EEBE stats

Alejandro Caceres

2023-09-17

Contents

Chapter 1

Objective

This is the introduction course to the statistics of the EEBE (UPC).

Statistics is a **language** that allows you to face new problems, on which we have no solution, and where the **randomness** plays a crucial role.

In this course we will discuss the **fundamental concepts** of statistics.

- 3 hours of **Theory** per week: we will explain the concepts, we will exercise.
- 6 hours of **Individual study** per week: notes of course notes and resources in Athena.
- 2 hours of problem solving with **R**: face-to-face sessions (practices).

Exam dates and additional study material can be found in **ATENEA metacurso**:

Activitat	Data	Pes
Q1 (T1 – T2)	**11/10/2023** (00:05) – **13/10/2023** (23:55)	10%
EP1 (T3 – T4)	**19/10/2023**, 15.30 h	25%
Q2 (T5 - T6)	**21/11/2023 – 4/12/2023** (en hora de clase)	20%
EP2 (T7 - T8)	**18/01/2024,** 16.00 h	40%
CG	**18/01/2024,** 16.00 h	5%

EP1: Evaluación presencial escrita
EP2: Evaluación presencial con ordenador o tablet que el estudiantado llevará a la prueba

Q1: Cuestionario asíncrono
Q2: Cuestionario síncrono

CG: Competencia Genérica

Evaluation objectives:

Q1 (10%): Test in computer Duration 2h on the indicated dates.

a. Basic command knowledge (practices)

b. Ability to calculate descriptive statistics and graphics, in specific situations (theory/practice)
c. Knowledge about linear regression (practices)

EP1 (25%): Written test (2-3 problems)

a. Capacity to interpret statements in probability formulas (theory).
b. Knowledge of the basic tools to solve problems of joint probability and conditional probability (theory).
c. Mathematical knowledge of probability functions to calculate its basic properties (theory).

Q2 (10%): Test in computer Duration 2h on the indicated dates

a. Capacity to identify probability models in concrete problems (theory/practice).
b. Use of R functions to calculate probabilities of probabilistic models (practice/theory)
c. Identification of a sampling statistic and its properties (theory/practice)
d. Knowledge of how to calculate the probability of sampling statistics (theory/practice)
e. Use of R commands to calculate probabilities and make random sampling simulations (practice)

EP2 (40%): Written test (2-3 problems)

a. Mathematical capacity to determine specific estimators of probability models.
b. Knowledge of the properties of specific estimators.
c. Knowledge of confidence intervals and their properties (theory).
d. Ability to identify the type of confidence interval in a specific problem (theory).
e. Knowledge of hypothesis types to be used in a specific problems (theory). f.Use of R commands to solve confidence intervals and hypothesis tests (practice).

CG (5%): Written test (2 questions about a text)

a. Written expression capacity on a subject related to statistics.

Coordinators:

- Luis Mujica (Luis.eduardo.mujica@upc.edu)
- Pablo Buenestado (Pablo.buenetado@upc.edu)

1.1 Recommended reading

- Class notes are our section will be accessible in Athena in PDF and HTML.

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

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 randomized experiments, observations, outcomes, and absolute and relative frequencies.

2.1 Scientific method

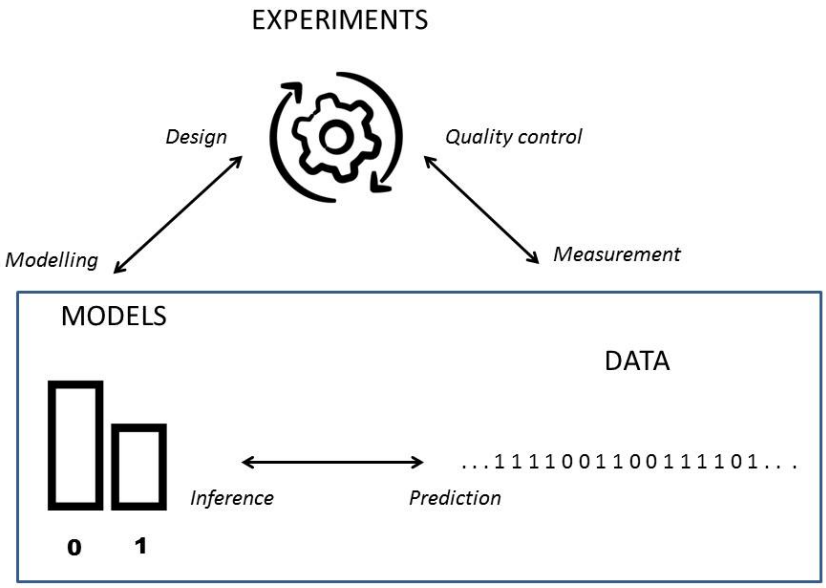
One of the goals of the scientific method is to provide a framework for solving problems that arise in the study of natural phenomena or in the design of new technologies.

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*.



2.2 Statistics

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?

2.3 Data

The data is presented in the form of observations.

An **Observation** or **Realization** is the acquisition of a number or characteristic of an experiment.

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** in a repeat of the experiment

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

1 is one result, **0** is the other result of the experiment.

Remember that the observation is **concrete** is the number you get one day in the laboratory. The **abstract** result is one of the characteristics of the type of experiment you are running.

2.4 Result types

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

- **Categorical**: If the result 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).
- **Numeric**: If the result 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).

2.5 Random experiments

It can be said that the subject of study of statistics is random experiments, the means by which we produce data.

Definition:

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

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

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

2.6 Absolute frequencies

When we repeat a randomized experiment with **categorical** results, we record a list of results.

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

Absolute frecuency:

$$n_i$$

is the number of times we observe the result i.

Example (leukocytes)

Let's take a leukocyte from **a** 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. The last **B cell** is observation number 119.

We can list the **results** (categories) in a **frequency table**:

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

From the table, we can say that, for example, $n_1 = 34$ is the total number of T cells observed in the repeat experiment. We also note that the total number of repetitions $N = \sum_i n_i = 119$.

2.7 Relative frequencies

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

$$f_i = n_i/N$$

where N is the total number of observations

In our example, $n_1 = 34$ T cells were recorded, so we asked about the proportion of T cells out of the total 119. We can add 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 result in relation to the other results. Later we will understand them as the observations of probabilities.

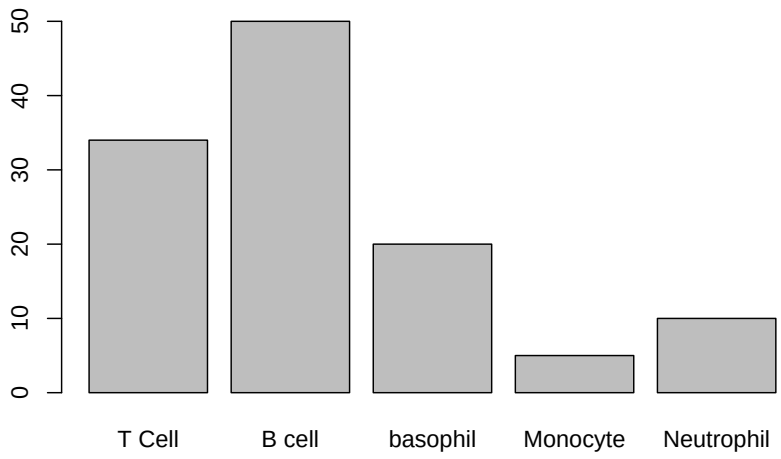
For absolute and relative frequencies we have the properties

- $\sum_{i=1..M} n_i = N$
- $\sum_{i=1..M} f_i = 1$

where M is the number of results.

2.8 Bar chart

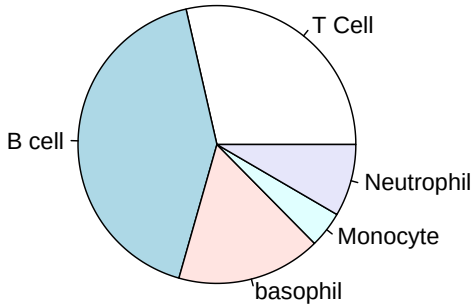
When we have a lot of results and want to see which ones are most likely, we can use a bar chart that is a number of n_i Vs the results.



2.9 Pie chart (pie)

We can also visualize the relative frequencies with a pie chart.

The area of the circle represents 100% of the observations (proportion = 1) and the sections the relative frequencies of each result.



2.10 Ordinal categorical variables

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

Sometimes **categorical** variables can be **sorted** when they meet 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 certain 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 a randomized experiment: measurement of the level of misophonia in a patient. This data series can be summarized in terms of the results in 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
```

2.11 Accumulated absolute and relative frequencies

Misophonia severity is **categorical ordinal** because its results can be ordered relative to its degree.

When the results can be ordered, it is useful to ask how many observations were obtained up to a given result. We call this number the **absolute cumulative frequency** up to the result i:

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

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

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

We can add these frequencies in 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
```

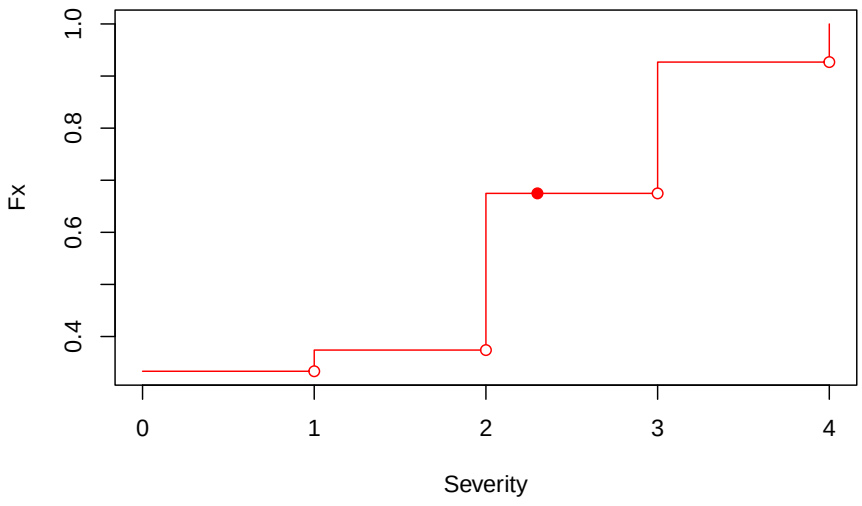
Therefore, **67%** of patients had misophonia up to severity **2** and **37%** of patients had severity less than or equal to **1**.

2.12 Cumulative frequency graph

F_i is an important quantity because it allows us to define the accumulation of probabilities down to intermediate levels.

The probability of an intermediate level x ($i \leq x < i+1$) is just the accumulation up to the lower level $F_x = F_i$.

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



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

2.13 Numeric variables

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

When the number is **continuous** the frequencies are not useful, we are most likely to observe or not observe a particular continuous number.

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 results (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
```

2.14 Transforming continuous data

Since continuous outcomes cannot be counted (informatively), we transform them into ordered categorical variables.

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 variable **ordered**; in this case with 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 between the interval (or **bin**) $(7.92, 12.4]$.

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

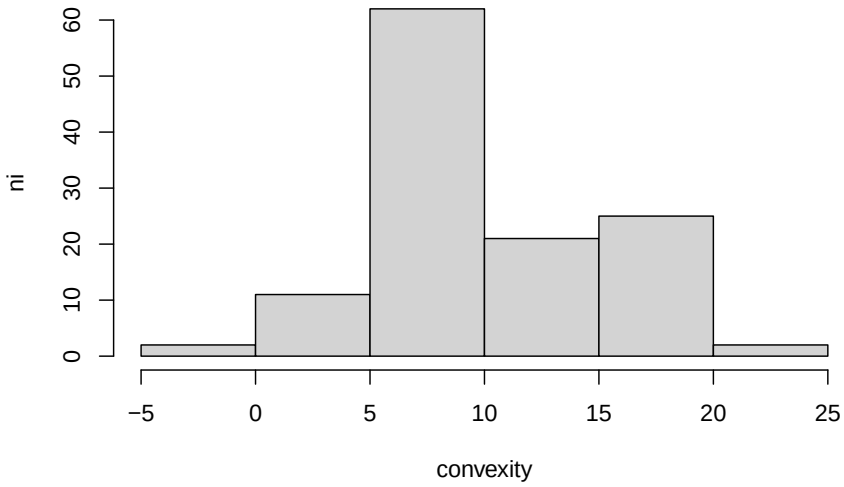
2.15 Frequency table for a continuous variable

For a given regular partition of the interval of results into intervals, we can produce a frequency table as before

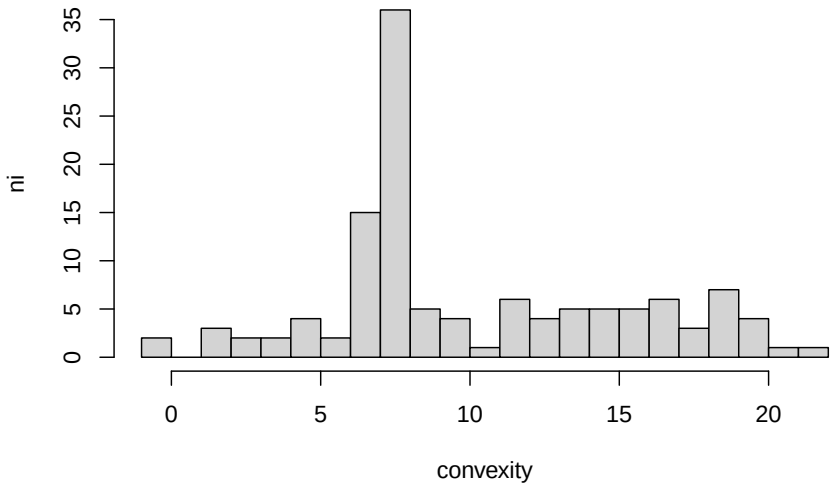
```
##        outcome ni         fi  Ni         Fi
## 1 [-1.02,3.46]  8 0.06504065   8 0.06504065
## 2  (3.46,7.92] 51 0.41463415  59 0.47967480
## 3  (7.92,12.4] 26 0.21138211  85 0.69105691
## 4  (12.4,16.8] 20 0.16260163 105 0.85365854
## 5  (16.8,21.3] 18 0.14634146 123 1.00000000
```

2.16 Histogram

The histogram is the graph of n_i or f_i Vs the results in intervals (bins). The histogram depends on the size of the bins .


ni
0
10
20
30
40
50
60
−5
0
5
10
15
20
25
convexity

This is a histogram with 20 bins .



We see that most people have angles within $(7, 8]$

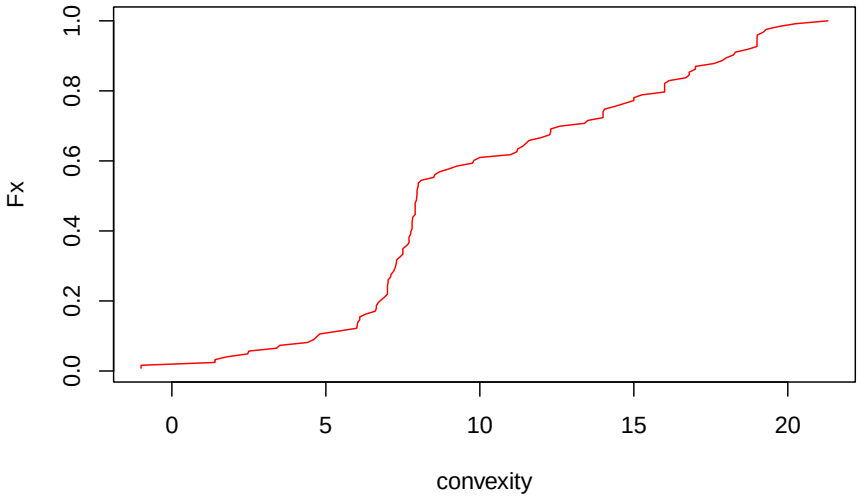
2.17 Cumulative frequency graph

We can also plot F_x against the results. Since F_x is of 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.



2.18 Summary Statistics

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

For example, we have statistics that describe extreme values:

- **minimum**: the minimum result observed
- **maximum**: the maximum result observed

2.19 Average (sample mean)

An important statistic that describes the central value of the results (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 out of a total of N.

Example (Misophonia)

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

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

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

For **categorically ordered** variables, 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 **results**.

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

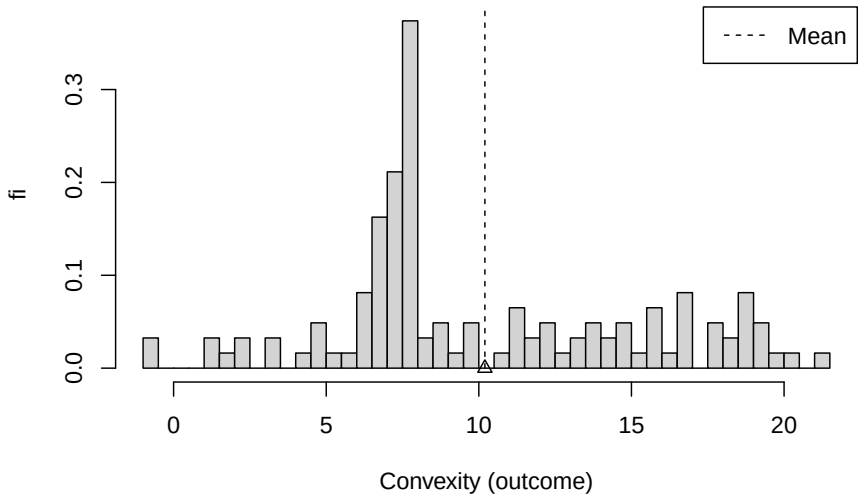
Example (Misophonia)

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
```

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

The average is also the center of gravity for continuous variables. That is the point where the relative frequencies balance.



2.20 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. 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, 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 cut the data at

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

to split them in half.

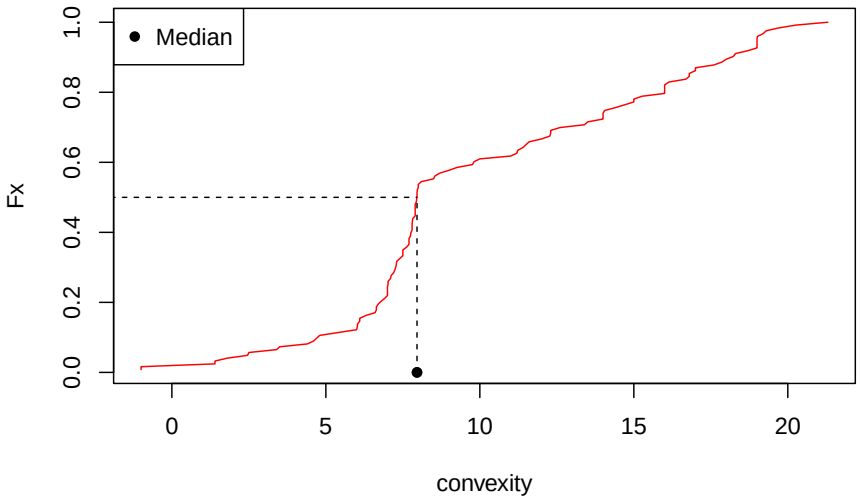
In terms of frequencies, $q_{0.5}$ makes the cumulative frequency F_x equal to 0.5

$$\sum_{i=0,\dots m} f_i = F_{q_{0.5}} = 0.5$$

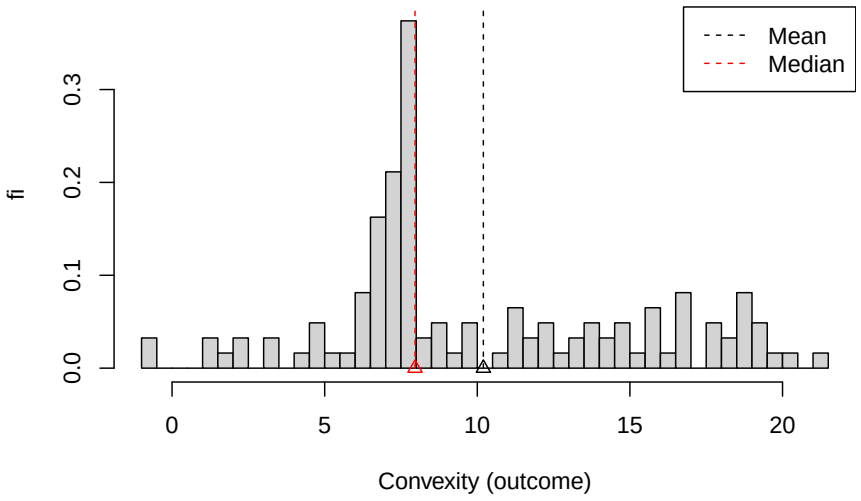
that is

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

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.



The mean and median are not always the same.

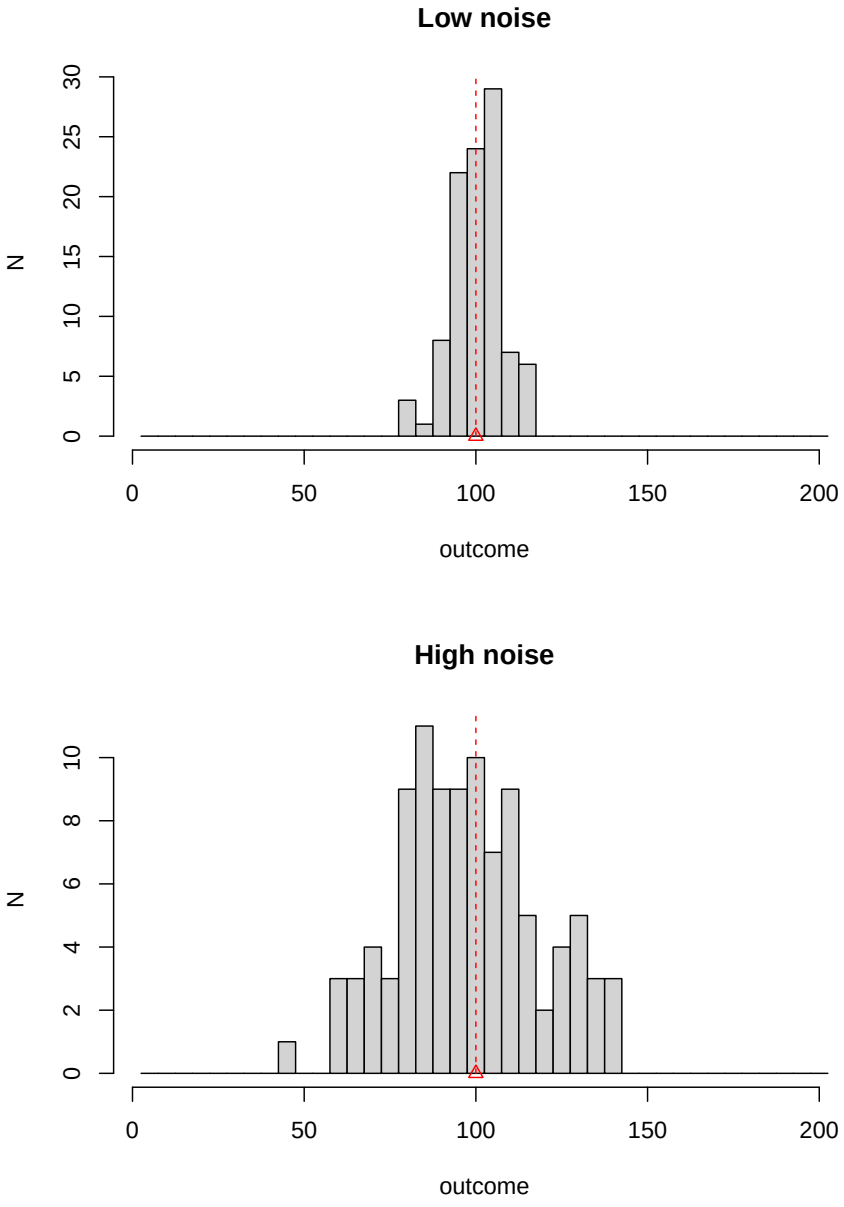


2.21 Dispersion

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

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

The dispersion of the observations is a measure of the **noise**.



2.22 Sample variance

The dispersion about the mean 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 variables that are **categorical and ordered**, we can **also** calculate the sample variance as

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

s^2 can be considered as the **moment of inertia** 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 angle of convexity is

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

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

The jaw convexity deviates from its mean by 5.086707.

2.23 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_m 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)

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

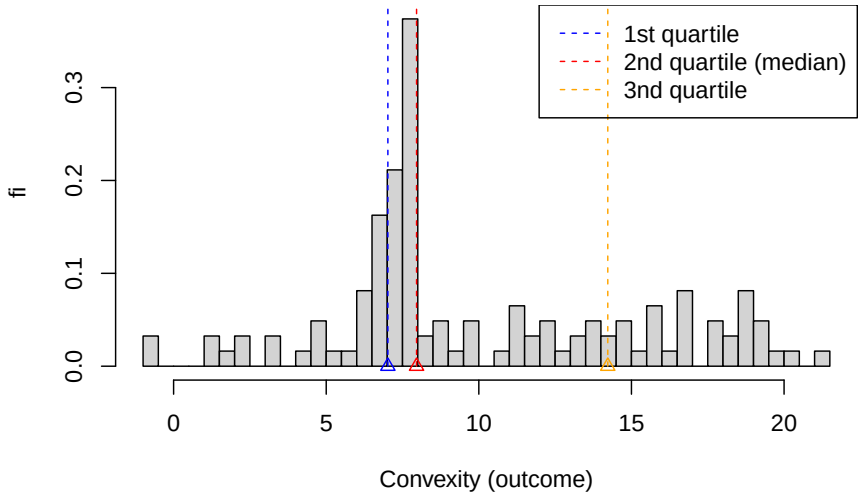
2) We define the **third** quartile as the value x_m that makes the cumulative frequency $F_{q_{0.75}}$ equal to 0.75 (or the value of x where we have accumulated three quarters of observations)

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

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

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

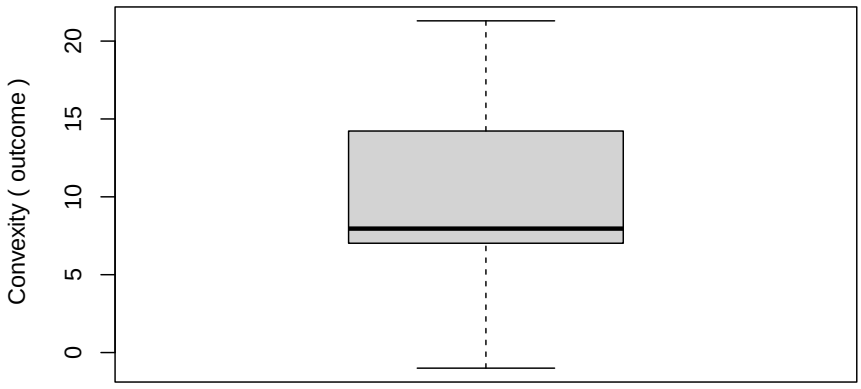
This is the distance between the third and first quartiles and captures the central 50% of the observations



2.24 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 results 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.



2.25 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 variables;

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.

2.26 Exercises

2.26.0.1 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?

2.26.0.2 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 .

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 **results**.

- 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 it is how can we know something if every time we look at it it changes?

3.2 Measurement probability

We would like to have a measure for the outcome of a randomized experiment that tells us **how sure** we are of observing the outcome when we perform a **future** randomized experiment.

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 happen.

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.

- We do not observe P_i
- We deduce P_i from our ratio 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?

$P_2 = 1/6 = 0.166666$.

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)

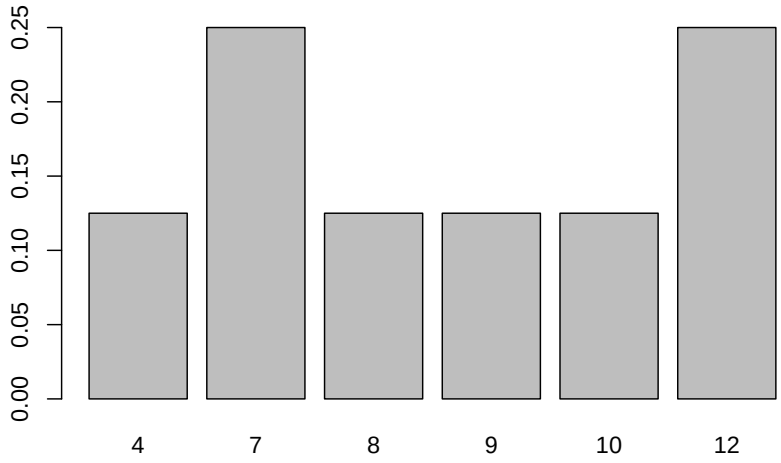
Imagine that we repeat a random experiment 8 times and obtain the following observations

8 4 12 7 10 7 9 12

- How sure are we of obtaining the result 12 in the following observation?

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

- is a number between 0 and 1.
- measures the proportion of the total number of observations that we observe of a particular result.

Since $f_{12} = 0.25$ then we would be one quarter sure, one out of every 4 observations, of getting 12.

Question: How good is f_i as a measure of certainty of the result i?

Example (random experiment with more repetitions)

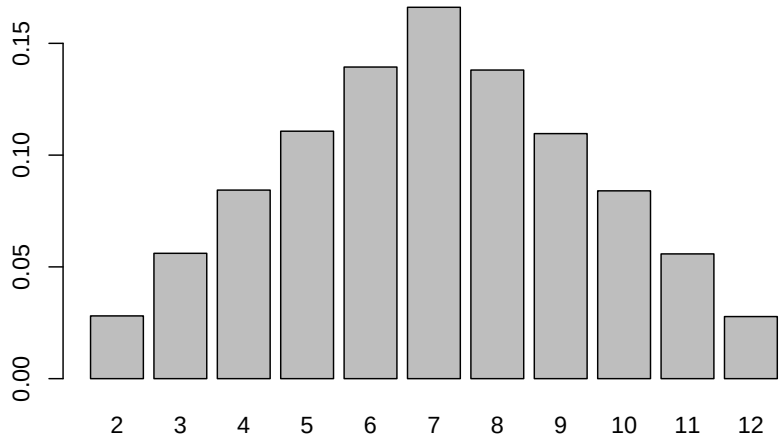
Let's say we repeat the experiment 100,000 more times:

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

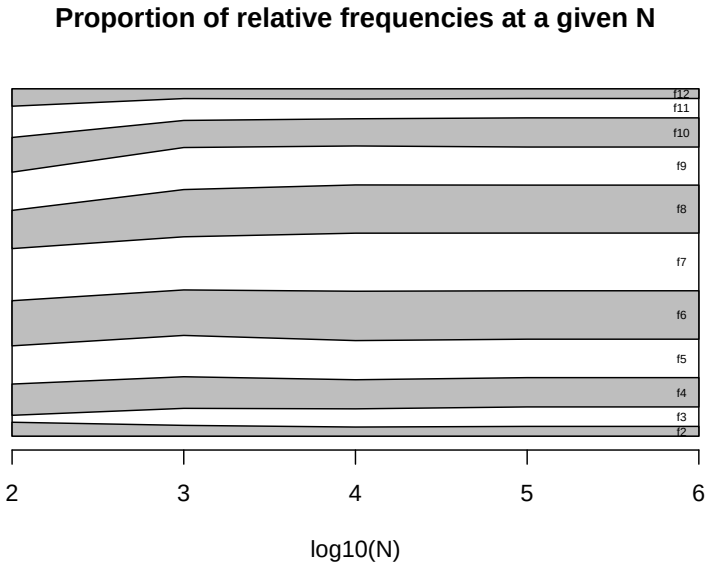


New results came out and f_{12} is now only 0.027, and so we are only $\sim 3\%$ sure to get 12 in the next experiment. The probabilities measured by f_i change with N.

3.5 Relative frequencies at infinity

A crucial observation is that if we measure the probabilities of f_i in increasing values of N they **converge**!

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



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

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

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.

- We do not observe P_i, we observe f_i
- **We estimate** P_i with f_i (usually when N is large), we 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.

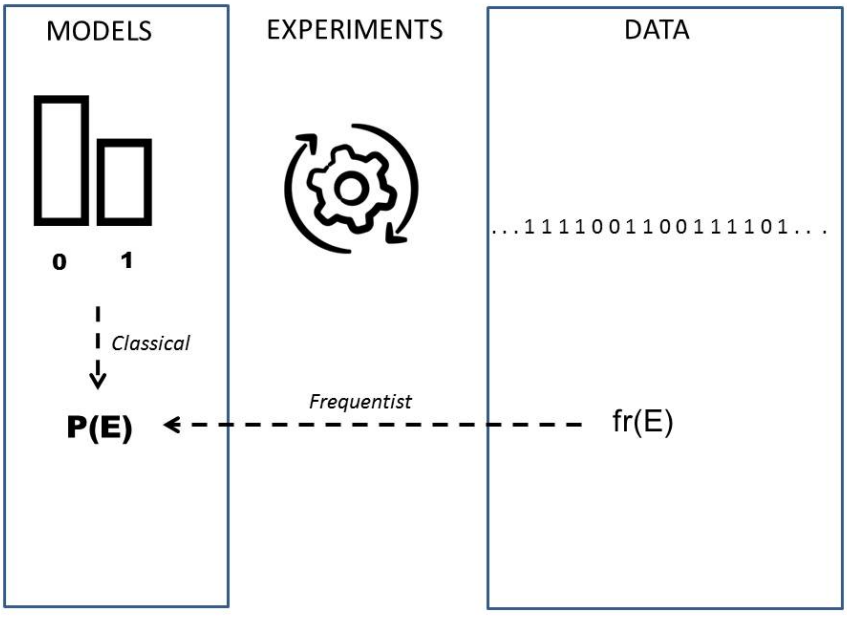
Example (sum of two dice)

Our previous example is based on the **sum of two dice**. Although we perform the experiment many times, write down the results, and calculate the **relative frequencies**, we can know the exact value of probability.

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)

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

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

3) The probabilities of each outcome are between 0 and 1.

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: $S = \{1, ...\infty\}$
- 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 posible results.

Examples of events:

- The event of a healthy temperature: $A = 37 - 38$ degrees Celsius
- The event of producing a screw with a size: $A = 71.5 mm$
- 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 roll of a says: $A = \{1, 2, 3\}$

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 derived 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 roll a die and look at the events (result set):

- 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

$$E_1 \cap E_2 = \emptyset$$

Examples:

- The result of having a misophonia severity of 1 and a severity of 4.
- The results of obtaining 12 and 5 by adding the throw of two dice.

According to the Kolmogorov 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's

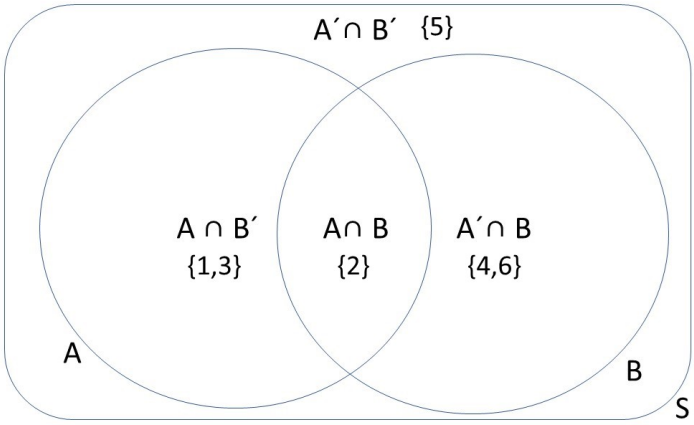
$$P(A \cap B)$$

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 Ven 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 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**

	B	B'
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 $(3, 4)$? (R:1/36)
- What is the probability of rolling 3 and 4 with any of the two dice? (R: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? (R: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 ? (R:31)
- How many individuals had grade 2 misophonia without depression? (R: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 ? (R: 0.3)
- What is the probability of not being misophonic **and** not being depressed? (R:0.284)
- What is the probability of being misophonic **or** depressed? (R: 0.715)
- What is the probability of being misophonic **and** being depressed? (R: 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 $P(A \cup B \cup C)$?