

# **Biostatistical Methods with JASP**

**UMG School of PhD Programmes Life Sciences and Technologies**

**Massimo Borelli**

Copyright © 2022 Massimo Borelli, Ph.D.

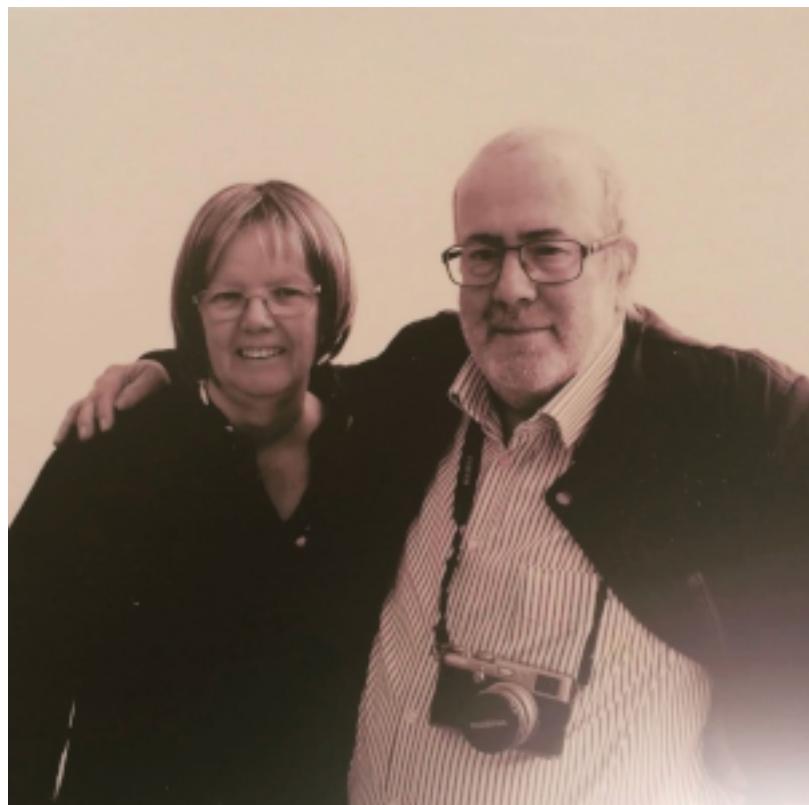
UNIVERSITY *Magna Græcia* SCHOOL OF PHD PROGRAMMES IN LIFE SCIENCES AND TECHNOLOGIES, CATANZARO, ITALY

<https://www.phdprogramme-scuoladottorati-umg.org>  
<https://github.com/MassimoBorelli/Germaneto>

Licensed under the Creative Commons Attribution-NonCommercial 3.0 Unported License (the “License”). You may not use this file except in compliance with the License. You may obtain a copy of the License at <http://creativecommons.org/licenses/by-nc/3.0>. Unless required by applicable law or agreed to in writing, software distributed under the License is distributed on an “AS IS” BASIS, WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied. See the License for the specific language governing permissions and limitations under the License.

To type this book we exploited, with deep thankfulness, the Mathias Legrand L<sup>A</sup>T<sub>E</sub>X template freely available at <https://www.latextemplates.com/cat/books>.

*First edition, July 2022*



*in loving memory*  
professors Heather Bond and Gianni Morrone



The 2022 Biostatistics 'A' Course in Germaneto

# Contents

<b>1</b>	<b>To start</b>	<b>9</b>
1.1	What are we talking about	9
1.2	Which is the 'best' software for medical statistics?	9
1.2.1	The R language	10
1.2.2	The R language user graphical interfaces	11
1.2.3	JASP	11
1.3	Activities	12
<b>2</b>	<b>Data Presentation</b>	<b>15</b>
2.1	Background	16
2.2	Descriptive Statistics in JASP	16
2.2.1	Numerical summaries	17
2.2.2	A picture is worth a thousand words	20
2.3	Which are 'the best' Descriptives?	22
2.4	Dealing with repeated measurements	24
2.5	Exercises	24
<b>3</b>	<b>Probability in medicine</b>	<b>27</b>
3.1	Brief recalls on random variables	27
3.2	Commonly used random variables	28
3.2.1	The Normal Distribution	28
3.2.2	The Lognormal Distribution	31
3.2.3	The Binomial Distribution	32
3.2.4	The Poisson Distribution	33

<b>3.3</b>	<b>Exercises</b>	<b>34</b>
<b>3.4</b>	<b>Evaluating odds and risks: Bayes theorem</b>	<b>35</b>
3.4.1	Bayes theorem . . . . .	36
3.4.2	The Bayes factor . . . . .	37
<b>3.5</b>	<b>Sample and population: approaching inference</b>	<b>38</b>
<b>3.6</b>	<b>Mismatching variability with reliability</b>	<b>39</b>
<b>3.7</b>	<b>Exercises</b>	<b>43</b>
<b>4</b>	<b>T-Test: the history of biostatistics</b> . . . . .	<b>45</b>
<b>4.1</b>	<b>Detecting a signal from noise</b>	<b>45</b>
4.1.1	Classical One-sample t test . . . . .	46
4.1.2	Classical Two-sample paired t test . . . . .	48
<b>4.2</b>	<b>Ronald Fisher's idea on significance level</b>	<b>48</b>
<b>4.3</b>	<b>Out of the frying pan into the fire: statistical or clinical significance?</b>	<b>50</b>
<b>4.4</b>	<b>Absence of evidence, or evidence of absence?</b>	<b>51</b>
4.4.1	Bayesian One-sample t test . . . . .	51
4.4.2	Bayesian Paired Samples T-Test . . . . .	54
<b>4.5</b>	<b>In conclusion</b>	<b>54</b>
<b>4.6</b>	<b>Exercises</b>	<b>54</b>
<b>5</b>	<b>Differences between groups</b> . . . . .	<b>55</b>
<b>5.1</b>	<b>Two groups</b>	<b>55</b>
5.1.1	The Student T-Test . . . . .	55
5.1.2	The Welch test . . . . .	57
5.1.3	The Mann - Whitney test . . . . .	58
5.1.4	In conclusion . . . . .	59
<b>5.2</b>	<b>Three or more groups</b>	<b>60</b>
5.2.1	The one-way Anova . . . . .	60
5.2.2	The multiple comparison issue . . . . .	62
5.2.3	How to mend heteroskedasticity . . . . .	63
<b>5.3</b>	<b>Exercises</b>	<b>63</b>
<b>6</b>	<b>Introducing the linear model</b> . . . . .	<b>65</b>
<b>6.1</b>	<b>Overview</b>	<b>65</b>
<b>6.2</b>	<b>The regression line</b>	<b>65</b>
6.2.1	Measuring point cloud disorder . . . . .	66
6.2.2	Ordinary least square fitting . . . . .	67
6.2.3	Toward linear modelling . . . . .	68
6.2.4	Understanding random effect . . . . .	68
<b>6.3</b>	<b>The diagnostic of a linear model</b>	<b>69</b>
<b>7</b>	<b>Introducing multivariable analysis</b> . . . . .	<b>71</b>
<b>7.1</b>	<b>Overview</b>	<b>71</b>
<b>7.2</b>	<b>The Wilkinson and Rogers notation</b>	<b>72</b>

<b>7.3</b>	<b>Ancova</b>	<b>73</b>
7.3.1	Ancova without interaction .....	73
7.3.2	Ancova with interaction .....	74
<b>7.4</b>	<b>The Model Selection</b>	<b>75</b>
7.4.1	The Akaike Information Criterion .....	76
7.4.2	Multiple comparison and AIC .....	76
<b>7.5</b>	<b>Exercises</b>	<b>78</b>
<b>8</b>	<b>The logistic regression</b>	<b>79</b>
<b>8.1</b>	<b>The generalized linear model</b>	<b>79</b>
<b>8.2</b>	<b>The logistic regression</b>	<b>80</b>
8.2.1	Analyzing the <code>roma</code> dataset .....	80
<b>9</b>	<b>Conclusions</b>	<b>83</b>
	<b>Bibliography</b>	<b>85</b>
	<b>Articles</b>	<b>85</b>
	<b>Books</b>	<b>87</b>
	<b>Index</b>	<b>89</b>





# 1. To start

## 1.1 What are we talking about

This book arranges the lectures of a short course in Medical Statistics held during the 2022 first semester of the Master in Medical Physics by the *Abdus Salam* International Center of Theoretical Physics in Miramare, Trieste, Italy with the on-line lessons provided to the U.M.G. Ph.D. student during the 2020 - 2021 pandemic period and a previous book entitled *Metodi di Biostatistica*, based on the R language and still published at <http://www.biostatisticaumg.it/biostatistica/main.pdf>.

The first chapter presents a number of softwares suitable to statistical analyses, focusing on the free open source JASP solution: its user-friendliness is presented in Chapter 2 discussing possible approaches to the descriptive statistics. In Chapter 3 a quick review on stochastic variables frequently used in biostatistics is presented, and the fundamental concept of the so called bayesian inference approach is recalled: the Bayes factor. On the contrary, the cornerstone of the frequentist inferential approach, and the standard error of the mean, is presented in Chapter 4, which is devoted to introduce the renowned T-Test. In Chapter 5 we go into further details on assessing differences between various groups introducing the 'Anova' methods. Chapter 6 deals with another historically important tool, the regression line, as a unifying introduction to the so called linear model. The multivariate analysis is discussed in Chapter 7, and the concept of the selection model is discussed. Lastly, the generalized linear model framework is presented in the emblematic case of the logistic regression in Chapter 8.

Copies of the present book and of the datasets can publicly be downloaded at <https://github.com/MassimoBorelli/Germaneto>.

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

Well, it depends. If you are required to perform 'heavy' computations, a programming language as Python will be needed - in this case, do not hesitate to contact professor Paolo Zaffino, at the Biomechatronics Lab, <http://dmse.unicz.it/personale/docente/paolozaffino>. The language R is more 'statistically' oriented, but require a certain learning curve. Often, a simpler

'statistical suite' like R Commander, Jamovi or JASP could be preferred. Let us spend few words to introduce the latters.

### 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://rdrr.io/snippets/>.



If you are interested in some historical details, Nick Thieme has published an article[50] 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 Grolemund textbook [58] *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 the R Studio <https://www.rstudio.com/> Integrated Development Environment (IDE).

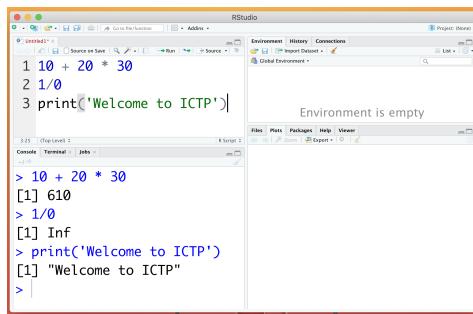


Figure 1.1: R Studio is preferred by many researchers and data analysts, ensuring a stable and well integrated programming and graphing environment. A fatal drawback is its 'steep learning curve': the newcomer has to practice quite a lot of time in managing syntaxes and commands – besides the effort 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 ~ 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 reason why in our short course, alas!, we skip the effort to learn it. But one possible recovery plan it exist: to adopt a G.U.I., a graphical user interface.

### 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/>.

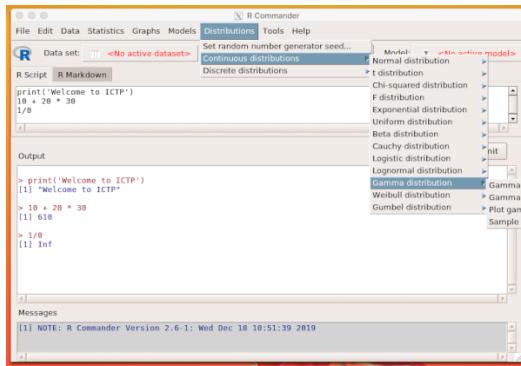


Figure 1.2: The R Commander appearance: you see a menu environment with an input section, named R Script, which has been created by the File | New Script procedure; and an Output section which lists the input commands and produces the outputs. Below, the gray backgrounded section provides Messages alerting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

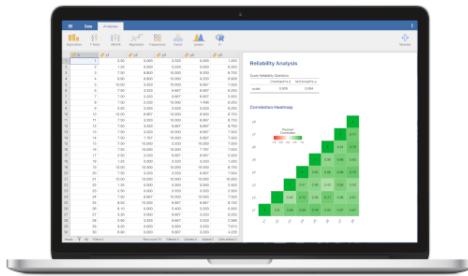


Figure 1.3: The Jamovi environment, which integrates the spreadsheet capabilities in managing raw data and a menu of typical R analysis commands. The user manual <https://www.jamovi.org/user-manual.html> helps the beginner to learn the basic procedures.

### 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’, the latter being 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/>

//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.

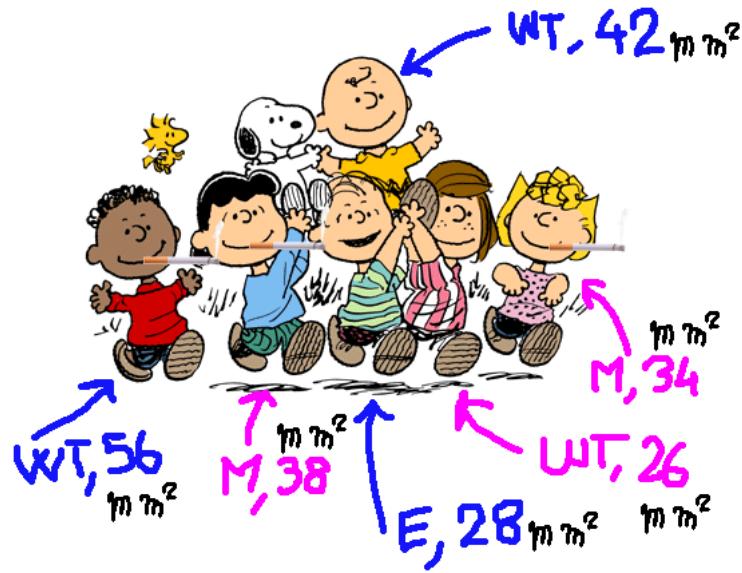


Figure 1.4: The official JASP web page, <https://jasp-stats.org/>

### 1.3 Activities

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	Inesah	Garcia	22	10	1970	

- **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. But, there, privacy is an important issue – but very often biostatisticians 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/Germaneto>), 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. ■



■ **Activity 1.2 — collecting data in a spreadsheet.** Let us consider three males (Charlie Brown, Linus van Pelt, Franklin) e three females (Lucy van Pelt, Sally Brown, Piperita Patty) suffering of gingival inflammation, measured by an image (numeric evaluation, expressed in square millimeters). Some of them are smokers, a known risk factor. Moreover they can be classified according to a genetic criterion into three groups: wild-type, mutated, heterozygotes. How do you think you collect these data into a spreadsheet? ■

**Discussion 1.3.1 — managing data collection in a multicenter experiment.** Have you got any experience about collecting data in a multicenter experiment? Can you report (or may you imagine) any possible issue in such situation? Discuss viable strategy to overcome such problems.



## 2. Data Presentation

**Problem 2.1** dear Massimo, we started to write our first paper, and we were required to prepare the notorious 'Table 1', in order to describe our data, but we are not sure about what to do ...

Let's start reviewing some examples of true papers published by our former UMG Ph.D. Students:

1. Roberta Venturella [53] in her paper *3 to 5 years later: long-term effects of prophylactic bilateral salpingectomy on ovarian function* compares the ovarian condition OvAge™ with her patients' ages, adopting the expression ' Mean Values  $\pm$  SD'. For instance, at the moment of the surgical intervention, patients ages were  $45.85 \pm 2.40$
2. Emanuela Chiarella [12] clarifies the Zinc Finger protein biological role in her paper *ZNF521 Represses Osteoblastic Differentiation in Human Adipose-Derived Stem Cells*, and to describe the experiments she uses the expression Means + SD.
3. Reading the Maria Teresa De Angelis *Short-term retinoic acid treatment sustains pluripotency and suppresses differentiation of human induced pluripotent stem cells* paper [17], we see that the adopted convention is mean  $\pm$  standard error of the mean (SEM).
4. Maria Vittoria Caruso [11] is interested in computational model in her paper *Influence of IABP-Induced Abdominal Occlusions on Aortic Hemodynamics: A Patient-Specific Computational Evaluation*, quoting the existing literature according to which *Furthermore, they also discovered that the distance between LSA and CT was  $241 \pm 23\text{mm}$* , taking for granted that readers are able to understand the meaning of those figures.
5. In order to describe the competing role of LDH1 and LDH3 biomarkers into uterine sarcomas, Annalisa Di Cello [18] in *A more accurate method to interpret lactate dehydrogenase isoenzymes' results in patients with uterine masses* uses the notation Mean [Median].
6. Last but non least, Paolo Zaffino [59], *Radiotherapy of hodgkin and non-hodgkin lymphoma: a nonrigid image-based registration method for automatic localization of prechemotherapy gross tumor volume*, in his tomographic image analyses describes certain similarity coefficients by medians and quartiles.

Who of them did the right choice?

## 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.

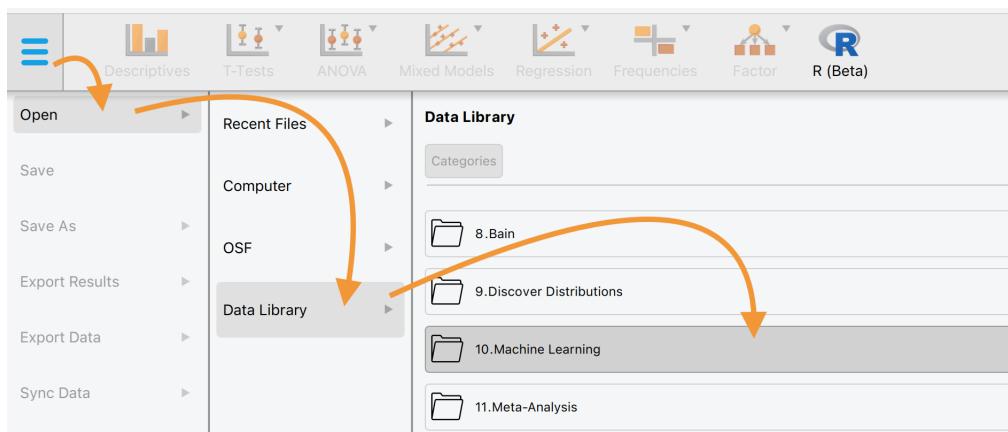


Alla Katsnelson. Colour me better: fixing figures for colour blindness  
<https://www.nature.com/articles/d41586-021-02696-z>

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 in 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, ...)

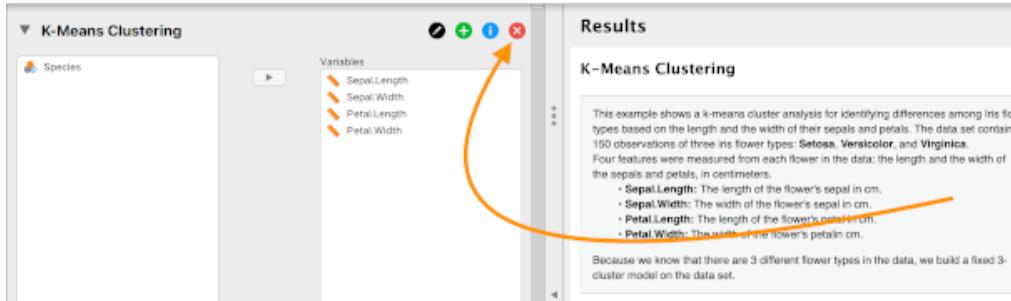
## 2.2 Descriptive Statistics in JASP

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 [23] 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:

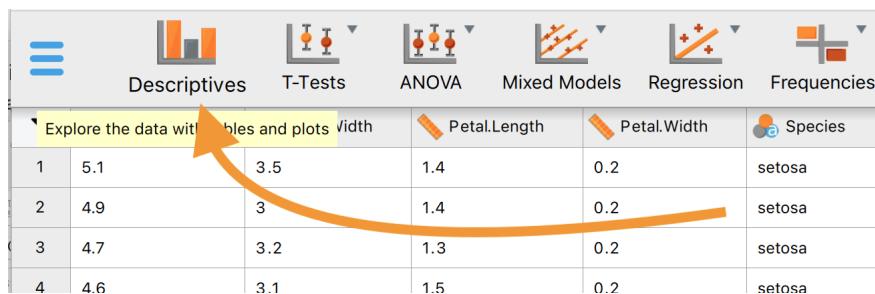


We are not interested now in discussing what is K-Means Clustering; but looking to the **Results** section on the right, we can read a description of the dataset, composed by 150 rows and 5 columns, named respectively Sepal.Length, Sepal.Width, Petal.Length, Petal.Width and Species. The first four columns 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:



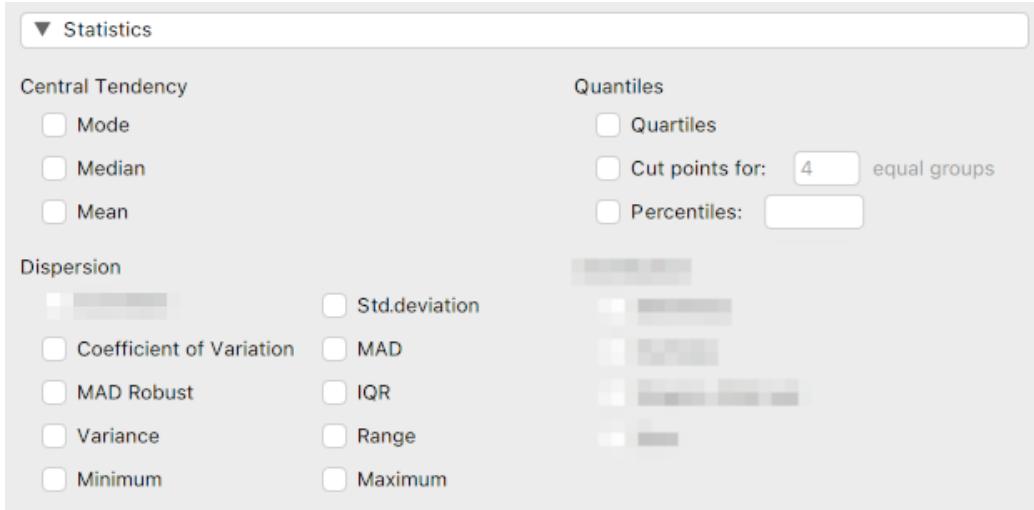
### 2.2.1 Numerical summaries

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*.



**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>

We need to precise a detail, which will be useful in the sequel. If you try to compute the standard deviation with a spreadsheet, for instance, with the italian version of Open Office Calc™, you are required to choose two possibilities: the **sample** standard deviation (DEV.ST) or the **population** one (DEV.ST.POP):

Categoria	DEV.ST.POP
Statistica	DEV.ST.POP( Numero 1; Numero 2; ... )
Funzione	
DEV.ST	
DEV.ST.POP	Calcola la deviazione standard sulla base dell'intera popolazione.
DEV ST POP VALORI	

We will clarify the terminology in Section 3.5, but for now we carefully look to their different algebraic formulas:

$$\text{population: } \sqrt{\frac{\sum_i^n (x_i - m)^2}{n}}, \text{ sample: } \sqrt{\frac{\sum_i^n (x_i - m)^2}{n-1}}$$

We may explain why in the latter case we divide by  $n - 1$ , without touching what textbooks (as e.g. [44]) explain by introducing the concepts of *unbiased estimators*. Suppose that Massimino goes to high school, and receive  $x$  and  $y$  in the first and second latin test. Can his mother deduce  $x$  and  $y$  knowing that the average is 6? No, of course, she lacks a piece of information. But, knowing the average 6 and the standard deviation 1.41, she could detect them:

$x = ? \quad y = ? \quad \sqrt{\frac{(x-6)^2 + (12-x-6)^2}{2-1}} = 1.41 \quad \sqrt{2(x-6)^2} = 1.41 \quad x = 1+6 = 7$

$\frac{x+y}{2} = 6 \quad y = 12 - x \quad \sqrt{2 \cdot (x-6)^2} = 1.41 \quad x-6 = 1$

$y = 12 - x \quad \sqrt{(x-6)^2 + (6-x)^2} = 1.41 \quad x-6 = 1$

So, if we know the mean  $m$ , we have 1 piece of information, and to disclose the  $n = 2$  marks we need  $(n - 1) = 2 - 1 = 1$  other information, or as they say, we have  $(n - 1)$  **degrees of freedom**. Therefore, when we have a sample  $x_1, x_2, \dots, x_n$  we convey into the sample mean  $m$  the first information on that data, and we have  $(n - 1)$  degrees of freedom left.

**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.

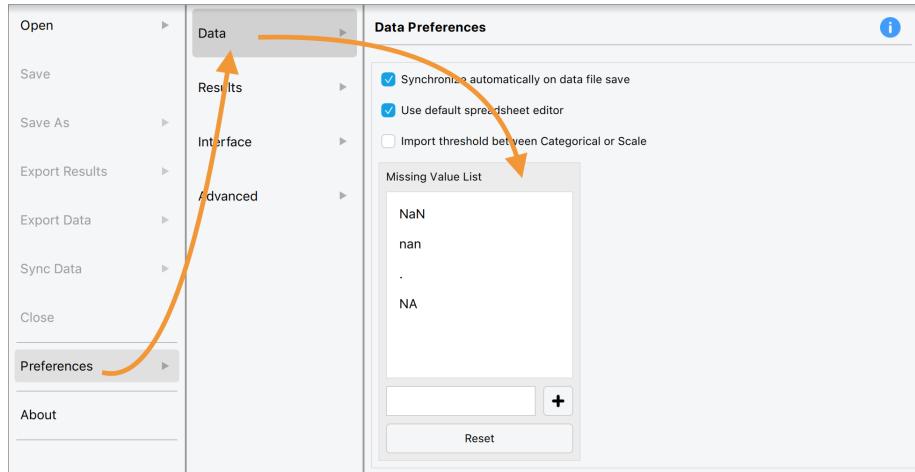
**Vocabulary 2.1 — balanced dataset.** The *iris* dataset is said to be **balanced** as we observe data with the same absolute frequencies in each group considered. In our example, fifty flowers belonging to each of the (levels of the) Species *setosa*, *versicolor* and *virginica* have been measured.

**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 [40])
- 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 [40] 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 [32] 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.

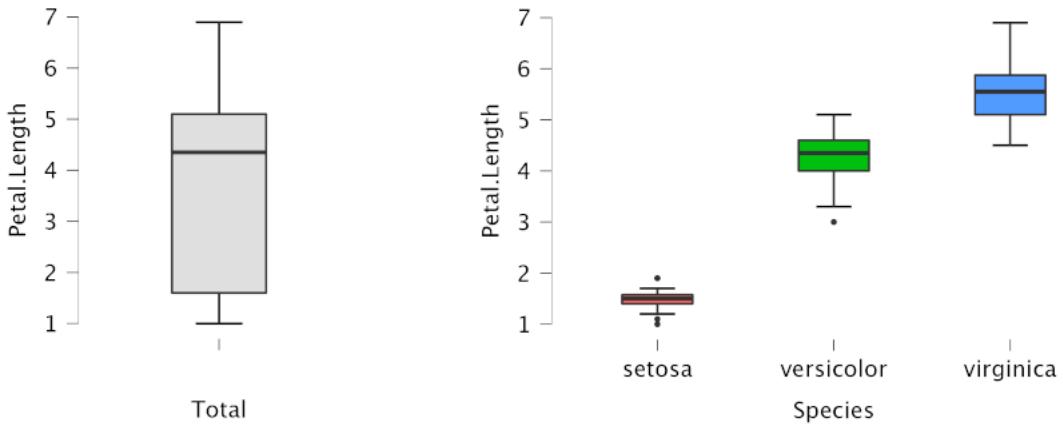


Figure 2.1: The boxplot.

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 — Quantiles.** 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**  $Q_1$  and the **third quartile**  $Q_3$  (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**,  $IRQ = Q_3 - Q_1$ , 'amplify' it by a 50%,  $1.5 \cdot IRQ$ , and search if there are points  $x_j \in x$  such that  $x_j < Q_1 - 1.5 \cdot IRQ$  or  $x_j > Q_3 + 1.5 \cdot IRQ$ . It can be shown (e.g. [31, 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' `ap1pack` [46].

## 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 [30], 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 [42] regarding the alcohol consumption in Estonia and Finland, <https://doi.org/10.1186/1471-2458-10-261>. Have a look to their Tables:

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

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	112 (110)	78

trapped the unaware reader in a pitfall, i.e. that in Estonia there might exists 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 Čebyšev 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 already mentioned (Section 2.3) 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 [47], <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 Pv events). ICCs for the NeuroSync<sub>MANU</sub>

## 2.4 Dealing with repeated measurements

Let us focus on the fact that up to now we were dealing with data collected along the 'easiest' study design, the **cross sectional** (rarely called **transversal**) design: in a undefined, not relevant, temporal point each subject is observed and data of our interest were collected. But, usually, in Life Sciences data are collected in a **longitudinal** scheme, along time – often one speaks of **repeated measurements**: for instance, when biologists work with qRT-PCR, evaluating the  $C_T$ 's in duplicate or in triplicate (technical or biological); or, when observing cellular growth day by day. Gynecologists follow month by month biometrical data of mother and fetus. Psychologists study many persons' behaviour (which can obviously varying time after time) about people belonging to different clusters ('groups', 'organizations', 'classes', ...), involving in this way also a more difficult **hierarchical data** structures.

Let us start from a controlled, balanced, and typical example occurred into the professor Gianni Cuda's lab, performed by Stefania Scalise e Maria Teresa De Angelis [17]; they were observing the oncosuppressor gene *p53* and a component of a certain interleukin (the p28 protein) with respect to a certain house-keeping gene. Have a look to their raw data:

<b>Id</b>	<b>Well</b>	<b>BiolRep</b>	<b>TechRep</b>	<b>CtGene</b>	<b>CtHK</b>
1	p28	a	1	22.76	17.20
2	p28	a	2	22.71	17.09
3	p28	b	1	22.46	16.95
4	p28	b	2	22.37	16.91
5	p28	c	1	22.69	17.09
6	p28	c	2	22.66	16.96
7	p53	d	1	24.65	17.08
8	p53	d	2	24.57	17.07
9	p53	e	1	24.54	17.26
10	p53	e	2	24.61	17.19
11	p53	f	1	24.76	17.33
12	p53	f	2	24.63	17.30

This sort of data representation is called **the long format** for repeated measures data and is particularly convenient if you use the programming languages like R to do some stat computations. Otherwise, JASP prefers **the short format**, in which every row reports the observed (repeated) measures.

	A	B	C	D	E
1	Well	CtGene1	CtHK1	CtGene2	CtHK2
2	p28	22.7585	17.1981	22.7101	17.0887
3	p28	22.4559	16.9521	22.3718	16.9078
4	p28	22.6911	17.088	22.658	16.9579
5	p53	24.651	17.0813	24.5663	17.0689
6	p53	24.5423	17.2591	24.6085	17.1872
7	p53	24.7599	17.328	24.6339	17.3011

## 2.5 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/Germaneto> 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*
-



### 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 [45, 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/Germaneto> repository, and draw a barplot as explained in subsection 2.2.2.

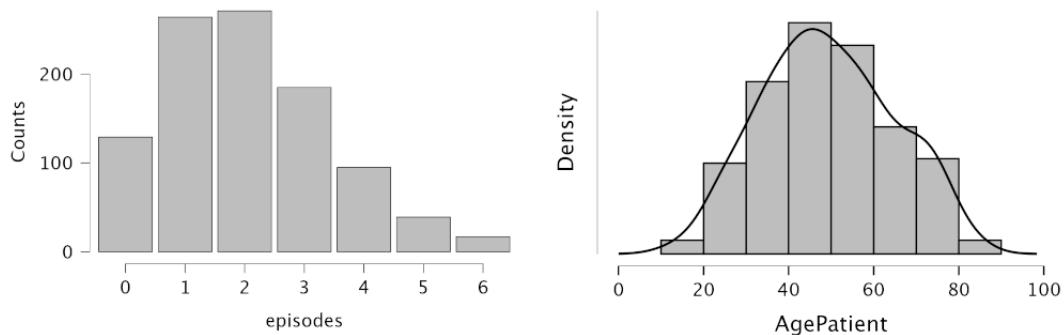
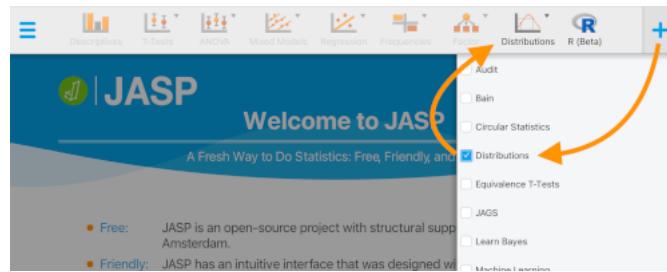


Figure 3.1: Estimating the probability density function of a continuous random variable

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

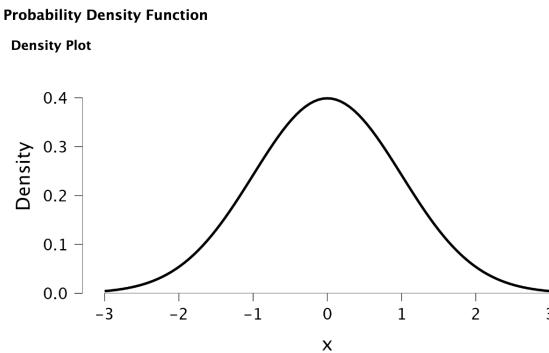
estimated density function, as in the right panel of Figure 3.1. Such curve relies on the histogram, which is of course an estimator[52] of the density function (which, in turns, depends on the starting point of the grid of bins – and the effect can be surprisingly large, as Venables and Ripley explain very well in their Figure 5.8 [52, 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

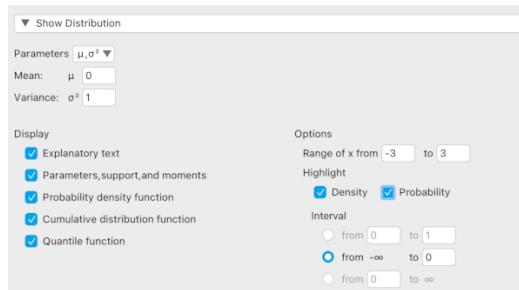


JASP possesses an additional menu which allows to study and to simulate the data random behaviour. Let us recap some basic facts on the most frequently used random variables in the medical field.

### 3.2.1 The Normal Distribution



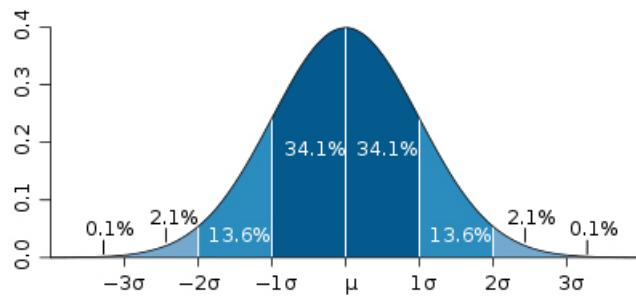
With JASP it is straightforward to perform calculations with the gaussian random distribution. The Show distribution menu provide to the user a nice way to reflect over the mathematical relations between the density function, the cumulative distribution and the quantile function, also highlighting the 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 [45, 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  $< 40$ ? ■

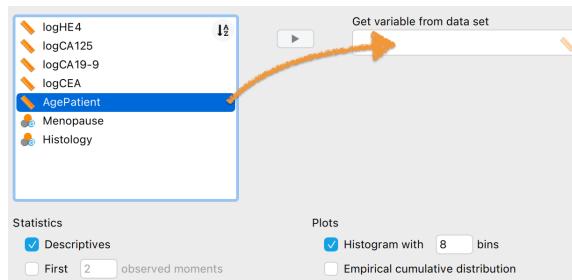
**Exercise 3.2** (B. Rosner, example 5.23 [45, page 132]) Glaucoma is characterized by intraocular pressure greater than 20 mmHg, while in normal population intraocular pressure  $X$  has mean  $\mu = 16$  and standard deviation  $\sigma = 3$ . How much it could be  $P(12 \leq X \leq 20)$ ? ■



**Exercise 3.3** (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/Germaneto> 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. [39], 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%20plot>, 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.

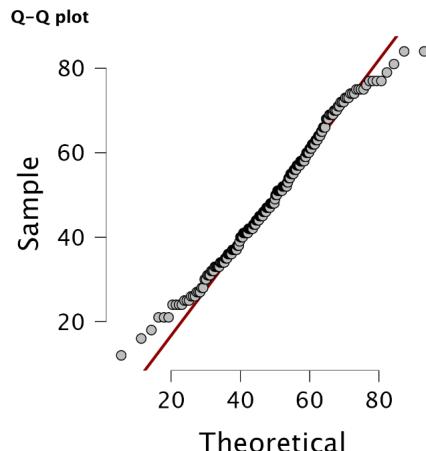


Figure 3.2: The quantile - quantile normal plot

### 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 [45, 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 [32, 35], in which the tallers (mostly, boys) stay on the right of the photo of the next page, while the smallers (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.

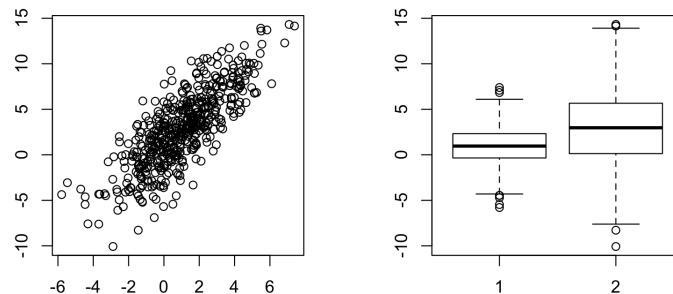


In particular, in a 1947 number of *Nature*, S. Vaswani [51] provide a counterexample, recalled and enlarged by C. Kowalski in his 1973 *Non-Normal Bivariate Distributions with Normal Marginals* [37]. And in 1982, E. Melnick and A. Tenenbein, with their *Misspecifications of the Normal Distribution* [38], 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.



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).

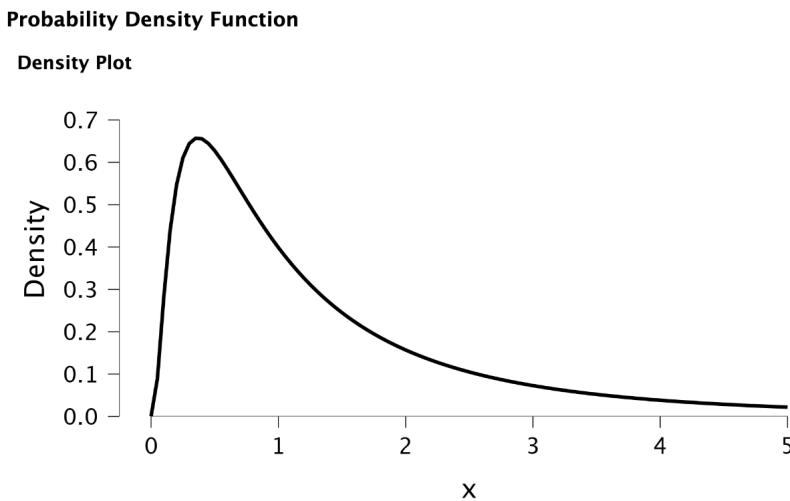


### 3.2.2 The Lognormal Distribution



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):



**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 [25]. 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  $\text{mean} \neq \text{median} \neq \text{mode}$ .

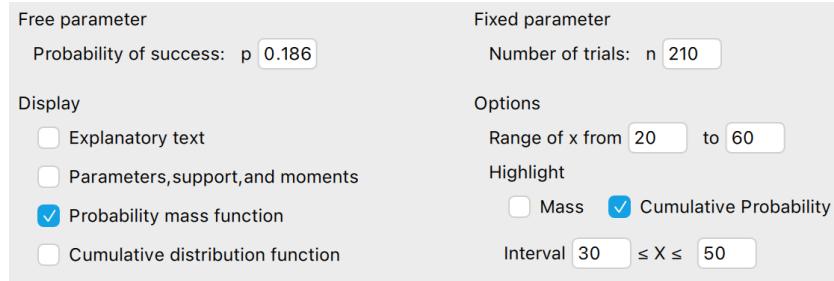
**Discussion 3.2.2 — summarizing body mass index.** Have a look to the body mass index histogram of more than  $10^5$  patients studied by Gregg Fonarow, <https://doi.org/10.1016/j.ahj.2006.09.007>. Suppose that you are required to lead a pilot study concerning radiation dosimetry in 25 obese patients. 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 3.4** 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*'.



**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 is it 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 [44], 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.

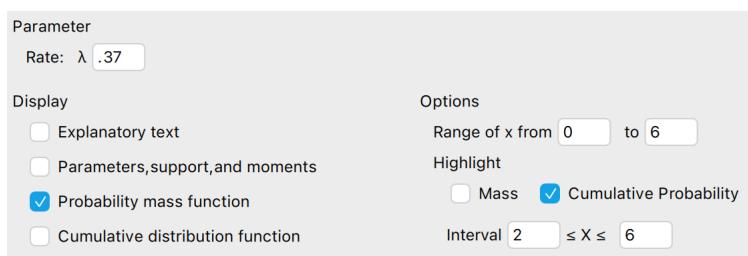
As an introductory example related to cancer, let us consider the Figure 7.4 of Daniel Zips, *Tumour growth and response to radiation*, collected in [36]. Let us read his words about the local tumour control:

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.

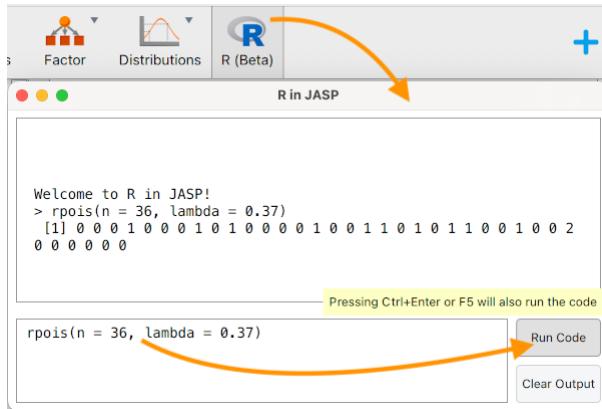
In that example, the Poisson distribution has the intensity (i.e. the mean, also called 'rate parameter')  $\lambda = 0.37$ .

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

**Exercise 3.5** 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.



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:



### 3.3 Exercises

■ **Activity 3.1 — describe in a proper way a dataset.** Search and read the paper by Annalisa Di Cello et al., entitled *A more accurate method to interpret lactate dehydrogenase (LDH) isoenzymes' results in patients with uterine masses* (PMID 30925450). Connect to <https://github.com/MassimoBorelli/Germaneto>, download the magnagraecia dataset and import it into JASP.

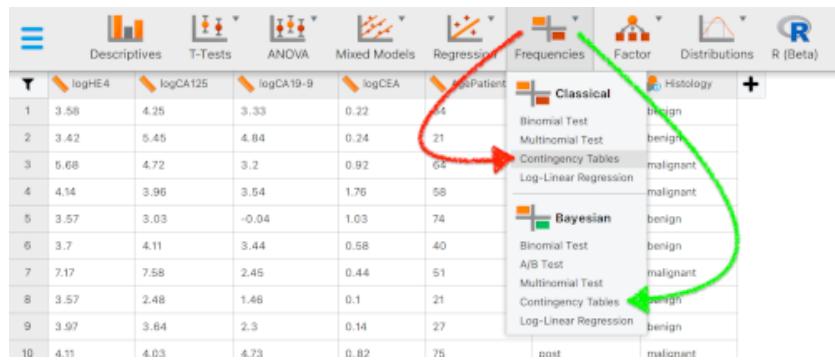
- Create the 'Table 1' of a paper manuscript, describing AGE, LDH1, LDH2, LDH3, LDH4, LDH5 and UMG.
- Provide a rebuttal answer to the cruel Referee which maintains that '*in Table 1 all variables should be described as mean  $\pm$  standard deviation.*'

### 3.4 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 the 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 ways to follow, the **Classical** and the **Bayesian** one. Let us start with the first one:



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?

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

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

**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 3.6** 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). ■

### 3.4.1 Bayes theorem

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

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

**Vocabulary 3.2 — Prior and posterior probability.** In the Bayes theorem, the marginal  $P(malignant)$  probability is called the **a priori probability**, while the conditional  $P(malignant|ante)$  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  $P(post|malignant)$  and  $P(ante|benign)$ ,  $Sens = 27/39 = 69\%$ , 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 [40]:

**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,  $PPV = P(malignant|post) = 27/92 = 29\%$  and  $NPV = P(benign|ante) = 106/118 = 90\%$ . 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>

### 3.4.2 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 3.4.2, the conditional probabilities:

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

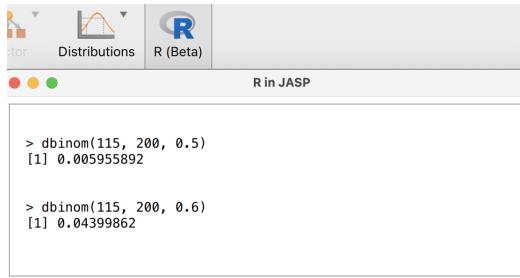
$$P(X = 115 | Bob) = \binom{200}{115} \cdot 0.6^{115} \cdot 0.4^{200-115} \approx 0.044$$

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.

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.



### 3.5 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.

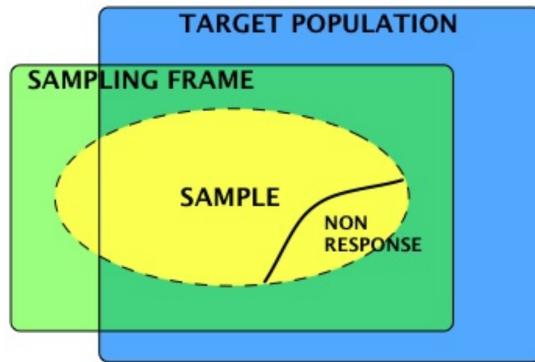
To obtain for example, an estimate of the population mean  $\mu$  of a certain characteristic  $x$  of the population, *sampling* must first take place because all the values of  $x$  for the entire population cannot be measured. Only a small part of the population can be surveyed and that part is called 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 3.5 by Stefano Panzeri [41], that our data could be affected by a not-random sort of 'distortion' 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.

### 3.6 Mismatching variability with reliability

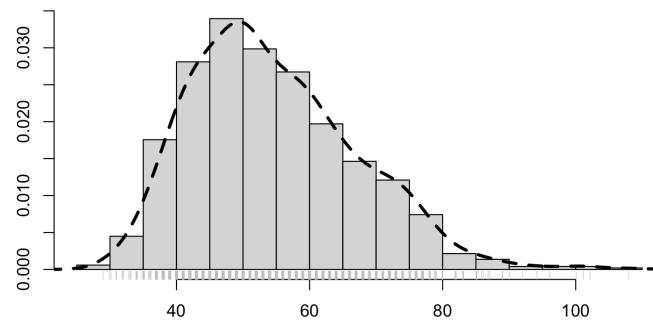
It is now the suitable moment to 'unblur' the image relative to Discussion 2.2.1: JASP collects under the descriptive menu the **standard error of the mean** index.

Dispersion	
<input checked="" type="checkbox"/> S.E.mean	<input checked="" type="checkbox"/> Std.deviation
<input type="checkbox"/> Coefficient of Variation	<input type="checkbox"/> MAD
<input type="checkbox"/> MAD Robust	<input type="checkbox"/> IQR
<input type="checkbox"/> Variance	<input type="checkbox"/> Range
<input type="checkbox"/> Minimum	<input type="checkbox"/> Maximum

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/Germaneto> 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$HDLcho1, 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$HDLcho1, 49)|
```

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$HDLcho1, 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, 15]. 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.

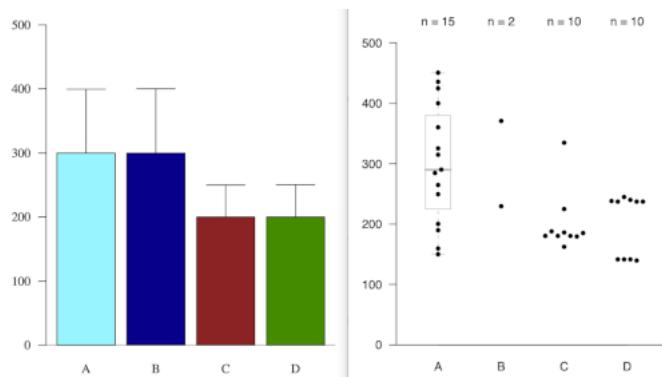
The error bars are very frequently exploited in biomedical literature to present experimental data collected with repeated measures. But many statisticians agree with Tatsuki Koyama, now

transfected enaotnenai cells. C shows migration assay for control *tac2* 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 0 and 10 minutes respectively. + indicates

at the Vanderbilt School of Medicine, which calls such very dangerous diagrams the **dynamite plots**: they do not convey important information and they are usually misleading. His poster is worth reading:



Tatsuki Koyama. Beware od 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* [55] and its natural sequel by Yu-Kang Tu and Mark Gilthorpe.

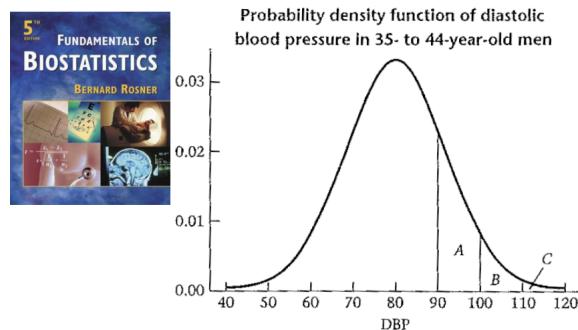


Richard Wainer. The most dangerous equation  
[https://www.researchgate.net/publication/255612702\\_The\\_Most\\_Dangerous\\_Equation](https://www.researchgate.net/publication/255612702_The_Most_Dangerous_Equation)



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>

### 3.7 Exercises



- **Activity 3.2 — 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 [49] 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 topic:

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 [61]:

<sup>1</sup>The quantity  $t$  is usually called **test statistic**, and this is a sort of pun, and source of confusion, in various language 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 synonymous 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 [49, 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

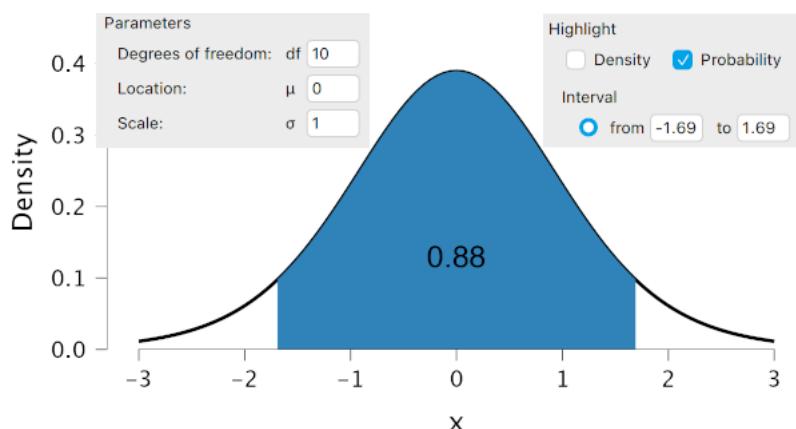
	difference
Valid	11
Missing	0
Mean	33.727
Std. Error of Mean	19.951
Std. Deviation	66.171

Let us import in JASP the gossett dataset, stored as usual in the <https://github.com/MassimoBorelli/Germaneto> 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:

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

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 ...

**Exercise 4.1 — Student is not Gauss.** From the Distributions menu, evaluate the white area outside interval from -1.69 to 1.69 under the normal distribution, and verify that it is approximately  $p = 0.090$ , and not  $p = 0.122$  as in the  $t$  distribution. This is the reason why

Gossett wrote in his paper:

Again, although it is well known that the method of using the normal curve is only trustworthy when the sample is 'large', no one has yet told us very clearly where the limit between 'large' and 'small' samples is to be drawn.



#### 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, as  $p = 0.122$ , what decision can we conclude? Here we go.

#### 4.2 Ronald Fisher's idea on significance level

It remains now to clarify how to draw a conclusion when knowing that  $p = 0.122$ : is it worth to kiln-dry the seeds, or it is uneuseful? In 1937 sir Ronald Aylmer Fisher started his fundamental book *The design of experiments* [24] presenting such a curious experiment.

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 this passage there are at least three relevant points which help to understand how  $p = 0.122$  can lead us to draw a decision. 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-value  $< \alpha \equiv 0.05$ ? Reject null hypotheses, there is an effect (kiln dried barley is different from not-kiln dried barley)
- p-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,[61]) and important debates have been rised (Ioannidis [33]), leading the American Statistical Association to release an official opinion with their *The ASA's statement on p-values* [56] (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 [16], 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 [40] 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 = (33.727 - 0),
                 sd = 66.171, 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 to discriminate if the kiln drying is useful or not.

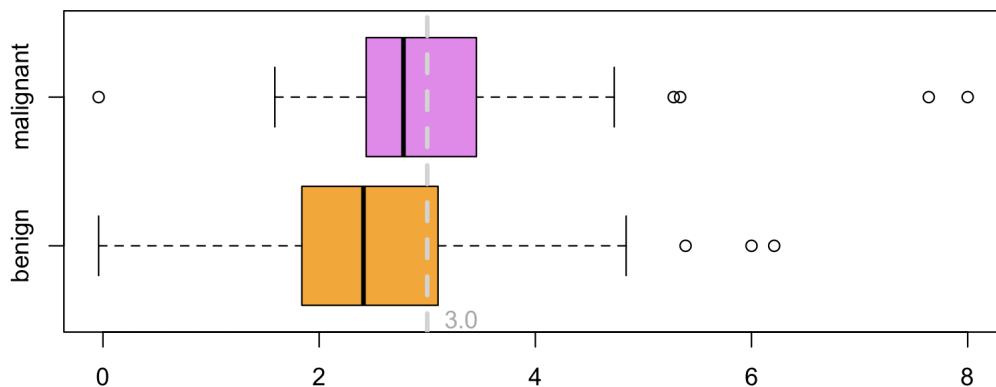


The power calculation here shown has only a didactical interest, but is is uneuseful – see John Hoenig and Dennis Heisey, *The Abuse of Power: The Pervasive Fallacy of Power Calculations for Data Analysis* [28].

### 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  $\log\text{CA19-9} = 3.0$ . 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 [40]:

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.bmjjournals.org/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:

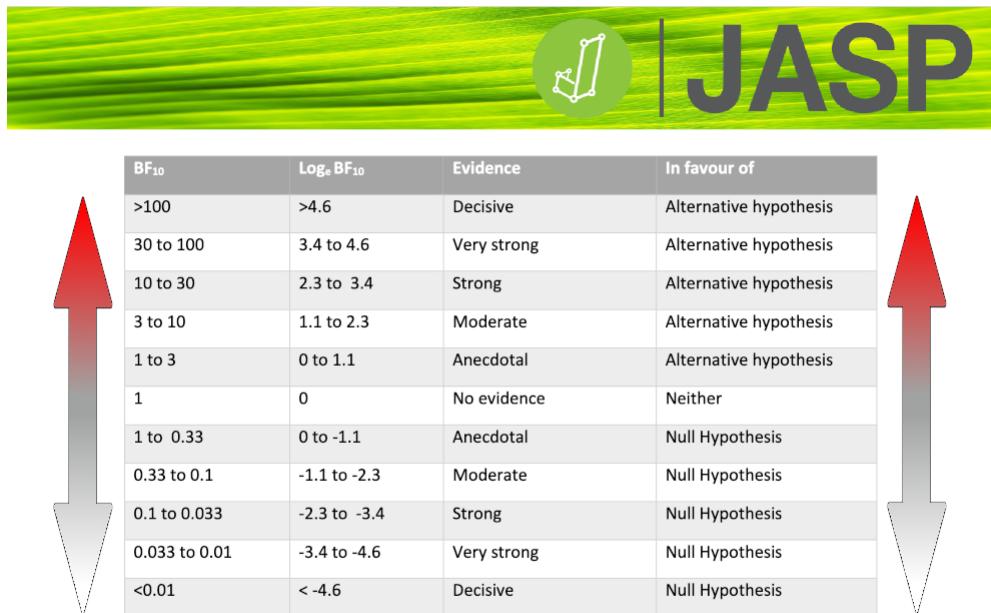
Table 4.5: Bayesian One Sample T-Test

	$BF_{10}$	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 4.4.1. 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[34] Jeffreys introduces the concept of the **non-informative prior distribution**: in fact, as recalled in Section 3.4.2, Gosset were observing eleven data  $D$  (the difference) having two different generative models:  $M_0$ , the normal distribution with  $\mu = 0$ , and  $M_1$  a normal distribution with  $\mu \neq 0$ :

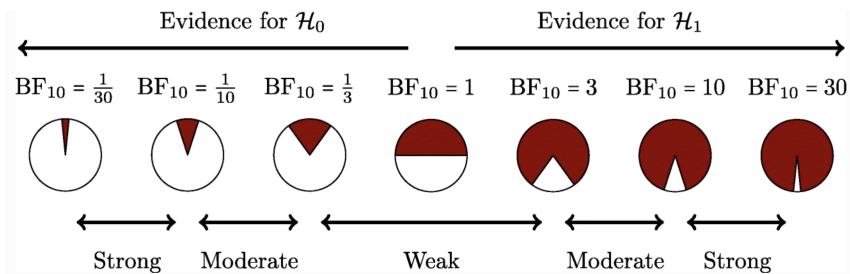
$$BF_{10} = \frac{P(D|M_1)}{P(D|M_0)} = 0.885$$



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

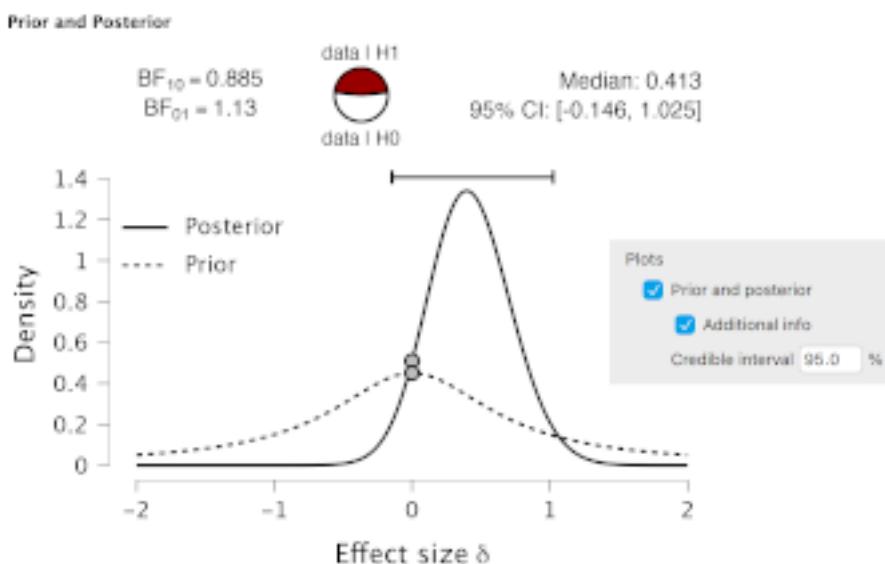
Here above we see the Mark Goss-Sampson[27] JASP table in evaluating Bayes Factor. As  $BF_{10} = 0.885$  is very close to 1 we should claim that Gosset experiment provided **absence of evidence** in favour of the null hypothesis  $\mu = 0$  (i.e. no difference between kiln dried and not-kiln dried seeds) 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 im-



mediately 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 %
kd	-	nkd	0.885    0.004

#### 4.5 In conclusion

Let us recap; JASP provides two possible approaches to statistical inference: the *frequentist* and the *bayesian* one. The latter, by exploiting the **Bayes Factor** as a summary, can provide (a continuous numerical) evidence toward one of the three possible decisions:

1. yes, something occurred
2. no, nothing happened
3. well, we can't say anything for sure

On the contrary the frequentist approach, relying on a probability **p-value**, 'melts' the second and the third decisions together and 'delegates' to the design of the experiment (i.e. choosing the proper sample size before collecting data, i.e. enhancing the test power) the faculty to cast away the unpleasant third possible decision.

#### 4.6 Exercises

*Additional hours' sleep gained by the use of hyoscyamine hydrobromide.*

Patient	1 (Dextro-)	2 (Laevo-)	Difference (2-1)
1.	+ .7	+ 1.9	+ 1.2
2.	- 1.6	+ .8	+ 2.4
3.	- .2	+ 1.1	+ 1.3
4.	- 1.2	+ .1	+ 1.3
5.	- 1	- .1	0
6.	+ 3.4	+ 4.4	+ 1.0
7.	+ 3.7	+ 5.5	+ 1.8
8.	+ .8	+ 1.6	+ .8
9.	0	+ 4.6	+ 4.6
10.	+ 2.0	+ 3.4	+ 1.4

■ **Activity 4.1 — the sleep dataset.** William Gosset [49] made it famous also the `sleep` dataset (you can find already stored inside the 1. Descriptives menu of the Data Library), concerning the 'soporific effect' of two optical isomers of a molecule. Analyze it and discuss the result. ■

## 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* [15, 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 Section 3.2.1 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

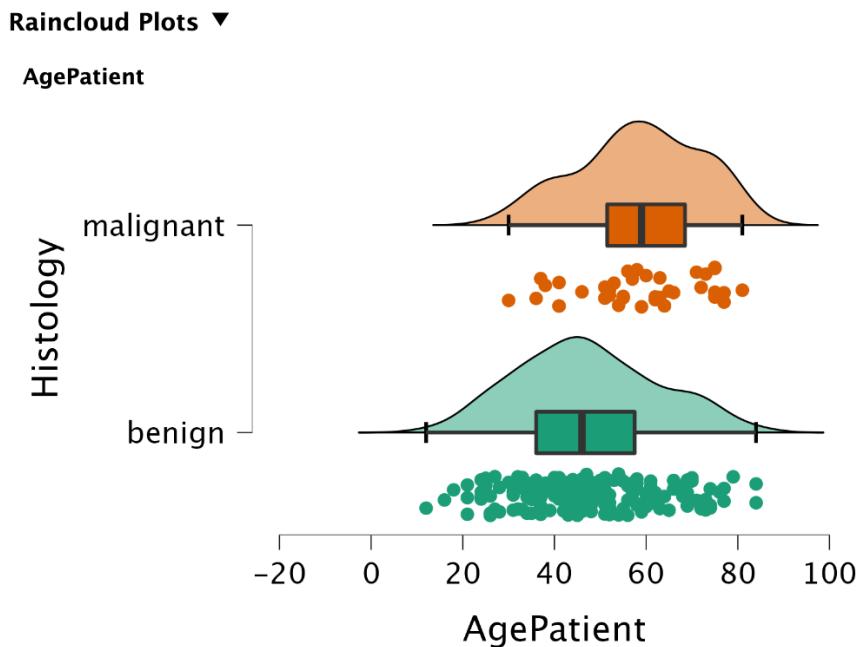
	$BF_{10}$	error %
AgePatient	652.530	$5.868e - 9$

and, being  $BF_{10}$  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:



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 sholud be better to depict two QQ-plot 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 pursue: the normality check is usually provided by applying the technique of Samuel Shapiro and Martin Wilk and their **normality 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.

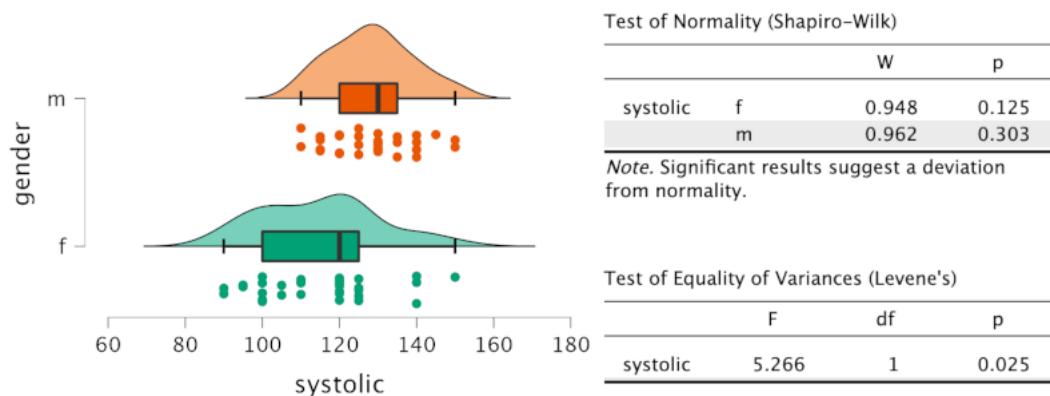
Normality (Shapiro-Wilk)		W	p	
AgePatient	benign	0.987	0.114	
	malignant	0.965	0.257	
Variances (Levene)		F	df	
AgePatient		1.049	1	0.307

### 5.1.2 The Welch test

Richard Mould [40] 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 [20] 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 originary Student T-Test are not fulfilled. Such mathematical questions have been explicated 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 convincing 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**:

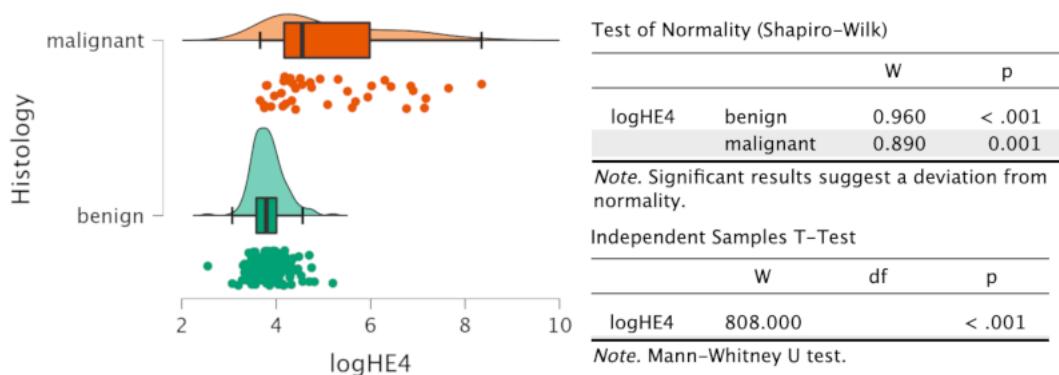


Wikipedia. Welch - Satterthwaite equation

[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  $\log\text{HE4}$  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 [14]. 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  $\text{BF}_{10}$  greater than one thousand, a decisive evidence in favour of the alternative hypothesis (i.e.  $\log\text{HE4}$  differs in benign and malignant ovarian lesions).

Table 5.2: Bayesian Mann-Whitney U Test

	$BF_{10}$	W	Rhat
logHE4	4073.742	808.000	1.087

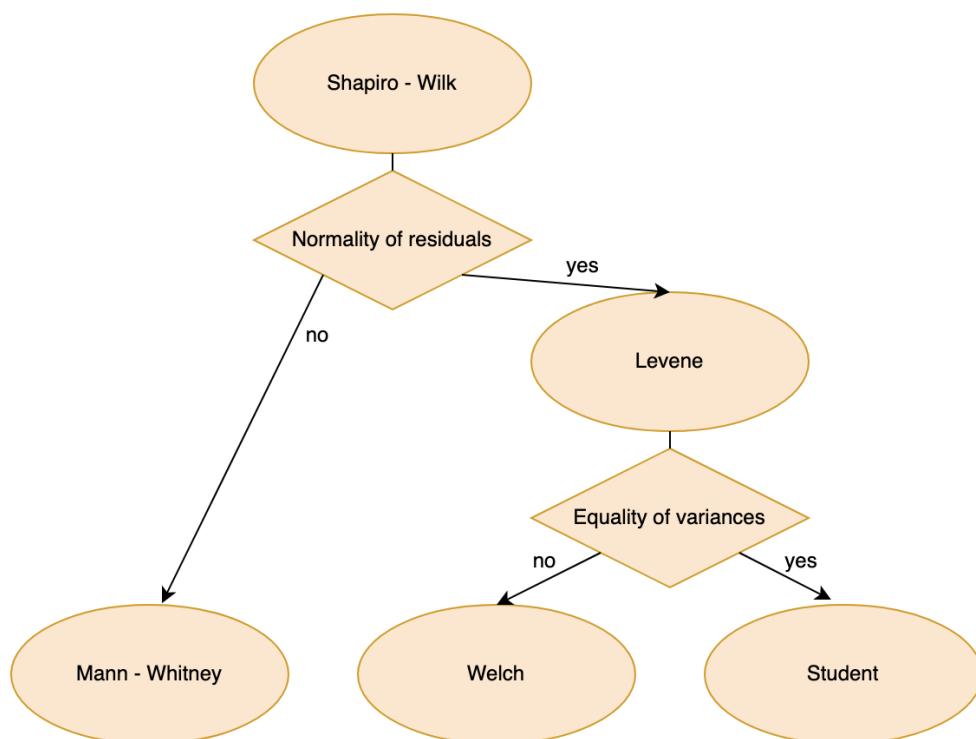
*Note.* Result based on data augmentation algorithm with 5 chains of 1000 iterations.

The output comes from a computational algorithm [19] 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>

#### 5.1.4 In conclusion

Let us recap with a flow chart what usually practitioners do when testing differences between two groups according to the frequentist approach.

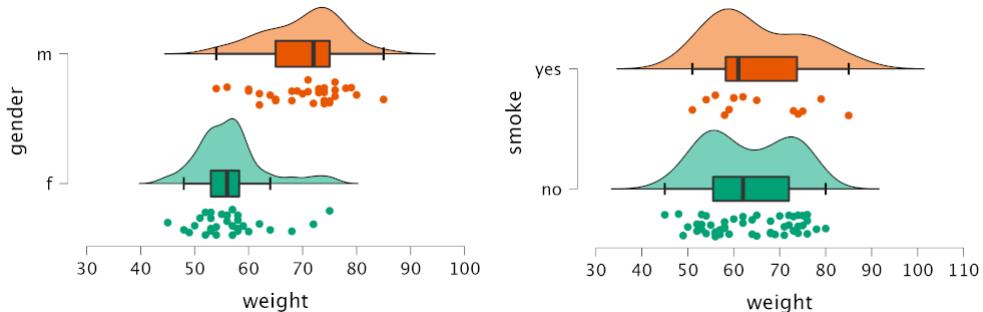


## 5.2 Three or more groups

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  $BF_{10}$ , 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.



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).

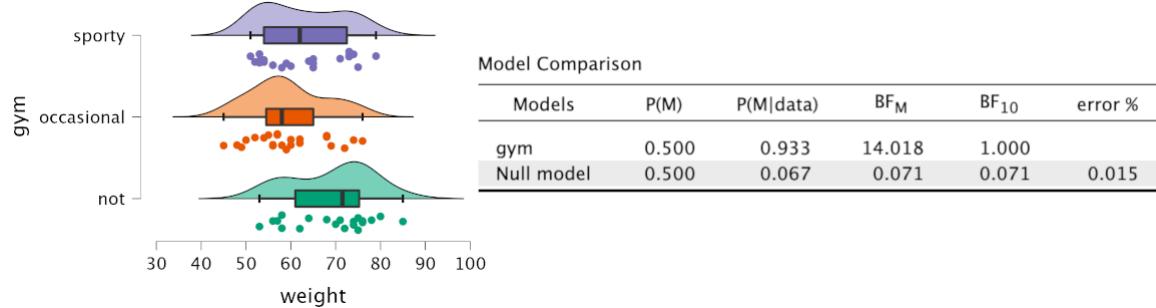
In conclusion, we have discovered the recovery plan: if we want to test differences between means, we have to test reduction in variances! And this is the reason why Anova (= *An.o.va.*, Analysis of Variance) has this strange name.

### 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 [44]:

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  $\sigma_1 = \sigma_2 = \sigma_3 = \dots$ .

**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**:





Wikipedia. Kruskal - Wallis one - way analysis of variance  
[https://en.wikipedia.org/wiki/Kruskal%20Wallis\\_one-way\\_analysis\\_of\\_variance](https://en.wikipedia.org/wiki/Kruskal%20Wallis_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 [40] 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.*

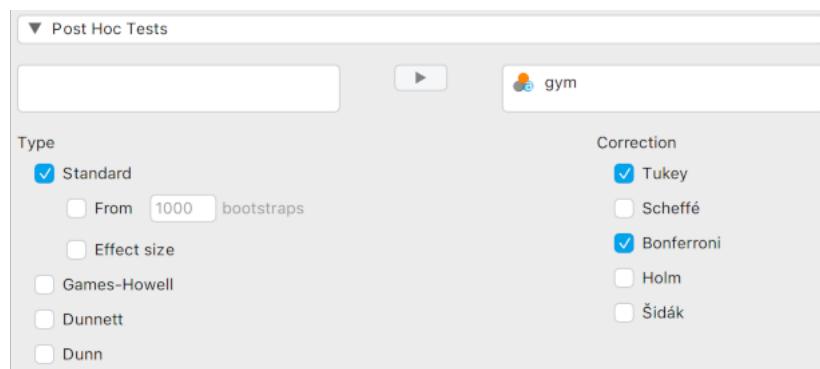
In fact, suppose to choose an  $\alpha$  level of 5%; then, the probability to commit an error of the first type is about the 14% (independent events, product of probabilities):

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

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**[43].



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**[15]:

		Mean Difference	SE	t	p <sub>tukey</sub>	p <sub>bonf</sub>
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:

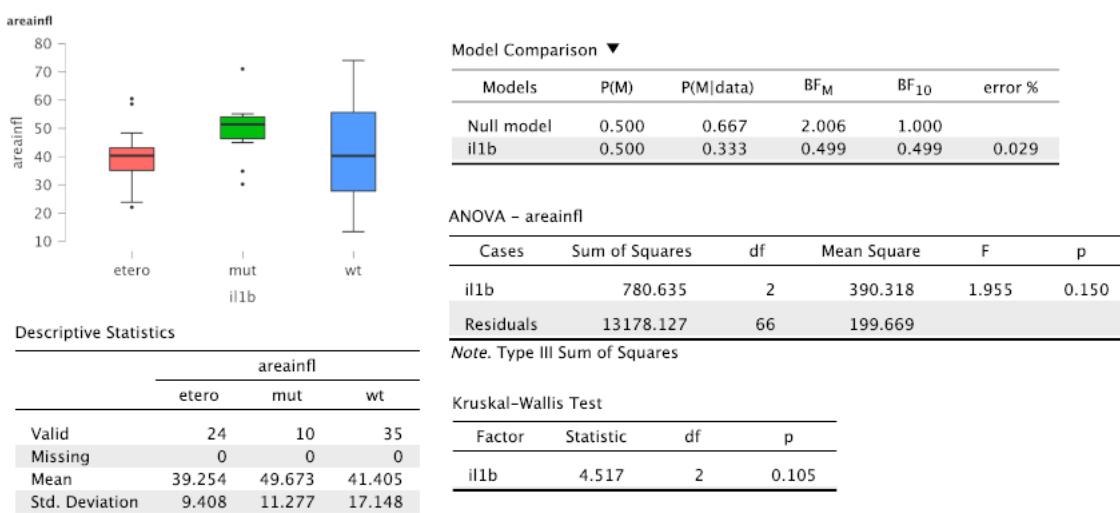
- not  $\neq$  occasional = sporty

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*'. We will return on the argument in Section 7.4.2, with a practical and viable solution.

### 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 sixtynine patients have been observed, measuring their gingival areainflammation 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 etero 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 [60] and multcomp [29] 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 7.4.2.



## 5.3 Exercises

- **Activity 5.1 — differences between groups.** 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/Germaneto> and import it into your JASP.

- Is the *Area outside shielding* different, in a statistical sense, with respect to the levels of *Angle*? Specify the proper test you chose.

■

■ **Activity 5.2 — differences between groups.** Reconsider the `fresher.ods` dataset. Choose the proper test.

- Is the systolic pressure different, in a statistical sense, with respect to the levels of the *gender*?
- Is the diastolic pressure different with respect to the levels of the *smoke*?
- Is the heartrate pressure different with respect to the levels of *gym*?

■



## 6. Introducing the linear model

### 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' [26], and at a first sight the two arguments perfectly overlap. We are going to discover that the linear model encompasses a variety of important and classical statistical tools (usually named **Ancova** methods, which will be discussed in the next Chapter) which encompasses also the **t-test** or the **Anova** we have just learnt.

### 6.2 The regression line

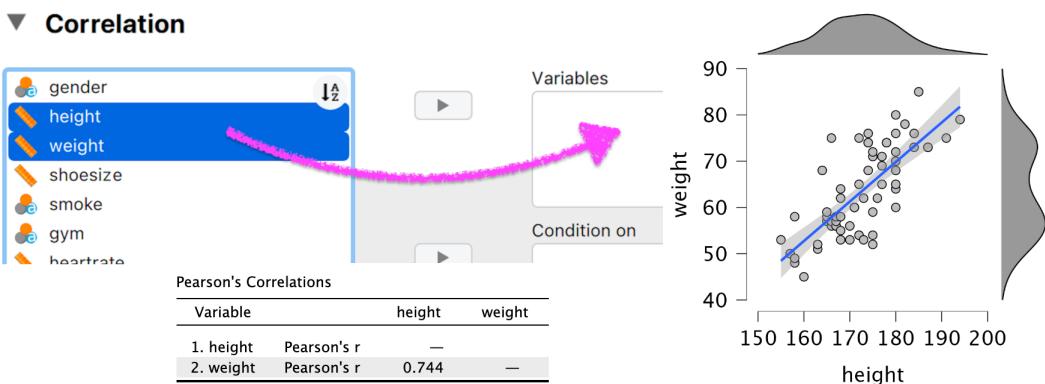
Suppose we are interested in assessing the possible relation that interlaces fresher's weights with their heights. It is a relation between two numeric variables, and we stress the role that height assumes as a possible **predictor** of (i.e. a dataset covariate significantly associated to) the **weight**. 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  $y$ .

### 6.2.1 Measuring point cloud disorder

A famous way to 'quantify the linear relationship' between two variables is the so called (**linear**) Auguste Bravais - Karl Pearson **correlation coefficient** (also called the product - moment correlation coefficient)  $1 \leq \rho \equiv 0.744 \leq 1$ . The squared value of the correlation coefficient,  $\rho^2$ , is called **coefficient of determination**, usually noted as  $R^2$ . The notion of coefficient of determination is linked to that of **Kullback - Leibler information measure**, which happens to be the negative of **Boltzmann's entropy** [10]. It is very interesting to observe that a computer science concept, a physics concept and statistics concept are so intimate related.



Here we provide a simple insight on  $\rho$  definition: 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:

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

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:

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

is dimensionless. And, ta-dah:

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

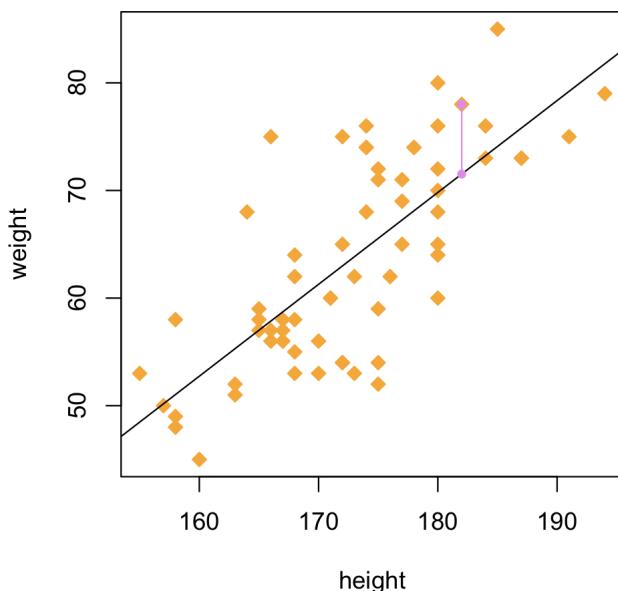
Before proceeding, we always remember that a statistical relation is **not** a cause-effect relation at all. Just for fun, look to the <http://www.tylervigen.com/spurious-correlations> in which for instance the divorce rate in Maine is put in relation with consumption of margarine.



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 [40], section 16.1)

### 6.2.2 Ordinary least square fitting



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. [22, 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 [48]. 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  $\varepsilon^T \cdot \varepsilon \equiv <\varepsilon|\varepsilon>$  and search the parameters  $a$  and  $b$  which minimize such scalar product.



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

**Exercise 6.1** Think a simple way to check that the regression line  $y = a + bx$  passes through the **mass center** of the point cloud, i.e. the point  $(\text{mean}(\text{height}), \text{mean}(\text{weight}))$ .

### 6.2.3 Toward linear modelling

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

Dependent Variable	Model Comparison - weight					
	Models	P(M)	P(M data)	BF <sub>M</sub>	BF <sub>10</sub>	R <sup>2</sup>
Covariates	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, remebering 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'. In fact, remembering the 'signal to noise ratio' discourse, we can use the  $t = \frac{x_m - \mu}{s/\sqrt{n}}$  relation, obtaining exactly the statistic  $t = \frac{0.85392 - 0}{0.09649} = 8.849\dots$ . Being more than 8 deviates far away from 0, we are sure (i.e. p value  $1.18e-12 < 0.001$ ) that the line has a not null slope, i.e. that weight is predicted by height.

### 6.2.4 Understanding random effect

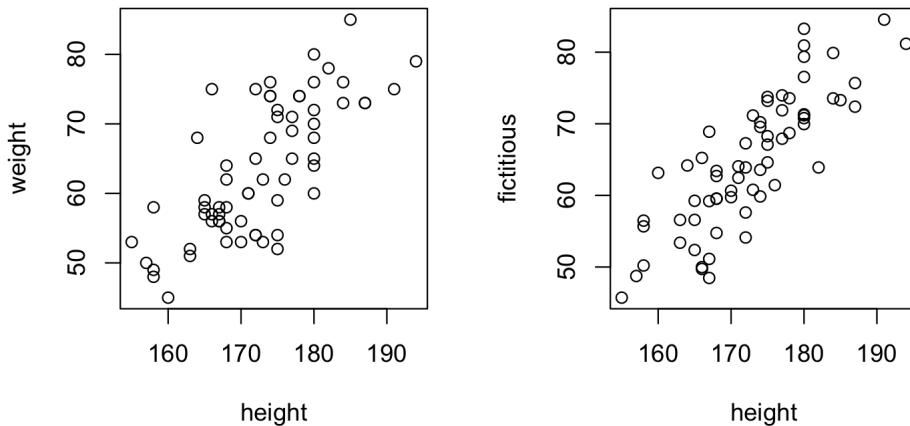
How can a thing develop out of its antithesis? For example, [...] truth from error?  
(Friedrich Nietzsche, *Menschliches, Allzumenschliches*)

While in naïve school algebra the straight line is fully described by two parameters (the intercept and the slope), the regression line is a statistical model which conveys three parameters – and a fourth parameter taken for granted: in our example, two of them are the **fixed effects**  $a = 0.854$  and  $b = -83.891$ , the intercept and the slope of the line. The fixed effects represent the first of the two **components** of the linear model; the second one is the so called **stochastic component** or **random effects** of the linear model, which has a 'taken for granted' parameter – the null mean of the residuals, that is the normal distribution  $N(\mu, \sigma)$  describing the residuals is always of the form  $N(0, \sigma)$  – and, lastly, the third parameter: the standard deviation  $\sigma$  quantifying the dispersion of the sampled residuals along the null mean (remember the standard error of the mean in section 3.6 and 5.1.3).

The above Model Summary - weight table in the **Classical** linear regression menu provides an estimate to  $\sigma$ : it is the **Root Mean Square Error** RMSE of the  $H_1$  model,  $\sigma = 6.46$ .

Model	R	$R^2$	Adjusted $R^2$	RMSE
$H_0$	0.000	0.000	0.000	9.598
$H_1$	0.744	0.554	0.547	6.459

Why is important to know both the fixed effects and the random component of a model? Naively, because one can judge if a model 'fit' in a proper way the data. Look to the below picture easily generated with few lines of an R code: in the left panel, the true weight vs. height of our students; the right panel instead shows a plot of 65 fictitious weights random generated, according to the estimated regression line perturbed by a `rnorm` normally distributed with null mean and 6.46 standard deviation. The fact that two panels resemble each other suggests that our linear model is well posed. We will discuss better this issue in the paragraph.



**Exercise 6.2** Try to re-plot the simulation in the right panel above, copying and pasting the following commands line into the R in JASP console:

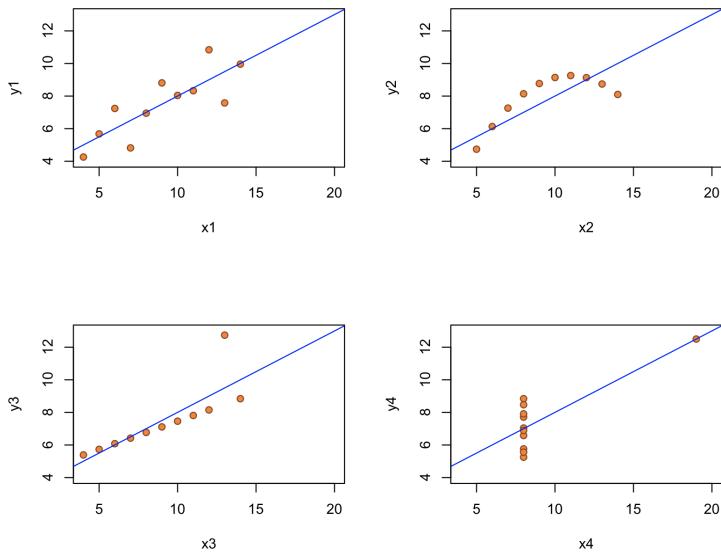
```
fictitious = -83.89 + 0.85 * data$height + rnorm(65, 0, 6.46)
plot(data$height, fictitious)
```

### 6.3 The diagnostic of a linear model

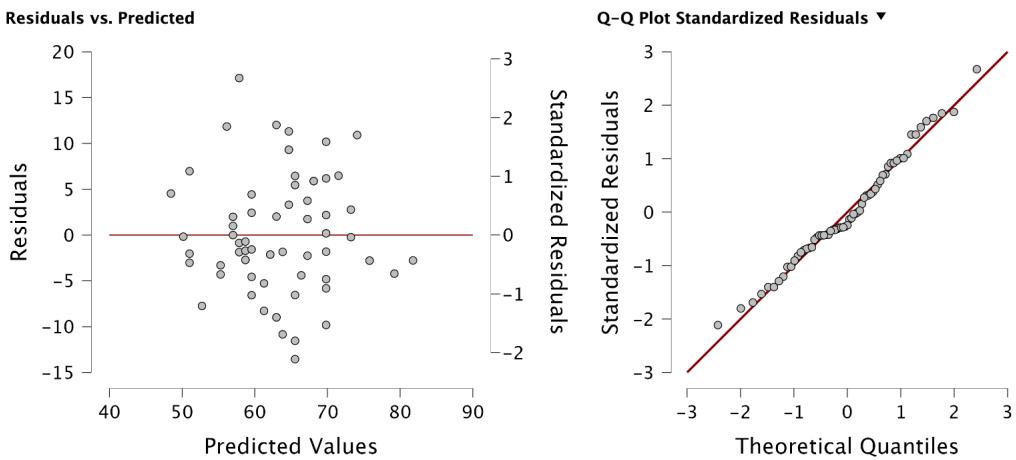


Wikipedia. Anscombe's quartet.  
[https://en.wikipedia.org/wiki/Anscombe%27s\\_quartet](https://en.wikipedia.org/wiki/Anscombe%27s_quartet)

In a remarkable 1973 paper [5], Francis Anscombe exhibited four artificial datasets, indeed very different one another, but characterized to have the same regression lines  $y = \frac{1}{2}x + 3$  and the same correlation (i.e. the same  $R^2 = 0.66$  determination coefficient). You can import them in JASP from the 17. Miscellaneous folder. It is clear that the blue regression line is a 'good model' only in the first of the four panels; but when leading a statistical analysis, in particular in datasets with many covariates, we need some tools to judge the 'quality' of our model, to decide if our linear model fits well data, or not. This can be done by means of the so-called **diagnostic plots**.



Many of them are listed, for example, in the .pdf *Data Quality Assessment Statistical Methods for Practitioners* provided by EPA (on <https://nepis.epa.gov>). With JASP, two of them are easily obtained, and we quote professor Michael Crawley [15, page 357] in explaining them.

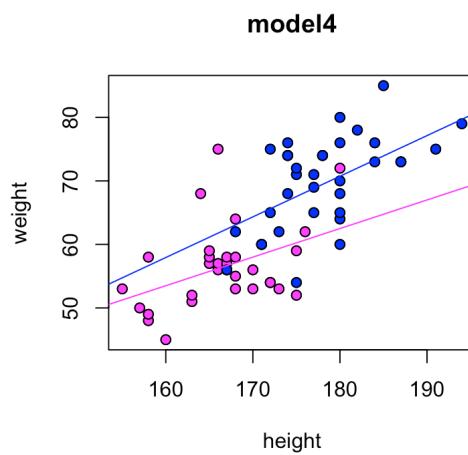


The first plot is called **Residuals vs. Predicted**. The point clouds represent nothing but the vector  $\varepsilon = (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_i, \dots, \varepsilon_n)$  along the predicted values  $y$ , as defined in previous section 6.2. We judge to be in a 'good model situation' if the cloud does not exhibit a pronounced curvature, or a 'wedge' shape (the latter being a hint for nonconstant dispersion  $\sigma$ ). In other words, this plot should look like the sky at night, with no pattern of any sort. The second plot, is the Q-Q plot as introduced in Section 3.2.1: as we desire to avoid the non-normality in the residuals, the plot should be straight as in this case, and not banana-shaped or humped. When the latter phenomena occur, it is necessary to modify the model: in next Chapter we will see a viable road.

## 7. Introducing multivariable analysis

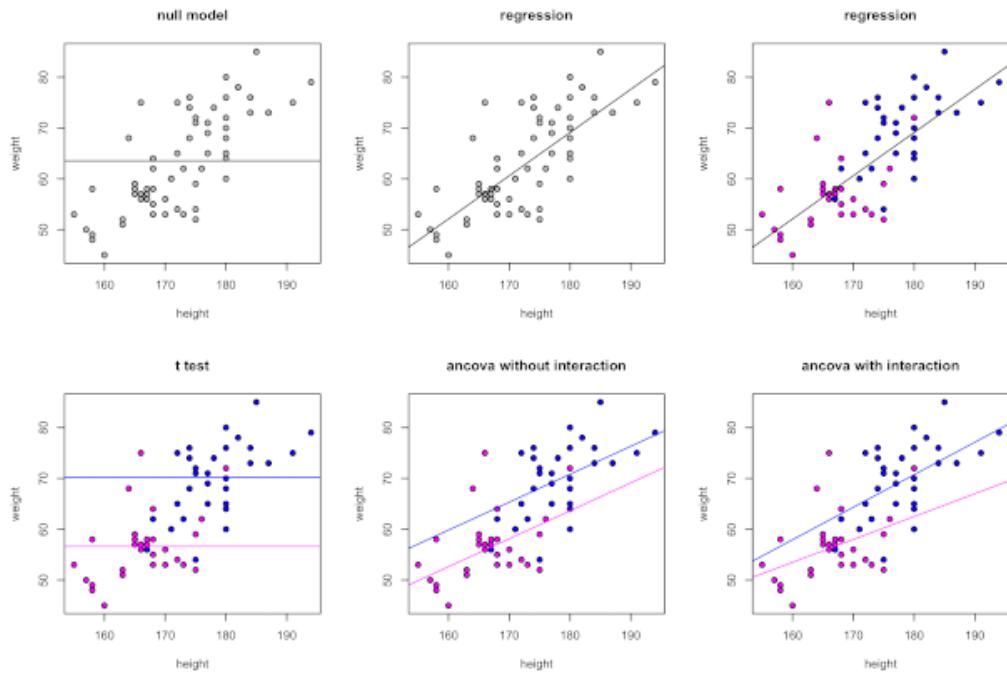
### 7.1 Overview

In the Exercise 5.1 we have observed that in the `fresher` dataset the weight can be predicted, in a statistical sense, by gender and in Paragraph 6.2.3 we saw that also height is a significant predictor. The *naive* question now is: if we 'melt' together those two predictors starting to perform a **multivariate** (or, better to say, **multivariable**) regression, we increase model's 'goodness-of-fit'?



The common approach to this question is known with the term **Ancova**, which stands for 'analysis of covariance'. To preform an Ancova in JASP one can choose to start from the ANOVA menu, or almost equivalently from the Regression menu. If you wonder why there is such a sort of 'duplication', and if you also wonder why in the **Bayesian** Regression the Logistic Regression is not (currently) contemplated, you will find an answer in the following pages.

## 7.2 The Wilkinson and Rogers notation



In the above picture, we depict different statistical situations and different possible linear models, investigating on weight and its possible relations with height and / or gender. Each of them can be denoted in a very compact way according to a notation formalized during Seventies by statisticians G. Wilkinson and C. Rogers.

1. In the first picture – which, actually, we already know to be unrealistic – the weight seems not to be related neither to height (in fact the regression line is flat) nor to gender (as all points are indistinguishably gray coloured). This is the **null model**  $H_0$ , and according to Wilkinson and Rogers notation it is noted:

$$\text{weight} \sim 1$$

Actually the flat black line passes through the mean of the weights, 63.5 Kg.

2. The second picture shows that the weight has a relation with the height, but neglecting the existence of the gender information (being all points are indistinguishably gray coloured). We are talking of the **regression line** and we use the symbolic writing:

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

3. The third picture shows that same regression line of the second drawing, but now the points are pink and blue coloured in order to enhance the gender information. Despite this, the regression line continues to be black painted, to say that weight has a significant relation with the height, but not significant with the gender. The notation is the previous one:

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

4. The fourth picture depicts the **T-Test** situation: the pink cloud and the blue cloud possess significantly different means (girls 56.6 Kg, boys 70.2 Kg), but the height does not convey any useful information to predict the persons' masses:

$$\text{weight} \sim \text{gender}$$

5. The fifth picture illustrates a novel situation: the pink cloud and the blue cloud possess significantly different means (girls 56.6 Kg, boys 70.2 Kg), and also the height conveys useful information to predict weight. The regression lines therefore have different intercepts but equal slopes (i.e. they are parallel); in statistical parlance this model is called **Ancova without interaction** and it will be presented in the next section. Its Wilkinson and Rogers notation is:

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

6. The last picture refers to the so called **Ancova with interaction** model: the regression lines have different intercepts and different slopes, and the admitted Wilkinson and Rogers notation is twofold:

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

or, equivalently,

$$\text{weight} \sim \text{gender} + \text{height} + \text{height:gender}$$

There exist also further possibilities, which are usually considered when analyses involve **standardized data**, i.e instead of considering raw data  $X$  one rescales all variables according to  $(X - \mu)/\sigma$  transformations, which are typical into psychometrics. According to those transformations the mass centers are translated into zero, and therefore intercepts are always null. The Wilkinson and Rogers notation in those cases are again twofold:

$$\text{response} \sim \text{some relation} - 1$$

or, equivalently,

$$\text{response} \sim \text{some realtion} + 0$$

In the following pages we will learn to decide which of the previous model is worth to be chosen. The key idea is that when we face competing models performing the same prediction, we shall select the model with the fewest assumptions, i.e. with less parameters, according to the latin motto '*Frustra fit per plura quod potest fieri per pauciora*'.



Wikipedia. Occam's Razor.  
[https://en.wikipedia.org/wiki/Occam%27s\\_razor](https://en.wikipedia.org/wiki/Occam%27s_razor)

## 7.3 Ancova

### 7.3.1 Ancova without interaction

We start making a first investigation with the default **Bayesian ANCOVA** menu, considering weight as Dependent Variable, gender as Fixed Factor and height as a Covariate. A smashing  $\text{BF}_M = 180.7$  dispels any doubt: both variables are to be considered as predictors.

Models	P(M)	P(M data)	$\text{BF}_M$	$\text{BF}_{10}$	error %
gender + height	0.250	0.984	<b>180.707</b>	1.000	
height	0.250	0.016	0.047	0.016	1.767
gender	0.250	$8.170e-4$	0.002	$8.305e-4$	1.767
Null model	0.250	$3.343e-12$	$1.003e-11$	$3.398e-12$	1.767

We can also pursuit the frequentist approach, to detect the same result:

Cases	Sum of Squares	df	Mean Square	F	p
gender	252.883	1	252.883	8.986	<b>0.004</b>
weight	556.453	1	556.453	19.772	< .001
Residuals	1744.880	62	28.143		

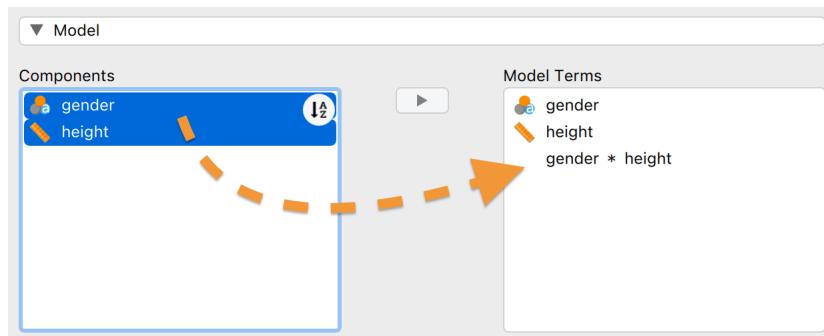
Now we are interested in estimating the *size effects*, i.e. *how much* gender and height influence the weight. So, we move from the ANOVA menu to **Classical** Linear Regression menu of the previous Chapter 6.2. The so called  $H_0$  model represents the **null model** as explained in Section 7.2, while the  $H_1$  is the fifth model we discussed there.

Model	Unstandardized	Standard Error	Standardized	t	p
$H_0$	(Intercept) 63.523	1.191		53.357	< .001
$H_1$	(Intercept) -35.373	20.716		-1.707	0.093
	gender (m) 7.196	2.061		3.492	< .001
	height 0.552	0.124	0.481	4.447	< .001

Let us correctly interpret the Unstandardized regression coefficients, as depicted in the fifth panel of the previous Section 7.2: the female pink regression line has slope 0.552 and intercept -35.373, i.e.  $y = 0.552 \cdot x - 35.373$ . The male blue regression line stands above the pink one, therefore the intercept shall have to be higher than -35.373. In fact the intercept becomes -35.373 + 7.196, i.e.  $y = 0.552 \cdot x - 28.177$ .

### 7.3.2 Ancova with interaction

Now we are called to decide about the the sixth panel, i.e. whether it is worth considering two different regression slopes: one steepest for the blue points, the boys, and one milder for the girls, the pink points. As usual, we start with the **Bayesian ANCOVA** menu, considering again weight as Dependent Variable, gender as Fixed Factor and height as a Covariate. We need now to learn a trick to insert the interaction term, `height * gender`. The trick is to move down until the drop down Model menu, to select **both** the Components and to drag and drop them into the Model Terms window, as shown by the orange dashed arrow here below:



We obtain a Model Comparison table, which summarize five different model assuming that they have the same prior probability to occur,  $P(M)=0.2$ .

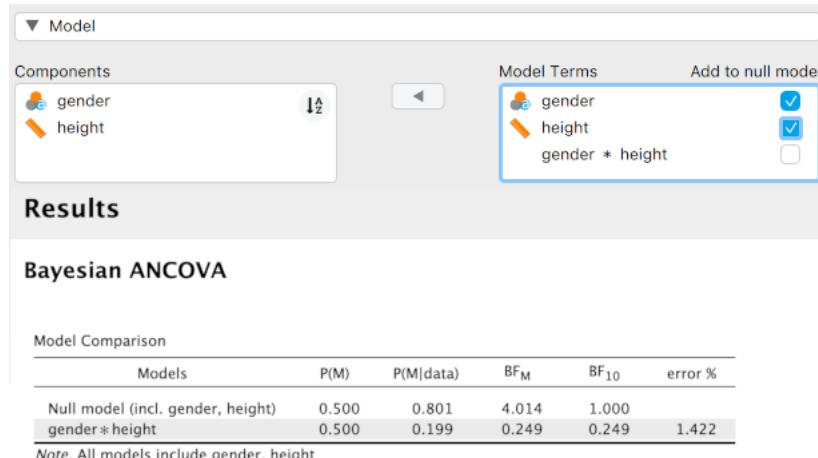
Models	P(M)	P(M data)	BF <sub>M</sub>	BF <sub>10</sub>	error %
gender + height	0.200	0.790	15.018	1.000	
gender + height + gender * height	0.200	0.197	0.980	<b>0.249</b>	1.422
height	0.200	0.013	0.052	0.016	0.778
gender	0.200	$6.802e-4$	0.003	$8.614e-4$	0.778
Null model	0.200	$2.783e-12$	$1.113e-11$	$3.525e-12$	0.778

Now, that 0.249 is not smashing anymore: the evidence supporting the necessity to 'pay for a new parameter' into the model seems to be only anecdotal, or maybe moderate.

**Exercise 7.1** Estimate the size effects of the Ancova with interaction model by means of the **Classical** Linear Regression menu and check that  $y = -18.50 + 0.45 \cdot x$  and  $y = -44.47 + 0.64 \cdot x$  are respectively the pink and the blue line equations. ■

## 7.4 The Model Selection

Of course, we urge to search for a reasonable criterion in order to decide whether the ancova with interaction or the ancova without interaction is the 'best' statistical relation to explain that point cloud. One bayesian possibility is the Add to null model option: we take for granted that gender and height are essential predictors, and we ask whether the \* operator gets better results:



We see again what we observed few minutes ago: a moderate evidence to null hypothesis, i.e. hard to tell if it is useful to take in account the interaction term. One can obviously resort the **Classical** Linear Regression menu obtaining a not significant 0.443 p-value:

Model	Unstandrd	Standard Error	Standardzd	t	p
H <sub>0</sub>	(Intercept)	63.523	1.191	53.357	< .001
H <sub>1</sub>	(Intercept)	-18.504	30.152	-0.614	0.542
	height	0.451	0.181	0.393	0.015
	gender (m)	-25.962	42.988	-0.604	0.548
	height * gender (m)	0.193	0.249	0.772	<b>0.443</b>

Nevertheless, many authors [3, 21] manteins that there are convincing reasons to avoid such an approach: why the conventional  $\alpha = 0.05$  level should by the criterion to select a proper model?

### 7.4.1 The Akaike Information Criterion



Stéphanie Portet. A primer on model selection using the Akaike Information Criterion  
<https://www.sciencedirect.com/science/article/pii/S2468042719300508>

In Section 5.2.3 we announced a difficult problem: to perform an anova in which data do not fulfill some necessary mathematical hypotheses, and we announced that the concept of entropy would be helpful. In 1974 a great intuition by professor Hirotugu Akaike[1] turned on a new reliable possibility in model selection: he was able to recognize that when statisticians evaluate the probability, or better the **likelihood**, to observe by chance certain model residuals, they are working with a function (the log-likelihood indeed) which can represents the 'cost' of the model if penalized by the number of parameters used (fixed effects + random effects), in agreement to the Kullback - Leibler theoretical framework [13, pages 28-30] on **divergence**, as an asymmetric distance (also called **relative entropy**) within two distribution probabilities. The professor Michael Crawley's crystal clear words [15, page 353] on the **Akaike information criterion** are illuminating:

The more parameters that there are in the model, the better the fit. You could obtain a perfect fit if you had a separate parameter for every data point, but this model would have absolutely no explanatory power. There is always going to be a trade-off between the goodness of fit and the number of parameters required by parsimony. AIC is useful because it explicitly penalizes any superfluous parameters in the model, by adding  $2(p + 1)$  to the deviance.

When comparing two models, the smaller the AIC, the better the fit.

Let us return to the unsolved question of section 7.3: we have to decide if it is proper to adopt the interaction model which 'costs' four fixed effects (two slopes and two intercepts), or the 'less expensive' ancovamodel02 with a common slope is sufficient. We compute their Akaike Information Criteria by means of the AIC function through the R in JASP window:

```
relationPlus = data$weight ~ data$gender + data$height
modelPlus = lm(relationPlus)
print(AIC(modelPlus))
relationCross = data$weight ~ data$gender * data$height
modelCross = lm(relationCross)
print(AIC(modelCross))
```

```
> relationPlus = data$weight ~ data$gender + data$height
modelPlus = lm(relationPlus)
print(AIC(modelPlus))
relationCross = data$weight ~ data$gender * data$height
modelCross = lm(relationCross)
print(AIC(modelCross))

[1] 421.278
[1] 422.6457
```

We are done: the modelCross has a higher cost, 422.6, than the modelPlus, 421.2, so we do not prefer it and we retain modelPlus as a minimal adequate model.

### 7.4.2 Multiple comparison and AIC

Let us recap what we saw in Section 5.2.2: we tried to predict weight in function of gym, a three level alphabetically ordered factor: not, occasional, sporty. Now with R in JASP we can set the linear model and evaluate its Akaike Information Criterion, which is equal to 473.1

```
relation3 = data$weight ~ data$gym
linearmodel3 = lm(relation3)
AIC(linearmodel3)
```

We have to decide (the 'multiple comparisons issue') if `linearmodel3` is the minimal adequate model and all the three different levels provide different information; or some levels can be joined together. Let us make some attempts, 'melting' together the `gym` factor levels in all the possible two by two manners, as shown in Table 7.1.

```
gymNO = data$gym
levels(gymNO) [1] = "notoccasional"
levels(gymNO) [2] = "notoccasional"
gymNS = data$gym
levels(gymNS) [1] = "notsporty"
levels(gymNS) [3] = "notsporty"
gymOS = data$gym
levels(gymOS) [2] = "occasionalsporty"
levels(gymOS) [3] = "occasionalsporty"
```

To be clear, this can be thought as a procedure which 'glues' to the dataset three new columns, `gymNO`, `gymNS` and `gymOS`; and all the new columns instead of having three levels has only two of them – see Table 7.1.

<b>weight</b>	<b>gym</b>	<b>gymNO</b>	<b>gymNS</b>	<b>gymOS</b>
53	not	notoccasional	notsporty	not
58	not	notoccasional	notsporty	not
50	occasional	notoccasional	occasional	occasionalsporty
49	occasional	notoccasional	occasional	occasionalsporty
73	sporty	sporty	notsporty	occasionalsporty
79	sporty	sporty	notsporty	occasionalsporty

Table 7.1: Example of six freshers whose gym activity has been re-grouped in a two by two manner.

Now, it is sufficient to compute the Akaike Information Criterion for all these new linear models, and to choose the smaller. And Bob's your uncle!

```
print(AIC(lm(data$weight ~ gymNO)))
print(AIC(lm(data$weight ~ gymNS)))
print(AIC(lm(data$weight ~ gymOS)))
```

We observe that the latter linear model, `weight` versus `gymOS`, has the lowest AIC = 472.7. Therefore it can be chosen as the minimal adequate model, and we interpret this saying that, in the fresher dataset, those who not practice gym has a `weight` significantly different from those who practice it in an occasional manner, or those who are sporty; and the latter two conditions do not differ between them.

**Exercise 7.2** Do you remember `iris` dataset of Section 2.2? Adapt the above code to decide whether `Sepal.Width` differs between `Species`. ■

## 7.5 Exercises

■ **Activity 7.1 — minimal adequate model by AIC.** Assume to be the systolic pressure the primary outcome (i.e. the response) of the `fresher.csv` dataset, and consider the following covariates as possible candidate predictors:

- the `heartrate`
- the `gender`
- the `smoke`
- the `diastolic` pressure

Explore several linear models involving one or more covariates and try to identify the possible minimal adequate model, exploiting the Akaike Information Criterion. ■

							
T	logHE4	logCA125	logCA19.9	logCEA	Classical	Menopause	Histology
1	3.59	4.26	3.22	0.22	Correlation	no	benign
2	3.43	5.45	4.04	0.24	Linear Regression	no	benign
3	5.60	4.73	3.2	0.92	Logistic Regression	yes	malign
4	-4.54	3.99	3.54	1.76	Bayesian	yes	malign
5	3.57	3.03	-0.34	1.03	Correlation	yes	benign
6	3.7	4.71	3.44	0.58	Linear Regression	yes	benign
7	7.17	2.58	3.45	0.48		no	benign
8	3.57	3.03	3.08	0.71		yes	benign
9	3.97	3.64	3.2	0.14	?	yes	benign

## 8. The logistic regression

### 8.1 The generalized linear model

Nowadays, biostatisticians often are consulted by biologists or physicians when seeking for reliable clinical biomarkers. In such a case, the typical response comes from a retrospective cross-section dataset whose response is of binomial type (benign/malignant, positive/negative, alive/dead, ...). In fact it was in Section 3.2.3 that we studied the Shadi Najaf Roma dataset, in which you remember 210 patients with a known Histology response (benign or malignant) were studied in association to four candidate biomarkers (logarithmic transformed) – logHE4, logCA125, logCA19.9 and logCEA –, along with their AgePatients and their Menopause status. And, as explained in subsection 3.2.3, the Histology response is not a numeric variable, but a factor response, mathematically modelled by a binomial random variable. Therefore, all we saw in the previous Chapters about the linear modelling machinery can not work at all. The same problems occur for instance when we have to model a count response, as described in the Poisson random variable 3.2.4 subsection. In situations like these, in which the response can not provide gaussian distributed residuals, we can exploit the **generalized linear models** theory.

**Vocabulary 8.1 — Generalized linear model.** A generalized linear model is a set of three 'statistical tools':

1. a **relation**, named the **linear predictor**, between the dataset response and one or more dataset covariates (as seen in previous Chapter 6).
2. a (family of) **random variable** able to model the response (or, to say better, to model the residuals)
3. a **link function** which transforms ('injects') the expected value of the random variable modelling the response into the mean of the linear predictor.

## 8.2 The logistic regression



Viv Bewick, Liz Cheek, Jonathan Ball. Statistics review 14: Logistic regression  
<https://ccforum.biomedcentral.com/articles/10.1186/cc3045>

In the `roma` dataset, a possible relation to investigate is suggested by Moore et al. [39], involving Menopause, logHE4 and logCA125 as predictors. But, being Menopause a binomial variable and logHE4, logCA125 numeric variable, we can imagine that in our case the linear predictor (often called also **predictive index**) maps the space  $\mathbb{Z}_2 \times \mathbb{R} \times \mathbb{R}$  into a real number  $y$ . Therefore, a suitable link function mapping such  $y \in \mathbb{R}$  into  $[0, 1] \supset \mathbb{Z}_2$  (the space of Histology) should be related to a probability evaluation. The standard choice is to start from the famous **logit** transformation, a bijective map from  $p \in [0, 1]$  into  $y \in \mathbb{R}$ :

$$y(p) = \text{logit}(p) \equiv \log\left(\frac{p}{1-p}\right)$$

and to compute the inverse function of the *logit*, in order to obtain the probability  $p$ . This inverse function is the famous **sigmoidal function**, also known as the **logistic** function (and this is the reason why often the binomial distributed generalized linear model is commonly called **logistic regression**).

$$p = \frac{e^y}{1 + e^y} \equiv \frac{1}{1 + \exp(-y)} \equiv \frac{1}{2} + \frac{1}{2} \tanh\left(\frac{y}{2}\right)$$

We have already noticed that JASP in its menu do not possess the bayesian counterpart of the **Classical Logistic Regression**. This fact mainly occurs because to fit the model is not an algebraic straightforward task, but one has to resort to an iterative method in order to maximize the likelihood on the model according to the Newton and Raphson's derivative method [21]. This is what is called the **Fisher Scoring** procedure.

### 8.2.1 Analyzing the `roma` dataset

We are interested to find which are the predictors, i.e. the statistically significant covariates, of the Histology within the `roma` dataset. We start to explore the **maximal additive model**, in which all the covariates are present, dragging and dropping all the icons in the proper windows:

`Histology ~ logHE4 + logCA125 + logCA19.9 + logCEA + AgePatient + Menopause`

	Estimate	Standard Error	z	Wald Test		
				Wald Statistic	df	p
(Intercept)	-14.921	2.872	-5.196	26.997	1	< .001
Menopause (post)	1.018	0.945	1.078	1.162	1	0.281
logHE4	2.410	0.704	3.423	11.719	1	< .001
logCA125	0.624	0.213	2.930	8.587	1	0.003
logCA19-9	0.222	0.202	1.101	1.211	1	0.271
logCEA	-0.183	0.426	-0.429	0.184	1	0.668
AgePatient	-0.001	0.030	-0.030	9.257e-4	1	0.976

We notice, trivially, that the maximal additive model has a very low AIC = 119.8, if compared to the null model (AIC = 203.578):

Model	Deviance	AIC	BIC	df	$\chi^2$	p
$H_0$	201.578	203.578	206.926	209		
$H_1$	105.827	119.827	143.257	203	95.752	< .001

Now, patiently, we remove the not significant ( $p = 0.976$ ) AgePatient covariate, and we see that the Akaike Information Criterion decreases, AIC = 117.8. Also, we continue removing logCEA (AIC = 116.0) and logCA19-9 (AIC = 115.3). After this simplification we arrive to the candidate **minimal adequate model**:

	Estimate	Standard Error	z	Wald Test		
				Wald Statistic	df	p
(Intercept)	-14.377	2.674	-5.377	28.910	1	< .001
logHE4	2.338	0.652	3.584	12.847	1	< .001
logCA125	0.684	0.203	3.374	11.384	1	< .001
Menopause (post)	0.938	0.570	1.645	2.707	1	0.100

It remains a little doubt: is it useful to consider the Menopause in the model or should we drop it away? Indeed, the  $p = 0.10$  can be seen as a 'borderline' value. Nevertheless, it is immediate to check that dropping away the Menopause from the model, the Akaike Information Criterion starts to increase (AIC = 116.0) and therefore we retain the Menopause term into the model. Clearly, we do not stop here exploring, but we have to investigate over possible interaction terms.

**Exercise 8.1** Discover how to insert various interaction terms – for instance, the difficultest one:  $\text{logHE4} * \text{logCA125}$ . And verify that the minimal adequate model is purely additive. ■

Now we explain how to use those Coefficients, taking as an example the third patient:

T	logHE4	logCA125	logCA19-9	logCEA	AgePatient	Menopause	Histology
1	3.58	4.25	3.33	0.22	34	ante	benign
2	3.42	5.45	4.84	0.24	21	ante	benign
3	5.68	4.72	3.2	0.92	64	post	malignant
4	4.14	3.96	3.54	1.76	58	post	malignant

Substituting the patient information into the model Estimate we evaluate the linear relation, obtaining the **predictive index, PI**:

$$y = P.I. = -14.377 + 2.338 \cdot 5.68 + 0.684 \cdot 4.72 + 0.938 \approx 3.07$$

Now, inverting the **logit** transformation, we obtain the link function, the **logistic** function with its peculiar sigmoidal behaviour:

$$p = \frac{e^{P.I.}}{1 + e^{P.I.}}$$

In the present example, the third patient had an estimated probability of being malignant  $p = \frac{e^{P.I.}}{1 + e^{P.I.}} = \frac{e^{3.07}}{1 + e^{3.07}} = 0.96 = 96\%$  (remember, R adopts the alphabetical order, therefore in the binomial (correctly, Bernoulli) random variable Histology benign is 0 and malignant is 1).

Also in the generalized linear model it is possible to make some diagnostic; to go into detail we recommend reading the milestone-book of Julian Faraway [21].

**Discussion 8.2.1 — overdispersion.** Another possible issue in estimating generalized linear models is represented by the phenomenon of **overdispersion**. Remember that the normal distribution depends on two 'free' parameters, the mean  $\mu$  and the standard deviation  $\sigma$ ; on the contrary, in binomial distribution ( $\text{mean} = n \cdot p \equiv \text{variance}/(1 - p)$ ) and in Poisson distribution ( $\text{mean} = \lambda \equiv \text{variance}$ ) variance and mean are algebraically related in a fixed manner: as a consequence the residual deviance has an implicit relation with the dimension  $n$  of the dataset, and therefore with the degrees of freedom of the model. Practically, one should check if the residual deviance is *less* than the model degrees of freedom. With the R language this is very easy to verify. And in JASP? Try to discover a possible strategy (*hint: squared Pearson residuals plot*).

## 9. Conclusions

Time has gone and our course has ended. But many other important topics should be covered. For instance, in JASP the **survival analysis** (the Kaplan-Meier curves, the Cox regression, ...) is not currently implemented. If you are interested, you can have a look to those general introductory surveys:



Viv Bewick, Liz Cheek, Jonathan Ball. Statistics review 12: Survival analysis  
<https://ccforum.biomedcentral.com/articles/10.1186/cc2955>



Isabella Zwiener, Maria Blettner, Gerhard Hommel. Survival Analysis  
<https://www.aerzteblatt.de/int/archive/article/81182>

Of course, with R this topic is well established, and there are many tutorials on the web; for instance:

- <https://www.datacamp.com/community/tutorials/survival-analysis-R>
- [https://www.emilyzabor.com/tutorials/survival\\_analysis\\_in\\_rTutorial.html](https://www.emilyzabor.com/tutorials/survival_analysis_in_rTutorial.html)
- <https://www.r-bloggers.com/steps-to-perform-survival-analysis-in-r/>

One other very important (and difficult, I say) argument concerns **longitudinal experimental design**, in which for instance we collect **repeated measures** on the same patient. We introduce the difficulty by a simple didactical example.

Alice and Ellen are twin, and they have a silly question: *have Alice and Ellen the same weight?* They decide to measure themselves each day at the same time with the same dress with the same weight scale, and the first day the situation is: Alice, 73.60; Ellen, 73.80. So, *have Alice and Ellen the same weight?* Well, **no**, from a pure mathematical point of view.

Alice and Ellen decide to annotate their weight, repeating for five days the same experiment. Let copy and paste into the R in JASP window the proper code:

```
alice = c(73.6, 73.4, 74.1, 73.5, 73.2)
ellen = c(73.8, 73.5, 74.6, 73.8, 73.6)
t.test(alice, ellen, var.equal = TRUE)
```

```
Two Sample t-test

data: alice and ellen
t = -1.2227, df = 8, p-value = 0.2562
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-0.865794 0.265794
sample estimates:
mean of x mean of y
73.56    73.86
```

Now, the mean weight of Alice is 73.56; and 73.86 for Ellen. The difference is about 0.30 Kg and it is not a significant difference ( $t = -1.2227$ ,  $df = 8$ ,  $p\text{-value} = 0.2562$ ). The common sense suggests that for Alice it would be enough to drink a glass of water to change the situation, and so our mind will be driven to decide that **yes**, Alice and Ellen have the same weight in a statistical sense.

Twist of fate: repeating for three weeks the measures, as reported in the table below, the 0.28 Kg difference becomes significant ( $t = -2.4594$ ,  $df = 40$ ,  $p\text{-value} = 0.01834$ ), and therefore **no**, Alice and Ellen have not the same weight in a statistical sense.

```
alice = c(73.6, 73.4, 74.1, 73.5, 73.2, 74.0, 73.6, 73.3, 74.2, 73.6,
73.4, 74.1, 73.6, 73.4, 74.1, 73.5, 73.2, 74.0, 73.6, 73.3, 74.2)
ellen = c(73.8, 73.5, 74.6, 73.8, 73.6, 74.4, 73.8, 73.5, 74.3, 73.9,
73.6, 74.6, 73.8, 73.6, 74.4, 73.7, 73.5, 74.4, 73.9, 73.6, 74.5)
t.test(alice, ellen, var.equal = TRUE)
```

```
Two Sample t-test

data: alice and ellen
t = -2.4594, df = 40, p-value = 0.01834
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-0.51183215 -0.05007261
sample estimates:
mean of x mean of y
73.66190 73.94286
```

Very strange? Well, the explanation is that Alice and Ellen's weights are represented by a time series, as it happened in the `airquality` dataset; but in the latter, the measurements appeared to be rather uncorrelated, while the Alice weight time series is obviously compounded by **correlated data** [54] (and the same obviously occurs for Ellen): it is natural to expect that tomorrow's Alice weight will resemble the current value.

The proper tools to manage in **JASP** these kind of data are the **linear mixed effects models** [21, 57], in which the pseudoreplication is managed adding a further random effect. In some easy situations one can resort the Repeated Measures ANOVA approach.



## Bibliography

### Articles

- [1] Hirotugu Akaike. “A new look at the statistical model identification”. In: *Automatic Control, IEEE Transactions on* 19.6 (1974), pages 716–723 (cited on page 76).
- [2] Douglas G Altman and J Martin Bland. “Statistics notes: Absence of evidence is not evidence of absence”. In: *Bmj* 311.7003 (1995), page 485 (cited on page 51).
- [3] DR Anderson and K Burnham. “Model selection and multi-model inference”. In: *Second. NY: Springer-Verlag* (2004) (cited on page 75).
- [4] Edgar Anderson. “The species problem in Iris”. In: *Annals of the Missouri Botanical Garden* 23.3 (1936), pages 457–509 (cited on page 16).
- [5] Francis J Anscombe. “Graphs in statistical analysis”. In: *The American Statistician* 27.1 (1973), pages 17–21 (cited on page 69).
- [6] Viv Bewick, Liz Cheek, and Jonathan Ball. “Statistics review 13: receiver operating characteristic curves”. In: *Critical care* 8.6 (2004), page 508 (cited on page 37).
- [11] Maria Vittoria Caruso, Attilio Renzulli, and Gionata Fragomeni. “Influence of IABP-Induced abdominal occlusions on aortic hemodynamics: a patient-specific computational evaluation”. In: *ASAIO Journal* 63.2 (2017), pages 161–167 (cited on page 15).
- [12] Emanuela Chiarella et al. “ZNF521 Represses Osteoblastic Differentiation in Human Adipose-Derived Stem Cells”. In: *International journal of molecular sciences* 19.12 (2018), page 4095 (cited on page 15).
- [17] Maria Teresa De Angelis et al. “Short-term retinoic acid treatment sustains pluripotency and suppresses differentiation of human induced pluripotent stem cells”. In: *Cell death & disease* 9.1 (2018), page 6 (cited on pages 15, 24).
- [18] Annalisa Di Cello et al. “A more accurate method to interpret lactate dehydrogenase (LDH) isoenzymes’ results in patients with uterine masses”. In: *European Journal of Obstetrics and Gynecology and Reproductive Biology* in press (2019). DOI: 10.1016/j.ejogrb.2019.03.017. URL: 10.1016/j.ejogrb.2019.03.017 (cited on page 15).

- [19] Johnny van Doorn et al. “Bayesian rank-based hypothesis testing for the rank sum test, the signed rank test, and Spearman’s  $\rho$ ”. In: *Journal of Applied Statistics* 47.16 (2020), pages 2984–3006 (cited on page 59).
- [20] Bradley Efron. “Student’s t-test under symmetry conditions”. In: *Journal of the American Statistical Association* 64.328 (1969), pages 1278–1302 (cited on page 57).
- [23] Ronald A Fisher. “The use of multiple measurements in taxonomic problems”. In: *Annals of eugenics* 7.2 (1936), pages 179–188 (cited on page 16).
- [24] Ronald Aylmer Fisher et al. “The design of experiments.” In: *The design of experiments*. 2nd Ed (1937) (cited on page 48).
- [25] Gregg C Fonarow et al. “An obesity paradox in acute heart failure: analysis of body mass index and inhospital mortality for 108 927 patients in the Acute Decompensated Heart Failure National Registry”. In: *American heart journal* 153.1 (2007), pages 74–81 (cited on page 32).
- [26] Francis Galton. “Regression towards mediocrity in hereditary stature.” In: *The Journal of the Anthropological Institute of Great Britain and Ireland* 15 (1886), pages 246–263 (cited on page 65).
- [27] MA Goss-Sampson, J van Doorn, and EJ Wagenmakers. “Bayesian inference in JASP: A guide for students”. In: *University of Amsterdam: JASP team* (2020) (cited on page 52).
- [28] John M Hoenig and Dennis M Heisey. “The abuse of power: the pervasive fallacy of power calculations for data analysis”. In: *The American Statistician* 55.1 (2001), pages 19–24 (cited on page 50).
- [29] Torsten Hothorn, Frank Bretz, and Peter Westfall. “Simultaneous Inference in General Parametric Models”. In: *Biometrical Journal* 50.3 (2008), pages 346–363 (cited on page 63).
- [30] Petteri Hovi et al. “Glucose regulation in young adults with very low birth weight”. In: *New England Journal of Medicine* 356.20 (2007), pages 2053–2063 (cited on page 22).
- [33] John PA Ioannidis. “Why most published research findings are false”. In: *PLoS medicine* 2.8 (2005), e124 (cited on page 49).
- [34] Harold Jeffreys. “An invariant form for the prior probability in estimation problems”. In: *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences* 186.1007 (1946), pages 453–461 (cited on page 52).
- [35] Brian L Joiner. “Living histograms”. In: *International Statistical Review/Revue Internationale de Statistique* (1975), pages 339–340 (cited on page 30).
- [37] Charles J Kowalski. “Non-normal bivariate distributions with normal marginals”. In: *The American Statistician* 27.3 (1973), pages 103–106 (cited on page 31).
- [38] Edward L Melnick and Aaron Tenenbein. “Misspecifications of the normal distribution”. In: *The American Statistician* 36.4 (1982), pages 372–373 (cited on page 31).
- [39] Richard G Moore et al. “Comparison of a novel multiple marker assay vs the Risk of Malignancy Index for the prediction of epithelial ovarian cancer in patients with a pelvic mass”. In: *American journal of obstetrics and gynecology* 203.3 (2010), 228–e1 (cited on pages 29, 80).
- [41] Stefano Panzeri, Cesare Magri, and Ludovico Carraro. “Sampling bias”. In: *Scholarpedia* 3.9 (2008), page 4258 (cited on page 38).
- [42] Kersti Pärna et al. “Alcohol consumption in Estonia and Finland: Finbalt survey 1994-2006”. In: *BMC Public Health* 10.1 (2010), pages 1–13 (cited on page 22).

- 
- [46] Peter J Rousseeuw, Ida Ruts, and John W Tukey. “The bagplot: a bivariate boxplot”. In: *The American Statistician* 53.4 (1999), pages 382–387 (cited on page 22).
- [47] Christer Sinderby et al. “An automated and standardized neural index to quantify patient-ventilator interaction”. In: *Critical Care* 17.5 (2013), pages 1–9 (cited on page 23).
- [49] Student. “The probable error of a mean”. In: *Biometrika* (1908), pages 1–25 (cited on pages 45, 46, 54).
- [50] Nick Thieme. “R generation”. In: *Significance* 15.4 (2018), pages 14–19 (cited on page 10).
- [51] Sundri P Vaswani. “A pitfall in correlation theory”. In: *Nature* 160 (1947), pages 405–406 (cited on page 31).
- [53] Roberta Venturella et al. “Three to five years later: long-term effects on ovarian function of prophylactic bilateral salpingectomy”. In: *Journal of Minimally Invasive Gynecology* 24.1 (2017). PMID: 27621194, pages 145–150 (cited on page 15).
- [55] Howard Wainer. “The most dangerous equation”. In: *American Scientist* 95.3 (2007), page 249 (cited on page 42).
- [56] Ronald L Wasserstein, Nicole A Lazar, et al. “The ASA’s statement on p-values: context, process, and purpose”. In: *The American Statistician* 70.2 (2016), pages 129–133 (cited on page 49).
- [59] P Zaffino et al. “Radiotherapy of Hodgkin and non-Hodgkin lymphoma: A nonrigid image-based registration method for automatic localization of prechemotherapy gross tumor volume”. In: *Technology in cancer research & treatment* 15.2 (2016), pages 355–364 (cited on page 15).
- [60] Achim Zeileis. “Econometric computing with HC and HAC covariance matrix estimators”. In: (2004) (cited on page 63).
- [61] Stephen T Ziliak and Deirdre N McCloskey. “The cult of statistical significance”. In: *Ann Arbor: University of Michigan Press* 27 (2008) (cited on pages 45, 49).

## Books

- [7] Martin Bland. *An introduction to medical statistics*. Ed. 3. Oxford University Press, 2000 (cited on pages 17, 30, 41).
- [8] Joseph K Blitzstein and Jessica Hwang. *Introduction to probability*. Chapman and Hall/CRC, 2019 (cited on page 18).
- [9] Frank Bretz, Torsten Hothorn, and Peter Westfall. *Multiple comparisons using R*. CRC Press, 2010 (cited on page 63).
- [10] Kenneth P Burnham and David R Anderson. *Model selection and multimodel inference: a practical information-theoretic approach*. Springer Science & Business Media, 2003 (cited on page 66).
- [13] Gerda Claeskens and Nils Lid Hjort. *Model selection and model averaging*. Cambridge University Press, 2008 (cited on page 76).
- [14] William Jay Conover. *Practical nonparametric statistics*. Wiley New York, 1980 (cited on page 58).
- [15] Michael J Crawley. *The R book*. John Wiley & Sons, 2012 (cited on pages 41, 55, 62, 70, 76).

- [21] Julian J Faraway. *Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models*. Chapman and Hall/CRC, 2016 (cited on pages 75, 80, 82, 84).
- [22] Julian J Faraway. *Linear models with R*. Chapman and Hall/CRC, 2016 (cited on page 67).
- [31] Sergio Invernizzi. *Matematica nelle Scienze Naturali*. Trieste: Edizioni Goliardiche, 1996. ISBN: 8886573170 (cited on page 22).
- [32] Sergio Invernizzi, Maurizio Rinaldi, and Federico Comoglio. *Moduli di matematica e statistica – Con l’uso di R*. Zanichelli, 2018 (cited on pages 21, 30).
- [36] Michael C Joiner and Albert Van der Kogel. *Basic clinical radiobiology*. Volume 1. CRC press, 2016 (cited on page 33).
- [40] Richard F Mould. *Introductory medical statistics*. CRC Press, 1998 (cited on pages 19, 20, 37, 49, 51, 57, 62, 67).
- [43] Dalgaard Peter. *Introductory statistics with R*. Springer Verlag New York Inc, 2002 (cited on page 62).
- [44] Vijay K Rohatgi. *Statistical inference*. Jonh Wiley & Sons, 1984 (cited on pages 18, 33, 61).
- [45] Bernard A Rosner. *Fundamentals of biostatistics*. Duxbury Press, 1995 (cited on pages 27, 29, 30).
- [48] Stephen M Stigler. *The history of statistics: The measurement of uncertainty before 1900*. Harvard University Press, 1986 (cited on page 67).
- [52] William N. Venables and Brian D. Ripley. *Modern Applied Statistics with S*. Fourth. ISBN 0-387-95457-0. New York: Springer, 2002. URL: <http://www.stats.ox.ac.uk/pub/MASS4> (cited on page 28).
- [54] Geert Verbeke and Geert Molenberghs. *Linear mixed models for longitudinal data*. Springer Science & Business Media, 2000 (cited on page 84).
- [57] Brady T West, Kathleen B Welch, and Andrzej T Galecki. *Linear mixed models: a practical guide using statistical software*. CRC Press, 2014 (cited on page 84).
- [58] Hadley Wickham and Garrett Grolemund. *R for data science: import, tidy, transform, visualize, and model data*. O'Reilly Media, Inc., 2016 (cited on page 10).

# Index

- Ancova, 71
  - with interaction, 73
  - without interaction, 73
- Anova, 60
  - one-way, 61
- Bayes factor, 37, 54
- bias, 38
- Bonferroni
  - correction, 62
- correlation
  - coefficient of, 66
- cut-off, 21
- cycle
  - for, 40
- data
  - factor, 17
  - nominal, 17
  - numeric, 17
  - quantitative, 17
- dataset
  - airquality, 84
  - cholesterol, 40
  - fresher, 60
  - iris, 16, 17, 77
  - roma, 29, 32, 35, 50, 55, 79, 80
  - sleep, 54
  - balanced, 19
  - complete, 19
- degree of freedom, 18
- design
  - cross section, 24
  - longitudinal, 48, 83
  - transversal, 24
- determination
  - coefficient of, 66
- distribution
  - binomial, 32
  - lognormal, 31
  - normal, 28
  - Poisson, 33
- effect
  - fixed, 68
  - random, 68
- entropy, 63, 76
  - Boltzmann, negative, 66
  - relative, 76
- error
  - root mean square, 68
  - type-I alpha, 50
  - type-II beta, 50
- events
  - associated, 36
  - independent, 36
- factor
  - levels, 17, 77
- frequence

- relative, 35
- heteroskedasticity, 57
- histogram, 21
  - bins, 21
  - classes, 21
  - relative frequencies, 21
- homoskedasticity, 57
- hypothesis
  - null, 46
- inclusion criteria, 35
- information
  - Akaike criterion, 76
  - Kullback - Leibler measure of, 66, 76
- kurtosis, 22
- likelihood, 76
  - ratio, 37
- linear predictor, 79
- link function, 79
- logistic, 80, 81
- logit, 80, 81
- matrix
  - matrix, 34
- measures
  - association, 36
  - central tendency, 18
  - dispersion, 18
  - location, 16, 18
  - reliability, 41
  - shape, 18
  - variability, 16
- missing values, 19
- mode, 19
- model
  - generalized linear, 79
  - linear, 65
- notation
  - Wilkinson and Rogers, 72
- odds ratio, 36
- outliers, 22
- overdispersion, 82
- p-value, 48, 54
- plot
  - bagplot, 22
  - barplot, 20
- diagnostic, 70
- dignostic, 70
- dot plot, 20
- dynamite, 42
- histogram, 20
- pie chart, 20
- pizza, 52
- quantile - quantile, 29
- scatter, 22
- population, 18, 38
- power, 50
- predictive index, 80, 81
- predictive value
  - negative, 37
  - positive, 37
- predictor, 65
- prevalence, 35
- probability
  - conditional, 35
  - frequentist, 35
  - marginal, 35
  - posterior, 36
  - prior, 36
- proxy, 35
- quantiles, 22
  - deciles, 22
  - interquartile range, 22
  - percentiles, 22
  - quartiles, 22
- random variable
  - bivariate normal, 31
  - event, 27
  - finite, 27
  - infinite, 27
  - normal, 28
  - Poisson, 33
- ratio
  - signal to noise, 45
- regression, 65
  - multivariable, 71
  - multivariate, 71
- relative risk, 36
- repeated measures, 83
- residual
  - influence, 82
  - residuals, 82
  - regression line, 67
- sample, 18, 38

mean, 41  
sampling  
    random, 38  
scoring  
    Fisher, 80  
sensitivity, 37  
significance  
    clinical, 51  
    statistical, 50  
skewness, 22  
specificity, 37  
standard deviation, 18  
standard error of the mean, 39  
standardization, 73  
statistic  
    t, 45, 68  
stochastic component, 68  
survival, 83  
  
table