Little e-book for MPH1 biostatistics

Nolwenn Le Meur, PhD - EHESP associate professor in Biostatistics and Bioinformatic

2022-01-25

Contents

Prerequisites

Not to be afraid of numbers ! Otherwise I will try to work around it.

Chapter 1

Introduction

You may wonder why you will need to learn some statistics as you do not plan to compute statistics yourself in your future job. But, as a public health professional, the decisions you may be brought to make based on data will be too important to delegate. You will want to be able to interpret the data that surrounds you and to come to your own conclusions [Sharpe et al., 2012].

At the end, I hope you will understand the importance of statistics in our complex world and even enjoy studying the discipline.

1.1 Lecture Tips

When you see a term written like `statistics` in the text it is a concept that you must understand the definition and the usage. A proper definition will often be highlighted in a box. No need to know the definition by heart.

Different boxes will highlight the concepts, warnings and small practical exercises.

The first learning objective is to relax and trust yourself on your capacities to enjoy the biostat class.

This will be the definition of a concept you should understand along with its usage

This is a warning

This is a small question you should try to answer

This little book is being developed as a support for my online classroom in the context of the SARS-Cov2 pandemic. The content is a mix of my usual slides with definitions and examples from the books by Sharpe et al. [2012] (second edition), by Diez et al. [2019] (fourth edition), and by Ancelle [2017, in French].

To illustrate the different concepts, I mostly use a sample of the 2006 French Health Behavior School-aged Children database (HBSC). Since 1982 HBSC has been a pioneer cross-national study gaining insight into young people's well-being, health behaviors and their social context. This research collaboration with the WHO Regional Office for Europe is conducted every four years in 50 countries and regions across Europe and North America. For more details please visit the HBSC website [Aarø et al., 1986]

The statistical software R is used to illustrate some of the statistical tests and modeling methods [R Core Team, 2020]. Commands and output examples are presented. R and the bookdown package were in fact used to create this little e-book [Xie, 2016].

Chapter 2

Data: Statistical units and Variables

The `W's` and their types are what you should be interested in and careful about in the data:

- Who
- What
- Where
- When

Statistics, from the latin `status` and same root as State, is the art of counting and classifying. At first, statistics were used to describe countable information on population. Rapidly they became essential to model and predict data from experiences to foresee outcomes and help decision making.

2.1 Statistical units

"Statistics is all about `variation`".

In Public Health, we are interested in population. In biostatistics, we like to compare groups in a population. To that aim we need to identify our groups which will be composed of `statistical units`. For example, the statistical units could be: - patients, schooled children - health care services - countries

The *Who* is the `statistical unit`, the unitary element of interest (individual, hospital, country…).

2.2 Variables

The statistical units are characterized by one or more `variables`, which by definition varies between statistical units.

The *What* are `variables`, the recorded characteristics of the statistical units.

For example, the variables characterizing the schooled children could be: age, height, weight…(Figure 2.1). For health care establishments, the characteristics could be the legal status, the number of beds, nurses, doctors, patients…(Figure 2.2).

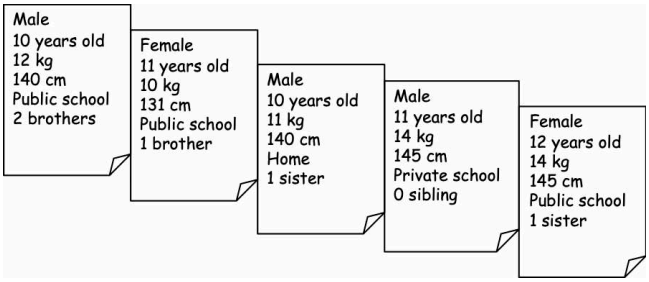


Figure 2.1: Individuals as statistical units

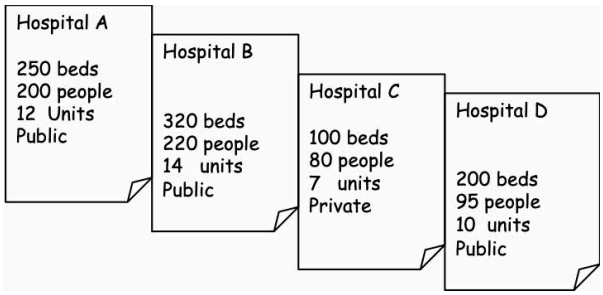


Figure 2.2: Hospitals as statistical units

2.3 Data storage

The information on each statistical unit can be stored and displayed in a data table. Typically, the *Who* of the table are found in the leftmost column and read row-wise. The *What* are stored in the remaining columns. Table 2.1 presents a snapshot of the HBSC data presented in the Introduction section (**??**).

Try to guess what these data represent and what information is available.

Hint: Do not forget to read the title of the table.

The *Where* and *When* are the context/location and time of the data collection.

Table 2.1: A table of the first 10 rows and first 8 columns of the HBSC data, France 2006.

ID	Grade.level	School.Status	Gender	Age	Weight
5617	7th grade	private	boy	13.58	50.0
4578	6th grade	public	boy	12.00	47.7
6512	6th grade	public	girl	11.42	28.7
5695	10th grade	private	girl	15.58	48.5
3906	8th grade	public	boy	13.75	48.5
6266	6th grade	public	boy	11.67	59.8
363	9th grade	public	girl	15.25	NA
1095	6th grade	public	boy	12.75	47.5
5388	8th grade	public	girl	13.50	52.0
6730	8th grade	public	girl	15.25	58.0

For instance in our HBSC the *When* is the year 2006 and the *Where* is in France. The scale of time and place are of great importance that need to be clearly defined. For example, the time can be a time point or a period of several months or years. As for France, it could be metropolitan France (excluding overseas departments) or France with all its departments. We could also look at different geographical levels like the city, the county, the state…

Those information have to be reported in every titles of every tables and plots you will create from the data along with the *Who* and *What*. A table or a figure should be `self-explanatory` (self-content).

2.4 Variable types

Public Health data may come from various sources. For instance, they can be collected via interviews, surveys,or health information systems.

In qualitative sciences, interviews are often based on open questions where answers are free text. We will not discuss that case in this class.

In quantitative sciences, surveys or records from health information systems are based on short queries where short answers with a finite range of possibilities are expected. For instance, let's say you are interested in tobacco consumption and plan a survey. You may ask the following questions with finite possibilities of answers:

Question	Possible answer	Variable type
How many cigarettes to you smoke per day?	1, 5, 10….	Quantitative discrete
Do you smoke?	yes or no	Qualitative binary
Are you a : (1) non-smoker, (2) a light smoker (0-5 cig./day), (3) moderate smoker (5-15cig./day) (4) heavy smoker (more than 15 cigarettes/day)?	light smoker	Qualitative ordinal

Variables are of different types (Figure 2.3). When a variable is allowed to takes a limited number of categorical values, or categories, and answers questions about how cases fall into those categories, we call it a `categorical`, or `qualitative`, variable. When a variable corresponds to measured numerical values with units and the variable tells us about the quantity of what is measured, we call it a `quantitative` variable [Sharpe et al., 2012]. The type of a variable will condition the statistical method chosen to summarize and describe your population of interest (Chapter 3).

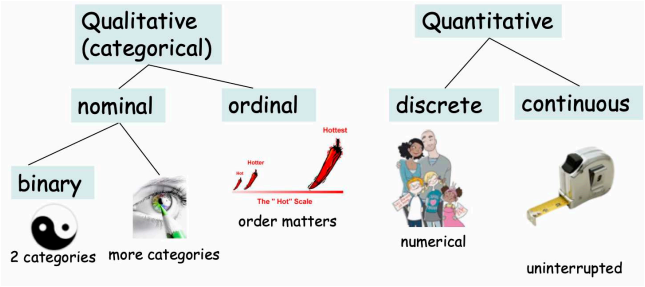


Figure 2.3: Different types of variables with example.

The `categorical`, or `qualitative`, variables can be of two sub-types: the `nominal` and the `ordinal` variables. The nominal variables are for instance the colors of the eyes or various professions that count many categories. The nominal variables can also be `binary` with only two categories like smoking/non smoking or boys/girls. The ordinal variables take into account an ordering between the possible categories of the variables. For instance, a ordinal variable could be a scale of spiciness: neutral, middle, hot, very hot ! Note that this is suggestive. You can come up with a ranking where the intervals between the categories are not of equal width.

The `quantitative` variables can be discrete or continuous. A quantitative discrete variable corresponds to numerical counts like the `number` of kids per household. A quantitative continuous variable corresponds to `measures` with potential decimals like the weight or height of pupils in the HBSC cohort.

Propose a set of variables, one of each type.

Chapter 3

Descriptive statistics

- Choose appropriate summary statistics (tables, graphics, parameters) to describe a population or a sample
- Interpret summary statistics

3.1 Frequency table

One type of data statistical summary are `frequency table`, or `contingency table`.

To summarize a `qualitative` variable, the frequency of the statistical units in each modality (category) of the variable is computed. The frequency can be reported as absolute count (absolute frequency) or proportion (relative frequency or count). For instance, in Table 3.1 the partition of students according to the school status is summarized (HBSC dataset).

In rows are the modalities (categories) of the variable "school status" in the original database and in columns are the frequency of students in each category. The first column is the `absolute frequency` or count and the second column is the `relative frequency` or proportion out of the total of students. The

Table 3.1: Absolute frequency and relative frequency of pupils in each school type in the French HBSC database in 2006.

	Count	Proportion (%)
private	115	23
public	385	77
Total	500	100

Table 3.2: Frequency table to summarize age distribution of children in the HBSC database, in France in 2006.

	Count	Proportion (%)
NA	1	0.2
[11-13[	173	34.6
[13-15[	178	35.6
[15-17[	148	29.6
Total	500	100.0

Table 3.3: Repartition of smokers and non smokers among age groups in the HSBC sample, in France in 2006

Age group	Smoking Status		
	No	Yes	Total
NA	1	0	1
[11-13[	166	6	172
[13-15[	161	17	178
[15-17[	104	44	148
Total	432	67	499

`margin total` is essential to display for quick assessment of potential mistake or missing values.

To summarize a `quantitative` variable into a contingency table, the numerical values first need to be grouped into classes, generating in fact like a categorical variable. Next, the frequency of the statistical units in each group (class) of the new variable is counted.

In Table 3.2, first the quantitative variable *age* was used to create age groups and, next, the partition of students according to age group was summarized.

One must read `[11-13[` as 11 years-old students being included (counted) in that group while 13 years-old students being excluded. `NA` stands for Not Attributed or missing values.

When summarizing two (or more) variables in two-way (or more) table using frequencies via a statistical software you might have to look for the term `pivot table`. Table 3.3 summarize in absolute frequencies the different age groups and the smoking status. For relative frequencies, you need to decide which way to count (Table 3.4 and 3.5). You need to ask yourself: Who are you interested in? What is your denominator?

Among the [15-17[years old, 29.7% smoke.

Table 3.4: Proportion of smokers and non smokers among age groups in the HSBC sample, in France in 2006

Age group	Smoking Status		
	No	Yes	Total
NA	100.00	0.00	100
[11-13[	96.51	3.49	100
[13-15[	90.45	9.55	100
[15-17[	70.27	29.73	100

Table 3.5: Relative distribution of age groups among smokers and non smokers in the HSBC sample, in France in 2006

Age group	Smoking Status	
	No	Yes
NA	0.23	0.00
[11-13[	38.43	8.96
[13-15[	37.27	25.37
[15-17[	24.07	65.67
Total	100.00	100.00

Among the smokers, 65% are aged [15-17[years old.

The proportion of smoker seems to increase with age. We will verify this later using inferential statistics (see Chapter 5)

3.2 Central parameters

A central parameter, or location parameter, is the numerical value around which are distributed most of the values of a serie of data.

3.2.1 Mean

The (arithmetic) mean is the most well known and commonly (but not always appropriately) used central parameter.

The arithmetic mean is the sum of the values divided by the number of values in the data serie.

Mathematically, in a population the equation is:

$\mu = (\sum_{i=1}^{i=N} X_i)/N$

In a sample, the equation is:

$m = (\sum_{i=1}^{i=n} x_i)/n$

Note: Greek letters are used for population and Roman letters for sample. The mean is also sometimes symbolized like $\bar{X}$ for population or $\bar{x}$ for sample

Figure 3.1 presents a `histogram` that summarize the distribution of weight of French student in the HBSC `sample` (for histogram definition see section 3.4 and sample definition see 4.1). From the graphic, the central point (pick) of the distribution, around which values of the data serie are spread is around the weight class [40-45[Kg.

On average, in 2006 the French students aged 11 to 16 weighted 48 Kg (dashed red color).

Why the mean is not between 40 and 45Kg?

The advantages:

- Easy to understand
- Easy to compute

The drawbacks:

- Sensitive to outlier: each value of the data serie count with the same weight
- Sensitive to the distribution shape

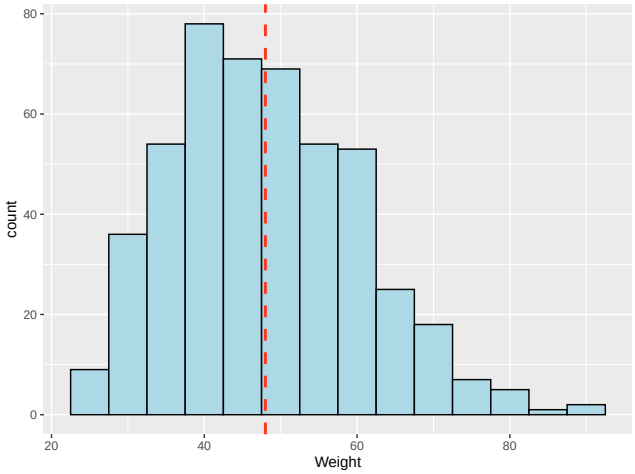
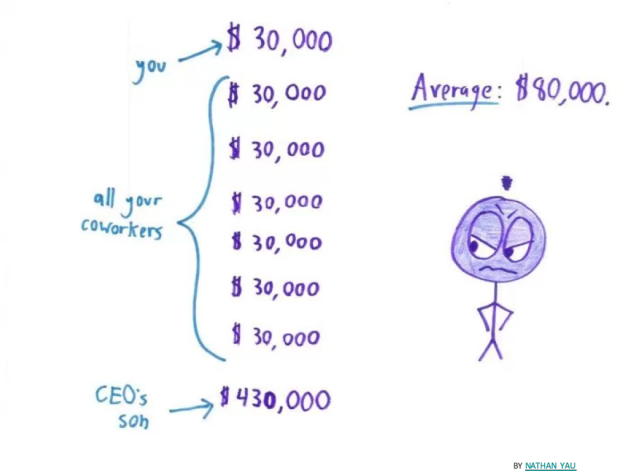


Figure 3.1: Distribution of weights (Kg) of 11 to 16 years-old students, in France in 2006



3.2.2 Median

The mean is not always the appropriate statistical indicator to summarize the distribution data and should be sometimes replaced by the median.

The median is the middle value of a ordered data serie. The median split the data serie in two part of equal number of data.

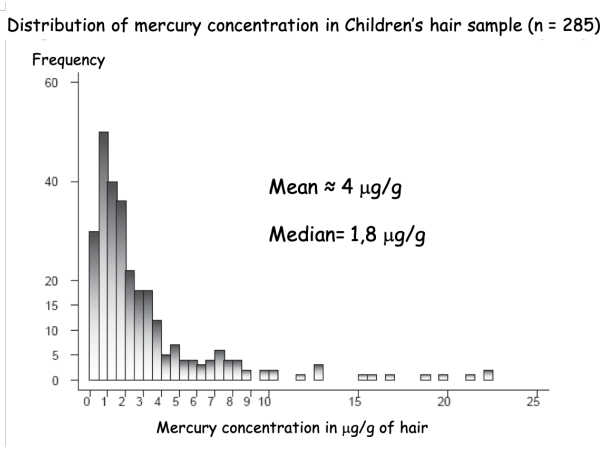


Figure 3.2: Mean or Median, that is the question?

In Figure 3.2, 50% of the statistical units, i.e 142 children hair sample, have mercury concentration below 1.8 $\mu g/g$ while 50%, i.e 142 children hair sample, have mercury concentration above 1.8 $\mu g/g$.

If you had rely on the mean you would have said that on average children hair contain 4 $\mu g/g$ of mercury which wrongly make you believe that represent most of the children case.

How to compute a median?

- 1) Sort values in increasing order
- 2) If there are an odd number of observations, find the middle value
- 2'. If there are an even number of observations, find the middle two values and average them

Would you use the median or the mean to compare French region rainfall?

The advantages:

- Easy to compute
- Not sensitive to outlier
- Less sensitive to skewed distribution than the mean
- Easy to understand

The drawbacks:

- Sensitive to the distribution shape
- No idea of the minimum and maximum
- Easy to understand but need to be explicitly exposed

3.2.3 Percentile and quantile

When looking at height distribution, the median is the exact middle value when people are ordered by height which correspond to the 50^{th} `percentile` or 50% below and above that value (Figure 3.3). But you could pick any percentile like the 80^{th} with 80% below and 20% above.

The nth percentile of a set of data is the value at which n percent of the data is below it.

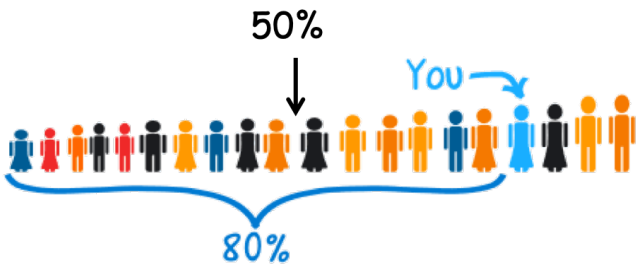


Figure 3.3: Percentile

Percentiles can be calculated using the formula $n = (P/100) * N$, where P = percentile, N = number of values in a data set (sorted from smallest to largest), and n = ordinal rank of a given value.

A student scores in the 75^{th} percentile of his class. What does that mean?

The 75^{th} percentile is also the 3rd `quartile`.

The quartile split the sorted data values into quarters.

The quartiles are the values that frame the middle 50% of the data (median or Q2). One quarter of the data lies below the lower quartile, Q1 (25% or 25^{th} percentile), and one quarter of the data lies above the upper quartile, Q3 (75% or 75^{th} percentile).

Using the R statistical software and the HBSC data set we can quickly describe the "Weight" variable of the French student aged 11 to 16 in 2006 with the `five-number summary`. The five-number summary provides a good overall look at the distribution of the data.

```
summary(hbsc$Weight)
```

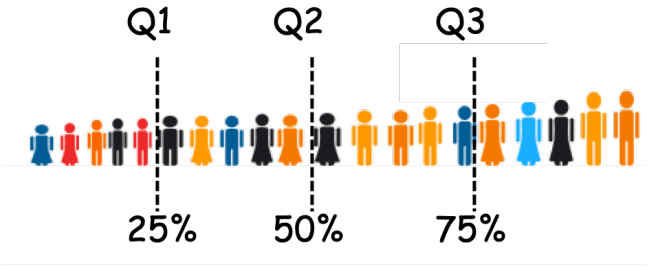


Figure 3.4: Quartiles

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   24.50   38.83   47.00   48.09   56.00   90.00      18
```

- There was 18 missing values (NA)
- The minimum weight was 24.5 Kg
- The maximum weight was 90 Kg
- The mean (average) is 48.09 Kg
- The median is 47 Kg meaning that 50% of the students weighted less than 47Kg and 50% of the students are heavier.
- The 1^{st} quartile is 38.8Kg meaning that 25% of the students weighted less than 38.8Kg and 75% weighted more.
- The 3^{rd} quartile is 56Kg meaning that 75% of the students weighted less than 56Kg and 25% weighted more.

Quantile algorithms

Several algorithms exit to compute quantile (for instance see ?quantile in the R staistical software). They rely on different definitions of the underlying distribution of the sample: discontinuous or continuous.

In your case no needs to go into the details but you should know how to interpret the values.

Mean versus Median

If the data is normally distributed, i.e. bell shape, as statisticians like it (bottom of Figure 3.5), feel free to use the mean.he mean is easier to communicate and so if you can use it, use it. In fact the value of the mean should be really close the value of the median and the mode (or modal class).

If your data is skewed (top of Figure 3.5), or there are large outliers, then use the median to find the centre of the data. Better yet, report both the mean and the median since any differences will reveal information about the presence of skew/outliers.

A more subtle rule: if you are more concerned with the total sum, rather than the typical value, use the mean. For instance, if you have a salary cap and you

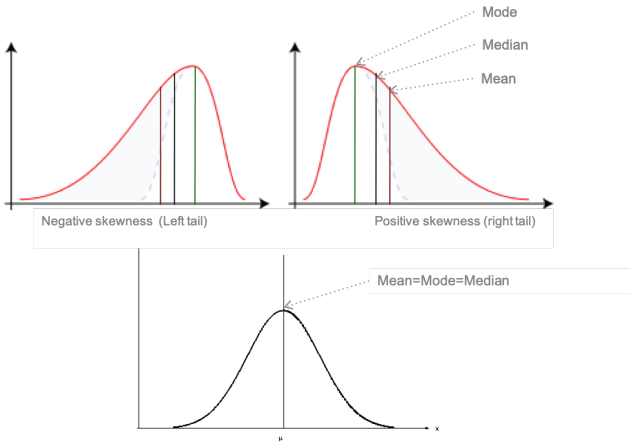


Figure 3.5: Bell shape distribution

Table 3.6: Frequency of physical activities (sport) in the HSBC sample, in France in 2006

Sports frequency	Number
NA	0
never	33
less than once a month	30
once a month	21
2 or 3 times a week	175
4 to 6 times a week	81
every day	63

are interested in the average salary of your players, use the mean. In this case, the mean is biased towards the high earners, and you really care about the high earners because they are the ones who are eating up your salary cap.

3.2.4 Mode

The most frequent value or modality

It is the only statistical parameter for the `qualitative` variables.

In table 3.6 the mode is the category "2 or 3 times a week" with 175 students out of 500 practicing that much sports.

For `quantitative` variables it could be a number or a class interval as in Figure 3.1 where the modal class is [40-45[Kg.

Likert scale data.

A Likert Scale is a type of rating scale used to measure attitudes or opinions. Five to seven items are usually used in the scale.

In a survey with a 1-5 scale of "1-Very bad", "2-Bad", "3-Neutral", "4-Good" and "5-Very Good" categories, the mean result across many participants came out to be 3.5. But what does 3.5 even mean in this context? Half way between Neutral and Good : Neutood? In terms of best practice, use the median when describing the centre of Likert data. Some may even argue for only using the mode on Likert data.

3.3 Variation parameters

A variation parameter is numerical value which describes the dispersion of all the values of a serie of data around its location parameter

3.3.1 Range and IQR

The `range` is the difference between the minimum and maximum value of a data serie. It is better to report the boundaries rather then the output of the difference because otherwise we do not know from where it starts.

In which hospital will you go for emergency care?

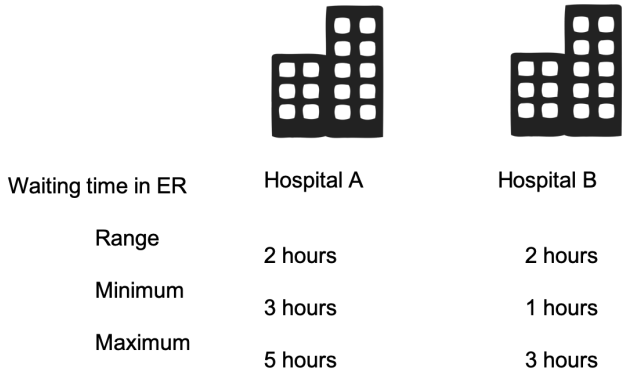


Figure 3.6: Hospitals ER

The `interquartile range (IQR)` summarizes the spread by focusing on the middle half of the data. It is defined as the difference between the two quartiles: $IQR = Q3 - Q1$. As the range, the IQR should be reported as an interval. For example, the five-number summary above on HBSC weight data show us that IQR is [38.8-56] Kg.

3.3.2 Variance and standard deviation

A powerful measure of spread is the distance of the values of a data serie to its mean.

The variance is the average squared differences to the mean.

Based on the mean, the `variance` is appropriate only for symmetric data and can be influenced by outlying observations.

Mathematically, if we tried to average the distances of all the values a data serie to its mean, the positive and negative differences would cancel each other out, giving an average deviation of 0-not very useful. Instead, we square each distance to get the `variance`.

In a population the equation is:

$$\sigma^2 = \frac{\sum_1^N (X_i - \bar{X})^2}{N}$$

where

- σ^2 is the variance of population
- X_i is the i^{th} value in the population
- $\bar{X}$ is the mean in the population
- N is the size of the population

In a sample, the formula is:

$$s^2 = \frac{\sum_1^n (x_i - \bar{x})^2}{(n-1)}$$

where

- s^2 is the variance of the sample
- x_i is the i^{th} value of the data serie
- $\bar{x}$ is the mean of the data serie
- n is number of values in the data serie

The variance plays an important role in statistics, but as a measure of spread, it has a problem. Whatever the units of the original data, the variance is in squared units. To express the spread to in the same units as the data we take the square root of the variance. That gives the `standard deviation`.

A standard deviation is the square root of the variance, the average differences to the mean.

Mathematically, in a sample the formula is:

$$s = \sqrt{\frac{\sum_1^n (x_i - \bar{x})^2}{(n-1)}}$$

Example 3.1. In the HBSC sample, one average pupils weight 48.09 Kg +/- 12.25Kg. It means that the **average variation** of the weights around the mean is of 12.25Kg. However some pupils can be lighter than 35.75kg (48-12.25 -

the minimum is in fact 24.5Kg) and some pupils can be heavier than 60.25kg (48+12.25 the minimum is in fact 90Kg).

3.4 Graphical summary

Graphical representations, as summary in tables, must be self-content and self-explanatory. They should be readable without text around.

Plot title should include the W's

Who are you representing, What characteristics, Where and When it is happening.

Note: I will be intransigent on that matter

3.4.1 Barplot

For one `qualitative` variable, the frequency `barplot`, or `barchart`, is the most appropriate plot. It could use the absolute or relative frequencies.

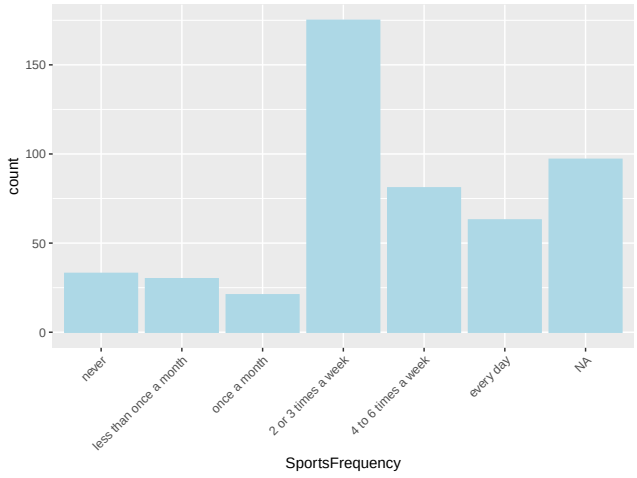


Figure 3.7: Repartition of students according to their sport activity level, in France in 2006 (source:HBSC)

Figure 3.7 represents the distribution of students (*Who*) according to their sport activity level (*What*), in France (*Where*) in 2006 (*When*) and if available the source should also be cited (here HSBC) in the caption.

It is a better practice to multiply plots that combining too many information in one plot. Imagine you would like to display the same information but for boys and girls. The below barplot is better than the next one as you can easily compare boys and girl but you also easily visualize the trend within each group.

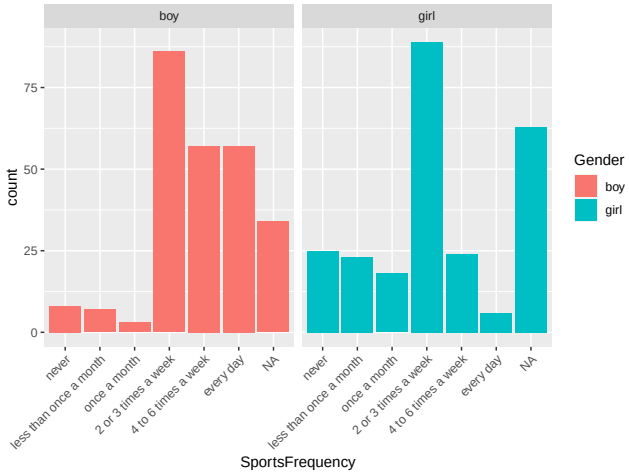


Figure 3.8: Repartition of students according to their sport activity level, in France in 2006 (source:HBSC)

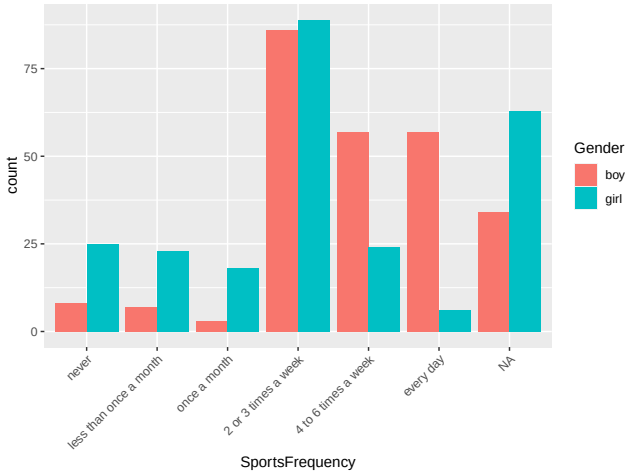


Figure 3.9: Repartition of students according to their sport activity level, in France in 2006 (source:HBSC)

3.4.2 Pie chart

For one `qualitative` variable, you can also use a `pie chart`. However you should be careful on the number of modalities the variable to display can have: too few, the plot take lots of room for few information but too many you will rapidly get "the wheel of fortune".

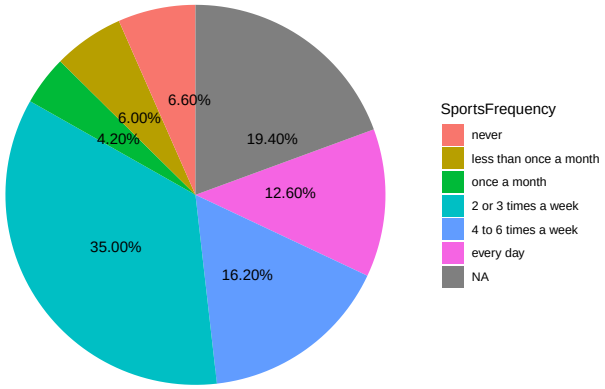


Figure 3.10: Repartition of students according to their sport activity level, in France in 2006 (source:HBSC)

My advice would be to avoid using that plot but if you absolutely need to use it do not forget to **display number and/or %** in the sectors of the plot so reader do not have to do the math.

3.4.3 Histogram

For one `quantitative` variable, the `histogram` is often used to visualize the shape of the distibution of the data serie. Although one have to be careful of the bin width (breaks) use to group the numerical values.

Figures 3.11 and 3.12 present the same data serie but with a bin width of 1 and then 10. The shape of the plots are different, the modal class varies. With a bin width of 1 you have no summary and too many details while a bin width of 10 may be compact.

There is no recipe, statistical software helps you with default algorithm but my advice would be to try a couple of bin widths and select the one you believe summarize the best the data

Barplot versus Histogram

Barplot are for `qualitative` data. The bars are separated.

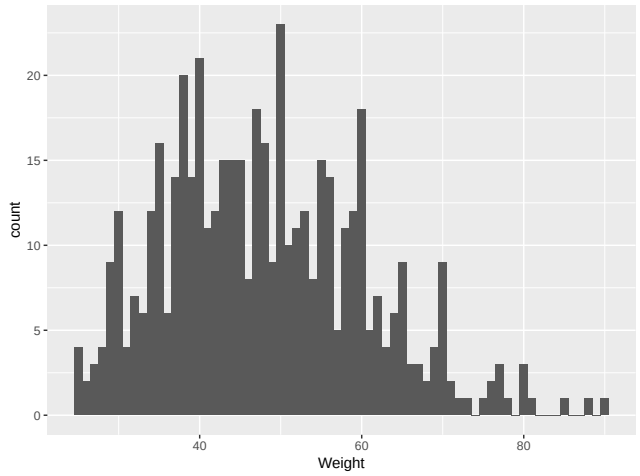


Figure 3.11: Distribution of students weight in France in 2006 (source: HBSC), binwidth=1

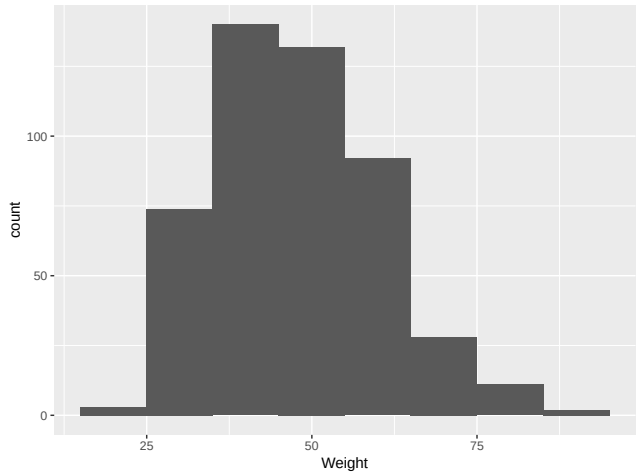


Figure 3.12: Distribution of students weight in France in 2006 (source: HBSC), binwidth=10

Histogram are for `quantitative` data. The bins are joined. Gaps may occurred when no one fall in a particular bin.

If you want to compare the distribution of `one quantitative variable between more than 2 groups (qualitative)` the histogram remain interesting, as shown below. But up to three the comparison start to be difficult. A boxplot is then more appropriate.

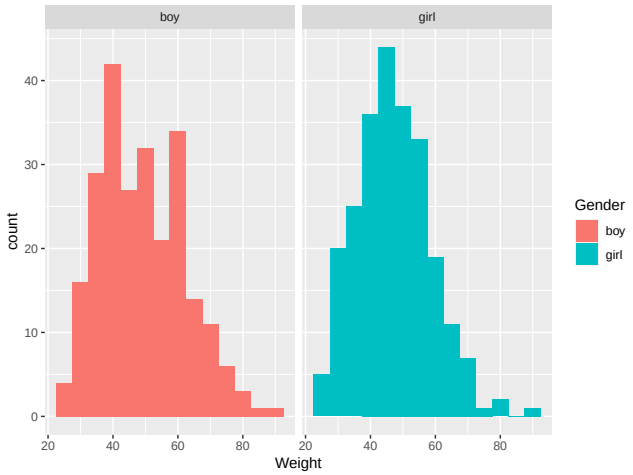


Figure 3.13: Distribution of students weight by gender group in France in 2006 (source: HBSC)

3.4.4 Boxplot

A boxplot helps presenting the five-number summary classically used to describe a sample.

The central box shows the middle half of the data, between the quartiles. The top of the box is at the third quartile (Q3) and the bottom is at Q1, the height of the box is equal to which is the IQR. The median is displayed as a horizontal line. If the median is roughly centered between the quartiles, then the middle half of the data is roughly symmetric. If it is not centered, the distribution is skewed. In extreme cases, the median can coincide with one of the quartiles.

The whiskers reach out from the box to the most extreme values that are not considered outliers according to John W. Tukey's rule. The boxplot nominates points as outliers if they fall farther than 1.5*IQRs beyond either quartile. They may be mistakes or they may be the most interesting cases in your data. This rule is not a definition of what makes a point an outlier. It just nominates cases for special attention.

In Figure 3.15 the distribution of students' weight vary by group of sport activity

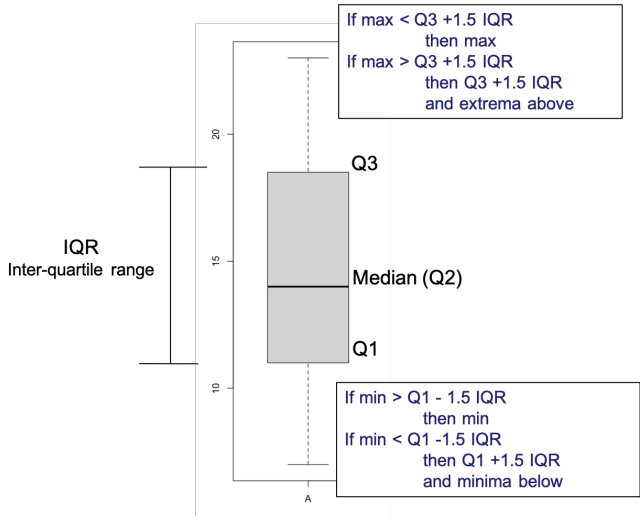


Figure 3.14: Boxplot elements

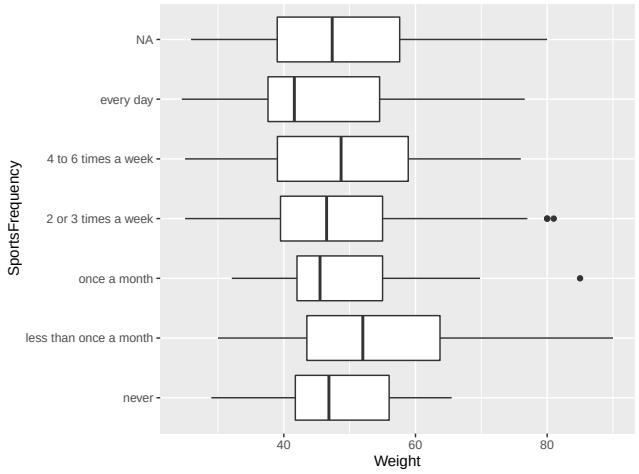


Figure 3.15: Distribution of students weight by group of sport activity level, in France in 2006 (source: HBSC)

level. We note 3 extrema: 2 in the "2 or 3 times a week" group and 1 in the "once a month" group. Overall the boxplots overlap suggesting than on average the weights between groups might not be statistically different (see Chapter 5).

Let's interpret the "never" group (bottom boxplot): around 50% of the students in that group weight 45Kg or less, 75% of the students weight 55Kg or less while 25% weight 55Kg or more

What proportion of students lies between 42 and 55 Kg?

3.4.5 Scatterplot

For two `quantitative` variables, a scatterplot is used to assess if there is a relationship between the two variables.

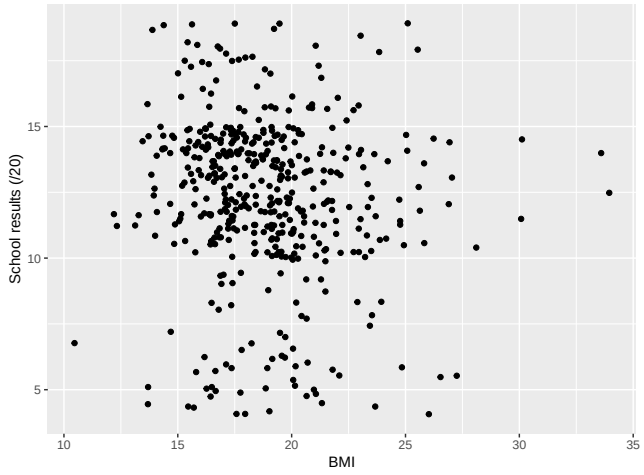


Figure 3.16: School results and BMI among students in France in 2006 (source: HBSC)

Figure 3.16 displays the school results and BMI among students in France in 2006. It appears that there is no linear relationship between the two variables (see Chapter 5 and 6.1).

3.4.6 Communication tips

Avoid 3D plots our eyes are not good at visualizing 3D and mathematically it's often wrong you do not manipulate volumes but numbers.

Use appropriate colors

- Proscribe rainbow plots
- Sequential scale (gradiant) are suited to ordered data

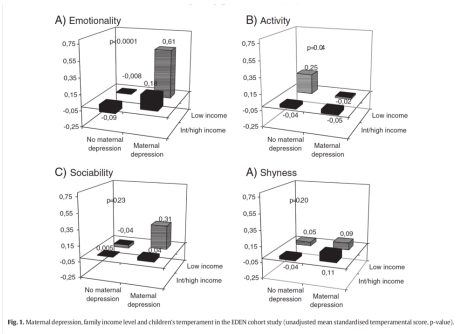


Fig. 1. Maternal depression, family income level and children's temperament in the EDEN cohort study (unadjusted mean standardised temperamental score, p-value).

Figure 3.17: Bad visual 1

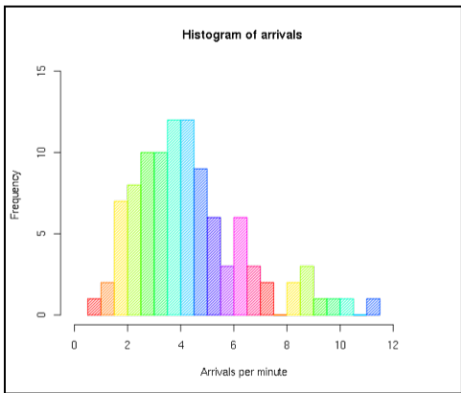


Figure 3.18: Bad visual 2

- Diverging palettes put equal emphasis on mid-range critical values and extremes at both ends of the data range -Qualitative palettes do not imply magnitude differences between legend classes, and hues are used to create the primary visual differences between classes.
- Do not forget our color-blind friends !

What summary statistics should you use: tables, graphics or statistical parameters? It all depends.

Depend on the objective

- Return the data
- “take home message”

Depend on the audience

- Expert
- Everybody else

Depend on the data type

- Qualitative
- Quantitative

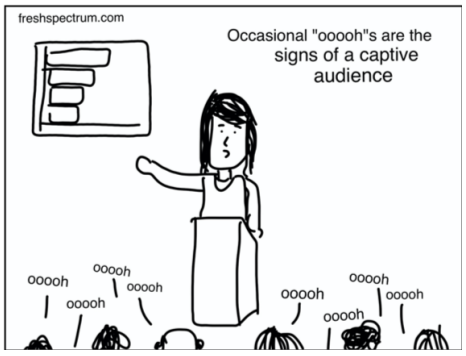


Figure 3.19: Communication tips

Chapter 4

Inference and sample

- Make the difference between individual variation and sample variation
- Make the difference between observed values and estimated parameters
- Interpret a confidence interval
- Interpret a statistical test

4.1 Sample

4.1.1 Population versus Sample

Why take a sample … to conclude on the population?

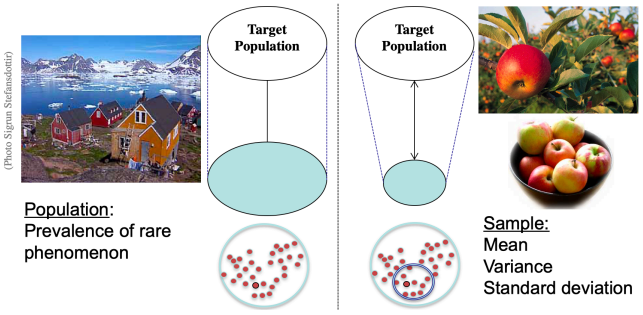


Figure 4.1: Population versus Sample

When the population is small and isolated, like an Inuit village in Greenland for instance, it is relevant to study/interrogate the entire population to get the *exact* prevalence of a rare disease (Figure 4.1).

However if the size of the population is large we might not be able (want) to get it all! It is time consuming, costs a lot, and we are smarter. Although we

will conclude on the population, a well design sample allows an *estimation* of the prevalence of the outcome of interest.

What is the main property of a good sample?

- Since we draw conclusions about a population based on the information obtained from a sample (subset of the population), it is important that the units of interest in the sample are **representative** of the entire population.
- A representative sample will allow to **confidently generalize** the results and the conclusion of your study

- Poor sampling designs can yield misleading conclusions

- Set of observations drawn from a population
- All the individuals of the study population should have the same probability of being drawn
- A `representative sample` should be an unbiased reflection of what the targeted population is like for the units of interest (outcome and covariates/determinants)
- A representative sample for a variable should reflect the variable distribution observed in the targeted population.

For example, you plan to determine the relationship between gratitude and job satisfaction in gynecologists. Your sample might consist of 30 to 40 gynecologists. Your population might be "gynecologists in the United States", or, if the scope of your study was more narrow, "gynecologists in New York City". So, if most gynecologists in the population are women, but your sample is all male, you do not have a good case for representativeness because your sample does not share the same characteristics as the larger population. In this case, you cannot generalize the results of your study to the population.

4.1.2 Sample designs

Element selection technique	*Probability sampling*	*Non-probability sampling*
Unrestricted sampling	Simple random sampling	Convenience sampling (Voluntary response)
Restricted sampling	Complex random sampling (systematic sampling, stratified sampling)	Purposive sample (such as quota sampling)

4.1.3 Probability sampling

1. **Simple random sample** of size n consists of n individuals from the population chosen in such a way that every set of n individuals has an

equal chance to be the sample actually selected.

The simple random sampling is **the gold standard**: All individuals are given an equal chance to be chosen to be in the sample.

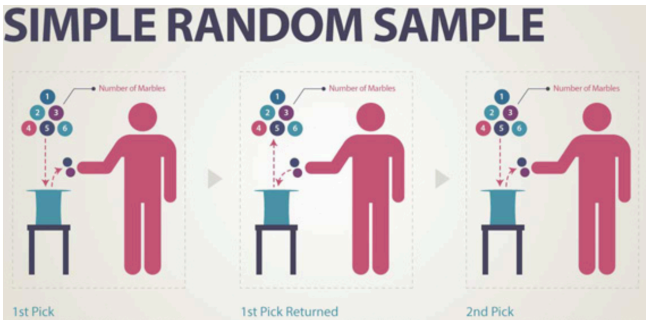


Figure 4.2: Simple random sample (source: Academic Web Services)

2. **Stratified random samples** (or block experiment) define by strata of similar characteristics and choose a separate simple random sample (SRS) in each stratum and combine these to form the full sample.

Figure 4.3: Stratified sampling (source: Academic Web Services)

3. **Cluster and multistage samples** consist of all or random units within clusters (hierarchical sampling)

For example, the HBSC study is based on a multistage sampling. In France, in each region some schools were randomly selected, within a school some classes were randomly selected and all the students of the selected class were interrogate.

The Dean of the School of public health wants to meet with several members of the faculty of the Department of Statistics to discuss concerns they might have about their department. He does not have time to talk to all of the faculty members, but wants to be sure that he meets with faculty with various levels of

Figure 4.4: Cluster sample (source: Academic Web Services)

seniority.
The Department of Statistics is composed of:

18 Professors
3 Associate Professors
5 Assistant Professors

The Dean decides to speak with 5 Professors, 2 Associate Professors, and 3 Assistant Professors. What sampling design can be used?

4. **Systematic sampling**

Sometimes we draw a sample by selecting individuals systematically. For example, a systematic sample might select every tenth person entering a school cafeteria (price, food quality, opening hours…). To make sure our sample is random, we still must start the systematic selection with a randomly selected individual, not necessarily the first person entering the cafeteria at lunchtime. When there is no reason to believe that the order of people entering the pool could be associated in any way with the responses measured.

4.1.4 Non-probability sampling

Voluntary response survey In a voluntary response sample, a large group of individuals is invited to respond, and all who do respond are counted. This method is used by call-in shows, Internet polls, Political party survey. Voluntary response samples are almost always biased, and so conclusions drawn from them are almost always wrong.

For example, the ABC news program Nightline once asked their viewers whether the United Nations should continue to have its headquarters in the United States. In order to have their opinions counted, viewers had to call a 1-900 number and pay a small fee. More than 186,000 callers responded and 67% said

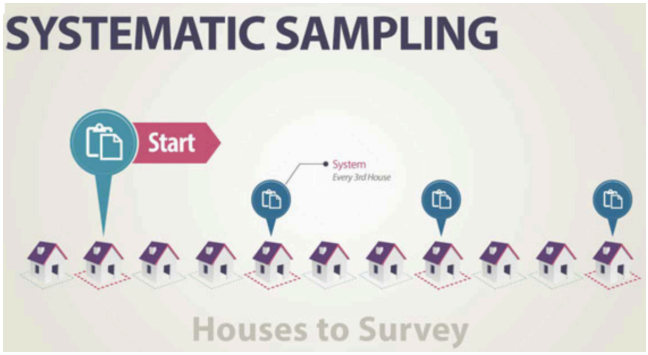


Figure 4.5: Systematic sampling (source: Academic Web Services)

"No." The callers of such program tend to be strong headed and often negative opinions. A nationwide poll with a proper sampling design found that less than 28 % of US adults want the UN to move out of the United States.

Convenience sampling chooses the individuals who are convenient, the easiest to reach to be in the sample. The group is probably not representative of the targeted population. This method is used in shopping area (street, mall). This population tend to be more affluent and include a larger percentage of teenagers and retirees than the population at large.

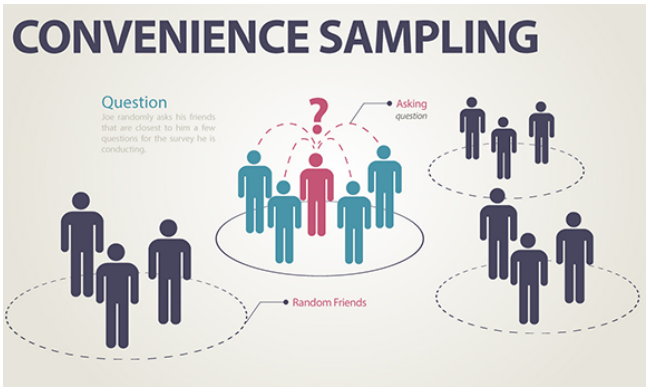


Figure 4.6: Convenience sample (source: Academic Web Services)

For example, you interview people at the mall or on a street corner about the French school policy that allows every cities to decide whether school schedule is spread over 4 or 5 days per week.

Other non-probability sampling methods exists but are largely bias. One has to be aware and truthful when concluding on such samples.

Method	Definition
Availability or Convenience Sampling	Cases selected because they're easy to find
Quota Sampling (proportional or not)	Groups defined by key characteristics; specified number of cases selected in each group.
Purposive or Expert Sampling	Individuals selected for sample because of their knowledge – "key informants"
Snowball Sampling	Start with initial sample, ask them to recommend other participants

Figure 4.7: Non-probability sampling methods

4.1.5 Sampling bias

When a sample is biased, the summary characteristics of a sample differ from the corresponding characteristics of the population it is trying to represent.

- Undercoverage: some portion of the population is not sampled at all or has a smaller representation in the sample than it has in the population.
- Non-response: those who don???t respond may differ from those who do
- Comparability
- Wording of questions: influence the answers by presenting one side of an issue consciously or not
- Response bias: tendency to please the interviewer

4.2 Confidence intervals

4.2.1 Within and between sample variation

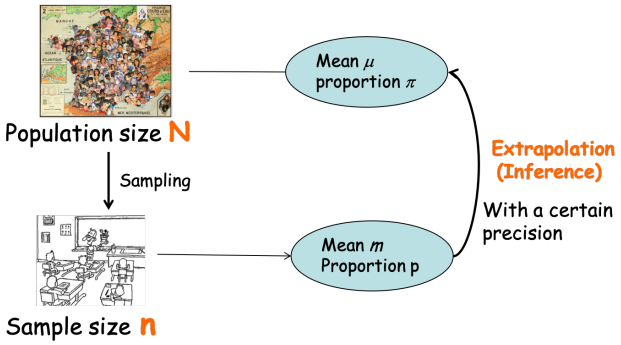


Figure 4.8: Parameter estimation from sample to population

As shown in Figure 4.9, individual measures vary with a sample. The standard deviation s is a indicator of that `individual fluctuation`. It is the average distance of the measures of the individuals to the mean.

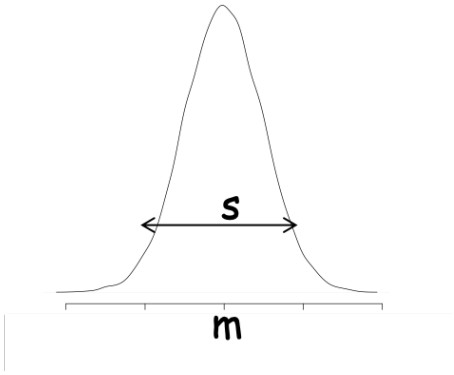


Figure 4.9: Variation of individuals within a sample

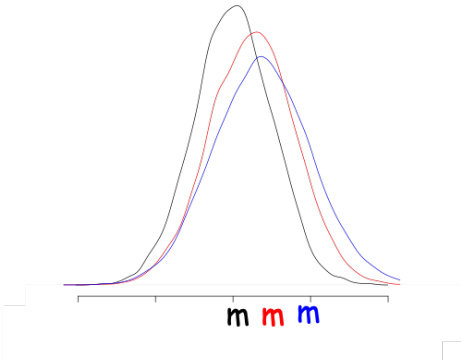


Figure 4.10: Variation between a samples drawn from the same population

If you drawn different samples from the same population you will get different means (Figure 4.10) due to `sampling fluctuation` or `sampling distribution of the mean`. This variance between the means of different samples can be estimated by the standard deviation of this sampling distribution and it is the standard error (SE) of the estimate of the mean.

The standard error (SE) is a type of standard deviation for the distribution of the means of samples drawn from the same population.

Standard Error:

$$\sigma_{\mu} = s/\sqrt{n}$$

where:

- σ_{μ} standard error of the mean
- s standard deviation of the sample (3.3.2)
- n size of the sample

4.2.2 The CLT and the confidence interval

The mean m in a sample of size n is a random variable, which varies between samples. This random variables should follow a Normal distribution centered around μ, the true mean of the population. This is the `central Limit Theorem`.

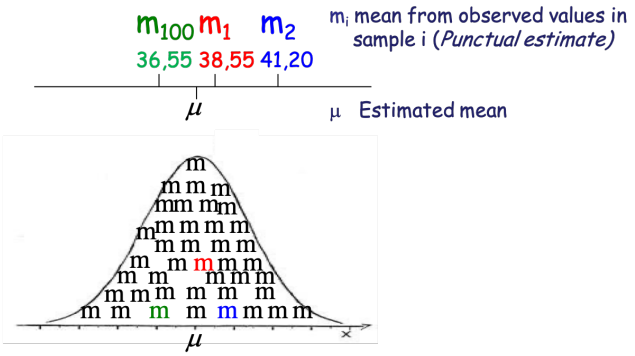


Figure 4.11: The observed mean is a random variable

Central Limit Theorem (CLT)

The mean of a random sample has a sampling distribution whose shape can be approximated by a Normal model. The larger the sample, the better the approximation will be.

For example, in the HBSC survey the observed mean of weights of 11 years-old French children in 2006 in our sample (n=500) is 38, 55 Kg (s = 8,10 Kg). How

do we estimate the **TRUE value** in the population of children aged 11 in 2006 in France?

Using central limit theorem, one can demonstrate that **we can be 95% confident that the true mean μ of the population from which is extracted our sample is within the interval**:

$m - 1.96 * \sigma_\mu \leq \mu \leq m + 1.96 * \sigma_\mu$

where:

- σ_μ standard error of the mean
- m mean of the sample
- 1.96 multiplier coefficient that depends on the confidence level expected

Following the example, the CI95% will be [37.40Kg-39.85Kg].

Similarly the proportion p in a sample of size n is a ramdon variable, which varies between samples. This random variables follows a Normal distribution centered around π.

For example, you are interested in the proportion of children (aged 2-3) with sleeping disorder in Isere, France (population: 14 000 children of 2 or 3 years old). You built a sample of 540 children among which 86 had sleeping disorder. The proportion of children with sleeping disorder in our sample is $p = 16\%$. But what is the TRUE proportion of children (aged 2-3) with sleeping disorder in Isere?

Using central limit theorem, one can demonstrate that **we can be 95% confident that the true proportion π of the population from which is extracted our sample is within the interval**:

$p - 1.96 * \sigma_\pi \leq \pi \leq p + 1.96 * \sigma_\pi$

where:

- $p = k/n$, k is the number of entities with the characteristic and n the size of the sample
- σ_π standard error of a proportion $\sqrt{p(1-p)/n}$
- 1.96 multiplier coefficient that depends on the confidence level expected

Following the example, the CI95% will be [12.9% ; 19.1%].

4.2.3 Interpretation of confidence intervals

The true mean is or is not in your estimated confidence interval. At 95% confidence, You took an 5%-risk of having the wrong confidence interval.

In Figure 4.12, 19 out of 20 samples (95%) from the same population will produce confidence intervals that contain the TRUE population parameter. There is a 95% probability that the calculated confidence interval encompasses the true value of the population parameter

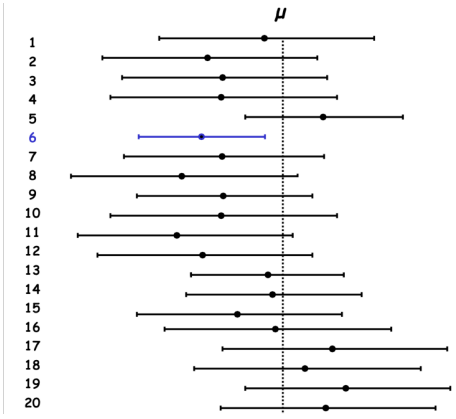


Figure 4.12: Confidence interval estimation from 20 samples drawn from the same population.

For example, in the HBSC sample and the age group [11-13[years old, the mean weight is 38.61Kg. But what is the true mean in the population?

```
##    n     mean       sd        se lower95ci upper95ci
##  167 38.60539 8.156085 0.6311368   37.3593  39.85148
```

where:

- n is the size of the sample
- sd standard deviation of the sample
- se standard error of population mean
- $lower95ci$ lower bound of the 95% confidence interval
- $upper95ci$ upper bound of the 95% confidence interval

We cannot give a unique true answer but only an interval (an estimation) with a certain confidence. We will say that in our sample the punctual estimate of the mean of weights for student aged 11 in France in 2006 was 38.60 kg and that we are 95% confident that in 2006, among the entire population of 11 year-old French student, the mean of weights is between 37.40Kg and 39.85Kg (CI95% [37.40Kg-39.85Kg]).

For the proportion example, we are 95% confident that the true proportion of children, aged 2-3, with sleeping disorder in Isere is between 12.9% and 19.1%.

4.2.4 Why 1.96?

We can generalize the equation above to:

$m - Z_\alpha \sigma_\mu \leq \mu \leq m + Z_\alpha \sigma_\mu$

where:

- Z_α is the critical value of the Normal distribution (centered on $\mu = 0$, reduced to $\sigma = 1$) where $100 - \alpha$ % of the values stand within Z standard deviations of the mean.

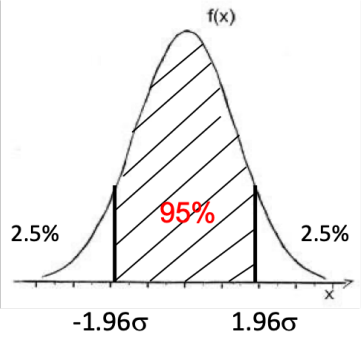


Figure 4.13: Centered reduced Normal distribution.

The critical value Z_α will be larger or smaller when computing a confidence interval at 90?

4.2.5 Precision or Margin error

The term $Z_\alpha \sigma_\mu$ or $Z_\alpha \sigma_\pi$ is noted i and named the `precision` or `margin error` of the confidence interval.

In a simple random sample, the precision is proportional to the square root of the inverse of the sample size. It is not linear!

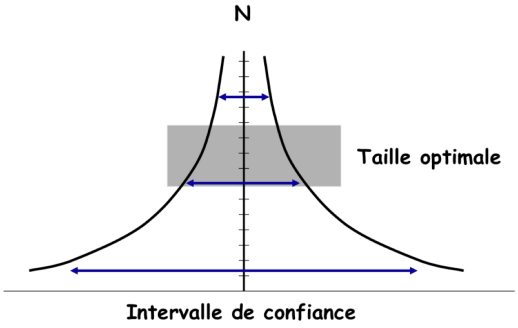


Figure 4.14: Precision to sample size relationship in a simple random sample.

For example, what should be the size of the sample to estimate the prevalence of overweight children aged 12, in Haute Savoie with a precision of i knowing from the literature that the estimate prevalence is 17%?

From the equation of the confidence interval of a proportion is $n = P(P - 1)Z_\alpha^2/i^2$

For a precision of 3%, $n = 0.17(1 - 0.17) * 1.96^2/0.03^2 = 627$

For a precision of 1%, $n = 0.17(1 - 0.17) * 1.96^2/0.01^2 = 5644$!!!

Chapter 5

Inference and statistical tests

- Understand statistical test reasoning
- Interpret results of a statistical test
- Discuss significance of a statistical test

The aim of a statistical test is to reach a scientific decision on a difference (or effect), on a probabilistic basis, based on observed data.

When assessing differences between groups (*Who*), you have to define *What* to compare. According to the type of the variable, you will choose a statistical parameter (mean, proportion, data distribution) to perform the comparison. The comparison will be based on hypothesis, with possible assumption to verify, and the associated statistical test.

In summary, the procedure is as follow:

1. Formulate hypothesis to be tested
2. Choose the appropriate statistical test
3. Calculate the appropriate statistic measure
4. Interpret the result

5.1 Formulate a hypothesis

In **hypothesis formulation** you always have two possibilities: **it is not different OR it is different**.

In the HBSC data, we are interested in the characteristics of the smoking students compare to the non-smoking. Do they differ by some characteristics?

Table 5.1: Description of the Height (cm) variable by smoking group (0=non-smoking, 1=smoking) in the French HBSC database in 2006.

SmokingStatus	Mean	SD	Median	Q1	Q3
0	157.89	12.00	159	149	166.25
1	166.26	11.39	168	159	173.00

Table 5.2: Proportion of students smoking (1) or non-smoking (0) by gender in the French HBSC database in 2006.

	Smoking status	
Gender	No	Yes
boy	85.26	14.74
girl	87.90	12.10

Example 5.1. Example 1: we would like to test whether there is, on average, a difference in height between smokers and non-smokers.

Example 5.2. Example 2: we would like to test if the proportion of smokers varies according to gender.

First, we describe the distribution of the variable between the groups.

In example 1, the height is a quantitative continuous variable which can be summarized by the mean (Table 5.1). In our sample, the students who do not smoke measure on average 157 cm while the students who smoke measure in average 166 cm. The question is "At the population level, is that different knowing that you do have individual variation (SD) and sample variations (SE)?".

In example 2, the gender is a qualitative variable which can be summarized into proportions (Table 5.2). In our sample, the proportion of students *seem* to different between groups. The question is "At the population level, are those proportions real different knowing that you do have individuals' variation (*sd*) and sampling variation (*se*)?".

In theory, we test whether the two groups (samples) come from the same population (Figure 5.1). For instance, sample 1 with mean *m1* from population 1 with μ_1 is coming from the same population as sample 2 with mean *m2* from population 1 with μ_2. Population 1 is equal to population 2.

When you state the hypothesis you should explicit H0 an H1.

Definition 5.1. H0 : The null hypothesis is that the parameters are EQUAL, i.e they are not different.

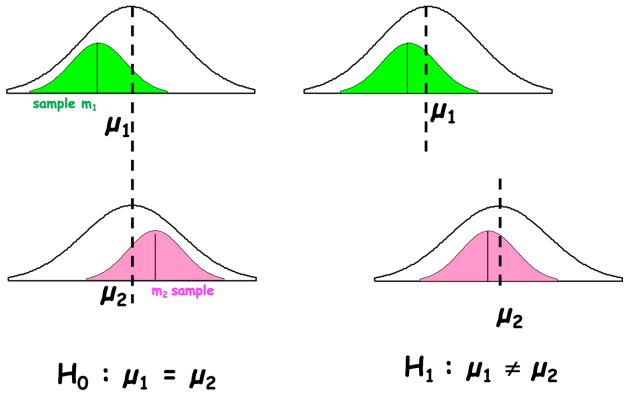


Figure 5.1: Population versus Sample

H1: The alternative hypothesis is that the parameters are NOT EQUAL, i.e they are different.

I like to write down the hypothesis with the $=$ sign and $\neq$ as I find easier to pick the test and interpret the results afterward.

Remember the **W's** they should appear in the hypothesis if you have the information.

Example 5.3. Example 1:

we would like to test whether there is, on average, a difference in height between smokers and non-smokers.

H0: In France in 2006, the mean height of the students 11-16 who smoke was equal to the mean height of the students 11-16 who do not smoke.

H1: In France in 2006, the mean height of the students 11-16 who smoke was NOT equal to the mean height of the students 11-16 who do not smoke.

Example 5.4. Example 2:

we would like to test if the proportion of smokers varies according to the groups of ages.

H0: In France in 2006, the proportion of student girls 11-16 who smoke was equal the proportion of student girls 11-16 who do not smoke and the proportion of student boys 11-16 who smoke was equal the proportion of student boys 11-16 who do not smoke.

H1: In France in 2006, the proportion of student girls 11-16 who smoke was NOT equal the proportion of student girls 11-16 who do not smoke and the proportion of student boys 11-16 who smoke was NOT equal the proportion of student boys 11-16 who do not smoke.

A statistical test is always performed to answer the H0 hypothesis (Figure 5.2).

- When we prove that the statistical parameters differ we reject H0 and accept H1. We say that we observed a statistically significant difference between the parameters.
- When we cannot prove that the statistical parameters differ, we stay under H0 and say that: we fail to reject H0 because we cannot show any statistically significant difference. H0 is never accepted as an error risk still exists that we can not compute.

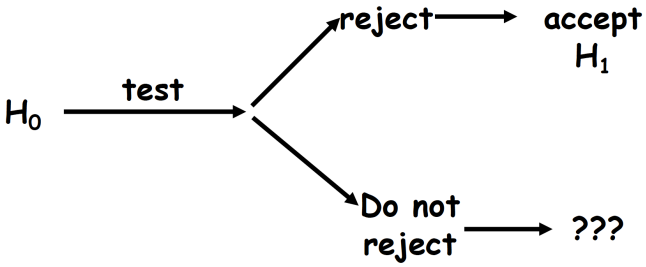


Figure 5.2: Statistical test interpretation

5.2 Comparison of two means

Once the hypothesis are stated, we choose a test. In the context if the comparison of means, the test will assess whether the observed difference (Δ) between the two groups is random (due to individuals' and sampling variations) or not (Figure 5.3).

In theory, if H0 is true $\Delta = m_1 - m_2 = 0$

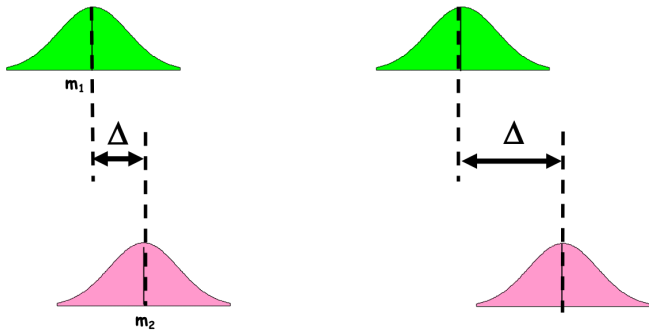


Figure 5.3: Population versus Sample

But we never compare Δ to 0 as we need to take into account the individuals' fluctuation (sd) and sampling (se) variation. What is the critical value, then ?

Can you guess ?

The critical value depends on the risk you are willing to take to conclude about a difference that does not exist in reality, the α risk.

When comparing means, you make the assumption that your sample's distributions are not too different from a Normal distribution. Therefore, the difference $\Delta = m_1 - m_2$ should follow a Normal distribution centered on 0. Then to take into account the individuals and sampling variation, the difference is standardized. The statistical value $(m_1 - m_2)/s_\Delta$ is computed and compare to the critical value Z_α of the centered reduced Normal distribution for a risk α. [Note s_Δ is function of the chosen test]

For a risk $\alpha = 5\%$ the critical value is $Z_\alpha = 1.96$.

If H0 is true, 95% of values of $(m_1 - m_2)/s_\Delta$ are between -1.96 and 1.96. If the statistical value $(m_1 - m_2)/s_\Delta$ returned by your test is above 1.96 or below -1.96, you reject H0 and accept H1.

In example 1, the statistical value for a risk $\alpha = 0.05$ is -5.4582. What is you conclusion?

5.3 Comparison of two proportions

In the context of proportion comparison like in example 2, the objective is to assess whether the proportion of cases among the exposed group is equal to the proportion of cases among the non-exposed group.

Example 5.5. Example 2:

we would like to test if the proportion of smokers varies according to the groups of ages.

H0: In France in 2006, the proportion of student girls 11-16 who smoke was equal the proportion of student girls 11-16 who do not smoke and the proportion of student boys 11-16 who smoke was equal the proportion of student boys 11-16 who do not smoke.

H1: In France in 2006, the proportion of student girls 11-16 who smoke was NOT equal the proportion of student girls 11-16 who do not smoke and the proportion of student boys 11-16 who smoke was NOT equal the proportion of student boys 11-16 who do not smoke.

To this aim, as for the comparison of means, we will compute the standardized differences (distances) between groups.

5.3.1 Chi-square test

The Chi-square is often the test used as rather intuitive and non computer greedy.

Table 5.3: Observed number of boys and girls students exposed or not to smoking in the French HBSC database in 2006.

	Smoking status		
Gender	No	Yes	Total
boy	214	37	251
girl	218	30	248
Total	432	67	499

Table 5.4: Expected number of boys and girls students exposed or not to smoking in the French HBSC database in 2006.

	Smoking status		
	No	Yes	Total
boy	?	?	251
girl	?	?	248
Total	432	67	499

First, we compute a two-way table with your `observed counts`:

Under H0 the absence of difference (independence assumption), we would expect to have the same proportion of patients among those exposed and those not exposed. So keeping the total margins what would be the expected counts? (Table 5.4)

To compute a theoretical two-way table with your `expected counts`:

- the proportion exposed student is: $p = 251/499 = 0.503$, *i.e* 50.3%.
- the number of boys students exposed would be: $p = (251/499) * 67 = 33.7$.

Then we compute the Chi-square (χ^2) statistic which is the standardized sum of the differences between the observed and the expected value.

Table 5.5: Expected number of boys and girls students exposed or not to smoking in the French HBSC database in 2006.

	Smoking status		
	No	Yes	Total
boy	?	(251/499)*67	251
girl	?	?	248
Total	432	67	499

Table 5.6: Expected number of boys and girls students exposed or not to smoking in the French HBSC database in 2006.

	Smoking status		
	No	Yes	Total
boy	217.3	33.7	251
girl	214.7	33.3	248
Total	432	67	499

$\chi^2_{Obs} = \sum(\frac{(Obs - Exp)^2}{Exp}$

Using the R statistical software, we have

```
chisq.test(hbsc$Gender, hbsc$SmokingStatus)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  hbsc$Gender and hbsc$SmokingStatus
## X-squared = 0.54013, df = 1, p-value = 0.4624
```

The $\chi^2_{Obs} = 0.54$. Now the question is:

"is it equal to 0?"

As for the comparison of means, in theory, if H0 is true $\chi^2_{Obs} \sim 0$ but it is never 0. There are variations and we test H0 with a α risk. Therefore, what is the threshold?

To define that threshold, we need to choose the correct statistical law. The χ^2 distribution depends on k, the number of degrees of freedom (df) which depends on the number of characteristics of the two variables we are comparing (Figure 5.4).

When the two-way table of the expected number is drawn, the degree of freedom is the number of values in the final calculation that are free to vary. For instance, in a 2x2 table once you have set one value the others cannot change. The quick formula to compute the degree of freedom (df) for the χ^2 distribution is the number of rows in the table minus 1 multiply by the number of columns in the table minus 1:

df = (#rows - 1) * (#cols - 1)

For a 2x3 table, what is the degree of freedom?

Next, we need to look at the statistical table of the χ^2 law (Figure 5.5). The threshold value also depends on the α risk you are willing to take.

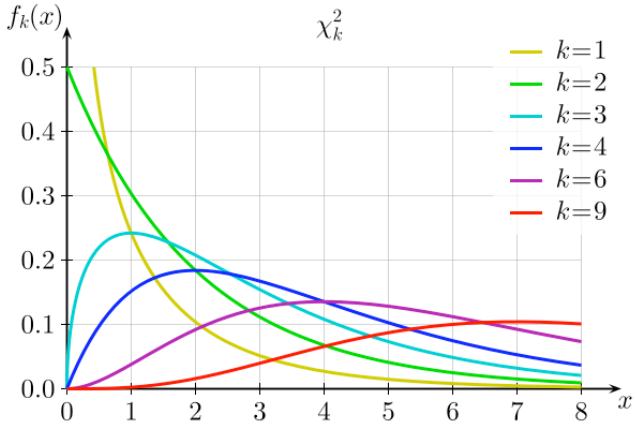


Figure 5.4: Chi2 distribution depends on the degree of fredom K

	α					
v	0.100	0.050	0.025	0.010	0.005	0.001
1	2.7055	3.8415	5.0239	6.6349	7.8794	10.8276
2	4.6052	5.9915	7.3778	9.2103	10.5966	13.8155
3	6.2514	7.8147	9.3484	11.3449	12.8382	16.2662
4	7.7794	9.4877	11.1433	13.2767	14.8603	18.4668
5	9.2364	11.0705	12.8325	15.0863	16.7496	20.5150
6	10.6446	12.5916	14.4494	16.8119	18.5476	22.4577
7	12.0170	14.0671	16.0128	18.4753	20.2777	24.3219
8	13.3616	15.5073	17.5345	20.0902	21.9550	26.1245
9	14.6837	16.9190	19.0228	21.6660	23.5894	27.8772
10	15.9872	18.3070	20.4832	23.2093	25.1882	29.5883
11	17.2750	19.6751	21.9200	24.7250	26.7568	31.2641
12	18.5493	21.0261	23.3367	26.2170	28.2995	32.9095
13	19.8119	22.3620	24.7356	27.6882	29.8195	34.5282
14	21.0641	23.6848	26.1189	29.1412	31.3193	36.1233
15	22.3071	24.9958	27.4884	30.5779	32.8013	37.6973

Figure 5.5: Chi2 distribution depends on the degree of fredom K

Table 5.7: Fisher Exact test principal.

	Smoking status		
	No	Yes	Total
boy	a	c	n1
girl	b	d	n2
Total	t1	t2	N

In the χ^2 table, for our 2x2 table the $df = 1$ and for a α risk of 5%, the $\chi^2_{Theo} = 3.84$.

Next the decision rule is the same as for the comparison of means.

In example 2, the statistical value for a risk $\alpha = 0.05$ and $df = 1$ is 3.84. The $\chi^2_{Obs} = 0.54$. What is you conclusion?

To correctly use the χ^2 test and have accurate estimation of the associated probabilities, we need to have at least **n=5** count in each cell of the **table of expected** numbers.

5.3.2 Fisher's Exact test

To compare proportions the Fisher's Exact test is even better than χ^2 as it computes the exact probability of obtaining a difference even greater if H0 is true. However the formula is more complex and difficult to compute by hand . A computer is highly recommended (Table 5.7).

$$p = \frac{n_1!n_2!t_1!t_2!}{a!b!c!d!}$$

Using the R statistical software, we have

```
fisher.test(hbsc$Gender, hbsc$SmokingStatus)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  hbsc$Gender and hbsc$SmokingStatus
## p-value = 0.4316
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.4571282 1.3779096
## sample estimates:
## odds ratio
##  0.7962999
```

How to read the output of the Fisher's Exact test?

1) We can look at the odds ratio and the confidence interval (95% CI)

As it is a ratio (numerator/denominator) if there is no difference (numerator=denominator) the $oddsratio \sim 1$. However it is never 1, we need to look at the confidence interval (at a certain risk level) to conclude.

- If the confidence interval includes 1, we fail to reject H0, we can not conclude that the proportions differ.
- If the confidence interval does not include 1, we reject H0, the proportions differ.

In example 2, the Fisher's exact test returns an odds ratio of 0.79 and 95% CI [0.45, 1.37]. What is your conclusion?

2) We can look at the $p-value$ but what is a $p-value$? See next section 5.4.

5.4 Risk α and $p-value$

The demonstrations above and the use of Z_α or χ^2_α values are valid for comparison of means or proportions under some assumptions and conditions (Z_α with sample size above 30 in each group; χ^2_α function df...). What is happening for other tests?

The philosophy is exactly the same but the critical value might come from other statistical laws than the Normal law (Chi-square, Binomial, Poisson...). It might be difficult to retrieve the critical value need to compare to your computed statistical value. There are more statistical tables that there are statistical tests.

The common practice is then to compare the risk α, defined *a priori*, to the $p-value$ returned by the test *a posteriori*.We want to know the ultimate risk that is taken. Meaning, the risk corresponding to the value found by the test.

α risk and $p-value$

α risk: *a priori* risk to conclude about a difference that does not exist in reality

$p-value$: *a posteriori* error risk that is taken knowing the result of the test

Statistical test's decision rule

When $p-value > \alpha$ risk, we FAIL to reject H0

When $p-value \leq \alpha$ risk, we reject H0

The $p-value$ is not synonymous of the importance of the possible difference between groups. In other words, a very small p-value means that the risk of making a mistake is very low. It does not mean that there is a huge difference between groups.

Example 5.6. Example 3: Using 2 studies we are assessing 2 new methods (A and B) for the prevention of surgical site infections (SSI) compared to a conventional method (0).

- In study A: method A shows 12% of SSI and the conventional method 24% of SSI. The test between A and 0 returns a $p-value \leq 0.05$.
- In study B: method B shows 12% of SSI and the conventional method 24% of SSI. The test between B and 0 returns a $p-value \leq 0.001$

Is method B better than method A in preventing SSI?

In the example above, we cannot tell if method B is better than method A : we did not test A versus B. The only information we get is that method A is different from method 0 and that method B is different from method 0. In fact method A and B present the same level of SSI. They might not be different but to conclude (with a certain level of confidence) we need to do a test. (see section 5.7 for comparison of multiple groups)

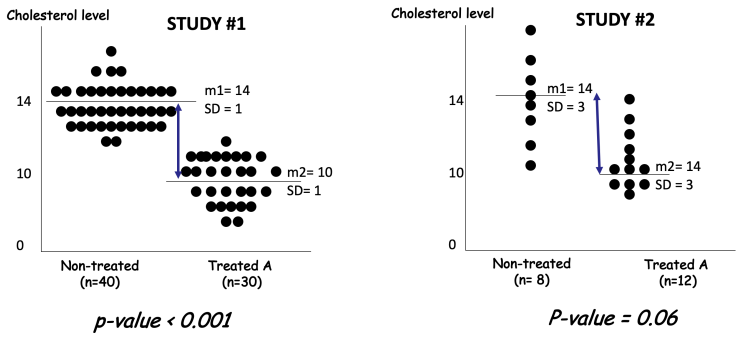
In example 1, we tested H0 with the Student's T test. The $p-value$ is 4.342e-07. What is your conclusion?

In example 2, we tested H0 with a χ^2 test. The $p-value$ is 0.462. What is your conclusion?

We also tested H0 with the Fisher's Exact test. The $p-value$ is 0.431. Do you reach the same conclusion than with the χ^2 test? Why?

If you try to compare the effects of different risk factors or compare similar studies but with different protocols, you sholud not compare the *p-values*. A *p-value* smaller than an other *p-value* does not mean that the observed difference is greater. It only means that your are more confident on the results of the test.

You should not compare p-values.



In the cholestrol studies presented above, the *p-values* are different, one is smaller than the other (even significant). Although the observed differences in mean are identical. The difference is the same in study 1 and 2 ($\delta = 4$). The differences between the two studies are on the standard deviations of the samples and the sample sizes which affect the t-statistics.

5.5 Risk α and risk β

Why do we not accept H0?

As mentioned earlier there is always a risk of being wrong (Figure 5.6) but that risk cannot be computed. It is β.

In row the unknown truth, in column the conclusion of the test.

	Reject H0	Do not reject H0
H0 is true	α	$1-\alpha$
H1 is true	$1-\beta$	β

Figure 5.6: Where do the risks stand?

- α is the probability of rejecting H0, when H0 is true (Figure 5.7)
- β is the probability of failing to reject H0, when H1 is true (Figure 5.8)

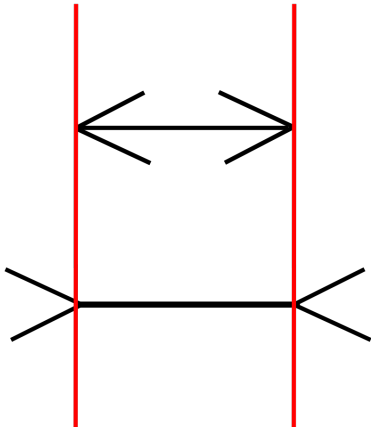


Figure 5.7: Population versus when it does not exist

Imagine that you are the "eyes" and you are unable to see the perspective above the horizon. You will believe that the 2 lines make 1 and that they are of the same length but in reality the further away is longer. That is the β risk.

What do you prefer α or β ?

It is a difficult question.

Example 5.7. Example 4: Hepatitis B vaccination and multiple sclerosis. Many study protocols have been conduct to assess an possible association between vaccination against Hepatitis B and the development of multiple sclero-

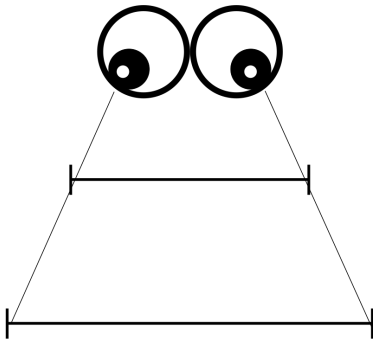


Figure 5.8: Not seeing a difference when it does exist.

sis. The null hypothesis was that prevalence of multiple sclerosis was the same among people vaccinated against Hepatitis B and people not vaccinated against Hepatitis B.

In that context, what would you favour α or β?

- Wrongly reject H0 and conclude that there is an effect : it could be catastrophic as the immunization coverage will fall down
- Not seeing an effect: people may not get Hepatitis B vaccination but multiple sclerosis !!!

Note that there have been many studies on the above question and that no effect as been seen so far.

You test the hypothesis of an absence of difference in school grades between gender. What is your conclusion if:

- $p - value = 0.049$
- $p - value = 0.051$

Comment on your conclusions.

5.6 Comparison of multiple groups

Example 5.8. Example 5: We are interested in the effect of 2 treatments to gain weight. The protocol include treatment A, treatment B, and a placebo group. We would like to compare the weights between the 3 groups.

How do we compare the means? Can we do 2 by 2 comparisons?

Table 1: Effects of 2 treatments to gain weight

Parameters	Treatment A	Treatment B	Placebo
sample size n	62	62	96
mean weight	66.06	63.83	62.32
sd weight	3.18	2.16	3.12

To the above question the answer is "No" : it increases the likelihood of incorrectly concluding that there are statistically significant differences, since each comparison adds to the probability of a type I error, α.

At the end, if k is the number of comparisons, the error rate becomes $1-(0.95)^k$

5.6.1 Graphical comparison

A boxplot and whisker plot (section 3.4) is an ideal graphical representation to compare data series of the same variable between different groups.

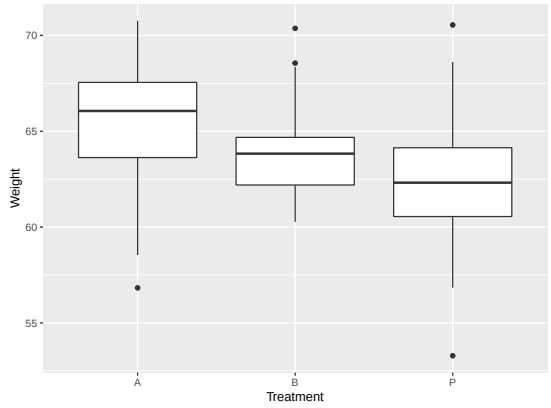


Figure 5.9: Boxplot and whisker plot of the effect of different treatments on gain weight

Figure 5.9 shows that the distributions seems to differ. The median (black line with the box) are located at differ weight.

What are the IQR of the 3 groups?

In Figure 5.9 the medians are not in the middle of the box. This suggest that the distributions might be skewed.

Figure 5.10 presents density plots, similar to histograms, and reveals the same thing.

The next step is thus to statistically test the hypothesis of equality of means.

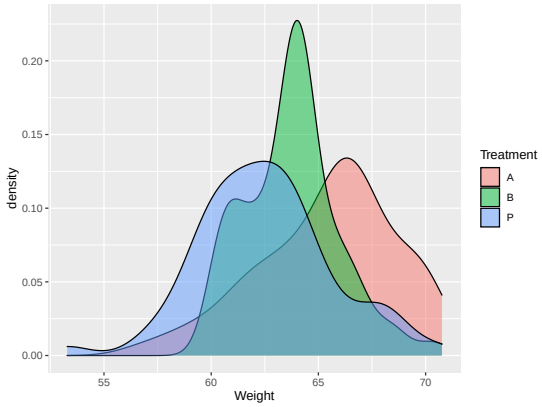


Figure 5.10: Density plots of the effect of different treatments on gain weight

5.6.2 Analysis Of Variance

The Analysis Of Variance or ANOVA allows comparing multiple groups

The Analysis Of Variance or ANOVA systematically compare variability *within* and *between* groups. When the variations observed between groups is greater than the within group variation, at least one group is differ from the other.

The statistical hypotheses are:
- H0: $\mu 1 = \mu 2 = \mu 3... = \mu k$ with α=5%
- H1: At least one mean is different from the other

where k is the number of independent groups

To that aim we use the F-test (named in honor of Sir Ronald Fisher). The F-statistic is a ratio of two variances that examine variability.

$F = \frac{MeanSquareBetween}{MeanSquareError} = \frac{MSB}{MSE}$

- between groups being compared (Mean Square Between or Mean Square Treatment)
- within the groups being compared (Mean Square Error or Mean Square Residuals)

Mean Squares	Sums of Squares (SS)	DF
MSB = SSB/(k-1)	$SSB = \sum_{i=1}^{k} n_i(\bar{X}_i - \bar{X})^2$	k-1
MSE = SSE/(N-k)	$SSE = \sum_{i=1}^{k} \sum_{j=1}^{n} (X_{ij} - \bar{X}_i)^2$	N-k
total	$SST = \sum_{i=1}^{k} \sum_{j=1}^{n} (X_{ij} - \bar{X})^2$	N-1

where

DF = degree of freedom X_{ij} = individual observation j in treatment i
$\bar{X}_i$ = sample mean of the i^{th} treatment (or group/sample)
$\bar{X}$ = overall sample mean
k = number of treatments or independent groups
n = number of observations in treatment i N = total number of observations or total sample size

In practice with R

```
res <- aov(Weight ~ Treatment, data=treatFrame)
summary(res)
```

```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## Treatment     2  270.5  135.26   16.48 2.62e-07 ***
## Residuals   183 1501.7    8.21
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The F-value is the F statistics. The F-value = 16.48 (=135.26/8.21). To conclude under H0 and a given α risk, we can try to look for the appropriate statistical law and its associated table or we can conclude using the p-value.

The $p - value = 2.62e^{-07}$. What is your conclusion?

5.6.3 Post-hoc analysis and ANOVA assumptions

The ANOVA results might help you conclude that a least one group varies differently that the others. However you will not know which group. To that aim you, need to perform a post-hoc analysis using the TukeyHSD's test.

In practice with R

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Weight ~ Treatment, data = treatFrame)
##
## $Treatment
##          diff       lwr         upr     p adj
## B-A -1.667323 -2.883055 -0.45159037 0.0040296
## P-A -2.945509 -4.161241 -1.72977619 0.0000001
## P-B -1.278186 -2.493918 -0.06245338 0.0367906
```

The output of the TukeyHSD's test presents adjusted p-values (p adj) for the two groups comparisons. In addition is displayed the difference in means (diff) and the lower (lwr) and upper (upr) bounds of the 95% CI on the difference in means.

In example 5, Owing the TukeyHSD's test, it seems that all means are differ from one another.

That is if we trust the appropriate use of the ANOVA and TukeyHSD's tests...

The ANOVA analysis relies on several **assumptions that need to be tested before** computing the ANOVA. The ANOVA formula is a based on mean and variance that can only be used if the distributions are not too different from the Normal distribution. It is a parametric test (see section 5.7). Before the ANOVA we need to verify with prior statistical tests that:

1. the outcome variable should be normally distributed within each group (*Shapiro test*)
2. the variance in each group should be similar (e.g. *Bartlett or Levene test*)
3. the observations are independent (not correlated or related to each other)

However, the F-test is fairly resistant or robust to violations of assumptions 1 and 2.

5.7 Parametric and non-parametric test

Parametric tests are those that make assumptions about the parameters of the population distribution from which the sample is drawn. This is often the assumption that the population data are normally distributed. Non-parametric tests are "distribution-free" and, as such, can be used for non-Normal variables. Non-parametric tests are often based on the ranking of the values in the data serie.

For the non-parametric the hypothesis are:

H0: The two samples are from the same distribution

H1: one distribution is shifted in location higher or lower than the other

While for the parametric interested in comparing means the hypothesis are:

H0: The two samples have the same mean

H1: The two samples do not have the same mean

5.7.1 Asessing Normality

1. Graphically

Normality can be assess using an histogram or the cumulative distribution function.

In figure @ref(fig:normality_plotfunction), the data normally distributed are on the left where the histogram is symmetric and the cumulative distribution function has a S shape. On the right, the data not normally distributed present a skewed histogram and cumulative distribution function different from a S shape.

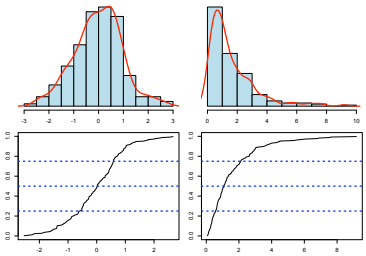


Figure 5.11: Assessing normality using graphics (#fig:distribution function)

Note that in the skewed distribution, the mean will be largely different from the median and the mode parameters.

2. Statistically

The Shapiro-Wilk test or Kolgomorov-Sminorv can be used do verify the normality a distribution. THe statistical hypothesis is:

H0: The sample distribution is equal to the Normal distribution

H1: The sample distribution is different from the Normal distribution

As an example, we randomly generate values drawn from a Normal distribution and test the Normality.

```
# Generating numbers following a Normal distribution
data <- rnorm(100)
# Using Shapiro-Wilk's test
shapiro.test(data)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data
## W = 0.99103, p-value = 0.7475
```

```
# Using Kolgomorov-Smirnov's test
ks.test(data, "pnorm")
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  data
## D = 0.094664, p-value = 0.3316
## alternative hypothesis: two-sided
```

What is your conclusion?

Note that the advantage of the Kolgomorov-Smirnov's test is that it can help assessing other type of distributions than the Normal distribution.

```
# Generating numbers following an Uniform distribution
data <- runif(100)
# Using Kolgomorov-Smirnov's test for Normality
ks.test(data, "pnorm")
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  data
## D = 0.50342, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

```
# Using Kolgomorov-Smirnov's test for Uniform distribution
ks.test(data, "punif")
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  data
## D = 0.050052, p-value = 0.9636
## alternative hypothesis: two-sided
```

What are your conclusions?

5.7.2 Two-sample Wilcoxon test (or Mann-Whitney U test)

- Given two samples X et Y
- Sort in increasing order the data from X and Y
- Give a rank to the values
- Compute the sum of rank for each sample R1 and R2
- Compute the random variable Un1,n2 = min(Un1, Un2) and the associated Z statistics

$$U_{n1} = n_1 \times n_2 + \frac{n_1(n_1+1)}{2} - R_1 \quad U_{n2} = n_1 \times n_2 + \frac{n_2(n_2+1)}{2} - R_2$$

$$Z_{n1,n2} = \frac{U_{n1,n2} - (n_1 \times n_2)/2}{\sqrt{(n1 \times n2)(n_1 + n_2 + 1)/12}}$$

If n1 and n2 < 20, Mann-Whitney's table

IF n1 and n2 > 20, Normal law table

Example 5.9. Example:

Among diabetes patients, is there a difference in age at diagnosis between men and women?

- Women: 20 11 17 12
- Men: 19 22 16 29 24

Sorting data: 11 12 16 17 19 20 22 24 29

R1 = 1+2+4+6 = 13 and Un1 = 17

R2 = 3+5+7+8+9 = 32 and Un2= 3

Table de Mann-Whitney p-value =0.11

What is you conlusion?

5.7.3 Which test to use?

Non-parametric tests are valid for both non-Normally distributed data and Normally distributed data, so why not use them all the time?

When it is possible to perform both a parametric and a non-parametric test (because we have quantitative measurements) reducing the data to ranks and using the Wilcoxon/Mann-Whitney test will have about 95% of the power of a corresponding two-sample t-test. And, as we have seen, often outliers are interesting in their own right. An analysis that simply ignores them may miss an important fact or instance.

Parametric	Non-Parametric equivalent
Paired t-test	Wilcoxon rank sum test
Unpaired t-test	Mann-Whitney U test
Pearson correlation	Spearmann Correlation
One analysis of variance	Kruskal Wallis test

Chapter 6

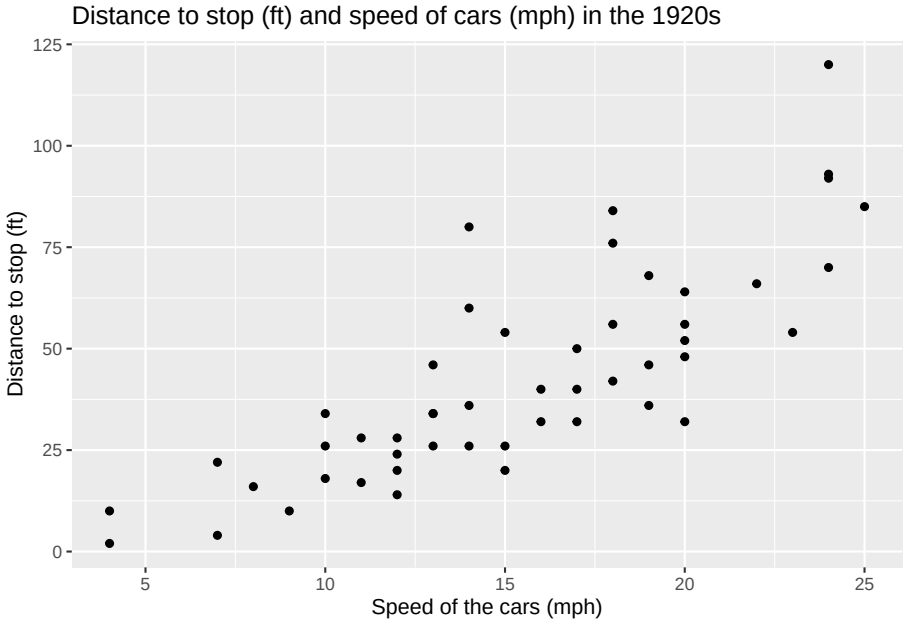
Introduction to regression modelling

- the step by step procedure to obtain an informative linear model, assess its validity and interpret it
- Interpret results of a simple linear model
- Discuss significance of a simple linear model

6.1 Simple linear regression

A regression attempt to explain the variations observed for a variable by other variables

In the plot below, we try to explain the observed distances to stop for cars by the speed of the cars. Here between the two variables, we see a rather linear diagonal trend. We may be able to fit a **simple linear regression** model.



A simple linear regression model is a statistical model that attempt to fit a linear relationship between two variables: one to explain and one explanatory variable

By convention:

- **Y axis** presents the variable to explain also named the dependent variable, the outcome variable or the response variable
- **X axis** presents the explanatory variable also named the independent variable or the predictor variable.

A first measure of linear relationship, often wrongly used, is the coefficient of correlation.

6.1.1 Pearson's coefficient of correlation

When we talk about correlation we often talk about the Pearson's coefficient of correlation that quantify the linear association between two quantitative variables.

Mathematically, the coefficient of correlation is the standardized covariance (Equation (6.1)). While the covariance is the average the product of the deviations of observed values to their mean (Equation (6.2)).

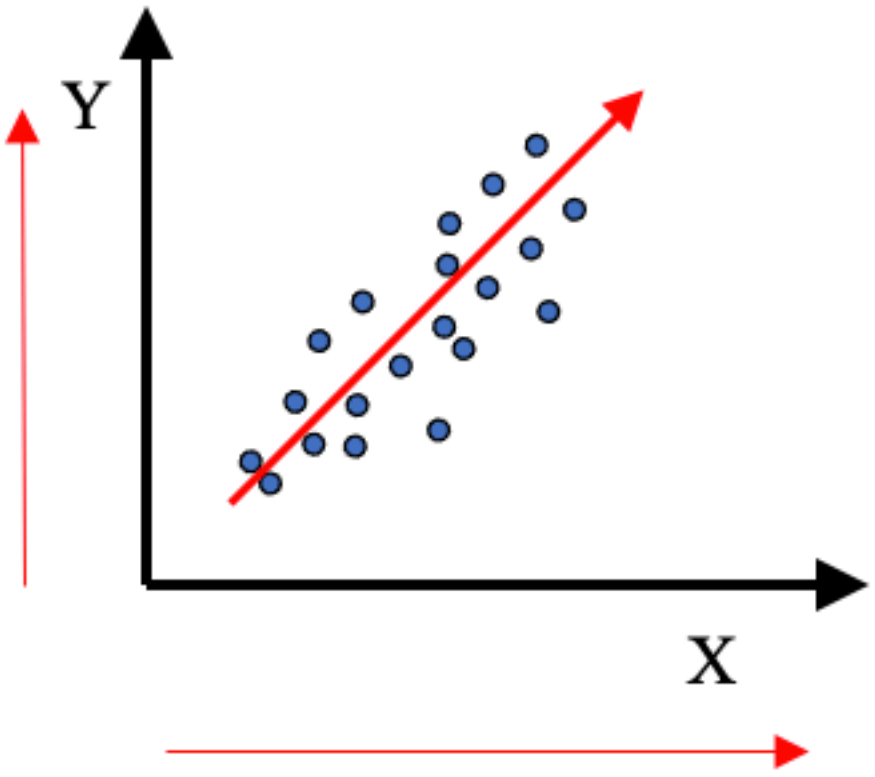
$$r = \frac{cov_{x,y}}{s_x s_y} \tag{6.1}$$

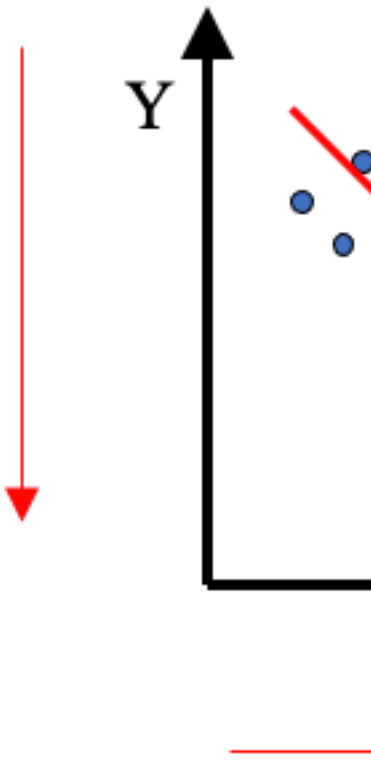
$$cov(x, y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{N - 1} \tag{6.2}$$

The coefficient of correlation r is :

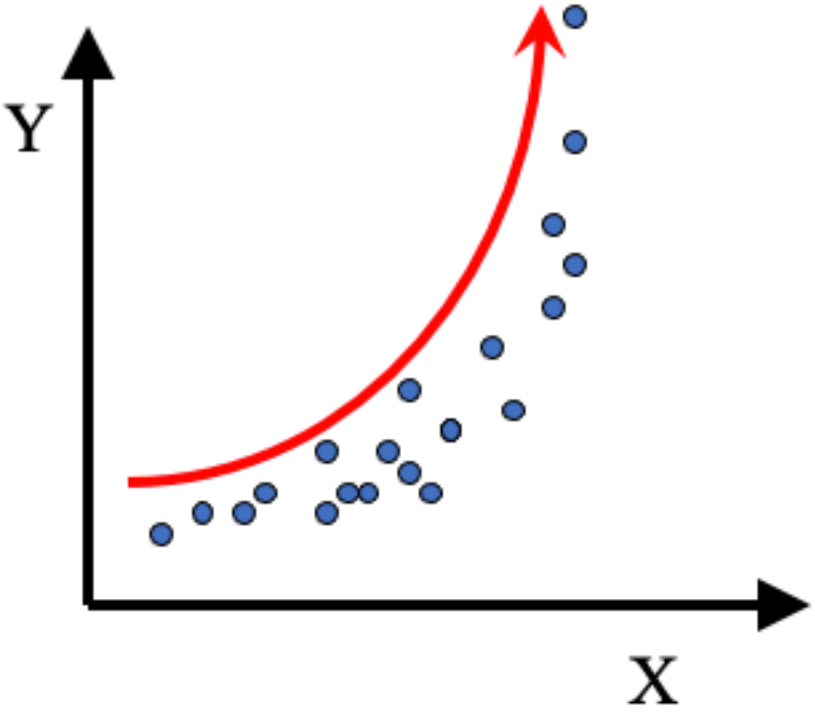
- Dimensionless
- Range between -1 and 1
- its absolute value presents the force of the relationship
- its sign presents the direction of the relationship

In the graphics below, what would be the estimated value of the Pearson's coefficient of correlation?

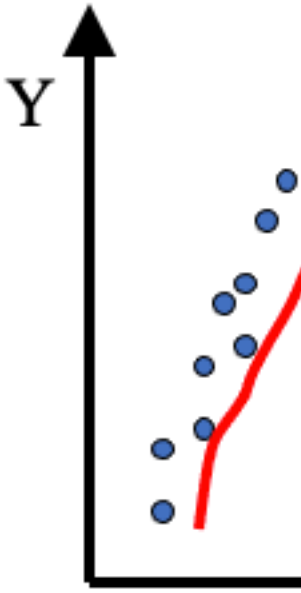




Absence of linear association (r=0) does not mean absence of relation.



Exponentional shape



Parabo

Correlation does not mean causation.

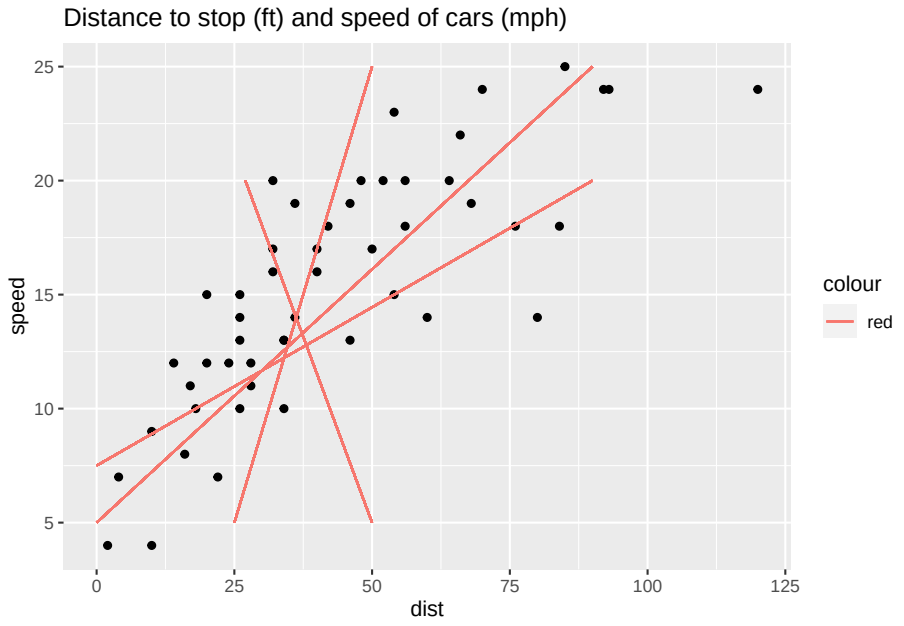
For details and other examples on too fast conclusions or spurious correlations see PennState Eberly College of Science [of Statistics PennState Eberly College of Science, 2021] and Tyler Vigen's Website [Vigen, 2021].

6.1.2 Simple linear regression model

A simple linear model fit a line between the points of the two variables. The final equation is of the form:

$$Y = \beta_1 X + \beta_0 + \epsilon \tag{6.3}$$

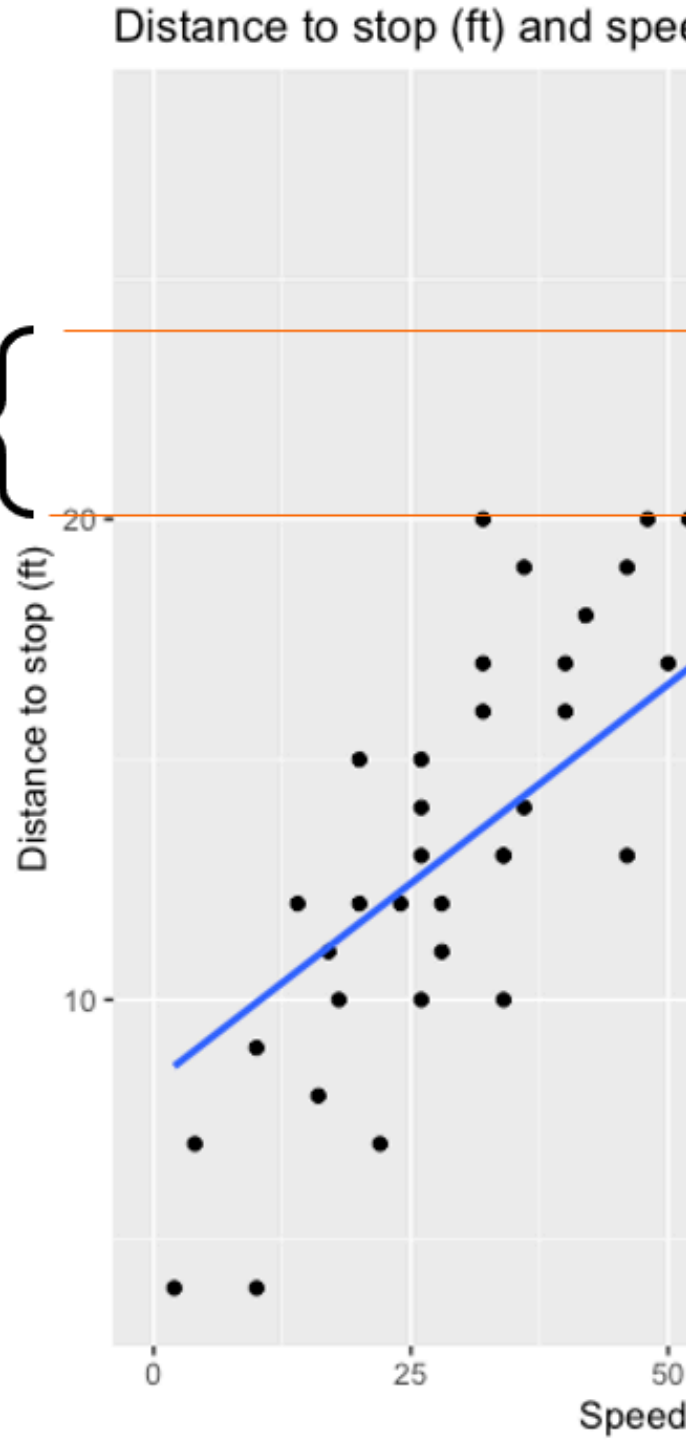
Which line fits best ?



Of all possible lines, the **regression line is the one that is the closest on average to all the points.**

$$\text{Residual} = Y_i - \hat{Y}_i$$

Where
Y_i : Observed value
$\hat{Y}_i$: Fitted value



The fitting of the regression line is based the least squares criteria. The aim is to minimize the sum of square distance between observed values and fitted values on the line (Equation (6.4)). The sum of square distance also named sum of square Error (SSE) or sum of square residuals (SSRes). When SSE is minimum it implies that the variance of the residuals that remain to be explained are fit to the minimum.

$$SSE = \sum_{i=1}^{n} (y_i - \widehat{y_i})^2 \tag{6.4}$$

To fit a simple linear regression with R follow the example below:

```
# load the data
data(cars)
# fit the line
m1 <- lm(dist ~ speed, data=cars)
# display the results
summary(m1)
```

```
##
## Call:
## lm(formula = dist ~ speed, data = cars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -29.069  -9.525  -2.272   9.215  43.201
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -17.5791     6.7584  -2.601   0.0123 *
## speed         3.9324     0.4155   9.464 1.49e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.38 on 48 degrees of freedom
## Multiple R-squared:  0.6511, Adjusted R-squared:  0.6438
## F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```

The output tells us that:

- Y(distance to stop) = 3.93*X(speed) -17.57(intercept) + ϵ
- If a car increases its speed by 1 mph, its distance to stop increases by 3.93 ft.
- This estimated coefficient is significantly different from 0 ($Pr(> |t|) = p - value = 1.49e^{-12}$)
- The model explain 64% of the observed variance in distances to stop for cars (Adjusted R-squared or coefficient of determination R^2).

The residuals are leftover of the outcome variance after fitting a model. They are use to:

- Verify if linear regression assumptions are met {6.1.3}
- Show how poorly a model represents data
- Reveal unexplained patterns in the data by the fitted model
- Could help improving the model in an exploratory way.

6.1.3 Post-hoc assumptions verification

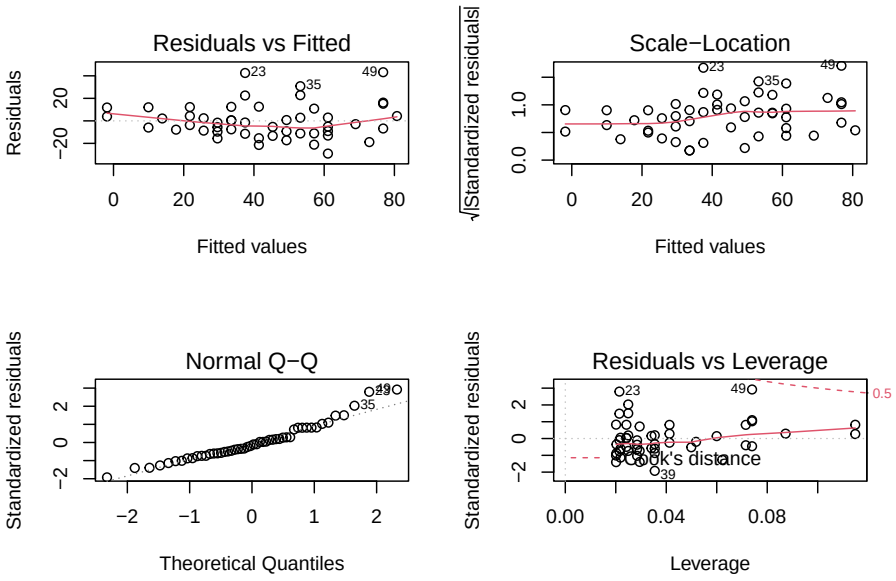
How to verify if a simple linear regression model was appropriate?

The method is descriptive and based on the distribution of the residuals that allows assessing:

- Linear and additivity of predictive relationships
- Homoscedasticity (constant variance) of residuals (errors)
- Normality of the residuals distribution
- Independence (lack of correlation) residuals (in particular, no correlation between consecutive errors in the case of time series data)

Using R:

```
# Define a plotting area with 4 panels
layout(matrix(1:4, nrow=2))
# Ask for diagnostic plots for linear regression
plot(m1)
```



```
# Set back the plotting area to 1 panel
dev.off()
```

```
FALSE null device
FALSE           1
```

From top to bottom and left to right the plots assess :

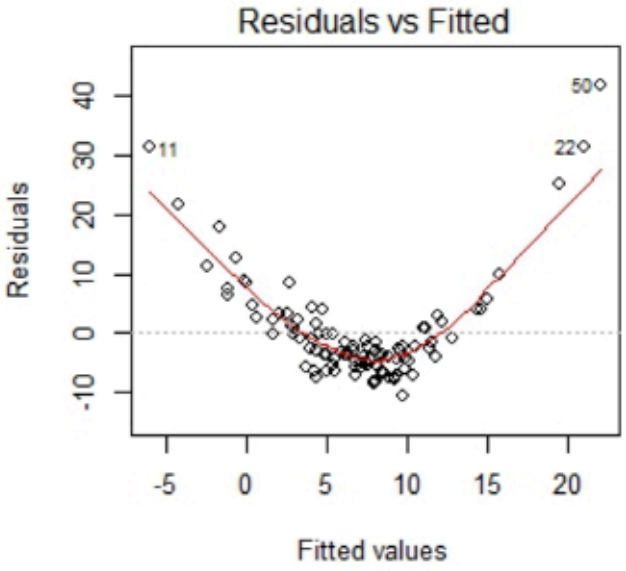
(1) Linearity and additivity of predictive relationships. A good fitting should show a red horizontal line, the variance of the residuals is randomly spread above and below the horizontal line.
(2) Homoscedasticity (*i.e.* constant variance) of residuals (errors). A good fitting should show a red horizontal line, the variance of the residuals is constant and do not depend on the fitted values.
(3) Normality of the residuals distribution. A good fitting should show a alignment of the dots on the diagonal of the Q-Q plot.
(4) Influential observations with Cook's distance: if dots appear outside the Cook's distance limits (red dashses) they are influential observations or even outliers (extreme). Their value need to be verified as they might negatively influence the fitting.

What do you think of the model above?

Below is an example of post-hoc assumptions verification of a linear model that shows problems. Here the model need to be refined or done differently.

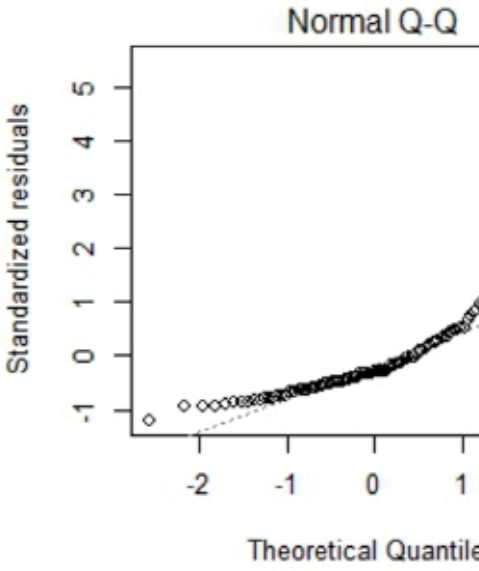
Linearity ?

More like curvarture

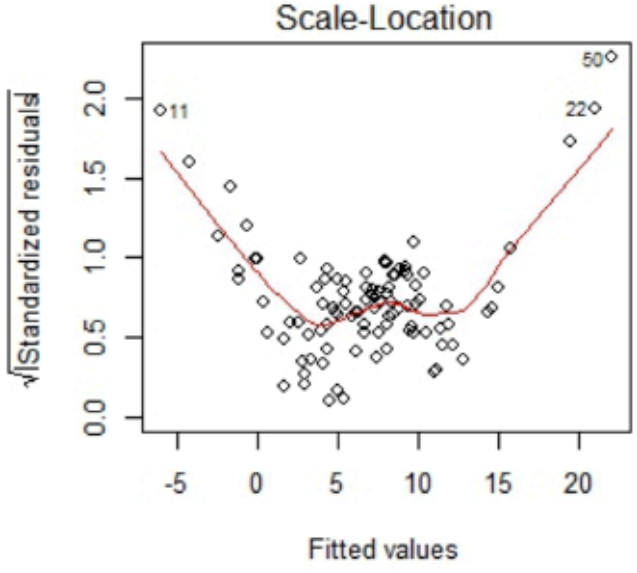


Normality of the e

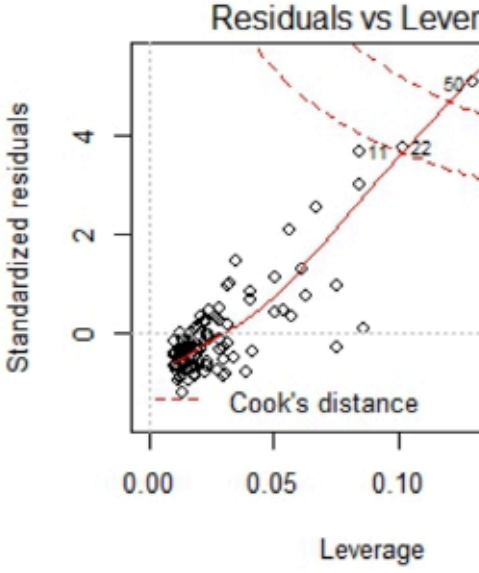
No



Homoscedasticity?



Influental point (c



6.2 Multiple linear regression model

A multiple (or multivariate) linear regression model is a statistical model that attempt to fit a linear relationship between one outcome variable to explain and several explanatory variables (determinants).

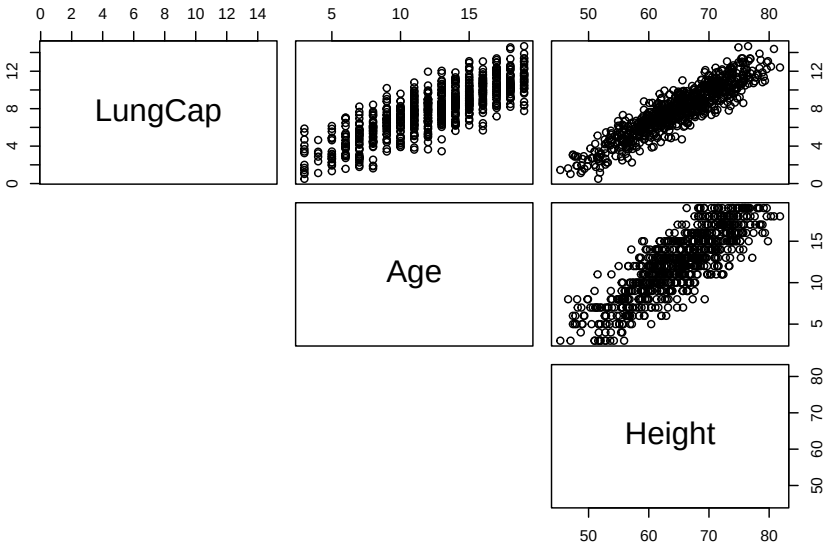
For instance, do your age and your height predict your lung capacity?

The table below presents the lung capacity dataset shared by Mike Marin and Ladan Hamadani and available at http://www.statslectures.com/ [Marin and Hamadani, 2021]. The first column is the lung capacity outcome variable and the other columns are some possible determinants.

LungCap	Age	Height	Smoke	Gender	Caesarean
6.475	6	62.1	no	male	no
10.125	18	74.7	yes	female	no
9.550	16	69.7	no	female	yes
11.125	14	71.0	no	male	no
4.800	5	56.9	no	male	no
6.225	11	58.7	no	female	no

First, we have a look at the two first covariates along with the outcome variable using a pair plot to visually assess any linear relationship.

```
pairs(LungCapData[,c("LungCap", "Age","Height")], lower.panel = NULL)
```



We could also look at each covariate against the outcome variable using a simple linear regression. However we will miss some information. Maybe it is better to

be tall than small to have a large lung capacity whatever your age. To test this hypothesis we should apply a multivariate regression on the ouctome variable.

Mathematically, **a multivariate regression is a generalization of a simple linear regression**.

From one predictor ...

$$y_i = \beta_0 + \beta_1 x_{i1} + \epsilon_i$$

... to two or more predictors

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} + \epsilon_i$$

and the independent error terms ϵ_i follow a normal distribution with mean 0 and equal variance σ^2.

The R function calls for a multiple regression is similar to a simple linear regression. First you build the model adding up the covariate on the right hand side of the equation. Next you display the summary and interpret the coefficients.

```
model3 = lm(LungCap ~ Age + Height, data= LungCapData)
summary(model3)
```

```
##
## Call:
## lm(formula = LungCap ~ Age + Height, data = LungCapData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4080 -0.7097 -0.0078  0.7167  3.1679
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -11.747065   0.476899 -24.632  < 2e-16 ***
## Age           0.126368   0.017851   7.079 3.45e-12 ***
## Height        0.278432   0.009926  28.051  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.056 on 722 degrees of freedom
## Multiple R-squared:  0.843,  Adjusted R-squared:  0.8425
## F-statistic:  1938 on 2 and 722 DF,  p-value: < 2.2e-16
```

- **P-values** for the *t-tests* appearing in the table of estimates suggest that the slope parameters for Age and Height are significantly different from 0 ($p - value < 0.05$).
- **Residual standard error** of 1.056 is rather small

- **F-statistic** (1938) is highly significant ($p - value :< 2.2e - 16$) implying that the model containing Age and Height is more useful in predicting lung capacity than not taking into account those 2 predictors. (Note that this does not tell us that the model with the 2 predictors is the best model!)
- **Multiple R squared**: R-squared measures the amount of variation in the response variable that can be explained by the predictor variable. The multiple R-squared is for models with multiple predictor variables. When adding up predictors, the multiple Rsquared increases, as a predictor always explain some portion of the variance.
- **Adjusted Rsquared**: is the R-squared corrected for multiple predictors' side-effect. It adds penalties for the number of predictors in the model. It shows a balance between the most parsimonious model, and the best fitting model.

$$adjR^2 = 1 - \frac{n-1}{n-(k+1)}.(1 - R^2)$$

where n: size of the sample and k: number of independent variables

Generally, if you have a large difference between your multiple and your adjusted R-squared that indicates you may have overfitted your model.

In model3, when age increases by one year, the lung capacity increases by 0.12, all other things (height) being equal for the persons in the population. When height increases by one unit, the lung capacity increases by 0.27, all other things (height) being equal for the persons in the population.

Once a model is built, we can use it for prediction.

You can **predict at the population level** with the estimation of the confidence interval for the **mean response**

```
new = data.frame("Age"=30, "Height"=70)
predict(model3, new, interval="confidence")
```

```
##        fit      lwr      upr
## 1 11.53421 10.99061 12.07781
```

We can be 95% confident that the average lung capacity score for all persons with age = 30 and height = 70 is between 10.99 and 12.07.

You can **predict at the individual level** using the prediction interval for a **new response**

```
predict(model3, new, interval="prediction")
```

```
##        fit      lwr      upr
## 1 11.53421 9.390385 13.67804
```

We can be 95% confident that the lung capacity score of an individual with age = 30 and height = 70 will be between 9.39 and 13.67

6.3 Logistic regression model

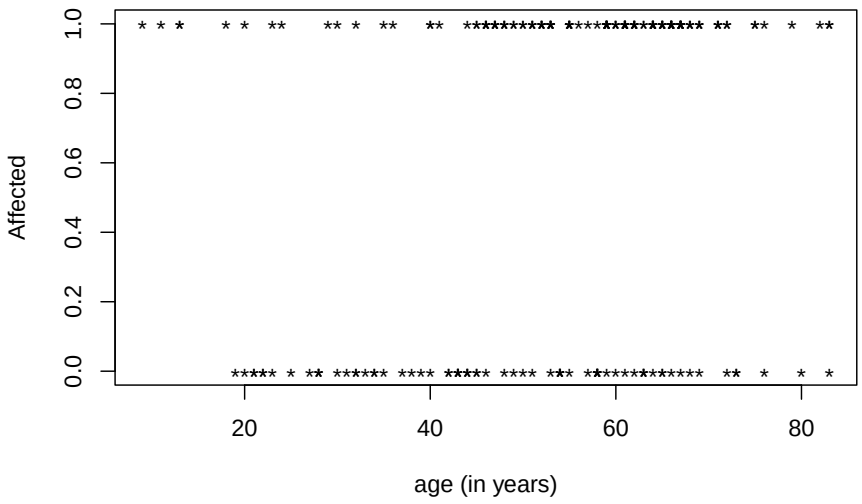
Logistic regression is a process of modeling the probability of a discrete outcome given an input variable.

The most common logistic regression is the binary logistic regression that models a binary outcome (ex: Dead/Alive, yes/no). Multinomial logistic regression can model scenarios where there are more than two possible discrete outcomes (ex: low, medium, high).

As an example, let's investigate the risk factor of chronic kidney disease (CKD).The response variable is either 1 (Affected) or 0 (Not affected) - a dichotomous response. By definition, that's a categorical response, so we can't use linear regression methods to predict it.

When coding the outcome variable 0 and 1 we can still represent what is happening in association with a covariate such as age but we cannot draw an simple line.

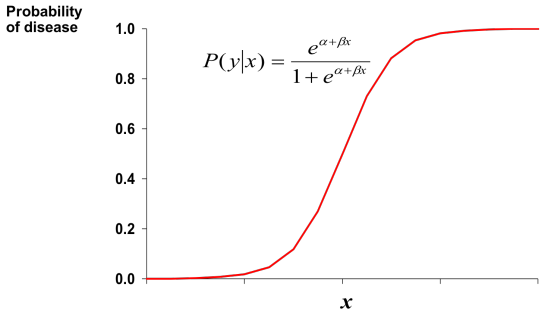
```
plot(ckd$age2, ckd$affected, ylab="Affected", xlab="age (in years)", pch="*")
```



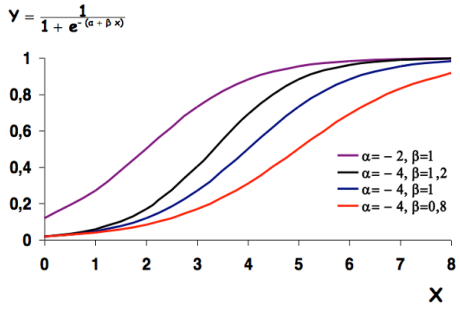
We can see that the proportions (or probabilities) of having the disease must lie between 0 and 1. It corresponds to the prevalence (%) of CKD according to age.

This relationship can be model using the Logistic function.

- $Y = \frac{1}{1+e^{-(\beta_0+\beta_1 X)}}$



- β_0 and β_1 give the shape of the curve

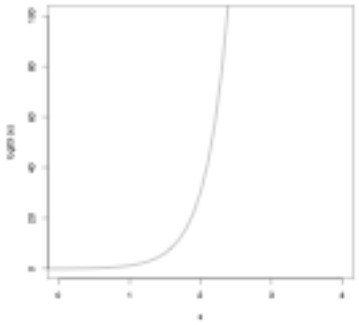


We note the the $\beta_0 + \beta_1 X$ part looks really similar a linear model. In fact, to estimate those coefficients the mathematical trick is to transform the above equation in order to get to a linear form.

By transformation:

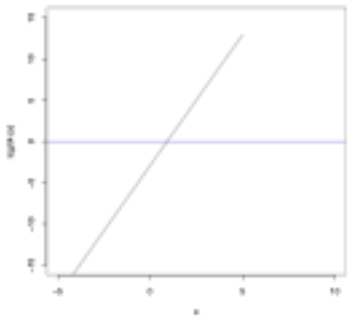
- First we compute the odds of the disease

$$Y/(1-Y) = \frac{1}{e^{-(\beta_0+\beta_1 X+\beta_2 x_2+\ldots+\beta_i x_i)}}$$



- Second we compute the Logit

$$ln(Y/1-Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ...\beta_i x_i$$



β_0 = log odds of disease in unexposed

β_i = log odds ratio associated with being exposed to i (age for instance)

The *Logit function* is defined as the natural log of the odds. A probability of 0.5 corresponds to a logit of 0, probabilities smaller than 0.5 correspond to negative logit values, and probabilities greater than 0.5 correspond to positive logit values.

After fit the model and estimating the β_0 and β_1 coefficients, we can retrieve the Odds Ratio (OR) which is the Probability of having the outcome / Probability of not having the outcome when exposed:

OR= e_i^{β} = odds ratio associated with being exposed to i

```
# Can age contribute to developpping the disease?
model_ckd <- glm(affected  ~ age2, family=binomial("logit"), data=ckd)
summary(model_ckd)
```

```
##
## Call:
## glm(formula = affected ~ age2, family = binomial("logit"), data = ckd)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8876  -1.2231   0.7832   0.9150   1.5734
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.198334   0.528806  -2.266 0.023444 *
## age2         0.033681   0.009818   3.431 0.000602 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

```
##     Null deviance: 259.57  on 197  degrees of freedom
## Residual deviance: 247.03  on 196  degrees of freedom
## AIC: 251.03
##
## Number of Fisher Scoring iterations: 4
```

The coefficient for age is positively link to the disease. When you increase in age you increase your risk of having disease. To compute the OR, we can use the $exp(0.039)$ but for a nicest output with confidence interval we can use the *epiDisplay* library and its *logistic.display()* function.

```
# Loading the epiDisplay library in the R environment
library(epiDisplay)
logistic.display(model_ckd)
```

```
##
## Logistic regression predicting affected
##
##                    OR(95%CI)         P(Wald's test) P(LR-test)
## age2 (cont. var.) 1.03 (1.01,1.05)  < 0.001        < 0.001
##
## Log-likelihood = -123.5132
## No. of observations = 198
## AIC value = 251.0263
```

The odd (risk) of having the disease is 1.04 times higher when people get older by 1 year (95%CI[1.02,1.06]). The P(Wald's test) tells us that the estimated coefficient is different from 0 (or the OR is different from 1). The P(LR-test) log-likelihood ratio test tells us that the model is better than the NULL model (without any covariate).

We can add other covariates and then use a stepwise approach to get the most parsimonious model.

```
model_ckd2 <- glm(affected  ~ age2 + bp.limit, family=binomial("logit"), data=ckd)
summary(model_ckd2)
```

```
##
## Call:
## glm(formula = affected ~ age2 + bp.limit, family = binomial("logit"),
##     data = ckd)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9805  -1.1433   0.5753   0.9494   1.8686
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.87178    0.59109  -3.167  0.00154 **
```

```
## age2         0.03529    0.01034   3.414  0.00064 ***
## bp.limit     0.85816    0.22052   3.891 9.96e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 259.57  on 197  degrees of freedom
## Residual deviance: 229.70  on 195  degrees of freedom
## AIC: 235.7
##
## Number of Fisher Scoring iterations: 4
```

```
model_ckd2<- step(model_ckd2)
```

```
## Start:  AIC=235.7
## affected ~ age2 + bp.limit
##
##            Df Deviance    AIC
## <none>          229.71 235.71
## - age2      1   242.42 246.42
## - bp.limit  1   247.03 251.03
```

```
# The best model is with the smallest AIC
# It is the model without removing any variable (<none>)
logistic.display(model_ckd)
```

```
##
## Logistic regression predicting affected
##
##                    OR(95%CI)         P(Wald's test) P(LR-test)
## age2 (cont. var.) 1.03 (1.01,1.05)  < 0.001        < 0.001
##
## Log-likelihood = -123.5132
## No. of observations = 198
## AIC value = 251.0263
```

6.4 Collinearity

Collinearity is the correlation between 2 independent variables (predictors).

Multi-collinearity is cross-correlation between multiple independent variables (predictors).

If the independent variables are perfectly correlated, you will face these problems:

- Inflated coefficients

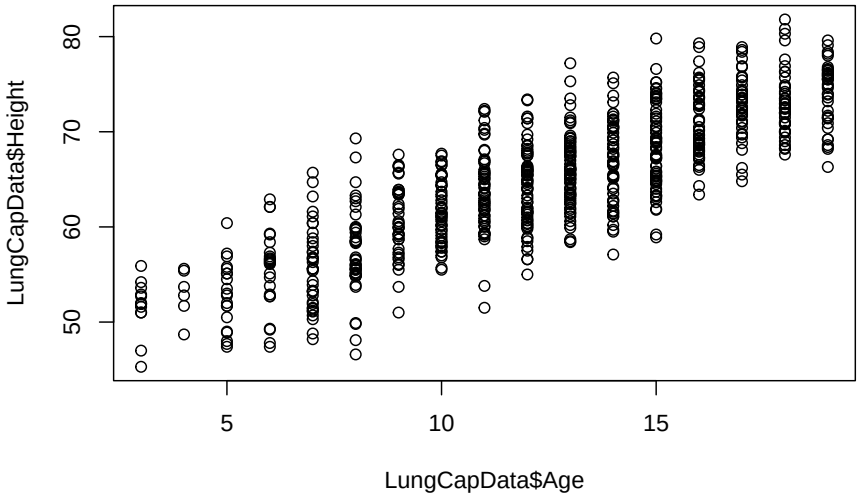
- Values and signs of the coefficients are incoherent with common knowledge
- Underestimated Student T test (non-significant p-values)
- Unsteady results, adding or deleting observation strongly modify the values and signs of the coefficients

6.5 Detecting (multi-)collinearity

6.5.1 Coefficient of correlation and visual assessment

Is there a problem in our example? Is Age too highly correlated with height?

```
plot(LungCapData$Age, LungCapData$Height)
```



A test over the coefficient of correlation can be performed where:

- H0: the coefficient of correlation r = 0
- H1: the coefficient of correlation r != 0

with $\alpha < 0.05$

```
cor.test(LungCapData$Age, LungCapData$Height)
```

```
##
##  Pearson's product-moment correlation
##
## data:  LungCapData$Age and LungCapData$Height
## t = 40.923, df = 723, p-value < 2.2e-16
```

```
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.8123578 0.8564338
## sample estimates:
##       cor
## 0.8357368
```

Note that even a small correlation between 2 variables and thus the test over the coefficient of correlation statistically significant.

6.5.2 Variance Inflation Factor

Therefore other metrics can be computed. The most commun one is the Variance Inflation Factor or **VIF** and its inverse the Tolerance indicator or **TOL**

Variance Inflation Factor (**VIF**): how much of the inflation of the standard error could be caused by collinearity

Tolerance indicator (**TOL**) = $\frac{1}{VIF}$

If all of the variables are orthogonal to each other (uncorrelated with each other) both indicators (TOL and VIF) are 1.

If a variable is very closely correlated to another variable, the TOL goes to 0, and the VIF gets very large. An independent variable with a VIF > 10 (empirical threshold) should be look at and some remedial measures might have to be taken.

Using the R *mctest* library and the *imcdiag()* function on your model you will get the VIF and TOL metrics and others. See help for interpretation of the others measures.

```
library(mctest)
imcdiag(model3)
```

```
##
## Call:
## imcdiag(mod = model3)
##
##
## All Individual Multicollinearity Diagnostics Result
##
##           VIF    TOL      Wi  Fi Leamer    CVIF Klein  IND1 IND2
## Age    3.3163 0.3015 1674.66 Inf 0.5491 -1.0332     0 4e-04    1
## Height 3.3163 0.3015 1674.66 Inf 0.5491 -1.0332     0 4e-04    1
##
## 1 --> COLLINEARITY is detected by the test
## 0 --> COLLINEARITY is not detected by the test
```

```
##
## * all coefficients have significant t-ratios
##
## R-square of y on all x: 0.843
##
## * use method argument to check which regressors may be the reason of collinearity
## ==================================
```

6.5.3 Remedial measures

The remedial measures to manage collinearity could be :

- Manual exclusion: VIF values above 10 and as well as the concerned variables logically seem to be redundant. You remove one of them.
- Aggregation: The two redundant variables could be summarized into an third variable using statistical methods such as Principal Component Analysis (PCA). That third variable will then be used in the model in place of the first two.
- Automatic variable selection: you let the modeling algorithm able the issue using stepwise regression which might remove variables that generate unstable modeling results

6.6 Explanatory variable selection

When defining a model you would like to either obtain the best explanatory model (goodness of fit) or the best predictive model (parsimonious). To this aim you need to select the most relevant variables. If you add all the variables you have in hand it might be one too many. The number of degrees of freedom will decrease and as the consequence the residual variance could increase.

How to select the contributing variables?

- Some variables are correlated with redundancy of information: the variables with the effect the easiest to explain/express should to be kept.
- Univariate tests (e.g., outcome vs exposition) are highly recommended prior building the model. Results with a p-value > 0.25 or 0.20 (empirical cut-off) is less likely to contribute to explain the outcome in a larger model. You might not want to include it in the full model. However, note that even so a variable shows a p-value > 0.25 in an univariate test you might still want to force it in the model because of inconsistency with your prior knowledge (literature, own studies).
- Automatic and iterative procedures for explanatory variable selection based on goodness of fit criterion (e.g., on Akaike's information criterion - AIC)

6.6.1 Iterative procedures for explanatory variable selection

These approaches introduced or suppressed variables based on results of nested models tests using goodness of fit criterion. The 3 variants are:

1. Backward procedure

- First, all of the predictors under consideration are in the model
- One at a time is deleted the less significant (the "worst")
- All the remaining variables make a significant contribution

2. Forward procedure

- Begins with none of the potential explanatory variables
- Adds one variable at a time
- Stops when no remaining variable makes a significant contribution

3. Stepwise procedure

- Modification of the forward procedure
- Drops variables if they lose their significance as other are added

The goodness of fit criteria is often the Akaike's information criterion or AIC. **The smaller the AIC, the better the fit**.

The R command for the stepwise approach is:

```
# Full model with all the variables you have in hand (or selected after univariate tes
modelFull <-  lm(LungCap ~ ., data= LungCapData)
# The argument direction="both" is for the procedure type. See the help page for more
modelStep <- step(modelFull, direction = "both")
```

```
## Start:  AIC=34.43
## LungCap ~ Age + Height + Smoke + Gender + Caesarean
##
##             Df Sum of Sq     RSS    AIC
## <none>                    747.78  34.43
## - Caesarean  1      5.80  753.57  38.03
## - Smoke      1     24.35  772.13  55.66
## - Gender     1     24.55  772.33  55.85
## - Age        1     82.65  830.43 108.44
## - Height     1    716.54 1464.32 519.66
```

Here at each iteration one variable is removed and compare to the full model. For example, the line -*Caesarean* indicates that the variable *Caesarean* is removed. The AIC of 38.03 is compared to the AIC 34.43 of the full model ($< none >$ removed). The full model is more informative than the model without the variable *Caesarean*. The interpretation is the same for the rest of the variables.

```
summary(modelStep)
```

```
##
## Call:
## lm(formula = LungCap ~ Age + Height + Smoke + Gender + Caesarean,
##     data = LungCapData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3388 -0.7200  0.0444  0.7093  3.0172
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -11.32249    0.47097 -24.041  < 2e-16 ***
## Age            0.16053    0.01801   8.915  < 2e-16 ***
## Height         0.26411    0.01006  26.248  < 2e-16 ***
## Smokeyes      -0.60956    0.12598  -4.839 1.60e-06 ***
## Gendermale     0.38701    0.07966   4.858 1.45e-06 ***
## Caesareanyes  -0.21422    0.09074  -2.361   0.0185 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.02 on 719 degrees of freedom
## Multiple R-squared:  0.8542, Adjusted R-squared:  0.8532
## F-statistic: 842.8 on 5 and 719 DF,  p-value: < 2.2e-16
```

6.6.2 Goodness of fit analysis

You can perform a goodness of fit analysis step by step using the *anova()* function call between 2 nested models [NOTE: do not mix-up anova() with aov() for the comparison of more than two means].

```
## Model without the Age variable
model4var <-  lm(LungCap ~ Height + Smoke + Gender + Caesarean, data= LungCapData)
summary(model4var)
```

```
##
## Call:
## lm(formula = LungCap ~ Height + Smoke + Gender + Caesarean, data = LungCapData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3631 -0.7360  0.0290  0.7529  3.0710
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)  -14.162235   0.365323 -38.766  < 2e-16 ***
## Height         0.339699   0.005706  59.536  < 2e-16 ***
## Smokeyes      -0.497653   0.132004  -3.770 0.000177 ***
## Gendermale     0.197766   0.080852   2.446 0.014683 *
## Caesareanyes  -0.206444   0.095549  -2.161 0.031055 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.074 on 720 degrees of freedom
## Multiple R-squared:  0.8381, Adjusted R-squared:  0.8372
## F-statistic: 932.1 on 4 and 720 DF,  p-value: < 2.2e-16
```

All the variables in the model4var are significant. The *anova()* function compare the full model to the model4var.

```
anova(modelFull, model4var)
```

```
## Analysis of Variance Table
##
## Model 1: LungCap ~ Age + Height + Smoke + Gender + Caesarean
## Model 2: LungCap ~ Height + Smoke + Gender + Caesarean
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    719 747.78
## 2    720 830.43 -1   -82.653 79.472 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The 2 models differ (p-value < 0.05). The AIC index highlight the model with the best fit (smallest AIC); here it is the full model.

```
AIC(modelFull, model4var)
```

```
##           df      AIC
## modelFull  7 2093.888
## model4var  6 2167.896
```

Chapter 7

Glossary

Bibliography

Leif Edvard Aarø, Bente Wold, Lasse Kannas, and Matti Rimpelä. Health behaviour in schoolchildren a who cross-national survey: A presentation of philosophy, methods and selected results of the first survey. *Health Promotion International*, 1(1):17–33, 1986.

Thierry Ancelle. *Statistique, Epidemiologie / Thierry Ancelle*. Sciences fondamentales. Editions Maloine, Paris, 4e edition edition, 2017. ISBN 978-2-2240-3522-8.

D.M. Diez, C.D. Barr, and M. Çetinkaya-Rundel. *OpenIntro Statistics*. OpenIntro, Incorporated, 2019. ISBN 9781943450039. URL https://leanpub.com/openintro-statistics.

M Marin and L Hamadani. Marinstatslectures online, 2021. URL https://www.statslectures.com/.

Department of Statistics PennState Eberly College of Science. Regression methods, 2021. URL https://online.stat.psu.edu/stat501/.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020. URL https://www.R-project.org/.

Norean Radke Sharpe, Richard D De Veaux, and Paul F Velleman. *Business statistics*. Pearson Education., Boston, 2012. ISBN 978-0-321-71609-5. OCLC: 960373192.

T Vigen. Spurious correlations, 2021. URL https://www.tylervigen.com/spurious-correlations.

Yihui Xie. *bookdown: Authoring Books and Technical Documents with R Markdown*. Chapman and Hall/CRC, Boca Raton, Florida, 2016. URL https://bookdown.org/yihui/bookdown. ISBN 978-1138700109.