1 The weights of six students (in kg) are:
50, 64, 70, 68, 55, 72
Find their mean weight.

The steps in answering the question are as follows:
Step 1: Compute the sum of the weights
i.e., 50 + 64 + 70 + 68 + 55 + 72 = 379

Step 2: Divide the sum by the number of observations

$$= \frac{379}{6}$$

= 63.17

The mean weight of the six students = **63.17kg**

Weighted arithmetic mean

This is used when observations have different weights attached to them.

If the values $x_1, x_2, x_3 \ldots\ldots x_n$ are associated with the weights $w_1, w_2, w_3 \ldots\ldots w_n$

The weighted arithmetic mean:

$$\overline{xw} = \frac{x_1w_1 + x_2w_2 + x_3w_3 \ldots\ldots x_nw_n}{n}$$

Example 2.2 If the mean PCV of 5 male patients is 45% and that of another 3 female patients is 40%, what is the mean PCV of the 8 patients?

The steps in answering the question are as follows:
Step 1: Compute the sum of the weighted PCV
i.e., (5 x 45) + (3 x 40) = 225 + 120 = 345
Step 2: Divide the sum by the number of observations

$$= \frac{345}{8} \quad = 43.13$$

The mean PCV of the 8 patients = **43.13%**

Geometric mean

This is the nth root of the product of the values, where n means the number of observations,

The geometric mean of the values $x_1, x_2, x_3 \ldots\ldots x_n$

$$= \sqrt[n]{(x_1)(x_2)(x_3) \ldots\ldots (x_n)}$$

Chapter 3

Principles of Parameters' Estimation

Statistical inference
Statistical inference refers to the procedure by which we reach a conclusion about a population on the basis of the information contained in a sample drawn from that population. This can be done by:

- Estimation
- Hypothesis testing

Estimation
Estimation involves calculating from the data of a **sample**, some **statistic** that is offered as an approximation of the corresponding **parameter** of the **population** from which the sample was drawn. The two types of estimates are:

- **Point estimate:** A point estimate is a single numerical value that is used to estimate the corresponding population parameter (e.g., arithmetic mean).
- **Interval estimate:** An interval estimate consists of two numerical values defining a range of values that, with a specified degree of confidence, we feel includes the parameter being estimated (e.g., confidence interval).

Estimator
An estimator is a rule (i.e., formula) that is used for computing an estimate. For example, the arithmetic mean is computed as the sum of all the observations divided by the number of observations. The mean for a population is denoted by "mu" (**μ**), while the mean for a sample is denoted by "x bar" ($\bar{x}$), and the formula for computing it is:

$$\bar{x} = \frac{x_1 + x_2 + x_3 \ldots\ldots x_n}{n}$$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i^n$$

The **formula** for calculating the sample mean ($\bar{x}$) is therefore an **estimator** of the population mean (μ), while the single numerical value that results from applying this formula is an **estimate** of the population **parameter**.

Sampled population refers to the population from which one actually draws a sample, while **target population** refers to the population about which one wishes to make an inference. Statistical inference procedures allow one to make inferences about sampled populations if probability sampling methods (i.e., **random sampling**) have been employed. Conclusions can be reached about the target population statistically only if the sampled and target populations are the same.

Confidence intervals
Although a sample mean ($\bar{x}$) can be calculated, the exact location of the population mean (μ) remains unknown. Confidence intervals (CI) are used to estimate how far away the population means (μ) are likely to be with a given degree of certainty. A confidence interval gives an estimated range of values which is likely to include the unknown population parameter, it is therefore an interval estimate of the population parameter.

The estimator (formula) for an interval estimate (confidence interval) is:

Interval estimate (**CI**) =
Point estimate (**mean**) ± (**reliability coefficient**) × (**standard error**)

The quantity obtained by multiplying the reliability coefficient by the standard error of the mean is called the **precision** of the estimate (or **margin of error**)

- *For a sample drawn from a normal distribution with known variance , an interval estimate (CI) of the population mean is expressed as:*

$$CI = \bar{x} \pm Z_{(1-\alpha/2)} * s.e$$

Where: CI = Confidence interval; $\bar{x}$ = sample mean
$z_{(1-\alpha/2)}$= reliability coefficient; $s.e$ = standard error

Chapter 4

Hypothesis Testing

Hypothesis

A hypothesis is a proposition that is assumed as a premise in an argument or claim, or set forth as an explanation for the occurrence of some specified group of phenomena. A research hypothesis is used to test the relationship between two or more variables.

Hypotheses are assertions that are capable of being proven false using a test of observed data. The process of proving assertions false using a test of observed data (sample data) is called ***Hypothesis Testing.***

Expected attributes of a hypothesis

- It should be clear and precise
- It should be capable of being tested
- It should be able to relate to a variable
- It should be limited in scope and specific
- It should be stated in very simple terms
- It should be consistent with most known facts
- It should be testable within a reasonable time
- It should explain the facts which most need explaining

Types of hypotheses

Null hypothesis (Ho)

A null hypothesis is a type of conjecture used in statistics that proposes that there is no difference between certain characteristics of a population or data generating process. It typically corresponds to a general or default position. Making this assertion will make no difference and hence it cannot be proved positively.

It is the null hypothesis that is being tested in a test of statistical significance. Hypothesis testing allows the researcher to reject or not reject a null hypothesis at a certain level of significance (expressed as ***p-value***) but it cannot prove the hypothesis.

For example:
Ho: "All the students attending the statistics lecture are males". Only one female student suffices to reject the hypothesis as they enter the class consecutively, but no number of male students can prove it since the next student could be a female.

Alternative hypothesis (Ha)

An alternative hypothesis asserts a rival relationship between the phenomena measured by the null hypothesis. The alternative hypothesis proposes that there is difference between certain characteristics of a population or data generating process. It needs not be a logical negation of the null hypothesis as it only helps in rejecting or not rejecting the null hypothesis.

Types of statistical hypotheses tests

One-tailed (or one-sided) test

A one-tailed test is a statistical hypothesis test in which the values for which one can reject the null hypothesis (Ho) are located entirely in one tail of the probability distribution. In other words, the critical region for a one-tailed test is the set of values less than the critical value of the test, or the set of values greater than the critical value of the test (Figure 4.1).This indicates that a one-tailed hypothesis specifies the direction of the association between the predictor and outcome variables. A one-tailed test is also referred to as a one-sided test of significance.

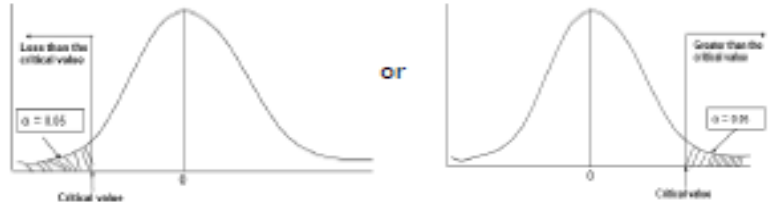


Figure 4.1: One-tailed (or one-sided) test

Two-tailed (or two-sided) test

A two-tailed (or two-sided) test is a statistical hypothesis test in which the values for which we can reject the null hypothesis (Ho) are located in both tails of the probability distribution.

In other words, the critical regions for a two-sided test is the set of values less than a first critical value of the test, and the set of values greater than a second critical value of the test (Figure 4.2). This indicates that a two-tailed hypothesis only states that an association exists between the predictor and outcome variables, but it does not specify the direction. A two-tailed test is also referred to as a two-sided test of significance.

Chapter 5

Chi-square (χ^2) Test

The chi-square (χ^2) test is a non-parametric test that measures how a model compares to actual observed data. It is performed to determine if there is a difference between the theoretical population parameter and the observed data. The chi-square statistic is a measure of the difference between the observed and expected frequencies of the outcomes of a set of events or variables.

Types of chi-square tests
The common types of chi-square tests include:

1. Pearson's chi-square test
2. Yate's continuity correction for Pearson's chi-square test (used in situations where the sample size is small, e.g., <40)
3. Fisher's exact test (used in situations where the assumptions for Pearson's chi-square test are not met, such as when more than 20% of the expected frequencies are less than 5)
4. McNemar's test (for paired observations)
5. Edward's continuity correction for McNemar's test (used when the sum of the discordant pair in the contingency table is less than 25)

1. Pearson's chi-square test
The Pearson's chi-square test is the most commonly performed chi-square test. It is used to determine whether there is a statistically significance difference between the expected frequencies and the observed frequencies in one or more categories of a contingency table.

The assumptions (i.e., conditions to be fulfilled) for Pearson's chi-square test are:

- Sample is a simple random sample from the population of interest (i.e., each individual in the population has an equal chance of being selected)
- Each observation is independent of all the others (i.e., one observation per subject)
- There are 2 variables, and both are measured at an ordinal or nominal level (i.e., categorical data)

- The data in the cells should be frequencies, or counts of cases rather than percentages, physical measurements or some other transformation of data (e.g., mean)
- No cell should have an expected frequency less than 1
- The expected frequencies should be 5 or more in at least 80% of the cells

If the Pearson's chi-square test conditions are not met, it is invalid; an option is the Fisher's exact test (for a 2 x 2 table).

The Pearson's chi-square test is used for three types of comparisons, these include:

a. Test of goodness of fit;
b. Test of independence; and,
c. Test of homogeneity

A: Chi-square goodness of fit test
The chi-square goodness of fit test is used to compare a randomly collected sample containing a single categorical variable to a larger population (most commonly the population from which it was potentially collected). The test enables us to observe how well the theoretical distribution (e.g., uniform, binomial, etc.) fits the observed distribution.

The hypothesis for the test is stated as:
Null hypothesis (H_0): The sample data is consistent with the theoretical distribution

Alternative hypothesis (H_A): The sample data is not consistent with the theoretical distribution

Example 5.1 The table below shows the number of new cases of COVID-19 on particular days in the week in a study conducted in a West African country.

Day	Mon	Tue	Wed	Thu	Fri	Sat	Sun
Number of new case of COVID-19	64	50	40	63	44	45	44

Can it be concluded that the data are a random sample from a uniform distribution?

Chapter 6

Comparison of Proportions

A proportion refers to the fraction of the total that possesses a certain attribute. If **x** is the number with a particular characteristic in a sample with size **n**, then the sample proportion (**p**) is given by the formula:

$$p = \frac{x}{n} \text{ i.e., } p = \frac{Number\ with\ the\ attribute\ or\ characteristic}{Sample\ size}$$

For example, if 30 of 60 women are married, the proportion of married women = 30/60 = 0.5.

A percentage is the amount, number or rate of something that is part of a total of 100. A percentage is often denoted by the character "%". To convert the proportion of married women to percentage requires multiplying it by 100.

The percentage of married women therefore equals
$\frac{30}{60} x\ 100 = 50\%$ (or 0.5 X 100 = 50%)

The proportion of married women = 30/60 = 0.5, while the percentage of married women = 50% (i.e., 50/100 when expressed as a fraction of 100). It therefore means that whereas, all percentages are proportions, it is not all proportions that are percentages (as it is only proportions that express fractions of a total of 100 that are percentages).

Comparison of a Sample Proportion with a Population Proportion
Comparison of a sample proportion with a population proportion can be done either by:

- Estimation of the population parameters (i.e., by estimating the Confidence Intervals of a proportion); or
- Hypothesis testing

Confidence Interval of a Proportion
Although a sample proportion (p) can be calculated, the exact location of the population proportion remains unknown. Confidence intervals (CI) are used to estimate how far away the population proportions are likely to be with a given degree of certainty. Confidence interval is therefore an interval estimate of the population parameter.

The estimator (formula) for an interval estimate (confidence interval) is:

Interval estimate (**CI**) =
Point estimate (**proportion**) $\pm$ (**reliability coefficient**) x (**standard error**)

The quantity obtained by multiplying the reliability coefficient by the standard error of the proportion is called the **precision** of the estimate (or **margin of error**)

- *For a sample drawn from a normal distribution with known variance, an interval estimate (CI) of the population proportion is expressed as:*

$CI = p \pm z_{(1-\alpha/2)} * s.e$

Where: CI = Confidence interval; p = sample proportion;
$z_{(1-\alpha/2)}$= reliability coefficient;
s.e = standard error of proportion

Conventionally, 95% confidence intervals are used, although they can be calculated for 90%, 99% or any other value.

Confidence interval = 1 – level of significance (α).

If the level of significance (α) is 0.05, then the corresponding confidence interval (CI) is 95%, i.e.,

CI = 1- 0.05 = 0.95 = 95%.

If the level of significance is 0.01, the CI = 99%; and if the level of significance is 0.1, the CI = 90%.

Chapter 7

Comparison of Means

The mean is the ideal measure of central tendency for a normally distributed data. Sets of normally distributed data can be compared by comparing their means. The main types of comparisons include:

1. Comparison of 2 means (using Z test or t-test)
2. Comparison of ≥3 means [using Analysis of Variance (ANOVA), in which the F-statistic is computed]

1. Comparison of 2 Means

The types of comparisons of 2 means include:

A. Comparison of a sample mean ($\bar{x}$) with a hypothetical population mean (μ) [using Z test or one sample t-test]
B. Comparison of the means of a paired observation in one sample (using paired t-test)
C. Comparison of the means of two independent samples (using independent t test)

1A. Comparison of a sample mean ($\bar{x}$) with a hypothetical population mean (μ) [using Z test or one sample t-test]

Comparison of a sample mean with a hypothetical population mean can be done by:

- Estimation of population parameter (using Z test or t test)
- Hypothesis testing (using Z test or t test)

CHOICE OF TEST (Z or t)

The choice of which one to use between Z test and t-test depends on the sample size (n) and whether or not the population standard deviation (σ) is known as shown in the reproduced Table 3.1 below.

Table 3.1: Choice of test (Z or t)

Population standard deviation (sd)	Sample size	
	n < 30	n ≥ 30
Known	Z test	Z test
Unknown	t test	t or Z test

Example 7.1 The medical risk associated with a certain occupation is being investigated and a random sample of 20 men aged 30 – 39 years has a mean systolic BP of 144.4mmHg with a standard deviation of 15.1mmHg. Does the evidence of our sample indicate if increased BP is associated with the occupation, if the true (population) mean systolic BP in such men is 133.2mmHg (Given that the level of significance α = 0.05)?

This question can be solved either by:

i. Estimation of confidence intervals, or
ii. Hypothesis testing

i. Answering the question by estimation of confidence intervals

To answer the question by estimation of confidence intervals, the first thing to do is to extract the relevant data in the question.

Sample size (n) = 20
Sample mean systolic blood pressure ($\bar{x}$) =144.4mmHg
Standard deviation for sample mean (sd) = 15.1mmHg
Population mean systolic blood pressure (μ) = 133.2mmHg

The next thing to do is to specify the appropriate test for estimating the confidence intervals, give the formula and the meaning of its components.

Since the population standard deviation is not known, and the sample size is less than 30 (i.e., n = 20), the appropriate test to use is the t-test.

$CI = \bar{x} \pm t_{(\alpha)} * s.e$

Where:
CI = Confidence interval
$\bar{x}$ = sample mean
$t_{(\alpha)}$= reliability coefficient
$s.e$ = standard error
sd = standard deviation
n = sample size

$$s.e = \frac{sd}{\sqrt{n}} \qquad s.e = \frac{15.1}{\sqrt{20}} \qquad s.e = 3.377$$

Given that the level of significance (α) is 0.05, then the corresponding confidence interval (CI) is 95%, i.e.,
CI = 1 – 0.05 = 0.95 = 95%

95% CI = $\bar{x} \pm t_{0.05} * s.e$

Chapter 8

Non Parametric Tests

Non parametric tests are methods of statistical analysis that do not require a distribution to meet the assumptions for a parametric test to be analyzed; as a result of this, they are sometimes called distribution free tests. The indications for using non-parametric tests include:

1. When the underlying data do not meet the assumptions about the population sample; for example, if the data is not normally distributed. However, in some cases, even if the data do not meet the assumptions for a parametric test, but the sample size is large enough, a parametric test can still be used instead of a non-parametric test (based on the central limit theorem).
2. When the sample size is small, in which case one may not be able to validate the distribution of the data, the only option therefore is to apply a non-parametric test.
3. When the data is nominal or ordinal; whereas, parametric tests are suitable for analyzing only continuous data, non-parametric tests are suitable for analyzing nominal or ordinal data.

The types of non-parametric tests and their corresponding parametric tests are shown in the reproduced Table 4.3.

A. One sample Wilcoxon signed-rank test

This is a non-parametric analogue to the one sample t-test. It is used when the data is not normally distributed. The one sample Wilcoxon signed rank test is used to compare the median of a sample with that of a hypothetical population.

Testing procedure

1. Calculate the differences between each individual value and the hypothesized median.
2. Rank the absolute values of the differences, from low to high, and affixes the sign of each difference to the corresponding rank (exclude those with differences = 0).
3. The rank assigned to tied observations is the mean of the ranks that would have been assigned to the observations had they not been tied.

4. Sum the ranks with a plus (+) sign, call it W_+
5. Sum the ranks with a minus (-) sign, call it W_-

Decision rule

Reject H_o if either W_+ or W_- is less than or equal to ($\leq$) the **critical value** $W_{\alpha(2)n}$ for a two tailed test or $W_{\alpha(1)n}$ for a one tailed test.

Table 4.3 Parametric and non-parametric statistical tests

Nature of groups	Type of variables	Parametric test (Purpose of test)	Non-parametric test (Purpose of test)
One group	Quantitative	Pearson's correlation (Test for relationship)	Kendall's tau, or Spearman rho correlation (Test for relationship)
One group compared with a population	Quantitative	1 sample t-test, or Z test (Compare means)	1 Sample Wilcoxon signed-rank test, or Sign test (Compare medians)
Two independent groups	Quantitative	Independent (or Unpaired) t-test (Compare means)	Mann-Whitney U test, or Wilcoxon rank sum test (Compare medians)
Two related Groups	Quantitative	Paired t-test (Compare means)	Wilcoxon matched pair signed-rank test (Compare medians)
Three or more Independent groups	Quantitative	ANOVA (Compare means)	Kruskal-Wallis rank sum H test (Compare medians)
Three or more repeated measures in one group	Quantitative	Repeated measures ANOVA (Compare means)	Friedman test (Compare medians)

Example 8.1 The table below shows the data from a study that assessed the systolic blood pressure of a sample of 10 patients attending the outpatient clinic of a teaching hospital in Northern Nigeria. Use the one sample Wilcoxon signed-rank test to demonstrate if the population median systolic blood pressure is greater than 120mmHg (Given that the level of significance α = 0.05).

Patient	1	2	3	4	5	6	7	8	9	10
Systolic BP (mmHg)	107	125	123	130	135	116	124	130	140	118

Chapter 9

Correlation and Regression

Correlation and linear regression are used for investigating the relationship between two quantitative variables. Correlation quantifies the strength of the linear relationship between a pair of variables, whereas regression expresses the relationship in the form of an equation. The first step in investigating a relationship between two quantitative variables is to show the data values graphically on a scatter diagram, but the strength of the relationship is measured by the means of an index known as **correlation coefficient** (represented by **r**).

In examining the relationship between 2 quantitative variables, it is necessary to know which one influences the other. For example in determining the association between height and weight, it is height that influences weight and not the other way round. It can thus be said that weight is dependent on height, but height does not depend on weight. Height is therefore the **independent variable** (and it is plotted on the **x axis**), while weight is the **dependent variable** (and it is plotted on the **y axis**) as shown in the scatter diagram (Figure 9.1).

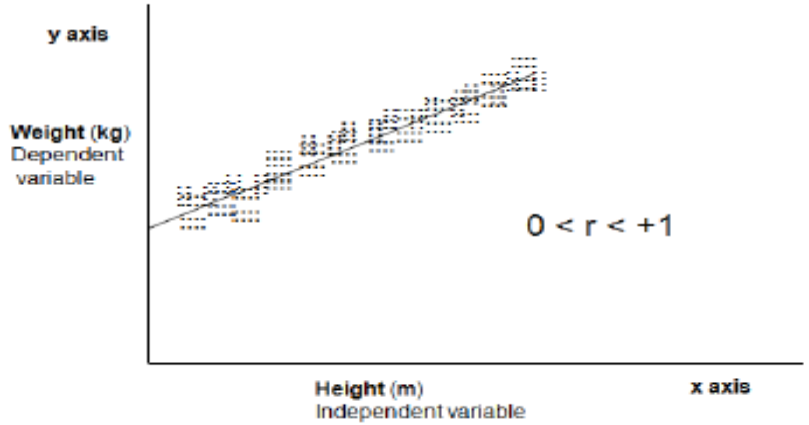


Figure 9.1: Positive correlation between weight and height

The correlation coefficient (r) is given by the formula:

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2] * [n\sum y^2 - (\sum y)^2]}}$$

Interpretation: (**r** ranges from **+1** to **0** and to **-1**)
r = +ve means: **Positive correlation** (or direct relationship). For example, there is a positive correlation between height and weight. This means that as the height increases, the weight also increases (Figure 9.1).

r = -ve means: **Negative correlation** (or inverse relationship). For example, there is a negative correlation between the immunity status and the risk of disease. This means that as the immunity status increases, the risk of disease decreases (Figure 9.2).

r = 0 means: **No correlation** (or no relationship). For example, there is no correlation between the students' admission numbers and their scores in an examination (Figure 9.3).

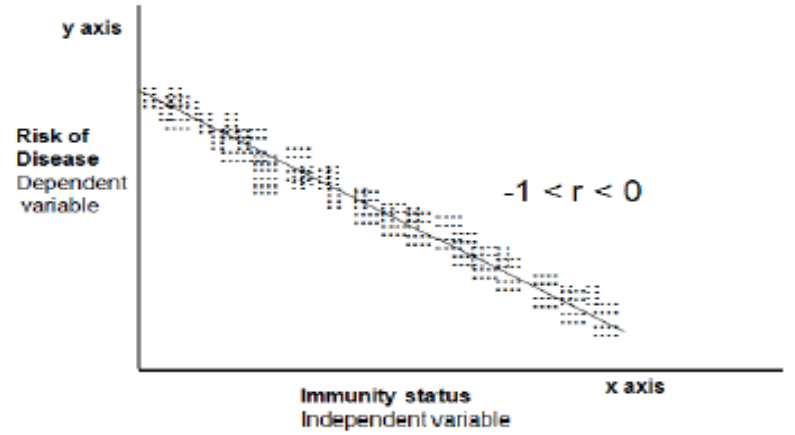


Figure 9.2: Negative correlation between immunity status and risk of disease

Chapter 10

Probability Theory

Probability is a mathematical technique for predicting outcomes. It predicts how likely it is that specific events will occur, and it is measured on a scale from 0 to 1. A probability can never be more than 1, nor can it be negative. The chance of obtaining a head when a coin is tossed is 50% (and the probability is expressed in decimal format as 0.5). The formula for determining the probability of an event occurring depends on the situation in which the event is expected to occur.

A. To determine the probability (P) of a single event (A) happening
The probability (P) of a single event (A) happening is given by the formula:

$$P(A) = \frac{the\ number\ of\ possible\ events}{the\ number\ of\ possible\ equally\ likely\ outcomes}$$

Example 10.1: Out of 25 students in a class, 5 are females. Find the probability of selecting a female in the class.

The number of possible events (i.e., the number of females available for selection) = 5

The number of possible equally likely outcomes (i.e., the total number of students available for selection) = 25

The probability of selecting a female in the class is computed as:

$$P(Female) = \frac{the\ number\ of\ possible\ events}{the\ number\ of\ possible\ equally\ likely\ outcomes}$$

$$P(Female) = \frac{5}{25}$$

$$P(Female) = \mathbf{0.2}$$

B. To determine the probability of event (A) and event (B) happening (independent events)
The probability of event (A) and event (B) happening (if they are independent) is given by the formula:

$P(A \cap B) = P(A) * P(B)$ [This is called "**the multiplication rule**"]

Example 10.2: Out of 25 students in a class, 5 are females. Find the probability of selecting a female and a male in the class.

Solution:
The number of possible events for females (i.e., the number of females available for selection) = 5

The number of possible events for males (i.e., the number of males available for selection) = 20

The number of possible equally likely outcomes (i.e., the total number of students available for selection) = 25

The probability of selecting a female in the class is computed as:

$$P(Female) = \frac{5}{25} = 0.2$$

The probability of selecting a male in the class is computed as:

$$P(Male) = \frac{20}{25} = 0.8$$

The probability of selecting a female and a male (independent events) is computed as:

$$P(Female \cap Male) = P(Female) * P(Male)$$

$$P(Female \cap Male) = 0.2 * 0.8$$

$$P(Female \cap Male) = \mathbf{0.16}$$

Example 10.3: If at age 60, the probability of surviving for the next five years is 0.7 for an African man and 0.8 for an African woman. For an African couple who are both aged 60, what is the probability that the wife will be a living widow five years later?

Chapter 11

Sample Size Estimation and Sampling Techniques

Sample size estimation

A sample is a part of the population that is being selected as a representative of the population. In some cases (e.g. census, very small finite populations, etc.) the whole population is studied, but in research due to time constraint and budget only a representative sample is needed. Sample size estimation is the mathematical computation of the number of subjects or units to be included in a study. When a representative sample is taken from a population and the subjects are selected by probability sampling techniques the findings of the study can be generalized to that population; and the study is said to have external validity.

For the findings of a study to be valid the sample size must not be too large or too small. If the sample size is too large, so much time and resources will be wasted, and the findings may still not be reliable as the risk of alpha error becomes high (in which case an effect that is not actually clinically significant appears significant statistically). On the other hand, if the sample size is too small the risk of beta error becomes high as the study may fail to detect an important effect or association because the power is poor. An optimum sample size is therefore required for a quantitative research as it allows for appropriate analysis, provides the desired level of accuracy and enables valid conclusions.

Factors influencing sample size estimation

The appropriate formula to be used in computing the optimum sample size for a study depends on many factors including the study design (cross-sectional case control, cohort, experimental, etc.), the nature of the outcome variable (qualitative or quantitative), consideration for type I and type II errors, the required margin of error (or precision), the minimum effect size considered to be significant, the standard deviation of a continuous variable, the odds ratio and relative risk in case control and cohort studies respectively, and the design effect.

1. Study design

A study design is a specific plan for conducting the study which allows the investigator to translate the conceptual hypothesis into an operational one. The various study designs can be categorized into 3 groups, namely observational studies (in which the researcher observes phenomena and takes records of the observations), experimental studies (in which the researcher manipulates independent or predictor variables and document their effects on dependent or outcome variables), and research synthesis (which includes systematic review and meta-analysis) as shown in Figure 11.1

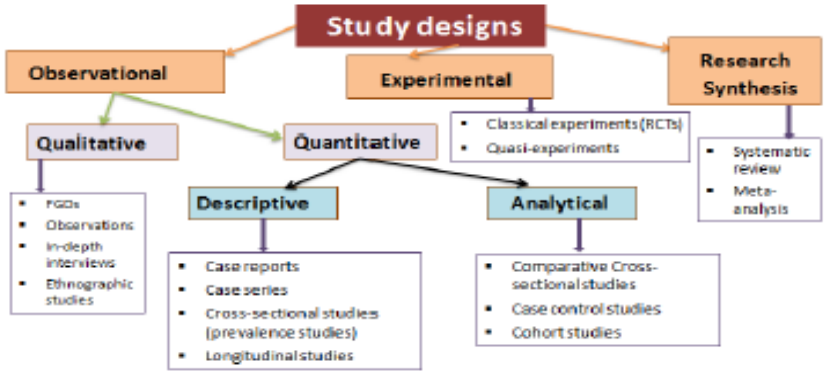


Figure 11.1: Study designs
Source: Author

2. Nature of the outcome variables

The outcome variables essentially belong to 2 groups, namely qualitative and quantitative. Qualitative variables are those that are classified by attributes or categories e.g., marital status (single, married, separated, etc.), and they are measured on nominal and ordinal (or ranking) scales. Quantitative variables are those that results from counting or measurement, e.g., number of students in a class, weight, height, etc., and they are measured on interval and ratio scales.

Chapter 12

Methods of Data Presentation

Data presentation refers to the methods of communicating the information generated following data analysis to other people. Data should be presented in concise, simple, and easy to understand forms, but must contain important details, and also stimulate interest in readers.

The methods of presenting data include:

- **Textual:** This involves presenting data using a combination of words and numbers
- **Tabular:** This involves presenting data using tables (with the contents arranged in columns and rows)
- **Graphical:** This involves presenting data using charts or figures

However, the appropriate data presentation method to be used depends on the study design that was employed, the types of variables involved, the readers concerned, and the purpose.

Textual Presentation of Data

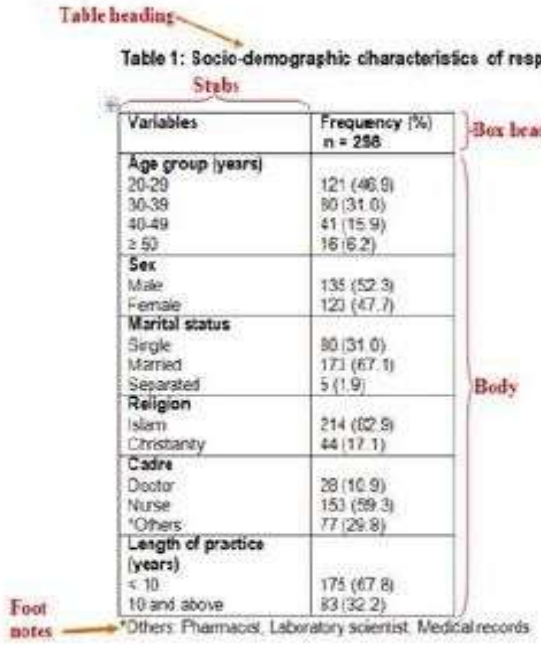
This is the most commonly used data presentation method in all types of studies. It involves presenting data using a combination of words and numbers in paragraphs or sentences. It is used to explain, interpret or emphasize the data being presented. Whereas, it is mostly used in combination with other methods (i.e., Tabular and Graphical) in quantitative studies, in qualitative studies it is usually the only data presentation method employed. In descriptive observational studies summary measures of central tendency and dispersion for quantitative variables are often presented textually, whereas frequencies and percentages for qualitative variables are often presented using tables and charts.

Tabular Presentation of Data

This is a popular method of presenting data in quantitative observational studies, experimental studies and research synthesis. It involves presenting data using tables (such as frequency distribution tables) and the contents are arranged in columns and rows. The tables contain the details of the data being presented in the text, while the text usually contains only the highlights of the data presented in the tables.

A frequency distribution table consists of rows that read from left to right and columns that read from top to bottom.
The parts of a table include (Figure 12.1):

- **Table heading:** This consists of the table number and title, and it should be above the table
- **Stubs:** This is the column that contains the main variables being considered in the table. It is usually at the extreme left and the heading of the column depends on whether the items are homogenous or not
- **Box head :** This refers to the column header
- **Body:** This refers to the main part of the table
- **Footnotes:** Additional details of symbols or abbreviations used in the table are inserted below the table as footnotes
- **Source note:** The source(s) of tables copied from somewhere else should also be indicated below the table



Variables	Frequency (%) n = 258
Age group (years)	
20-29	121 (46.9)
30-39	80 (31.0)
40-49	41 (15.9)
≥ 50	16 (6.2)
Sex	
Male	135 (52.3)
Female	123 (47.7)
Marital status	
Single	80 (31.0)
Married	173 (67.1)
Separated	5 (1.9)
Religion	
Islam	214 (82.9)
Christianity	44 (17.1)
Cadre	
Doctor	28 (10.9)
Nurse	153 (59.3)
*Others	77 (29.8)
Length of practice (years)	
< 10	175 (67.8)
10 and above	83 (32.2)

Figure 12.1: The Components of a Table
Source: Author

Table of Random Numbers

7579	2550	2487	9477	0864	2349	1012	8250
3554	5080	9074	7001	6249	3224	6368	9102
6895	3371	3196	7231	2918	7380	0438	7547
5634	5323	2623	7803	8374	2191	0464	0696
7803	8832	5119	6350	0120	5026	3684	5657
1428	1796	8447	0503	5654	3254	7336	9536
4534	2105	0368	7890	2473	4240	8652	9435
5144	7649	8638	6137	8070	5345	4865	2456
1277	6316	1013	2867	9938	3930	3203	5696
0951	5991	5245	5700	5564	7352	0891	6249
2179	4554	9083	2254	2435	2965	5154	1209
2972	9885	0275	0144	8034	8122	3213	7666
1341	9860	6565	6981	9842	0171	2284	2707
5291	2354	5694	0377	5336	6460	9585	3415
2626	5238	5402	7937	1993	4932	2327	0875
1947	6380	3425	7267	7285	1130	7722	0164
0653	3645	7497	5969	8682	4191	2976	0361
6938	4899	5348	1641	3652	0852	5296	4538
8797	8000	4707	1880	9660	8446	1883	9768
4219	0807	3301	4279	4168	4305	9937	3120
1192	1175	8851	6432	4635	5757	6656	1660
7702	6958	9080	5925	8519	0127	9233	2452
1730	5005	1704	0345	3275	4738	4862	2556
1257	6163	4439	7276	6353	6912	0731	9033
4260	5277	4998	4298	5204	3965	4028	8936

Chi-square (χ^2) Distribution

(For χ^2 test and Kruskal-Wallis H-test)

Degrees of freedom	Two-tailed probability (P)					
	0.2	*0.1*	*0.05*	*0.02*	*0.01*	*0.001*
1	1.642	2.706	3.841	5.412	6.635	10.827
2	3.219	4.605	5.991	7.824	9.210	13.815
3	4.642	6.251	7.815	9.837	11.345	16.268
4	5.989	7.779	9.488	11.668	13.277	18.465
5	7.289	9.236	11.070	13.388	15.086	20.517
6	8.558	10.645	12.592	15.033	16.812	22.457
7	9.803	12.017	14.067	16.622	18.475	24.322
8	11.030	13.362	15.507	18.168	20.090	26.125
9	12.242	14.684	16.919	19.679	21.666	27.877
10	13.442	15.987	18.307	21.161	23.209	29.588
11	14.631	17.275	19.675	22.618	24.725	31.264
12	15.812	18.549	21.026	24.054	26.217	32.909
13	16.985	19.812	22.362	25.472	27.688	34.528
14	18.151	21.064	23.685	26.873	29.141	36.123
15	19.311	22.307	24.996	28.259	30.578	37.697
16	20.465	23.542	26.296	29.633	32.000	39.252
17	21.615	24.769	27.587	30.995	33.409	40.790
18	22.760	25.989	28.869	32.346	34.805	42.312
19	23.900	27.204	30.144	33.687	36.191	43.820
20	25.038	28.412	31.410	35.020	37.566	45.315
21	26.171	29.615	32.671	36.343	38.932	46.797
22	27.301	30.813	33.924	37.659	40.289	48.268
23	28.429	32.007	35.172	38.968	41.638	49.728
24	29.553	33.196	36.415	40.270	42.980	51.179
25	30.675	34.382	37.652	41.566	44.314	52.620

NB: If the distribution being used in a test is symmetric, then one-sided corresponds with one-tailed. In the case of distributions which are not symmetric (such as Chi-square and F tests), the standard tests use only one tail, but are two sided or non-directional. All chi-square tests with degrees of freedom (df) > 1 are two sided or non-directional.

The "Two-tailed probability" indicated on the table is actually the "Upper tail probability" (i.e., for the areas to the right of the critical value). If df =1, and the alternative hypothesis (H_A) is directional (e.g., Upper tailed), the corresponding "Two-tailed probability" obtained should be multiplied by 0.5 when determining the p value for the test. To obtain the "Lower tail probability" subtract the Upper tail probability from 1 (e.g., if the Upper or Rt tail probability = 0.05, the Lower or Lt tail probability = 1-0.05 = 0.95)