Statistical Data Analysis

Alejandro Caceres

2024-08-30

Contents

Chapter 1

About

This book serves as the course material for statistical data analysis (SDA). The course is designed for students enrolled in the Master of Multidisciplinary Research in Experimental Sciences program at the Barcelona Institute of Science and Technology (BIST).

The primary objective of the course is to provide students with a comprehensive and unified understanding of statistical thinking. Emphasis is placed on both conceptual and know-how levels of comprehension. Students are expected to have prior exposure to basic statistical concepts at the degree level and/or related mathematical courses.

Statistics is expressed in mathematical language and implemented using programming languages. While we will utilize both languages, this is not a mathematics or programming course. The main goal is to grasp the application of statistical thinking and its solutions in scientific inquiry. Having clarified this, the course aims to help develop a proficient use of both mathematical concepts and programming functions without focusing on theorem demonstrations or code compilations.

Our focus lies in understanding nature, and we need the right prose to effectively communicate about it.

1.1 How

The course is typically divided into **theory** and **practical** classes (Bootcamps). The classes on theory are subdivided into statistics (the present material) and complemented with introduction to machine learning, and Bayesian inference.

The statistics theory is covered in three weeks of intense work. They comprise a total of 30 hours: 24 plenary lectures (24 hours) divided in

1. Descriptive statistics and probability (4 days)

2. Group work session 1 (3 hours)

3. Inference (4 days)

4. Group work session 2 (3 hours)

Evaluation: There will be a written exam (60%) on the **31st of October** and a report of group practicals (40%). These grades will constitute the 64% of the SDA grade. Machine learning (18%) and Bayesian statistics (18%) modules comprise the rest.

Evaluation objectives:

1. Knowledge on how to describe and interpret data in tables and figures using Python.

2. Understanding of joint probability, conditional probability and Bayes theorem. Ability to read off information from contingency and conditional tables.

3. Knowledge on how to identify probability models and ability to compute their probabilities in Python.

4. Understanding the distributions of the sample mean and sample sum, and the concept of estimators. knowledge of the central limit theorem and when to apply it.

5. Understanding and know how to compute confidence intervals in python.

6. Ability to formulate hypothesis tests in different situations. Interpretation of the p-value.

7. Ability to perform and interpret the results using Python commands and output for different statistical tests depending of the type of data.

The theory test will combine multiple choice questions and the solution of exercises.

1.2 Schedule

Statistics theory classes are shown in the orange blocks with the chapters within.

Schedule	SEPTEMBER						
	09	10	11	12	13	14	15
9:00-12:00		MMRES Annual Ceremony	Holiday	SDA Bootcamp	SDA Chp2-3		
	16	17	18	19	20	21	22
9:00-12:00	SDA Chp 4-7	SDA Bootcamp	SDA Chp 8-9	SDA Bootcamp	SDA Chp 10-11		
12:30-14:30	SAR	RRSC	SAR	RRSC			
	23	24	25	26	27	28	29
9:00-12:00	SDA Group Work	Holiday	SDA Chp 13-14	SDA Bootcamp	SDA Chp 15-16		
12:30-14:30	SAR		SAR	RRSC			
	30						
9:00-12:00	SDA Chp 17						
12:30-14:30	SAR						
	OCTOBER						
		1	2	3	4	5	6
9:00-12:00		RRSC	SDA Presentations	RRSC	SDA Chp18		
12:30-14:30	SAR	SDA Bootcamp	SAR	SDA Bootcamp			
	7	8	9	10	11	12	13
9:00-12:00	SDA	SDA	SDA	SDA			
12:30-14:30	SAR	RRSC	SAR	RRSC			
	14	15	16	17	18	19	20
9:00-12:00	SDA	SDA	SDA				
12:30-14:30	RRSC	RRSC	RRSC				

1.3 Recommended reading list

If further reading is desired for more applications with good mathematical background, I suggest:

- Douglas C. Montgomery and George C. Runger. "Applied Statistics and Probability for Engineers" 4th Edition. Wiley 2007.

If you are looking for mathematical demonstrations but keeping an eye on applications:

- Irwin Miller and Marylees Miller. "John E. Freund's Mathematical Statistics" 8th Edition.

Chapter 2

Data description

In this chapter, we will introduce tools for describing data.

We will do so using tables, figures, and descriptive statistics of central tendency and dispersion.

We will also introduce key concepts in statistics such as random experiments, observations, outcomes, and absolute and relative frequencies.

2.1 Scientific method

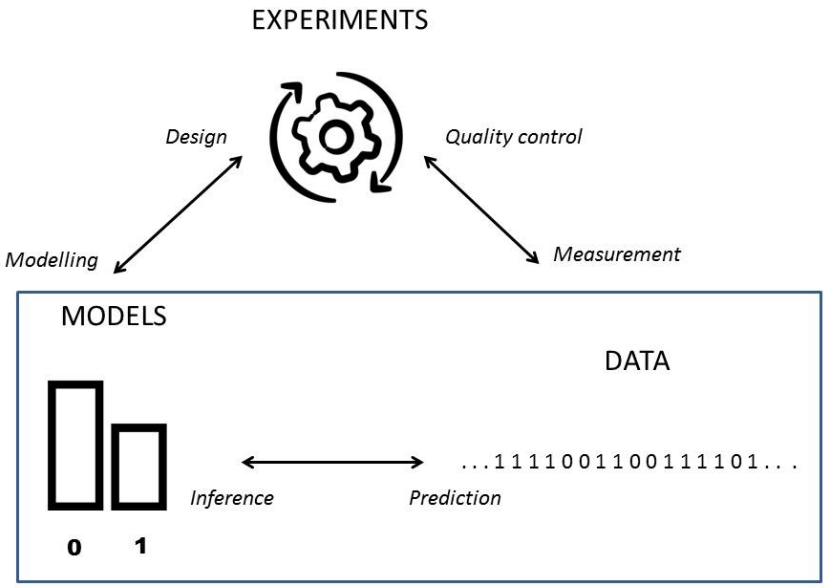
One of the goals of the scientific method is to provide a framework for solving problems that arise in the study of natural phenomena or in the design of new technologies.

Modern humans have developed a **method** over thousands of years that is still in development.

The method has three main human activities:

- *Observation* characterized by the acquisition of **data**
- *Reason* characterized by the development of mathematical **models**
- *Action* characterized by the development of new **experiments** (technology)

Their complex interaction and results are the basis of *scientific activity*.



Statistics deals with the interaction between *models* and *data* (the bottom part of the figure).

The statistical questions are:

- What is the best model for my data (inference)?
- What are the data that a certain model (prediction) would produce?

2.2 Data

The data is presented in the form of observations.

An **observation** or **realization** is the acquisition of a number or characteristic from an experimental run.

For example, let's take the series of numbers produced by repeating an experiment (1: success, 0: failure).

… 1 0 0 1 0 **1** 0 1 1 …

The number in bold is **an observation** from a repeat of the experiment.Remember that the observation is **concrete** is the number you get one day in the laboratory. An observation is a particular entity.

By contrast, an **outcome** is a **possible** observation that could result from the experiment.

In the example, **1** is one possible result, **0** is the other possible result of the experiment. The outcome of an experiment is **abstract**. You can obtain them by reasoning, no need to go to the lab. As such, the outcomes are characteristics of the experiments you are running. An outcome is a universal entity.

2.3 Result types

In statistics we are mainly interested in two types of outcomes.

- **Categorical**: If the outcome of an experiment is a quality. They can be nominal (binary: yes, no; multiple: colors) or ordinal, when the qualities can be ranked (severity of a disease, emergency grades).
- **Numeric**: If the outcome of an experiment is a number. The number can be discrete (number of emails received in an hour, number of leukocytes in the blood) or continuous (battery charge status, engine temperature).

2.4 Random experiments

We may boldly say that the subject matter of statistics is to gain knowledge from random experiments, the means by which we produce data.

Definition:

A **random experiment** is an experiment that gives a different result when repeated in the same way.

Random experiments are of different types, depending on how they are conducted:

- on the same object (person): temperature, sugar levels.
- on different objects (animals): the weight of an animal.
- on events (climate phenomena): the number of hurricanes per year.

2.5 Absolute frequencies

When we repeat a random experiment with **categorical** outcomes, we make a list of the results.

We summarize the observations by counting how many times we saw a particular outcome.

The **absolute frequency**:

$$n_i$$

is the number of times we observe the result i.

Example (leukocytes)

Consider the following random experiment. Let's extract one leukocyte from a blood sample of one donor and write down its type. Let's repeat the experiment $N = 119$ times.

```
(T cell, T cell, Neutrophil, ..., B cell)
```

The second **T cell** in bold is the second observation (extraction). The last **B cell** is observation number 119.

We can list the observations using a **frequency table** of the outcomes (categories):

```
##      outcome ni
## 1     T Cell 34
## 2     B cell 50
## 3   basophil 20
## 4   Monocyte  5
## 5 Neutrophil 10
```

The table is putting together the outcomes (first column) and the observations (second column). From the table, we can say that, for example, $n_1 = 34$ is the total number of T cells observed in the repetition of the experiment. We also note that the total number of repetitions is

$$\sum_{i=1}^{M} n_i = N = 119$$

, where M is the number of outcomes (number of rows in the table) and N the number of observations (repetitions of the random experiment). The table may be regarded as the immune profile of the donor.

2.6 Relative frequencies

We can also summarize the observations by calculating the **proportion** of how many times we saw a particular result.

$$f_i = \frac{n_i}{N}$$

In our example, $n_1 = 34$ T cells were recorded, so we can ask about the proportion of T cells from the total number of leukocytes observed: 119. We can copy these proportions f_i in the frequency table.

```
##      outcome ni         fi
## 1     T Cell 34 0.28571429
## 2     B cell 50 0.42016807
## 3   basophil 20 0.16806723
```

```
## 4   Monocyte  5 0.04201681
## 5 Neutrophil 10 0.08403361
```

Relative frequencies are **fundamental** in statistics. They give the proportion of one outcome in relation to the other outcomes. Later we will understand them as the observations of probabilities; or more accurately probability estimators.

For absolute and relative frequencies we have the properties

$$\sum_{i=1}^{M} n_i = N$$

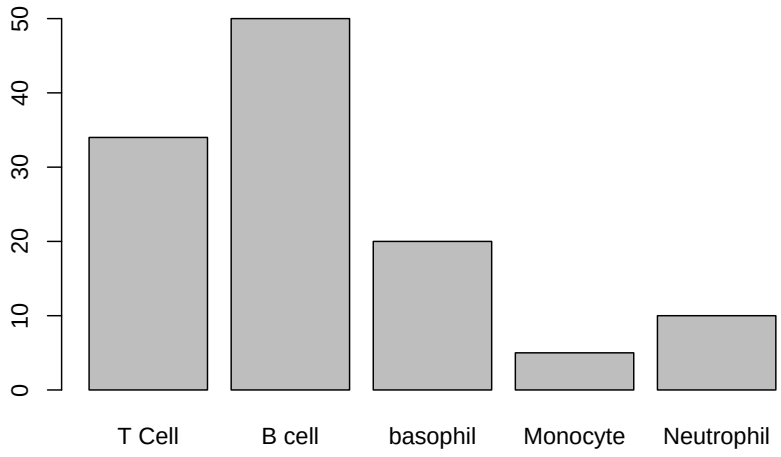
and

$$\sum_{i=1}^{M} f_i = 1$$

where M is the number of outcomes.

2.7 Bar chart

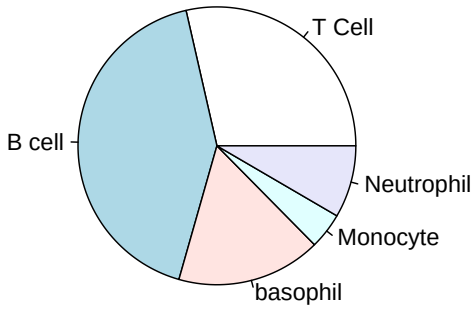
When we have a lot of outcomes and want to see which ones are most likely, we can use a bar chart. This is a plot of the frequencies n_i Vs the outcomes i.



2.8 Pie chart (pie)

We can also visualize the relative frequencies f_i using a pie chart.

In the pie chart, the area of the circle represents the 100% of the observations (proportion = 1) and the sections represent the relative frequencies of each outcome.



2.9 Ordinal categorical outcomes

The leukocyte type in the examples above is a **categorical** nominal variable. Each observation belongs to a category (quality). The categories do not always have a certain order.

Sometimes **categorical** outcomes can be **sorted** when they have a natural ranking. This allows you to compute **cumulative frequencies**.

Example (misophonia)

This is a clinical study on 123 patients who were examined for their degree of misophonia. Misophnia is uncontrolled anxiety/anger produced by particular sounds.

Each patient was evaluated with a questionnaire (AMISO) and they were classified into 4 different groups according to severity.

The results of the study are

```
##   [1] 4 2 0 3 0 0 2 3 0 3 0 2 2 0 2 0 0 3 3 0 3 3 2 0 0 0 4 2 2 0 2 0 0 0 3 0 2
##  [38] 3 2 2 0 2 3 0 0 2 2 3 3 0 0 4 3 3 2 0 2 0 0 0 2 2 0 0 2 3 0 1 3 2 4 3 2 3
##  [75] 0 2 3 2 4 1 2 0 2 0 2 0 2 2 4 3 0 3 0 0 0 2 2 1 3 0 0 3 2 1 3 0 4 4 2 3 3
## [112] 3 0 3 2 1 2 3 3 4 2 3 2
```

Each observation is the result of the run of one particular random experiment: the measurement of the level of misophonia in a patient. This data series can be summarized using the frequency table

```
##   outcome ni         fi
## 1       0 41 0.33333333
## 2       1  5 0.04065041
## 3       2 37 0.30081301
## 4       3 31 0.25203252
## 5       4  9 0.07317073
```

About 25% of the patients had misophonia grade 3. Note that the rows of table are **ordered** by the severity of the disease.

2.10 Absolute and relative cumulative frequencies

misophonia severity is **categorical ordinal** because its outcomes can be meaningfully ordered by their degree.

When the outcomes of a random experiment can be ordered, it is useful to ask how many observations were obtained up to a given outcome. We call this number the **absolute cumulative frequency** up to the outcome i, and compute it with sum the relative frequencies up to i

$$N_i = \sum_{k=1}^{i} n_k$$

It is also useful to calculate the **proportion** of observations up to a given result.

$$F_i = \sum_{k=1}^{i} f_k$$

This is called the **relative cumulative frequency**, or **frequency distribution** of the data.

We can add these frequencies to the **frequency table**

```
##   outcome ni         fi  Ni        Fi
## 0       0 41 0.33333333  41 0.3333333
## 1       1  5 0.04065041  46 0.3739837
## 2       2 37 0.30081301  83 0.6747967
```

```
## 3       3 31 0.25203252 114 0.9268293
## 4       4  9 0.07317073 123 1.0000000
```

Therefore, **67%** of patients had misophonia up to severity **2**, and **37%** of patients had severity less than or equal to **1**.

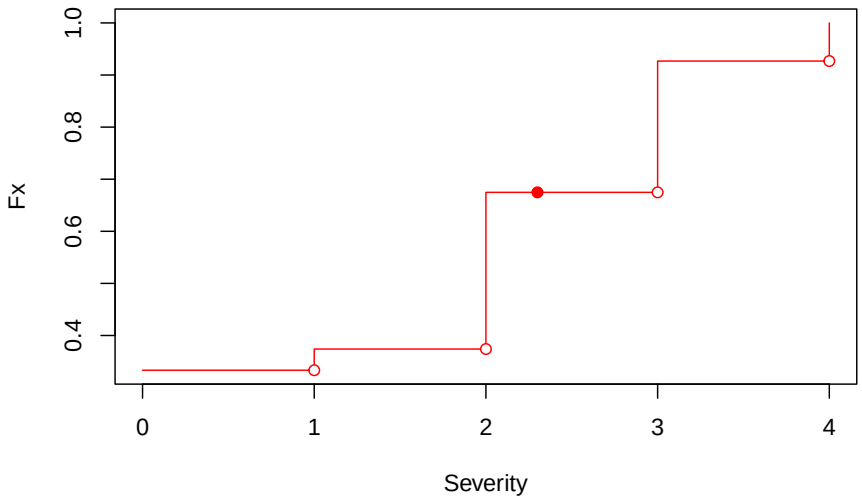
2.11 Cumulative frequency graph

F_i is an important quantity because it allows us to define the accumulation of frequencies up to **intermediate** levels of the outcomes.

The frequency up to an intermediate level x $(i \leq x < i+1)$ is just the accumulation up to the lower possible outcome of the experiment

$$F_x = F_i$$

F_x is therefore a function on a **continuous** range of values. We can draw it with respect to the outcomes



Therefore, we can say that **67%** of the patients had misophonia up to severity 2.3, although 2.3 is not an observed outcome.

2.12 Numerical outcomes

The result of a random experiment can produce a number. If the number is **discrete**, we can generate a frequency table, with absolute, relative, and cumu-

lative frequencies, and illustrate them with bar, pie, and cumulative charts.

When the outcome is **continuous** the frequency table is not so useful, because it is unlikely to obtain a repetition of the same measurement.

Example (misophonia)

The researchers wondered if the convexity of the jaw would affect the severity of misophonia. The scientific hypothesis is that the angle of convexity of the jaw can influence hearing and its sensitivity. These are the mandibular convexity observations in degrees for each patient:

```
##   [1]  7.97 18.23 12.27  7.81  9.81 13.50 19.30  7.70 12.30  7.90 12.60 19.00
##  [13]  7.27 14.00  5.40  8.00 11.20  7.75  7.94 16.69  7.62  7.02  7.00 19.20
##  [25]  7.96 14.70  7.24  7.80  7.90  4.70  4.40 14.00 14.40 16.00  1.40  9.76
##  [37]  7.90  7.90  7.40  6.30  7.76  7.30  7.00 11.23 16.00  7.90  7.29  6.91
##  [49]  7.10 13.40 11.60 -1.00  6.00  7.82  4.80 11.00  9.00 11.50 16.00 15.00
##  [61]  1.40 16.80  7.70 16.14  7.12 -1.00 17.00  9.26 18.70  3.40 21.30  7.50
##  [73]  6.03  7.50 19.00 19.01  8.10  7.80  6.10 15.26  7.95 18.00  4.60 15.00
##  [85]  7.50  8.00 16.80  8.54  7.00 18.30  7.80 16.00 14.00 12.30 11.40  8.50
##  [97]  7.00  7.96 17.60 10.00  3.50  6.70 17.00 20.26  6.64  1.80  7.02  2.46
## [109] 19.00 17.86  6.10  6.64 12.00  6.60  8.70 14.05  7.20 19.70  7.70  6.02
## [121]  2.50 19.00  6.80
```

while, in these set of observations, we see the repetition of the convexity degree 7.90, this is a result of measurement rounding or limited instrument resolution. Degrees are continuous and it is unlikely that to patients have the same value of convexity, if we had sufficient resolution.

2.13 Transforming continuous data

Since observations of continuous outcomes cannot be counted in frequencies (at least by definition), we may transform them into ordered categorical outcomes.

1) First we cover the range of observations in regular intervals of the same size (bins)

```
## [1] "[-1.02,3.46]" "(3.46,7.92]"  "(7.92,12.4]"  "(12.4,16.8]"  "(16.8,21.3]"
```

2) Then we map each observation to its interval: creating a categorical **ordered** outcomes; in this case we have 5 possible outcomes

```
##   [1] "(7.92,12.4]"  "(16.8,21.3]"  "(7.92,12.4]"  "(3.46,7.92]"  "(7.92,12.4]"
##   [6] "(12.4,16.8]"  "(16.8,21.3]"  "(3.46,7.92]"  "(7.92,12.4]"  "(3.46,7.92]"
##  [11] "(12.4,16.8]"  "(16.8,21.3]"  "(3.46,7.92]"  "(12.4,16.8]"  "(3.46,7.92]"
##  [16] "(7.92,12.4]"  "(7.92,12.4]"  "(3.46,7.92]"  "(7.92,12.4]"  "(12.4,16.8]"
##  [21] "(3.46,7.92]"  "(3.46,7.92]"  "(3.46,7.92]"  "(16.8,21.3]"  "(7.92,12.4]"
##  [26] "(12.4,16.8]"  "(3.46,7.92]"  "(3.46,7.92]"  "(3.46,7.92]"  "(3.46,7.92]"
##  [31] "(3.46,7.92]"  "(12.4,16.8]"  "(12.4,16.8]"  "(12.4,16.8]"  "[-1.02,3.46]"
```

```
##  [36] "(7.92,12.4]"  "(3.46,7.92]"  "(3.46,7.92]"  "(3.46,7.92]"  "(3.46,7.92]"
##  [41] "(3.46,7.92]"  "(3.46,7.92]"  "(3.46,7.92]"  "(7.92,12.4]"  "(12.4,16.8]"
##  [46] "(3.46,7.92]"  "(3.46,7.92]"  "(3.46,7.92]"  "(3.46,7.92]"  "(12.4,16.8]"
##  [51] "(7.92,12.4]"  "[-1.02,3.46]" "(3.46,7.92]"  "(3.46,7.92]"  "(3.46,7.92]"
##  [56] "(7.92,12.4]"  "(7.92,12.4]"  "(7.92,12.4]"  "(12.4,16.8]"  "(12.4,16.8]"
##  [61] "[-1.02,3.46]" "(12.4,16.8]"  "(3.46,7.92]"  "(12.4,16.8]"  "(3.46,7.92]"
##  [66] "[-1.02,3.46]" "(16.8,21.3]"  "(7.92,12.4]"  "(16.8,21.3]"  "[-1.02,3.46]"
##  [71] "(16.8,21.3]"  "(3.46,7.92]"  "(3.46,7.92]"  "(3.46,7.92]"  "(16.8,21.3]"
##  [76] "(16.8,21.3]"  "(7.92,12.4]"  "(3.46,7.92]"  "(3.46,7.92]"  "(12.4,16.8]"
##  [81] "(7.92,12.4]"  "(16.8,21.3]"  "(3.46,7.92]"  "(12.4,16.8]"  "(3.46,7.92]"
##  [86] "(7.92,12.4]"  "(12.4,16.8]"  "(7.92,12.4]"  "(3.46,7.92]"  "(16.8,21.3]"
##  [91] "(3.46,7.92]"  "(12.4,16.8]"  "(12.4,16.8]"  "(7.92,12.4]"  "(7.92,12.4]"
##  [96] "(7.92,12.4]"  "(3.46,7.92]"  "(7.92,12.4]"  "(16.8,21.3]"  "(7.92,12.4]"
## [101] "(3.46,7.92]"  "(3.46,7.92]"  "(16.8,21.3]"  "(16.8,21.3]"  "(3.46,7.92]"
## [106] "[-1.02,3.46]" "(3.46,7.92]"  "[-1.02,3.46]" "(16.8,21.3]"  "(16.8,21.3]"
## [111] "(3.46,7.92]"  "(3.46,7.92]"  "(7.92,12.4]"  "(3.46,7.92]"  "(7.92,12.4]"
## [116] "(12.4,16.8]"  "(3.46,7.92]"  "(16.8,21.3]"  "(3.46,7.92]"  "(3.46,7.92]"
## [121] "[-1.02,3.46]" "(16.8,21.3]"  "(3.46,7.92]"
```

Therefore, instead of saying that the first patient had an angle of convexity of 7.97, we say that his angle was in the interval (or **bin**) between $(7.92, 12.4]$.

No other patients had an angle of 7.97 degrees, but many had angles between $(7.92, 12.4]$.

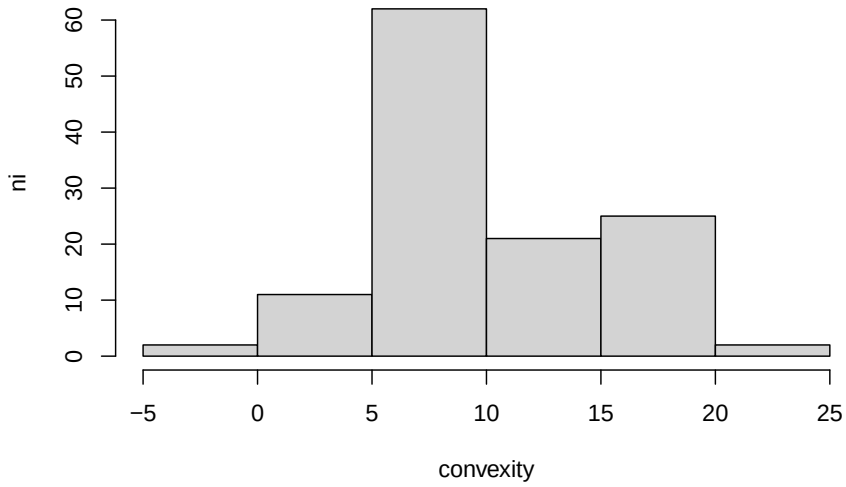
2.14 Frequency table for a continuous variable

For a given regular partition of the range of the continuous outcomes into intervals, we can produce a frequency table as before

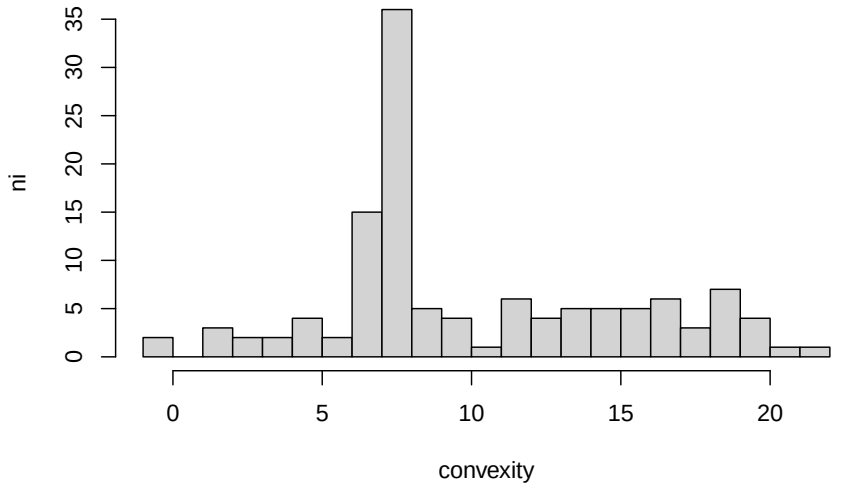
```
##        outcome ni         fi  Ni         Fi
## 1 [-1.02,3.46]  8 0.06504065   8 0.06504065
## 2  (3.46,7.92] 51 0.41463415  59 0.47967480
## 3  (7.92,12.4] 26 0.21138211  85 0.69105691
## 4  (12.4,16.8] 20 0.16260163 105 0.85365854
## 5  (16.8,21.3] 18 0.14634146 123 1.00000000
```

2.15 Histogram

The histogram is the graph of n_i or f_i Vs the interval outcomes (bins). The histogram depends on the size of the bin.



This is a histogram with 20 bins .



We see that most people have angles in the interval $(7, 8]$.

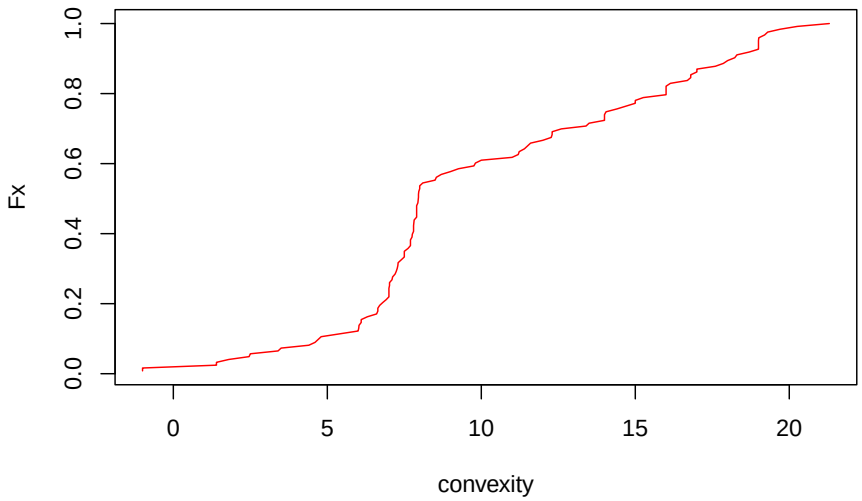
2.16 Cumulative frequency graph

We can also plot F_x against the outcomes. Since F_x has a continuous range, we can order the observations $(x_1 < ...x_j < x_{j+1} < x_n)$ and therefore

$$F_x = \frac{k}{n}$$

for $x_k \leq x < x_{k+1}$.

F_x is known as the **distribution** of the data. F_x does not depend on the size of the bin. However, its **resolution** depends on the amount of data.



2.17 Summary Statistics

Summary statistics are numbers calculated from the data that tell us important characteristics of the numerical outcomes (discrete or continuous).

For example, we have statistics that describe extreme values:

- **minimum**: the minimum observation.
- **maximum**: the maximum observation.

2.18 Average (sample mean)

An important statistic that describes the central value of the observations (where to expect most observations) is the **average**

$$\bar{x} = \frac{1}{N}\sum_{j=1}^{N} x_j$$

where x_j is the **observation** j from a total of N observations. We also call the average the **sample mean**. The sample is the set of observations, or the data obtained from the repetition of the random experiment.

Example (misophonia)

The average convexity can be calculated directly from the **observations** in the usual way

$\bar{x} = \frac{1}{N}\sum_j x_j$

$= \frac{1}{123}(7.97 + 18.23 + 12.27... + 6.80) = 10.19894$

For **categorically ordered** outcomes, we can **also** use the relative frequencies to calculate the average

$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_j = \frac{1}{N}\sum_{i=1}^{M} x_i n_i =$

$$\sum_{i=1}^{M} x_i f_i$$

where we went from adding N **observations** to adding M **outcomes**.

The form $\bar{x} = \sum_{i=1}^{M} x_i f_i$ shows that the average is the **center of mass** of the data. As if each observation had a mass density given by f_i.

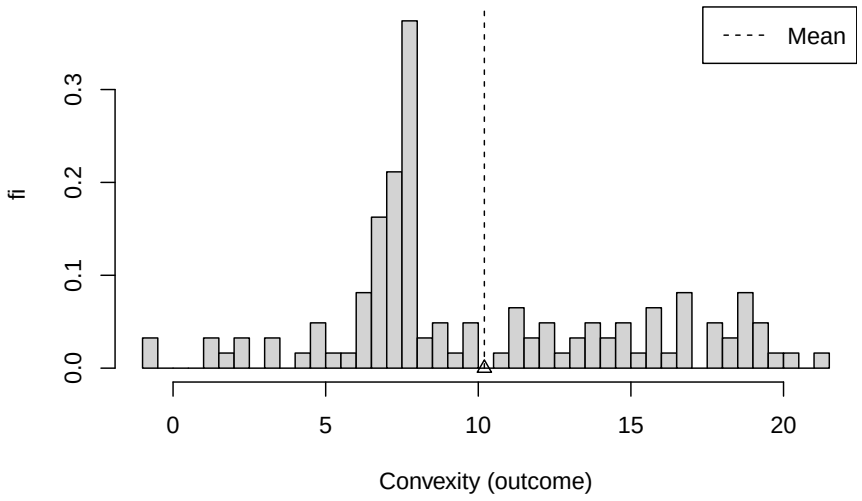
Example (misophonia)

The average **severity** of misophonia in the study can be calculated from the relative frequencies of the **outcomes**

```
##   outcome ni         fi
## 1       0 41 0.33333333
## 2       1  5 0.04065041
## 3       2 37 0.30081301
## 4       3 31 0.25203252
## 5       4  9 0.07317073
```

$\bar{x} = 0 \times f_0 + 1 \times f_1 + 2 \times f_2 + 3 \times f_3 + 4 \times f_4 = 1.691057$

The average is also the center of mass for continuous outcomes; that is, the point where the relative frequencies of the bins balance.



2.19 Median

Another measure of centrality is the median. The median x_m, or $q_{0.5}$, is the value below which we find half of the observations. When we order the observations $x_1 < ...x_j < x_{j+1} < x_N$, we count them until we find half of them, then we report that value x_m. Therefore, x_m is the observation such that m satisfies

$$\sum_{i \leq m} 1 = \frac{N}{2}$$

Example (misophonia)

If we order the angles of convexity and cut them in half, we see that 62 observations (individuals) ($N/2 \sim 123/2$) are below 7.96. The **median convexity** is therefore $q_{0.5} = x_{62} = 7.96$

```
##  [1] -1.00 -1.00  1.40  1.40  1.80  2.46  2.50  3.40  3.50  4.40  4.60  4.70
## [13]  4.80  5.40  6.00  6.02  6.03  6.10  6.10  6.30  6.60  6.64  6.64  6.70
## [25]  6.80  6.91  7.00  7.00  7.00  7.00  7.02  7.02  7.10  7.12  7.20  7.24
## [37]  7.27  7.29  7.30  7.40  7.50  7.50  7.50  7.62  7.70  7.70  7.70  7.75
## [49]  7.76  7.80  7.80  7.80  7.81  7.82  7.90  7.90  7.90  7.90  7.90  7.94
## [61]  7.95  7.96
```

```
##  [1]  7.96  7.97  8.00  8.00  8.10  8.50  8.54  8.70  9.00  9.26  9.76  9.81
## [13] 10.00 11.00 11.20 11.23 11.40 11.50 11.60 12.00 12.27 12.30 12.30 12.60
## [25] 13.40 13.50 14.00 14.00 14.00 14.05 14.40 14.70 15.00 15.00 15.26 16.00
```

```
## [37] 16.00 16.00 16.00 16.14 16.69 16.80 16.80 17.00 17.00 17.60 17.86 18.00
## [49] 18.23 18.30 18.70 19.00 19.00 19.00 19.00 19.01 19.20 19.30 19.70 20.26
## [61] 21.30
```

We thus cut the data at

```
## [1] 7.96
```

to split them in half.

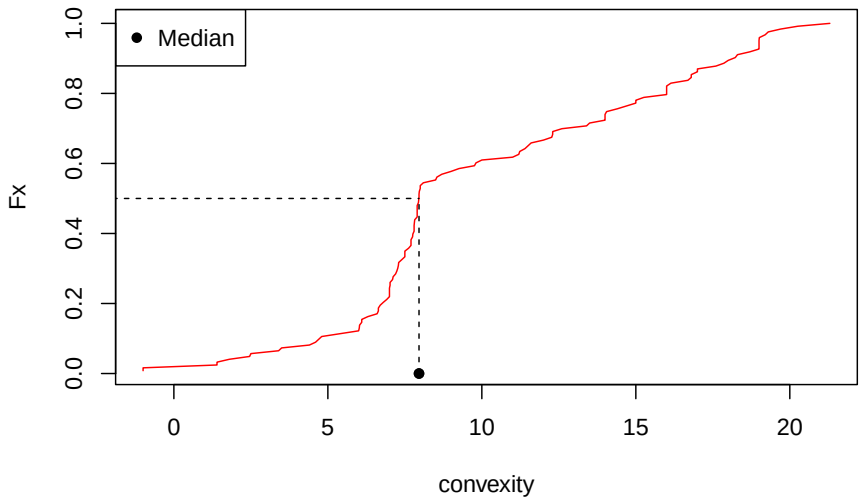
In terms of frequencies, $q_{0.5}$ makes the cumulative frequency F_x equal to 0.5

$$\sum_{i=1}^{m} f_i = F_{q_{0.5}} = 0.5$$

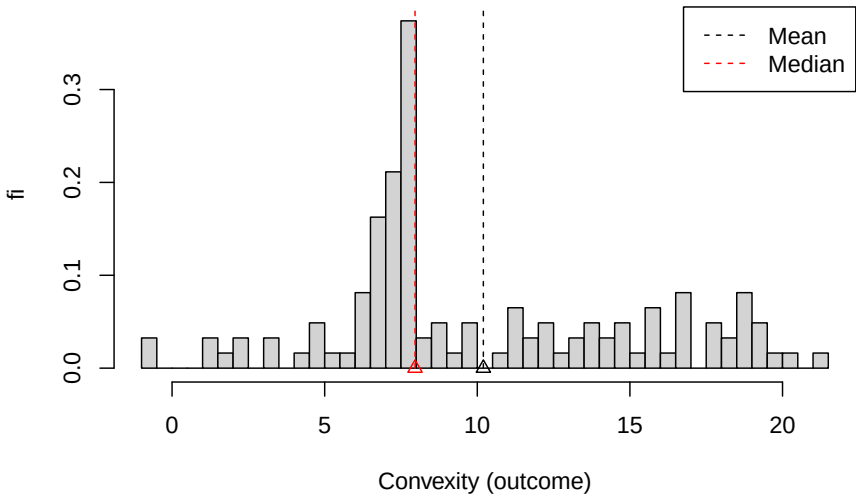
that is

$$q_{0.5} = F^{-1}(0.5)$$

This last equation means that, in the distribution graph, the median $q_{0.5}$ is the value of x at which we have climbed half of the total height of F_x. We have accumulated half of the observations.



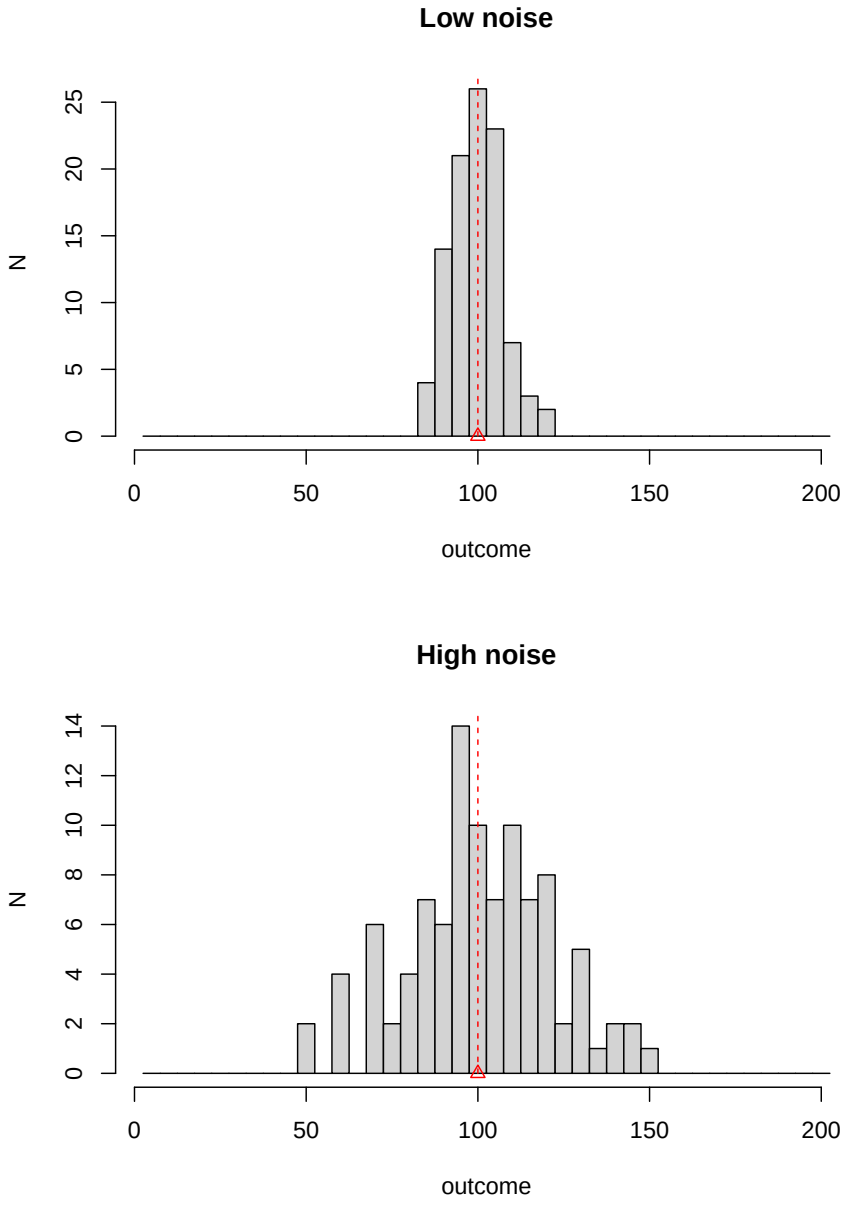
The average and median are not always the same.



2.20 Dispersion

Other important summary statistics for observations are the **spread** statistics.

Many experiments may share their average, but differ in how **sparse** the values are.



The dispersion of the observations is a measure of the **noise** or randomness of the experiment. If there is no spread, the random experiment gives always the same outcome and there is no need to learn statistics.

2.21 Sample variance

The dispersion about the average is measured by the sample variance

$$s^2 = \frac{1}{N-1}\sum_{j=1}^{N}(x_j - \bar{x})^2$$

This number measures the average squared distance of the **observations** from the average. The reason for $N-1$ will be explained when we talk about inference, when we study the spread of $\bar{x}$, as well as the spread of the observations.

In terms of the frequencies of the outcomes that are **categorical and ordered**, we can **also** calculate the sample variance as

$$s^2 = \frac{N}{N-1}\sum_{i=1}^{M}(x_i - \bar{x})^2 f_i$$

Take many observations and then make $N/(N+1)$ close to a, then s^2 can be considered as the **moment of inertia** about the average of the observations.

The square root of the sample variance, s, is called **standard deviation** of the sample.

Example (misophonia)

The standard deviation of the convexity angle is

$$s = [\tfrac{1}{123-1}((7.97 - 10.19894)^2 + (18.23 - 10.19894)^2$$

$$+(12.27 - 10.19894)^2 + ...)]^{1/2} = 5.086707$$

The jaw convexity deviates from its average by 5.086707.

2.22 Interquartile range (IQR)

The spread of the data can also be measured with respect to the median using the **interquartile range**:

1) We define the **first** quartile as the value x that makes the cumulative frequency $F_{q_{0.25}}$ equal to 0.25, or the value of x where we have accumulated a quarter of the observations, or the value that splits the first quarter of the observations.

$$q_{0.25} = F^{-1}(0.25)$$

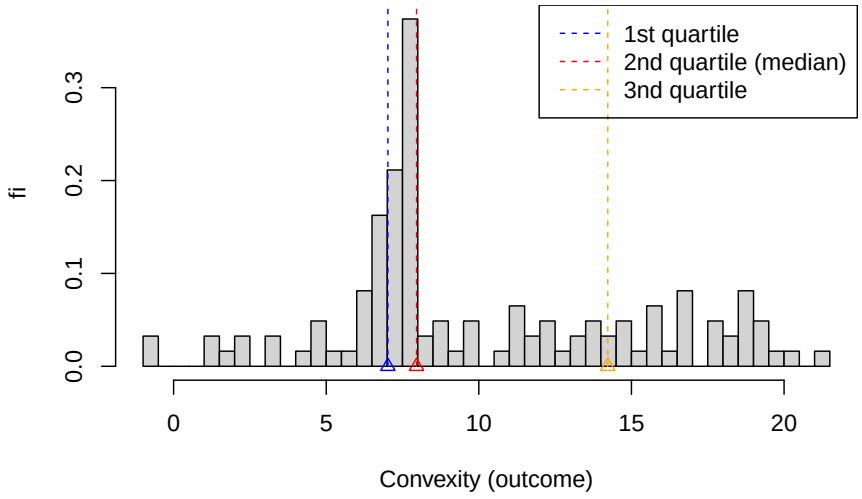
2) We define the **third** quartile as the value x that makes the cumulative frequency $F_{q_{0.75}}$ equal to 0.75, or the value of x where we have accumulated three quarters of observations.

$$q_{0.75} = F^{-1}(0.75)$$

3) The **interquartile range** (IQR) is

$$IQR = q_{0.75} - q_{0.25}$$

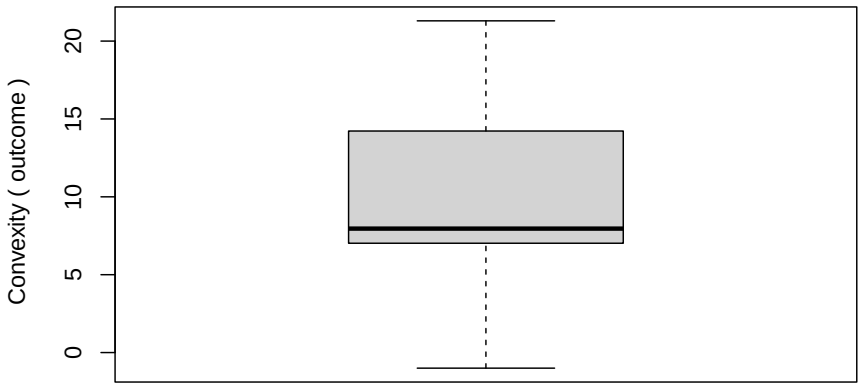
This is the distance between the third and first quartiles and captures the central 50% of the observations



2.23 Boxplot

The interquartile range, median, and 5% and 95% of the data can be displayed in a **box plot**.

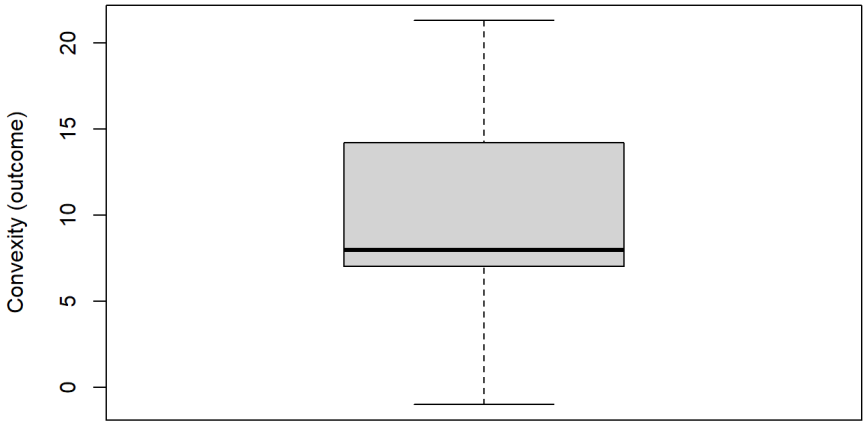
In the boxplot, the values of the outcomes are on the y-axis. The IQR is the box, the median is the middle line, and the whiskers mark the 5% and 95% of the data.



2.24 Questions

1) In the following boxplot, the first quartile and second quartile of the data are:

a: $(-1.00, 21.30)$; **b:** $(-1.00, 7.02)$; **c:** $(7.02, 7.96)$; **d:** $(7.02, 14.22)$



2) The main disadvantage of a histogram is that:

a : Depends on the size of the bin ; **b :** Cannot be used for categorical outcome; **c :** Cannot be used when the bin size is small; **d :** Used only for relative frequencies;

3) If the relative cumulative frequencies of a random experiment with outcomes $\{1, 2, 3, 4\}$ are: $F(1) = 0.15$, $F(2) = 0.60$, $F(3) = 0.85$, $F(4) = 1$.

Then the relative frequency for the outcome 3 is

a: 0.15; **b:** 0.85; **c:** 0.45; **d:** 0.25

4) In a sample of size 10 from a random experiment we obtained the following data:

8, 3, 3, 7, 3, 6, 5, 10, 3, 8.

The first quartile of the data is:

a: 3.5; **b:** 4; **c:** 5; **d:** 3

5) Imagine that we collect data for two quantities that are not mutually exclusive, for example, the gender and nationality of passengers on a flight. If we want to make a single pie chart for the data, which of these statements is true?

a : We can **only** make a nationality pie chart because it has more than two possible outcomes;

b : We can make a pie graph for a new variable marking gender **and** nationality;

c : We can make a pie chart for the variable sex **or** the variable nationality;

d : We can only choose **whether** to make a pie chart for gender **or** a pie chart for nationality.

2.25 Exercises

2.25.0.1 Exercise 1

We have performed an experiment 8 times with the following results

```
## [1]  3  3 10  2  6 11  5  4
```

Answer the following questions:

- Calculate the relative frequencies of each result.
- Calculate the cumulative frequencies of each result.
- What is the average of the observations?
- What is the median?
- What is the third quartile?
- What is the first quartile?

2.25.0.2 Exercise 2

We have performed an experiment 10 times with the following results

```
##  [1] 2.875775 7.883051 4.089769 8.830174 9.404673 0.455565 5.281055 8.924190
##  [9] 5.514350 4.566147
```

Consider 10 bins of size 1: [0,1], (1,2] ...(9,10).

Answer the following questions:

- Calculate the relative frequencies of each result and draw the histogram
- Calculate the cumulative frequencies of each result and draw the cumulative graph.
- Draw a box plot .

2.26 Practice

Load misophonia data from https://alejandro-isglobal.github.io/SDA/data/data_0.txt

1. Extract misophonia variable (Misofonia.dic)

- Do a bar plot and a pie chart

2. Extract convexity angle variable (Angulo_convexidad)

- Calculate its sample mean (average), standard deviation and make a histogram
- Calculate its median and inter-quartile range
- Draw a boxplot

Solutions

Chapter 3

Probability

In this chapter we will introduce the concept of probability from relative frequencies.

We will define the events as the elements on which the probability is applied. Composite events will be defined using set algebra.

Then we will discuss the concept of conditional probability derived from the joint probability of two events.

3.1 Random experiments

Let's remember the basic objective of statistics. Statistics deals with data that is presented in the form of observations.

- An **observation** is the acquisition of a number or characteristic from an experiment

Observations are realizations of **outcomes**.

- An **outcome** is a possible observation that is the result of an experiment.

When conducting experiments, we often get different results. The description of the variability of the results is one of the objectives of statistics.

- A **random experiment** is an experiment that gives different results when repeated in the same way.

The philosophical question behind statistics is how can we know anything at all if every time we look at an experiment, its result changes?

3.2 Measurement probability

We would like to have a measure for **how sure** we are of a particular outcome in the **future** run of a random experiment. Probabilities are statements of future actions.

We will call this measure the probability of the outcome and assign values to it:

- 0, when we are sure that the observation **will not** occur.
- 1, when we are sure that the observation **will** occur.

3.3 Classical probability

As long as a random experiment has M possible outcomes that are all **equally likely**, the probability of each i outcome is

$$P_i = \frac{1}{M}$$

.

Classical probability was defended by Laplace (1814).

Since every outcome is **equally likely** in this type of experiment, we declare complete ignorance and the best we can do is equally distribute the same probability for each outcome. In other words, there are **some random experiments** for which we have no reason to prefer one outcome than another.

Note that:

- We do not observe P_i.
- We deduce P_i from the ratio above and we don't need to carry out any experiment to know it.

Example (dice):

What is the probability that we will get 2 on the roll of a die?

$$P_2 = 1/6 = 0.166666$$

We reason that all other 5 numbers in the dice are equally likely to 2.

3.4 Relative frequencies

What about random experiments whose possible outcomes are **not** equally likely?

How then can we define the probabilities of the outcomes?

Example (random experiment)

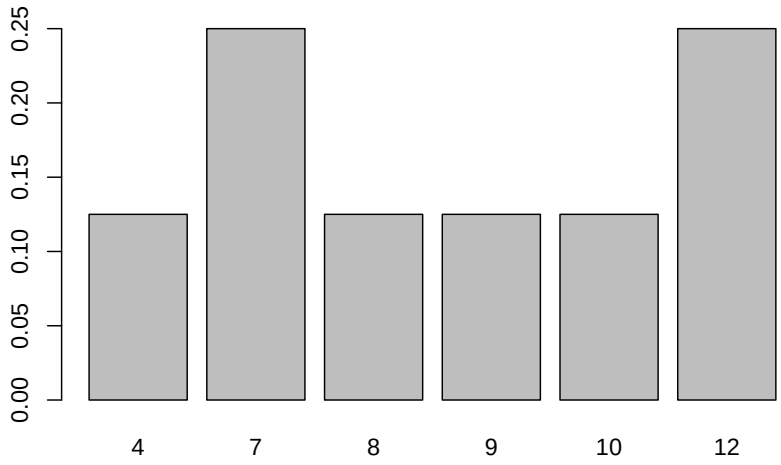
Imagine that we repeat a random experiment 8 times and obtain the following observations

8 4 12 7 10 7 9 12

- How sure are we of obtaining the result 12 from a future repetition of the random experiment that produced the data above?

The frequency table is

```
##   outcome ni    fi
## 1       4  1 0.125
## 2       7  2 0.250
## 3       8  1 0.125
## 4       9  1 0.125
## 5      10  1 0.125
## 6      12  2 0.250
```



The **relative frequency** $f_i = \frac{n_i}{N}$ seems like a reasonable probability measure because

- it is a number between 0 and 1; and
- it measures the proportion of the total number of observations that we obtained for a particular result.

Since $f_{12} = 0.25$ then we would be one quarter sure, one out of every four observations, of getting 12.

Question: How good is f_i as a measure of how sure we are of the result i?

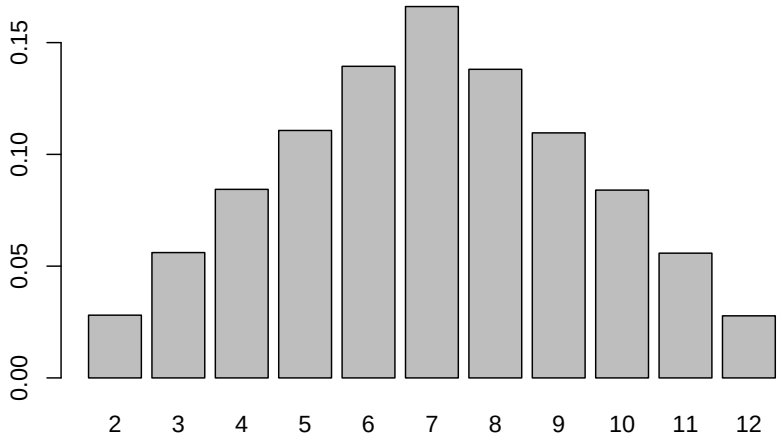
Example (random experiment with more repetitions)

Let's say we repeat the previous experiment 100,000 more times:

The frequency table is now

```
##    outcome    ni      fi
## 1        2  2807 0.02807
## 2        3  5607 0.05607
## 3        4  8435 0.08435
## 4        5 11070 0.11070
## 5        6 13940 0.13940
## 6        7 16613 0.16613
## 7        8 13806 0.13806
## 8        9 10962 0.10962
## 9       10  8402 0.08402
## 10      11  5581 0.05581
## 11      12  2777 0.02777
```

and the barplot is

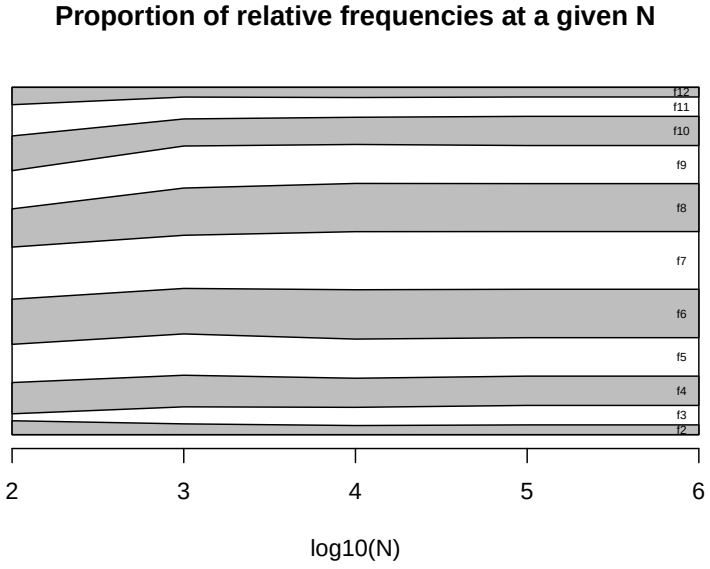


New results came out and f_{12} is now only 0.027, and so we are only $\sim 3\%$ sure to get 12 in the next experiment. The probabilities measured by f_i change with N.

3.5 Relative frequencies at infinity

A crucial fact is that when we compute f_i with increasing values of N, f_i **converges**!

In this graph each vertical section gives the relative frequency of each observation. We see that after $N = 1000$ ($log10(N) = 3$) the sections' proportions hardly change with more N.



We find that each of the relative frequencies f_i converges to a constant value

$$lim_{N\to\infty} f_i = P_i$$

Note that f_i is a quantity derived from concrete observations that in infinity becomes an abstract quantity P_i. This relationship is fundamental. It tell us that we can use experience to access general knowledge.

3.6 Frequentist probability

We call **Probability** P_i the limit as $N \to \infty$ of the **relative frequency** of observing the outcome i in a random experiment.

Defended by Venn (1876), the frequentist definition of probability is derived from (empirical) data/experience.

Note that:

- We do not observe P_i, we observe f_i; and
- **We estimate** P_i with f_i (usually when N is large), and write:

$$\hat{P}_i = f_i$$

Similar to the relationship between **observation** and **result**, we have the relationship between **relative frequency** and **probability** as a concrete value of an abstract quantity.

3.7 Classical and frequentist probabilities

We have situations where classical probability can be used to find the limit of relative frequencies:

- If the results are **equally probable**, the classical probability gives us the limit:

$$P_i = lim_{N\to\infty} \frac{n_i}{N} = \frac{1}{M}$$

- If the results in which we are interested can be derived from other **equally probable** results. We will see more about this when we study probability models.

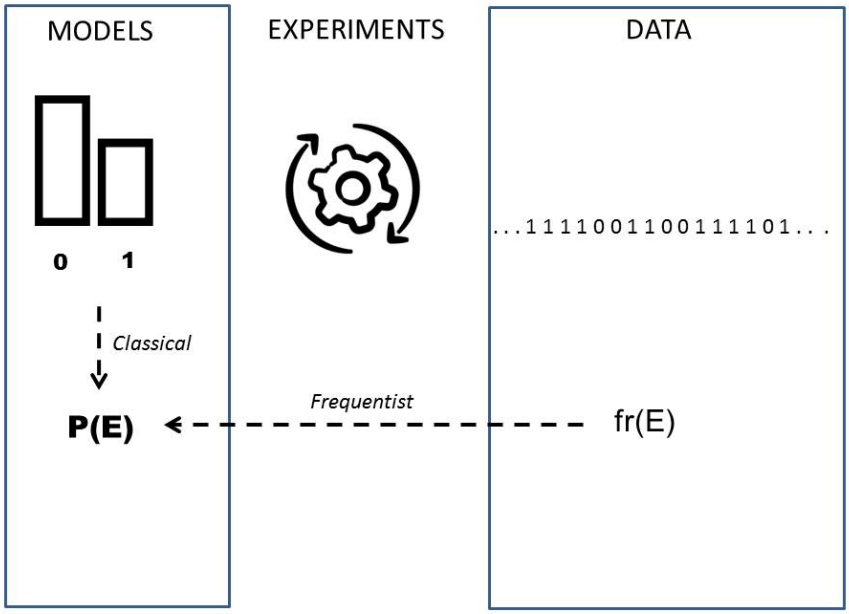
Example (sum of two dice)

Our previous example is based on the **sum of two dice** (done by a computer). Although we could perform the experiment many times, write down the results, and calculate the **relative frequencies**, we can know the exact value of the probability whitout performing any experiment.

This probability **follows** from the fact that the outcome of each die is **equally likely**. From this assumption, we can find that (Exercise 1)

$$P_i = \begin{cases} \frac{i-1}{36}, & i \in \{2,3,4,5,6,7\} \\ \frac{13-i}{36}, & i \in \{8,9,10,11,12\} \end{cases}$$

The motivation of the frequentist definition is **empirical** (data) while that of the classical definition is **rational** (models). We often combine both approaches (inference and deduction) to find out the probabilities of our random experiment.



3.8 Definition of probability

A probability is a number that is assigned to each possible outcome of a random experiment and satisfies the following properties or **axioms**:

1) when the results E_1 and E_2 are mutually exclusive; that is, only one of them can occur, so the probability of observing E_1 **or** E_1, written as $E_1 \cup E_2$, is their sum:

$$P(E_1 \cup E_2) = P(E_1) + P(E_2)$$

2) when S is the set of all possible outcomes, then its probability is 1 (at least something is observed):

$$P(S) = 1$$

3) The probability of any outcome is between 0 and 1

$$P(E) \in [0, 1]$$

Proposed by Kolmogorov's less than 100 years ago (1933), even after the formulation of statistical mechanics and quantum mechanics; two prominent physics theories based on probability concepts.

3.9 Probabilities Table

Kolmogorov properties are the basic rules for building a **probability table**, similar to the relative frequency table.

Example (dice)

The probability table for the throw of a dice

result	probability
1	1/6
2	1/6
3	1/6
4	1/6
5	1/6
6	1/6
$P(1 \cup 2 \cup ... \cup 6)$	1

Let's verify the axioms:

1) Where $1 \cup 2$ is, for example, the **event** of rolling a 1 **or** a 2. So

$$P(1 \cup 2) = P(1) + P(2) = 2/6$$

2) Since $S = \{1, 2, 3, 4, 5, 6\}$ is made up of **mutually exclusive** outcomes, then

$$P(S) = P(1 \cup 2 \cup ... \cup 6) = P(1) + P(2) + ... + P(6) = 1$$

3) The probabilities of each outcome are between 0 and 1. This can be seen in the table.

3.10 Sample space

The set of all possible outcomes of a random experiment is called the **sample space** and is denoted S.

The sample space can be made up of categorical or numerical outcomes.

For example:

- human temperature: $S = (36, 42)$ degrees Celsius.
- sugar levels in humans: $S = (70 - 80)mg/dL$
- the size of a production line screw: $S = (70 - 72)mm$
- number of emails received in an hour: $S = \{0, ...\infty\}$
- the throw of a dice: $S = \{1, 2, 3, 4, 5, 6\}$

3.11 Events

An **event** A is a **subset** of the sample space. It is a **collection** of possible results.

Examples of events:

- The event of a healthy temperature: $A = 37 - 38$ degrees Celsius
- The event of producing a screw with a size: $A = 71.5mm$
- The event of receiving more than 4 emails in an hour: $A = \{4, \infty\}$
- The event of obtaining a number less than or equal to 3 in the throw of a dice: $A = \{1, 2, 3\}$

An event refers to a possible set of **outcomes**.

3.12 Algebra of events

For two events A and B, we can construct the following **composite events** using the basic set operations:

- Complement A': the event of **not** A
- Union $A \cup B$: the event of A **or** B
- Intersection $A \cap B$: the event of A **and** B

Example (dice)

Let's imagine we intend to roll a die but first we want look at a range of interesting events (composite outcomes):

- a number less than or equal to three $A : \{1, 2, 3\}$
- an even number $B : \{2, 4, 6\}$

Let's see how we can build new events with set operations:

- a number not less than three: $A' : \{4, 5, 6\}$
- a number less than or equal to three **or** even: $A \cup B : \{1, 2, 3, 4, 6\}$
- a number less than or equal to three **and** even $A \cap B : \{2\}$

3.13 Mutually exclusive results

Outcomes like rolling 1 and 2 on a die are events that cannot occur at the same time. We say that they are **mutually exclusive**.

In general, two events denoted as E_1 and E_2 are mutually exclusive when they have no element in common

$$E_1 \cap E_2 = \emptyset$$

Examples:

The following events are mutually exclusive.

- The result of having a misophonia severity 1 and severity 4. Only one severity is possible.
- The results of obtaining 12 and 5 in the roll of two dice. If we get 12 we do not get 5.

According to the Kolmogorov's properties, only **mutually exclusive** outcomes can be arranged in **probability tables**, as in relative frequency tables.

3.14 Joint probabilities

The **joint probability** of A and B is the probability of A and B. That is

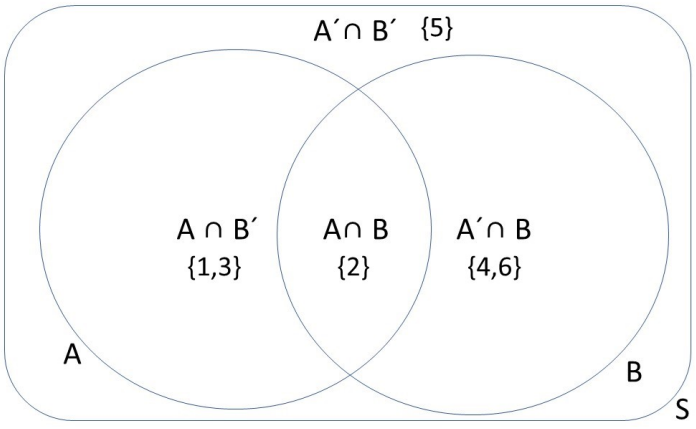
$$P(A \cap B)$$

or $P(A, B)$.

To write joint probabilities of non mutually exclusive events ($A \cap B \neq \emptyset$) into a probability table, we note that we can always decompose the sample space into **mutually exclusive** sets involving the intersections:

$S = \{A \cap B, A \cap B', A' \cap B, A' \cap B'\}$

Let's consider the Venn diagram for the example where A is the event that corresponds to drawing a number less than or equal to 3 and B corresponds to an even number:



The **marginals** of A and B are the probability of A and the probability of B, respectively:

- $P(A) = P(A \cap B') + P(A \cap B) = 2/6 + 1/6 = 3/6$

- $P(B) = P(A' \cap B) + P(A \cap B) = 2/6 + 1/6 = 3/6$

We can now write the **probability table** for the joint probabilities

Result	probability
$(A \cap B)$	$P(A \cap B) = 1/6$
$(A \cap B')$	$P(A \cap B') = 2/6$
$(A' \cap B)$	$P(A' \cap B) = 2/6$
$(A' \cap B')$	$P(A' \cap B') = 1/6$
sum	1

Each result has *two* values (one for the feature of type A and one for type B)

3.15 Contingency table

The joint probability table can also be written in a **contingency table**

	B	B'	sum
A	$P(A \cap B)$	$P(A \cap B')$	$P(A)$
A'	$P(A' \cap B)$	$P(A' \cap B')$	$P(A')$
sum	$P(B)$	$P(B')$	1

Where the marginals are the sums in the margins of the table, for example:

- $P(A) = P(A \cap B') + P(A \cap B)$
- $P(B) = P(A' \cap B) + P(A \cap B)$

In our example, the contingency table is

	B	B'	sum
A	1/6	2/6	3/6
A'	2/6	1/6	3/6
sum	3/6	3/6	1

3.16 The addition rule:

The addition rule allows us to calculate the probability of A or B, $P(A \cup B)$, in terms of the probability of A and B, $P(A \cup B)$. We can do this in three equivalent ways:

1) Using the marginals and the joint probability

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

2) Using only joint probabilities

$$P(A \cup B) = P(A \cap B) + P(A \cap B') + P(A' \cap B)$$

3) Using the complement of joint probability

$$P(A \cup B) = 1 - P(A' \cap B')$$

Example (dice)

Take the events $A : \{1, 2, 3\}$, rolling a number less than or equal to 3, and $B : \{2, 4, 6\}$, rolling an even number on the roll of a dice.

Therefore:

1) $P(A \cup B) = P(A) + P(B) - P(A \cap B) = 3/6 + 3/6 - 1/6 = 5/6$
2) $P(A \cup B) = P(A \cap B) + P(A \cap B') + P(A' \cap B) = 1/6 + 2/6 + 2/6 = 5/6$
3) $P(A \cup B) = 1 - P(A' \cap B') = 1 - 1/6 = 5/6$

In the contingency table $P(A \cup B)$ corresponds to the three cells in bold (method 2 above). That is: all cells but 1/6 from the bottom right (method 3).

	B	B'
A	**1/6**	**2/6**
A'	**2/6**	*1/6*

3.17 Questions

We collect the age and category of 100 athletes in a competition

	$age : junior$	$age : senior$
$category : 1st$	14	12
$category : 2nd$	21	18
$category : 3rd$	22	13

1) What is the estimated probability that an athlete is 2nd category and senior?

a: 18/100; **b:** 18/43; **c:** 18; **d:** 18/39

2) What is the estimated probability that the athlete is not in the third category and is senior?

a: 35/100; **b:** 30/100; **c:** 22/100; **d:** 13/100

3) What is the marginal probability of the third category?

a: 13/100; **b:** 35/100; **c:** 22/100; **d:** 13/22

4) What is the marginal probability of being senior?

a: 13/100; **b:** 43/100; **c:** 43/57; **d:** 57/100

5) What is the probability of being senior or third category?

a: 65/100; **b:** 86/100; **c:** 78/100; **d:** 13/100

3.18 Exercises

3.18.0.1 Classical probability: Exercise 1

- Write the table of **joint probability** for the **results** of rolling two dice; In the rows write the results of the first die and in the columns the results of the second die.
- What is the probability of drawing $(3, 4)$? (A:1/36)
- What is the probability of rolling 3 and 4 with any of the two dice? (A:2/36)
- What is the probability of rolling 3 on the first die or 4 on the second? (To:11/36)
- What is the probability of rolling 3 or 4 with any dice? (A:20/36)
- Write the **probability table** for the result of the **add** of two dice. Assume that the outcome of each die is **equally likely**. Verify that it is:

$$P_i = \begin{cases} \frac{i-1}{36}, & i \in \{2, 3, 4, 5, 6, 7\} \\ \frac{13-i}{36}, & i \in \{8, 9, 10, 11, 12\} \end{cases}$$

3.18.0.2 Frequentist probability: Exercise 2

The result of a randomized experiment is to measure the severity of misophonia **and** the state of depression of a patient.

Misophonia

- severity: $S_M : \{M_0, M_1, M_2, M_3, M_4\}$
- Depression: $S_D : \{D', D\}$)

Write the contingency table for the absolute frequencies ($n_{M,D}$) for a study on a total of 123 patients in which it was observed

- 100 individuals did not have depression.
- No individual with misophonia 4 and without depression.
- 5 individuals with grade 1 misophonia and no depression.
- The same number as the previous case for individuals with depression and without misophonia .

- 25 individuals without depression and grade 3 misophonia .
- The number of misophonics without depression for grades 2 and 0 were distributed equally .
- The number of individuals with depression and misophonia increased progressively in multiples of three, starting at 0 individuals for grade 1.

Answer the following questions:

- How many individuals had misophonia ? (A:83)
- How many individuals had grade 3 misophonia ? (A:31)
- How many individuals had grade 2 misophonia without depression? (A:35)

Write down the contingency table for relative frequencies $f_{M,D}$. Suppose N is large and the absolute frequencies **estimate** the probabilities $f_{M,D} = \hat{P}(M \cap D)$. Answer the following questions:

- What is the marginal probability of severity 2 misophonia ? (A: 0.3)
- What is the probability of not being misophonic **and** not being depressed? (A:0.284)
- What is the probability of being misophonic **or** depressed? (A: 0.715)
- What is the probability of being misophonic **and** being depressed? (A: 0.146)
- Describe in spoken language the results with probability 0.

3.18.0.3 Exercise 3

We have carried out a randomized experiment 10 times, which consists of recording the sex and vital status of patients with some type of cancer after 10 years of diagnosis. We got the following results

```
##           A     B
## 1     male   dead
## 2     male   dead
## 3     male   dead
## 4   female  alive
## 5     male   dead
## 6   female  alive
## 7   female   dead
## 8   female  alive
## 9     male  alive
## 10    male  alive
```

- Create the contingency table for the number $(n_{i,j})$ of observations of each result (A, B)
- Create the contingency table for the relative frequency $(f_{i,j})$ of the results
- What is the marginal frequency of being a man? (R/0.6)
- What is the marginal frequency of being alive? (R/0.5)
- What is the frequency of being alive **or** being a woman? (R/0.6)

3.18.0.4 Theory: Exercise 4

- From the second form of the addition rule, obtain the first and the third form.
- What is the third form addition rule for the probability of three events $P(A \cup B \cup C)$?

3.19 Practice

Load misophonia data https://alejandro-isglobal.github.io/SDA/data/data_0.txt

- Compute the contingency table of absolute frequencies for misophonia diagnosis (Misofonia.dic) and depression (depresion.dic)
- Compute the contingency table of relative frequencies for misophonia diagnosis (Misofonia.dic) and depression (depresion.dic)
- Compare the differences with exercise 2.

Solutions

Chapter 4

Conditional probability

In this chapter, we will introduce conditional probability.

We will use conditional probability to define statistical independence.

We will discuss Bayes' theorem and we will discuss one of its main applications, which is the predictive efficiency of a diagnostic tool.

4.1 Joint probability

Recall that the joint probability of two events A **and** B is defined as the probability of their intersection

$$P(A, B) = P(A \cap B)$$

Now imagine random experiments that measure simultaneously two different outcomes. Simultaneous means that the order in which they are measured is not a condition of the experiment.

- height and weight of an individual: (h, w)
- time and position of an electric charge: (p, t)
- the throw of two dice: $(n_1,\, n_2)$
- cross two green traffic lights in green (not red): $(R'_1,\, R'_2)$

In this last case, while we may cross the traffic lights always in the same order, a run of the experiment may ask first whether the driver crossed traffic light 2 in green and then ask if the driver crossed traffic light 1 in green. Clearly this is the same as asking the other way round: $(R'_2,\, R'_1)$=$(R'_1,\, R'_2)$. That is, the joint probability of two events is commutative.

We are often interested in whether the values of one result **condition** the values of the other.

4.2 Statistical independence

In many cases, we are interested in whether two outcomes often occur together. We want to be able to discern between two cases.

- **Independence** between events. For example, rolling a 1 on one die does not make it more likely to roll another 1 on a second die.
- **Correlation** between events. For example, if a man is tall, he is probably heavy.

We may not observe what is ''first": whether the man is first tall and then heavy, or first heavy and then tall. The correlation does not tell us which of the two options is the most likely or what is **the causal relationship** between them. We need more information, for instance, that being heavy and short is not uncommon. Being heavy is observed with different heights while being tall and light is rarely observed. Therefore, it is more likely that the man is heavy because he is tall, than the opposite. In other words, being tall conditions the weight. Let us remark that point, the correlation between to outcomes does not give the direction of causality, we need more.

Example (conductor)

We conducted an experiment to find out if observing structural flaws in a material affects its electrical conductivity. We repeated the experiment n times and measured both quantities.

The data would look like

Conductor	Structure	Conductivity
c_1	flaws	low
c_2	no flaws	high
c_3	flaws	low
...	...	...
c_i	no flaws	low*
...	...	...
...	...	...
c_n	flaws	high*

We can expect low conductivity to occur more often with flaws than without flaws if the flaws affect conductivity.

Let's imagine that from the data we obtain the following contingency table of **estimated joint probabilities**

	with flaws (F)	no flaws (F')	sum
low (L)	0.005	0.045	0.05
high (L')	0.095	0.855	0.95
sum	0.1	0.9	1

where, for example, the joint probability of L and F is

- $P(L, F) = 0.005$

and the marginal probabilities are

- $P(L) = P(L, F) + P(L, F') = 0.05$
- $P(F) = P(L, F) + P(L', F) = 0.1.$

4.3 The conditional probability

Low conductivity is **independent** of having structural flaws if the probability of having low conductivity (L) is the same **whether** it has flaws (F) or not (F') .

Let us first consider only the materials that have flaws.

Among those materials that have flaws (F), what is the estimated probability that they have low conductivity?

$\hat{P}(L|F) = \frac{n_{L,F}}{n_F} = \frac{n_{L,F}/n}{n_F/n} = \frac{f_{L,F}}{f_F}$

$$= \frac{\hat{P}(L, F)}{\hat{P}(F)}$$

Therefore, in the limit when $N \to \infty$, we have

$$P(L|F) = \frac{P(L, F)}{P(F)} = \frac{P(L \cap F)}{P(F)}$$

Definition:

The **conditional probability** of an event A given an event B, denoted $P(B|A)$, is

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

We can prove that conditional probability satisfies the axioms of probability. The conditional probability can be understood as a probability with a sample space given by B: S_B. In our example, looking only at the materials with structural flaws.

4.4 Conditional contingency table

If we divide the columns of the joint probability table by the marginal probabilities of the conditioning effects (F and F'), we can write **a conditional contingency table**

F

F'

L

P(L | F)

P(L | F')

L'

P(L' | F)

P(L' | F')

sum

1

1

where the column probabilities sum to one. The first column shows the probabilities of low conductivity or not only of the materials that have flaws (first condition: F). The second column shows the probabilities only for the materials that have no flaws (second condition: F').

Conditional probabilities are the probabilities of the event within each condition. We read them as:

- $P(L|F)$: Probability of having low conductivity **if** it has flaws
- $P(L'|F)$: Probability of not having low conductivity **if** it has flaws
- $P(L|F')$: Probability of having low conductivity **if** it has no flaws
- $P(L'|F')$: Probability of not having low conductivity **if** it has no flaws

4.5 Statistical independence

In our example, the conditional contingency table is

F

F'

L

P(L | F)= 0.05

P(L | F')= 0.05

L’

P(L’ | F)= 0.95

P(L’ | F’)= 0.95

sum

1

1

We note that the marginals from the joint probability table from before (Section 4.2) and the conditional probabilities in this table are equal

- $P(L) = P(L|F) = P(L|F')$
- $P(L') = P(L'|F) = P(L'|F')$

This means that the probability of observing a low conductivity is **not** dependent on having a structural flaw or not.

We conclude that for the physical situation represented by this experiment, low conductivity in the material is not affected by having a structural flaw.

Definition

Two events A and B are statistically independent if either of the equivalent cases occurs:

1) $P(A|B) = P(A)$; A is independent of B
2) $P(B|A) = P(B)$; B is independent of A

and by the definition of conditional probability

3) $P(A \cap B) = P(A|B)P(B) = P(A)P(B)$

This third form is a statement about joint probabilities. It says that we can obtain joint probabilities by multiplying the marginal probabilities.

In our original joint probability table

	F	F’	sum
L	0.005	0.045	0.05
L’	0.095	0.855	0.95
sum	0.1	0.9	1

we can confirm that all the entries of the matrix are indeed the product of the marginal probabilities. For example: $P(L \cap F) = P(F)P(L)$ and $P(L' \cap F') = P(L')P(F')$. Therefore, in our experiment, low conductivity is independent of having a structural flaw because the joint probability is the product of the marginals.

Example (Coins)

We want to confirm that the results of tossing two coins are independent. We consider all outcomes to be equally likely:

Putcome	Probability
(H, T)	1/4
(H, H)	1/4
(T, T)	1/4
(T, H)	1/4
sum	1

where (H, T) is, for example, the event of heads on the first coin and tails on the second coin. The contingency table for the joint probabilities is:

	H	T	sum
H	1/4	1/4	1/2
T	1/4	1/4	1/2
sum	1/2	1/2	1

From this table, we see that the probability of getting a head and then a tail is the product of the marginals $P(H, T) = P(H)P(T) = 1/4$. Therefore, the events of heads in the first coin and tails in the second are independent.

If we build the conditional contingency table on the toss of the first coin, we will see that obtaining tails in the second coin is not conditioned by having obtained heads in the first coin: $P(T|H) = P(T) = 1/2$

4.6 Statistical dependency

An important example of statistical dependency is found in the performance of **diagnostic tools**, where we want to determine the state of a system with possible outcomes

- satisfactory (yes)
- unsatisfactory (not)

using a test with results

- positive
- negative

For example, we test a battery to see how long it can last. We load a cable to find out if it resists carrying a certain load. We run a PCR to see if someone has an infecteion.

4.7 Diagnostic test

Let's consider diagnosing an infection with a new test. The infection status has two possible outcomes:

- yes (the patient is infected)
- no (the patient is not infected)

The test has two possible outcomes:

- positive (the test detects the infection)
- negative (the test does not detect the infection)

The **conditional contingency table** is the data we collect in a controlled environment (laboratory)

Infection: yes

Infection: no

Test: positive

P(pos | yes)

P(pos | no)

Test: negative

P(neg | yes)

P(neg | no)

sum

1

1

The conditional table tells us that we are running an experiment where **we know** that the patient either has the disease or not. These are the **controlled conditions** of the experiment. Think for instance that you test the diagnostic tool in the hospital where you know the patients are infected. You also test the tool in healthy individuals who you know they are not infected.

Let's look at the table entries

1) $P(pos|yes)$ is called the **sensitivity** of the tool or the true positive rate: The probability of testing positive **if** a patient has the disease.

2) $P(neg|no)$ is called the **specificity** of the tool. or the true negative rate: The probability of testing negative **if** a patient does not have the disease.

3) $P(pos|no)$ is called the false positive rate: the probability of testing positive **if** the patient does not have the disease.

4) $P(neg|yes)$ is called the false negative rate: the probability of testing negative **if** the patient has the disease.

High correlation (statistical dependence) between test and infection means high values for probabilities 1 and 2 (successes) **and** low values for probabilities 3 and 4 (errors).

Example (COVID)

Now let's consider a real situation. In the early days of the coronavirus pandemic, there was no measure of the effectiveness of PCRs in detecting the virus. One of the first published studies (https://www.nejm.org/doi/full/10.1056/NEJMp2015897) found that

- The PCR had a sensitivity of 70%, in infection condition.
- The PCR had a specificity of 94%, in non-infected condition.

Therefore, the conditional probability table for this study was

Infection: yes

Infection: no

Test: positive

P(pos | yes)=0.7

P(pos | no)=0.06

Test: negative

P(neg | yes)=0.3

P(neg | no)=0.94

sum

1

1

Therefore, the errors in the diagnostic tests were:

- The false positive rate: $P(pos|no) = 0.06$
- The false negative rate: $P(neg|yes) = 0.3$

Can we say with this data that the test is useful to detect the infection?

4.8 Inverse probabilities

We are really interested in finding the probability of being infected if the test is positive:

$$P(yes|pos)$$

In other words, you may want to collect the people with positive test and compute the fraction of those that are really infected. Note that this experiment is not in a controlled environment of the lab, it is rather a surveying campaign on the population.

Alternatively, we can:

1. Recover the contingency table for **joint probabilities**, multiplying by the marginal $P(yes)$ and $P(no)$ that we need to know

	Infection: yes	Infection: No	sum
Test: positive	P(pos | yes)P(yes)	P(pos | no)P(no)	P(pos)
Test: negative	P(neg | yes)P(yes)	P(neg | no) P(no)	P(neg)
sum	P(yes)	P(no)	1

2. Obtain the **conditional probability table** for rows.

Infection: yes

Infection: No

sum

Test: positive

P(yes|pos)

P(no|pos)

1

Test: negative

P(yes|neg)

P(no|neg)

1

To compute these probabilities, we use the definition of conditional probabilities for rows instead of columns. We divide the rows of the joint probability table in step 1 by the marginals of the test outcomes: $P(pos)$ and $P(neg)$.

For example for the first cell of the conditional probability table we obtain:

$$P(yes|pos) = \frac{P(pos|yes)P(yes)}{P(pos)}$$

$P(pos|yes)$ was the result of the study (0.7). However, to apply this formula we still need to know the probability of infection $P(yes)$ (prevalence) and of no

infection $P(pos)$. We also need the probability that a random subject in the poulations tests positive $P(pos)$.

- The prevalence $P(yes)$ **needs to be given**. In real life it is obtained from another study. The first prevalence study in Spain showed that during confinement $P(yes) = 0.05$, $P(no) = 0.95$, before the summer of 2020.
- To find the marginal of positive tests $P(pos)$, we can use the definition of marginal and conditional probability:

$$P(pos) = P(pos \cap yes) + P(pos \cap no)$$

$$= P(pos|yes)P(yes) + P(pos|no)P(no)$$

This last relation of the marginals is called **total probability rule**.

4.9 Bayes' Theorem

After substituting the total probability rule into $P(yes|pos)$, we have

$$P(yes|pos) = \frac{P(pos|yes)P(yes)}{P(pos|yes)P(yes) + P(pos|no)P(no)}$$

This expression is known as **Bayes' theorem**. It allows us to reverse the conditionals:

$$P(pos|yes) \rightarrow P(yes|pos)$$

This result is important. It allows us to **assess** a diagnostic tool in a controlled condition (infection status is a lab) and then use it to **infer** the probability of the condition (infection) when the test is positive.

Example (COVID):

The test performance was:

- Sensitivity: $P(positive|yes) = 0.70$
- False positive rate: $P(positive|no) = 1 - P(neg|no) = 0.06$

The study in the Spanish population gave:

- $P(yes) = 0.05$
- $P(no) = 1 - P(yes) = 0.95$.

Therefore, the probability of being infected in case of testing positive was:

$$P(yes|pos) = 0.38$$

We concluded that at that time PCR was not very good at **confirming** infections.

However, let us now apply Bayes' theorem to the probability of not being infected if the test was negative.

$$P(no|neg) = \frac{P(neg|no)P(no)}{P(neg|no)P(no) + P(neg|yes)P(yes)}$$

Substituting all values gives

$$P(no|neg) = 0.98$$

So the tests were good for **ruling out** infections and a fair requirement for travelling.

Example (Mosophonia)

In general, we can have more than two conditioning events, or controlled environments of our experiment. Think for instance on mispophonia categories. Therefore, Bayes' theorem says:

$$P(E_i|B) = \frac{P(B|E_i)P(E_i)}{P(B|E_0)P(E_0) + ... + P(B|E_k)P(E_k)}$$

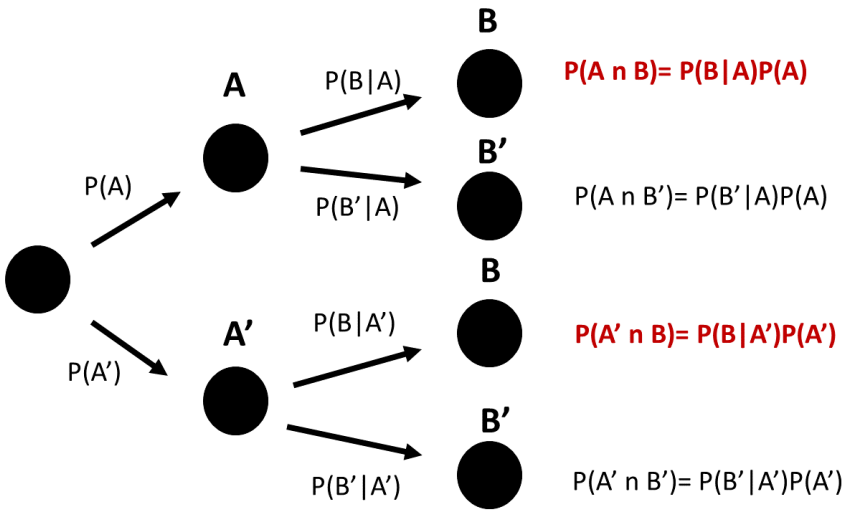
when $E_0, E_1, ..., E_k$ are k mutually exclusive and exhaustive events (misophonia 0, 1, 2, 3, 4) and B is an outcome of interest (depression). Then the inverse probability $P(E_i|B)$ will give us the probability that patient has mosophinia grade i if he has depression. We do not know what was first, depression or misophonia, but this computation may be useful for finding potential causes of depression. But for this, we need **to know**, by running a campaign, the prevalence of the various degrees of misophania in the population $P(E_0), P(E_1), ...P(E_k)$.

Remember that the denominator, or the prevalence of depression $P(B)$, can be obtained from the total probability rule:

$$P(B) = P(B|E_0)P(E_0) + ... + P(B|E_k)P(E_k)$$

Conditional tree

The terms in the total probability rule can also be **organized** in a conditional tree.



The **total probability rule** tells us in how many ways I can get the result B from the outcomes A or A' (illustrated in red above).

$$P(B) = P(B|A)P(A) + P(B|A')P(A')$$

4.10 Questions

We collect the age and category of 100 athletes in a competition

	junior	*senior*
$1st$	14	12
$2nd$	21	18
$3rd$	22	13

1) What is the estimated probability that the athlete is in the third category if the athlete is a junior?

a: 22; **b:** 22/100; **c:** 22/57; **d:** 22/35;

2) What is the estimated probability that the athlete is a junior and is in the 1st category if the athlete is not in the 3rd category?

a: 14/35; **b:** 14/65; **c:** 14/100; **d:** 14/26

3) A diagnostic test has a probability of 8/9 of detecting a disease if the patients are sick and a probability of 3/9 of detecting the disease if the patients are healthy. If the probability of being sick is 1/9. What is the probability that a patient is sick if a test detects the disease?

a: $\frac{8/9}{8/9+3/9} * 1/9$; **b:** $\frac{3/9}{8/9+3/9} * 1/9$; **c:** $\frac{3/9*8/9}{8/9*1/9+3/9*8/9}$; **d:** $\frac{8/9*1/9}{8/9*1/9+3/9*8/9}$;

4) As discussed in the notes, a PCR test for coronavirus had a sensitivity of 70% and a specificity of 94% and in Spain during confinement there was an incidence of 5%. With these data, what was the probability of testing positive in Spain ($P(positive)$)

a: 0.035; **b:** 0.092; **c:** 0.908; **d:** 0.95

5) With the same data as in question 4, testing positive in the PCR and being infected are not independent events because:

a: Sensitivity is 70%; **b:** Sensitivity and false positive rate are different; **c:** The false positive rate is 0.06%; **d:** the specificity is 96%

4.11 Exercises

4.11.0.1 Exercise 1

A machine is tested for its performance in producing high-quality turning rods. These are the test results

	Rounded: yes	Rounded: No
smooth surface: yes	200	1
smooth surface: no	4	2

- What is the estimated probability that the machine will produce a rod that does not satisfy any quality control? (A: 2/207)
- What is the estimated probability that the machine will produce a rod that fails at least one quality check? (A: 7/207)
- What is the estimated probability that the machine will produce rods with a rounded and smooth surface? (A: 200/207)
- What is the estimated probability that the bar is rounded if the bar is smooth? (A: 200/201)
- What is the estimated probability that the rod is smooth if it is rounded? (A: 200/204)
- What is the estimated probability that the rod is neither smooth nor rounded if it does not satisfy at least one quality check? (A: 2/7)
- Are smoothness and roundness independent events? (No)

4.11.0.2 Exercise 2

We developed a test to detect the presence of bacteria in a lake. We found that if the lake contains the bacteria, the test is positive 70% of the time. If there are no bacteria, the test is negative 60% of the time. We implemented the test in a region where we know that 20% of the lakes have bacteria.

- What is the probability that a lake that tests positive is contaminated with bacteria? (A: 0.30)

4.11.0.3 Exercise 3

Two machines are tested for their performance in producing high-quality turning rods. These are the test results

Machine 1

	Rounded: yes	Rounded: No
smooth surface: yes	200	1
smooth surface: no	4	2

Machine 2

	Rounded: yes	Rounded: No
smooth surface: yes	145	4
smooth surface: no	8	6

- What is the probability that the bar is rounded? (A: 357/370)
- What is the probability that the rod was produced by machine 1? (A: 207/370)
- What is the probability that the rod is not smooth? (A: 20/370)
- What is the probability that the rod is smooth or rounded or produced by machine 1? (A: 364/370)
- What is the probability that the rod will be rounded if it is smoothed and from machine 1? (A: 200/201)
- What is the probability that the rod is not rounded if it is not smooth and it is from machine 2? (A: 6/14)
- What is the probability that the rod has come out of machine 1 if it is smooth and rounded? (A: 200/345)
- What is the probability that the rod came from machine 2 if it fails at least one of the quality controls? (A: 0.72)

4.11.0.4 Exercise 4

A quality test on a random brick is defined by the events:

- Pass the quality test: E, fail the quality test: E'
- Defective: D, non-defective: D'

If the diagnostic test has sensitivity $P(E|D') = 0.99$ and specificity $P(E'|D) = 0.98$, and the probability of passing the test is $P(E) = 0.893$ then

- What is the probability that a randomly chosen brick is defective $P(D)$? (A: 0.1)
- What is the probability that a brick that has passed the test is actually defective? (A: 0.022)
- The probability that a brick is not defective **and** that it fails the test (A: 0.009)
- Are D and E' statistically independent? (No)

4.12 Practice

Load misophonia data from https://alejandro-isglobal.github.io/SDA/data/data_0.txt

- Compute the conditional probability table of misophophonia (Misofonia.dic) given marital status (Estado). What is the estimated probability of having misophonia if the patient is married?
- Compute the conditional probability table of marital status (Estado) given misophonia (Misofonia.dic). What is the estimated probability of being married if the patient is misophonic?

Solutions

Chapter 5

Discrete Random Variables

5.1 Objective

In this chapter we will define a random variables and study discrete random variables.

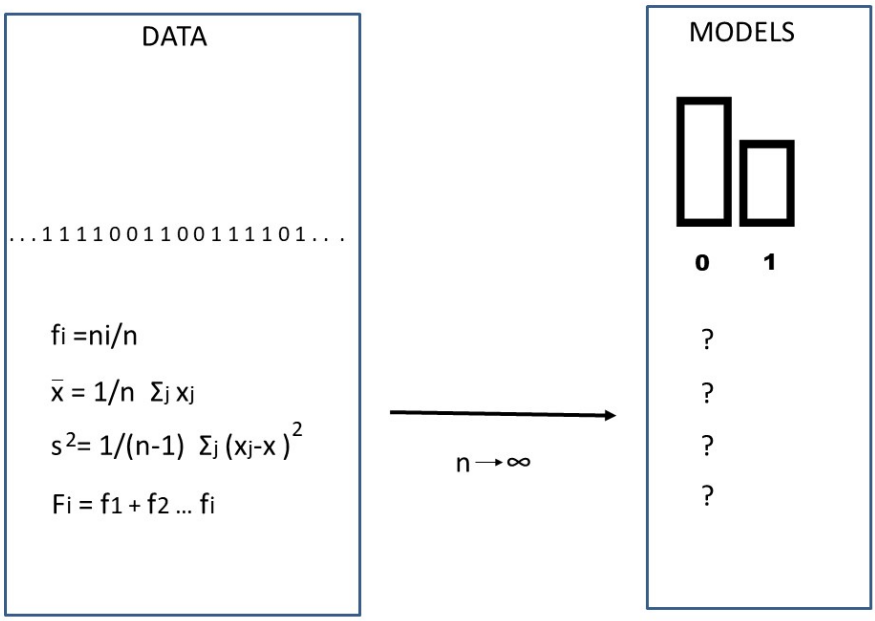
We will define the probability mass function and its main properties of mean and variance. Following the abstraction process of the relative frequencies into probabilities, we also define the probability distribution as the limiting case of the relative cumulative frequency.

5.2 Relative frequencies

The relative frequencies of the a random experiment are observations of the outcomes propensities (Popper, 1957). We can used them as estimators of their probabilities, when we repeat the random experiment a lot of times ($n \to \infty$).

We defined central tendency (average), dispersion (sample variance) and the frequency distribution the **data** (F_i).

In terms of **probabilities**, how are this quantities defined?



5.3 Random variable

We defined the relative frequencies based on the **observations** obtained from several runs of the random experiment. We now define the equivalent quantities for probabilities in terms of the **outcomes** of the experiments. We will deal with numerical outcomes only.

A **random variable** is a symbol that represents a **numerical outcome** of a random experiment. We write the random variable in **capitals** (i.e. X). Think of it as the a recipe to obtain an observation from the experiment. It is not the observation is the procedure to obtain it. In mathemaics procedures are called functions.

Definition:

A **random variable** is a function that assigns a real **number** to an **event** from the sample space of a random experiment.

Remember than an event can be an outcome or a collection of outcomes.

When the random variable takes a **value**, it indicates the realization of an **event** and the production of a number from the random experiment.

Example (Switch)

If $X \in \{0, 1\}$, we then say X is a random variable that can take the values 0 or

1. X may be the state of a switch. The switch can be on if there observe light in the room 1 or off we see nothing 0.

5.4 The value of a random variable

We make the distinction between variables in the model space with capital letters, as abstract entities (the state of a switch, the color of a car, the sex of a patient), and the realization of a particular outcome (whether the light of my office is on right now, the red color of my car, the first patient on the hospital today was male). For instance, we say that

- $X = 1$ is the **event** of observing the random variable X taking the outcome value of 1
- $X = 2$ is the **event** of observing the random variable X taking the outcome value of 2

...

In general:

- $X = x$ is the **event** of observing the random variable X (big X) with value x (little x).

We use lower case letters to indicate the numerical **outcome** of an experiment that is obtained one day in the lab.

5.5 Probability of random variables

We are interested in assigning probabilities to the outcome values of a random variable.

For instance for the dice we will write the probability table as

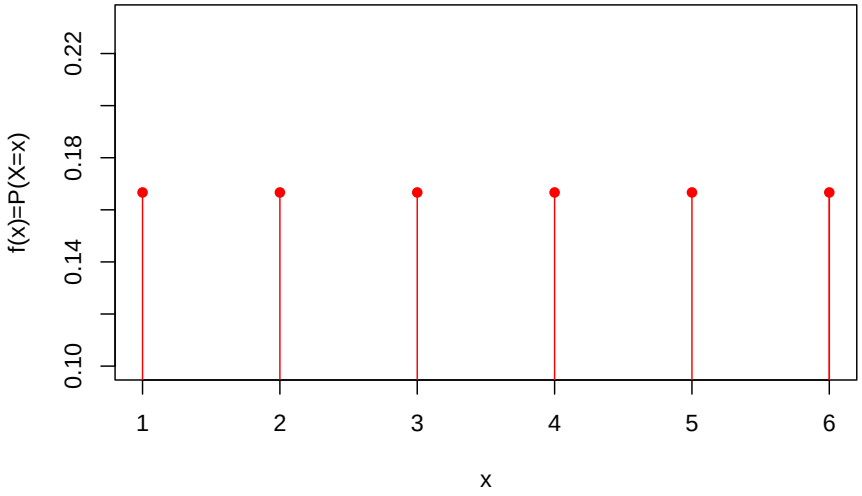
X	Probability
1	$P(X = 1) = 1/6$
2	$P(X = 2) = 1/6$
3	$P(X = 3) = 1/6$
4	$P(X = 4) = 1/6$
5	$P(X = 5) = 1/6$
6	$P(X = 6) = 1/6$

where we make explicit the events that the variable takes a given outcome value $X = x$.

If X is the roll of the dice, then $P(X = 1)$ is the probability that the roll of a dice gives an outcome value of 1.

5.6 Probability functions

Because (little) x is a numerical quantity, the probabilities of the random variable can be plotted against the outcomes x



or written as the mathematical function

$$f(x) = P(X = x) = 1/6$$

Therefore, the probabilities of a random experiment can be given in table, in a plot or as a **function**.

5.7 Probability mass functions

We can **create** any type of probability function if we satisfy Kolmogorov's probability rules.

For a discrete random variable $X \in \{x_1, x_2, .., x_M\}$, a **probability mass function** that is used to compute probabilities

- $f(x_i) = P(X = x_i)$

is always positive

- $f(x_i) \geq 0$

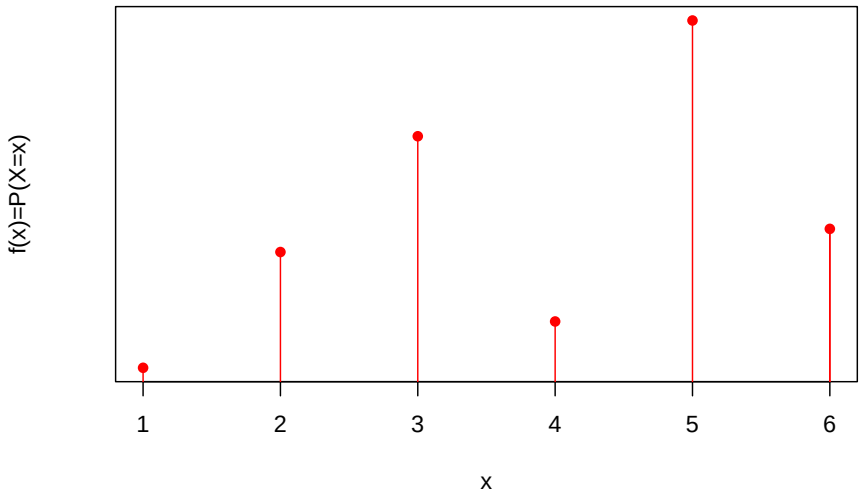
and its sum over all the values of the variable is 1:

- $\sum_{i=1}^{M} f(x_i) = 1$

Where M is the number of possible outcomes.

Note that the definition of X and its probability mass function is general **without reference** to any experiment. The functions live in the model (abstract) space.

Here is an example



X and $f(x)$ are abstract objects that may or may not describe real random experiment and its probabilities. We have the freedom to construct them as we want as long as we respect their definition.

Probability mass functions have some **properties** that are derived exclusively from their definition. Before, let us look in further detail the relationship between probabilities and relative frequencies.

5.8 Probabilities and relative frequencies

Example (Urn)

In one urn put 8 balls following the instructions:

- mark 1 ball with number -2
- mark 2 balls with number -1
- mark 2 balls with number 0
- mark 2 balls with number 1

- mark 1 ball with number 2

And consider performing the following random **experiment:** Take one ball and read the number.

From the classical probability, we can write the probability table, for which we do not need to run any experiment

X	$P(X = x)$
-2	$1/8 = 0.125$
-1	$2/8 = 0.25$
0	$2/8 = 0.25$
1	$2/8 = 0.25$
2	$1/8 = 0.125$

Now, let's perform the experiment 30 times (in a computer) and write the frequency table. For a particular set of runs we obtained

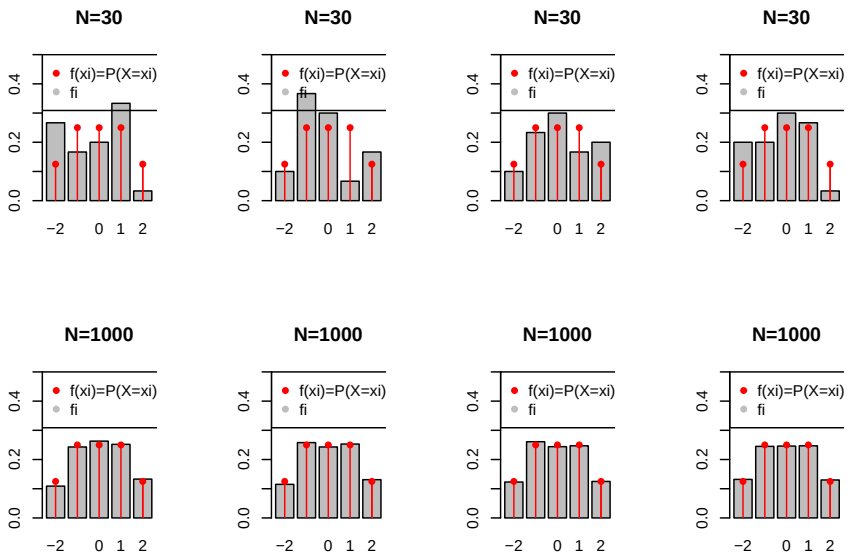
X	f_i
-2	0.132
-1	0.262
0	0.240
1	0.248
2	0.118

The frequentist statistician tells us that that the relative frequencies in the limit is the probability mass function

$$lim_{N\to\infty} f_i = f(x_i) = P(X = x_i)$$

Then, if we did not know the set up of the experiment (black box), the best we can do is to **estimate** the probabilities with the frequencies, obtained from N repetitions of the random experiment:

$$f_i = \hat{P}_i$$



Everytime we estimate the probabilities, our estimates $\hat{P}_i = f_i$ change (bar plots). But P_i is an abstract quantity that never changes (red rods). As N increases, however, the bars get closer to the rods.

5.9 Mean or expected value

When we discussed summary statistics of data, we defined the center of the observations as a value around which the outcome frequencies are concentrated.

We used the **average** to measure the center of mass of the **data**. In terms of the relative frequencies of the values of ordinal categorical outcomes, we wrote the average as

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i = \sum_{i=1}^{M} x_i \frac{n_i}{N} =$$

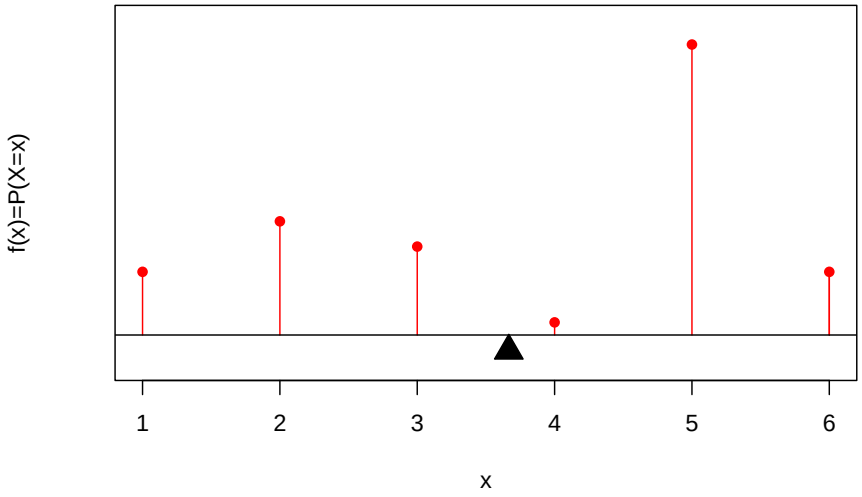
$$\sum_{i=1}^{M} x_i f_i$$

Note the change from the number of observations N to the number of outcomes M in the summation.

Definition

The **mean** (μ) or expected value of a discrete random variable X, $E(X)$, with mass function $f(x)$ is given by

$$\mu = E(X) = \sum_{i=1}^{M} x_i f(x_i)$$



It is the center of mass of the **probabilities**: The point where the probability loadings on a road are balanced.

From the definition we have

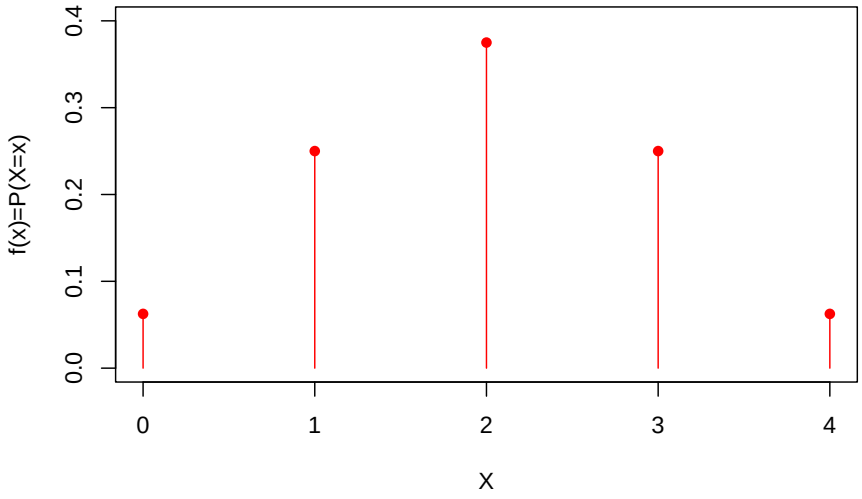
$$\bar{x} \to \mu$$

in the **limit** when $N \to \infty$ as the frequency tends to the probability mass fucntion $f_i \to f(x_i)$.

Example

What is the mean of X if its probability mass function $f(x)$ is given by

X	$f(x) = P(X = x)$
0	1/16
1	4/16
2	6/16
3	4/16
4	1/16



$$\mu = E(X) = \sum_{i=1}^{m} x_i f(x_i)$$

$E(X)$ =**0** * 1/16 + **1** * 4/16 + **2** * 6/16 + **3** * 4/16 + **4** * 1/16 =2

The **mean** μ is the center of mass of the probability mass function. **it does not change**. However, the **average** $\bar{x}$ is the center of mass of the observations (relative frequencies). It **changes** with different data.

5.10 Variance

When we discussed summary statistics of data, we also defined the spread of the observations as an average distance from the data average.

Definition

The variance, written as σ^2 or $V(X)$, of a discrete random variable X with mass function $f(x)$ is given by

$$\sigma^2 = V(X) = \sum_{i=1}^{M} (x_i - \mu)^2 f(x_i)$$

$\sigma = \sqrt{V(X)}$ is called the **standard deviation** of the random variable.

The variance is the spread of the **probabilities** about the mean. That is the moment of inertia of probability loadings about the center of mass.

Example

What is the variance of X if its probability mass function $f(x)$ is given by

X	$f(x) = P(X = x)$
0	1/16
1	4/16
2	6/16
3	4/16
4	1/16

$$\sigma^2 = V(X) = \sum_{i=1}^{m} (x_i - \mu)^2 f(x_i)$$

$V(X) =$**(0-2)**^{2*} 1/16 + **(1-2)**^{2*} 4/16 + **(2-2)**^{2*} 6/16 + **(3-2)**^{2*} 4/16 + **(4-2)**^{2*} 1/16 =1

$$V(X) = \sigma^2 = 1$$
$$\sigma = 1$$

5.11 Probability functions for functions of X

In many occasions, we will be interested in outcomes that are function of the random variables. Perhaps, we are interested in the square of the number of flu infections, or on the square root of the number of emails in an hour.

Definition

For any function h of a random variable X, with mass function $f(x)$, its expected value is given by

$$E[h(X)] = \sum_{i=1}^{M} h(x_i) f(x_i)$$

This is an important definition that allows us to prove three frequently used properties of the mean and variance:

1) The mean of a linear function is the linear function fo the mean:

$$E(a \times X + b) = a \times E(X) + b$$

for a and b scalars (numbers).

2) The variance of a linear function of X is:

$$V(a \times X + b) = a^2 \times V(X)$$

3) The variance **about the origin** is the variance **about the mean** plus the mean squared:

$$E(X^2) = V(X) + E(X)^2$$

Example

What is the variance X about the origin, $E(X^2)$, if its probability mass function $f(x)$ is given by

X	$f(x) = P(X = x)$
0	1/16
1	4/16
2	6/16
3	4/16
4	1/16

$$E(X^2) = \sum_{i=1}^{m} x_i^2 f(x_i)$$

$E(X^2) =$**(0)**^{2*} 1/16 + **(1)**^{2*} 4/16 + **(2)**^{2*} 6/16 + **(3)**^{2*} 4/16 + **(4)**^{2*} 1/16 =5

We can also verify:

$$E(X^2) = V(X) + E(X)^2$$

$5 = 1 + 2^2$

5.12 Probability distribution

When we discussed summary statistics of data, we also defined the frequency distribution (or the relative cumulative frequency) F_x. F_x is an important quantity because it is a function on the continuous range of x, even if the outcomes are discrete.

Definition:

The **probability distribution** function of a random variable X is defined as

$$F(x) = P(X \leq x) = \sum_{x_i \leq x} f(x_i)$$

That is the accumulated probability up to a given value x

$F(x)$ satisfies therefore satisifies:

1) $0 \leq F(x) \leq 1$; and
2) If $x \leq y$, then $F(x) \leq F(y)$

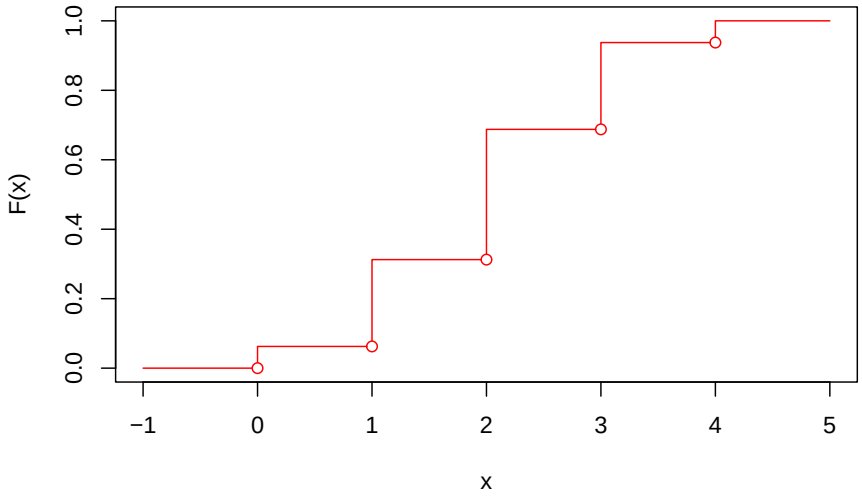
For the probability mass function:

X	$f(x) = P(X = x)$
0	1/16
1	4/16
2	6/16
3	4/16
4	1/16

The probability distribution is:

$$F(x) = \begin{cases} 0, & x \leq 0 \\ 1/16, & 0 \leq x < 1 \\ 5/16, & 1 \leq x < 2 \\ 11/16, & 2 \leq x < 3 \\ 15/16, & 3 \leq x < 4 \\ 16/16, & 4 \leq x \end{cases}$$

For$X \in \mathbb{R}$



5.13 Probability function and probability distribution

The probability function and distribution are equivalent. They encode the same information. We can get one from the other and vice-versa

$$f(x_i) = F(x_i) - F(x_{i-1})$$

with

$$f(x_1) = F(x_1)$$

for X taking values in $x_1 \leq x_2 \leq ... \leq x_n$

Example

From probability distribution:

$$F(x) = \begin{cases} 1/16, & \text{if } 0 \leq x < 1 \\ 5/16, & 1 \leq x < 2 \\ 11/16, & 2 \leq x < 3 \\ 15/16, & 4 \leq x < 5 \\ 16/16, & x \leq 5 \end{cases}$$

We can obtain the probability mas function.

$f(0) = F(0) = 1/16$ $f(1) = F(1) - f(0) = 5/32 - 1/32 = 4/16$ $f(2) = F(2) - f(1) - f(0) = F(2) - F(1) = 6/16$ $f(3) = F(3) - f(2) - f(1) - f(0) = F(3) - F(2) = 4/16$ $f(4) = F(4) - F(3) = 1/16$

5.14 Quantiles

Finally, we can use the probability distribution $F(x)$ to define the median and the quartiles of the random variable X.

In general, we define the **q-quantile** as the value x_q **under** which we have accumulated $q \times 100\%$ of the probability

$$q = \sum_{x_i \leq x_q} f(x_i) = F(x_q)$$

- The **median** is value $x_{0.5}$ such that $q = 0.5$

$$F(x_{0.5}) = 0.5$$

- The 0.05-quantile is the value $x_{0.05}$ such that $q = 0.05$

$$F(x_{0.05}) = 0.05$$

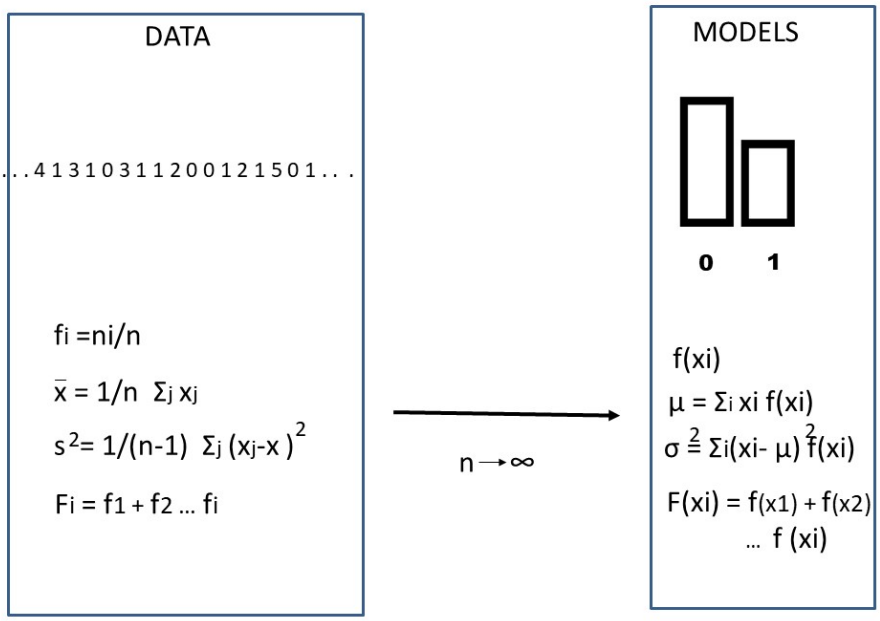
- The 0.25-quantile is **first quartile** the value $x_{0.25}$ such that $q = 0.25$

$$F(x_{0.25}) = 0.25$$

In the plots for the probability distribution $F(X)$, the median is the distance x at which $F(x)$ has gone up 50% of the height. And the first quartile is the distance at wich $F(x)$ has gone up 25% of the height.

5.15 Summary

This is a graphical summary



quantity names	model (unobserved)	data (observed)
probability mass function // relative frequency	$f(x_i) = P(X = x_i)$	$f_i = \frac{n_i}{N}$
probability distribution // cumulative relative frequency	$F(x_i) = P(X \leq x_i)$	$F_i = \sum_{k \leq i} f_k$
mean // average	$\mu = E(X) = \sum_{i=1}^{M} x_i f(x_i)$	$\bar{x} = \sum_{j=1}^{N} x_j / N$
variance // sample variance	$\sigma^2 = V(X) = \sum_{i=1}^{M} (x_i - \mu)^2 f(x_i)$	$s^2 = \sum_{j=1}^{N} (x_j - \bar{x})^2 / (N-1)$
standard deviation // sample sd	$\sigma = \sqrt{V(X)}$	s
variance about the origin // 2nd sample moment	$E(X^2) = \sum_{i=1}^{M} x_i^2 f(x_i)$	$m_2 = \sum_{j=1}^{N} x_j^2 / n$

Note that:

- $i = 1...M$ is an **outcome** of the random variable X.
- $j = 1...N$ is an **observation** of the random variable X.

Properties:

- $\sum_{i=1...N} f(x_i) = 1$

- $f(x_i) = F(x_i) - F(x_{i-1})$
- $E(a \times X + b) = a \times E(X) + b$; for a and b scalars.
- $V(a \times X + b) = a^2 \times V(X)$
- $E(X^2) = V(X) + E(X)^2$

5.16 Questions

1) For a probability mass function is not true that

a: the addition of their image values is 1; **b:** its values can be interpreted as probabilities of events; **c:** it is always positive; **d:** cannot take value 1;

2) A value of a random variable is

a: an observation of a random experiment; **b:** the frequency of an outcome of a random experiment; **c:** an outcome of a random experiment; **d:** a probability of an outcome;

3) The estimated value of a probability $\hat{P}_i$ is equal to the probability P_i when the number of repetitions of the random experiment is

a: large; **b:** infinite; **c:** small **d:** zero;

4) If a probability mass function is symmetric around $x = 0$

a: The mean is lower than the median; **b:** The mean is greater than the median; **c:** The mean and the median are equal; **d:** The mean and the median are different from 0;

5) The mean and variance

a: are inversely proportional; **b:** are expected values of functions of X; **c:** of a linear function are the linear function of the mean and the linear function of the variance; **d:** change when we repeat the random experiment;

5.17 Exercises

5.17.0.1 Exercise 1

Consider the following random variable X over the outcomes

outcome	X
a	0
b	0
c	1.5
d	1.5
e	2
f	3

outcome	X

a) If each outcome is equally probable then what is the probability mass function of x?

b) Find:

- $P(X > 3)$
- $P(X = 0 \cup X = 2)$
- $P(X \leq 2)$

5.17.0.2 Exercise 2

Given the probability mass function

x	$f(x) = P(X = x)$
10	0.1
12	0.3
14	0.25
15	0.15
17	?
20	0.15

- what is its expected value and standard deviation? (A: 14.2; 2.95)

5.17.0.3 Exercise 3

Given the probability distribution for a discrete variable X

$$F(x) = \begin{cases} 0, & x < -1 \\ 0.2, & x \in [-1, 0) \\ 0.35, & x \in [0, 1) \\ 0.45, & x \in [1, 2) \\ 1, & x \geq 2 \end{cases}$$

- find $f(x)$
- find $E(X)$ and $V(X)$ (A:1; 1.5)
- what is the expected value and variance of $Y = 2X + 3$ (A: 6)
- what is the median and the first and third quartiles of X? (A:2,0,2)

5.17.0.4 Exercise 4

We are testing a system to transmit digital pictures. We first consider the experiment of sending 3 pixels and having as **possible** outcomes events such like $(0, 1, 1)$. This is the event of receiving the first pixel with no error, the second with error and third with error.

- List in one column the sample space of the random experiment.
- In the a second column assign the random variable that counts the number of errors transmitted for each outcome

Consider that we have a totally noisy channel, that is any outcome of three pixels is equally likely.

- What is the probability of receiving 0, 1, 2, or 3 errors in the transmission of 3 pixels? (A: 1/8; 3/8; 3/8; 1/8)
- Sketch the probability mass function for the number of errors
- What is the expected value for the number of errors? (A:1.5)
- What is its variance? (A: 0.75)
- Sketch the probability distribution
- What is the probability of transmitting at least 1 error? (A:7/8)

Chapter 6

Continous Random Variables

6.1 Objective

In this chapter we will study continuous random variables.

We will define the probability density function, its mean and variance and, similar to discrete random variables, we will define the probability distribution function.

6.2 Continuous random variables

In the last chapter, we used the probabilities of discrete random variables to define the probability mass function

$$f(x) = P(X = x)$$

Where the probability that the random variable takes the value x is understood as the value of its relative frequency, when the number of repetitions of the random experiment tends to infinity.

When we talked about continuous data, we saw that we had to transform them into discrete variables (bins) to produce relative frequency tables or histograms. Let's see how to define the probabilities of continuous variables taking these partitions into account.

Example (misophonia)

Let us reconsider the angle of convexity of patients with misophonia (Section 2.21). The angle of convexity of 123 patients was measured. We understood

each measurement as the result of a random experiment that we repeated 123 times and that we could describe in a frequency table or in a histogram.

To do this, we redefine the results as small regular intervals (bins) and calculate the relative frequency of each interval.

```
##        outcome ni         fi
## 1 [-1.02,3.46]  8 0.06504065
## 2  (3.46,7.92] 51 0.41463415
## 3  (7.92,12.4] 26 0.21138211
## 4  (12.4,16.8] 20 0.16260163
## 5  (16.8,21.3] 18 0.14634146
```

6.3 Relative frequencies

Relative frequency for a bin $(x_i, x_i + \Delta x)$ when $N \to \infty$ is the probability that the random variable is observed between x_i and $x_i + \Delta x$

$$f_i = \frac{n_i}{N} \to P(x_i \leq X \leq x_i + \Delta x)$$

The probability depends now on the length of the bins Δx. If we make the bins smaller and smaller then the frequencies get smaller and smaller because it is unlikely to find any observation is such a small bin; that is, $n_i \to 0$. As a consequence, the probability also vanishes

$$P(x_i \leq X \leq x_i + \Delta x) \to 0$$

when $\Delta x \to 0$.

Let's see how the frequencies get smaller when we divide the range of X into 20 bins

```
##          outcome ni         fi
## 1  [-1.02,0.115]  2 0.01626016
## 2   (0.115,1.23]  0 0.00000000
## 3    (1.23,2.34]  3 0.02439024
## 4    (2.34,3.46]  3 0.02439024
## 5    (3.46,4.58]  2 0.01626016
## 6    (4.58,5.69]  4 0.03252033
## 7     (5.69,6.8] 11 0.08943089
## 8     (6.8,7.92] 34 0.27642276
## 9    (7.92,9.04] 12 0.09756098
## 10   (9.04,10.2]  4 0.03252033
## 11   (10.2,11.3]  3 0.02439024
## 12   (11.3,12.4]  7 0.05691057
## 13   (12.4,13.5]  2 0.01626016
```

```
## 14   (13.5,14.6]  6 0.04878049
## 15   (14.6,15.7]  4 0.03252033
## 16   (15.7,16.8]  8 0.06504065
## 17     (16.8,18]  4 0.03252033
## 18     (18,19.1]  9 0.07317073
## 19   (19.1,20.2]  3 0.02439024
## 20   (20.2,21.3]  2 0.01626016
```

6.4 Probability Density Function

We define a quantity at a point x that is the amount of probability per unit distance that we would find in an **infinitesimal** bin dx at x

$$f(x) = \frac{P(x \leq X \leq x + dx)}{dx}$$

$f(x)$ is called the probability **density** function.

Therefore, the probability of observing the random variable X between the observations x and $x + dx$ is given by

$$P(x \leq X \leq x + dx) = f(x)dx$$

Definition

For a continuous random variable X, a **probability density** function is such that

1) The function is positive:

$$f(x) \geq 0$$

2) The probability that X takes **any** value is 1:

$$\int_{-\infty}^{\infty} f(x)dx = 1$$

3) The probability that X is within an interval is the **area under the curve**:

$$P(a \leq X \leq b) = \int_{a}^{b} f(x)dx$$

The properties make sure that $f(x)dx$ satisfy Kolmogorov's properties of a probability measure.

The probability density function is a step forward in the abstraction of probabilities: we add the continuous limit

$$dx \to 0$$

to the already limit of infinite repetitions $N \to \infty$.

All the properties of probabilities are translated in terms of densities

$$\sum \to \int$$

Similar to probability mass functions for discrete variables, probability densities are mathematical quantities that do not necessarily represent random experiments. We are free to define them as long as we respect their properties.

However, a fundamental interest in statistics is to select the densities that better describe our particular random experiment.

6.5 Total area under the curve

Example (raindrop fall)

Take the **probability density** that may describe the random variable that measures where a raindrop falls in a rain gutter of length $100cm$.

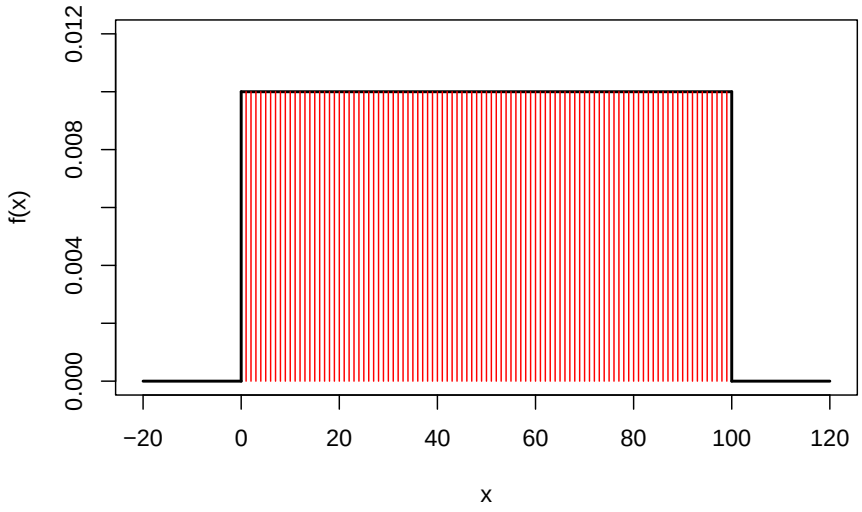
$$f(x) = \begin{cases} \frac{1}{100}, & \text{if } x \in (0, 100) \\ 0, & otherwise \end{cases}$$

Let us verify that the function satisfies the three properties of a probability density.

1) It is evident from the definition that $f(x) \geq 0$

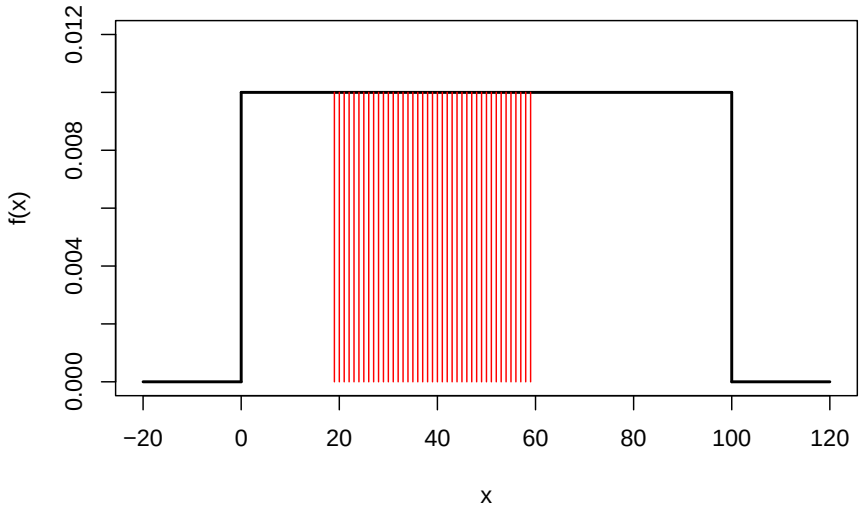
2) The probability of observing **anything**; that is, that X takes any value, is the total **area under the curve**

$P(-\infty \leq X \leq \infty) = \int_{-\infty}^{\infty} f(x)dx = 100 * 0.01 = 1$



3) The probability of observing x in an interval is the **area under the curve** within the interval

- $P(20 \leq X \leq 60) = \int_{20}^{60} f(x)dx = (60 - 20) * 0.01 = 0.4$



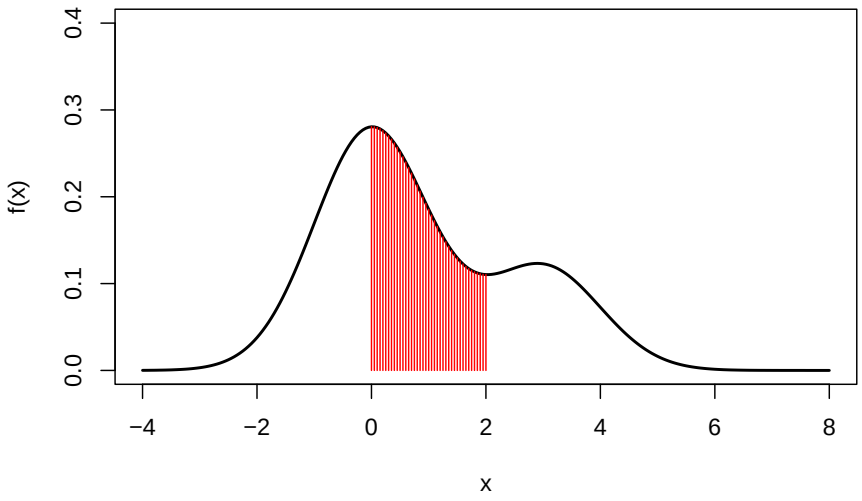
6.6 Probabilities of continous variables

For continuous variables, we compute the probability that the variable is between a and b. That is

$$P(a \leq X \leq b)$$

We saw that for continuous variables, the probability that the experiment gives us a particular real number is zero: $P(X = a) = 0$. The probability $P(a \leq X \leq b)$ is the area under the curve of $f(x)$ between a and b

- $P(a \leq X \leq b) = \int_a^b f(x)dx$



6.7 Probability distribution

The **probability distribution** $F(c)$ is defined as the accumulation of probability up to the outcome c

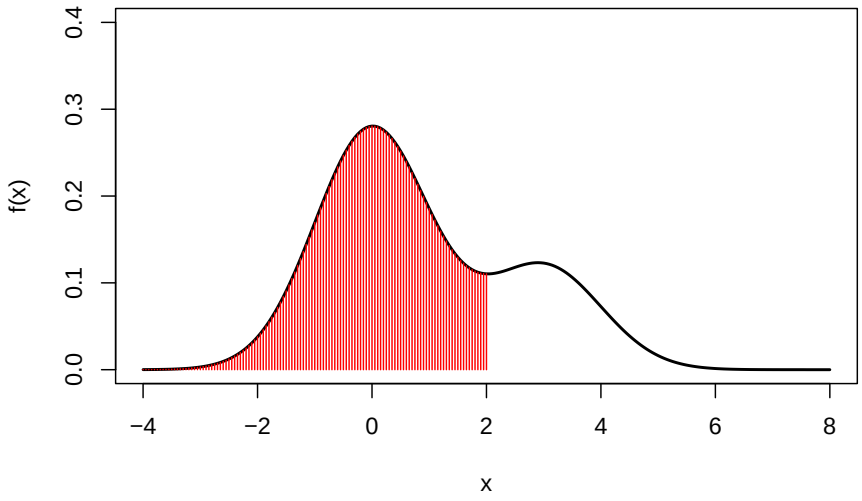
$$F(c) = P(X \leq c)$$

That is the probability that the random variable is less or equal to c. We can use $F(c)$ to compute $P(a \leq X \leq b)$.

Consider that:

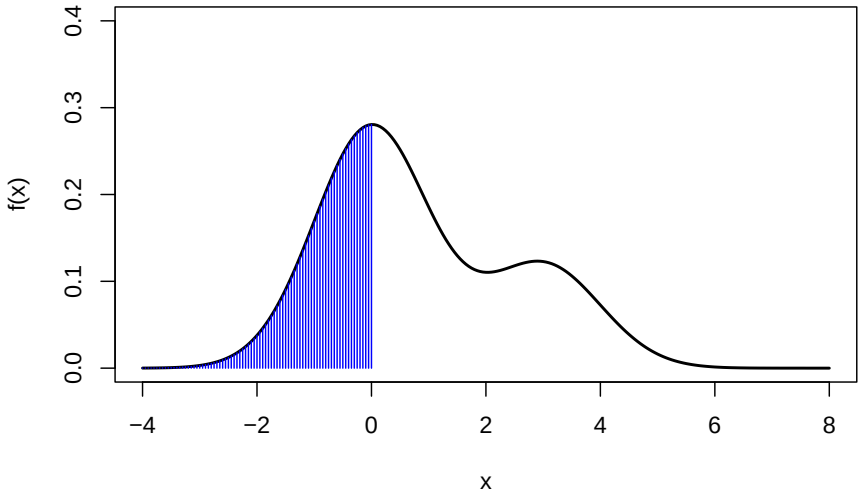
1) The probability accumulated up to b is given by

- $F(b) = P(X \leq b) = \int_{-\infty}^{b} f(x)dx$



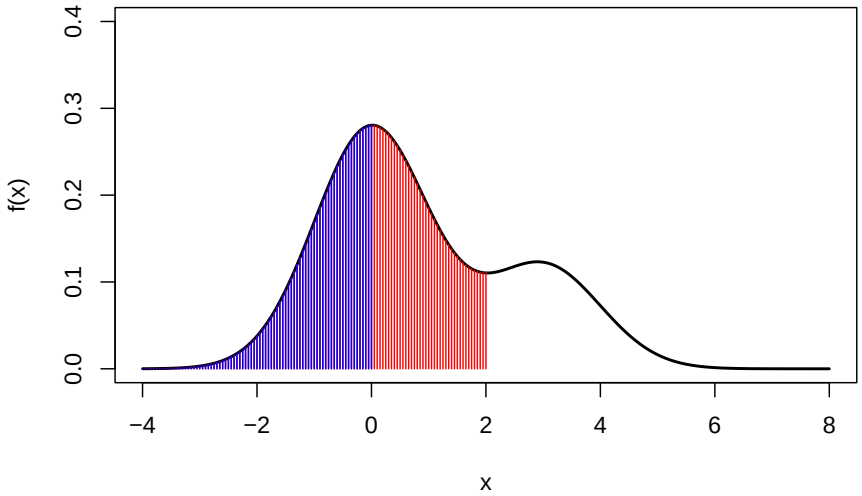
2) That the probability accumulated up to a is

- $F(a) = P(X \leq a)$



Then the probability that the random variable is between a and b is given by the difference in the probability distribution between a and b

- $P(a \leq X \leq b) = \int_a^b f(x)dx = F(b) - F(a)$



Definition

The **probability distribution** of a continuous random variable is defined as the area under the curve up to a

$$F(a) = P(X \leq a) = \int_{-\infty}^{a} f(x)dx$$

$F(a)$ has the properties:

1) It is between 0 and 1:

$$F(-\infty) = 0 \; and \; F(\infty) = 1$$

2) It always increases:

$$F(a) \leq F(b)$$

if $a \leq b$

3) It can be used to compute probabilities:

$$P(a \leq X \leq b) = F(b) - F(a)$$

4) It recovers the probability density:

$$f(x) = \frac{dF(x)}{dx}$$

We use **probability distributions** to **compute probabilities** of a random variable within intervals. $F(x)$ is derivative is the probability density function $f(x)$.

Example (raindrop fall)

For the uniform density function:

$$f(x) = \begin{cases} \frac{1}{100}, & \text{if } x \in (0, 100) \\ 0, & otherwise \end{cases}$$

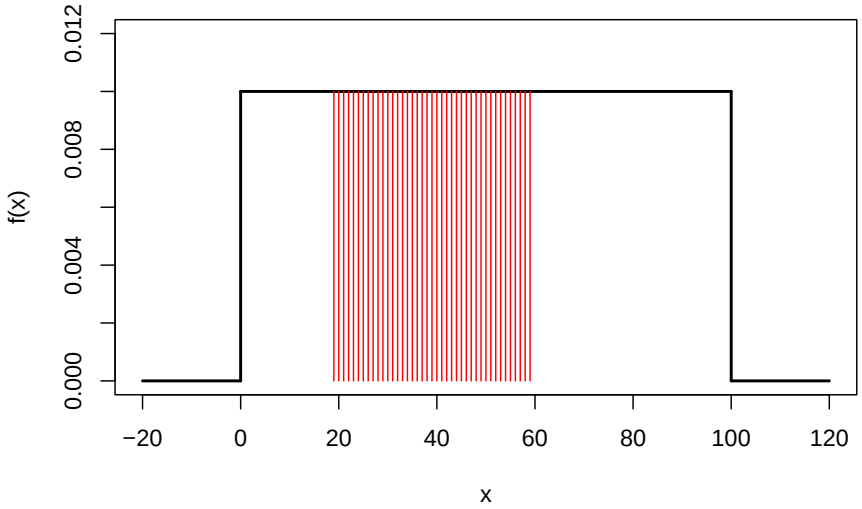
We find that the probability distribution is

$$F(x) = \begin{cases} 0, & x \leq 0 \\ \frac{x}{100}, & \text{if } x \in (0, 100) \\ 1, & 100 \leq x \end{cases}$$

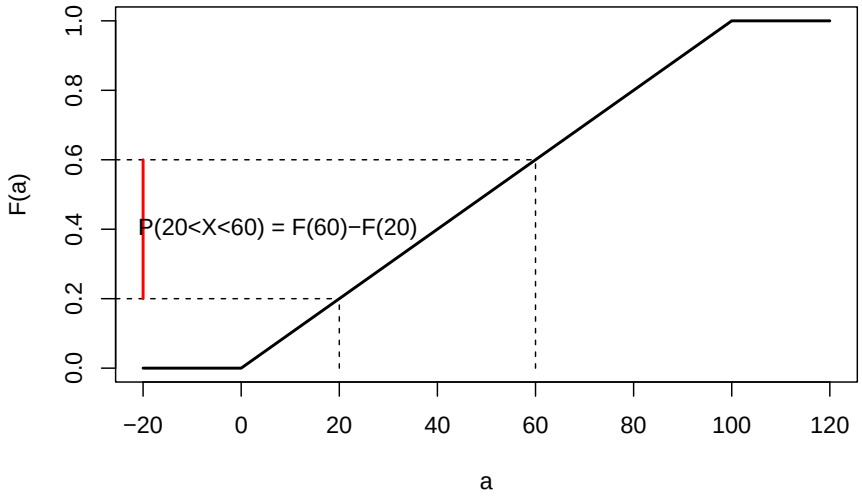
6.8 Probability plots

We can plot the the probability of a random variable in an interval as the *area* under the **density** curve. For instance

$$P(20 < X < 60)$$



Or, equivalenty, we can plot the probability $P(20 < X < 60)$ as the *difference* in **distribution** values



6.9 Mean

As in the discrete case, the **mean** measures the center of mass of probabilities

Definition

Suppose X is a continuous random variable with probability **density** function $f(x)$. The mean or expected value of X, denoted as μ or $E(X)$, is

$$\mu = E(X) = \int_{-\infty}^{\infty} x f(x) dx$$

It is the continuous version of the center of mass.

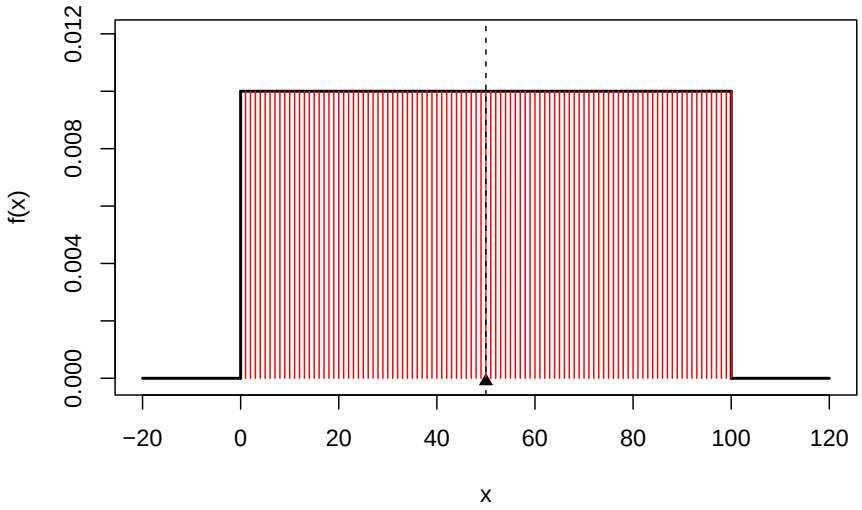
Example (raindrop fall)

The random variable with probability density

$$f(x) = \begin{cases} \frac{1}{100}, & \text{if } x \in (0, 100) \\ 0, & otherwise \end{cases}$$

Has an expected value at

$$E(X) = \int_0^{100} \frac{x}{100} dx = \frac{1}{2} \frac{x^2}{100} |_0^{100} = 50$$



6.10 Variance

As in the discrete case, the variance measures the dispersion of probabilities about the mean

Definition

Suppose X is a continuous random variable with probability density function $f(x)$. The variance of X, denoted as σ^2 or $V(X)$, is

$$\sigma^2 = V(X) = \int_{-\infty}^{\infty} (x-\mu)^2 f(x) dx$$

It is the continuous version of the moment of inertia.

6.11 Functions of X

In many occasions, we will be interested in outcomes that are function of the random variables. Perhaps, we are interested in the square of the elongation of a spring, or on the square root of the temperature of an engine.

Definition

For any function h of a random variable X, with mass function $f(x)$, its expected value is given by

$$E[h(X)] = \int_{-\infty}^{\infty} h(x)f(x)dx$$

From this definition we recover the same properties as in the discrete case:

1) The mean of a linear function is the linear function fo the mean:

$$E(a \times X + b) = a \times E(X) + b$$

for a and b scalars.

2) The variance of a linear function of X is:

$$V(a \times X + b) = a^2 \times V(X)$$

3) The variance about the origin is the variance about the mean plus the mean squared:

$$E(X^2) = V(X) + E(X)^2$$

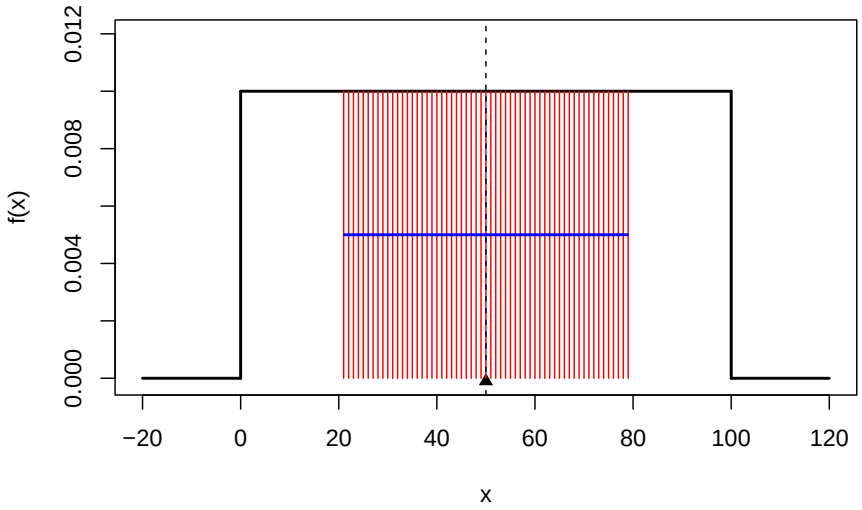
6.12 Exercises

6.12.0.1 Exercise 1

For the probability density

$$f(x) = \begin{cases} \frac{1}{100}, & \text{if } x \in (0, 100) \\ 0, & otherwise \end{cases}$$

- compute the mean (A:50)
- compute the variance using $E(X^2) = V(X) + E(X)^2$ (A:100^2/12)
- compute $P(\mu - \sigma \leq X \leq \mu + \sigma)$ (A: 0.57)
- What are the first and third quartiles? (A: 25; 75)



6.12.0.2 Exercise 2

Given

$$f(x) = \begin{cases} 0, & x < 0 \\ ax, & x \in [0,3] \\ b, & x \in (3,5) \\ \frac{b}{3}(8-x), & x \in [5,8] \\ 0, & x > 8 \end{cases}$$

- What are the values of a and b such that $f(x)$ is a continous probability density function? (A: 1/15; 1/5)
- what is the mean of X? (A:4)

6.12.0.3 Exercise 3

For the probability density

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & \text{if } x \geq 0 \\ 0, & otherwise \end{cases}$$

- Confirm that this is a probability density
- Compute the mean (A: $1/\lambda$)
- Compute the expected value of X^2 (A: $2/\lambda^2$)
- Compute variance (A: $1/\lambda^2$)

- Find the probability distribution $F(a)$ (A: $1 - exp(-\lambda a)$)
- Find the median (A: $\log 2/\lambda$)

6.12.0.4 Exercise 4

Given the cumulative distribution for a random variable X

$$F(x) = \begin{cases} 0, & x < -1 \\ \frac{1}{80}(17 + 16x - x^2), & x \in [-1, 7) \\ 1, & x \geq 7 \end{cases}$$

compute:

- $P(X > 0)$ (A:63/80)
- $E(X)$ (A:1.93)
- $P(X > 0 | X < 2)$ (A:28/45)

Chapter 7

Discrete Probability Models

7.1 Objective

In this chapter we will see some probability mass functions that are used to describe common random experiments.

We will introduce the concept of parameter and of parametric models.

In particular, we will discuss the uniform and Bernoulli probability functions and how they are used to derive the binomial and negative binomial probability models.

7.2 Probability mass function

Let us remember that a probability mass function of a **discrete random variable** X with possible values $x_1, x_2, .., x_M$ is **any function** such that

1) It allows us to compute probabilities for all outcomes

$$f(x_i) = P(X = x_i)$$

2) It is always positive:

$$f(x_i) \geq 0$$

3) The probability that the random variable takes any outcome in a run of the random experiment is 1

$$\sum_{i=1}^{M} f(x_i) = 1$$

We studied two important **properties:**

1) The mean as a measure of central tendency:

$$E(X) = \sum_{i=1}^{M} x_i f(x_i)$$

2) The variance as a measure of dispersion:

$$V(X) = \sum_{i=1}^{M} (x_i - \mu)^2 f(x_i)$$

7.3 Probability model

A **probability model** is a probability mass function that may represent the probabilities of a random experiment.

Examples

1) The probability mass function defined by

X	$f(x)$
-2	$1/8$
-1	$2/8$
0	$2/8$
1	$2/8$
2	$1/8$

Represents the probability of drawing **one** ball from an urn where there are two balls with labels: $-1, 0, 1$ and one ball with labels: $-2, 2$.

2) $f(x) = P(X = x) = 1/6$ represents the probability of the outcomes of **one** throw of a dice.

7.4 Parametric models

When we have a random experiment with M possible outcomes, we need to find M numbers to determine the probability mass function. As in the first example above, we needed 5 values in the column $f(x)$ of the probability table.

However, **in many cases**, we can formulate probability functions $f(x)$ that depend on **very few** numbers only. As in the second example above, we only needed to know how many possible outcomes the throw of a dice has.

Example (classical probability)

A random experiment with M equally likely outcomes has a probability mass function:

$$f(x) = P(X = x) = 1/M$$

We only need to know M.

The numbers we **need to know** to fully determine a probability function are called **parameters**.

7.5 Uniform distribution (one parameter)

The previous example is the classical interpretation of probability, and defines our first parametric model.

Definition

A random variable X with outcomes $\{1, ...M\}$ has a discrete **uniform distribution** if all its M outcomes have the same probability

$$f(x) = \frac{1}{M}$$

M is the natural parameter of the model. Once we define M for an experiment, we choose a particular probability mass function. The function above is really a family of probability mass functions that depend on M: $f(x; M)$.

The mean and variance of a variable that follows a uniform distribution are:

$$E(X) = \frac{M+1}{2}$$

and

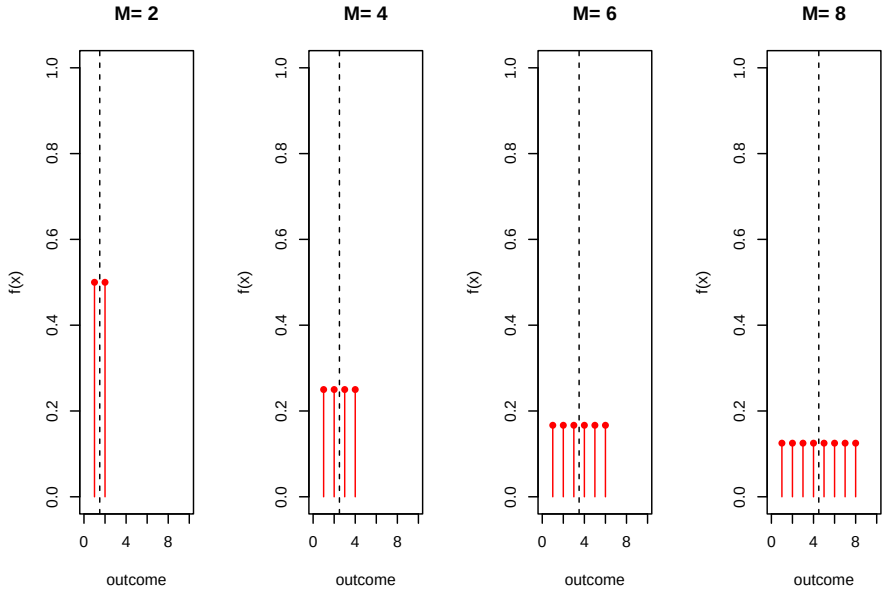
$$V(X) = \frac{M^2 - 1}{12}$$

which are derived from the definitions.

Note that $E(X)$ and $V(X)$ are also **parameters**. If we know any of them then we can fully determine the distribution. Using the equations above, we can have three different parametrizations of the uniform distribution

$$f(x) = \frac{1}{M} = \frac{1}{2E(X) - 1} = \frac{1}{\sqrt{12V(X) + 1}}$$

The two last are cumbersome. The first one is natural and the parameter M has the simple interpretation of the number of possible outcomes. Let's look at some probability mass functions in the family of uniform parametric models. Here are four members of the family, each characterized by a different M



7.6 Uniform distribution (two parameters)

Let's introduce a new **uniform** probability model with **two parameters**: The minimum and maximum outcomes.

If the random variable takes values in $\{a, a+1, ...b\}$, where a and b are integers and all the outcomes are equally probable then

$$f(x) = \frac{1}{b-a+1}$$

as the total number of outcomes is $M = b - a + 1$.

If the random variable has the probability mass function $f(x)$ above, we then say that X distributes uniformly between a and b and write

$$X \to Unif(a, b)$$

Properties:

If X distributes uniformly between a and b

$$X \to Unif(a, b)$$

1) Its mean is

$$E(X) = \frac{b+a}{2}$$

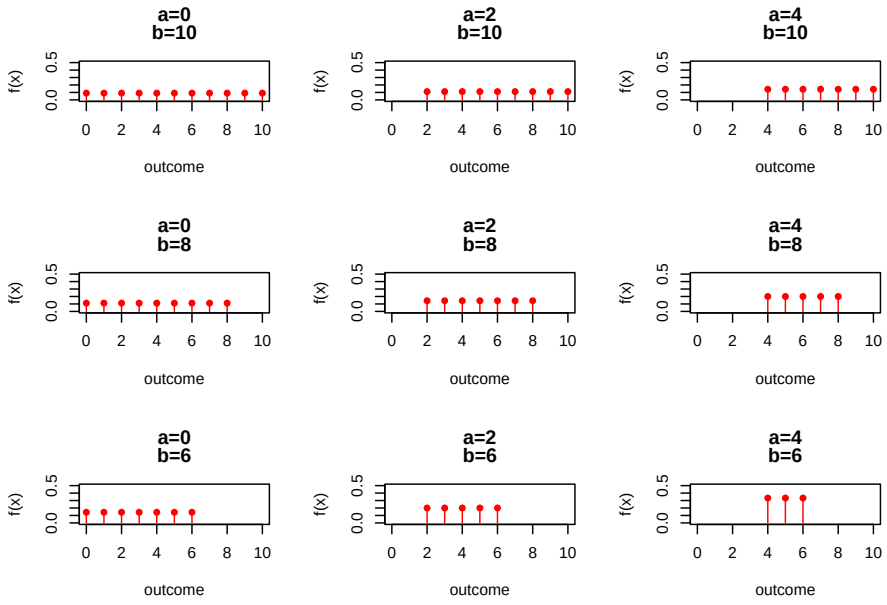
2) Its variance is

$$V(X) = \frac{(b-a+1)^2 - 1}{12}$$

To prove this, change variables $X = Y + a - 1$, $y \in \{1, ...M\}$, and apply the properties of the mean and variance of a a linear function.

Probability mass functions

Let's look at some probability mass functions in the family of uniform parametric models:



Example (school classes)

What is the probability of observing a child of a particular age in a primary school (if all classes have the same amount of children)?

From the set up of the experiment we know: $a = 6$ and $b = 11$ then

$$X \to Unif(a = 6, b = 11)$$

that is

$$f(x) = \frac{1}{6}$$

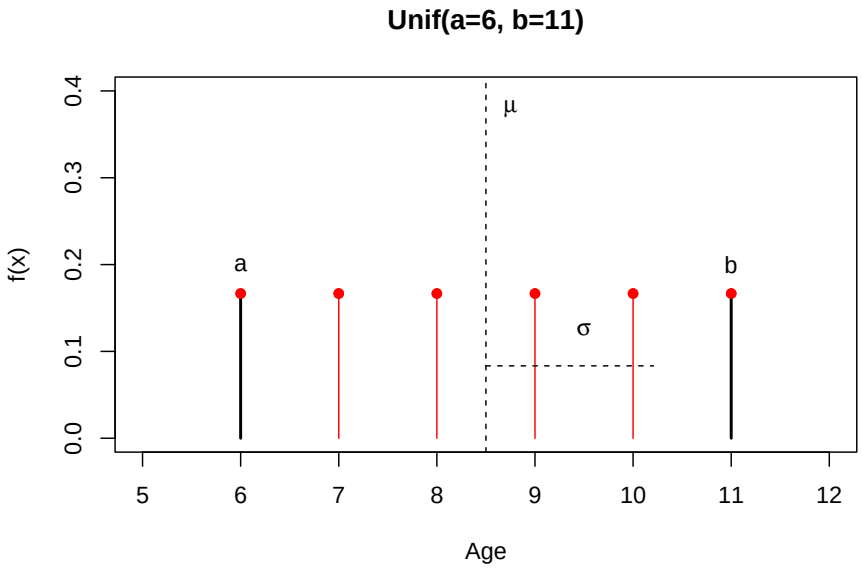
for $x \in \{6, 7, 8, 9, 10, 11\}$, and 0 otherwise.

The mean and variance for this probability mass function is:

- $E(X) = 8.5$
- $V(X) = 2.916667$

Remember that

- The expected value is the **mean** $\mu = 8.5$
- The **standard deviation** $\sigma = 1.707825$ is the average distance from the mean and is computed from the square root of the variance.



Parameters and Models:

A **model** is a particular function $f(x)$ that **describes** our experiment.

If the model is a **known** function that depends on a few parameters then changing the values of the parameters, we produce a **family of models**: $f(x; a, b)$.

Knowledge of $f(x)$ is reduced to the knowledge of the value of the parameters a, b.

Ideally, the model and the parameters are **interpretable**.

In our example, a represents the the minimum age at school and b the maximum age. They can be considered as the **physical properties** of the experiment.

7.7 Bernoulli trial

Let's try to advance from the equal probability case and suppose a model with only two possible outcomes (A and A') that have **unequal** probabilities

Examples:

- Writing down the sex of a patient who goes into an emergency room of a hospital ($A : male$ and $A' : female$).
- Recording whether a manufactured machine is defective or not ($A : defective$ and $A' : not\ defective$).
- Hitting a target ($A : success$ and $A' : failure$).
- Transmitting one pixel correctly ($A : yes$ and $A' : no$).

In these examples, the probability of outcome A is usually **unknown**.

Probability model:

We will introduce the probability of an outcome (A) as the **parameter** of the model. The model can be written in different forms

1) As a probability table:

Outcome	*Probability*
A'	$1-p$
A	p

- outcome A' (failure/outcome of no interest): has a probability $1-p$
- outcome A (success/outcome of interest): has probability p (parameter)

2) As a probability mass function of the random variable K taking values $\{0, 1\}$ for A' and A, respectively.

$$f(k) = \begin{cases} 1-p, & k = 0\,(event\ A') \\ p, & k = 1\,(event\ A) \end{cases}$$

3) As a concise probability mass function

$$f(k; p) = p^k (1-p)^{1-k}$$

for $k \in \{0, 1\}$.

We then say that X follows a Bernoulli distribution with parameter p

$$X \to Bernoulli(p)$$

Properties:

If X follows a Bernoulli distribution then

1) its mean is

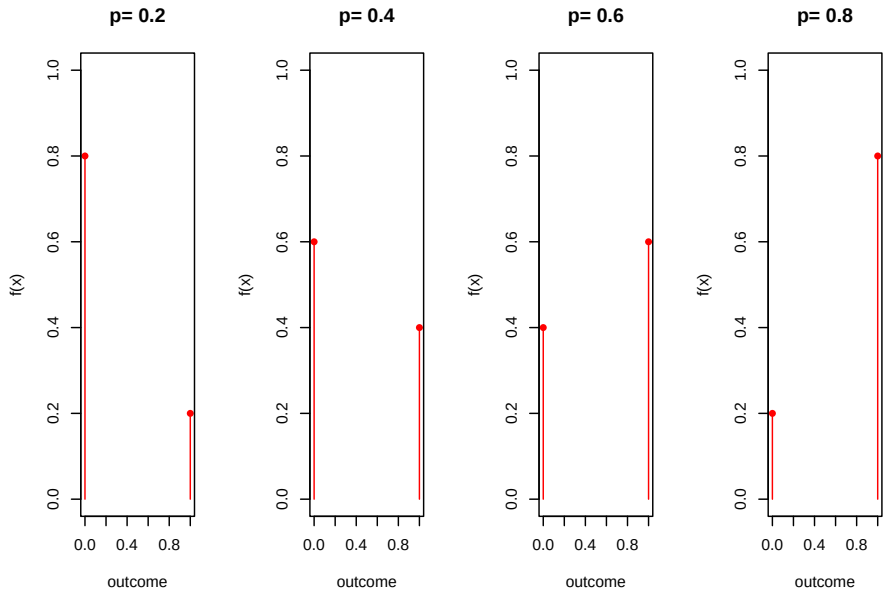
$$E(K) = p$$

2) its variance is

$$V(K) = (1-p)p$$

Note that the parameter p is the probability of the outcome A, which is the same as the value of the probability mass function at 1: $p = P(A) = P(K = 1) = f(1)$. The parameter fully determines the probability mass function, including its mean and variance.

Let's look at some probability mass functions in the family of uniform parametric models:



7.8 Binomial experiment

When we are interested in predicting **absolute frequencies** of the event A, and we know the probability of A; that is the parameter p of the Bernoulli trial, we

1) **Repeat** the Bernoulli trial n times and count how many times we obtained A using the absolute frequency of A: N_A.

2) Define the **Random variable** $X = N_A$ taking values $x \in 0, 1, ...n$

When we repeat n times a Bernoulli trial, we obtain one value for n_A. If we perform other n Bernoulli trials then n_A changes its value. Therefore, $X = N_A$ is a random variable and $x = n_A$ is its observation. We are interested in the probability that the absolute frequency takes a given value.

Examples (Some binomial experiments):

- Writing down the sex of $n = 10$ patients who go into an emergency room of a hospital. What is the probability that $X = 9$ patients are men when $p = 0.8$?
- Trying $n = 5$ times to hit a target ($A : success$ and $A' : failure$). What is the probability that I hit the target $X = 5$ times when I usually hit it 25% of the times ($p = 0.25$)?
- Transmitting $n = 100$ pixels with errors ($A : error$ and $A' : correct$). What is the probability that $X = 2$ pixels are errors, when the probability of error is $p = 0.1$?

7.9 Binomial probability function

Let us suppose that **we know** the real value of the parameter of the Bernoulli trial p.

When we repeat a Bernoulli trial and stop at n, is the value x, the total number of events A, common or rare? what is its probability mass function $P(X = x) = f(x)$?

Example (transmission of pixels)

What is the probability of observing $X = 2$ errors when transmitting $n = 4$ pixels, if the probability of a single error is $p = 0.1$?

Let us consider that

1) A random variable of the **transmission experiment** is the vector

$$(K_1, K_2, K_3, K_4)$$

where one observation may be $(K_1 = 0, K_2 = 1, K_3 = 0, K_4 = 1)$ or $(0, 1, 0, 1)$. That is a particular transmission with an error in the second and in the fourth pixels, and no errors in the first and third pixels. If we repeat the transmission, we will obtain another 4-element set of zeros and ones.

2) The transmission of each pixel is a Bernoulli trial

$$K_i \rightarrow Bernoulli(p)$$

$k_i \in \{0, 1\}$

3) $X = N_A$ can be computed as the sum

$$X = \sum_{i=1}^{4} K_i$$

Therefore, the possible number of errors in the transmission are values $x \in \{0, 1, 2, 3, 4\}$. For example X takes the value 2 ($X = 2$) for the outcome $(0, 1, 0, 1)$ because $x = 0 + 1 + 0 + 1$.

Now let's see the probabilities of some **number of errors** and then we will generalize them.

1) What is the probability of observing 4 **errors**?

The probability of observing 4 errors is the probability of observing an error in the 1^{st} **and** 2^{nd} **and** 3^{rd} **and** 4^{th} pixel:

$$P(X = 4) = P(1, 1, 1, 1) = p \times p \times p \times p = p^4$$

because K_i are **independent** the probabilities of the errors at each pixel multiply.

2) What is the probability of observing 0 **errors**?

The probability of 0 errors is the joint probability of observing **no error** in **any** transmission:

$$P(X = 0) = P(0, 0, 0, 0) = (1 - p)(1 - p)(1 - p)(1 - p) = (1 - p)^4$$

3) What is the probability of observing 3 **errors**?

The probability of 3 errors is the **addition** of the probabilities of observing 3 errors in **different transmissions**:

$$P(X = 3) = P(0, 1, 1, 1) + P(1, 0, 1, 1) + P(1, 1, 0, 1) + P(1, 1, 1, 0) = 4p^3(1 - p)^1$$

because all off these events are **mutually exclusive**.

4) Therefore the probability of x **errors** is

$$f(x) = \begin{cases} 1 \times p^0(1 - p)^4, & x = 0 \\ 4 \times p^1(1 - p)^3, & x = 1 \\ 6 \times p^2(1 - p)^2, & x = 2 \\ 4 \times p^3(1 - p)^1, & x = 3 \\ 1 \times p^4(1 - p)^0, & x = 4 \end{cases}$$

or more shortly

$$f(x) = \binom{4}{x} p^x (1-p)^{4-x}$$

for $x = 0, 1, 2, 3, 4$

where $\binom{4}{x}$ is the number of **possible outcomes** (transmissions of 4 pixels) with x errors.

Definition:

The **binomial probability** function is the probability mass function of observing x outcomes of type A in n independent Bernoulli trials, where A has the same probability p in each trial.

The function is given by

$f(x) = \binom{n}{x} p^x (1-p)^{n-x}$, $x = 0, 1, ...n$

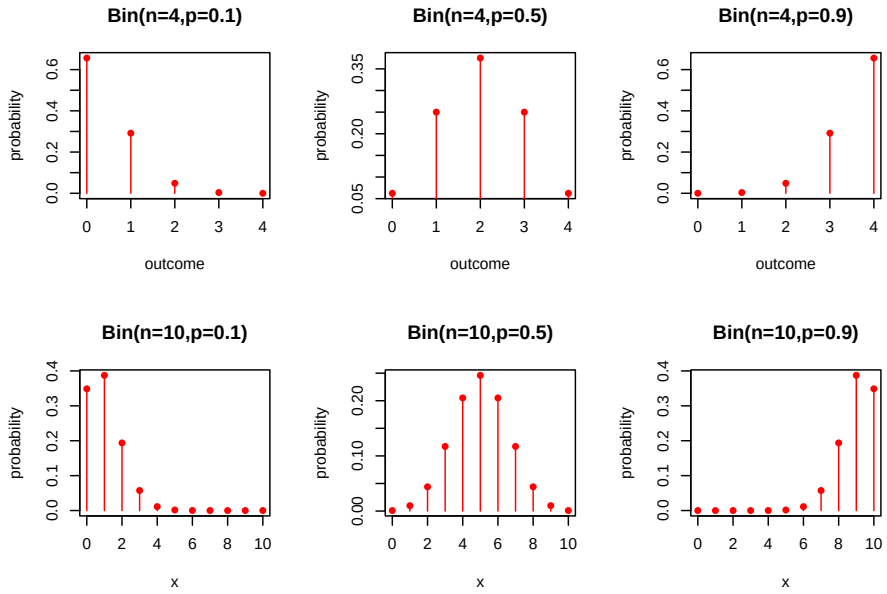
$\binom{n}{x} = \frac{n!}{x!(n-x)!}$ is called **the binomial coefficient** and gives the number of ways one can obtain x events of type A in a set of n.

When a variable X has a binomial probability function we say it distributes binomially and write

$$X \to Bin(n, p)$$

where n and p are parameters.

Let's look at some probability mass functions in the family of binomial parametric models:



Properties:

If a random variable distributes binomially $X \to Bin(n, p)$ then

1) its mean is

$$E(X) = np$$

2) its variance is

$$V(X) = np(1-p)$$

These properties can be demonstrated by the fact that X is the sum of n independent Bernoulli variables. Therefore,

$E(X) = E(\sum_{i=1}^{n} K_i) = np$

and

$V(X) = V(\sum_{i=1}^{n} K_i) = n(1-p)p$

The last equation requieres independence of the Bernoulli trials.

Example (transmission of pixels)

- The expected value for the number of errors in the transmission of 4 pixels is $np = 4 \times 0.1 = 0.4$ when the probability of an error is 0.1.
- The variance is $n(1-p)p = 0.36$

- What is the probability of observing 4 errors?

Since we are repeating a Bernoulli trial $n = 4$ times and counting the number of events of type A (errors), when $P(A) = p = 0.1$ then

$$X \to Bin(n = 4, p = 0.1)$$

That is

$$f(x) = \binom{4}{x} 0.1^x (1 - 0.1)^{4-x}$$

$P(X = 4) = f(4) = \binom{4}{4} 0.1^4 0.9^0 = 0.1^4 = 10^{-4}$

In Python:

```
from scipy.stats import binom
binom.pmf(4,4,0.1)
```

- What is the probability of observing 2 errors?

$P(X = 2) = \binom{4}{2} 0.1^2 0.9^2 = 0.0486$

In Python:

```
binom.pmf(2,4,0.1)
```

Example (opinion polls)

- What is the probability of observing **at most** 8 voters of the ruling party in an election poll of size 10, if the probability of a positive vote for the party is 0.9

For this case

$$X \to Bin(n = 10, p = 0.9)$$

That is

$$f(x) = \binom{10}{x} 0.9^x (0.1)^{10-x}$$

We want to compute: $P(X \leq 8) = F(8) = \sum_{i=1}^{8} f(x_i) = 0.2639011$

in Python binom.cdf(8,10, 0.9)

7.10 Negative binomial probability function

Now let us imagine that we are interested in counting the well-transmitted pixels (A') before a **given number** of errors occur. Say we can **tolerate** r errors in the transmission.

Our random experiment is now: Repeat Bernoulli trials until we observe the outcome A appears r times. Stop and count how many A' you have got.

The outcome of the experiment is the number of A' events before r A's occur, that is the frequency $n_{A'} = y$

We are interested in finding the probability of observing a particular number of events A', $P(Y = y)$, where $Y = N_{A'}$ is the random variable.

Example (transmission of pixels)

What is the probability of observing y well-transmitted (A') pixels before r errors (A)?

Let's first find the probability of **one particular transmission event** with y number of correct pixels (A') and r number of errors (A).

$$(0, 0, 1, ., 0, 1, ...0, 1)$$

where we consider that there are y zeros and r ones. Therefore, we observe y correct pixels in a total of $y + r$ pixels.

The probability of this event is:

$$P(0, 0, 1, ., 0, 1, ...0, 1) = p^r(1 - p)^y$$

Remember that p is the probability of error (A).

How many **transmissions events** can have y correct pixels (0's) before r errors (1's)?

Note that

1) The last pixel is fixed (marks the end of transmission)
2) The total number of ways in which y number of zeros can be allocated in $y + r - 1$ pixels (the last pixel is fixed with value 1) is: $\binom{y+r-1}{y}$

Therefore, the probability of observing y 1's before r 0's (each 1 with probability p) is

$$P(Y = y) = f(y) = \binom{y + r - 1}{y} p^r(1 - p)^y$$

for $y = 0, 1, ...$

We then say that Y follows a negative binomial distribution and we write

$$Y \to NB(r, p)$$

where r and p are parameters representing the tolerance and the probability of a single error (event A).

Properties:

A random variable $Y \to NB(r, p)$ has

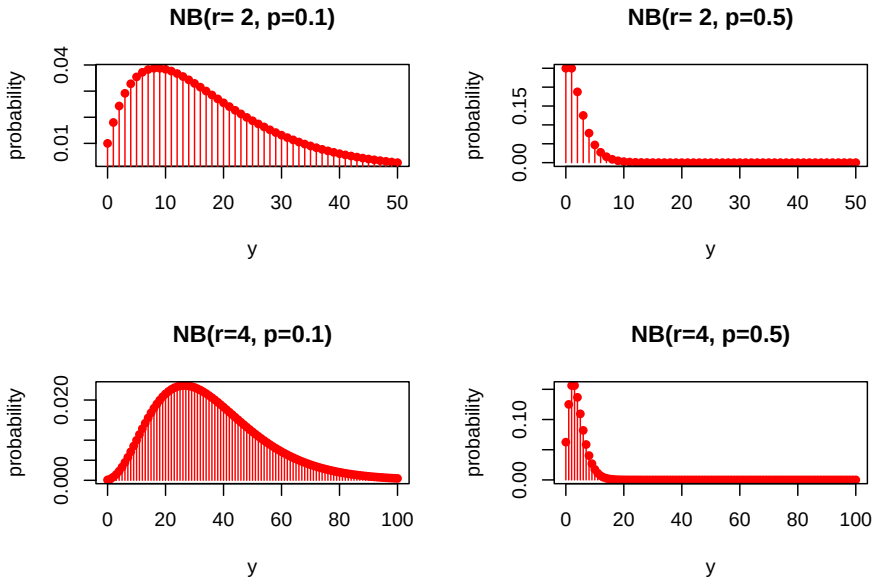
1) mean

$$E(Y) = r\frac{1-p}{p}$$

2) and variance

$$V(Y) = r\frac{1-p}{p^2}$$

Let's look at some probability mass functions in the family of negative binomial parametric models:



Example (website)

A website has three servers. One server operates at a time and, only when a request fails, another server is used.

If the probability of failure for a request is known to be $p = 0.0005$ then

- What is the expected number of successful requests before the three computers fail?

Since we are repeating a Bernoulli trial until $r = 3$ events of type A (failure) are observed (each with $P(A) = p = 0.0005$) and are counting the number of events of type A' (successful requests) then

$$Y \to NB(r = 3, p = 0.0005)$$

Therefore, the expected number of requests before the system fails is:

$E(Y) = r\frac{1-p}{p} = 3\frac{1-0.0005}{0.0005} = 5997$

Note that there are actually 6000 trials.

- What is the probability of dealing with at most 5 successful requests before the system fails?

We therefore want to compute the probability distribution at 5:

$F(5) = P(Y \leq 5) = \Sigma_{y=0}^{5} f(y)$

$= \sum_{y=0}^{5} \binom{y+2}{y} 0.0005^r 0.9995^y$

$= \binom{2}{0}0.0005^3 0.9995^0 + \binom{3}{1}0.0005^3 0.9995^1$

$+\binom{4}{2}0.0005^3 0.9995^2 + \binom{5}{3}0.0005^3 0.9995^3$

$+\binom{6}{4}0.0005^3 0.9995^4 + \binom{7}{5}0.0005^3 0.9995^5$

$= 6.9 \times 10^{-9}$

In Python this is computed with

```
from scipy.stats import nbinom
nbinom.cdf(5,3,0.0005)
```

Examples

- What is the probability of observing 10 correct pixels before 2 errors, if the probability of an error is 0.1?

$f(10; r = 2, p = 0.1) = 0.03835463$

in Python: nbinom.pmf(10, 2, 0.1)

- What is the probability that 2 girls enter the class before 4 boys if the probability that a girl enters is 0.5?

$f(2; r = 4, p = 0.5) = 0.15625$

in Python: nbinom.pmf(2, 4, 0.5)

7.11 Geometric distribution

We call **geometric distribution** to the **negative binomial** distribution with $r = 1$

The probability of observing A' events before observing the **first** event of type A is

$$P(Y = y) = f(y) = p(1-p)^y$$

$$Y \to Geom(p)$$

which has

1) mean
$$E(Y) = \frac{1-p}{p}$$

2) and variance
$$V(Y) = \frac{1-p}{p^2}$$

7.12 Hypergeometric model

The **hypergeometric model** arises when we want to count the number of events of type A that are drawn from a finite population.

The general model is to consider N total balls in a urn. Mark K with label A and $N - K$ with label A'. Take out n balls one by one with no replacement in the urn, and then count how many A's you obtained.

The **Binomial** model can be derived from the **Hypergeometric** model when we consider that either N is infinite, or that every time we draw a ball we replace it back in the urn.

Example (chickenpox)

A school of $N = 600$ children has an epidemic of chickenpox. We tested $n = 200$ children and observed that $x = 17$ were positive. If we knew that a total of $K = 64$ were really infected in the school, what is the probability of our observation?

Definition:

The probability of obtaining x cases (type A) in a sample of n drawn from a population of N where K are cases (type A).

$P(X = x) = P(one\,sample) \times (Number\,of\,ways\,of\,obtaining\,x)$

$$= \frac{1}{\binom{N}{n}} \binom{K}{x} \binom{N-K}{n-x}$$

where $k \in \{\max(0, n+K-N), \ldots \min(K, n)\}$

We then say that X follows a hypergeometric distribution and write

$$X \to Hypergeometric(N, K, n)$$

The hypergeometric model has three parameters.

Properties:

If $X \to Hypergeometric(N, K, n)$ then it has

1) mean

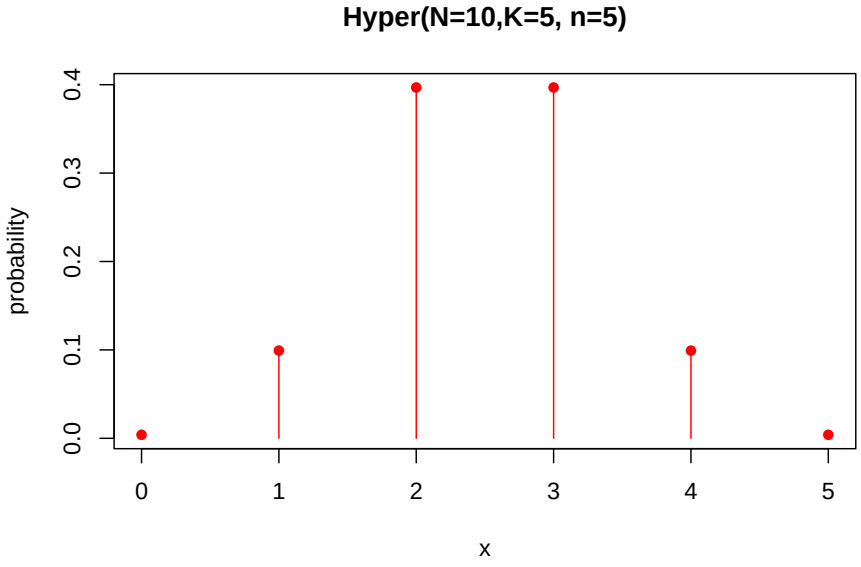
$$E(X) = n\frac{K}{N} = np$$

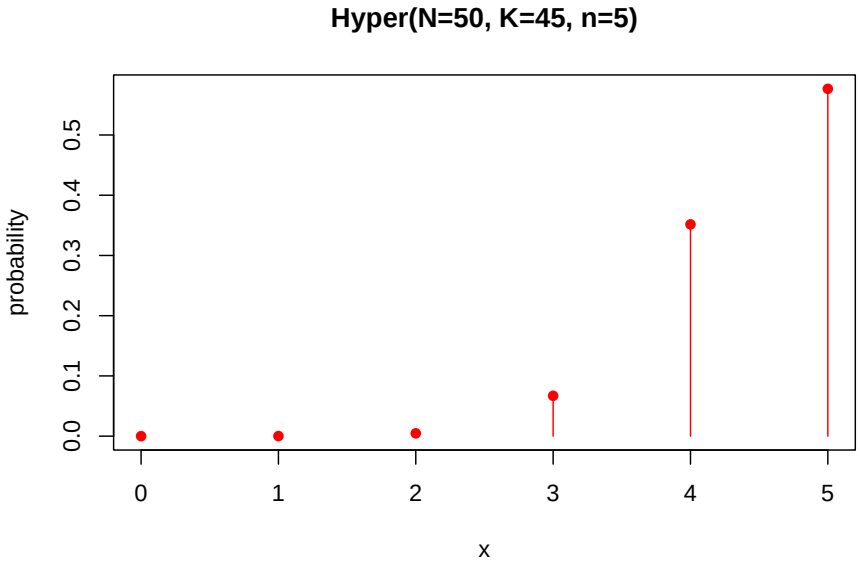
2) and variance

$$V(X) = np(1-p)\frac{N-n}{N-1}$$

when $p = \frac{K}{N}$ is the proportion of hepatitis C in a population of size N. Note that when $N \to \infty$ then we recover the binomial properties.

Let's look at some probability mass functions in the family of hypergeometric parametric models:





Example (chickenpox)

- what is the probability of infections less or equal than 17 children in a sample of 200, drawn from a population of 600 where 64 are infected?

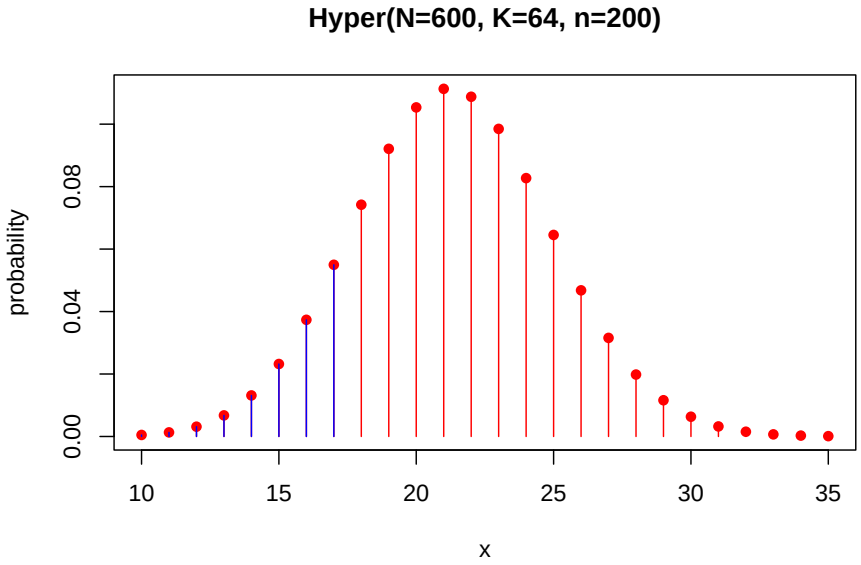
The probability that we need to compute is $P(X \leq 17) = F(17)$

where $X \rightarrow Hypergeometric(N = 600, K = 64, n = 200)$

in Python:

```
from scipy.stats import hypergeom
hypergeom.cdf(17, 600, 64, 200)
```

The solution is the addition of the blue needles in the plot.



7.13 Questions

1) What is the expected value and the variance of the number of failures in 100 prototypes, when the probability of a failure is 0.25

a: 0.25, 0.1875; **b:** 25, 0.1875; **c:** 0.25, 18.75; **d:** 25, 18.75

2) The number of available tables at lunch time in a restaurant is better described by which probability model

a: Binomial; **b:** Uniform; **c:** Negative Binomial; **d:** Hypergeometric

3) The expected value of a Binomial distribution is not

a: n times the expected value of a Bernoulli; **b:** the expected value of a Hypergeometric, when the population is very big; **c:** np; **d:** the limit of the relative frequency when the number of repetitions is large

4) Opinion polls for the USA 2022 election give a probability of 0.55 that a voter favors the republican party. If we conduct our own poll and ask 100 random people on the street, How would you compute the probability that in our poll democrats win the election?

a:binom.cdf(49, 100, 0.55)=0.13; **b:**1-binom.cdf(49, 100, 0.55)=0.86; **c:**binom.cdf(51, 100, 0.45)=0.90; **d:**1-binom.cdf(51, 100, 0.45)=0.095

5) In an exam a student chooses at random one of the four answers that he does not know. If he doesn't know 10 questions what is the probability that more than 5 (> 5) questions are correct?

a:binom.pmf(5, 10, 0.25)~ 0.05; **b:**binom.cdf(5, 10, 0.75)~ 0.07; **c:**binom.pmf(5, 10, 0.75)~ 0.05; **d:**1-binom.cdf(5, 10, 0.25)~ 0.01

7.14 Exercises

7.14.0.1 Exercise 1

If 20% of the bolts produced by a machine are defective, determine the probability that, out of 4 bolts chosen at random

- 1 bolts will be defective (A:0:4096)
- 0 bolts will be defective (A::4096)
- at most 2 bolts will be defective (A:0:9728)

7.14.0.2 Exercise 2

In a population, the probability that a baby boy is born is $p = 0.51$. Consider a family of 4 children

- What is the probability that a family has only one boy?(A: 0.240)
- What is the probability that a family has only one girl?(A: 0.259)
- What is the probability that a family has only one boy or only one girl?(A: 0.4999)
- What is the probability that the family has at least two boys?(A: 0.7023)
- What is the number of children that a family should have such that the probability of having at least one girl is more than 0.75?(A:$n = 3 > \log(0.25)/\log(0.51)$)

7.14.0.3 Exercise 3

A search engine fails to retrieve information with a probability 0.1

- If we system receives 50 search requests, what is the probability that the system fails to answer three of them?(A:0.1385651)
- What is the probability that the engine successfully completes 15 searches before the first failure?(A:0.020)
- We consider that a search engine works sufficiently well when it is able to find information for moe than 10 requests for every 2 failures. What is the probability that in a reliability trial our search engine is satisfactory?(A: 0.697)

Chapter 8

Poisson and Exponential Models

8.1 Objective

In this chapter we will see two tightly related probability models: the **Poisson** and the **exponential** models.

The Poisson model is for **discrete** random variables while the exponential function is **continuous** random variables. They are closely related. We will see that from the same random experiment we may ask different questions that will lead to either one or the other model.

8.2 Discrete probability models

In the previous chapter we built complex models from simple ones. At each stage, we introduced some novel concept:

Uniform: Classical interpretation of probability

↓

Bernoulli: Introduction of the probability p as **parameter** (family of models)

↓

Binomial: Introduction of the **repetition** of a random experiment (n-times Bernoulli trials)

↓

Poisson: Repetition of random experiment within a continuous interval, having **no control** on when/where the Bernoulli trial occurs.

The last stage is the Poisson process that describes a the repetition of a random experiment with additional randomness at the time of repetition.

8.3 Poissson experiment

Imagine that we are observing events that **depend** on time or distance **intervals**.

for example:

- cars arriving at a traffic light
- getting messages on a mobile phone
- impurities occurring at random in a copper wire

Suppose that the events are outcomes of **independent** Bernoulli trials each appearing randomly on a continuous interval, and we want to **count** them.

What is the probability of observing X events in an interval's unit (time or distance)?

Example (Impurities in a wire):

Imagine that some impurities deposit randomly along a copper wire. You want to count the number of impurities (X) in one given centimeter of the wire.

Consider that you know that on average there are 10 impurities per centimeter $\lambda = 10/cm$.

What is the probability of observing $X = 5$ impurities in one particular wire specimen of one centimeter?

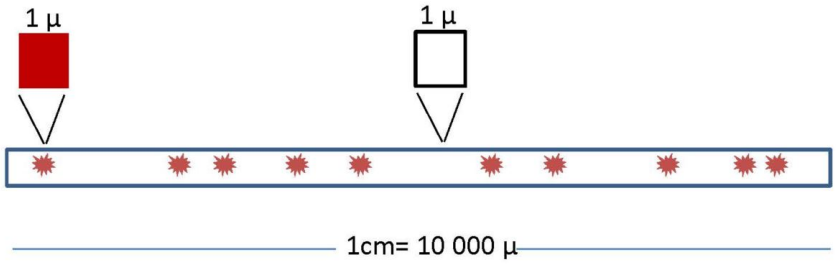
8.4 Poisson probability mass function

To calculate the probability mass function $f(x) = P(X = x)$ of the previous example we divide the centimeter into micrometers ($0.0001cm$).

Micrometers are small enough so

1) either there is or there is not an impurity in each micrometer

2) each micrometer can be then be considered a **Bernoulli trial**



From the Binomial to the Poisson probability function

The probability of observing X impurities in $n = 10,000\mu$ (1cm) approximately follows a binomial distribution

$$f(x) \sim \binom{n}{x} p^x (1-p)^{n-x}$$

where p is the probability of finding an impurity in a micrometer.

Since the expected value of a Binomial variable is: $E(X) = np$. This is the average number of impurities per 1cm or $\lambda = np$ which we know, and consider that it is a property of the material. Therefore, substitute $p = \lambda/n$

$$f(x) \sim \binom{n}{x} (\frac{\lambda}{n})^x (1 - \frac{\lambda}{n})^{n-x}$$

Since, there **could** still be two impurities in a micrometer, we need to increase the partition of the wire and $n \to \infty$.

Then in the limit:

$$f(x) = \frac{e^{-\lambda}\lambda^x}{x!}$$

λ is constant. It is the density of impurities per centimeter, a **physical property** of the system. λ is therefore the **parameter** of the probability model.

Derivation details:

For $f(x) \sim \binom{n}{x}(\frac{\lambda}{n})^x(1 - \frac{\lambda}{n})^{n-x}$

in the limit $(n \to \infty)$

1) $\frac{1}{n^x}\binom{n}{x} = \frac{1}{n^x}\frac{n!}{x!(n-x)!} = \frac{(n-x)!(n-x+1)\ldots(n-1)n}{n^x x!(n-x)!} = \frac{n(n-1)..(n-x+1)}{n^x x!} \to \frac{1}{x!}$
2) $(1 - \frac{\lambda}{n})^n \to e^{-\lambda}$ (definition of exponential)
3) $(1 - \frac{\lambda}{n})^{-x} \to 1$

Putting everything together then:

$f(x) = \frac{e^{-\lambda}\lambda^x}{x!}$

Definition

Given

1) an interval in the real numbers
2) events occur at random in the interval
3) the average number of events on the interval is known (λ)
4) if one can find a small regular partition of the interval such that each of them can be considered a Bernoulli trial.

Then the random variable X that counts events across the interval is a **Poisson** variable with probability mass function

$$f(x) = \frac{e^{-\lambda}\lambda^x}{x!}, \lambda > 0$$

which we write as

$$X \to Poiss(\lambda)$$

Properties: When X follows a Poisson model, it has

1) mean
$$E(X) = \lambda$$
2) and variance
$$V(X) = \lambda$$

Examples

1) What is the probability of receiving four emails in an hour, when the average number of emails in an hour is 1?

We have that the random variable that counts emails follows a Poisson model

$$X \to Poiss(\lambda)$$

with $\lambda = 1$. That is; it has probability mass function

$$f(x) = \frac{e^{-1}1^x}{x!}$$

Therefore, the probability that the variable takes value 4 is $P(X = 4)$ is

$$f(4; \lambda = 1) = \frac{e^{-1}1^4}{4!} = 0.01532831$$

in Python poisson.pmf(4,1)

2) What is the probability of receiving 4 emails in **three hours**, when the average number of emails in an hour is 1?

The unit in which we do the counts has changed from 1 hour to 2 hours, so we have to **re-scale** λ. If before the average number of emails in one hour was $\lambda = 1$, the average number of emails in three hours is now 3: $\lambda_{3h} = 3\lambda_{1h} = 3 * 1 = 3$

We have that the variable is Poisson: $X \to Poiss(\lambda_{3h})$ with $\lambda_{3h} = 3$ and its probability mass function is:

$$f(x) = \frac{e^{-3}3^x}{x!}$$

Therefore the probability that the variable takes value 4 is $P(X = 4)$:

$$f(4; \lambda = 3) = \frac{e^{-3}3^4}{4!} = 0.1680314$$

in Python poisson.pmf(4,3)

3) What is the probability of counting **at least** 10 cars arriving at a toll booth in one minute, when the average number of cars arriving at a toll booth in one minute is 5;

We have that the variable is Poisson: $X \to Poiss(\lambda)$ with $\lambda = 5$ and its probability mass function is:

$$f(x) = \frac{e^{-5}5^x}{x!}$$

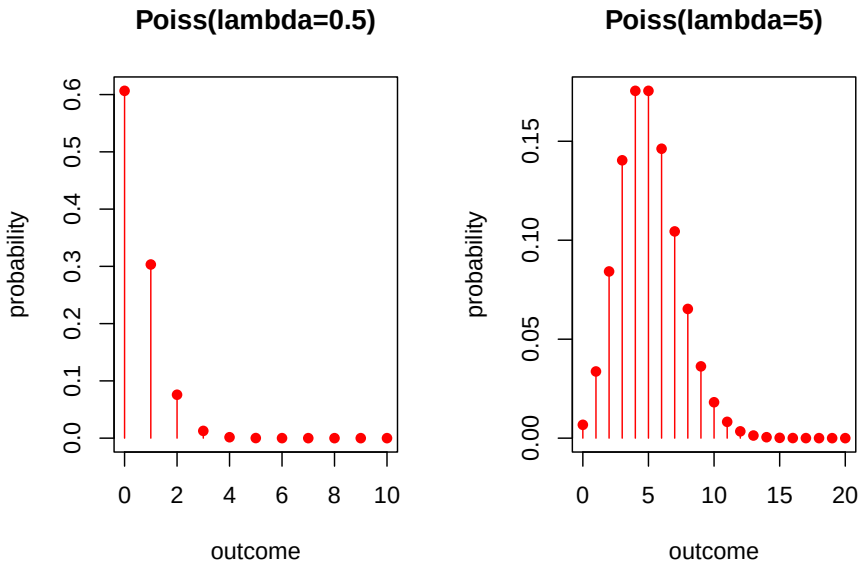
and then, we want to compute

$P(X \geq 10) = 1 - P(X < 10) = 1 - P(X \leq 9) = 1 - F(9; \lambda = 5) = 0.03182806$

where we use the probability distribution $F(x, \lambda)$ of the Poisson model.

In Python 1-poisson.cdf(9,5)

Let's look at some probability mass functions in the family of parametric Poisson models:



8.5 Continuous probability models

Continuous probability models are **probability density functions** $f(x)$ of a continuous random variables that we **believe** describe real random experiments.

A probability density function $f(x)$

1) is positive

$$f(x) \geq 0$$

2) allows us to compute probabilities using the area under the curve

$$P(a \leq X \leq b) = \int_a^b f(x)dx$$

3) is such that the probability of anything is 1

$$\int_{-\infty}^{\infty} f(x)dx = 1$$

8.6 Exponential process

Let's go back to a **Poisson process** defined by probability

$$f(k) = \frac{e^{-\lambda}\lambda^k}{k!}, \lambda > 0$$

for the number of events (k) in an interval.

Let us now consider that we are interested in the length/time we should wait for the **first** event to occur.

We can ask about the probability that the first event occurs after length/time X.

Therefore, since X is a **continuous** random variable, we are looking for its probability density function $f(x)$.

8.7 Exponential probability density

The probability of 0 counts **if** an interval has **unit** x is

$$f(0|x) = \frac{e^{-x\lambda}x\lambda^0}{0!}$$

(look at example 2 above, where we re-escale the units of λ) or

$$f(0|x) = e^{-x\lambda}$$

We can treat this as the conditional probability of 0 events given a distance x: $f(K = 0|X = x)$ and apply the Bayes theorem to reverse it:

$$f(x|0) = Cf(0|x) = Ce^{-x\lambda}$$

C is a constant that collects the marginal divided by total probability rule. So we can calculate the **probability of observing a distance** x that has 0 counts. This is the probability that we measure a length/time x until the first event. Since we can cut the wire at any point and measure the distance until the next event, if we decide to cut at one event, this probability is also for the distance between any two consecutive events.

Definition

In a Poisson process with parameter λ the probability of measuring a length/time X between any two consecutive events is given by the **probability density**

$$f(x) = Ce^{-x\lambda}$$

where C is a constant that ensures $\int_{-\infty}^{\infty} f(x)dx = 1$. By integration we find $C = \lambda$, and therefore

$$f(x) = \lambda e^{-\lambda x}, x \geq 0$$

λ is the paramenter of the model also known as a **decay rate**.

Properties:

When $X follows an exponential model $\$X \rightarrow Exp(\lambda)$ then

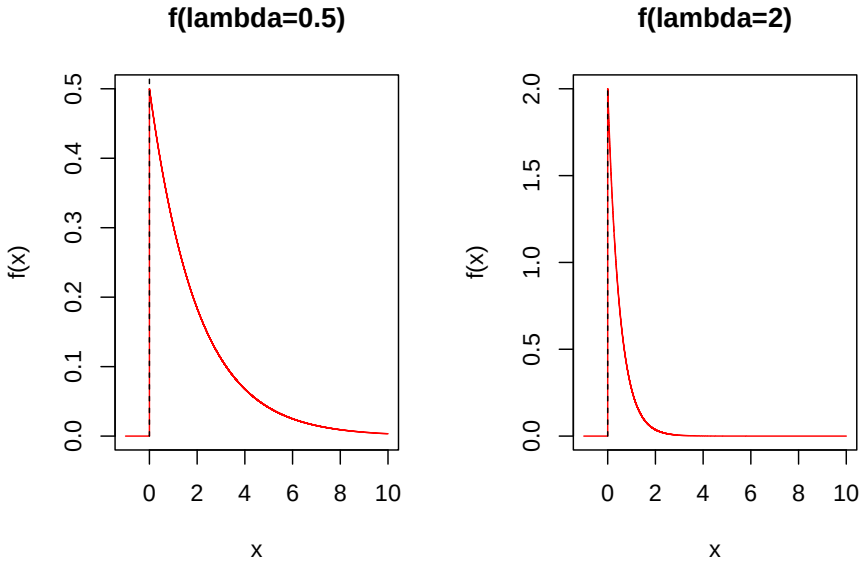
1) It has mean

$$E(X) = \frac{1}{\lambda}$$

2) and variance

$$V(Y) = \frac{1}{\lambda^2}$$

Let's look at a couple of the probability densities in the exponential family



8.8 Exponential Distribution

Consider the following questions:

1) In a Poisson process ¿What is the probability of observing an interval **smaller** than size a until the first event?

We have shown that if we are counting Poisson events $K \rightarrow Poiss(\lambda)$ then the distance until the first interval follows an exponential model $X \rightarrow Exp(\lambda)$. Now, remember that the probability of observing and interval smaller than a is computed with the probability distribution $F(a) = P(X \leq a)$

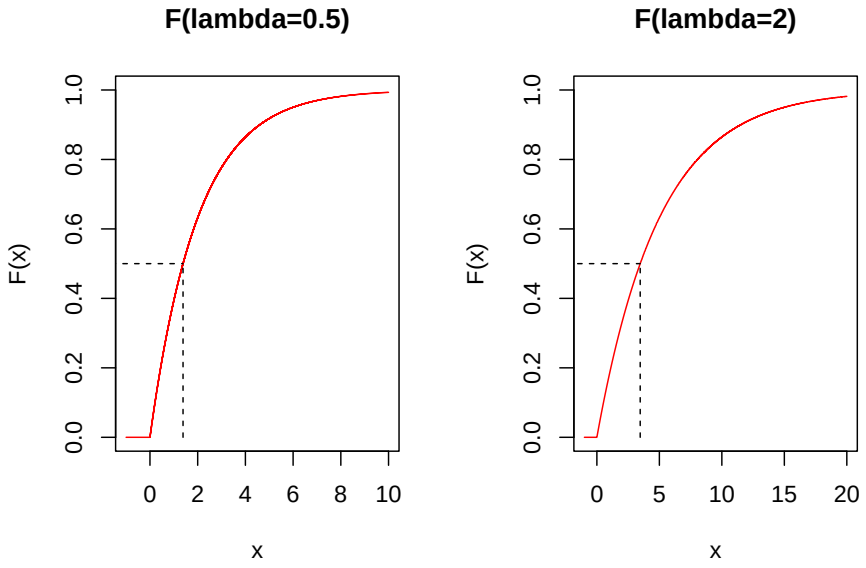
$$F(a) = \lambda \int_{\infty}^{a} e^{-x\lambda} dx = 1 - e^{-a\lambda}$$

2) In a Poisson process ¿What is the probability of observing an interval **larger** than size a until the first event?

Using the probability distribution this can be computed as:

$$P(X > a) = 1 - P(X \leq a) = 1 - F(a) = e^{-a\lambda}$$

Let's look at a couple of exponential distributions from the exponential family



The median x_m is such that $F(x_m) = 0.5$. That is $x_m = \frac{\log(2)}{\lambda}$.

Examples

1) What is the probability that we have to wait for a bus for more than 1 hour when on average there are two buses per hour?

X is the time for the first Poisson event (bus arrival) and therefore $X \rightarrow Exp(\lambda = 1)$. Threfore, we can compute

$$P(X > 1) = 1 - P(X \leq 1) = 1 - F(1, \lambda = 2) = 0.1353353$$

Where $F(1)$ is the esponential distribution function above and in Python 1-expon.cdf(1,0,1/2). Note than Pyhton uses $1/\lambda$ as argument (the half life).

2) What is the probability of having to wait less than 2 seconds to detect one particle when the radioactive decay rate is 2 particles each second; $F(2, \lambda = 2)$

$$P(X \leq 2) = F(2, \lambda = 2) = 0.9816844$$

in Python expon.cdf(2,0,1/2)

8.9 Questions

1) During WWII in London, the expected number of bombs that hit an area of $3km^2$ was 0.92. The probability that, in one day, one area received two bombs was

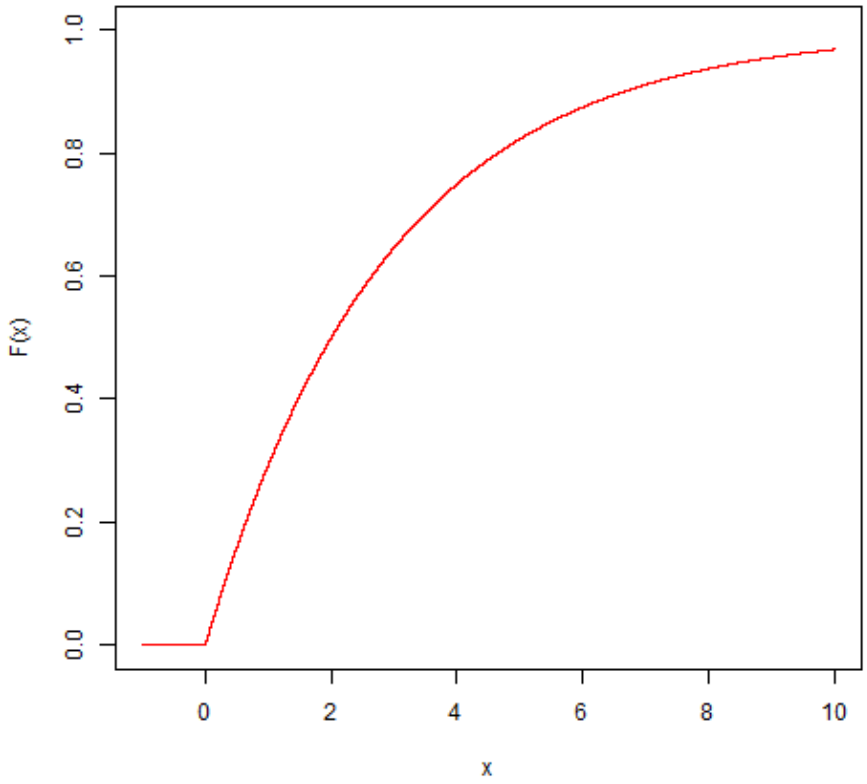
a:1-poisson.cdf(x=2, lambda=0.92); **b:**poisson.cdf(x=2, lambda=0.92); **c:**1-poisson.pmf(x=2, lambda=0.92); **d:**poisson.pmf(x=2, lambda=0.92)

2) The probability that a passenger has to wait less than 20 minutes until the next bus arrives at her stop is better described by

a: A poisson model on the number of buses per 20 minutes;
b: An exponential distribution at 20 minutes with a given expectation of buses per minute; **c:** A binomial model that counts the number of buses per 20 minutes **d:** A normal distribution with an average of buses per 20 minutes and a given standard deviation;

3) From the exponential probability distribution in the figure below, what is the most likely value of the median?

a: 2; **b:** 3; **c:** 4; **d:** 5



8.10 Exercises

8.10.0.1 Exercise 1

The average number of phone calls per hour coming into the switchboard of a company is 150. Find the probability that during one particular minute there will be

- 0 phone calls (R:0.082)
- 1 phone call (R:0.205)
- 4 or fewer calls (R:0.891)
- more than 6 phone calls (R:0.0141)

8.10.0.2 Exercise 2

The average number of radioactive particles hitting a Geiger counter in a nuclear energy plant under control is 2.3 per minute.

- What is the probability of counting exactly 2 particles in a minute? (R:0.265)
- What is the probability of detecting exactly 10 particles in 5 minutes? (R:0.112)
- What is the probability of at least one count in two minutes? (R:0.9899)
- What is the probability of having to wait less than 1 second to detect a radioactive particle, after we switch on the detector? (R:0.037)
- We suspect that a nuclear plant has a radioactive leak if we wait less than 1 second to detect a radioactive particle, after we switch on the detector. What is the probability that when we visit in 5 plants that are under control, we suspect that at least one has a leak? (R:0.1744).

Chapter 9

Normal Distribution

9.1 Objective

In this chapter we will introduce the normal probability distribution.

We will discuss its origin and its main properties.

9.2 History

In 1801 Gauss analyzed the data that obtained Giuseppe Piazzi for the position of the Ceres, a large asteroid between Mars and Jupiter.

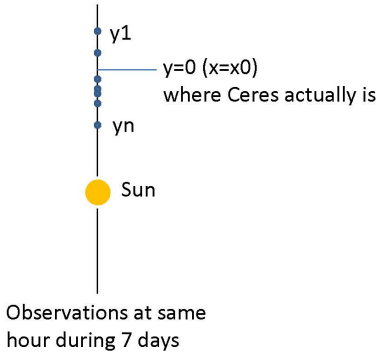
At the time Piazzi suspected it was a new planet, as it moved day to day against the fixed stars. In January, it could be seen at the horizon just before dawn. However, as days passed Ceres will rise latter and latter until it could not longer be seen because of the Sun rise. The Sun caught up with it.

Gauss understood that the measurements for the position of Ceres had errors.

He was therefore interested in finding how the observations were **distributed** so he could find the most **likely** orbit. With the orbit, we could derive the mass of the object and then decide whether it was massive as a planet or just a big asteroid.

Data was available only for the month of January. After which Ceres would disappear. He wanted to **predict** six moths latter, once it had slowly transited behind the Sun, where astronomers should point their telescopes just after at dusk.

Gauss had to account for the errors in the position of Ceres at a given day due to measurement



Gauss supposed that

1) small errors were more likely than large errors.

2) error were symmetrical. The error at a distance $-\epsilon$ from the real position of Ceres was equally probable than the error at $+\epsilon$.

3) the most **likely** position (the one that we believe the most) of Ceres at any given time in the sky was the **average** of multiple measurements.

That was enough to show that the deviations of the observations y **from the orbit** satisfied the equation*

$$\frac{df(y)}{dy} = -Cyf(y)$$

with C a positive constant. The solution of this differential equation is:

$$f(y) = \frac{\sqrt{C}}{\sqrt{2\pi}} e^{-\frac{Cy^2}{2}}$$

*The evolution of the normal distribution, Saul Stahl, Mathemetics Magazine, 2006.

9.3 normal density

Gaussian probability density gives the distribution of measurement errors from the **actual** but **unknown** position of Ceres in the sky. Let's make a couple of changes to the function.

1) Let's write the error density from the horizon using the random variable X, that is, $y = x - \mu$. μ is the **actual** but **unknown** position of Ceres from the horizon. After a change of variable we find the probability density function:

$$f(x) = \frac{\sqrt{C}}{\sqrt{2\pi}} e^{-C(x-\mu)^2}$$

2) Let's rename the variable C to $\frac{1}{\sigma^2}$

So, we arrive at the following definition.

9.4 Definition

A random variable X defined on the real numbers has a density **Normal** if it takes the form

$$f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, x \in \mathbb{R}$$

The variable has

1) mean

$$E(X) = \mu$$

which for Gauss represented the real position of Ceres from the horizon.

2) and variance

$$V(X) = \sigma^2$$

which represented the dispersion of the error in the observations that depended on the quality of the telescope and the skill of Piazzi.

μ and σ are the **two parameters** that completely describe the normal density function and their **interpretation** depends on the random experiment.

It is important underline right away that the parameters are **not observable**, they are abstract quantities. This is true not only for the normal model but for every model. The real unique position of Ceres (μ) was unknown, Gauss instead had repetitions of a random experiment that measured the position of Ceres

$(x_1, x_2, ...x_n)$. How can we get to know, or at least learn μ from $(x_1, x_2, ...x_n)$? This is the central question from chapter 10 on wards.

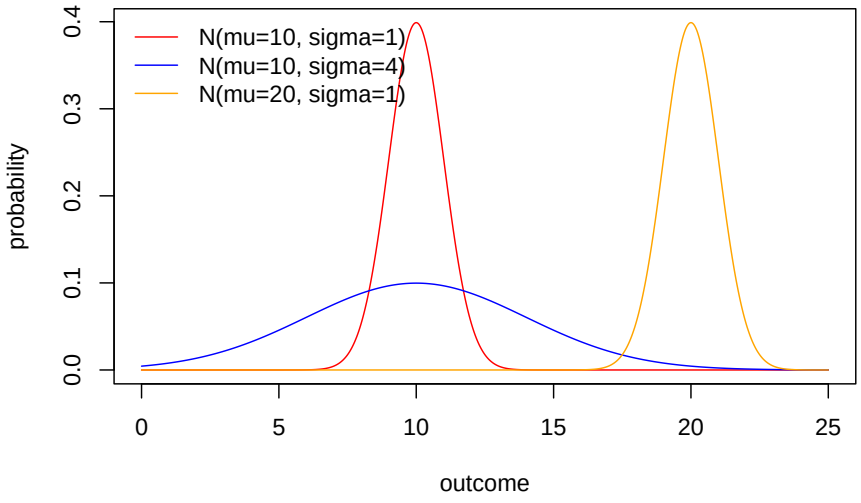
Spoiler: the best we can do is to take the average $\bar{x}$. But before we get there, we are going to **pretend** that we know the parameters μ and σ.

Let me just point out how deep the problem is. For empiricists, any knowledge comes from experience thus μ is an abstraction from data. For idealist, knowledge comes from intellectual insight or reasoning, so μ represents a fundamental aspect of reality and the observations are imperfect representations of it. The experimental scientist finds relief knowing that $\bar{x}$ approaches to μ when n is large.

When X follows a normal probability density; that is, When X is normally distributed, we write

$$X \to N(\mu, \sigma^2)$$

Let's look at some probability densities in the normal parametric model



For Gauss, the change from the red to the yellow curve means that Ceres moved, and the change from the red to the blue curve means that the telescope was less accurate.

9.5 Probability distribution

The probability distribution of the normal density:

$$F(a) = P(Z \leq a)$$

is the **error** function defined by the area under the curve from $-\infty$ to a

$$F(a) = \int_{-\infty}^{a} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$

The function is found in most computer programs, and it does not have a closed form of known functions.

Examples (female height)

1) What is the probability that a woman in the population is at most $150cm$ tall if women have a mean height of $165cm$ with standard deviation of $8cm$?

$P(X \leq 150) = F(150, \mu = 165, \sigma = 8) = 0.03039636$

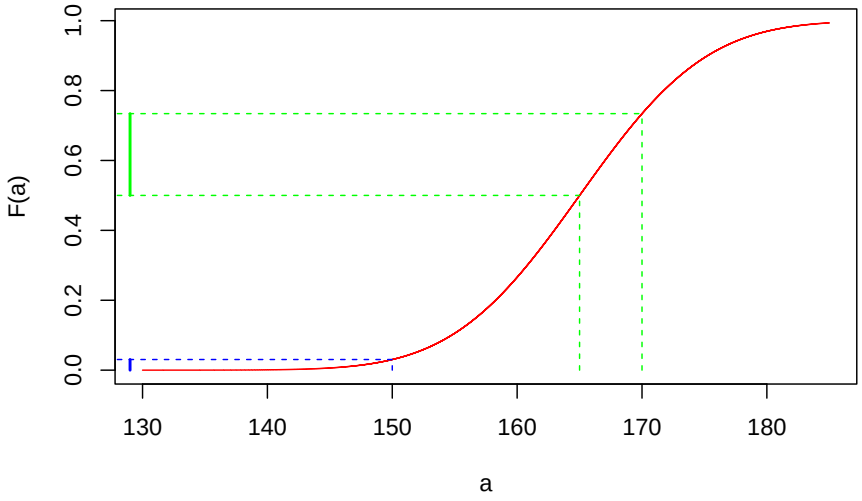
In Python norm.cdf(150, 165, 8)

2) What is the probability that a woman's height in the population is between $165cm$ and $170cm$?

$P(165 \leq X \leq 170) = F(170, \mu = 165, \sigma = 8) - F(165, \mu = 165, \sigma = 8) = 0.2340145$

In Python norm.cdf(170, 165, 8)-norm.cdf(165, 165, 8)

Let's look at the probability distribution function



The blue bar in the left is the probability in example 1 above. The green bar is the probability in example 2.

3) What is the first quartile for female height?

The first quartile is defined as:

$$F(x_{0.25}, \mu = 165, \sigma = 8) = 0.25$$

or

$$x_{0.25} = F^{-1}(0.25, = 165, \sigma = 8) = 159.6041$$

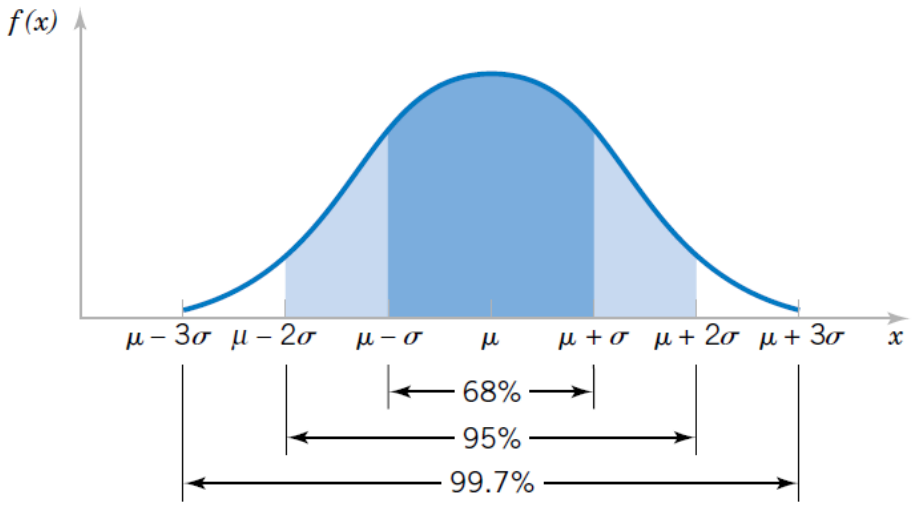
In Python norm.ppf(0.25, 165, 8). Let's make a prediction. If we measure many women, about 25% of them would be at most $160cm$ tall.

Properties of the Normal distribution

1) the mean μ is also the median as it splits the measurements in two.

2) x values that fall farther than 2σ are considered **rare** (less than 5%)

3) x values that fall farther than 3σ are considered **extremely rare** (less than 0.2%).



Example (female height)

We can define the limits of **common outcomes** for the distribution of female height in the population.

1) at a distance of one standard deviation from the mean, we find 68% of the population

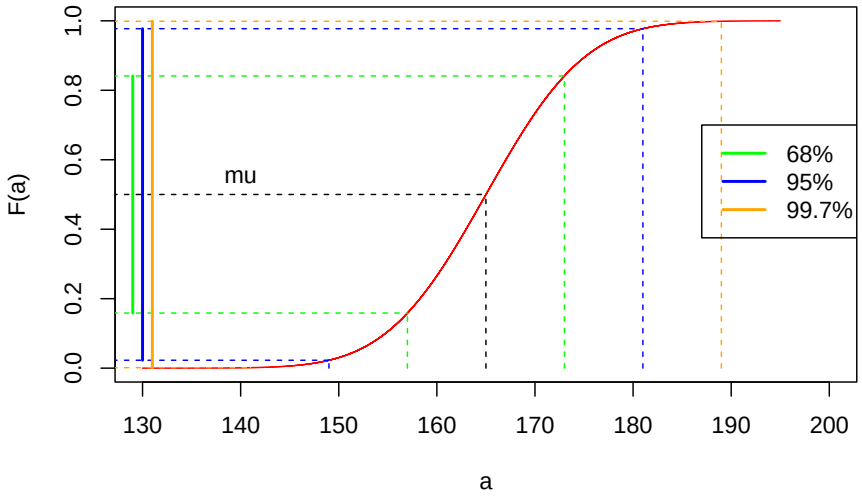
$$P(165 - 8 \leq X \leq 165 + 8) = P(157 \leq X \leq 173) = F(173) - F(157) = 0.68$$

2) at a distance of two standard deviations from the mean, we find 95% of the population

$$P(165 - 2 \times 8 \leq X \leq 165 + 2 \times 8) = F(181) - F(149) = 0.95$$

3) at a distance of three standard deviations from the mean, we find 99.7% of the population

$$P(165 - 3 \times 8 \leq X \leq 165 + 3 \times 8) = F(189) - F(141) = 0.997$$



We often say that the distribution of the random variable X is the **population distribution**. When we talk about a **population**, we really mean the repetition of the random experiment many, many times; like when we made the prediction in the example 3, before this one. Of course, if we could measure the entire population of women, the histogram of heights from the population would be very close to $f(x)$. But again, this is the difference between observation (histogram) and abstraction (probability density). The population distribution is then close, if not identical, to the random variable distribution. Consider, however, that it makes sense to talk about the probability that a woman in a century's time has a given height, even if she has not been born, never mind the population, as probabilities are statements about future events.

9.6 Standard normal density

The **standard** normal density is one member of the normal family, such that

$$f(x; \mu = 0, \sigma^2 = 1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x)^2}{2}}, x \in \mathbb{R}$$

It is therefore the density with

1) mean

$$E(X) = \mu = 0$$

2) and variance

$$V(X) = \sigma^2 = 1$$

When a random variable follows a normal probability density, we say that the variable is **standard normal** and write

$$X \to N(0, 1)$$

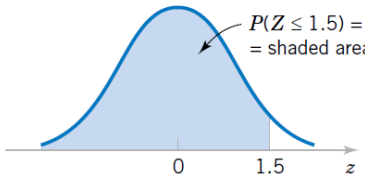
9.7 Standard distribution

The probability distribution of a standard normal variable is

$$\phi(x) = F(x) = P(X \leq a)$$

Because the **standard** distribution is special and will appear often, we use the letter ϕ for it. This is the **error** function defined by

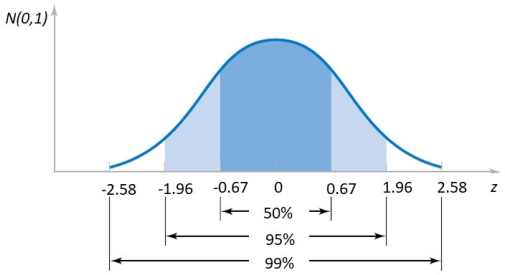
$$\phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz$$



z	0.00	0.01	0.02	0.03
0	0.50000	0.50399	0.50398	0.51197
⋮		⋮		
1.5	0.93319	0.93448	0.93574	0.93699

You can find it in most computer programs. In Python is norm.cdf(x) with the default parameters, 0 and 1.

We usually define the limits of the **most common outcomes** for the normal standard variable



1) The interquartile range

$$P(-0.67 \leq X \leq 0.67) = 0.50$$

In Python: [norm.ppf(0.25), norm.ppf(0.75)]

2) 95% range

$$P(-1.96 \leq X \leq 1.96) = 0.95$$

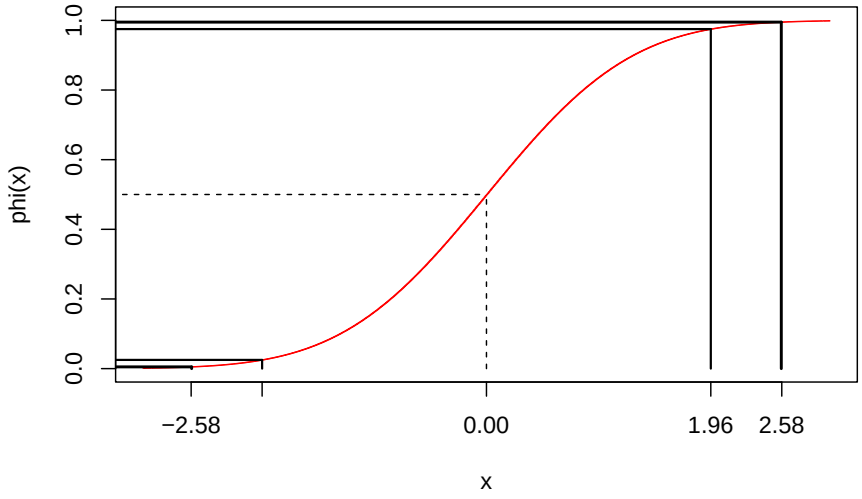
In Python: [norm.ppf(0.025), norm.ppf(0.975)]

3) 99% range

$$P(-2.58 \leq X \leq 2.58) = 0.99$$

In Python: [norm.ppf(0.005), norm.ppf(0.995)]

It is less common but we can also see the ranges in the plot for $\phi(x)$



9.8 Standardization

All normal variables can be **standardized**. This means that if $X \to N(\mu, \sigma^2)$, then we can transform the variable to a **standardized variable**

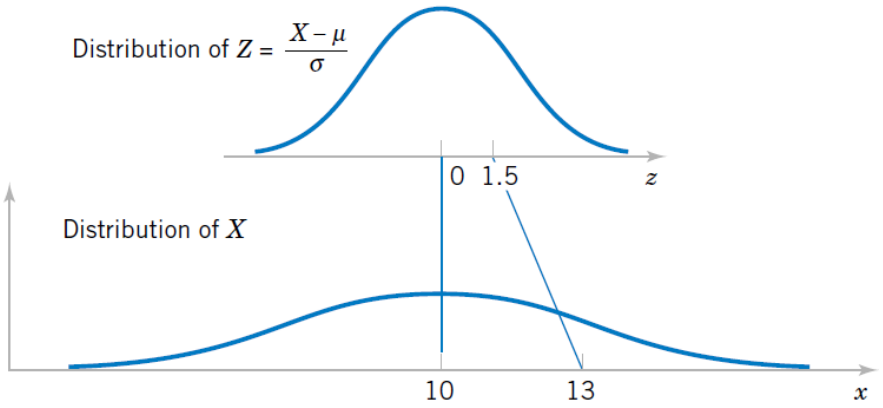
$$Z = \frac{X - \mu}{\sigma}$$

which will have density:

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$

Therefore, for any $X \to N(\mu, \sigma^2)$

$$Z = \frac{X - \mu}{\sigma} \to N(0, 1)$$



You can demonstrate this by replacing $x = \sigma z + \mu$ and $dx = \sigma dz$ in the probability equation

$$P(x \leq X \leq x + dx) = P(z \leq Z \leq z + dz)$$

$$= \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$

$$= \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz$$

We can compute the probability of **any normal variable** $X \to N(\mu, \sigma^2)$ using the standard distribution

$$F(a) = P(X < a) = P(\tfrac{X-\mu}{\sigma} < \tfrac{a-\mu}{\sigma})$$

$$= P(Z < \frac{a-\mu}{\sigma}) = \phi(\frac{a-\mu}{\sigma})$$

For computing $P(a \leq X \leq b)$, we use the property of the probability distributions

$$F(b) - F(a) = P(X \leq b) - P(X \leq a)$$

$$= \phi(\frac{b-\mu}{\sigma}) - \phi(\frac{a-\mu}{\sigma})$$

Therefore, $\phi(x)$ is the only function we need to compute probabilities for any normal variable. This is an important property that allowed all statisticians before the computer to look at a single table, and that helps concentrate all the randomness of the experiment in a probability function that does not depend on the parameters.

9.9 Summary of probability models

Model	Experiment	range of X	f(x)	E(X)	V(X)
Uniform	Measuring an integer or real number	$[a, b]$	$\frac{1}{n}$	$\frac{b+a}{2}$	$\frac{(b-a+1)^2-1}{12}$
Bernoulli	Observing A	$(0, 1)$	$(1-p)^{1-x}p^x$	p	$p(1-p)$
Binomial	Counting # of A events in n repetitions of Bernoulli trials	$(0, 1, .., n)$	$\binom{n}{x}(1-p)^{n-x}p^x$	np	$np(1-p)$
Negative Binomial for events	Counting # of A' events in repetitions of Bernoulli trials until r events A are observed	$(0, 1, ..)$	$\binom{x+r-1}{x}(1-p)^x p^r$	$\frac{r(1-p)}{p}$	$\frac{r(1-p)}{p^2}$
Hypergeom	Counting # of A events in a sample n from population N with K # of A events	$\max(0, n+K-N)$, ... $\min(K, n)$	$\frac{1}{\binom{N}{n}}\binom{K}{x}\binom{N-K}{n-x}$	$n * \frac{N}{K}$	$n\frac{N}{K}(1-\frac{N}{K})\frac{N-n}{N-1}$
Poisson	Counting # A events in an interval	0,1, ..	$\frac{e^{-\lambda}\lambda^x}{x!}$	λ	λ
Exponential	Measuring an interval between two events A	$[0, \infty)$	$\lambda e^{-\lambda x}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$

Model	Experiment	range of X	f(x)	E(X)	V(X)
Normal	Measuring values with symmetric errors whose most likely value is the average	$(-\infty, \infty)$	$\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	μ	σ^2

9.10 Python functions of probability models

Model	Python
Uniform (continuous)	uniform.pdf(x, a, b)
Binomial	binom.pmf(x,n,p)
Negative Binomial for events	nbinom.pmf(x,r,p)
Hypergeom	hypergeom.pmf(x, N, K, n)
Poisson	poisson.pmf(x, lambda)
Exponential	expon.pdf(x,0,1/lambda)
Normal	norm.pdf(x, mu, sigma)

9.11 Questions

1) It is not true that for a normally distributed variable

a: its mean and median are the same; **b:** the standard probability distribution can be used to compute its probabilities; **c:** its interquartile range is twice its standard deviation; **d:** 5% of its observations are a distance greater than twice its standard deviation

2) For a normal standard variable

a: 50% of its observations are between $(-0.67, 0.67)$; **b:** 2% of its observations are lower than -2.58; **c:** 5% of its observations are greater than 1.96; **d:** 25% of its observations are between $(-1.96, -0.67)$

3) if we know that $\phi(-0.8416212) = 0.2$ then what is $\phi(0.8416212)$

a: 0.1; **b:** 0.2; **c:** 0.8; **d:** 0.9

4) the third quartile of a normal variable with mean 10 and standard deviation 2 is

a: norm.ppf(1/3, 10, 2)=9.138545; **b:** norm.ppf(1-0.75, 10, 2)=8.65102 ; **c:** norm.ppf(1-1/3, 10, 2)=10.86145 ; **d:** norm.ppf(0.75, 10, 2)= 11.34898

5) probability that a normal variable with mean 10 and standard deviation 2 is in $(-\infty, 10)$ is

a: 0.25; **b:** 0.5; **c:** 0.75; **d:** 1:

9.12 Exercises

9.12.0.1 Exercise 1

Find the area under the standard normal curve in the following cases:

- Between $z = 0.81$ and $z = 1.94$ (A:0.182)
- To the right of $z = -1.28$ (A:0.899)
- To the right of $z = 2.05$ or to the left of $z = -1.44$ (A:0.0951)

9.12.0.2 Exercise 2

- What is the probability that a man's height is at least 165cm if the population mean is 175cm y the standard deviation is 10cm? (A:0.841)
- What is the probability that a man's height is between 165cm and 185cm? (A:0.682)
- What is the height that defines the 5% of the smallest men? (A:158.55)

Chapter 10

Sampling distributions

10.1 Objective

In this chapter, we will study estimates of the mean and variance of normal distributions using **random samples**.

We will introduce the **sample mean** and the **sample variance** as random variables that estimate the parameters of the normal distribution.

The sample mean and the sample variance have probability density functions, these are called **sample density functions**.

10.2 Random sample

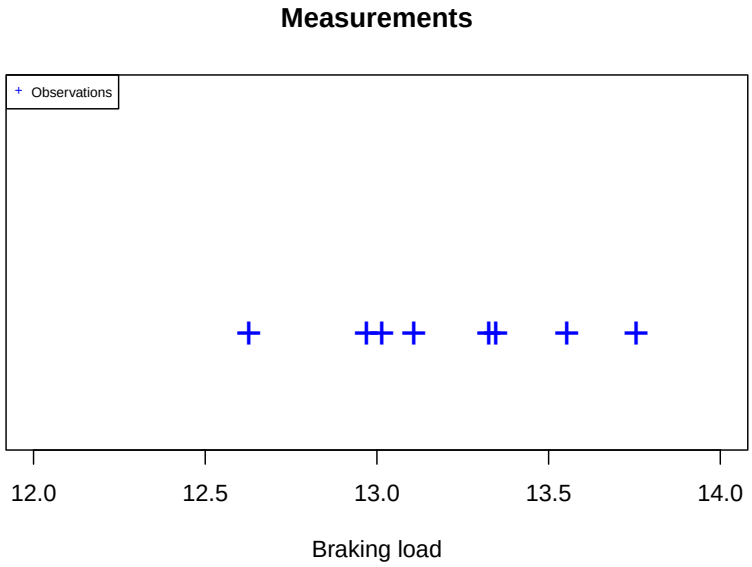
Example (Cables)

Imagine that a client asks your metallurgical company to sell them 8 cables that can carry together up to 96 Tons; that's 12 Tons each. You must guarantee that none of them will break when they are loaded with this weight.

In **stock**, there are many cables, all produced under the same specifications, that might work, but you're not sure. So you take 8 of those cables at random, and load them until they break.

We say that you take a **random sample** of size 8, which means that you repeat the random experiment 8 times. Here are the results

```
## [1] 13.34642 13.32620 13.01459 13.10811 12.96999 13.55309 13.75557 12.62747
```



Definition:

A **random sample** of size n is the **repetition** of a random experiment n **independent** times.

- A random sample is an n-dimensional **random variable**

$$(X_1, X_2, ...X_n)$$

where X_i is the *i-th* iteration of the random experiment with identical probability density function $f(x)$ for all i

- **An observation** of a random sample is the set of n values obtained from the experiments

$$(x_1, x_2, ...x_n)$$

Example (Cables)

Our **observation** of the sample of size 8 cables was

```
## [1] 13.34642 13.32620 13.01459 13.10811 12.96999 13.55309 13.75557 12.62747
```

In this sample, we found that

1) No cable broke 12 Tons.
2) There was one that broke at 12.62747 Tons.

Do you take a chance and sell a random sample of 8 cables from your stock?

What happens if your company is liable for a cable break and has to pay a large fine?

To assure the customer that none of the cables will break at 12 Tons, we would like to see that $P(X \leq 12)$ is reasonably low.

10.3 Calculation of probabilities

To calculate probabilities we need:

1. A probability model (probability function)
2. The parameters of the model (the values of the probability function)

Let's **assume** that the breaking load of the cables follows a **normal** probability density function.

$$X \to N(x; \mu, \sigma^2)$$

To compute $P(X \leq 12)$, we need the parameters μ and σ^2. How can we estimate the parameters from the observed sample?

10.4 Parameter estimation

To find likely values for the parameters, we use data. Therefore, we take a **random sample**. That is, we repeat the experiment n times, collect data, and use it to **estimate** the parameters.

Estimate of the mean and variance

Recall that for a discrete random variable, we define the mean as

$$\mu = \sum_{i}^{m} x_i f(x_i)$$

which is the center of mass of the **probabilities**, where $f(x_i)$ is the probability function. This definition was motivated by the center of mass of the **observations**

$$\bar{x} = \frac{1}{n} \sum_{i}^{n} x_i = \sum_{i}^{m} x_i f_i$$

that we defined as the **average**, and where f_i are the relative frequencies. Remember that n is the number of observations (it can be as large as we want)

and m is the number of possible outcomes (usually fixed by the sample space). We have argued that when $n \to \infty$ then

$$\hat{P}(X = x) = f_i$$

This means that the probabilities can be **estimated** (or dressed with a **hat**) by the relative frequencies when n is large, because $lim_{n\to\infty} f_i = f(x_i)$. To estimate simply means to assign a number to an unknown quantity. Therefore, we should also have that the **mean** μ can be estimated by the **average** $\bar{x}$

$$\hat{\mu} = \bar{x} = \sum_i^m x_i f_i \to \mu = \sum_i^m x_i f(x_i)$$

when $n \to \infty$. Thus, for large samples, we may take the center of mass of the data $\bar{x}$, as the center of mass of the probability function μ. Doing this we know will make an error (we are not at infinity) but we will try to deal with it as we will discuss later on.

With the variance

$$\sigma^2 = \sum_i^m (x_i - \mu)^2 f(x_i)$$

we have a similar situation. In the limit when $n \to \infty$

$$\hat{\sigma}^2 = s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{n}{n-1} \sum_{i=1}^m (x_i - \bar{x})^2 f_i \to \sigma^2$$

and we assume that the moment of inertia of the data is close to the moment of inertia of the probabilities.

Example (Cables)

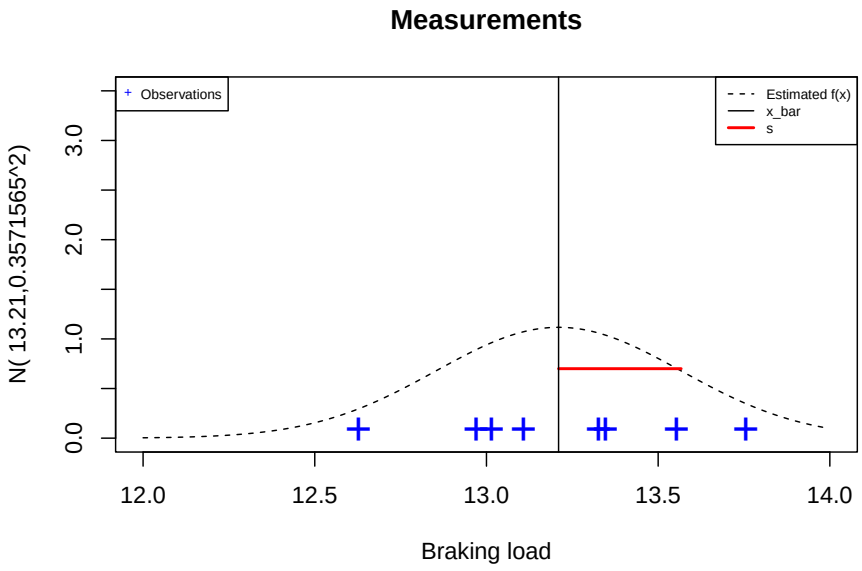
Assuming that the breaking load of our cable is a normal random variable

$$X \to N(x; \mu, \sigma^2)$$

we use the estimates $\bar{x} = 13.21$ (np.mean(x)) and $s^2 = 0.3571565^2$ (np.var(x)) as the values of μ and σ^2. So the **fitted** model is

$$X \to N(x; \mu = 13.21, \sigma^2 = 0.3571565^2)$$

In this problem we **didn't know** μ or σ and therefore we are guessing their values and the underlying model



What is the probability that the cable will break at 12 Tons?

Since

$$X \to N(x; \mu = 13.21, \sigma^2 = 0.3571565^2)$$

so

$$P(X \leq 12) = F(12; \mu = 13.21, \sigma^2 = 0.1275608)$$

In Python norm.cdf(12,13.21, 0.3571565)= 0.000352188

Given the **observed** sample, there is an estimated probability of 0.03% that a single cable breaks at 12 Tons. We have a probabilistic argument for selling the cables. If a cable breaks, we can argue that while accidents happen, we sold a product that had lower probability of failure than dying from a car accident (1/100). So our risk management is reasonable in relation to other risky activities.

10.5 Law of large numbers

When we estimate parameters using data, such as by taking the value of

$$\widehat{\mu} = \bar{x}$$

for the value of μ; and the value of

$$\hat{\sigma}^2 = s^2$$

for the value of σ^2, we know that we are **making a mistake**. We know that if we take another sample of cables of size 8 **the estimate will change**, because the average $\bar{x}$ will change.

Can we get an idea of how big the error is in our estimate? This is important because if a cable finally breaks, we may be accused for "selecting" the best cables to run the experiment and then overestimate μ in order to sell the cables. The situation would be different if the average had been 12.5 Tons. The probability of failure would be then 8%, unacceptably high.

The first thing to realize is that the numerical value we get for

$$\bar{x}$$

is the observation of a **random variable**

$$\bar{X}$$

Definition

The **sample mean** (or average) of a random sample of size n is defined as

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

The average is a **random variable** that in our sample of size 8 took the value

$$\bar{x}_{stock} = 13.21$$

If we take another sample, this number will change.

Mean as estimator

The number $\bar{x}$ can be used to **estimate** the unknown parameter μ because the random variable $\bar{X}$ satisfies these two important and **general** properties

1) The expected value of $\bar{X}$ is precisely the parameter μ:

$$E(\bar{X}) = \mu$$

2) The variance of $\bar{X}$ vanishes if the sample is very large:

$$lim_{n\to\infty} V(\bar{X}) = 0$$

The first property holds because

$$E(\bar{X}) = E(\frac{1}{n}\sum_{i=1}^{n} X_i) = E(X) = \mu$$

The second property holds because

$$V(\bar{X}) = V(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{V(\sum_{i=1}^{n} X_i)}{n^2} = \frac{V(X)}{n} = \frac{\sigma^2}{n}$$

Which uses the fact that each random experiment in the sample is independent and therefore $V(\sum_{i=1}^{n} X_i) = nV(X)$.

Estimating μ

As a consequence of properties 1 and 2, we understand that the value $\bar{x}$ **concentrates closer and closer** to μ as n increases. This is called the **law of large numbers**. And it is a law because it is true **for any** probability models of $\bar{X}$. This means that the error we make when we take a value of $\bar{x}$ as the estimate of μ

$$\bar{x} = \hat{\mu}$$

gets smaller and smaller as the sample gets larger and larger because the variance of $\bar{X}$ gets smaller with large n. At least we know that if the sample of broken cables is very large, our mistake is very small. However, we have only got a sample of size 8. We need a better argument.

10.6 Inference

We know that when we take large samples, our error is small. However, for a given value of n, we want to have a measure of the **error**. Therefore, we may ask for the **probability of making an error** of a given size when estimating μ with $\bar{x}$.

When we calculate probabilities of an estimator, we say that we are making an **inference**. Inference problems often arise when we are interested in calculating the probability of making an error of a given size m when we estimate μ with $\bar{x}$. In mathematical terms we are interested in the probability

$$P(-m \leq \bar{X} - \mu \leq m)$$

That is the probability that the difference between μ and the average $\bar{X}$ is within a distance m. To calculate probabilities of $\bar{X}$, we need

1. A probability model for $\bar{X}$ (probability function)

2. The parameters of the model (the values of the probability function)

What is the probability functions of $\bar{X}$ and S^2 so that we can calculate their probabilities?

These probability functions are called **sampling probability functions**, because they are derived from a sampling experiment.

Example (Cables)

Let's ask an inference question. Imagine our cables are **certified** to break with an average load of $\mu = 13$ Tons with variance $\sigma^2 = 0.35^2$. We are now going to pretend that we know the parameters of the probability function.

If this is the case and we take a random sample of 8 cables, what is the probability that the sample mean $\bar{X}$ will be within a **margin of error** of 0.25 Tons from the mean μ?

$$P(-0.25 \leq \bar{X} - \mu \leq 0.25)$$

To calculate this probability, we need to know the probability function of $\bar{X}$.

10.7 Sample mean

Theorem: If X follows a normal distribution

$$X \to N(\mu, \sigma^2)$$

then $\bar{X}$ is normal

$$\bar{X} \to N(\mu, \frac{\sigma^2}{n})$$

This particular case, fully agrees with the law of large numbers, as it should, since $\bar{X}$ has

1) mean
$$E(\bar{X}) = \mu$$

2) variance
$$V(\bar{X}) = \frac{\sigma^2}{n}$$

We say that $\bar{X}$ is an **unbiased** estimator of μ because its expected value is exactly what it wants to estimate: μ; a dart player who on average hits the bull's-eye. We also say that $\bar{X}$ is a **consistent** estimator because its variance gets smaller with large n; a dart's player with better darts hits closer the bull's-eye.

We call

$$se = \sqrt{V(\bar{X})} = \sigma/\sqrt{n}$$

the **standard error** of the sample mean. The standard error is also written as $\sigma_{\bar{x}}$, to emphasize that it is the standard deviation of the sample mean. Note that this is the error we expect when using $\bar{x}$ as the value of μ, and it is the bias we need to correct when estimating the variance (Section 10.9).

So, if we **know** μ and σ, we can calculate the **probabilities of** $\bar{X}$ using the normal distribution.

Remember that we have **two probability functions**:

1. The probability function of X that is also known as the probability function of the **population**.

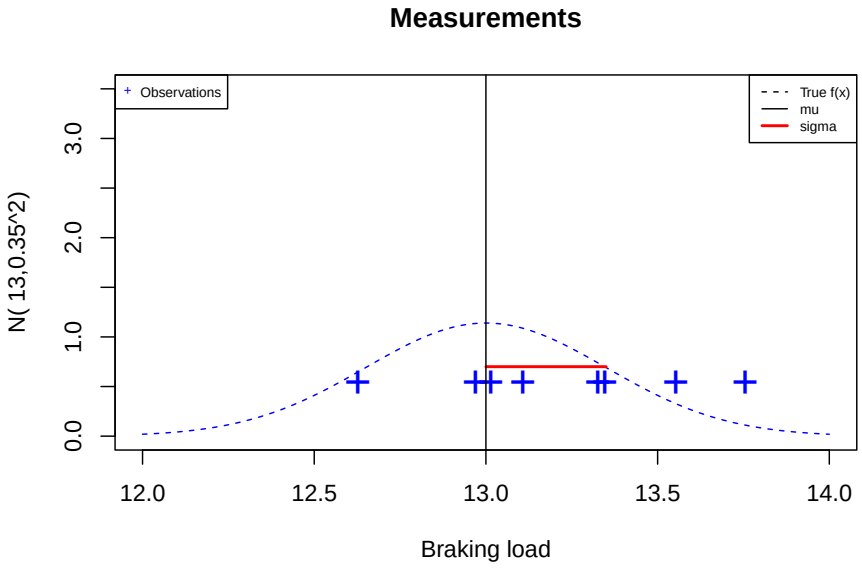
2. The probability function of $\bar{X}$ that is probability function of the **sample**.

Example (Cables)

Probability densities for X and $\bar{X}$

In the new problem of calculating the probability of an error, we **need know** (or pretend we know) the true values μ and σ^2 and the probability function of the **population**. Let's assume that the our cables have been **certified** such that their breaking load is

$$X \to N(\mu = 13, \sigma^2 = 0.35^2)$$

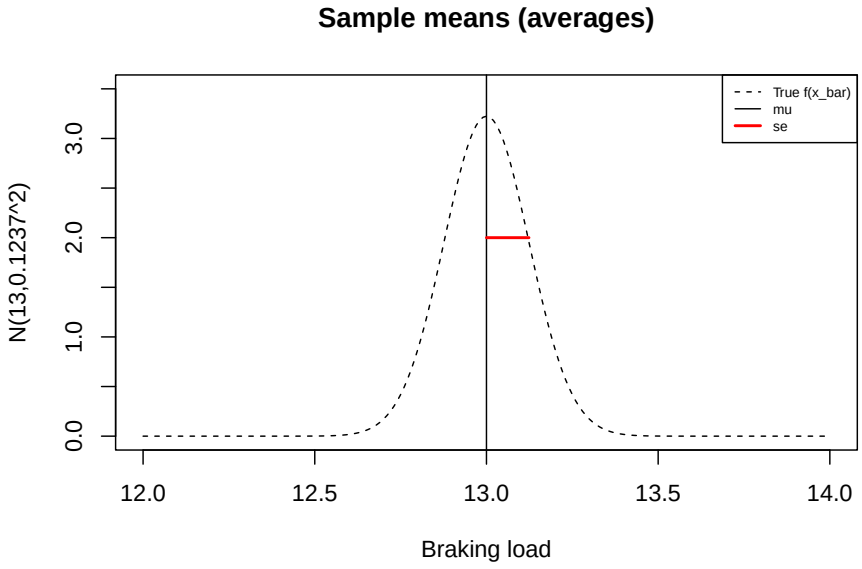


Since X is normal, then $\bar{X}$ is normal, and therefore we also know the probability function of the sample mean $\bar{X}$

$$\bar{X} \to N(13, \frac{0.35^2}{8})$$

which has mean and variance

1) $E(\bar{X}) = \mu = 13$
2) $V(\bar{X}) = \frac{\sigma^2}{n} = \frac{0.35^2}{8} = 0.01530169$



We can see that the outcomes of $\bar{X}$ concentrate more closely to μ than the random variable X (blue line above, in latter plot). In fact, if n is big the averages will be so close, and the distribution so peaky that a single value of the average is a good enough value of μ.

Finally, we want to calculate **the probability** that our estimate is within a margin of error of 0.25 Tons. That is a distance of 0.25 from the mean, or

$$P(-0.25 \leq \bar{X} - 13 \leq 0.25) = P(12.75 \leq \bar{X} \leq 13.25)$$

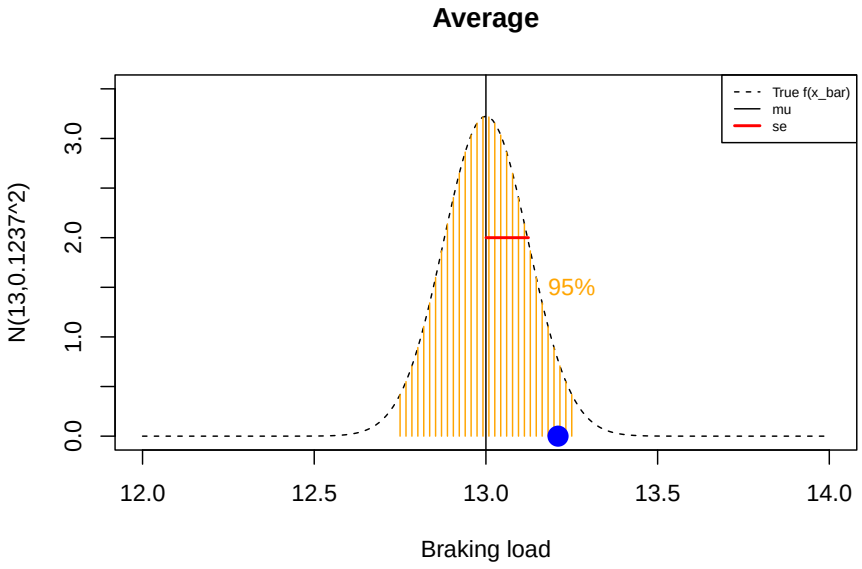
$= F(13.25; \mu, \sigma^2/n) - F(12.75; \mu, \sigma^2/n)$

In Python we can calculate it as:

norm.cdf(13.25, 13, 0.1237)-norm.cdf(12.75, 13, 0.1237)=0.956.

Remember: $se = \sigma_{\bar{x}} = \sqrt{0.01530169} = 0.1237$.

We can, therefore, expect that 95.6% of the averages of samples of size 8 to fall between $(12.75, 13.25)$ if cables are manufactured with an expected value for the breaking load of $\mu = 13$.



The **observed average** (the blue point above) of our sample was $\bar{x}_{stock} = 13.21$, which is within the margin of error of 2.5 Tons. While the estimate was certainly high, it was not unusually high, and we can defend that the error of estimation was within our expectations.

10.8 Sample sum

If we are interested in using all 8 cables at the same time to carry a total of 96 Tons, then we should consider **adding** their individual contributions.

The **sample sum** is the **statistic**

$$Y = n\bar{X} = \sum_{i=1}^{n} X_i$$

A **statistic** is any function from the random sample $(X_1, ...X_n)$.

Theorem: if X follows a normal distribution

$$X \to N(\mu, \sigma^2)$$

then Y is normal

$$Y \to N(n\mu, n\sigma^2)$$

and Y has

1) mean
$$E(Y) = n\mu$$
2) variance
$$V(Y) = n\sigma^2$$

Example (sum of cables)

What is the probability that when we put all the cables together, they can carry a total weight between $102 = 8(13 - 0.25)$ and $106 = 8(13 + 0.25)$ Tons?

We know that for our Cables

$$X \to N(\mu = 13, \sigma^2 = 0.35^2)$$

then

$$Y \to N(n\mu = 104, n\sigma^2 = 8 \times 0.35^2)$$

with mean and variance

1) $E(Y) = n\mu = 104$
2) $V(Y) = n\sigma^2 = 8 \times 0.35^2 = 0.98$; $\sqrt{V(Y)} = 0.9899495$

We want to calculate
$$P(102 \leq Y \leq 106)$$

$= F(102; n\mu, n\sigma^2) - F(106; n\mu, n\sigma^2)$

In Python we can calculate it as:

norm.cdf(106, 104, 0.9899495)-norm.cdf(102, 104, 0.9899495)=0.956.

Therefore, 95.6% of the total weight that 8 cables can carry is between 102 and 106 Tons, or a distance of $8 \times 0.25 = 2$ Tons from the total mean $n\mu = 104$. This result offers an insight. If the cables are independent, the collective effect of many cables is determined by n-times the properties of a single cable. Here, we have a strong argument to study the collective from the individual **if** individuals do not interact. Think, for instance, of a gas made of non-interacting particles.

10.9 Sample variance

By estimating the variance

$$s^2 = \hat{\sigma}^2$$

We also make a mistake. How can we estimate the error we make?

Definition

The **sample variance** S^2 of a random sample of size n

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

is the dispersion of the measurements around $\bar{X}$. In our sample of size 8, S^2 took the value

$$s^2_{stock} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = 0.1275608$$

S^2 is

1) unbiased because its expected value is the parameter: $E(S^2) = V(X) = \sigma^2$
2) and consistent because its variance vanishes with large samples: $V(\bar{S^2}) \to 0$ when $n \to \infty$

and therefore, we technically say that S^2 is and unbiased and consistent estimator of σ^2. We can therefore take one observation s^2 as an estimate for σ^2. That is

$$s^2 = \hat{\sigma}^2$$

Similar to $\hat{\mu}$, the error of this estimate gets smaller and smaller as n gets bigger and bigger.

The unbiased sample variance (why do we divide by n-1?)

We could propose to estimate σ^2 by dividing the squared differences of $\bar{X}$ by n

$$S^2_n = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

S^2_n is therefore

1) **biased**: $E(S^2_n) = \sigma^2 - \frac{\sigma^2}{n} \neq \sigma^2$
2) but consistent $V(S^2_n) \to 0$ when $n \to \infty$

The bias term $\frac{\sigma^2}{n}$ arises because S^2_n measures the spread around $\bar{X}$ and not around μ. Remember that the error we make when we substitute $\bar{x}$ for μ is the variance of $\bar{X}$: σ^2/n. Let us correct for the bias, writing equation 1 above as:

$$E(\frac{n}{n-1}S^2_n) = \sigma^2$$

We can define the **sample variance** (corrected)

$$S^2 = \frac{n}{n-1} S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

which is an unbiased estimator of σ^2 because $E(S^2) = \sigma^2$.

Example (quality control)

We have inference problems when we are interested in the probability of the **sample variance** S^2.

Consider a quality control process that requires that the cables are produced close to the specified value μ. We don't want cables that break too far from the mean.

If a sample of size 8 cables is very scattered ($S^2 > 0.3$), we stop production and we declare that the process is out of control.

What is the probability that the sample variance of a sample of size 8 Cables will be greater than the required 0.3?

10.10 Probabilities of the sample variance

Theorem: If X follows a normal distribution

$$X \to N(\mu, \sigma^2)$$

The **statistic**:

$$W = \frac{(n-1)S^2}{\sigma^2} \to \chi^2(n-1)$$

has a distribution χ^2 (chi-square) with $df = n - 1$ degrees of freedom given by

$$f(w) = C_n w^{\frac{n-3}{2}} e^{-\frac{w}{2}}$$

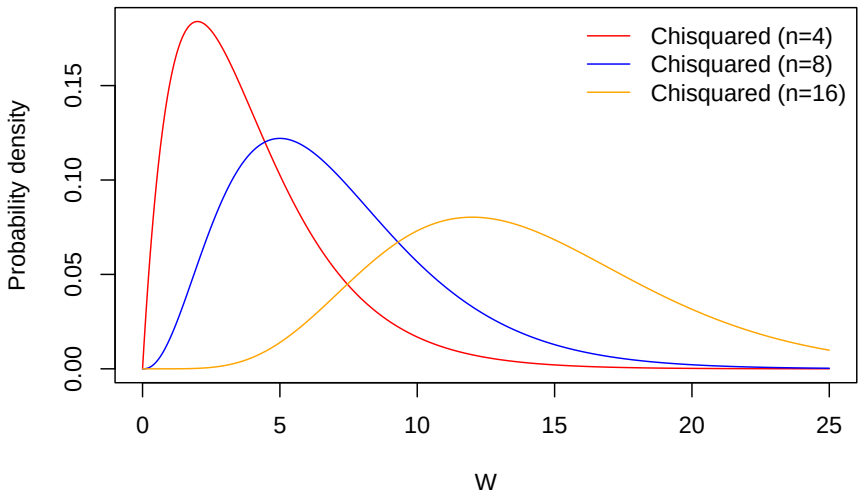
where:

1) $C_n = \frac{1}{2^{(n-1)/2}\sqrt{\pi(n-1)}}$ ensures $\int_{-\infty}^{\infty} f(t)dt = 1$
2) $\Gamma(x)$ is the Euler factorial for real numbers

If we **know** the value of σ, we can calculate the probabilities of S^2 using the χ^2 distribution for W.

10.11 χ^2-statistic

The probability density χ^2 has a parameter $df = n - 1$, called the number of degrees of freedom. Let's look at some probability densities in the family of probability models χ^2.



Example (variations in cable break)

If we **know** that our cables are certified as

$$X \to N(\mu = 13, \sigma^2 = 0.35^2)$$

so

$$W = \frac{(n-1)S^2}{\sigma^2} = \frac{7S^2}{0.35^2} \to \chi^2(n-1)$$

we can calculate

$$P(S^2 > 0.3) = P(\frac{(n-1)S^2}{\sigma^2} > \frac{(n-1)0.3}{\sigma^2})$$

$= P(W > \frac{7 \times 0.3}{0.35^2}) = P(W > 17.14286)$

$= 1 - P(W \leq 17.14286)$

$= 1 - F_{\chi^2, df=7}(17.14286) = 0.016$

in Python 1-chi2.cdf(17.14286, df=7)=0.016

There is only a 1% chance of getting a value greater than $s^2 = 0.3$. So $s^2 > 0.3$ seems to be a good criteria to stop production and review the process.

Let's imagine that the sample of 8 cables that we took was to perform a quality control test of the production line. Therefore, for our observations

```
## [1] 13.34642 13.32620 13.01459 13.10811 12.96999 13.55309 13.75557 12.62747
```

the observed sample variance was $s^2_{stock} = 0.1275608$. Thereofre, the sample is not very dispersed because $s^2_{stock} < 0.3$ and we believe that all is well and manufacturing is under control.

10.12 Questions

1) The sample mean is an unbiased estimator of the population mean because

a: The expected value of the sample mean is the population mean; **b:** The expected value of the population mean is the sample mean; **c:** The standard error approaches zero as n approaches infinity; **d:** The variance of the sample mean approaches zero as n approaches infinity;

2) Why is the statistic $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ used? instead of $S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$ to estimate the variance of a random variable?

a: because its variance is 0; **b:** because it is a consistent estimator of σ^2; **c:** because it is an unbiased estimator of σ^2; **d:** because it is the mean square distance to the sample mean ($\bar{X}$);

3) What is the variance of the sample mean $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$?

a:σ; **b:**$\frac{\sigma}{\sqrt{n}}$; **c:**σ^2; **d:**$\frac{\sigma^2}{n}$;

4) What is the mean and variance of the sample sum?

a:μ, $n\sigma$; **b:**$n\mu$,$n\sigma$; **c:**μ, $n\sigma^2$; **d:**$n\mu$, $n\sigma^2$;

5) An inference question requires to

a: calculate the expected value of an estimator; **b:** estimate the value of a parameter; **c:** calculate a probability of an estimator; **d:** fit a probability model;

10.13 Exercises

10.13.0.1 Exercise 1

An electronics company manufactures resistors that have an average resistance of 100 ohms and a standard deviation of 10 ohms. The resistance distribution is normal.

- What is the sample mean of $n = 25$ resistors? (R:100)
- What is the variance of the sample mean of $n = 25$ resistors? (R:4)
- What is the standard error of the sample mean of $n = 25$ resistors? (R:2)
- Find the probability that a random sample of $n = 25$ resistors have an average resistance of less than 95 ohms (R: 0.0062)

10.13.0.2 Exercise 2

A battery model charges an average of 75% of its capacity in one hour with a standard deviation of 15%.

- If the battery charge is a normal variable, what is the probability that the charge difference between the sample mean of 25 batteries and the mean charge is at most 5%? (R:0.9044)
- If we charge 100 batteries, what is that probability? (R:0.9991)
- If instead we only charge 9 batteries, what charge c is exceeded by the sample mean with probability 0.015? (A:85.850)

Chapter 11

Central limit theorem

11.1 Objective

In this chapter we will discuss more fully the **margin of errors** when estimating the mean of the population distribution with the average.

We will discuss how the **central limit theorem** will allow us to compute the margin of error for any type of distribution if the sample is large.

We will also introduce the t-statistic, for computing the margin of error when the sample is small but the population distribution is normal.

11.2 Margin of error

When deciding whether the error of estimation of μ by the sample mean $\bar{x}$ is large or not, we usually compare it with a **predefined** tolerance.

The **margin of error** at 5% level is the distance from μ that captures 95% of the averages $\bar{X}$:

$$P(-m \leq \bar{X} - \mu \leq m) = P(\mu - m \leq \bar{X} \leq \mu + m) = 0.95$$

This means that 95% of the outcomes of $\bar{X}$ from a random sample are a distance m from μ.

11.3 Averages of normal variables

We want to know the number m in the equation

$$P(\mu - m \leq \bar{X} \leq \mu + m) = 0.95$$

To solve this equation we need two steps. **First**, we need to know the **distribution** of $\bar{X}$.

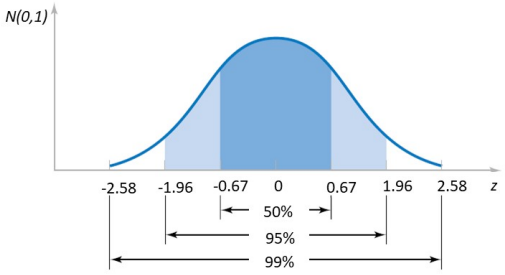
1. When X is normal ($X \to N(\mu, \sigma^2)$) then

$$\bar{X} \to N(\mu, \frac{\sigma^2}{n})$$

We **then** need to **standardize** $\bar{X}$. Remember that to standardize a normal variable, we subtract its mean an divide it by its standard deviation.

$$Z = \frac{\bar{X} - E(\bar{X})}{\sqrt{V(\bar{X})}} = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \to N(0,1)$$

Z is then standard normal. Remember the probability density function for a standard normal variable and its quantiles



2. Substituting the mean of $\bar{X}$ and its standard deviation into the equation for the margin of error, we have:

$$P(\mu - m \leq \bar{X} \leq \mu + m) = P(-\tfrac{m}{\sigma/\sqrt{n}} \leq \tfrac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \leq \tfrac{m}{\sigma/\sqrt{n}})$$

$$= P(-\frac{m}{\sigma/\sqrt{n}} \leq Z \leq \frac{m}{\sigma/\sqrt{n}}) = 0.95$$

Compare it with the plot above, where

$$P(-1.96 \leq Z \leq 1.96) = 0.95$$

Therefore,

$$\frac{m}{\sigma/\sqrt{n}} = 1.96$$

is the distance from $z = 0$ that captures 95% of the distribution of the standard normal variable. Threfore, the margin of error at 5% is

$$m = 1.96\frac{\sigma}{\sqrt{n}}$$

The limits of the interval $(-1.96, 1.96)$ for Z are computed from

$$\phi^{-1}(0.975) = 1.96$$

where ϕ^{-1} is the inverse of the standard normal distribution (norm.ppf(0.975)), and retrieves the value of z that has accumulated 97.5% of probability. 2.5% are left out at the higher tail of the distribution, that add to the other 2.5% left out at the left of $\phi^{-1}(0.025) = -1.96$, because the distribution is symmetric.

Example (cables)

If we take a sample of 8 cables from a population of cables whose breaking load follows a normal distribution with **known** parameters $\mu = 13$ and $\sigma^2 = 0.35^2$,

$$X \to N(\mu = 13, \sigma^2 = 0.35^2)$$

What is the margin of error of the average breaking load of a sample of size 8, when we estimate μ by $\bar{x}$?

The sample mean $\bar{X}$ has:

1. mean $E(\bar{X}) = \mu$

and

2. standard error $se = \sqrt{V(\bar{X})} = \frac{\sigma}{\sqrt{n}} = \frac{0.35}{\sqrt{8}}$

Then the margin of error at 5% is:

$$m = 1.96\frac{0.35}{\sqrt{8}} = 0.24$$

We can expect that 95% of the averages ($\bar{x}$) for the breaking load of 8 cables fall between

$$(13 - 0.24, 13 + 0.24) = (12.76, 13.24)$$

11.4 Central Limit Theorem

We could solve the margin of error because we assumed that that variable X was normal. What if X follows any other probability distribution?

Theorem: For any random variable X with any type of probability function

$$X \to f(x)$$

the standardized statistic

$$Z = \frac{\bar{X} - E(\bar{X})}{\sqrt{V(\bar{X})}}$$

approximates to a standard distribution

$$Z \to_d N(0,1)$$

when $n \to \infty$

Consequence: We can compute probabilities for $\bar{X}$ if n is large, using the normal distribution:

$$\bar{X} \sim_{aprox} N(\mu, \frac{\sigma^2}{n})$$

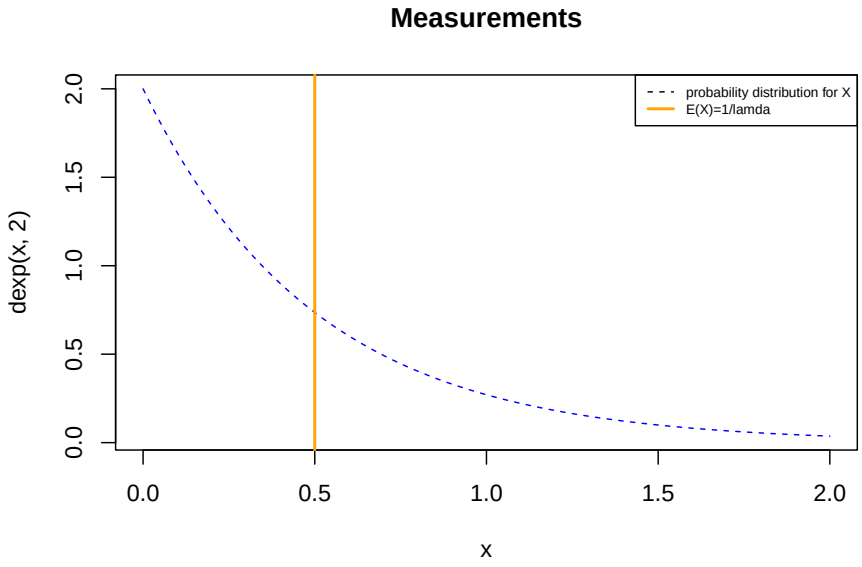
This is topically a good approximation when $n \geq 30$.

Example (drug in blood concentration)

Consider an experiment where we want to measure the concentration in blood of a drug after 10-hour administration in 30 patients.

If we **know** that levels follow an exponential distribution

$$X \to exp(\lambda = 2)$$



The mean and variance are:

- $E(X) = \frac{1}{\lambda} = 0.5$
- $V(X) = \frac{1}{\lambda^2} = 0.25$

Therefore the mean and the estandard error of $\bar{X}$ are:

- $E(\bar{X}) = \frac{1}{\lambda} = 0.5$
- $se = \sqrt{\frac{V(X)}{n}} = \sqrt{\frac{1}{n\lambda^2}} = 0.091$

As $n \geq 30$

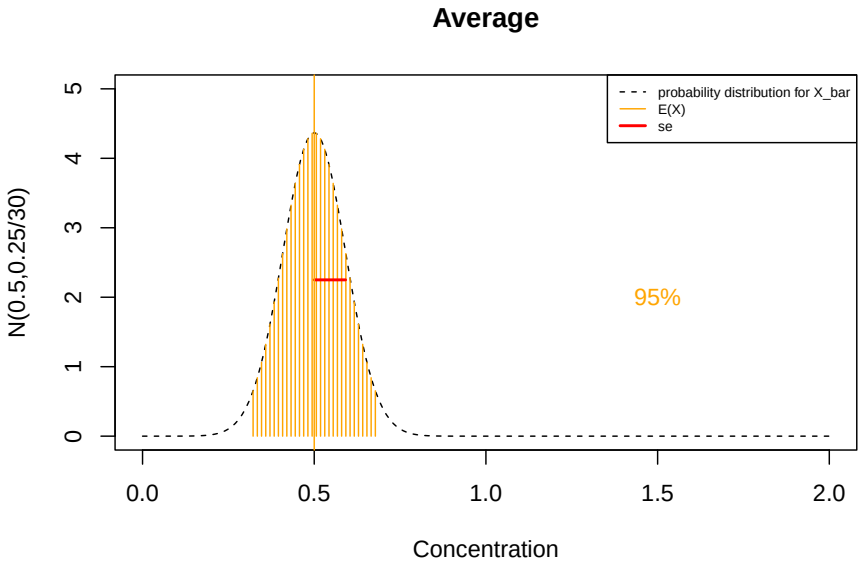
$$Z = \frac{\bar{X} - \lambda}{\sqrt{\frac{1}{n\lambda^2}}}$$

is almost standard normal variable and: $\bar{X} \sim_{aprox} N(\lambda, \frac{1}{n\lambda^2})$

The margin of error at 5% level can be computed again with the standard distribution

$$m = \phi^{-1}(0.975)\sqrt{\frac{V(X)}{n}} = 1.96\sqrt{\frac{0.25}{30}} = 0.1789227$$

We can expect 95% of the averages of 30 patient samples to fall between $(0.5 - 0.178, 0.5 + 0.178) = (0.322, 0.678)$



11.5 Sample sum and CLT

The **sample sum** is the **statistic**

$$Y = X_1 + X_2 + ...X_n = \sum_{i=1}^{n} X_i = n\bar{X}$$

with

1) mean

$$E(Y) = n\mu$$

2) variance

$$V(Y) = n\sigma^2$$

The CLT tells us that for any random variable X with **unknown** (any type of) distribution

$$X \to f(x)$$

the standardized statistic

$$Z = \frac{\bar{X} - E(\bar{X})}{\sqrt{V(\bar{X})}}$$

approximates to a standard distribution

$$Z \to_d N(0,1)$$

when $n \to \infty$. Z can also be written as

$$Z = \frac{n\bar{X} - nE(\bar{X})}{\sqrt{n^2 V(\bar{X})}} = \frac{Y - E(Y)}{\sqrt{V(Y)}} = \frac{Y - n\mu}{\sqrt{n}\sigma}$$

Consequence: We can compute probabilities for the sample sum $Y = n\bar{X}$ if n is large, using the normal distribution:

$$Y \sim_{aprox} N(n\mu, n\sigma^2)$$

Example (Bernoulli trial)

For a Bernoulli trial $X \to Bernoulli(p)$, its mean is $\mu = p$ and variance $\sigma^2 = p(1-p)$.

Now, the sample sum $Y = n\bar{X}$ is a random variable that counts the number of events with probability p in a repetition of n trials, therefore

$$Y \to Binom(n,p)$$

with mean $E(Y) = np$ and variance $V(Y) = np(1-p)$. If we apply the CLT for the sample sum of Brenouilli trials then we have

$$Y \to Binom(n,p) \sim_{aprox} N(np, np(1-p))$$

we can then proximate the binomial probability mass function with the normal probability density when n is big. This approximation is good when both np and $n(1-p)$ are greater than 5.

11.6 Questions

1) A magnetic resonance imaging of the brain's hippocampus has 100 pixels. We expect 90% of the pixels to be white (brain tissue). According to the central limit theorem, what is the probability that the scanning of a patient has at most 85% of white pixels?

a:norm.cdf(0.9, 0.85, sqrt(0.85*0.15)/10); **b:**norm.pdf(0.85, 0.9, sqrt(0.9*0.1)/10); **c:**norm.cdf(0.85, 0.9, sqrt(0.9*0.1)/10); **d:**norm.pdf(0.9, 0.85, sqrt(0.85*0.15)/10)

2) For a standard normal variable, if we the number $z_{0.025}$ in the definition of the margin of error $m = z_{0.025}\frac{\sigma}{\sqrt{n}}$, then it will refer to

a: The first quartile; **b:** The number at which the distribution has accumulated 0.975 of probability; **c:** The number at which the distribution has accumulated 0.025 probability; **d:** The third quartile;

3) The importance of the central limit theorem is that is applyes to the standardization of

a: A random variable; **b:** The sample mean of a normal variable; **c:** The sample mean of a random variable; **d:** A normal variable;

11.7 Exercises

11.7.0.1 Exercise 1

An electronic component is needed for the correct functioning of a telescope. It needs to be replaced immediately when it wears out.

The mean life of the component (μ) is 100 hours and its standard deviation σ is 30 hours.

- what is the probability that the average of the mean life of 50 components is within 1 hour from the mean life of a single component? (A: 0.1863)
- How many components do we need such that the telescope is operational 2750 consecutive hours with at least 0.95 probability? (A: 31)

11.7.0.2 Exercise 2

The probability that a particular mutation is found in the population is 0.4. If we test 2000 people for the mutation:

- What is the probability that the total number of people with the mutation is between 791 and 809? (A: 0.31)

hint: Use the CLT with a sample of 2000 Bernoulli trials. This is known as the normal approximation of the binomial distribution.

11.7.0.3 Exercise 3

An automated machine fills test tubes with biological samples with mean $\mu =$ 130mg and a standard deviation of $\sigma =$ 5mg. For a random sample of size 50

- What is the probability that the sample mean (average) is between 128 and 132gr? (A: 0.995)
- what is the margin of error at 5%? (A: 1.385929)
- what should be the size of the sample n such that the margin of error at 5% is 1? (A: 97)

11.7.0.4 Exercise 4

In the Caribbean, there appears to be an average of 6 hurricanes per year. Considering that hurricane formation is a Poisson process, meteorologists plan to estimate the mean time between the formation of two hurricanes. They plan to collect a sample of size 36 for the times between two hurricanes.

- What is the probability that their sample average is between 45 and 60 days? (A: 0.39)
- Which should be the sample size such that they have a probability of 0.025 that the sample mean is greater than 70 days? (A: 169)

Chapter 12

Maximum likelihood

12.1 Objective

In this chapter we will further discuss what an **estimator** is and give some examples. Then we will introduce a general method for obtaining the **estimators** of model parameters. This is the **maximum likelihood** method.

12.2 Statistic

Definition

A **statistic** is any function of a **random sample**

$$T(X_1, X_2, ..., X_n)$$

It usually returns a number.

Statistics are **random variables** and their **probability distributions** are called **sampling distributions**.

Statistics have different functions:

1. **Description** of a sample's data

- location: $\bar{X}$
- Minimum: $\min\{X_i\}$
- Maximum: $\max\{X_i\}$

2. **Estimation** of a probability model's **parameters**

- We use $\bar{X}$ to estimate μ
- We use S^2 to estimate σ^2

3. **Inference** to say something about the parameters given the data

- for the mean we will use the statistics Z, T
- for the variance we will use χ^2

This last point will be developed from chapter 14 on wards.

Remember that statistics are **random variables**. Every time we take a new sample their value change.

Definition of estimators

An **estimator** is a statistic whose observed values are used to estimate the **parameters** of the population distribution on which the sample is defined.

If we write the probability model of the population as the probability function

$$X \to f(x;\theta)$$

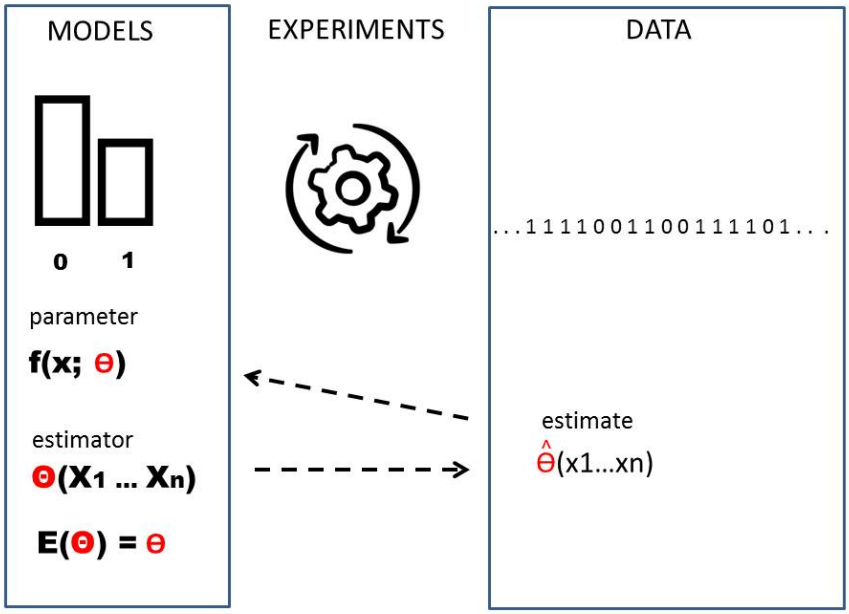
then θ is a parameter and the statistics Θ is a random variable whose observations $\widehat{\theta}$ we take as estimations of θ.

$$\widehat{\theta} \sim \theta$$

Let's unpack this statement. There are three different quantities that we must consider:

1. θ is a **parameter** of the population distribution $f(x;\theta)$.
2. Θ is an **estimator** of θ: A random variable.
3. $\widehat{\theta}$ is the **estimate** of θ: An outcome, a realized value of Θ.

Let's see where the different quantities live



Example (Sample mean)

When we have a normal random variable

$$X \to N(\mu, \sigma^2)$$

we can identify the three different quantities:

1. The mean: μ is a **parameter** of the **population** distribution $N(\mu, \sigma^2)$.
2. The average: $\bar{X}$ is an **estimator** of μ.
3. The point estimate of the mean: $\bar{x} = \hat{\mu}$ is the **estimate** of μ.

Example (Sample variance)

When we have a normal random variable

$$X \to N(\mu, \sigma^2)$$

1. σ^2 is a **parameter** of the population distribution
2. S^2 is an **estimator** of σ^2
3. $s^2 = \hat{\sigma}^2$ is the **estimate** of σ^2

12.3 Properties

1. An estimator Θ is **unbiased** if its expected value is the parameter

$$E(\Theta) = \theta$$

For example:

- $\bar{X}$ is an **unbiased** estimator of μ because $E(\bar{X}) = \mu$
- S^2 is an **unbiased** estimator of σ^2 because $E(S^2) = \sigma^2$

2. An estimator is **consistent** when its observed values get closer and closer as the sample size is increased

$$lim_{n\to\infty} V(\Theta) = 0$$

For example:

- $\bar{X}$ is **consistent** because $V(\bar{X}) = \frac{\sigma^2}{n} \to 0$ when $n \to \infty$.

3. The mean squared error mse of Θ is its expected squared difference from the parameter

$$mse(\Theta) = E([\Theta - \theta]^2)$$

or equivalently is the sum of the errors in consistency and bias

$$mse(\Theta) = se^2 + bias^2$$

where $se = \sqrt{V(\Theta)}$ is the standard error.

12.4 Maximum likelihood

How can we obtain **estimators** of the parameters of **any** probability model?

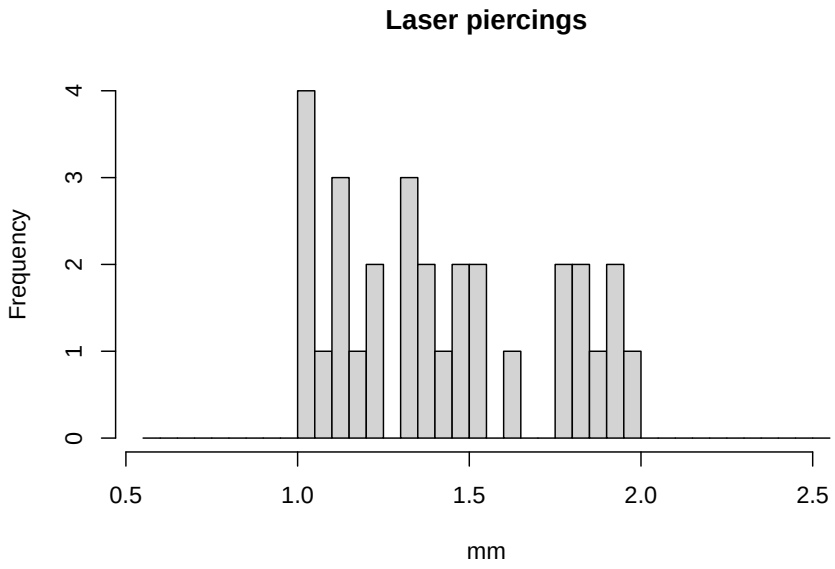
Example (Laser)

Imagine we design a laser with a diameter of $1mm$ that we want to use for clinical applications.

We want to characterize the diameter of a piercing in a tissue made with the laser and take a random sample of 30 cuts made with the laser. Here are the results

```
##  [1] 1.11 1.64 1.20 1.79 1.89 1.01 1.31 1.81 1.34 1.25 1.92 1.24 1.49 1.36 1.03
## [16] 1.82 1.09 1.01 1.14 1.91 1.80 1.51 1.44 1.98 1.46 1.53 1.33 1.39 1.12 1.04
```

and the histogram



What is a probability function that can describe the data?

For this we follow the following process:

1. we propose **a model** that depends on parameters,
2. we derive the **estimators** for the parameters, by maximum likelihood or the method of moments,
3. finally we use the estimator to **estimate the parameters** with the data.

Proposing a probability density

In many applications, we can propose the form of a probability density that depends on some parameters. Proposing a probability model is done by following **general properties** of the observations, or what we expect to observe. Modelling requires experience, skill and knowledge of several mathematical functions. However, in most cases **well known models** are typically applied.

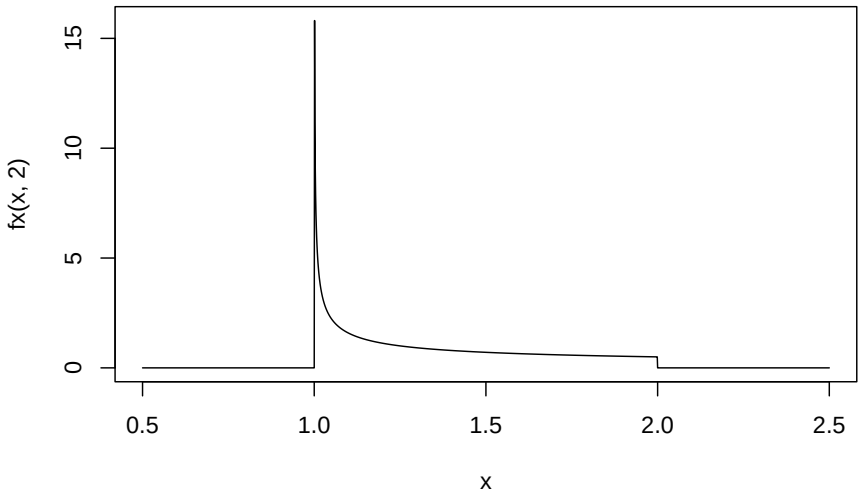
Example (Laser)

In our example, we may consider for example that maximum probability should be given to diameters of $x = 1mm$, and that the diameters should decrease as the inverse power of some **unknown** parameter α, with a limit of $2mm$ beyond which the probability is 0.

A suitable probability density distribution is

$$f(x) = \begin{cases} \frac{1}{\alpha}(x-1)^{\frac{1}{\alpha}-1}, & \text{if } x \in (1,2) \\ 0, & x \notin (1,2) \end{cases}$$

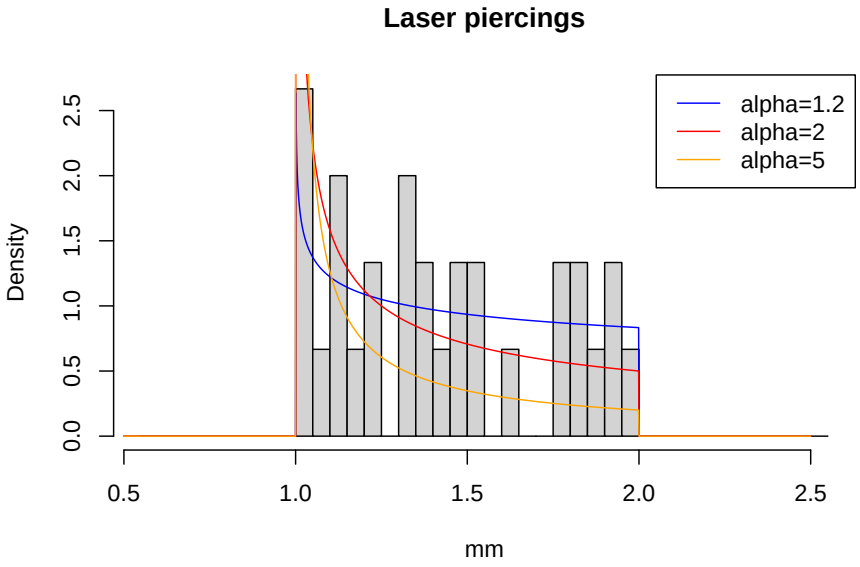
Where α is a parameter. This is a probability density because it integrates to one and it is positive. In particular, for $\alpha = 2$ we can plot it



Deriving the estimators

If we perform a n-sample: $X_1, ...X_n$, how should we combine the data for obtaining the best value of α?

Many values of the parameters could explain the data. We are interested in **one criterion** to choose one particular value.



The **maximum likelihood** method will gives us the estimator for α

$$\hat{\alpha}_{ml}$$

using a probabilistic argument.

12.5 Maximum likelihood

The objective is to find the value of the parameter that we **believe** can **best** represent the data.

The method of maximum likelihood is based on the search for the parameter value that makes the **observation** of the sample the most **probable**.

Remember, a random sample is a random variable

$$M = (X_1, ...X_n)$$

an observed sample is an outcome of M, a set of numerical values that we actually obtained when we repeated the random experiment

$$m = (x_1, ...x_{30}) = c(1.11, 1.64, ...1.04)$$

Maximum likelihood step 1

We firt calculate the probability of having observed the particular n-sample: $x_1, ...x_n$. This is the product of probabilities for each observation because the repetitions of the random experiment are independent of one another:

$$P(M = x_1, ...x_n) = P(X = x_1)P(X = x_2)...P(X = x_n)$$

$$= f(x_1; \alpha)f(x_2; \alpha)...f(x_n; \alpha)$$

We call this function the **likelihood function** and we consider that:

- Once the data are observed, they are **fixed**
- The unknown is α

$$L(\alpha) = \Pi_{i=1..n} f(x_i; \alpha)$$

Example (Laser)

For the laser experiment the likelihood is

$$L(\alpha; x_1, ..x_n) = \frac{1}{\alpha^n}\Pi_{i=1..n}(x_i - 1)^{\frac{1-\alpha}{\alpha}} = \frac{1}{\alpha^n}\{(x_1 - 1)(x_2 - 1)...(x_n - 1)\}^{\frac{1-\alpha}{\alpha}}$$

Maximum likelihood step 2

We then ask: what is the value of α that makes the observed sample the most probable event? We thus want to maximize $L(\alpha)$ with respect to α. Since we have the multiplication of many factors, it is easier to maximize the logarithm of $L(\alpha)$. This is called the the log-likelihood function:

$$\ln L(\alpha; x_1, ..x_n)$$

Example (Laser)

In the laser example, we therefore take the logarithm and obtain the **Log-likelihood**

$$\ln L(\alpha; x_1, ..x_n) = -n \ln(\alpha) + \frac{1-\alpha}{\alpha}\Sigma_{i=1...n} \ln(x_i - 1)$$

Maximum likelihood step 3

Finally we **maximize** the log-likelihood with respect to the parameter. Therefore, we differentiate the log-likelihood with respect to the parameter α, equate to zero and solve for the maximum.

$$\frac{d \ln L(\alpha)}{d\alpha}\Big|_{\hat{\alpha}} = 0$$

The value of the parameter at the maximum is called the **maximum likelihood estimate** for the parameter and it is written with a hat $\hat{\alpha}$.

Example (Laser)

We derive the log-likelihood

$$\frac{d \ln L(\alpha)}{d\alpha} = -\frac{n}{\alpha} - \frac{1}{\alpha^2} \Sigma_{i=1\ldots n} \ln(x_i)$$

The maximum is where the derivative is 0. This maximum is the value of our estimator $\hat{\alpha}_{ml}$.

$$\hat{\alpha}_{ml} = -\frac{1}{n} \sum_{i=1}^{n} \ln(x_i - 1)$$

The estimator of the parameter is therefore (note the capital letters)

$$A = -\frac{1}{n} \sum_{i=1}^{n} \ln(X_i - 1)$$

Which is a random variable, function of the random sample

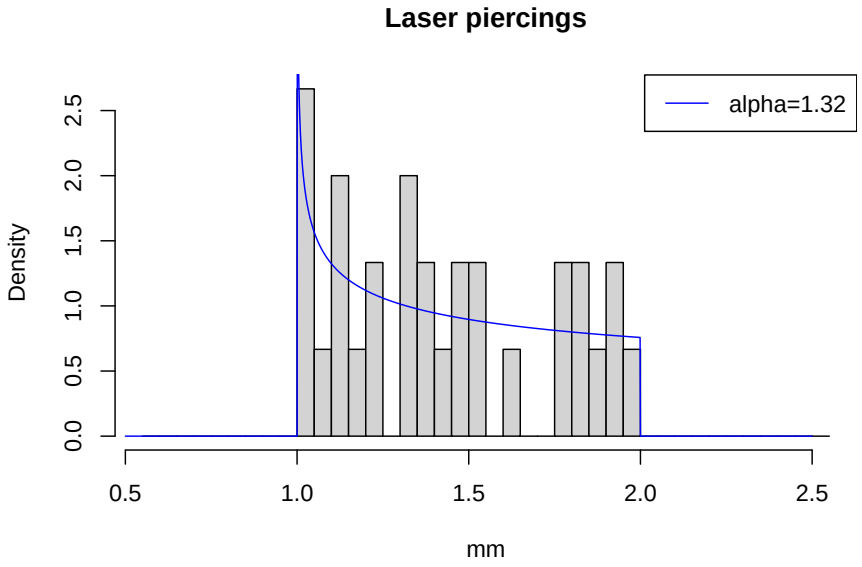
$$(X_1, X_2, \ldots X_n)$$

Estimating the parameters with the data

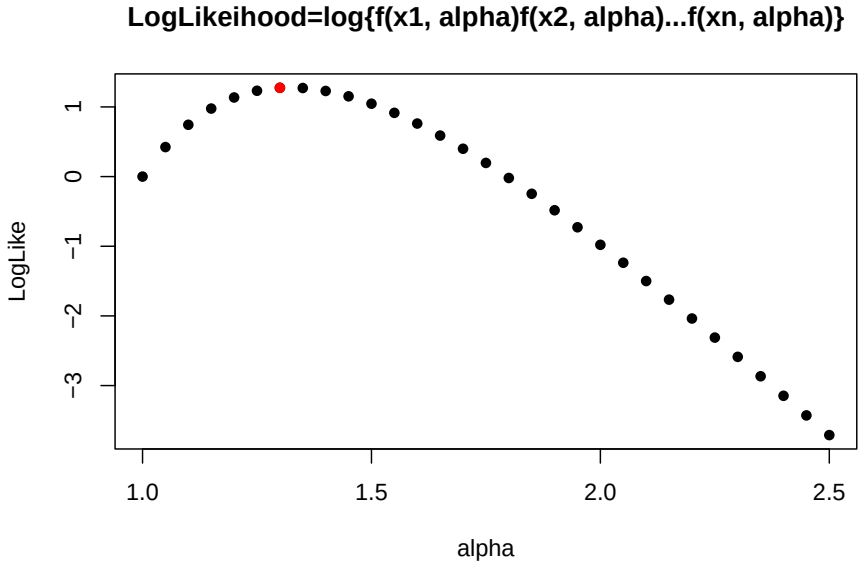
In our example, we then have the observation of the random sample as a set of 30 numbers $(x_1, x_2, \ldots x_{30})$, we therefore substitute the numbers in the estimator and this will give us its observed value.

$\hat{\alpha}_{ml} = -\frac{1}{n}\{\ln(1.11 - 1) + \ln(1.64 - 1) + \ldots \ln(1.04 - 1)\} = 1.320$

Therefore the maximum likelihood estimate of the parameter is 1.320. If we substitute this value in the probability function, and overlay it with the histogram, we can see that it gives us a suitable description of the data.



Let's look at the log-likelihood function for our 30 laser cuts. Remember, data is fixed by our experiment and α varies. The function has a maximum. However, if we take another sample this function changes and so does its maximum.



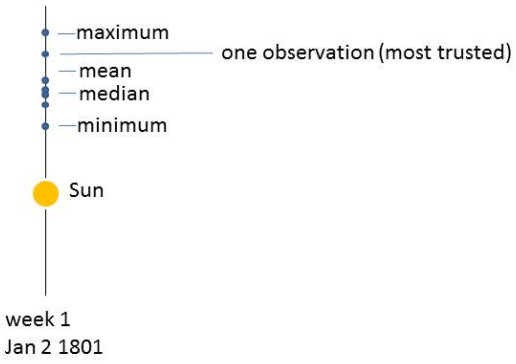
Maximum likelihood: History

To infer the true position of Ceres at a given time, Gauss derived the error function

$$f(x;\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

Where the **true** position of Ceres was the mean μ. How can we combine the data for having the best estimate for the position of Ceres?

What is the statistic that can describe best its position?



This question can be formulated as: What is the maximum likelihood estimate of μ for a random normal variable?

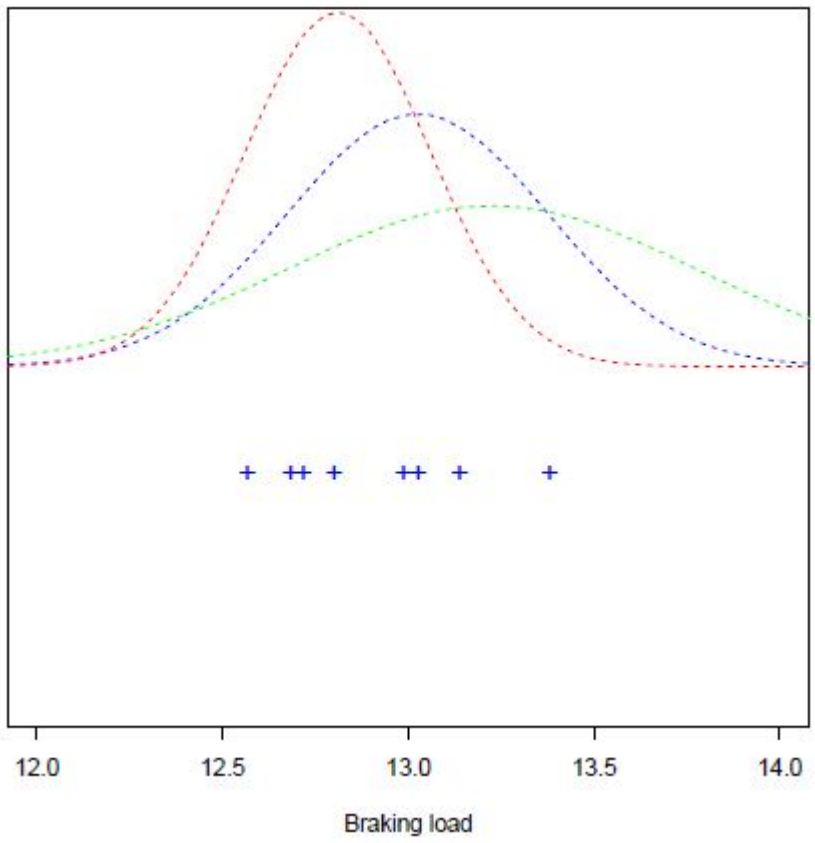
Maximum likelihood of the normal distribution

For a random normal variable

$$X \to N(\mu,\sigma^2)$$

.

What are the estimators of μ and σ^2 that maximize the probability of the observed data?



We follow the maximum likelihhod method:

1. The likelihood function, or the probability of having observed the sample $(x_1, ...x_n)$ is

$L(\mu, \sigma^2) = \Pi_{i=1..n} f(x_i; \mu, \sigma)$

$$= (\frac{1}{\sigma\sqrt{2\pi}})^n e^{-\frac{1}{2\sigma^2}\sum_i (x_i-\mu)^2}$$

2. We take the log of L, and compute the **log-likelihood**

$$\ln L(\mu, \sigma^2) = -n \ln(\sigma\sqrt{2\pi}) - \frac{1}{2\sigma^2}\Sigma_i (x_i - \mu)^2$$

The estimates of μ and σ^2 are where the likelihood is maximum. They give the highest probability for the data.

3. We differentiate with respect to μ and σ^2. These two derivatives give us two equations, one for each of the parameters. For deriving respect to σ^2, it is easier to make a substitution $t = \sigma^2$.

a) $\frac{d \ln L(\mu,\sigma^2)}{d\mu} = \frac{1}{\sigma^2}\sum_i (x_i - \mu)$

b) $\frac{d \ln L(\mu,\sigma^2)}{d\sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_i (x_i - \mu)^2$

The derivatives are 0 at the maxima

a) $\frac{1}{\hat{\sigma}^2}\sum_i (x_i - \hat{\mu}) = 0$
b) $-\frac{n}{2\hat{\sigma}^2} + \frac{1}{2\hat{\sigma}^4}\sum_i (x_i - \hat{\mu})^2 = 0$

solving both equations for the parameters we find for μ

$$\hat{\mu}_{ml} = \frac{1}{n}\sum_i x_i = \bar{x}$$

and for σ^2

$$\hat{\sigma}^2_{ml} = \frac{1}{n}\sum_i (x_i - \bar{x})^2$$

Therefore the average $\bar{X}$ is the maximum likelihood estimator of the mean μ. Gauss showed that the statistics that we should trust most (that with highest likelihood) for the real position of the Ceres was the **average**. Gauss solving the position of Ceres, not only discovered the normal distribution, but also created the regression analysis and showed the importance of the average. It is due to him that we use the average for many things, and not some other statistic.

In addition, the maximum likelihood estimator of σ^2 is a **biased** estimator because it can be shown that

$$E(\hat{\sigma}^2_{ml}) = \sigma^2 + \frac{\sigma^2}{n} \neq \sigma^2$$

It was Fisher who showed that this estimator was important, as he used it to generalize the central limit theorem

12.6 Questions

1) An estimator is not

a: a statistic; **b:** a random variable; **c:** discrete; **d:** an observation of the parameter;

2) An estimator is unbiased if

a: it is the parameter that it estimates; **b:** depends on $1/n$; **c:** its variance is small; **d:** its expected value is the parameter it estimates;

3) An estimator is consistent if

a: it is the parameter that it estimates; **b:** depends on $1/n$; **c:** its variance is small; **d:** its expected value is the parameter it estimates;

4) The maximum likelihood method

a: Produces estimators based on the probability of the observations; **b:** produces unbiased estimators; **c:** produces consistent estimators; **d:** produces estimators equal to those of the method of moments;

12.7 Exercises

12.7.0.1 Exercise 1

Take a random variable with the following probability density function

$$f(x) = \begin{cases} (1+\theta)x^{\theta}, & \text{if } x \in (0,1) \\ 0, & x \notin (0,1) \end{cases}$$

- What is the maximum likelihood estimate for θ?
- If we take a 5-sample with observations $x_1 = 0.92$; $x_2 = 0.79$; $x_3 = 0.90$; $x_4 = 0.65$; $x_5 = 0.86$

What is the estimated value of the parameter θ?

- Compute $E(X) = \mu$ as a function of θ. What is the maximum likelihood estimate for μ?

12.7.0.2 Exercise 2

For a random variable with a binomial probability function

$$f(x;p) = \binom{n}{x} p^x (1-p)^{n-x}$$

- What is the maximum-likelihood estimator of p for a sample of size 1 of this random variable?
- In **one** exam of 100 students we observed $x_1 = 68$ students that passed the exam. What is the estimate of the p?

12.7.0.3 Exercise 3

Take a random variable with the following probability density function

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & \text{if } 0 \leq x \\ 0, & otherwise \end{cases}$$

- What is the maximum likelihood estimate for λ?
- If we take a 5-sample with observations $x_1 = 0.223 \quad x_2 = 0.681; \quad x_3 = 0.117; \quad x_4 = 0.150; \quad x_5 = 0.520$

What is the estimated value of the parameter λ?

- What is the maximum likelihood estimate of the parameter $\alpha = \frac{n}{\lambda}$
- Is α an unbiased and consistent estimator of the mean of the sample sum $E(Y)$, where $Y = \sum_1^n X_i$?

Chapter 13

Interval estimation

13.1 Objective

In this chapter, we will introduce the concept of **confidence intervals** for means, proportions and variances.

We will derive the formulas for the confidence intervals under different conditions, like normality with known and unknown variance, and large n.

13.2 Estimation of the mean

We have seen that whenever we take a random sample $(X_1, X_2, ...X_n)$, the sample mean

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

is an estimator of the mean μ of the random variable X. The estimator is **unbiased** because its values center about the parameter it is estimating

- $E(\bar{X}) = \mu$

and **consistent** because as n increases then it is closer to the parameter because its variance gets smaller

- $V(\bar{X}) = \frac{\sigma^2}{n}$

where σ^2 is the variance of X. We call the quantity $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ the **standard error** (se).

These properties justify that we take the value $\bar{x}$ derived from **one particular sample** as the the value of μ, which is a characteristic of the **population**. That is

$$\bar{x} = \hat{\mu}$$

Since $\bar{X}$ is a random variable the estimation of the mean changes when we take another sample. Therefore, we know that we are making an error every time we take $\bar{x}$ for μ.

13.3 Margin of error

When deciding whether the **error** in estimation

$$\bar{X} - \mu$$

is large or not, we usually compare it with a predefined tolerance. **If we know** that the distribution of X is normal, $X \to N(\mu, \sigma^2)$, and we also know the values of the parameters μ and σ^2, we can then assess how far the estimation of $\bar{x}$ would falls from μ.

We defined the **margin of error** at 5% level as the distance m such that distribution of $\bar{X}$ captures 95% of the estimations:

$$P(-m \leq \bar{X} - \mu \leq m) = P(\mu - m \leq \bar{X} \leq \mu + m) = 0.95$$

If $\bar{X}$ is normally distributed then the marging of error is

$$m = z_{0.025}\frac{\sigma}{\sqrt{n}} = 1.96 \times se$$

where $z_{0.025} = \phi^{-1}(0.975) = \text{norm.ppf}(0.975)$. It is the value of z that leaves out to the right 2.5% of the probability function.

Example (cables):

We want to perform take a random sample of size 8: Load a cable until it breaks and record the breaking load.

Before we take the sample, **if we know** that our cables truly distribute as

$$X \to N(\mu = 13, \sigma^2 = 0.35^2)$$

then

$$\bar{X} \to N(13, \frac{0.35^2}{8})$$

With mean $E(\bar{X}) = 13$ and standard error $se = \frac{0.35}{\sqrt{8}} = 0.1237$

Therefore the margin of error at 95% is

$$m = z_{0.025}\frac{\sigma}{\sqrt{n}} = 1.96 \times se = 1.96\frac{0.35}{\sqrt{8}} = 0.24$$

We therefore expect that %95 of the averages will be a distance of 0.24 from the mean. Note that, in this case, the margin of error only depends on σ and therefore is a measure of **precision**. Furthermore, it does not care where μ really is, and therefore is **not** a measure of **accuracy**.

Now, let us take a random sample for which we find the results

```
## [1] 13.34642 13.32620 13.01459 13.10811 12.96999 13.55309 13.75557 12.62747
```

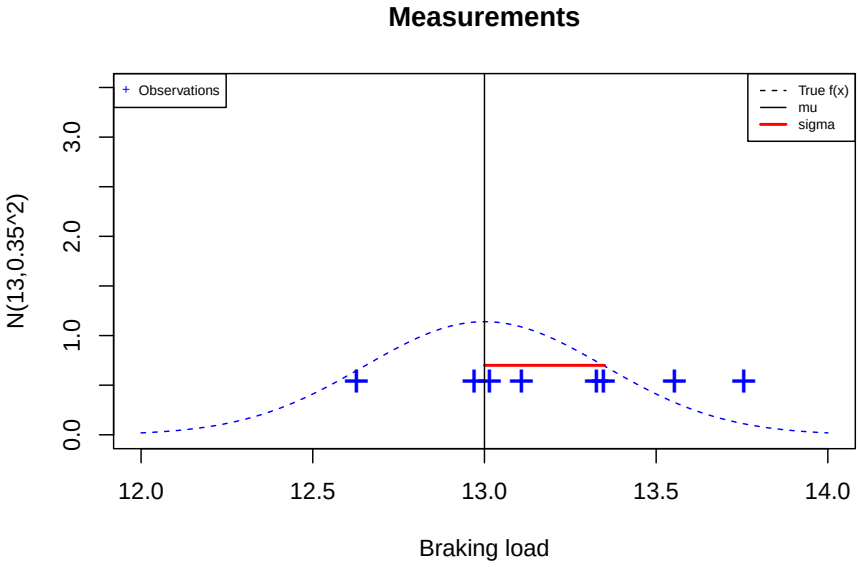
The **observed average** is $\bar{x} = 13.21$, and therefore the error we would make if we replaced μ by $\bar{x}$, would be

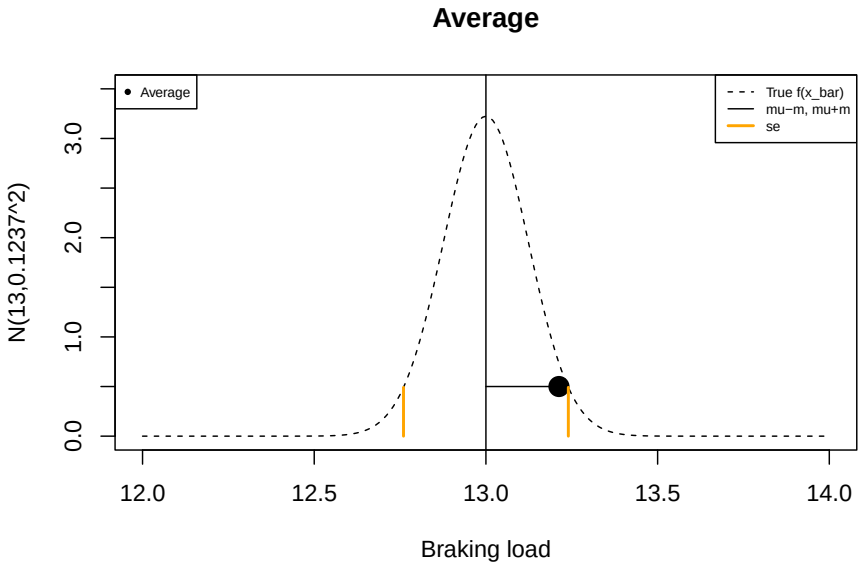
$$\bar{x} - \mu = 13.21 - 13 = 0.21$$

The **observed error** is within the margin of error

$$\bar{x} - \mu < m$$

When we use $\bar{x}$ for the value of μ, we say that we are performing a **point estimation** of the parameter. We are interested in such replacement in the cases where we do not know μ.



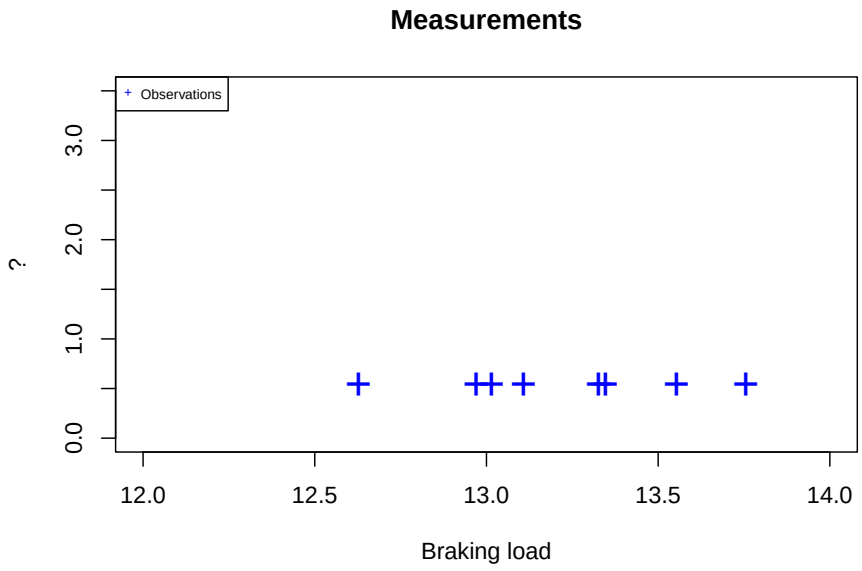


13.4 Interval estimation for the mean

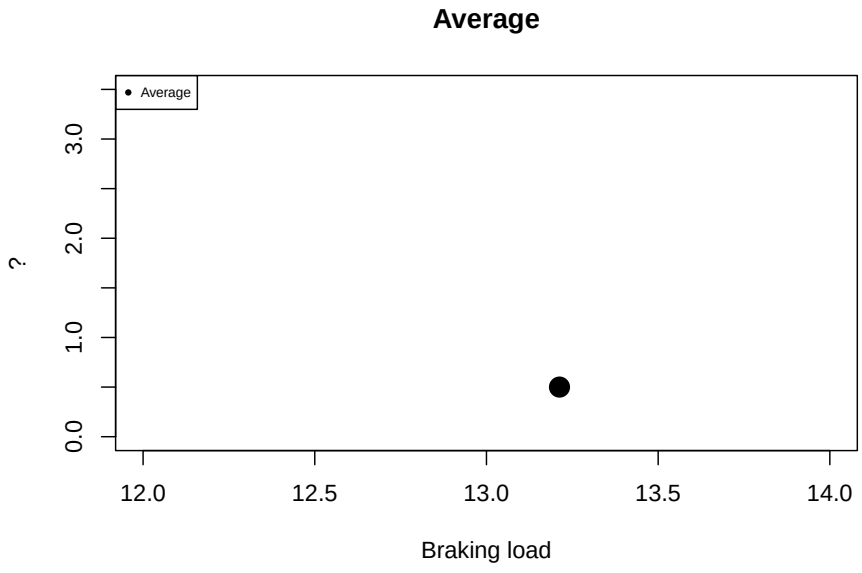
The problem is that in real life we **do not know** the real values of μ or σ for X given

$$X \to N(\mu, \sigma^2)$$

What we do is to start by taking a sample



we then compute the mean



and that's it. What is then the value of μ?

Our data suggest that μ could be $\bar{x} = 13.21$. But, how **confident** are we? after

all, we know that when we take $\bar{x}$ for μ, we are making a mistake and do not how big it actually is.

To address this question, we are going to define the **confidence interval** for μ. From the margin of error equation

$$P(-m \leq \bar{X} - \mu \leq m) = 0.95$$

let's solve for μ that is indeed **the real unknown**

$$P(\bar{X} - m \leq \mu \leq \bar{X} + m) = 0.95$$

The left and right limits of the inequality are random variables which motivate the definition for the **random confidence interval at** 95%**:**

$$(L, U) = (\bar{X} - m, \bar{X} + m)$$

This interval is a new **random variable** and it has by definition a probability of 0.95 to contain μ.

The **observed interval** that we obtain from the experiment is (lower case)

$$(l, u) = (\bar{x} - m, \bar{x} + m)$$

This interval either contains or does not the parameter μ: we will **never know**!

However, we can still say that we have a confidence of 95% that the interval (l, u) has captured the true unknown parameter μ. Think of buying a lottery scratch ticket that we cannot scratch to see the prize. The ticket either has or does not have the prize, only that we do not know which case it is.

13.4.1 Case 1 (known variance)

Confidence intervals can be estimated in different cases. The first case is when

1. X is a normal variable, and
2. we know the value of σ

the confidence interval at 95% is

$$(l, u) = (\bar{x} - m, \bar{x} + m)$$

where

$$m = z_{0.025} \frac{\sigma}{\sqrt{n}}$$

That is:

$$(l, u) = (\bar{x} - z_{0.025}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{0.025}\frac{\sigma}{\sqrt{n}})$$

Example (cables):

In our example, we assume that X is normally distributed and that we know $\sigma^2 = 0.35^2$.

1. Since $\bar{X}$ is normal, the margin of error is

$$m = z_{0.025}\frac{\sigma}{\sqrt{n}}$$

2. Since we know $\sigma^2 = 0.35^2$, then the 95% confidence interval is

$(l, u) = (\bar{x} - m, \bar{x} + m) =$

$$(\bar{x} - z_{0.025}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{0.025}\frac{\sigma}{\sqrt{n}}) = (12.97, 13.45)$$

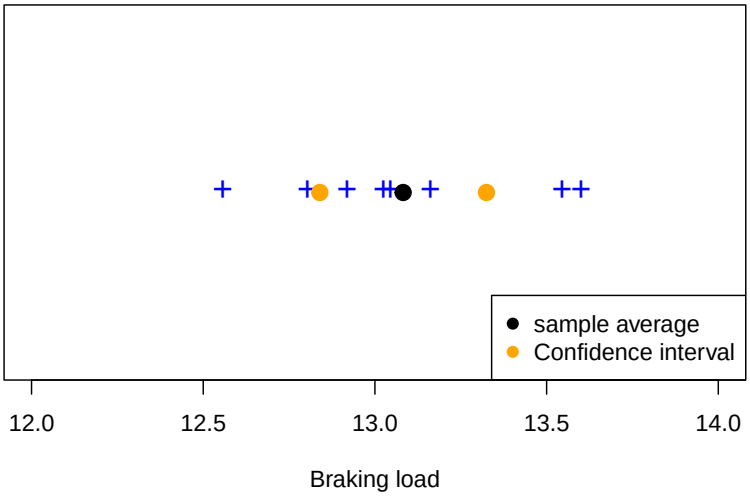
In Python we can compute the confidence interval with

```
# https://pypi.org/project/bioinfokit/

!pip install bioinfokit
from bioinfokit.analys import stat
import pandas as pd

data = {'x': [13.34642, 13.32620, 13.01459, 13.10811, 12.96999, 13.55309, 13.75557, 12.62747]}
df = pd.DataFrame(data)

res = stat()
res.ztest(df=df, x='x', mu=13, x_std=0.35, test_type=1)
print(res.summary)
```

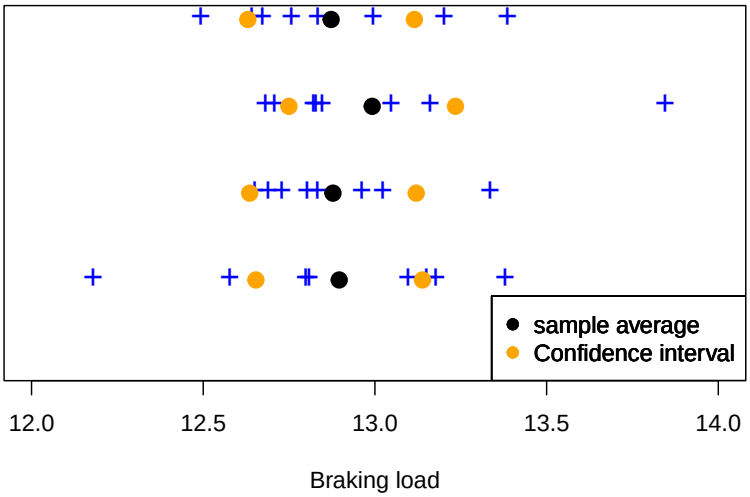
We can also write the interval as

$$\hat{\mu} = \bar{x} \pm m = 13.21 \pm 0.24$$

This means that, when estimating the mean by the average, we are confident up to the number 3 in the units and not so much about the decimal places.



Remember that the confidence interval (l, u) is an observation of the random confidence interval (L, U). Therefore, every time that we obtain a new sample then (l, u) changes. If we perform 100 samples of size n then about 95 of the confidence intervals will contain μ, we just don't know which!



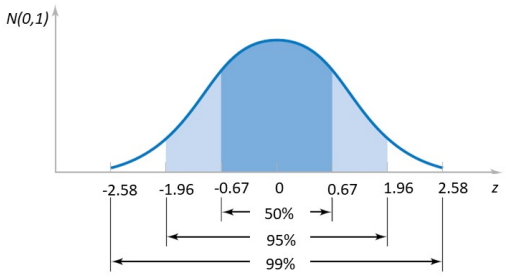
13.4.2 Confidence level

We can change our confidence from 95% to 99%. When we computed the margin of error at 95%, we left out $\alpha = 0.05$ probability, 0.025 on each side.

Now, we can leave out $\alpha = 0.01$ probability, 0.005 on each side. Therefore the 99% confidence interval is

$$(l, u) = (\bar{x} - z_{0.005}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{0.005}\frac{\sigma}{\sqrt{n}})$$

$$= (\bar{x} - 2.58\frac{\sigma}{\sqrt{n}}, \bar{x} + 2.58\frac{\sigma}{\sqrt{n}})$$



where $z_{0.005} = \phi^{-1}(0.995) = \text{norm.ppf}(0.995)$. We can also write it as

$$\hat{\mu} = \bar{x} \pm 2.58\frac{\sigma}{\sqrt{n}}$$

For our cables, the 99% confidence interval is

$$\hat{\mu} = 13.21 \pm 0.31$$

If we want to be more confident then we need larger confidence intervals. The more confidence we demand the less precision we obtain.

Example (Impact energy)

A metallic material is tested for impact to measure the energy required to cut it at a given temperature. Ten specimens of A238 steel were cut at 60°C at the following impact energies (J):

64.1, 64.7, 64.5, 64.6, 64.5, 64.3, 64.6, 64.8, 64.2, 64.3

If we **asume** that the impact energy is normally distributed with $\sigma = 1J$ what is the 95% confidence interval for the mean of these data?

We know

1. $X \to N(\mu, \sigma^2)$
2. $\sigma = 1J$
3. $\alpha = 0.05$ (the confidence limit)

The 95% confidence interval is then

$(l, u) = (\bar{x} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96\frac{\sigma}{\sqrt{n}})$

$$= (64.46 - 1.96\frac{1}{\sqrt{10}}, 64.46 + 1.96\frac{1}{\sqrt{10}}) = (63.84, 65.08)$$

or

$$\hat{\mu} = 64.46 \pm 0.61$$

this tells us that we can be sure on the first digit (6), somewhat confident on the second (4), and unsure on the decimals (46).

What if σ^2 is **unknown**?

13.5 Marging of error for unkown variance

We were able to compute the confidence interval $(l, u) = (\bar{x} - m, \bar{x} - m)$ because we were able to find the margin of error

$$m = 1.96\frac{\sigma}{\sqrt{n}}$$

since **we knew** σ. σ is a parameter of the distribution that we usually **do not know**, likewise μ. To find the margin of error with **unknown** variance, we need the following theorem by Gosset.

13.5.1 Theorem (T-statistic)

When X is normal then the standardized statistic

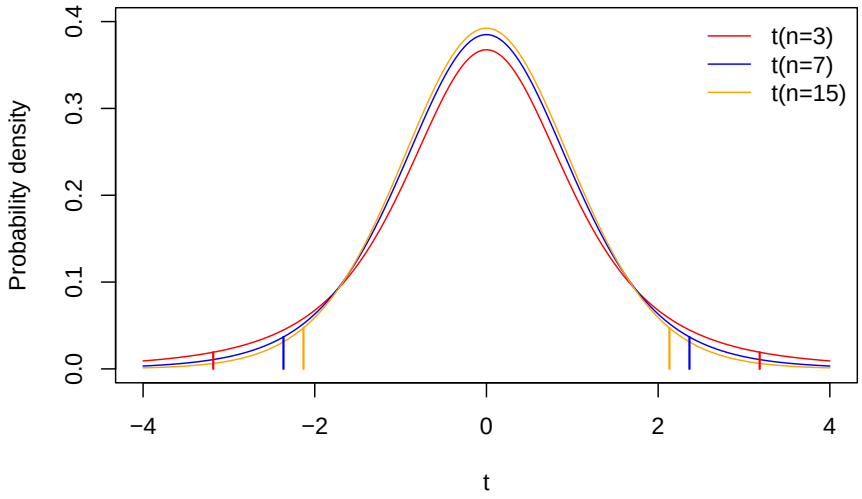
$$T = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}}$$

follows a t-distribution with $n - 1$ degrees of freedom

$$T \to t_{n-1}$$

where $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ is the sample variance. We can, therefore, compute probabilities for $\bar{X}$, even if we do not know σ.

Let's look at some probability densities in the family of the t distributions.



Now we need to recompute the the margin of error m at 5% level when we use the t-distribution

$P(\mu - m \leq \bar{X} \leq \mu + m)$

$$= P(-\frac{m}{s/\sqrt{n}} \leq T \leq \frac{m}{s/\sqrt{n}}) = 0.95$$

The margin of error then is computed from the values of tT that contain 95% of the t-distribution $(-t_{0.025,n-1}, t_{0.025,n-1})$. The vertical lines in the plot above, for different values of n. Then

$$m = t_{0.025,n-1}\frac{s}{\sqrt{n}}$$

$t_{0.025,n-1}$ is the value of T that leaves 2.5% of probability at the right hand side of the t-distribution with $n-1$ degrees of freedom. In short, if we do not know σ, we can replace it with s but need to use a different distribution in the computation of the margin of error.

13.5.2 Case 2 (unknown variance)

The second case to compute confidence intervals is more realistic. If

1. X is a normal variable

the confidence interval at 95% is

$$(l, u) = (\bar{x} - m, \bar{x} + m)$$

where

$$m = t_{0.025,n-1}\frac{s}{\sqrt{n}}$$

That is:

$$(l, u) = (\bar{x} - t_{0.025,n-1}\frac{s}{\sqrt{n}}, \bar{x} + t_{0.025,n-1}\frac{s}{\sqrt{n}})$$

where $t_{0.025,n-1} = F^{-1}(0.975)$=t.ppf(0.975, n-1)

Example (Impact energy)

A metallic material is tested for impact to measure the energy required to cut it at a given temperature. Ten specimens of A238 steel were cut at 60ºC at the following impact energies (J):

64.1, 64.7, 64.5, 64.6, 64.5, 64.3, 64.6, 64.8, 64.2, 64.3

If we **asumme** that the impact energy is normally distributed but we **do not know** the variance, what is the 95% confidence interval for the mean of these data?

We compute from the data

- $\bar{x} = 64.46$
- $s = 0.227$

we assume

- $\alpha = 0.05$ (the confidence limit)
- $t_{0.025,9} = 2.26$ obtained from $t_{0.025,9}$ = t.ppf(0.975, 9)

The confidence interval is then

$(l, u) = (\bar{x} - t_{0.025,9}\frac{s}{\sqrt{n}}, \bar{x} + t_{0.025,9}\frac{s}{\sqrt{n}})$

$$= (64.46 - 2.26\frac{0.227}{\sqrt{10}}, 64.46 + 2.26\frac{0.227}{\sqrt{10}})$$

$$= (64.29, 64.62)$$

Note that when we assumed $\sigma = 1$ the confidence interval $(63.84, 65.08)$ was larger. Data therefore suggests that $\sigma < 1$ as $s = 0.227$.

In Python, we can compute the confidence interval with:

```
from scipy import stats
x = [64.1,64.7,64.5,64.6,64.5,64.3,64.6,64.8,64.2,64.3]
res = stats.ttest_1samp(x, popmean=0)
res.confidence_interval(confidence_level=0.95)
```

13.6 Estimation of proportions

Example (vaccine)

A random sample of 400 patients was selected for testing a new vaccine for the influenza virus, after 6 months of vaccination 134 were ill. What is the expected efficacy of the vaccine?

Since each vaccination X_i is a Bernoulli trial

$$X \to Bernoulli(p)$$

with mean $\mu = p$ and variance $\sigma^2 = p(1-p)$.

The sample is something like

$$(x_1, x_2, x_3, ...x_n) = (0, 1, 0, ..1, 0)_{400}$$

with 136 ones and in a total of 400 repetitions. The sample has an average

$$\bar{x} = \frac{1}{400}\sum_{i}^{400} x_i = 134/400 = 0.34$$

Since the sample mean is an unbiased estimator of μ, then we can have a point estimate for p

$$\hat{p} = \bar{x} = 134/400 = 0.34$$

This makes sense because $\bar{x}$ is the observed relative frequency of ones f_1 in the sample. And as such it is an estimator of the probability of observing a one in a Bernoulli trial

$$f_1 = \hat{P}(X = 1)$$

This is is consistent with frequentist definition of probability that we saw in chapter 2. However, how confident are we about this estimation? That is, how confident are we to take the relative frequency as the value of the probability? We want a confidence interval for p.

13.6.1 Case 3 (proportions)

When $\hat{p}n > 5$ and $(\hat{p}-1)n > 5$, the **standardized statistic** of $\bar{X}$ can be approximated to a standard normal variable by the CLT

$$Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} = \frac{\bar{X}-p}{\left[\frac{p(1-p)}{n}\right]^{1/2}} \to N(0,1)$$

and the 95% CI interval of p is:

$$CI = (l,u) = (\bar{x} - z_{0.025}\left[\frac{\bar{x}(1-\bar{x})}{n}\right]^{1/2}, \bar{x} + z_{0.025}\left[\frac{\bar{x}(1-\bar{x})}{n}\right]^{1/2})$$

Where we estimate the Bernoulli variance $\sigma = p(1-p)$ by $\hat{\sigma} = \bar{x}(1-\bar{x})$. That is $\hat{\sigma} = \sqrt{\bar{x}(1-\bar{x})}$.

Example (vaccine)

In our case, we are counting failures on vaccinations 134 in 400 trials.

We compute from the data

- $\bar{x} = 134/400 = 0.34$

we assume

- $z_{0.025} = 1.96$ (the confidence limit)

Therefore the 95% confidence interval for p is

$(l,u) = (\bar{x} - 1.96\left[\frac{\bar{x}(1-\bar{x})}{n}\right]^{1/2}, \bar{x} + 1.96\left[\frac{\bar{x}(1-\bar{x})}{n}\right]^{1/2})$

$$= (0.29, 0.38)$$

The estimated probability of failure of the vaccine is

$$\hat{p} = 0.34 \pm 0.05$$

Note: Voting intention surveys (Bernoulli test) in a sample of n individuals report this type of estimation with its **margin of error**. It does not mean that the **true value** of p falls within this interval with probability 95%. It means that we are 95% confident that we have caught the p that this particular sample represents.

In Python

```
from statsmodels.stats.proportion import proportion_confint
proportion_confint(134, 400)
```

13.7 Estimation of the variance

We have seen that whenever we take a random sample $(X_1, X_2, ...X_n)$, the sample variance

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

is an estimator of the mean σ^2 of the random variable X. The estimator is **unbiased** because its values are centered about the parameter it is estimating

$$E(S^2) = \sigma^2$$

and it is also **consistent**. We can then take a the value of s^2 from a particular sample as the value of σ^2, which is a characteristic of the whole population. That is

$$s^2 = \hat{\sigma}^2$$

Since S^2 is a random variable the estimation of the variance changes when we take another sample.

Example (impact energy)

A metallic material is tested for impact to measure the energy required to cut it at a given temperature. Ten specimens of A238 steel were cut at 60ºC at the following impact energies (J):

64.1, 64.7, 64.5, 64.6, 64.5, 64.3, 64.6, 64.8, 64.2, 64.3

what is estimation of the variance of these data?

$$s^2 = 0.05155556$$

In Python:

```
import numpy as np
x = [64.1, 64.7, 64.5, 64.6, 64.5, 64.3, 64.6, 64.8, 64.2, 64.3]
np.var(x)
```

How confident are we on the decimals of the estimation?

13.8 Confidence interval for the variance

To compute a confidence interval of the variance, we need a statistics that is a function of S^2 and allows us to compute probabilities. We will use the following theorem

13.8.1 Theorem (χ^2):

When X is normal then the standardized statistic

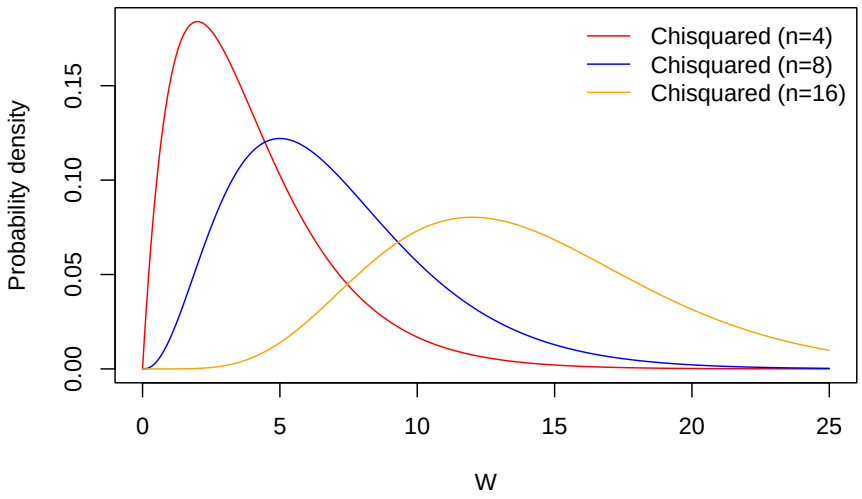
$$W = \frac{S^2(n-1)}{\sigma^2}$$

follows a χ^2 distribution with $n-1$ degrees of freedom

$$W \to \chi^2_{n-1}$$

We can therefore compute probabilities for W.

Let's look at some probability densities in the family of the χ^2 distriutions.



13.8.2 Confidence interval for the variance

We look for confidence interval of σ^2 at confidence 95% (L, U) such that

$$P(L \leq \sigma^2 \leq U) = 0.95$$

We start by determining the values that capture the 95% of the χ^2-distribution

$$P(\chi^2_{0.975,n-1} \leq W \leq \chi^2_{0.025,n-1}) = 0.95$$

Replacing the value of W

$$P(\chi^2_{0.975,n-1} \leq \frac{S^2}{\sigma^2}(n-1) \leq \chi^2_{0.025,n-1}) = 0.95$$

and solving for σ^2

$$P(\frac{S^2(n-1)}{\chi^2_{0.025,n-1}} \leq \sigma^2 \leq \frac{S^2(n-1)}{\chi^2_{0.975,n-1}}) = 0.95$$

We find a random interval that captures σ^2 with 95% confidence

$$(L, U) = (\frac{S^2(n-1)}{\chi^2_{0.025,n-1}}, \frac{S^2(n-1)}{\chi^2_{0.975,n-1}})$$

13.8.3 Case 4 (variance)

1. When X is a normal variable

The **observed** 95% confidence interval (script size) is

$$(l, u) = (\frac{s^2(n-1)}{\chi^2_{0.025,n-1}}, \frac{s^2(n-1)}{\chi^2_{0.975,n-1}})$$

where

- $\chi^2_{0.975,n-1} = F^{-1}(0.025)$ = chi2.ppf(0.025, df=n-1)
- $\chi^2_{0.025,n-1} = F^{-1}(0.975)$ =chi2.ppf(0.975, df=n-1) for $n = 10$ or $df = n - 1 = 9$

Example (impact energy)

A metallic material is tested for impact to measure the energy required to cut it at a given temperature. Ten specimens of A238 steel were cut at 60ºC at the following impact energies (J):

64.1, 64.7, 64.5, 64.6, 64.5, 64.3, 64.6, 64.8, 64.2, 64.3

what is confidence interval for the variance of these data?

$$(l, u) = (\frac{s^2(n-1)}{\chi^2_{0.025,n-1}}, \frac{s^2(n-1)}{\chi^2_{0.975,n-1}})$$

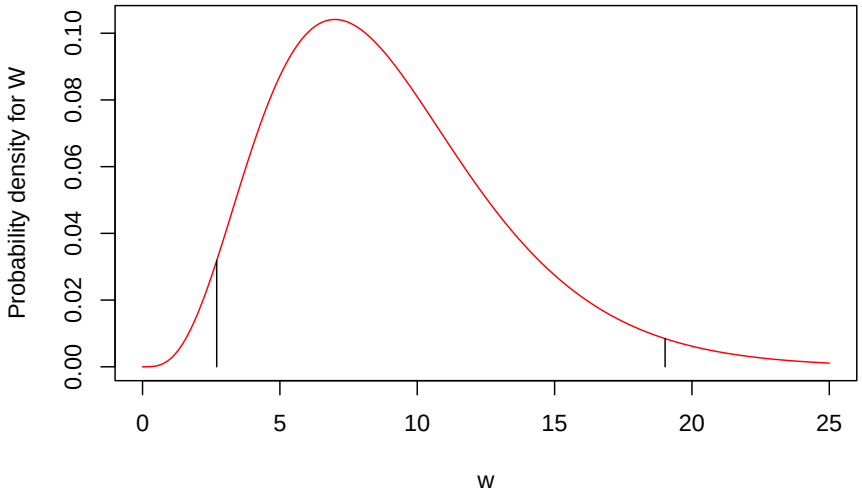
1. The standard deviation of the data is $s^2 = 0.05155556$
2. $n = 10$
3. We then compute $\chi^2_{0.025,n-1}$ and $\chi^2_{0.975,n-1}$

```
from scipy.stats import chi2
chi2.ppf(0.025, df=9)
```

```
2.700389
```

```
from scipy.stats import chi2
chi2.ppf(0.975, df=9)
```

```
19.02277
```

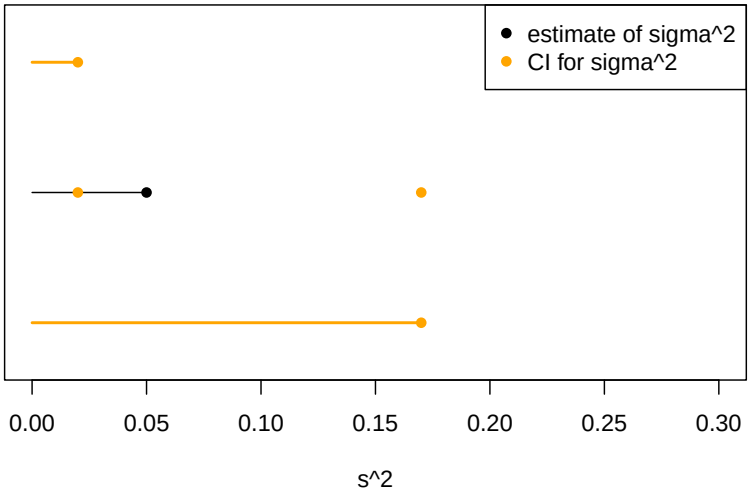


Therefore

$$(l, u) = (\frac{0.227^2(10-1)}{19.02277}, \frac{0.227^2(10-1)}{2.700389}) = (0.02, 0.17)$$

Note that when we computed the confidence interval for the mean and assumed $\sigma^2 = 1$ (case 1), this was not consistent with the data because the confidence interval does not contain the value $\sigma^2 = 1$. According to the data $\sigma^2 \neq 1$ at 95% confidence.

The interval for the variance is **not symmetric** and we cannot formulate it as an estimate $\pm$ margin of error.



13.9 Questions

1) The margin of error at 95% confidence of a normal variable is

a: $\frac{s}{\sqrt{n}}$; **b:** $1.96 \times se$; **c:** $\frac{\sigma}{\sqrt{n}}$; **d:** σ

2) when we talk about $z_{0.025}$ we mean:

a: The value of a normal standard variable that has accumulated up to 99.75% of probability; **b:** The value of a normal standard variable that has accumulates up to 0.25% of probability; **c:** The probability of a standard variable up to 99.75%; **d:** The probability of a standard variable up to 0.25%

3) The random confident interval (L, U) for the mean at 95%

a: is a two dimensional parameter of the sample distribution; **b:** gives the limits where μ has a probability of occurring 95% of the times; **c:** is an estimate of the average; **d:** captures μ 95% of the times

4) A confidence interval for the mean written as $\hat{\mu} = 56.99 \pm 0.01$

a: indicates that we are %99 confident that the mean is 56.99; **b:** indicates that we cannot trust the last decimal place on the estimation of the mean; **c:** indicates that the mean of the population is at 56.99 with error 0.01; **d:** indicates that we can trust the unit figure (6) on the estimation of the mean

5)If we know the value of μ and find that the confidence interval did not catch it then

a: the confidence interval is not well computed; **b:** it is a rare observation of the confidence interval; **c:** the confidence interval does not estimate the mean; **f:** there is little probability of finding the mean in the confidence interval

13.10 Exercises

13.10.0.1 Exercise 1

In a scientific paper, the authors report a 95% confidence interval of $(228, 232)$ for the natural frequency (Hz) of a metallic beam. They used a sample of size 25 and considered that the measurements were distributed normally.

- What is the mean and the standard deviation of the measurements?
- Compute the 99% confidence interval.

hints:

- in R $t_{0.025,24}$ = t.ppf(0.975, 24)$\sim$ 2
- in R $t_{0.005,24}$ =t.ppf(0.995, 24)$\sim$ 2.8

13.10.0.2 Exercise 2

compute 95% CI the mean of a normal variable with known variance $\sigma^2 = 9$ and $\bar{x} = 22$, using a sample of size 36.

13.10.0.3 Exercise 3

This year, 17 of 1000 of patients with influenza developed complications.

- Compute the 99% confidence interval for the proportion of complications.
- The previous year 2% showed complications. Can we say with 99% confidence that this year there is a significant drop in influenza complications?

13.10.0.4 Exercise 4

What is the confidence interval for population variance of a normal variable if we take a random sample of size $n = 10$ and observe a sample variance of 0.5?

13.11 Practice

Load misophonia data https://alejandro-isglobal.github.io/SDA/data/data_0.txt

- Compute the confidence interval for the mean of the cephalometric measures. ("Angulo_convexidad", "protusion.mandibular", "Angulo_cuelloYtercio", "Subnasal_H")
- Compute the confidence interval for the proportion of misophonic ("Misofonia"), and depression ("depresion.dic").
- Compute the confidence interval for the variance of the age ("Edad"). What is the confidence interval for the standard deviation of the population?

Solutions

Chapter 14

Hypothesis testing

14.1 Objective

In this chapter we will study **hypothesis testing** of means and proportions. We will define the null and the alternative hypotheses and how to use data to choose between both.

We will also introduce hypothesis testing of variances. Finally we will describe the errors that are made when hypotheses are tested. These errors are known as false positives and false negatives.

14.2 Hypothesis

When we perform an experiment, we often want to test whether the changes we make to the experiment have a real effect. We want, for example, to know if we are able to influence the experiment. Or, if we submit the experiment under a new condition, we want to know if that condition affects the results of the experiment. We usually have **an idea** of what the data should look like when the new conditions are **not present**. Since the results of the experiment are random in any circumstance, how can we tell the difference in the experiment when changing conditions?

The strategy is to formulate a probability model for the experiment and estimate the change in the model **parameters** between conditions. We then use observations from taking random samples of the experiment under the different conditions to assess the change in the parameters and thus provide evidence about the expected change in the experiment.

Examples (Tyres)

The mean life of a standard tyre is $20,000$ km. A tyre manufacturer wants

to know if their improved tyres run longer than the standard type. Let's try to translate their interest into statistical terms. Imagine a random experiment that consists of measuring how long the life of a particular tyre is. Therefore, the manufacturer is interested in knowing if the mean life of a new tyre is more than $20,000$ km.

Let us formulate two dichotomous statements, that is, two mutually exclusive situations:

either

a. The mean life of the new tyre may be **less** than $20,000$ km

or

b. The mean life of the new tyre may be **greater** than $20,000$ km

Only one can be true. The question is then how we can use experiments to decide between situation a. or situation b.

Note than when running the experiment several times, some tyres will run for more than $20,000$ km and some for less than $20,000$ km. Our question is then translated as whether **the mean**, as a parameter, is higher than $20,000$ km. The question is not about a single observation. The statements a. and b. are general statements.

Let's consider that μ is the mean of the random variable that measures the life of a new tyre. Therefore, the statements a. and b. can also be written as

a. $H_0 : \mu \leq 20000$ km
b. $H_1 : \mu > 20000$ km

Where the statement H_0 states that the mean life of the new tyre is not the desired $20,000$ km while statement H_1 is the desired case. H_0 and H_1 are called hypotheses.

Definition

In statistics, a statement (conjecture) about the probability function of a random variable is called a **hypothesis**.

The hypothesis is usually written in two dichotomous statements

a. The **null** hypothesis H_0: when the conjecture is false. It usually refers to the **status quo**. The data may be explained by the satus quo.

b. The **alternative** hypothesis H_1: when the conjecture is true. It usually refers to **research hypothesis**. The data may be explained by the alternative to the status quo.

Example (Fertilizer)

What are the null and the alternative hypothesis for the following situation?

Fertilizer developers want to test whether their new product has a real effect on the growth of plants.

Being μ_0 the mean growth of the plants **without** fertilizer (known) and μ the mean growth of the plants with the fertilizer (unknown)

a. $H_0 : \mu \leq \mu_0$ (The fertilizer may do nothing: status quo)
b. $H_1 : \mu > \mu_0$ (The fertilizer may have the desired effect: research interest)

Example (chemotherapy)

Pharmaceutical companies need to know if a novel chemotherapy can cure 90% of cancer patients.

Being p_0 the proportion of patients that are cured **without** the chemotherapy (known) and p the proportion that are cures **with** the chemotherapy (unknown)

a. $H_0 : p \leq p_0$ (The chemotherapy may do nothing: status quo)
b. $H_1 : p > p_0$ (The chemotherapy may have the desired effect: research interest)

Note that our new improved experiment has the parameter p and we want to know how it compares to the experiment without **any improvement** that has the parameter p_0.

We want to decide between H_0 and H_1. There are two options:

1. We **reject** the alternative hypothesis H_1; that is, we accept the null hypothesis H_0.
2. We **accept** the alternative hypothesis H_1 (our interest); that is, we reject the null hypothesis H_0.

Example (Microprocessors)

We want to produce computers with a certain type of microprocessor. The design requires that the **mean width** of a microprocessor is about 26mm. We buy 8 microprocessors from a manufacturing company that claims that they produce them with **mean** with of $\mu_0 = 26$mm. We are not sure and want to decide on whether the microprocessors of the company do measure **on average** 26mm or not. We formulate the hypothesis contrast

a. $H_0 : \mu = \mu_0$ (The manufacturer claim: status quo)
b. $H_1 : \mu \neq \mu_0$ (Our doubt that the manufacturer is not right: research interest)

To decide between H_0 or H_1, we measure the width of the purchased sample of 8 microprocessors.

```
## [1] 26.69284 26.65240 26.02918 26.21622 25.93998 27.10618 27.51114 25.25494
```

The idea is simple. The average of the sample width is $\bar{x} = 26.42536$ which would make us believe that the microprocessors are too thick. But the manufacturer would argue that $\bar{x}$ is the observation of a random variable and that $\bar{x} > 26$

is not a sufficient argument that they are not right (H_0 is not true) because sometimes $\bar{X}$ would be higher than 26 and sometimes it will me lower. Being more technical, the manufacturer asks us to account for the variability of $\bar{X}$ as an estimator of the $\mu_0 = 26$.

We would then ask ourselves: Is $\bar{x} = 26.42536$ what we would typically observe when we estimate $\mu_0 = 26$ by $\bar{x}$? or are those 0.42536 decimal figures too high to believe that the microprocessors are build with of expected value of precisely 26mm?

To test these ideas, we need a decision process that we describe now.

14.3 Hypothesis testing

Let us summarize the different cases, ways and types for testing hypothesis. We will then discuss each case with a particular example.

Hypotheses can be tested, or decided, using confidence intervals. Therefore, we are going to test hypothesis in the four **cases** that we saw for confidence intervals, namely:

- **Case 1**: Hypothesis test for the mean μ, when $X \to N(\mu, \sigma^2)$ and we know σ
- **Case 2**: Hypothesis test for the mean μ, when $X \to N(\mu, \sigma^2)$ and we do not know σ
- **Case 3**: Hypothesis test for the proportion p when $X \to Bernoulli(p)$ and both np and $n(1-p) > 5$.
- **Case 4**: Hypothesis test for the variance σ^2, when $X \to N(\mu, \sigma^2)$

There are three **ways** to test the hypotheses:

1. Using **confidence intervals**
2. using a **rejection zone**
3. using a $p - value$.

All three options are equivalent. Finally, there are three **types** of hypothesis that we can test:

1. **Two** tailed
2. **Upper** tailed
3. **Lower** tailed

14.4 Case 1 (known variance)

A **tow tailed** hypothesis contrast is of the form

a. $H_0 : \mu = \mu_0$ (status quo)

b. $H_1 : \mu \neq \mu_0$ (research interest)

This is called two tailed because the alternative hypothesis H_1 requires that the mean μ is either lower or higher than μ_0. This hypothesis can be tested in different cases. The **case 1** is when

1. X is a normal variable, and
2. we **know** the value of σ^2

14.4.1 Hypothesis test with a confidence interval

For **case 1** the confidence interval at 95% is

$$(l, u) = (\bar{x} - z_{0.025}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{0.025}\frac{\sigma}{\sqrt{n}})$$

Testing Criteria:

- If the confidence interval **contains** the null hypothesis

$$\mu_0 \in (l, u)$$

then we **accept** H_0 with 95% confidence.

- If the confidence interval does **not contain** the null hypothesis

$$\mu_0 \notin (l, u)$$

then we **reject** H_0 with 95% confidence.

Example (Microprocessors)

We want to know whether the microprocessors on average 26mm or not. Therefore, we test the following hypotheses

a. $H_0 : \mu = 26$ (The microprocessors **may** have the width that the manufacturer claim: status quo)
b. $H_1 : \mu \neq 26$ (The microprocessors **may not** have the width that the manufacturer claim: research interest)

Since we do no know which one is true, let's start by estimating the mean of our sample (μ). We then use $\bar{x} = \hat{\mu}$. We do this as before, and consider **case 1** where the model of the widths are normal

1. $X \to N(\mu = 26, \sigma^2 = 0.7^2)$
2. and, somehow we know $\sigma^2 = 0.7^2$ (perhaps given by the manufacturer)

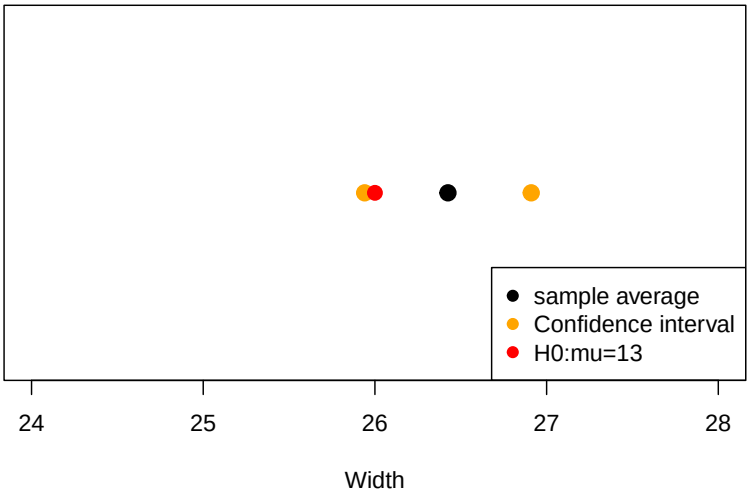
The confidence interval for the mean of our sample μ is

$$(l, u) = (\bar{x} - z_{0.025}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{0.025}\frac{\sigma}{\sqrt{n}}) = (25.94, 26.91)$$

The confidence interval tells us that we trust with 95% confidence that the true width μ is in the interval. We don't know the true value of μ but we see that $\mu = 26$mm could be it. Since the interval caught μ_0 (the manufacturer's claim)

$$\mu_0 \in (25.94, 26.91)$$

our conclusion is to accept that H_0 could have produced our **observed interval**. We also say that the data supports the manufacturer's claim. More technically, we say that we **do not reject** H_0.



14.4.2 Hypothesis test with acceptance/rejection zones

An equivalent way to test the hypothesis is to see if our set of observations are either common or rare **if** we assume that the **null hypothesis is true**. Let's remember the hypothesis contrast

a. $H_0 : \mu = \mu_0$ (status quo)
b. $H_1 : \mu \neq \mu_0$ (research interest)

To test the hypothesis with a **rejection zone** we compute the standardized statistic

$$Z = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

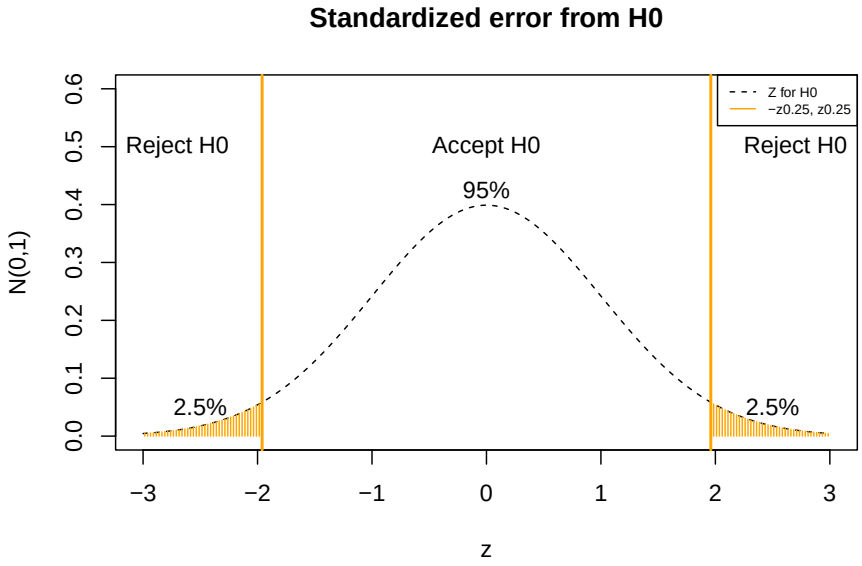
when the null hypothesis is true. Note that we are standardizing with μ_0 (the null hypothesis). We then see if the observed value of Z is within the interval

$$(-z_{0.025}, z_{0.025})$$

Remember that this interval defines the most common values of Z since

$$P(-z_{0.025} \leq Z \leq z_{0.025}) = 0.95$$

The interval $(-z_{0.025}, z_{0.025})$ is called **acceptance interval** of H_0 at 95% confidence level.



Testing criteria:

- If the observed statistics z_{obs} under the null hypothesis **is** in the acceptance region

$$z_{obs} = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \in (-z_{0.025}, z_{0.025})$$

then we **accept** H_0 with 95% confidence.

- If the observed statistics z_{obs} under the null hypothesis **is not** in the acceptance region

$$z_{obs} \notin (-z_{0.025}, z_{0.025})$$

then we **reject** H_0 with 95% confidence.

The region $(-z_{0.025}] \cup [z_{0.025})$ is called the **rejection zone**.

Example (Microprocessors)

We want to test whether the manufacturer of microprocessors produce them with mean $\mu_0 = 26$. Let's take the point of view of the manufacturer and ask whether the observation $\bar{x} = 26.42536$ is an expected average of a sample of 8 microprocessors when the expected value of the with of a microprocessor is $\mu = 26$. If H_0 is true then $\bar{X}$ is an estimator of μ_0 and therefore Z

$$Z = \frac{\bar{X} - 26}{\frac{0.7}{\sqrt{8}}} \to N(0, 1)$$

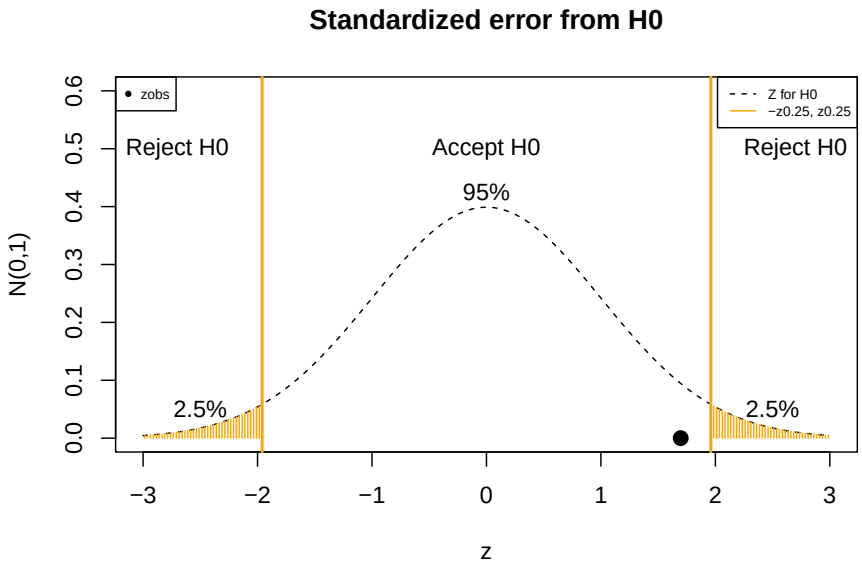
is **the standardized error** that we make when we estimate μ_0 with $\bar{X}$, as the manufacturer claims. Because we are in **case 1**, Z is standard normal Since

$$\bar{X} \to N(26, \frac{0.7^2}{8})$$

For **our data** the standardized **observed error** is in the acceptance region

$$z_{obs} = \frac{26.42536 - 26}{\frac{0.7}{\sqrt{8}}} = 1.7187 \in (-z_{0.025}, z_{0.025})$$

Think of this regions as a region of tolerance for the error. Because our observation is within then everything is ok for the manufacturer. Had it not, we would have had have enough evidence to distrust the manufacturer's claim. We conclude that our observed average is a typical observation of $\bar{x}$ when the null hypothesis μ_0 is true. Therefore, we again accept that the data is consistent with the manufacturer's claim. We also say that we **do not reject** H_0.



14.4.3 Hypothesis test with a P-value

We can also contrast the **two tail** hypothesis by calculating the probability that the average of another sample from the null hypothesis will be even rarer than the average we just observed. Because we are in **case 1**, We know that the standardized statistics Z is standard normal variable then we define the *pvalue* as

$$pvalue = P(Z \leq -z_{obs}) + P(z_{obs} \geq Z) = 2(1 - \phi(|z_{obs}|))$$

That is the probability that when we take another sample of the status quo of the same size, we are able to obtain an even rarer observation. If our observation is rare for the status quo then this value will be small.

Testing criteria:

- If the observed *pvalue* is

$$pvalue \geq \alpha = 1 - 0.95 = 0.05$$

then we **accept** the status quo H_0 with 95% confidence.

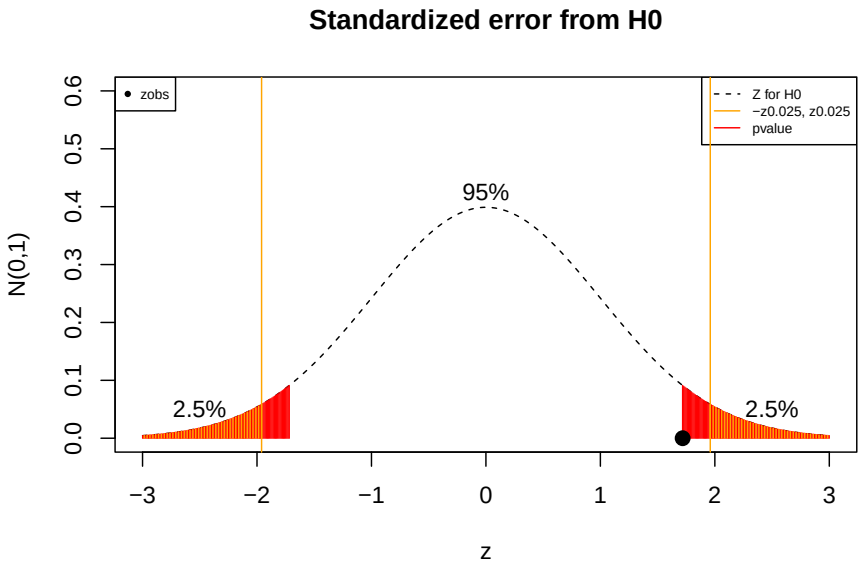
- If the observed *pvalue* is

$$pvalue < \alpha = 1 - 0.95 = 0.05$$

then we **reject** H_0 and accept our research question with 95% confidence.

α is the significance level. It gives us how much of the distribution we are leaving out, and defines the region that we consider as rare observations.

Remember: We always trust our data. If the null hypothesis says that our data is a **rare** observation we then distrust the null hypothesis, and reject it.



Example (Microprocessors)

Fie, let's think that the average $\bar{x} = 26.42536$ was produced under the manufacturer's claim. Perhaps, we were unlucky and happened to buy a sample made of thick microprocessors. If that is so then how unlucky were we? For our data the observed statistic $z_{obs} = 1.718714$ and its **p-value** is then

$$pvalue = 2(1 - \phi(1.718714)) = 0.08567$$

Python: 2*(1- norm.cdf(1.718714))

We conclude that if we purchase another set of 8 microprocessors, it is likely to obtain averages this far from the 26mm in 8.5% of the 8-microprocessor-set purchases. Being 8.5% unlucky is within the margin. Lower than 5 unlucky is the limit where we start to believe that it was not that we were unlucky but rather that the manufacturer is wrong. In this example the null hypothesis can tolerate the observed error with 95% confidence. We, therefore, accept that the status quo could have produced our data and conclude again that the data is consistent with the manufacturer's claim.

In Python the entire hypothesis testing can be performed with the function ztest from the library bioinfokit (that needs to be previously installed)

```
# https://pypi.org/project/bioinfokit/

!pip install bioinfokit
from bioinfokit.analys import stat
import pandas as pd

data = {'x': [26.69284, 26.65240, 26.02918, 26.21622,
              25.93998, 27.10618, 27.51114, 25.25494]}

df = pd.DataFrame(data)

res = stat()
res.ztest(df=df, x='x', mu=26, x_std=0.7, test_type=1)
print(res.summary)
```

```
One Sample Z-test

------------------  ----------
Sample size          8
Mean                26.42536
Z value              1.71871
p value (one-tail)   0.0428332
p value (two-tail)   0.0856665
Lower 95.0%         25.94029
Upper 95.0%         26.91043
------------------  ----------
```

14.4.4 Upper tail hypothesis

We may be interested in only testing for the fact that our experiment's mean has a higher mean than the null's mean.

Upper-tailed test:

- $H_0 : \mu \leq 26$ (**at most** microprocessors have this mean width)
- $H_1 : \mu > 26$ (microprocessors have **higher** mean width)

Perhaps, our design needs an upper limit to the with that the microprocessors can have, such that if it is not stisfy we may look for another provider.

This called **upper-tailed** because the alternative hypothesis H_1 requires that the mean μ is **higher** than μ_0. This hypothesis can be tested in different cases. The **case 1** is when

1. X is a normal variable, and
2. we know the value of σ

Testing criteria:

1. *Confidence interval:* If the **upper-tailed** confidence interval **contains** the null hypothesis

$$\mu_0 \in (l, u) = (\bar{x} - z_{0.05}\frac{\sigma}{\sqrt{n}}, \infty)$$

where $z_{0.05} = \phi^{-1}(0.95)$ =norm.ppf(0.95), then we **accept** H_0 with 95% confidence. Note that this test is from the point of view of the **data**, we are not centering the confidence interval around $\bar{x}$, instead we are leaving all the 5% of the rare cases to the left of the average. We are therefore asking if the null hypothesis is lower than the average.

2. *Rejection/acceptance region:* If the observed statistics z_{obs} under the null hypothesis **is** in the acceptance region

$$z_{obs} = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \in (-\infty, z_{0.05})$$

then we **accept** H_0 with 95% confidence. Note that this test is from the point of view of the **null hypothesis**. We are leaving all the 5% of the rare averages to the right of the null hypothesis and therefore ask if the average is higher than the null hypothesis.

3. *pvalue*: If the observed **upper-tailed**

$$pvalue = 1 - \phi(z_{obs})$$

1-norm.cdf(zobs) is greater than $\alpha = 1 - 0.95 = 0.05$

$$pvalue \geq \alpha = 0.05$$

then we **accept** H_0 with 95% confidence. Note that this test is again from the point of view of the **null hypothesis**. We are asking: if we were to take another average what is the probability that is higher than the observed one?

Example (Microprocessors)

In the example of the microprocessors, we may be interested to choose another manufacturer only in the case that the mean width is too high. Therefore the upper-tailed hypothesis is

a. $H_0 : \mu \leq 26$ (**at most** microprocessors have this mean width: **status quo**)
b. $H_1 : \mu > 26$ (microprocessors have **higher** mean width: **research interest**)

We will test the higher tail of the distribution. For the data that we discussed before, we then **reject** H_0 (i.e. the manufacturer's claim) at 95% confidence because of any of the three equivalent contrasts:

1. The **upper tailed** confidence interval does not contain the null hypothesis $\mu_0 = 13$

$$\mu_0 = 26 \notin (\bar{x} - z_{0.05}\frac{\sigma}{\sqrt{n}}, \infty) = (26.01828, \infty)$$

where $z_{0.05}$ =norm.ppf(0.95)=1.644854

2. We have that the acceptance region for H_0 is:

$$(-\infty, z_{0.05}) = (-\infty, 1.644854)$$

and that the observed standardized error is not in the region

$$z_{obs} = \frac{26.42536 - 26}{\frac{0.7}{\sqrt{8}}} = 1.7187 \notin (-\infty, 1.644854)$$

3. The upper tail *pvalue* is lower than $\alpha = 0.05$

$$pvalue = 1 - \phi(1.7187) = 0.04283451 < 0.05$$

where *pvalue* =1-norm.cdf(1.7187).

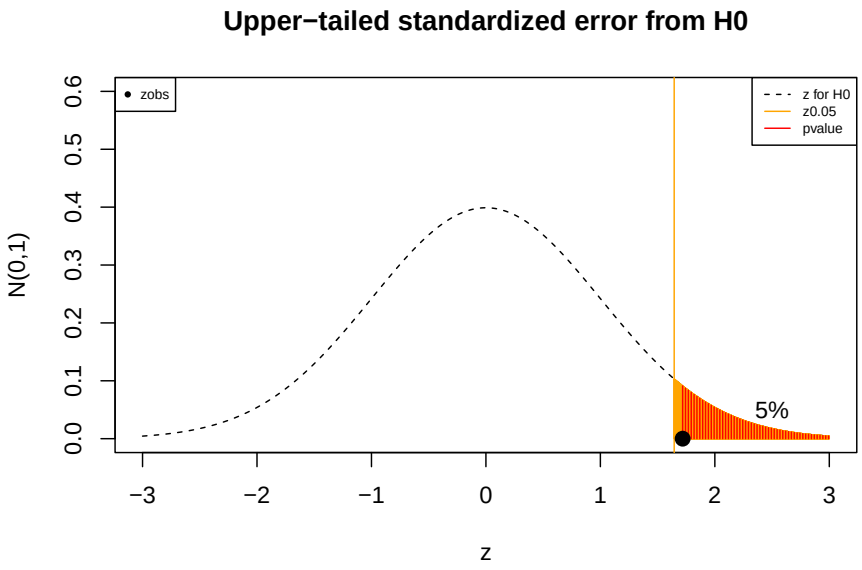
The hypothesis test is performed in Python under the label p value (one-tail)

```
# https://pypi.org/project/bioinfokit/

!pip install bioinfokit
from bioinfokit.analys import stat
import pandas as pd

data = {'x': [26.69284, 26.65240, 26.02918, 26.21622,
              25.93998, 27.10618, 27.51114, 25.25494]}

df = pd.DataFrame(data)

res = stat()
res.ztest(df=df, x='x', mu=26, x_std=0.7, test_type=1)
print(res.summary)
```

```
One Sample Z-test

-----------------  ----------
Sample size          8
Mean                26.42536
```

```
Z value               1.71871
p value (one-tail)    0.0428332
p value (two-tail)    0.0856665
Lower 95.0%          25.94029
Upper 95.0%          26.91043
------------------   ----------
```

The result can be seen in the *pvalue* for one tail. Note that the *pvalue* for one tail is half of the *pvalue* for two tails.



We then conclude that the microprocessors of the manufacturer are too thick for our specifications and have evidence that support changing from provider.

14.5 Case 2 (unknown variance)

A **two tailed** hypothesis contrast of the form

a. $H_0 : \mu = \mu_0$ (status quo)
b. $H_1 : \mu \neq \mu_0$ (research interest)

can be tested when we do not know σ^2 using **case 2**, namely when

1. X is a normal variable, $X \to N(\mu, \sigma^2)$, and
2. we do **not** know the value of σ^2

Let's remember that in this case, the **standardized error** with respect to the **sample standard deviation** S

$$T = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}}$$

Follows a t-distribution with $n-1$ degrees of freedom. Therefore, we can apply **all three criteria** as in **case 1** but making the substitution of s for σ and Z for T.

Testing criteria:

1. *Confidence interval:* If the confidence interval **contains** the null hypothesis

$$\mu_0 \in (l, u) = (\bar{x} - t_{0.025,n-1}\frac{s}{\sqrt{n}}, \bar{x} + t_{0.025,n-1}\frac{s}{\sqrt{n}})$$

then we **accept** H_0 with 95% confidence.

2. *Rejection/acceptance region:* If the observed statistics t_{obs} under the null hypothesis **is** in the acceptance region

$$t_{obs} = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} \in (-t_{0.025}, t_{0.025})$$

where $t_{0.025} = F_t^{-1}(0.975, n-1) =$ t.ppf(0.975, n-1), then we **accept** H_0 with 95% confidence.

3. *pvalue*: If the observed $pvalue = 2(1 - F_t(|t_{obs}|)) =$2*(1- t.cdf(abs(tobs), n-1)) is

$$pvalue \geq \alpha = 1 - 0.95 = 0.05$$

then we **accept** H_0 with 95% confidence.

Example (Microprocessors)

For the hypothesis contrast for the microprocessors width

a. $H_0 : \mu = 26$
b. $H_1 : \mu \neq 26$

We will only assume that the mean width of a microprocessor is normally distributed

1. $X \to N(\mu = 26, \sigma^2 = ?)$
2. We do not know σ^2

Having obtained the sample

```
## [1] 26.69284 26.65240 26.02918 26.21622 25.93998 27.10618 27.51114 25.25494
```

with this data, we accept that H_0 (the manufacturer's claim) at 95% significance because of any of following equivalent contrasts:

1. The confidence interval

$$(\bar{x} - t_{0.025,n-1}\frac{s}{\sqrt{n}}, \bar{x} + t_{0.025,n-1}\frac{s}{\sqrt{n}}) = (25.82818, 27.02254)$$

contains $H_0 : \mu = 26$.

2. The acceptance region for H_0 is:

$$(-t_{0.025,7}, t_{0.025,7}) = (-2.36, 2.36)$$

and the observed standardized error from H_0 is

$$t_{obs} = \frac{26.42536 - 26}{\frac{0.714313}{\sqrt{8}}} = 1.6843$$

within the acceptance region.

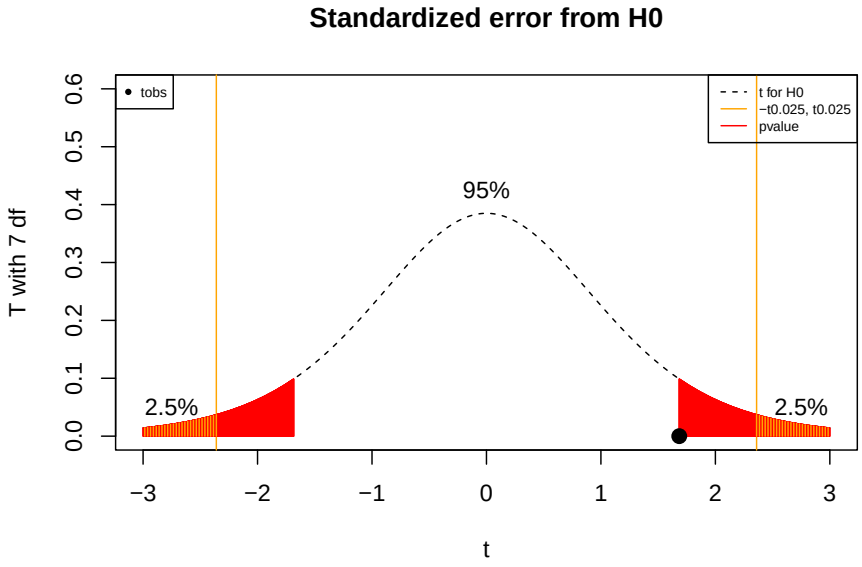
3. The

$$pvalue = 2(1 - F_{t,7}(1.6843)) = 0.136$$

is greater then $\alpha = 0.05$. The *pvalue* is computed R like 2*(1- t.cdf(1.6843,7))

In Python this contrasts are performed with the function stats.ttest_1samp:

```
from scipy import stats

data = [26.69284, 26.65240, 26.02918, 26.21622,
        25.93998, 27.10618, 27.51114, 25.25494]

res=stats.ttest_1samp(data, popmean=13)
print (res)

TtestResult(statistic=1.68427527723504, pvalue=0.1360005269730744, df=7)
```



If we use an upper tailed hypothesis, we would obtain half of the *pvalue* for the two tail test, that it $pvalue = 0.068$ and, therefore, we still would not reject the manufacturer's claim. The upper tail *pvalue* is greater than the significance limit $\alpha = 0.05$. With a t-test we do not look for another manufacturer in any case.

Note that in the case 1, we have **assumed** that $\sigma = 0.7$. In case 2, for the same data we computed $s = 0.714313$. Therefore the data suggest that the widths are more dispersed than what we have assumed, giving more benefit of the doubt to the manufacturer. This difference in dispersion lead to two different decisions for the upper-tail test, to replace the manufacturer (case 1) or not (case 2).

Example (NaCl)

$11.6g$ of NaCl is dissolved in $100g$ of water and has a molar concentration of $1.92mol/L$

We design a process to remove salt from this concentration and obtain the following results

```
## [1] 1.716 1.889 1.783 1.849 1.891
```

We first want to test at 0.95% confidence if the process changes the salt concentration in any direction. Therefore we propose a two-tail hypothesis:

a. $H_0 : \mu = 1.92$
b. $H_1 : \mu \neq 1.92$

We will assume that X is normal and that we do not know the variance σ^2. Therefore, we are in **case 2** that we test with a stats.ttest_1samp

```
from scipy import stats

data = [1.716, 1.889, 1.783, 1.849, 1.891]

res=stats.ttest_1samp(data, popmean=1.92, alternative='two-sided')
print (res)
res.confidence_interval(confidence_level=0.95)
```

```
TtestResult(statistic=-2.8038174739802226, pvalue=0.048622042176166516, df=4)
ConfidenceInterval(low=1.7321215833901604, high=1.9190784166098398)
```

Since $pvalue < 0.05$, we conclude that the molar concentration has **significantly** changed after the process.

14.5.1 Lower tail hypothesis

If we are interested only in the case that we are able to remove salt from the concentration then we rather propose a **lower tail** hypothesis:

a. $H_0 : \mu \geq 1.92$ (After the desalinization process the concentration o salt is at least the initial one: status quo)
b. $H_1 : \mu < 1.92$ (After the desalinization process the concentration is lower the initial one: research interest)

Note that the lower tail is given by the alternative H_1. We want to test that the average concentration after the process is lower than the initial concentration. The contrast criteria are the same as for the other types of hypothesis. For this type of hypothesis, we will accept the null hypothesis if

1. μ_0 is in the confidence interval:

$$\mu_0 \in (l, u) = (-\infty, \bar{x} + t_{0.05,n-1}\frac{s}{\sqrt{n}})$$

2. or, t_{obs} is in the aceptance region:

$$t_{obs} \in (t_{0.05,n-1}, \infty)$$

3. or, the *pvalue* on the lower tail of the distribution.

$$pvalue = F_t(t_{obs}, n-1)$$

is higher than $\alpha = 0.05$

In any other case, we reject H_0 and accept the alternative hypothesis.

Example (NaCl)

For the lower tail contrast

a. $H_0 : \mu \geq 1.92$
b. $H_1 : \mu < 1.92$

We can assume that the concentration is normal and that we do not know σ^2. Therefore, we are in **case 2** for which we only need to change the argument alternative to less in the function t.test.

```
from scipy import stats

data = [1.716, 1.889, 1.783, 1.849, 1.891]

res=stats.ttest_1samp(data, popmean=1.92, alternative='less')
print (res)
res.confidence_interval(confidence_level=0.95)
```

```
TtestResult(statistic=-2.8038174739802226, pvalue=0.024311021088083258, df=4)
ConfidenceInterval(low=-inf, high=1.8973758334933486)
```

We see that the *pvalue* is reduced in half, and therefore we have more confidence in rejecting the lower tail hypothesis than the two-sided hypothesis.

Example 2 (soporific)

In some cases, we are not sure about the numerical value of the hypothesis to test, but we know that we want to improve the value of a parameter in two different conditions.

In the original paper of Gosset, he analyzed the effect of two soporific medicines.

- 10 individuals were given **soporific 1** and wrote down the additional hours slept under treatment, with a mean 0.75

```
import numpy as np

medicine1 = np.array([0.7,-1.6,-0.2,-1.2,-0.1,3.4,3.7,0.8,0,2])
```

- The same 10 individuals were given **soporific 2** and wrote down the additional hours slept under treatment, with a mean 2.33

```
medicine2 = np.array([1.9,0.8,1.1,0.1,-0.1,4.4,5.5,1.6,4.6,3.4])
```

The scientific hypothesis was that soporific 2 was better than soporific 1. For each individual, Gosset computed the difference between the treatments. Taking X as the **difference** between treatments, this was the sample observed for X

```
x = medicine2-medicine1
x
```

```
##  [1] 1.2 2.4 1.3 1.3 0.0 1.0 1.8 0.8 4.6 1.4
```

The average hours gained by soporific 2 with respect to soporific 1 was 1.58, and $s = 1.229995$.

The scientific question can be stated as **upper-tailed** paired t-test:

a. $H_0 : \mu \leq 0$ (no treatment difference: $\mu_2 - \mu_1 = 0$)
b. $H_1 : \mu > 0$ (treatment 2 higher then treatment 1: $\mu_2 - \mu_1 > 0$)

Where μ is the mean of the **differences** between treatments, and the null hypothesis states that there is no difference.

If we suppose that X is normal and we do not know σ^2 then we are in **case 2**. The **standardized error** is:

$$T = \frac{\bar{X}}{\frac{S}{\sqrt{n}}}$$

and its observation

$$t_{obs} = \frac{\bar{x}}{\frac{s}{\sqrt{n}}}$$

which is also known as the **signal** to **noise** ratio.

we can test the hypothesis for the difference $X = medicine_1 - medicine_2$

```
from scipy import stats

res=stats.ttest_1samp(x, popmean=0, alternative='greater')
print (res)
res.confidence_interval(confidence_level=0.95)
```

```
TtestResult(statistic=4.062127683382037, pvalue=0.001416445098692135, df=9)
```

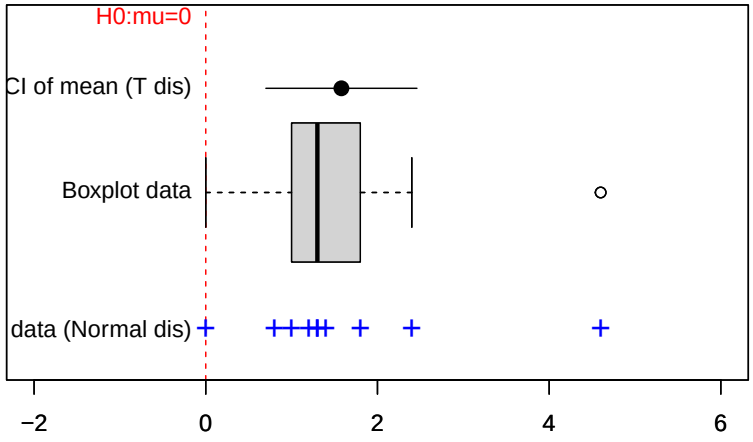
showing a significant gain by soporific 2 (rejection of the null).

Equivalently we can test the hypothesis using a paired t-test, where we introduce the observation for each separate condition, and state that the observations are paired

```
t_statistic, p_value = stats.ttest_rel(medicine2, medicine1, alternative='greater')
```

```
TtestResult(statistic=4.062127683382037, pvalue=0.001416445098692135, df=9)
```

In this plot, we show all the statistical elements for this example. In red, we show the null hypothesis of no difference between treatments. The 95% confidence interval for the difference is shown in the upper part. The CI does not contain the null. In the second row, we see the data represented in a box plot. The 5%-quantile of the data is on 0. At the bottom row, we see the raw data, that is the individual observations for the change in hours for each patient.



14.5.2 Hypothesis testing with large n and any distribution

On many occasions, X is not normally distributed but if we can take large samples $n \geq 30$ then we can use the CLT:

Then the **standardized error** from the null hypothesis can be approximated to a standard distribution

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \to N(0,1)$$

and then we proceed as in **case 1**. If σ is unknown we then replace it with its estimate s and proceed as in **case 2** using the t-statistic

$$T = \frac{\bar{X} - \mu_0}{\frac{S}{\sqrt{n}}}$$

14.6 Case 3 (proportions)

If our random experiment is a Bernoulli trial $X \to Bernoulli(p)$, we can formulate hypothesis contrasts for the probability p of an event in the trial. Consider an upper tailed hypothesis

a. $H_0 : p \leq p_0$ (status quo)
b. $H_1 : p > p_0$ (research interest)

In this **case 3**, we test a hypothesis for the proportion if

1. X is a Bernoulli trial, and
2. np, $n(1-p)$ are both greater than 5, so we can apply the central limit theorem.

Remember that if we take a the sample of n Bernoulli trials $(1, 0, 1, ...0)$,

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

is the relative frequency for the ''ones" in the sample. This is an estimator of p.

If we assume that the null hypothesis is true then $X \to Bernoulli(p_0)$ and the standardized error that we make when we estimate p_0 with $\bar{X}$ is

$$Z = \frac{\bar{X} - p_0}{\frac{\sqrt{p_0(1-p_0)}}{\sqrt{n}}} \to N(0,1)$$

$\sigma = \sqrt{p_0(1-p_0)}$ is the standard deviation of X when the null hypothesis is true: $V(X) = \sigma^2 = p_0(1-p_0)$. With this Z statistic, we can accept or reject the null hypothesis using any of the three testing criteria.

Example (process improvement)

We may be satisfied with a new process if 90% of the times we improve the previous process.

If we run a sample of 200 new processes and find that 188 times we improved the previous process, can we be satisfied with the new process at 95% confidence?

We then formulate an upper-tailed hypothesis contrast for the null hypothesis $p_0 = 0.9$. Therefore, the null and alternative hypotheses are

a. $H_0 : p \leq p_0 = 0.9$ (The new process is not satisfactory: status quo)
b. $H_1 : p > p_0 = 0.9$ (The new process is satisfactory: research interest)

We assume that if the null hypothesis is true then

1. The probability model of a random experiment is

$$X \to Bernoulli(p_0)$$

2. and check that when $np_0 = 180 > 5$ and $n(1-p_0) = 20 > 5$

Therefore, we can apply **case 3**.

We can use any of the three criteria to test the hypothesis. For this example, we believe that have a satisfactory process because, we **reject** H_0 at 95% following

1. The upper tail confidence interval for the p does not include p_0

$$p_0 = 0.9 \notin (\bar{x} - z_{0.05}\left[\tfrac{p_0(1-p_0)}{n}\right]^{1/2}, 1) = (0.905, 1)$$

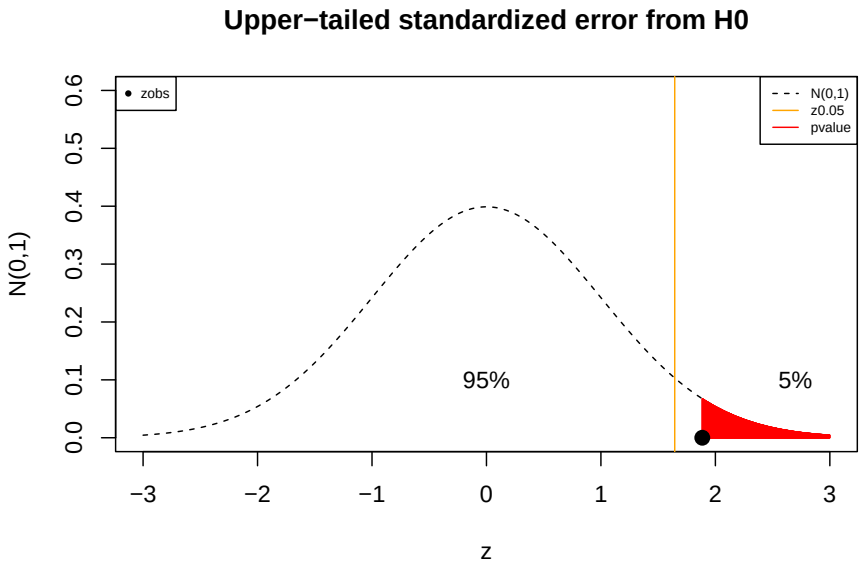
2. The observed standardized error from the null is not in the acceptance region

$$z_{obs} = \frac{\bar{X} - p_0}{\left[\tfrac{p_0(1-p_0)}{n}\right]^{1/2}} = \frac{0.94 - 0.90}{\sqrt{0.00045}} = 1.88563 \notin (-\infty, z_{0.05}) = (-\infty, 1.644)$$

3. The upper tail *pvalue* is lower than $\alpha = 0.05$:

$$pvalue = 1 - \phi(1.885618) = 0.02967323 < 0.05$$

in Python: 1-norm.cdf(1.885618)



The test can be performed in Python using the proportions_ztest function

```
from statsmodels.stats.proportion import proportions_ztest
proportions_ztest(count=0.9*200, nobs=200, value=188/200, alternative='smaller')
```

```
(-1.8856180831641234, 0.029673219395960154)
```

Note that the function is programmed such that the null hypothesis is in the count argument, which is used to compute the standard deviation of the sample. This is why the value of the statisitcs is negative and the test i lower tailed.

14.7 Case 4 (variances)

In many cases, experiments are run to test the dispersion of data.

Such as

- when complying with strict design standards where measurements must be between certain values.
- when different treatments are applied to different groups, we want to see the dispersion of outcomes between the groups.

A **tow tailed** hypothesis contrast for the variance is of the form

a. $H_0 : \sigma = \sigma_0$ (status quo)
b. $H_1 : \sigma \neq \sigma_0$ (research interest)

This hypothesis for σ (**case 4**) can be tested when

1. X is a normal variable

Remember that if we take a random sample

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

is the sample variance. This is an estimator of σ^2.

If we assume that the null hypothesis is true and $X \to N(\mu, \sigma_0)$ then the **error ratio** that we make when we estimate σ^2 with s^2 is

$$W = \frac{(n-1)S^2}{\sigma_0^2}$$

Note that when $W = 1$, we make no error. W follows a χ^2 (chi-squared) distribution with $n-1$ degrees of freedom.

$$W \to \chi^2(n-1)$$

With W, we can accept or reject the null hypothesis using any of the three testing criteria.

Example (Semiconductor)

The production of a semiconductor chip is regulated by a process that requires that the thickness of a particular layer does not vary in more than $\sigma_0 = 0.6mm$, from its mean of $25mm$.

To keep control of the process every so often a sample of 20 specimens is taken.

On one occasion a sample of the thickness of 20 semiconductors was

```
##  [1] 24.51239 24.79975 26.35608 25.06134 25.11248 26.49211 25.40100 23.89940
##  [9] 24.40244 24.61227 26.06495 25.31304 25.34867 25.09629 24.51642 26.55461
## [17] 25.43313 23.28904 25.61018 24.58867
```

The estimated standard deviation for this data is $s = 0.8462188$ was the process out of control at 99% confidence and should be stopped?

We therefore want to contrast the upper tail hypotheses

a. $H_0 : \sigma^2 \leq \sigma_0^2 = 0.6^2$ (Process is **under** control)
b. $H_1 : \sigma^2 > \sigma_0^2 = 0.6^2$ (Process is **out of** control)

Let's test the hypothesis using the **acceptance region**.

The contrast statistics is

$$W = \frac{(n-1)S^2}{\sigma_0^2} \rightarrow \chi^2(n-1)$$

and the threshold limit $\alpha = 0.01 = 0 - 0.99$. Therefore, the acceptance region $P(W \leq \chi^2_{0.01,19}) = 0.99$ is

$$(0, \chi^2_{0.01,19}) = (0, 36.19)$$

In Python: $\chi^2_{0.01,19}$ =chi2.ppf(0.99,19)= 36.19

For our data, the observed **standardized error ratio** is:

$$w_{obs} = \frac{19(0.8462188)^2}{0.60^2} = 37.79344$$

That falls outside the acceptance region

$$w_{obs} = 37.79344 \notin (0, 36.19)$$

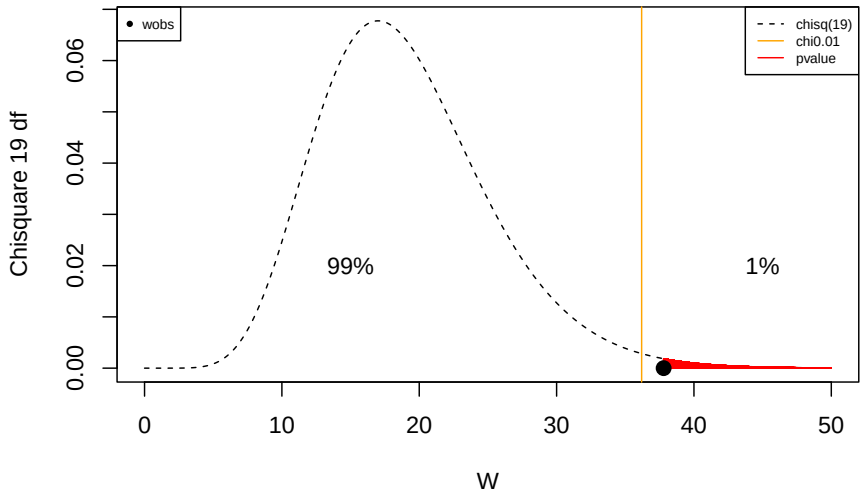
Therefore, we reject the null hypothesis and conclude that that yes! the process is out of control.

If we, alternativelly, calculate the upper tailed *pvalue*

$$pvalue = 1 - F_{\chi^2,19}(37.79344) = 0.006$$

we see that it is lower than $\alpha = 0.01$ and reject the null hypothesis.

R: 1-chi2.cdf(37.79344, 19)



For testing the hypothesis, we can use the following code in Python

```
import numpy as np
from scipy import stats

# Sample data
x = np.array([24.51239, 24.79975, 26.35608, 25.06134, 25.11248,
              26.49211, 25.40100, 23.89940, 24.40244, 24.61227,
              26.06495, 25.31304, 25.34867, 25.09629, 24.51642,
              26.55461, 25.43313, 23.28904, 25.61018, 24.58867])

# Hypothesized variance
hypothesized_variance = 0.6**2

# Degrees of freedom
df = len(x) - 1

# Calculate the test statistic
chi_square = (len(x) - 1) * np.var(x,ddof=1) / hypothesized_variance

# Calculate the critical value from the Chi-square distribution
alpha = 1 - 0.99
critical_value = stats.chi2.ppf(1 - alpha, df)

# Perform the variance test
```

```
p_value = 1 - stats.chi2.cdf(chi_square, df)

print("Variance Test:")
print("Chi-Square Statistic:", chi_square)
print("Critical Value:", critical_value)
print("P-Value:", p_value)

Variance Test:
Chi-Square Statistic: 37.79345223227778
Critical Value: 36.19086912927004
P-Value: 0.006304215036982974
```

14.8 Errors in hypothesis testing

The result of an upper tail hypothesis test may be to **reject** the null hypothesis:

$$H_0 : \mu \leq \mu_0$$

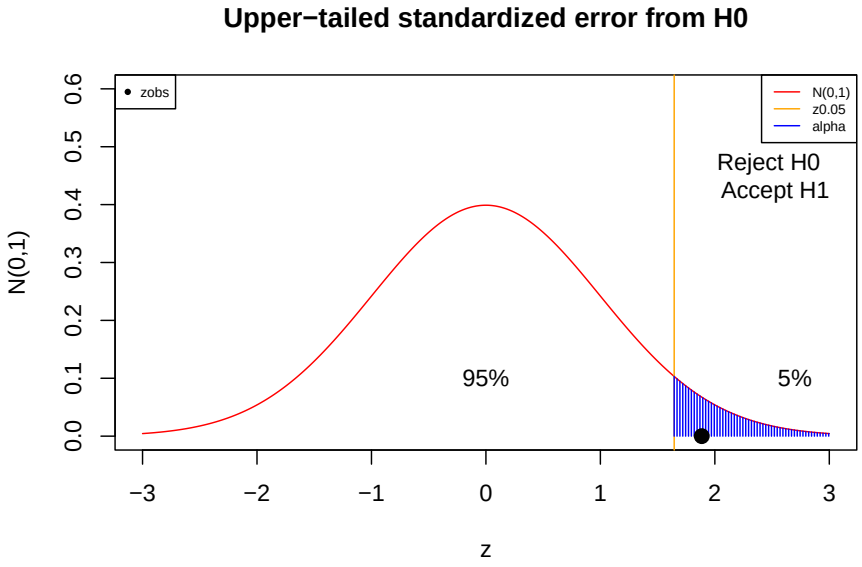
when H_0 is actually **true**. In the case of the microprocessors, imagine, for example that we rejected the manufacturer's claim (H_0) when they actually do produce microprocessors of widths with mean 26mm. Data made us believe that they produce thick microprocessors. But it was only because we bought a sample of 8 microprocessors that by chance were too thick.

We must bear in mind that the decision is made based on the data. It may well be that the observed statistic has fallen, by chance, far from the null hypothesis, in the rejection zone of H_0 even when this hypothesis is true. The statistic is a random variable and one observation can have a large value by chance.

When we perform the hypothesis test, we don't know if H_0 is true. Let us assume that we found by other means that H_0 is really true. The probability of rejecting the truth (H_0) is precisely the level of statistical significance α. We call this probability the probability of making a **type 1** error. Taking the example for **case 1**, an upper tail test, and a confidence of 95% we have that

$$\alpha = P(Z > z_{0.05}) = 0.05$$

where $z_{0.05} = \phi^{-1}(0.95) =$ norm.ppf(0.95)=1.644



A type 1 error is also called a **false positive** because our research interest is in H_1. When we reject H_0, we accept H_1 and say that our test is **positive**. Accepting H_1 translates to announcing a discovery, so the type 1 error is announcing a discovery that is not true: we falsely claimed a discovery because the data suggested it, we were fooled by it.

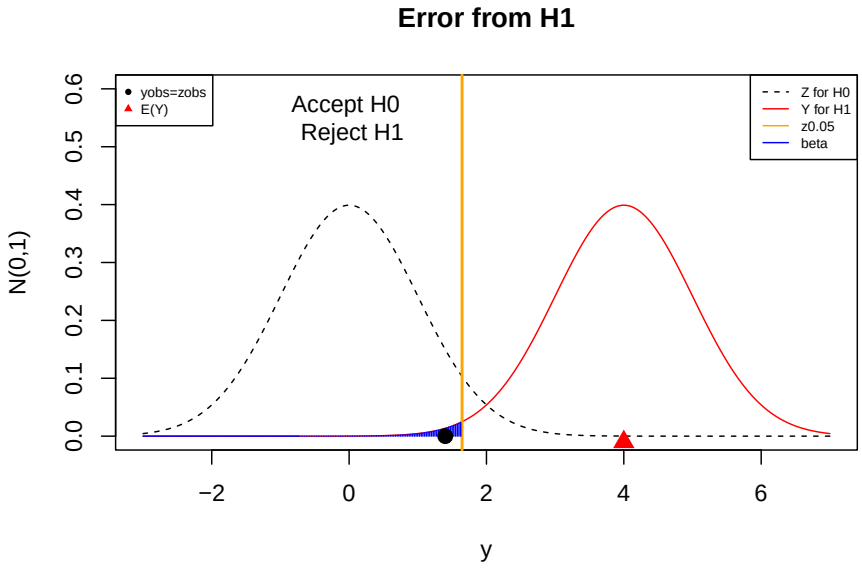
There is another type of error. The result of an upper tail hypothesis test can be **accepting** the null hypothesis:

$$H_0 : \mu \leq \mu_0$$

when this is **not true**. Again for the microprocessors, imagine now that our data made us accept the manufacturer's claim (H_0) when they actually do produce microprocessors that are too thick. Data made us believe that they produce acceptable microprocessors. But it was only because we bought a sample of 8 microprocessors that by chance were too thin.

In this case, it may be that the observed statistic has fallen, due to randomness, close to the null hypothesis, in the zone of acceptance of H_0, when really H_1 is true. If we found out somehow that, for example, μ really does have a value of μ_1 then the alternative hypothesis would be exactly:

$$H_1 : \mu = \mu_1$$



If H_1 is indeed true (red line, which we don't know when we perform the hypothesis test) then the statistic is **really** a random variable Y that has mean (case 1)

$$E(Y) = \frac{\mu_1 - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

and most of them will fall close to this value, therefore, in the rejection area of H_0, validating the hypothesis test. However, there are cases in which the observed statistic falls within the acceptance zone of H_0 due to randomness, despite the fact that the statistics are produced by H_1. In these cases we accept H_0 when it is not true. This error is called **a type 2 error** or a **false negative**. Since our research interest is in H_1 we failed to accept it. rejecting H_1 translates to discarding a discovery, so the type 2 error is ignoring a discovery that is actually true.

For case 1, with an upper tailed test and a confidence level of 95%, this is

$$\beta = P(Y < z_{0.05})$$

Where $Y \to N(\frac{\mu_1-\mu_0}{\sigma/\sqrt{n}}, 1)$ is the **true distribution** of the observed statistics.

Example (Light bulb)

The energy efficiency of a new light bulb is a normal random variable with a standard deviation of 5 watts. We consider that the light bulbs we produce are

efficient if their average does not exceed 80 watts, so we propose the hypothesis test

a. $H_0 : \mu \geq 80$ (not efficient)
b. $H_1 : \mu < 80$ (efficient)

We want to prove that we produce efficient light bulbs and we plan to draw a random sample of size 100, with a statistical significance level of 5%. If previous studies have suggested that light bulbs may average $\mu = \mu_1 = 79$ watts, what type 2 error do we expect?

The contrast is for **case 1** with a **lower tail**. Therefore the probability of accepting the null hypothesis (we do not produce efficient light bulbs) is

$$\alpha = P(Z < z_{0.95}) = 0.05$$

and $z_{0.95}$ =norm.ppf(0.05)= -1.644.

The type 2 error is therefore

$$\beta = P(Y > -1.644)$$

that is, the probability of accepting that we do not produce efficient light bulbs when in fact we do. Because we are in **case 1**, as we know σ and the variable is normal, the observed statistics are actually distributed as

$$Y \to N(\frac{79-80}{5/\sqrt{100}} = -2, 1)$$

and the type 2 error is

$$\beta = 1 - F(-1.644) = 0.36$$

computed in Python as 1-norm.cdf(-1.644,-2,1)=0.36. Therefore, only $\alpha = 5\%$ of the times we would announce that we produce efficient light bulbs when they really are not, while $\beta = 36\%$ of the times we would announce that we have a production that is not useful when it really does work.

When we carry out a hypothesis test we have two possibilities for each condition

- H_1 is actually : **true** ($\mu = \mu_1$) or **false** ($\mu = \mu_0$)
- The test for H_1 is: **positive** (z_{obs} in the acceptance zone of H_1) or **negative** (z_{obs} in the acceptance zone of H_0)

Example (PCR)

We do **a** PCR to test for an infection. The hypothesis test is

a. H_0 no infection

b. H_1 there is infection

We do the PCR test and it gives us

i. negative: we reject the infection (H_1)
ii. positive: we accept the infection (H_1)

We can write the contingency table for the probabilities of the results of the hypothesis test as

	H_1 is true	H_1 is false
The test on H_1 is positive	$1-\beta$	α
The test on H_1 is negative	β	$1-\alpha$
sum	1	1

we therefore have

1. The **type 2 error** rate: probability of a false negative (ignore a finding when it is true)

$$\beta = P(negative|H_1)$$

2. The **True positive** rate: This is the power or sensitivity of a test (claiming a discovery when it is true, the main objective)

$$1-\beta = P(positive|H_1)$$

3. The **Type 1 error** rate: probability of a false positive (state a discovery when it is false)

$$\alpha = P(positive|H_0)$$

4. The **True Negative** rate: This is the specificity of a test (ignore a finding when it is false)

$$1-\alpha = P(negative|H_0)$$

14.9 Exercises

14.9.0.1 Exercise 1

Imagine we take a random sample of size $n = 41$ of a normal random variable X, and find that the sample average is 10 and the sample variance is 1.5.

- What is then the confidence interval for the mean of X at 95% confidence level?

Consider that $t_{0.025,40}$ = t.ppf(0.975, 40) ~ 2.

- Test the hypothesis that the mean of X is **different** than 10.5, using a 5% significance threshold.
- Write the code to calculate the P-value to test the hypothesis that the mean of μ is **lower** than 10.5, using a 5% significance threshold.

Consider that the code for the T probability distribution with $n-1$ degrees of freedom is t.cdf(tobs, n-1).

14.9.0.2 Exercise 2

10 gas condensates showed the following concentrations of mercury (in ng/ml):

23.3, 22.5, 21.9, 21.5, 19.9, 21.3, 21.7, 23.8, 22.6, 24.7

Assuming that the mercury concentration is distributed normally across gas condensates, test the hypothesis that condensate does not surpass the toxicity limit established at $24ng/ml$.

14.9.0.3 Exercise 3

The manufacturer of gene expression microarrays guarantees that at least 97% of the microarrays they produce have high-quality signals. A customer receives a batch of 200 pieces and finds that 8 unperformed.

Should the customer return the lot due to poor quality?

14.10 Practice

Load misophonia data https://alejandro-isglobal.github.io/SDA/data/data_0.txt

We have four measures of anxiety:

- Trait: ansiedad.rasgo (are you an anxious person?) continuous:0-100
- State: ansiedad.estado (are you currently feeling anxious?) continuous:0-100
- Diagnosed: ansiedad.medicada (have you been diagnosed with an anxiety disorder?) binary (si, no)
- Excess: ansiedad.dif (difference between State and Trait)

We are interested in the variable misofonia.dif, that is the observed **excess** of anxiety from the trait

$excess = state - trait$

Test the following hypotheses

a. Is excess in anxiety different from 0? is it higher?

b. Is excess in anxiety higher than 0 for men and women separately?

c. Is the proportion of anxious patients different from 0.03?

Solutions

Chapter 15

Contingency tables

15.1 Objective

In this chapter we will see how to test statistical independence between **two discrete** variables.

We will use the contingency table of conditional probabilities to derive the null hypothesis and test it with a χ^2 statistics.

We will also introduce the Fisher exact test for contingency tables.

15.2 Difference between proportions

Example (Hepatitis C)

For disease surveillance, we want to know if more hepatitis C patients are being observed in hospital A than in hospital B.

Let's take a sample of patients arriving to the hospital. We write down the diagnosis of hepatitis C of a patient who is treated at **hospital A**. In hospital A, we observe

$$1, 0, 0, 1, 0, ...0$$

where the first patient has hepatitis and the last one does not.

Hepatitis C diagnosis is a Bernoulli trial X with outcomes (0:no hepatitis and 1:hepatitis) that has a probability mass function

$$X_A \to Bernoulli(p_A)$$

The parameter p_A is the probability of hepatitis at hospital A.

We also write down the hepatitis status of a patient who goes to **hospital B**, and observe

$$0, 0, 1, 0, 0, ...1$$

Diagnosis at this hospital is a Bernoulli trial with probability p_B

$$K_B \rightarrow Bernoulli(p_B)$$

15.3 Difference between proportions

In this study **one** random experiment has a two-value outcome: $(disease, hospital)$. That is the observation of a patient is determined by two variables **both of which are categorical**

These are the possible outcomes of each variable:

- $Disease \in \{no, yes\}$
- $Hospital \in \{A, B\}$

Repeating the experiment n times, the data for the first five patients look like

```
##   Hospital Disease
## 1        A     yes
## 2        A      no
## 3        B      no
## 4        A     yes
## 5        A      no
```

The first patient was from hospital A and had hepatitis C.

Question:

Our research question is to find whether $Disease$ and $Hospital$ are **statistically dependent** variables. We want to know if there is more infections in one hospital than another and perhaps emit an alarm to the affected hospital. Let's formulate the hypothesis contrast.

15.4 Contingency table of conditional probabilities

Our experiment is such that we make diagnosis **conditioned** to each hospital. We then have the conditional probability table

Hospital: A

Hospital: B

No Hepatitis

P(no | A)=1-pA

P(no | B)=1-pB

Yes Hepatitis

P(yes | A)=pA

P(yes | B)=pB

sum

1

1

We want to know if for example the probability of infection in hospital A is different from the probability of infection in hospital B, and therefore the hypothesis contrast is

a. $H_0 : p_A = p_B$ (status quo), no differences of infection probability between hospitals.
b. $H_1 : p_A \neq p_B$ (research question), the infection probabilities are different between hospitals.

15.5 Test for the difference between proportions

If we consider that the **null hypothesis** is true, then the hospital tag ("A" or "B") is statistically independent from whether a diagnosed of hepatitis C was observed. Remember the definition of statistical independence: if infection is independence of hospitals then

$$p = P(yes) = P(yes|A) = P(yes|B)$$

Therefore the hypothesis contrast is:

a. $H_0 : p = p_A = p_B$
b. $H_1 : p \neq p_A \neq p_B$

We do **not know** the values of p, p_A and p_B. Therefore we need to **estimate** them from the data.

Example (Hepatitis C)

We perform diagnostic surveillance in both hospitals and found

- Hospital A included in the study a total of $n_A = 200$ and observed 18 infections.
- Hospital B included in the study a total of $n_B = 400$ and observed 46 infections.

For hospital A we have that the estimated probability of infection is the **relative frequency** of infections, that is

- $\bar{x}_A = \hat{p}_A = 18/200 = 0.09$

Consequently, for hospital B, we have that the estimated probability of infection is

- $\bar{x}_B = \hat{p}_B = 46/400 = 0.115$

If the null hypothesis is true then the rate of infection p can be estimated from either hospital, or from the two hospitals **taken together**, as if they were same hospital:

$$\hat{p} = \frac{n_A \bar{x}_A + n_B \bar{x}_B}{n_A + n_B} = \frac{18 + 46}{200 + 400} = 0.1066667$$

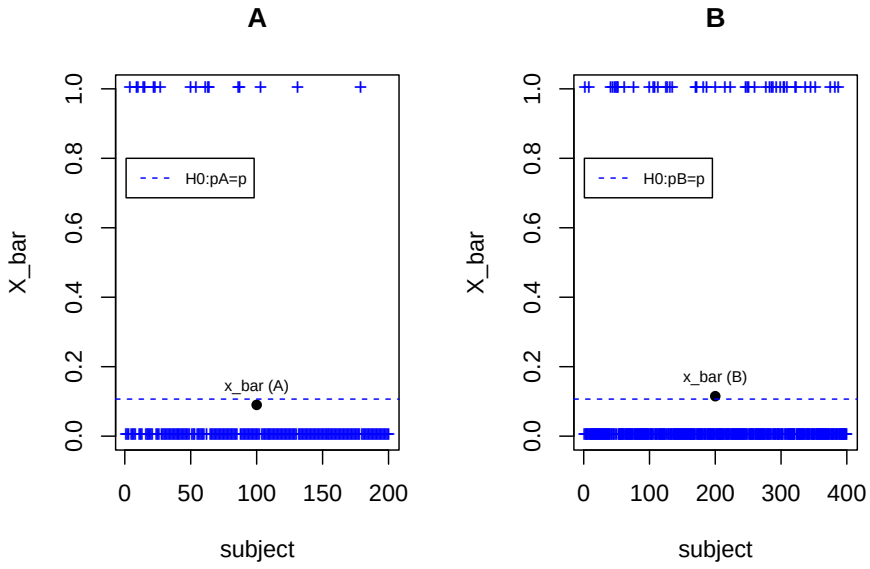
15.6 χ^2 test

For the first hospital the **standardized error** observed when we estimate the rate of infection $\hat{p}$ using only the average rate of infections at this hospital is

$$Z_A = \frac{\bar{X}_A - \hat{p}}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n_A}}}$$

The error made in hospital B is

$$Z_B = \frac{\bar{X}_B - \hat{p}}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n_B}}}$$



In the plot, each cross is a diagnosis (either 1 or 0) the dotted line is the estimated probability $\hat{p}$ of both hospitals taken together and the dots are the estimated probabilities of each separate hospital.

The conditional probability table by hospital is

Hospital: A

Hospital: B

No Hepatitis

0.910

0.885

Yes Hepatitis

0.090

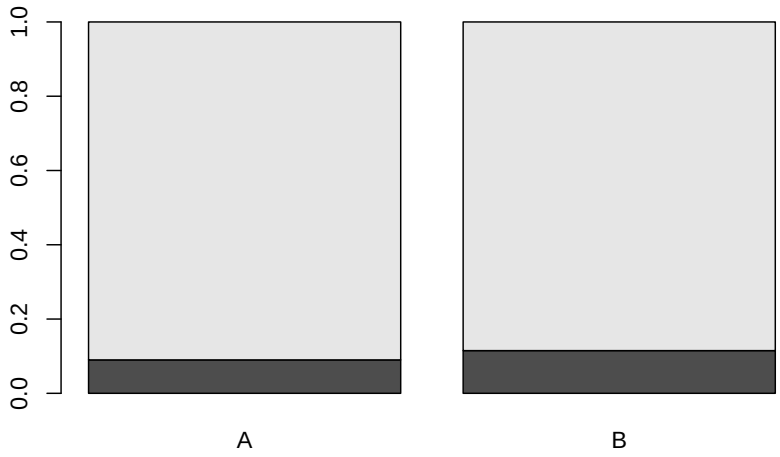
0.115

sum

1

1

Which we can illustrate in a barplot



The probabilities of infection within each hospital are different but look very similar. Is the difference significant, although small?

Using the CLT, the total **standardized squared error** from both hospitals is the sum of the errors

$$W = Z_A^2 + Z_B^2 \to \chi^2(1)$$

which is a random variable that follows a χ^2 distribution with 1 degree of freedom.

Using the marginals

We formulated the null hypothesis from the statistical independence given by conditional probabilities. **Another alternative** is to formulate it from the product of the marginals.

Therefore, the hypothesis test can **also** be written as

- $H_0 : P(yes, A) = P(yes)P(A)$ (Disease and Hospital A are statistically **independent** as the joint probability is the product of the marginals)
- $H_1 : P(yes, A) \neq P(yes)P(A)$ (Disease and Hospital A are statistically **dependent**)

If hospitals and infection status are truly independent from each other then one can show that the variable W above **can also** be computed as the sum of the squared errors that we make when we estimate the joint probabilities using the marginals (i.e. $\hat{p}_{yes,A} = \hat{p}_{yes}\hat{p}_A$), which we are allowed if the null hypothesis

is true. Therefore the total **standardized squared error** of before can be re-written as:

$$W = \frac{(\hat{p}_{no,A} - \hat{p}_{no}\hat{p}_A)^2}{\hat{p}_{no}\hat{p}_A} + \frac{(\hat{p}_{no,B} - \hat{p}_{no}\hat{p}_B)^2}{\hat{p}_{no}\hat{p}_B} + \frac{(\hat{p}_{yes,A} - \hat{p}_{yes}\hat{p}_A)^2}{\hat{p}_{yes}\hat{p}_A} + \frac{(\hat{p}_{yes,B} - \hat{p}_{yes}\hat{p}_B)^2}{\hat{p}_{yes}\hat{p}_B} \to \chi^2(1)$$

If the observed value for W is a rare error from the null hypothesis, for a χ^2 variable, we then reject the null hypothesis. The value of W is exactly the same however we compute it.

Example (Hepatitis C)

When we compute the observed value of W, using the first form, we find

$$w_{obs} = \frac{0.09 - 0.1066667}{\sqrt{\frac{0.1066667(1-0.1066667)}{200}}} + \frac{0.115 - 0.1066667}{\sqrt{\frac{0.1066667(1-0.1066667)}{400}}} = 0.87453$$

We can test whether this value is large under the null hypothesis. If the null is true this statistics is close to zero, otherwise is large. We can then compute the *pvalue* given by

$$pvalue = P(W \geq w_{obs}) = 0.3497$$

or in Python

```
import numpy as np
from scipy.stats import chi2_contingency

# Create the 2x2 contingency table
observed = np.array([[182, 18], [354, 46]])

# Perform the chi-squared test
chi2, p, dof, expected = chi2_contingency(observed, correction=False)
print(f"Chi-squared statistic: {chi2:.4f}")
print(f"P-value: {p:.4f}")
```

```
Chi-squared statistic: 0.8745
P-value: 0.3497
```

This *pvalue* is not lower than the significance level $\alpha = 0.05$ and therefore we **do not** reject H_0 and acept either

- that the frequencies of hepatitis C **are equal** between hospitals,
- or, equivalently, that the frequency of hepatitis C **is independent** from hospital,

- or, equivalently, that the frequency of hepatitis C **is not significantly associated** with the hospital.

15.7 Fisher's exact test

Another approach is **Fisher's exact test**: Take a ballot for each of the $N = 600$ patients of both hospitals into an urn:

There are a total of N ballots:

- There are $K = 64$ ballots that have hepatitis C
- $N - K = 536$ ballots do not have hepatitis C

Then, if we take a $n = 200$ randomly selected ballots (similar in number to hospital A), we may ask: What is the probability of observing more than 18 patients with hepatitis C? Remember that 18 infections were actually observed. Therefore, we are asking how rare that observation is in the case where the Hospital plays no role. That is the probability a more extreme observation under the null hypothesis, or put simply the *pvalue*. More explicitly the hypothesis contrast is

a. $H_0 : p_A \geq 64/600$

The null hypothesis (status quo) assumes that hospital A has at least the same parameter of both hospitals together.

b. $H_1 : p_A < 64/600$

The alternative hypothesis assumes that the parameter of hospital A is lower than the parameter for both hospitals together, indicating that there is a reduction in the number of infections in this hospital.

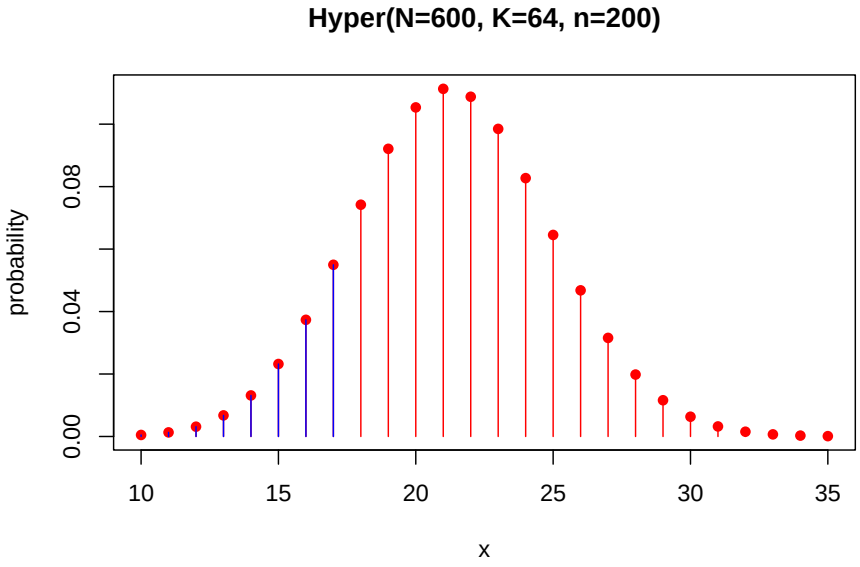
15.8 Hypergeometric distribution

The probability of obtaining x hepatitis C cases in a sample of n drawn from a population of N where K have hepatitis C is

$$P(X = x) = P(one\,sample) \times (Number\,of\,ways\,of\,obtaining\,x)$$

$$= \frac{1}{\binom{N}{n}} \binom{K}{x} \binom{N-K}{n-x}$$

$$X \to Hypergeometric(N, K, n)$$



If the observed value for $x = 18$ is a rare low **observation** from the null hypothesis, we then reject the null hypothesis. The lower tail *pvalue* for an observation $X = 18$ is

$$pvalue = P_{hyper}(X \leq 18) = 0.2147683$$

In Python

```
from scipy.stats import hypergeom

# Calculate the hypergeometric probability
hypergeom.cdf(18, 600, 64, 200)
```

Which is not lower than the significance level $\alpha = 0.05$ and therefore we **do not** reject H_0 and conclude:

- that the frequency of hepatitis C **is not significantly associated** with the hospital.

The **odds ratio** gives us the **strength** of the of the change in the in the infection proportion between hospitals:

$$OR = \frac{p_A/(1-p_A)}{p_B/(1-p_B)} = 0.76$$

It is a magnitude of the **observed statistical association** between hospital and disease. We use the OR to distinguish between the different cases

- When p_A is equal to p_B then $OR = 1$ (idependence of infection and hospital)
- When p_A is lower to p_B then $OR < 1$
- When p_A is greater than p_B then $OR > 1$

The way we report the OR in the case of the hospitals is: There was **a decrease** of 24% $(1 - 0.76)$ in the risk of hepatitis C for hospital A in relation to hospital B but it was not statistically significant ($pvalue = 0.21$).

This is how we compute it:

```
import numpy as np
from scipy.stats import fisher_exact

# Create the 2x2 contingency table
observed = np.array([[18, 182], [46, 354]])

# Perform the Fisher's exact test with "less" alternative hypothesis
odds_ratio, p_value = fisher_exact(observed, alternative="less")

print(f"Odds ratio: {odds_ratio:.4f}")
print(f"P-value: {p_value:.4f}")
```

```
Odds ratio: 0.7611
P-value: 0.2148
```

15.9 Difference between several proportions

Now, we want to know if the frequency of hepatitis C is different across 5 difference hospitals.

We then formulate the hypothesis contrast:

a. $H_0 : p = p_A = p_B = p_C = p_D = p_E$

The null hypothesis (status quo) assumes that hospital ($i = \{A, B, C, D, E\}$) and disease ($j = \{yes, no\}$) are all independent and then they have the same infection rate.

b. $H_1 : p \neq p_{A \cup B \cup C \cup D \cup E}$

The alternative hypothesis assumes that **at least one hospital** is not independent from infection.

When we collect the data we can build a contingency table for the observations

	A	B	C	D	E
Hepatitis (no)	182	354	375	85	90
Hepatitis (yes)	18	46	25	15	10

	A	B	C	D	E
sum	200	400	400	100	100

We have that the **standardized squared error** from the null hypothesis can be written as:

$$W = Z_A^2 + Z_B^2 + Z_C^2 + Z_D^2 + Z_E^2 \to \chi^2(4)$$

which is a random variable that follows a χ^2 distribution with $4 = 5 - 1$ degrees of freedom (number of hospitals -1).

Each term in W is the squared **standardized error** observed when we estimate the probability of infection $\hat{p}$ using only the frequency of infections at each hospital. The estimated probability of proportion across hospitals is

$$\hat{p} = \frac{n_A\bar{x}_A + n_B\bar{x}_B + n_C\bar{x}_C + n_D\bar{x}_D + n_E\bar{x}_E}{n_A + n_B + n_C + n_D + n_E} = \frac{18 + 46 + 25 + 15 + 10}{200 + 400 + 400 + 100 + 100} = 0.095$$

The computation of the observed **standardized squared error** is

$$w_{obs} = 10.381$$

For which we can test how rare an observation is, using a *pvalue*

$$pvalue = P(W \geq w_{obs}) = 0.03448$$

These values can be obtained in Python with the following code

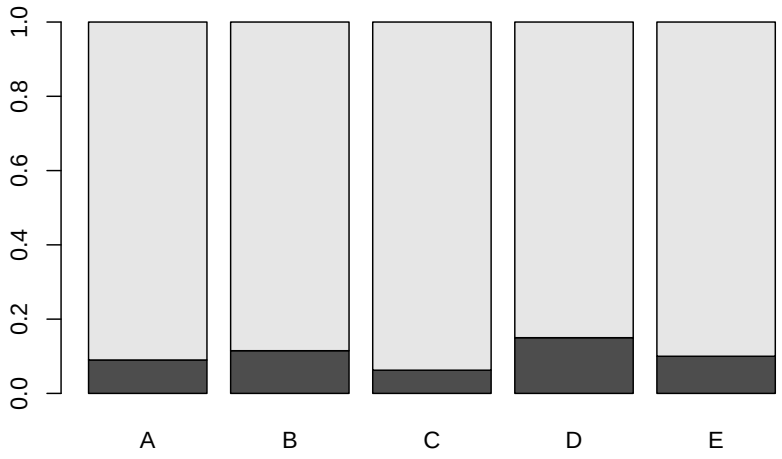
```
import numpy as np
from scipy.stats import chi2_contingency

# Create the 2x5 contingency table (adjust the data as needed)
observed = np.array([[18, 46, 25, 15, 10], [182, 354, 375, 85, 90]])

# Perform the chi-squared test
chi2_stat, p_value, dof, expected = chi2_contingency(observed)

print(f"Chi-squared statistic: {chi2_stat:.4f}")
print(f"P-value: {p_value:.4f}")
print(f"Degrees of freedom: {dof}")
```

```
Chi-squared statistic: 10.3809
P-value: 0.0345
```

The *pvalue* is lower than the significance level $\alpha = 0.05$ and therefore we **reject** H_0 and conclude that: The frequency of hepatitis C **is significantly associated** with the hospital.

We can illustrate the result a conditional frequency table by hospital or with a bar plot. We see that Hospital C and D may have too low and too large infection rates, respectively

```
##
##          A      B      C      D      E
##   0 0.9100 0.8850 0.9375 0.8500 0.9000
##   1 0.0900 0.1150 0.0625 0.1500 0.1000
```



15.10 Questions

1) The χ^2 test for contingency table is used to

a: test the associations of a categorial variable; **b:** test statistical independence between two categorical variables; **c:** test statistical independence between a continuous and a categorical variable; **d:** test significance of a categorial variable

2) In the Fisher test we compute the *pvalue* using

a: the central limit theorem; **b:** a conditional probability; **c:** the relative frequencies; **d:** a parameric model

3) The probability to heal from an injury without treatment is 0.6. The probability to heal from an injury with treatment is 0.8. What is the odd ratio of healing with treatment with respect of healing without it.

a: 0.376; **b:** 1.5; **c:** 2.666; **d:** 0.6666

4) The null hypothesis in testing for differences in proportions across different conditions assume

a: different conditions for different proportions; **b:** the same condition across different proportion; **c:** different proportions across the conditions; **d:** the same proportion across the conditions

5) if the *pvalue* is lower than 0.05 in a test of several proportions then

a: we accept that at least one proportion is different from the rest; **b:** We accept that all proportions are different between each other; **c:** We accept the all proportions are equal; **d:** We accept that there is one single proportion

15.11 Practice

Load hospital data (https://alejandro-isglobal.github.io/SDA/data/hospital.txt)

- Test the hypothesis that the infection proportion at hospitals C and D are different. Use a χ^2 test and a exact Fisher test. Interpret the OR.
- Make a bar plot
- Test the hypothesis that the infection proportion at hospitals A, B and E are different.
- Make a bar plot

Solutions

Chapter 16

Mean differences between two samples

16.1 Objective

In this chapter we will see how to test statistical independence of **one continuous** variable and two **discrete** conditions or groups.

We will see the case where n is large, where we can use a z-test. We will also introduce the t-test that allows testing for the cases of small n with equal and unequal variances.

We will introduce the concepts with the reanalysis of **real data** from reported studies relating sex differences in leptin levels.

16.2 Differece in means between two groups

We will consider an random variable of interest that follows a normal probability model

$$Y \to N(\mu, \sigma^2)$$

which we aim to measure several times under **two different conditions** or groups called A and B. We want to determine if the expected value (mean) of the random variable changes between the conditions.

When testing a hypothesis for the mean, that is a change of mean in a new condition, we often do not know its value in the status quo μ_0. Therefore, we often infer the mean in a condition A as **the null** (controlled condition) and then compare it with the mean in a **new condition**, where we have our research

interest, condition B (treatment condition). This is the case of testing for example differences in clinical outcomes between cases and controls, or between administration of a drug and a placebo.

Example (leptin)

Leptin is an adipose tissue hormone that creates the sensation of satiety after eating. We want to study the serum leptin levels in obese children (PMID: 18755049) under different conditions, such as sex. We therefore assume two conditions: $A : female$ and $B : male$ and assume that

1) the level of leptin in a randomly chosen girl has a probability density

$$Y_A \to N(\mu_A, \sigma_A^2)$$

2) the level in a boy has a probability density

$$Y_B \to N(\mu_B, \sigma_B^2)$$

16.3 Data

One random experiment has two variables: (y, c).

1. Y a continuous variable, whose observation produces the outcome of interest
2. $C \in \{A, B\}$ is a discrete variable with two possible outcomes, one for each experimental condition.

The repetition of the random experiment n times is therefore a table such as

subject	Y	C
1	1.2	A
2	0.9	A
...	...	...
n	1.4	B

For example, subject 1 had outcome 1.2 for Y under condition A.

Example (leptin)

In the leptin example, we have that the continuous outcome is the level of leptin in a subject

- $leptin \in (0, 200)$

And that the condition is the sex of the subject:

- $sex \in \{girl : A, boy : B\}$

190 girls were tested for the level of leptin, while 166 boys were included in the studied. The first six subjects of the studied were

```
##     leptin  sex
## 1 32.81111 girl
## 2 41.76219 girl
## 3 90.24100 girl
## 4 49.91078 girl
## 5 51.50370 girl
## 6 94.47826 girl
```

While the last subjects were

```
##        leptin sex
## 351  57.13473 boy
## 352 -10.23757 boy
## 353  24.82894 boy
## 354  37.01268 boy
## 355  14.39215 boy
## 356  41.87943 boy
```

We want to know whether boys have different levels of leptin than girls. In other words, if the mean levels of leptin change between sexes.

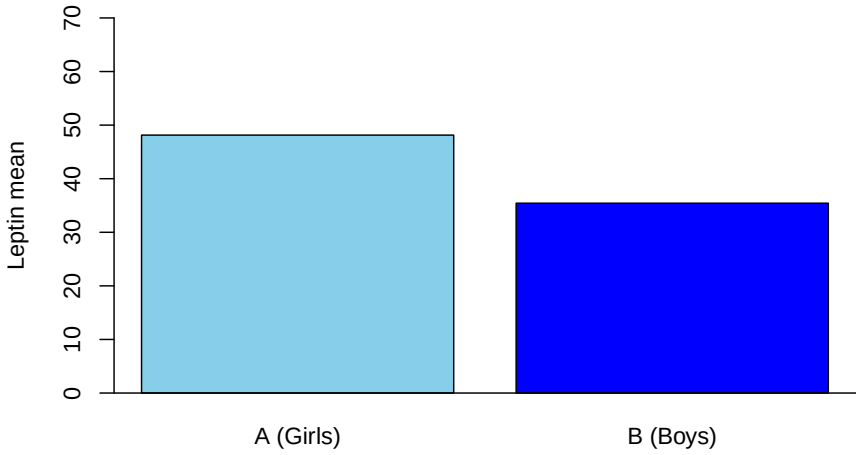
Research question: is **leptin** statistically dependent from **Sex**?

16.4 Difference between means

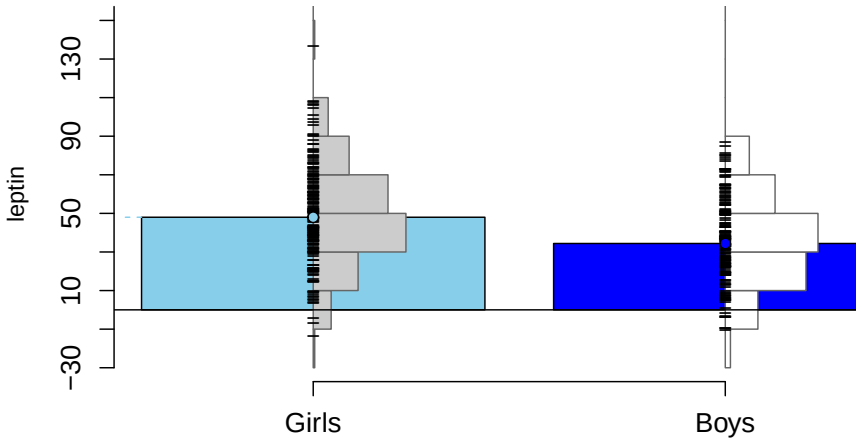
We took leptin levels **conditioned to** each sex, and observed:

- $n_A = 190$ girls had a mean of $\bar{y}_A = 48.13$ and $s = 25.44$
- $n_B = 166$ boys had a mean of $\bar{y}_B = 35.42$ and $s = 21.99$

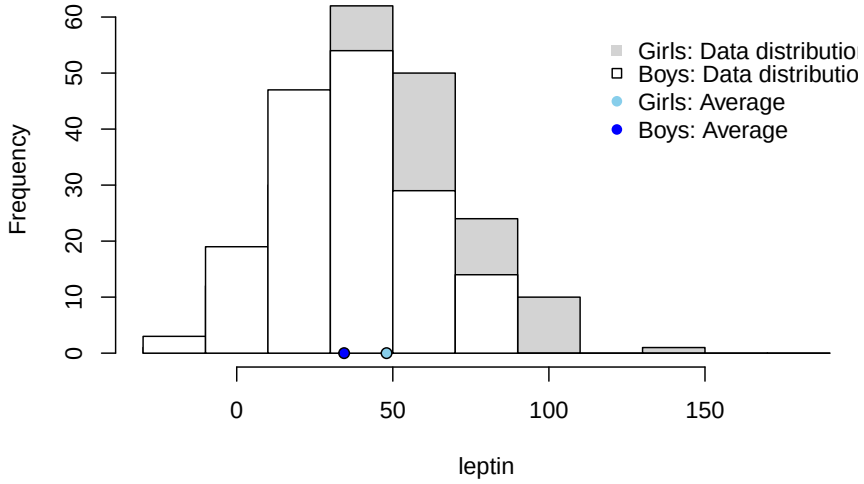
Bar plots for the means are popular, although they do not show the spread of the data



We can draw histograms of leptin for each group to have a better idea of the distributions



Or we can overlay the histograms horizontally



The plots suggest that boys have lower levels of leptin **on average** but is the difference between the averages $\bar{y}_A$ and $\bar{y}_B$ **statistically significant**?

How confident are we on the magnitude of the difference? That is, could have we got this lower levels of lepting in boys by chance alone (null hypothesis) or there is difference that cannot be fully explained by chance (alternative hypothesis)?

16.5 Hypothesis test

Let us formulate the hypothesis contrast

a. $H_0 : \mu_A = \mu_B$ The null hypothesis (status quo) is that both groups have the same mean.

b. $H_1 : \mu_A \neq \mu_B$ The alternative hypothesis (research interest) is that the means are different between groups.

Only one can be true. How would the data decide? To make things easier let us redefine the contrast.

If we consider δ, the difference between means $\delta = \mu_A - \mu_B$, then the hypotheses can be written as

a. $H_0 : \delta = 0$
b. $H_1 : \delta \neq 0$

Either the difference between the means is zero or not. δ is the **parameter of interest** in our study, and to test hypotheses on it, we need to find an

estimator for it.

16.6 Estiamtor of the mean difference

The statistic $D = \bar{Y}_A - \bar{Y}_B$, that is the difference in averages, is an estimator of δ. In particular, we can show that the the estimator is

- unbiased because its expected value is the parameter

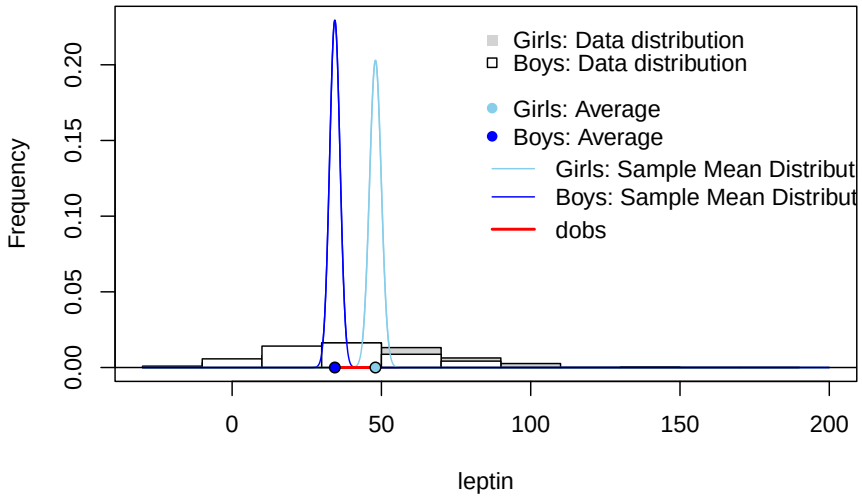
$$E(D) = E(\bar{Y}_A - \bar{Y}_B) = \mu_A - \mu_B = \delta$$

- and consistent because its variance gets smaller when $n = n_A + n_B$ gets bigger

$$V(D) = V(\bar{Y}_A - \bar{Y}_B) = \frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}$$

This means that we can take one observed value of D, that is d_{obs}, for the estimation of the parameter δ.

It is important to have a clear idea of the different components in the estimation and the hypothesis test, as they are shown bellow



Remember that, for instance, the average $\bar{Y}_A$ is also called the sample mean, and it has a distribution that is centered on μ_A and has variance $\frac{\sigma_A^2}{n_A}$.

16.7 Standardized error

As both Y_A and Y_B are normal and independent variables, their averages are normal and so it is their difference D. Therefore, the standardization of D is arrpoximately standard normal

$$Z = \frac{D - \delta}{\sqrt{V(D)}} = \frac{\bar{Y}_A - \bar{Y}_B - \delta}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}} \rightarrow_{approx} N(0,1)$$

when both n_A and n_B are large. It is approximate because as we do not know the variances σ_A^2 and σ_B^2, we estimated them with s_A^2 and s_B^2. We can also use this approximation when we do not know the distributions of Y_A and Y_B, as a result of the central limit theorem.

16.8 Standardized error for the null

If the null hypothesis is true ($\delta = 0$), then when we estimate δ with D we make an error. The **standardized error** from the null hypothesis is the statistic

$$Z = \frac{\bar{Y}_A - \bar{Y}_B}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}}$$

that follows a standard normal distribution.

To test the hypothesis we then ask if the observed z_{obs} is within the acceptance region of the null hypothesis.

In particular, we ask: is the probability of observing a more extreme value than z_{obs} lower than $\alpha = 0.05$ if the null hypothesis is true?

The value of z_{obs} is the standardized value of the observed difference between the averages:

$$z_{obs} = \frac{d_{obs}}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}}$$

For our data, we have that the **observed** standardized mean difference is

$$z_{obs} = \frac{\bar{y}_A - \bar{y}_B}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}} = \frac{48.13 - 34.42}{\sqrt{\frac{647.58}{190} + \frac{483.62}{166}}} = 5.053$$

Since our hypothesis is two tailed then the the two-tailed *pvalue* is

$$pvalue = 2(1 - \phi(5.053)) = 4.32 \times 10^{-7}$$

which is much lower than α.

Therefore, we reject the null hypothesis that $H_0 : \delta = 0$ that is that the leptin levels in obese children are equal between boys and girls. In other words, we have strong evidence that the mean leptin levels between sexes are different.

In Python

```
import pandas as pd
import math
from scipy.stats import norm

data = pd.read_csv('https://alejandro-isglobal.github.io/SDA/data/leptin_sex.txt',
sep='\t')

groupA = data[data['sex']=="girl"]['leptin']
groupB = data[data['sex']=="boy"]['leptin']

yA_bar = np.mean(groupA)
yB_bar = np.mean(groupB)

s2A = np.var(groupA, ddof=1)
s2B = np.var(groupB, ddof=1)

d = (yA_bar-yB_bar)/math.sqrt(s2A/len(groupA)+s2B/len(groupB))
pval= 2*(1-norm.cdf(d))

d, pval

(5.053890638648715, 4.32899547764265e-07)
```

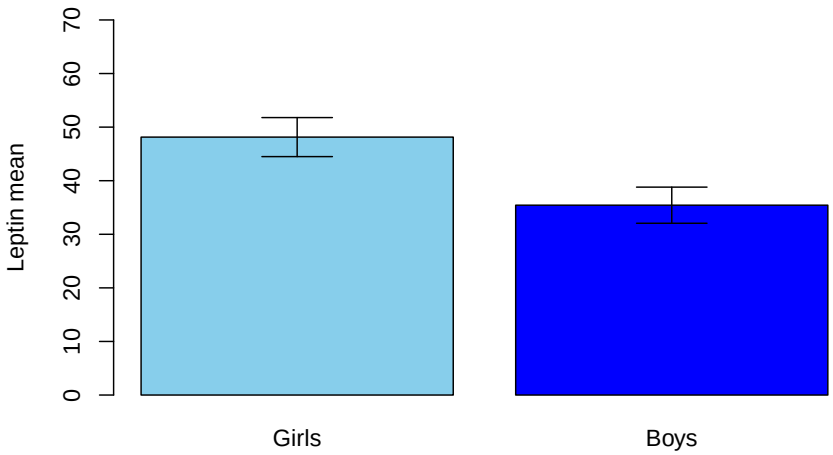
Example (Abstract)

Abstract:

Obesity rates are different between boys and girls, suggesting that the physiopathology of the disease is different between the sexes.

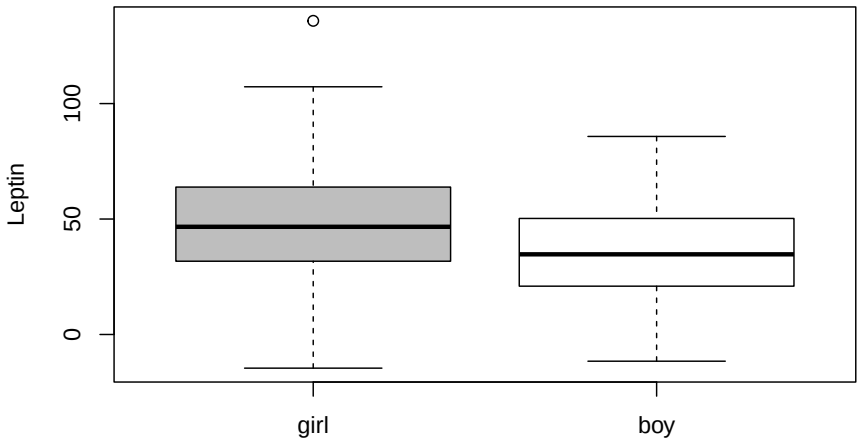
In this study, we tested the hypothesis that the leptin levels in serum are different between boys and girls.

We analyzed data from 190 obese girls and 166 obese boys and found a significant difference in leptin between sexes (mean difference 13.6*,* $P = 2.195757 \times 10^{-7}$*)*



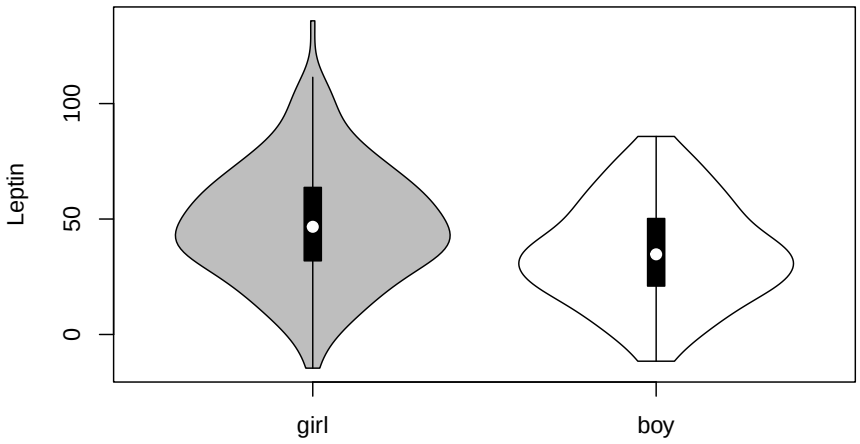
Note that in a report we usually add the confidence intervals for the means. Non-overlapping confidence intervals also indicate that the difference between means is **statistically significant** (reject the null).

Another popular way to visualize the distributional differences between the groups is to use a box plot



In this plot, we do not show the means but a summary of the distribution properties of the data at each condition. Remember that the properties are the median (the middle line), the quartiles (the box edges) and the 5% and 95% quantiles (the wiskers).

Finally, you would see that violin plots are also widely used. These are boxplots with the (smoothed) mirrowed histograms overlayed. Here, you do not see the everages of each group but the modes (peaks of the hitograms). However, you get the idea that the distribution for girls higher than the distribution for boys.



16.9 Mean differences when n is small

In a study that wanted to test the effect of leptin in neurodevelopment, 7 male mice had their leptin gene knocked out. And, they could not produce the horomone. While 16 mice were left with normal leptin function (PMID:30694175). These are called wild type. An initial question was to test the effect of leptin on the body weight of the animals. Is the mean weight of the animals different between wild types and knock-outs?

We assume that

1. the weight of the control animals (wild type) has a probability density

$$Y_A \to N(\mu_A, \sigma^2)$$

2. the weight of the animals with no leptin gene has a probability density

$$Y_B \to N(\mu_B, \sigma^2)$$

3. both distributions have the same variance σ^2.

16.10 Data

One random experiment in this study has two outcomes: $(weight, leptin)$.

Continuous variable (outcome of interest)

- $weigth \in (20, 60)$

Categorical variable:

- $leptin \in \{control : A, knockout : B\}$

The data looks like

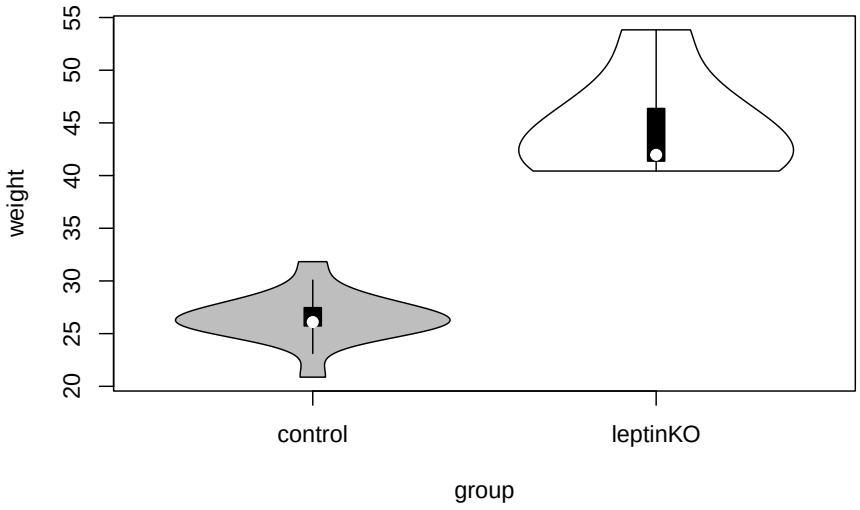
```
##    weight    group
## 1   27.67  Control
## 2   27.40  Control
## 3   25.77  Control
## 4   25.60  Control
## 5   25.03  Control
## 6   25.90  Control
## 7   26.67  Control
## 8   25.60  Control
## 9   28.93  Control
## 10  31.83  Control
## 11  25.90  Control
## 12  26.30  Control
## 13  27.90  Control
## 14  26.77  Control
## 15  25.83  Control
## 16  20.87  Control
## 17  46.57 leptinKO
## 18  40.43 leptinKO
## 19  41.97 leptinKO
## 20  41.17 leptinKO
## 21  41.57 leptinKO
## 22  46.17 leptinKO
## 23  53.83 leptinKO
```

16.11 Difference between means

We take mice weights **conditioned to** each lepting group, and observed:

- $n_A = 16$ control mice had a weight mean of $\bar{y}_A = 26.49$ and $s_A = 2.24$
- $n_B = 7$ leptin KO mice had a weight mean of $\bar{y}_B = 44.53$ and $s_B = 4.77$

We can draw violin plots per group



We see that the distributions are different and in particular that mice without leptin (leptin KO) have on average higher weights. Can we have a statistical test for this difference?

16.12 Hypothesis test

We can again formulate the hypothesis on the difference between means $\delta = \mu_{control} - \mu_{leptin}$

a. $H_0 : \delta = 0$. The null hypothesis (status quo) assumes that both mice (control, leptin KO) have the same mean weight.

b. $H_1 : \delta \neq 0$. Therefore, the alternative hypothesis is that the mean weigth is different between groups.

16.13 Estimator of the mean difference

The statistic $D = \bar{Y}_A - \bar{Y}_B$ is again an unbiased estimator of δ

$$E(D) = \delta$$

Since the weights for A and B are normal variables with the same variance σ^2 each sample variance in each group is an estimator of σ^2. That is

$$\hat{\sigma}^2 = s_A^2 = s_B^2$$

then (Theorem) the **standardized error**

$$T = \frac{\bar{Y}_A - \bar{Y}_B - \delta}{\sqrt{s_p^2(\frac{1}{n_A} + \frac{1}{n_B})}} \to T(n_A + n_B - 2)$$

follows exactly a T-distribution with $n_A + n_B - 2$ degrees of freedom.

The **pooled variance** s_p^2, is an estimator of σ^2

$$\hat{\sigma}^2 = s_p^2 = \frac{(n_A - 1)s_A^2 + (n_B - 1)s_B^2}{n_A + n_B - 2}$$

16.14 Standardized error for the null

If the null hypothesis is true ($\delta = 0$), then when we estimate δ with D we make an error. The **standardized error** from the null hypothesis is the statistic

$$T = \frac{\bar{Y}_A - \bar{Y}_B}{\sqrt{s_p^2(\frac{1}{n_A} + \frac{1}{n_B})}} \to T(n_A + n_B - 2)$$

that follows a T-distribution.

To test the hypothesis we then ask if the observed t_{obs} falls within the acceptance region of the null hypothesis.

In particular, we ask: is the probability of observing a more extreme value than t_{obs} lower than $\alpha = 0.05$, if the null hypothesis is true?

The value of t_{obs} is the standardized value of the observed difference between the averages:

$$t_{obs} = \frac{d_{obs}}{\sqrt{s_p^2(\frac{1}{n_A} + \frac{1}{n_B})}}$$

$$= \frac{\bar{y}_A - \bar{y}_B}{\sqrt{\frac{s_p^2}{n_A} + \frac{s_p^2}{n_B}}} = \frac{26.49 - 44.53}{\sqrt{\frac{10.12}{16} + \frac{10.12}{7}}} = -12.508$$

The two-tailed *pvalue* of t_{obs} is

$$pvalue = 2(1 - F_{t,21}(12.508)) = 3.37 \times 10^{-11}$$

which is lower than $\alpha = 0.05$.

Therefore, the data shows a very significant increase in 18.03gr ($P = 3.376854 \times 10^{-11}$) in weight between the wild-type mice and leptin knockouts. "Absence

of leptin signaling in early life alters the energy balance and predisposes the animals to obesity", as quoted from the original article.

in Python

```
import numpy as np
from scipy import stats

data = pd.read_csv('https://alejandro-isglobal.github.io/SDA/data/leptin.txt',
sep='\t')

groupA = data["Control"]
groupB = data["LepNull"]

stats.ttest_ind(groupA, groupB, equal_var=True, nan_policy='omit')
```

```
Ttest_indResult(statistic=-12.507930602554847, pvalue=3.377211689857233e-11)
```

16.15 Unequal variances

The box plot suggests that the variances for each group are different because the box for the control group is thinner (less dispersed) than the box for the leptin KO group. Therefore it is better to assume that

1. the weight of the control animals (wild type) has a probability density

$$Y_A \to N(\mu_A, \sigma_A^2)$$

2. the weight of the animals with no leptin has a probability density

$$Y_B \to N(\mu_B, \sigma_B^2)$$

The **standardized error** for the null hypothesis ($\delta = 0$)

$$T = \frac{\bar{Y}_A - \bar{Y}_B}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}} \to_{aprox} T(\nu)$$

approximately follows a t-distribution with

$$\nu = \frac{(\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B})^2}{\frac{(s_A^2/n_B)^2}{n_A - 1} + \frac{(s_B^2/n_B)^2}{n_B - 1}}$$

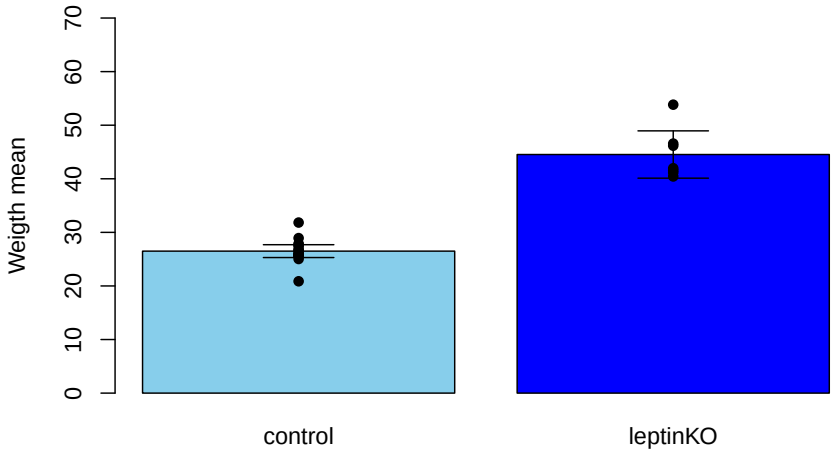
degrees of freedom. The brilliant idea of Welsh was to fix the form of the variance of D (denominator of T), and then ask which was the closest t-distribution that would describe T. For that he "only" needed to adjust the degrees of freedom. This is called a Welsh test.

In Python

```
import numpy as np
from scipy import stats

data = pd.read_csv('https://alejandro-isglobal.github.io/SDA/data/leptin.txt',
sep='\t')

groupA = data["Control"]
groupB = data["LepNull"]

stats.ttest_ind(groupA, groupB, equal_var=False, nan_policy='omit')
```

```
Ttest_indResult(statistic=-9.541047987836473, pvalue=2.4443518964606636e-05)
```

Therefore, the data **still** shows a very significant increase in 18.03gr ($P = 2.444 \times 10^{-5}$) in weight between the wild-type mice and leptin knockouts (under a more appropriate test).



16.16 Questions

1) We test for the difference between the means of two random variables when we have measured

a: two continuous random variables; **b:** two categorical random variables; **c:** one dichotomic variable and one continuous random variable;

d: any categorical variable and one continuous random variable;

2) The statistic $D = \bar{Y}_A - \bar{Y}_B$ estimates

a: The difference between the averages of Y_A and Y_B; **b:** The difference between the means of Y_A and Y_B; **c:** The variation of $\bar{Y}_A$ with respect to $\bar{Y}_B$; **d:** d_{obs}

3) We can use a Z-test only when

a: we have a large sample of the continuous variable in one of the conditions; **b:** we have large samples of the continuous variable in each condition; **c:** the variances of the continuous variable are equal in each conditions; **d:** when the distributions of the continuous variable in each condition are normal

4) We can use a t-test only when

a: we have small samples of the continuous variable in each condition; **b:** the distributions of the continuous variable in each condition are normal; **c:** D is unbiased; **d:** we cannot use a Z-test

5) For the data

subject	Y	C
1	1.1	A
2	0.9	A
3	0.8	B
4	0.6	B

assuming that Y is normally distributed, which is the best test?

a: t-test with equal variances **b:** t-test with unequal variances **c:** z-test with equal variances **d:** z-test with unequal variances

16.17 Practice

Load leptin data https://alejandro-isglobal.github.io/SDA/data/dataleptin.txt

- Test the hypothesis that in the control animals, the weight of females is different than the weight of males
- Test the hypothesis that the leptin KO animals have higher weight than the KO animals with supplemented leptin
- Make a bar plot and a boxplot for the latter case. Compute the confidence intervals for the weight, in each group.

Solutions

Chapter 17

Mean differences across several groups

17.1 Objective

In this chapter, we will generalize the test for difference between means among many groups. That is many levels of a categorical variables. We will do this by using the analysis of variance (ANOVA) that compares the variance within the groups and the variance between the groups.

We will see that ANOVA can be further used to describe the mean of a continuous variable across two different categorical variables with many levels each. This is called a 2-way ANOVA.

Finally we will use the ANOVA to test that at least one combination of the two categorical variables can not be explained by the independent contributions of each condition. This is called a 2-way ANOVA plus interaction.

17.2 Different means among several conditions

In a study that wanted to test the effect of leptin in neurodevelopment, 7 male mice had their leptin gene knocked out. Whereas, 16 mice were left with normal leptin function. In a third group, 10 mice with knocked out leptin were injected leptin protein (PMID:30694175). An initial question was to test the effect of the leptin group on the body weight of the animals. We therefore assume three experimental conditions and that

1. the weight of the control animals (wild type) has a probability density

$$Y_A \to N(\mu_A, \sigma^2)$$

2. the weight of the animals with no leptin gene (KO) but were injected leptin protein has a probability density

$$Y_B \to N(\mu_B, \sigma^2)$$

3. the weight of the animals with no leptin gene (KO) has a probability density

$$Y_C \to N(\mu_C, \sigma^2)$$

17.3 Data

One random experiment has two outcomes: $(weight, leptin)$.

A continuous variable (outcome of interest)

- $weight \in (20, 60)$

A categorical variable with three conditions (levels)

- $leptin \in \{control : A, KOplus : B, leptinKO : C\}$

The data looks like

```
##    weight    group
## 1   27.67  Control
## 2   27.40  Control
## 3   25.77  Control
## 4   25.60  Control
## 5   25.03  Control
## 6   25.90  Control
## 7   26.67  Control
## 8   25.60  Control
## 9   28.93  Control
## 10  31.83  Control
## 11  25.90  Control
## 12  26.30  Control
## 13  27.90  Control
## 14  26.77  Control
## 15  25.83  Control
## 16  20.87  Control
## 17  46.57 leptinKO
## 18  40.43 leptinKO
## 19  41.97 leptinKO
## 20  41.17 leptinKO
## 21  41.57 leptinKO
## 22  46.17 leptinKO
## 23  53.83 leptinKO
```

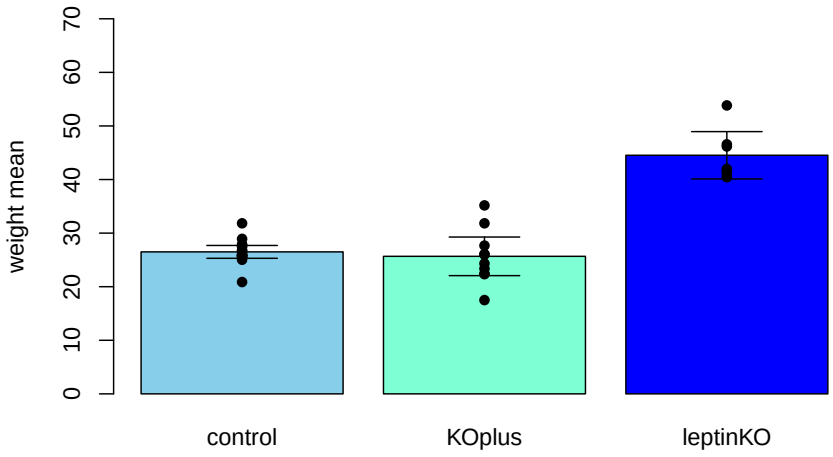
```
## 24  24.33   KOplus
## 25  22.37   KOplus
## 26  26.10   KOplus
## 27  17.50   KOplus
## 28  35.17   KOplus
## 29  25.97   KOplus
## 30  27.67   KOplus
## 31  23.37   KOplus
## 32  31.83   KOplus
## 33  22.37   KOplus
```

17.4 Difference between means

We take weights **conditioned to** each leptin group, and observed:

- $n_A = 16$ control mice had a weight mean of $\bar{y}_A = 26.49813$ and $s_A = 2.247577$
- $n_B = 10$ leptin KO mice with leptin replacement had a weight mean of $\bar{y}_B = 25.668$ and $s_B = 5.034161$
- $n_C = 7$ leptin KO mice had a weight mean of $\bar{y}_C = 44.53$ and $s_C = 4.774167$

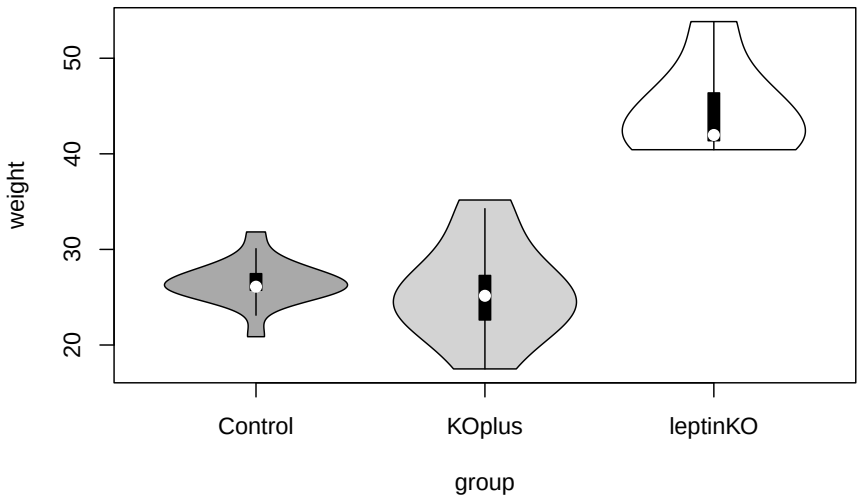
We can draw a barplot for each mean with a confidence interval



and we see that the wild type (control) has a similar mean weight to the mice

without the leptin gene but which have been injected leptin (KOplus). The mice without leptin gene and no leptin protein (leptinKO) have mean higher mean weight. As the confidence intervals for this group does not overlap with the other two, we say that the mean of the knock outs is significant higher than for the other two cases.

We can see the differences in distributions with a violin plot



When we test for the difference in means between **two groups**, the null hypothesis assumes that the means are equal. When we consider more than two groups, we ask if **at least one** of the means is different.

17.5 Hypothesis test

Let us formulate the hypothesis contrast

a. $H_0 : \mu_A = \mu_B = \mu_C$ The null hypothesis (status quo) assumes that there are no differences between group means.

b. H_1: at least one mean $\mu_i$is different from the rest (research interest).

How can we test this hypothesis? Is there a statistic that we can formulate such that its value give us information on whether we should accept or reject H_0?

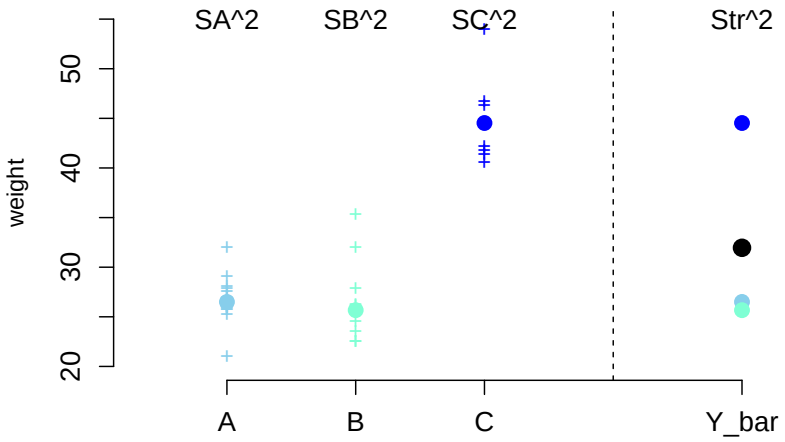
Let's look at the dispersion of **all the data points** at wach group. Within each group the data disperses about their averages. Therefore an estimator of the variance for each group, for example for group A, the sample variance

$$S_A^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_{Ai} - \bar{Y}_A)^2$$

is an estimator of the variance σ^2 of the weights in group A. n is the number of observations made in group A. Since we are assuming that in all groups the data disperses with the same variance σ^2 then the average of the sample variances

$$Se^2 = \frac{1}{3}(S_A^2 + S_B^2 + S_C^2)$$

is another estimator of the variance σ^2 that takes into account the estimators at each group. S_e^2 is called the **mean square error**.



Let's now look at the dispersion of **averages** of each group, instead. That is the right part of the plot above (Y_bar), where we look at how the averages disperse about the overall mean of the data. We know, for instance, that $\bar{Y}_A$ is normal with mean μ_A and variance $\frac{\sigma^2}{n}$.

If the **null hypothesis is true** there are no differences between the means and then

$$\mu_A = \mu_B = \mu_C = \mu$$

Therefore $\bar{Y}_A, \bar{Y}_B, \bar{Y}_C \rightarrow N(\mu, \sigma^2/n)$ and $(\bar{y}_A, \bar{y}_B, \bar{y}_C)$ is the observed sample of size $k = 3$ of the same random experiment: one that takes averages on the

weights of animals in groups that are identical between each other (no differences between groups). In this experiments of averages, the sample variance

$$S_{tr}^2 = \frac{1}{k-1}\sum_{j=1}^{k}(\bar{Y}_j - \bar{Y})^2$$

estimates the sample mean variance σ^2/n. j indicates the groups (A, B, C) and $\bar{Y}$ is the average of the averages or put simply: the average for all observations (the black dot in the figure)

$$\bar{Y} = \frac{1}{nk}\sum_{j}\sum_{i} Y_{ji}$$

If the null **hypothesis is true** and we take more and more observations in each group then all the group averages get closer to the overall average, and the overall average gets closer to the grand mean μ (which is unknown). Since S_{tr}^2 estimates σ^2/n and S_e^2 estimates σ^2 then the ratio

$$F = \frac{nS_{tr}^2}{S_e^2}$$

is a statistic that is close to 1, when H_0 is true.

If the null **hypothesis is not true** and we take more and more observations in each group then at least one average will not get close to the other averages, for instance $\mu_A = \mu_B = \mu \neq \mu_C$. Therefore the expected value of nS_{tr}^2 will not get close to σ^2, as there is an irreducible variance given by the group that is different, for instance group C. We also say that there is an additional variance in the data given by the treatment. In this case the statistic

$$F = \frac{nS_{tr}^2}{S_e^2} = \frac{MST}{MSE}$$

will be greater than 1 with high probability. The numerator MST is called the mean sum of squares of treatments (groups) and the denominator MSE is the sum of squares of errors (residuals). Since the variables within each group are considered normal, the variance estimators follow χ^2 distributions with their respective degrees of freedom, and the ratio F follows an (F) Fisher distribution with $k-1$ and $kn-1$) degrees of freedom

$$F \to Fisher(k-1, k(n-1))$$

where k is the number of groups and n is the average number of observations per group.

17.6 Analysis of variance (ANOVA)

The hypothesis test for the difference in means across several groups is then

- $H_0 : \mu_1 = \mu_2, ... = \mu_k$ There are no difference between group means

Then, the observed value of F, *e.g* f_{obs} will be near 1.

- H_1 at least one μ_i is different

Then, the observed value of F, *e.g* f_{obs} will be **greater** than 1.

Example (mice knockouts)

The value of the f_{obs} computed from the group (treatments) mean and residual (error) mean squares can be calculated in any statistical software that usually outputs the ANOVA table

```
import pandas as pd
import statsmodels.api as sm
from statsmodels.formula.api import ols

data = pd.read_csv('https://alejandro-isglobal.github.io/SDA/data/dataleptin.txt',
sep=' ')

filtsex = data['sex']=='M'
filtered_data = data[filtsex]

model = ols('weight ~ group', data=filtered_data).fit()

anova_table = sm.stats.anova_lm(model, typ=1)

print(anova_table)
```

```
             df       sum_sq     mean_sq          F        PR(>F)
group       2.0  1861.549820  930.774910  63.373347  1.693725e-11
Residual   30.0   440.615004   14.687167        NaN           NaN
```

In this table MST is the mean squares for the group (treatment), computed as the corresponding sum of squares divided by the degrees of freedom ($k-1=2$).

$$MST = \frac{1}{k-1} SSq_{group} = 930.77$$

MSE is the mean squares for the residuals (error), computed as the corresponding sum of squares divided by the degrees on freedom ($k(n-1) = 30$), where $n = (16+7+10)/3 = 11$ is the mean number of observations in each group

$$MSE = \frac{1}{k(n-1)} SSq_{Residual} = 14.69$$

The observed value of the statistics is

$$f_{obs} = \frac{MST}{MSE} = 63.373$$

To test the the null hypothesis, we then compute the probability that in a future experiment, if the null hypothesis is true, we observe a higher value than 63.373: $P(F > 63.373)$. This probability is the **upper tailed p-value**

$$pvalue = 1 - F_{Fisher,2,30}(63.373) = 1.694 \times 10^{-11}$$

which is much lower than $\alpha = 0.05$, suggesting significant differences in at least one group mean. Therefore we reject the null hypothesis and accept that at least one of the groups has a different mean from the rest.

Conclusion:

We observed a significant difference between groups (ANOVA test $F(2, 30) = 63.373, P = 1.69 \times 10^{-11}$), due to the higher gain in weight of the knockout mice. Note that knocked-out mice with replacement recovered wild-type weight (t-test difference between means -0.83, $P = 0.63$)

17.7 ANOVA for two groups

We can explicitly compute the observed value of F when we have only two groups:

$$f_{obs} = \frac{n(\bar{y}_A - \bar{y}_B)^2}{s_p^2}$$

where s_p^2 is the pooled variance:

$$s_p^2 = \frac{s_A^2 + s_B^2}{2}$$

taking the square root of f_{obs} and rearranging we find

$$t_{obs} = \sqrt{f_{obs}} = \frac{(\bar{y}_A - \bar{y}_B)}{\frac{s_p}{\sqrt{n}}}$$

and thus recover the t-test for the difference between two groups. ANOVA is the generalization of a two-sample t-test, indeed.

Example (wild type and knockout mice)

When we compared the weights between wild types (group A) and knock out mice (group C), we performed a t-test. The observed statistic was

$$t_{obs} = \frac{\bar{y}_A - \bar{y}_B}{\sqrt{\frac{s_p^2}{n_A} + \frac{s_p^2}{n_B}}} = \frac{26.49813 - 44.53}{\sqrt{\frac{3.18127^2}{16} + \frac{3.18127^2}{7}}} = -12.508$$

Therefore the observed F is

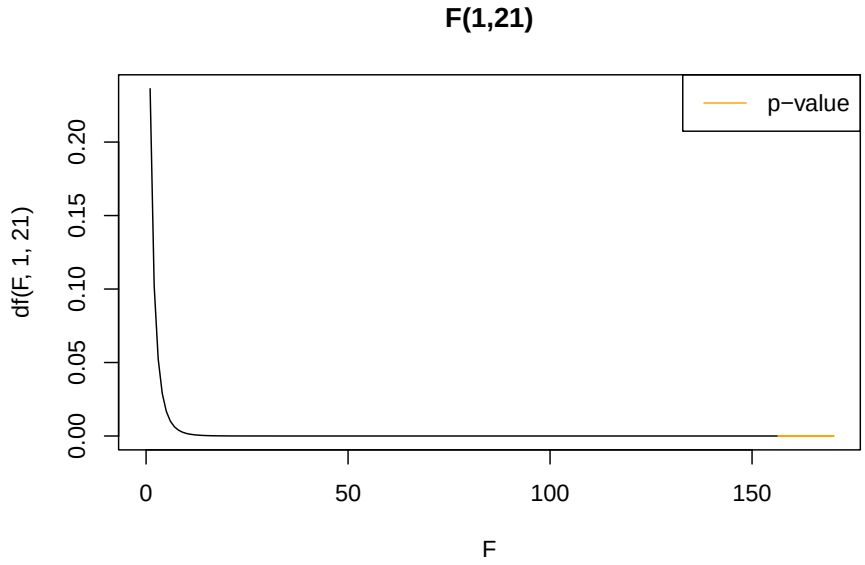
$$f_{obs} = t_{obs}^2 = (-12.508)^2 = 156.45$$

The upper tailed p-value for f_{obs} using the Fisher distribution is

$$pvalue = 1 - F_{Fisher,1,21}(156.45) = 3.377 \times 10^{-11}$$

with $k = 2$ and $n = 11.5$ is the average number of observations per group.

in R: 1-pf(156.45, 1,21)



This *pvalue* (the yellow area at the far right of the plot) is exactly the same as the one we obtained with the t-test with equal variances. Therefore, we reject the hypothesis that the weights between wild types and leptin knock outs are the same.

We can check the ANOVA table

```
## Analysis of Variance Table
##
## Response: weight
##           Df  Sum Sq Mean Sq F value    Pr(>F)
```

```
## group      1 1583.33 1583.33  156.45 3.377e-11 ***
## Residuals 21  212.53   10.12
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
import pandas as pd
import statsmodels.api as sm
from statsmodels.formula.api import ols

data = pd.read_csv('https://alejandro-isglobal.github.io/SDA/data/dataleptin.txt',
sep=' ')

filtsex = data['sex']=='M'
filtgroups = (data['group']=='Control') | (data['group']=='leptinKO')

filtered_data = data[filtsex & filtgroups]

model = ols('weight ~ group', data=filtered_data).fit()

anova_table = sm.stats.anova_lm(model, typ=1)

print(anova_table)
```

```
            df       sum_sq      mean_sq           F        PR(>F)
group      1.0  1583.331904  1583.331904  156.448328  3.377212e-11
Residual  21.0   212.530044    10.120478         NaN           NaN
```

and compare it with the t-test

```
import numpy as np
from scipy import stats

data = pd.read_csv('https://alejandro-isglobal.github.io/SDA/data/leptin.txt',
sep='\t')

groupA = data["Control"]
groupB = data["LepNull"]

stats.ttest_ind(groupA, groupB, equal_var=True, nan_policy='omit')
```

```
Ttest_indResult(statistic=-12.507930602554847, pvalue=3.377211689857233e-11)
```

17.8 Linear model

The following formulation is useful to integrate the different types of analysis. Let's classify the observations of the random variable Y using the group j and the particular observation i. Let's assume that we can separate the **random**

error

$$Y_{ji} = \mu + \alpha_j + E_{ji}$$

- E_{ji} is new **random variable** with expected value $E(E_{ji}) = 0$ and variance $V(E_{ji}) = \sigma^2$

Fixed parameters:

- μ is the overall mean
- α_j is the deviation of group j to the overall mean: $j \in (A, B, C, ...)$ and $\alpha_A = \mu_A - \mu$, $\alpha_B = \mu_B - \mu$, ...

Random error within each group:

Then the expected value of Y in the group j is

$$E(Y|j) = \mu_j = \mu + \alpha_j$$

and the variance of Y in the group j is

$$V(Y|j) = \sigma^2$$

Hypothesis contrast

In this formulation the hypothesis contrast can be simply written as

a. $H_0 : \alpha_j = 0$ for all j. There are no difference between the group means and the overall mean μ.

b. $H_1 : \alpha_j \neq 0$ for at least one j. There is at least one group whose mean is different from the overall mean μ.

This in only a formulation, a useful and popular formulation. We again use the F statistics to decide between the two hypothesis.

17.9 2-way ANOVA

The ANOVA approach allows for the analysis of the joint effect of **two** or more different factors each with different numner of groups.

Let us include an additional factor given by the sex of the mice and ask: Is there an effect of sex when leptin groups are taken into account?

Here there are two simultaneous questions:

- Is there an effect of **sex** on the weight of the mice?
- Is that effect different between leptin groups?

17.10 Data

We will introduce an additional outcome. One random experiment has three outcomes: $(weight, leptin, sex)$.

Continuous variable (outcome of interest)

- $weight \in (20, 60)$

Categorical variable:

- $leptin \in \{KOplus : A, knockout : B\}$

Categorical variable:

- $sex \in \{male : a, female : b\}$

The data looks like

```
##    weight    group sex
## 17  46.57 leptinKO   M
## 18  40.43 leptinKO   M
## 19  41.97 leptinKO   M
## 20  41.17 leptinKO   M
## 21  41.57 leptinKO   M
## 22  46.17 leptinKO   M
## 23  53.83 leptinKO   M
## 24  24.33   KOplus   M
## 25  22.37   KOplus   M
## 26  26.10   KOplus   M
## 27  17.50   KOplus   M
## 28  35.17   KOplus   M
## 29  25.97   KOplus   M
## 30  27.67   KOplus   M
## 31  23.37   KOplus   M
## 32  31.83   KOplus   M
## 33  22.37   KOplus   M
## 58  65.80 leptinKO   F
## 59  51.40 leptinKO   F
## 60  54.60 leptinKO   F
## 61  48.30 leptinKO   F
## 62  50.60 leptinKO   F
## 63  48.90 leptinKO   F
## 64  51.20 leptinKO   F
## 65  46.80 leptinKO   F
## 66  50.90 leptinKO   F
## 67  42.70 leptinKO   F
## 68  28.70   KOplus   F
## 69  25.60   KOplus   F
## 70  26.40   KOplus   F
```

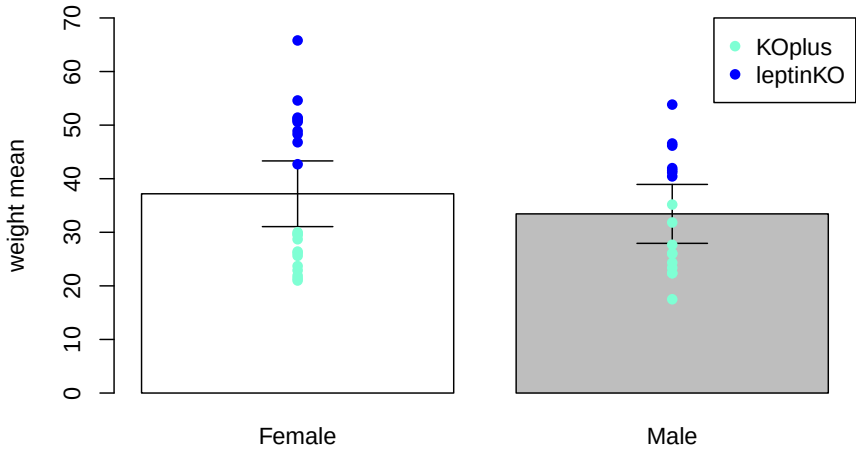
```
## 71  22.90   KOplus   F
## 72  30.00   KOplus   F
## 73  29.70   KOplus   F
## 74  26.10   KOplus   F
## 75  21.40   KOplus   F
## 76  29.50   KOplus   F
## 77  21.90   KOplus   F
## 78  23.70   KOplus   F
## 79  21.00   KOplus   F
```

17.11 Modeling residuals

We see that there is no effect of sex when the leptin groups are ignored. No significant differences in means between sexes is observed

```
import pandas as pd
import statsmodels.api as sm
from statsmodels.formula.api import ols


data = pd.read_csv('https://alejandro-isglobal.github.io/SDA/data/dataleptin.txt',
sep=' ')

filtgroups = (data['group']=='leptinKO') | (data['group']=='KOplus')

filtered_data = data[filtgroups]

model = ols('weight ~ sex', data=filtered_data).fit()

anova_table = sm.stats.anova_lm(model, typ=1)

print(anova_table)
```

```
            df       sum_sq     mean_sq         F    PR(>F)
sex        1.0   134.975026  134.975026  0.854439  0.361289
Residual  37.0  5844.862933  157.969268       NaN       NaN
```



We see however that the points are clustered by leptin group, suggesting that we should adjust by the differences in the lepting before we test the differences in sex. It is possible that when we subtract the differences in lepting the mean weight of males is lower than that of females.

The way we subtract the effect of leptin is to perform an ANVOVA of mice weight on lepting group and then to obtain the residuals of each observations. The residuals are the observed values of the error E_{ij} and are computed as

$$r_{ji} = y_{ji} - \bar{y}_j$$

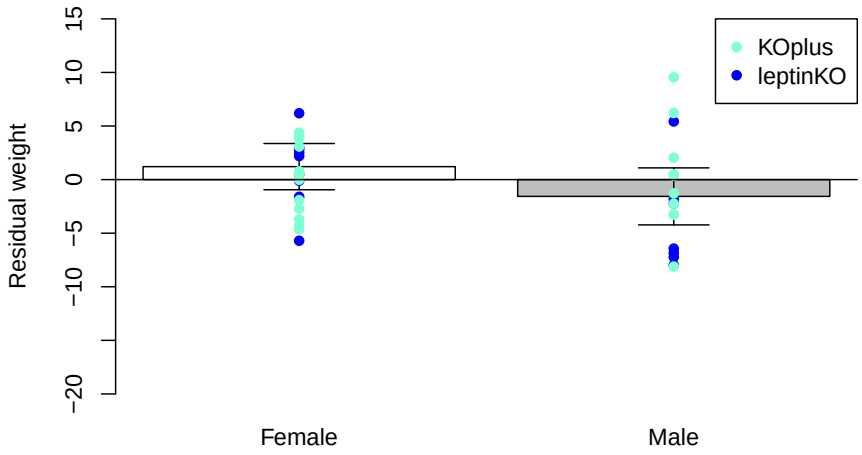
That is, we subtract the mean of the group j to each observation i in leptin group j. In this example we only consider two groups of leptin (KOplus y leptinKO).

Then we can perform another ANOVA test for the residuals on sex differences

```
import pandas as pd
import statsmodels.api as sm
from statsmodels.formula.api import ols

data = pd.read_csv('https://alejandro-isglobal.github.io/SDA/data/dataleptin.txt',
sep=' ')

filtgroups = (data['group']=='leptinKO') | (data['group']=='KOplus')

filtered_data = data[filtgroups]
```

```
model1 = sm.OLS.from_formula('weight ~ group', data=filtered_data).fit()

filtered_data["residuals"] = model1.resid

model2 = ols('residuals ~ sex', data=filtered_data).fit()

anova_table = sm.stats.anova_lm(model2, typ=1)

print(anova_table)
```

```
              sum_sq    df         F    PR(>F)
sex        73.938528   1.0  2.955976  0.093917
Residual  925.489697  37.0       NaN       NaN
```



We see that we have adjusted by the differences in leptin, as the dots are mixed together. The lower weight in males is almost significant for the residuals. However, from this model we do not know the contribution of the leptin group to the variability of the data (points). Also, if n is large, it is possible that even if the null hypothesis is true, the means of each sex do not get closer to the grand mean if there is a significant effect of leptin.

The ANOVA can be used to explain the simultaneous contributions of sex and leptin groups.

17.12 Linear model

The ANOVA approach allows for the simultaneous analysis of two factors (each with many groups or levels).

Let's consider the **linear model**

$$Y_{jri} = \mu + \alpha_j + \beta_r + E_{jri}$$

with

- **grand mean**:

$$E(Y_{jri}) = \mu$$

that is the expected value of all the observations.

- **random error**:

$$E_{jri}$$

that is a **random variable** with mean $E(E_{jri}) = 0$, and variance $V(E_{jri}) = \sigma^2$

- k **deviations** of the group means to the the grand mean for **factor 1** (leptin):

$$\alpha_j = \mu_{j\cdot} - \mu$$

for each group $j \in \{j...k\}$, and $\sum_j \alpha_j = 0$

- m **deviations** of the group means to the the grand mean for **factor 2** (sex):

$$\beta_r = \mu_{\cdot r} - \mu$$

for each group $r \in \{j...m\}$, and $\sum_r \beta_r = 0$

The dot indicates that we ignore the other factor.

The model can also be written as

$$Y_{jri} = \mu_{jr} + E_{jri}$$

where

$$\mu_{jr} = \mu_{j\cdot} + \mu_{\cdot r} - \mu$$

is the mean in the **condition** given by the group j of factor 1 **and** group r of factor 2. Think for instance of the condition *male* and *leptinKO*. Our

leptin example will have four of those groups, corresponding to all possible combinations of the sex and leptin group of mice.

17.13 Hypothesis test

ANOVA allows testing simultaneously for two hypothesis tests

First

- $H_0 : \alpha_1 = \alpha_2, ... = \alpha_k = 0$ There are no difference between group means for the fist factor
- H_1 at least one α_i is different

Second

- $H_0 : \beta_1 = \beta_2, ... = \beta_m$ There are no difference between group means for the second factor
- H_1 at least one β_i is different

17.14 Variance components

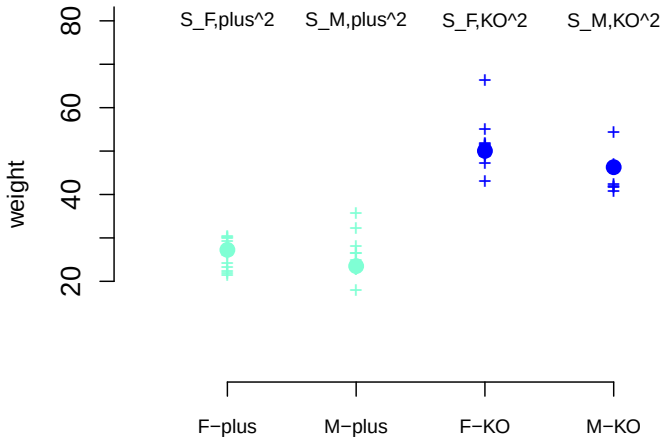
To decide in each hypothesis contrast, we need test statistics. We can again estimate the dispersion of the outcomes from their means in each **condition**, determined by both factors. For instance the mean for knock out males

$$\mu_{M,KO} = \mu_{M\cdot} + \mu_{\cdot KO} - \mu$$

can be estimated by the averages

$$\hat{\mu}_{M,KO} = \bar{Y}_{M\cdot} + \bar{Y}_{\cdot KO} - \bar{Y}$$

and represent the dots in the following plot



The dispersion of the knock-out male observations about the average $\hat{\mu}_{M,KO}$ is given by the sample mean statistic

$$S^2_{M,KO} = \frac{1}{n-1}\sum_{i=1}^{n}(Y_{i,M,KO} - \hat{\mu}_{M,KO})^2$$

If we assume that the variance of the weight in all sex-leptin conditions is σ^2 then we can estimate it with the sample variance

$$Se^2 = \frac{1}{km}\sum_j\sum_r S^2_{jr}$$

which is the sum of the depressions across the sex-leptin conditions. S^2_e is again an estimator for the σ^2.

Now, we can also estimate the variance of the averages in each group of factor 1 about the overall average

$$S^2_{tr1} = \frac{1}{k-1}\sum_{j=1}^{k}(\bar{Y}_{j\cdot} - \bar{Y})^2$$

That is an estimator of σ^2/k under the null hypothesis. Again we define the F statistics

$$F_1 = \frac{kmS^2_{tr1}}{S^2_e} \rightarrow Fisher(k-1, (kmn-1)-(k-1)-(m-1))$$

that will be close to 1 if the first null hypothesis is true: There is no difference between means $\mu_{j\cdot}$ across the groups of factor 1. In this case the sample variance of treatments (groups) for factor 1 will reduce to zero with large n, because all the treatment means of factor 1 are the same as the grand mean ($\mu_{j\cdot} = \mu$). If the alternative is true then at least one treatment mean is different from the grand mean and the statistic will be far from 1 (at least one $\mu_{j\cdot} \neq \mu$).

We can also define another statistic that test the second hypothesis contrast. For factor 2, we can estimate the dispersion of the means of each group of the factor to the overall average, using the statistics

$$S^2_{tr2} = \frac{1}{m-1}\sum_{r=1}^{m}(\bar{Y}_{\cdot r} - \bar{Y})^2$$

and we can define

$$F_2 = \frac{kmS^2_{tr2}}{S^2_e} \rightarrow Fisher(m-1, (kmn-1)-(k-1)-(m-1))$$

Then, observed values of F_2 far from 1 suggest that the means between groups of the second factor significantly **differ**, rejecting the null.

All this is summarized in the two-way ANOVA table

```
import pandas as pd
import statsmodels.api as sm
from statsmodels.formula.api import ols

data = pd.read_csv('https://alejandro-isglobal.github.io/SDA/data/dataleptin.txt',
sep=' ')

filtgroups = (data['group']=='leptinKO') | (data['group']=='KOplus')

filtered_data = data[filtgroups]

model = ols('weight ~ sex + group', data=filtered_data).fit()

anova_table = sm.stats.anova_lm(model, typ=1)

print(anova_table)
```

```
            df       sum_sq      mean_sq           F        PR(>F)
sex        1.0   134.975026   134.975026    5.251072  2.788981e-02
```

```
group      1.0  4919.508806  4919.508806  191.388693  5.593450e-16
Residual  36.0   925.354127    25.704281         NaN           NaN
```

where for example the observed value of F_1 is

$$f_{1obs} = \frac{MST1}{MSE} = \frac{\frac{1}{k-1}SSq_{sex}}{\frac{1}{(kmn-1)-(k-1)-(m-1)}SSq_{Residual}} = 5.2511$$

We can see that this value is large indicating that the null hypothesis shoud be rejected $pvalue = 0.02 < \alpha = 0.05$. Therefore we say that there is a **significant effect** of sex when we take into account the effect of leptin, or correct for leptin differences. From the dispersion plot above, we see that males seem to have lower weights than females, whether they are in the group with supplemented leptin (KOplus) or with out it (leptinKO).

17.15 2-way ANOVA with interaction

In the previous ANOVA model, we computed the means at each sex-lepting condition (j, r) by the contribution of the group j from factor 1 (sex) and group r from factor 2 (leptin) to the overall mean

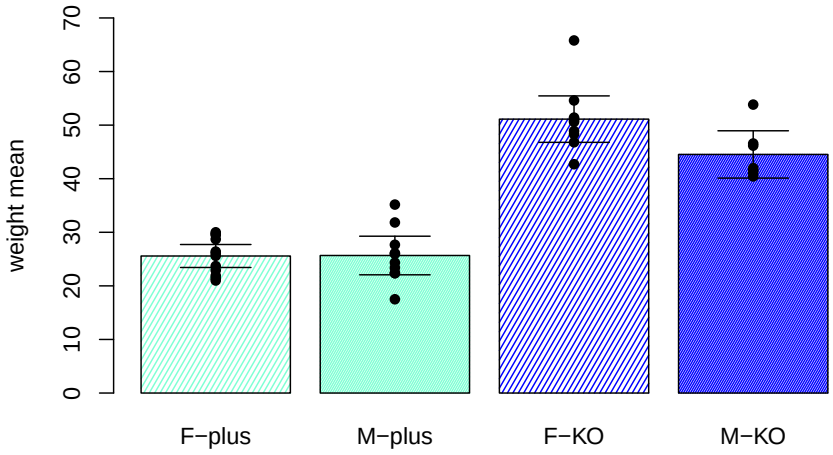
$$\mu_{jr} = \alpha_j + \beta_r + \mu$$

However, we can see that the distribution of the data about these means is not symmetrical because these are not the means computed within each condition.

When we take weights **conditioned to** each leptin by sex group we observed:

- $n_{F,plus} = 12$ supplemented leptin **female** mice had a weight average of $\bar{y}_{F,plus} = 25.575$.
- $n_{M,plus} = 10$ supplemented leptin **male** mice had a weight average of $\bar{y}_{M,plus} = 25.668$.
- $n_{F,KO} = 10$ control **female** mice had a weight average of $\bar{y}_{F,KO} = 51.120$.
- $n_{M,KO} = 7$ leptin KO **male** mice had a weight average of $\bar{y}_{M,KO} = 44.530$.

For these estimations we can draw a barplot with confidence intervals.



From this plot we can see that most of the difference between male and female weights comes from the knock out condition. In the supplemented leptin groups sexes seem the have similar weights. However, we see that the differences between sexes seem to be bigger in the knock-out mice than in the supplemented-with-leptin mice. How can we test the hypothesis that there is a particular combination of conditions that has an effect on the weight of mice? For instance, how can we test that male knock-out seem to be less sensitive in weight gains than female knock-outs.

The apparent less-than-expected gain in weight of leptin KO males seems like a specific interaction between leptin KO and being male.

17.16 Linear model

We then formulate the **linear model** with an **interaction term** that takes into account the specific contribution of the condition given by the groups (j, r) of each factor

$$Y_{jri} = \mu + \alpha_j + \beta_r + (\alpha\beta)_{jr} + E_{jri}$$

Such that each observation in the condition, given by the groups (j, r), has a specific contribution given by the condition

$$E(Y|j, r) = \mu + \alpha_j + \beta_r + (\alpha\beta)_{jr}$$

This is the mean **conditioned** to the groups (j, r). The condition mean can be computed as the mean given the independent contributions of each group (μ_{jr}) plus a specific contribution of the condition $(\alpha\beta)_{jr}$

$$E(Y|jr) = \mu_{jr} + (\alpha\beta)_{jr}$$

17.17 Hypothesis test

This ANOVA model allows testing three null hypothesis

First

- $H_0 : \alpha_1 = \alpha_2, ... = \alpha_k = 0$. There are no difference between group means in the fist factor

Second

- $H_0 : \beta_1 = \beta_2, ... = \beta_m = 0$. There are no difference between group means in second factor

Third

- $H_0 : (\alpha\beta)_{jr} = 0$. There is no difference between specific condition means

And the alternatives are that at least one of the terms is different from 0

17.18 Variance components

The first two hypothesis contrasts can be tested as shown in the previous section. To test the first contrast, we define the F_1 statistic that measures the dispersion of the group means of factor 1 with respect the dispersion of the residuals

$$F_1 = \frac{MS1}{MSE} \rightarrow Fisher(k-1, (n-1)-(k-1)-(m-1))$$

To test the second contrast, we define the corresponding F_2 statistic for factor 2

$$F_2 = \frac{MS2}{MSE} \rightarrow Fisher(m-1, (n-1)-(k-1)-(m-1))$$

Finally, to test the hypothesis contrast for the interaction term, we compute the F_I statistic given by the dispersion of group averages to the overall average with respect the the dispersion of the residuals

$$F_I = \frac{MSI}{MSE} \rightarrow Fisher((k-1)(m-1), (nkm-1)-(k-1)-(m-1)-(k-1)(m-1))$$

These statistics are given in the ANOVA table with interaction

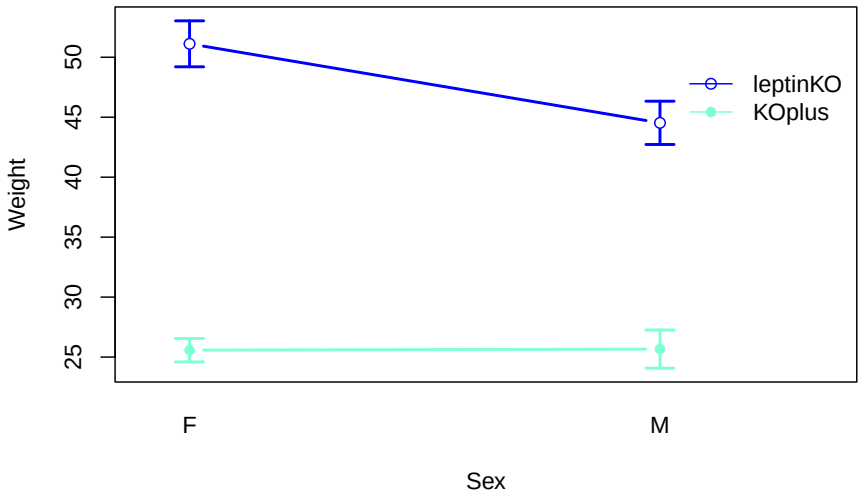
```
## Analysis of Variance Table
##
## Response: weight
##           Df Sum Sq Mean Sq  F value  Pr(>F)
## group      1 4980.4  4980.4 212.4335 < 2e-16 ***
## sex        1   74.1    74.1   3.1595 0.08418 .
## group:sex  1  104.8   104.8   4.4699 0.04169 *
## Residuals 35  820.6    23.4
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
import pandas as pd
import statsmodels.api as sm
from statsmodels.formula.api import ols

data = pd.read_csv('https://alejandro-isglobal.github.io/SDA/data/dataleptin.txt',
sep=' ')

filtgroups = (data['group']=='leptinKO') | (data['group']=='KOplus')

filtered_data = data[filtgroups]

model = ols('weight ~ sex*group', data=filtered_data).fit()

anova_table = sm.stats.anova_lm(model, typ=1)

print(anova_table)
```

```
              df       sum_sq      mean_sq           F        PR(>F)
sex          1.0   134.975026   134.975026    5.757201  2.187963e-02
group        1.0  4919.508806  4919.508806  209.835870  2.367238e-16
sex:group    1.0   104.794667   104.794667    4.469893  4.169367e-02
Residual    35.0   820.559460    23.444556         NaN           NaN
```

As we observed from the bar plots, the statistical inference confirms that there are significant differences in weight between leptin groups, significant differences between sexes, and significant interactions between sex and leptin groups. In particular, the effect of being male is to reduce weight in knock-out mice, in contrast to gain weight in mice with leptin supplmentation. The effect of sex chanes with leptin context.

Interactions are better represented in plots displaying the conditioned means across both factors. Significant interactions are therefore given by non-parallel lines.



17.19 Questions

1) We test ANOVA when we have

a: several continuous random variables; **b:** several categorical variables; **c:** several continuous random variables and several categorical variables; **d:** a continuous random variable and several categorical outcomes;

2) The statistic $F = \frac{MST}{MSE}$ is

a: the estimation of σ^2 by the residuals; **b:** the estimation of σ^2 by the group means; **c:** the estimation of σ^2 by the group means minus by the estimation of σ^2 by the residuals; **d:** the estimation of σ^2 by the group means divided by the estimation of σ^2 by the residuals

3) We accept the null hypothesis that the groups have equal means when the value of $F = \frac{MST}{MSE}$ is

a: close to 1; **b:** different from 1; **c:** small; **d:** large;

4) ANOVA assumes that the observations in each condition have

a: the same mean μ; **b:** the same variance σ^2; **c:** different variances; **d:** different means;

5) The interaction term in a an ANOVA tests that

a: the group means of factor 1 are different from the group means of factor

2 ; **b:** the mean of a group of factor 1 is different from the mean of the other groups of factor 1; **c:** the means of the groups across both factors are not added contributions of each factor; **d:** the means of the groups across both factors are different;

17.20 Practice

Load leptin data (https://alejandro-isglobal.github.io/SDA/data/dataleptin.txt)

- Use a t-test to test the hypothesis that the weight between sexes is different.
- Use an ANOVA to test that the weight between sexes is different.
- Extract residuals from the ANOVA on leptin group and do a second anova on sex.
- Use an ANOVA on sex and group to test the whether sex and leptin groups are significantly associated with weight
- Include an interaction term in the previous ANOVA
- Make a bar plot of weight across all conditions given by both factors

Solutions

Chapter 18

Regression and correlation

18.1 Objective

In this chapter we will see how to test statistical dependence of two continuous random variables. We will introduce the **correlation** as a parameter of a two dimensional normal distribution. We will perform statistical test on the correlation.

We will also see how to test statistical independence when we consider one of the variables as an outcome (dependent) and another as a condition (independent). We will introduce **regression analysis** to test the linear dependence between the variables. Finally, we will talk bout multiple regression, when we want to adjust the associations by other variables.

18.2 Correlations

Leptin is a hormone produced by adipose tissue. We want to study the serum leptin levels in the adult population (PMID: 23628382, data available at GEO:GSE45987) under a **continuous** condition, such as the amount in kilograms of body fat. We assume that

- The levels of leptin in blood have a probability density

$$Y \to N(\mu_y, \sigma_y^2)$$

- The levels of fat mass have a probability density

$$X \to N(\mu_x, \sigma_x^2)$$

18.3 Data

One random experiment in our study has two outcomes: $(leptin, fatmass)$.

Continuous variable (outcome of interest)

- $lepting \in (0, 5)$

Continuous variable (explanatory variable)

- $fatmass \in (20, 80)$

Repeating the experiment n times, the data for the first five repetitions look like

```
##     leptin fatmass
## 1 3.355677  45.721
## 2 2.272126  43.895
## 3 1.071584  47.871
## 4 3.921082  65.801
## 5 1.536867  56.644
## 6 1.177115  56.355
```

In this, study we are interested in finding out whether individuals with higher fat mass will have higher circulating levels of leptin. In statistical terms we are asking whether fat mass and leptin are statistically independent. How can we test this?

We need to formulate a hypothesis test on a parameter of a distribution.

18.4 Normal bivariate

Let's consider that one random experiment of the study consists on drawing a pair of measurements of both leptin levels **and** a fat mass of an individual. We therefore have a random variable that is a pair of measurements that follows a probability distribution. We will assume that the distribution is the **two-dimensional** version of a normal probability density

$$(Y, X) \to N(\mu_y, \sigma_y^2, \mu_x, \sigma_x^2, \rho)$$

This is function has five parameters. More explicitly the function is of the form

$$f(y, x) = \frac{1}{2\pi\sigma_y\sigma_x\sqrt{1-\rho^2}} e^{\frac{(y-\mu_y)^2}{\sigma_y^2} - \frac{2\rho(y-\mu_y)(x-\mu_x)}{\sigma_y\sigma_x} + \frac{(x-\mu_x)^2}{\sigma_x^2}}$$

and the parameters are $\mu_y, \mu_x, \sigma_y^2, \sigma_x^2, \rho$. These are

The **marginal mean** and variance of Y

- $\mu_y = E(Y)$, $\sigma_y^2 = V(Y)$

The **marginal mean** and variance of X

- $\mu_x = E(X)$, $\sigma_x^2 = V(X)$

The **correlation** between X and Y

- $\rho = \frac{E[(Y-\mu_y)(X-\mu_x)]}{\sigma_y \sigma_x}$

ρ is called the **correlation coefficient**.

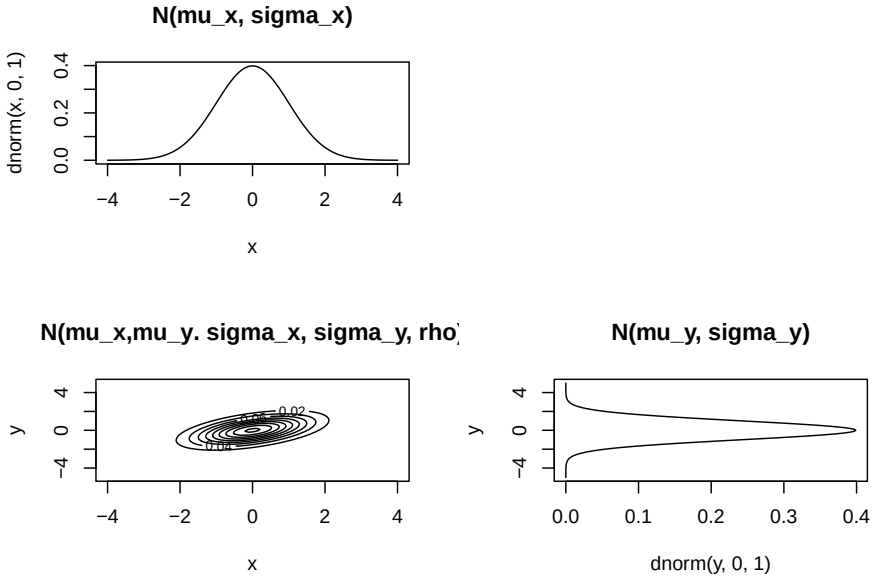
When we draw the probability density in two dimensions, we see that the marginal densities are the projections of the densities on each axis. Let us define the densities for each variable

$$X \to N(\mu_x, \sigma_x^2)$$

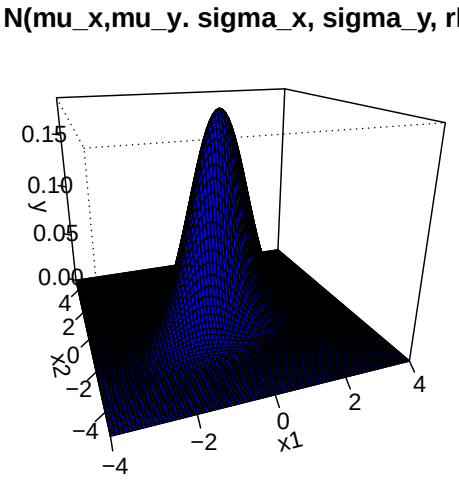
and

$$Y \to N(\mu_y, \sigma_y^2)$$

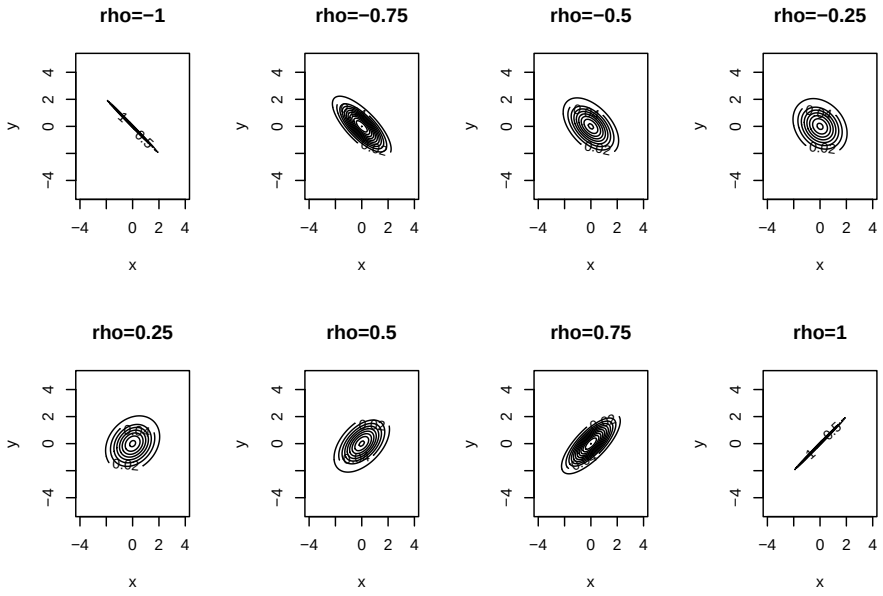
with its parameters. We can plot the 2D density as contour lines and the marginal densities on each axis



We can also plot the density in a 3D plot, where the Z axis is the value of the probability density.



If we keep the contour plots of the 2D distribution, we see that the fifth parameter ρ defines the direction on which the ellipses are elongated (mayor axis).



The density is entirely determined by 5 parameters.

18.5 Estimators

We can derive the estimator of all 5 parameters, if we formulate the likelihood function

$$L = \Pi_{i=1}^{n} f(y_i, x_i; \mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho)$$

and maximize it for each parameter. As a result, we obtain the following estimators of the parameters.

The estimators for the means are the usual averages

- $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ estimates μ_y
- $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$ estimates μ_x

The estimators for the variances are

- $S_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$ estimates σ_y^2
- $S_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ estimates σ_x^2

The estimator for the correlation

- $R = \frac{\sum_{i=0}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$ estimates ρ.

18.6 Correlation coefficient

R is then a **statistics** that can be computed from the data and we can take one value as an estimate of ρ. The sampling distribution of R can be obtained if we transform it into a new variable (Fisher's z transformation)

$$Z = \frac{1}{2}\ln(\frac{1+R}{1-R})$$

The new variable Z is a normal variable

$$Z \rightarrow_{aprox} N(\frac{1}{2}\ln(\frac{1+\rho}{1-\rho}), \frac{1}{n-3})$$

with mean $\frac{1}{2}\ln(\frac{1+\rho}{1-\rho})$ and variance $\frac{1}{n-3}$.

As R estimates ρ we have that

- If R is near 0 then there is no linear relationship between y and x (the mayor and minor axes of the 2D plot are the x and y axes).
- If R is near 1 there is strong evidence of a linear relationship between y and x (the mayor axis does not align with neither the x or the y axis).

R is then a measure of the statistical dependence between Y and X. We can use it to test that hypothesis.

18.7 Hypothesis contrast

The hypothesis contrast for the independence on the relationship between Y and X can be formulated as

a. $H_0 : \rho = 0$ (null hypothesis). Therefore, Y and X are consider statistically **independent** and consequently $f(y, x) = f(x)f(y)$ (the joint probability is the product of the marginals)

b. $H_1 : \rho \neq 0$ (alternative hypothesis). Therefore, Y and X are statistically **dependent**.

We then use the statistic R to test whether there is a dependency between y and x. For testing the hypothesis contrast, we take the observed value of R

$$r_{obs} = \hat{\rho} = \frac{\sum_{i=0}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

and compute its *pvalue*. That is the probability that we obtain a larger value if we repeat the sample when the **null hypothesis** is true. Under the null hypothesis $\rho = 0$, we then we compute the two-tailed *pvalue*

$$pvalue = 2(1 - F(|r_{obs}|) = 2(1 - F_{normal}(|z_{obs}|)$$

where

$$Z \rightarrow_{aprox} N(0, \frac{1}{n-3})$$

Example (leptin and fat mass)

```
import pandas as pd
from scipy.stats import pearsonr

data = pd.read_csv('https://alejandro-isglobal.github.io/SDA/data/leptinFatmass.txt',
sep=' ')

leptin = data['leptin']
fatmass = data['fatmass']

pearsonr(leptin, fatmass)
```

```
PearsonRResult(statistic=-0.27664924595293827, pvalue=0.00012143760304939414)
```

For our data, we have that the observed correlation coefficient is $r_{obs} = -0.2766492$ and its transformed value

$$z_{obs} = ln((1 - 0.2766492)/(1 + 0.2766492))/2 = -0.2840499$$

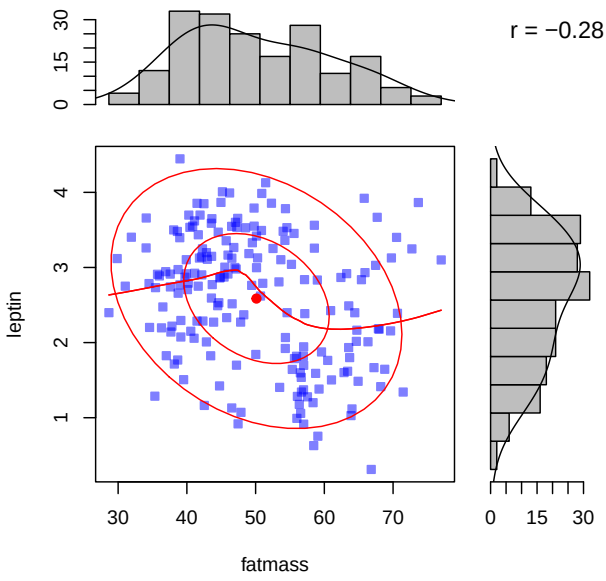
The *pvalue* is therefore

$$pvalue = 2(1 - F_{N(0,1/185)}(0.2840499)) = 0.0001$$

Therefore, since it is lower than $\alpha = 0.05$, we reject the null hypothesis that leptin and fat mass are independent. Note that leptin and fat mass are **weakly correlated** because $r = -0.28$ but the correlation is **highly significant** because $pvalue = 0.0001$.

We should take this result with caution:

- Looking at the data, we see that the fitted 2D probability function and we see that the **marginal** histograms are not quite normally distributed.
- From literature, we know that as we increase fat mass, we should obtain **more** leptin, not the contrary, as it is released from adipose tissue.



How can we improve our analysis, so that our results are as expected?

18.8 Regression analysis

To determine if *leptin* and $fatmass$ are statistically independent, we can rather ask: What is the probability density of leptin at a **given value** of fat mass.

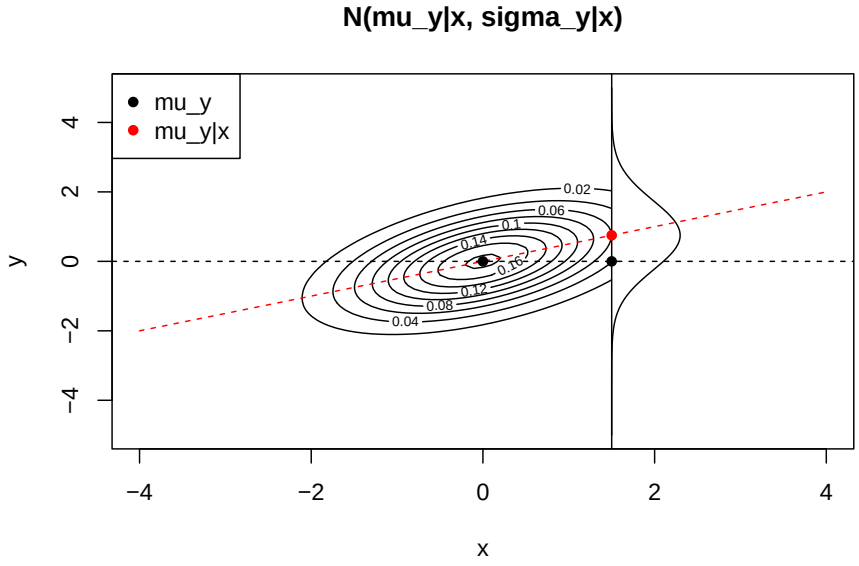
Remember, two variables Y and X are independent if

$$P(Y|X) = P(Y)$$

Therefore, we first calculate the **conditional probability density** of Y (leptin) given X (fat mass) from the definition

$$f(y|x) = \frac{f(y,x)}{f(y)} = N(\mu_{y|x}, \sigma^2_{y|x})$$

This turns up to be a normal distribution. The conditional probability is the profile, or a slice, of the 2D distribution at a given value of x.



The mean of the conditional distribution is (the red dot)

$$\mu_{y|x} = \mu_y + \frac{\sigma_y \rho}{\sigma_x}(x - \mu_x)$$

If we manipulate this equation, we see that the **conditional mean** is a linear function of x

$$\mu_{y|x} = (\mu_y - \mu_x \frac{\sigma_y \rho}{\sigma_x}) + \frac{\sigma_y \rho}{\sigma_x} x = \alpha + \beta x$$

where $\alpha = \mu_y - \mu_x\beta$ and $\beta = \frac{\sigma_y\rho}{\sigma_x}$. This line is called a **regression** line (red dotted line in the plot). That is, when we move along the X axis, the conditional mean of Y moves linearly.

The variance of the conditional density is

$$\sigma^2_{y|x} = \sigma^2_y(1-\rho^2)$$

Solving for ρ^2, we have

$$\rho^2 = \frac{\sigma^2_y - \sigma^2_{y|x}}{\sigma^2_y}$$

ρ^2 is the proportion of the total variance that is explained by the regression line. For instance, $\rho^2 = 1$ when the conditional variability $\sigma^2_{y|x}$ is zero; that is, all the observations fall on the regression line.

18.9 Linear model

Consider the **linear model** for the values of Y **conditioned** to the value of x_i

$$Y_{x_i} = \alpha + \beta x_i + E_i$$

with **Random error**:

- E_i is a **random variable** with expected value $E(E_i) = 0$ and variance $V(E_i) = \sigma^2_{y|x}$
- i is the index of the observation: $1...n$ (typically one for every x_i as x_i is continuous)

and the expected value of Y_{x_i} is the regression line

$$\mu_{y|x_i} = \alpha + \beta x_i$$

with

$$\alpha = \mu_y - \beta\mu_x$$

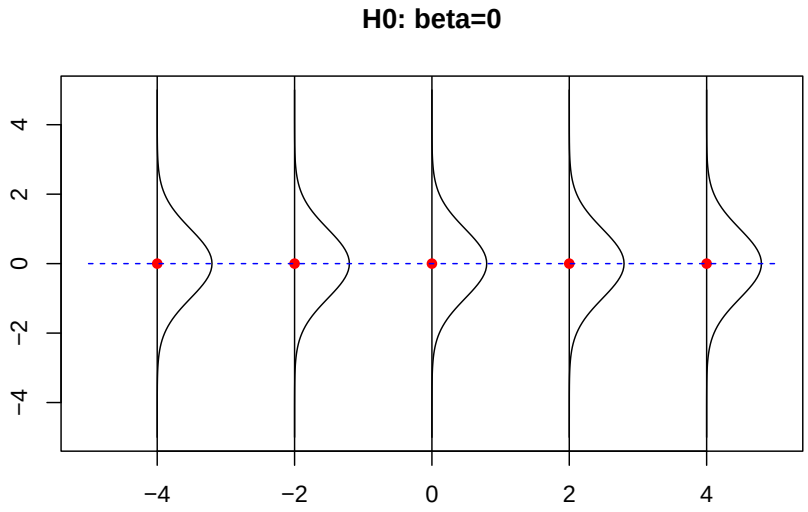
and

$$\beta = \rho\frac{\sigma_y}{\sigma_x}$$

18.10 Hypothesis contrast

Null hypothesis:

a. $H_0 : \beta = 0$. Y and X are statistically **independent**, and therefore $f(y|x) = f(y)$. If this is the case then the regression line is flat and we have the following model

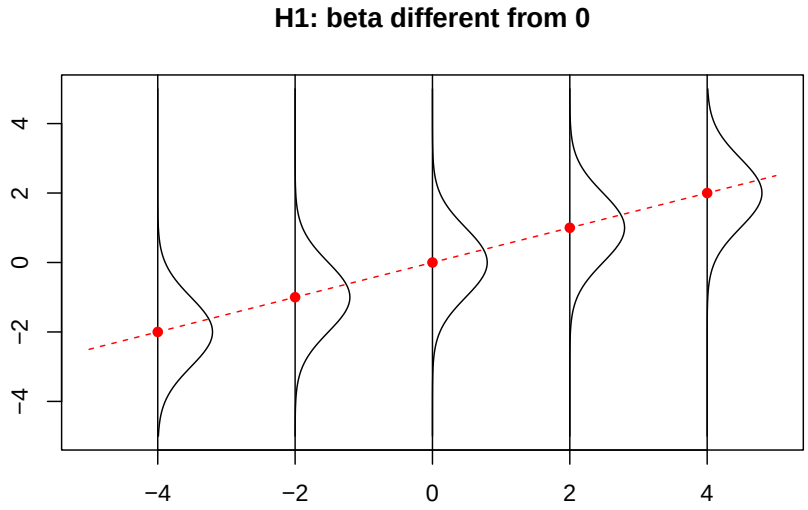


Note that all marginal probability (projection of the probability on the y axis) would be identical to all the conditional probabilities.

Alternative hypothesis:

b. $H_1 : \beta \neq 0$. Y and X are statistically **dependent**. If this is the case then we have the following model

Note that all marginal probability would be the combination of the conditional probabilities.



18.11 Estimators

β is therefore our parameter of interest and we need a probability distribution to test its hypothesis contrast

- $\beta = \rho\frac{\sigma_y}{\sigma_x}$ suggest the estimator for β

$$B = \frac{\sum_{i=1}^{m}(x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- $\alpha = \mu_y - \beta\mu_x$ suggests the estimator for α

$$A = \bar{Y} - \hat{\beta}\bar{x}$$

The estimators A and B for α and β can formally be derived from **minimizing the sum of squares** for the error given x_i

$$SSE = \sum_{i=1}^{n}(Y_i - \bar{Y_i})^2 = \sum_{i=1}^{n}(Y_i - \alpha + \beta x_i)^2$$

with respect to α and β. If $\bar{Y_i}$ is normal then

$$B \to N(\beta, \frac{n\sigma_y^2}{(n-2)s_x^2})$$

with mean $E(B) = \beta$ and, therefore, it is an unbiased estimator.

18.12 Hypothesis testing

Under the null hypothesis, $\beta = 0$ and the standardized error from the null hypothesis are

$$\frac{\hat{\beta}}{\sqrt{\frac{ns_y^2}{(n-2)s_x^2}}} \to T(n-2)$$

follows a t-distribution with $n-2$ degrees of freedom.

Example (leptin and fatmass)

We can perform the statistical test

```
import pandas as pd
from scipy.stats import pearsonr

data = pd.read_csv('https://alejandro-isglobal.github.io/SDA/data/leptinFatmass.txt',
sep='\t')

model = ols('leptin ~ fatmass', data=data).fit()

print(model.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                 leptin   R-squared:                       0.077
Model:                            OLS   Adj. R-squared:                  0.072
Method:                 Least Squares   F-statistic:                     15.42
Date:                Mon, 04 Sep 2023   Prob (F-statistic):           0.000121
Time:                        09:33:40   Log-Likelihood:                -231.59
No. Observations:                 188   AIC:                             467.2
Df Residuals:                     186   BIC:                             473.7
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      3.7201      0.295     12.604      0.000       3.138       4.302
fatmass       -0.0226      0.006     -3.926      0.000      -0.034      -0.011
==============================================================================
Omnibus:                       10.660   Durbin-Watson:                   1.753
Prob(Omnibus):                  0.005   Jarque-Bera (JB):                4.983
Skew:                          -0.141   Prob(JB):                       0.0828
Kurtosis:                       2.254   Cond. No.                         249.
```

==========

where the observed value of B is

$$\beta_{obs} = \frac{\sum_{i=1}^{m}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = -0.02262$$

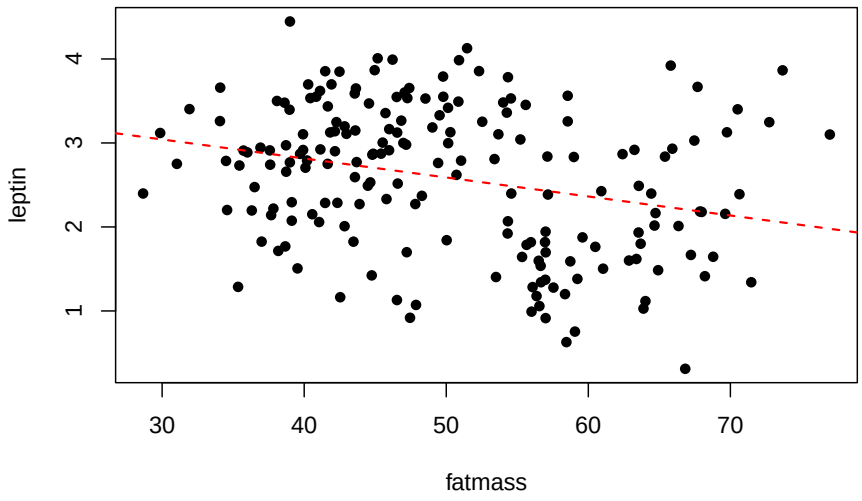
and

$$\alpha_{obs} = \bar{y} - \beta_{obs}\bar{x} = 3.72012$$

and only 7% of the variability is explained by the regression line

$$R^2 = 0.07653$$

We can draw the regression line



We can see that the regression line is negative and there is quite a bit of dispersion about it. Remember that the correct way to interpret the line is to think that these are the values of Y **conditioned** on X.

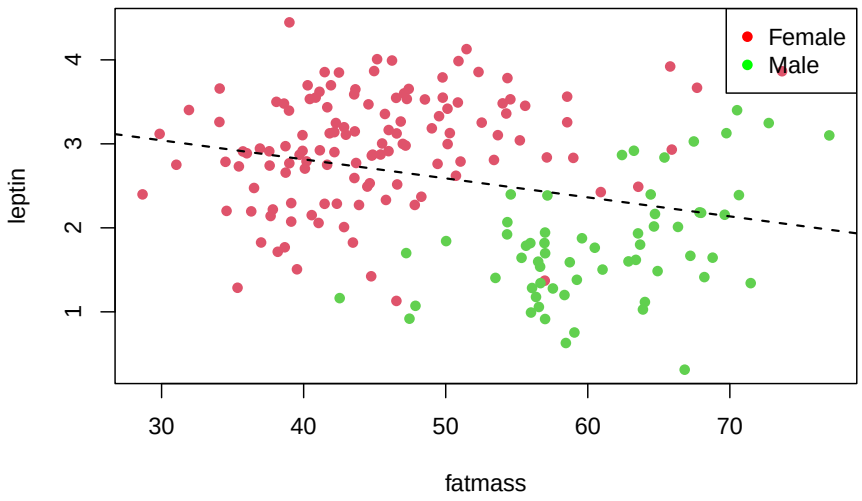
18.13 Stratified analysis

We can include other conditions in the regression, such as sex or age. In general, it is important **to adjust** for other factors that we believe are correlated with the outcome Y and the condition x, as they may explain part of the association.

Our study then includes the outcome Y with multiple regressors, or explanatory variables. For instance, the first observations of the complete dataset look like

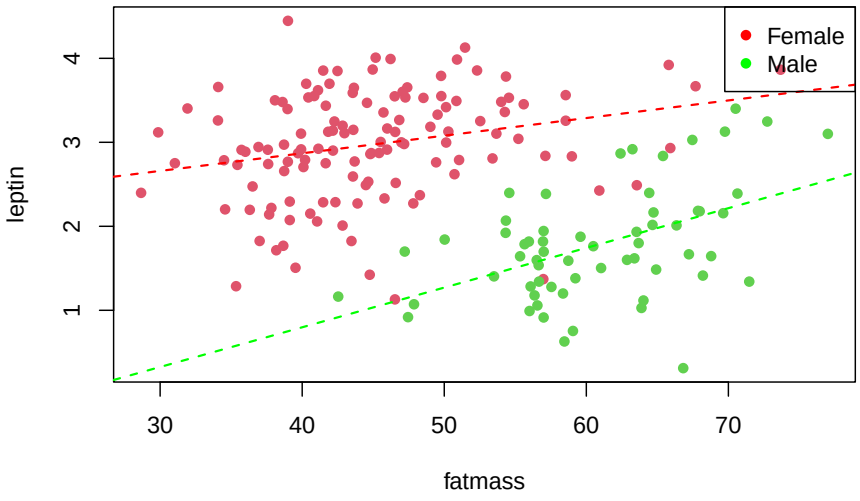
```
##     leptin fatmass sex age
## 1 3.355677  45.721   F  45
## 2 2.272126  43.895   F  77
## 3 1.071584  47.871   M  79
## 4 3.921082  65.801   F  58
## 5 1.536867  56.644   M  42
## 6 1.177115  56.355   M  75
```

Consider the previous regression and color the points according to their sex



We clearly see that the negative association between leptin and fat mass is given by the **effect of sex**. As males have lower leptin levels and higher fat mass then the negative correlation between leptin and fat mass is due to sex.

If we run regression **separating**, or stratifying, by sex we find a positive association between leptin and fat mass, as it is expected from the literature.



18.14 Multiple Regression

The linear regression can be extended to include several number of factors. Consider the **linear model** with one additional factor

$$Y_{jr} = \alpha + \beta x_j + \gamma z_r + E_{jr}$$

Here, our effect of interest is the coefficient β of x_i (leptin) and we want to adjust for the effect of z_j (sex). The mode is similar to the linear model we used for ANOVA, only that now the treatment effects are continous and depend on the variable X.

When we include both sex and age as additional factors in the regression, we observe that there is a positive increase in leptin when we correct for sex. Males have a negative assotiation with leptin, compared with women, and age has an almost significant association with leptin. Now, everything fits.

```
import pandas as pd
from scipy.stats import pearsonr

data = pd.read_csv('https://alejandro-isglobal.github.io/SDA/data/leptinFatmass.txt',
sep='\t')

model = ols('leptin ~ fatmass + sex + age', data=data).fit()
```

```
print(model.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                 leptin   R-squared:                       0.491
Model:                            OLS   Adj. R-squared:                  0.482
Method:                 Least Squares   F-statistic:                     59.09
Date:                Mon, 04 Sep 2023   Prob (F-statistic):           8.47e-27
Time:                        09:40:31   Log-Likelihood:                -175.66
No. Observations:                 188   AIC:                             359.3
Df Residuals:                     184   BIC:                             372.3
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      1.4826      0.304      4.870      0.000       0.882       2.083
sex[T.M]      -1.6120      0.136    -11.864      0.000      -1.880      -1.344
fatmass        0.0276      0.006      4.596      0.000       0.016       0.039
age            0.0054      0.003      1.811      0.072      -0.000       0.011
==============================================================================
Omnibus:                        4.457   Durbin-Watson:                   1.685
Prob(Omnibus):                  0.108   Jarque-Bera (JB):                4.067
Skew:                          -0.345   Prob(JB):                        0.131
Kurtosis:                       3.207   Cond. No.                         475.
==============================================================================
```

18.15 Multiple Regression interaction

We can furthermore include interactions between conditions in the regression.

Consider the **linear model**

$$Y_{jr} = \alpha + \beta x_j + \gamma z_r + \delta (xz)_{jr} + E_{jr}$$

The parameter δ will add a contribution to the slope β that is **specific** to the condition r. For instance, if $z_i \in (0, 1)$ then when $z = 0$ the coefficient of x_i is β. But when $z = 1$ the coefficient of x_i is $\beta + \gamma$.

In our example, if z is sex, then γ will test the differences in βs between males and females. We can furhter adjust by age.

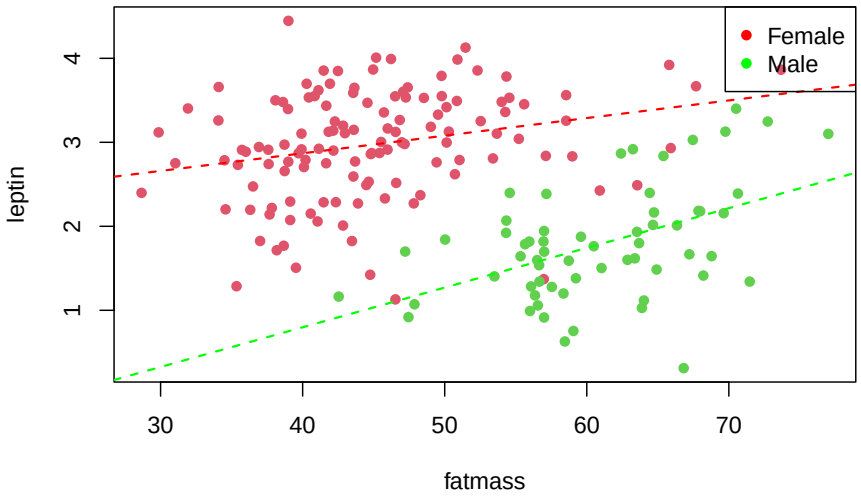
$$Y_{jrs} = \alpha + \beta \times fatmass_j + \gamma \times sex_r + \delta \times (fatmass \times sex)_{jr} + \epsilon_s age + E_{jrs}$$

```
import pandas as pd
from scipy.stats import pearsonr
```

```
data = pd.read_csv('https://alejandro-isglobal.github.io/SDA/data/leptinFatmass.txt',
sep='\t')

model = ols('leptin ~ fatmass*sex + age', data=data).fit()

print(model.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                 leptin   R-squared:                       0.503
Model:                            OLS   Adj. R-squared:                  0.492
Method:                 Least Squares   F-statistic:                     46.25
Date:                Mon, 04 Sep 2023   Prob (F-statistic):           8.19e-27
Time:                        09:43:38   Log-Likelihood:                -173.41
No. Observations:                 188   AIC:                             356.8
Df Residuals:                     183   BIC:                             373.0
Df Model:                           4
Covariance Type:            nonrobust
================================================================================
                   coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
Intercept        1.8001      0.337      5.337      0.000       1.135       2.466
sex[T.M]        -3.2180      0.775     -4.151      0.000      -4.748      -1.688
fatmass          0.0200      0.007      2.881      0.004       0.006       0.034
fatmass:sex[T.M] 0.0284      0.014      2.104      0.037       0.002       0.055
age              0.0058      0.003      1.989      0.048    4.81e-05       0.012
==============================================================================
Omnibus:                        7.164   Durbin-Watson:                   1.699
Prob(Omnibus):                  0.028   Jarque-Bera (JB):                6.943
Skew:                          -0.459   Prob(JB):                       0.0311
Kurtosis:                       3.209   Cond. No.                     1.27e+03
==============================================================================
```

Our data suggest a steeper increase of leptin with body fat in males than in females (interaction: 0.028427, $pvalue = 0.03$), as we saw in the figures of the stratified analysis. A positive and significant interaction means that the slope for males is higher than the slope for females.



The model now explains 50% of the variance $R^2 = 0.5$

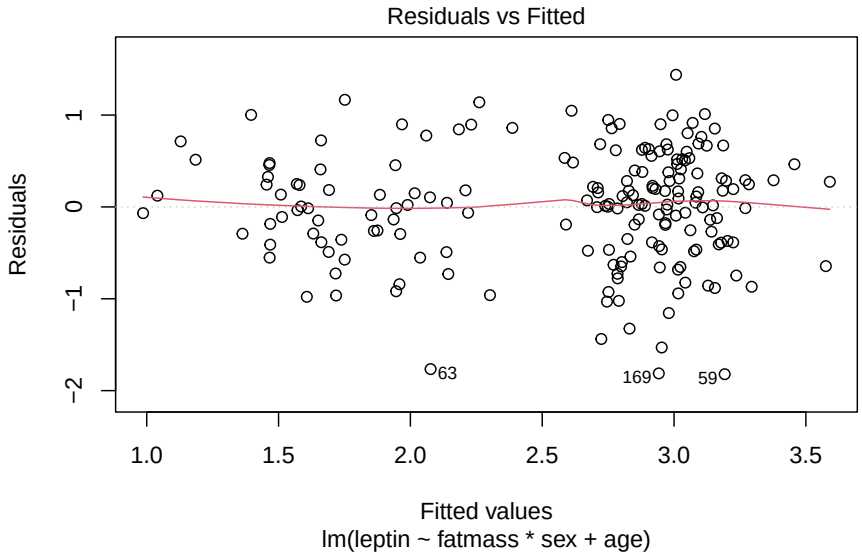
18.16 Model diagnostics

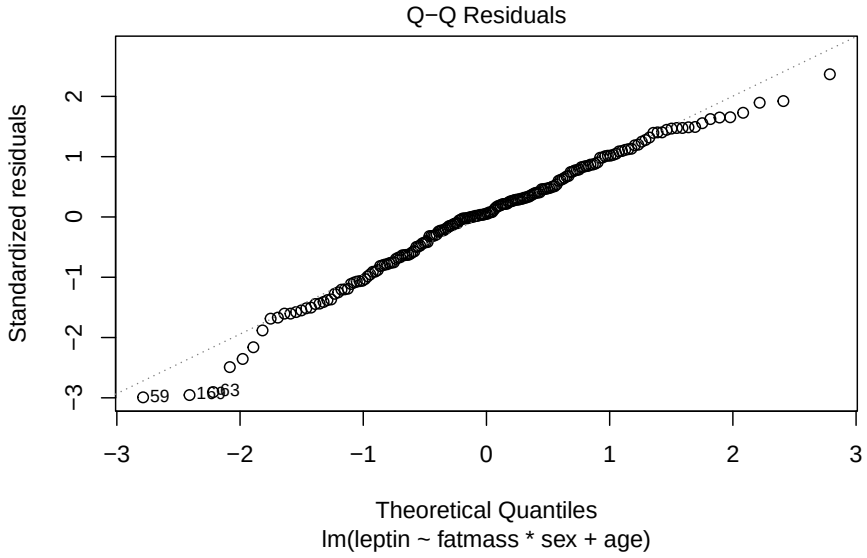
All linear models have been made on the supposition that

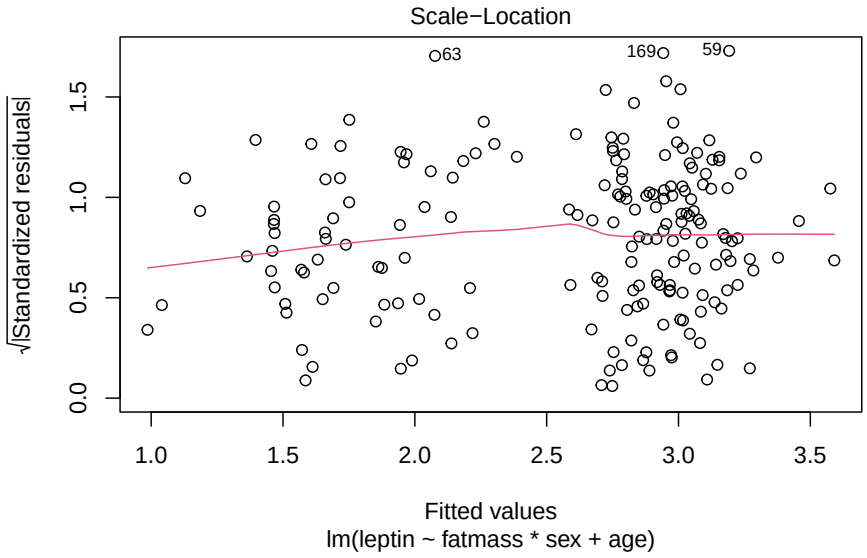
1. Errors are distributed normally

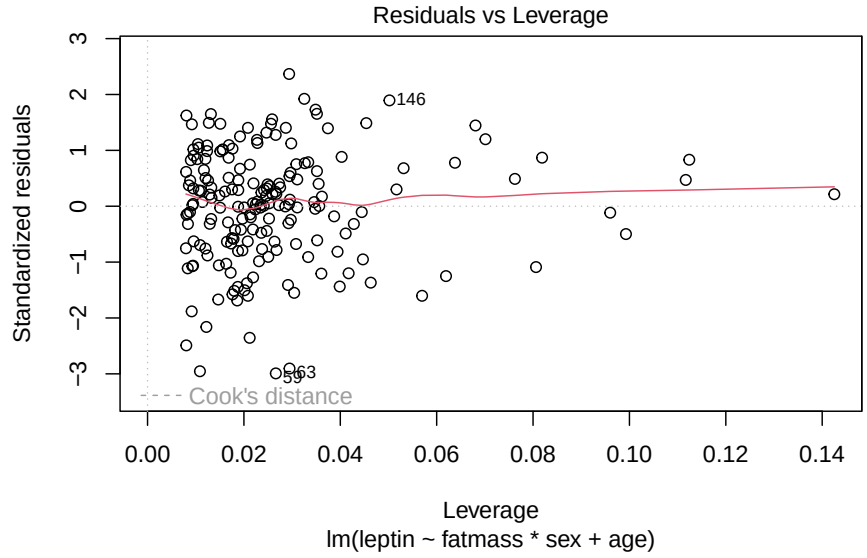
2. Errors have the same variance

There are a number of plots to check that at least the data is consistent with these suppositions. In particular, we are interested to see that the residuals distribute symmetrically against the fitted values, revealing equal variance (homoscedasticity). We also want to check that the points in a quantile-quantile (QQ) plot fall in the line, revealing how close the rediduals distribute normally.


Residuals vs Fitted
Residuals
1
0
−1
−2
63
169
59
1.0
1.5
2.0
2.5
3.0
3.5
Fitted values
lm(leptin ~ fatmass * sex + age)


Q−Q Residuals
Standardized residuals
2
1
0
−1
−2
−3
59
169
63
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(leptin ~ fatmass * sex + age)





18.17 Questions

1) We perform a regression analysis when we have

a: one categorical variable; **b:** two categorical variables; **c:** one continuous random variable; **d:** two continuous random variables;

2) The correlation coefficient ρ is

a: a parameter of a 2D probability density; **b:** a statistic of the relationship between two continuous variables; **c:** the linear coefficient btween of two variables; **d:** the variance explained by one continuous variable on other;

3) β in the linear model

$$Y_{x_i} = \alpha + \beta x_i + E_i$$

a: is a statistic that measure the linear relationship between x and y; **b:** is zero for the null hypothesis; **c:** is a parameter with expected value of zero; **d:** is the correlation between x and y

4) Why do we adjust for a variable in a regression coefficient?

a: to test the interaction between the variable and x; **b:** to stratify the relationship between x and y; **c:** to remove confounding in the relationship between x and y; **d:** to improve the significance of the relationship between x and y;

5) The regression analysis assumes that

a: the errors have mean 0 and equal variances; **b:** the errors distribute normally with mean 0 and equal variances; **c:** the errors distribute normally with different mean and equal variances; **d:** the errors distribute normally with different mean and different variances;

18.18 Practice

18.18.0.1 Practice

Load misophonia data https://alejandro-isglobal.github.io/SDA/data/data_0.txt

We have four measures of anxiety:

- Trait: ansiedad.rasgo (are you an anxious person?) continuous:0-100
- State: ansiedad.estado (are you currently feeling anxious?) continuous:0-100
- Diagnosed: ansiedad.medicada (have you been diagnosed with an anxiety disorder?) binary (si, no)
- Excess: ansiedad.dif (difference between State and Trait)

We formulate the following hypothesis:

Participants who enrolled in the study had an increased level of anxiety from their baseline (trait) that is related to their:

- age
- sex
- anxiety state.

We are interested in the variable anxiety.dif, that is the observed **excess** of anxiety from the trait

$excess = state - trait$

Answer the following questions:

1. Are the state and trait of anxiety correlated?
2. Is excess in anxiety higher in older people?
3. Is excess in anxiety higher in older people after adjusting by sex?
4. Is the interaction between age and sex significant on excess anxiety?

Solutions

Chapter 19

Group Work sessions

19.1 Objectives

- The objective of the work sessions is to work together with a student of a **different background** to perform a full analysis of the **misophonia dataset**.
- The analysis is **open**. You can formulate the analysis you consider interesting, trying to cover as much as possible the material we have seen in theory and bootcamps.
- **Justify** your analysis and **discuss** them.
- We will have **two sessions** to perform the report that will be done in colab and handed in through **google classroom**.
- *Work together, follow your interests and have fun!*

Next, we **describe** the data and show an **example** of the kind of analysis that can be performed in both group sessions.

19.2 Misophonia dataset

Misophonia is a recently described neurological condition whereby patients feel strong anxiety when hearing particular noises (someone blowing their nose, mobile ringing, trains passing, etc..). It is believed that 5% of the population suffers from this condition without knowing it, likely blaming their anxiety on other causes.

The misophonia dataset is from a recent (unpublished) study that aimed to describe the relationships between misophonia and anxiety, depression, and cephalometric measures (shape of the jaw).

```
##   Misofonia Misofonia.dic     Estado Estado.dic ansiedad.rasgo
## 1        si             4 divorciado          2             99
## 2        si             2     casado          1             75
## 3        no             0 divorciado          2             77
## 4        si             3     casado          1             95
## 5        no             0     casado          1             30
## 6        no             0     casado          1             30
##   ansiedad.rasgo.dic ansiedad.estado ansiedad.estado.dic ansiedad.medicada
## 1                  1              99                   1                no
## 2                  1              75                   1                no
## 3                  1              55                   0                no
## 4                  1              99                   1                no
## 5                  0              40                   0                no
## 6                  0              30                   0                no
##   ansiedad.medicada.dic depresion depresion.dic Sexo Edad CLASE
## 1                     0     33.65             1    M   44   III
## 2                     0     19.77             0    M   43    II
## 3                     0     29.57             0    M   24     I
## 4                     0      1.40             0    M   33   III
## 5                     0      5.98             0    H   41     I
## 6                     0     13.87             0    H   35     I
##   Angulo_convexidad protusion.mandibular Angulo_cuelloYtercio Subnasal_H
## 1              7.97                 13.0                 89.6        1.5
## 2             18.23                 -5.0                107.2        7.3
## 3             12.27                 11.5                101.4        5.0
## 4              7.81                 16.8                 75.3        2.7
## 5              9.81                 33.0                105.5        6.0
## 6             13.50                  2.0                105.0        7.0
##   cambio.autoconcepto Misofonia.post Misofonia.pre ansiedad.dif
## 1                   1             21            14            0
## 2                   0             14            13            0
## 3                  NA             NA            NA          -22
## 4                   1             NA            NA            4
## 5                  NA             NA            NA           10
## 6                  NA             NA            NA            0
```

Here is the description of the variables

[1] “Misofonia”: Binary (si: misophinic, no: no misophinic)
[2] “Misofonia.dic”: Categorical (0: no misophinic, 1: severity 1, 2: severity 2, 3: severity 3, 4: severity 4)
[3] “Estado”: Marital status (casado: married, soltero: single, viuda: widow, divorciado:divorced)
[4] “Estado.dic”: Numeric Marital status
[5] “ansiedad.rasgo”: Score from 0-100 with anxiety personality trait
[6] “ansiedad.rasgo.dic”: Binary score (0,1) of anxiety personality trait
[7] “ansiedad.estado”: Score from 0-100 with current state of anxiety

[8] “ansiedad.estado.dic”: Binary score (0,1) with current state of anxiety
[9] “ansiedad.medicada”: Diagnosed with anxiety disorder (si, no)
[10] “ansiedad.medicada.dic”: Diagnosed with anxiety disorder (1, 0)
[11] “depresion”: Score from 0-50 with current state of depression
[12] “depresion.dic” : Binary score (0,1) with current state of depression
[13] “Sexo”: Male=H, Female:M
[14] “Edad”: Age
[15] “CLASE”: Type of jaw
[16] “Angulo_convexidad”: convexity angle
[17] “protusion.mandibular”: Projection of the jaw [18] “Angulo_cuelloYtercio”: angle between jaw and neck [19] “Subnasal_H”: Nasal angle
[20] “cambio.autoconcepto”: Whether people changed their self-concept after treatment.
[21] “Misofonia.post”: Misophionia diagnosed (A-MISO) after an educational program, where patients were made aware of a condition called misophonia.
[22] “Misofonia.pre”: Misophionia diagnosed (A-MISO) before an educational program, where patients were made aware of a condition called misophonia
[23] “ansiedad.dif”: Difference between anxiety state and anxiety trait scores

19.3 Group Work session 1: Data description

When reporting the results of a study, we first describe the variables of interest in tables and figures.

- We describe demographics (sex, age, marital status, etc..)
- We describe outcome variables (misophonia)
- We describe explanatory variables (cephalometric measures, anxiety, depression)

Example:

Imagine we want to study the anxiety of participants in the misophonia study

We load the data

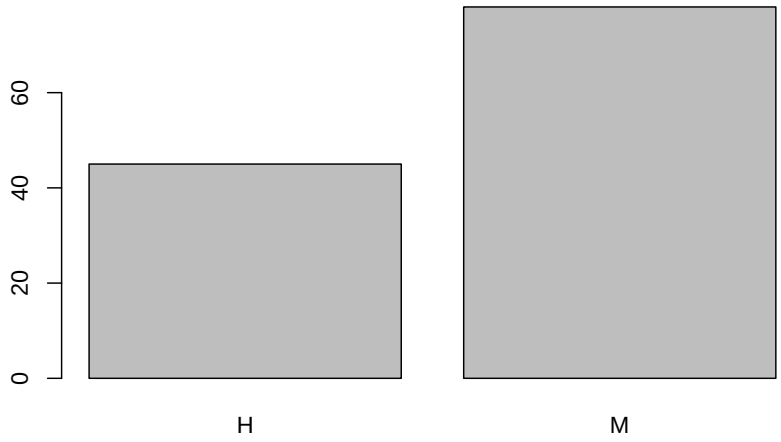
```
##   Misofonia Misofonia.dic     Estado Estado.dic ansiedad.rasgo
## 1        si             4 divorciado          2             99
## 2        si             2     casado          1             75
## 3        no             0 divorciado          2             77
## 4        si             3     casado          1             95
## 5        no             0     casado          1             30
## 6        no             0     casado          1             30
##   ansiedad.rasgo.dic ansiedad.estado ansiedad.estado.dic ansiedad.medicada
## 1                  1              99                   1                no
## 2                  1              75                   1                no
## 3                  1              55                   0                no
## 4                  1              99                   1                no
```

```
## 5                   0             40                   0                no
## 6                   0             30                   0                no
##   ansiedad.medicada.dic depresion depresion.dic Sexo Edad CLASE
## 1                     0     33.65             1    M   44   III
## 2                     0     19.77             0    M   43    II
## 3                     0     29.57             0    M   24     I
## 4                     0      1.40             0    M   33   III
## 5                     0      5.98             0    H   41     I
## 6                     0     13.87             0    H   35     I
##   Angulo_convexidad protusion.mandibular Angulo_cuelloYtercio Subnasal_H
## 1              7.97                 13.0                 89.6        1.5
## 2             18.23                 -5.0                107.2        7.3
## 3             12.27                 11.5                101.4        5.0
## 4              7.81                 16.8                 75.3        2.7
## 5              9.81                 33.0                105.5        6.0
## 6             13.50                  2.0                105.0        7.0
##   cambio.autoconcepto Misofonia.post Misofonia.pre ansiedad.dif
## 1                   1             21            14            0
## 2                   0             14            13            0
## 3                  NA             NA            NA          -22
## 4                   1             NA            NA            4
## 5                  NA             NA            NA           10
## 6                  NA             NA            NA            0
```

1. We describe the participants' sex, age, and marital status

   a. Sex

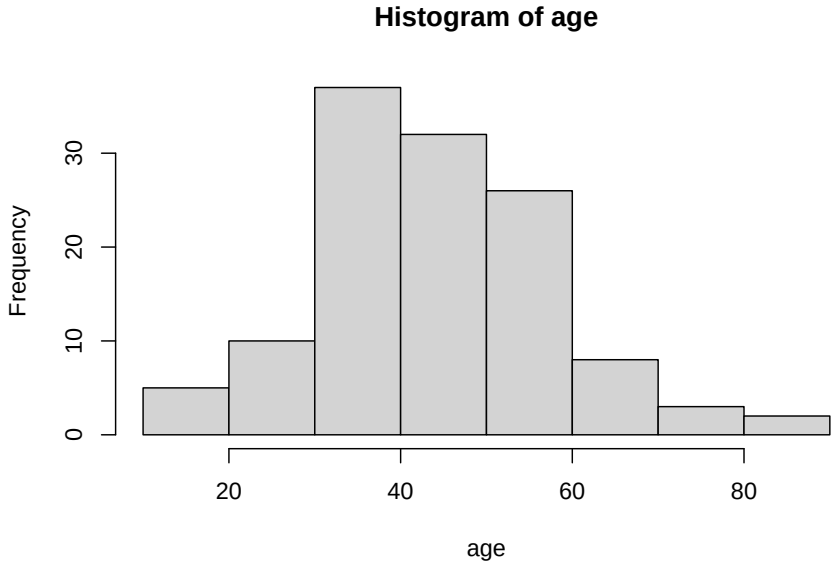
```
## sex
##         H         M
## 0.3658537 0.6341463
```

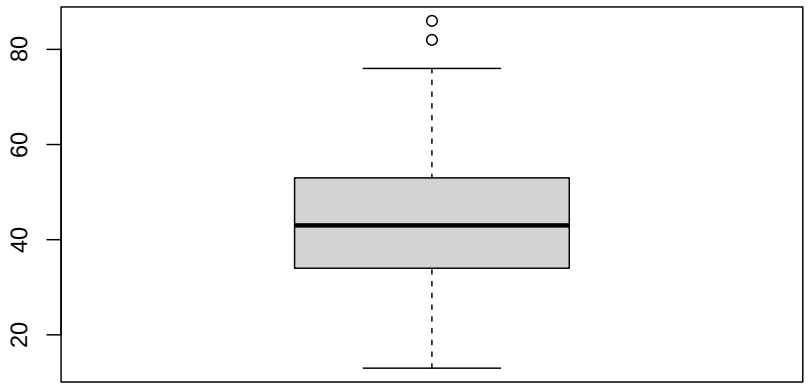


b. Age, mean and standard deviation

```
## [1] 43.93496
```

```
## [1] 14.18654
```





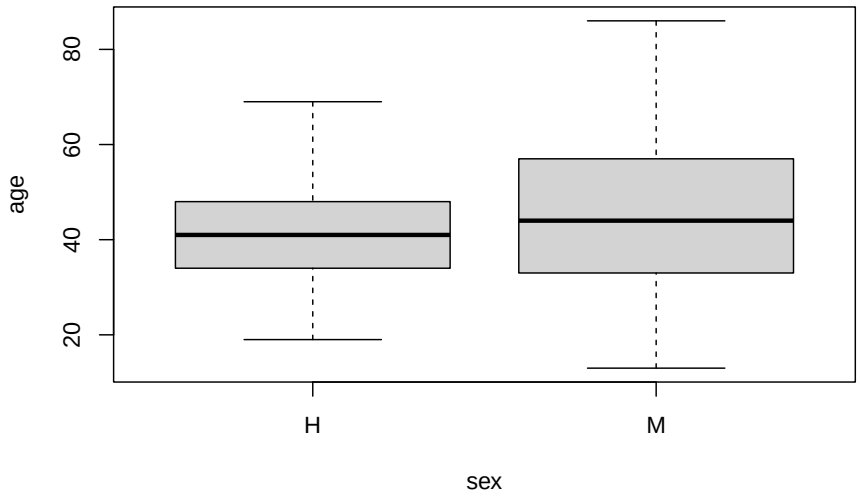
c. Age by sex, mean and standard deviation for males, and mean and standard deviation for females

```
## [1] 40.64444
```

```
## [1] 10.75165
```

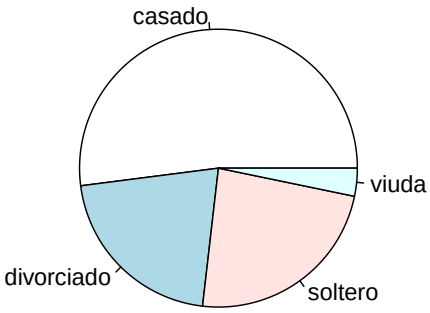
```
## [1] 45.83333
```

```
## [1] 15.58339
```



d. Marital status

```
## Mstate
##     casado divorciado    soltero      viuda
## 0.52032520 0.21138211 0.23577236 0.03252033
```



2. We describe the clinical outcome, for example, anxiety.

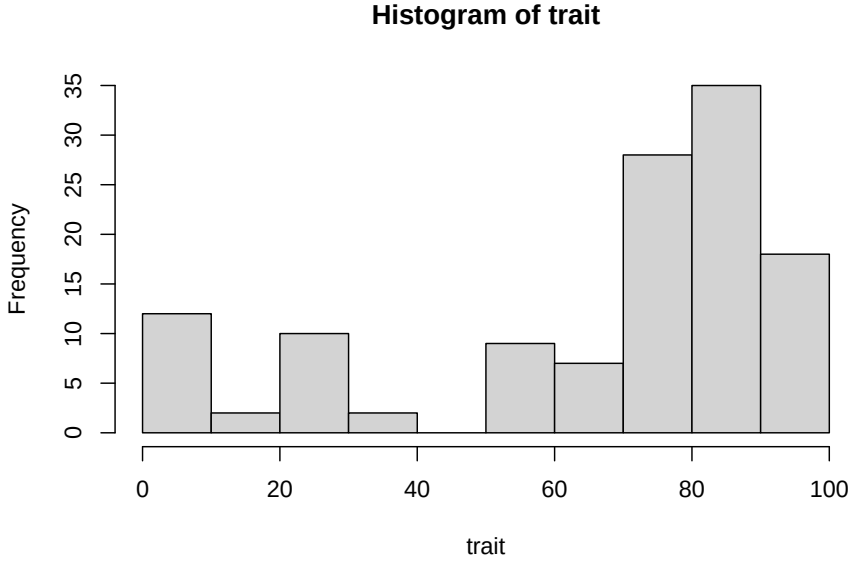
We have four measures of anxiety:

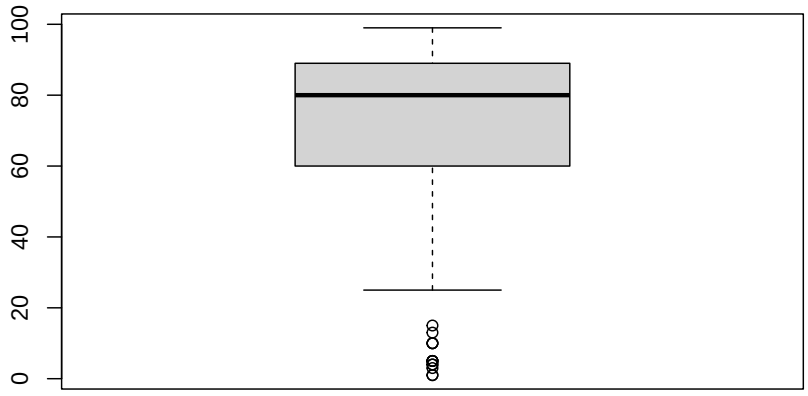
- Trait: ansiedad.rasgo (are you an anxious person?) continuous:0-100
- State: ansiedad.estado (are you currently feeling anxious?) continuous:0-100
- Diagnosed: ansiedad.medicada (have you been diagnosed with an anxiety disorder?) binary (si, no)
- Excess: ansiedad.dif (difference between State and Trait)

we describe these clinical outcomes

a. Trait (min, max, quantiles, median)

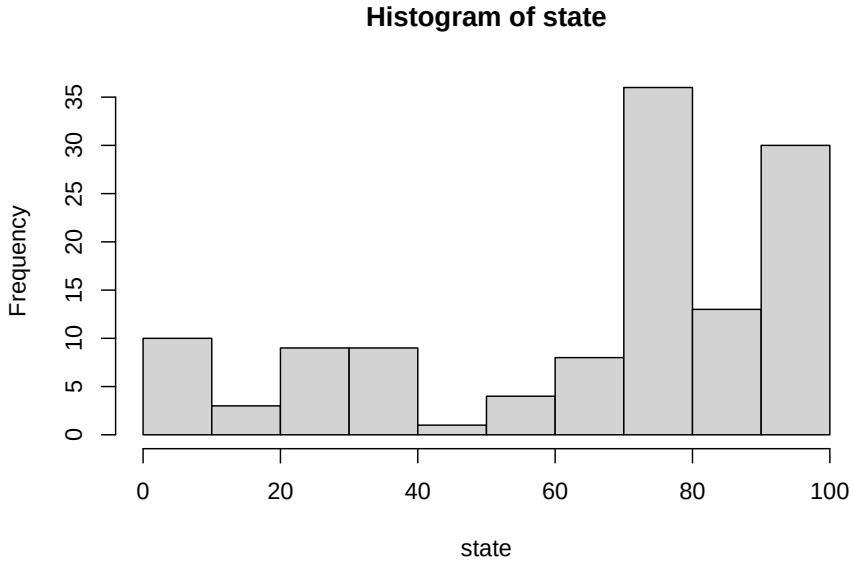
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    1.00   60.00   80.00   68.77   89.00   99.00      15
```

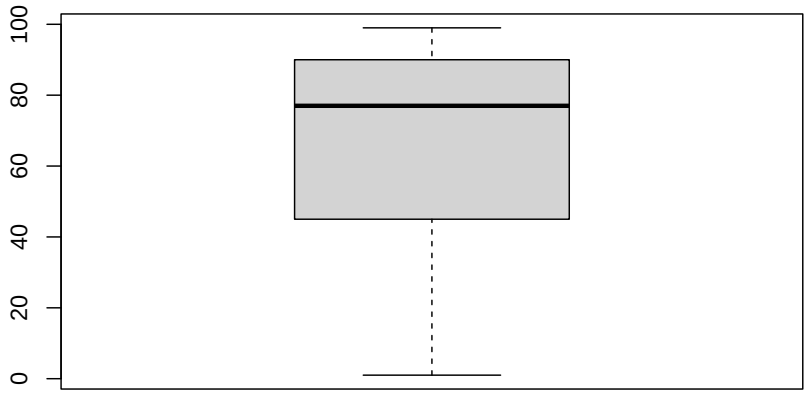




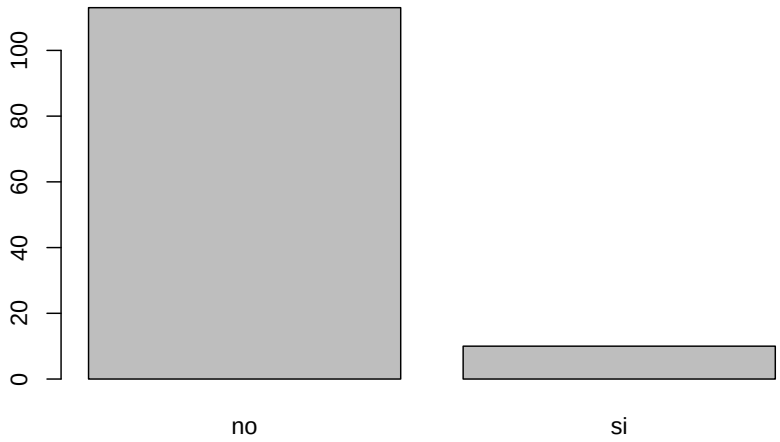
b. State (min, max, quantiles, median)

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    1.00   45.00   77.00   67.85   90.00   99.00      15
```



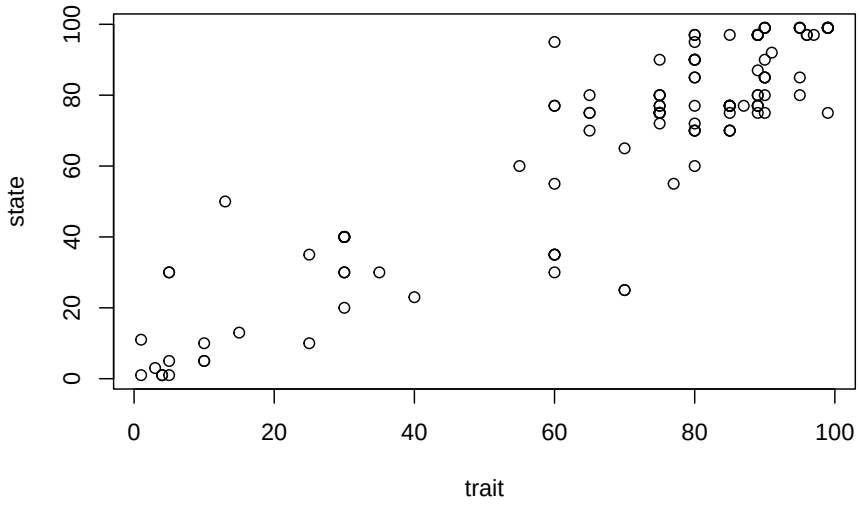


c. Diagnosed



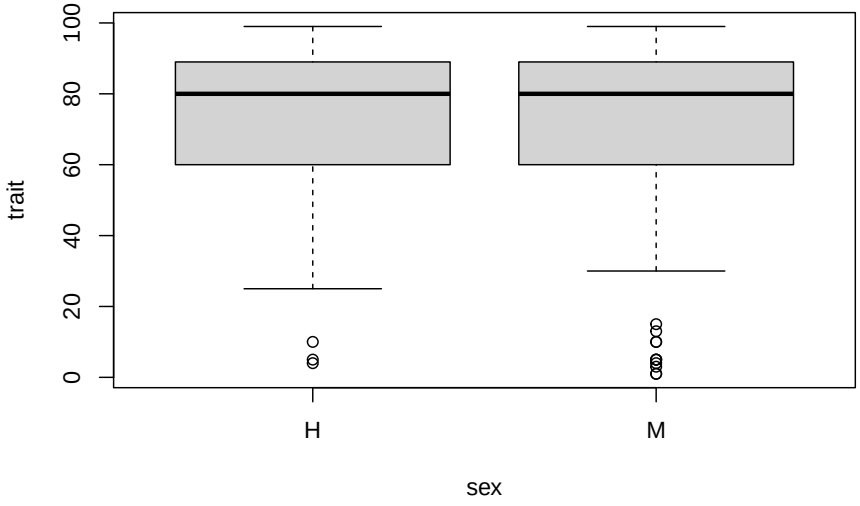
We can look at relationships between outcomes

d. Trait Vs Estate

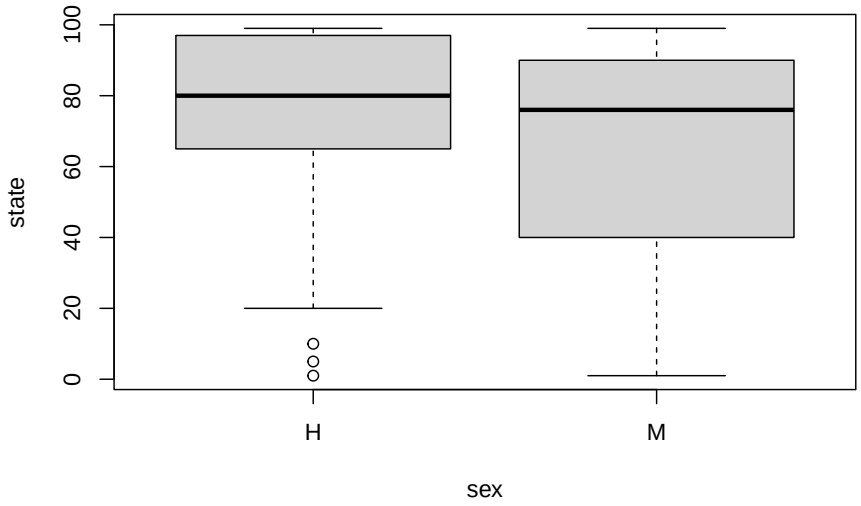


We can also look at the relationships between the clinical outcomes and the features of the participants

e. Trait by sex

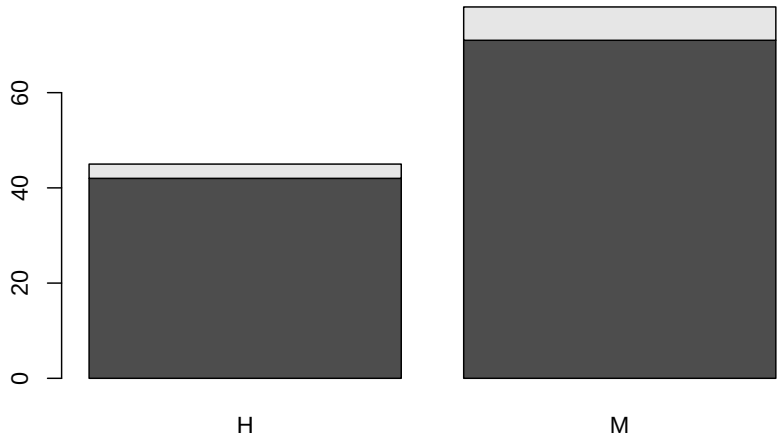


f. State by sex



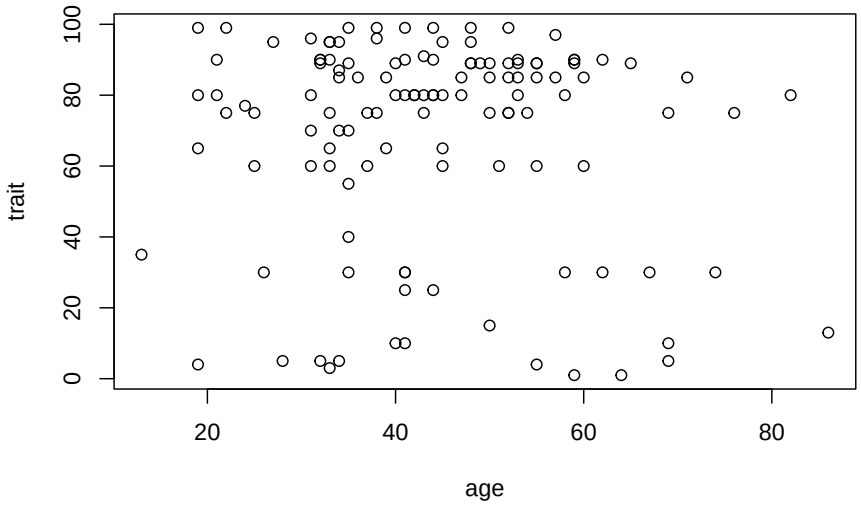
g. Diagnosed by sex

```
##          sex
## diagnosed  H  M
##        no 42 71
##        si  3  7
```

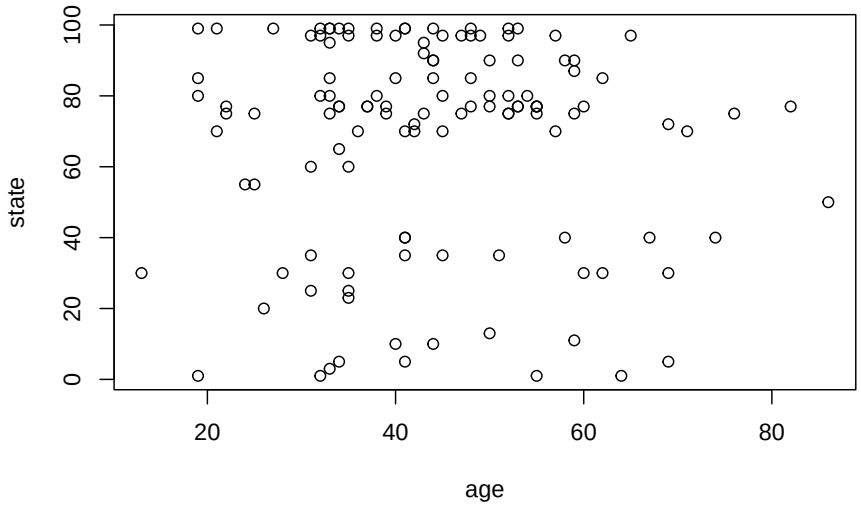



```
##          sex
## diagnosed          H          M
##        no 0.93333333 0.91025641
##        si 0.06666667 0.08974359
```

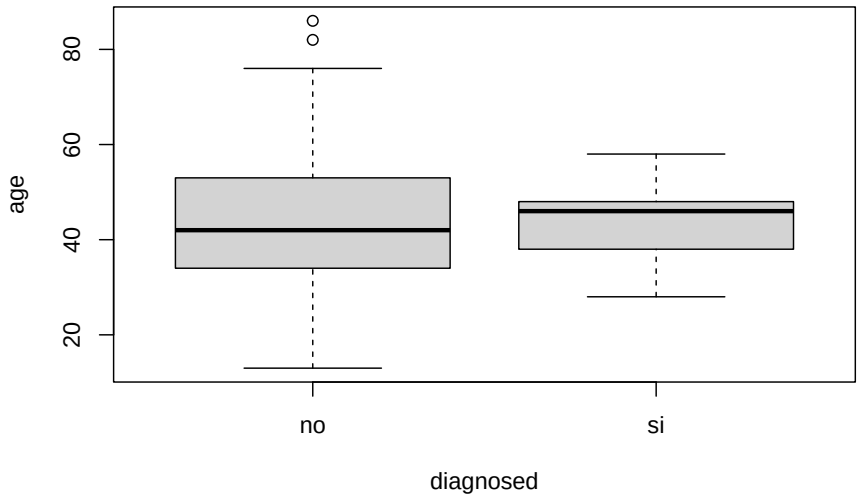
h. Trait Vs age



i. State Vs age



j. age by diagnosis



19.4 Group Work session 2: Inference

When reporting the results of a study, we first describe the variables of interest in tables and figures.

- We describe demographics (sex, age, marital status, etc..)
- We describe outcome variables (misophonia/axiety/depression/etc..)
- We describe explanatory variables (cephalometric measures, anxiety, depression)

We then test the main hypotheses of the study.

- We state the main relationships we want to study and formulate the statistical hypothesis (Introduction)
- We describe how the study was performed and the statistical methods to test the hypothesis (Methods)
- We describe the results of the hypothesis tests with statistics, and significance measures.
- We illustrate the results with figures.

Example:

Imagine we want to study the anxiety of participants in the misophonia study.

We formulate the following hypothesis:

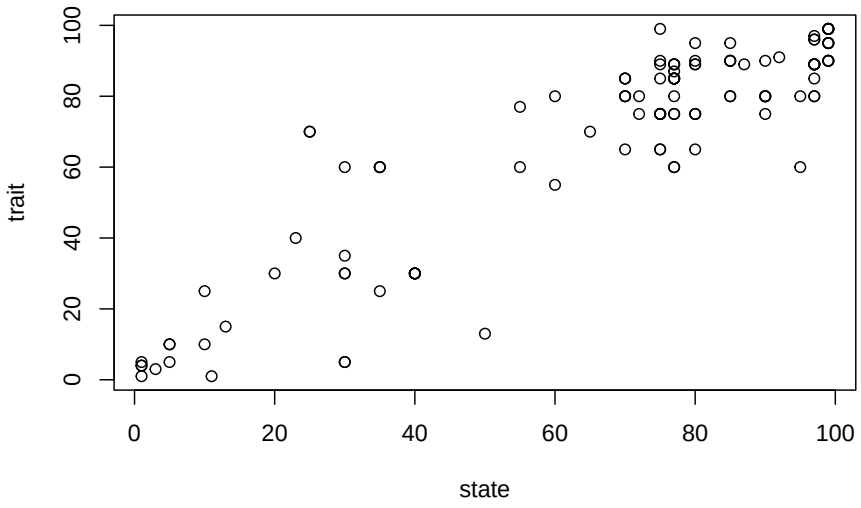
Participants who enrolled in the study had an increased level of anxiety from their baseline (trait) that is related to their:

- age
- sex
- anxiety state.

We are interested in the variable anxiety.dif, that is the observed **excess** of anxiety from the trait

$excess = state - trait$

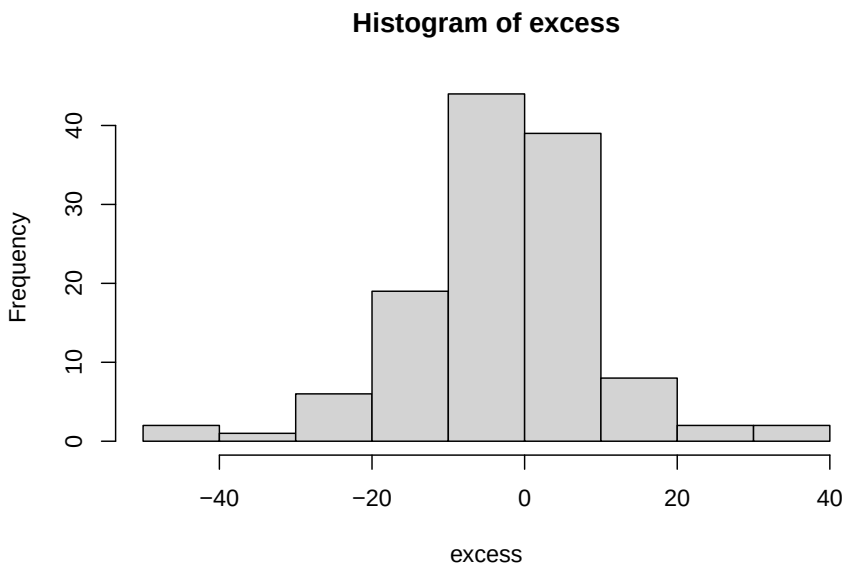
1. Are the state and trait of anxiety correlated?




```
##
##  Pearson's product-moment correlation
##
## data:  state and trait
## t = 23.282, df = 121, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.8656964 0.9320106
## sample estimates:
##       cor
## 0.9041609
```

2. Is excess in anxiety higher than 0?

a. We describe the Excess variable with summary statistics and figures (histogram)




```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's
## -45.0000  -8.0000   0.0000  -0.9187   8.0000  37.0000       15
```

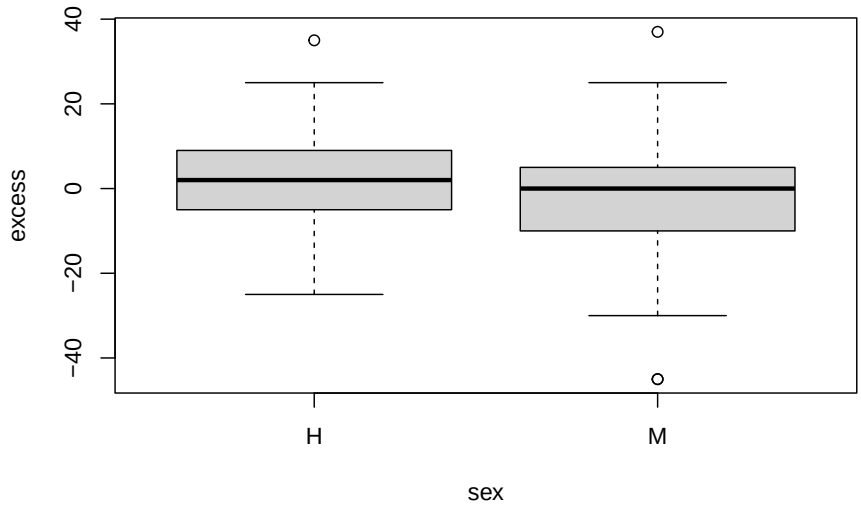
b. We then perform a hypothesis test for the mean of anxiety excess H_0 : $\mu = 0$ against $H_1 : \mu \neq 0$.

```
##
##  One Sample t-test
##
## data:  excess
## t = -0.79192, df = 122, p-value = 0.4299
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -3.215212  1.377814
## sample estimates:
##  mean of x
## -0.9186992
```

c. We conclude: We do not see significant large values of the difference in anxiety; Enrollment in the study does not seem to detect individuals with an excess of anxiety.

2. Is excess in anxiety higher than 0 for men and women separately?

a. We first describe the conditional distributions



b. We perform the hypothesis test for each sex separately

```
##
##  One Sample t-test
##
## data:  excess[sex == "M"]
## t = -1.6994, df = 77, p-value = 0.09328
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -5.5685793  0.4403741
## sample estimates:
## mean of x
## -2.564103
```

```
##
##  One Sample t-test
##
## data:  excess[sex == "H"]
## t = 1.1158, df = 44, p-value = 0.2706
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -1.558796  5.425462
## sample estimates:
## mean of x
##  1.933333
```

c. We conclude: We see that women (M) have a reduction in the excess of anxiety (almost significant), while men (H) had an increase (no significant). Why? perhaps because females tend to consult doctors before men do.

3. Is the excess of anxiety significantly different between the sexes?

a. We test the hypothesis $H_0 : \mu_{men} = \mu_{women}$ against $H_1 : \mu_{men} \neq \mu_{women}$ using a group t.test

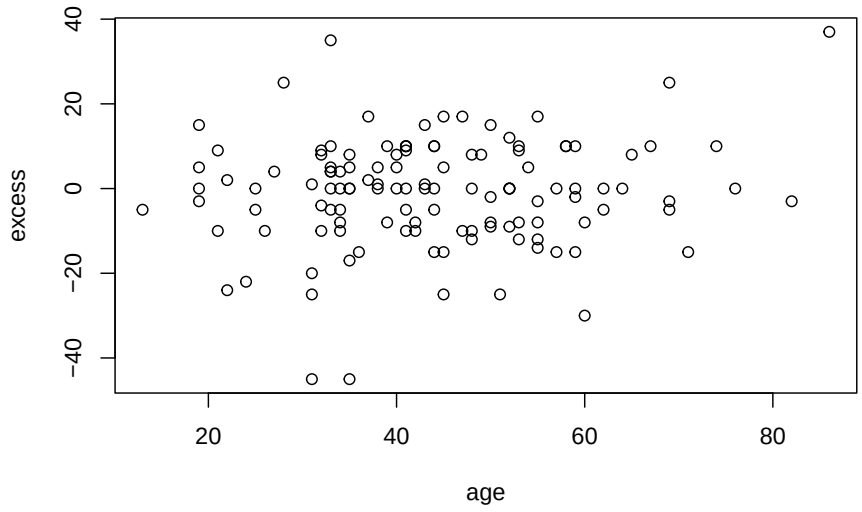
```
##
##  Welch Two Sample t-test
##
## data:  excess[sex == "M"] and excess[sex == "H"]
## t = -1.9574, df = 102.39, p-value = 0.05302
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -9.05452801  0.05965621
## sample estimates:
## mean of x mean of y
## -2.564103  1.933333
```

b. We conclude: we see that the difference between the group means is within the limit of significance with women having less excess anxiety than men.

```
##
## Call:
## lm(formula = excess ~ sex)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -42.436  -7.436   2.067   7.564  39.564
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.933      1.898   1.019   0.3105
## sexM          -4.497      2.384  -1.887   0.0616 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.73 on 121 degrees of freedom
##   (15 observations deleted due to missingness)
## Multiple R-squared:  0.02858,    Adjusted R-squared:  0.02055
## F-statistic:  3.56 on 1 and 121 DF,  p-value: 0.06158
```

4. Is excess in anxiety higher in older people?

a. We make a plot between anxiety and age

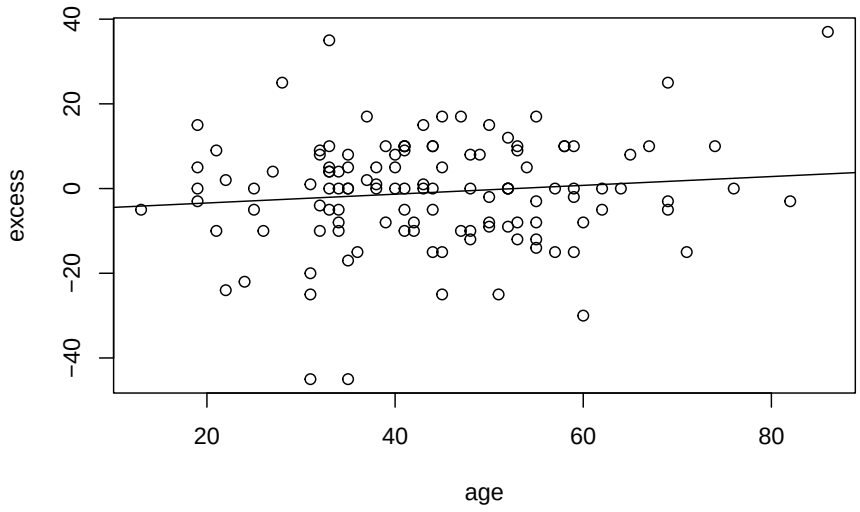


b. We fit the regression model

$$excess = \alpha + \beta * age + \epsilon$$

and test the hypothesis $H_0 : \beta = 0$ against $H_1 : \beta \neq 0$

```
##
## Call:
## lm(formula = excess ~ age)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -43.151  -7.776   0.912   8.516  37.057
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -5.4917     3.7799  -1.453    0.149
## age           0.1041     0.0819   1.271    0.206
##
## Residual standard error: 12.83 on 121 degrees of freedom
##   (15 observations deleted due to missingness)
## Multiple R-squared:  0.01317,    Adjusted R-squared:  0.005016
## F-statistic: 1.615 on 1 and 121 DF,  p-value: 0.2062
```

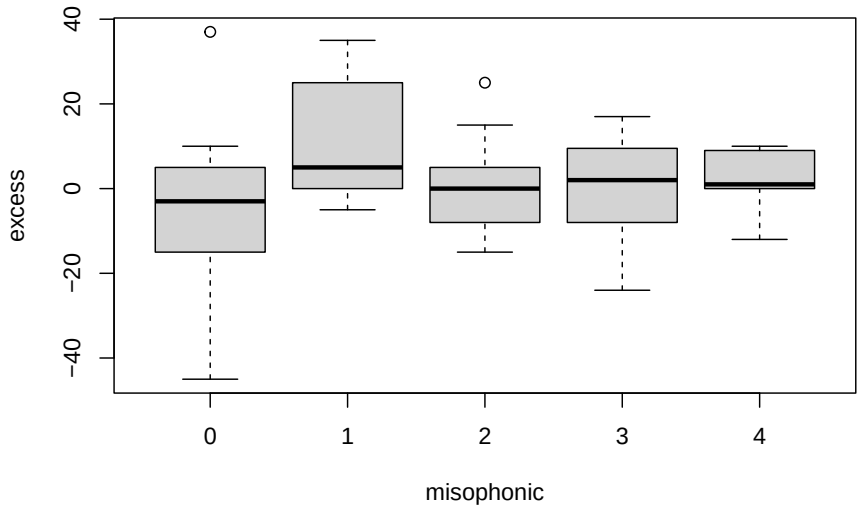


c. We conclude: The association, while positive it is not significant. If we adjust by sex the association is a bit stronger but still not significant.

```
##
## Call:
## lm(formula = excess ~ age + sex)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -40.969  -6.849   0.781   8.019  34.124
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.57179    3.82807  -0.933   0.3527
## age          0.13545    0.08198   1.652   0.1011
## sexM        -5.20025    2.40467  -2.163   0.0326 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.64 on 120 degrees of freedom
##   (15 observations deleted due to missingness)
## Multiple R-squared:  0.05019,    Adjusted R-squared:  0.03436
## F-statistic:  3.17 on 2 and 120 DF,  p-value: 0.04553
```

5. Is excess in anxiety different between misophonic grades?

a. We plot the excess anxiety across groups (boxplot)



b. We test the hypothesizes $H_0 : \mu_0 = \mu_1 ... = \mu_4$ against $H_1 :$ at least one of them is different. We fit an ANOVA model.

```
##
## Call:
## lm(formula = excess ~ misophonic)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -39.902  -8.257   1.243   7.152  42.098
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -5.098      1.944  -2.622  0.00988 **
## misophonic1   17.098      5.896   2.900  0.00445 **
## misophonic2    3.854      2.822   1.366  0.17464
## misophonic3    6.904      2.962   2.331  0.02148 *
## misophonic4    7.986      4.582   1.743  0.08391 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.45 on 118 degrees of freedom
##   (15 observations deleted due to missingness)
## Multiple R-squared:  0.09483,    Adjusted R-squared:  0.06414
```

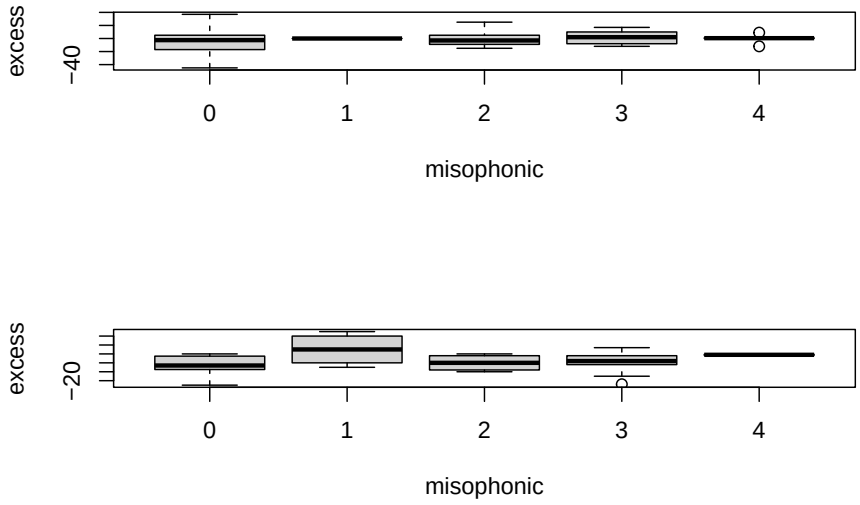
```
## F-statistic:  3.09 on 4 and 118 DF,  p-value: 0.01847
```

```
## Analysis of Variance Table
##
## Response: excess
##             Df Sum Sq Mean Sq F value  Pr(>F)
## misophonic   4   1915  478.76  3.0904 0.01847 *
## Residuals  118  18280  154.92
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

c. We conclude: We see that anxiety excess of misophonia grade 1 is significantly higher than misophonia grade 0 (no misophonia), as it is grade 3. The ANOVA table shows that we accept the alternative hypothesis, where the differences between groups are significantly higher than within groups.

6. Are the differences in excess anxiety between monophonic grades modulated by sex?

a. We plot excess anxiety for each misophonic group, for men and women separately



b. We perform an ANOVA test for the interaction

```
## Analysis of Variance Table
##
## Response: excess
##                 Df  Sum Sq Mean Sq F value  Pr(>F)
```

```
## misophonic       4  1915.0  478.76  3.0366 0.02026 *
## sex              1   179.7  179.74  1.1400 0.28792
## misophonic:sex   4   284.5   71.13  0.4512 0.77137
## Residuals      113 17815.9  157.66
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

c. We conclude: We do not see a significant interaction (modulation) of the effect of sex on the group differences. We cannot say that the profiles of anxiety excess across misophonia grades are different between sexes.

Chapter 20

Solutions to Questions

Chapter 2

1.c; 2.a; 3.d; 4.d; 5.b

Chapter 3

1.a; 2.b; 3.b; 4.b; 5.a

Chapter 4

1.c; 2.b; 3.d; 4.b; 5.b

Chapter 5

1.d; 2.c; 3.b; 4.c; 5.b

Chapter 7

1.d; 2.a; 3.d; 4.a; 5.d

Chapter 8 1.d; 2.b; 3.a

Chapter 9

1.c; 2.a; 3.c; 4.d; 5.b

Chapter 10

1.a; 2.c; 3.d; 4.d; 5.c

Chapter 11

1.c; 2.b; 3.c

Chapter 12

1.d; 2.d; 3.b; 4.a

Chapter 13

1.b; 2.a; 3.d; 4.b; 5.b

Chapter 15

1.b; 2.d; 3.c; 4.d; 5.a

Chapter 16

1.c; 2.b; 3.b; 4.b; 5.a

Chapter 17

1.d; 2.d; 3.a; 4.b; 5.c

Chapter 18

1.d; 2.a; 3.b; 4.c; 5.b