APPLIED STATISTICS: Data Analysis

VOLUME I: ANALYSIS

Master data analysis with a simple and effective method that allows for fast results and autonomy in your studies

1. **Simplified Essential Concepts**
2. **Illustrated step-by-step data analysis**
3. **The best free software for your analysis**

LEARN STATISTICS EASILY

MASTER DATA ANALYSIS QUICKLY, EFFORTLESSLY, AND WITH UNDENIABLE RESULTS

Discover our SIMPLE teaching method that will empower you to analyze your data on your own in no time.

We cover ALL the essential steps and only what's truly necessary for analyzing your data.

Built on the proven principle that it's entirely possible to accurately analyze data without complex concepts or formulas.

This book will serve you well, even if you have no prior knowledge of statistics. *All 3 volumes are included in this offer.*

GET YOURS BY CLICKING HERE!

FORGET EVERYTHING YOU HAVE EVER LEARNED ABOUT STATISTICS TO ANALYZE YOUR DATA

Applied Statistics: Data Analysis (Volume I - Analyses) / Alves, Ana. - Learn Statistics Easily, 2023. 196 p. 2nd ed.

1. Statistics. 2. Biostatistics. 3. Learn Statistics Easily. 4. Applied statistics. 5. Data analysis. 6. Inferential statistics. 7. Descriptive statistics. 8. Graphs. 9. Sample size.

Learn Statistics Easily
statisticseasily.com

PREFACE

Welcome to your ultimate guide to unlocking the power of data analysis – quickly, easily, and confidently.

This book presents a groundbreaking teaching method that empowers you to independently analyze your data with speed and precision.

We've distilled the essentials, providing only the necessary information to conquer data analysis without getting lost in complexities.

Say goodbye to intimidating concepts, formulas, and tables. This guide is designed to benefit you, even if your knowledge of statistics is limited.

Our innovative approach to "*learning data analysis quickly, easily, independently, and with confidence*" sets this book apart from the rest.

Let this guide be your invaluable companion as you embark on the exciting data analysis journey.

MASTERING OUR METHODOLOGY

(a) We distill only the most vital concepts, making them effortlessly understandable.

(b) Crystal-clear examples and diagrams bring each concept to life.

(c) Our algorithm for selecting statistical analyses and graphs is streamlined and straightforward.

(d) We tackle the most prevalent statistical analyses, covering 99% of real-world scenarios.

(e) Our detailed, step-by-step instructions, paired with vivid illustrations, make data analysis a breeze to grasp.

(f) Experience the ultimate in user-friendly, comprehensive, and intuitive free statistical software.

SUMMARY

1. **GETTING STARTED: ESSENTIAL KNOWLEDGE**
Grasp the key concepts in a simplified and accessible manner.

2. **TOP FREE STATISTICAL SOFTWARE**
Discover premier tools for data analysis, graphing, spreadsheets, and sample size calculations.

3. **DESCRIPTIVE STATISTICS: SUMMARY MEASURES**
Dive into the most vital measures for summarizing and showcasing your data.

4. **INFERENTIAL STATISTICS: UNLEASHING DATA ANALYSIS**
Learn to choose the right analysis and apply it with precision.

5. **PICKING THE PERFECT GRAPH (VOL. II)**
Follow a step-by-step guide to selecting and creating the ideal graph for your data.

6. **BONUS CONTENT & ADVANCED TOPICS (VOL. III)**
Delve into extra tips and explore slightly more sophisticated subjects.

TABLE OF CONTENTS - VOLUME I

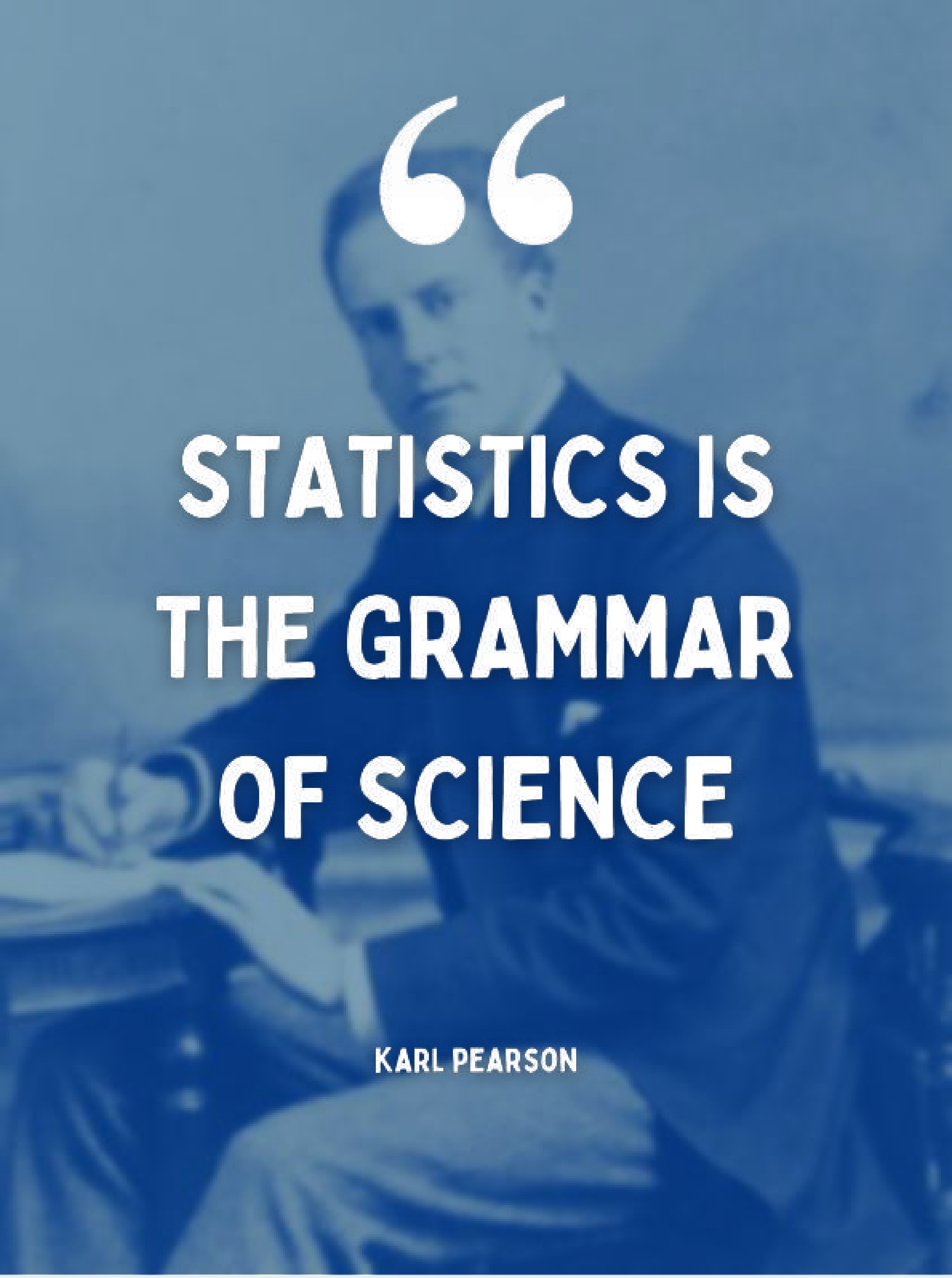
TABLE OF CONTENTS – VOLUME I

CHAPTER 4: INFERENTIAL STATISTICS ANALYSIS

LOOKING FOR DIFFERENCES BETWEEN UNPAIRED GROUPS

LOOKING FOR DIFFERENCES BETWEEN PAIRED GROUPS

LOOKING FOR RELATIONSHIPS BETWEEN VARIABLES


STATISTICS IS THE GRAMMAR OF SCIENCE
KARL PEARSON

CHAPTER 1

WHAT SHOULD I KNOW TO GET STARTED?

1. THE STATISTICS

Statistics is a science that deals with collecting, analyzing, interpreting, and presenting data.

It helps in **decision-making**, even under conditions of uncertainty.

There are several **areas of knowledge** whose statistical methods are widely used, such as biostatistics, which applies statistical tools to problems related to life and health sciences, such as medicine, biology, ecology, etc.

A **fundamental rule** to follow is that *statistics should simplify, not complicate, the interpretation of data*.

For **example**, suppose your analyses complicate the understanding of the data — in that case, something needs to be corrected and revisited.

5. VARIABLE & ITS TYPES

The elements of a population may exhibit numerous **conditions** or **characteristics** that we can observe, count, or measure.

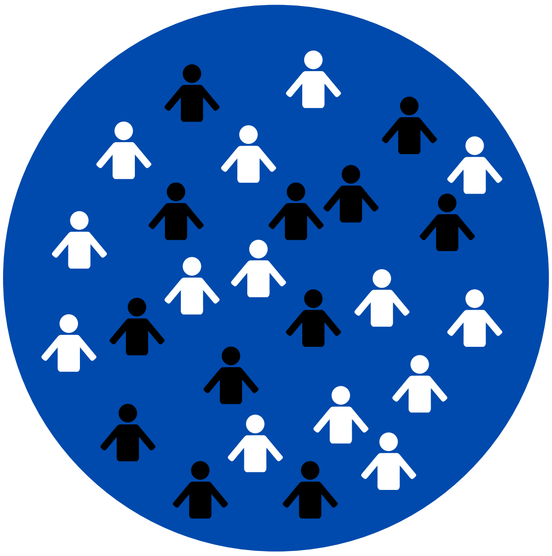
These conditions or features are called **variables**.

The following are some **examples** of variables:

(a) Eye color *in individuals of the human species.*
(b) Weight *in individuals of the species Canis familiaris.*
(c) The offspring number *in a nest of a particular bird species.*
(d) The mean temperature *in cities of Texas.*
(e) Marital status *of individuals in New York.*
(f) Schooling level *of North American adults.*

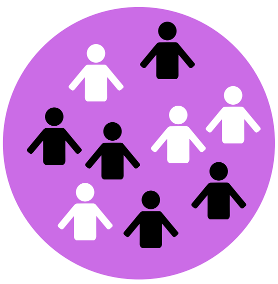
RELATIONSHIP

POPULATION: SAMPLE: ELEMENT: VARIABLE



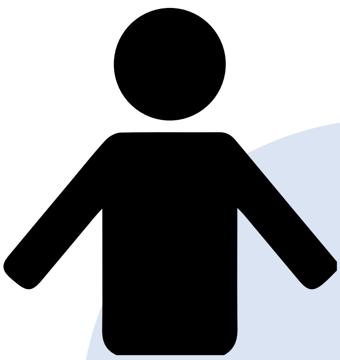
POPULATION

The total group of elements



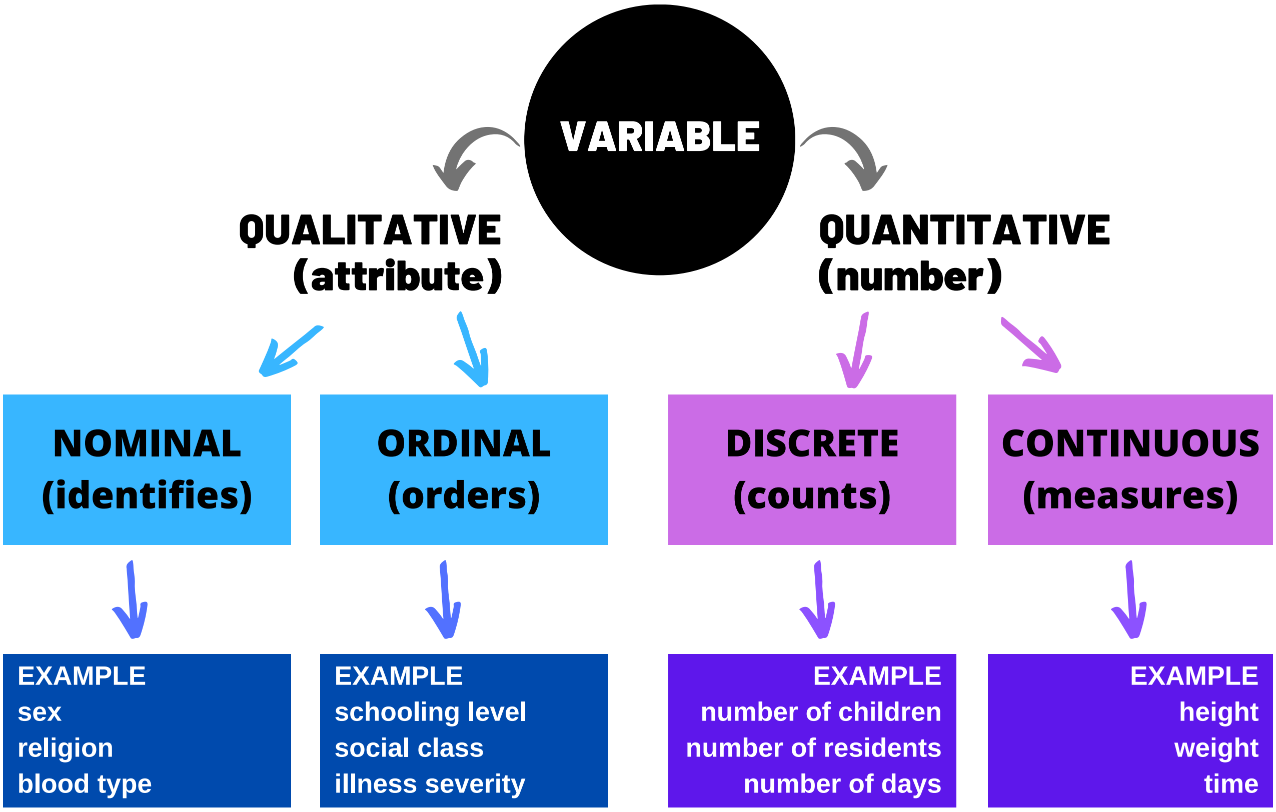
SAMPLE

A subset of elements selected from the population for analysis



ELEMENT

From each individual or unit in the sample, we collect information on observable characteristics or conditions, such as weight, height, eye color, age, body temperature, days of hospitalization, etc. These characteristics or conditions are known as variables.



Although not always explicitly stated, quantitative variables can be **transformed** into qualitative ones.

For **example**, consider the height of adult males, which can range from 4.0 to 7.0 ft. When data is collected from the population, any value within this range can be found.

However, suppose the raw, quantitative data is not desired. In that case, the variable can be made qualitative by categorizing it into groups, such as short (4.0-5.0 ft), medium (5.0-6.0 ft), and tall (6.0-7.0 ft).

It's important to note that this transformation results in a **loss of information.**

7. NORMALITY & PARAMETRIC TESTS

The **normal (or Gaussian) distribution** represents one of the most widely used statistical probability distributions.

It is used because many **natural phenomena** behave similarly to it.

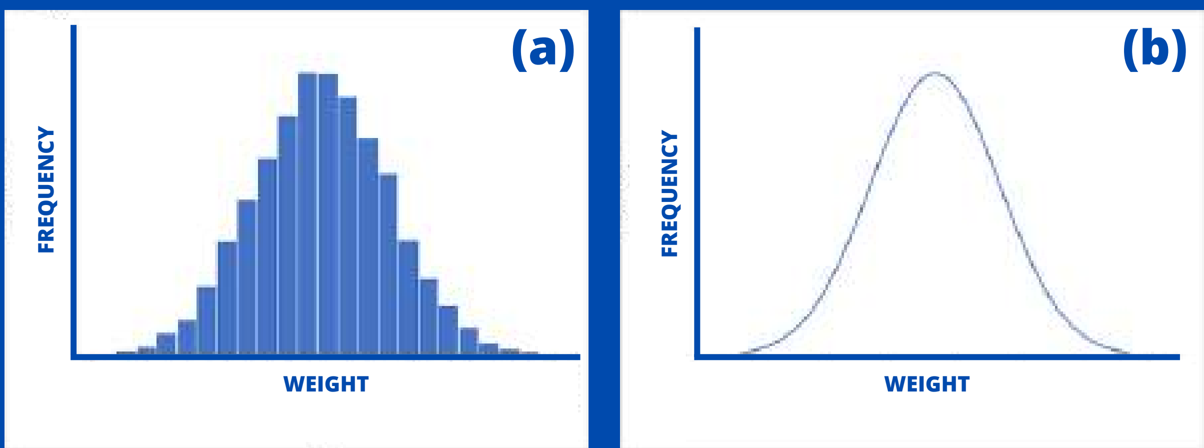
What happens if we plot a normally distributed quantitative variable's frequencies?

When plotted, a variable's frequencies that follow the normal distribution will have a **bell-shaped curve**.

See the following **example**.

EXAMPLE: NORMAL DISTRIBUTION

Weight, in lbs, of 10,000 randomly selected persons.
(a) Histogram showing the frequency of the weights.
(b) Normal curve fitted from the weight data.



Parametric analyses assume that the data or residuals are adjusted to the normal distribution.

If this assumption is violated, **nonparametric** analyses, which do not have the normality prerequisite, should be used instead.

Generally speaking, there is an **equivalent** nonparametric analysis for every parametric analysis.

For **example**, the Mann-Whitney U-test is the nonparametric analysis corresponding to the parametric independent samples Student's t-test.

Specific inferential tests can also be used to **check** if the normality assumption is violated.

EXAMPLE #2: ESTABLISHING THE CONCEPTS

DO MEN HAVE AN AVERAGE IQ (INTELLIGENCE QUOTIENT) DIFFERENT FROM WOMEN IN CANADA?

In this example, each person in Canada represents one **element** of the target population. So, we would have everyone living in Canada as our target **population**.

The **variables** of interest to be extracted from the sampled elements will be sex and IQ. **Sex** represents a **qualitative** variable with two attributes (male and female), and **IQ** represents a **quantitative** variable.

Since sex would define IQ and not the other way around, the **independent variable (X)**, the cause, is sex. The **dependent variable (y)**, the effect, is IQ.

We must obtain both variables from **each** sampled person.

Since we have an independent qualitative variable (X) with two groups and a quantitative dependent variable (y) that we want to check for differences between groups, we can use an independent samples Student's t-test as the **statistical analysis**.

We need to define the **sampling method** and make a **sample size** calculation to determine the appropriate number of elements to sample.

We need to check the normality assumption for the correct definition of the statistical analysis.

CHAPTER 2

THE BEST FREE STATISTICAL SOFTWARE

1. THE SECRET

After years of research and testing, we have identified **the best** free statistical software for data analysis, sample size calculation, spreadsheeting, and graphing. Our recommendations are based on our experience testing a wide range of software.

The importance of these recommendations has grown over time and is now more valuable than ever. The best part is that all the software we will use in this guide is **free**.

All of the software we recommend has a **user-friendly** interface and does not require any knowledge of command lines, making it easy to use.

With the software we recommend in this guide, you will have everything you need to effectively analyze your data. In addition, we provide tools for data analysis, sample size calculation, spreadsheeting, and graphing, and these programs will cover **99%** of your data analysis needs.

3. SAMPLE SIZE

G*Power is a powerful and free statistical tool used for sample size calculation.

In addition to other functions, such as power and effect size analysis, GPower is the ideal tool for determining your study's **sample size**.

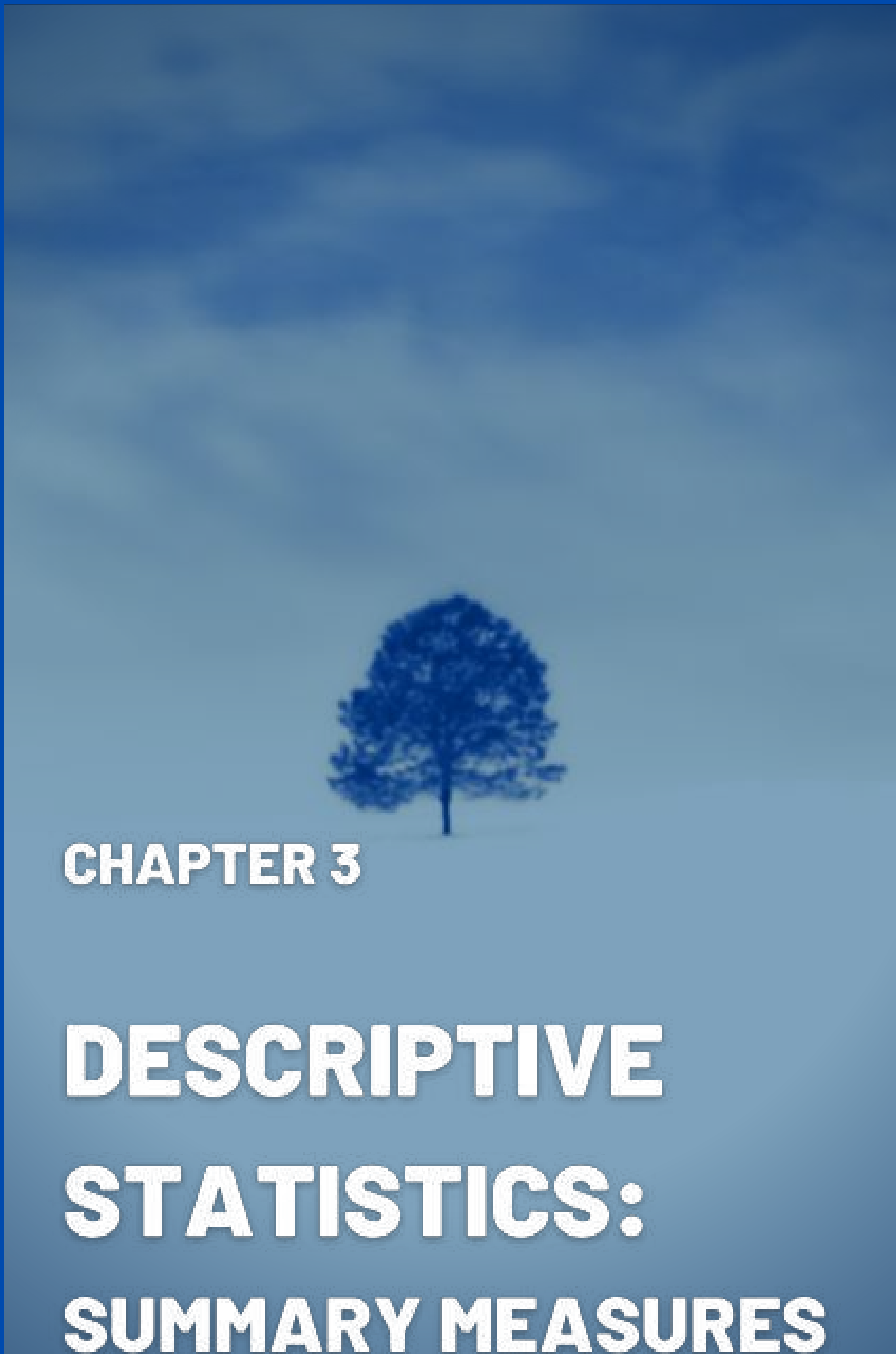
It includes all the **analyses** we will cover in this guide, such as t-tests, ANOVA, linear regression, etc.

To access G*Power, visit its official **website**: bit.ly/gpowerstat

Versions are available for Windows and macOS.

Once you are on the **website**, scroll down to the "Download" section, and select the compatible version with your operating system.

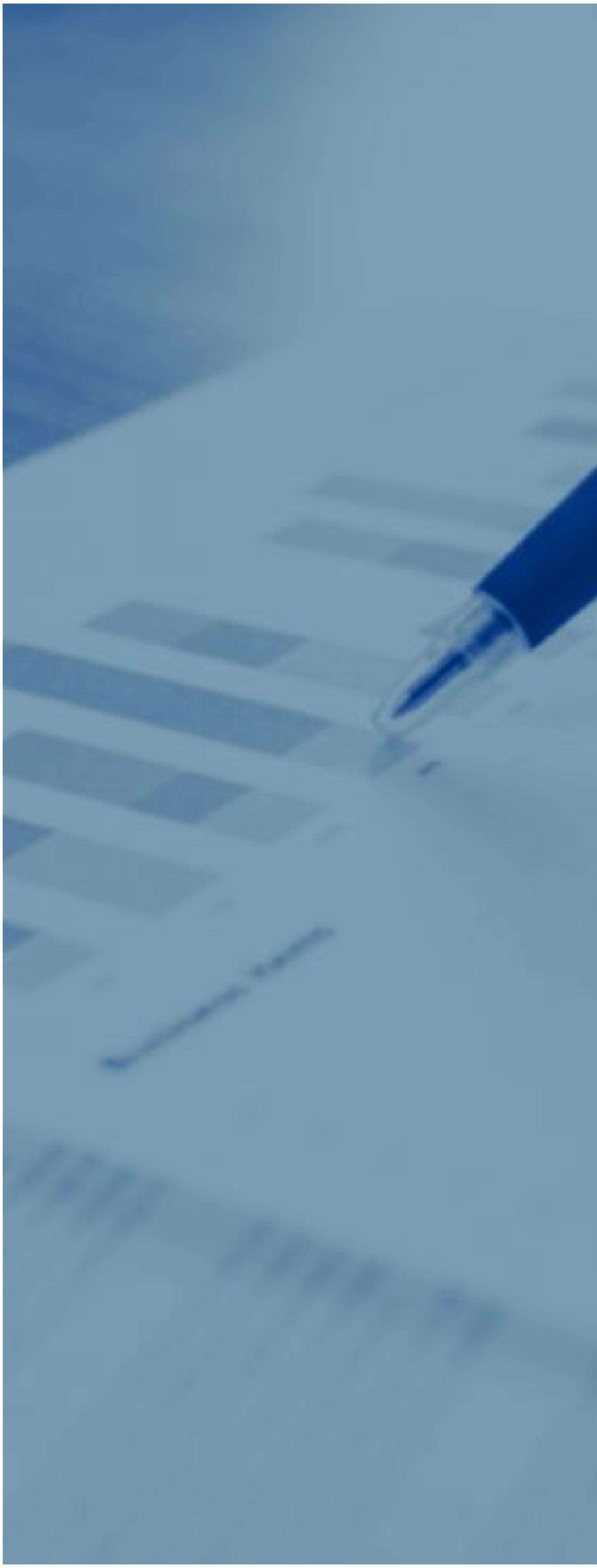
Download and **install** G*Power, and we will use it in the guide's following sections.



CHAPTER 3

DESCRIPTIVE STATISTICS: SUMMARY MEASURES

1. DESCRIPTIVE STATISTICS



Descriptive statistics is a **set of methods**, also known as simple statistics, that aims to make collected data **easier to understand** by:

(a) organizing,

(b) simplifying,

(c) describing, and

(d) presenting the data.

It uses tables, graphs, and measures that summarize the raw data.

2. MEASURES OF CENTRAL TENDENCY

Measures of central tendency are used to find the value representing the **center or middle** of a dataset — they aim to identify the closest value to all other values in the dataset. These measures are also referred to as measures of central **location** or **position**.

The two most commonly used for data analysis are:

(a) the simple arithmetic mean and

(b) the median.

While there are other measures of central tendency, this guide will only cover the most important ones.

2.2
MEDIAN

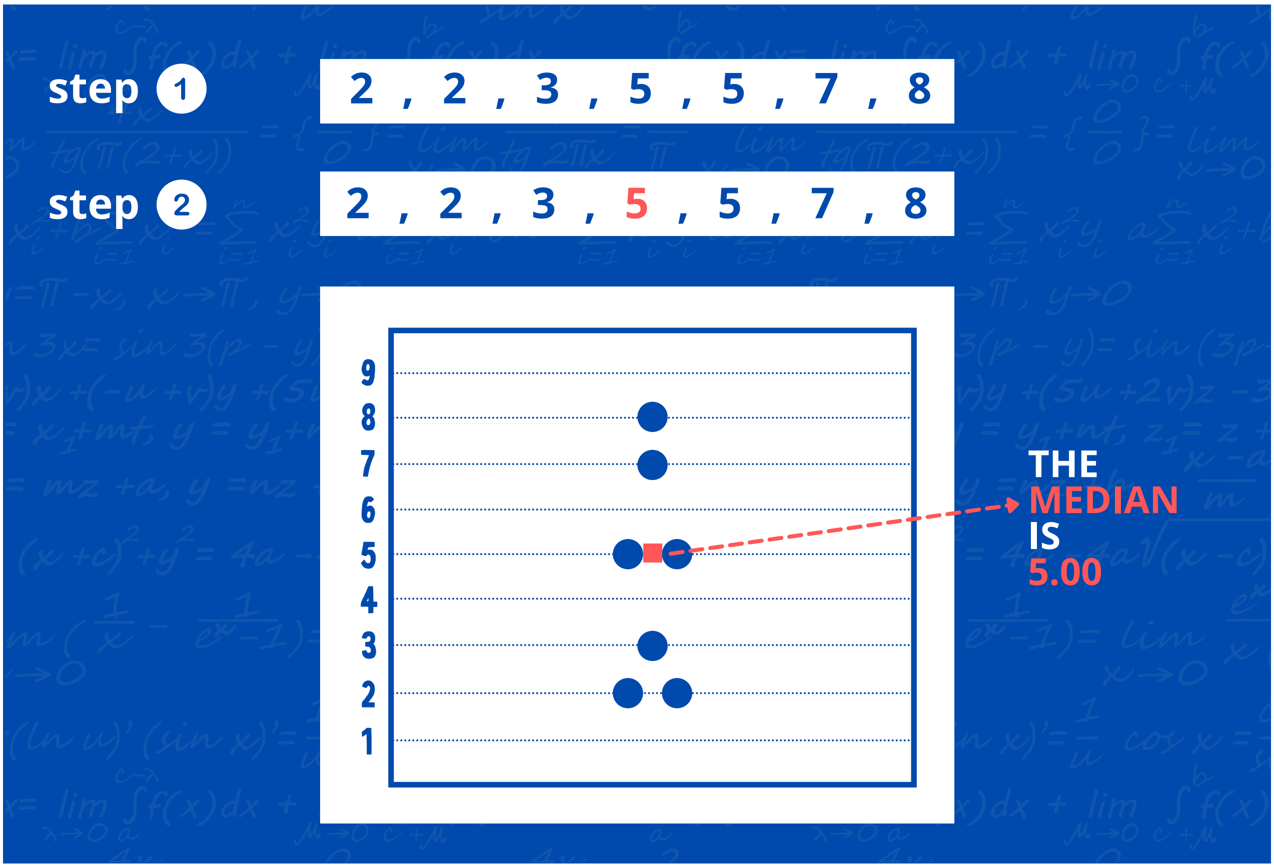
The median represents the middle value in an ordered series of observations.

It is often used in place of the simple arithmetic mean when the data does not fit a normal distribution.

THE STEPS ARE AS FOLLOWS:

1. Sort the observations in order.

2. The median is the center value (*if there are two center values, their mean is the median*).

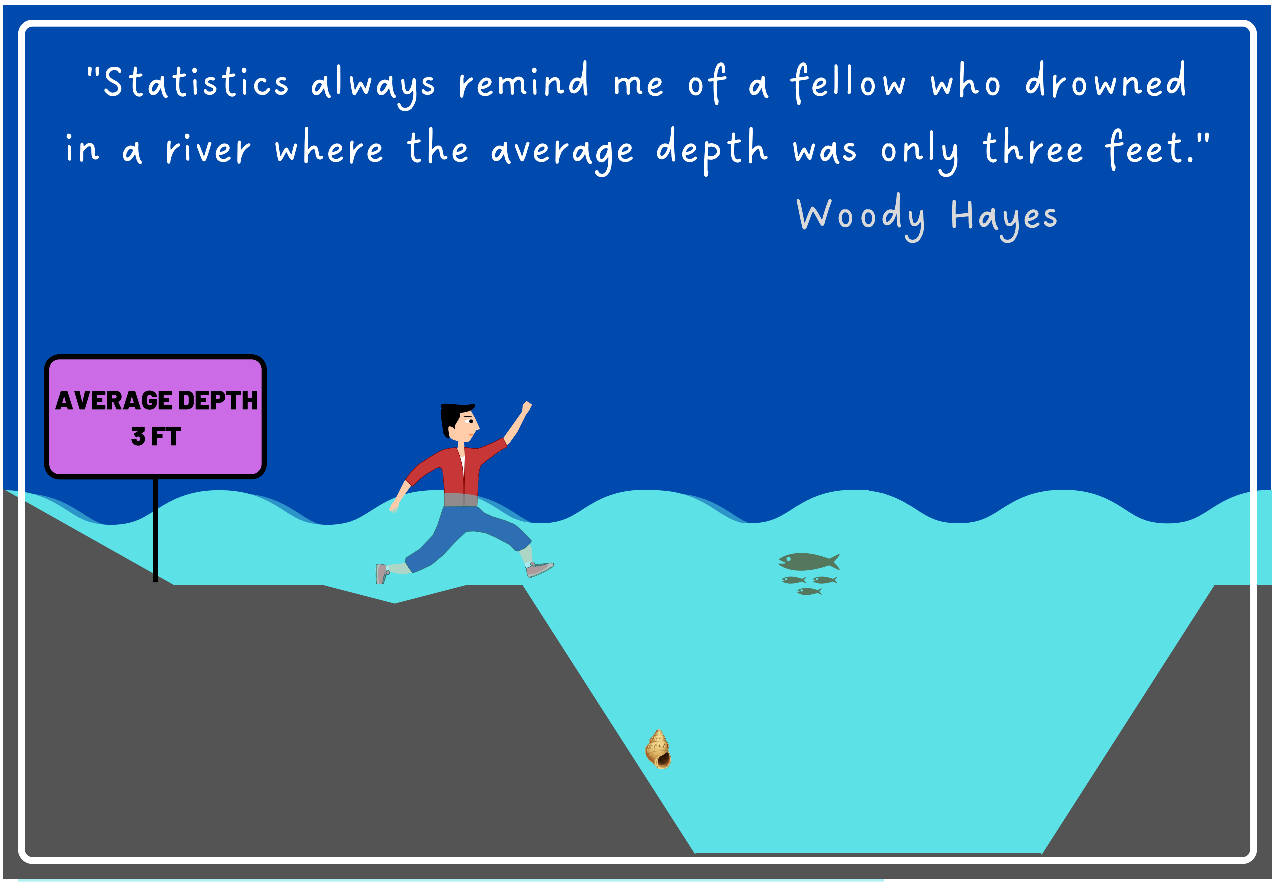


3. MEASURES OF DISPERSION/VARIABILITY

Measures of **central tendency** provide a single value that represents a dataset.

Still, they do not show how much the values in the dataset **vary** from each other.

That's why we use measures of variability to show the **difference** between all the values in a given dataset.



The **importance** of using a variability measure with a central tendency measure becomes evident in the following example.

Consider the following **sets**:

X = {50, 50, 50, 50, 50}

Y = {48, 49, 50, 51, 52}

Z = {10, 20, 50, 80, 90}

The simple arithmetic mean of each set is the same (50), but the variability is quite **different**.

Therefore, it is essential to **use both** measures together as they provide a more comprehensive understanding of the data.

3.1
RANGE

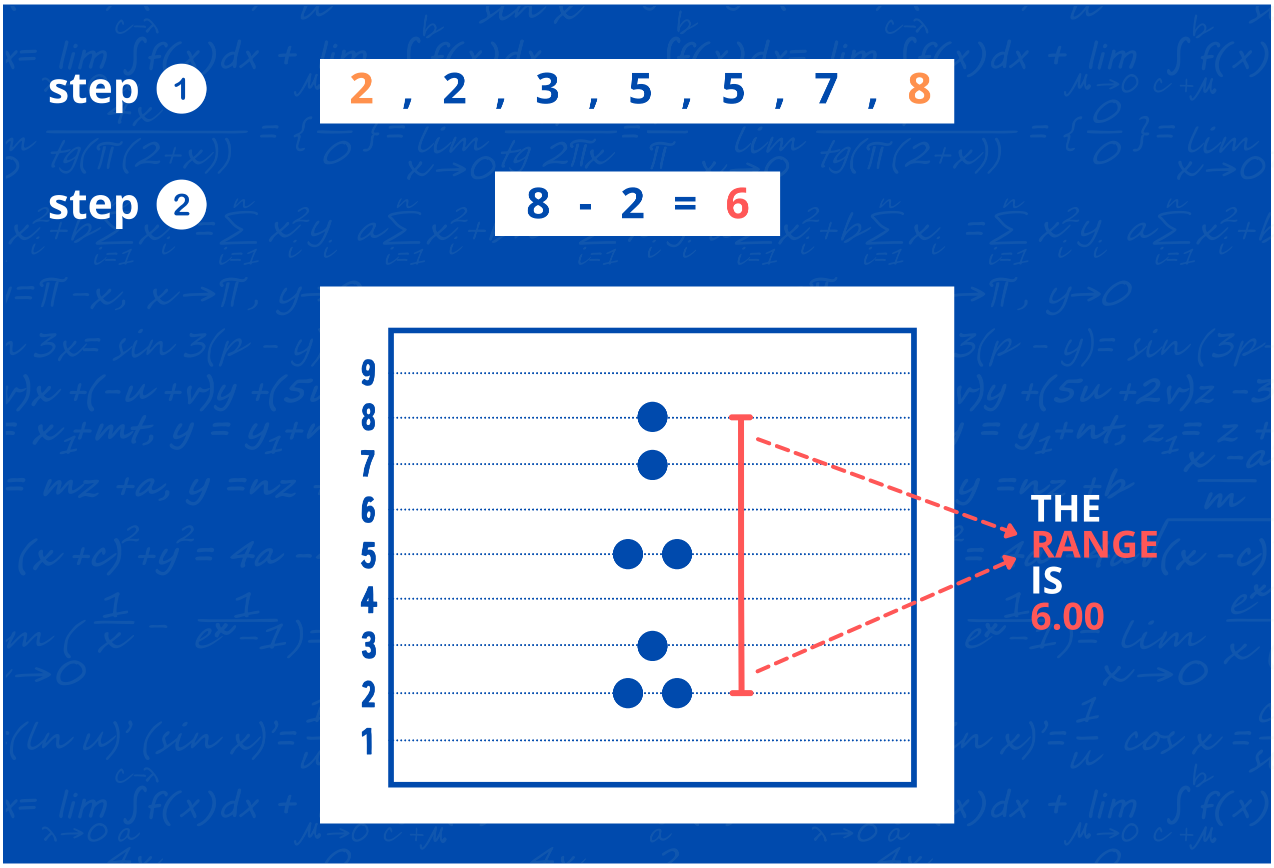
The range is a straightforward and intuitive measure of variability.

It is calculated as the difference between the highest and lowest values in a series of observations. *It is widely used in box plots with the median and other quartiles.*

THE STEPS ARE AS FOLLOWS:

1. Identify the highest and lowest values in the dataset.

2. Subtract the lowest value from the highest value.



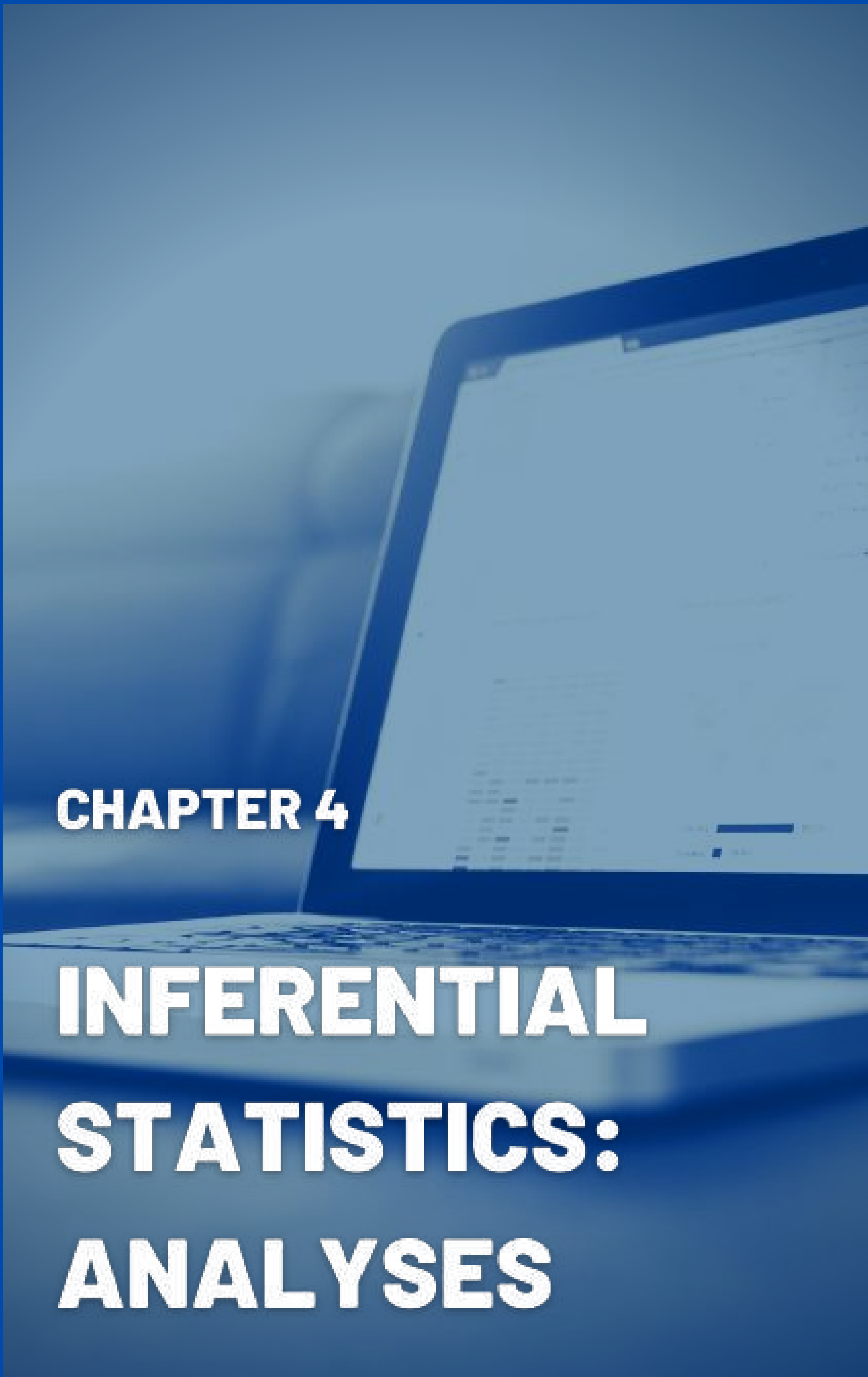
4. SUMMARIZING

This chapter covered the basics of descriptive statistics and the most commonly used **summary measures**. These include measures of central tendency, such as the mean and median, and measures of dispersion or variability, such as the range, standard deviation, and interquartile range.

It's important to note that while we have discussed the **most widely used** measures, there are other less commonly used measures, such as the mode. Therefore, we have chosen to focus on the most important and widely used measures for simplicity and practicality.

When analyzing data, it's essential to use both central tendency and dispersion measures. The mean and standard deviation are typically used for data that fit a **normal distribution**. For non-normal data, the median, interquartile range, and amplitude are more appropriate.

learn statistics easily



CHAPTER 4

INFERENTIAL STATISTICS: ANALYSES

1. INFERENTIAL STATISTICS

Inferential statistics is a set of methods that enable us to estimate and draw conclusions about a population based on a sample of data.

These methods allow us to make **better decisions** and **predictions**.

Some **examples** of inferential analyses or hypothesis tests include:

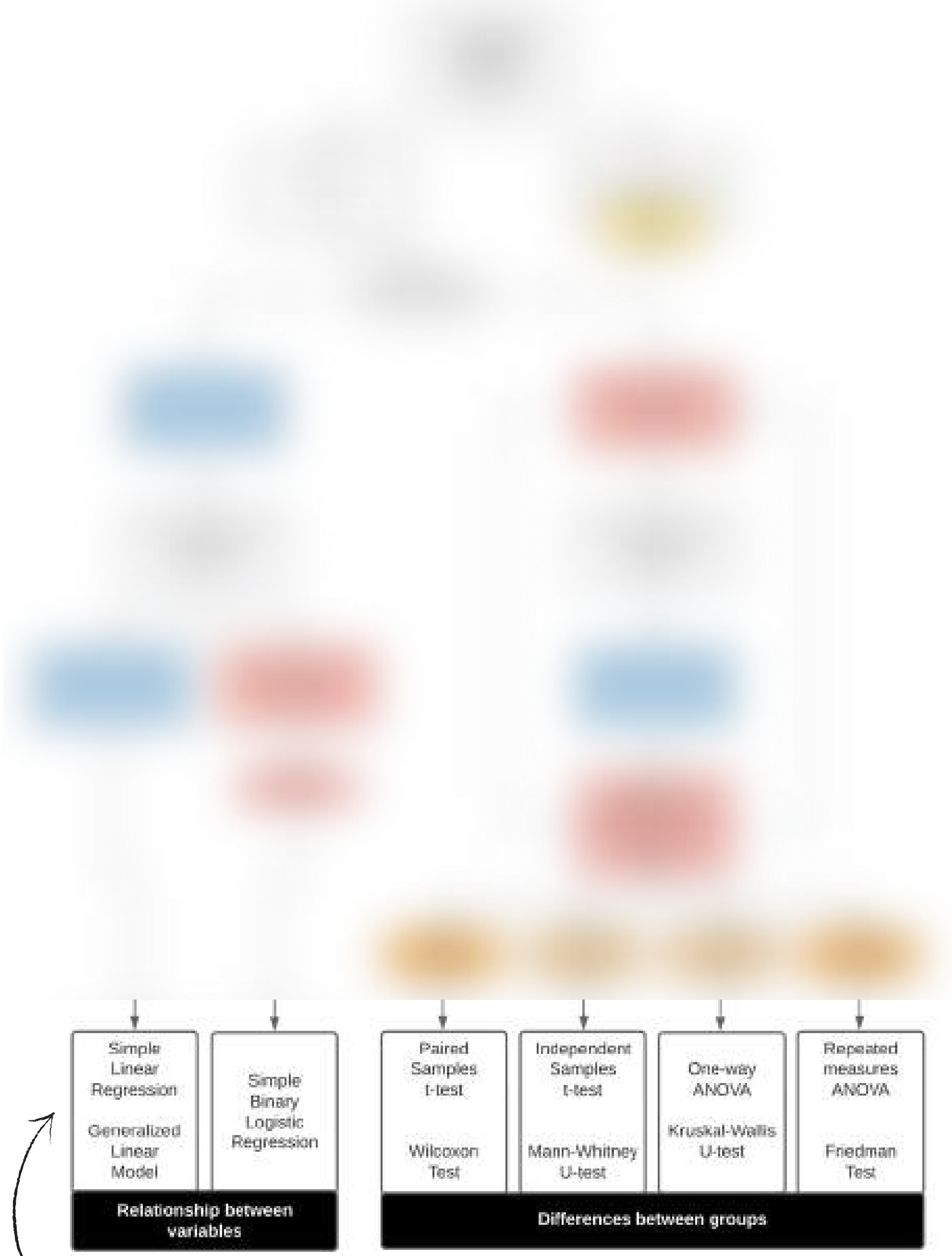
(a) chi-square,
(b) ANOVA,
(c) t-test,
(d) correlation, and
(e) linear regression.

2. BEFORE PROCEEDING

Before proceeding with inferential statistics, it is important to clearly understand the essential concepts presented in **Volume I: Chapter 1**:

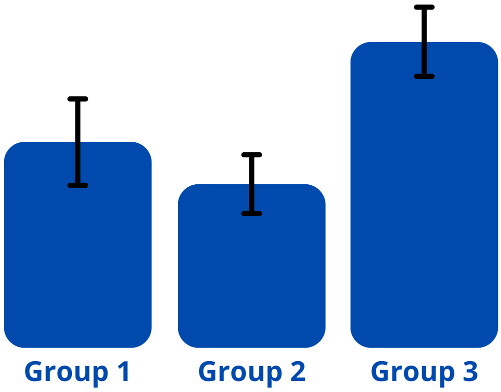
(a) population
(b) sample
(c) element/observation
(d) variable and its types (quantitative and qualitative)
(e) independent and dependent variables (cause and effect)
(f) sampling error
(g) normality (parametric and nonparametric tests)

THE FLOWCHART!®
(PART 1)



**Correlation analyses (Pearson, Spearman, and Kendall) could be here; however, for these analyses, we do not define independent and dependent variables.*

***To analyze two nominal qualitative variables, we use the chi-square test.*



COMPARING UNPAIRED GROUPS

LOOKING FOR DIFFERENCES BETWEEN UNPAIRED GROUPS

(ARE THESE UNPAIRED GROUPS DIFFERENT?)

16.1 INDEPENDENT SAMPLES T-TEST
(Nonparametric alternative: Mann-Whitney's U-test)

16.2 ONE-WAY ANOVA
(Nonparametric alternative: Kruskal-Wallis H-test)

16.3 TWO-WAY ANOVA (FACTORIAL)

16.1
INDEPENDENT SAMPLES T-TEST

OBJECTIVE

This analysis checks whether the means of 2 unpaired groups are significantly different — it assesses the effect of one between-subjects factor (w/ 2 groups) on a quantitative outcome variable.

INDEPENDENT/EXPLANATORY VARIABLE (CAUSE)

1 x nominal qualitative variable with two unpaired groups

DEPENDENT/RESPONSE VARIABLE (EFFECT)

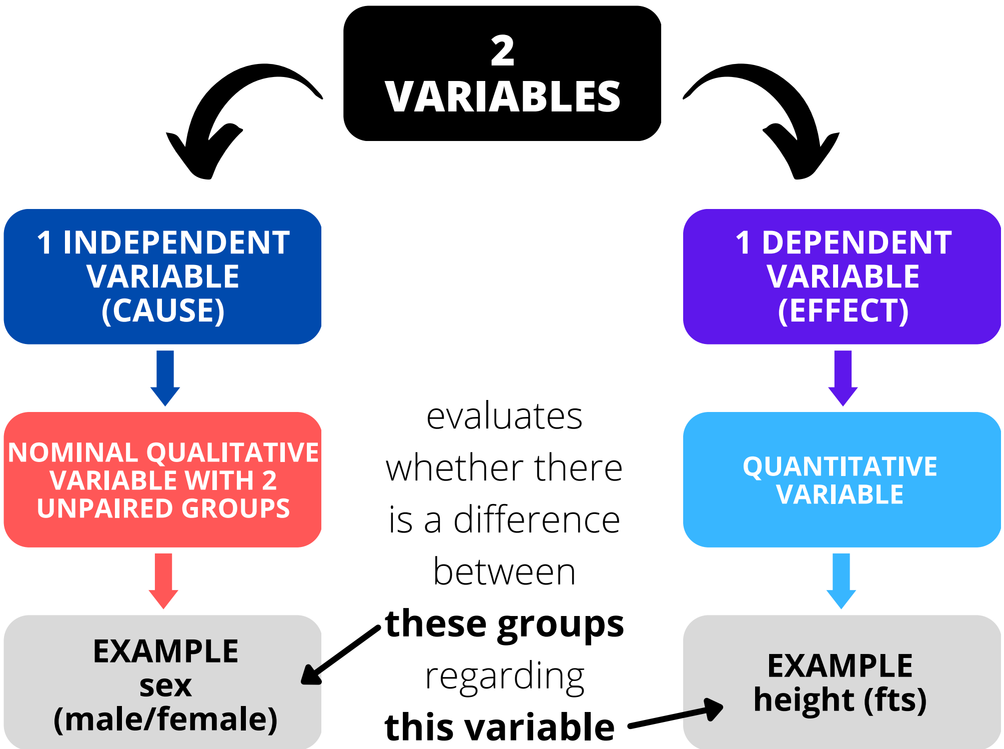
1 x quantitative variable

ASSUMPTIONS

Independence of observations, normality, homoscedasticity

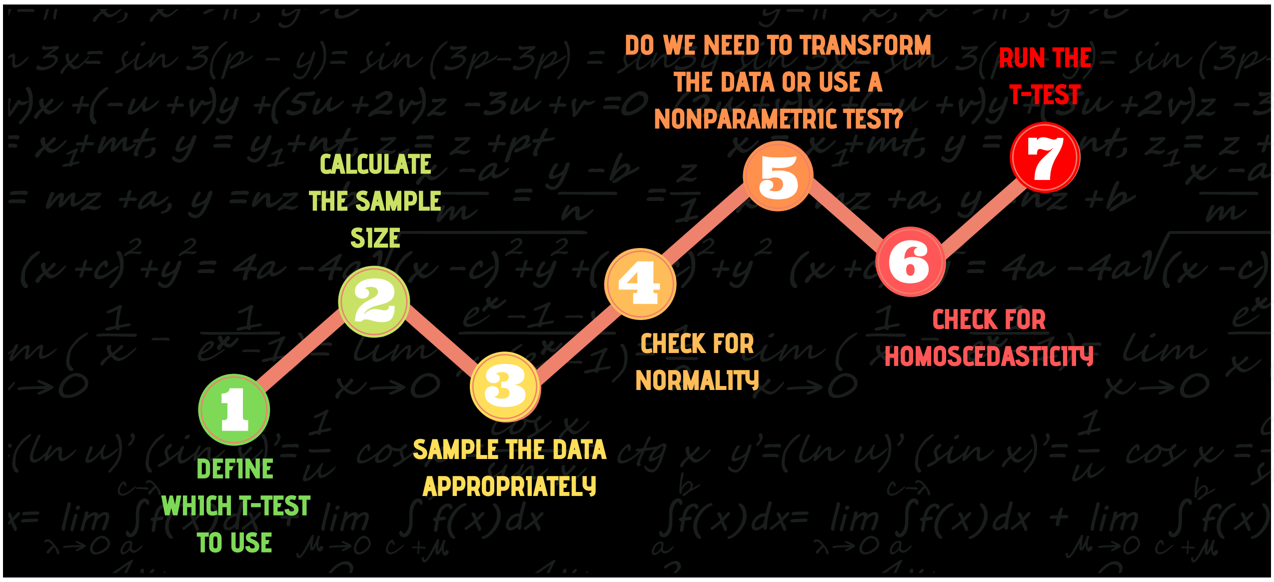
EQUIVALENT NONPARAMETRIC ANALYSIS

Mann-Whitney U-test

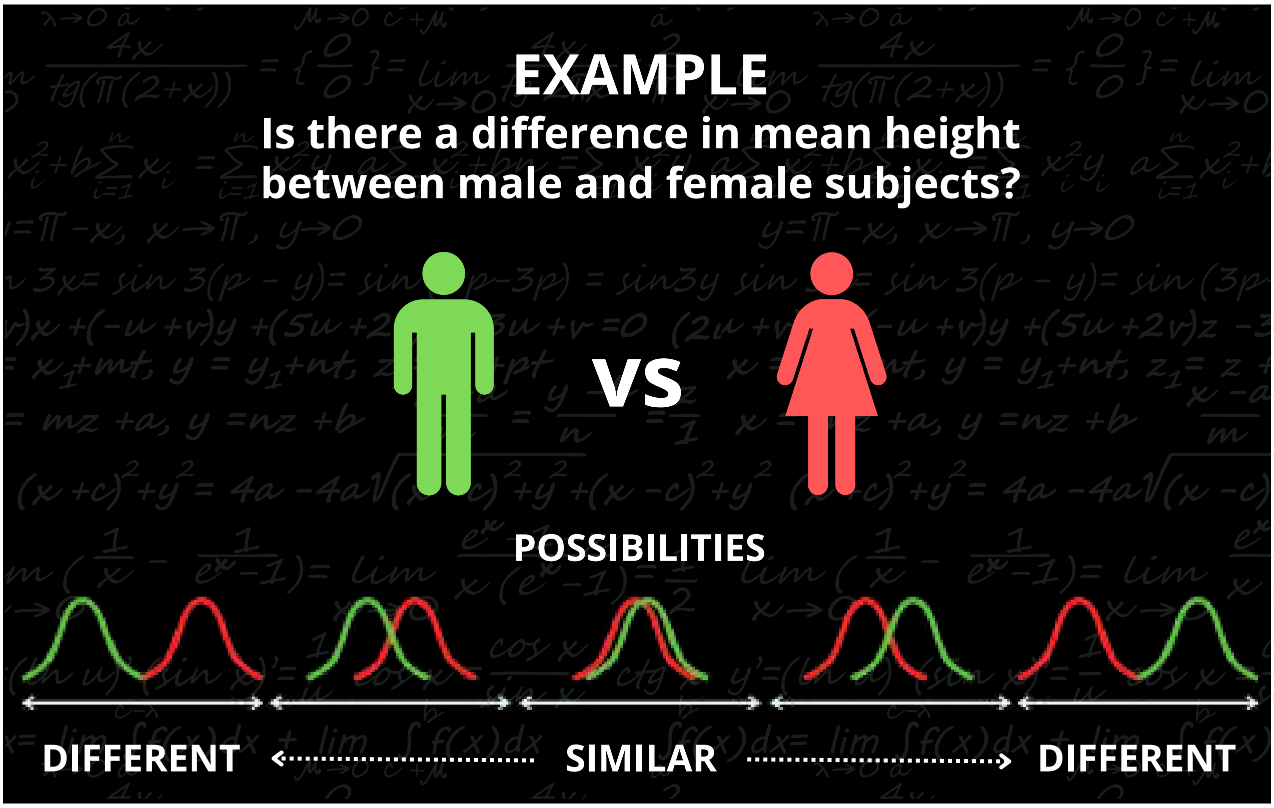


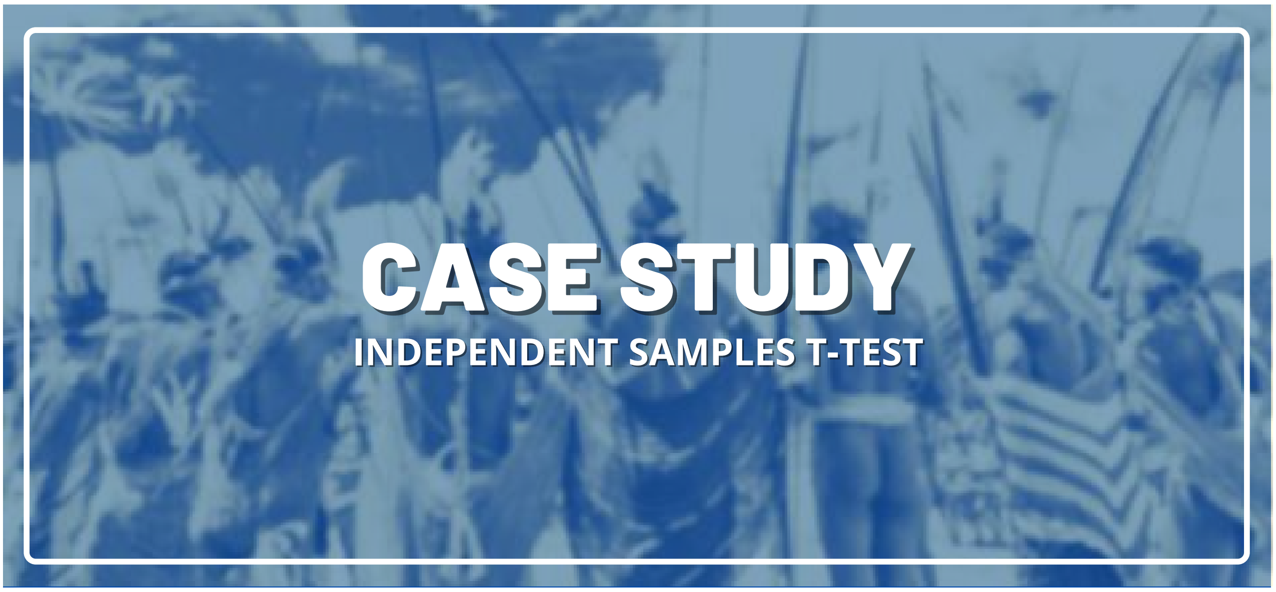
SCHEMATICS FOR BETTER UNDERSTANDING

There are three types of Student t-tests. The most widely used is the one discussed in this topic, the **independent samples t-test**. The steps below indicate the main points to consider in the analysis.



The independent samples t-test will check whether two unpaired groups differ concerning some quantitative characteristic. In the diagram below, we present the possible results of a two-group comparison. The two groups can be similar or different. There are two extreme possibilities if they are different: one group is larger than the other, or the opposite.





CASE STUDY

INDEPENDENT SAMPLES T-TEST

OBJECTIVE

Test whether there is a difference in height between male and female individuals in an indigenous tribe.

INDEPENDENT VARIABLE (X), THE CAUSE

Sex (male/female).

DEPENDENT VARIABLE (y), THE EFFECT

Height.

SET THE SIGNIFICANCE LEVEL

$\alpha = 0.05$

MY HYPOTHESIS

Males are taller on average.

CALCULATING THE SAMPLE SIZE

G*Power (Volume III).

ASSUMPTIONS

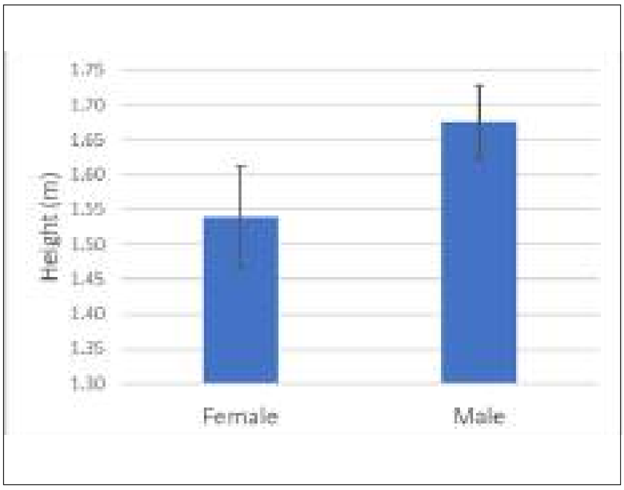
independence of observations, normality, and homoscedasticity.

ANALYSIS

Independent samples t-test (parametric) or Mann-Whitney U-test (nonparametric).

The data should be arranged in the **spreadsheet** as follows. A suitable graph representing this data analysis is the **barplot with error bars**.

	A	B	C
1	ELEMENT	SEX (X)	HEIGHT (y)
2	#001	MALE	1.70
3	#002	MALE	1.64
4	#003	FEMALE	1.45
5	#004	MALE	1.68
6	#005	MALE	1.67
7	#006	MALE	1.69
8	#007	MALE	1.64
9	#008	MALE	1.58



REPORTING RESULTS

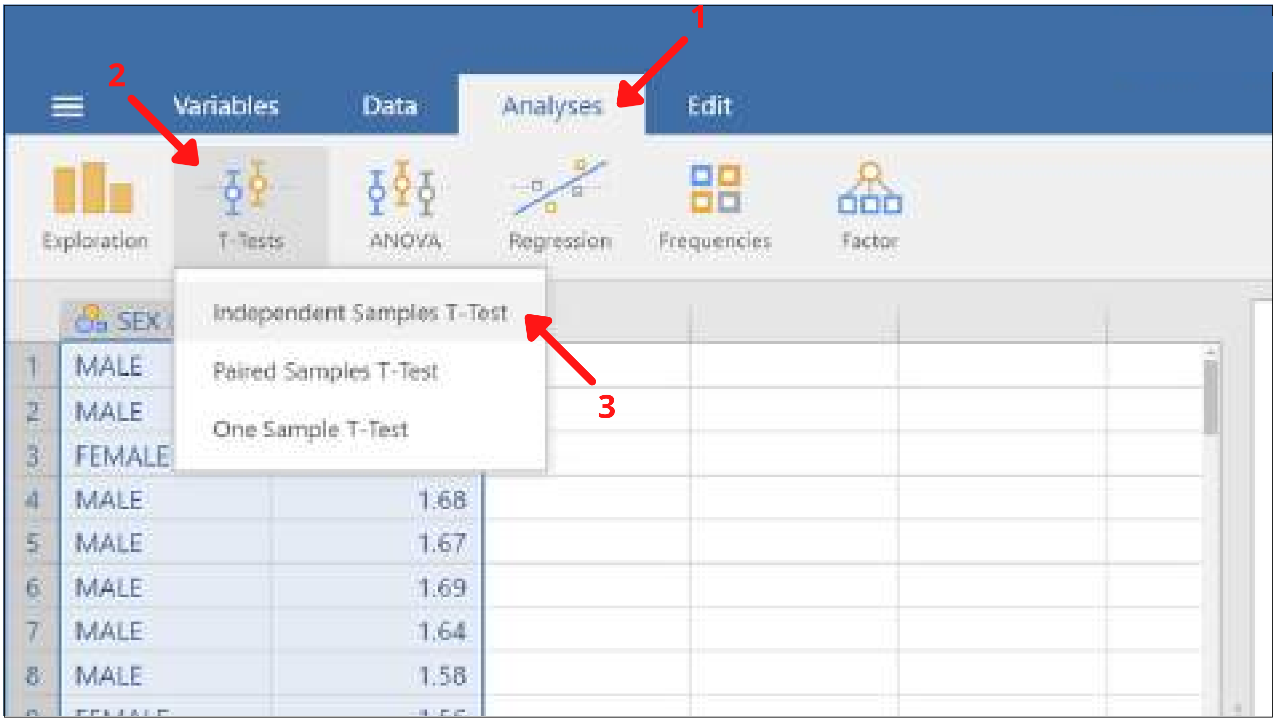
According to the Shapiro-Wilk and Levene tests, the normality (W = 0.977, p = 0.078) and homoscedasticity (F(1,98) = 2.428, p = 0.122) **assumptions** were not violated.

Thus, we used the parametric analysis **independent samples t-test** without corrections, which indicated a significant difference between the groups (t(98) = 10.542, p < 0.001, d = 2.108).

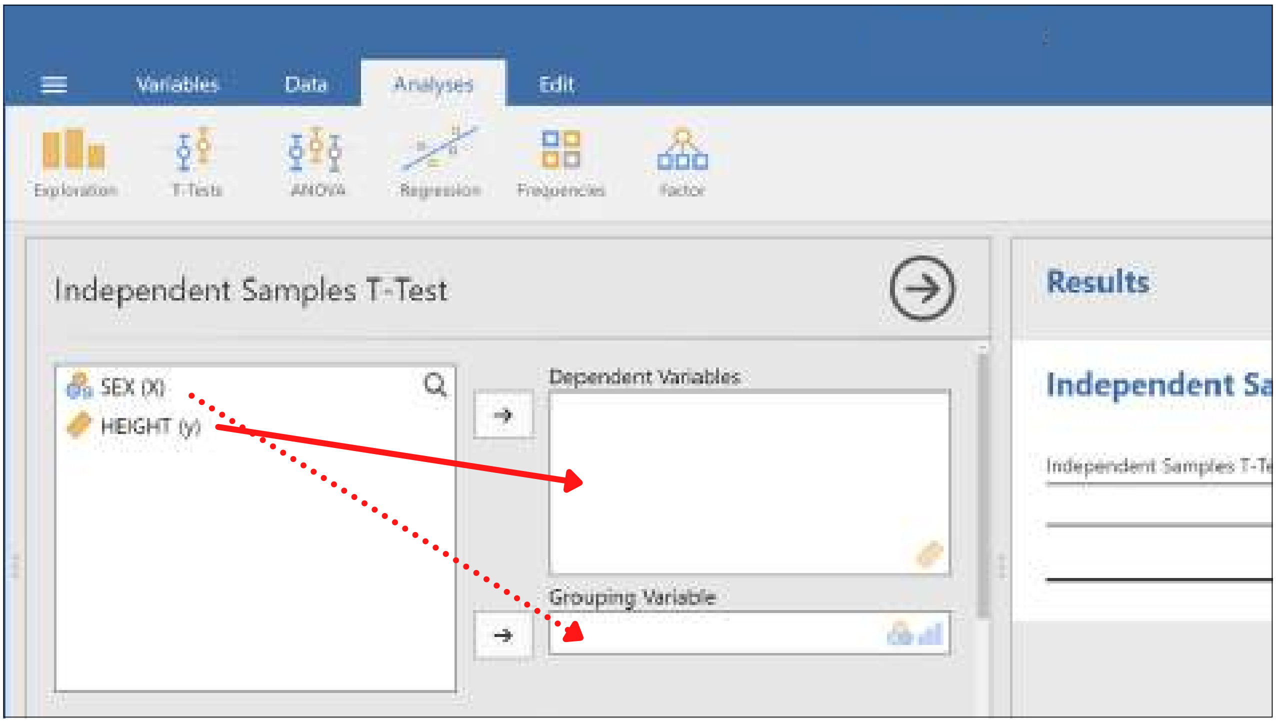
In the indigenous tribe, **MALE** individuals (M = 1.675, SD = 0.073) were taller than **FEMALE** individuals (M = 1.540, SD = 0.054) on average.

STEP-BY-STEP

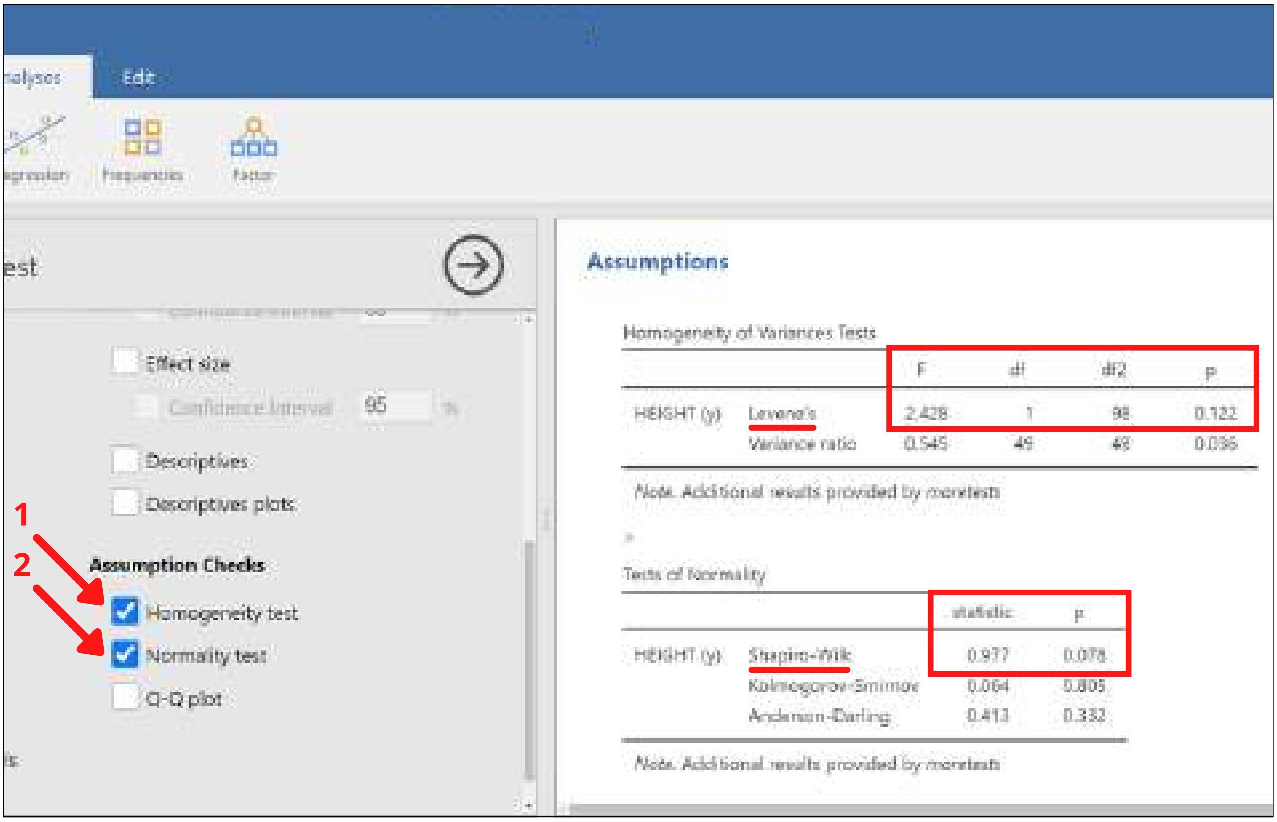
1. After importing or pasting the data to be analyzed, click on **Analyses: T-Tests: Independent Samples T-Test**.



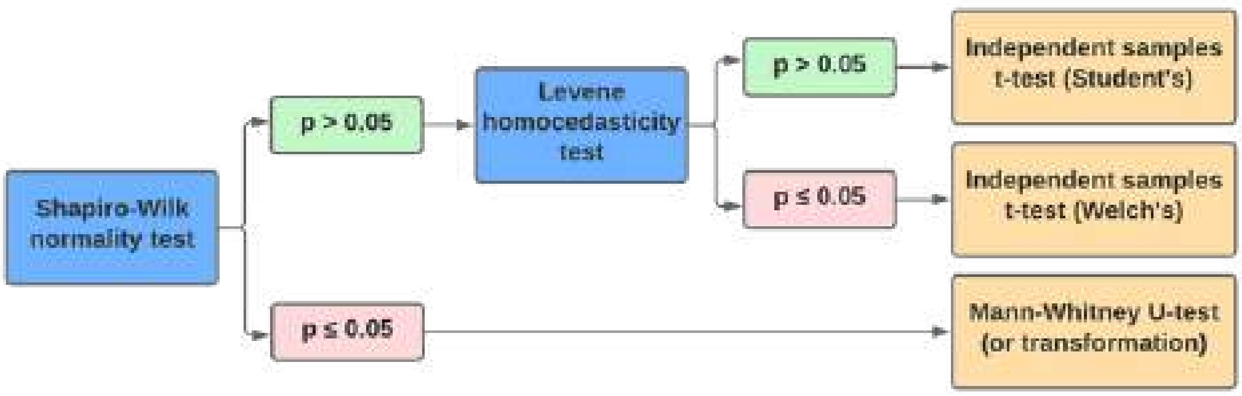
2. Drag the quantitative dependent variable **HEIGHT** to the **Dependent Variables** box and the qualitative independent variable with two groups **SEX** to the **Grouping Variable** box.



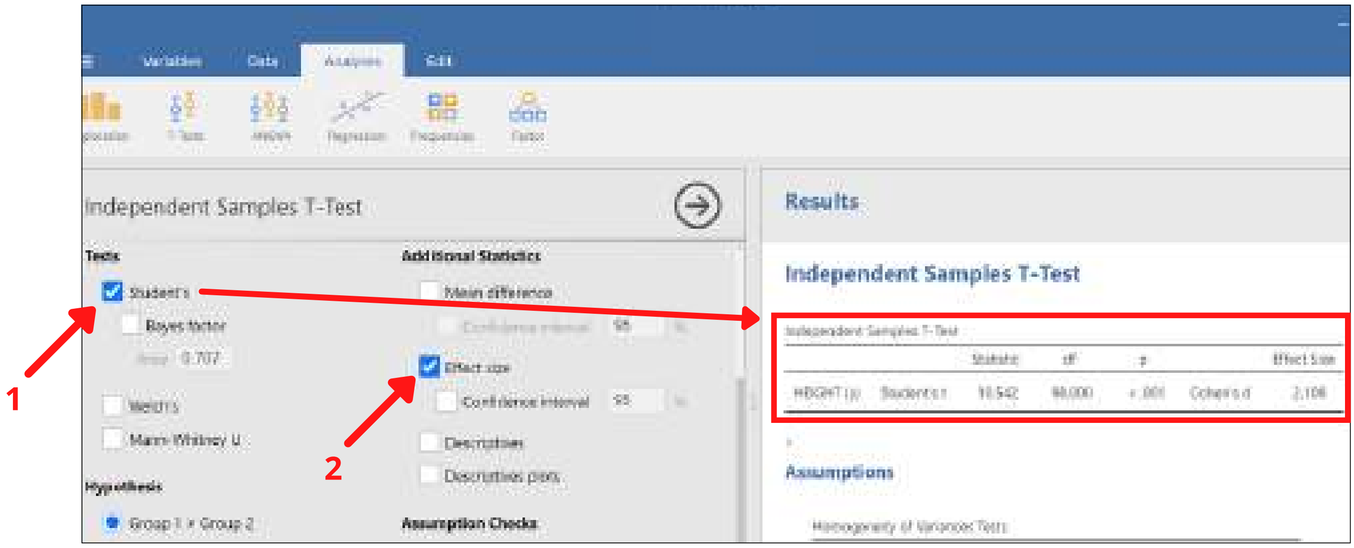
3. In the **Assumption Checks** options, check the **Homogeneity test** and **Normality test** boxes.



4. Now, look at the p-value from the Shapiro-Wilk normality test. Suppose it is less than or equal to 0.05. In that case, we should use the nonparametric analysis **Mann-Whitney U**-test (or try data transformation, **Volume I: Chapter 4: Topic 15: Subtopic 8**). But suppose the normality test p-value is above 0.05. In that case, the next step is to check the p-value of Levene's homoscedasticity test. If it is less than or equal to 0.05, we should use the parametric t-test for independent samples with **Welch's** correction. But if it is above 0.05, we should use the parametric independent samples **Student's** t-test. *We defined the significance level (0.05) a priori.*



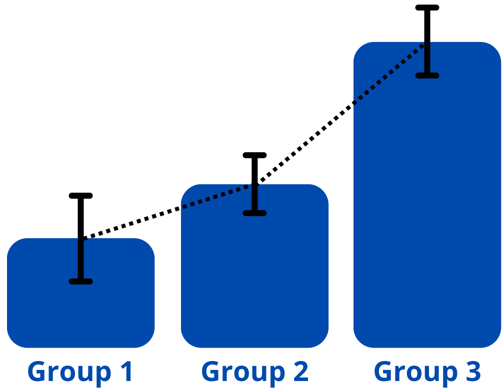
5. In this example, the residuals of the analysis showed normal distribution (W = 0.977; p = 0.078) and homoscedasticity (F(1,98) = 2.428; p = 0.122). Thus, we can use the parametric analysis independent samples **Student's** t-test without corrections. The groups showed significant differences. MALEs are, on average, taller than FEMALEs (t(98) = 10.542; p < 0.001; d = 2.108). *Check the **Effect Size** box under **Additional Statistics** — we cover this crucial measure in detail in **Volume III**.*



(a) If you need to use the parametric test with Welch's correction for heteroscedasticity or the nonparametric Mann-Whitney U-test, uncheck the **Student's** t-test and check the option for **Welch's** t-test or **Mann-Whitney U**-test, respectively.

(b) You can obtain descriptive statistics by checking the option for **Descriptives** under **Additional Statistics**. You will see a comparison of the groups (male and female) with the sample size, the mean, median, standard deviation, and standard error. Additionally, you can obtain a comparative graph from the **Descriptives plots** option.

(c) In the **Hypothesis** section, you can select the expected statistical hypothesis based on the theoretical background of the subject. We choose the hypothesis "**Group 1 ≠ Group 2**" if we think there is a difference between the groups, but without defining which group would be the highest/lowest. Selecting the hypotheses "**Group 1 > Group 2**" or "**Group 1 < Group 2**" would determine which group is the highest. If you are unsure, it is better to keep the first option.



COMPARING PAIRED GROUPS

LOOKING FOR DIFFERENCES BETWEEN PAIRED GROUPS

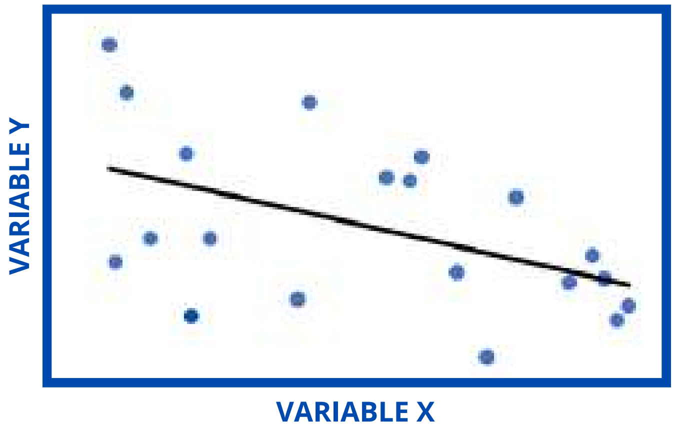
(ARE THESE PAIRED GROUPS DIFFERENT?)

16.4 PAIRED SAMPLES T-TEST
(Nonparametric alternative: Wilcoxon Signed-RankTest)

16.5 ONE-WAY REPEATED MEASURES ANOVA
(Nonparametric alternative: Friedman test)

16.6 TWO-WAY REPEATED MEASURES ANOVA

16.7 MIXED-DESIGN ANOVA



RELATING VARIABLES

LOOKING FOR RELATIONSHIP BETWEEN VARIABLES

(ARE THESE VARIABLES RELATED?)

16.8 PEARSON CORRELATION
(Nonparametric alt.: Kendall's & Spearman's Correlation)

16.9 SIMPLE & 16.10 MULTIPLE LINEAR REGRESSION
(Nonparametric alternative: Generalized Linear Models)

16.11 SIMPLE & 16.12 MULTIPLE LOGISTIC REGRESSION

16.13 QUI-SQUARE TEST OF INDEPENDENCE

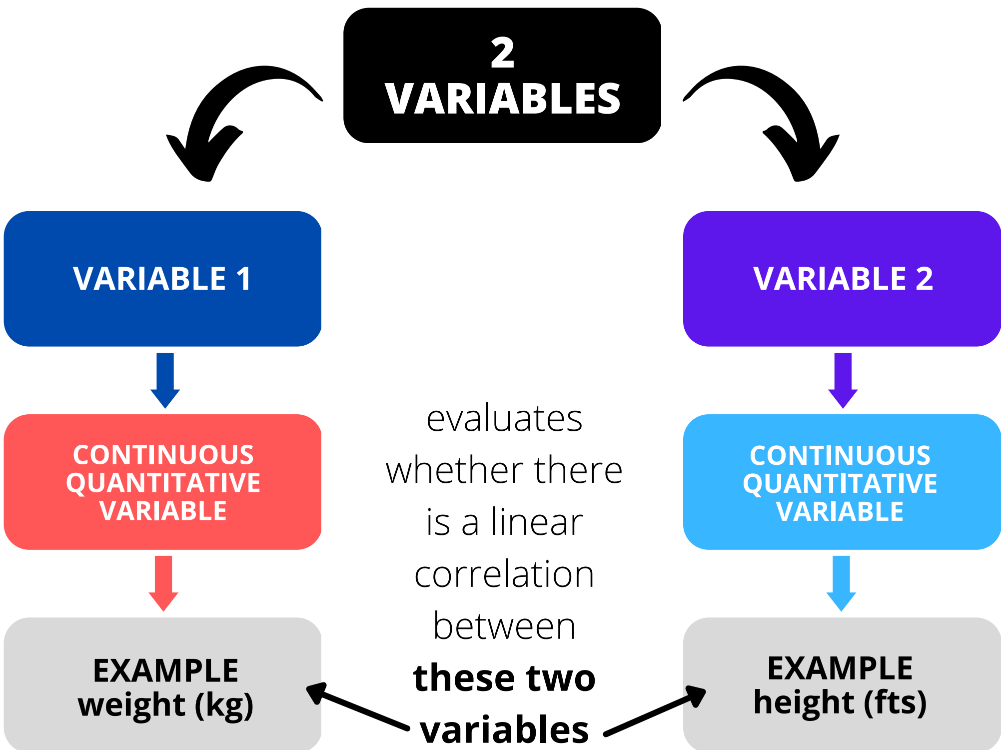
16.8
PEARSON PRODUCT-MOMENT CORRELATION

OBJECTIVE
This analysis is performed to determine if there is a linear relationship between two quantitative variables. If so, it measures the strength and direction of the relationship.

INDEPENDENT & DEPENDENT VARIABLES
In this analysis, defining which variable is independent or dependent is not necessary.

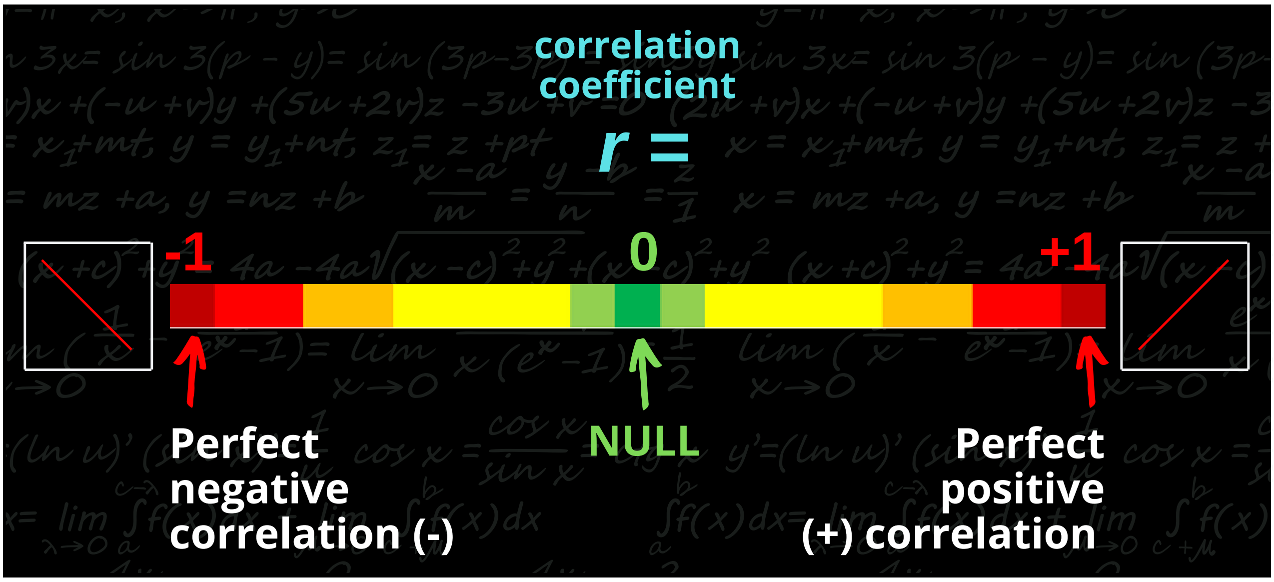
ASSUMPTIONS
Independence of observations, normality, linearity, no significant outliers

EQUIVALENT NONPARAMETRIC ANALYSIS
Kendall's and Spearman's rank correlation

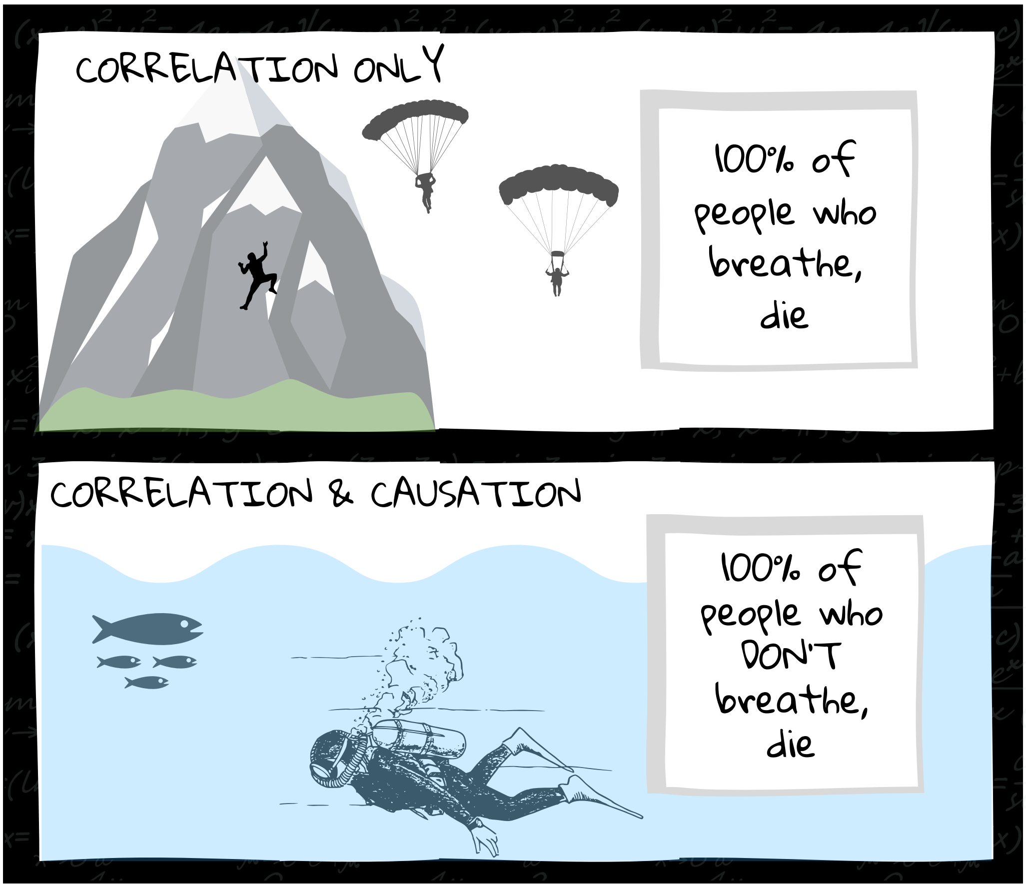


SCHEMATICS FOR BETTER UNDERSTANDING

Pearson's correlation coefficient (r) measures the **strength** and **direction** of a linear relationship between two quantitative variables. The value of r ranges from -1 to +1.

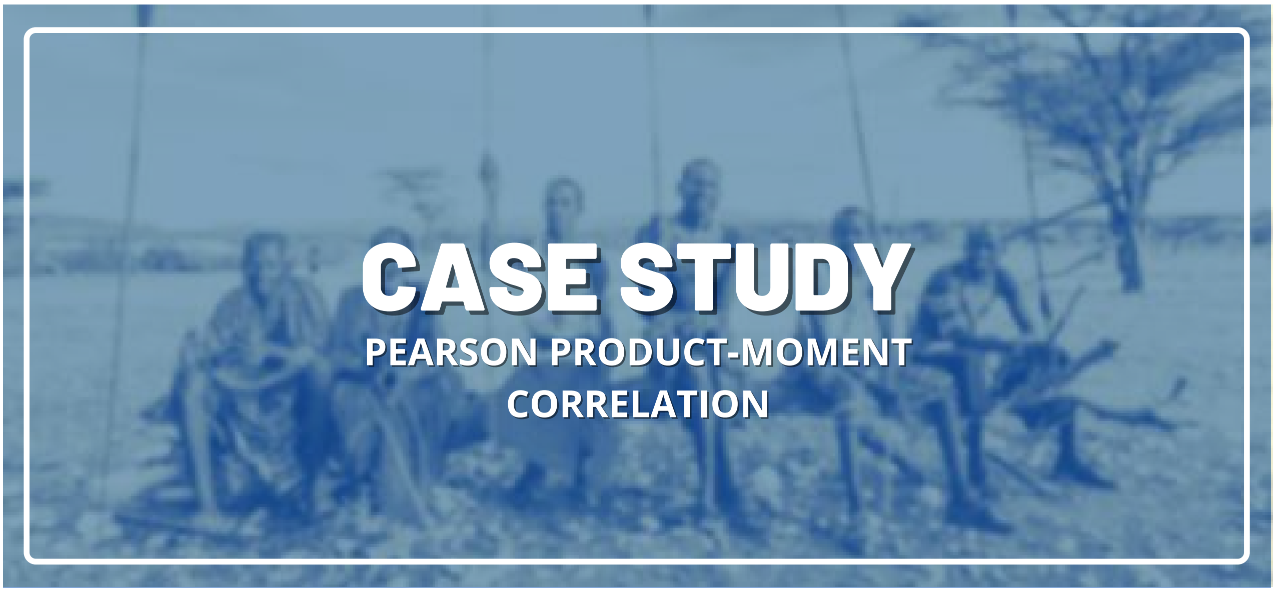


It's important to note that **correlation does not imply causation**. Thus, a significant correlation between two variables does not necessarily mean that one variable is causing the other to vary.



1. People who breathe are alive, but breathing is not what causes death.

2. Conversely, stopping breathing will cause a person's death.



OBJECTIVE

Test whether the quantitative variables weight and height in an indigenous tribe are correlated.

VARIABLE 1

Height (ft).

VARIABLE 2

Weight (lbs).

SET THE SIGNIFICANCE LEVEL

$\alpha = 0.05$

MY HYPOTHESIS

Height and weight are positively correlated.

CALCULATING THE SAMPLE SIZE

G*Power (Volume III).

ASSUMPTIONS

independence of observations, normality, linearity, no significant outliers.

ANALYSIS

Pearson Product-Moment Correlation (parametric) or Kendall's and Spearman's rank correlation (nonparametric).

WAIT!

IT`S NOT OVER YET!

DOWNLOAD VOLUMES II & III ON THE PLATFORM!

@learnstatisticseasily

MORE INFORMATION ABOUT VOLUME III

The **Volume III** initiative is one-of-a-kind.

It allows us to incorporate topics suggested by our followers in each new edition.

And those who have already purchased the package will have lifetime access to these updates.

Volume III covers more advanced topics than those covered in the first two volumes.

As we plan to regularly update Volume III with new topics, we encourage you to send us your suggestions via our Instagram profile.

@learnstatisticseasily


APPLIED STATISTICS: Data Analysis
3 VOLUMES
Master data analysis with a simple and effective method that allows for fast results and autonomy in your studies
$27
3 VOLUMES INCLUDED
ADD TO CART
VISA
$47 $27

Unlock the secrets to analyzing your data swiftly, effortlessly, and confidently.

Our SIMPLE approach focuses on teaching you precisely what you need to know to conquer data analysis.

Leave behind the complexities of concepts, formulas, and tables – this course proves that accurate data analysis is achievable for everyone.

This accessible resource is tailored for those with little or no prior knowledge of statistics.

Discover our unparalleled method for "*fast, easy, and confident data analysis*" – a game-changer you won't find anywhere else.

LEARN STATISTICS EASILY