APPLIED
MEDICAL STATISTICS
FOR BEGINNERS

First edition
2021

Prepared by:
Dr.Mohamed Elsherif

الإحصاء للأطباء
/statistics.for.doctors

About the Author:

Dr. Mohamed Elsherif is an Egyptian physician specialized in medical statistics, public health, epidemiology, and healthcare management. He has over 7 years of experience in performing statistical analysis for medical and non-medical research. He performed statistical analysis for more than 250 different research projects. He experienced in statistical analysis using SPSS, Stata, and R.

Education and certificates:

- ☐ Master of Public Health (MPH) from the American University of Beirut (AUB) with a concentration in epidemiology and biostatistics, 2020.
- ☐ Introduction to Clinical Research Training Certificate from Harvard Medical school, 2018.
- ☐ Egyptian Board in Healthcare Management from the Egyptian Ministry of Health and Population, 2011.
- ☐ Certified Professional in Healthcare Quality (CPHQ), 2010
- ☐ Diploma of Total Quality Management for Healthcare Reform from the American University in Cairo (AUC), 2009.
- ☐ Bachelor of Medicine and Bachelor of Surgery (MBBCh) from Fayoum University, 2007.

Online Certificates:

- Ask Questions to Make Data-Driven Decisions
- Data Analysis with R Programming
- Data Analyst Track, 1 Million Arab Coders Initiative
- Design and Interpretation of Clinical Trials
- Foundations: Data, Data, Everywhere
- Genomic and Precision Medicine
- Hypothesis Testing in Public Health
- Introduction to Statistics & Data Analysis in Public Health
- Introduction to Probability and Data with R
- Linear Regression in R for Public Health
- Reasoning, Data Analysis, and Writing
- Summary Statistics in Public Health
- The Data Scientist's Toolbox
- Understanding Medical Research: Your Facebook Friend is Wrong

Contact Information:

Dr. Mohamed Elsherif

Official website: www.stats4drs.com

Email: contact@stats4drs.com, dr.m.elsherif@gmail.com

Mobile/WhatsApp: (+20) 01029418284

Social media:

Facebook page: https://www.facebook.com/statistics.for.doctors

Facebook account: https://www.facebook.com/dr.mohamed.elsherif/

LinkedIn: https://www.linkedin.com/in/mohamed-elsherif/

YouTube: https://www.youtube.com/channel/UCQ7UciCJJa-x42uyLPHHQfg

Useful links:

- Udemy courses by the author (more than 13000 students):
  - SPSS 26 for Beginners (Arabic) → click here
  - Survival Analysis using SPSS, Simplified in Arabic → Click here
  - Other courses are coming soon …
- Booking or getting information about the medical statistics course and other courses → Click here
- Requesting the statistical analysis service or other related services:

contact@stats4drs.com 00201029418284

Services we provide:

- Statistical Analysis
- Applied medical statistics and basics of research courses
- Sample Size Calculation
- Study design
- Data Visualization
- Tutoring postgraduate students in statistics, research methodology and public health

Before you start

- ☐ This book is written in a simplified way that is suitable for absolute beginners.
- ☐ Mathematical equations and theoretical issues are avoided whenever possible.
- ☐ Illustrations are used extensively to simplify the topics.
- ☐ This book covers the basic statistical concepts and methods, but it does not cover the application using statistical software. To be capable of doing statistical analysis you have to learn using at least one statistical software.
- ☐ Some topics are illustrated extensively, some are explained superficially, and other topics were ignored intentionally.
- ☐ For extra statistical topics written in simplified Arabic, please visit our Facebook page.
- ☐ The author made every possible effort to make the information in this book as accurate as possible, however in case you find anything that you think incorrect or vague, please don't hesitate to contact the author.
- ☐ If you have any suggestions or recommendations for the coming editions, please contact the author.

Contents

Part 1 Basic Statistical Concepts

Introduction

What is statistics?

Statistics is the science concerned with developing and studying methods for collecting, analyzing, interpreting, and presenting **data**.

What is biostatistics?

Biostatistics is the application of statistical principles to questions and problems in medicine, public health, or biology.

What studying biostatistics is useful for?

- Design and analysis of research studies.
- Describe and summarize the data we have.
- Analyze data to measure the association or difference.
- To conclude if an observation is of real significance or just due to chance.
- To understand and evaluate published scientific research papers.

The statistical analysis journey:

The statistical analysis journey goes through the following steps:

- Transforming the research idea into a research question.
- Choosing the proper study design and selecting a suitable sample.
- Performing the study and collecting data.
- Analyzing data (using the appropriate test).
- Getting and interpreting the p-value.
- Reaching a conclusion (answer) regarding the research question.

We are covering this journey in the different parts of this book.

Types of data variables

A **data variable** is "something that varies" or differs from person to person or group to group.

Data variables are the items that we collect data about.

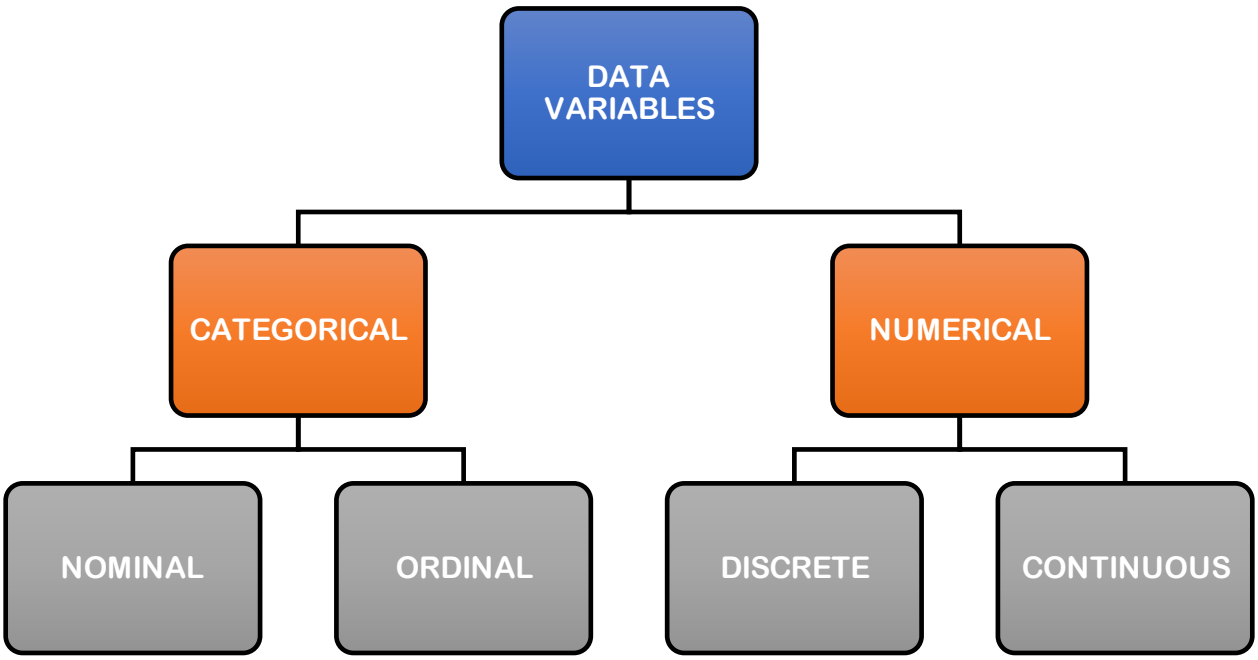
Examples for data variables are sex, age, weight, marital status, satisfaction rate, etc.

When dealing with data, it is important to recognize the type of each data variable for the following reasons:

- **Summarizing data:** describing a variable in mean with standard deviation or in frequency with percentage depends on the type of data variable.
- **Graphical presentation:** choosing the proper graph to represent the data depends on the type of data variable.
- **Analyzing data:** choosing the suitable statistical tests also depends on the type of data variables.

Data variables are classified generally into the following 2 types:

A. **Categorical variables:** which are either nominal or ordinal.
B. **Numerical Variables:** which are either discrete or continuous.



A. Categorical variables:

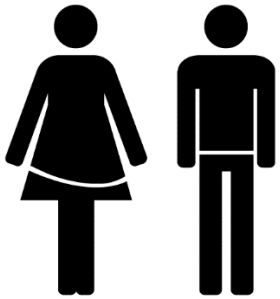
They are also known as qualitative or nominal data; they have NO unit of measurement.

Individuals are described as belonging to any of the categories of this variable.

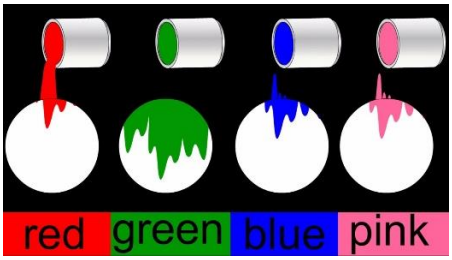
Examples:

Satisfaction status: (satisfied, neutral, not satisfied)

Sex: (female, male)



Colors: (red, green, blue, pink)



Nationality: (all countries)

We can describe one patient as belonging to the males' group or the females' group, and we can describe one customer as belonging to the satisfied group, the neutral group, or the unsatisfied group.

Sometimes, categorical variables are coded in numbers like:
1 for females and 2 for males, or 0 for No, and 1 for yes, and so on.
Even if they are coded or represented as numbers, they are still categories, and the data type is categorical.
The number here is just a code.

1- **Nominal variables:** those are categorical variables that have no intrinsic order.

Examples:

Sex: (female, male), can also be presented as (male, female)
Blood groups: (A, B, AB, O) can also be presented as (A, B, O, AB) or any other order.
Nationality: can be presented in any way; there is no order for the countries.

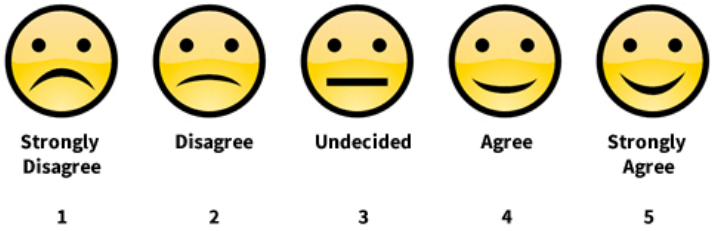
If the nominal variable has only two groups as sex (male, female), an answer to a question (Yes, No), or a disease status (diseased, not diseased), we call it a **dichotomous** variable, or a **binomial** variable.

2- **Ordinal variables:** those are categorical variables that have an order, and that order has a meaning.

Examples:

BMI status: (underweight, normal, overweight, obese, extremely obese)

Agreement level: (strongly disagree, disagree, undecided, agree, strongly agree)



Even if this variable is coded in numbers from 1 to 5, it is still an ordinal variable that is categorical and not numerical.

B. Numerical variables:

Those variables are either measured or counted, represented in numbers, and have a measurement unit.

Examples:

- Height (in cm)
- Weight (in kg)
- Blood glucose level (in mg/dL)
- Number of kids in the family (4 kids, 2 kids, one kid, etc.)

Numerical variables are either discrete or continuous.

1- Discrete variables:

They take only integer numbers (no decimals) such as 0,1,2,3,4…
They usually represent a count of something.

Examples:

- Number of kids in a family.
- Number of stents inserted into the coronaries.
- Number of patient visits to the hospital.

The unit of measurement represents what we are counting (as kid, stent, visit, respectively)

2- Continuous variables:

They can take any real numerical value, including decimals (as 14.55, 48.8, 178.2).
They involve measurement and have measurement units.

Examples:

- Weight (in kg)
- Height (in cm)
- Blood glucose level (in mg/dL)

How to differentiate between types of data variables:

Step 1: Is there a unit of measurement?

If No, it is categorical, and if Yes, it is numerical.

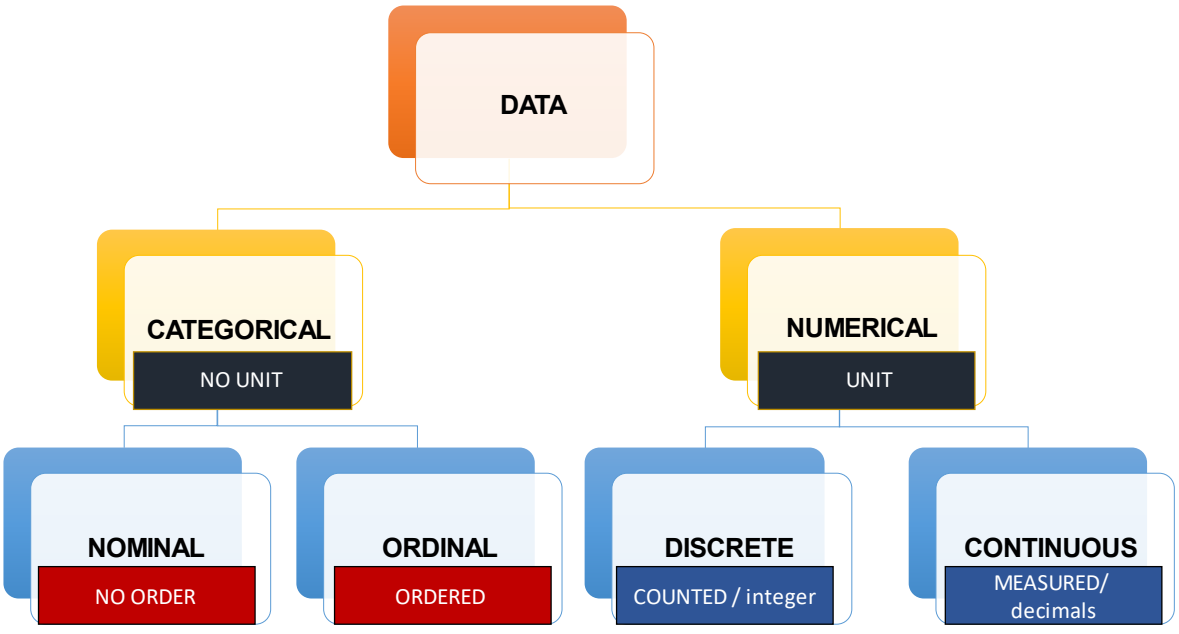
Step 2:

For the categorical variables: Is there an order?

If No, it is nominal, and if Yes, it is ordinal.

For the numerical variables: Is it counted or measured?

If counted, it is discrete, and if measured, it is continuous.



Data are usually presented as follows:

Student No	sex	Blood group	BMI	BMI group	Number of courses	Body Temp
1	male	O	17.8	Underweight	4	36.6
2	female	B	26	Overweight	5	37.1
3	male	AB	24.5	Healthy weight	4	36.9
4	male	A	31.6	Obese	4	36.8
5	female	A	33.4	Obese	5	36.6
6	female	B	27.5	Overweight	6	37
7	female	O	26.8	Overweight	7	37.2

Types of data variables in this dataset are:

- Sex: nominal (dichotomous), categorical
- Blood group: nominal, categorical
- BMI: continuous, numerical
- BMI group: ordinal, categorical
- Number of courses: discrete, numerical
- Body temperature: continuous, numerical

Some more ideas:

- Some textbooks classify numerical data into interval variables and ratio variables. **Ratio variables** are variables that have true zero, such as weight. When we say the weight is zero, this means the complete absence of weight, and a weight of 30 kgs is twice as heavy as 15 kgs.
  While in **interval variables** as temperature, there is no true zero. A temperature of 0^0C does not mean the absence of heat, and a temperature of 30^0C is not twice as hot as 15^0C.
- When data is ordinal in nature with a large number of levels as a pain score measured on a 10 levels scale, it can be treated as a discrete variable.
- Some variables that are continuous in nature are sometimes measured as discrete; age is an example as it is usually reported as the number of years instead of the exact age.

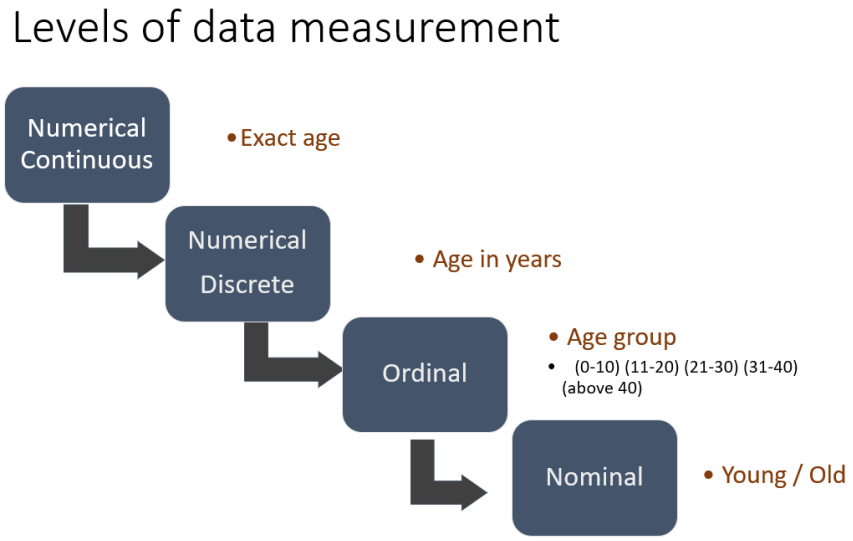
Levels of data measurement:

It is possible to change the type of data variable into another one, but only in one direction:

numerical continuous → numerical discrete → ordinal → nominal

- We can change the age from a numerical variable to an ordinal variable if we categorize it into different age groups.
- Also, we can change the age from an ordinal variable as age groups into a nominal variable of two levels (young, old).
- But if we collect the data in a categorical form, we cannot transform it into a numerical form.

> Whenever possible, collect your data at the highest level, numerical continuous or numerical discrete, as it is more accurate and can be categorized easily later on.



Data entry

Sometimes, data is collected on paper forms, and we need to do data entry into a computer file in preparation for the data analysis.

The goal of any data entry process is to have data arranged in a spreadsheet, like this one:

	A	B	C	D	E
1	child_ID	Age	Gender	intervention_control	Family_financial_status
2	1	11	2	2	3
3	2	10	1	2	3
4	3	10	1	2	3
5	4	10	1	2	3
6	5	11	2	2	4
7	6	10	1	2	3
8	7	10	2	2	3
9	8	10	2	2	3
10	9	10	2	2	3
11	10	9	1	2	2
12	11	11	1	2	4

A well-arranged datasheet should satisfy the following characteristics:

1- **Each column represents one variable.**
   If one variable is measured twice (as before and after an experiment), then it should be recorded in two columns.
   If a variable consists of 2 elements (as blood pressure consisting of systolic and diastolic blood pressure), then each element should be recorded in a single column.

2- **The unit of measurement is unified in each column.**
   Height is measured either in meter or in cm, can't be in meter for some patients, and in cm for others.

3- **Each row represents a case**
   The case is the unit of which we collect data, as a patient, a rat, a village, a hospital, etc., depending on each study.

4- **Each cell contains only one data point.**
   It can't include both systolic and diastolic blood pressure, or gestational age in weeks and days.

5- **Nominal and ordinal data are coded using numeric codes.**
   We use numbers as codes for each category instead of writing the name of the category. For example, we may use 1 as code for males and 2 as code for females. Always keep a codebook for your coded variables where you can find the codes and corresponding values.

Coding of categorical data:

It is better to use numeric codes when entering categorical data, easier, less prone to typing mistakes, and more suitable for the statistical software packages.

It is better to use reasonable codes for each variable as in the following examples:

Severity of disease:

- Mild → 1
- Moderate → 2
- Severe → 3

Severity of Pain:

- No pain → 0
- Mild pain → 1
- Moderate pain → 2
- Severe pain → 3

If binary (Yes/No)

- Yes → 1
- No → 0

- If multiple answers are allowed for one question, use a column for each choice and code it as 1/0 representing Yes/No.
  In the data collection form, asking about chronic conditions may be in this way:

Do you have any of the following Chronic diseases?

- ☐ DM
- ☐ Hypertension
- ☐ CVD
- ☐ Hypothyroidism

But in the data entry, it should be like this:

D	E	F	G
DM	Hypertension	CVD	Hypothyroidism
1	0	0	1
1	1	0	0
0	0	1	1
1	0	1	0

- ✓ If there is a variable with open answers or a large number of possible answers, we have to evaluate those answers and categorize them into a limited number of categories, so that we can include them in the statistical analysis.

Tips for data entry of numeric variables

• **Be precise**	1.56, not 1.5 or 1.6
• **Only numbers**	2, not two
• **Keep consistent units**	m or cm / kg or pound, not both
• **Don't write the unit**	2, not 2 times, or 2 years
• **Use basic measurements**	as weight and height, not BMI (it can be calculated later)
• **Don't categorize**	Collect the exact age, not 20-25 years
• **Only one data element**	not as gestational age 20+2, representing 20 weeks and 2 days (it should be in days only or weeks only)

Coding of missing data

It is better to use codes for missing data instead of leaving the cells empty so that we are sure that it is a missing value and not a data entry mistake.

- ➢ Use impossible values (as codes) that can't be correct for this variable.

For example:

If binary variable as Yes/No coded as 1,0	we can use 9
If categorical variable with three categories coded 1,2,3	we can use 9
If age of a child (in years)	we can use 99
If weight (in kg)	we can use 999

Note that: Refused to answer and Not applicable are not considered the same as missing (we give them other codes such as 998, 997).

Exploring data for errors:

Before running the statistical analysis, we need to explore the data to make sure that there are no data entry errors.

This can be done using many techniques:

- **Check the range (minimum and maximum)**

  Are there any incorrect extreme values? Are they consistent with other data values?

- **Check the frequency distribution for categorical variables**

  Are there any typing mistakes or unusual codes or groups?

- **Check the missing values**

  Are they really not available? Or we just forgot them during data entry?

- **Checking the consistency of data**

  For example, a man can't be pregnant, disease duration can't be larger than age, and diastolic blood pressure can't be larger than systolic blood pressure.

- **Graphically checking the data**

  A histogram or a boxplot for a single numeric variable, and a scatterplot for two related variables as weight and waist circumference may be helpful to explore possible errors.

Descriptive statistics

It is important to learn how to describe our data and present them correctly using numbers (in the proper table format) or graphs.

The first table in most scientific research papers shows descriptive statistics of the study subjects.

As in the following table:

Table 1. Baseline Characteristics of the Study Participants*

Characteristics	Active Treatment (n = 817)	Placebo (n = 813)
Age, mean (SD), y	42.1 (9.0)	42.4 (9.1)
Sex		
Men	440	440
Women	377	373
Daily smoking	203 (25)	198 (24)
Alcohol use	194 (24)	160 (20)
Dyspepsia symptoms	417 (51)	409 (50)
Dietary intake ≥2 times/wk		
Green tea	205 (25)	181 (22)
Preserved vegetables	144 (18)	132 (16)
Salty fish	364 (45)	372 (46)
Fish sauce	172 (21)	241 (30)
Fruit	112 (14)	83 (10)
Fresh vegetables	275 (34)	253 (31)
Histopathologic test results		
Chronic active gastritis	485 (59.4)	503 (61.9)
Gastric atrophy	72 (8.8)	57 (7.0)
Intestinal metaplasia	243 (29.7)	234 (28.8)
Gastric dysplasia	4 (0.5)	5 (0.6)
Unclassified†	13 (1.6)	14 (1.7)

*Data are expressed as No. (%) of participants unless otherwise indicated.
†Histology slides were uninterpretable or no definite conclusions could be drawn.

There are different ways of numerically describing data based on the type of the variable.

1- Describing categorical variables

Categorical variables such as sex, smoking status, and disease severity are described as:

- **Frequencies (numbers):** which is the number of participants in each category, as the number of males and the number of females.
- **Relative frequencies (percentages):** which is the percentage of participants in each category.

For example:

If you have 200 participants, 120 are males and 80 are females.

We can express the frequencies and percentages as follows:

Males: 120 (60%)

Females: 80 (40%)

Percentages can be calculated easily by dividing the number of that category by the total number and multiplying it by 100.

For the males, it is: $\frac{120}{200} \times 100 = 60\%$.

2- Describing numeric variables

Numerical variables are usually described using two numbers, one represents the center of the data (**central tendency**), and the other represents the spread of the data (**dispersion**).

- **Measures of central tendency**

The most common measures for the center of the data are the mean, median, and mode.

a- Mean

- The mean of a variable can be computed as the sum of the observed values divided by the number of observations.
  For example: if we want to calculate the mean for the age of 7 children; 7,5,6,8,2,9,3

  it will be: $\frac{7+5+6+8+2+9+3}{7} = \frac{40}{7} = 5.71$ years.

- The mean is easily affected by extreme values.
  If we add one adult whose age is 64 years to this group and try to calculate the mean again it will be:

  $\frac{7+5+6+8+2+9+3+64}{8} = \frac{104}{8} = 13$ years.

- We can see that the mean age has changed obviously from 5.71 to 13 years by adding only one value and that the new value (13) is even larger than the age of all the 7 children. The mean here is not a good representative of our data.
- The mean is also called the average or the arithmetic mean.

b- Median

- The median is the point at the center of the data, where half of the values are above, and half are below it.
- To calculate the median, we first arrange (order) our data from the smallest value to the largest value. Then, the median is the value in the middle.

  For example: if we want to calculate the median for the age of 7 children mentioned above; 7,5,6,8,2,9,3
  First, we order the data:

  2,3,5,6,7,8,9

  Then it is obvious that the center of it is the number 6, where 3 values are below, and 3 values are above it:

  2,3,5,⑥,7,8,9

  So, the median= 6 years
- What if we try to add the adult with the age of 64 years old?
  Then the ordered data will be:

  2,3,5,6,7,8,9,64

  Here, we can't see one value in the middle with half the values above and half below it. In this case, we will take the average of the two values in the middle:

  2,3,5,⑥,⑦,8,9,64

  So, the median = $\frac{6+7}{2} = 6.5$ years

As we notice, the median didn't change much when that extreme value was added.

c- Mode

- Simply, the mode is the most frequently occurring value in the dataset.

So , if you have a data set like: 2,3,5,6,7,8,9,64,3,4,5,3
Then the mode is 3.

- It can be also calculated for categorical variables as it depends only on the frequency of each value.
- The mode can be more than one value; if two values have the same highest frequency, then, both are the modes, and data is called bimodal.
  The mode is rarely reported in scientific research.

	Advantages	Disadvantages
Mean	Uses all data values Algebraically defined	Distorted by outliers Distorted by skewed data
Median	Not distorted by outliers Not distorted by skewed data	Ignores most of the information Not algebraically defined
Mode	Easily determined for categorical data	Ignores most of the information Not algebraically defined

The five-number summary

If we arrange our values from lowest to highest and choose five points on the arranged data to divide the variable into 4 quarters, those five points (numbers) will be:

- **The minimum value**
- **The maximum value**
- **The median:** which is the point at the center of the data where half of the values above and half are below it.
- **The first quartile (lower quartile):** where 25% of the data are below it, it is the center point for the lower half of the data. It is also called the 25th percentile.
- **The third quartile (upper quartile):** where 75% of the data are below it, it is the center point for the upper half of the data. It is also called the 75th percentile.

If you have the following values for a variable:

8, 10, 10, 10, 12, 14, 15, 15, 18, 23, 25, 27

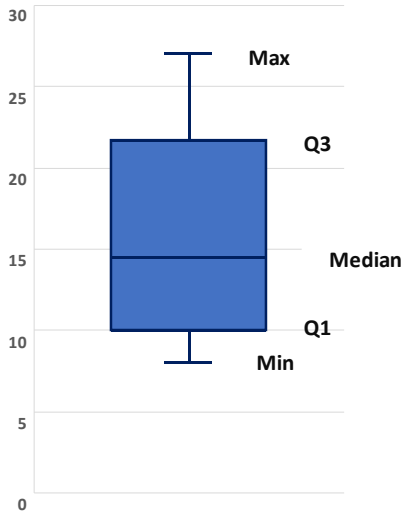
The five-number summary will be:

Min: 8 **Q1:** 10 **Median:** 14.5 **Q3:** 21.75 **Max:** 27

It is easily calculated using computer software. The following is an SPSS output:

N	Valid	12
	Missing	0
Minimum		8
Maximum		27
Percentiles	25	10.00
	50	14.50
	75	21.75

The example is graphically presented in a graph called the box-plot as follows:



- **Measures of dispersion**

The most commonly used measures of dispersion (spread of the data) are range, inter-quartile range, variance, and standard deviation.

a- Range:

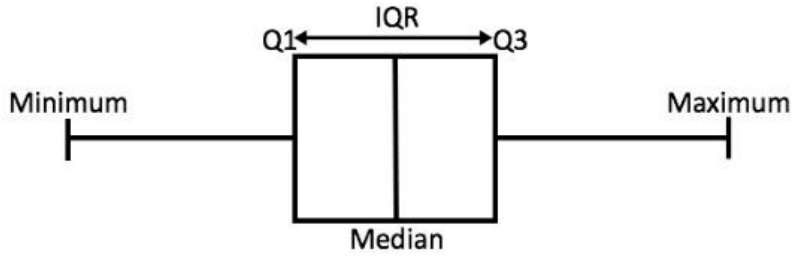
- The range is simply the difference between the largest and smallest values.

If you have the following values for the age variable: 8,10,10,10,12,14,15,15,18,23,25,27

- The lowest value is 8, the highest value is 27, so the range is 27-8= 19 years.
- It is obvious that the range is affected by any extreme values.
- Adding one adult aged 64 to this group will increase the range significantly. The range becomes 64-8= 56 years.

b- Inter-quartile range (IQR):

- The inter-quartile range is simply the difference between the upper quartile and the lower quartile = Q3-Q1
- It represents the middle 50% of the data, where 25% of the data are below it, and 25% are above it.



For those values: 8,10,10,10,12,14,15,15,18,23,25,27
Q1=10, Q3= 21.75
The IQR = Q3-Q1 = 21.75-10 = 11.75
The IQR is not calculated using the minimum or the maximum values, so it is not affected by extreme values.

c- Variance

- The variance is a measure of spread that takes all data points in the calculation. It represents the distance of all data points from the mean.
- We calculate it as in the following steps:

1- Calculate the mean.
2- Calculate the difference between each data point and the mean, then square it (not to have negative values).
3- Sum all the squared differences calculated in step 2.
4- Divide this sum by the number of observations -1 (n-1)
This is the variance; it is in square units (as we squared the difference!).
This means that if the mean height in m, then the variance is in m^2

Example:

We have a group of 7 children, and their age in years is; 7,5,6,8,4,9,3, let's calculate the variance.

1- Calculate the **mean**: $\frac{7+5+6+8+4+9+3}{7} = \frac{42}{7} = 6$ years.
2- Calculate the **difference** between each data point and the mean, then **square** it.
$(7-6)^2, (5-6)^2, (6-6)^2, (8-6)^2, (4-6)^2, (9-6)^2, (3-6)^2$
$= 1^2,-1^2,0^2,2^2,-2^2,3^2,-3^2$
$=1,1,0,4,4,9,9$
3- **Sum** all the squared differences =28
4- **Divide** this sum by the number of datapoints -1 (n-1)
$s^2=\frac{28}{7-1}$= 4.67 years^2

So, the variance = 4.67 years^2

But the interpretation of variance of age with a squared unit (years^2) is not easy to understand.

So, we take the square root of the variance to have the standard deviation (s), which is now of the same unit as the mean.

$$s=\sqrt{s^2} = \sqrt{4.67} = 2.16 \text{ years}$$

d- Standard deviation

- The standard deviation is a measure of spread that represents the average distance of the data values from their mean.
- It is calculated as the square root of the variance that has been calculated before.

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

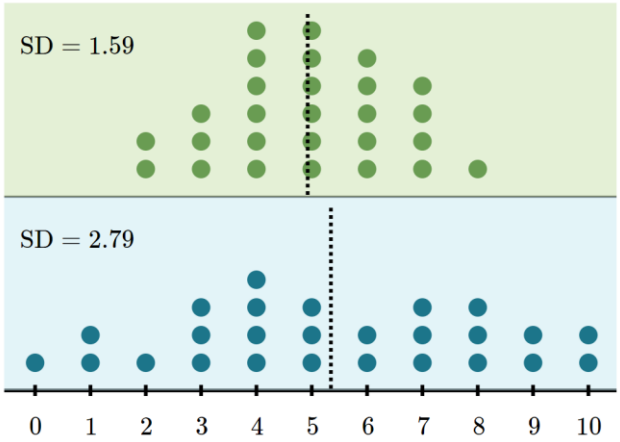
In the previous example, the variance = 4.67 years^2

So, the standard deviation, $s=\sqrt{s^2} = \sqrt{4.67} = 2.16$ years

If the data values are widely spread, the average distance of the values from their mean will be large, and the standard deviation will be large.

If the values are narrowly spread, this average distance will be small, and the standard deviation will be small.

The figure below shows how the spread of data affects the value of the standard deviation.

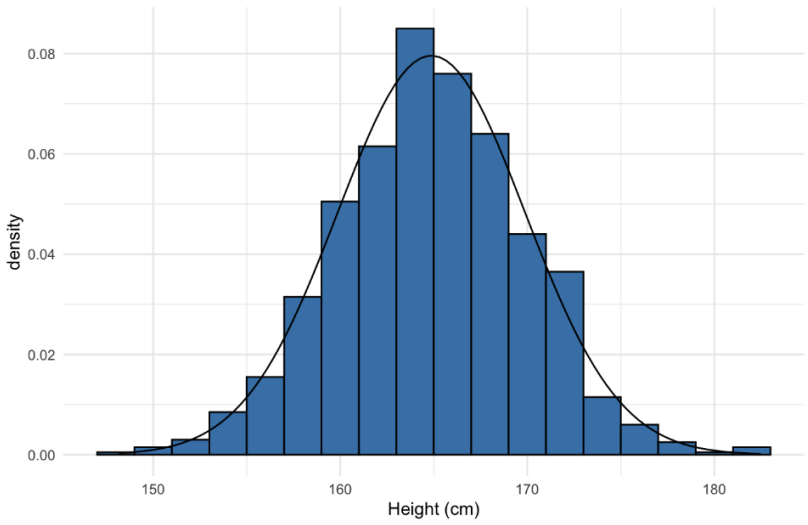


Combining measures of central tendency and measures of dispersion:

When summarizing a numerical variable, we present it using two measures; one for central tendency and one for dispersion.

- For the normally distributed data, we use the mean and standard deviation.
- For the non-normally distributed data, we use the median and inter-quartile range (IQR).

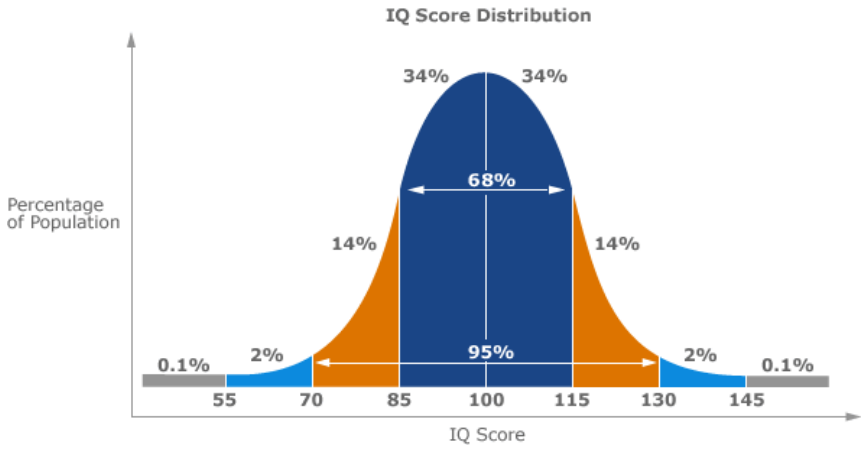
What is normally distributed data?



Normally distributed variables are common in biological measurements and have the following characteristics:

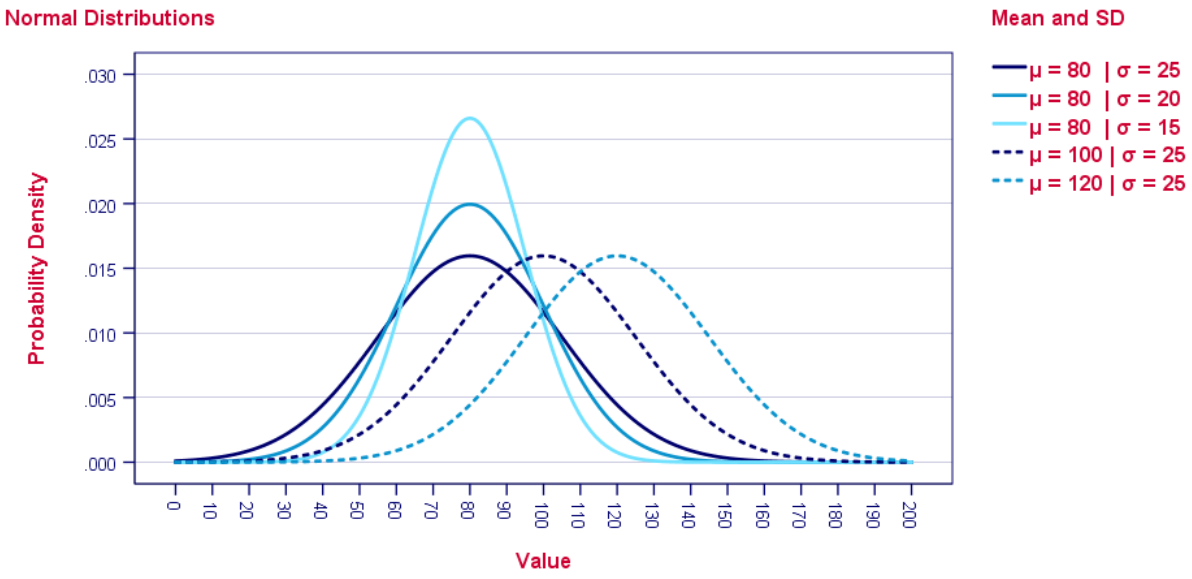
- Symmetric around the mean.
- The mean, median, and mode of a normal distribution are equal.
- Normal distributions are denser in the center and less dense in the tails (bell shape).
- 50% of values less than the mean and 50% greater than the mean
- Normal distributions are defined by two parameters, the mean (μ) and the standard deviation (σ).

- 68% of the area of a normal distribution is within one standard deviation of the mean.
- Approximately 95% of the area of a normal distribution is within two standard deviations of the mean.
- Approximately 99.7% of the area of a normal distribution is within three standard deviations of the mean.

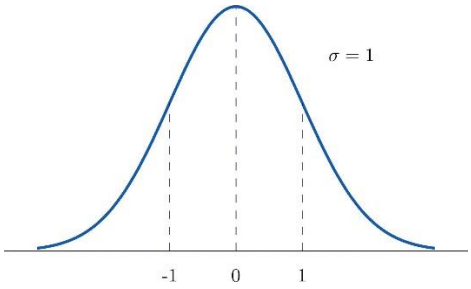


Examples for normally distributed data: height, blood pressure, IQ, …

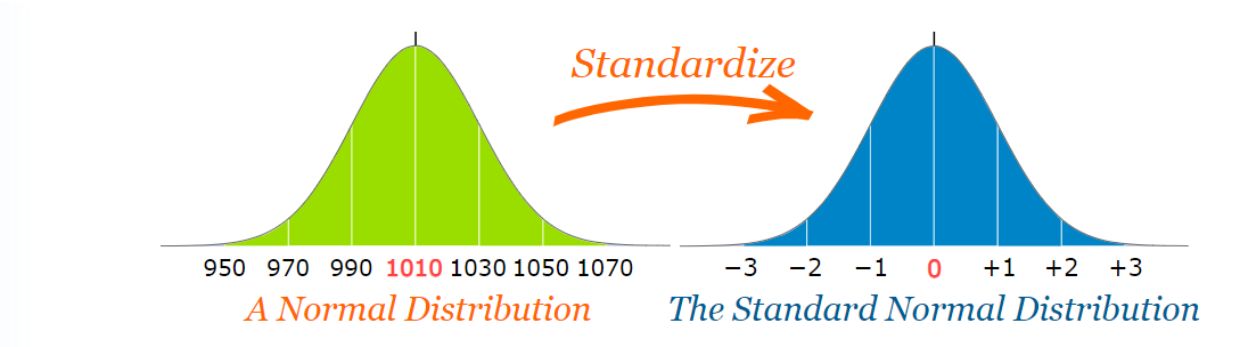
The following graph represents normal distributions with different means and standard deviations:



The normal distribution with mean μ = 0 and standard deviation σ = 1 is called the **standard normal distribution.**



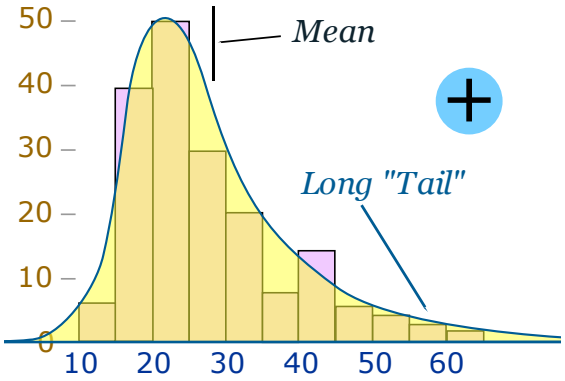
Any normal distribution values can be standardized (transferred to a standardized Z value)



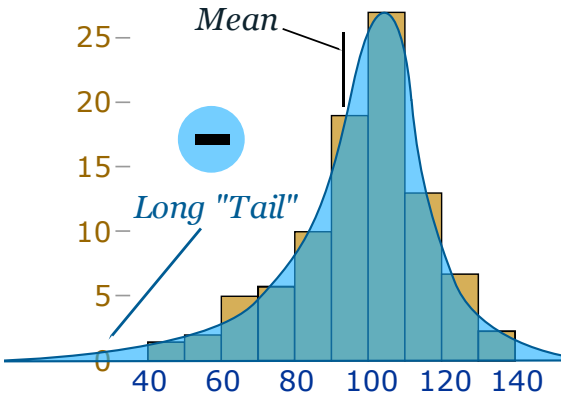
Examples of non-normally distributed data:

Data can be "skewed", meaning it tends to have a long tail on one side or the other.

Positive skew is when the long tail is on the positive side and is skewed to the **right**.



Negative skew is when the long tail is on the negative side and is skewed to the **left**.



Note that the mean is nearer to the tail (it is affected by the extreme values).

Tabular presentation of data

It is important to know how to present data in meaningful tables that are easy to understand.

Nominal Variables

- **Nominal variables: Frequency**
  We can present them as frequencies, the number of individuals in each category.
  For example, the nationalities of participants:

NATIONALITY	FREQUENCY (N= 180)
Bahraini	22
Egyptian	42
Iraqi	36
Lebanese	17
Qatari	8
Saudi	55

Here, the categories are arranged alphabetically, but as they don't have an order, it may be more comfortable for the reader to arrange them according to the frequencies.

We start with the nationality with the highest frequency to the lowest as follows:

NATIONALITY	FREQUENCY (N= 180)
Saudi	55
Egyptian	42
Iraqi	36
Bahraini	22
Lebanese	17
Qatari	8

- **Nominal variables: Relative frequency**
  Although reporting of frequencies is easy to understand, reporting the percentages (relative frequencies) is more comfortable for most people to get a sense of the data.
  It is calculated easily by dividing the number of individuals in each category and dividing it by the total number. Then we multiply it by 100 to get the percentage as follows:

NATIONALITY	FREQUENCY (N= 180)	RELATIVE FREQUENCY	HOW TO CALCULATE?
Saudi	55	30.6	$= \frac{55}{180} X\ 100$
Egyptian	42	23.3	$= \frac{42}{180} X\ 100$
Iraqi	36	20.0	$= \frac{36}{180} X\ 100$
Bahraini	22	12.2	$= \frac{22}{180} X\ 100$
Lebanese	17	9.4	$= \frac{17}{180} X\ 100$
Qatari	8	4.4	$= \frac{8}{180} X\ 100$

Ordinal Variables

- **Ordinal variables: Frequency**
  The same as nominal variables, ordinal variables are presented as frequencies:

SATISFACTION LEVEL	FREQUENCY (N= 140)
Very satisfied	43
Satisfied	55
Neutral	15
Dissatisfied	19
Very dissatisfied	8

But we have to respect the order of categories. Presenting them in a different order will confuse the readers.

- **Ordinal Variables: relative frequency**
  The same as nominal variables, percentages (relative frequencies) are calculated and presented as follows:

SATISFACTION LEVEL	FREQUENCY (N= 140)	RELATIVE FREQUENCY	HOW TO CALCULATE?
Very satisfied	43	30.7	$= \frac{43}{140}$ X 100
Satisfied	55	39.3	$= \frac{55}{140}$ X 100
Neutral	15	10.7	$= \frac{15}{140}$ X 100
Dissatisfied	19	13.6	$= \frac{19}{140}$ X 100
Very dissatisfied	8	5.7	$= \frac{8}{140}$ X 100

- **Ordinal Variables: Cumulative relative frequency**

Sometimes we use the cumulative relative frequency to present the ordinal variables making benefit from the order. They are presented and calculated as follows:

SATISFACTION LEVEL	FREQUENCY (N= 140)	RELATIVE FREQUENCY	CUMULATIVE RELATIVE FREQUENCY	HOW TO CALCULATE?
Very satisfied	43	30.7	30.7	30.7
Satisfied	55	39.3	70.0	30.7+39.3=70
Neutral	15	10.7	80.7	70.0+10.7=80.7
Dissatisfied	19	13.6	94.3	80.7+13.6=94.3
Very dissatisfied	8	5.7	100.0	94.3+5.7=100

The cumulative relative frequency at one level is calculated simply by adding the relative frequency at this level to all relative frequencies before this level.

For example, if the cumulative relative frequency at the "satisfied" level is 70%, this means that 70% of the individuals are either satisfied or very satisfied. While the cumulative relative frequency at the "neutral" level is 80.7% meaning that 80.7% of the participants are very satisfied, satisfied, or neutral.

Numerical Discrete Variables

- **Numerical Discrete Variables: Frequency, relative frequency, and cumulative relative frequency**

If the numerical discrete variable is of few levels, we can represent it in frequencies, relative frequencies, and cumulative relative frequencies in the same way as in ordinal variables.

For example, the number of kids in the family:

NUMBER OF KIDS	FREQUENCY (N= 240)	RELATIVE FREQUENCY	CUMULATIVE RELATIVE FREQUENCY
0	32	13.3	13.3
1	64	26.7	40.0
2	83	34.6	74.6
3	42	17.5	92.1
4	13	5.4	97.5
5	6	2.5	100.0

Here, for example, 74.6% of the families have two kids or less (2, 1, or 0).

Numerical Continuous Variables

- **Numerical Continuous Variables: Frequency, relative frequency, and cumulative relative frequency**

If we are dealing with a continuous variable as the birth weight in grams, it is impractical and useless to present the frequencies for each birthweight we observe in grams.

Instead, we can group the variable into groups of equal width: (2000-2499, 2500-2999, 3000-3499, 3500-3999, and 4000-4500).

For those groups, we can present the frequency, relative frequency, and cumulative relative frequency as we did before.

BIRTHWEIGHT (G)	FREQUENCY (N= 45)	RELATIVE FREQUENCY	CUMULATIVE RELATIVE FREQUENCY
2000-2499	3	6.7	6.7
2500-2999	13	28.9	35.6
3000-3499	18	40.0	75.6
3500-3999	7	15.6	91.1
4000-4499	4	8.9	100.0

Sometimes, instead of having some groups with very few frequencies at the lower or the upper end, we group them into one group less than a specific value, or one group that is higher than a specific value and call them "open-ended groups "as in the following table representing the age of patients:

AGE OF PATIENT	FREQUENCY (N= 120)
≤19	**5**
20-24	42
25-29	36
30-34	30
≥ 35	**7**

We notice that the first and last groups are open-ended.

Two Categorical Variables

- **Cross- tabulation: two-way table**

Sometimes we are interested in presenting two categorical variables in the same table, which we call the two-way table (as we have two variables).

A table representing the relationship between sex and the disease status can be as flows:

		Sex		
		Male	Female	total
Disease	Diseased	24	18	42
	Not diseased	41	35	76
	total	65	53	118

From this table we can get the following information:

- Total number of participants: 118 (cell in the right lower corner)
- Total number of males: 65 (lower margin)
- Total number of females: 53 (lower margin)
- Total number of diseased: 42 (right margin)
- Total number of not diseased: 76 (right margin)
- Males and diseased: 24
- Females and diseased: 18
- Males and not diseased: 41
- Females and not diseased: 35

We can even make the table more informative by adding percentages by rows or columns.

Adding percentages by rows gives us the following table:

		Sex		
		Male	Female	total
Disease	Diseased	24 57%	18 43%	42 100%
	Not diseased	41 54%	35 46%	76 100%
	total	65 55%	53 45%	118 100%

From the percentages presented in the table we can see that:

- The total percentage of males is 55% while that of females is 45% (last row)
- The percentage of males among diseased is 57% while that of females is 43%.
- The percentage of males among not diseased is 54% while that of females is 46%.

Adding percentages by columns gives us the following table:

		Sex		
		Male	Female	total
Disease	Diseased	24 37%	18 34%	42 36%
	Not diseased	41 63%	35 66%	76 64%
	total	65 100%	53 100%	118 100%

From the percentages presented in the table we can see that:

- The total percentage of diseased is 36% while that of not diseased is 64% (last column).
- The percentage of diseased among males is 37% while that of not diseased is 63%.
- The percentage of diseased among females is 34% while that of not diseased is 66%.

Three Categorical Variables

- **Cross- tabulation: Three-way table**

Three categorical variables can be presented in the same table such as sex, disease status, and smoking status as follows:

		Sex	
		Male	Female
Smoker	Diseased	36	42
	Not diseased	22	18
Non-Smoker	Diseased	24	18
	Not diseased	41	35

In this table the three variables are presented, we can add more numbers as total numbers and percentages, but we prefer to keep it simple. The arrangement of the variables can be also changed.

It all depends on what information we want to tell the reader.

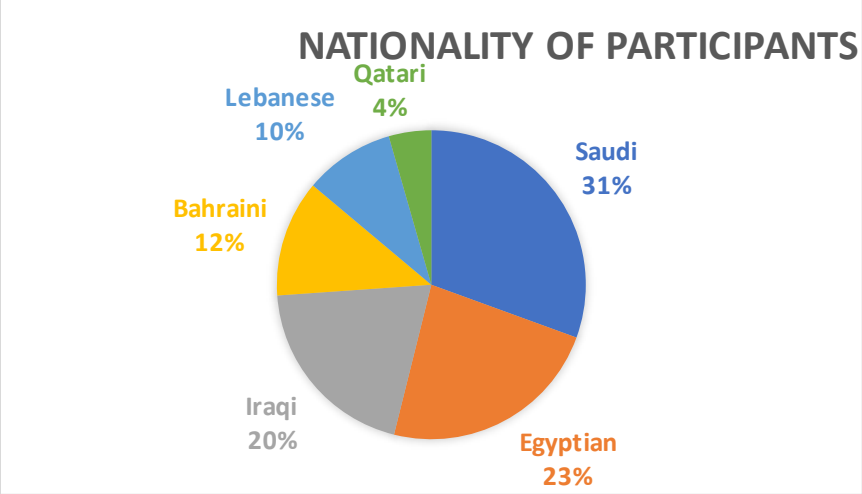
Graphical presentation of data

It is important to use the appropriate graph for each data type that clearly delivers the meaning.

We will illustrate each type of data variable with the appropriate graphs that can represent it.

Nominal Variables

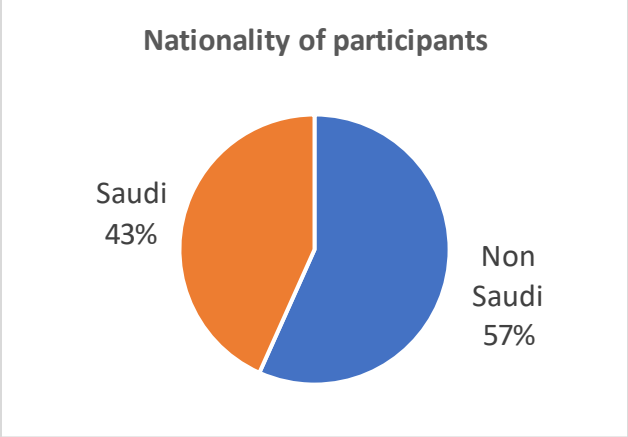
- **Nominal variables: Pie chart**
  The pie (circle) represents 100% of the variable and is divided into sectors. The area of each sector represents the frequency of each category in the variable it represents as follows:



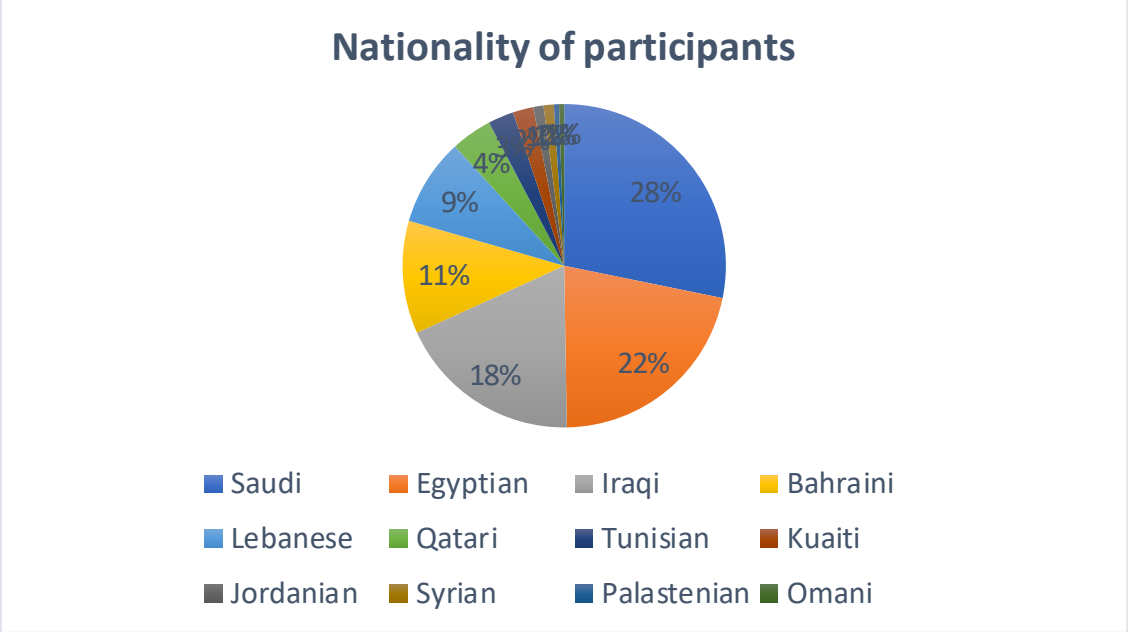
The pie chart is less commonly used in scientific papers due to its limitations. It can present only one variable.

If the categorical variable is binary (dichotomous), it will not be that informative and if the number of categories is large, the graph will not be that clear as follows:

A pie graph of a binary variable:

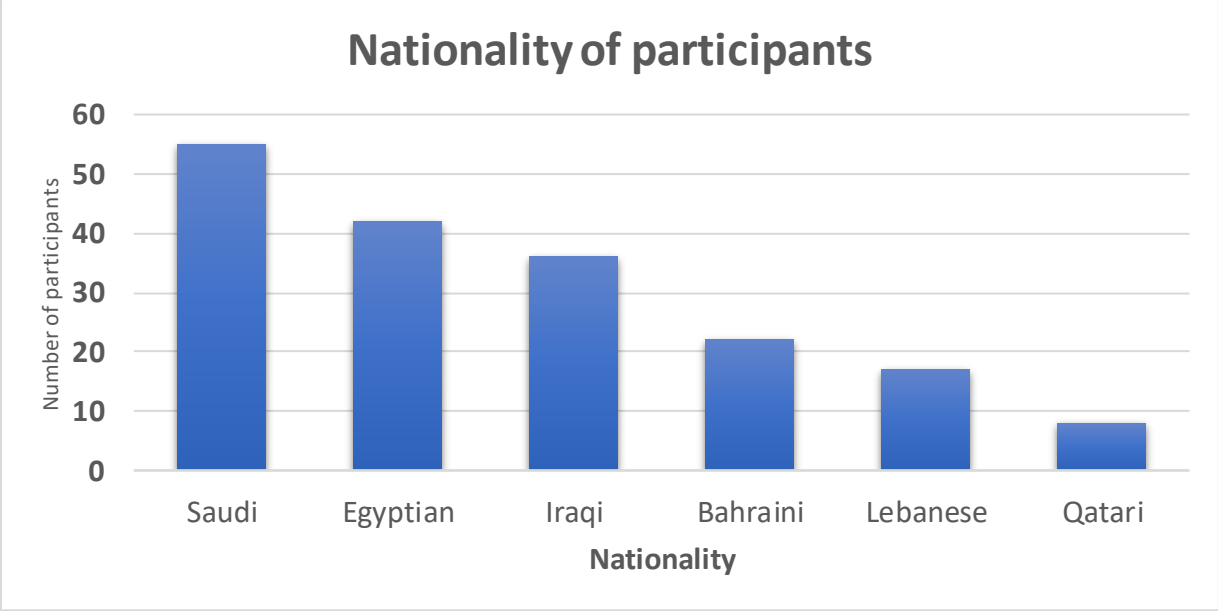


A pie graph of a variable with a large number of categories that is **not clear**



- **Nominal variables: Bar graph**

Bar graphs are more commonly used to represent categorical variables. It can be vertical or horizontal graphs and can show the frequency or the percentage of each category.

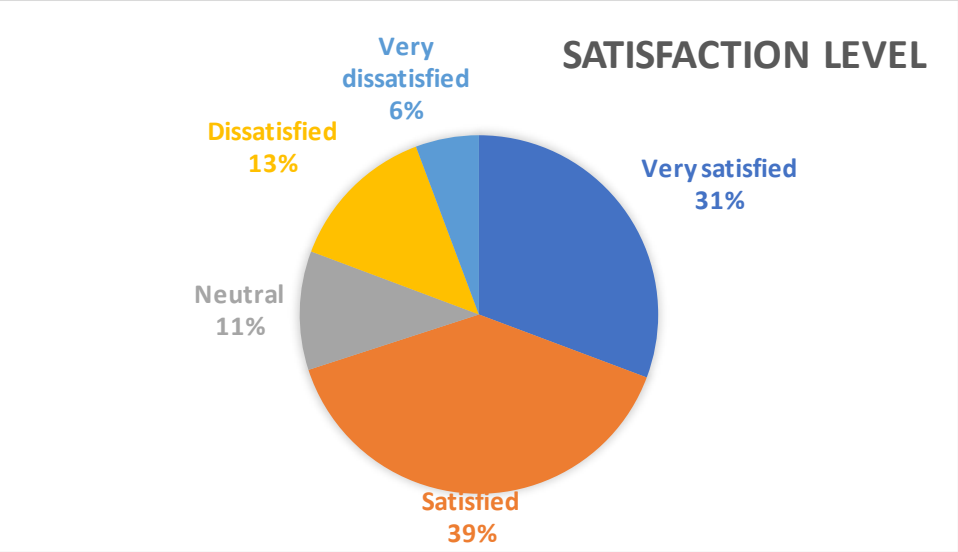


As the nominal variables have no meaningful order, it may be better looking to rearrange the categories based on their frequencies as in the graph above.

Ordinal Variables

- **Ordinal variables: Pie chart**

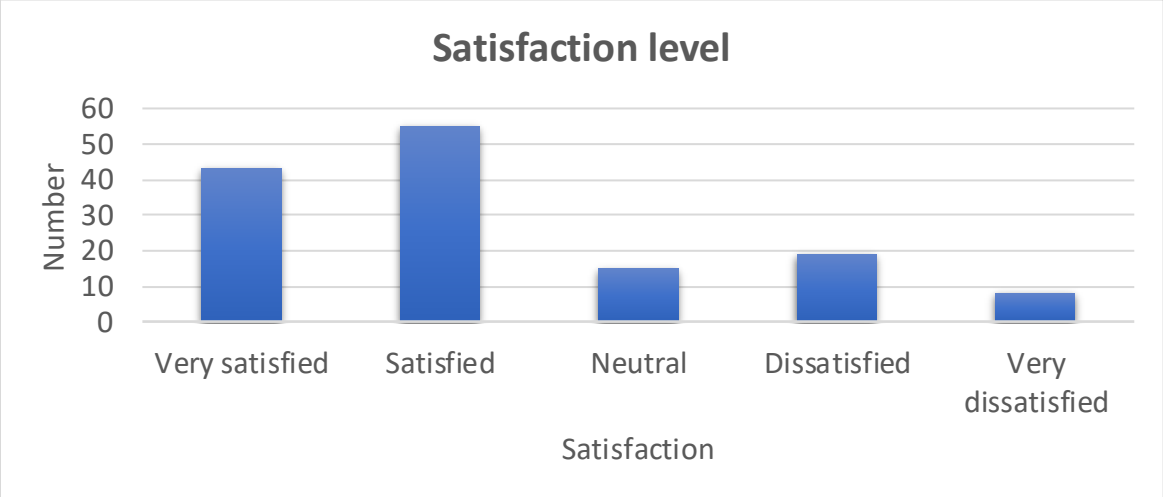
  Ordinal variables can be presented using the pie chart as in the nominal variables.



- **Ordinal variables: Bar graph**

  Bar graphs may be the best way to present an ordinal variable. As in nominal variables, it can present either the frequency or the percentage.

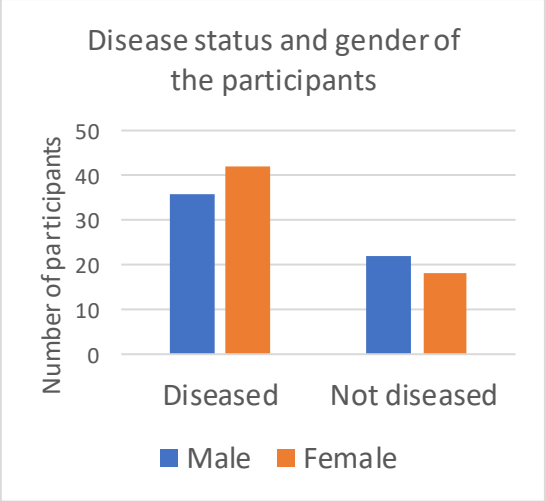
  Here, we can't change the order of the categories, otherwise, the reader may get confused.

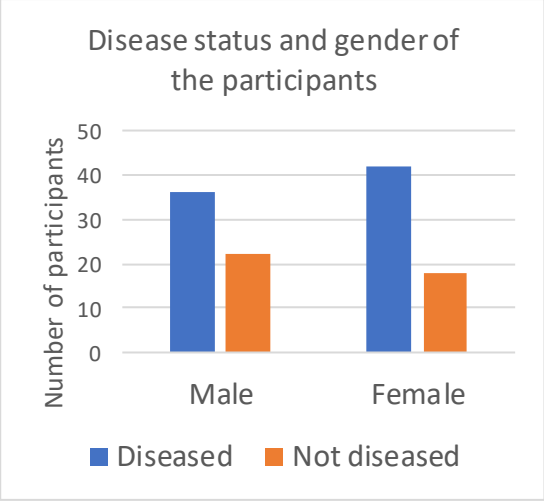


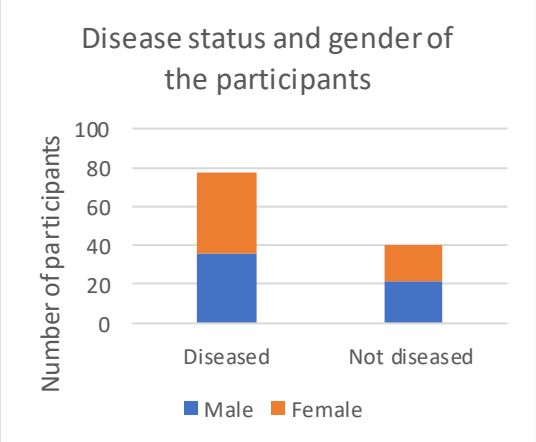
- As a general concept, we should use the graph the best demonstrates our data.
- Pie charts are not commonly used in scientific papers, they are usually used for presentations.

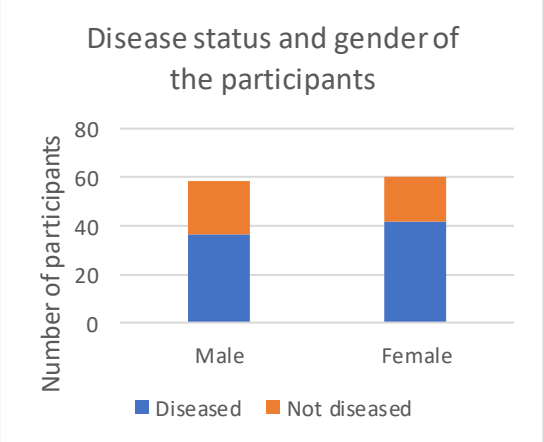
Two Categorical Variables

- **Two categorical variables: Bar graphs**
  Presenting two categorical variables in the same chart can be done using bar graphs. Either **segmented bar charts** or **side-by-side bar charts** can be used.
  The following four graphs present the same data of the two variables, gender and disease status.
  We can choose any of them based on which presents our results the best.









Numerical Variables

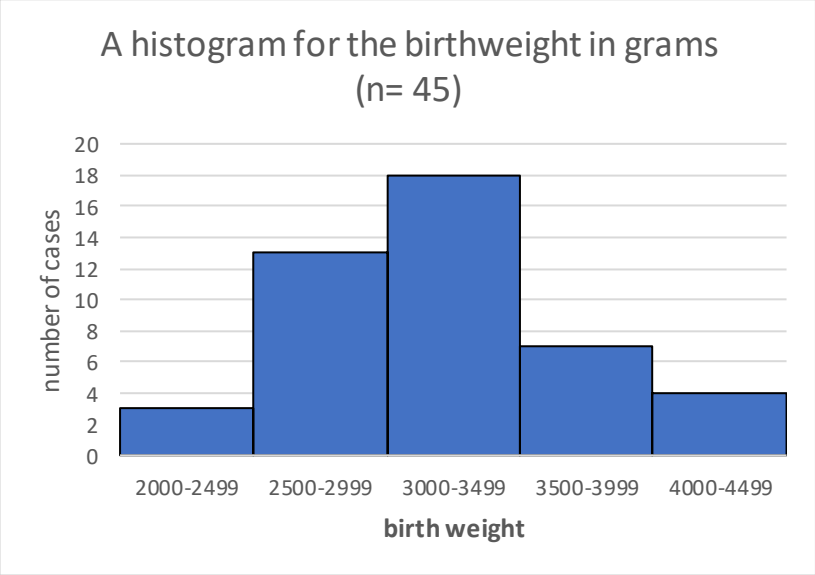
- **Numerical variables: Histogram**

  It is similar to the bar chart, but there are no gaps between the bars as the variable is continuous.

  The width of each bar of the histogram relates to a range of values for the variable, but in most cases, the width is kept the same.

  For example, a numerical variable as the birth weight in grams can be presented in the following groups (2000-2499, 2500-2999, 3000-3499, 3500-3999, and 4000-4500) with each group represented by a column.

  The height of the column represents the frequency of cases in this group.

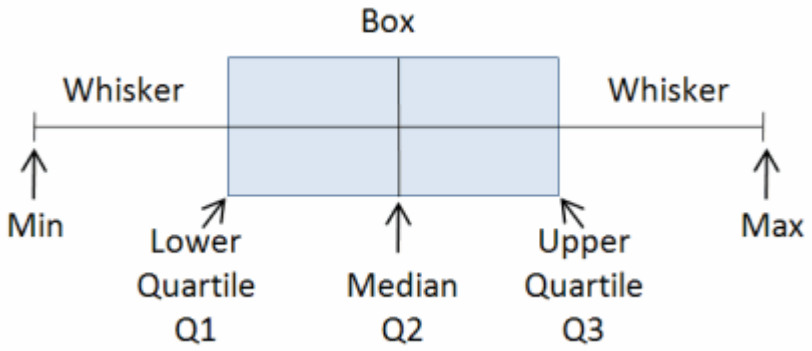


- **Numerical variables: Box plot**

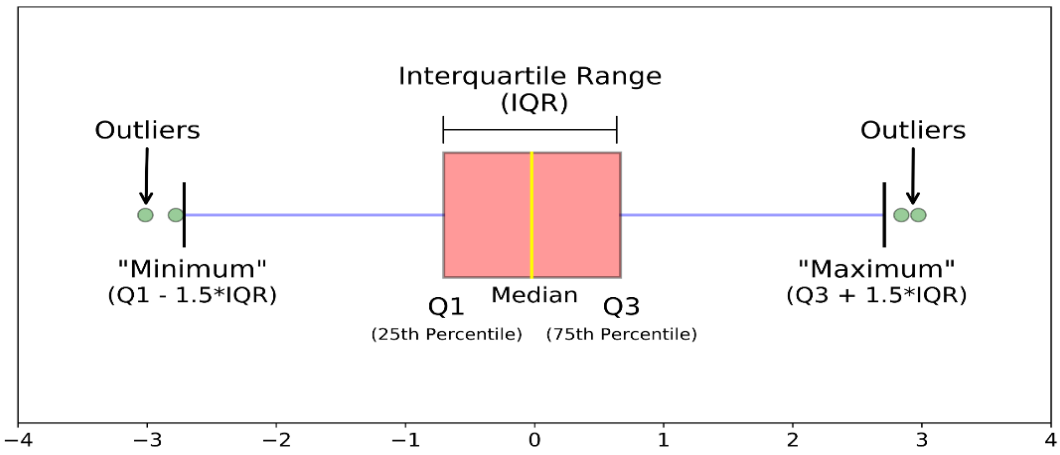
  The boxplot (also called Box and whisker plot) is used to summarize numerical variables based on the five-number summary.

  Those five numbers are minimum, maximum, median, upper quartile, and lower quartile.

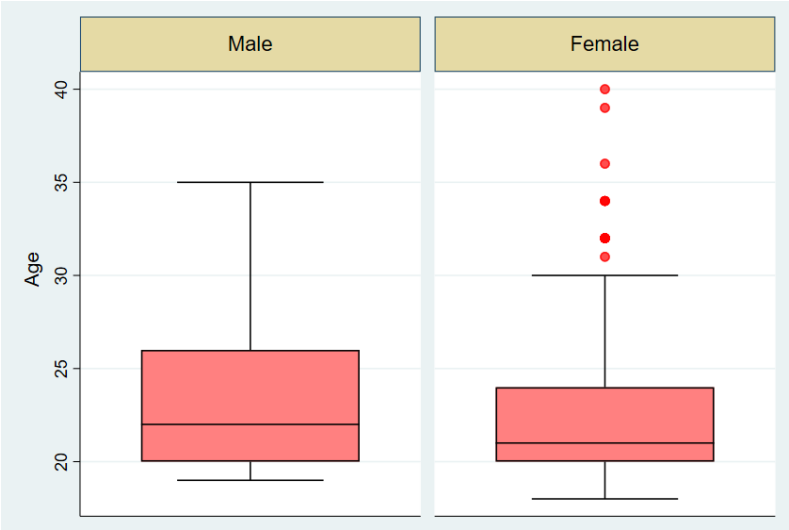
  • Median = horizontal line in the box
  • Upper quartile = top edge of the box
  • Lower quartile = lower edge of the box
  • Maximum = top of 'whisker'
  • Minimum = bottom of 'whisker'



The whiskers are limited to outside 1.5 times the interquartile range above the upper quartile and below the lower quartile (Q1 - 1.5 * IQR or Q3 + 1.5 * IQR).
Boxplot is useful in showing the outliers (presented as dots outside the limits of the whiskers).



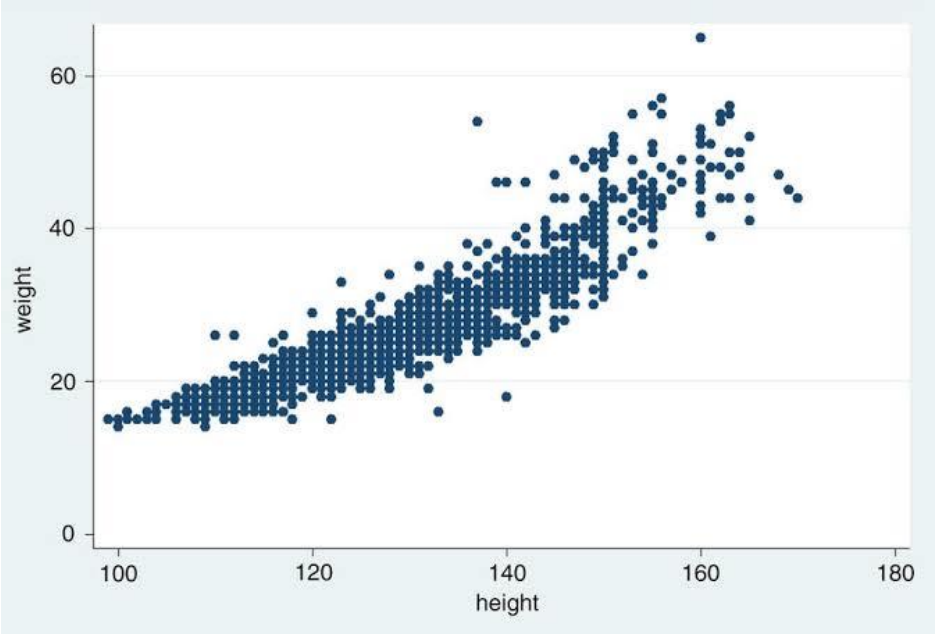
It is useful in comparing the same numeric variable across different groups as comparing a score between men and women.
The following graph shows a boxplot for men and a boxplot for women allowing us to compare the same variable (age) between the two groups.



Two Numerical Variables

- **Two numerical variables: Scatter plot**
  - If we have two variables that are numerical (or ordinal), the relationship between them can be illustrated using a scatter diagram.
  - It plots one variable against the other in a two-way diagram.
  - One variable is represented on the horizontal axis and the other is plotted on the vertical axis with each dot representing one case.

The following scatter plot represents the relationship between weight and height.



Hypothesis testing

The research question:

In any research project, it is not enough for the researcher to have an idea, we need to formulate it into a research question.

A research question should be:

- A question
- Specific (time/place/subjects/condition)
- Answerable
- Novel
- Relevant to medical knowledge

Examples of research questions:

- What is the prevalence of diabetes mellitus in Egypt?
- Does lowering blood pressure reduce the risk of coronary heart disease in diabetic patients?
- Is prognosis following stroke dependent on age at the time of the event?
- Is drug A better than drug B in lowering blood pressure?
- Is there a difference between males and females regarding response to drug X?

If we have a research question and we want to reach a conclusion about it, we do what is called hypothesis testing.

Steps for hypothesis testing

1. Define the null and alternative hypotheses.
2. Choose the level of significance.
3. Pick and compute the test statistic.
4. Compute the p-value
5. Check whether to reject the null hypothesis by comparing the p-value to the level of significance.
6. Draw conclusion from the test.

We will go through those steps in detail.

The null and alternative hypotheses

Before starting with any statistical analysis, we begin with defining the "hypotheses" based on the research question we are trying to answer.

For each research question, we define two types of hypotheses; the **null hypothesis (H_o)** and the **alternative hypothesis (H_1)**.

Both are mutually exclusive (not overlapping). Only one of them is true!

H_o: Null hypothesis	H_1 /H_a: Alternative hypothesis
▪ Is the currently accepted belief/ idea /parameter ▪ Nothing is happening / there is no difference / there is no association ▪ The researcher doubts it to be true	▪ Is the researcher's idea ▪ something is happening/ there is a difference/ there is an association ▪ The researcher believesit to be true and wishes to prove

Examples:

1)

Research question	**Is there a difference in the exam scores between males and females?**
The null hypothesis (H_o)	The difference between two groups' means is zero Mean score of the males = mean score of the females Mean score of the males – mean score of the females= 0
The alternative hypothesis (H_1)	The difference between the two groups' means is not zero Mean score of the males ≠ mean score of the females Mean score of the males – mean score of the females≠ 0

2)

Research question	**Does a new diabetes treatment reduce blood glucose different than an existing treatment?**
The null hypothesis (H_o)	The mean reduction in blood glucose level is the same in the two treatment groups.
The alternative hypothesis (H_1)	the mean reduction in blood glucose is different in the two treatment groups.

3)

Research question	Is there an association between smoking and the risk of cardiovascular diseases?
The null hypothesis (H_o)	There is no association between smoking and the risk of cardiovascular disease.
The alternative hypothesis (H_1)	There is an association between smoking and the risk of cardiovascular disease.

We perform a statistical analysis to test our hypotheses and reach a conclusion regarding the null and alternative hypotheses.

The conclusion is either :

- Fail to reject the null hypothesis (accept the null hypothesis) and conclude that nothing is happening / no difference / no association.
- Reject the null hypothesis (accept the alternative hypothesis) and conclude that something is happening/ there is a difference/ there is an association.

The decision is done regarding the null hypothesis

We use REJECTING & FAILING TO REJECT (not accepting)
• Acceptance implies that the null hypothesis is true.
• Failure to reject implies that the data are not sufficiently persuasive for us to prefer the alternative hypothesis over the null hypothesis

One-tailed and two-tailed tests:

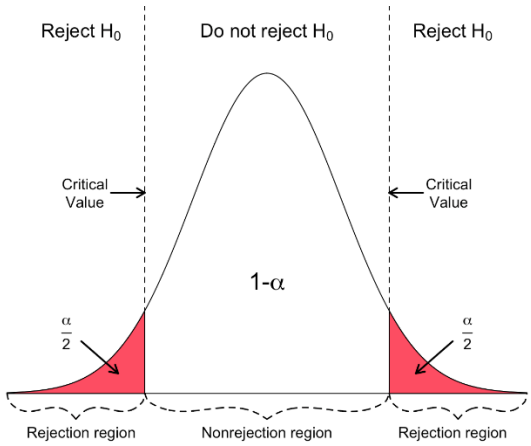
Based on the alternative hypothesis, the used statistical test can be either a two-tailed test or a one-tailed test

The two-tailed tests	The one-tailed tests
▪ The alternative hypothesis allows for the difference **to be in either of the two directions**. ▪ As in previous examples: Exam scores could be higher for males or females, and any of the two drugs may reduce blood glucose more than the other. **H_o: drug A = drug B** **H_1: drug A ≠ drug B**	▪ The alternative hypothesis is **in one direction only.** For example: ▪ We have a new drug A, and we want to examine only if it is **better** than standard drug B. **H_o: drug A ≤ drug B** **H_1: drug A > drug B**

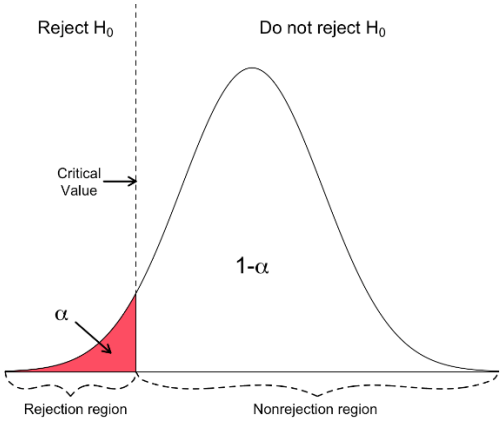
In most cases, we use the two-tailed tests unless there is a clear justification for using a one-tailed test.

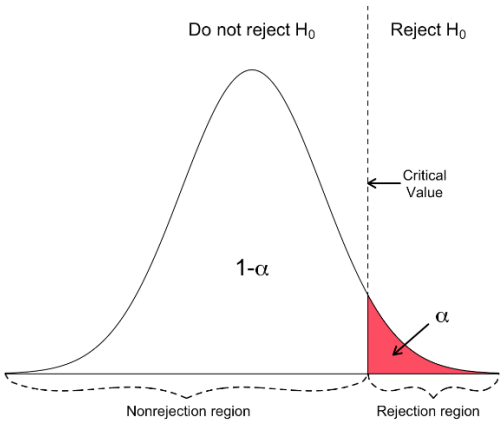
The following graphs illustrate the difference between the two types :

Two-tailed tests:



One tailed tests:





Type 1 and type 2 errors

There are two types of errors that can be committed while performing statistical analysis (hypothesis testing), type 1 errors and type 2 errors.

They are illustrated using the following 2 cases:

Case 1:

In the real world (population): Smoking rate in males = smoking rate in females

We have a research idea (question) that smoking rates are different in the two groups

As before,

H_0: Null hypothesis: nothing is happening / no difference

H_1: Alternative hypothesis: the idea of the researcher

So,

H_0: Smoking rate in males = smoking rate in females

H_1: Smoking rate in males ≠ smoking rate in females

Decisions based on the statistical test:

1- Accept the null hypothesis and conclude that smoking rate in males = smoking rate in females.

(which is the correct decision)

2- Reject the null hypothesis, accept the alternative hypothesis and conclude that smoking rate in males ≠ smoking rate in females.

(here we made a mistake)

Type I error / false positive / α

Here, we made a mistake by rejecting a true null hypothesis and is called type I error, or α. We reached a false positive conclusion.

This type of error is serious, as we reach a false positive conclusion (we accept the alternative hypothesis of the researcher). This (false) conclusion may be that a drug is effective while it is not, or something is a risk factor for a disease while it is not!

Researchers tend to keep the probability of this type of error as low as possible.

It is usually set at 5% (α=0.05) and sometimes they are even more conservative and make it 1% (α=0.01).

Case 2:

In the real world (population): Smoking rate in males ≠ smoking rate in females

We have a research idea (question) that smoking rates are different in the two groups

As before,

H_0: Null hypothesis: nothing is happening / no difference

H_1: Alternative hypothesis: the idea of the researcher

So, **H_0:** Smoking rate in males = smoking rate in females

H_1: Smoking rate in males ≠ smoking rate in females

Decisions based on the statistical test:

1- Reject the null hypothesis, accept the alternative hypothesis and conclude that smoking rate in males ≠ smoking rate in females

(which is the correct decision)

2- Accept the null hypothesis and conclude that smoking rate in males = smoking rate in females

(here we made a mistake)

Type II error / false negative / β

Here, we made a mistake by accepting a false null hypothesis and is called type II error, or β. We reached a false negative conclusion.

This type of error is less serious than the first one as we here reach a false negative conclusion which means that we conclude that a drug is not effective while it is truly effective.

The probability of this type of error is usually set at 20% (β=0.2).

The following graph illustrates type 1 and type 2 errors.

		The Truth (Based on the entire population)	
		Nothing is there **H_0 is True**	Something is there **H_0 is False**
Your conclusion (Based on your sample)	I do not see anything (Non-significant)	✓	✗ **Type II error**
	I see something (Significant)	✗ **Type I error**	✓

The probabilities of the two errors are inversely related.

Level of significance

The level of significance (α) is the maximum allowed probability of committing a Type I error.

The smaller the value of α, the lower the risk of committing a Type I error.

Hence, we choose a level of significance depending on the consequence of committing a Type I error.

Common values for α are 0.05 and 0.01 indicating 5% and 1%, respectively.

Power

The probability of **not** committing a Type II error is called the **power** of a hypothesis test.

Power = 1-β

The statistical power of a study is the power (or ability) of the study to detect a difference if a difference really exists.

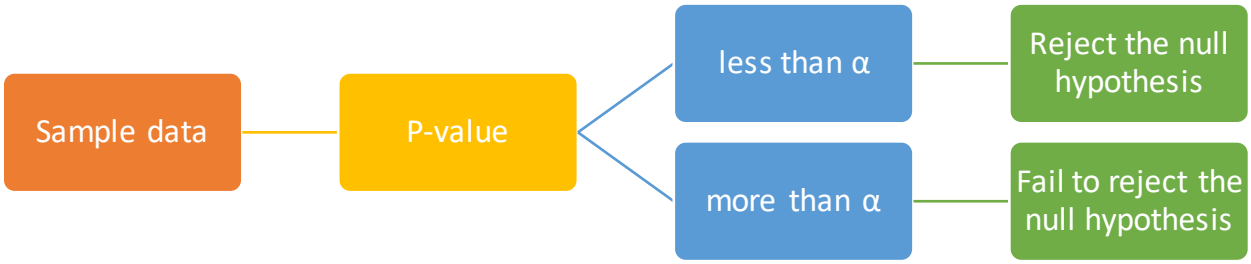
- In practice, β is usually set at 0.2.
- This provides a power value of 0.8 (80%).
- If there is a difference, then the probability of the test detecting it is 80%.

P-value

If we have a research question, we collect data to test the related hypothesis.

We perform the proper statistical test based on the nature of the data and the research question.

The end result of any statistical test is a **p-value**. The '*P* ' stands for probability. We use this p-value to make a decision about the null hypothesis by comparing its value to the level of significance (α).



What is the P-value?

If the null hypothesis is true, the p-value is the probability of obtaining this result (or something more extreme).

In other words, the p-value is the probability of seeing the observed difference (in the collected data), or greater, just by chance if the null hypothesis is true.

For simplicity (but less accurate scientifically) think of the p-value as:

P-value: The **probability** that random **chance** generated the data or something else that is equal or rarer.

• A P-value is a probability and therefore lies between 0 and 1.

• It expresses the weight of evidence in favor of or against the stated null hypothesis.

• 0.05 or 5% is commonly used as a cut-off (significance level).

• If the observed p-value is less than this ($p<0.05$), we consider that there is good evidence that the null hypothesis is not true. So, we reject the null hypothesis (and accept the alternative hypothesis).

• $P<0.05$ is described as statistically significant and $P\geq0.05$ is described as not statistically significant.

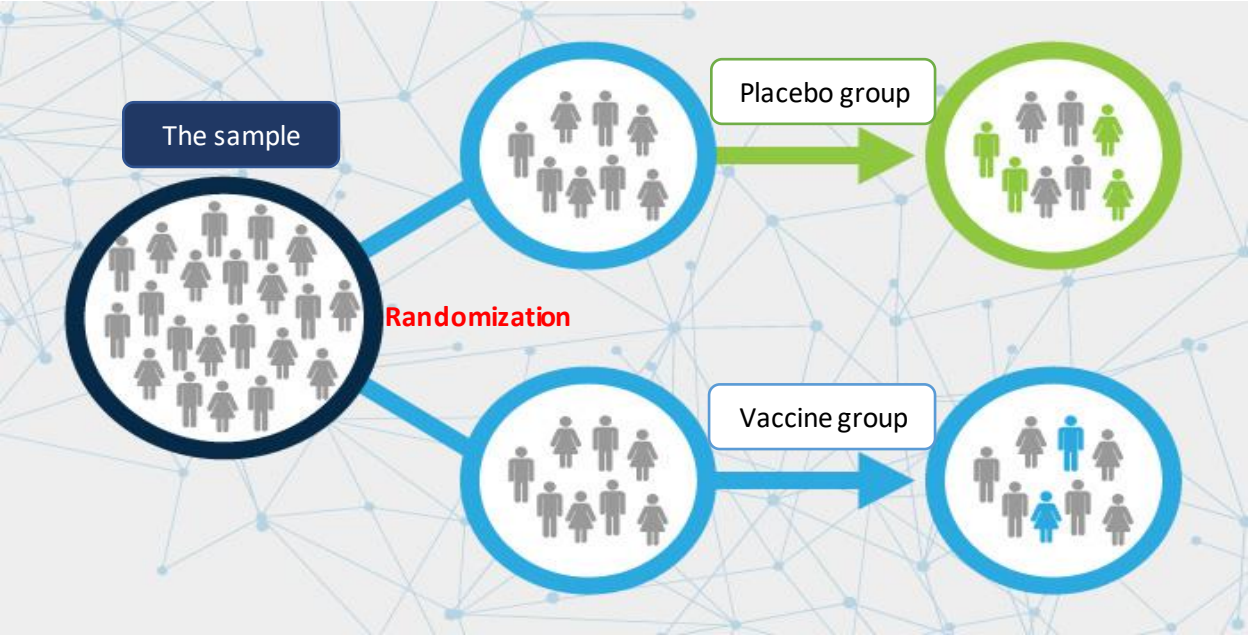
• Exact p-values should be reported rather than reporting findings as P<0.05 or P≥0.05 or worse "P=NS" (meaning non-significant).

• The p-value in some statistical programs may be very small and appear in the output as 0.000 this value is very small but not equal to zero. It should be reported as p < 0.001.

Example:

If we have a new vaccine and want to test if it is working. We design an experiment to give the vaccine to one group and a placebo to the control group. We collect data about the incidence of the disease in both groups and perform the proper statistical test to get the p-value.

(0.05 is used as the level of significance)



There are two possibilities:

A significant p-value (P<0.05)	A non-significant p-value (P≥0.05)
▪ If the obtained p-value is 0.02. ▪ This means that if the null hypothesis is true (there is no effect of the vaccine), the probability of getting this result (the observed one) or even more extreme result is 2%. ▪ This small p-value indicates that the observed data is not consistent with the null hypothesis – they are unlikely to have occurred if the null hypothesis was really true. ▪ This small p-value provides evidence against the null hypothesis. **So, we reject the null hypothesis and conclude that there is a strong evidence that the vaccine is effective.**	▪ If the obtained p-value is 0.21. ▪ This means that if the null hypothesis is true (there is no effect of the vaccine), the probability of getting this result (the observed one) or even more extreme result is 21%. ▪ This p-value indicates that the observed data is consistent with the null hypothesis – they are likely to have occurred if the null hypothesis was really true. ▪ This p-value provides evidence in favor of the null hypothesis. **So, we fail to reject the null hypothesis and conclude that there is insufficient evidence that the vaccine is effective.**

So, in the end, the decision rule is simple:

• Get the p-value for the data you have collected (using the computer software).

• Compare it with the critical value (usually 0.05 or 0.01). It is called the significance level and is denoted as α (alpha).

- If the p-value is **greater** than or equal to the level of significance α, then we **do not reject** the null hypothesis.
- If the p-value is **less** than the level of significance, then we **reject** the null hypothesis.

Confidence Interval

To understand the concept of the confidence interval, let's think of the following example:

Suppose a researcher wanted to know the mean weight of students in one university. As it is not practical to measure the weight of all students, he chose a random sample of 50 students and measured their weight.

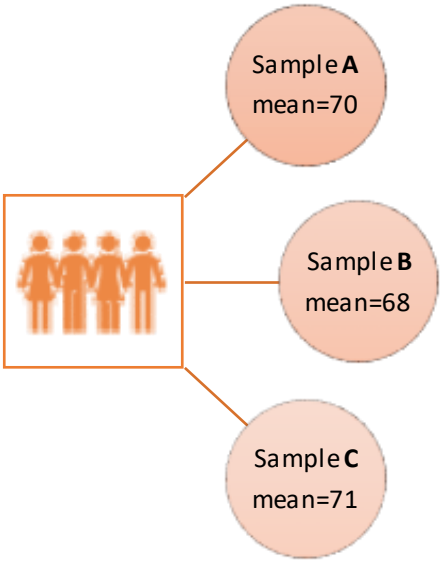
The mean weight that he got was 70 kg. As the sample is representative of the university students, we expect this number to represent the whole university students.

- But, what if we measure the weight of all the students, will we get this mean of 70kg?

Maybe and maybe not.

- What if we take another random sample of 50 students, will we get this mean of 70kg?

  Maybe and maybe not.



So, it is better to think of a range of values that most probably includes the true mean of the population (university students).

This range of values is called the confidence interval.

A Confidence Interval (CI) is a range of values we are fairly sure the true value lies in.

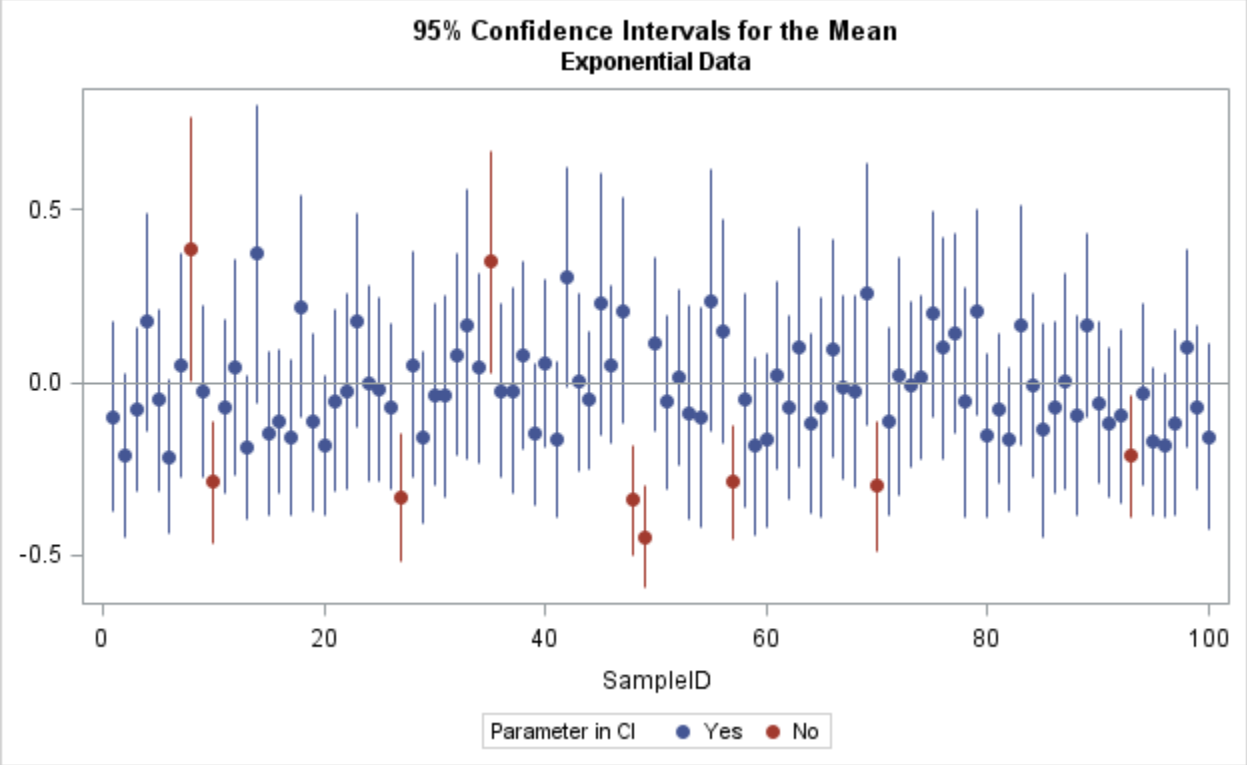
The confidence interval is usually expressed as **95% CI**, but we can see a **99% CI** or **90% CI** or any other percentage confidence interval.

In our example, if the 95% CI of the mean weight of students is 67 and 73 kgs. This is interpreted as:

We are 95% confident that the true population mean (mean of all university students' weight) lies between 67 and 73 kg. (confidence interval is sometimes interpreted in this way as it is easier to understand although this is not scientifically accurate).

A more scientifically accurate interpretation is that if we repeat the experiment a large number of times or infinite times, **95% of all confidence intervals constructed using this procedure should contain the true population mean**.

The figure below shows a simulation of different constructed confidence intervals. Those in blue contain the true mean (0), while those in brown don't include the true mean.



Example:

The average BMI before intervention in a study, of a group of 200 obese patients, was 29 kg/m^2. After being on a specific diet for 6 months, the mean BMI dropped by 2.5 kg/m^2

If the 95% CI is 1.5–3.5 kg/m^2, this means we can be 95% confident that the true effect of this diet is to lower the BMI by 1.5–3.5 kg/m^2.

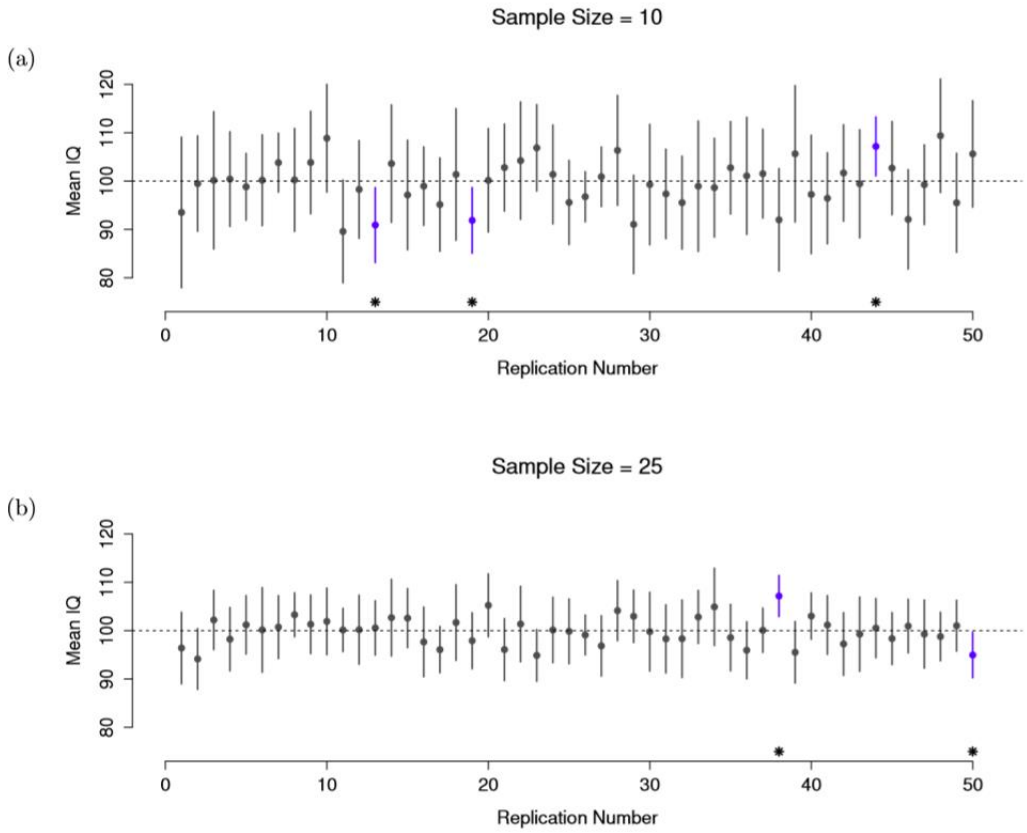
Some important notes:

1- Confidence Interval and sample size

If we take a sample of 500 students instead of 50, we expect that the mean of this sample to be more accurate in presenting the whole population than the smaller sample and the 95% CI will be nearer to the mean we calculated (which means it is narrower). If the sample mean is 70kg in both samples, the 95% CI for the small sample may be 67 and 73 kg, but for the large sample, it will be 69 and 71.

The larger the sample size, the narrower the confidence interval

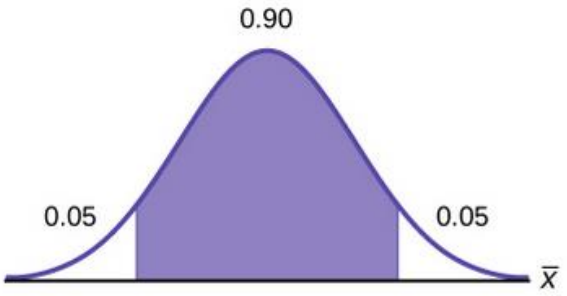
The following figure shows 95% confidence intervals of 50 simulated replications of an experiment in which the IQs are measured for 10 people (top), and 25 people (bottom)



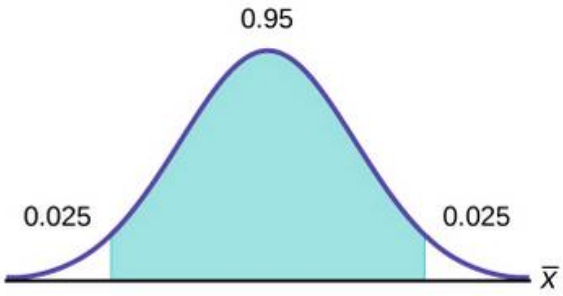
2- Confidence Interval and confidence level

If we are interested only in 90% CI (allowing for an error of 10%), then the calculated CI will be narrower (68 and 72 for example).

This narrowing allows an error of 10% compared to an error of 5% only in the 95% CI.

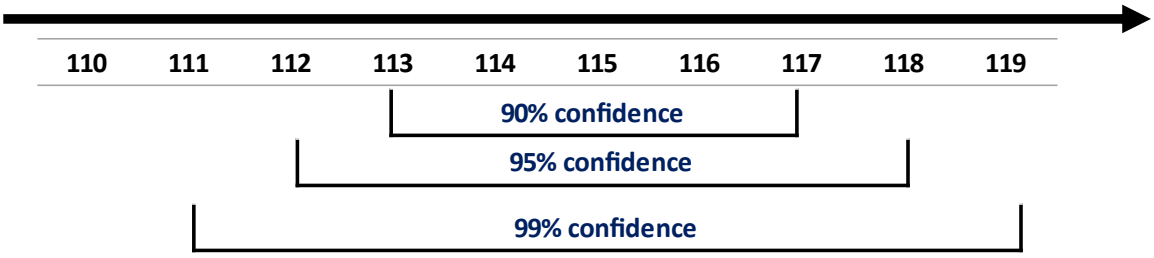


90% CI



95% CI

The following figure illustrates the 90%, 95%, and 99% confidence intervals of the IQ levels of students of a university where the mean IQ for a random sample of 100 students was 115.



3- Confidence Interval for a proportion

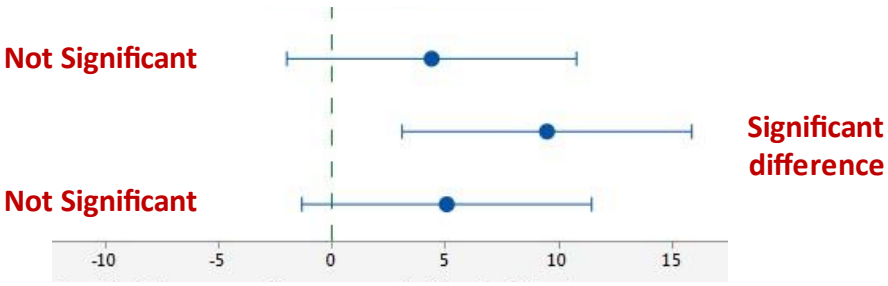
The confidence interval can be also calculated for a proportion.

For example, if we take a sample of university students to estimate the prevalence of smoking. We may get a prevalence of smoking of 12% with a 95% CI of 10% to 14%.

4- Confidence Interval of a difference

If we are calculating the confidence interval for the difference between two groups' means representing two populations and this confidence interval included zero (-2 to 6 for example), this means that the difference between those two means can be zero, and we will conclude that there is no significant difference between the two populations' means.

- The result of a (1-α)100% interval estimation is consistent with the result of the corresponding 2-tailed test at α level of significance.
  If the 95% CI is containing 0 (no significant difference), the p-value is not significant at α=0.05



If the confidence interval of the difference doesn't contain zero, the means are significantly different.

Example:

The following table represents two cases of comparing two groups with the mean difference and the 95% CI of the difference.

Case 1	Case 2
Sample 1 Mean (M_1): 70 Sample 1 Size (n_1): 50 Standard Deviation 1 (s_1): 10 Sample 2 Mean (M_2): 68 Sample 2 Size (n_2): 50 Standard Deviation 2 (s_2): 12 Confidence Level: 95% **Result** $\mu_1 - \mu_2 = (M_1 - M_2) = 2$, 95% CI [-2.38, 6.38]. You can be 95% confident that the difference between your two population means ($\mu_1 - \mu_2$) lies between -2.38 and 6.38.	Sample 1 Mean (M_1): 70 Sample 1 Size (n_1): 50 Standard Deviation 1 (s_1): 10 Sample 2 Mean (M_2): 62 Sample 2 Size (n_2): 50 Standard Deviation 2 (s_2): 12 Confidence Level: 95% **Result** $\mu_1 - \mu_2 = (M_1 - M_2) = 8$, 95% CI [3.62, 12.38]. You can be 95% confident that the difference between your two population means ($\mu_1 - \mu_2$) lies between 3.62 and 12.38.
The 95% CI is containing 0 (no significant difference), the p-value is not significant at α=0.05	The 95% CI is not containing 0 (a significant difference), the p-value is significant at α=0.05

Those two cases of comparing two groups showed a significant difference and a non-significant difference with the corresponding CI.

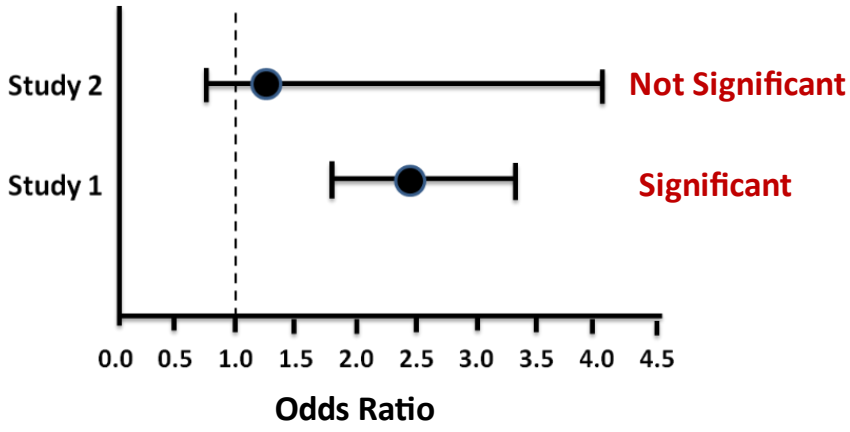
This website is used for producing the results:

https://www.socscistatistics.com/confidenceinterval/default4.aspx

5- Confidence Interval of ratios

In the case of the risk ratios (RR) and odds ratio (OR), we are interested in knowing if the confidence interval contains one. If it contains 1, there is no difference in the risk/odds between the two groups.

Both RR and OR are ratios and calculated by dividing the risk/odds of one group by the risk/odds of the other group. A value of 1 means that there is no difference between the groups.



If the confidence interval of the ratios doesn't contain one, the odds (or risks) are significantly different.

Example:

Case 1

Odds ratio calculator

Cases with positive (bad) outcome

Number in exposed group: 15 *a*

Number in control group: 8 *c*

Cases with negative (good) outcome

Number in exposed group: 20 *b*

Number in control group: 20 *d*

Test

Results

Odds ratio	1.8750
95 % CI:	0.6506 to 5.4039
z statistic	1.164
Significance level	P = 0.2444

The 95% CI is containing 1 (no significant difference),
the p-value is not significant at α=0.05

Case 2

Odds ratio calculator

Cases with positive (bad) outcome

Number in exposed group: 15 *a*

Number in control group: 8 *c*

Cases with negative (good) outcome

Number in exposed group: 20 *b*

Number in control group: 40 *d*

Test

Results

Odds ratio	3.7500
95 % CI:	1.3629 to 10.3181
z statistic	2.560
Significance level	P = 0.0105

The 95% CI is not containing 1 (a significant difference),
the p-value is significant at α=0.05

Those two cases of comparing two groups show a significant and a non-significant association based on OR with the corresponding CI.

This website is used for producing the results:

https://www.medcalc.org/calc/odds_ratio.php

Standard error

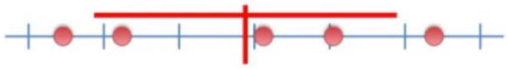
- Calculation of the confidence interval depends on the standard error.
- If we take enough samples from a population, the means will be arranged into a distribution around the true population mean.
- The standard deviation of this distribution, i.e. the standard deviation of sample means, is called the standard error.

The standard error tells us how accurate the mean of any sample is likely to be compared to the true population mean.

When the standard error increases, i.e. the means are more spread out, it becomes more likely that any given mean is an inaccurate representation of the true population mean.
Remember: Standard Deviation is the measure of how much the data is spread out from their mean/average to both sides of the number line.

Standard Deviation and Standard Error:

Standard deviation quantifies the variation within a sample



Standard error quantifies the variation in the means from multiple samples

For calculating the standard error for the mean we divide the standard deviation by the square root of the sample size:

$$SE = \frac{\sigma}{\sqrt{n}}$$

- Standard error increases when standard deviation, i.e. the variance of the population, increases.
- Standard error decreases when sample size increases – as the sample size gets closer to the true size of the population, the sample means cluster more and more around the true population mean.

The standard error can be calculated for:

- Sample mean.
- Sample proportion.
- Difference between means.
- Difference between proportions.

Confidence interval and standard error

The calculation of the confidence interval depends on the standard error.

For example: **95% Confidence interval (CI) for the mean** is

Sample mean – (1.96 X SE) to Sample mean + (1.96 X SE)

SE= standard error of the mean

So, the confidence interval depends on both the **variability** in the data and the **sample size**.

- A higher variability will result in a wider CI, and smaller sample size will result in a wider CI also.
- A wide interval indicates that the estimate is not precise; a narrow one indicates a precise estimate.

Summary points

- Interpretation of 95% CI : we are 95% confident that the true population mean lies between …. & ….
- As the sample size increases, the confidence interval is narrower.
- If we want a more precise confidence level (for example 99% instead of 95%), the CI will be wider.
- If the confidence interval for the difference between the two groups contains 0, the difference is not significant.
- If the confidence interval for the risk ratios (RR) and odds ratio (OR) contains 1, the difference is not significant.
- **It is always better to report the CI with the p-value rather than the p-value alone.**

Confidence interval calculation

The confidence interval can be calculated through:

- Using the equation.
- Using websites calculators, for example:

  Single-Sample Confidence Interval Calculator:
  https://www.socscistatistics.com/confidenceinterval/default3.aspx

  Independent Samples Confidence Interval Calculator

  https://www.socscistatistics.com/confidenceinterval/default4.aspx

- From the statistical software (SPSS for example)

  When doing any statistical test on SPSS, the CI is produced as part of the output, either as a default option or by choosing that option.

Part 2 Study design

Observational and interventional studies

Generally, medical studies are divided into observational studies and interventional (experimental) studies.

Non-experimental (observational) studies	Experimental studies
The researchers just **observe**, measure, or collect data without intervening with the study objects. - **Cross-sectional studies** - **Cohort studies** - **Case-control studies**	The researchers apply an **intervention** that may be a new drug, a surgical technique, an educational program, or any other intervention. - **Randomized controlled trials** (gold standard)

• **Quasi-experimental studies**

"Almost" experiments (but lacks randomization) as pre-post studies and interrupted time series studies.

Cross-sectional studies

- In a cross-sectional study, a sample is chosen and data on each individual (or case) is collected at one point in time.
- It is usually looked at as a snapshot.
- Examples: Surveys of prevalence (point prevalence studies) and surveys of individuals' beliefs or attitudes towards a particular issue. Prevalence of anxiety among medical students in Egypt, and knowledge attitude and practice of breast self-examination among Egyptian women.

Notes:

- We should not confuse the period of the study, which is the data collection period with the fact that we collect data only once from each individual (a snapshot). A survey may be conducted over weeks or months and each individual fills in the survey only once.

- We are collecting data at one point of time, so inferring temporal relationships should be dealt with carefully. We do not know when the events occurred prior to the study. So, we can only say that there is an association between the factor of interest and disease, but we cannot say that the factor is likely to have caused disease.
  For example, if we design a cross-sectional study the presence of an association between obesity and depression, we are not sure which problem started first and we cannot conclude that one of them caused the other.

- Cross-sectional studies can be used to estimate the **prevalence** of a disease in the population but not the incidence.

Example:

Research Question: Is the regular consumption of coffee associated with improved academic performance among Cairo University medical students?

Study Design: a cross-sectional study

A questionnaire is administered after receiving the final exam results.

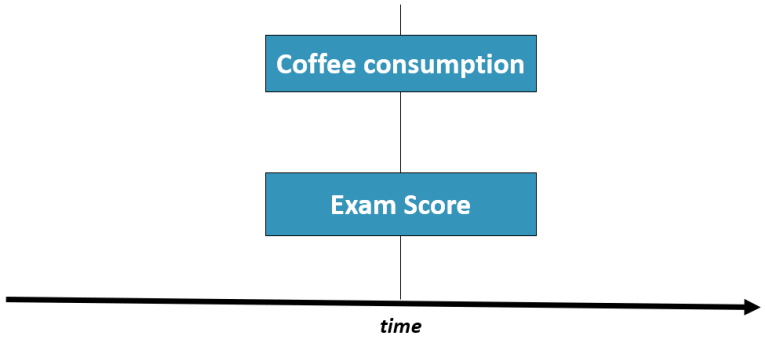
We collect data regarding:

Exposure: Self-reporting of >3 coffee drinks daily for the previous year (or the average number of daily coffee drinks).

Outcome: Score of final exams.

Covariates: Age, sex, residency, educational year, smoking status, socioeconomic level

This figure shows that we collect data about the exposure and outcome at the same time.



Importance of cross-sectional studies:

Descriptive value:

- What is the percentage of Cairo University medical students who drink coffee (prevalence)?
- What are the age and sex distributions of Cairo University medical students who drink coffee?

Analytic value:

- Is there an association between coffee consumption and exam scores among Cairo University medical students?

The analysis can be in the form of bivariate analysis (studying the association between the two variables), or multivariable analysis (studying this association while controlling for other variables "confounders").

Multivariable analysis can be used to study the association between multiple variables (possible risk factors) and a specific disease or outcome.

Advantages:

- Quick and inexpensive (No waiting for the occurrence of outcome)
- Easy and feasible
- No loss to follow up (there is no follow up)
- Used for determining prevalence (but not incidence)
- Associations can be studied
- Help in hypothesis generation (possible risk factors)

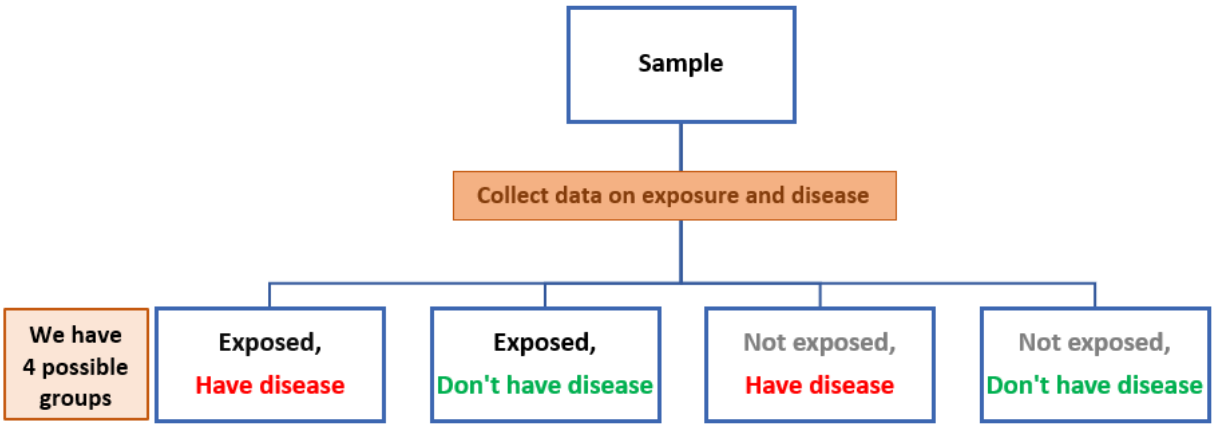
- Repeated cross-sectional studies on the same population at different points of time help in the evaluation of **trends** in the prevalence of the disease or risk factor.

Disadvantages

- Cannot determine causality (temporal relationship can't be determined)
- Not suitable for rare outcomes (diseases) or diseases of short duration.
- High refusal or non-response can cause bias.

Measures of association

When collecting data on the exposure and the outcome we have 4 possible groups as in the graph:



Two possible approaches to the analysis of results:

(A) Calculating the **prevalence of the disease** in exposed persons compared to the prevalence of the disease in nonexposed persons.
or

(B) Calculating the **prevalence of exposure** in persons with the disease compared to the prevalence of exposure in persons without the disease.

	Diseased	Not Diseased
Exposed	a	b
Not exposed	c	d

(A) Prevalence of the disease in exposed persons compared to the prevalence of the disease in nonexposed persons

	Diseased	Not Diseased
Exposed	a	b
Not exposed	c	d

$\frac{a}{a+b}$ Prevalence of the disease among exposed

$\frac{c}{c+d}$ Prevalence of the disease among the non-exposed

(B) Prevalence of exposure in persons with the disease compared to the prevalence of exposure in persons without the disease

	Diseased	Not Diseased
Exposed	a	b
Not exposed	c	d

$\frac{a}{a+c}$ Prevalence of exposure among the diseased

$\frac{b}{b+d}$ Prevalence of exposure among the non-diseased

Case-control studies

- They are "Observational" studies in which a group of patients (**cases**) is compared to a group of individuals who are free of this disease (**controls**) as regard exposure to a suspected agent or factor.
- We start with cases and controls, then we look for the past exposure.
- We work "backward" **(from outcome to exposure)**, data related to risk factors are collected after the disease has been identified.
- It determines the strength of the association between the risk factor (exposure) and the presence or absence of disease (outcome).
- It cannot yield estimates of incidence or prevalence of disease in the population (if we select 50 cases and 50 controls, we cannot assume that the prevalence of the disease is 50%).

Example: If we are interested in a rare outcome as: The association between regular coffee consumption and getting a score over 90% in the medical school.

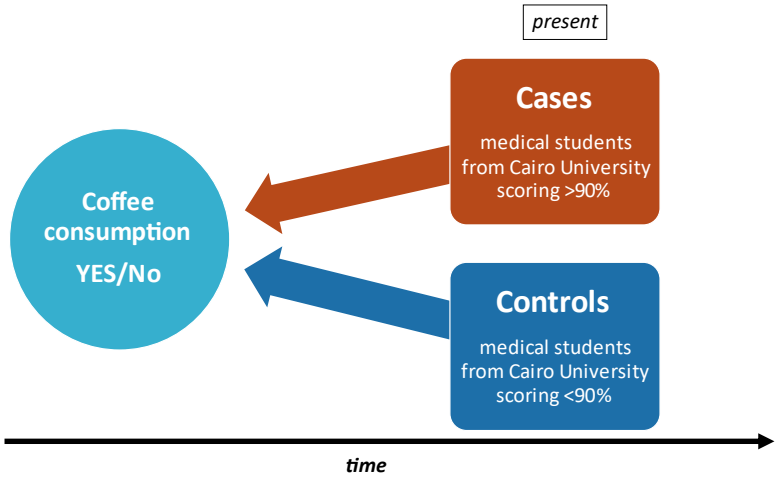
We use: **A Case-Control study**

Cases: medical students from Cairo University who scored over 90%.

Controls: medical students from Cairo University who scored less than 90%.

Exposure: Self-report of >3 coffee drinks daily for the previous year (or the average number of daily coffee drinks).

Additional covariates: Age, sex, residency, socioeconomic level, educational year, smoking status.



Important design considerations

- **Formulation of a clearly defined hypothesis**
- **Case definition** (diagnostic criteria for the cases)
- **Source of cases** (should be representative of all cases of the disease in the population)
- **Selection of cases** (incident or prevalent cases, incident cases are better as they are the new cases and are expected to better remember the exposure status)
- **Source and selection of controls** (controls should meet all the criteria for cases, except having the disease)

  Possible sources of controls:
  1. General population
  2. Hospital controls (patients in the same hospital who have diseases that are unrelated to the exposure being studied)
  3. Special controls (Friends, neighbors, peers, family members)
- **Measuring exposure status** (prone to recall and observer bias)

  Methods that can be used to ascertain exposure status include:
  - Standardized questionnaires
  - Biological samples
  - Interviews with the subject
  - Interviews with spouse or other family members
  - Medical records
  - Employment records
- **Ratio of controls to cases = 1:1 up to 4:1** to increase the statistical power (ratios greater than 4:1 have little additional impact on the power)

Matching

There are two types of matching controls and cases:

1- **Group matching or frequency matching** on a group basis
The average value of each of the relevant potential risk factors of the whole group of cases should be similar to that of the whole group of controls.

2- **Pairwise matching** on an individual basis
Each case is matched individually to one control (or more) who has similar characteristics, such as, socioeconomic status, or environment (factors that can be confounders).

Advantages

- Useful with rare diseases or diseases with a long latent period
- Inexpensive and efficient (may be the only feasible option)
- Establishes association (Odds ratios)
- Useful for generating hypotheses (multiple risk factors can be explored in one study)

Disadvantages

- Causality is still not easy to establish
- Selection bias: (if not choosing appropriate controls)
- Recall bias: (the study is retrospective and participants may not report their exposure accurately, especially the controls)
- Cannot give incidence or prevalence
- Accuracy and validity of information collected

Measures of association

		Disease: Yes	Disease: No
Exposure	Yes	a	b
	No	c	d

The odds ratio (OR) is used in case-control studies to estimate the strength of the association between exposure and outcome.

OR is the odds of having the disease among the exposed divided by the odds of having the disease among the non-exposed.

$$\text{Odds} = \frac{\text{Number who have the evet (disease)}}{\text{Number who do not have the evet (disease)}}$$

Odds ratio calculation:

$$\text{OR} = \frac{\text{Odds of having the disease among exposed } \frac{a}{b}}{\text{Odds of having the disease among non exposed } \frac{c}{d}}$$

$$\text{OR} = \frac{a/b}{c/d} = \frac{ad}{bc}$$

A confidence interval for the odds ratio should be reported. A confidence interval not containing 1 is considered statistically significant.

Cohort studies

Preliminary results from the cross-sectional and case-control studies suggest an association between the exposure (coffee consumption) and the outcome (improved academic performance) among medical students at Cairo University.

So, what is missing?

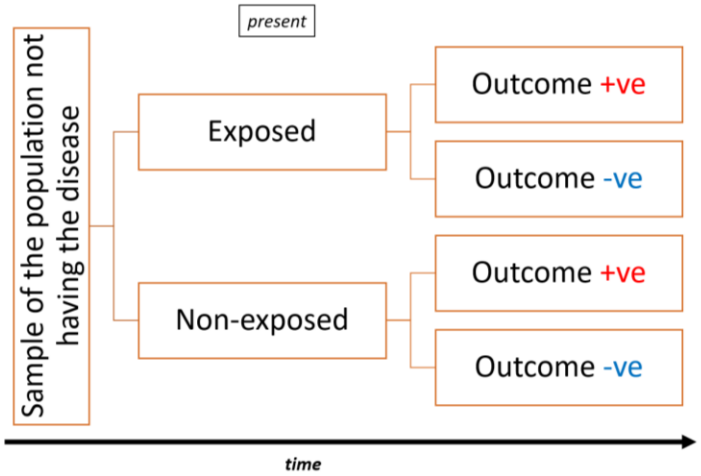
- Strengthening the evidence for a causal link between coffee consumption and academic performance.
- We use the results from our previous studies to propose for a cohort study.
- Cohort studies evaluate a possible association between exposure and outcome by **following two groups of individuals** (exposed and unexposed) over a period of time (often years) to see whether they develop the disease or outcome of interest.
- The rates of disease incidence among the exposed and unexposed groups are determined and compared.

⚐ Subjects should not have the outcome variable (should be disease-free) on entry and should have the potential to develop the outcome.

Cohort studies may be **prospective or retrospective**, but both types define the cohorts **on the basis of exposure**, not the outcome.

Elements of a cohort study

- Selection of sample from the population or self-allocated groups.
- Measuring the exposure variable in the sample.
- Ensuring that the outcome is not present (and participants can develop it).
- Follow up the population (the different exposure groups) for a period of time.
- Measure the occurrence of the outcome variable.



Advantages

- Knowing that a predictor variable (exposure) was present before outcome variable occurred (some evidence of causality).
- Valuable in studying rare exposures.
- Directly measures the incidence of a disease or an outcome.
- Can study multiple outcomes of a single exposure.
- Relative risk (RR) is the measure of association.

Disadvantages

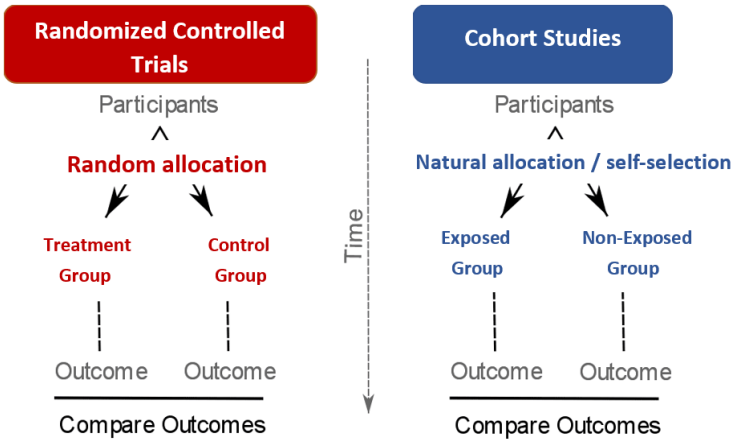
- Expensive and inefficient for studying rare outcomes.
- A large number of subjects is usually needed.
- Often needs a long follow-up period or a very large population.
- Loss to follow-up can affect the validity of findings.
- Retrospective cohort studies need complete and accurate records.

Experimental clinical trials (RCT) and observational (cohort) studies

There are two main differences between clinical trials and cohort studies:

1- In the cohort studies, no intervention, treatment, or exposure is given to participants, we just **observe** the exposure and the occurrence of the outcome.

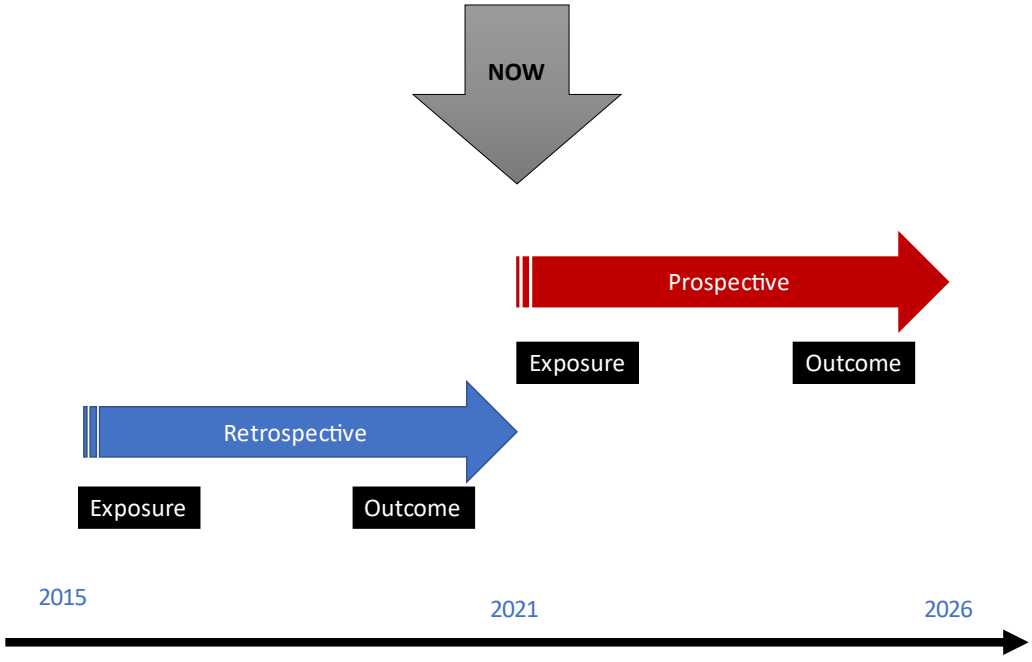
2- In RCTs, participants are randomly allocated to either receive the treatment/intervention or not, while in cohort studies, there is **no randomization**, we do not allocate participants to groups, there is either pre-existing exposure, or the participants self-select themselves (for example, some women might choose to be on hormonal contraception while others choose the mechanical contraception).



Prospective and retrospective cohort studies

- **Prospective cohort studies**: participants are identified and followed up over time until the outcome of interest has occurred, or the time limit for the study has been reached.

- **Retrospective cohort studies**: exposure and outcome have already occurred at the start of the study (current time). Pre-existing data, such as medical records or employee files, can be used to assess exposure status in the past.
  This type of cohort study is less time-consuming and less costly, but it is more susceptible to the effects of bias. Information on exposure and confounding variables may be unreliable, unavailable, inadequate, or difficult to collect.



- **Retrospective-Prospective Study:**

A cohort study may combine both retrospective and prospective data.

Measures of association

		Disease: Yes	Disease: No
Exposure	Yes	a	b
	No	c	d

The relative risk (RR) is used in cohort studies to estimate the strength of the association between exposure and outcome. It is also called Risk Ratio.

RR is the risk (incidence) of having the disease among the exposed divided by the risk of having the disease among the non-exposed.

$$\text{Risk (incidence)} = \frac{\text{Number who have the disease}}{\text{Total group}}$$

Relative risk (Risk ratio) calculation:

$$\text{RR} = \frac{\text{Incidence among exposed } \frac{a}{a+b}}{\text{Incidence among non exposed } \frac{c}{c+d}}$$

$$\text{RR} = \frac{a/(a+b)}{c/(c+d)}$$

A confidence interval for the relative risk should be reported. A confidence interval not containing 1 is considered statistically significant.

Interventional RCTs

- **A randomized controlled trial (RCT)** is an **interventional** study in which subjects are **randomly allocated** to different treatment options. Randomized controlled trials (RCTs) are the accepted **'gold standard'** of individual research studies.
- The comparison is done against an active agent or with an inert substance (placebo).
- It is a common practice that the comparison group receive the usual care.

Example:

Randomized controlled trial of the effect of daily coffee consumption on exam scores among Cairo University medical students.

A **sample** of the population (Cairo University medical students) is selected.

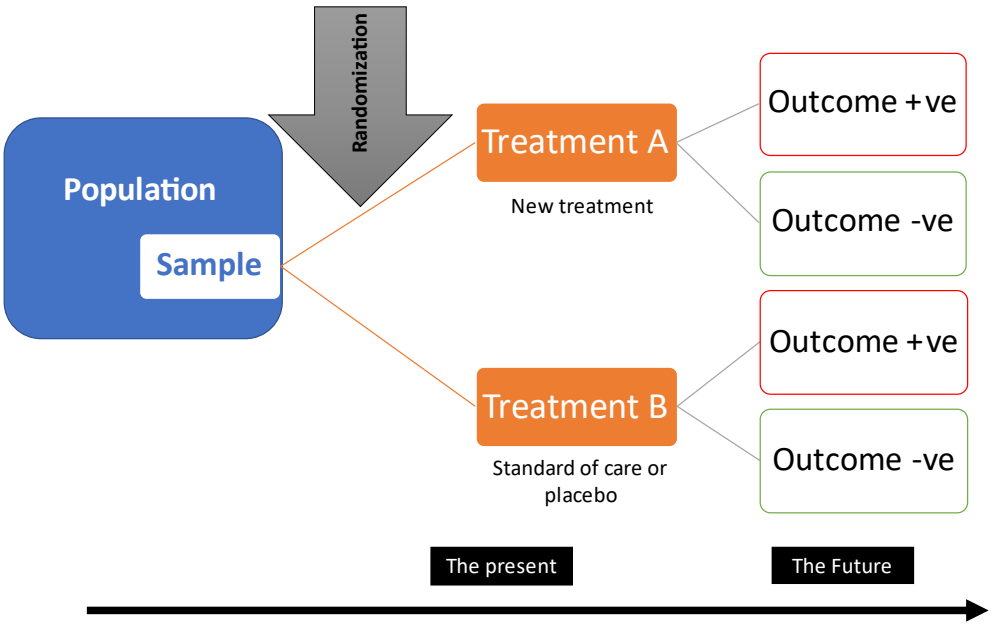
Participants are **randomized** to daily consumption of coffee (intervention group) vs. daily consumption of placebo (control group)

Outcome: Final exam scores

Example:

Randomized controlled trial of the effect of daily coffee consumption on exam scores among Cairo University medical students.

Design



Clinical trials have to be:

- Randomized (participants are randomly allocated to each group).
- Controlled (have a control arm)
- With adequate power to detect clinically meaningful differences between the study groups.

Steps

1. **Selecting participants:**
    - High-risk for the outcome (high incidence).
    - Likely to benefit and not be harmed.
    - Likely to adhere.
2. **Measuring baseline variables**
3. **Randomization**
    - Eliminates baseline confounding.
    - Types (simple, stratified, block).
4. **Blinding the intervention**
    - As important as randomization.
    - Eliminates bias.
5. **Follow subjects**
    - Adherence to protocol.
    - Lost to follow up.
6. **Measure outcome**
    - Clinically important measures.
    - Adverse events.

Participants

What is our target population?

Inclusion and exclusion criteria for participants should be considered carefully. Participants might be:

- Participants with certain diseases.
- Participants with certain stages of a disease.
- Healthy subjects.

Controls

Controls are either given a placebo or an active material.

Placebo-Controlled: control group is given:

- No intervention
- Inert material manufactured to resemble the new treatment in shape, texture, …etc.
- Sham surgery (fake surgery that is an imitation of the surgical process in the intervention group).

In clinical trials, a response is observed in the placebo group and is called the placebo response. The placebo effect is the difference between that response and no treatment. We compare the active group to the placebo group to account for this response.

Active-Controlled: control group is given an active drug (usually the Standard treatment)

The intervention

The intervention given to the participants might be:

- Drug
- Intervention
- Surgical technique
- Device
- Diet

Examples:

- **Drug Trials**: most common example of randomized clinical trials. It provides the strongest evidence for concluding that the results obtained were due to the therapy given.
- **Prevention Trials:** vaccination trials, prostate cancer prevention trial (assessed the effect of a drug vs. placebo on preventing the occurrence of prostate cancer)
- **Screening Trials**: try to find new ways of detecting the disease earlier and see if such earlier detection will lead to better health outcomes.

Advantages of clinical trials

- Stronger evidence over observational studies.
- Ability to demonstrate causality.
- Randomization controls for unmeasured confounding variables.
- Comparison between the new drug and the current treatment (or a placebo).
- The gold standard for generating scientific evidence for new drugs or interventions.

Single site Vs Multicenter studies

Single-site study:

Single center studies are usually set up in a specific hospital or clinic.
– Done if:

- The center can provide a representative sample of the population.
- The center has adequate resources.

– Advantages:

- Consistency in assessment (better quality and reliability of data collected.
- Easier to conduct and usually less expensive.

– Disadvantage

- Might take a longer time to complete the study (slow recruitment).
- Might fail to achieve the needed sample size.
- Not suitable for rare diseases.

Multicenter study:

- Involves more than one study center (hospital, clinic, country, etc).
- Identical protocols should be followed in all centers.
- Data intended to be analyzed as a whole.
- Provides replication and generalizability.
- Shorter time to complete the study (in case of rare disease, and developing new technologies)
- A central laboratory for assessments of laboratory tests is preferred.
- More complicated in terms of co-ordination, quality control, data management
- Problematic if there is treatment-by-center-interaction.

Treatment-by-center interaction:

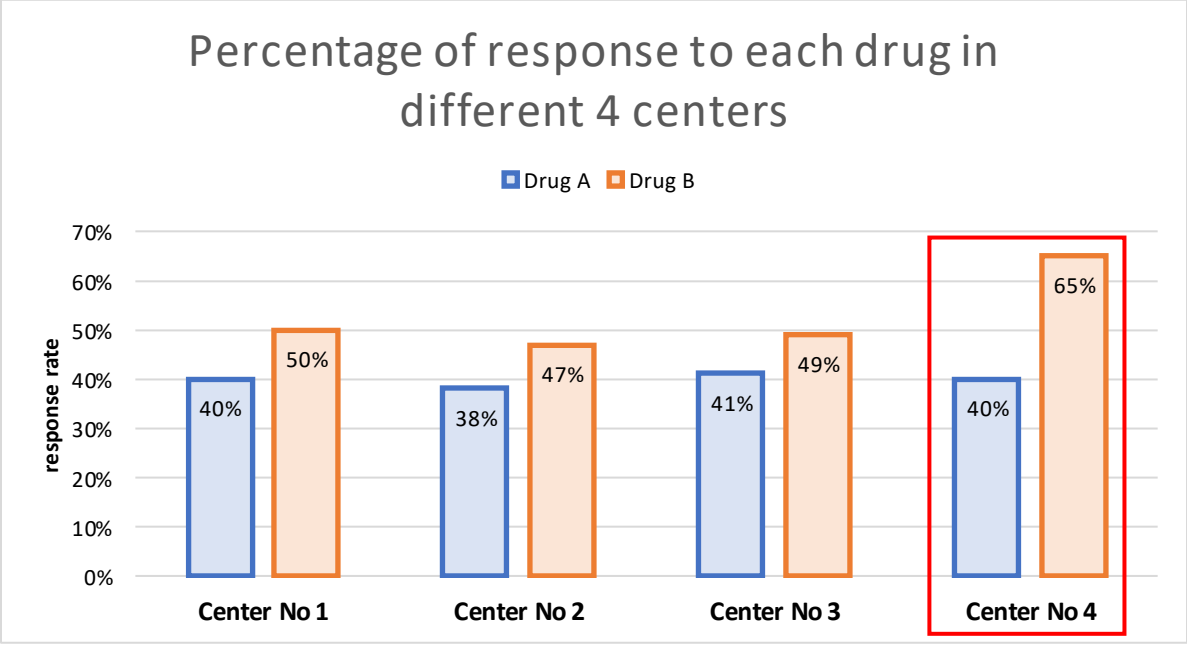
The results of the comparison between the two or more study arms differ per different centers.

The difference may be in the same direction or different directions:

• **Same direction** of the differences but different magnitudes (not very harmful)

The following figure shows the response to each drug in different 4 centers.

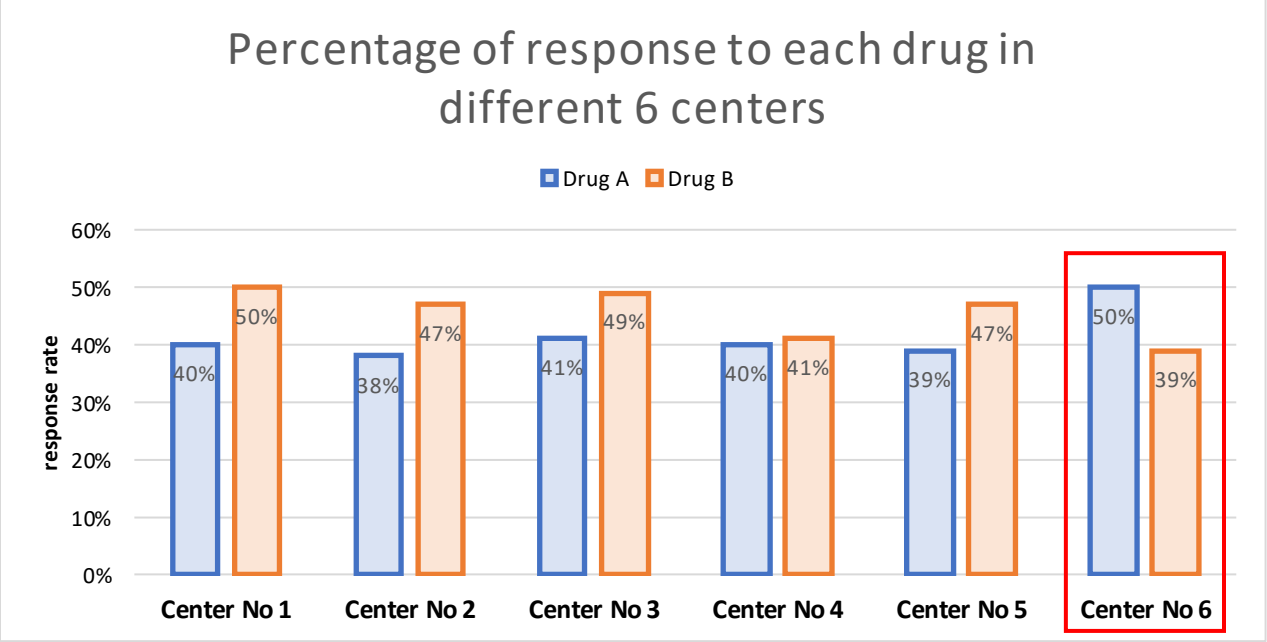
- In centers No 1, 2,3, we can see almost a similar response.
- In center No 4, the difference between the response of drug A and drug B is much higher than in the other three centers.



• **Different directions** of the differences (questioning the reproducibility of the study)

The following figure shows the response to each drug in different 6 centers.

- In centers No 1, 2,3,5, we can see almost a similar response.
- In center No 4, there is almost no difference between the response of drug A and drug B.
- In center No 6, the response of drug A and drug B is reversed.

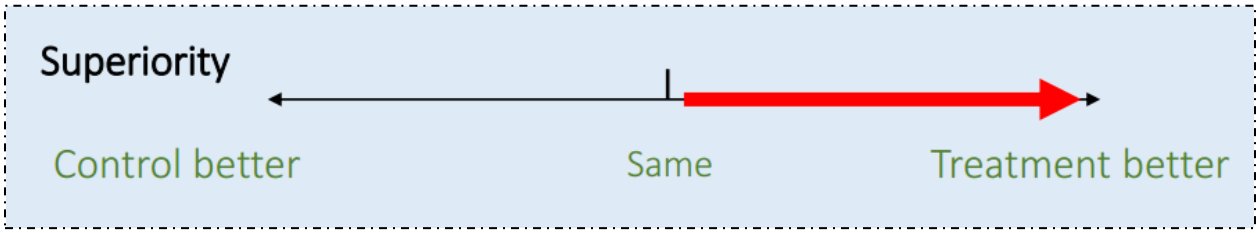


• this situation needs to be investigated to find the reasons such as different types of patients (high risk vs. low risk in different centers, health professionals not following the protocol correctly …etc).

Superiority, Noninferiority/Equivalence Trials:

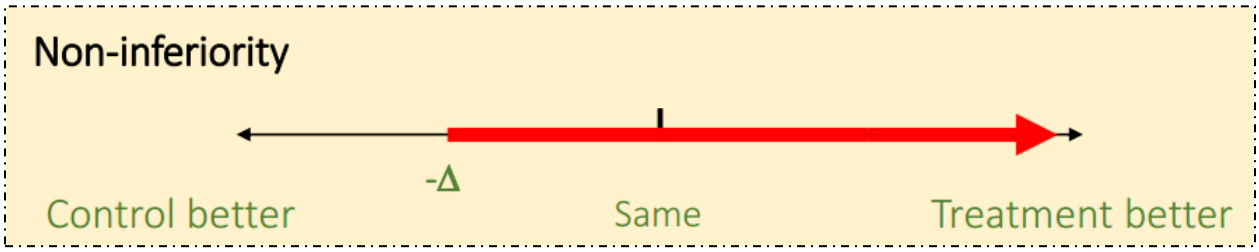
Superiority Trial: the aim is to prove that the new drug is **better** (superior, more efficacious) than a placebo or current treatment.

It should have the power to detect a clinically meaningful difference between the two treatments.



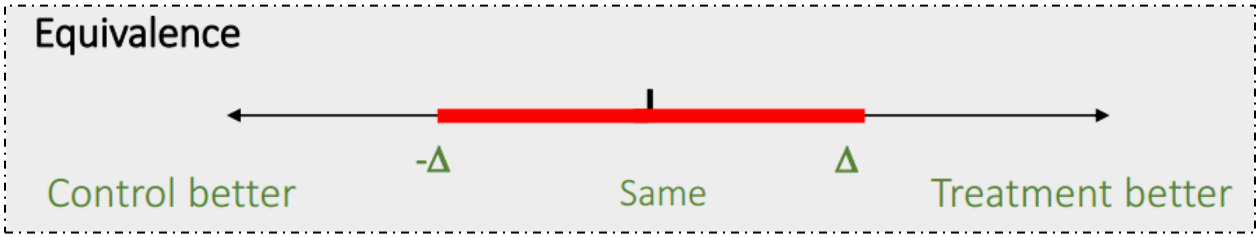
• **Non-inferiority Trial:** The aim is to prove that the **new drug is no worse** (but can be better) than the current treatment (control).

The treatment is not worse than the control by more than a small pre-specified amount referred to as Δ and is called the non-inferiority margin.



• **Equivalence Trial:** the aim is to determine whether one intervention is therapeutically "similar" (i.e. neither better nor worse) to another.

The hypothesis intends to test whether the treatment effect lies within a pre-defined set of limits (called equivalence margin).



In non-inferiority or equivalence trials, the new drug or intervention might be characterized by:

- Less toxicity
- Easier administration route
- Less frequently taken
- Better quality of life
- Shorter duration of action
- Cheaper
- Minimal invasive

⚐ Obtaining a non-significant result from a standard superiority trial does not mean we can infer equivalence. It means that there is not enough evidence to show that treatment is better than control.

Blinding (Masking)

It refers to withholding information about the treatment allocation from patients, treating physicians/investigators, sponsors and/or biostatisticians.

The goal is to **reduce bias** due to the subjectivity in: reporting (by patient), evaluating (by physician), and analyzing the data that might occur when the statistician knows who is getting what.

• **It is not always feasible.** For ethical reasons or due to the nature of the intervention (it is not possible to blind for an educational program for example).
• Can be breached due to severe side effect or because of the taste of the medicine.
• In very extreme situations such as the occurrence of extreme side effects, unblinding might be necessary.

Types of blinding

- **Open label (no blinding):** the patient and physicians know which treatment/intervention the patient is receiving.
- **Single blinded:** the patient does not know what drug he/she is taking.
- **Double-blinded**: both the patient and the investigator do not know which patient is receiving which treatment.
- **Triple blinded**: patient, investigator, and data analysts (or the sponsor) do not know the treatment allocations to the patients.

How to Blind?

- **Pharmacists/companies can prepare the drugs** (which should be identical in size, texture, packaging, taste,..etc) and label them as A or B.
- The physician/investigator will receive a **randomization sequence** such as AABABBA…. etc.
  So, the first patient receives drug A, the second patient receives drug A, the third receives B, …etc.
- The biostatistician should only get codes A and B for the drugs and will not know what A and B are (in case of triple blinding).
- At the end of the study, the pharmacist or the principal investigator who is not involved in the recruitment of patients declares what is drug A and what is drug B.

Prospective Randomized Open Label Blinded Endpoint (PROBE)

- It is open-label trials where both the patient and the physician know the treatment allocations but blinded to the endpoint of interest.
- The endpoint(s) are assessed by an independent committee that is blinded to the treatment allocation.

Randomization

- It refers to the **random** assignment of patients to the different treatments considered in the study.
- It removes **bias due to subjectivity** of assignment of patients.
- It produces **balanced groups,** i.e., the **measured and unknown** prognostic factors and other characteristics of the participants at the time of randomization will be, on average, evenly balanced between the intervention and control group.
- Successful randomization ensures the internal validity of the study (that is the comparison between the two groups is not biased).
- Randomization should be done by an **independent body** that is not with the recruitment of the patients.
- That independent body creates the **randomization sequence.**

Once an eligible patient consents to be part of the study, then the investigator will contact that independent body gives him/her/it the id number and other information such as age, then the independent body will record it and give back the randomization allocation.

The person who generates the allocation sequence should not be the person who determines the eligibility and entry of patients.
This last process can be done through computers, sealed envelopes, private companies, phone calls…etc

Randomization outcome is the randomization sequence:

Patient ID	Drug
1	A
2	B
3	B
4	A
5	A
6	B
7	A
8	A
9	B
10	A
11	B
…	
n	B

Some Randomization Methods:

1- Simple Randomization

- For each patient who is eligible and consents to the study, it is like **flipping a coin.** (for example, tails = treatment, heads = control)
- Each patient has a 50/50 chance of receiving the test drug or the placebo.
- Randomization is performed **independently** for each patient.
- **Strength:** future allocations cannot be predicted from previous ones, good for large trials.
- **Weakness:** This can result in an unequal number of participants in each group or an imbalance in key patient characteristics especially for studies with small sample sizes, it is highly likely to have an imbalance in size between the two groups (40%/60% or worse).

2- Random allocation or permutation

- To overcome the problem of imbalance of number between groups, this technique is used since the sample size of the study is usually computed before the start.
- Randomization sequences could be generated ahead of time.

Suppose that we want to recruit 40 patients and randomly allocate 20 for treatment A and 20 for treatment B.

We can do the following:

- Prepare identical sheets numbered 1 to 40 mixed in a bowl.
- Randomly pick 20 of those sheets.
- The picked numbers will receive treatment A and the rest will receive treatment B (or vice versa).
- This guarantees equal sample sizes in the two treatment groups.

This can be done using some statistical software or online programs as:

https://www.graphpad.com/quickcalcs/randomize1/

3- Block Randomization

Even with complete randomization or random allocation methods, we might see sequences or chains with the same allocation such as ABBBBAABAA.

Those can create problems of:

- Covariate imbalance: If there is a change in demographic or any other factor of patients over time (suppose the first 5 patients were young and the last 5 patients were older).
- If the trial had to be stopped for one reason or another before the full recruitment (suppose we stop after the 6th recruited patient, there will be an imbalance)

As a solution: we can divide patients into **blocks with equal or usually unequal sizes.** Then, independently randomize patients within each block.

For example: if we have a block of 4 , the 6 possible allocations to treatment A or B are:

AABB ABAB ABBA BBAA BABA BAAB

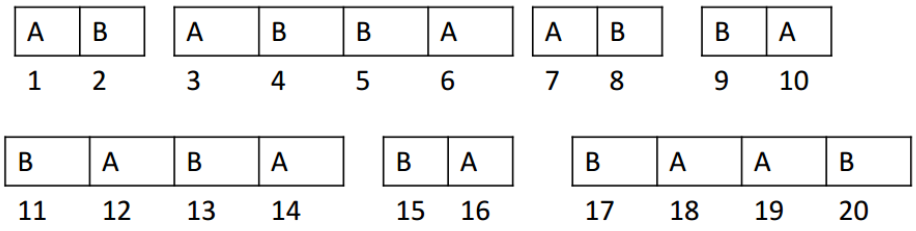
- **Advantage:**
  Insures treatment and covariate balance during randomization.
- **Disadvantage:**
  – When the block size is **not blinded and small**, there is a possibility of bias by investigators who can correctly guess the treatments (never disclose the block size).
  – If the block size is **too large** then there is a possibility of a large imbalance (due to incomplete blocks) if the trial is stopped early.

How can we perform permuted block randomization?

• Suppose we want to randomize 20 patients (10 for treatment A and 10 for treatment B).
• We can imagine that we have for example 7 blocks of sizes 2,4,2,2,4,2,4.

For blocks of size 2 we can flip a coin and for blocks of size 4 we can role a die as we have 6 possible allocations.

Results of the block randomization



This can be done using the same website, or the following website:

http://www.jerrydallal.com/random/random_block_size_r.htm

Can be also reached through http://www.randomization.com

The randomization sequence produced using this website is reproducible as there is a seed number with each generated sequence.

For 20 subjects randomized into blocks of: 2 4 2 2 2 4 4

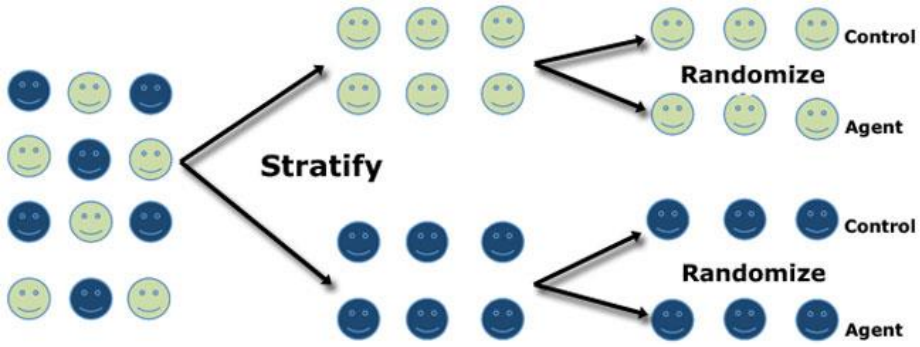
We may reach the following sequence:

1. B
2. A
3. A
4. B
5. A
6. B
7. A
8. B
9. A
10. B
11. B
12. A
13. A
14. B
15. B
16. A
17. B
18. A
19. A
20. B

4- Stratified Randomization

- Although balance in sample size can be achieved with block randomization, the groups may not be comparable in terms of covariates (e.g. one group may have older, more severe disease, more likely to smoke, more participants with secondary diseases, which confound the outcome)
  Covariates (such as age, gender, race, underlying disease severity…etc) that are directly related to the major outcome variable of interest can affect the reliability and accuracy of the results of a clinical trial.
- Stratified randomization **reduces the risk of imbalance** due to pre-specified prognostic patient characteristics
  Patients are grouped according to covariate values to form strata.
- Separate block randomization is carried out for each stratum (e.g. if gender is a potential confounder, a separate randomization list would be produced for males and a list for females).

Stratify then randomize!



- Clinics (hospitals) should be used for stratification in multicenter trials.
- This will control for differences in the study population due to environmental, social, demographic, and other factors related to the clinic.
- It is not recommended to use more than 2 covariates for stratification.

Randomization in the analysis of a clinical trial

- The first table in any paper describing a randomized clinical trial intends to prove that randomization was successful by **comparing all demographic and possible confounders** between the two study groups hoping that they are similar.
  So that the authors can claim that the difference in the outcome between the two groups (arms) of the study is due to the actual treatments assigned to those groups.

- If they could not prove that randomization worked, then they need to adjust for imbalances in such variables in the analysis phase (multivariate analysis)

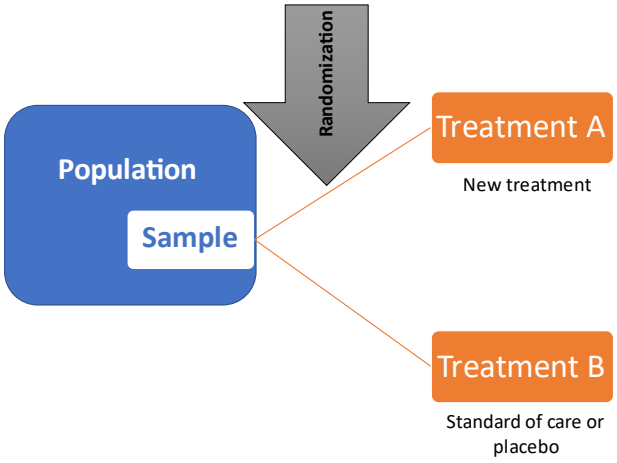
Common Designs of randomized clinical trials

Parallel design

A **parallel design** is where patients are randomized to one of the two groups of treatments, A and B, and each patient receives only one type of treatment.

Participants are randomized into different study arms (sometimes more than two arms), but there is approximately an equal number in each group.

The parallel design is the **most common** design for clinical trials.

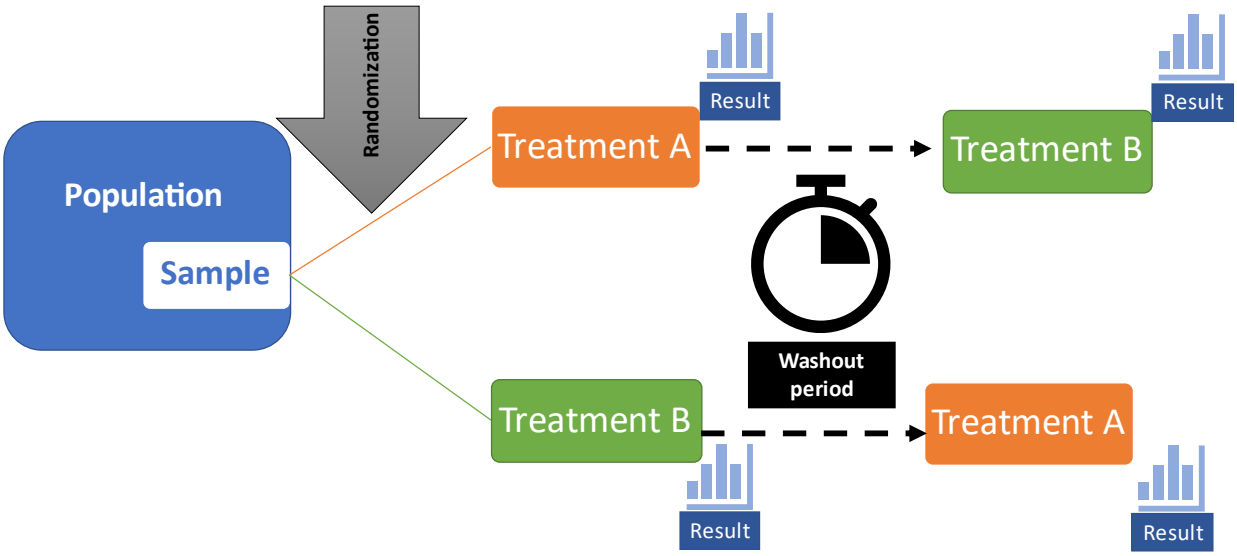


Cross-over design

- Each participant is exposed to both the control and the intervention, in a sequence.

For example: Patient X and Y are randomized into two different treatment groups. Patient X receives Treatment A during the first period of the study; Patient Y receives Treatment B. After the first period is over, there is a washout period. Patient X then receives Treatment B for the second period of the study while Patient Y receives Treatment A.

- Patient condition must be chronic and stable.
- Each individual serves as his/her own control.
- Main disadvantage: Carryover effects which may affect the direct intervention effect (if the effect of the drug in the first period affects that in the second period).
- It also allows fewer patients, but the study must run for a longer period of time, and patients may drop out after the 1st treatment period.
- Analysis is not straightforward.



Factorial design

- Factorial clinical trials test the effect of **more than one treatment**.
- It allows the assessment of **potential interactions** among the treatments.
- Smaller sample size is needed than individual trials.
- The simplest design is 2X2 factorial design:

If treatments A and B are to be compared to control, 4 treatment groups would be formed:

A only **B only** **A + B** **neither**

Treatment B \ Treatment A	Yes	No
Yes	**AB** Both	**B** only
No	**A** only	**-** Neither

It allows for answering the following questions:

- Effect of treatment A
- Effect of treatment B

- Interaction effect between treatment A and B (Whether the effect of treatment A is influenced by the presence or absence of treatment B)

Cluster randomized trials

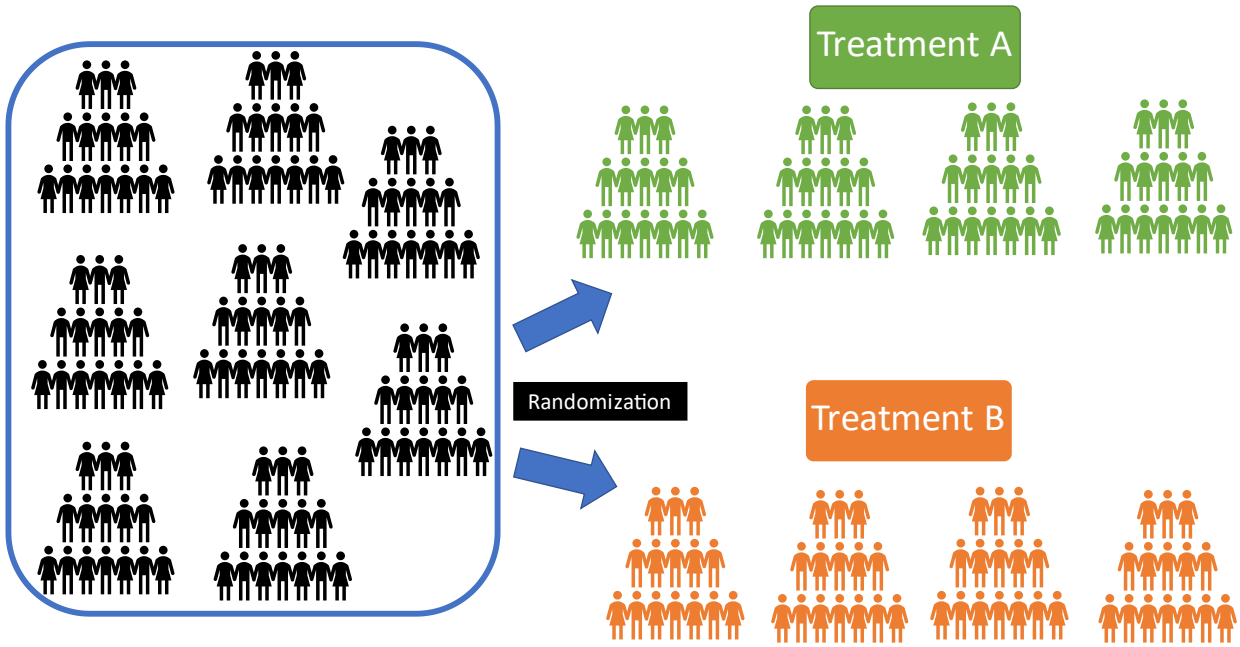
Cluster RCTs are trials in which groups/clusters (of individuals), rather than individuals, are randomly allocated to intervention groups.

• Unit of randomization (clusters) can be families, geographical areas, maternity units, hospitals, communities, schools….

• Randomization is at the group (cluster) level, but the analysis is at the individual level.

• It is used to avoid treatment contamination, for administrative convenience, and sometimes, there is no alternative as in the case of interventions affecting the whole area (as in water fluoridation).

• Sample size tends to be larger, and statistical analysis is more complex as it has to account for the clustering effect.



Statistical analysis of Clinical Trials

A statistical analysis plan should be provided in the proposal or the protocol before starting a clinical trial.

The statistical analysis plan usually includes:

1- Comparison of baseline characteristics:

- We provide summary measures of the baseline variables.
- This baseline table should be provided showing summary statistics by group and overall participants.
- The aim is to ensure that we have balance in these variables between our randomized groups and that randomization was successful.
- Clinical comparison and not statistical is now recommended (no need to test for the differences between groups at baseline, no p-value is provided)
  – To avoid false positive results.
  – The Study is not powered for comparing those variables.
  – To adjust for any imbalances likely to have a possible impact on the main/primary analysis.

2- Primary analysis

Comparing the primary endpoints between the two (or more) groups (not adjusted for anything).

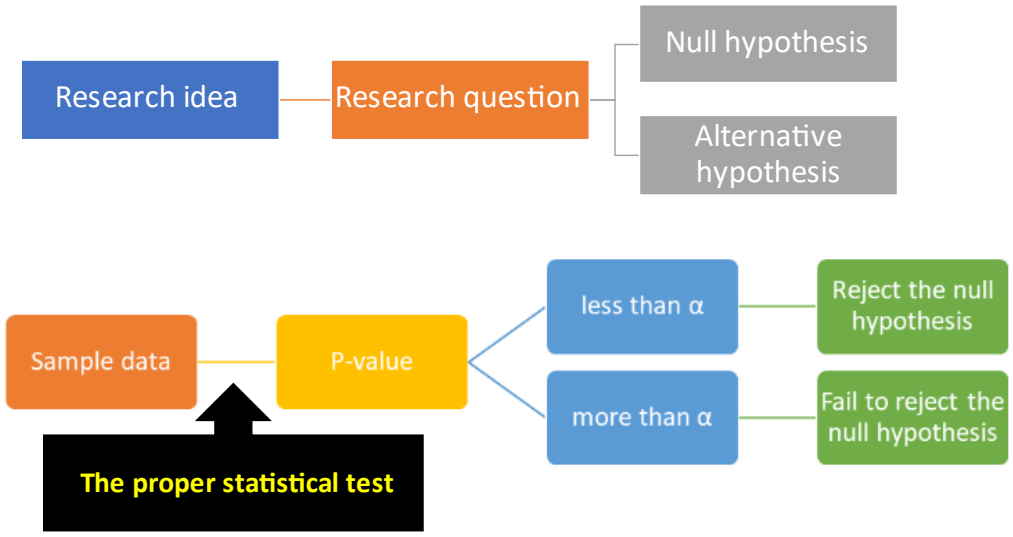
3- Secondary analysis

- Comparing secondary outcomes (not adjusting for anything).
- Comparing primary variables adjusted (multivariate) for imbalances in baseline characteristics.
- Comparing baseline characteristics between those lost to follow-up in the groups.
- Secondary analysis for exploratory purposes such as subgroup analysis.
- Analysis of adverse events.

Part 3 Choosing the Suitable Statistical Test

Steps of statistical test selection

After finishing the descriptive statistics, we move to the analytical part. Basic skills in analytical statistics include choosing the proper statistical test that is suitable for the data and helps to answer the research question and reaching a conclusion regarding the null hypothesis (reject or fail to reject the null hypothesis).



There are multiple flowcharts to guide us to a suitable test. They are of variable degree of complexity and the number of tests included.

Here, a simplified guide is proposed based on 5 steps to guide us to the most common statistical tests.

Choosing the proper statistical test, the five steps (questions):

Q1: Bivariate Vs Multivariable

The first question we need to ask is whether we are dealing with bivariate analysis or multivariable analysis.

Bivariate analysis: studying the relationship between two variables.

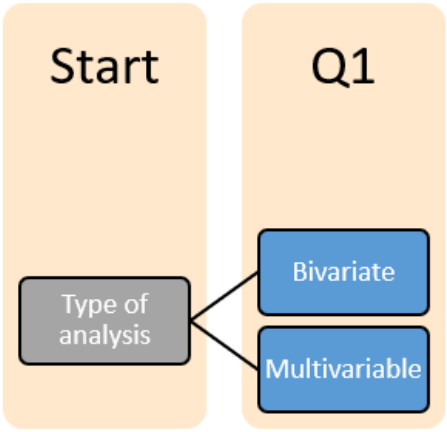
For example:

- Age and height
- Type of treatment and complication
- Sex and smoking
- Smoking and coffee consumption

Multivariable (regression modelling/analysis): studying the effect of multiple variables on an outcome variable.

For example:

- Effect of smoking, sex, coffee consumption on blood pressure.
- Effect of smoking, sex, coffee consumption on having a heart attack.



🏳 Regression can be used for bivariate analysis if it is used to study the effect of one variable only on the outcome.

Q2: Difference Vs Correlation

If we are doing bivariate analysis, we have to ask if we are studying a difference or a correlation.

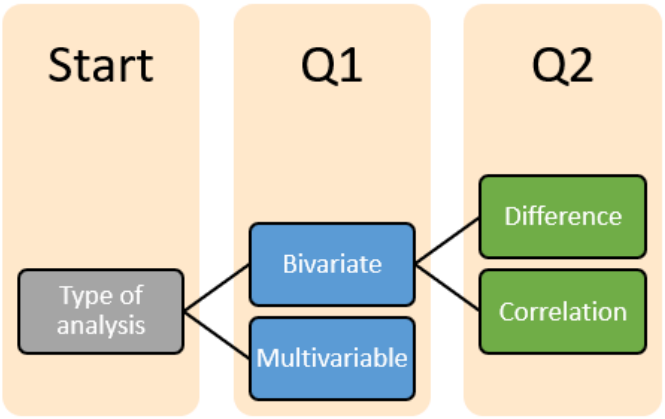
Difference: to study the difference between two or more groups, or two or more conditions

For example:

- The difference between males and females regarding coffee consumption
- The difference in body weight before and after being on a specific diet.

Correlation: to study the association between two variables

- The association between age and weight
- The association between coffee consumption and the number of sleeping hours.



Q3: Independent Vs Paired data

If we are doing bivariate analysis, we have to ask if we are working with independent data or paired data

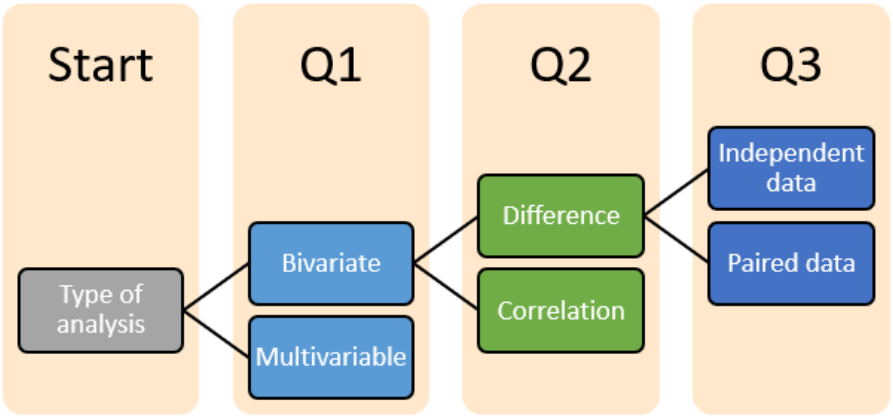
Independent (unpaired) The observations in each sample are not related

There is no relationship between the subjects in each sample.

- Subjects in the first group cannot also be in the second group
- No subject in either group can influence subjects in the other group
- No group can influence the other group

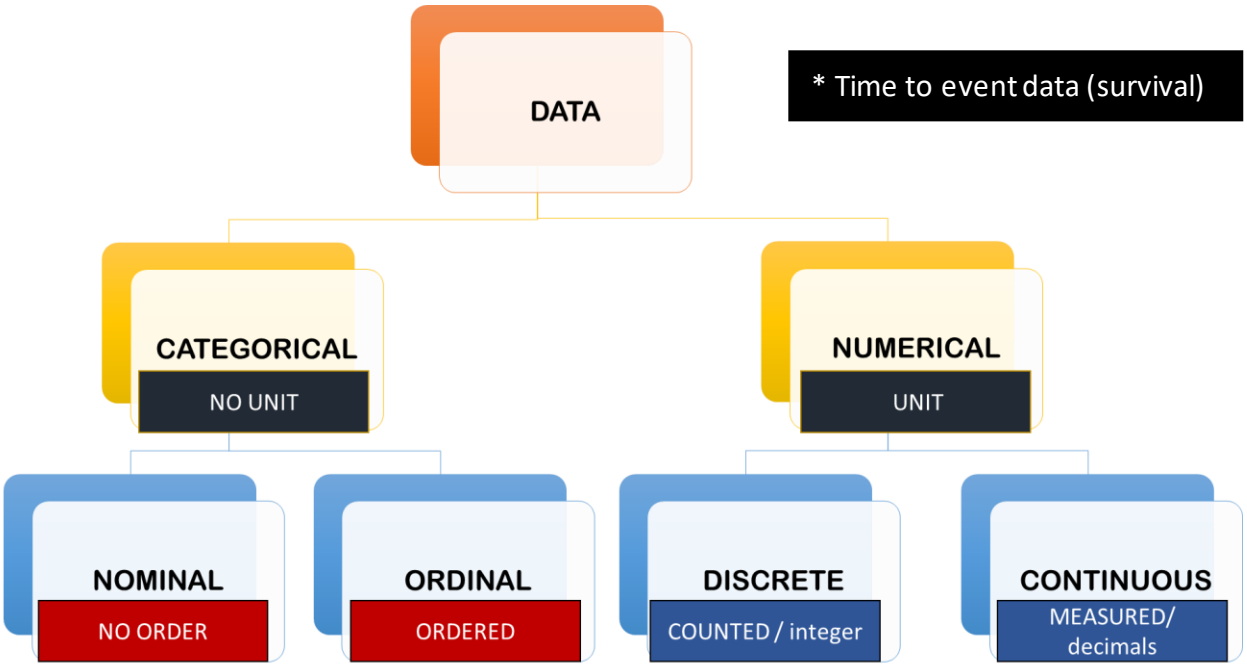
Dependent (Paired): paired samples include:

- Pre-test/post-test samples (a variable is measured before and after an intervention)
- Cross-over trials
- Matched samples
- When a variable is measured twice or more on the same individual



Q4: Type of outcome and normality of distribution

Whatever the analysis we are doing, it is important to identify the types of data variables we are studying. The type of data variables is very important in choosing the suitable test. The following chart helps to distinguish between different types of data variables.

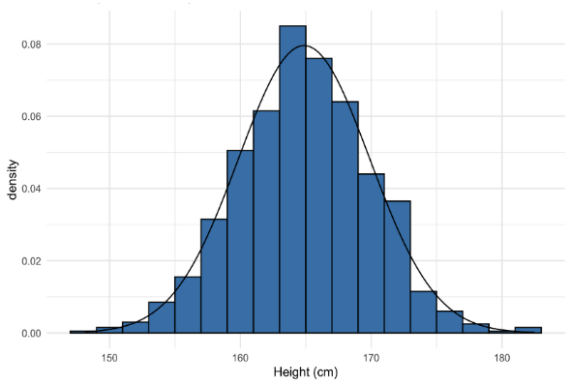


* **Time to event data (survival data):** This is a special data type that will be discussed in survival analysis.

Normality of distribution

It is important before doing some statistical tests to determine if a numeric variable is normally distributed or not.

This histogram shows normally distributed variable.



Q5: Number of groups /conditions

It is important to ask if we are comparing two groups (conditions) or more than two groups (conditions).

For example:

Are we comparing two groups (diseased, not diseased), or three groups (normal, osteopenia, osteoporosis)?

Are we comparing two conditions (pre-test, post-test), or three conditions (before the operation, during the operation, after the operation)?

Choosing the most common statistical tests guide

All the questions are arranged together in the following guide to direct us to the suitable test.

A guide for choosing the most common statistical tests

Q1	Q2	Q3	Q4	Q5	Statistical test
Bivariate / Multivariable	Difference / Correlation	Independent / Paired	Type of outcome (and Normality)	No of groups (conditions)	
Bivariate	Difference	Independent (un-paired)	Continuous (Normal)	2	Student's t-test
				>2	One-way ANOVA
			Continuous (Non-normal)/ Ordinal	2	Mann-Whitney U test
				>2	Kruskal-Wallis H test
			Nominal	2	Chi-square test/ Fisher's exact test
				>2	Chi-square test
			Time to Event (survival)		Log-Rank test (Kaplan-Meier plot)
		Dependent (paired)	Continuous (Normal)	2	Paired t-test
				>2	Repeated measured ANOVA
			Continuous (Non-normal)/ Ordinal	2	Wilcoxon signed-rank test
				>2	Friedman test
			Nominal	2	McNemar's test
	Correlation		Continuous (Normal)		Pearson's correlation (r)
			Continuous (Non-normal)/ Ordinal		Spearman's correlation (ρ)
			Nominal (2 levels)	2	Spearman/Kappa (Agreement)
Multivariable			Continuous		Linear Regression
			Ordinal		Ordered Logistic Regression
			Nominal	(2 levels)	Binary Logistic Regression
			Nominal	(>2levels)	Multinomial Logistic Regression
			Time to Event (survival)		Cox Regression
			Count variable		Poisson regression

We start from left to right, a step by step till reaching the suitable statistical test in the last column.

Normality and homogeneity of variance assumptions

Assumption of normality

For some tests to be done, data needs to be approximately normally distributed.

- **How to test for normality?**

1- Plotting a **histogram or QQ plot**
2- Using a statistical test

The statistical tests for normality are the **Shapiro-Wilk** and **Kolmogorov-Smirnoff** tests
We usually do both, the graph and the statistical tests.

- **The hypotheses of the Shapiro-Wilk and Kolmogorov-Smirnoff tests**

H_o: the variable is normally distributed

H_1: the variable is not normally distributed

We accept the null hypothesis (say that it is normally distributed) if the P-value > 0.05.

For large samples (more than 30), we don't worry much about the normality assumption (based on the central limit theorem)

If the normality assumption is not met, then we can't use the parametric tests, and we have to uses the non-parametric tests.

Assumption of Homogeneity of variances

Homogeneity of variances (similar standard deviations) means that the variable we are studying has the same variance across groups.

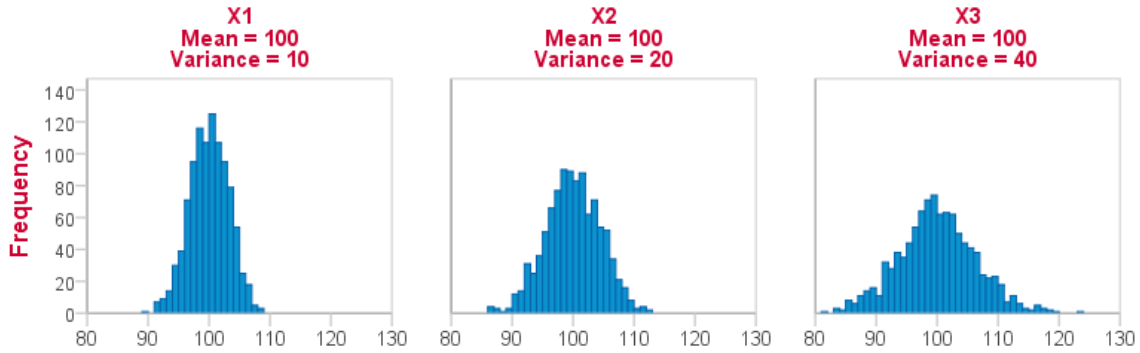
We need to test for the equality of variances between groups when using some statistical tests, e.g. **Independent t-tests** and **one-way ANOVA**.

Homogeneity of variances is tested using **Levene's test.**

Interpretation of the test result:
If the p-value is < 0.05 reject H0 and conclude that the assumption of equal variances has not been met.

We accept the null hypothesis (say that there is equal variance) if the P-value > 0.05.

The following graph shows the distribution of three groups that have equal means but not equal variances.



If the homogeneity of variance assumption was not met, the standard tests cannot be done, and modified tests can be used (will be discussed with the relevant tests).

For more details regarding learning how to use SPSS intended for beginners, you can join the Udemy course prepared by Dr. Mohamed Elsherif:

Available through the following link:

https://www.udemy.com/course/spss-26-for-beginners-arabic/?couponCode=ELSHERIF

Part 4 Numerical data analysis

Numerical data analysis

In this part, we will discuss some statistical tests that are used for the analysis of numeric variables.

Statistical tests discussed in this part are:

Independent Sample t test, Man-Whitney U test, Paired samples t test, Wilcoxon Signed Rank test, one-way ANOVA, Kruskal Wallis test

Parametric and non-parametric tests:

Statistical tests are either parametric or non-parametric tests:

- Parametric tests are used to compare means of the groups while non-parametric tests are used to compare the medians.
- Parametric tests are used to compare samples with normally distributed numeric data.
- Non-parametric tests are used to compare samples with non-normally distributed numeric data, or with ordinal data.
- Parametric tests use the actual values of the variable.
- Non-parametric tests use the ranks of the values.

Parametric tests	Non-parametric tests
Independent Sample t test	Man-Whitney U Test
Paired samples t test	Wilcoxon Signed Rank Test
One way ANOVA	Kruskal Wallis test

The following table shows how the raw data (used in the parametric tests) are transformed to ranked data (used for the non-parametric tests)

Raw data	Sorted data	Ranks		Ranked data
15	8	1		1
8	10	2	→	3
27	10	3	→	3
25	10	4	→	3
10	12	5		5
23	14	6		6
12	15	7	→	7.5
18	15	8	→	7.5
14	18	9		9
10	23	10		10
15	25	11		11
10	27	12		12

Independent (student) t test

Q1	Q2	Q3	Q4	Q5	Statistical test
Bivariate /Multivariable	**Difference /Correlation**	**Independent / Paired**	**Type of outcome (and Normality)**	**No of groups**	
Bivariate	Difference	Independent (un-paired)	Continuous (Normal)	2	Student's t-test

Dependent variable: Continuous numeric variable

Independent variable: Binary (2 Groups)

Usage: An independent t-test is used to compare the means of two independent groups.

"Independent groups" means that each individual belongs to one of the groups.

Assumptions to be satisfied:

Assumptions	How to check	What to do if the assumption is not met
Normality: dependent variables should be normally distributed within each group	Histograms of dependent variable per group Tests of normality (Shapiro-Wilk, Kolmogorov-Smirnov)	Use Mann-Whitney U test
Homogeneity of variance (standard deviation)	Levene's test (part of standard SPSS output)	Use bottom row of t test output in SPSS "equal variances not assumed"

Where to find in SPSS:

Analyze → Compare Means → Independent-Samples T Test.

Example:

Comparing hemoglobin level of patients in the treatment and control groups.

Steps:

Step 1: We test if hemoglobin is normally distributed in both groups using Shapiro-Wilk test, or Kolmogorov-Smirnov test. We should also have a look at the histogram.

Step 2: After confirmation that hemoglobin is normally distributed in both groups, we use the independent sample t test.

Step 3: We check the result of Levene's test for the homogeneity of variance which is part of the output in SPSS to decide which row should be used for reporting the result (The first row is for equal variance, and the second is for the non-equal variance).

⚐ If you are using software other than SPSS, run Levene's test before running the independent sample t test.

Interpretation of the result:

If the p-value < 0.05 (or another chosen significance level), there is a statistically significant difference between the means of the two groups.

How to report the result:

Report the means of the two groups (with the standard deviation) or the mean difference with the confidence interval for the difference and the p-value.

Table presentation of the result:

	Mean (SD)		Difference (95% CI)	P-value
	Treatment group (N=20)	Control Group (N=20)		
Age in years	32.55 (5.60)	30.15 (5.69)	2.4 (-1.21, 6.01)	0.187
Hemoglobin mg/dl	12.86 (1.69)	11.37 (1.26)	1.49 (0.53, 2.45)	0.003

Reporting significant results:

An independent-samples t-test was done to determine if there were differences in hemoglobin level between treatment and control groups. The hemoglobin level was higher in the treatment group (12.86 ± 1.69) than the control group (11.37 ± 1.26), a statistically significant difference of 1.49 (95%CI: 0.53, 2.45) was found, p = .003.

Reporting non-significant results:

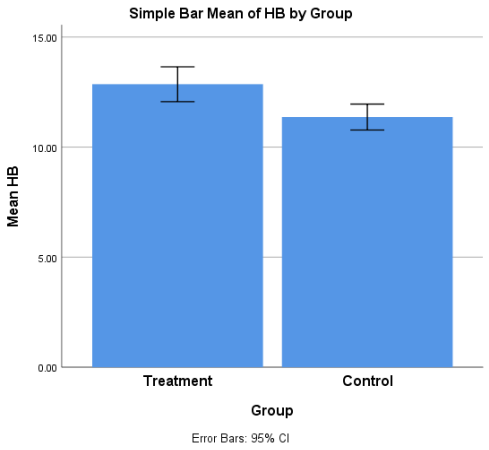
An independent-samples t-test was done to determine if there were differences in age between treatment and control groups. The age was not different in the treatment group (32.55 ± 5.60) from the control group (30.15 ± 5.69), p = .187.

Graphing the output:

The aim of graphing the difference between the two groups is either to show the mean with 95% CI for each group (bar graph or dot plot) or to show the distribution of the variable in the two groups (boxplot).

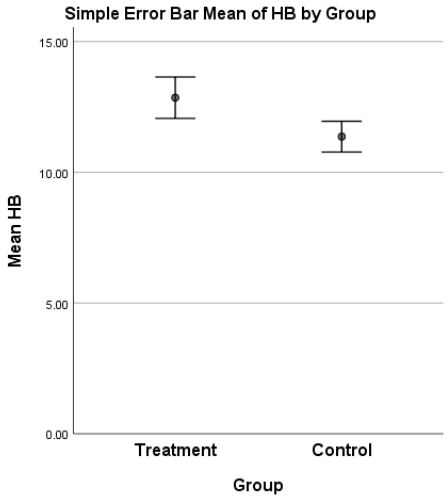
- Bar graph for the means with error bars

Each bar represents the mean of this group and the small error bars represent the 95% CI of the mean.

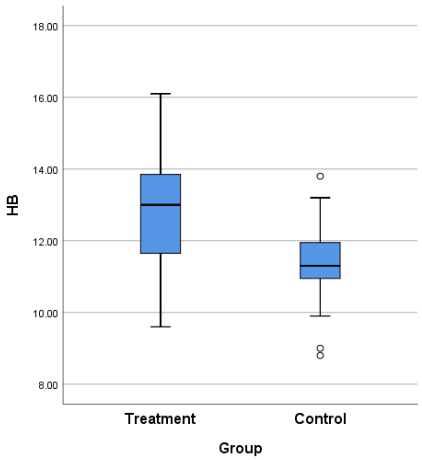


- Dot plot for the means with error bars

Each dot represents the mean of this group and the error bars represent the 95% CI of the mean.



- Boxplot for the distribution of the studied variable in the two groups



Online calculator:

- Using raw data:

  https://www.socscistatistics.com/tests/studentttest/default2.aspx

- Using summary data (means and standard deviations)

  https://www.usablestats.com/calcs/2samplet&summary=1

Mann-Whitney test

Q1	Q2	Q3	Q4	Q5	
Bivariate /Multivariable	Difference /Correlation	Independent /Paired	Type of outcome (and Normality)	No of groups	Statistical test
Bivariate	Difference	Independent (un-paired)	Continuous (Non-normal)/ Ordinal	2	Mann-Whitney U test

It is the non-parametric equivalent to the independent t-test

Dependent variable: Continuous (Non-normal)/ Ordinal

Independent variable: Binary(Group)

Usage: It is used to compare whether two groups containing different people are the same or not. The Mann-Whitney test ranks all of the data and then compares the sum of the ranks for each group to determine whether the groups are the same or not.

Where to find in SPSS:

Analyze → Nonparametric Tests → Independent Samples

Example:

We want to compare the hospital length of stay in the treatment and control groups.

Steps:

Step 1: We test if the length of stay is normally distributed in both groups using Shapiro-Wilk test, or Kolmogorov-Smirnov test. We should also have a look at the histogram.

Step 2: If we find that the length of stay is not normally distributed in both groups, we use the Mann-Whitney U test.

* If the outcome we are studying is an ordinal variable, we move directly to the Mann-Whitney U test.

Interpretation of the result:

If the p-value < 0.05 (or another chosen significance level), there is a statistically significant difference between the ranks of the two groups.

How to report the result:

Report the medians of the two groups with the interquartile range, and the p-value.

Table presentation of the result:

	Median (IQR)		P-value
	Treatment group (N=20)	Control Group (N=20)	
Length of stay in days	2 (2)	3 (3)	0.012
No of outpatient visits	4 (2)	4 (2)	0.947

Reporting significant results:

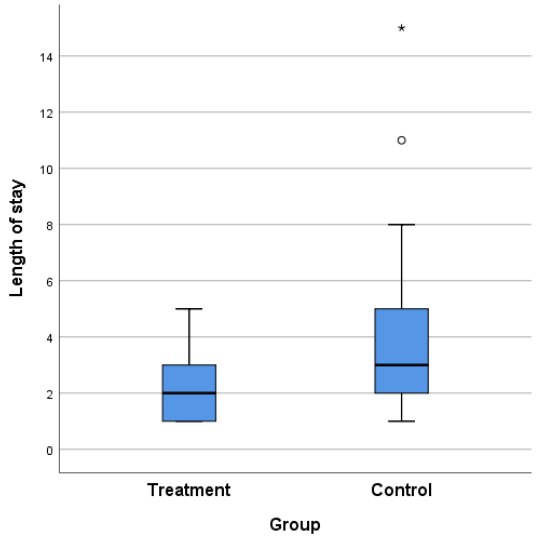
A Mann-Whitney U test was used to examine if there is a difference between the treatment and control groups regarding hospital length of stay. Patients in the treatment group stayed less in the hospital (median = 2) than patients of the control group (median = 3), p =0.012.

Reporting non-significant results:

A Mann-Whitney U test was used to examine the difference in the number of outpatient visits between the treatment and control groups. No significant difference in the number of outpatient visits was found p=0.947. The median number of outpatient visits in both groups is 4 visits.

Graphing the output:

The aim of graphing is to show the distribution of the variable in the two groups (boxplot).



Online calculator:

Using raw data:

https://www.socscistatistics.com/tests/mannwhitney/default2.aspx

another link:

https://astatsa.com/WilcoxonTest/

Paired t-test

Q1	Q2	Q3	Q4	Q5	Statistical test
Bivariate /Multivariable	Difference /Correlation	Independent / Paired	Type of outcome (and Normality)	No of groups	
Bivariate	Difference	Dependent (paired)	Continuous (Normal)	2	Paired t-test

Dependent variable: Continuous

Independent variable: 2 time points/ pre-post testing / 2 conditions

Usage: A paired samples t-test can only be used when the data is paired or matched. Either there are before/after measurements of the same variable or the t-test can be used to compare how a group of subjects perform under two different test conditions. The test assesses whether the mean of the paired differences is zero.

Assumptions:

Assumptions	How to check	What to do if the assumption is not met
Normality: paired differences should be normally distributed	Histograms of the difference / tests of normality (Shapiro-Wilk, Kolmogorov-Smirnov)	Wilcoxon signed rank test

Where to find in SPSS:

Analyze → Compare means → Paired-samples T-test

Example:

Comparing the weight of a group of individuals before and after being on a specific diet to see if there is any difference.

Steps:

Step 1: We calculate the difference between the two readings.

Step 2: We test if the difference is normally distributed using Shapiro-Wilk test, or Kolmogorov-Smirnov test. We should also have a look at the histogram.

Step 3: After confirmation that the difference is normally distributed, we use the paired t test.

Interpretation of the result:

If the p-value <0.05 (or another chosen significance level), then there is a statistically significant difference between the two time points/ measurements /experiments.

How to report the result:

Report the mean and standard deviation of the two conditions, the mean difference, the confidence interval of the difference, and the p-value.

Table presentation of the result:

	Mean (SD)		Difference (95% CI) After-before	P-value
	Before the program	After the program		
Weight in Kg	71.61 (12.31)	63.79 (10.95)	-7.82 (-13.63, -2.01)	0.011
Hemoglobin mg/dl	11.37 (1.26)	11.98 (1.53)	0.61 (-0.28, 1.50)	0.168

Reporting significant results:

A paired-samples t test was used to compare the mean weight before the program to the mean weight after the program. The mean weight before the program was 71.61 (SD=12.31), and the mean weight after the program was 63.79 (SD=10.95). A statistically significant decrease of -7.82 kg (95%CI, -13.63, -2.01) was found, p =0.011.

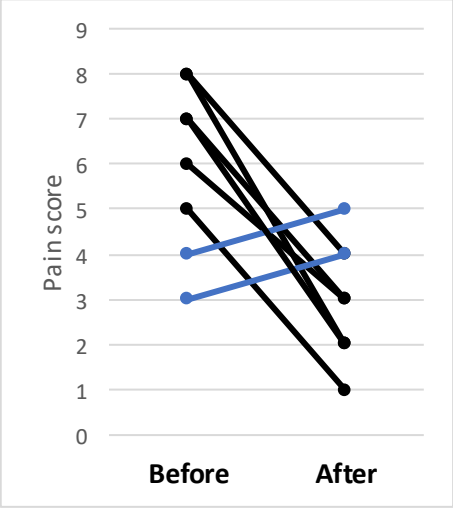
Reporting non-significant results:

A paired-samples t test was used to compare the mean hemoglobin before the program to the mean hemoglobin after the program. The mean hemoglobin before the program was 11.37 (SD=1.26), and the hemoglobin after the program was 11.98 (SD=1.53). No statistically significant difference was found, p=0.168.

Graphing the output:

The aim of graphing the difference between the two conditions is either to show the mean with 95% CI for each condition **(bar graph or dot plot)** or to show the distribution of the variable in the two conditions **(boxplot)**.

If the number of cases is small (10 or less for example), a graph showing lines connecting the pre and post conditions of each case can be used.



Online calculator:

Using raw data:

https://www.socscistatistics.com/tests/ttestdependent/default2.aspx

one other calculator:

https://mathcracker.com/t-test-for-paired-samples

Wilcoxon Signed Rank test

Q1	Q2	Q3	Q4	Q5	
Bivariate /Multivariable	Difference /Correlation	Independent / Paired	Type of outcome (and Normality)	No of groups	Statistical test
Bivariate	Difference	Dependent (paired)	Continuous (Non-normal)/ Ordinal	2	Wilcoxon signed-rank test

It is the non-parametric equivalent to the paired t-test

Dependent variable: Continuous (Non-normal)/ Ordinal

Independent variable: 2 time points / pre-post testing / 2 conditions

Usage: The Wilcoxon signed-rank test is used to compare two related samples, matched samples, or repeated measurements on a single sample to assess whether the two mean ranks differ. It is a paired difference test and the absolute differences are ranked, then the signs of the actual differences are used to add the negative and positive ranks.

Where to find in SPSS:

Analyze → Nonparametric tests → 2 related samples

Example:

Comparing the pain score of a group of individuals before and after receiving a specific medication to see if there is any difference.

Steps:

Step 1: We calculate the difference between the two conditions.

Step 2: We test if the difference is normally distributed using Shapiro-Wilk test, or Kolmogorov-Smirnov test. We should also have a look at the histogram.

Step 3: If the difference is not normally distributed, we use the Wilcoxon signed-rank test.

* If the outcome we are studying is an ordinal variable, we move directly to the Wilcoxon signed-rank test.

Interpretation of the result:

If p-value < 0.05 (or another chosen significance level), then there is evidence that the mean ranks differ (a change has occurred between the two conditions).

How to report the result

Report the medians and interquartile range of the two sets of measurements, and the p-value.

The percentage of positive and negative ranks (increased or decreased scores) might also be reported.

Table presentation of the result:

	Median (IQR)		P-value
	Before the medication	After the medication	
Pain score	6.5 (3.5)	2.5 (1.75)	0.034
Satisfaction level	2 (1.75)	2.5 (1)	0.272

Reporting significant results:

A Wilcoxon signed-rank test was used to compare the median pain score before and after the medication. A significant difference was found in the results, p = 0.034. The median pain score after the medication (median =2.5) was lower than the median pain score before the medication (median =6.5).

Reporting non-significant results:

A Wilcoxon signed-rank test was used to compare the median satisfaction level before and after the medication. No significant difference was found in the results, p = 0.272. The median satisfaction level after the medication was not different from the median satisfaction level before the medication.

Graphing the output:

The aim of graphing the difference between the two conditions is to show the distribution of the variable in the two conditions (boxplot).

If the number of cases is small (10 or less for example), a graph showing lines connecting the two conditions of each case can be used.

Online calculator:

Using raw data:

https://www.socscistatistics.com/tests/signedranks/default2.aspx

one other calculator:

https://astatsa.com/WilcoxonTest/

One-way ANOVA

Q1	Q2	Q3	Q4	Q5	
Bivariate /Multivariable	Difference /Correlation	Independent / Paired	Type of outcome (and Normality)	No of groups	Statistical test
Bivariate	Difference	Independent (un-paired)	Continuous (Normal)	>2	One-way ANOVA

It can be thought of as an extension of the t-test for 3 or more independent groups.

Dependent variable: Continuous

Independent variable: Categorical (at least 3 categories)

Usage: Used to examine the difference in means of 3 or more independent groups. ANOVA uses the ratio of the "between-group variance" to the "within-group variance" to decide whether there are statistically significant differences between the groups or not.

Assumptions:

Assumptions	How to check	What to do if the assumption is not met
Normality: dependent variables should be normally distributed within each group	Histograms / Tests of normality (Shapiro-Wilk, Kolmogorov-Smirnov)	Kruskall-Wallis test (non-parametric)
Homogeneity of variance	Levene's test	Welch test instead of ANOVA (adjusted for the differences in variance) or Kruskal-Wallis test

Where to find in SPSS:

Analyze → Compare means → One-Way ANOVA

Example:

Comparing the birthweight of a group of infants of mothers with different smoking status (never smoke, quit before pregnancy, smoke during pregnancy).

Steps:

Step 1: We test if birth weight is normally distributed in the three groups using Shapiro-Wilk test, or Kolmogorov-Smirnov test. We should also have a look at the histograms.

Step 2: After confirmation that birth weight is normally distributed in the three groups, we run the one-way ANOVA test and the Levene's test for homogeneity of variance.

Step 3: We check the result of Levene's test for the homogeneity of variance, if there is no homogeneity of variance, we need to run the Welsh test from which we report the result.

Step 4: If the result of the one-way ANOVA is statistically significant ($p<0.05$), we need to do a post hoc test.

Interpretation of the result:

If the p-value ≥ 0.05 (or another chosen significance level), we conclude that there is no significant difference between the groups.

If the p-value < 0.05 (or another chosen significance level), we conclude that there is a significant difference between at least one pair of the groups. Post-hoc tests are used to test where the pairwise differences are.

Post-hoc testing:

We use the post-hoc tests to make comparisons between each pair of the groups while adjusting the p-value. The aim is to reduce the probability of occurrence of type 1 error (the probability of having type 1 error increases as the comparisons increase).

Bonferroni, Tukey, or Scheffe are commonly used post-hoc tests (adjustments). Some other tests can be used for different conditions.

How to report the result:

Report the mean and standard deviation of each group, the p-value for one-way ANOVA, and the significant pairwise differences.

Table presentation of the result:

	Mean (SD)			**P-value**
	Smoke during pregnancy **(N=10)**	Quit before pregnancy **(N=10)**	Never smoke **(N=10)**	
Birth weight (g)	2606 (334)	2959 (490)	3101 (411)	0.037*
Age (years)	31.5 (6.6)	29.8 (4.4)	31.3 (6.4)	0.782

* Post-hoc testing using Bonferroni adjustment showed that mean birth weight for the never smoke group is higher than that who smoke during pregnancy.

Reporting significant results:

We conducted a one-way ANOVA test to compare the birth weight of infants of mothers with different smoking behavior (never smoke, quit before pregnancy, smoke during pregnancy). A significant difference was found among the groups, p=0.037. Bonferroni test was used to determine the nature of the differences between those groups. This analysis revealed that the birth weight of infants to mothers who smoke during pregnancy was lower (M = 2606, sd = 334) than that of infants to mothers who never smoke (M = 3101, sd = 411). The birth weight of infants to mothers who quit smoking before pregnancy (M = 2959, sd = 490) was not significantly different from either of the other two groups.

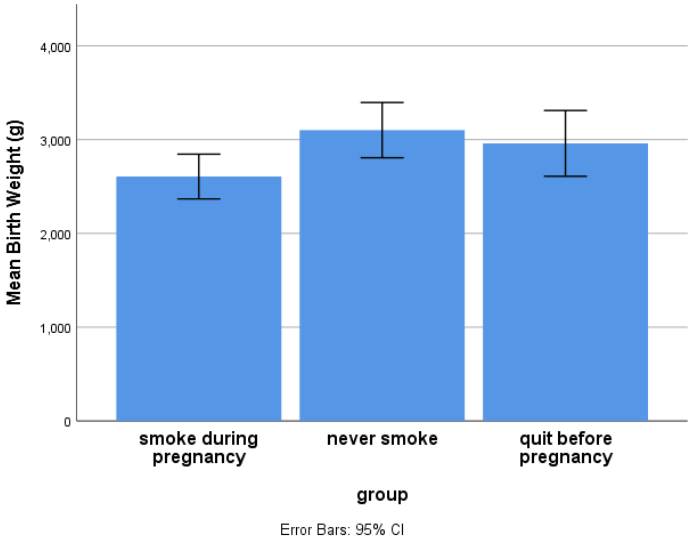
Reporting non-significant results:

We conducted a one-way ANOVA test to compare the age of mothers with different smoking behavior (never smoke, quit before pregnancy, smoke during pregnancy). No statistically significant difference was found among the three groups, p=0.782.

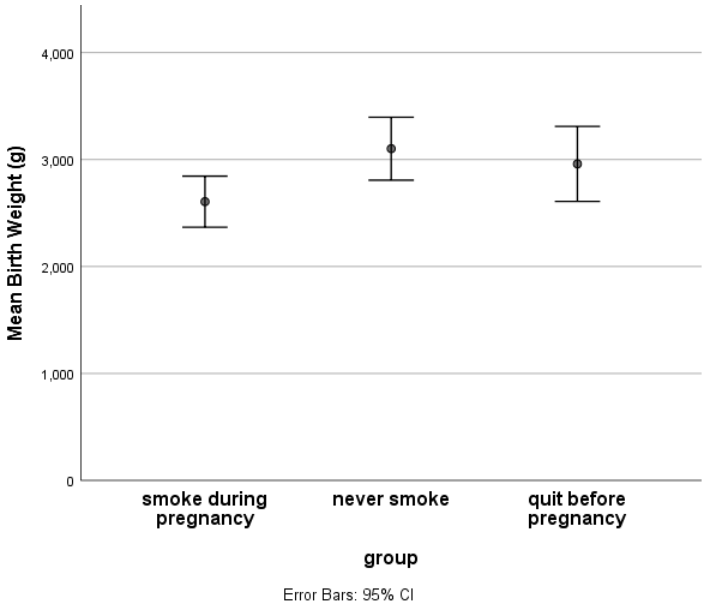
Graphing the output:

The aim of graphing the difference between the groups is either to show the mean with 95% CI for each group (bar graph or dot plot) or to show the distribution of the variable in the groups (boxplot).

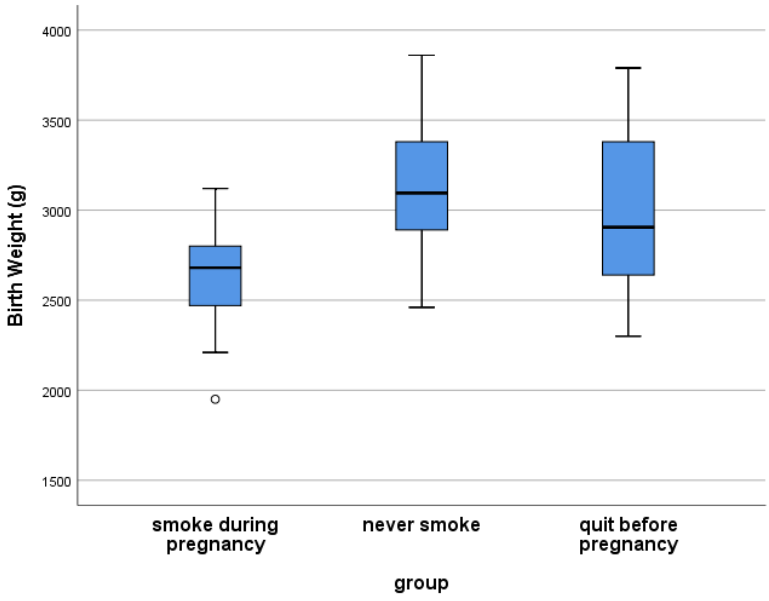
- Bar graph for the means with error bars



- Dot plot for the means with error bars



- Boxplots



Online calculator:

Using raw data:

https://www.socscistatistics.com/tests/anova/default2.aspx

one other calculator:

https://goodcalculators.com/one-way-anova-calculator/

Kruskal-Wallis test

Q1	Q2	Q3	Q4	Q5	
Bivariate /Multivariable	Difference /Correlation	Independent / Paired	Type of outcome (and Normality)	No of groups	Statistical test
Bivariate	Difference	Independent (un-paired)	Continuous (Non-normal)/ Ordinal	>2	Kruskal-Wallis test

It is the non-parametric equivalent to the one-way ANOVA

Dependent variable: Continuous (Non-normal)/ Ordinal

Independent variable: Categorical (at least 3 categories)

Usage: Kruskal-Wallis compares the medians of three or more samples to determine if the samples came from different populations.
It is an extension of the Mann–Whitney U test to 3 or more groups.

Where to find in SPSS:

Analyze → Nonparametric tests → Independent samples

Example:

Comparing the neonatal intensive care unit (NICU) length of stay for a group of infants of mothers with different smoking status (never smoke, quit before pregnancy, smoke during pregnancy).

Steps:

Step 1: We test if NICU length of stay is normally distributed in the three groups using Shapiro-Wilk test, or Kolmogorov-Smirnov test. We should also have a look at the histograms.

Step 2: If the NICU length of stay is not normally distributed in the three groups, we run the Kruskal-Wallis test.

Step 4: If the result of the Kruskal-Wallis test is statistically significant ($p<0.05$), we need to do the post hoc testing using Mann–Whitney U test with Bonferroni correction.

Interpretation of the result:

If the p-value ≥ 0.05(or another chosen significance level), we conclude that there is no significant difference between the groups.

If the p-value < 0.05 (or another chosen significance level), we conclude that there is a significant difference between at least one pair of the groups. Post-hoc test is done to test where the pairwise differences are.

How to report the result:

Report the median and IQR of each group, the p-value for the Kruskal-Wallis test, and the significant pairwise differences.

Table presentation of the result:

	Median (IQR)			P-value
	Smoke during pregnancy **(N=10)**	Quit before pregnancy **(N=10)**	Never smoke **(N=10)**	
NICU length of stay (days)	4 (4)	2 (1)	2 (1.25)	0.004*

* Post-hoc testing using Bonferroni adjustment showed that median NICU length of stay for infants of mothers who smoke during pregnancy was higher than the other two groups.

Reporting significant results:

We computed Kruskal-Wallis test to compare the median NICU length of stay for infants of mothers with different smoking behavior (never smoke, quit before pregnancy, smoke during pregnancy). A significant difference was found among the groups, p=0.004. Post-hoc testing was done using Mann–Whitney U test with Bonferroni correction to determine the nature of the differences between the groups. This analysis revealed that the median NICU length of stay of infants to mothers who smoke during pregnancy was higher (Median = 4) than that for infants to mothers who never smoke (Median = 2), and for infants to mothers who quit before pregnancy (Median = 2). The median NICU length of stay of infants to mothers who quit smoking before pregnancy was not significantly different from that of infants to mothers who never smoke.

Reporting non-significant results:

We computed the Kruskal-Wallis test to compare the median gestational age for infants of mothers with different smoking behavior (never smoke, quit before pregnancy, smoke during pregnancy). No statistically significant difference was found among the three groups, p=0.699.

Graphing the output:

The aim of graphing is to show the distribution of the variable in the different groups (**boxplot**).

Online calculator:

Using raw data:

https://www.socscistatistics.com/tests/kruskal/default.aspx

one other calculator:

https://mathcracker.com/kruskal-wallis

Part 5 Categorical data analysis

Relative Risk and Odds Ratio

Risk Ratio (Relative Risk, RR) and Odds Ratio (OR) are different measures of association. We have to know the difference between them.

To understand how they are calculated, let us consider this hypothetical example:

A cohort study was done to follow up 800 individuals for 5 years period, 400 are smokers, and 400 are non-smokers. They were followed up for the occurrence of coronary heart disease.

The result is presented in the following table:

	Diseased	Not Diseased	Total
Smokers	40 a	360 b	400 a+b
Non-smokers	20 c	380 d	400 c+d
Total	60 a+c	740 b+d	800 a+b+c+d

Relative risk calculation	Odds ratio calculation
• The **risk** (incidence) of having coronary heart disease **among smokers** = a/a+b =40/400= 10% • The **risk** (incidence) of having coronary heart disease **among non-smokers** =c/c+d =20/400= 5% • **Relative risk** = incidence among exposed/ incidence among non-exposed = $\frac{a/(a+b)}{c/(c+d)}$ • **RR**= 0.1/0.05=10/5=**2** • The risk of having coronary heart disease among smokers is 2 times the risk of having heart disease among non-smokers. • It can be also interpreted as: The risk of coronary heart disease among smokers is 200% of that among non-smokers. • Smoking is associated with 100% increase in the risk of coronary heart disease	• The **odds** of having coronary heart disease **among smokers** =a/b = 40/360= 0.11 • The **odds** of having heart diseases **among non-smokers** = c/d = 20/380= 0.053 • **Odds ratio** = odds of the disease among exposed/ odds of the disease among non-exposed = $\frac{a/b}{c/d} = \frac{a\,d}{bc}$ • **OR**= 0.11/0.053= **2.08** • The odds of having heart diseases among smokers is 2.08 times the odds of having heart diseases among non-smokers. • OR is less easy to express in English, and less easy to be understood than RR. • It can be also interpreted as: for every 208 persons who experience the event in the exposed group, 100 persons will experience the event in the non-exposed group.
Interpretation • RR of 1.00 means that the risk of the event is identical in the exposed and non-exposed groups. • RR that is less than 1.00 means that the risk is lower in the exposed group. • RR that is greater than 1.00 means that the risk is higher in the exposed group.	**Interpretation** ORs are interpreted in the same way as RRs. • OR of 1.00 means that there is no increase or decrease in risk • OR that is less than 1.00 means that the risk is lower in the exposed group. • OR that is greater than 1.00 means that the risk is higher in the exposed group. • We mentioned here "risk" although we are talking about the odds as it is easier to understand.

Relative Risk and Odds Ratio Calculator:

https://www.socscistatistics.com/biostatistics/default2.aspx

When using this Relative Risk and Odds Ratio Calculator with the same numbers of the previous examples, we get the following output:

Result					
	Bad Outcome	Good Outcome	Total	Risk	Odds
Group 1	40	360	400	0.1	0.11
Group 2	20	380	400	0.05	0.05
Total	60	740			

Result

Relative Risk = 2.

Odds Ratio = 2.11.

Calculate Reset

Summary of Calculation

Group 1 Risk = 40 ÷ 400 = 0.1

Group 2 Risk = 20 ÷ 400 = 0.05

Relative Risk 1 = 0.1 ÷ 0.05 = 2

Relative Risk 2 = 0.05 ÷ 0.1 = 0.5

Group 1 Odds = 0.1 ÷ (1 - 0.1) = 0.11

Group 2 Odds = 0.05 ÷ (1 - 0.05) = 0.05

Odds Ratio 1 = 0.11 ÷ 0.05 = 2.11

Odds Ratio 2 = 0.05 ÷ 0.11 = 0.47

We notice that there are two Relative risks and two Odds ratios calculated. This is because we can calculate the relative risk of having coronary heart disease among smokers compared to non-smokers which is 2, or we can calculate the relative risk of having coronary heart disease among non-smokers compared to smokers which is 0.5.

The same concept can be applied to calculating the odds ratios.

Important notes:

• Odds Ratio is different from Risk Ratio (relative risk).

• Only in rare diseases the value of odds ratio and relative risk will be almost numerically similar.

• **Odds** is calculated by dividing part/part as diseased/not diseased, while **risk** is calculated by dividing part/total as diseased/all exposed (diseased and not diseased).

• **Relative risk** is calculated in cohort studies but not in case-control studies.

• **Odds ratios** are important as they are used for the interpretation of logistic regression and are the only suitable measure of association in case-control studies.

• Relative risk and Odds ratio should always be reported with the 95% confidence interval.

• If the confidence interval includes (1), then there is no significant association between exposure and the outcome.

Chi-square test and Fisher's exact test

Q1	Q2	Q3	Q4	Q5	
Bivariate /Multivariable	Difference /Correlation	Independent /Paired	Type of outcome (and Normality)	No of groups	Statistical test
Bivariate	Difference	Independent (un-paired)	Nominal	2	Chi-square test/ Fisher's exact test
				>2	Chi-square test

The chi-square test for independence (also called Pearson's chi-square test or the chi-square test of association) is used to study if there is a relationship between two categorical variables. It is a non-parametric test.

Dependent variable: Categorical (nominal)

Independent variable: Categorical (nominal)

Usage: used to study if there is a relationship/association between two categorical variables. relationship/association between the two categorical variables. The chi-squared test compares expected frequencies, assuming the null is true, with the observed frequencies from the study.

How expected values are calculated?

1- We make a 2*2 table for the 2 variables (sex, preferred drink) with the observed (actual values)

	Prefer coffee	Prefer tea
Men	207	282
Women	231	242

2- We add the totals to the rows and columns

	Prefer coffee	Prefer tea	total
Men	207	282	489
Women	231	242	473
total	438	524	962

3- We calculate the "Expected Value" for each cell. This is done by multiplying each row total by each column total and dividing that by the overall total.

	Prefer coffee	Prefer tea	total
Men	$\frac{489 \times 438}{962}$	$\frac{489 \times 524}{962}$	489
Women	$\frac{473 \times 438}{962}$	$\frac{473 \times 524}{962}$	473
total	438	524	962

4- We get the following expected values (if there is no association between the two variables):

	Prefer coffee	Prefer tea	total
Men	222.64	266.36	489
Women	215.36	257.64	473
total	438	524	962

The chi-square test works by comparing the observed values (actual data) to the expected data (if there is no association).

Assumptions:

Assumption	How to check	What to do if the assumption is not met
80% of expected cell counts >5 (less than 20% of cells have expected count less than 5)	SPSS gives this in the output	Fisher's exact (for 2X2 tables) or exact test (for more than 2X2) or merge categories where possible

Where to find in SPSS:

Analyze → Descriptive → Crosstabs → Statistics

Example:

Comparing males and females regarding the preferred drink (tea or coffee). Is there an association between sex (male and female), and the preferred drink (tea or coffee).

Steps:

Step 1: We run the chi-square test on SPSS.

Step 2: We check if we have expected counts less than 5, and what is the percentage of those cells.

Step 3: If < 20% of expected cell counts is < 5, we report the p-value of the Chi Square test.

If > 20% of expected cell counts is < 5, we need to use another test.

- If 2X2 (each variable has only 2 categories), we use Fisher's exact test.

- If more than 2X2, use the exact test

Interpretation of the result:

If p< 0.05, there is a significant relationship between the two variables.

How to report the result:

We use percentages to describe what the relationship is. We present the result in a table to summarise where the differences between the groups are using numbers and percentages.

Table presentation of the result:

	Prefer coffee N (%)	Prefer tea N (%)	P-value
Males	207 (42.3)	282 (57.7)	0.043
Females	231 (48.8)	242 (51.2)	

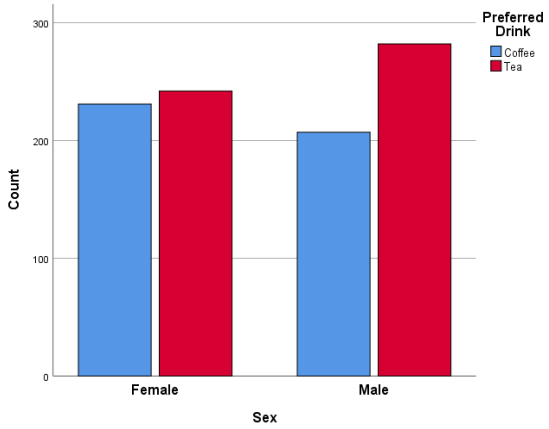
Reporting significant results:

A chi-square test was conducted between sex and preferred drink. There was a statistically significant association between sex and the preferred drink, p = 0.043. A higher percentage of females (48.8%) prefer coffee as compared to males (42.3%).

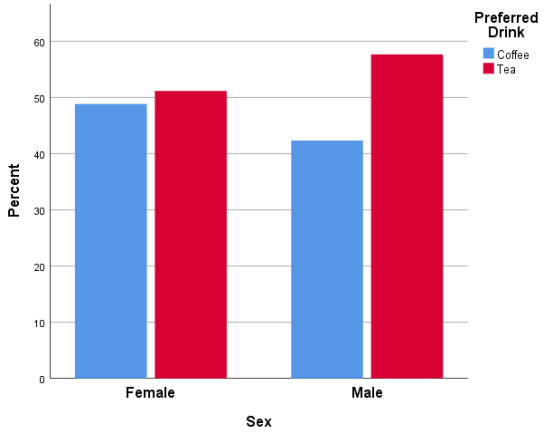
Graphing the output:

Clustered bar charts are used to present the two variables using frequencies or percentages.

- Bar chart with count:



- Bar chart with percentage:



Online calculator:

Using summary data:

https://www.socscistatistics.com/tests/chisquare2/default2.aspx

one other calculator:

http://quantpsy.org/chisq/chisq.htm

Fisher's exact test:

It is used when we have a 2 x 2 crosstabulation and the assumption that "less than 20% of cells have expected count less than 5" is not met.

It is done in the same way as the Chi-square test and is interpreted in the same way. The difference is that we report the P-value from Fisher's exact test part in the output.

Online calculator:

https://www.socscistatistics.com/tests/fisher/default2.aspx

Exact test:

It is used when we have more than 2 x 2 crosstabulation (one or two of the variables have more than 2 categories) and the assumption that "less than 20% of cells have expected count less than 5" is not met. To get this result, we use the exact option in the exact button while doing crosstabulation in SPSS.

It is done in the same way as the Chi-square test and is interpreted in the same way. The difference is that we report the P-value from the exact column in the output.

Other statistical tests (not covered in this book)

Repeated measures ANOVA

- It is used to compare a continuous variable that is measured in three or more conditions/ time points.
- It is similar to one-way ANOVA but used for repeated samples.
- It is considered an extension of a paired-samples t-test, but for more than 2 conditions/time points.

Friedman test

- It is the non-parametric equivalent to repeated measures ANOVA.
- It is used to compare an ordinal variable that is measured in three or more conditions/ time points.
- It is similar to the Kruskal Wallis test but used for repeated samples.
- It is considered an extension of the Wilcoxon signed-rank test, but for more than 2 conditions/time points.

Two-way ANOVA

- It is also called factorial ANOVA.
- It is used to determine if there is an interaction effect between two independent variables on a continuous variable (for example if there is an interaction between the effects of Drug A and Drug B on Blood pressure).

McNemar's test

- McNemar test is used to determine if there are differences in a binary dependent variable between two time points/conditions.
- It is similar to the paired-samples t-test, but for a binary rather than a continuous variable.
- For example, a group of students were assessed at the beginning of the academic year whether they were smokers or non-smokers. Then, they are assessed again at the end of the year whether they are smokers or not. McNemar's test is used to determine whether the proportion of smokers has changed by the end of the year.

Part 6 Additional topics

Correlation

Q1	Q2	Q3	Q4	Q5	
Bivariate /Multivariable	Difference /Correlation	Independent / Paired	Type of outcome (and Normality)	No of groups	Statistical test
Bivariate	Correlation		Continuous (Normal)		Pearson's correlation (r)
			Continuous (Non-normal)/ Ordinal		Spearman's correlation (ρ)

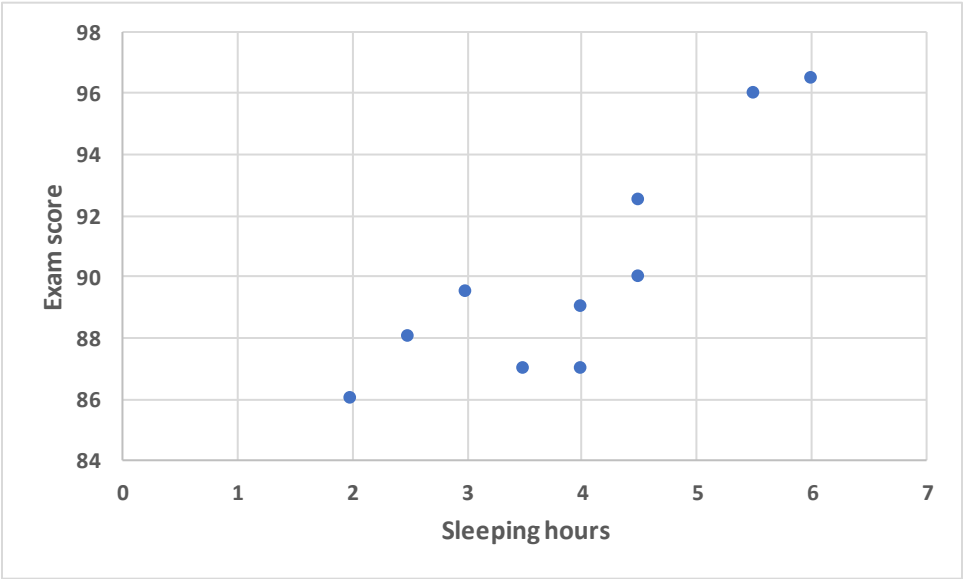
- Correlation is used to assess the relationship between two variables (continuous or ordinal).
- It gives the following information:
    - **Direction** of relationship (positive/negative)
    - **Strength** of relationship (weak / medium / strong)
- Correlation is often explored using scatterplot diagram.

One variable is presented on the X-axis, and the other on the Y-axis.

Example: If we have data for some students regarding the number of sleep hours before the exam and the exam score and we want to check if there is a correlation between the two variables.

Student No.	Number of hours of sleep (X)	Exam score (Y)
1	3.0	89.5
2	2.0	86
3	4.5	92.5
4	5.5	96
5	3.5	87
6	4.0	87
7	2.5	88
8	4.5	90
9	4.0	89
10	6	96.5

The scatterplot looks like this:



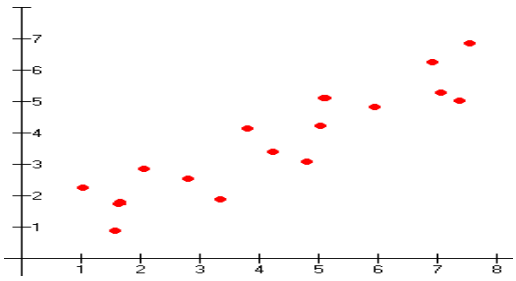
The number of sleeping hours is represented on the X-axis, and exam scores are represented on the Y-axis.

Each point represents a student and his corresponding sleeping hours and exam score.

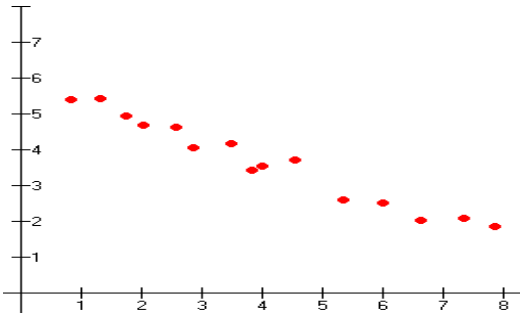
By visual inspection, we can say that the number of hours of sleep (X) and score in the exam (Y) are possibly linearly related to each other.

Direction of the relationship:

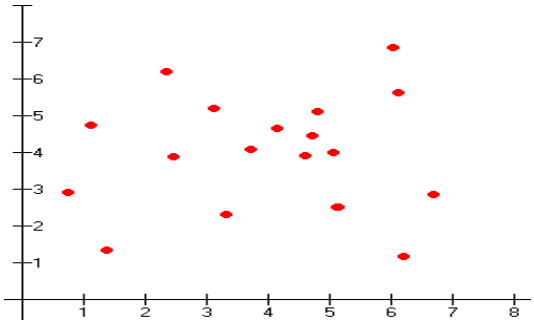
- ☐ **Positively** related – when one increases, the other increases.



- ☐ **Negatively** related – when one goes up, the other goes down.



☐ No relationship between variables



Strength of the relationship:

Can be shown either graphically using the scatterplot or statistically using the correlation coefficient.

☐ **The correlation coefficient (r)**: shows the **strength** and **direction** of the relationship between the two variables.

- It ranges between **-1 and 1**

- Direction of the relationship:

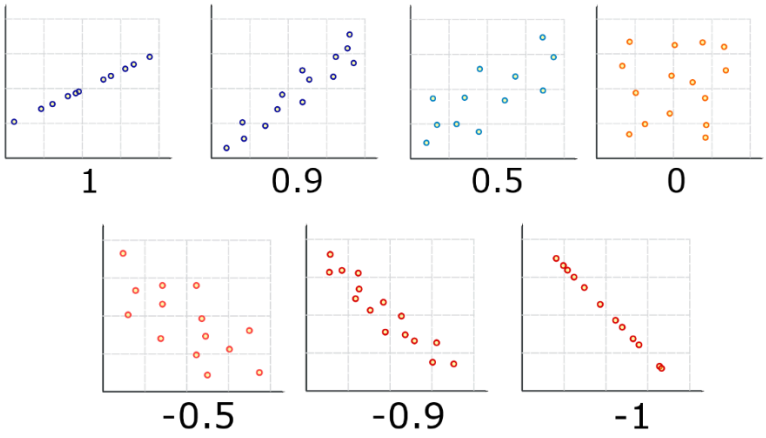
If **positive** : positive association , if **negative** : negative association , if **zero**: no association

- Strength of the relationship:

The nearer the number to 1, or -1, the stronger the relationship. Different cut values are presented in different textbooks, for example, the correlation can be:

- **Perfect:** If the value is **near ± 1**, then it is said to be a **perfect** correlation.
- **High degree:** If the coefficient value lies between **± 0.50 and ± 1**, then it is said to be a **strong** correlation.
- **Moderate degree:** If the value lies between **± 0.30 and ± 0.49**, then it is said to be a **medium** correlation.
- **Low degree:** When the value lies **below ± 0.29**, then it is said to be a **small** correlation.
- **No correlation:** When the value is **zero**.

The following figure shows different scatterplots and the corresponding correlation coefficients:



- ☐ Correlation is significant if p-value < 0.05
- ☐ **Coefficient of determination (R^2):** explains the proportion of variance in one variable as a result of the other variable. It is the square of the correlation coefficient (r).

For the above-mentioned example of the correlation between sleeping hours and exam scores, r= 0.87 indicating a strong positive correlation, and R^2=0.75.

Types of correlation:

- ☐ **Pearson's correlation** is only used for parametric data (numeric normally distributed variables).

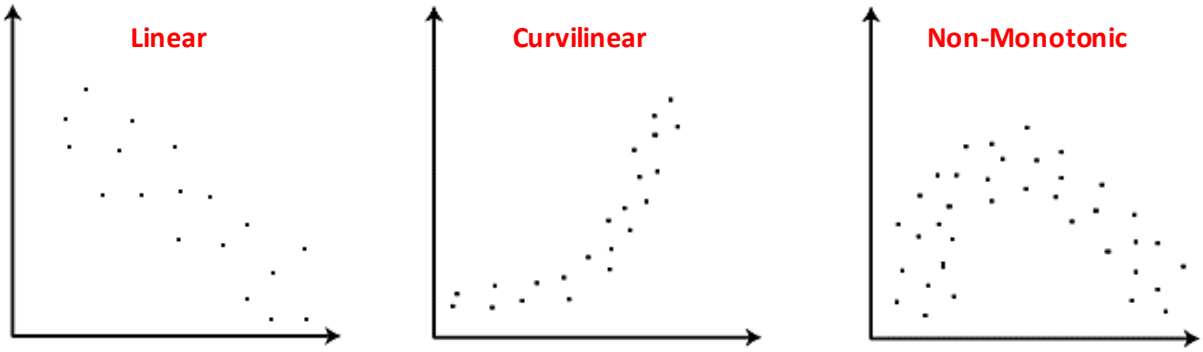
If the data is ordinal (ranked), then we can't use the Pearson's correlation coefficient.

- ☐ **Spearman's rho coefficient** is used for ordinal data, or if the assumptions of numeric data not satisfied.

It is the non-parametric equivalent of Pearson's correlation.

Pearson's correlation (*r*) (parametric)	Spearman's correlation (*p*) (non-parametric)
Two numeric variables	Ordinal or numeric variables
Linear relationship	Monotonic (linear or curvilinear) relationship
Normal distribution (at least for one of the two variables)	

It is important always to plot the data when doing a correlation analysis to check that the relationship is linear.



Pearson's correlation can be used	Spearman's correlation can be used	None of them can be used

☐ **Correlation does not imply causation!**
The presence of correlation between two variables does not mean that one of them causes the other.

Correlation matrix

The correlation coefficient can be used to summarize the relationship between several pairs of continuous variables in the form of a matrix. This allows us to see which pairs have the highest correlation.

The diagonal of the table always contains one, because the correlation between a variable and itself is always 1.

The following table represents a correlation matrix for the exam score in each subject. The numbers in the table represent the correlation coefficient between each pair of variables.

	English	Math	Writing	Reading
English	1			
Math	0.271	1		
Writing	0.366	0.149	1	
Reading	0.386	0.52	0.152	1

From the correlation matrix we can see that the strongest correlation is observed between math and reading scores (r=0.52), while the weakest correlation is observed between math and writing scores (r=0.149).

Online correlation calculators

Pearson's correlation: https://www.socscistatistics.com/tests/pearson/default2.aspx

Spearman's correlation: https://www.socscistatistics.com/tests/spearman/default2.aspx

Simple linear regression

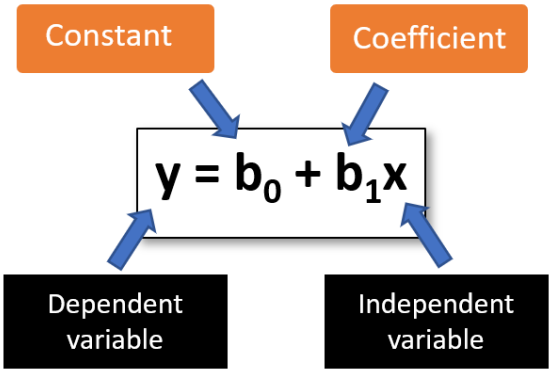
Simple linear regression is used for:

- Studying the **association** between two variables (as the correlation)
- Not just looking at the **direction** and **strength** of the relationship, but also **quantifies** the relationship between the variables (generates an equation).

Types of variables:

- Dependent variable / outcome / y **(should be numeric variable)**
- Independent variable / predictor / x **(can be numeric, ordinal, categorical)**

It generates a regression equation:



Uses of regression:

- To evaluate the **impact** of an independent variable on the outcome.
- It can be used for the **prediction** of the outcome variable using the independent variable.

Example:

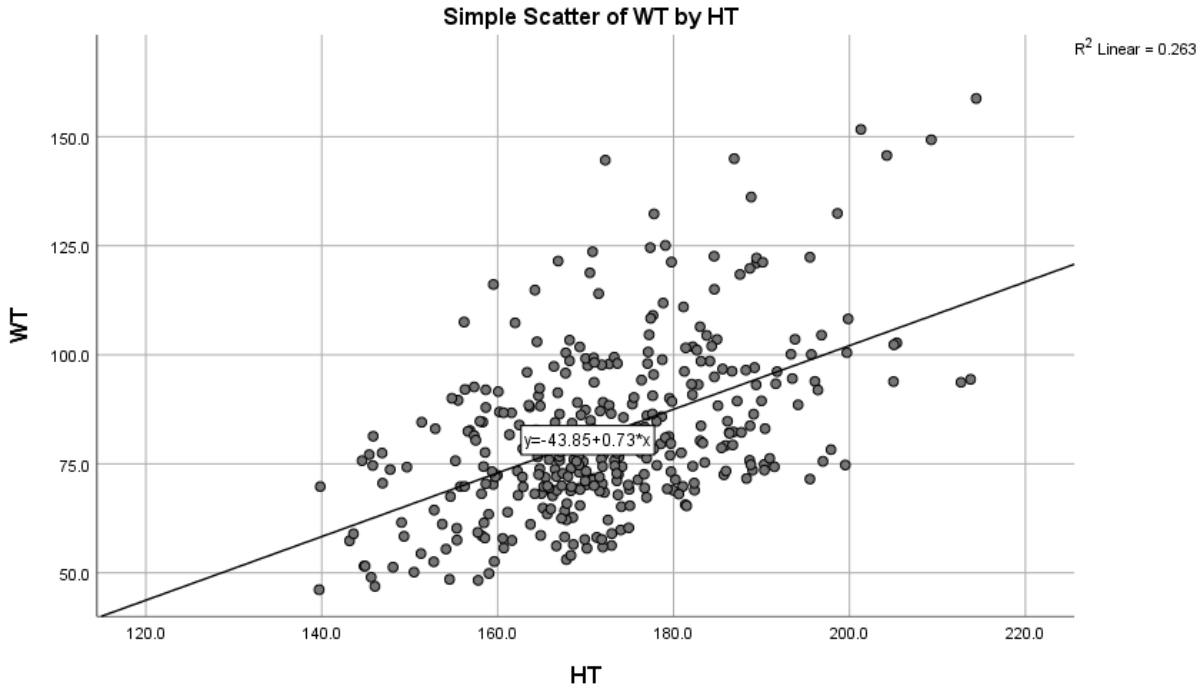
- Using fasting blood glucose to predict HbA1c level in some patients
- Dependent variable: HbA1c (%)
- Independent variable: Blood glucose (mmol/l)
- Regression equation:

HBA1c = 3.2+ (0.45*Blood glucose)

Example:

- Using height to predict weight for some university students.
- Dependent variable: weight in Kg (WT)
- Independent variable: height in cm (HT)
- Regression equation:

WT = -43.8 + (0.73*HT)



This equation can be used to predict the weight: for a student 170cm, we can predict the weight:

WT= -43.8 + (0.73*170)=80.3 kg

Components of the regression model (equation):

$$y = b_0 + b_1 x$$

Where:

y: the outcome (dependent) variable

x: the predictor (independent) variable

b_0 is called the **intercept** or constant

It is the value of y when x = 0, hence the name intercept.

b_1 is the **slope** of the line

It is the amount of change in Y (whether positive or negative, depending on the sign) for each unit increase in the value of X, hence the name slope.

Example:

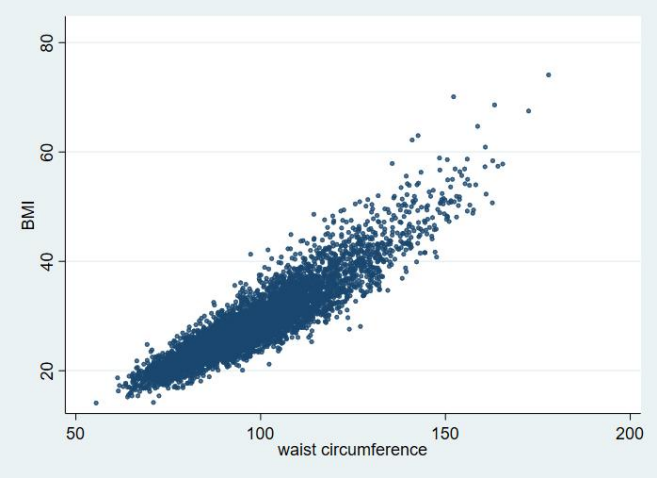
- Using years of experience to predict salary
- Dependent variable: salary in $ (salary)
- Independent variable: years of experience (Exp)
- Regression equation:

salary = 1500 + (250*Exp)

The intercept = 1500 (the value of y when x = 0).
When an individual is a fresh graduate, with no years of experience (x=0), the expected salary is 1500$.
The slope= 250 (the amount of change in y for each unit increase in the value of x).
An individual with one year of experience is expected to have a salary of 1750$, while an individual with two years of experience will have an expected salary of 2000$.
For each year increase in experience, the salary is expected to increase by 250$.

Example:

What is the estimated regression equation of the data on the **waist circumference (X)** and **body mass index (Y)**?



The estimated regression equation from the regression output is:

BMI = -8.5 + 0.38 (waist circumference)

Interpretation:

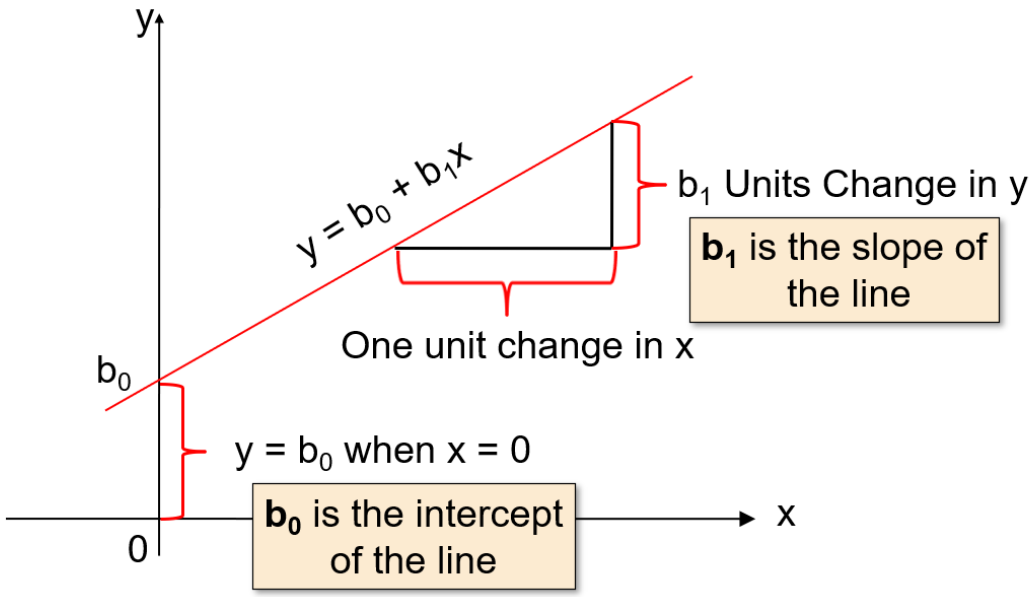
For every unit increase in the waist circumference, there is a 0.38 unit increase in the BMI.

Interpretation of the intercept, in this case, is not possible as the waist circumference can't be zero.

The predicted BMI of a patient having a waist circumference of 110 cm is given by:

BMI = -8.5 + 0.38(110)= 33.3

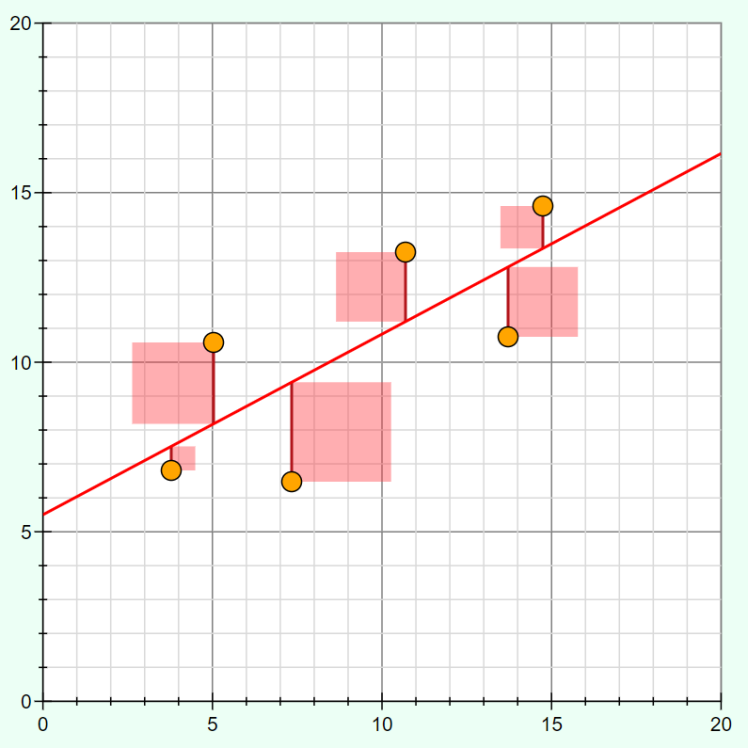
Graphical presentation of the intercept and the slope:



Linear regression fits the 'best' straight line to the data.

The most common method to fit the line is called the **least squares method**.

It is the line that makes the sum of the square of the distance of each point and the line as small as possible as presented in the following graph.



Residuals:

For the "real" datasets, there will always be a difference between what we observe and what our model predicts.

We adjust for this difference by adding an **error** term in the model (**e**):

$$y = b_0 + b_1 x + e$$

- **Residuals** are then defined as the difference between the predicted and observed values:

$e = y - (b_0 + b_1 x)$

= **observed value** (real data) - **predicted value** (from the equation)

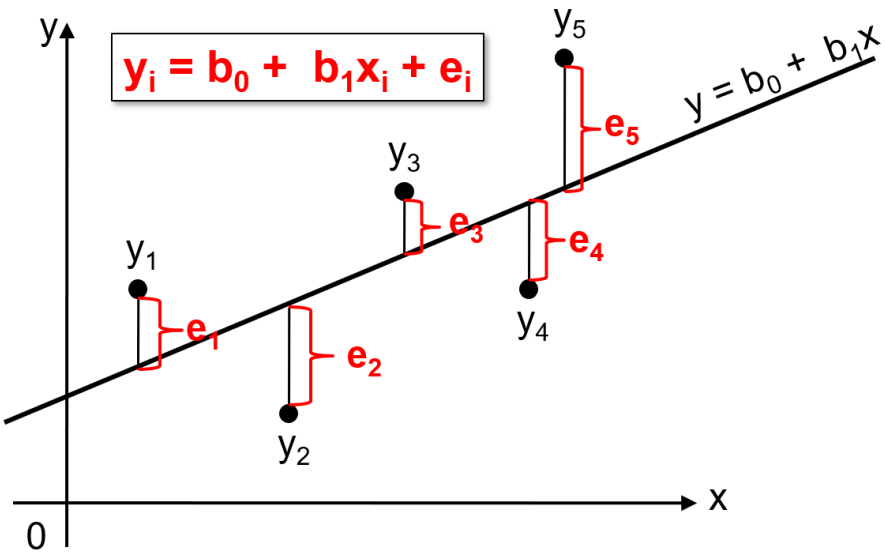
= **error**

It represents the vertical distance between each observation and the regression line.

In the previous example of the relationship between salary and years of experience, the predicted salary for an employee with 2 years of experience is 2000$. The actual salary the employee is receiving may be 2000, 2100, or 1800, with residuals of 0, 100, -200 respectively.

A good regression model will have small residuals (errors).

The following figure shows a graphical presentation of the residuals.



Coefficient of determination (R^2):

The coefficient of determination, denoted by **R^2**, is the proportion of the variability in the outcome (response) variable that can be explained by the explanatory variable through their linear relationship.

- The Pearson's correlation coefficient between two variables X and Y may be used in simple linear regression as a descriptive statistic to measure the strength of the linear relationship between two variables.
- However, a more meaningful descriptive statistic that may be used to assess the goodness-of-fit of the linear regression model is obtained by squaring the Pearson correlation (r) to get the **coefficient of determination (R^2)**.
- This value is expressed in terms of **percentage** so that we may interpret the value to be the percentage of variability in the response variable that is explained by the explanatory variable through the model.

Although the term "explained" may seem to imply causality, we clarify that the relationship between the variables may not be causal.

$$\mathbf{0 \leq R^2 \leq 1.}$$

• If a model has perfect predictability, then $R^2 = 1$.

• If a model has no predictive capability, then $R^2 = 0$

Adjusted R Squared (R^2adj): Adjusted R Squared (R^2adj) is the percentage of variation explained by the model, adjusted for the sample size and the number of coefficients in the regression model (used in the multiple linear regression instead of the R^2).

Example:

In the previous example of the relationship between waist circumference (X) and BMI (Y), the computed Pearson's correlation is 0.91.

Squaring it to obtain the coefficient of determination, $R^2 = 0.83$.

Interpretation: 83% of the variability in the BMI can be explained by the waist circumference through this model.

Online simple linear regression calculators

https://www.socscistatistics.com/tests/regression/default.aspx

another calculator:

https://stats.blue/Stats_Suite/correlation_regression_calculator.html

Checking the model fit:

How to know that our model is a good one that can predict our outcome?

1- Checking the R^2 (or the adjusted R^2 in multiple regression).

The higher the R Square, the better the predictive power of the model.

2- Check the significance of the ANOVA model (part of the regression output).

A significant p-value means that the model is significant (has a good fit).

3- Checking the model assumptions.

Assumptions:

Assumptions of the simple linear regression that should be satisfied for the model to have a good fit:

1. **The relationship is linear**

There should be a linear relationship between the two variables. As with correlation, it is important to plot the data before doing a regression modelling to check that the relationship is linear. This is done using a scatterplot.

2. **There is no significant outliers**

The presence of outliers can have a negative effect on the regression model reducing its predictive ability. An outlier is a point on a scatterplot that is far away from the regression line.

3. **The observations are independent**

There is no more than one pair of observations on each individual.

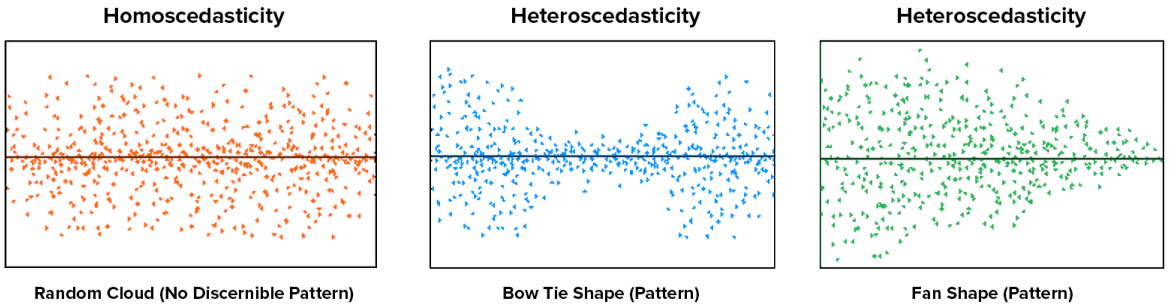
4. **The residuals are approximately normally distributed**

Residuals (errors) should be normally distributed. This can be tested using a histogram or doing a normal plot of the residuals.

5. **Homoscedasticity (The variance of the outcome y is constant line of best fit)**

This can be checked from the scatterplot where we plot the residuals against the predicted values to see if the spread of the residuals varies across the range of the predictor (non-constant variance).

The following figure shows three examples: one that meets this assumption (called homoscedasticity), and two that do not meet the assumption (called heteroscedasticity).



There are some advanced techniques for checking the assumptions but are not covered in this book.

Don't extrapolate beyond the scope of the model.

This means that the model should not be used to predict the outcome variable if the predictor variable is outside the range of the values used for generating the model.

For example: if we used the data of employees that have 0-10 years of experience to make a model that predicts salary, this model should not be used to predict the salary of an employee with 15 years of experience.

This is more obvious for the BMI example. Just think of using a very small waist circumference (of a child) to predict BMI.

Multiple linear regression

It is similar to simple linear regression with the exception that we have more than one predictor variable.

Types of variables:

- Dependent variable / outcome / y **(should be numeric variable), one variable**
- Independent variable / predictor / x **(can be numeric, ordinal, categorical), multiple variables**

It generates a regression equation:

$$y = b_0 + b_1 x_1 + b_2 x_2 + \ldots$$

y is the dependent variable

$\mathbf{x_1}$ and $\mathbf{x_2}$ are the independent variables

$\mathbf{b_0}$ is the intercept coefficient

$\mathbf{b_1}$ and $\mathbf{b_2}$ are the slope coefficients

Example:

In the simple linear regression, we used the following equation to estimate the BMI using waist circumference.

BMI = -8.5 + 0.38 (waist circumference)

We can add other predictive variables to this equation, let us start with **adding age**. The regression equation changes to:

BMI = -7.53 + 0.39 (waist circumference) -0.05(age)

The interpretation is similar to the simple linear regression.

For every unit increase in the waist circumference (in cm), there is 0.39 units increase in the mean BMI while controlling for age.

For every unit increase in the age (in years), there is 0.05 units decrease in the mean BMI while controlling for waist circumference.

Interpretation of the intercept, in this case, is not possible as the waist circumference can't be zero.

The predicted BMI of a patient having a waist circumference of 110 cm and age of 50 years is given by: BMI = -7.53 + 0.39 (110) -0.05(50) =32.9

What if we add another variable to the equation? let it be the **gender** (coded as 1 for males and 2 for females). Now the regression equation changes to:

BMI = -10.99 + 0.40 (waist circumference) -0.05(age) + 2.13(gender)

The interpretation of the constant, the coefficients of waist circumference and age are the same as before. But what about the coefficient of gender?

Gender is a **binary categorical variable** that has two levels only, and this coefficient represents the difference between the two levels). One of the groups is a reference category (here, the male category is the reference one), and the coefficient represents the difference between the other category (females) and the reference one (males).

Let's consider predicting the BMI for two patients, one is having a waist circumference of 110 cm and age of 50 years, and the other is a female with the same sex and waist circumference.

For the male patient: BMI = -10.99 + 0.40 (110) -0.05(50) + 2.13(1) = 32.64

For the female patient: BMI = -10.99 + 0.40 (110) -0.05(50) + 2.13(2) = 34.77

We notice that the only difference is that the female patient has a BMI that is 2.13 higher than the male patient of the same waist circumference and age, this value is the coefficient of gender.

The interpretation of the gender coefficient is: the mean BMI for female patients is 2.13 higher than the mean BMI of male patients controlling for waist circumference and age.

If we have a predictor categorical variable with multiple categories (for example socioeconomic status; low, intermediate, and high). One of the categories will be chosen as a reference category (low status for example), and each of the other two categories will have a coefficient that represents the difference between this category and the chosen reference category.

Online calculator:

Regression techniques are usually complicated and need statistical software, however, some simple online calculators are available with limited options:

Two predictor variables:
https://stats.blue/Stats_Suite/multiple_linear_regression_calculator.html

Up to four predictor variables:

http://home.ubalt.edu/NTSBARSH/Business-stat/otherapplets/MultRgression.htm

Checking the model fit:

How to know that our model is a good one that can predict our outcome?

- Checking the value of the adjusted R^2.
  The higher the Adjusted R Square, the better the predictive power of the model.
- Check the significance of the ANOVA model (part of the regression output).
  A significant p-value means that the model is significant (has a good fit).
- Checking the model assumptions.

Assumptions:

Assumptions of the multiple linear regression are **the same** as those of the simple linear regression with the addition of:

- **There is no multicollinearity:**

Multicollinearity means that two or more of the independent variables are highly correlated with each other. We can check for the presence of multicollinearity by checking the correlation coefficients between all predictor variables.

There are some advanced techniques for checking the assumptions but are not covered in this book.

What to report from the regression output?

1- The constant (in the case of using the model for a predictive purpose)

2- The coefficients (for the interpretation)

3- The p-value for each predictor (for the significance)

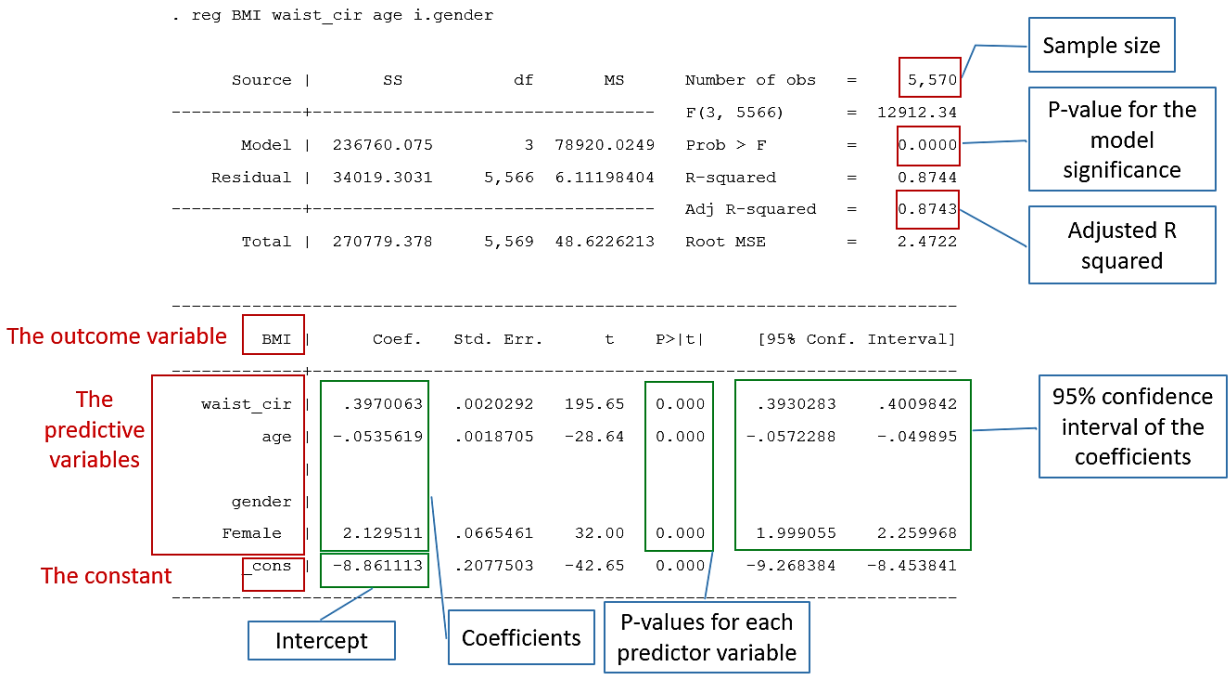
4- The 95% CI of the coefficients (not significant if contains zero)

5- The model fit (adjusted R squared, p-value) and model diagnostics (for the assumptions) are usually reported for predictive models.

For example:

The following is a Stata output of a multiple regression model to predict BMI using waist circumference, age, and gender.

(different programs produce different arrangements of the output)



We present the output in the following way:

Variables	Coefficients	P-value	95% CI of the coefficients
Waist circumference	0.397	<0.001	0.393, 0.401
Age	-0.054	<0.001	-0.057, -0.050
Gender			
Male	Reference		
Female	2.130	<0.001	1.999, 2.260
Constant	-8.861	<0.001	-9.268, -8.454

Simple logistic regression

A test of association that is similar to the simple linear regression but the outcome variable must be binary (dichotomous).

Types of variables:

Dependent variable / outcome / y **(must be binary variable)**

Independent variable / predictor / x **(can be numeric, ordinal, categorical)**

Generates regression equation (rarely used in medical practice)

- It is common in medicine and epidemiological studies as in the assessment of possible risk factors for a specific disease or complication (the outcome variable is binary as: disease/no disease, complication/no complication, recurrence/no recurrence, etc..)
- The most common use is to study the effect of several predictor variables on a binary outcome while adjusting for all other variables in the model.
- The equation from the logistic regression is not straightforward or easy to understand as in the case of linear regression. It uses logarithmic transformation.

Here the equation looks like:

$$\log\left(\frac{p}{1-p}\right) = b_0 + b_1 x$$

p is the probability of occurrence of the outcome variable (e.g. the disease). In other words, it is the proportion who has the outcome.

$1 - p$ is the probability that the outcome variable doesn't occur (no disease). In other words, it is the proportion who does not have the outcome.

b_0 is the intercept.

b_1 is the regression coefficient for the variable **x**. The value of this coefficient is back-transformed from the log scale to the natural scale to give the odds ratio.

In the logistic regression, we have to be cautious when reading the output and make sure if we are dealing with the **coefficient** (b), or **the exponential of the coefficient** exp(b) which is the odds ratio. It is an indicator of the change in odds resulting from a unit change in the predictor.

- If there is no association between the predictor and the outcome, the value of the coefficient (b) will be 0, and the value of the exp(b) will be 1.

Examples:

Predictor continuous variable:

If we study the association between **waist circumference** (continuous variable) and having diabetes (binary variable), we may have the following logistic regression result:

Predictor variable	Odds ratio	95% CI of OR	P-value
Waist circumference	1.04	1.03, 1.05	<0.001

The interpretation of the OR value is as follows:

For each unit increase in waist circumference (1cm), the odds of being diabetic increases multiplicatively by 1.04.

Note that if the waist circumference increases by 3 units (3cm), the odds of being diabetic increases by 1.04 X 1.04 X 1.04. (It does not increase by 1.04 X3).

Predictor binary variable:

If we study the association between **hypertension** (binary variable) and having diabetes (binary variable), we may have the following logistic regression result:

Predictor variable	Odds ratio	95% CI of OR	P-value
Hypertension	2.30	1.90, 2.78	<0.001

The hypertension variable is coded 0=no, 1=yes, and the odds ratio is the odds of the outcome (diabetes) in the 'yes' group divided by the odds of the outcome in the 'no' group (yes/no).

For patients with hypertension, the odds of having diabetes is 2.3 times the odds of having diabetes among patients who don't have hypertension.

Predictor categorical variable:

If we study the association between **smoking status** (categorical variable) and having bladder cancer (binary variable), we may have the following logistic regression result:

Predictor variable	Odds ratio	95% CI of OR
Smoking		
Never smoker	1	
Occasional smoker	1.5	0.9, 2.6
Former smoker	2.3	1.9, 2.8
Current smoker	5.2	4.0, 6.6

For the smoking variable, the "never smokers" group is considered the reference category. All other smoking categories are compared to this group.

For occasional smokers, the odds of having bladder cancer is 1.5 times the odds of having bladder cancer among never smokers (this relationship is not statistically significant as the confidence interval is containing 1).

For former smokers, the odds of having bladder cancer is 2.3 times the odds of having bladder cancer among never smokers. For current smokers, the odds of having bladder cancer is 5.2 times the odds of having bladder cancer among never smokers. There is an association between being a former smoker or a current smoker and having bladder cancer.

Even though the p-values are not reported in the table above, it is still clear which relationship is significant and which is not, based on the reported CI.

If the confidence interval is containing 1, there is no statistically significant association.

Odds ratio interpretation

Interpretation of the OR from the logistic regression for a continuous variable:

If the OR value is greater than 1: as the predictor increases, the odds of the outcome occurring increase.

If the OR value is less than 1: as the predictor increases, the odds of the outcome occurring decrease.

If the OR value is 1: no change (no association).

Interpretation of the OR from the logistic regression for a binary variable:

The OR compares the odds of occurrence of the outcome in the higher coded group (1) to the odds of the lower coded group (0).

If the OR value is greater than 1: the odds of the outcome occurring are higher in the higher coded group (coded as 1).

If the OR value is less than 1: the odds of the outcome occurring are lower in the higher coded group (coded as 1).

If the OR value is 1: no association.

It is always important to recognize the reference category. If the sex is coded 0 for males and 1 for females and the resulting OR=1.5. This means that the odds of having the outcome in females is 1.5 times that of males.

If the coding is reversed, 0 for females and 1 for males, the resulting OR is 0.67. This means that the odds of having the outcome in males is 0.67 that of the females. The two results are the same, the difference is only which one is used as a reference group.

Interpretation of the OR from the logistic regression for categorical variable:

The OR compares the odds of occurrence of the outcome in each category compared to the reference category. The number of reported odds ratios is equal to the number of categories -1. Sometimes, OR of 1 is written in the row of the reference category.

If the OR value is greater than 1: the odds of the outcome occurring are higher in this category as compared to the reference category.

If the OR value is less than 1: the odds of the outcome occurring are lower in this category as compared to the reference category.

If the OR value is 1: no difference from the reference category.

Differences from linear regression:

- The outcome variable in the linear regression is **numeric (continuous)**, while in the logistic regression the outcome is **binary**.
- In the linear regression, we interpret the **coefficient (b)**, while in the logistic regression we interpret the exponential of the coefficient **exp(b)**, which represents the change in the odds ratio (OR)
- **Interpretation of 95% CI:**

In linear regression, if the **95% CI of the coefficient crosses 0**, the result is statistically non-significant (true value may be 0, indicating no relationship)

In logistic regression, if the **95% CI of the OR crosses 1**, the result is statistically non-significant (true value may be 1, indicating no difference/change)

Example:

Linear regression:

b=1.08 (95% CI 0.77 to 1.40) = statistically significant

b=1.08 (95% CI -0.98 to 1.36) = statistically non-significant

Logistic regression:

Exp(b) = OR = 0.60 (95% CI 0.80 to 0.98) = statistically significant

Exp(b) = OR = 0.80 (95% CI 0.89 to 5.55) = statistically non-significant

Online logistic regression calculator :

https://stats.blue/Stats_Suite/logistic_regression_calculator.html

Multiple logistic regression

It is similar to the simple logistic regression with the exception that we have more than one predictor variable.

Types of variables:

- Dependent variable / outcome / y **(should be a binary variable), one variable**
- Independent variable / predictor / x **(can be numeric, ordinal, categorical), multiple variables**

The regression equation:

$$\log\left(\frac{p}{1-p}\right) = b_0 + b_1 x_1 + b_2 x_2 + \ldots$$

p is the probability of occurrence of the outcome variable (e.g. the disease). In other words, it is the proportion who has the outcome.

$1 - p$ is the probability that the outcome variable doesn't occur (no disease). In other words, it is the proportion who does not have the outcome.

b_0 is the intercept.

b_1, b_2, are the regression coefficients for the variables x_1, x_2. The values of those coefficients are back-transformed from the log scale to the natural scale to give the odds ratios. Odds ratios are interpreted in the same way as in the simple logistic regression.

Crude and adjusted odds ratios

- The odds ratios resulting from the simple logistic regression are called **crude** or **unadjusted odds ratios** meaning that this OR measures the association between the two variables without adjusting (controlling) for any other variables.
- The odds ratios resulting from the multiple logistic regression are called **adjusted odds ratios** meaning that this OR measures the association between the two variables while adjusting (controlling) for other variables in the model.
- Sometimes, both the unadjusted and the adjusted ORs are presented. This is useful to show how estimates change after adjustment. We have to be cautious if the OR changes greatly after adjustment as this might indicate the presence of confounding factors or effect modifiers.
- When reporting the crude and adjusted ORs, we have to report the 95% CI for the adjusted ORs, but we might or might not report the 95% CI for the crude ORs.

Example:

Suppose we are studying the effect of Age, BMI, hypertension on having diabetes in a group of patients:

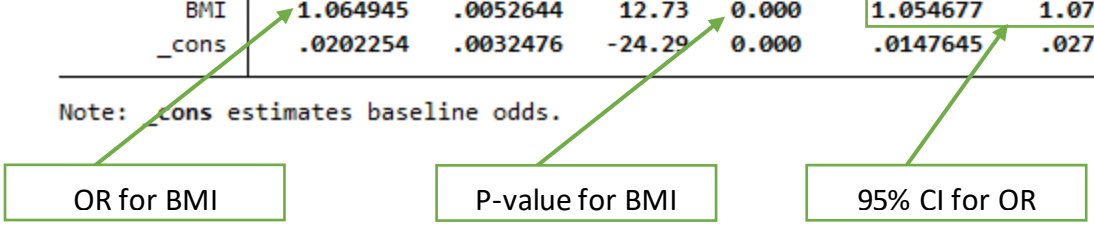
Here is the Stata output for :

1- Simple logistic regression for the **BMI** giving unadjusted OR

```
. logistic diabetes BMI

Logistic regression                             Number of obs     =      5,844
                                                LR chi2(1)        =     156.34
                                                Prob > chi2       =     0.0000
Log likelihood = -2055.5954                     Pseudo R2         =     0.0366

------------------------------------------------------------------------------
    diabetes | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         BMI |   1.064945   .0052644    12.73   0.000     1.054677    1.075313
       _cons |   .0202254   .0032476   -24.29   0.000     .0147645    .0277059
------------------------------------------------------------------------------
Note: _cons estimates baseline odds.
```



2- Simple logistic regression for the **age** giving unadjusted OR

```
. logistic diabetes age

Logistic regression                             Number of obs     =      6,110
                                                LR chi2(1)        =     481.04
                                                Prob > chi2       =     0.0000
Log likelihood = -1987.0074                     Pseudo R2         =     0.1080

------------------------------------------------------------------------------
    diabetes | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   1.052106   .0026665    20.04   0.000     1.046893    1.057345
       _cons |   .0088304   .0013759   -30.35   0.000     .0065065    .0119843
------------------------------------------------------------------------------
Note: _cons estimates baseline odds.
```

3- Simple logistic regression for **hypertension** giving unadjusted OR

```
. logistic diabetes i.hypertension

Logistic regression                             Number of obs     =      5,423
                                                LR chi2(1)        =      65.75
                                                Prob > chi2       =     0.0000
Log likelihood = -1931.3952                     Pseudo R2         =     0.0167

------------------------------------------------------------------------------
    diabetes | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
hypertension |
         Yes |   2.297452   .2250656     8.49   0.000     1.896095    2.783767
       _cons |   .1126795   .0055238   -44.53   0.000     .1023568    .1240432
------------------------------------------------------------------------------
Note: _cons estimates baseline odds.
```

4- **Multiple** logistic regression for the three variables giving adjusted ORs

```
. logistic diabetes BMI age i.hypertension

Logistic regression                             Number of obs     =      5,364
                                                LR chi2(3)        =     576.34
                                                Prob > chi2       =     0.0000
Log likelihood = -1648.5561                     Pseudo R2         =     0.1488

------------------------------------------------------------------------------
    diabetes | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         BMI |   1.076396   .006439    12.31   0.000     1.063849     1.08909
         age |   1.056309   .0032016   18.07   0.000     1.050053    1.062603
             |
hypertension |
         Yes |   1.075347   .115927     0.67   0.500     .8705332    1.328348
       _cons |   .0007521   .000212   -25.52   0.000     .0004328    .0013068
------------------------------------------------------------------------------
Note: _cons estimates baseline odds.
```

Adjusted ORs

What to report from the regression output?

1- The ORs (unadjusted) (from the simple regression)

2- The ORs (adjusted) (from the multiple regression)

3- The 95% CI of the adjusted OR (not significant if contains 1)

4- The p-value (for the significance of association)

Predictor variable	Unadjusted OR	Adjusted OR	95% CI of OR	P-value
BMI	1.06	1.08	1.06, 1.09	<0.001
Age	1.05	1.06	1.05, 1.06	<0.001
Hypertension	2.30	1.08	0.87, 1.33	0.500

From the table above, we can see that only Age and BMI are associated with diabetes.

In simple regression, hypertension was significantly associated with diabetes with OR=2.3, but in the multiple regression the relationship is no more significant and the OR dropped to 1.08.

This result suggests that part of the observed hypertension effect when examined alone was in fact due to the other variables in the model (BMI and age).

Diagnostic tests: Sensitivity, specificity, and predictive values

If a group of researchers comes up with a new diagnostic test (e.g. blood test) to diagnose the presence of a certain disease (e.g. presence of cancer), they will have to run an experiment to see how good is this new diagnostic test (which may be cheaper, easier, less invasive than the standard test).

We need to compare this **new diagnostic test** to the **gold standard test** that provides a definitive diagnosis of a particular condition (it may be a histopathology exam in the condition of cancer).
So, we apply this new test and the gold standard test (true diagnosis) to a group of individuals who might have the disease or not.

Based on the results, we will have 4 groups:

A. **Positive** for the blood test and **positive** for the histopathology test

True positive = correctly identified

B. **Positive** for the blood test and **negative** for the histopathology test

False positive = incorrectly identified

C. **Negative** for the blood test and **positive** for the histopathology test

False negative = incorrectly rejected

D. **Negative** for the blood test and **negative** for the histopathology test

True negative = correctly rejected

Those results are presented in a table as follows:

		Based on the gold standard test	
		Disease present	Disease absent
The new diagnostic test	Test positive	True positive (a)	False positive (b)
	Test negative	False negative (c)	True negative (d)

After adding the totals:

		Based on the gold standard test		**Total**
		Disease present	Disease absent	
The new diagnostic test	Test positive	True positive (a)	False positive (b)	Total test positive **(a+b)**
	Test negative	False negative (c)	True negative (d)	Total test negative **(c+d)**
Total		Total diseased **(a+c)**	Total normal **(b+d)**	total population **(a+b+c+d)**

The new test (blood test) would be ideal if it gives exactly the same result of the gold standard test (the histopathology test). But this is not usually the case, there is a margin of error with some false positive and some false negative test results.

So, we use the sensitivity and specificity to describe the accuracy of a diagnostic test.

For example:
If 1000 individuals were exposed to the two tests and the result is summarized as follows:

		Based on the gold standard test	
		Disease present	Disease absent
The new diagnostic test	Test positive	**180** True positive **(a)**	**80** False positive **(b)**
	Test negative	**20** False negative **(c)**	**720** True negative **(d)**

a. True positive = 180
b. False positive = 80
c. False negative = 20
d. True negative = 720

✓ **Sensitivity:**
Sensitivity is the percentage of true positives,
i.e. the proportion of those who have the disease who are correctly identified by the test as positive.
In other words: the probability that a test result will be positive when the disease is present.

$$\textbf{Sensitivity} = \frac{a}{a+c} = \frac{\text{number of true positive (a)}}{\text{number of true positive (a)+number of false negative (c)}}$$

$$\text{Sensitivity} = \frac{180}{180+20} = 0.9 = 90\%$$

This 90% sensitivity means that if we are sure that 100 patients have the disease (based on the gold standard test), the new diagnostic test will be positive in 90 cases.

- 90% of people who have the target disease will test positive

✓ **Specificity:**
Specificity is the percentage of true negatives,
i.e. the proportion of those who don't have the disease who are correctly identified by the test as negative.
In other words: the probability that a test result will be negative when the disease is absent.

$$\textbf{Specificity} = \frac{d}{b+d} = \frac{\text{number of true negative (d)}}{\text{number of false positive (b)+number of true negative (d)}}$$

Specificity $= \frac{720}{80+720} = 0.9 = 90\%$

This 90% specificity means that if we are sure that 100 individuals don't have the disease (based on the gold standard test), the new diagnostic test will be negative in 90 cases.

- 90% of people who do not have the target disease will test negative

Sensitivity and specificity are characteristics of the test. But the physician and the patient may have a different question: what is the chance that a person with a positive test truly has the disease?

In other words, if the subject is in the first row in the table above, what is the probability of being in cell A as compared to cell B?

We have here two new calculations:

✓ **Positive Predictive Value (PPV) =** $\frac{a}{a+b}$

Positive predictive value is the probability that when having a positive test result, that individual will truly have that specific disease.

✓ **Negative Predictive Value (NPV) =** $\frac{d}{c+d}$

Negative predictive value is the probability that when having a negative test result, that individual will truly be free of the disease.

(PPV) = $\frac{a}{a+b} = \frac{180}{180+80} = 0.69 = 69\%$

For those who test positive, 69% are having the disease

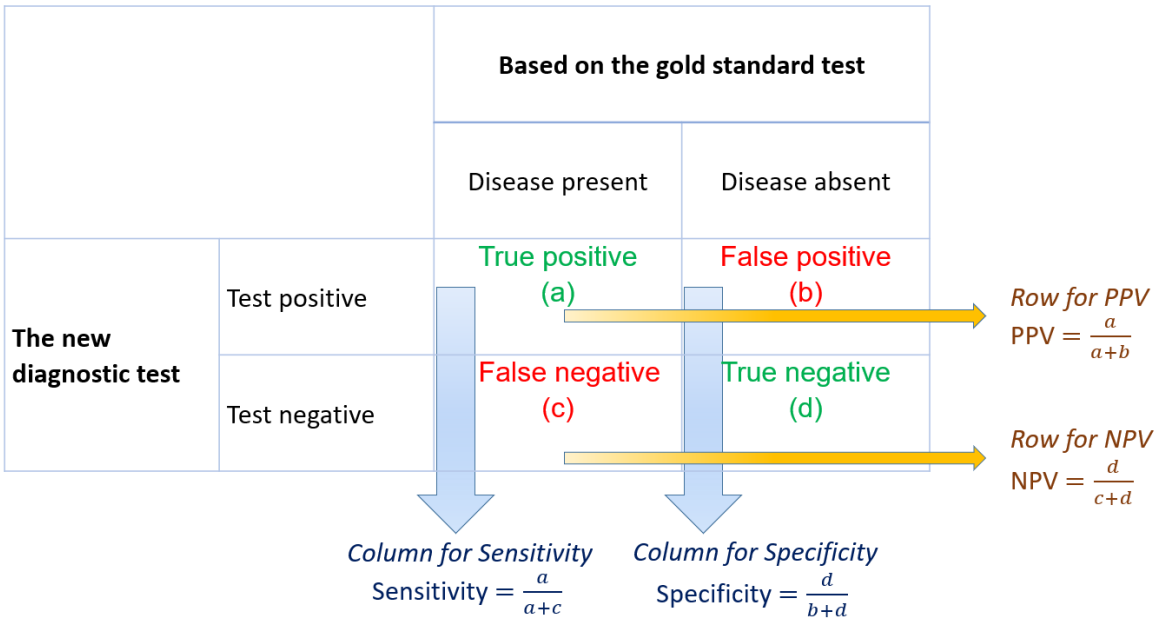
(NPV) = $\frac{d}{c+d} = \frac{720}{20+720} = 0.97 = 97\%$

For those who test negative, 97% are having the disease

Online sensitivity, specificity, PPV and NPV calculator :

https://www.medcalc.org/calc/diagnostic_test.php

A simple guide for calculating sensitivity, specificity, PPV, and NPV



For simplicity, we can remember it as:

Sensitivity is: true positive / diseased

Specificity is: true negative / non-diseased

PPV is: true positive / testing positive

NPV is: true negative / testing negative

What about the prevalence of the disease?

The prevalence of a disease is the percentage of the population who (truly) have this disease.

✓ **Prevalence =** $\frac{a+c}{a+b+c+d} = \frac{180+20}{180+80+20+720} = 0.2 = 20\%$

This 20% prevalence means that out of 100 individuals from this population, 20% are diseased.

Disease prevalence and sensitivity, specificity, PPV and NPV

- Sensitivity and specificity are characteristics of the test. The population prevalence does not affect the results.
- Positive and negative predictive values are influenced by the prevalence of the disease in the population.

As prevalence decreases, the NPV increases because there will be more true negatives for every false negative. This is because a false negative would mean that a person actually has the disease, which is unlikely because the disease is rare (low prevalence). Also, the PPV decreases because there will be more false positives for every true positive.

ROC curve

- Sometimes when developing a diagnostic or screening test, we are concerned about choosing the appropriate **cut-off value** for a numeric measurement.
- The receiver operating characteristic **(ROC) curve** is used to compare the sensitivity and specificity for all possible cut-offs.

To understand the idea of the importance of this cut-off value, let's consider the following **example**.

- A researcher is concerned about the diagnosis of depression among university students.
- To reach the diagnosis of depression, a psychiatrist needs a one-hour session with each student.
- What if we want to screen all the university students for the presence of depression? It is not feasible to make a psychiatric consultation for each student.
- Instead, we can use a **screening tool** (a questionnaire for example), that can be applied easily to a large number of students.
- We suggested a tool consisting of 21 questions, each is scored 0-3.
- The highest possible score is 63, and **the higher the score the higher the possibility of being depressed**.
- The question here is: **what is the cut-off point** we can use to refer a student to the psychiatrist?
- To determine this cut-off point, we use the **ROC curve**, and the following **experiment** is performed.

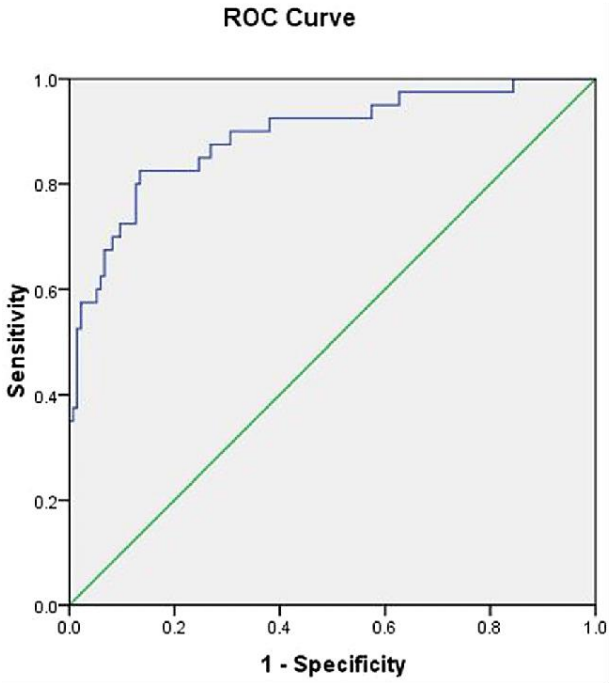
- We select a **sample** of the students (100 students, for example). Each is subjected to the **screening tool** (the questionnaire), and a **one-hour session** with the psychiatrist to decide if the student is depressed or not.
- The resulting data consists of **pairs**, one numeric variable **(test score)**, and one binary variable for the diagnosis **(diseased/not diseased)**.
- We plot a ROC curve to present the sensitivity and specificity of each possible cut-off point.
- We choose **a suitable cut-off point** based on its corresponding sensitivity and specificity.
- A cut-off point of 21 for example might be chosen. Any student who scores 21 or higher is referred to the psychiatrist for assessment.

The ROC curve has the following characteristics

The true positive rate (**Sensitivity**) is plotted on **Y-axis**.

The false positive rate (**1-Specificity**) is plotted on **X-axis**.

- **Each point** on the ROC curve represents a sensitivity/specificity pair corresponding to a particular possible cut-off point.
- Sensitivity and specificity are **inversely related**; if we change the cut-off for better sensitivity, this will reduce the specificity.
- We can choose the optimal cut-off point from this graph depending on the implications of false positive and false negative results, and the prevalence of the condition.



For example, when screening for a deadly disease that is curable, it may be desirable to accept more false positives (lower specificity) in return for fewer false negatives (higher sensitivity).

- A perfect test has a ROC curve that passes through the upper left corner (100% sensitivity, 100% specificity) and the closer the ROC curve is to the upper left corner, the higher the overall accuracy of the test.
- A diagnostic test can have two cut-offs: one to rule out the disease and another to rule in disease. The values in between are inconclusive.

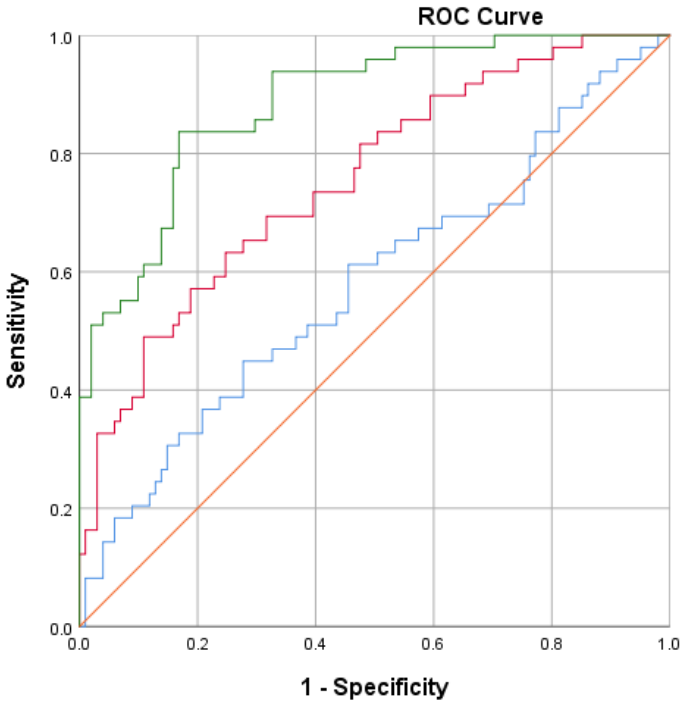
In the previous example, we might consider those below 21 as normal, those over 40 as depressed, and those between 21 and 40 as inconclusive and need further assessment.

Area Under the Curve (AUC)

- We can compare **the accuracy of two or more tests** by considering the **Area Under the Curve (AUC)**, the curve with higher AUC represents a better discriminating test.
- The AUC ranges **from 0.5 to 1**. A test that is perfect at discriminating between the disease outcomes has AUC = 1 and a non-discriminating test that performs just like chance has AUC = 0.5.
- The nearer the curve to the diagonal line (drawn at 45°), the poorer the test is.

Example:

Let's examine the following graph:



Area Under the Curve

Test Result Variable(s)	Area
BiochemA	.580
BiochemB	.755
BiochemC	.886

Those are three ROC curves for three different Biochemical tests for the detection of a specific condition.

It is obvious from the curve and from the value of AUC that test No. C gives the highest accuracy compared to the other two tests.

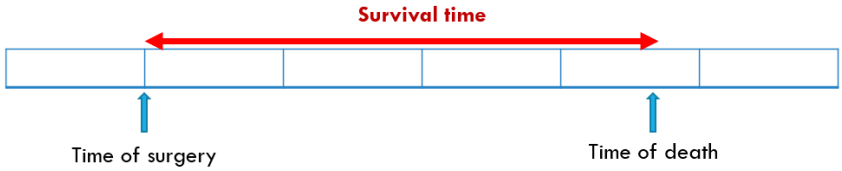
Survival analysis

Survival analysis is concerned with the time until an event occurs (**time to event**). This event is usually death, as survival after breast cancer, but can be any other event.
Examples:

- Time from operation to death
- Time from response till the recurrence of a tumour
- Time from operation to discharge from the hospital.

Survival time

- Survival times are calculated from baseline time to the endpoint.
- The baseline time reflects a natural 'starting point' for the study (e.g. time of surgery or diagnosis of a condition).
- Until the time that a patient reaches the endpoint of interest (event).



Objectives of survival analysis

- **To estimate the time to event for a group of individuals**, such as the time until the second heart attack for a group of myocardial infarction patients.
- **To compare the time to event between two or more groups**, such as comparing time to second heart-attack between male and female MI patients (or two treatment groups).
- **To calculate the survival probability at a certain time**, the probability that patients will survive for 1 or 5 years (after diagnosis of lung cancer).

Characteristics of survival data:

- Individuals do not enter the study at the same time.
- When the study ends, some individuals still haven't had the event yet.
- Other individuals drop out or get lost in the middle of the study, and all we know about them is the last time they were still 'free' of the event.

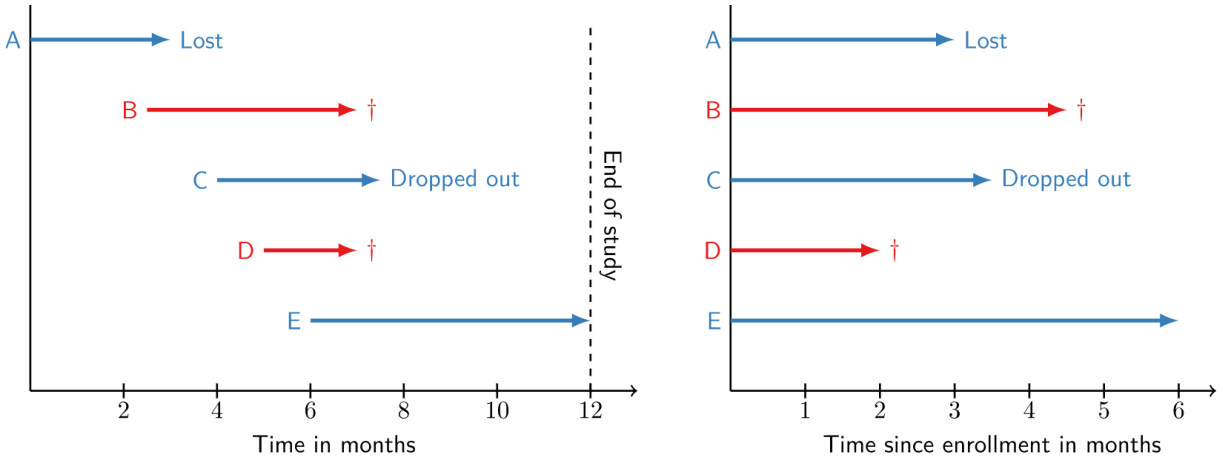
Survival analysis terms:

✓ **Time to event:** The time from entry into a study until a subject has a particular event (outcome).

✓ **Censoring (no event):** Subjects are said to be censored if they are **lost to follow up** or drop out of the study, or if the **study ends** before they die or have an outcome of interest. They are counted as alive or disease-free for the time they were enrolled in the study.

Example:

Consider a clinical study that has been carried out over 1 year as in the figure below.



Patient **A** was lost to follow-up after **3 months** with no event.

Patient **B** had an event **4.5 months** after enrollment.

Patient **C** withdrew from the study **3.5 months** after enrollment.

Patient **D** experienced an event **2 months** after enrollment.

Patient **E** did not have any event before the study ended (was followed up for **6 months**).

The exact time of an **event** could only be recorded for patients **B** and **D**; their records are ***uncensored***.

For the remaining patients, it is unknown whether they did or did not experience an event after the termination of the study. The only valid information that is available for patients **A**, **C**, and **E** is that they were **event-free** up to their last follow-up. Therefore, their records are ***censored***.

Notes:

- We can't just take the average time for the individuals as this will be an underestimation (due to the censored data).
- We can't just consider the percentage of occurrence of an event.

For example,

If we are comparing two chemotherapy treatments and at end of the study,

40% of patients in each of the two chemotherapy groups had died. (Exactly the same proportion)

What about the timing?

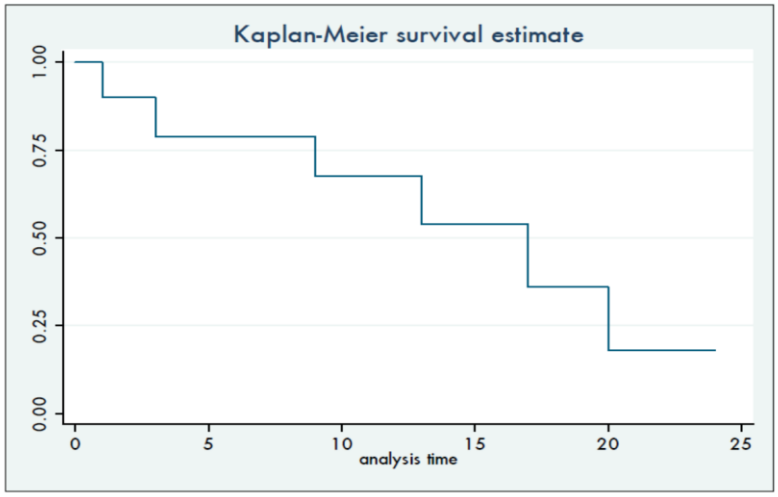
If in the first chemotherapy group, most of the 40% died within a year of starting the treatment; while in the second group, most of the 40% died between 5-6 years after starting treatment.

Then, the timing of events is very different between the two groups and is important.

Displaying survival data

A survival curve, usually calculated by the Kaplan–Meier method, displays the cumulative probability (the survival probability) of an individual remaining free of the event at any time after baseline.

The Kaplan–Meier curve will look like this:



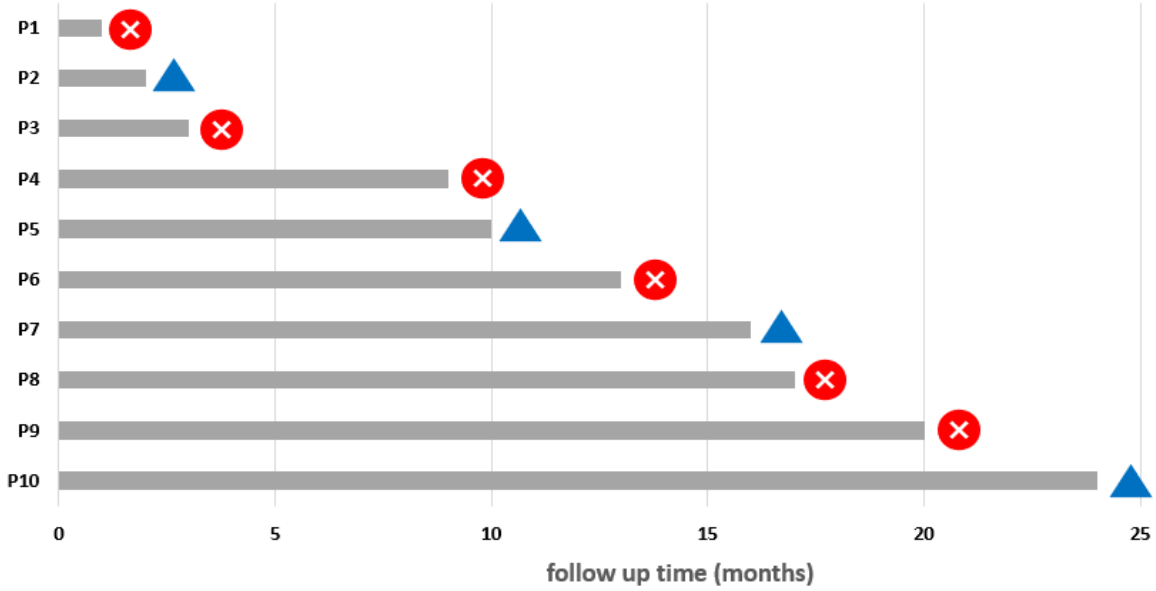
- The vertical axis shows the probability of surviving or the proportion of people surviving.
- The horizontal axis represents the time in months.

The curve moves down at the occurrence of every event.

How Kaplan–Meier curve is produced?

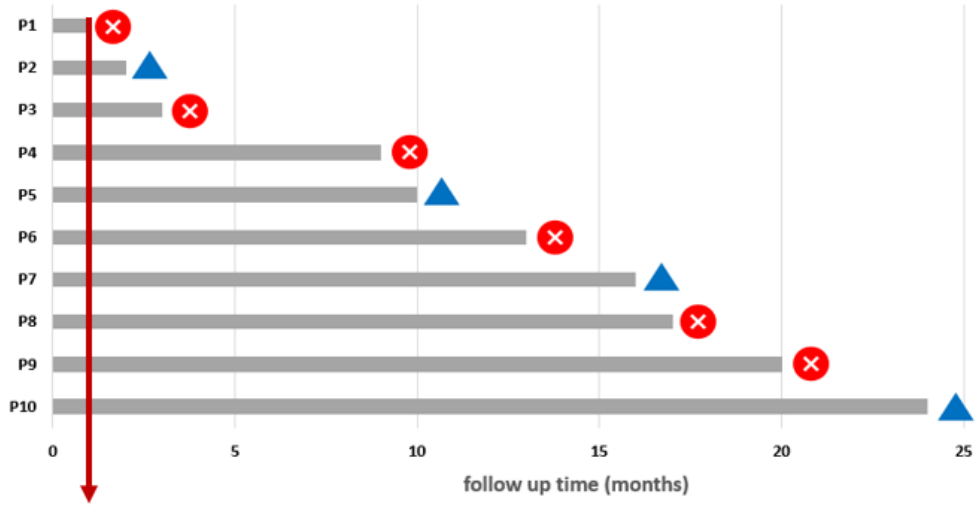
Here are survival data for 10 patients, 6 of them died ⊗ and 4 were censored ▲

1- Sort the survival times from shortest to longest



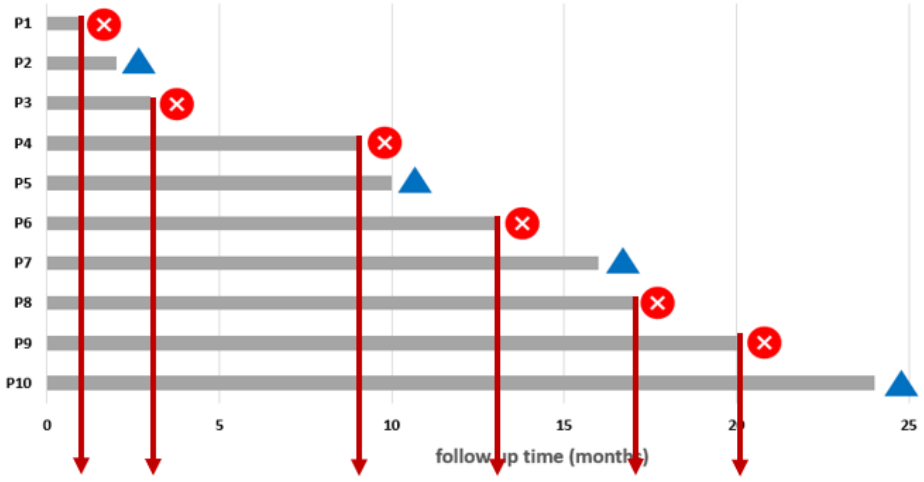
2- Calculate the conditional survival at each event

- At the beginning of the study, all patients are alive and the survival probability is 100%
- When the first event occurs 1 month after the beginning of follow up, one patient dies and 9 survive, so survival probability beyond 1 month is 9/10= 90%.



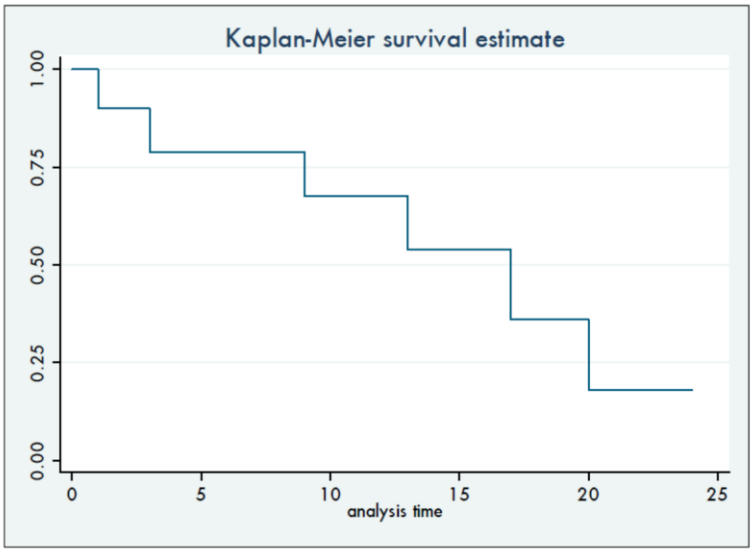
- One patient is lost to follow-up at 2 months of follow-up, but the survival probability is still the same (no events yet).

- The next event occurred at 3 months of follow-up. Here, one patient dies and 7 survive (notice that 2 patients are no longer available, one event and one censored), so survival probability beyond 3 months is 7/8= 87.5%.

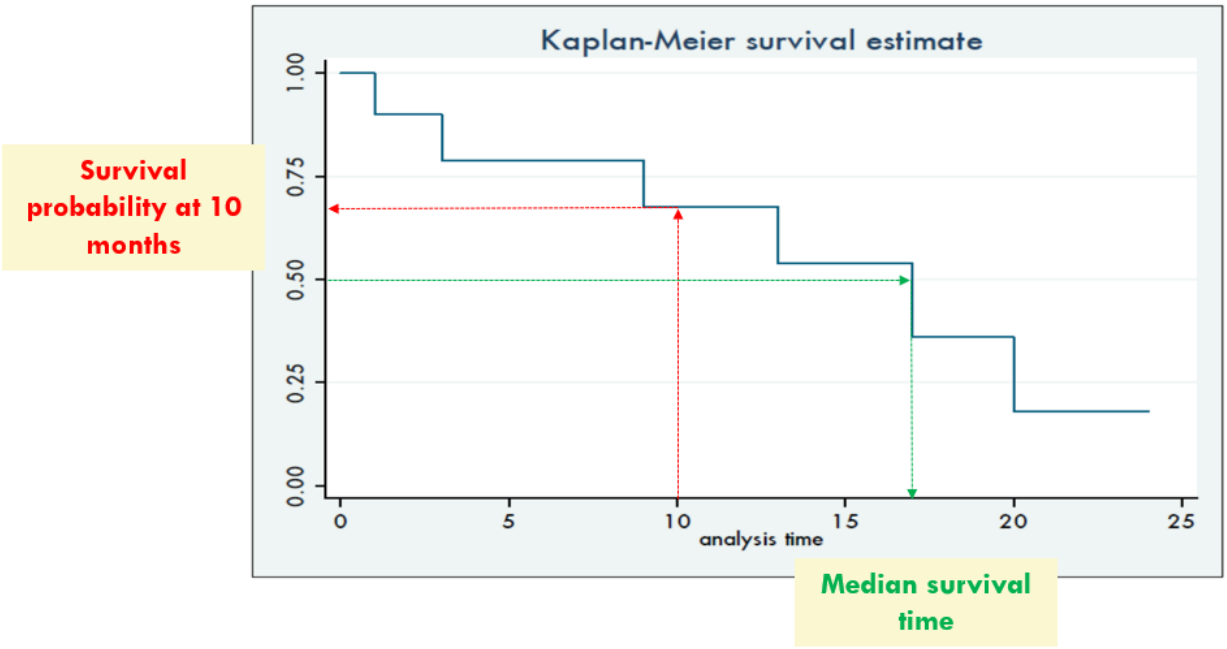


- We continue calculating the survival probability at each event the same way.

The Kaplan–Meier curve will look like this:

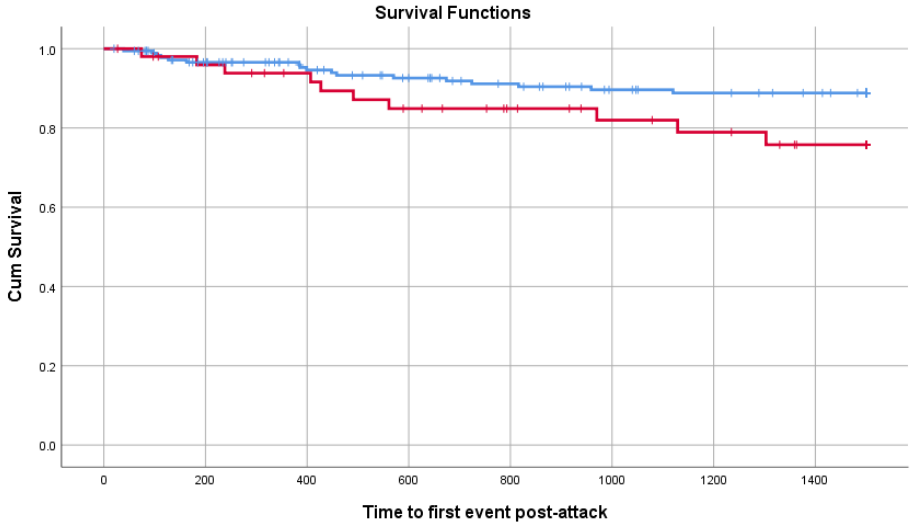


- The Kaplan-Meier Curve can be used to estimate the **probability of survival** at a specific time and can be used to estimate the **median survival time** which is the time at which half the patients are expected to be alive.



- Kaplan-Meier Curve can be used to **compare the survival in two groups**. If the curve goes down rapidly, the occurrence of the event is at a higher rate in this group.

The following curve compares the time to the occurrence of a second heart attack after the occurrence of the first in two groups; **smokers** (red curve) and **non-smokers** (Blue curve). Time here is presented in days.



The statistical test used to compare survival in the two groups is called the **log-rank test**, and the regression model used to assess factors affecting survival is called **cox regression**.

For more details regarding doing the survival analysis, Kaplan-Meier Curve, the log-rank test, and the cox regression on SPSS, you can join the udemy course prepared by Dr. Mohamed Elsherif:

Available on the following link:

https://www.udemy.com/course/survival-analysis-using-spss/?couponCode=ELSHERIF

Sample Size and Power Analysis

- For any study design, one of the most important questions is how many participants we should have in the study **(sample size)**.
- A study with a small sample size will not have enough power to answer the research question and will have a negative result and a true difference may be missed (due to the low power).
- On the other side, doing a study with more sample size than what is needed is a waste of time and resources and may be unethical as more people are exposed to dangers and side effects.
- So, it is important to calculate the sample size with a proper power capable of answering the research question.
- One of the wrong ideas about sample size calculation is using 60 patients for clinical trials (30 for the treatment and 30 for the control group), considering that this number is large enough.
- Some researchers tend to use the same sample size of a similar study, or just ask a supervisor how many participants should I include in my research and this is another mistake.
  Calculating the sample size has its roles.

Sample size calculation for clinical trials (RCTs)

If we are designing a randomized control trial of one treatment group and one control group (parallel study), then we need all the following pieces of information for our sample size calculation:

- **Power of the study:** is the ability of the study to prove that the new drug is better than the standard drug/placebo when in reality it is, or the ability of the study to detect the difference between the two groups when in reality there is a difference.
  - The power of the study is usually set at 80% or 90%.
  - The power of study means not committing type 2 error (not having a false negative result).

  So,

  Power = 1-β

- **Level of statistical significance:** which is called **α** or **type I error**.
  - This level is set at 0.05 in most trials, but it can be 0.01.
  - This number represents the cut-off level for the p-value, below which we will consider the p-value significant.
  - This number represents the probability of committing type 1 error.

- **Enrollment ratio:** it means the ratio of participants in the control group to the treatment group.

In most studies, it is 1:1, which means an equal size of the treatment and control groups.

- **Expected effect size** (the minimum clinically important difference)**:** this is the most important item that we have to define.
  - The expected effect size is the expected difference between the two groups, which is of clinical significance.
  - To reach the correct expected effect size, we have to answer some questions:

  1- What is the primary outcome (variable) that I am looking for?
    - Is it a change in the blood glucose level or blood pressure?
    - Is it the number of cured patients in each group?
    - Is it time to an event, as the time to discharge from the hospital or time till the recurrence of the tumor?

  2- What is the type of this outcome variable?
    - Is it a change in the blood glucose level or blood pressure? So we are comparing two means.
    - Is it the number of cured patients in each group? So we are comparing two proportions.
    - Is it time to an event, as the time to discharge from the hospital or time till the recurrence of the tumor? So we are comparing two median survival times or survival rates.

  3- What are the expected numbers from this study?
    - If we are comparing two drugs, the standard drug has a 70% cure rate, what is the expected cure rate of the new drug so that it is considered a good one? Is it 75%? 80%?
    - If the outcome variable is the change in glycated hemoglobin level or the change in blood pressure, we need to know the expected mean and standard deviation for this change in each of the groups.

  4- Where can I get those numbers?

  Those numbers can be estimated based on previous studies (literature review), a pilot study, or the opinion of experts.

All those information are inputs for the equations used for the calculation of the sample size.

Factors affecting sample size:

1- **Power of the study:** the higher the required power (smaller β), the larger the needed sample size.
2- **Level of statistical significance:** the lower the level of statistical significance (smaller α, lower probability of type I error), the larger the needed sample size.

3- **Expected effect size:** the lower the effect size (small difference between the two treatments), the larger the needed sample size.

Adjustment for loss of follow up

The calculated sample size should be adjusted for the expected loss of follow-up cases.

For example, if the calculated required sample size is 80 in total and it is expected that 20% of those recruited will not complete the study, then 100 patients should be recruited to ensure that 80 will complete it.

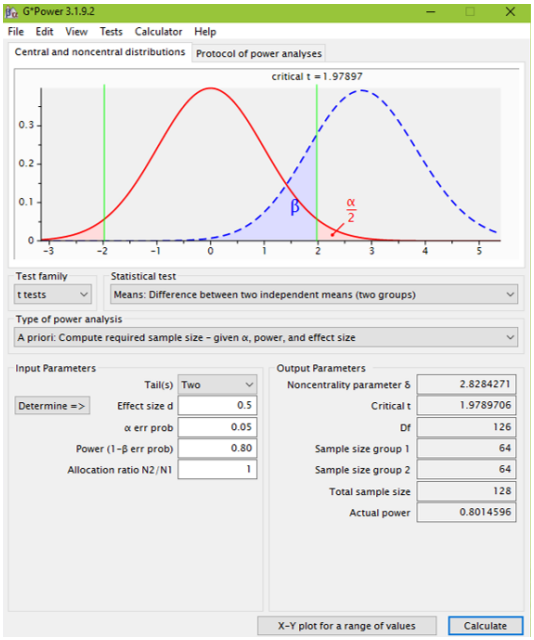
It is easily calculated by **dividing the calculated sample size by (1- % expected to be lost)**. If the calculated sample size is 150 and the expected loss to follow-up = 15%, we can calculate the number needed as: 150/0.85 = 176.5. This is rounded up to 177 individuals.

Sample size and power calculation tools:

Free software:

The following software programs are designed for sample size calculation and are freely available

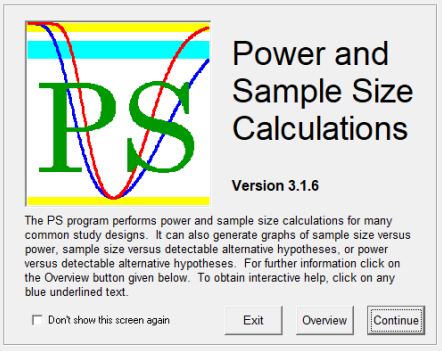
1- G*Power software



Available for download at:

https://www.psychologie.hhu.de/arbeitsgruppen/allgemeine-psychologie-und-arbeitspsychologie/gpower.html

2- PS: Power and Sample Size Calculations



Available for download at:

https://ps-power-and-sample-size-calculation.software.informer.com/3.1/

Stata software (not free)

It is a famous statistical analysis software that can also be used for sample size and power calculations.

Some user-friendly websites for sample size calculation (free):

- Online sample size calculator for means and proportions

https://clincalc.com/stats/samplesize.aspx

- Online sample size calculator for means

https://www.stat.ubc.ca/~rollin/stats/ssize/n2.html

- Online sample size calculator for proportions

https://www.stat.ubc.ca/~rollin/stats/ssize/b2.html

- Online sample size and power calculator for many conditions

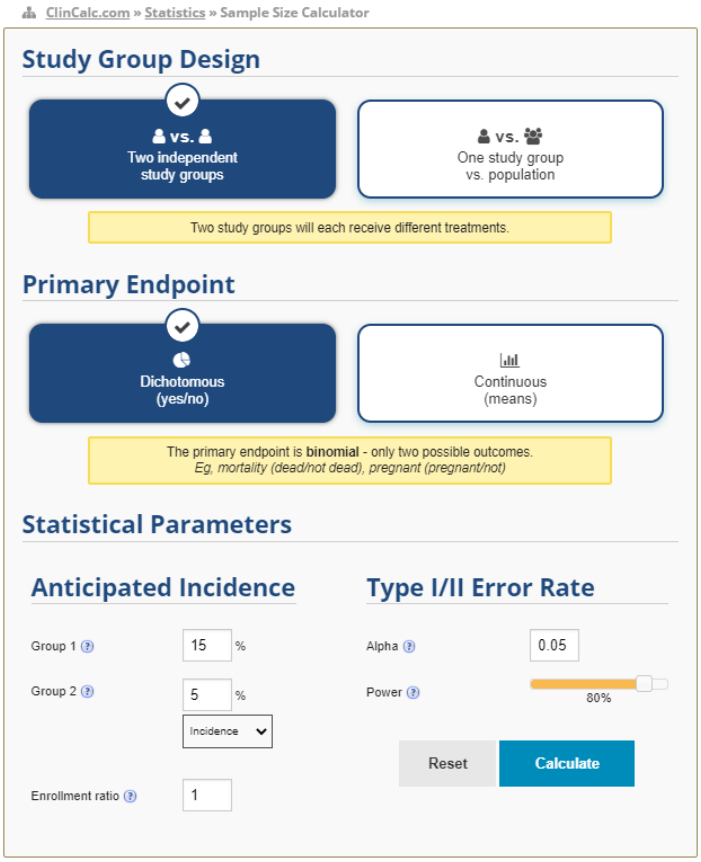
http://powerandsamplesize.com/

Example of a sample size calculation:

If we want to compare a new type of surgery to the current one. The proportion of patients who develop complications after the current surgery is 15%, and it is expected that the complication rate for the new surgery will be 5%. What is the needed sample size using a significance level of 0.05 and power of 80% and equal size for both groups?

If we use the website: https://clincalc.com/stats/samplesize.aspx

The input is:



And the result is:

RESULTS

Dichotomous Endpoint, Two Independent Sample Study

Sample Size	
Group 1	140
Group 2	140
Total	280

Study Parameters	
Incidence, group 1	15%
Incidence, group 2	5%
Alpha	0.05
Beta	0.2
Power	0.8

So, 140 patients are needed in each group.

What if we expect 10% of patients to be lost for follow-up?

We will adjust the number by dividing it by (1- % expected to be lost)= 140/0.9= 155.6

So, 156 patients are needed in each group.

Sample size calculation tools for observational studies

Many software programs and some online websites can be used:

1- Epi Info™

A software program freely available from CDC in formats suitable for windows and mobile phones. It is available for download here:

https://www.cdc.gov/epiinfo/support/downloads.html

2- Stata

This is statistical software that can be used for sample size calculation.

3- Online calculators:

Different calculators are available with varying options and requirements. As in the following examples:

https://riskcalc.org/samplesize/
https://epitools.ausvet.com.au/samplesize

It is always recommended to consult a statistician for sample size calculation.

Post-study power analysis

Sometimes clinical trials do not go as designed. There might be more dropouts, fewer people recruited, the actual effect size is smaller or bigger than what we used in the sample size calculation.

So, power analysis is necessary to make sure that our study is still powered enough to answer the research question.

We can estimate the power of a study to check whether the negative result (non-significant result) was really a true negative one, or maybe just due to the low power of the study resulting from the small sample size.

The **same equations** for sample size calculations are used to calculate the power. Some websites or statistical programs provide the power calculator in the same section with the sample size calculator.

Example of power analysis:

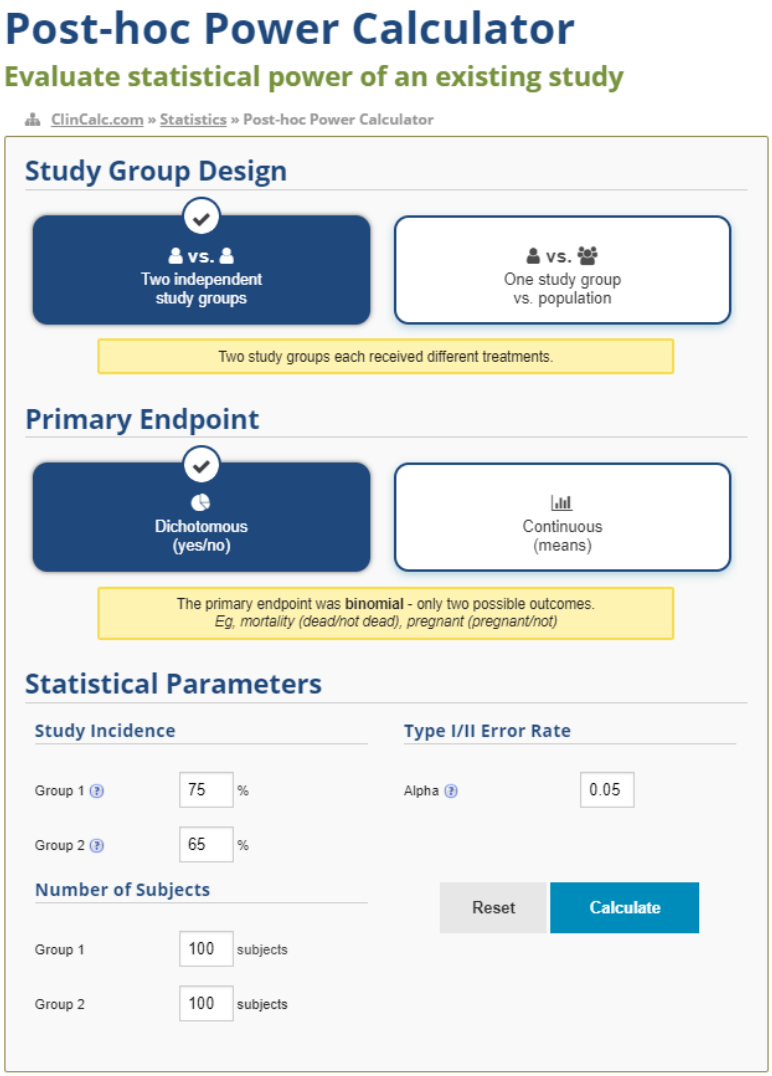
A study comparing the percentage of response between two drugs: A (new drug) and B (standard drug) using a parallel arms study reported the below results:

	Drug A	Drug B
Responders	75	65
Non-responders	25	35

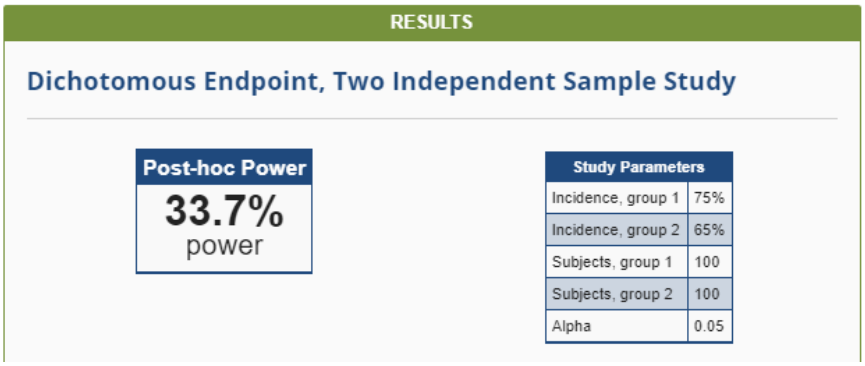
What is the current power of this study?

If we use the website: https://clincalc.com/Stats/Power.aspx

The input is:



And the result:



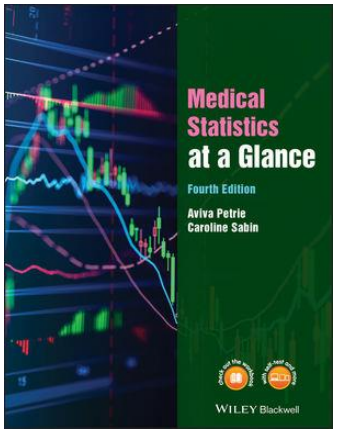
So, the power of the study is 33.7%.

If we conduct a chi-square test and get a non-significant result, this is probably a type II error due to the small sample size.

Recommended resources for further readings

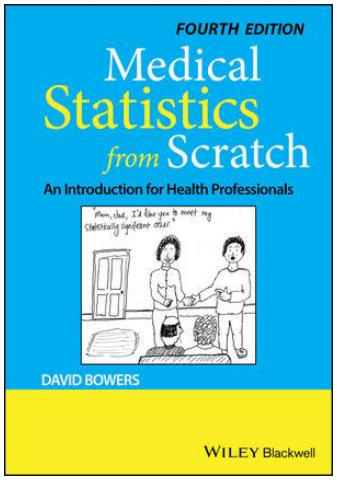
The following resources are recommended as next step readings based on our experience. We relied on those resources and tens of other resources to come up with the material of this book.

1- Medical statistics at a glance, 4th edition 2019. 211 pages



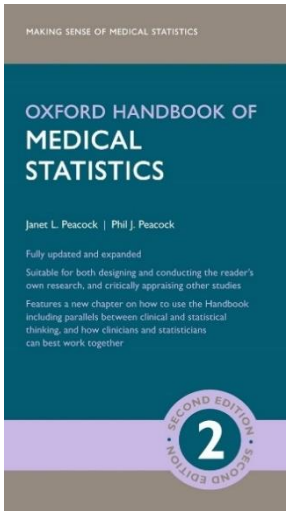
https://www.amazon.com/Medical-Statistics-Glance-Aviva-Petrie/dp/1119167817

2- Medical Statistics from Scratch: An Introduction for Health Professionals, 4th edition 2020, 496 pages



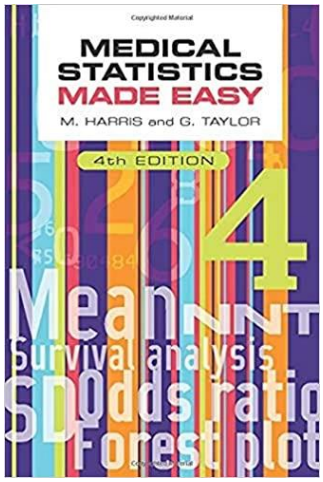
https://www.amazon.com/Medical-Statistics-Scratch-Introduction-Professionals/dp/1119523885

3- Oxford handbook for medical statistics, 2nd edition 2020, 640 pages



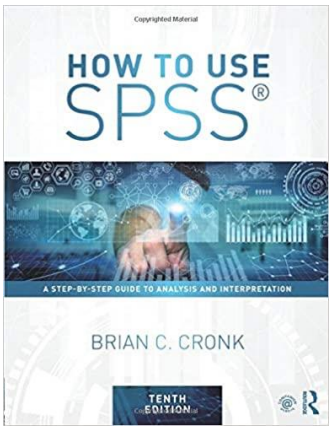
https://www.amazon.com/Oxford-Handbook-Medical-Statistics-Handbooks/dp/0198743580

4- Medical statistics made easy, 4th edition 2020, 140 pages



https://www.amazon.com/Medical-Statistics-Made-Easy-4th-dp-1911510630/dp/1911510630/

5- How to Use SPSS®: A Step-By-Step Guide to Analysis and Interpretation, 11th Edition 2019, 228 pages



https://www.amazon.com/How-SPSS-C2-AE-Step-Step-Interpretation-dp-0367355698/dp/0367355698