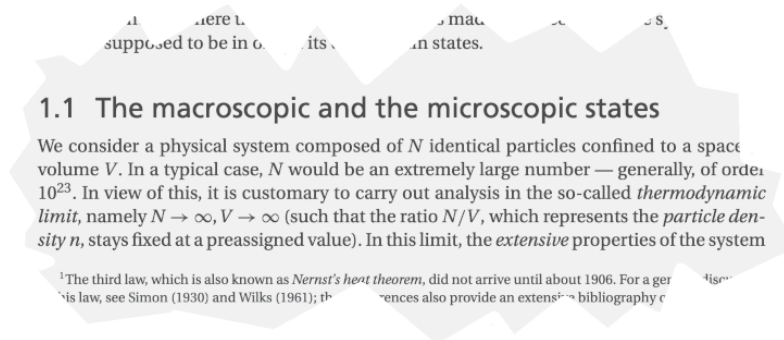


A copy of these lecture note can be downloaded from <https://www.researchgate.net/profile/Massimo-Borelli>, or from the course public website <https://ictpmmp.weebly.com/>. All the dataset are publicly available at <https://github.com/MassimoBorelli/Miramare>, both in .csv and in .ods format.

### 1.1.1 Shifting Statistics from Physics to Medicine



Many of us have a more or less solid background in calculus and in probability; terms like *average*, *moment*, *expected value*, *finite integral* are familiar. And if we open whatever book of, for instance, *Statistical Mechanics* (here above the first page of Pathria and Beale [40]), we find no difficulty in dreaming (*'it is customary'*) about those  $N$  particles growing and growing, until reaching a so huge number called  $\infty$ . On the contrary, in medical statistics  $N$  can be ridiculously small – this is the reason why in medical statistics hardly we can perform 'experiments' to grab an unknown Nature's variability; hardly we can distinguish a 'confounding factor' from the 'reality'; hardly we can 'improve the sample size'. Another important topic (the most important one would say) is that when we talk about  $j \in \{1, \dots, N\}$ , that  $j$  could our mother, our son, ourselves: medical statistics therefore involves important ethical questions and requires important privacy laws respect.

### 1.2 Which is the 'best' software for medical statistics?

Well, it depends. If you are required to perform 'heavy' computation, a programming language like R or Python will be needed. Otherwise, a simpler 'statistical suite' like R Commander, Jamovi or JASP could be preferred. Let us spend a few word to introduce them.

#### 1.2.1 The R language

R is an open source software environment for statistical computing and graphics, which can be freely downloaded from the so-called CRAN (the Comprehensive R Archive Network) world-wide mirrors: <https://cran.r-project.org/mirrors.html>. R runs on UNIX/Linux, Windows and MacOS platforms. You can also exploit the cloud computing facilities, and compile online your script into <https://rdr.io/snippets/>.



If you are interested in some historical details, Nick Thieme has published an article[48] which recalls the astonishing success of R, born more or less twenty five years ago in Auckland University by the ideas of two statistics professors: Ross Ihaka and Robert Gentleman. Other details are provided by Carlos Alberto Gómez Grajales in his *Created by statisticians for statisticians: How R took the world of statistics by storm* appeared on <http://www.statisticsviews.com/view/index.html>.

Of course, R is very well documented; for instance, you can find free on line introductory books, as the Hadley Wickham and Garrett Golemund textbook [53] *R for data science*,

available at <https://r4ds.had.co.nz/>, or as the Kim Seefeld and Ernst Linder textbook *Statistics Using R with Biological Examples*, available at [https://cran.r-project.org/doc/contrib/Seefeld\\_StatsRBio.pdf](https://cran.r-project.org/doc/contrib/Seefeld_StatsRBio.pdf). There are also lots of webpages, blogs and Moocs concerning R; for instance:

- [http://ncss-tech.github.io/stats\\_for\\_soil\\_survey/chapters/](http://ncss-tech.github.io/stats_for_soil_survey/chapters/)
- <http://www.sthda.com/english/wiki/r-software>
- Quick-R, <https://www.statmethods.net/>

Many video tutorials are also available on YouTube, following the query [https://www.youtube.com/results?search\\_query=R+tutorial](https://www.youtube.com/results?search_query=R+tutorial).

Instead of working directly on the R Console, many scientists prefer to use R Studio <https://www.rstudio.com/> Integrated Development Environment (IDE).

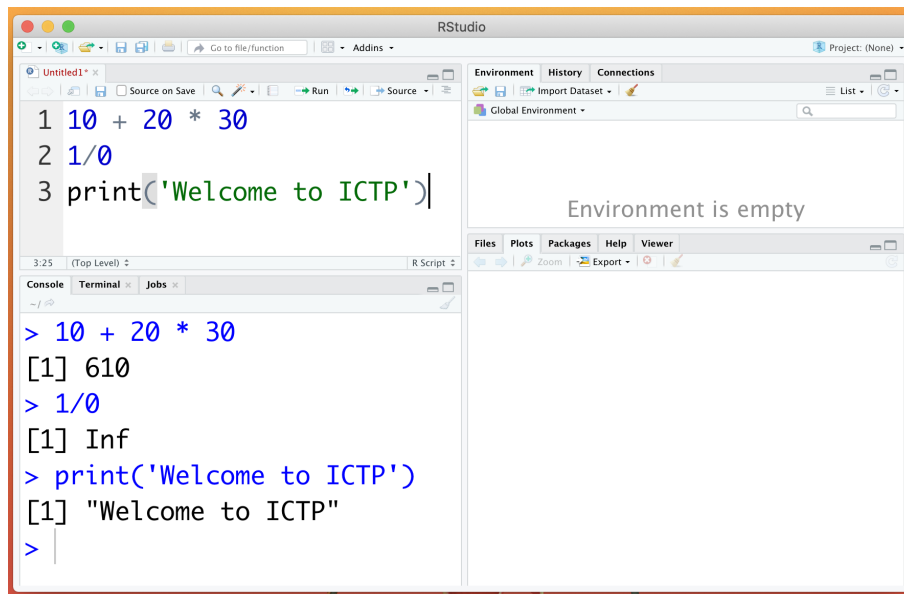


Figure 1.1: R Studio is preferred by many researchers and data analysts, ensuring a stable and well integrated programming and graphing environment. A fatal drawback is its 'steep learning curve': the newcomer has to practice quite a lot of time in managing syntaxes and commands – besides the effort in learning Statistics.

Being R a programming language, of course, you can start copying and pasting code chunks from all around the web, just 'googling' what you need. But in order to master the language you have to spend a lot of time to practice: newcomers find frustrating to search for the  $\sim$  symbol on the keyboard, or feel stuck when they copy some code from a pdf, in which it is written  $x - y$  (with the four point 'en dash') but the software needs to read  $x - y$  (with the minus, i.e. the three point 'hyphen'). These are just two of the main reasons why in our short course, alas!, we skip the effort to learn it. But one possible recovery plan it exists: to adopt a G.U.I., a graphical user interface – let us see in the next page.

## 1.2.2 The R language user graphical interfaces

Beginners often find sufficient to access to a selection of commonly-used R commands using 'familiar' graphical user interfaces, as R Commander, <https://www.rcommander.com/>, or as Jamovi, <https://www.jamovi.org/>.

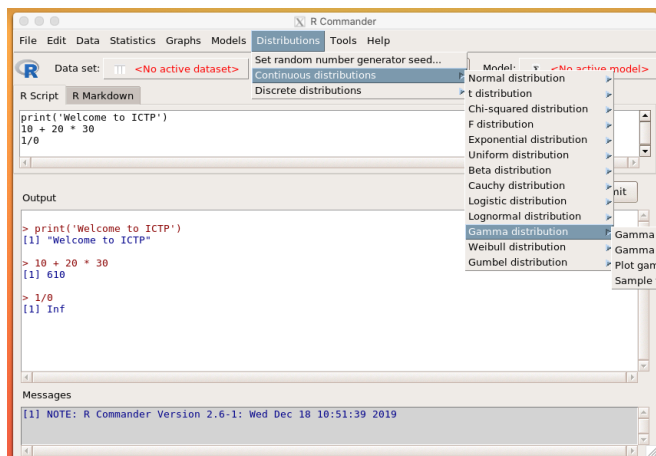


Figure 1.2: The R Commander appearance: you see a menu environment with an input section, named R Script, which has been created by the File | New Script procedure; and an Output section which lists the input commands and produces the outputs. Below, the gray backgrounded section provides Messages alerting for possible mismatches.

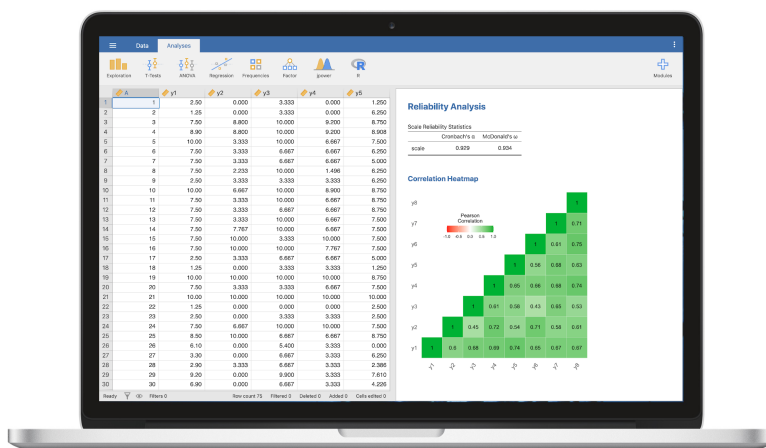


Figure 1.3: The JAMovi environment, which integrates the spreadsheet capabilities in managing raw data and a menu of typical R analysis commands. The user manual <https://www.jamovi.org/user-manual.html> helps the beginner to learn the basic procedures.

## 1.2.3 JASP

About ten years ago, a group of people belonging to social research areas (mainly from Amsterdam University, <https://cordis.europa.eu/project/id/283876>) started to work on a sort of 'free and open SPSS', which has been a sort of *lingua franca* spoken by psychometrists. The idea was brilliant: to use R as an hidden engine (in particular, to exploit the package **BayesFactor**) and to pack it with a 'drag and drop' interface: their result was the creation of JASP, which

can be freely downloaded from: <https://jasp-stats.org/team/>. Their original goal was to promote the Bayesian hypothesis testing approach in social sciences, recognising that major advances in computational statistics should have had a positive impact over the old-fashioned (or, as they said, even inappropriate) psychometric methodologies. JASP is also very well documented, and newcomers can start reading <https://jasp-stats.org/getting-started/>, or <https://jasp-stats.org/how-to-use-jasp/>; very valuable are also the free manuals, <https://jasp-stats.org/jasp-materials/>. We will discuss better the details along these lecture notes.



Figure 1.4: The official JASP web page, <https://jasp-stats.org/>

### 1.3 Exercises.

■ **Activity 1.1 — protecting privacy in a spreadsheet.** In hospitals, to use the spreadsheet (Microsoft Excel, Libre Office Calc, Google Sheets, iOS Numbers, ...) in order to collect data is routinary. Remembering what announced in section 1.1.1, the privacy is an important issue – but very often biostatistician are required to analyse data not properly masked, in which private information (e.g. name, surname, date of birth, ...) are disclosed. As an exercise, download on your computer the privacy dataset (at <https://github.com/MassimoBorelli/Miramare>), explore it with your favourite spreadsheet and create a new column of data by means of a text function (or joining together the outputs of different text functions) in order to provide a unique identifier for each row ('record') of the dataset. ■

Timestamp	Name	Surname	Daybirth	Monthbirth	Yearbirth	Id
28/11/2021 10.55.31	James	Wang	29	12	1966	
28/11/2021 10.56.53	Mary	Chen	19	10	1978	
28/11/2021 10.56.59	Robert	Singh	9	7	1957	
28/11/2021 10.58.00	Patricia	Kumar	12	8	1980	
28/11/2021 11.01.35	John	Ali	11	11	1976	
28/11/2021 11.03.07	Jennifer	Nguyen	7	12	1968	
28/11/2021 11.04.33	Michael	Khan	11	9	1977	
28/11/2021 11.05.04	Linda	Ahmed	26	1	1982	
28/11/2021 11.05.55	William	Khatun	22	1	1960	
28/11/2021 11.06.14	Elizabeth	Silva	18	3	1980	
28/11/2021 11.07.27	David	Tang	13	9	1983	
28/11/2021 11.07.47	Barbara	Mohamed	2	5	1962	
28/11/2021 11.07.47	Richard	Xie	23	8	1966	
28/11/2021 11.08.19	Susan	Han	20	4	1972	
28/11/2021 11.11.10	Isaiah	Garcia	22	10	1970	









## 2. Data Presentation

### 2.1 Background

 Elise Whitley, Jonathan Ball. Statistics review 1: Presenting and summarising data  
<https://ccforum.biomedcentral.com/articles/10.1186/cc1455>

The first goal to achieve in any data analysis is to 'understand' them, to describe and to summarize them in a proper way (being not too much verbose; or not too much cryptic). Such analysis may enlight 'strange' values (outliers), which very high or very low with respect to the rest of the data. Tables and graphs are the usual way to summarize large amounts of information and the above review recalls the basics, providing examples of qualitative data (unordered and ordered) and quantitative data (discrete and continuous). In their review, Elise Whitley and Jonathan Ball recalls in which way the previous types of data can be depicted, enhancing the two important features of a quantitative dataset: the **location** of the data and their **variability**. Common measures of location (mean, median and mode) and of variability (range, interquartile range, standard deviation and variance) are revised.

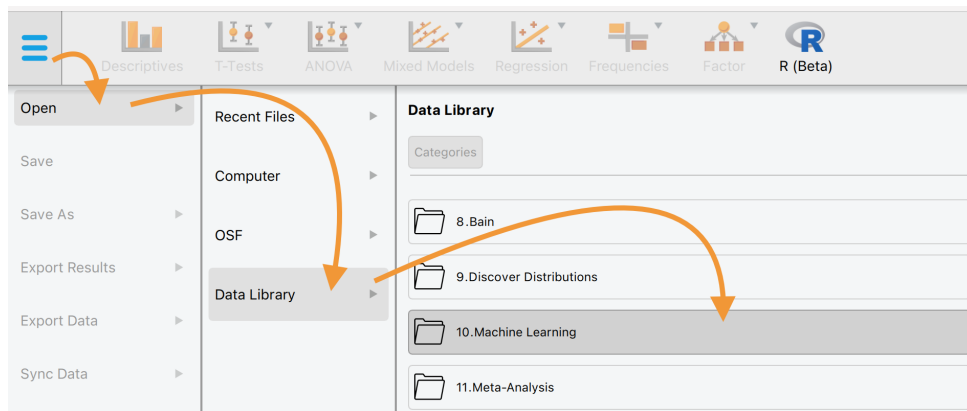
 Alla Katsnelson. Colour me better: fixing figures for colour blindness  
<https://www.nature.com/articles/d41586-021-02696-z>

We do not forget that, all around the world, the color vision deficiency in male is estimated to be around the 5 – 10 percent of the population: this is an invitation to prefer, in any possible occasion, to adopt the so called *viridis* color palette in your graphs, and to enhance different informations also by means of different graphical coding (solid, dashed, dotted, ...)

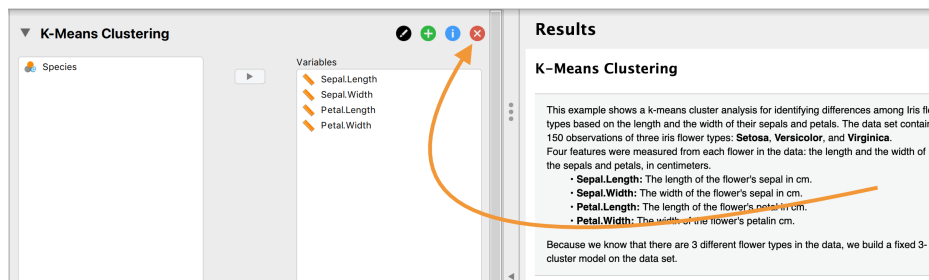
### 2.2 Descriptive Statistics in JASP

Let us start exploring JASP capabilities in summarizing and presenting data. For simplicity we refer to a very famous example, the *iris* dataset by Ronald Fisher [19] and Edgar Anderson [4]. Nowadays the *iris* dataset is commonly used by computer scientists when they want to test their

softwares' performances in supervised learning, and this is probably the reason why JASP stores it into the 'Machine Learning' folder:



We are not interested now in discussing what is K-Means Clustering; but looking to the **Results** section on the right, we can read a description of the dataset, composed by 150 rows and 5 columns, named respectively `Sepal.Length`, `Sepal.Width`, `Petal.Length`, `Petal.Width` and `Species`. The first four columns provide numerical data, while the last column provide qualitative information about the three different species (*Setosa*, *Versicolor* and *Virginica*) of flowers considered. Scrolling down the Results section we can immediately see a set of nice coloured graphs, depicting certain function densities and a scatterplot with three coloured point clouds.



Acting on the 'Remove this analysis' red button, we can start our first exploration. We recognize the dataset, we observe that `Sepal.Length`, `Sepal.Width`, `Petal.Length`, `Petal.Width` are signed with an orange **Scale** ruler, while `Species` has three Venn diagrams, identifying the **Nominal** variables. The software suggest this classification as a default, but we can modify it simply clicking on the icons. Being satisfied of the situation, we can start the analysis clicking the **Descriptives** menu:

	width	Petal.Length	Petal.Width	Species	
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa

### 2.2.1 Numerical summaries

Now let us practice with JASP Descriptives menu to provide answers to the following requests:

**Exercise 2.1 — location measures.** Exploring the iris dataset, say:

- how much is the mean of `Sepal.Length`?
- how much are the medians of `Sepal.Width`, distinguishing between the three `Species`?

Previous exercise allows us to discover how to 'split' by means of a nominal variable a numeric variable, and to verify that the tables produced by JASP are ready to be copy and pasted both in 'Word' and  $\LaTeX$  format. We want to recall that the Scale / Ordinal / Nominal variable taxonomy is not universally accepted. The R language calls numeric what Martin Bland [7] defines to be a **quantitative** variable. On the contrary, an R factor (i.e. a **qualitative** or a **nominal** variable, according to Martin Bland), is a list of different 'groups' which are called the `levels` (ordered or unordered) inside the factor.

The screenshot shows the 'Statistics' menu in JASP. It is divided into several sections:

- Central Tendency:** Mode, Median, Mean.
- Dispersion:** Std.deviation, Coefficient of Variation, MAD Robust, Variance, Minimum, Std.deviation, MAD, IQR, Range, Maximum.
- Quantiles:** Quantiles, Cut points for: 4 equal groups, Percentiles.

Most options are currently unchecked.

**Discussion 2.2.1 — other position and dispersion measures.** Look at the picture above. Are you able to define all the **measures of central tendency** (or **measures of location**)? And can you define all the (not blurred) **measures of shapes**, or **measures of dispersions** calculated by JASP? We will discuss better the concepts of quantiles but, if you need a refresh, a recommended book may be that by professor Joe Blitzstein (Harvard University) and Jessica Hwang (Stanford University), entitled *Introduction to Probability* [8]. Professor Blitzstein also offers a free edX course and a free copy of his must-read book:



Jonathan Blitzstein, Jessica Hwang. Introduction to Probability.  
<https://projects.iq.harvard.edu/stat110/home>

**Exercise 2.2 — frequencies.** Create the following frequencies table:

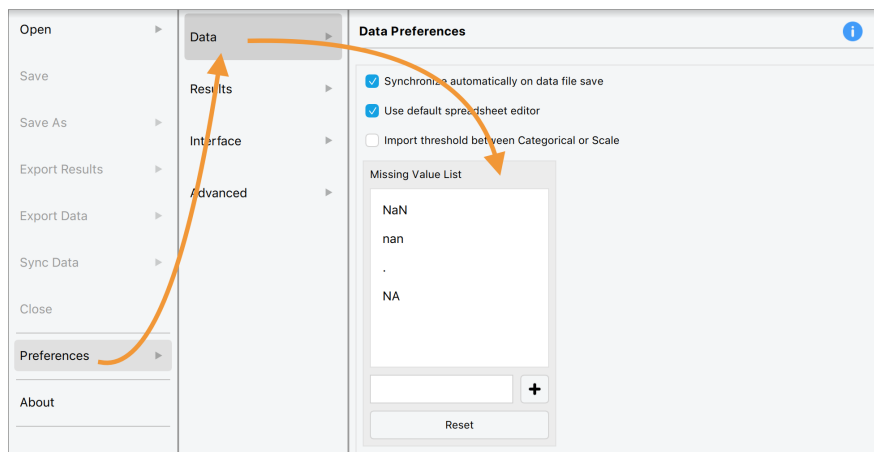
Species	Frequency	Percent	Valid Percent	Cumulative Percent
setosa	50	33.333	33.333	33.333
versicolor	50	33.333	33.333	66.667
virginica	50	33.333	33.333	100.000
Missing	0	0.000		
Total	150	100.000		

Looking to the frequencies, we can not see the typical central tendency measure of nominal data: the **mode** of the distribution (i.e. the group having the highest frequency). Nevertheless, it is correct to exploit the mode also with ordinal and scale variables: as an example, in the sequel we will discuss of the famous bimodal female vs. male height distribution.

**Vocabulary 2.1 — balanced dataset.** The iris dataset is said to be **balanced** as we observe data with the same absolute frequencies in each group considered. In our example, fifty flowers belonging to each of the (levels of the) Species setosa, versicolor and virginica have been measured.

**Vocabulary 2.2 — complete dataset.** A dataset is said to be **complete** when we do not observe any **missing data**, or **missing values**, usually represented with the symbol NA.

It may happen that different systems or different researchers adopt various way to code the missing information. While NA is the preferred one, the symbol NaN (= 'not a number', e.g. 0/0). JASP allows to manage this modifying the Preferences. A caveat: always avoid to use 'blank cells' when having missing information.



Not so often, other descriptive measures implemented in JASP are evaluated:

- the **coefficients of variation**, which is the ratio between the mean and the standard deviation. *'Coefficients of variation are particularly useful when observations with different dimensions are being compared, such as UK sterling and US Dollars. A dimensionless measure of dispersion is then very convenient.'* (R. Mould, 2.5 [37])
- the **median absolute deviation**, which is – as a word pun – the median of the absolute deviation from the median, [https://en.wikipedia.org/wiki/Median\\_absolute\\_deviation](https://en.wikipedia.org/wiki/Median_absolute_deviation)

## 2.2.2 A picture is worth a thousand words



Yan Holtz. The R Graph Gallery.  
<https://www.r-graph-gallery.com/>

Thanks to the the powerful graphical capabilities of R, JASP allows to easily depict data distributions and summaries. Let us see them in a brief review, having in mind that different types of variables (nominal, ordinal, scale) requires different graphics.

### Pie charts

**Exercise 2.3** Depict the frequencies of Species by a pie chart, choosing the viridis palette. ■

### Dot plots

**Discussion 2.2.2 — what is an 'informative' picture?.** We are not able to provide a mathematical definition of what is an 'informative' drawing; anyway, when we try to depict the dotplots of the Sepal.Length splitted over the three Species we can not 'easily grab' what is happening. Do you agree?

### Distribution plots

Let us spend a couple of minutes to clarify the difference between the **barplot** and the **histogram**: both of them fall inside the 'Distribution plots' denomination adopted by JASP. But the former is properly named when we are drawing (nominal or) ordinal data, while the latter requires data collected along a continuous scale of measure. In fact, talking about histogram, Richard Mould [37] writes in his 1.4 paragraph:

In a histogram, the height of each vertical block does not always represent the value of the variable of interest (unless the width of the block is unity), as is the case of a bar in a bar chart. Also, in a histogram, the horizontal scale is continuous and not, like the bar charts, discrete. Also, unlike a bar chart width, a histogram block width *does have a meaning*.

Therefore let us explain in a precise way [28] the idea of relative frequency histogram, which is a central concept naturally linked to 'probability density function' concept. Let  $x = (x_1, x_2, \dots, x_n)$  be the  $n$  numeric data considered and let  $c_1 < c_2 < c_3 < \dots < c_r$ ,  $2 \leq r < n$ , a class partition with **cut-off**  $c_j$ 's, such that  $c_1 = \min(x)$  and  $c_r = \max(x)$ . We obtain  $r - 1$  limited disjoint **classes** (or **bins**):

$$C_1 = [c_1, c_2], C_2 = (c_2, c_3], C_3 = (c_3, c_4], \dots, C_{r-1} = (c_{r-1}, c_r]$$

Denote with  $n_j$  the absolute frequencies of the  $x$  data falling into each class  $C_j$ , and let  $f_j = n_j/n$  the relative frequencies ( $1 \leq j \leq r - 1$ ). With these choices, the **relative frequency histogram** is made by  $r - 1$  rectangles of bases  $C_j$  and heights:

$$h_j = \frac{n_j/n}{c_{j+1} - c_j}$$

**Discussion 2.2.3** Draw the distribution plot of `Petal.Length`. Is it a barplot? Is it a relative frequency histogram? Tick the box `Display density`. Is now the picture a relative frequency histogram?

### the Boxplot and the Quartiles

**Exercise 2.4** Draw the boxplot of `Petal.Length`, in grey color. Then draw the boxplots of `Petal.Length` splitted by `Species` according to `ggplot2` palette. ■

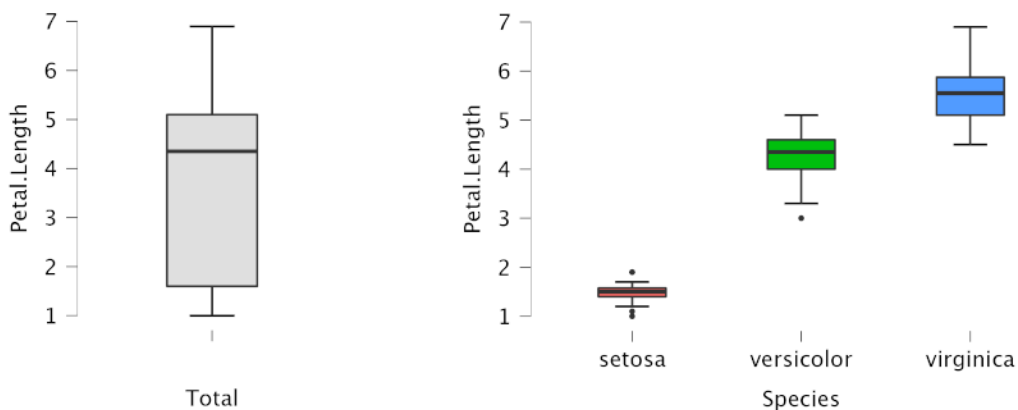


Figure 2.1: The boxplot.

The legendary chemist, mathematician and statistician John W. Tukey ([https://en.wikipedia.org/wiki/John\\_Tukey](https://en.wikipedia.org/wiki/John_Tukey)) introduced this type of data visualization, providing the so called **five point summary**. In fact, when we deal with ordered (or scale) data, as in `Petal.Length` variable, we can suppose without loss of generality that the sample  $x = (x_1, x_2, \dots, x_n)$  is already ordered,  $x_1 \leq x_2 \leq \dots \leq x_n$ . Obviously,  $x_1$  is the **minimum** and  $x_n$  is the **maximum**. Now we can consider the index  $n/2$ , which is integer in  $n$  is even (but if  $n$  is odd we can arrange the situation a little, chopping or rounding away the decimal, eventually averaging the  $x$ 's): we are now in presence of the **median**,  $x_{n/2}$ .

**Vocabulary 2.3 — Quartiles.** Let us denote with  $L$  the median of  $x$ :  $L$  divide the sample  $x$  into two subsets, the first half and the second half. If we compute the medians of those two halves we obtain respectively the **first quartile**  $Q1$  and the **third quartile**  $Q3$  (being the median  $L$  the second quartile,  $\min(x)$  the zeroth quartile and  $\max(x)$  the fourth quartile). If we split  $x$  in ten sections instead of two, one can define the first, second, ... **deciles**. And again, splitting  $x$  in one hundred sections, we compute the **percentiles**. Quartiles, deciles and percentiles are examples of **quantiles**.

The spacings between the different parts of the coloured box (which, of course, encompasses the 50 per cent of the data) indicate the 'degree' of dispersion (spread) and the 'skewness' in the data. The two whiskers describe the **tails** of the distribution, 'short' or 'long'.

As we can see, in the red `setosa` and in the green `versicolor` boxplots some isolated points appear. They are the so-called **outliers**, as defined by Tukey himself: consider the **interquartile range**,  $IRQ = Q3 - Q1$ , 'amplify' it by a 50%,  $1.5 \cdot IRQ$ , and search if there are points  $x_j \in x$  such that  $x_j < Q1 - 1.5 \cdot IRQ$  or  $x_j > Q3 + 1.5 \cdot IRQ$ . It can be shown (e.g. [27, page 29]) that outliers are not so rare in experimental measures: asymptotically, 0.7% of data.



## Scatter Diagrams

**Exercise 2.5** Draw the scatter plot of Petal.Length versus Petal.Width. ■

JASP offers two possibilities to draw a cartesian x-y **scatter plot**: the 'basic' one (called the Correlation plot) and the 'customizable' variant. We will discuss the details in the next chapters.



There exists – although not so frequently used (unfortunately, I say) – the Rousseeuw & Ruts & Tukey bidimensional version of the boxplot, which is called the **bagplot**, <https://en.wikipedia.org/wiki/Bagplot>. In JASP it is not currently implemented, but in R you have it, available in the 'Another PLOt PACKAge' aplpack [44].

## 2.3 Which are 'the best' Descriptives?

Once upon a time, the **skewness** (<https://en.wikipedia.org/wiki/Skewness>) measure of asymmetry and the **kurtosis** (<https://en.wikipedia.org/wiki/Kurtosis>) measure of 'fat tails' were commonly calculated and used in literature to describe data distribution. Nowadays these concepts seems to be buried in dust, even if JASP allows you to calculate them. Nevertheless, skewness plays an important role in data description – and a boxplot reveals it immediately. In fact, when our mind try to perceive the data distribution only knowing some numerical descriptive statistics, some pitfalls can occur. To be more clear, let us make some examples caught from literature.

Consider for instance two studies: the first of Petteri Hovi and his colleagues [26], on glucose regulation in young adults with very low birth weight <https://www.nejm.org/doi/pdf/10.1056/nejmoa067187>; the second of professor Kersti Pärna and her colleagues [39] regarding the alcohol consumption in Estonia and Finland, <https://doi.org/10.1186/1471-2458-10-261>. Have a look to their Tables:

Characteristic	Study Participants	Study Nonparticipants
<b>Very low birth weight</b>		
No. of subjects	166	89
Gestational age — wk	29.17±2.22	29.17±2.68
Birth weight — g	1120±221	1130±209

	n	Mean (SD) g/ week	Median g/ week
1994	362	128 (147)	79
1996	363	117 (110)	78

Maybe, the researchers, after having watched the shape of the data distribution, have decided that in the first study the numbers behave in a symmetric and unimodal way, and therefore the symbol  $\mu \pm \sigma$  (i.e. mean plus or minus standard deviation) to summarize data distribution can

be a proper choice. And, very likely, the second team realized that the weekly mean of alcohol consumption had a very long right tail, and they avoid the symbology  $\mu \pm \sigma$  which should have trapped the unaware reader in a pitfall, i.e. that in Estonia there might exist some drinkers whose body do not consume, but 'produce' alcohol during the weekend (as  $128 - 147 = -19!$ ).

This is the reason why, when data are skewed, many authors recommend to avoid to describe them using the mean and the standard deviation, and to prefer using the Tukey five numbers summary. There is also a well-posed mathematical reason to prefer such a recommended choice: the Čebišev inequality. In fact, [https://en.wikipedia.org/wiki/Chebyshev%27s\\_inequality#Probabilistic\\_statement](https://en.wikipedia.org/wiki/Chebyshev%27s_inequality#Probabilistic_statement), it is possible to create a set of artificial data  $x$ , all of them extremely far away from the mean, such that  $P(|x - M| \geq S) = 1$ .

Another pivotal point in the correct reporting of statistical facts concerns the well known Occam's Razor principle – *Frustra fit per plura quod potest fieri per pauciora*: it is not worth to provide a number of statistics greater than the collected data dimension. Here you have a funny example: suppose that Expert A checks 4439 images, and Expert B checks 4686. Suppose you want to communicate these **two** pieces of information: how would you write it in a paper? Have a look to **three** information solution chosen by Christer Sinderby and colleagues [45], <https://ccforum.biomedcentral.com/track/pdf/10.1186/cc13063.pdf>

### Results

#### Reliability of automated analysis

For the analysis of the datasets, the two expert analysts manually detected, on average, 4,562 (range 4,439 to 4,686) events (EAdi or P<sub>v</sub> events). ICCs for the NeuroSync<sub>MANU</sub>

## 2.4 Exercises

■ **Activity 2.1 — describe a dataset.** Search and read the paper by Mara Severgnini, Mario de Denaro et al., entitled *In vivo dosimetry and shielding disk alignment verification by EBT3 ...* (PMID 25679150). Read and understand the data of their Table 1 (page 118). Download the dataset `breastioert` from the repository <https://github.com/MassimoBorelli/Miramare> and import it into your JASP.

- obtain a table reporting absolute frequencies and relative frequencies of *Energy*
- obtain the median and compute the interquartile range of the *Collimator Diameter*
- obtain a boxplot of the *Area outside shielding*
- obtain a cartesian x-y scatter plot of the *Area outside shielding* versus the *Difference Expected Dose and Measured Dose*

Report the four outputs obtained, eventually arranged in a more readable form and go to <https://ictpmmp.weebly.com/assignments.html> in order to upload your report, complete with your name and surname, in a .pdf document.

■

## 3. Probability in medicine

### 3.1 Brief recalls on random variables

In medical statistics very often one deals with **finite random variables**. As an example (Table 4.3 in Bernard Rosner [43, page 84]) consider the number of episodes of otitis media in the first two years of life:

$$\begin{pmatrix} 0 & 1 & 2 & 3 & 4 & 5 & 6 \\ 0.129 & 0.264 & 0.271 & 0.185 & 0.095 & 0.039 & 0.017 \end{pmatrix}$$

The first row describe all the possible **events**, while the second row precise their single success probability; and the function which associates the event to its probability is called **probability mass function**, or **discrete density function**. In effect, those probabilities are simply a frequencies distribution, as we were dealing in Exercise 2.2: you can verify it by loading the `otitis` dataset from the <https://github.com/MassimoBorelli/Miramare> repository, and draw a barplot as explained in subsection 2.2.2.

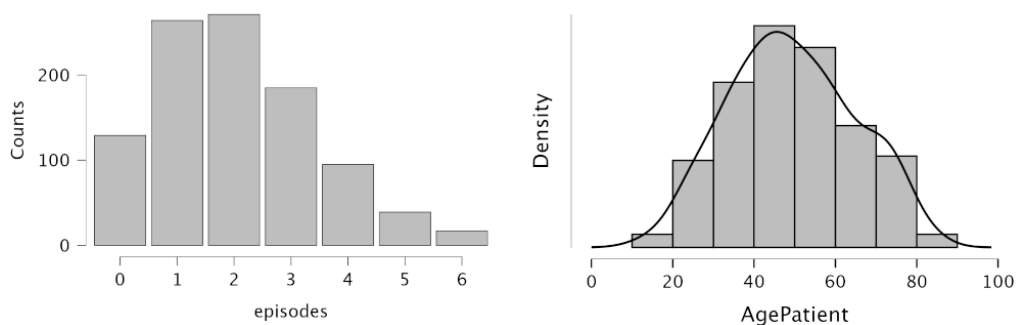
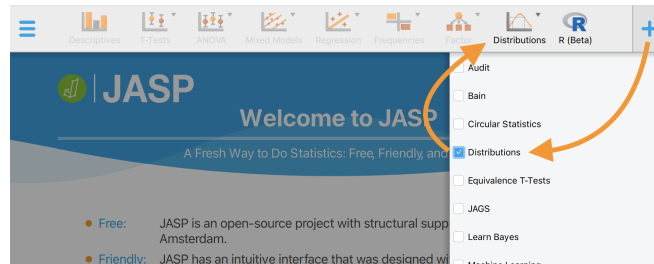


Figure 3.1: Estimating the probability density function of a continuous random variable

Moving to **infinite random variables**, JASP (or, better, the R language) possesses an inner algorithm which (depending on the user's choice of a bandwidth and of a kernel) fits a numerically

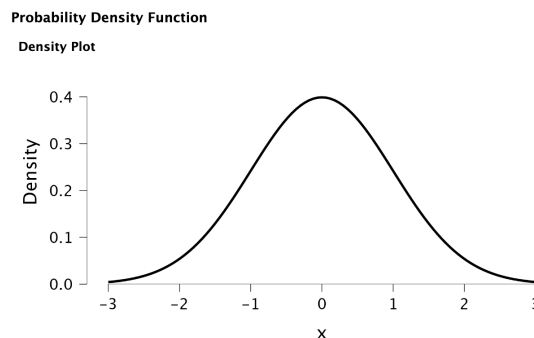
estimated density function, as in the right panel of Figure 3.1. Such curve relies on the histogram, which is of course an estimator[50] of the density function (which, in turns, depends on the starting point of the grid of bins – and the effect can be surprisingly large, as Venables and Ripley explain very well in their Figure 5.8 [50, pages 127-128]). The figure here depicted in right panel represents the AgePatient of the roma dataset, which will be presented in a while.

### 3.2 Commonly used random variables

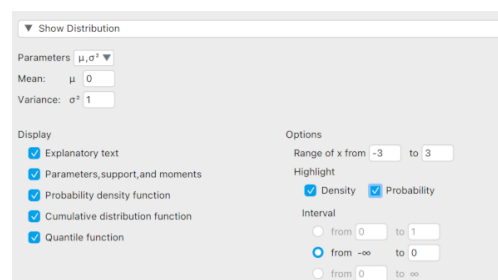


JASP possesses an additional menu which allows to study and to simulate the data random behaviour. Let us recap some basic facts on the most frequently used random variables in the medical field.

#### 3.2.1 The Normal Distribution



With JASP it is straightforward to perform calculations with the gaussian random distribution. The Show distribution menu provide to the user a nice way to reflect over the mathematical relations between the density function, the cumulative distribution and the quantile function, also highlighting the density and the probability evaluated over an interval, bounded or unbounded.

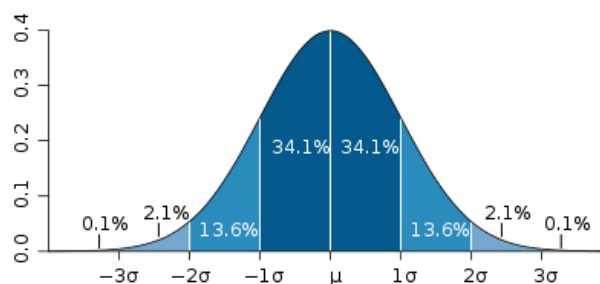


Let us try to move the parameters  $\mu$  and  $\sigma^2$  of the distribution in order to solve some typical textbook exercises.

**Exercise 3.1** (B. Rosner, example 5.22 [43, page 131]) The cerebral blood flow (CBF) in the general population is, approximately, normally distributed with mean  $\mu = 75$  and standard deviation  $\sigma = 17$ . Which could be the percentage of persons having a CBF  $< 40$ ? ■

**Exercise 3.2** (B. Rosner, example 5.23 [43, page 132]) Glaucoma is characterized by intraocular pressure greater than 20 mmHg, while in normal population intraocular pressure  $X$  has mean  $\mu = 16$  and standard deviation  $\sigma = 3$ . How much it could be  $P(12 \leq X \leq 20)$ ? ■

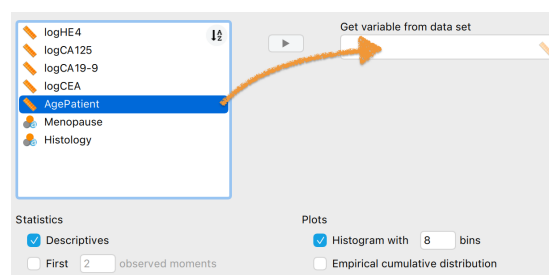
**Exercise 3.3** Can you find the upper and the lower fifth percentile of the intraocular pressure, as above defined? ■



**Exercise 3.4** (the 'three sigma' property) Can you 'explain' with JASP the above picture: [https://en.wikipedia.org/wiki/File:Standard\\_deviation\\_diagram.svg](https://en.wikipedia.org/wiki/File:Standard_deviation_diagram.svg) ■

Now, we can apply the Distribution menu possibilities to a real dataset: let us connect to the <https://github.com/MassimoBorelli/Miramare> repository and exploit the roma dataset. Actually, this name does not indicate the city, but the acronym of 'Risk of Ovarian Malignancy Algorithm', a method introduced more or less fifteen years ago by Richard Moore et al. [36], in order to estimate benign vs. malignant probability in an ovarian cancer. Doctor Shadi Najaf, a gynaecologist now at the Kantonsspital Baden, Zürich (Swiss), explored the possibility to enhance their algorithm, collecting data on 210 patients with an ovarian mass. She was seeking to know whether the Histology may be associated, in a statistical sense that will be precised better, to AgePatient, to their Menopause status, and to four candidate biomarkers (logarithmic transformed): logHE4, logCA125, logCA19.9 and logCEA.

Let us open roma into JASP and drag-and-drop the AgePatient variable into the Get variable from data set box, activating the histogram with 8 bins in order to compare it with the right panel of Figure 3.1.



Loosely speaking, it might seem that data behaves like a gaussian bell, with a  $\mu \approx 49.3$  and  $\sigma = 15.5$ . But a more efficient way to check it, is to introduce a very useful graph called the **quantile - quantile plot** (i.e. the **Q-Q plot**). When data are normally distributed, they (approximately) tends to lay on the 'diagonal' of the Q-Q plot (i.e. the red line intersecting the first and third quartile of the gray bullet shaped sample). To read in deep the details, see for instance <https://en.wikipedia.org/wiki/Q%E2%80%93plot>, or refer to our previous Lecture Notes [https://www.researchgate.net/publication/331571258\\_Medical\\_Statistics\\_with\\_R](https://www.researchgate.net/publication/331571258_Medical_Statistics_with_R). The Q-Q plot will be very useful in assessing the 'quality' of the linear models in the forthcoming pages.

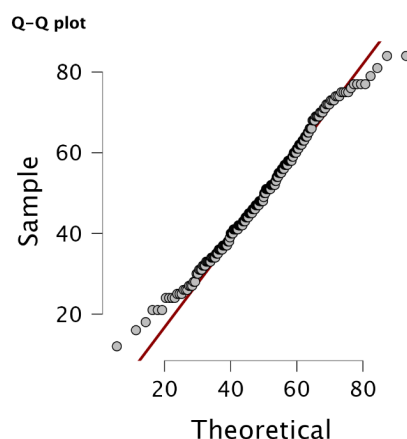


Figure 3.2: The quantile - quantile normal plot

### The sum of normal variables is, or is not, normal?

Do two dromedaries make a camel? It's a funny question, but there is in literature a bit of mess about the 'sum' of two normal variables. Let us read the authoritative Bernard Rosner [43, page 135]

.. linear combination of normal random variables are often of specific concern. It can be shown that any linear combination of normal random variables is itself normally distributed.

And now, let us move to Martin Bland [7, page 111]:

... If we add two variables from Normal distributions together, even with different means and variances, the sum follows a Normal distribution.

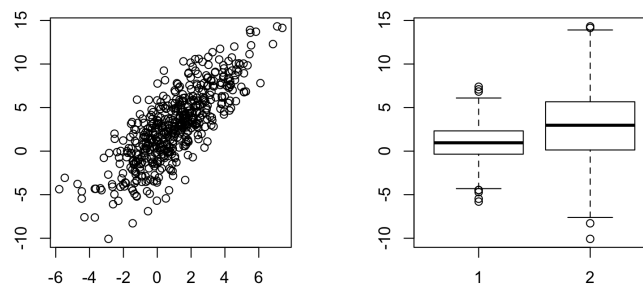
The two statements are misleading; it seems that there is a confusion between things happening  $\mathbb{R} \times \mathbb{R} = \mathbb{R}^2$  or in  $\mathbb{R}$ . As a famous counterexample, we recall the *Living histograms* of Brian Joiner [28, 31], in which the taller (mostly, boys) stay on the right of the photo of the next page, while the smaller (mostly, girls) are on the left: the distribution suggests an immediate bimodality, and therefore normality is clearly excluded (i.e. two dromedaries do not make a camel). We will discuss again such important case.



In particular, in a 1947 number of *Nature*, S. Vaswani [49] provide a counterexample, recalled and enlarged by C. Kowalski in his 1973 *Non-Normal Bivariate Distributions with Normal Marginals* [33]. And in 1982, E. Melnick and A. Tenenbein, with their *Misspecifications of the Normal Distribution* [34], provide a clear response:

Question 3: are linear combinations of normally distributed random variables always normal? The answer to this question is no and it can be illustrated by using the example in Question 2 ... linear combinations of normal random variables need not themselves be normal. The correct statement is that any linear combination of random variables from a multivariate normal distribution is normally distributed.

**www** In our previous Lecture Notes [https://www.researchgate.net/publication/331571258\\_Medical\\_Statistics\\_with\\_R](https://www.researchgate.net/publication/331571258_Medical_Statistics_with_R) one can find a simple code to generate one-dimensional and two-dimensional normal data. The picture below depicts a **bivariate normally distributed** cloud of 500 random points, respectively of mean 1 and 3, and standard deviation 2 and 4, on the x and y axes, with correlation of 75% (and we will discuss it better in the sequel).



### 3.2.2 The Lognormal Distribution

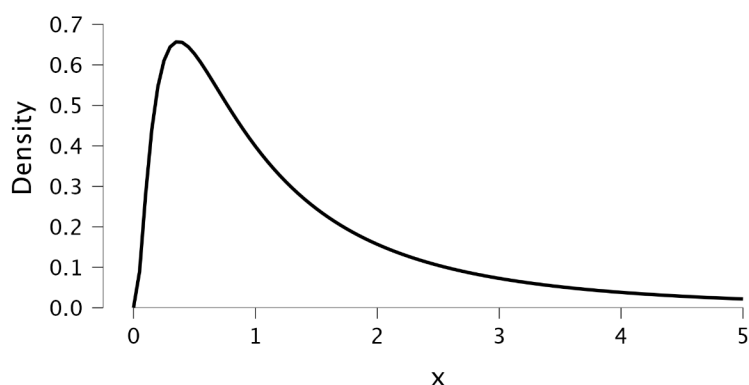
**www** Eckhard Limpert, et al. Log-normal Distributions across the Sciences: Keys and Clues <https://academic.oup.com/bioscience/article/51/5/341/243981>

Let us start recalling a fundamental result, the renowned 'Central Limit theorem' [https://en.wikipedia.org/wiki/Central\\_limit\\_theorem](https://en.wikipedia.org/wiki/Central_limit_theorem):



## Probability Density Function

Density Plot



**Theorem 3.2.1 — Lindenberg-Lévy Central Limit Theorem.** Suppose  $(X_i)_{i \in \mathbb{N}}$  is a sequence of independent and identically distributed random variables with  $E[X_i] = \mu$  and  $\text{Var}[X_i] = \sigma^2 < +\infty$ . Then as  $n$  approaches infinity, the random variables  $\sqrt{n}(S_n - \mu)$  converge in distribution to a normal  $N(0, \sigma^2)$ .

Now we can easily guess that multiplying (instead of adding) repeatedly the result of a random variable, the logarithm of the standardized distribution will be approximately normal (as an example, imagine to throw many dices and consider the product of the results). This is an insight to explain why many biological phenomena are modelled by a log-normal distribution: for instance, patients' body mass indexes [21]. Again, JASP allows to recap all the basic facts checking all the boxes of the Show distribution menu. We observe that, in general, in the log-normal distribution mean  $\neq$  median  $\neq$  mode.

**Discussion 3.2.2 — summarizing body mass index.** Have a look to the body mass index histogram of more than  $10^5$  patients studied by Gregg Fonarow, <https://doi.org/10.1016/j.ahj.2006.09.007>. Suppose that you are required to lead a pilot study concerning radiation dosimetry in 25 obese patients. How do you think you are going to describe the data? Using the mean and the standard deviation, or the median and the quartiles? What are here the difficulties?

### 3.2.3 The Binomial Distribution

Instead of speaking of tossing fair coins or picking balls from the urn, let us refer again to the Shadi Najaf roma dataset. We see that the Histology collects 39 malignant cancer over 210 patients (i.e.  $p \approx 39/210 = 0.186$ ).

**Exercise 3.5** Suppose that you collect a new sample of 210 women with the same symptoms of those enrolled in roma. Obviously, only by chance you will observe exactly '39' malignancies. Can you compute the probability to observe a number of malignancy between 30 and 50? ■

It is important to note that when the statistician seeks to fit a gaussian distribution on her/his data, there are two independent 'radio knob' to 'tune': the mean  $\mu$  and the standard deviation  $\sigma$ . With the binomial, on the contrary, there is a compulsory constraint which links the mean  $\mu$  to the variance  $\sigma^2 \equiv \mu \cdot (1 - p)$ , being  $p$  the elementary probability of success. This is the reason why often in papers you will read the sentence '*accounting for overdispersion*'.

Free parameter	Fixed parameter
Probability of success: $p$ <input type="text" value="0.186"/>	Number of trials: $n$ <input type="text" value="210"/>
Display	Options
<input type="checkbox"/> Explanatory text	Range of $x$ from <input type="text" value="20"/> to <input type="text" value="60"/>
<input type="checkbox"/> Parameters, support, and moments	Highlight
<input checked="" type="checkbox"/> Probability mass function	<input type="checkbox"/> Mass <input checked="" type="checkbox"/> Cumulative Probability
<input type="checkbox"/> Cumulative distribution function	Interval <input type="text" value="30"/> $\leq X \leq$ <input type="text" value="50"/>

**Discussion 3.2.3 — smallpox vaccine.** In Mould's 6.3 paragraph we read: *A binomial situation of historical importance is the work of Sir Edward Jenner on smallpox vaccination (an enquiry into the causes and effects of the variolae vaccinae, 1798). A sample of 23 people was infected with cowpox ( $n = 23$ ). The probability of contracting smallpox when inoculated with the virus was some 90% ( $p = 0.9$ ), but none of the previously vaccinated 23 people did in fact contract smallpox ( $r = 0$ ). The binomial probability of such an event occurring is exceedingly small, and the observations are therefore definitely not random. While with a programming language as R it is straightforward to compute such 'exceedingly small' probability, have you any idea on how to do it with JASP?*

### 3.2.4 The Poisson Distribution



Susan Holmes, Wolfgang Huber. *Modern Statistics for Modern Biology*  
<https://www.huber.embl.de/msmb/Chap-Generative.html>

Born as a distribution of the number of occurrences of a rare event, i.e. with 'small' probability  $p$  in  $n$  independent trials and closely connected to the binomial distribution [42], the Poisson distribution is nowadays applied not only to rare events but to generic 'count' problems. Indeed, Susan Holmes and Wolfgang Huber in their *Modern Statistics for Modern Biology* fantastic textbook introduce the discourse in Chapter 1 by means of such random variable.

As an introductory example related to cancer, let us consider the Figure 7.4 of Daniel Zips, *Tumour growth and response to radiation*, collected in [32]. Let us read his words about the local tumour control:

If not a single tumour but a group of tumours (or patients) is considered, the local tumour control probability (TCP) as a function of radiation dose can be described statistically by a Poisson distribution of the number of surviving clonogenic tumour cells (...). As an illustration, one might imagine that a given radiation dose causes a certain amount of 'lethal hits' randomly distributed within the cell population. Some cells will receive one 'lethal hit' and will subsequently die. Other cells will receive two or more 'lethal hits' and will also die. However, some cells will not be hit, will therefore survive and subsequently cause a local failure. According to Poisson statistics, a radiation dose sufficient to inflict on average one 'lethal hit' to each clonogenic cell in a tumour (number of 'lethal hits' per cell,  $m, = 1$ ) will result in 37 per cent surviving clonogenic cells.

In that example, the Poisson distribution has the intensity (i.e. the mean, also called 'rate parameter')  $\lambda = 0.37$ .

1				1	2
2	3	1	2	1	
	1		2		1
1	1	2		4	1
1		1		3	1
	2	1	1		

**Exercise 3.6** Use JASP to discover in a  $\lambda = 0.37$  Poisson distribution how many, in probability, cells could have a value greater or equal than 2.

Parameter  
Rate:  $\lambda$  .37

Display  
 Explanatory text  
 Parameters, support, and moments  
 Probability mass function  
 Cumulative distribution function

Options  
 Range of x from 0 to 6  
 Highlight  
 Mass  Cumulative Probability  
 Interval 2  $\leq X \leq$  6

For simulation purpose, JASP possesses some limited capabilities in managing the output when generating random numbers, being the latter possibility associated to a new column added to the current dataset. Here, only for didactical purpose, we show how it is possible to generate a sequence of Poisson distributed counts exploiting the R language 'inside' JASP:

The screenshot shows the 'R in JASP' window. At the top, there are icons for Factor, Distributions, and R (Beta). The main window contains a terminal with the following text:

```
Welcome to R in JASP!
> rpois(n = 36, lambda = 0.37)
[1] 0 0 0 1 0 0 0 1 0 1 0 0 0 0 1 0 0 1 1 1 0 1 0 1 1 1 0 0 1 0 0 2
0 0 0 0 0 0
```

Below the terminal, there is a text input field containing the code `rpois(n = 36, lambda = 0.37)`. To the right of this field are two buttons: 'Run Code' and 'Clear Output'. A yellow tooltip above the 'Run Code' button says 'Pressing Ctrl+Enter or F5 will also run the code'. An orange arrow points from the 'R (Beta)' icon to the terminal window, and another orange arrow points from the 'Run Code' button to the code input field.

### 3.3 Evaluating odds and risks: Bayes theorem



Viv Bewick, Liz Cheek, Jonathan Ball. Statistics review 11: Assessing risk  
<https://ccforum.biomedcentral.com/articles/10.1186/cc2908>

Aging is recognized to be a risk factor for the ovarian cancer; therefore, not surprisingly, in roma dataset a **contingency table** exploring joint frequencies of Menopause and Histology could provide some clues: menopausal status indeed is a (coarse) statistical **proxy** of age. But JASP reveals two possible way to follow, the **Classical** and the **Bayesian** one. Let us start with the first one:

In Table 3.1 we see that 39 women over 210 has been diagnosed with a malignant ovarian tumor; so one could estimate the **relative frequency**, i.e. an estimate of the disease (**frequentist**) **probability**

	logHE4	logCA125	logCA19-9	logCEA	AgePatient	Histology
1	3.58	4.25	3.33	0.22	64	benign
2	3.42	5.45	4.84	0.24	21	benign
3	5.68	4.72	3.2	0.92	64	malignant
4	4.14	3.96	3.54	1.76	58	malignant
5	3.57	3.03	-0.04	1.03	74	benign
6	3.7	4.11	3.44	0.58	40	benign
7	7.17	7.58	2.45	0.44	51	malignant
8	3.57	2.48	1.46	0.1	21	benign
9	3.97	3.64	2.3	0.14	27	benign
10	4.11	4.03	4.73	0.82	75	post

Histology	Menopause		Total
	ante	post	
benign	106	65	171
malignant	12	27	39
Total	118	92	210

Table 3.1: Menopausal status is a predictor, or a confounder, of malignancy in ovarian cancer?

to be around the 19 percent (of course not within the whole healthy population, but within women with certain precise symptoms known to the gynæcologists):

$$P(\text{malignant}) = \frac{39}{210} = 0.186\dots$$

**Vocabulary 3.1 — Prevalence.** In a cross-section design, the **prevalence** of the disease into a selected subpopulation described by some precise **inclusion criteria** is represented by its (frequentist marginal) probability.

Such marginal probability does not distinguish whether women are in their ante-menopausal or post-menopausal status. So we look to the inner columns of the table, i.e. we estimate the **conditional probability**:

$$Pr(\text{malignant}|\text{ante}) = \frac{12}{118} = 0.102\dots$$

$$Pr(\text{malignant}|\text{post}) = \frac{27}{92} = 0.293\dots$$

Those numbers appears to be different in a pure mathematical sense: a post-menopausal woman appears to have a triple risk than an ante-menopausal woman. Therefore, we can argue that Menopause and Histology are not **independent events**, but they are (in a statistical sense to be better precised later) **associate events**.

By the way, we recall here two commonly used **association measure**; the first is the **odds ratio**:

$$O.R. = \frac{106 \cdot 27}{65 \cdot 12} = 3.67$$

and when O.R. is 'far away from' 1 (i.e. close to 0 or to  $+\infty$ ), then rows – and columns – are 'far away' from proportionality, and therefore one event (e.g. menopausal status ante / post) provide 'a

certain quantity of information' to the other event (e.g. to be ante / post inform us on benign / malignant response). Another common association measure is the **relative risk** (i.e. the ratio of the conditional probabilities):

$$RR = \frac{\frac{27}{92}}{\frac{12}{118}} = \frac{27}{92} \cdot \frac{118}{12} = 2.89$$

**Exercise 3.7** Explore the output of the Odds Ratio ( $2 \times 2$  only) checkbox in the Statistics menu of the contingency table of Histology (Rows) versus Menopause (Columns). ■

### 3.3.1 Bayes theorem

In a contingency table, marginal probabilities and conditional probabilities are ruled by the famous **Bayes theorem**:

$$P(\text{malignant}|\text{ante}) = \frac{P(\text{ante}|\text{malignant})}{P(\text{ante})} \cdot P(\text{malignant})$$

**Vocabulary 3.2 — Prior and posterior probability.** In the Bayes theorem, the marginal  $P(\text{malignant})$  probability is called the **a priori probability**, while the conditional  $P(\text{malignant}|\text{ante})$  probability is the **a posteriori probability**.

Although the proof is straightforward, we do not spend time in this task, but simply we check the relation with our example:

$$\begin{aligned} \frac{12}{118} &? \frac{(12/39)}{(118/210)} \cdot \frac{39}{210} \\ \frac{12}{118} &? \frac{12}{39} \cdot \frac{210}{118} \cdot \frac{39}{210} \\ \frac{12}{118} &\equiv \frac{12}{118} \end{aligned}$$

We will discuss in detail why Bayes theorem is so important in statistical inference. Let us conclude this section recalling some relevant concepts in medical statistics, when we are required to evaluate the 'performance of a diagnostic test'.

**Vocabulary 3.3 — Sensitivity and specificity.** In a cross-section design, the **sensitivity** is the probability of a positive test in people with the disease, while **specificity** is the probability of a negative test in people without the disease.

In our Table 3.1, sensitivity and specificity are the conditional probabilities  $P(\text{post}|\text{malignant})$  and  $P(\text{ante}|\text{benign})$ ,  $Sens = 27/39 = 69\%$ , while  $Spec = 106/171 = 62\%$ . Sensitivity and specificity are characteristics of a test and are not affected by the *prevalence* of the disease [6].

Nevertheless, those two quantities are not suitable in assessing the 'quality', the 'usefulness' of a clinical test (i.e to answer to the question 'is it relevant to know about the menopausal status in order to foresee malignancy?'). Therefore one considers [37]:

**Vocabulary 3.4 — Predictive values.** In a cross-section design, the **positive predictive value** (PPV) is the probability of the person having the disease when the test is positive, while the **negative predictive value** (NPV) is the probability of the person not having the disease when the test is negative.

In our Table 3.1,  $PPV = P(\text{malignant}|\text{post}) = 27/92 = 29\%$  and  $NPV = P(\text{benign}|\text{ante}) = 106/118 = 90\%$ . Unfortunately, although the PPV and NPV give a direct assessment of the usefulness of the test, they are affected by the prevalence of the disease [6]. This is the reason why often researchers move to the **likelihood ratios** [6]. For these and other concepts as likelihood ratios, pre-test probability, post-test odds, Youden's index see:



Viv Bewick, Liz Cheek and Jonathan Ball. Statistics review 13: Receiver operating characteristic curves  
<https://ccforum.biomedcentral.com/articles/10.1186/cc3000>

### 3.3.2 The Bayes factor



Wikipedia. Bayes factor  
[https://en.wikipedia.org/wiki/Bayes\\_factor](https://en.wikipedia.org/wiki/Bayes_factor)

We need to introduce an important concept, the **Bayes factor**, and we do it with a simple, artificial, example, similar to the one presented in Wikipedia. Alice has a balanced urn with 5 winning black balls and 5 white balls ( $p = 0.5$ ), Bob has a tricky urn with 6 winning black balls and 4 white balls ( $p = 0.6$ ). Suppose that, in a pure binomial scheme, the extractions with replacement, we observe 115 successes over 200 draws, but without knowing if they are generated from Alice's or Bob's urn.

If we compute with JASP, as shown in Figure 3.3.2, the conditional probabilities:

$$P(X = 115 | Alice) = \binom{200}{115} \cdot 0.5^{115} \cdot 0.5^{200-115} \approx 0.006$$

$$P(X = 115 | Bob) = \binom{200}{115} \cdot 0.6^{115} \cdot 0.4^{200-115} \approx 0.044$$

we observe that it is much more likely that the balls have been drawn by Bob's urn: its probability is about seven times higher than Alice's one. The ratio  $P(X = 115 | Alice) / P(X = 115 | Bob)$  represents what is called the Bayes factor.

```

> dbinom(115, 200, 0.5)
[1] 0.005955892

> dbinom(115, 200, 0.6)
[1] 0.04399862

```

More formally, if we have observe some data  $D$  and we have two different generative models  $M_1$  and  $M_2$  and we desire to quantify the 'plausibility', the 'preferability' for a model over another, the Bayes factor is defined to be:

$$\frac{P(D|M_1)}{P(D|M_2)} = \frac{P(M_1|D)}{P(M_2|D)} \cdot \frac{P(M_2)}{P(M_1)}$$

In next chapter we will appreciate the importance of evaluating the Bayes factor as a foundations of the JASP software.

### 3.4 Sample and population: approaching inference



Elise Whitley, Jonathan Ball. Statistics review 2: Samples and populations  
<https://ccforum.biomedcentral.com/articles/10.1186/cc1473>

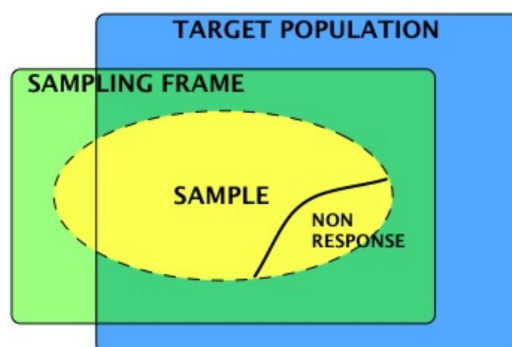
In medical (and other) research there is generally some population that is ultimately of interest to the investigator (...). It is seldom possible to obtain information from every individual in the population, however, and attention is more commonly restricted to a sample drawn from it. The question of how best to obtain such a sample is a subject worthy of discussion in its own right and is not covered here. Nevertheless, it is essential that any sample is as representative as possible of the population from which it is drawn, and the best means of obtaining such a sample is generally through random sampling.

The above quotation, from Elise Whitley and Jonathan Ball, clearly introduces the matter: we collect data from a **sample** of patients and we are required to analyse them in order to provide some general conclusions, possibly valid for the whole **population** whose that sample belongs to. Richard Mould's words depicts even better the situation:

In statistical parlance the term population refers to the group of objects, events, results of procedures or observations (rather than the geographical connotation of population relating only to persons in a country or state etc) which is so large a group that usually it cannot be given exact numerical values for statistics such as the population mean  $\mu$  or the population standard deviation  $\sigma$ . These statistics therefore can only be estimated.

To obtain for example, an estimate of the population mean  $\mu$  of a certain characteristic  $x$  of the population, *sampling* must first take place because all the values of  $x$  for the entire population cannot be measured. Only a small part of the population can be surveyed and that part is called a *sample*.

There are various methods of sampling, including *random sampling*, which for clinical trials is discussed in a later chapter as simple randomisation, stratified randomisation and balanced randomisation.



The random sampling is a sort of 'life insurance' against the **sampling bias** issue: we have to be aware that, as shown in the above Figure 3.4 by Stefano Panzeri [38], that our data could be affected by a not-random sort of 'distorsion' and, in the typical research framework of medical statistics, when data are already collected we can neither detect it nor fix it.





Stefano Panzeri, Cesare Magri and Ludovico Carraro. Sampling bias.  
[http://www.scholarpedia.org/article/Sampling\\_bias](http://www.scholarpedia.org/article/Sampling_bias)

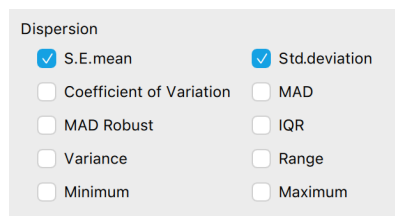
Statistical inference relies on two different perspectives, which have been established during the decades on sound mathematical foundations by, among others, Bruno de Finetti (*'probability does not exist'*) for the concept of subjective probability and Richard von Mises (*'probability theory is long sequences of experiments or observations repeated very often and under a set of invariable conditions'*) for the frequentist definition of probability.



Figure 3.3: Bruno de Finetti, from Trieste, and Richard von Mises, from Lviv, two borderline cities at the Austro-Ungarian empire at the end of nineteenth century. Source: Wikipedia.

### 3.5 Mismatching variability with reliability

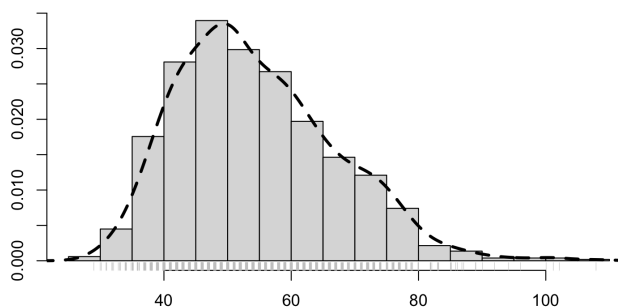
It is now the suitable moment to 'unblur' the image relative to Discussion 2.2.1: JASP collects under the descriptive menu the **standard error of the mean** index.



One might wonder if this is a proper choice. Let us carefully read R. Mould's words written in his 4.1 paragraph:

The standard deviation  $s_m$  of the sample mean  $x_m$  tells you about the spread of the measured sample values  $x_1, x_2, \dots, x_i, \dots$  (...) If the *sampling experiment* to measure  $x_m$  is then repeated  $N$  times, with the sample size  $n$  always remaining the same, a total of  $N$  values of  $x_m$  will be obtained. If these are then averaged, then  $M$ , which is the *mean of means* or *grand mean* is obtained. The standard deviation of the mean of means  $M$  is given a special name: standard error of the mean, where  $SE = \text{Sample Standard Deviation} / \sqrt{n}$

To clarify the concept, we try a simulation. Let us import into JASP the `cholesterol` dataset concerning 1025 Triestiners healthy blood donors, from the <https://github.com/MassimoBorelli/Miramare> repository.



The picture above depicts their HDLcho1 high density lipo-protein cholesterol levels skewed distribution, whose mean  $m$  is approximately 54.7. We are interested in estimating the unknown HDL cholesterol mean level  $\mu$  of the whole Triestine healthy population: could be  $m = 54.7$  a plausible candidate? Well, naively, we can suspect that blood donors represents a biased random sample of the overall target population (which comprises also not donors: babies, elderlies and diseased people). Nevertheless, for exercise, we try a simulation.

We activate the R in JASP window: in this environment to the active dataset the standard name `data` is attributed, so the variable of our interest is coded as `data$HDLcho1`. As a start try, let us extract 49 random values (why 49? Only because it is something squared,  $49 = 7^2$ ):

```
sample(data$HDLcho1, 49)
```

```
Cleared...
> sample(data$HDLchol, 49)
[1] 43 53 46 59 54 40 65 42 45 70 72 51 70 37 38 54 41 57 57 44 76 45 50 46 53
[26] 55 55 68 53 67 40 54 63 39 64 60 63 61 43 70 72 50 42 38 33 49 46 41 63
```

```
sample(data$HDLchol, 49)|
```

Run Code

Clear Output

The idea is to compute the mean of this sample, to store it into a memory numeric vector of dimension, say, 1000 and to repeat such calculation for 1000 times by means of a for cycle:

```
memory = numeric(1000); for(i in 1:1000){memory[i] = mean(sample(data$HDLchol, 49))}
```



```
mean(memory)
sd(memory)
sd(data$HDLchol)/7
```

We observe one good thing: the mean of memory, i.e. the *mean of means* in Mould's world, is 54.6 and it appears to be very similar to the mean  $m = 54.7$  of the HDLchol data. But what about variability?

Table 3.2: Descriptive Statistics

	HDLchol
Valid	1025
Mean	54.685
Std. Error of Mean	0.387
Std. Deviation	12.392

Originally, the standard deviation of HDLchol was 12.39, while now the standard deviation of memory is very different, 1.71. Does it exist any relation between those two numbers? Well, the first 'relation' is that they have the same name, because they measure the variability of their data. But the second relation is that 1.71 measures the variability of a well defined statistics estimator, the **sample mean**. And, not surprisingly, the Jakob Bernoulli Weak Law of Large Numbers Theorem states that the standard deviation of the sample mean is exactly  $\sigma/\sqrt{n}$  and in fact:

$$\frac{\sigma}{\sqrt{n}} = \frac{12.39}{\sqrt{49}} = \frac{12.4}{7} \approx 1.77$$

and such result is really close to 1.71, the standard deviation of memory, which is indeed the **standard error of the mean**  $\sigma/\sqrt{n}$ , which is a **measure of reliability** [7, 13]. of estimating the unknown parameter  $\mu$ , the mean of the high density lipo-protein within the target population (incidentally, observe the elegant bell shape of memory: this is a consequence of the Central limit theorem 3.2.1).

In conclusion: do not confuse variability with reliability and do not confuse standard deviation with standard error. Sukhbir Kaur et al. in their repeated measurement experiments concerning certain gene silencing, curiously perform some experiments three times, and other in a fourfold replicate. And much more curiously, in the former cases they summarize data variability with the standard deviation, and in the latter with standard errors... very mysterious.

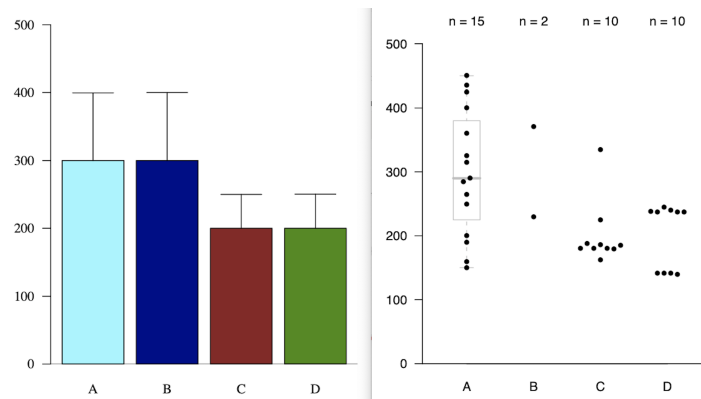
transfected endothelial cells. **C** shows migration assay for control *lacZ* and *robo4* siRNA transfected cells to Serum or AP-Slit2N in either upper (U), lower (L) or both chambers as indicated. Error bars in **A** (n = 3), and **B** (n = 3) represent SD while in **C** represent SEM (n = 4). **D** shows pulldown analysis of Cdc42-GTP levels in AP and AP-Slit2N (25 ng/ml) treated endothelial cell lysates for 5 and 15 minutes respectively. **+** indicates

The error bars are very frequently exploited in biomedical literature to present experimental data collected with repeated measures. But many statisticians agree with Tatsuki Koyama, now at the Vanderbilt School of Medicine, which calls such very dangerous diagrams the **dynamite plots**: the do not convey important information and they are usually misleading. His poster is worth reading:



Tatsuki Koyama. Beware of Dynamite

<https://biostat.app.vumc.org/wiki/pub/Main/TatsukiRcode/Poster3.pdf>



And if you are delighted about such foggy world and want to discover further 'epic fails' concerning the London Royal Mint and its six centuries mistake, or the 1.7 USD billion badly spent by Bill and Melinda Gates Foundation in wrong support to schools, refer to Richard Wainer tells in his *The most dangerous equation* [51] and its natural sequel by Yu-Kang Tu and Mark Gilthorpe.

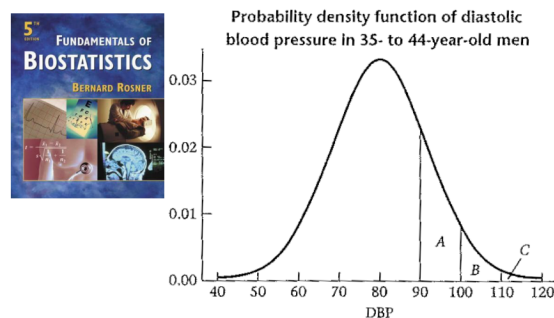


Richard Wainer. The most dangerous equation  
[https://www.researchgate.net/publication/255612702\\_The\\_Most\\_Dangerous\\_Equation](https://www.researchgate.net/publication/255612702_The_Most_Dangerous_Equation)



Yu-Kang Tu and Mark Gilthorpe. The most dangerous hospital or the most dangerous equation?  
<https://bmchealthservres.biomedcentral.com/articles/10.1186/1472-6963-7-185>

### 3.6 Exercises




■ **Activity 3.1 — the normal distribution.** Simply referring to the above graph as proposed by Bernard Rosner, concerning the normally distributed diastolic blood pressure, are you able to evaluate by means of JASP the probabilities of region *A*, *B* and *C*? ■





## 4. T-Test: the history of biostatistics

### 4.1 Detecting a signal from noise

 Student. The probable error of a mean.  
[http://seismo.berkeley.edu/~kirchner/eps\\_120/Odds\\_n\\_ends/Students\\_original\\_paper.pdf](http://seismo.berkeley.edu/~kirchner/eps_120/Odds_n_ends/Students_original_paper.pdf)

In 1908 it appeared on a newly trendy journal called *Biometrika*, <https://en.wikipedia.org/wiki/Biometrika>, a fundamental paper [47] signed by an anonymous author called Student. For decades the mysterious halo surrounded the identity of the author, which actually was the mathematician and chemist William Gosset, head of the experimental department of the Guinness brewery in Dublin (for other fascinating details, consult: [https://en.wikipedia.org/wiki/William\\_Sealy\\_Gosset](https://en.wikipedia.org/wiki/William_Sealy_Gosset)). The paper clarifies two very important topics:

1. in a random sample from a gaussian distribution  $N(\mu, \sigma)$ , estimating the sample mean  $m$  do not convey any information in estimating the sample standard deviation  $s$ , and vice versa.
2. the random variable  $t = \frac{m - \mu}{s/\sqrt{n}}$  possesses an explicit density function, which is not a gaussian, but can be numerically computed.

Although more than a century has elapsed, the paper is a masterpiece still worth reading. Here, first of all, we need to precise why the quantity

$$t = \frac{m - \mu}{s/\sqrt{n}}$$

is of our interest<sup>1</sup>. To do it, let us exploit the concept of **signal to noise ratio**, with the words of Stephen Ziliak and Deirdre McCloskey in their *The Cult of Statistical Significance* magistral paper [55]:

<sup>1</sup>The quantity  $t$  is usually called **test statistic**, and this is a sort of pun, and source of confusion, in various languages of the World: while in English and in Spanish the words 'Statistics' and 'Estadística' means the science, and 'the test statistic' and 'el estadístico de test' means the  $t$  – and the word 'statistic' is a synonym of 'summary' –, in French and in Italian 'Statistique' and 'Statistica' do not differ from 'la statistique test' and 'la statistica test'. Very confusing!



The signal to noise ratio is calculated by dividing a measure of what the investigator is curious about – the sound of a Miles Davis number, the losing of body fat, the yield of a barley variety, the impact of the interest rate on capital investment – by a measure of the uncertainty of the signal, such as the variability caused by static interference on the radio or the random variation from a smallish sample.

In the final pages, William Gosset illustrates its method providing concrete examples; in particular one question is to decide whether an *ante-litteram* 'agricultural biotechnology' treatment is useful, or not, in increasing the production of beer, i.e. to dry seeds into a special oven before seeding them. Here Gosset's words:

To test whether it is advantage to kiln-dry barley seed before sowing, seven varieties of barley were sown (both kiln-dried and not kiln-dried) in 1899 and four in 1900; the results are given in the table (4.1), expressed in Lbs. head corn per acre.

Not Kiln-Dried	Kiln-Dried	Difference
1903	2009	+106
1935	1915	-20
1910	2011	+101
2496	2463	-33
2108	2180	+72
1961	1925	-36
2060	2122	+62
1444	1482	+38
1612	1542	-70
1316	1443	+127
1511	1535	+24

Table 4.1: The original data of Student published in *Biometrika* [47, page 24].

#### 4.1.1 Classical One-sample t test



Elise Whitley, Jonathan Ball. Statistics review 5: Comparison of means  
<https://ccforum.biomedcentral.com/articles/10.1186/cc1548>

Table 4.2: Descriptive Statistics	
	difference
Valid	11
Missing	0
Mean	33.727
Std. Error of Mean	19.951
Std. Deviation	66.171

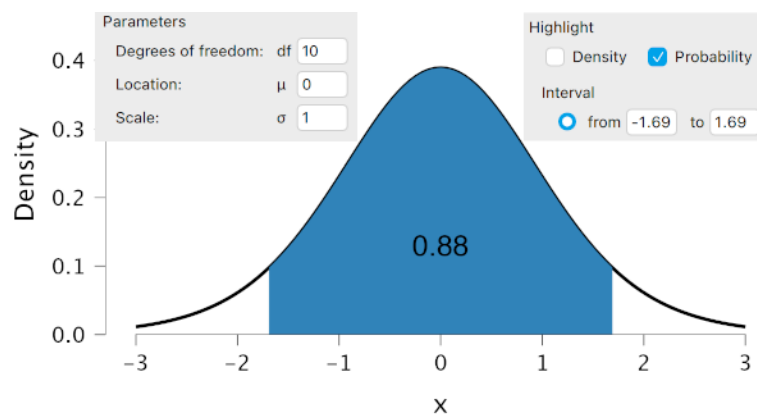
Let us import in JASP the `gossett` dataset, stored as usual in the <https://github.com/MassimoBorelli/Miramare> repository: the difference between `nkd` not treated (not kiln-dried) and `kd` treated (kiln-dried) seeds collects eleven data. The goal is to compare the **experimental**

**result**  $m = 33.7$  with the theoretical hypothesis that to treat or not to treat provide the same effect: this is the so-called **null hypothesis**, i.e.  $\mu = 0$ .

One therefore is interested in evaluating the 'distance' of these two quantities,  $x_m - \mu$ , from a statistical point of view; that is, to decide if  $|x_m - \mu|$  could be considered a null distance, or not. In other words, if the signal  $|x_m - \mu|$  differs from the noise  $s/\sqrt{n}$ . We could proceed by hand, with chalk and blackboard:

$$t = \frac{33.727 - 0}{66.171 / \sqrt{11}} = \frac{33.727}{19.951} \approx 1.690$$

Now,  $t = 1.69$  represent a quantile, but of what random variable? The one studied by William Gosset, nowadays simply called  $t$ . If we search within the Distributions menu, we may compute the probability to observe a signal to noise ratio smaller than 1.69 with respect to the  $t$  distribution with 10 degrees of freedom (why 10 degrees of freedom? Because 11 are the numbers, but 1 information has already been 'consumed' in order to compute the sample mean  $m = 33.727$ ):



As a trivial consequence, the white area outside is approximately equal to 0.12: this is exactly what we can immediately read when performing the **Classical** One Sample T-Test in the JASP menu:

Table 4.3: One Sample T-Test

	t	df	p
difference	1.690	10	0.122

So, what we can conclude? What decision do we make? A bit of suspense ...

### 4.1.2 Classical Two-sample paired t test

There exists another proper methodology to achieve the previous result: to perform the **Classical Paired Samples T-Test**, a typical statistical procedure exploited in the **longitudinal** experimental design, where (a couple of) repeated measures are collected on the same subject. Dragging and dropping *kd* and *nkd* into the Variable Pairs slot, we obtain the same previous result:

Table 4.4: Paired Samples T-Test

Measure 1	Measure 2	t	df	p
kd	- nkd	1.690	10	0.122

And, again, what decision can we conclude? Here we go.

## 4.2 Ronald Fisher's idea on significance level

A Lady declares that by tasting a cup of tea made with milk she can discriminate whether the milk or the tea infusion was first added to the cup. We will consider the problem of designing an experiment by means of which this assertion can be tested. (...) Our experiment consists in mixing eight cups of tea, four in a way and four in the other, and presenting them to the subject for judgement in a random order. (...) It is open to the experimenter to be more or less exacting in respect of the smallness of the probability he would require before he would be willing to admit that his observation have demonstrated a positive result. (...) Thus, if he wishes to ignore results having probabilities as high as 1 in 20 ...

In 1937 sir Ronald Aylmer Fisher started his fundamental book *The design of experiments* [20] presenting such a curious experiment. In this passage there are at least three relevant points. Let us discuss them briefly.

### 1. The conventional significance level of 5%.



Elise Whitley, Jonathan Ball. Statistics review 3: Hypothesis testing and P values  
<https://ccforum.biomedcentral.com/articles/10.1186/cc1493>

Fisher considered reasonable that 1/20, i.e. 5%, might be a critical level of probability, usually called **the  $\alpha$  significance level**, convincing you that what has happened is not 'chance'. Therefore, turning back to the Gosset data, we computed a probability of 12.2% that the observed effect on dried barley  $m = 33.727$  is simply due to chance. This  $p = 0.122$  probability is named the **p-value** of the test with respect to the so-called **null hypothesis  $H_0$** . In detail, in the One-Sample T-Test the null hypothesis is  $H_0 = \{\mu = 0\}$ , while in the Two-sample paired T-Test  $H_0 = \{\mu_{nkd} = \mu_{kd}\}$ . So, practically, the decision is:

- $p\text{-value} < \alpha \equiv 0.05$ ? Reject null hypotheses, there is an effect (kiln dried barley is different from not-kiln dried barley)
- $p\text{-value} > \alpha \equiv 0.05$ ? Do not reject hypotheses, we are not sure there is an effect (maybe no effect at all?)

### 2. The freedom to choose the significance level.



Douglas Curran-Everett and Dale Benos. Guidelines for reporting statistics in journals published by the American Physiological Society  
<https://journals.physiology.org/doi/full/10.1152/japplphysiol.00513.2004>

Fisher's sentence '*It is open to the experimenter to be more or less exacting in respect of the smallness of the probability he would require*' clearly leaves open to the researchers the choice about how much it has to be the  $\alpha$  significance level. During the decades putting  $\alpha = 0.05$  has become a sort of mystic cult (Ziliak and McCloskey,[55]) and important debates have been raised (Ioannidis [29]), leading the American Statistical Association to release an official opinion with their *The ASA's statement on p-values* [52] (to quickly access to such free papers see [https://padlet.com/massimo\\_borelli/sxa0vfqojwx1](https://padlet.com/massimo_borelli/sxa0vfqojwx1)).

As a rule, we can follow Douglas Curran-Everett [14], when defining and justifying a critical significance level appropriate to the goals of the study:

For any statistical test, if the achieved significance level  $P$  is less than the critical significance level  $\alpha$ , defined before any data are collected, then the experimental effect is likely to be real (...). By tradition, most researchers define  $\alpha$  to be 0.05: that is, 5% of the time they are willing to declare an effect exists when it does not. These examples illustrate that  $\alpha = 0.05$  is sometimes inappropriate.

If you plan a study in the hopes of finding an effect that could lead to a promising scientific discovery, then  $\alpha = 0.10$  is appropriate. Why? When you define  $\alpha$  to be 0.10, you increase the probability that you find the effect if it exists.

In contrast, if you want to be especially confident of a possible scientific discovery, then  $\alpha = 0.01$  is appropriate: only 1% of the time are you willing to declare an effect exists when it does not.

So, again turning back to the Gosset data, it would be wise to state that being a pilot study we a priori decided to set an  $\alpha = 0.10$  significance level – and being  $p = 0.122$  – that the experiment does not reach the statistical significance, i.e. we can not exclude that the difference in drying barley or not is due to chance.

### 3. significance level and sample size impact on the test power



Elise Whitley, Jonathan Ball. Statistics review 4: Sample size calculations  
<https://ccforum.biomedcentral.com/articles/10.1186/cc1521>

If one makes a little of combinatorics [https://en.wikipedia.org/wiki/Lady\\_tasting\\_tea](https://en.wikipedia.org/wiki/Lady_tasting_tea) one discover that the probability that the Lady correctly guesses the tasting cups is  $1/70 \approx 0.014 < 1/20 = 0.05$ : therefore implicitly recognise that obtaining a  $p < \alpha$  is equivalent to a 'zero error' situation. But changing the number of cups, i.e. changing all the  $\binom{n}{k}$  necessarily would move that 'zero error' situation, possibly admitting, one, two and even more errors as negligible. In fact, the  $p$  value depends on  $N$ , and in a slight complicate manner.

Let us quote Mould's paragraph 8.4 [37] words:

There are two types of error which can be made in arriving at a decision about the null hypothesis,  $H_0$ . A type-I error is to *reject  $H_0$  when in fact it is true* and a type-II error is to *accept  $H_0$  when in fact it is false*. By convention the probability of a type-I error is usually denoted by  $\alpha$  and the probability of a type-II error by  $\beta$ . (...) The probability  $1 - \beta$  is defined as the *power* of the test of the hypothesis  $H_0$  against an alternative hypothesis.

By analogy, a judge starts from the hypothesis  $H_0 =$  'this defendant is innocent'; the type-I error is to *reject innocence when in fact it is true* and to imprison an innocent. And a type-II error is to *accept innocence when in fact it is false*, i.e. to release a culprit. Usually, in practice, many researchers as a default put  $\alpha = 0.05$  and  $\beta = 0.20$ , i.e the power  $1 - \beta = 0.80$ .

The R language possesses a particular function which is able to compute any one of the quantity desired; here, in the Gosset example of the dried barley, the sample size is so 'limited' (with respect to the variability exhibited) that the power is about 33%, far away from common accepted limit of 80%: so Gosset had a very high probability to decide that the drying was unuseful when in effect the truth was just the opposite. Here the proper syntax:

```
> power.t.test(n = 11, delta = (mean(Difference) - 0),
              sd = sd(Difference), sig.level = 0.05,
              power = NULL, type = "one.sample")
```

```
> power.t.test(n = 11, delta = (33.727 - 0), sd = 66.171,
              sig.level = 0.05, power = NULL, type = "one.sample")

One-sample t test power calculation

      n = 11
  delta = 33.727
     sd = 66.171
sig.level = 0.05
  power = 0.3334406
alternative = two.sided
```

```
power.t.test(n = 11, delta = (33.727 - 0), sd = 66.171,
              sig.level = 0.05, power = NULL, type = "one.sample")
```

Run Code

Clear Output

Therefore, we have a clue: the experiment has been performed in a 'paucity of data' condition, i.e. with a too small sample size.



The power calculation here shown has only a didactical interest, but is is unuseful – see John Hoening and Dennis Heisey, *The Abuse of Power: The Pervasive Fallacy of Power Calculations for Data Analysis* [24].

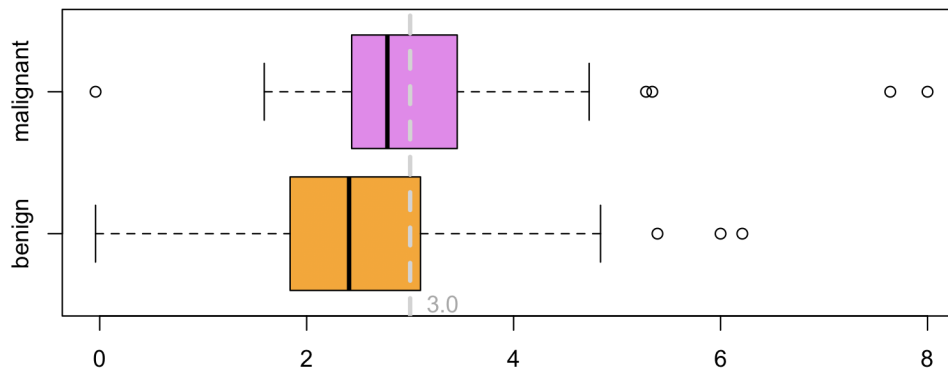
### 4.3 Out of the frying pan into the fire: statistical or clinical significance?

We try to clarify the point with an example. Suppose that we want to assess the role of the carbohydrate antigen 19-9, logCA19.9, as a predictor of the ovarian cancer in the **roma** dataset. In the next Chapter we will discuss the details, but suppose to know that the proper test shows no doubt about its *statistical significance*, exhibiting a smashing p-value = 0.004.

Nevertheless, a simple boxplot enlightens the fact that although CA19-9 may be 'significant' it is not 'useful', i.e. *clinically significant* in detecting ovarian pathology. Suppose for instance that a woman with symptoms has logCA19.9 = 3.0. Of course, such a value is closer to the malignant group mean 3.2 than to the benign group mean 2.4, but basing on the 3.0 information to guess histology is nothing more than looking into a crystal ball:

Let us in conclusion read what Richard Mould claims in his 8.3.2 paragraph [37]:

One of the problems encountered by those involved with statistics is how, and with what accuracy, inferences can be drawn about the nature of a population when the only evidence which exists is that from samples of the population. In order to solve this problem an understanding of *statistical significance* is essential and it should be immediately recognised that this is not necessarily the same as *clinical significance* when the statistics refer to medicine. (...) It is an absolute priority for those using tests



for statistical significance that they understand the conditions which must apply for a particular test to be valid and that they have a clear understanding of the hypotheses which are being tested.

#### 4.4 Absence of evidence, or evidence of absence?



Douglas Altman, Martin Bland. Absence of evidence is not evidence of absence  
<https://www.bmj.com/content/311/7003/485>

The two famous statisticians Doug Altman and Martin Bland in their paper [2] clearly depict our situation: the classical one-sample T-test applied to the Gosset kiln-drying seeds experiment is not able to reveal us the **evidence of absence**, i.e. that the data support the hypothesis that there is no effect (i.e., the two conditions kiln dried and not-kiln dried do not differ); or the **absence of evidence**, i.e. that the data are inconclusive (i.e. we have few data to distinguish the truth). Such a trouble generally affects the 'p-value methodology' in null-hypothesis significance testing. Let us discover why Bayesian approach may help to overcome such *impasse*.

##### 4.4.1 Bayesian One-sample t test



Mark Goss-Sampson. Bayesian Inference in JASP: A Guide for Students  
[http://static.jasp-stats.org/Manuals/Bayesian\\_Guide\\_v0\\_12\\_2\\_1.pdf](http://static.jasp-stats.org/Manuals/Bayesian_Guide_v0_12_2_1.pdf)

Let us now explore the **Bayesian** One Sample T-Test in the JASP menu. Leaving untouched the defaults, on obtain the following table:

Table 4.5: Bayesian One Sample T-Test

	BF <sub>10</sub>	error %
difference	0.885	0.004

*Note.* For all tests, the alternative hypothesis specifies that the population mean differs from 0.

We see that the Bayes Factor is close to 0.89. What can we deduce? We may refer to the table in Figure ???. The left column lists in order the Bayes Factors according to the proposal of the British astronomer and mathematician Harold Jeffreys (the J in JASP!). In his seminal 1946 paper[30] Jeffreys introduces the concept of the **non-informative prior distribution**: in fact, as recalled in Section 3.3.2, Gosset were observing eleven data  $D$  (the Differences) having two different

generative models:  $M_0$ , the normal distribution with  $\mu = 0$ , and  $M_1$  a normal distribution with  $\mu \neq 0$ :

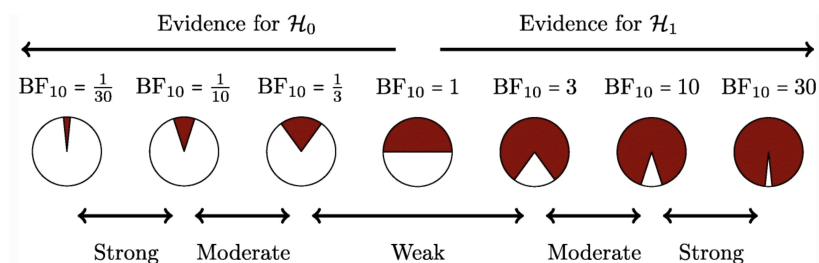
$$BF_{10} = \frac{P(D|M_1)}{P(D|M_0)} = 0.885$$



$BF_{10}$	$\text{Log}_e BF_{10}$	Evidence	In favour of
>100	>4.6	Decisive	Alternative hypothesis
30 to 100	3.4 to 4.6	Very strong	Alternative hypothesis
10 to 30	2.3 to 3.4	Strong	Alternative hypothesis
3 to 10	1.1 to 2.3	Moderate	Alternative hypothesis
1 to 3	0 to 1.1	Anecdotal	Alternative hypothesis
1	0	No evidence	Neither
1 to 0.33	0 to -1.1	Anecdotal	Null Hypothesis
0.33 to 0.1	-1.1 to -2.3	Moderate	Null Hypothesis
0.1 to 0.033	-2.3 to -3.4	Strong	Null Hypothesis
0.033 to 0.01	-3.4 to -4.6	Very strong	Null Hypothesis
<0.01	< -4.6	Decisive	Null Hypothesis

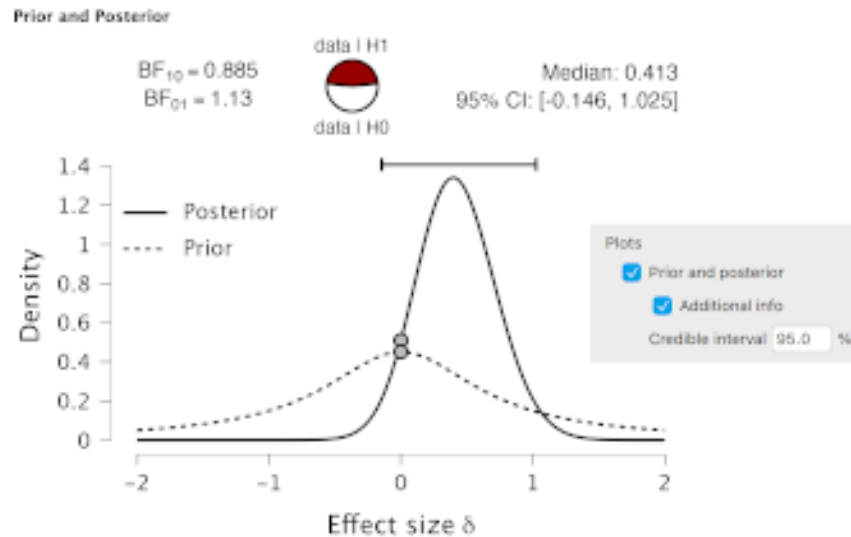
*However, these are merely a simplified heuristic for interpreting Bayes factors, but that the Bayes factor really is a continuous metric of evidence.*

Here above we see the Mark Goss-Sampson[23] JASP table in evaluating Bayes Factor. As  $BF_{10} = 0.885$  is very close to 1 we should claim that Gosset experiment provided **absence of evidence** in favour of the null hypothesis  $\mu = 0$  (i.e. no difference between kiln dried and not-kiln dried seeds) or the alternative hypothesis  $\mu \neq 0$  (i.e. there is a certain difference when kiln drying the seeds). Actually, as  $BF_{10} < 1$ , one can say that it could be an **anecdotal evidence** toward the null hypothesis (i.e. a faint clue toward 'evidence of absence'). The double red / gray / white arrows in the table recalls a useful graphic tool called the **pizza plot**, which use the red tomato and the white mozzarella cheese to enhance the evidence for  $H_1$  versus  $H_0$ .



In particular, selecting the Additional Info in Plots Prior and posterior menu, we immediately see that the pizza plot is nearly half tomato and half mozzarella. The two random distributions are the default prior (which is a Cauchy distribution, [https://en.wikipedia.org/wiki/Cauchy\\_distribution](https://en.wikipedia.org/wiki/Cauchy_distribution)).

In conclusion, how could William Gosset had reported such a finding? Saying that a 2-sided Bayesian one-sample t-test comparing the sample population difference ( $m = 33.7$ ) to the null mean ( $\mu = 0$ ) returns a  $BF_{01}$  of 0.885 suggesting anecdotal evidence in favour of the alternative hypothesis. Equivalently, this means that the data is 1.13 times more likely to have occurred under the null than under the alternative hypothesis.



#### 4.4.2 Bayesian Paired Samples T-Test

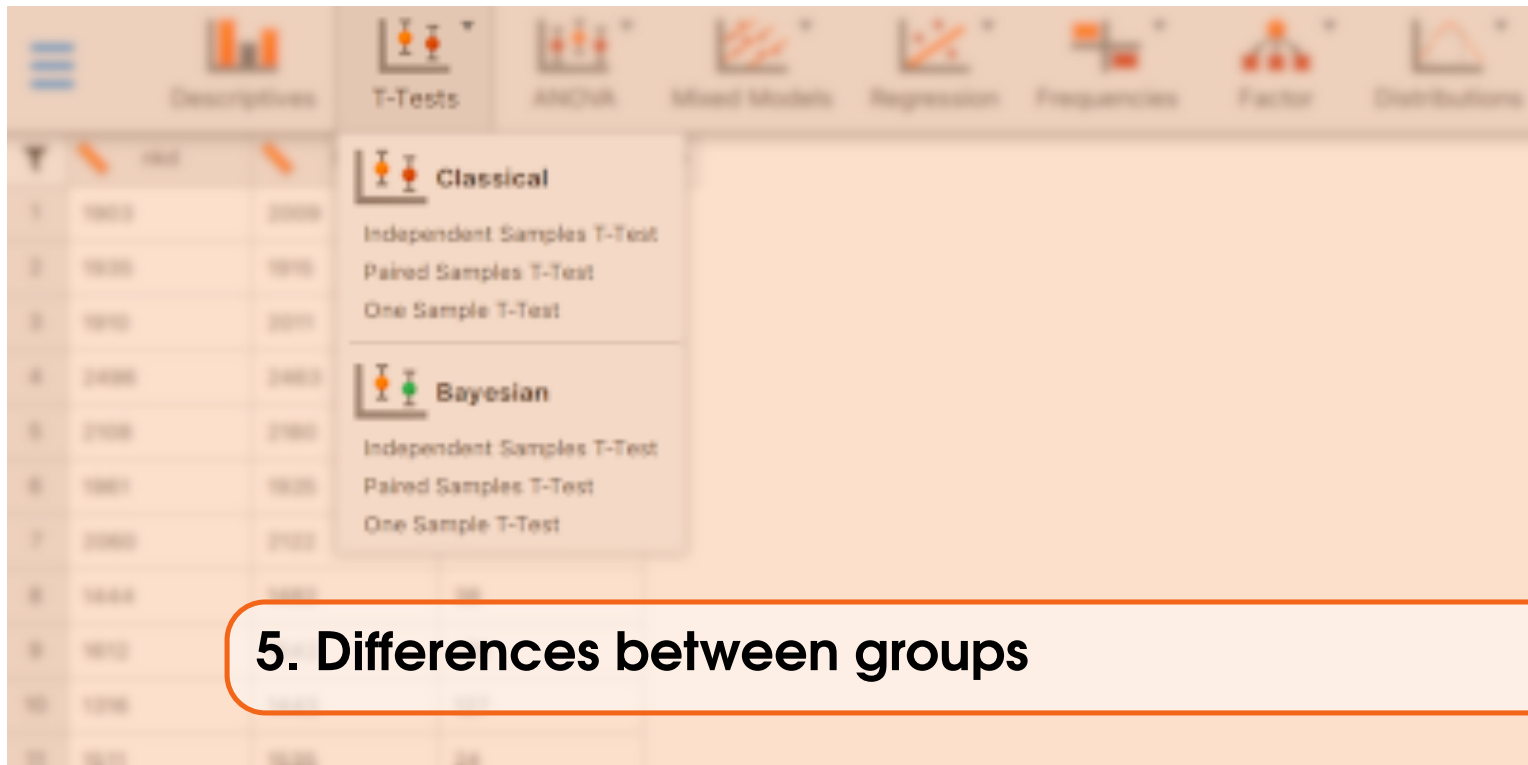
No surprise: we obtain the same conclusion when performing the **Bayesian** Paired Samples T-Test, dragging and dropping *kd* and *nkd* into the Variable Pairs slot:

Table 4.6: Bayesian Paired Samples T-Test

Measure 1	Measure 2	$BF_{10}$	error %
<i>kd</i>	- <i>nkd</i>	0.885	0.004







## 5. Differences between groups

### 5.1 Two groups

We provide here a brief survey of some classical tests concerning two independent samples, adapting the Michael Crawley comprehensive *The R Book* [13, pages 289-298]. We are interested in two main questions:

1. comparing two (unpaired) sample means with normal errors
2. comparing two means with non-normal errors

In the first case, the main tool is again the **Student T-Test** introduced in the previous Chapter. The frequentist approach demands to distinguish two further items:

- comparing two (unpaired) sample means with normal errors and similar dispersion (the proper Student's t test)
- comparing two (unpaired) sample means with normal errors but different dispersion (the so called **Welch test**)

and, to achieve such decision - in the frequentist framework - one has to be able to

- assess normality in data (**Shapiro - Wilk test**)
- compare data dispersion (i.e. the variances, with the **Levene test**)

In the second case, when non-normal errors appears, the straightforward application of the **Wilcoxon - Mann - Whitney test** is recommended. Let us see some example.

#### 5.1.1 The Student T-Test



Elise Whitley, Jonathan Ball. Statistics review 5: Comparison of means  
<https://ccforum.biomedcentral.com/articles/10.1186/cc1548>

We refer again to the ovarian cancer roma dataset. We observed in Section3.2.1 that data appears to be normally distributed. We know that aging is a risk factor for the tumor, so the question

is: do AgePatient differs, in a statistical sense, between the benign and malignant groups, i.e. with respect to Histology? The descriptive analysis shows that the mean age of the 171 women with benign pathology is more or less eleven years younger than the 39 with malignant cancer. But there is a certain dispersion, of more than a dozen of years, measured by the standard deviation: can we say that the mean ages are different in a statistical sense?

We resort to the **Bayesian** Independent Samples T-Test:

Table 5.1: Bayesian Independent Samples T-Test

	BF <sub>10</sub>	error %
AgePatient	652.530	5.868e-9

and, being BF<sub>10</sub> greater than 100 we have a **decisive evidence** in favour of the alternative hypothesis, i.e. that ages are different between the two women groups.

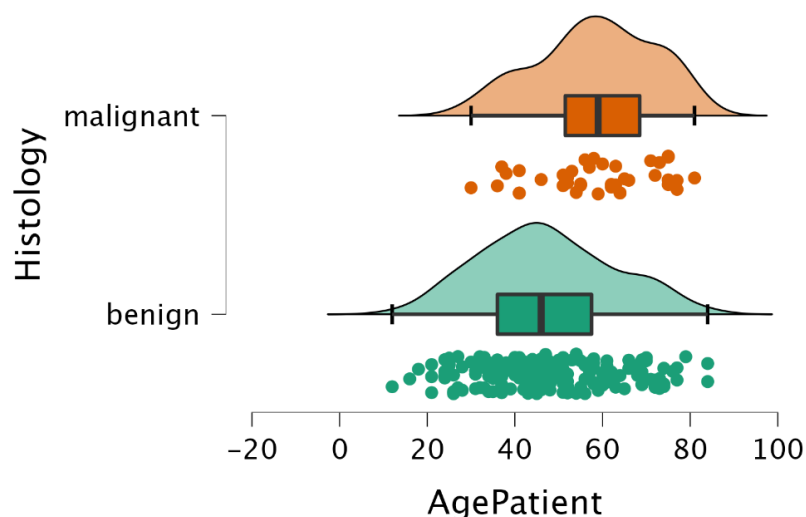
Now, looking to the **Classical** Independent Samples T-Test, we observe that the test statistic  $t$  is much more than 4 deviates away from zero, i.e. the p-value is practically zero: we say that a **very high significant** difference has occurred.

T-Test	t	df	p
AgePatient	-4.282	208	< .001

But we have to verify also two Assumption Checks: normality of errors and homogeneity in error dispersion. Have a look to the Raincloud Plots:

#### Raincloud Plots ▼

AgePatient



To assess if the orange and the green dots are possible outcomes of the gaussian distribution we could try to evaluate the shapes of the orange and green densities, recognizing a bell shape (or examining the symmetry of the boxplots). But this road is skittish, it should be better to depict two QQ-plots in order to visually assess normality. The latter hypothesis, i.e. homogeneity in error dispersion, could be evaluated looking to the boxes in the boxplot: if they have approximately the

same length, good news, we are in presence of **homoskedasticity**, i.e. equal dispersion of errors in terms of variance.

Besides visual inspection, one has also formal test to pursuit: the normality check is usually provided by applying the technique of Samuel Shapiro and Martin Wilk and their **Shapiro-Wilk test**: in this example, being  $p = 0.114$  and  $p = 0.257$  we do not claim evident departure from normality (according to the typical  $\alpha = 0.05$  significance level). The second check involves the **Levene test**, whose significant response leads to an **heteroskedasticity** condition, i.e. different variance of errors. In the present example, a  $p = 0.307$  convinces ourselves that no violation occurs.

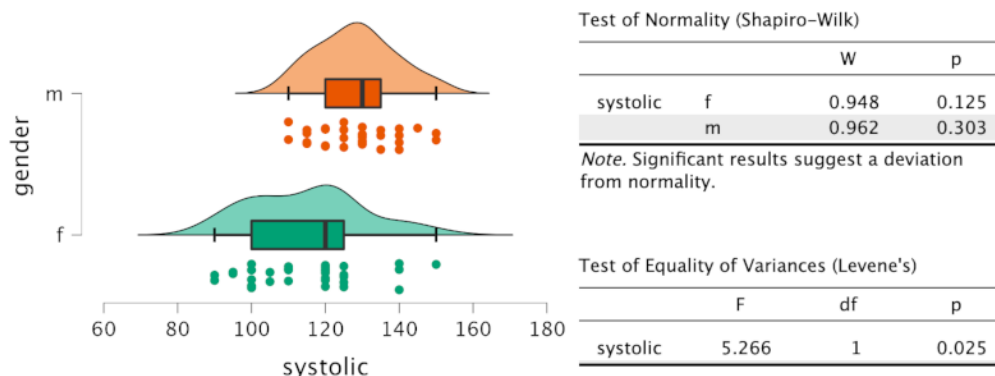
Normality (Shapiro-Wilk)		W	p	
AgePatient	benign	0.987	0.114	
	malignant	0.965	0.257	
Variances (Levene)		F	df	p
AgePatient		1.049	1	0.307

### 5.1.2 The Welch test

Richard Mould [37] recalls in his Table 11.1 that in order to properly apply the t-test, several hypotheses have to be fulfilled:

1. The observations must be independent in order to avoid bias
2. The observations must be drawn from normal populations
3. These normal populations must have the same variance (or in special circumstances, a known ratio of variances)
4. The variables involved must have been measured in an interval scale, so that it is possible to use arithmetical operations (e.g. add, divide, obtain means) on the values of the variables

Despite the fact that in 1969 Bradley Efron [16] has proved that some mild 'orthant symmetry condition' instead of normality and homoskedasticity can be sufficient, have a look to the following situation, concerning the systolic pressure measured on some male and some female students (we will introduce better the dataset in the next Chapter):



As you see, the Shapiro-Wilk test do not suggest violations to normality ( $p = 0.125 > 5\%$ ;  $p = 0.303 > 5\%$ ), but we might have a problem of heteroskedasticity: the Levene's test could have a significant  $p = 0.025 < 5\%$ .

Therefore we can suspect to be in presence of two normal distribution with different dispersions; and if we seek to test two (unpaired) sample means with errors modelled by heteroskedastic normal distributions, the mathematical hypotheses of the ordinary Student T-Test are not fulfilled. Such mathematical questions have been explicit in the famous 'Walter Behrens and Ronald Fisher problem'.



Wikipedia. Behrens - Fisher problem  
[https://en.wikipedia.org/wiki/Behrens%E2%80%93Fisher\\_problem](https://en.wikipedia.org/wiki/Behrens%E2%80%93Fisher_problem)

To overcome the difficulty, JASP implements the Bernard Lewis **Welch test**. It is a two-sample location test used to test the hypothesis that two unpaired populations have equal means, but in a situation in which the two samples have unequal variances and/or unequal sample sizes.

	t	df	p
systolic	-4.110	55.153	< .001

*Note.* Welch's t-test.

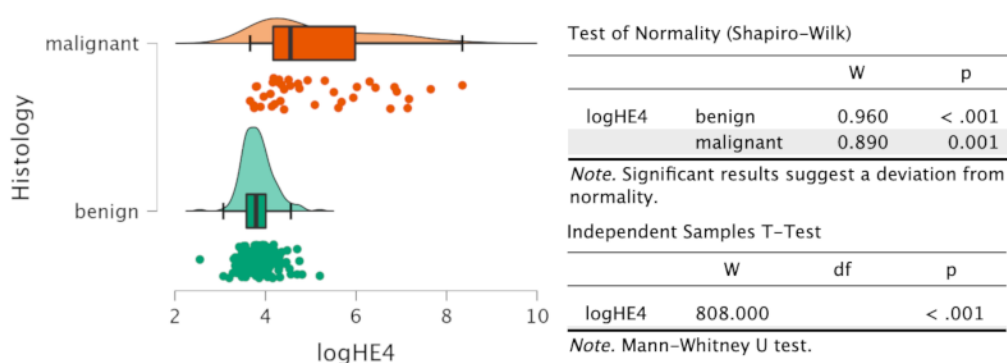
The  $p < .001$  response is a convincent proof to decide for difference in mean systolic pressure between girls and boys. If you notice, the degree of freedom  $df = 55.153$  is not an integer number – this is a consequence of the so called **Welch - Satterthwaite relation**:



Wikipedia. Welch - Satterthwaite equation  
[https://en.wikipedia.org/wiki/Welch%E2%80%93Satterthwaite\\_equation](https://en.wikipedia.org/wiki/Welch%E2%80%93Satterthwaite_equation)

### 5.1.3 The Mann - Whitney test

Suppose now to be interested to confirm the biomarker logHE4 ability in predicting Histology outcome. The orange boxplot exhibits a skewed distribution, with a long whisker, and we are surely doubtful about normality: the Shapiro - Wilk test in both group is very highly significant.



In this case, i.e. testing two (unpaired) sample means with non-normal errors, it is proper to resort to the non-parametric **Wilcoxon - Mann - Whitney U test**, which considers data ordered along their ranks [12]. No doubt, here: a so small p-value  $< .001$  confirms our expectation. We can also approach this issue by means of the **Bayesian** Independent Samples T-Test, obtaining a  $BF_{10}$  greater than one thousand, a decisive evidence in favour of the alternative hypothesis (i.e. logHE4 differs in benign and malignant ovarian lesions).

Table 5.2: Bayesian Mann-Whitney U Test

	BF <sub>10</sub>	W	Rhat
logHE4	4073.742	808.000	1.087

*Note.* Result based on data augmentation algorithm with 5 chains of 1000 iterations.

The output comes from a computational algorithm [15] known as *data augmentation*, which relies on the Markov chain Monte Carlo (MCMC) sampling method.



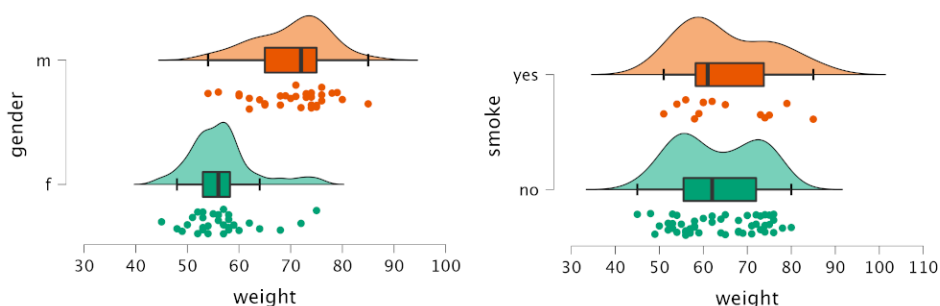
Johnny van Doorn et al. Bayesian rank-based hypothesis testing for the rank sum test, the signed rank test, and Spearman's  $\rho$   
<https://www.tandfonline.com/doi/pdf/10.1080/02664763.2019.1709053>

## 5.2 Three or more groups

We introduce now a new dataset, named *fresher*. It is a cross-section dataset, relative to a cohort of medicine and surgery first year Trieste university students: they were 65, and we collected their gender (a factor variable with *f* and *m* levels), their height, weight and shoesize (numeric variables), along with their smoke habits (a factor with levels *no* and *yes*), and their gym physical activity (classified as a three level alphabetically ordered factor *not* < *occasional* < *sporty*).

**Exercise 5.1** Do weight differ, in a statistical sense, with respect to gender? And, is smoke a predictor of weight? ■

From a bayesian perspective, while the former question has a crystal clear answer with a ludicrously high BF<sub>10</sub>, the latter has an anecdotal or moderate evidence toward the null: we are not so sure, but smoking might not be a reliable predictor of weight at all. If you prefer the classical approach, you will find that both the Student and the Mann - Whitney test are in the first question very highly significant, and in the second question close to one half.



Now we recall that the Welch test is able to detect differences in means between two groups, as its test statistic is defined as:

$$t = \frac{m_1 - m_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Suppose that we have to test, for instance, three groups: how could you modify that statistic? Well, it would be easy to modify the denominator adding a term,  $\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} + \frac{s_3^2}{n_3}}$ . But the numerator

would remain undefined:  $m_1 - m_2 - m_3$ ?  $m_1 + m_2 - m_3$ ?  $m_1 - m_2 + m_3$ ? This simple algebraic observation is the main reason why the T-Test can not be extended to three or more groups.

It is possible to overcome this difficulty observing that when differences in mean are present, also the data dispersions, i.e. the variances, decrease. Have a look:

**Exercise 5.2** Compute weight's variance. Then split weight with respect to gender and to smoke, and compute again the variance. What do we observe? ■

We know that gender is a predictor of weight, and the above Exercise shows that the weight variances of girls and boys are, respectively, 41.1 and 50.7: a great reduction with respect to the 92.1 variance of the whole weight data. On the contrary, splitting the weight into the two groups of smokers ( $\sigma^2 = 106.4$ ) and not smokers ( $\sigma^2 = 89.4$ ) do not reduce the 92.2 variance (actually, in one group there is an increase).

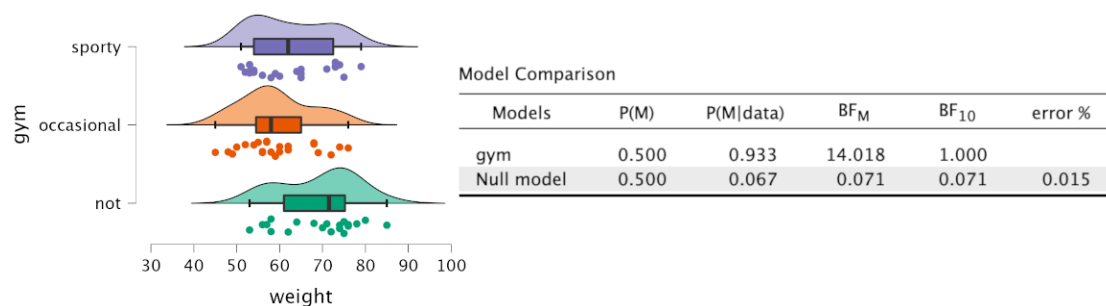
In conclusion, we have discovered the recovery plan: if we want to test differences between means, we have to test reduction in variances! And this is the reason why Anova (= *An.o.va.*, *Analysis of Variance*) has this strange name.

### 5.2.1 The one-way Anova



Viv Bewick et al. Statistics review 9: One-way analysis of variance  
<https://ccforum.biomedcentral.com/articles/10.1186/cc2836>

The **one-way Anova** analysis in JASP can be performed into bayesian or into classical frequentis approach. As an example, we consider as a Fixed Factor the gym physical activity (ordered according the three levels not < occasional < sporty and as Dependent Variable the weight:



The **Bayesian ANOVA** can be interpreted reading the  $BF_M = 14.02$ , which provides a **strong evidence** in favour of the alternative hypothesis: some of the groups is different in mean from some of the other. A  $BF_M = 14.02$  implies that the data have increased the prior model odds of more than ten times. We can also examine the **Classical ANOVA**, yielding a highly significant  $p = 0.003$ :

Table 5.3: ANOVA - weight

Cases	Sum of Squares	df	Mean Square	F	p
gym	1020.400	2	510.200	6.488	0.003
Residuals	4875.816	62	78.642		

But, simply looking to the purple sporty distribution, we get the impression not to be in presence of a gaussian distribution, which is mathematically required as correctly stated by Vijay Rohatgi [42]:

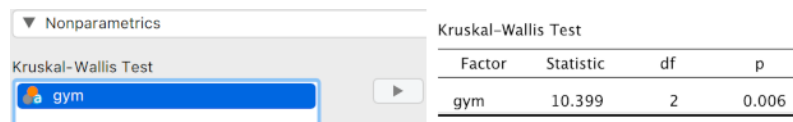
Let  $X_{11}, X_{12}, \dots, X_{1n_1}, X_{21}, X_{22}, \dots, X_{2n_2}$  and  $X_{31}, X_{32}, \dots, X_{3n_3}$  be independent random samples from three normal populations with respective parameters  $\mu_1$  and  $\sigma_1^2$ ,  $\mu_2$  and  $\sigma_2^2$  and  $\mu_3$  and  $\sigma_3^2$ . Suppose  $\sigma_1 = \sigma_2 = \sigma_3$ . ...

**Exercise 5.3** Go to the Descriptive menu and make a Q-Q plot of weight splitted on gym. What do you think about normality? ■

Therefore if one wants to perform an Anova according the traditional way, it is required to check whether in weight:

1. all three groups not, occasional, sporty are normally distributed
2. their dispersions are homoskedastic, i.e. in statistical sense  $\sigma_1 = \sigma_2 = \sigma_3$ .

So, in this particular case, it is proper to move away from parametric approach resorting nonparametric methods, i.e. the William Kruskal and Wilson Wallis **Kruskal - Wallis test**:



Kruskal-Wallis Test			
Factor	Statistic	df	p
gym	10.399	2	0.006



Wikipedia. Kruskal - Wallis one - way analysis of variance  
[https://en.wikipedia.org/wiki/Kruskal%E2%80%93Wallis\\_one-way\\_analysis\\_of\\_variance](https://en.wikipedia.org/wiki/Kruskal%E2%80%93Wallis_one-way_analysis_of_variance)

### 5.2.2 The multiple comparison issue

We saw that in weight versus gym, the Anova p-value is significant. But such a p-value do not disclose which group is different from the other, and many possibilities are plausible, and we are required to choose one of them:

- not = occasional  $\neq$  sporty
- not  $\neq$  occasional = sporty
- not = sporty  $\neq$  occasional
- not  $\neq$  occasional  $\neq$  sporty  $\neq$  not

Richard Mould's words in his chapter 17.1 [37] are clear:

With more than two means it is of course technically possible to make multiple t-tests on all possible pairs of means, but *making multiple tests increases the probability of making a type I error*.

In fact, suppose to choose an  $\alpha$  level of 5%; then, the probability to commit an error of the first type is about the 14% (independent events, product of probabilities):

$$1 - \left(1 - \frac{5}{100}\right) \cdot \left(1 - \frac{5}{100}\right) \cdot \left(1 - \frac{5}{100}\right) = 1 - \left(1 - \frac{5}{100}\right)^3 = 0.143$$

One 'radical' solution is to exploit the Bernoulli inequality  $1 + nh < (1 + h)^n$ , i.e. if we have  $n = 3$  groups and therefore  $n \cdot (n - 1)/2 = 3$  comparisons, then one fix  $h = \alpha/3$ , i.e.  $\alpha = 0.05/3 = 0.017$  instead of the common choice  $\alpha = 0.05$ . This is the famous **Carlo Bonferroni correction**[41].





Wikipedia. Bonferroni correction

[https://en.wikipedia.org/wiki/Bonferroni\\_correction](https://en.wikipedia.org/wiki/Bonferroni_correction)

One milder and elegant approach is to trust in John Tukey and adopt his **Honest Significant Differences multiple comparison test**[13]:

		Mean Difference	SE	t	$P_{tukey}$	$P_{bonf}$
not	occasional	9.665	2.711	3.565	0.002	0.002
	sporty	6.373	2.740	2.326	0.060	0.070
occasional	sporty	-3.292	2.645	-1.245	0.432	0.653

From the table, we are convinced that `occasional` and `sporty` has not different means, in a statistical sense. Therefore we are led to decide that:

- `not`  $\neq$  `occasional` = `sporty`

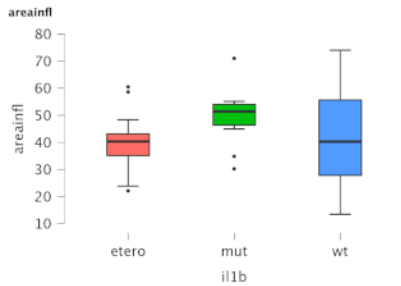
Nevertheless, observe this strange fact: a p-value = 0.060 (Tukey) or 0.070 (Bonferroni) could wrongly suggest that `not` = `sporty`. As you see, everything could appear to be shaky and slippery, if you forget that *'absence of evidence is not evidence of absence'*.

### 5.2.3 How to mend heteroskedasticity

When you try to perform an ANOVA with JASP (and with most of all others statistical softwares) in a heteroskedastic situation, the things can be really bad. Consider as an example the `tooth` dataset, in which sixty-nine patients have been observed, measuring their gingival area inflammation and considering their gender and their different attitude toward smoke (yes or no). The main goal was to discover a statistical relation with a particular cytokine mediating inflammatory response named Interleukin-1 beta ( $IL-1\beta$ ), `il1b`, expressed in three levels: mutated, heterozygotes or wild-type. We see from the boxplots that red hetero patients has, on average, a lower inflammation area than the green mut patients; the observation is confirmed by the descriptive statistics. But performing an ANOVA, the software detects only a faint anecdotal evidence  $BF_{10} = 2.0$  of effect, and not significant p values. This contradiction between descriptive and inferential result is `wt`'s fault: the blue boxplot has a dispersion that is approximately the double of the other two groups.

Unfortunately, JASP has not any valid tool to manage the impasse. If we move to R language we have two effective strategy: the first, to continue within the ANOVA framework and resort to the `sandwich` [54] and `multcomp` [25] packages, as magistrally explained in the *Multiple comparisons using R* book written by Bretz, Hothorn and Westfall [9]. In that case, one can discover that hetero

versus mut has p-value = 0.024. The second approach is even simpler and involves the well-known concept of information **entropy** applied to the so called linear model: we discuss it in the next Chapter.



Descriptive Statistics

	areainfl		
	etero	mut	wt
Valid	24	10	35
Missing	0	0	0
Mean	39.254	49.673	41.405
Std. Deviation	9.408	11.277	17.148

Model Comparison ▼

Models	P(M)	P(M data)	BF <sub>M</sub>	BF <sub>10</sub>	error %
Null model	0.500	0.667	2.006	1.000	
il1b	0.500	0.333	0.499	0.499	0.029

ANOVA - areainfl

Cases	Sum of Squares	df	Mean Square	F	p
il1b	780.635	2	390.318	1.955	0.150
Residuals	13178.127	66	199.669		

Note. Type III Sum of Squares

Kruskal-Wallis Test

Factor	Statistic	df	p
il1b	4.517	2	0.105



## 6. Regression

### 6.1 Overview



Viv Bewick et al. Statistics review 7: Correlation and regression  
<https://ccforum.biomedcentral.com/articles/10.1186/cc2401>

In the previous Chapter we were interested in assessing differences in the **numeric** (or **scale** in JASP language) **weight** variable with respect to the **nominal** **gender** factor within our **fresher** students dataset, resorting the JASP T-Test menu; and, when referring to the three level **gym** factor, we addressed the ANOVA menu. In this Chapter we introduce a modern and powerful statistical tool widely used in the cross-sectional studies: the **linear model**.



Francis Galton. Regression towards Mediocrity in Hereditary Stature  
<https://galton.org/essays/1880-1889/galton-1886-jaigi-regression-stature.pdf>

Typically, in the statistical textbooks, this argument is introduced talking about the sir Francis Galton **regression** 'towards mediocrity' **line** 'in hereditary stature' [22], and at a first sight the two arguments perfectly overlap. We are going here to show that the linear model encompasses a variety of important and classical statistical tools, usually named **Ancova** methods – and we are going to show that the **t-test** or the **Anova** we have just learnt are trivial consequences of this method.

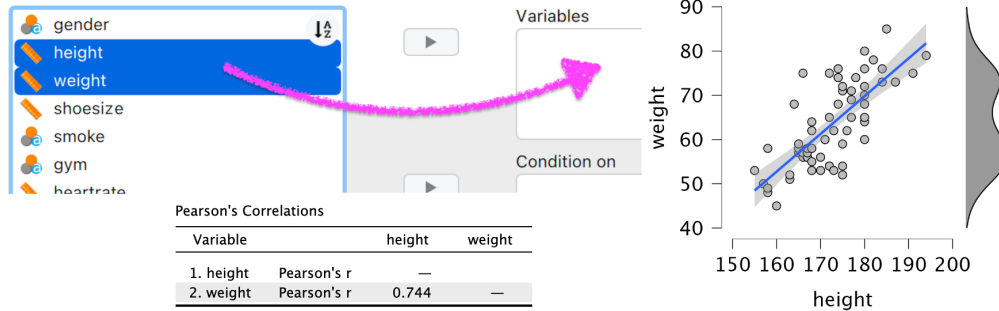
### 6.2 The regression line

Suppose we are interested in assessing the possible relation that interlaces **fresher's weights** with their heights. It is a relation between two numeric variables, and we stress the role that height assumes as a possible **predictor** of (i.e. a dataset covariate significantly associated to) the **weight**. In this sense, using the symbolic **Wilkinson and Rogers notation** we pose the following relation:

$$\text{weight} \sim \text{height}$$

This position implies that `height` represents the input, the independent variable located on the abscissa  $x$ , while `weight` is thought to be the output, the dependent variable located on the ordinate  $y$ .

### ▼ Correlation



A famous way to 'quantify the linear relationship' between two variables is the so called Karl Pearson's correlation  $1 \leq \rho \equiv 0.744 \leq 1$ . There are many straightforward or clever way to explain its definition, but we will provide a simple explanation: suppose that the blue line in the figure has equation  $y = a + b \cdot x$ , with  $b$  the slope. The dimensional analysis lead us to deduce that:

$$[b] = \frac{[\Delta \text{weight}]}{[\Delta \text{height}]} = \frac{[Kg]}{[m]}$$

but reasonable proxy of  $\Delta \text{weight}$  and  $\Delta \text{height}$  are, respectively, the standard deviations  $\sigma_{\text{weight}}$  and  $\sigma_{\text{height}}$ . So, no surprise, the quantity:

$$[b] \cdot \frac{[\sigma_{\text{height}}]}{[\sigma_{\text{weight}}]}$$

is dimensionless. And, ta-dah:

$$\rho = b \cdot \frac{\sigma_x}{\sigma_y} ; b = \rho \cdot \frac{\sigma_y}{\sigma_x}$$

Before proceeding, we always remember that a statistical relation **is not** a cause-effect relation at all. Just for fun, look to the <http://www.tylervigen.com/spurious-correlations> in which for instance the divorce rate in Maine is put in relation with consumption of margarine.



Tyler Vigen. Spurious Correlations  
<https://galton.org/essays/1880-1889/galton-1886-jaigi-regression-stature.pdf>

More seriously, remember that:

The objective (..) is to show that a relationship exists between these two variables, so that having demonstrated the existence of this relationship, it can be used within some theoretical framework. Blind use of regression formulae, just because they exist, can be very misleading. If  $Y =$  a cause and  $X =$  an effect, one must be careful not to draw too many conclusions if there may be several other possible causes. Cause and effect in medicine are seldom so simple as to be explained by a single straight line. (R. Mould [37], section 16.1)

When looking for a regression line  $y = a + b \cdot x$  we need to precise how to choose the intercept  $a$  and the slope  $b$ , in a way that the line crosses the point cloud in the 'best possible way'. This can always be achieved as demonstrated in the **Gauss - Markov theorem** (e.g. [18, page 18]): the regression line is the Best Linear Unbiased Estimate ('BLUE') according to the Ordinary Least Square (OLS) estimation, a method explored since 1755 by the dalmatian Ruggero Boscovich / Ruđer Bošković [46]. Simply, one consider all the **residuals** (one of them is the violet segment in the above figure) and, likewise in the Pythagorean theorem, one consider the sum of the squared residuals (i.e the sum of the squared 'vertical' distances from each cloud point  $(x_i, y_i)$  and its vertical projection of the line, i.e. the point  $(x_i, a + b \cdot x_i)$ ). In other words, the residuals are defined as  $\varepsilon_i = y_i - (a + b \cdot x_i)$  and defining the vector  $\varepsilon = (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_i, \dots, \varepsilon_n)$  one computes the scalar product  $\varepsilon^T \cdot \varepsilon \equiv \langle \varepsilon | \varepsilon \rangle$  and search the parameters  $a$  and  $b$  which minimize such scalar product.



Daniel Kunin et al. Seeing Theory.  
<https://seeing-theory.brown.edu>

Let us start the analysis with the **Bayesian** Linear Regression:

Dependent Variable		Model Comparison - weight					
Covariates		Models	P(M)	P(M data)	BF <sub>M</sub>	BF <sub>10</sub>	R <sup>2</sup>
weight	height	height	0.500	1.000	4.641e+9	1.000	0.554
		Null model	0.500	2.155e-10	2.155e-10	2.155e-10	0.000

We observe that the regression line, or better, the **linear model** we are considering possesses a **decisive evidence** against the **null hypothesis**, to be precised in a while. We also read the **determination coefficient**  $R^2 = 0.554$  value, remembering that this is the squared value of Pearson's  $\rho = 0.744$ .

If we move to the **Classical** Linear Regression menu, we focus the attention on the Coefficients table:

Model		Unstandardized	Standard Error	Standardized	t	p
H <sub>0</sub>	(Intercept)	63.523	1.191		53.357	< .001
H <sub>1</sub>	(Intercept)	-83.891	16.677		-5.030	< .001
	height	0.854	0.096	0.744	8.850	< .001

The table allow us to discover the coefficients of the regression line  $y = a + b \cdot x$ , i.e.  $a = -83.891$  and  $b = 0.854$ ; and allow us to precise the meaning of the null model  $H_0$ , which is that particular 'flat' horizontal line having  $b = 0$  slope and  $a = \text{mean}(\text{weight}) = 63.523$  intercept. We are very highly confident that all these three numbers are different from zero: in fact  $p < .001$  for each of them. The pivotal role, anyway, is played by the  $p < .001$  of the `height` term, which is considered 'the p of the model'.

After having discovered  $a$  and  $b$ , which represent the **fixed effects** of the linear model, we have to describe also the **stochastic component** of the linear model, or **random effect**. The theory requires in fact that residuals  $\varepsilon_i$  have to be independent and normally distributed, with zero mean, and with a constant standard deviation  $\sigma$  (remember section 5.1.3). To estimate that  $\sigma = 6.46$ , we can type:

```
> summary(model)$sigma
```