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.

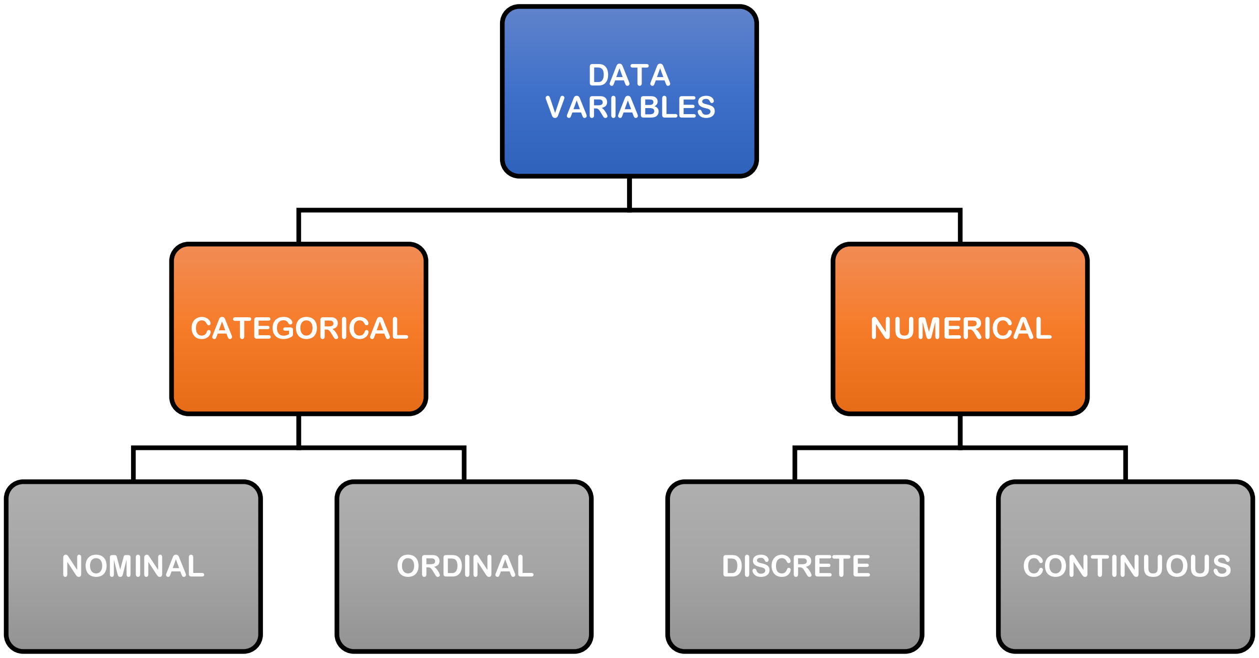
Examples for data variables are sex, age, weight, marital status, satisfaction rate, etc.

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:

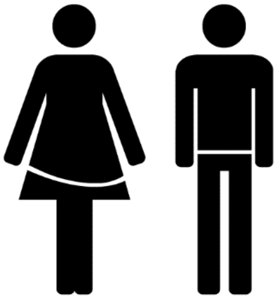
They are also known as qualitative or nominal data; they have NO unit of measurement.

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

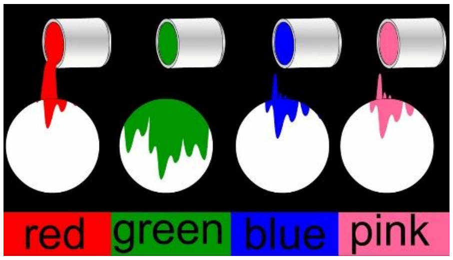
Examples:

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.

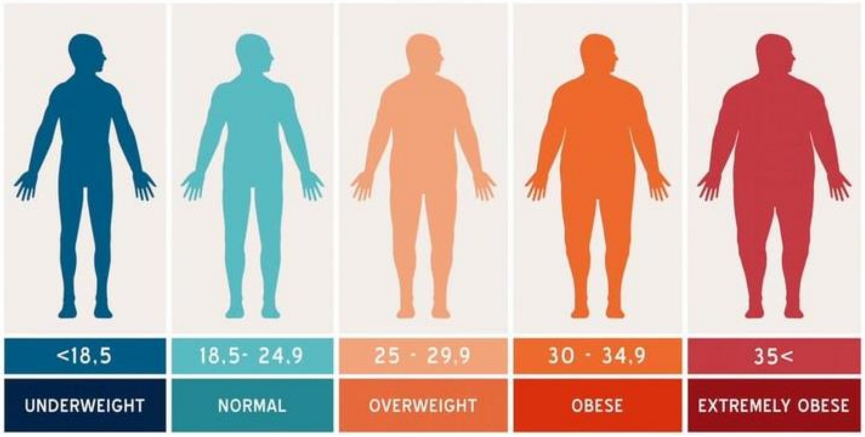
Nationality: can be presented in any way; there is no order for the countries.

> 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.

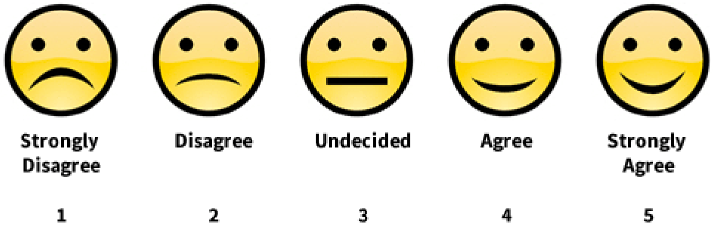
2- **Ordinal variables:** those are categorical variables that have an order, and that order has a meaning.

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.

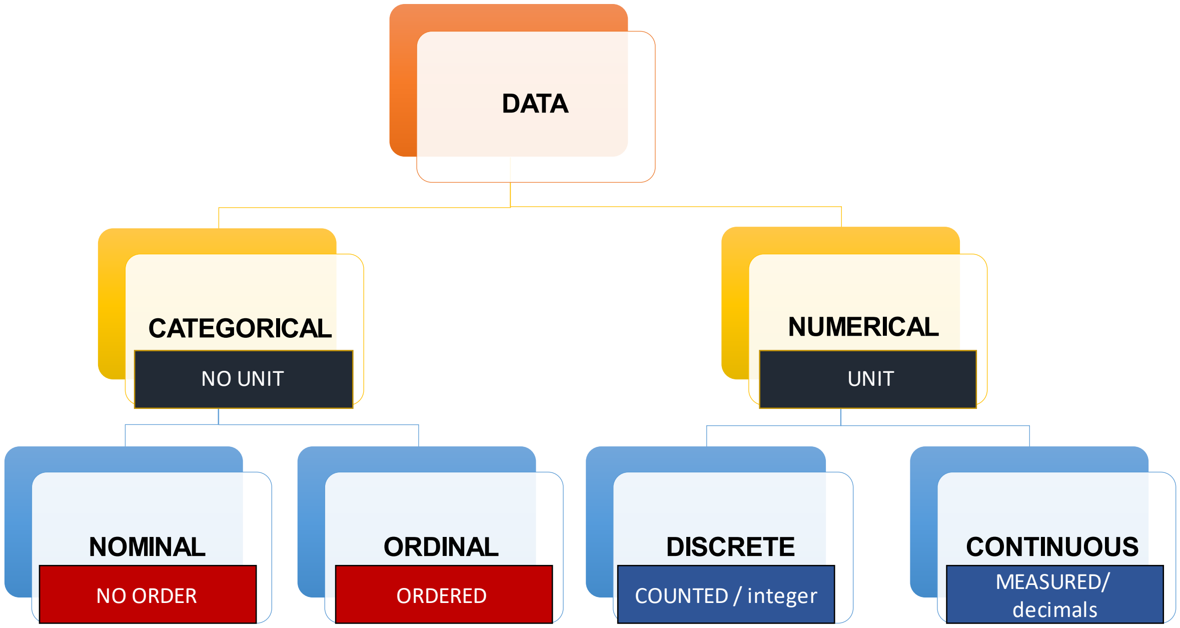
Step 2:

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^{0}C does not mean the absence of heat, and a temperature of 30^{0}C is not twice as hot as 15^{0}C.
- 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.

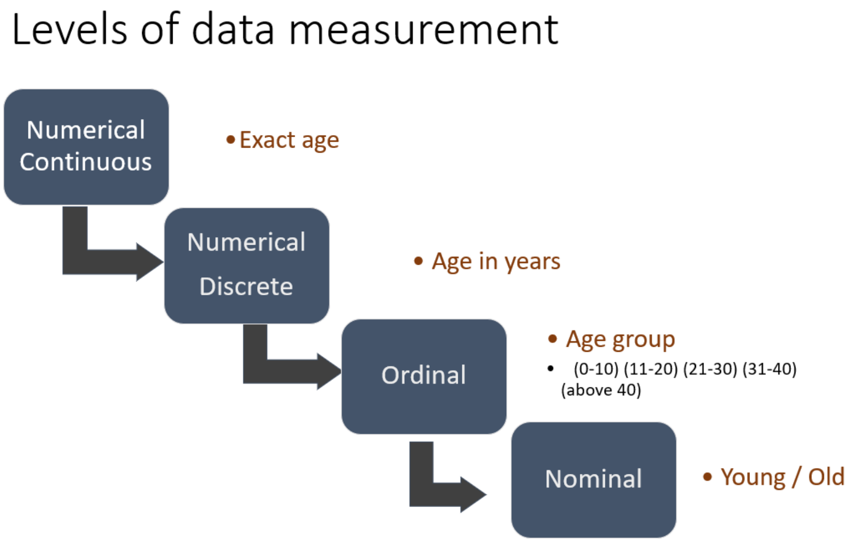
Levels of data measurement:

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.



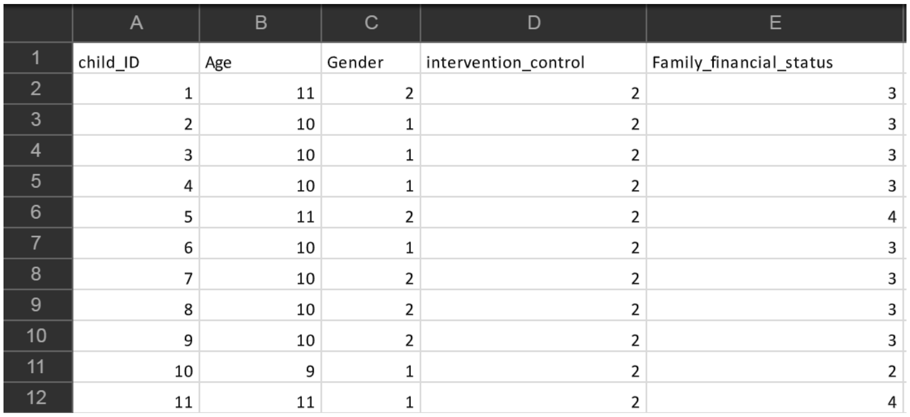
Levels of data measurement

- Numerical Continuous
  - Exact age
- Numerical Discrete
  - Age in years
- Ordinal
  - Age group
  - (0-10) (11-20) (21-30) (31-40) (above 40)
- Nominal
  - Young / Old

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)

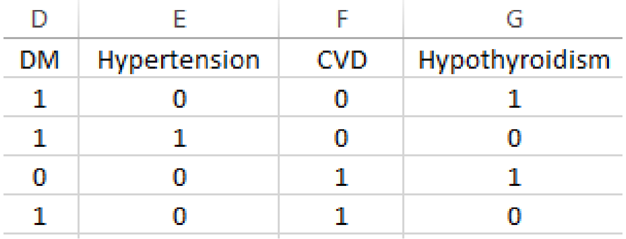
- Yes → 1
- No → 0

➢ 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:

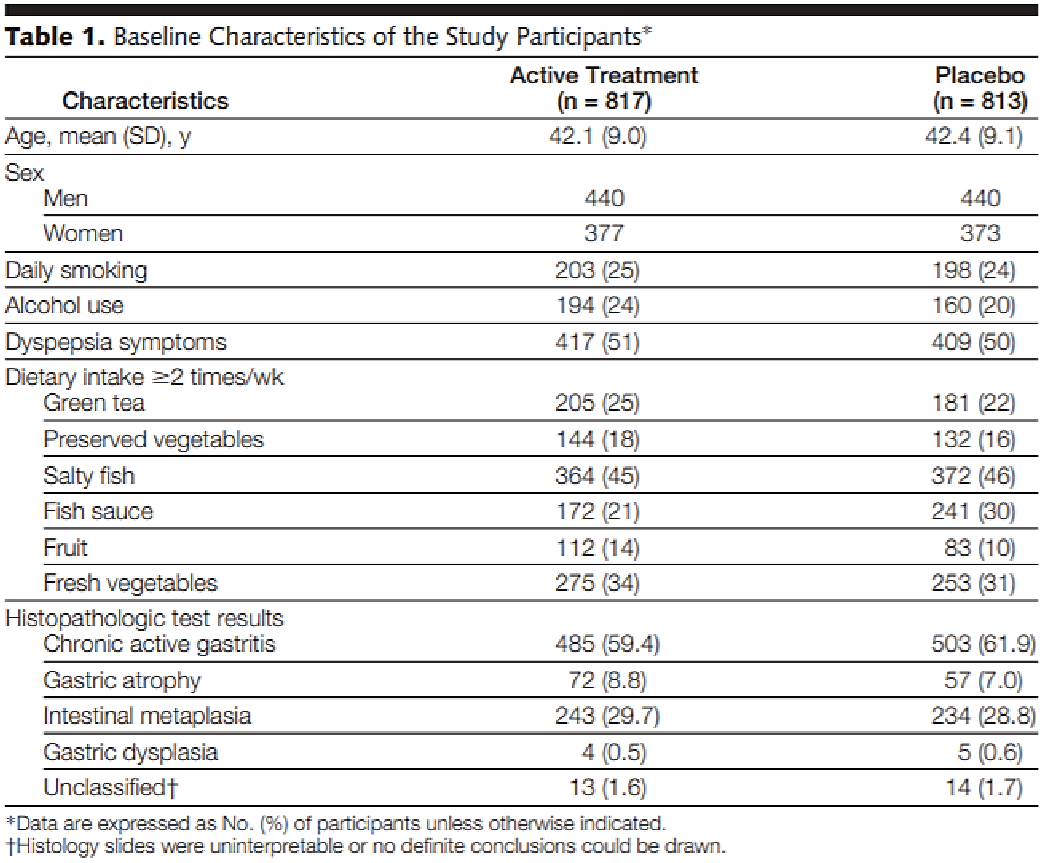


Table 1. Baseline Characteristics of the Study Participants*

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.

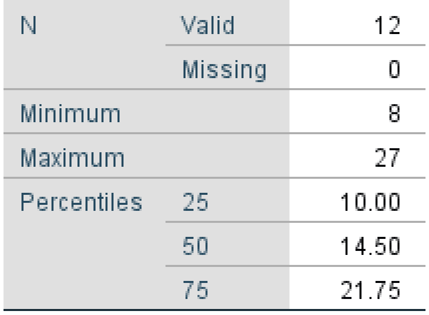
If you have the following values for a variable:

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

The five-number summary will be:

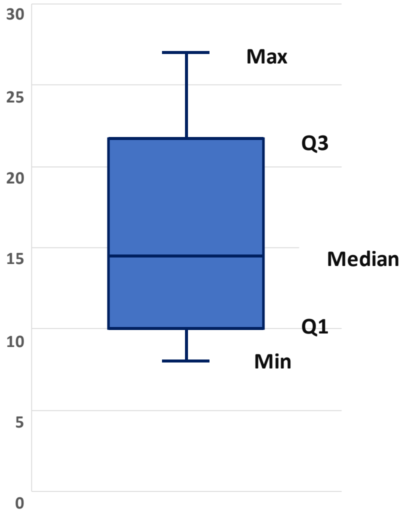
Min: 8 **Q1:** 10 **Median:** 14.5 **Q3:** 21.75 **Max:** 27

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:

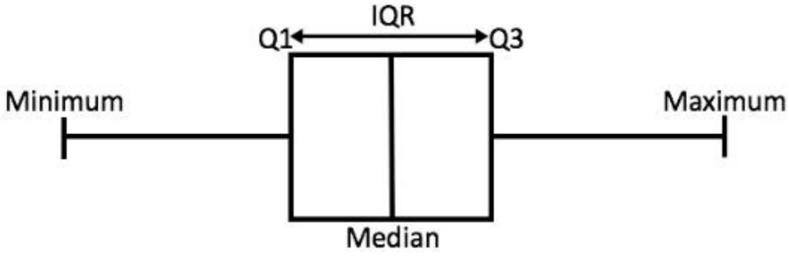
- The range is simply the difference between the largest and smallest values.

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.
$(7\text{-}6)^2$, $(5\text{-}6)^2$, $(6\text{-}6)^2$, $(8\text{-}6)^2$, $(4\text{-}6)^2$, $(9\text{-}6)^2$, $(3\text{-}6)^2$
$= 1^2, -1^2, 0^2, 2^2, -2^2, 3^2, -3^2$
=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.

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

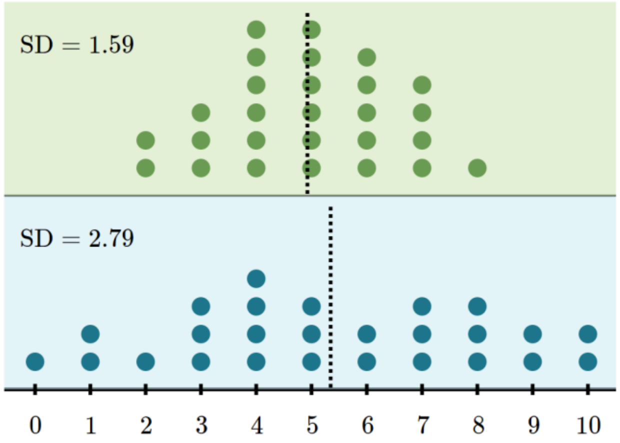
In the previous example, the variance = 4.67 years^2

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.

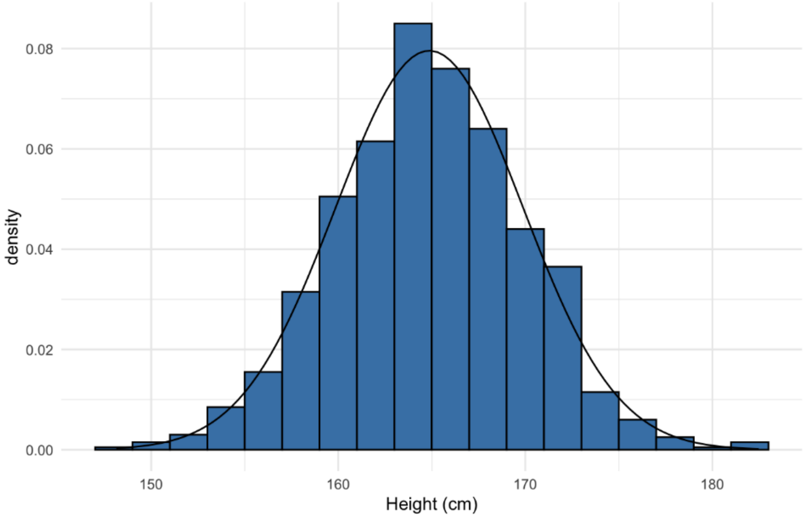


Combining measures of central tendency and measures of dispersion:

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).

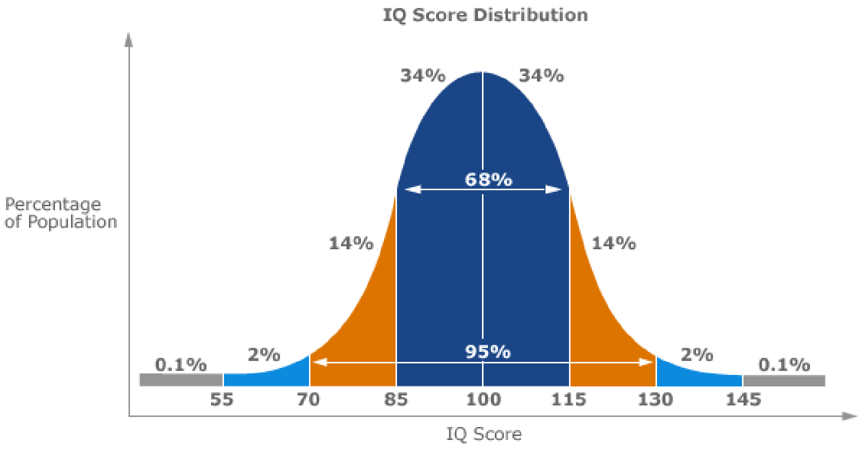
What is normally distributed data?



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

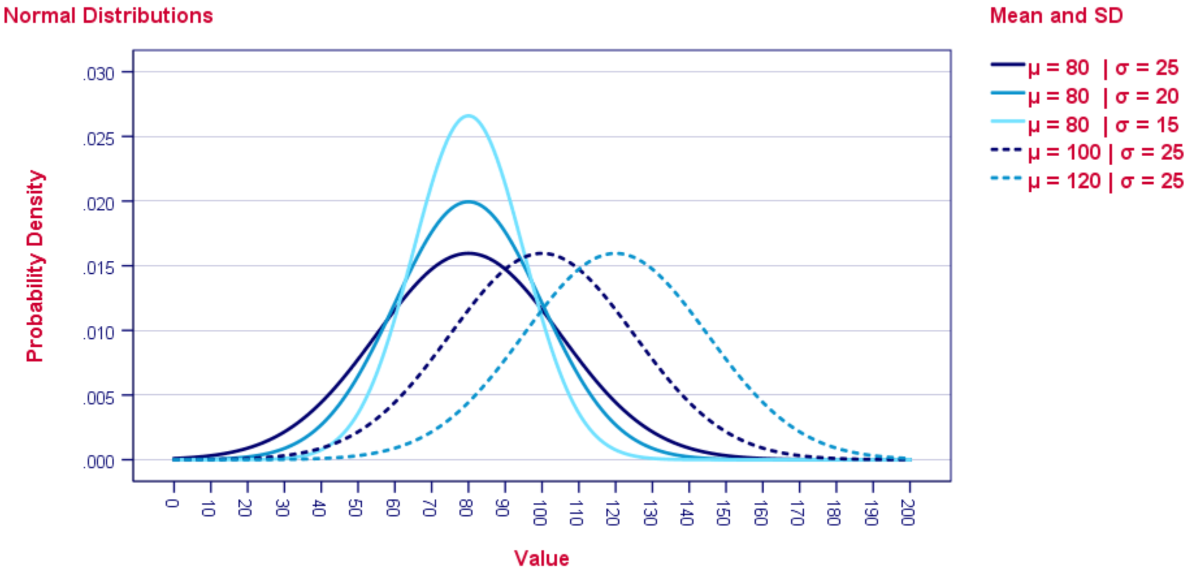
- Symmetric around the mean.
- The mean, median, and mode of a normal distribution are equal.
- Normal distributions are denser in the center and less dense in the tails (bell shape).
- 50% of values less than the mean and 50% greater than the mean
- Normal distributions are defined by two parameters,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

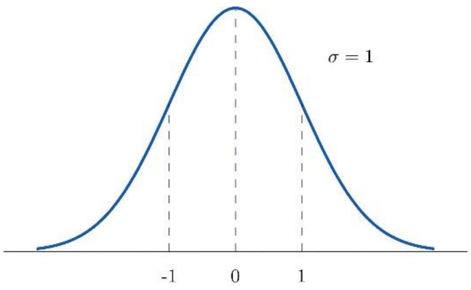


Examples for normally distributed data: height, blood pressure, IQ, …

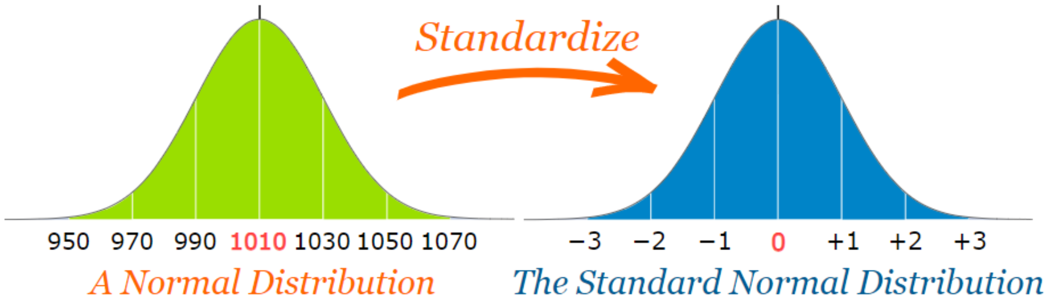
The following graph represents normal distributions with different means and standard deviations:



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



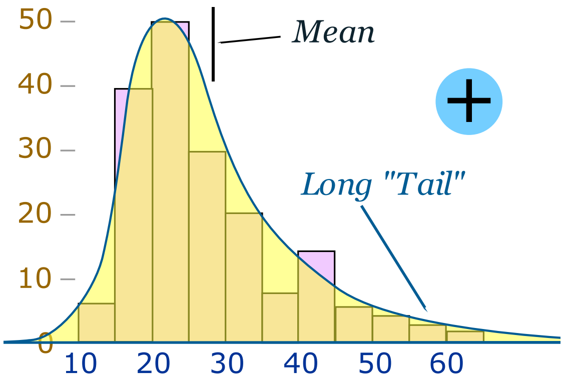
Any normal distribution values can be standardized (transferred to a standardized Z value)



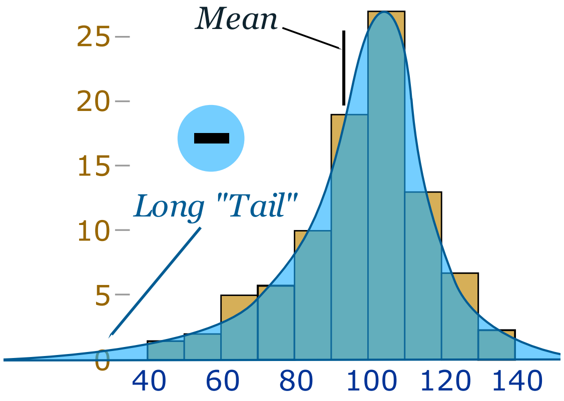
Examples of non-normally distributed data:

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.

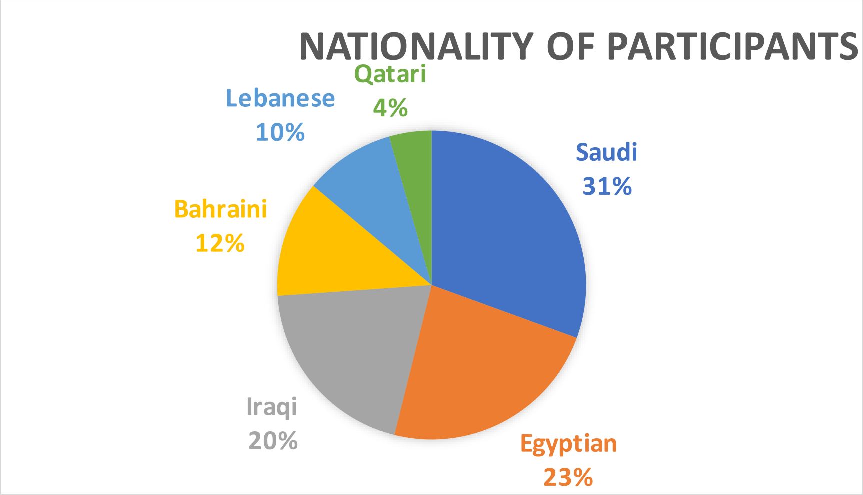
Graphical presentation of data

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

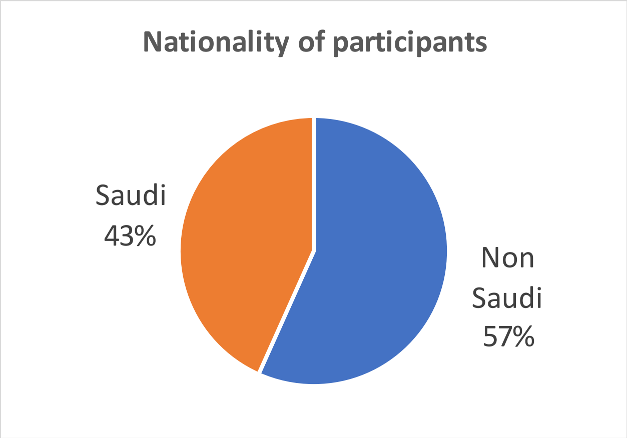
- **Nominal variables: Pie chart**
  The pie (circle) represents 100% of the variable and is divided into sectors. The area of each sector represents the frequency of each category in the variable it represents as follows:



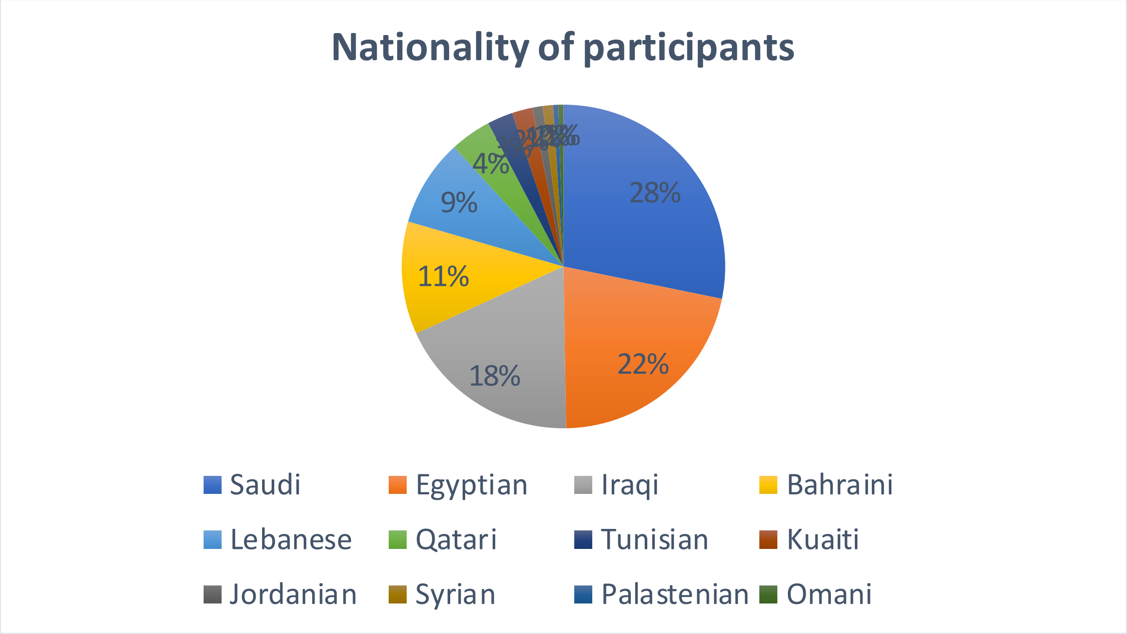
The pie chart is less commonly used in scientific papers due to its limitations. It can present only one variable.

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:

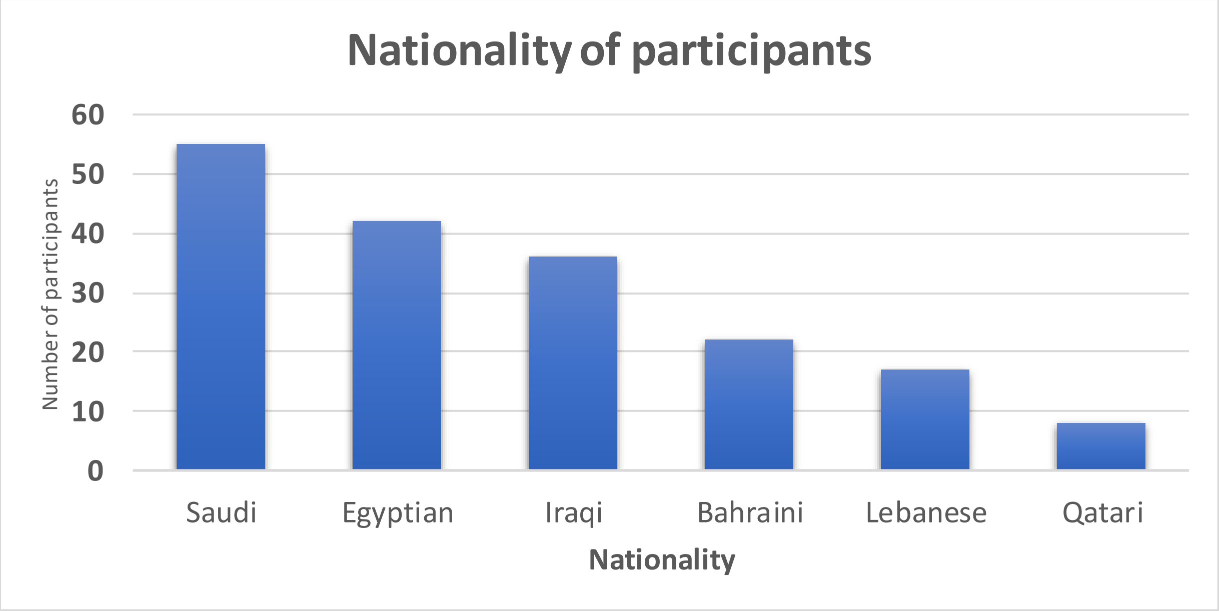
A pie graph of a binary variable:



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.

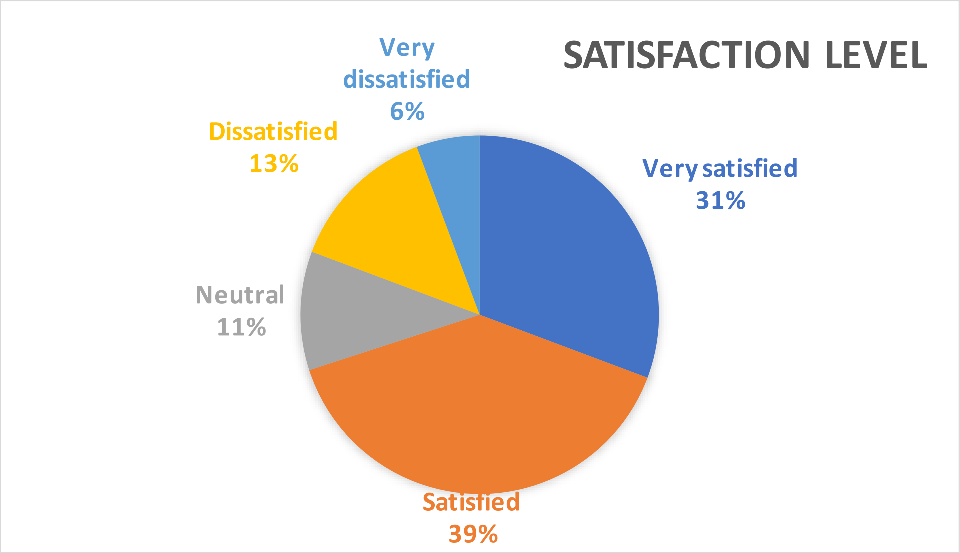


As the nominal variables have no meaningful order, it may be better looking to rearrange the categories based on their frequencies as in the graph above.

Ordinal Variables

- **Ordinal variables: Pie chart**

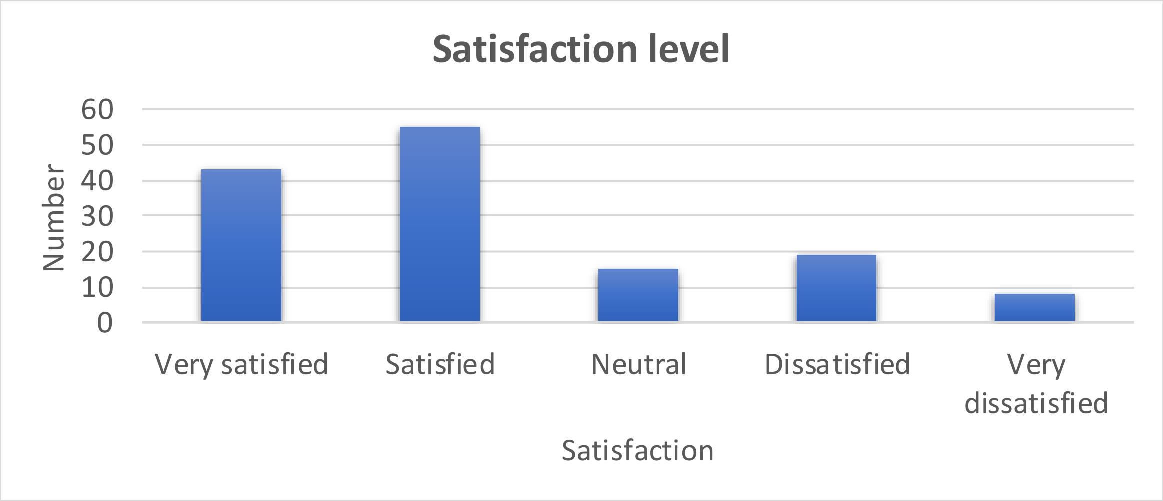
  Ordinal variables can be presented using the pie chart as in the nominal variables.



- **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.

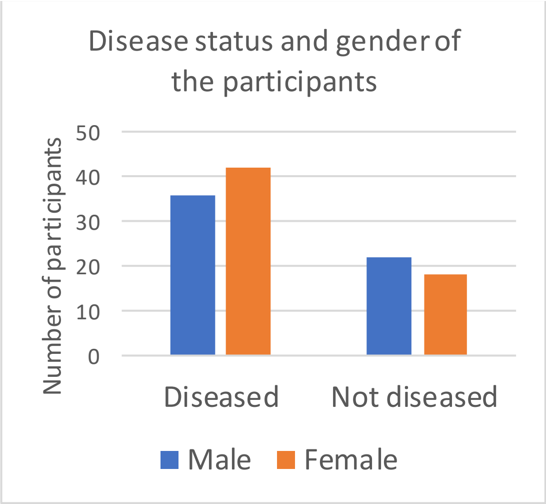
  Here, we can't change the order of the categories, otherwise, the reader may get confused.

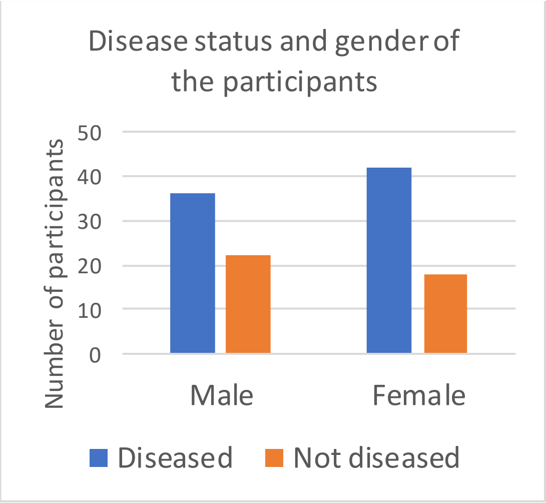


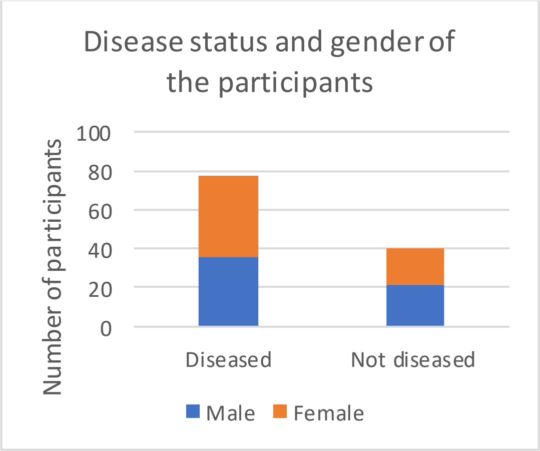
- As a general concept, we should use the graph the best demonstrates our data.
- Pie charts are not commonly used in scientific papers, they are usually used for presentations.

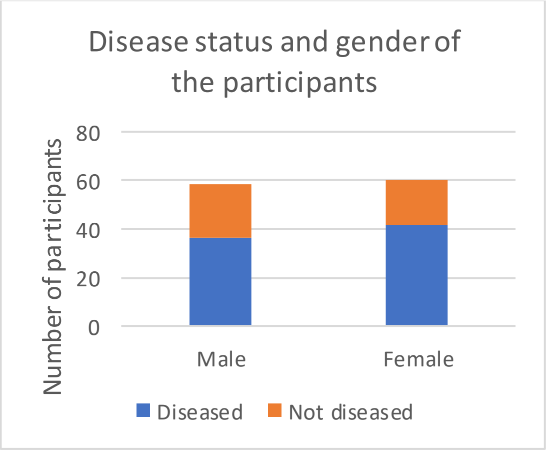
Two Categorical Variables

- **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

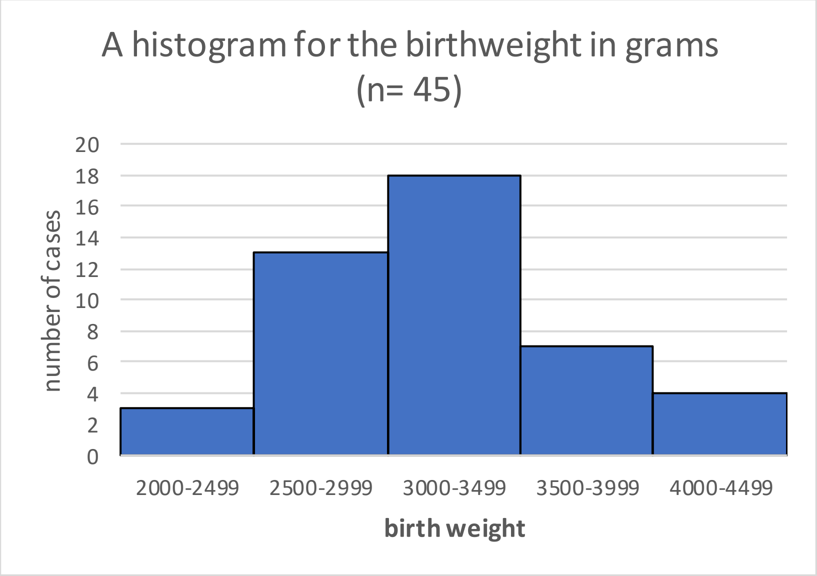
- **Numerical variables: Histogram**

  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.

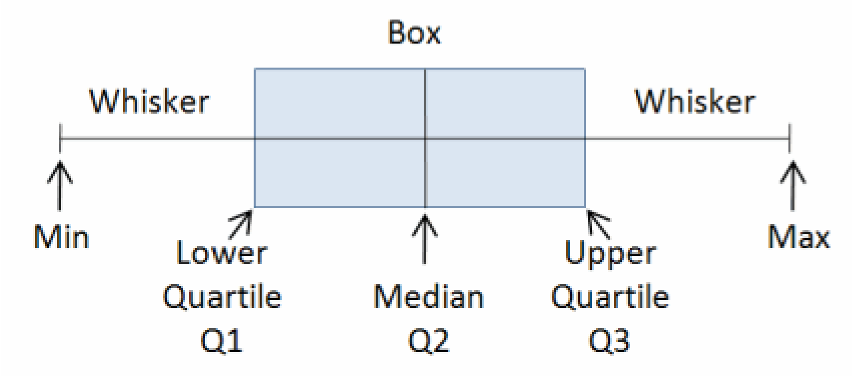


- **Numerical variables: Box plot**

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

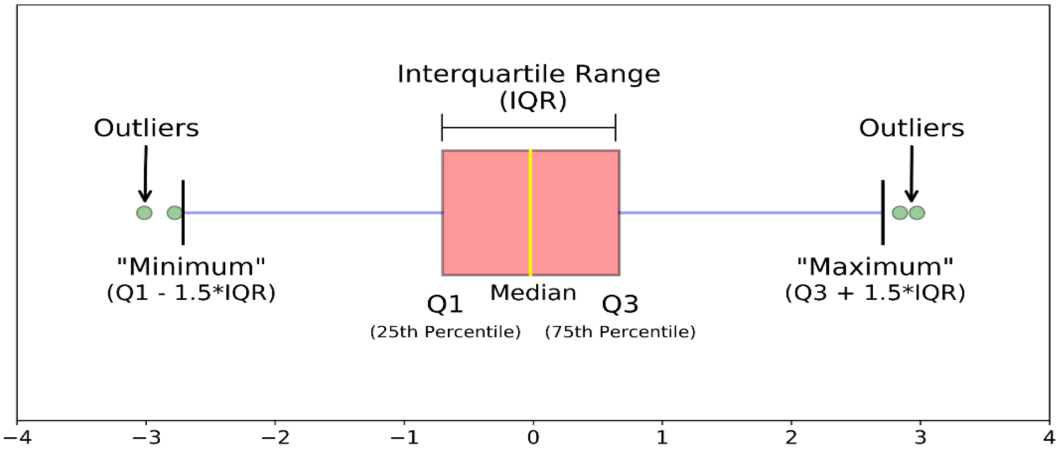
  Those five numbers are minimum, maximum, median, upper quartile, and lower quartile.

  - 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'



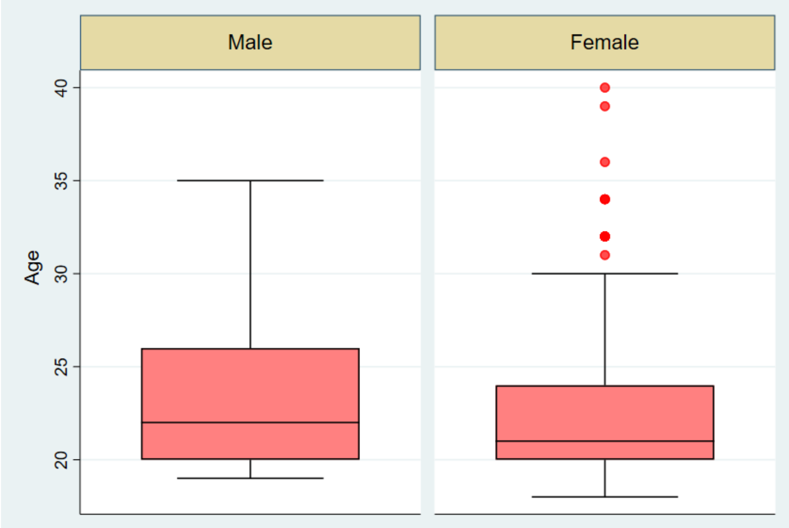
The whiskers are limited to outside 1.5 times the interquartile range above the upper quartile and below the lower quartile (Q1 - 1.5 * IQR or Q3 + 1.5 * IQR).

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.

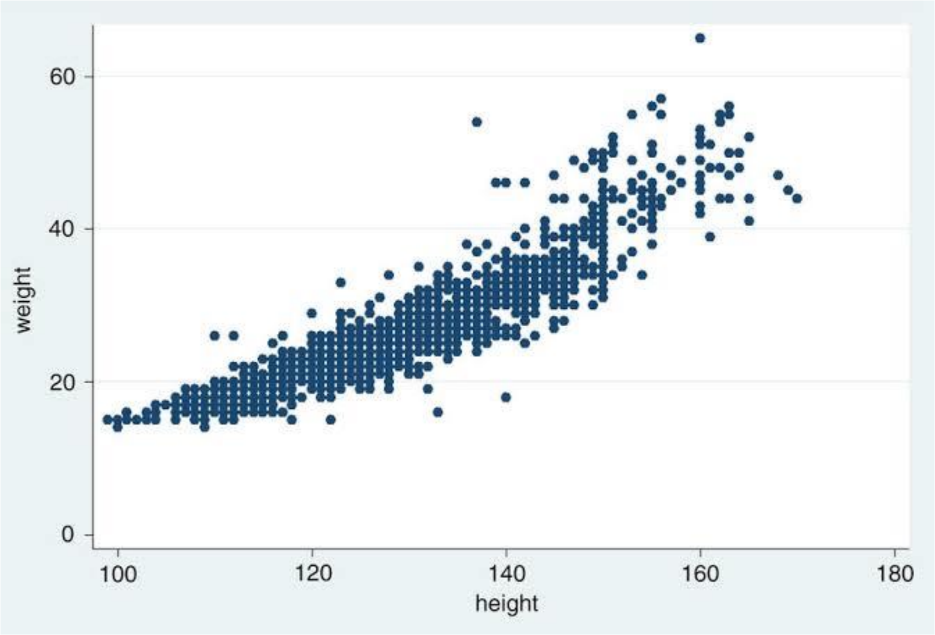
The following graph shows a boxplot for men and a boxplot for women allowing us to compare the same variable (age) between the two groups.



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.


weight
height