EEBE stats

Alejandro Caceres

2023-11-10

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

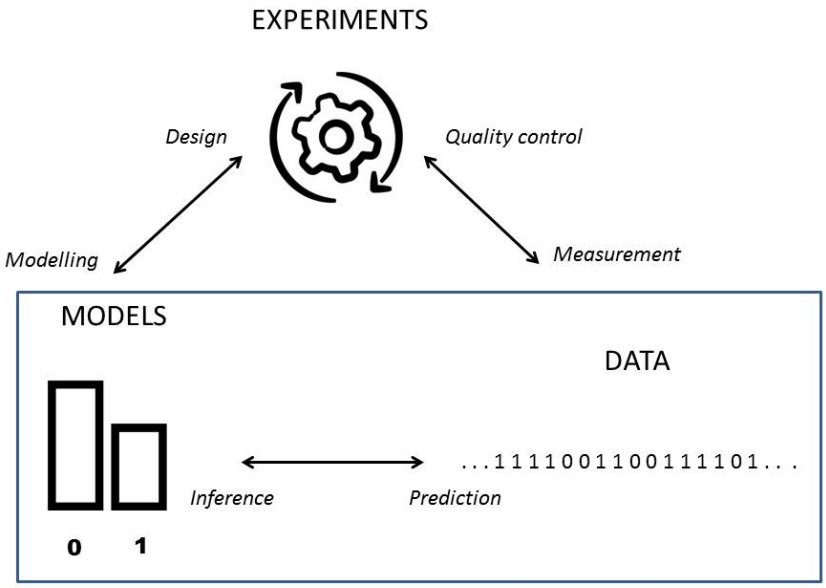
One of the goals of the scientific method is to provide a framework for solving problems that arise in the study of natural phenomena or in the design of new technologies.

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.

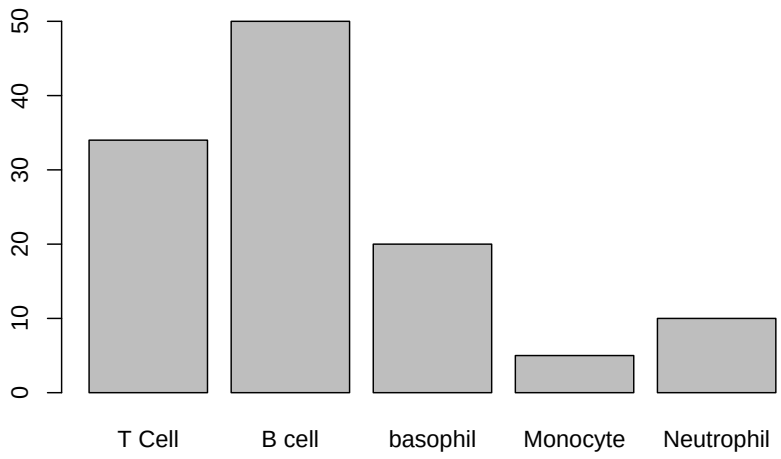
For absolute and relative frequencies we have the properties

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

where M is the number of results.

2.8 Bar chart

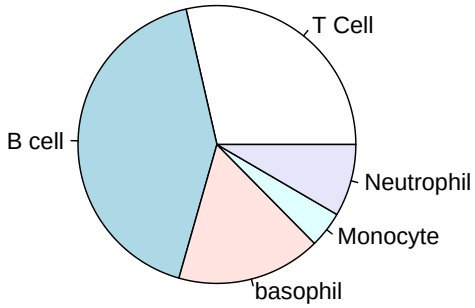
When we have a lot of results and want to see which ones are most likely, we can use a bar chart that is a number of n_i Vs the results.



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
```

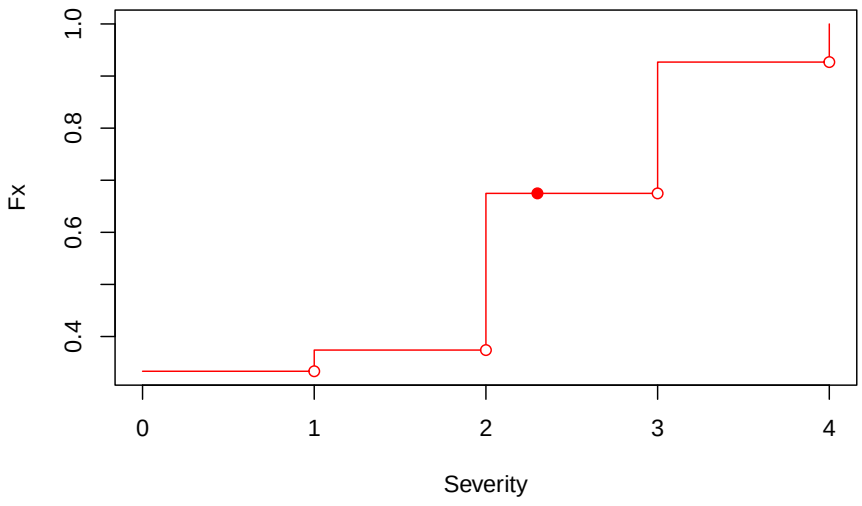
Therefore, **67%** of patients had misophonia up to severity **2** and **37%** of patients had severity less than or equal to **1**.

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 $(7.92, 12.4]$.

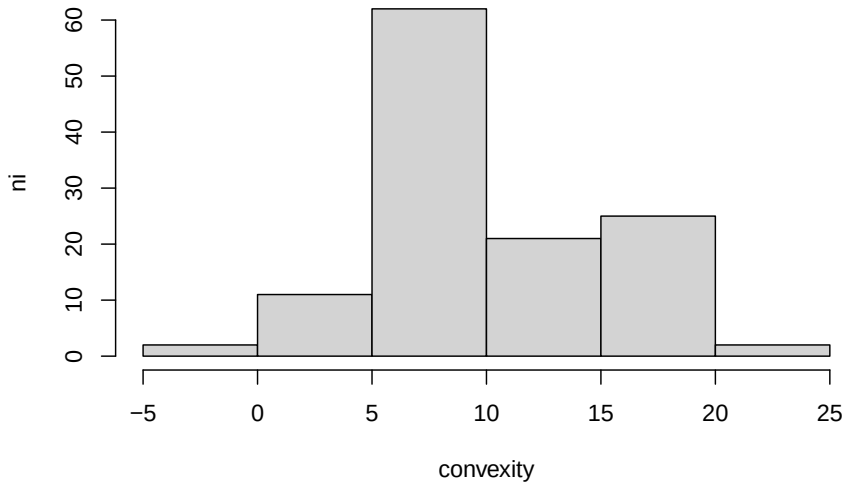
2.15 Frequency table for a continuous variable

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

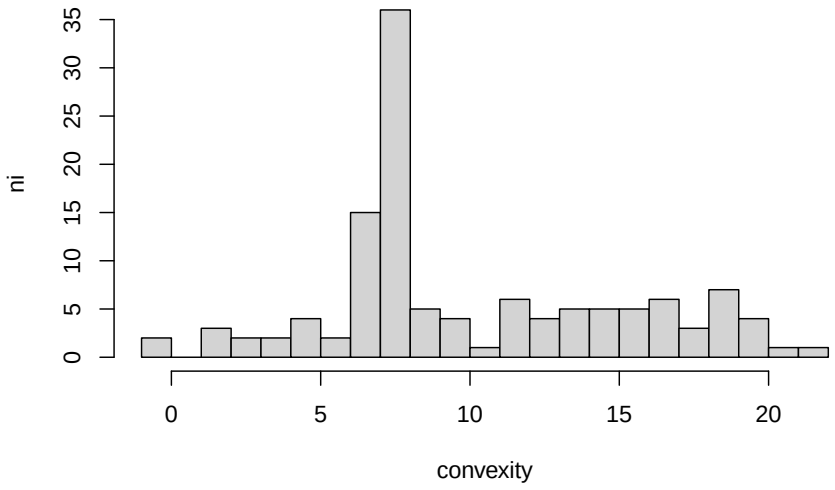
```
##        outcome ni         fi  Ni         Fi
## 1 [-1.02,3.46]  8 0.06504065   8 0.06504065
## 2  (3.46,7.92] 51 0.41463415  59 0.47967480
## 3  (7.92,12.4] 26 0.21138211  85 0.69105691
## 4  (12.4,16.8] 20 0.16260163 105 0.85365854
## 5  (16.8,21.3] 18 0.14634146 123 1.00000000
```

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 $(7, 8]$

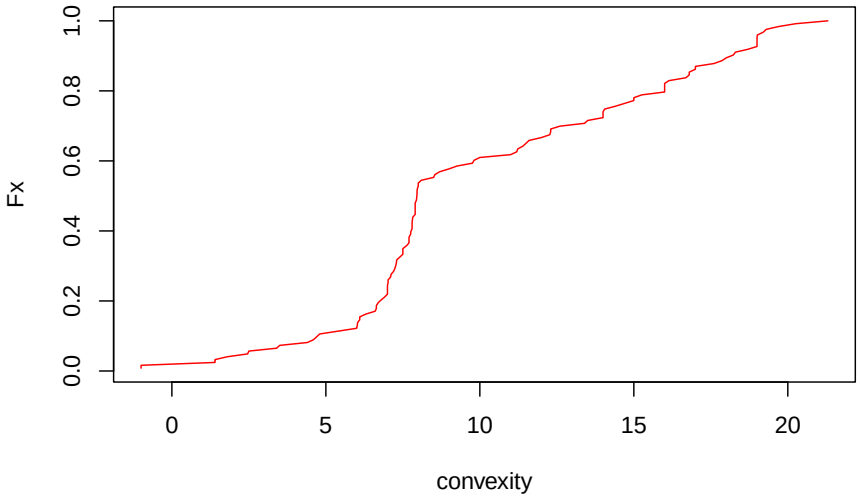
2.17 Cumulative frequency graph

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.

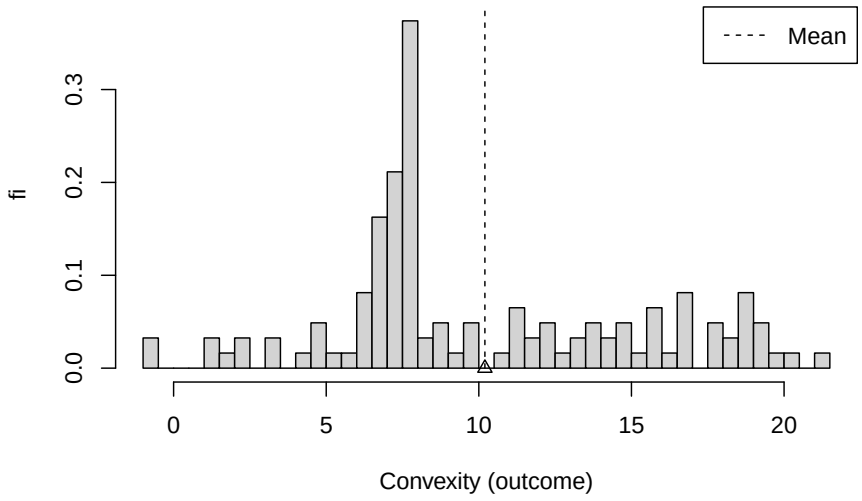
Example (Misophonia)

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.

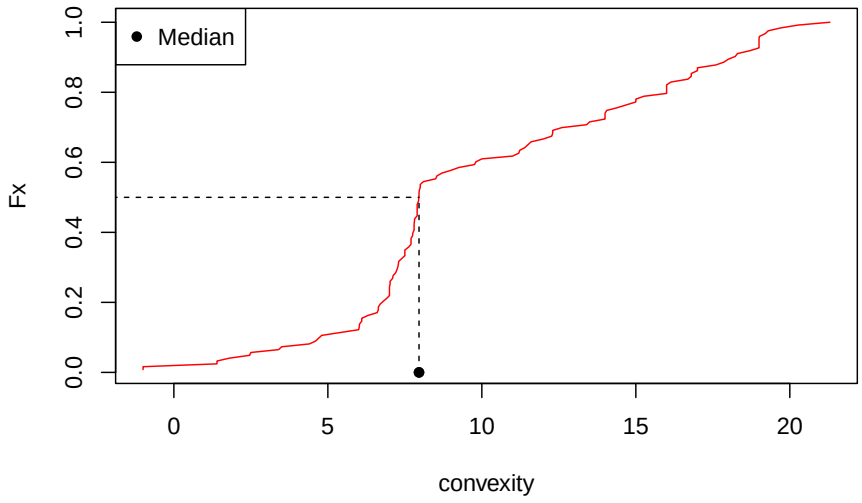
In terms of frequencies, $q_{0.5}$ makes the cumulative frequency F_x equal to 0.5

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

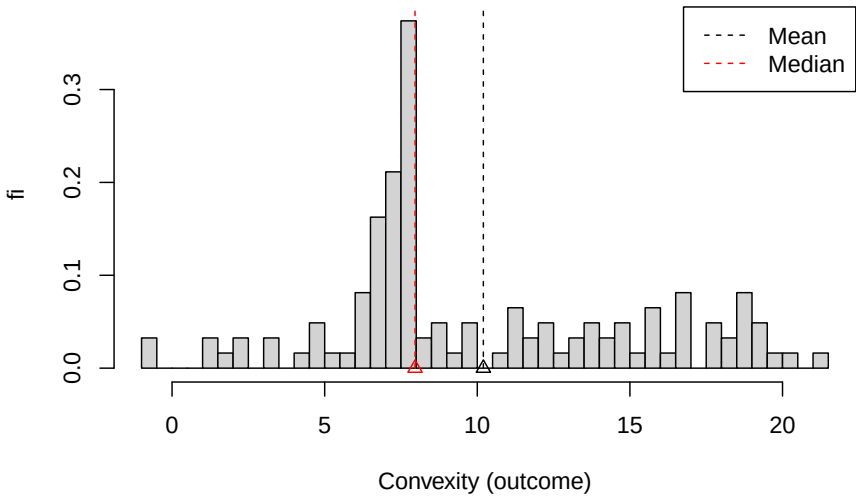
that is

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

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.

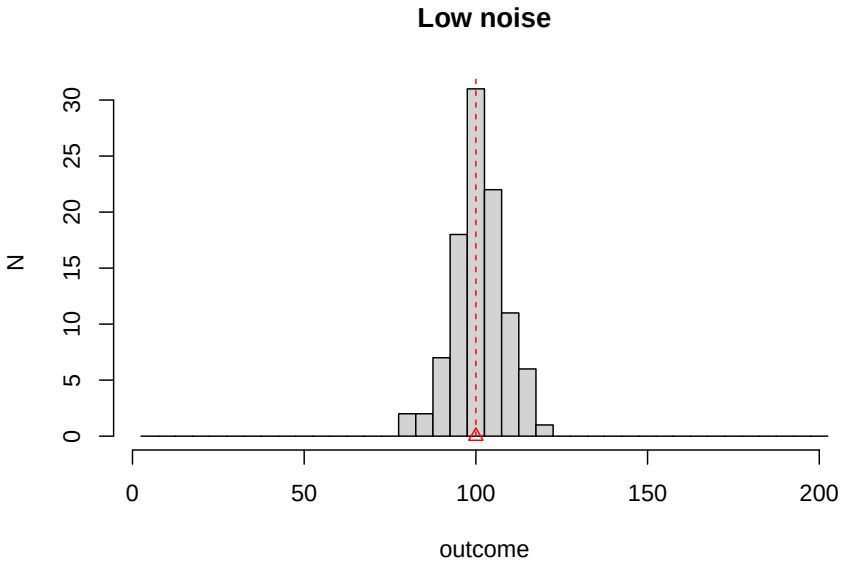


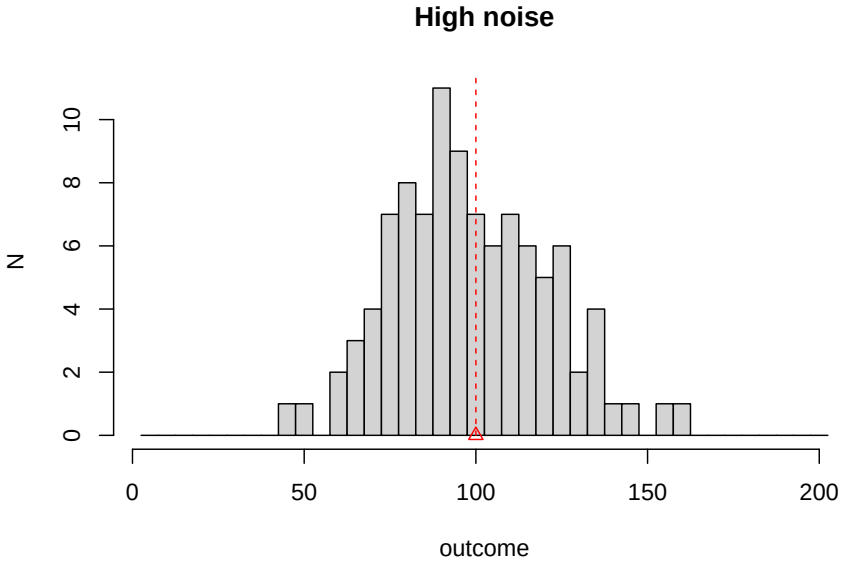
2.21 Dispersion

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)

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

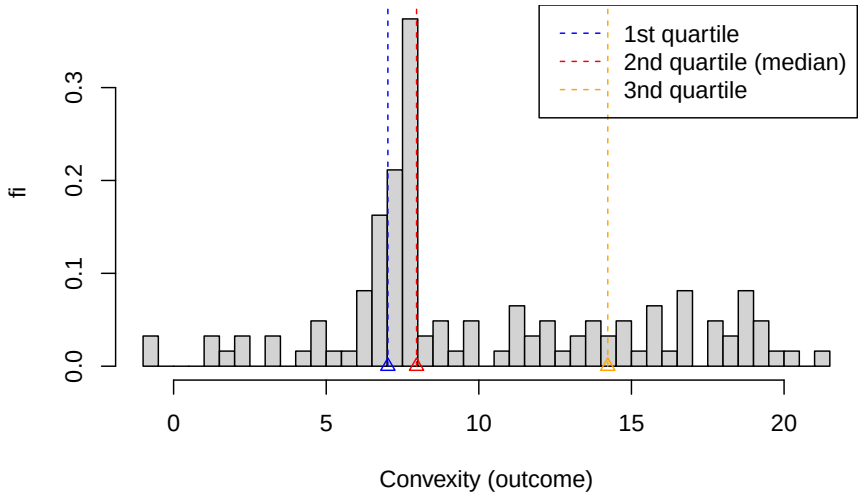
2) We define the **third** quartile as the value x_m that makes the cumulative frequency $F_{q_{0.75}}$ equal to 0.75 (or the value of x where we have accumulated three quarters of observations)

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

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

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

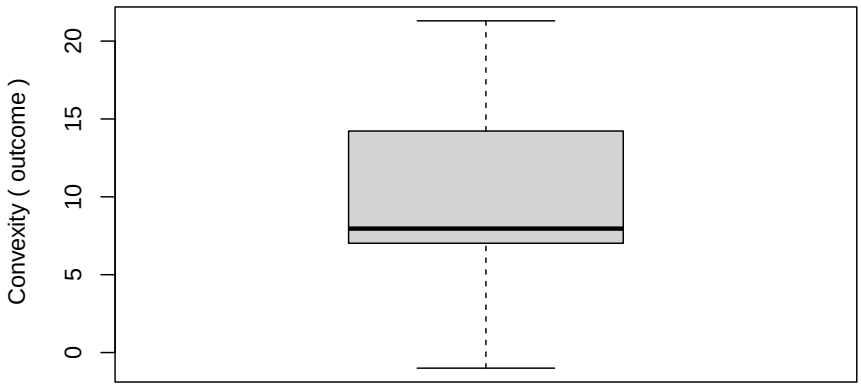
This is the distance between the third and first quartiles and captures the central 50% of the observations



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)

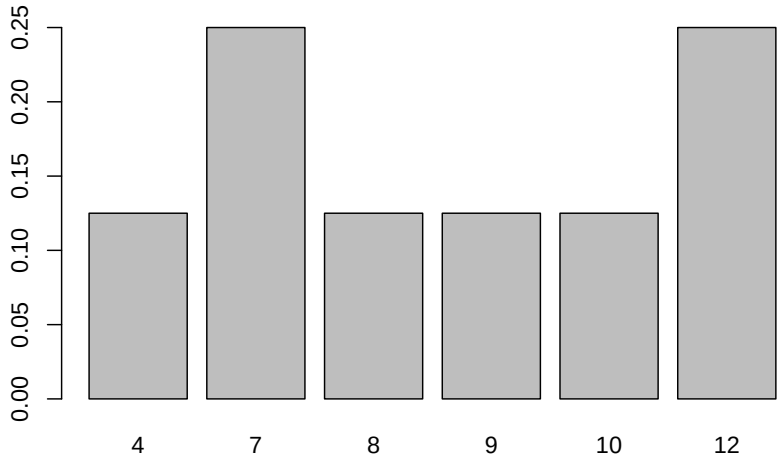
Imagine that we repeat a random experiment 8 times and obtain the following observations

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)

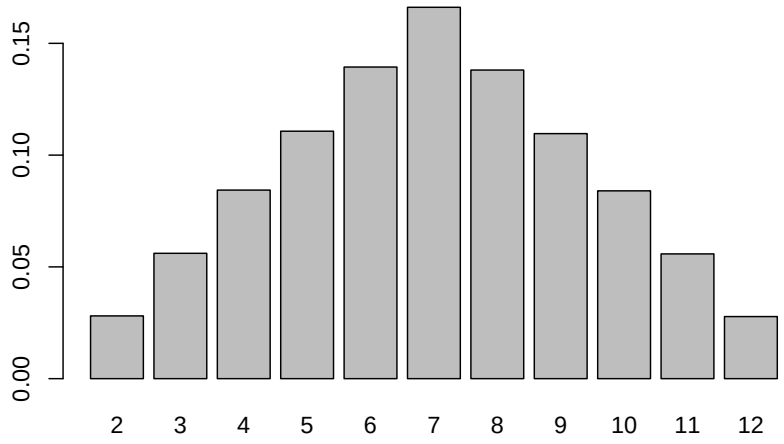
Let's say we repeat the experiment 100,000 more times:

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

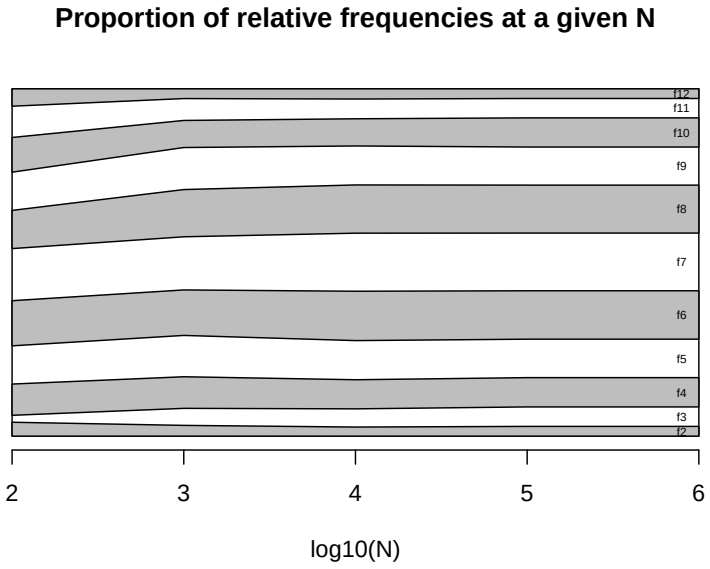


New results came out and f_{12} is now only 0.027, and so we are only $\sim 3\%$ sure to get 12 in the next experiment. The probabilities measured by f_i change with N.

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.

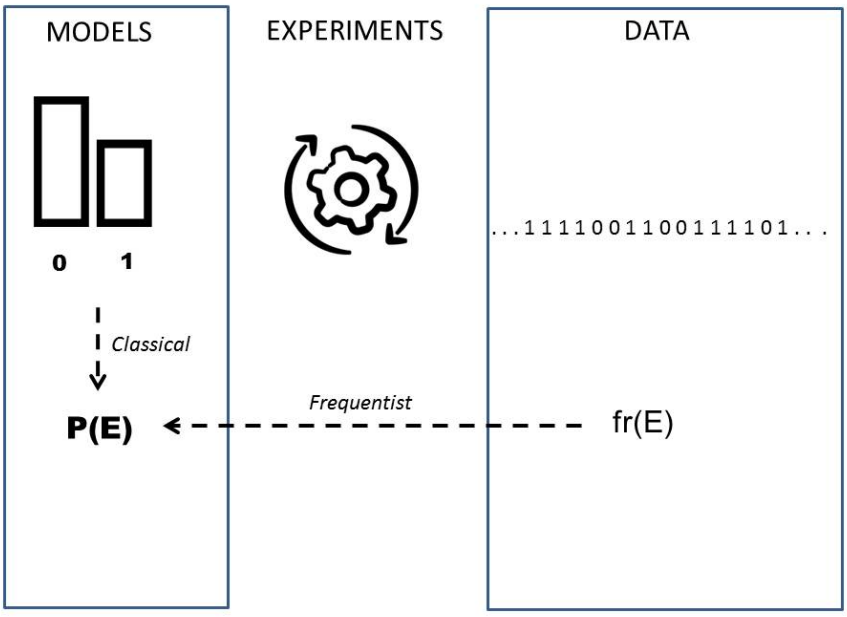
Example (sum of two dice)

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

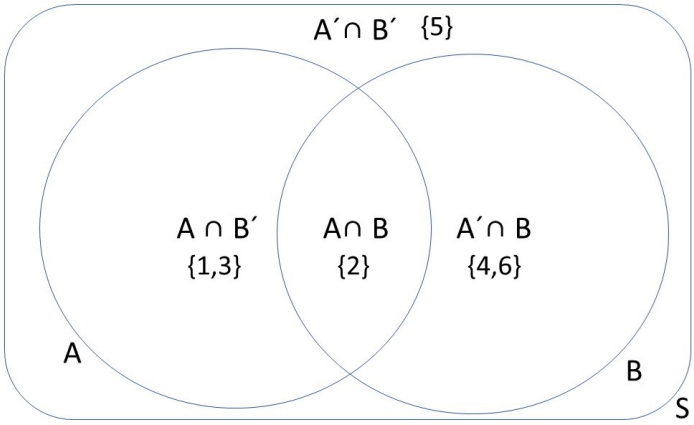
$$P(A \cap B)$$

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 $P(A \cup B \cup C)$?

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 their intersection

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

Now imagine randomized experiments that measure two different types of outcomes.

- 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: $(\bar{R}_1 , \bar{R}_2)$

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 events often tend to 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.

Example (conductor)

We conducted an experiment to find out if observing structural flaws in a material affects its conductivity.

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

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

Definition:

The *conditional probability* of an event B given an event A, denoted $P(A|B)$, 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, the materials with stuctural flaw.

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 marginal and conditional probabilities are the same!

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

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

We conclude that low conductivity 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 ones.

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 the product of the marginal ones. For example: $P(F)P(L) = P(L \cap F)$ and $P(L')P(F') = P(L' \cap F')$. Therefore, low conductivity is independent of having a structural flaw.

Example (Coins)

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

result	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(s) with results

- satisfactory (yes)
- unsatisfactory (not)

with a test (t) 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 is infected.

4.7 Diagnostic test

Let's consider diagnosing an infection with a new test. Infection status:

- yes (infected)
- no (not infected)

Test:

- positive
- negative

The **conditional contingency table** is what we get in a controlled environment (laboratory)

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

Let's look at the table entries 1) Rate of true positives (Sensitivity): The probability of testing positive **if** you have the disease $P(positive|yes)$

2) Rate of true negatives (Specificity): The probability of testing negative **if** you do not have the disease $P(negative|no)$

3) False positive rate: the probability of testing positive **if** you do not have the disease $P(positive|no)$

4) False negative rate: the probability of testing negative **if** you have the disease $P(negative|yes)$

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

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.

The conditional contingency table is

	Infection: yes	Infection: No
Test: positive	P(positive| yes)= 0.7	P(positive|no)=0.06
Test: negative	P(negative| yes)= 0.3	P(negative|no)=0.94
sum	1	1

Therefore, the errors in the diagnostic tests were:

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

4.8 Inverse probabilities

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

$$P(si|positive)$$

For that:

1. We 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(positive | yes)P(yes)	P(positive | no)P(no)	P(positive)
Test: negative	P(negative | yes)P(yes)	P(negative | no) P(no)	P(negative)
sum	P(yes)	P(no)	1

2. We use the definition of conditional probabilities for rows instead of columns (we divide by the marginal of the test results)

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

	Infection: yes	Infection: No	sum

For example:

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

To apply this formula we need the marginals $P(yes)$ (prevalence) and $P(positive)$.

- The prevalence $P(yes)$ needs to be given 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 positives $P(positive)$, we can then use the definition of marginal and conditional probability:

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

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

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

4.9 Bayes' Theorem

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

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

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

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

Or **assess** a test in a controlled condition (infection) and then use it to **infer** the probability of the condition when the test is positive.

Example (COVID):

The test performance was:

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

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:

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

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.

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

Substituting all values gives

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

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

Bayes's theorem

In general, we can have more than two conditioning events. Therefore, Baye's theorem says:

If $E1, E2, ..., Ek$ are k mutually exclusive and exhaustive events and B is any event, then the probability inverse $P(Ei|B)$ is

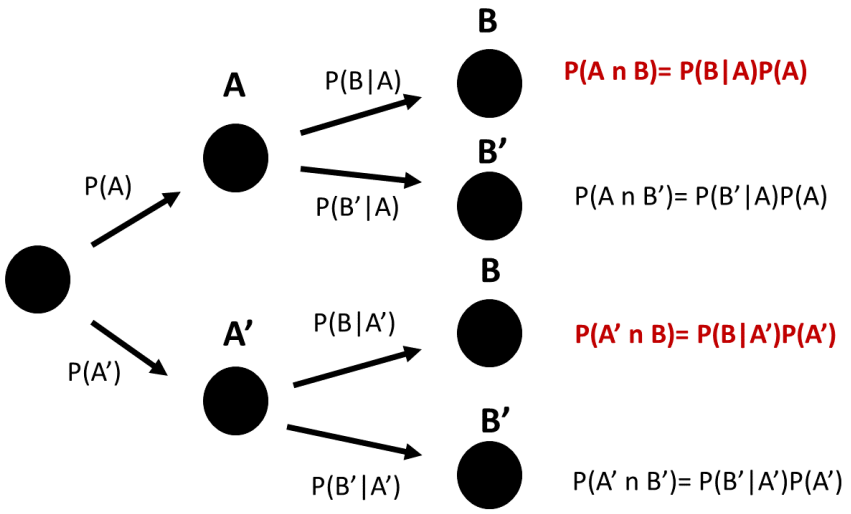
$$P(Ei|B) = \frac{P(B|Ei)P(Ei)}{P(B|E1)P(E1) + ... + P(B|Ek)P(Ek)}$$

The denominator is the total probability rule for the marginal $P(B)$, in terms of the marginals $P(E1), P(E2), ...P(Ek)$.

$$P(B) = P(B|E1)P(E1) + ... + P(B|Ek)P(Ek)$$

Conditional tree

The total probability rule can also be illustrated using a **conditional** tree.



Rule of total probability for the marginal of B: In how many ways can I get the result B?

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

4.10 Questions

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

	junior	*senior*
1*st*	14	12
2*nd*	21	18
3*rd*	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? (R: 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? (R: 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? (R: 200/345)
- What is the probability that the rod came from machine 2 if it fails at least one of the quality controls? (R:0.72)

4.11.0.4 Exercise 4

We want to cross an avenue with two traffic lights. The probability of finding the first red light is 0.6. If we stop at the first traffic light, the probability of

stopping at the second is 0.15. While the probability of stopping at the second if we don't stop at the first is 0.25.

When we try to cross both traffic lights:

- What is the probability of having to stop at each traffic light? (R:0.09)
- What is the probability of having to stop at at least one traffic light? (R:0.7)
- What is the probability of having to stop at a single traffic light? (R:0.61)
- If I stopped at the second traffic light, what is the probability that I would have to stop at the first? (R: 0.47)
- If you were to stop at any traffic light, what is the probability that you would have to stop twice? (R: 0.12)
- Is stopping at the first traffic light an independent event from stopping at the second traffic light? (No)

Now, we want to cross an avenue with three traffic lights. The probability of encountering the first red light is 0.6, and the probability of encountering a red light at the second signal depends solely on the probability of the first light. Similarly, the probability of encountering a red light at the third signal depends only on the probabilities of the second light. As previously mentioned, the probability of stopping at a traffic light is 0.15 if we stopped at the previous light. If we didn't stop at the previous one, the probability of stopping at a traffic light is 0.25.

- What is the probability of having to stop at each traffic light? (R:0.013)
- What is the probability of having to stop at at least one traffic light? (R:0.775)
- What is the probability of having to stop at a single traffic light? (R:0.5425)

tips:

- If the probability that a traffic light is red depends only on the previous one, then

$P(R_3|R_2, R_1) = P(R_3|R_2, \bar{R}_1) = P(R_3|R_2)$ and $P(R_3|\bar{R}_2, R_1) = P(R_3|\bar{R}_2, \bar{R}_1) = P(R_3|\bar{R}_2)$

- The joint probability of finding three red lights can be written as:

$P(R_1, R_2, R_3) = P(R_3|R_2)P(R_2|R_1)P(R_1)$

4.11.0.5 Exercise 5

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

- Pass the quality test: E, fail the quality test: $\bar{E}$
- Defective: D, non-defective: $\bar{D}$

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

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

Chapter 5

Discrete Random Variables

5.1 Objective

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

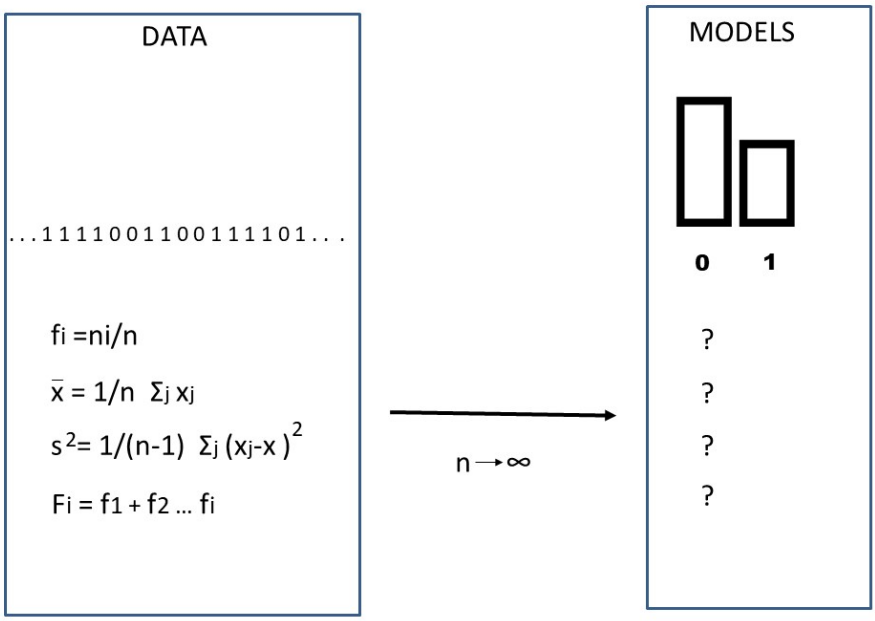
We will define the probability mass function and its main properties of mean and variance. Following the abstraction process of the relative frequencies into probabilities we also define the probability distribution as the limiting case of the relative cumulative frequency.

5.2 Relative frequencies

Relative frequencies of the outcomes of a random experiment are a measure of their propensity. We can used them as estimators of their probabilities, when the 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 on the **observations** of the experiments. 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).

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** of a random experiment.

Example:

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

5.4 Events of observing a random variable

We make the distinction between variables in the model space with capital letters, as abstract entities, and the realization of a particular event or outcome. For instance:

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

...

In general:

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

5.5 Probability of random variables

We are interested in assigning probabilities to the events of observing a particular value of a random variable.

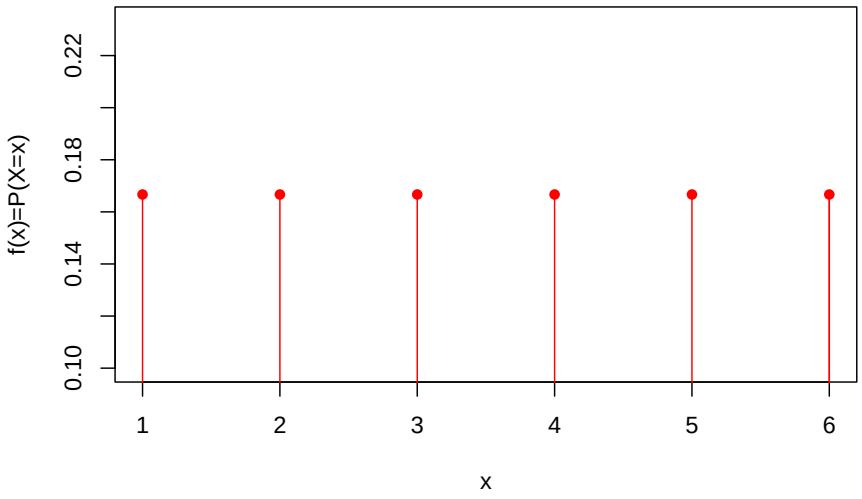
For instance for the dice we will write the probability table as

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 $X = x$.

5.6 Probability functions

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



or written as the mathematical function

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

5.7 Probability 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

- $f(x_i) \geq 0$

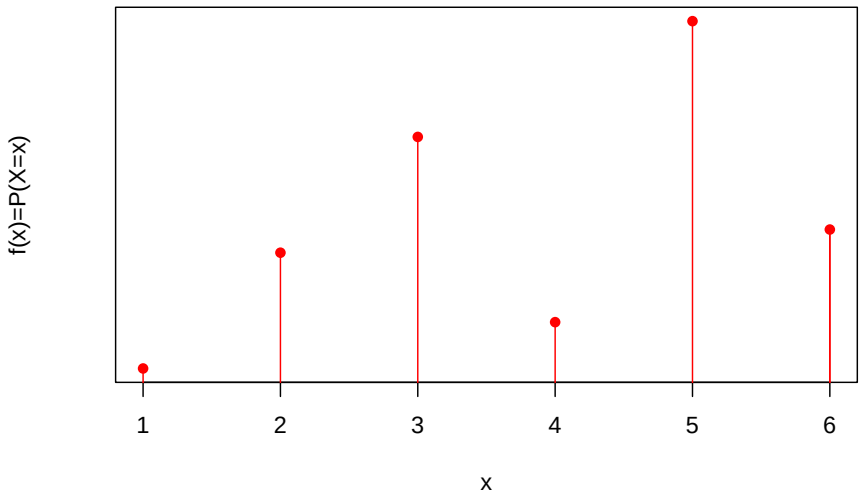
and its sum over all the values of the variable is 1:

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

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 map to an experiment. 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.

5.8 Probabilities and relative frequencies

Consider the example

Make following set-up: In one urn put 8 balls and:

- mark 1 ball with -2
- mark 2 balls with -1
- mark 2 balls with 0
- mark 2 balls with 1
- mark 1 ball with 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 and write the frequency table

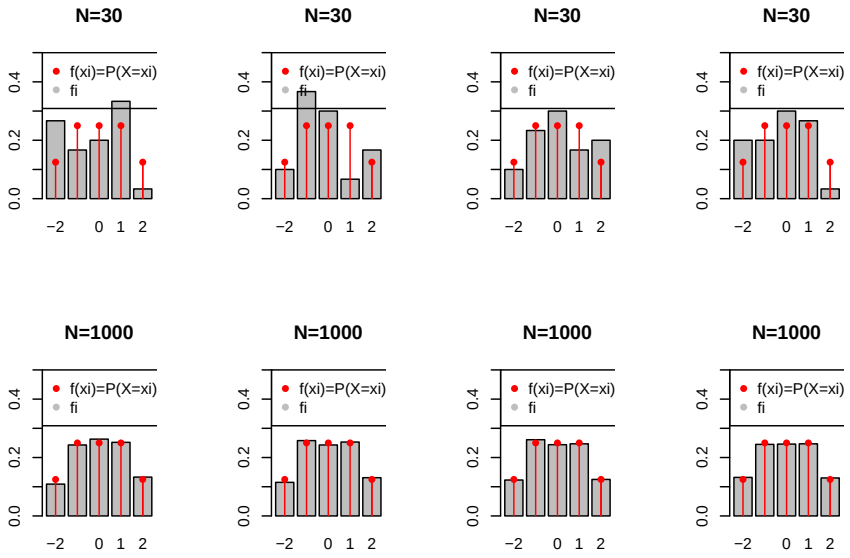
X	f_i
-2	0.132
-1	0.262
0	0.240
1	0.248
2	0.118

The frequentist probability tells us

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

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. But P_i is an abstract quantity that never changes. As N increases we get closer to it.

5.9 Mean or expected value

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

We used the **average** to measure the centre of gravity of the **data**. In terms of the relative frequencies of the values of discrete outcomes, we wrote the average as

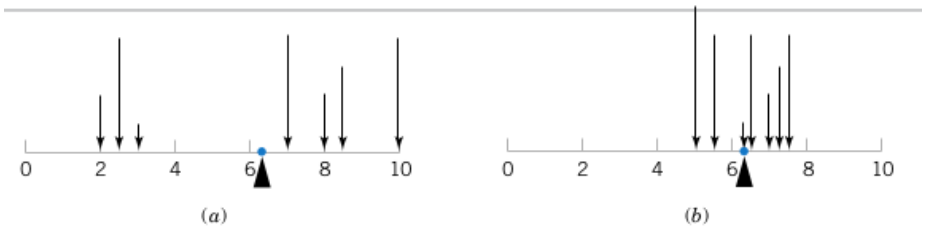
$\bar{x} = \sum_{i=1}^{M} x_i \frac{n_i}{N} =$

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

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 gravity of the **probabilities**: The point where probability loading on a road are balanced.

From the definition we have

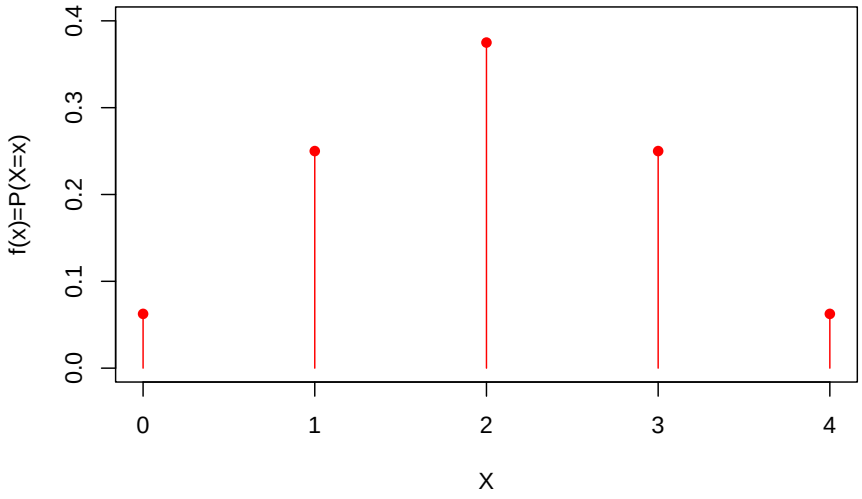
$$\bar{x} \to \mu$$

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 centre of gravity of the probability mass function **it does not change**. However, the average $\bar{x}$ is the centre of gravity 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.

Teh variance is the spread of the **probabilities** about the mean: The moment of inertia of probabilities about the mean.

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_i. F_i is an important quantity because it is a continuous function F_x is therefore a **continuous** function, even if the outcomes are discrete.

Definition:

The **probability distribution** function 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) $0 \leq F(x) \leq 1$
2) If $x \leq y$, then $F(x) \leq F(y)$

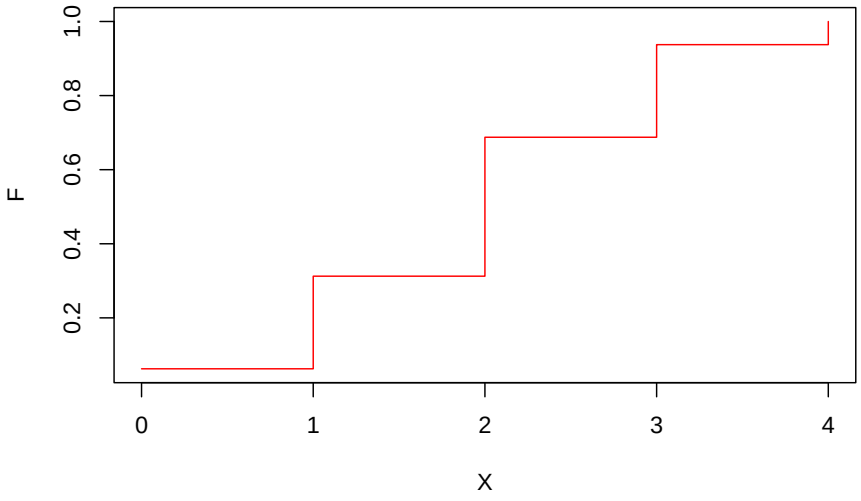
For the probability mass function:

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:

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

For$X \in \mathbb{Z}$



5.13 Probability function and probability distribution

The probability function and distribution are equivalent. 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 radom variable X.

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

$$q = \sum_{i=1}^{p} f(x_i) = F(x_p)$$

- The **median** is value x_m such that $q = 0.5$

$$F(x_m) = 0.5$$

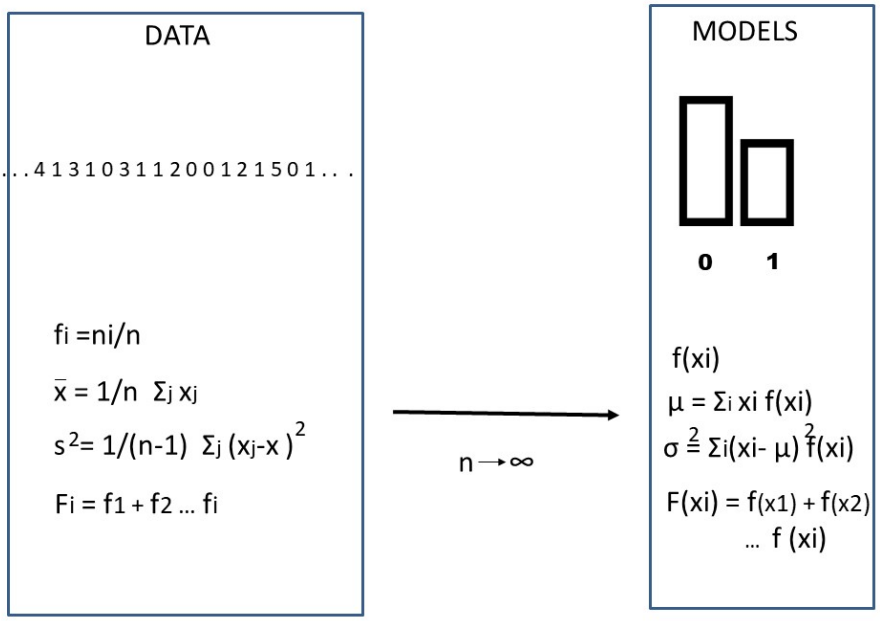
- The 0.05-quantile is the value x_r such that $q = 0.05$

$$F(x_r) = 0.05$$

- The 0.25-quantile is **first quartile** the value x_s such that $q = 0.25$

$$F(x_s) = 0.25$$

5.15 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	$\sigma^2 = V(X) = \sum_{i=1}^{M} (x_i - \mu)^2 f(x_i)$	$s^2 = \sum_{j=1}^{N} (x_j - \bar{x})^2 / (N-1)$

quantity names	model (unobserved)	data (observed)
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

We place ballots with letters from a to f in an urn. Consider the drawing that gives 0 euros to the first two letters of the alphabet, 1.5 euros to the next two, and 2 and 3 euros to the following ones.

a) What are the probability mass function and the probability distribution for the money prizes in the game?

b) What is the expected value of the prize? (R: 1.3)

c) What is the variance of the prize? (R: 1.13)

d) What is the probability of winning 2 or more euros? (R: 2/6)

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? (R: 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 $V(X)$ (R:1; 1.5)
- what is the expected value and variance of $Y = 2X + 3$ (R: 6)
- what is the median and the first and third quartiles of X? (R: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? (R: 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? (R:1.5)
- What is its variance? (R: 0.75)
- Sketch the probability distribution
- What is the probability of transmitting at least 1 error? (R:7/8)

Chapter 6

Continous Random Variables

6.1 Objective

In this chapter we will study continous 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 frequencies for the intervals are the probabilities when $N \to \infty$

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

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 therefore

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

when $\Delta x \to 0$, because $n_i \to 0$

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 x between 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 of observing **any** value is 1:

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

3) The probability of observing a value 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

$$dx \to 0$$

All the properties of probabilities are translated in terms of densities

$$\sum \to \int$$

Probability densities are mathematical quantities that do not necessarily represent random experiments.

A fundamental interest in statistics is to describe the densities that 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 $100cm$.

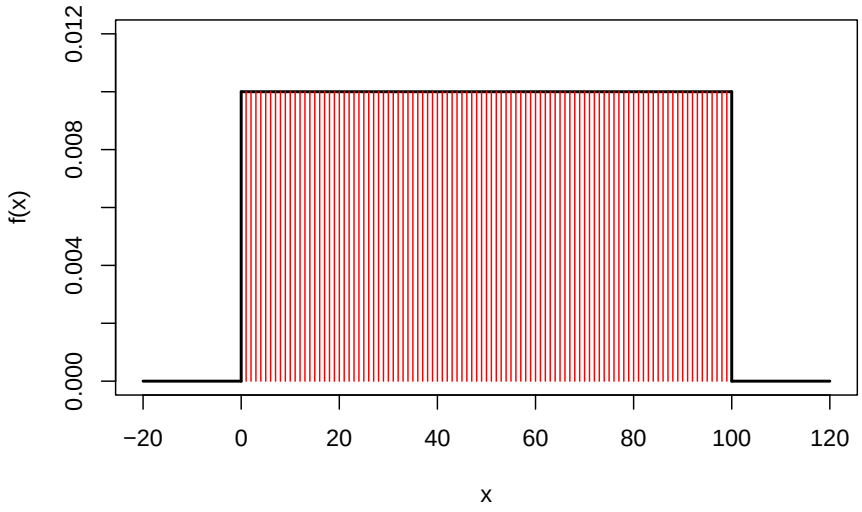
$$f(x) = \begin{cases} \frac{1}{100}, & \text{if } x \in (0, 100) \\ 0, & otherwise \end{cases}$$

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

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

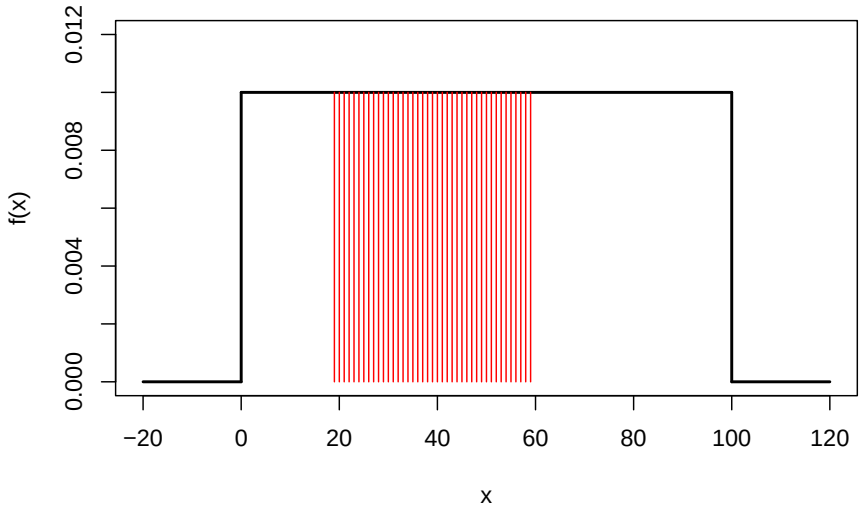
2) The probability of observing **anything** is the total **area under the curve**

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



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

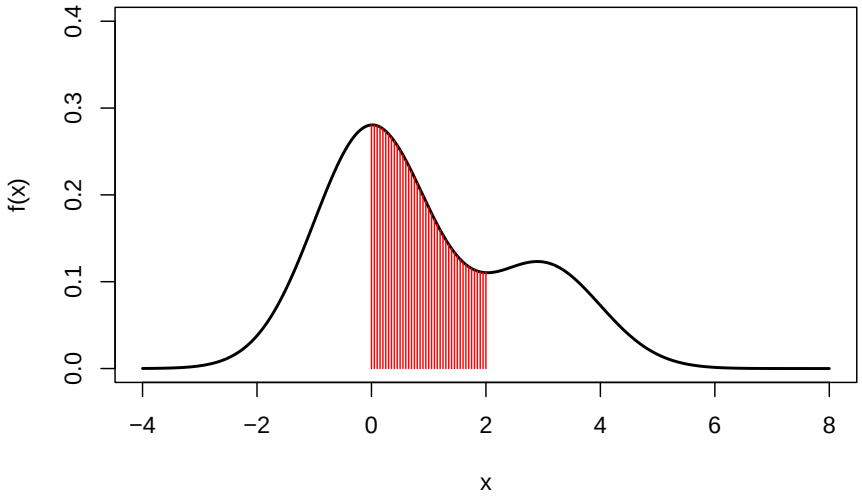
For continuous variables, we compute the probability that the variable is in a given interval. That is

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

We saw that for continuous variables, the probability that the experiment gives us a particular real number is zero: $P(X = a) = 0$

The $P(a \leq X \leq b)$ 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** $F(c)$ defined as the accumulation of probability up to the outcome C

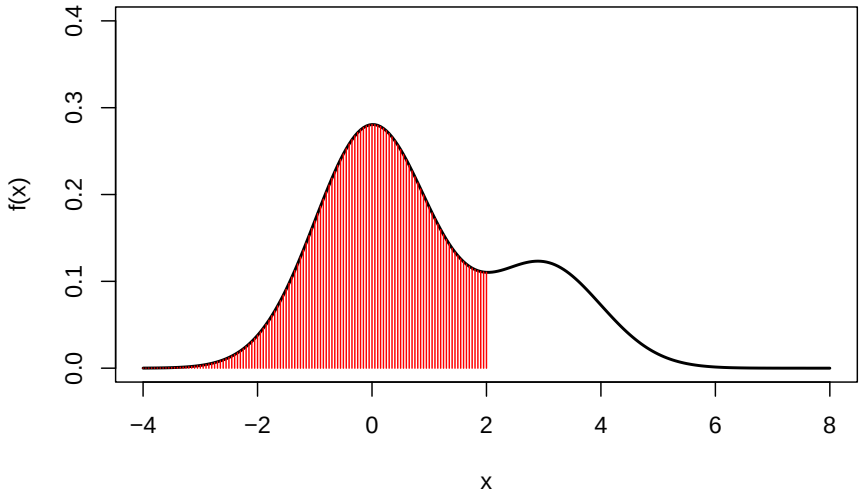
$F(c) = P(X \leq c)$

can be used to compute the probability $P(a \leq X \leq b)$.

Consider:

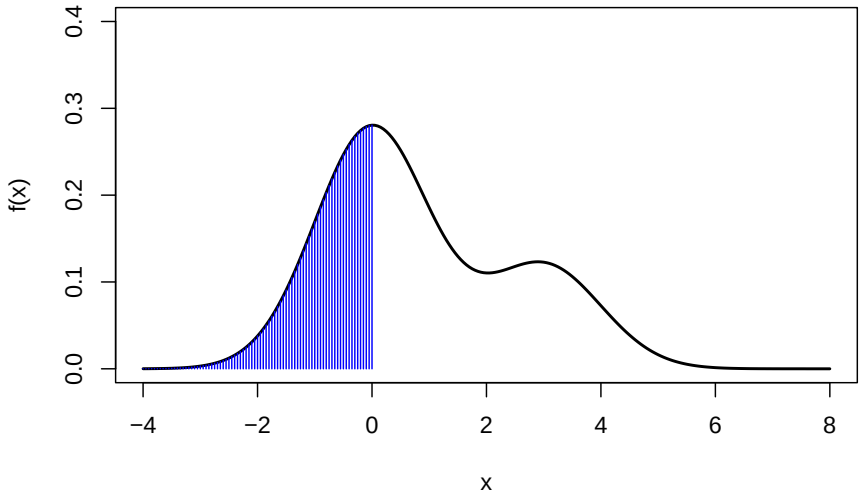
1) that the probability accumulated up to b is given by

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



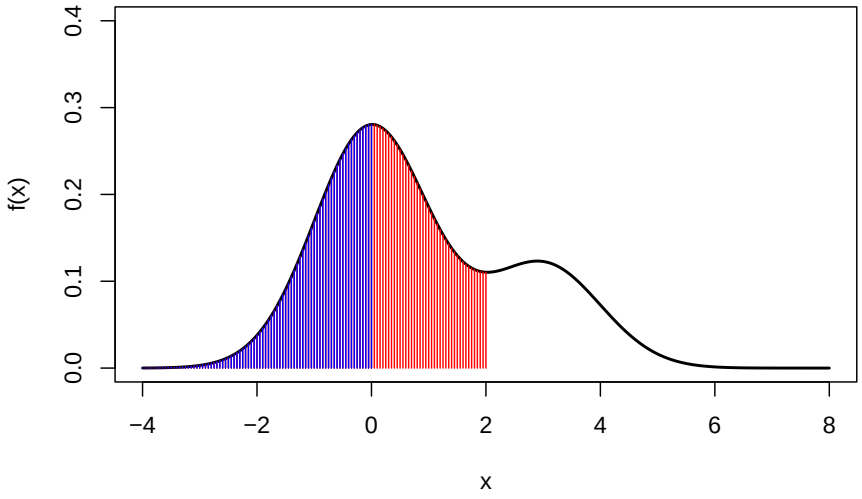
2) that the probability accumulated up to a is

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



Then the probability between a and b is given by the difference in the value of the probability distribution

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



Definition

The **probability distribution** of a continuous random variable is defined as

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

and have 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, and its derivative is the probability density function.

Example (raindrop fall)

For the uniform density function:

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

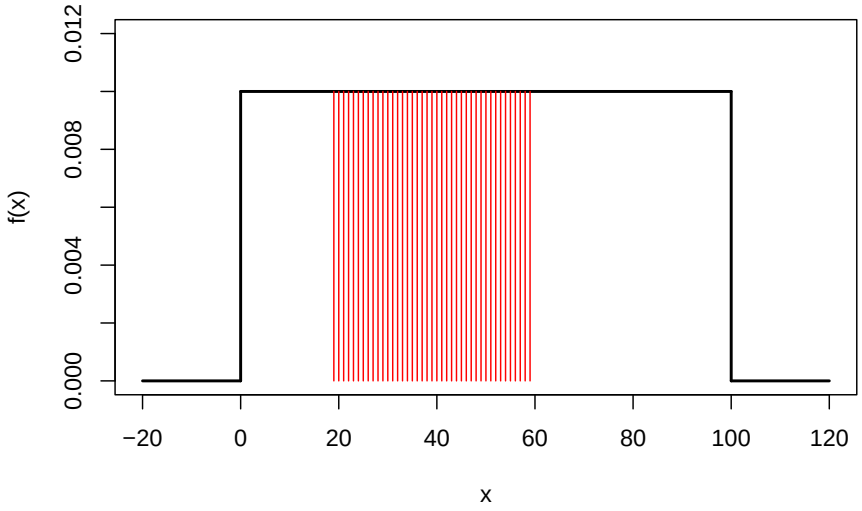
We find that the probability distribution is

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

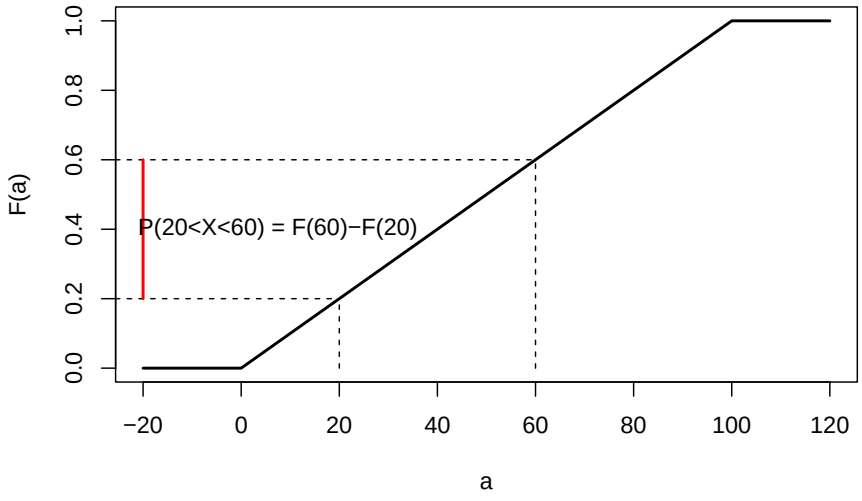
6.8 Probability plots

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

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



2) 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.

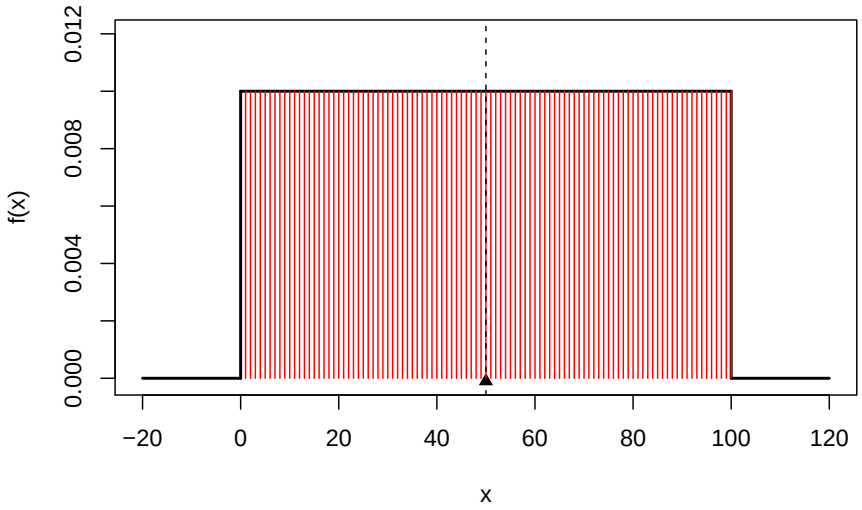
Example (raindrop fall)

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

$$E(X) = 50$$



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:

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

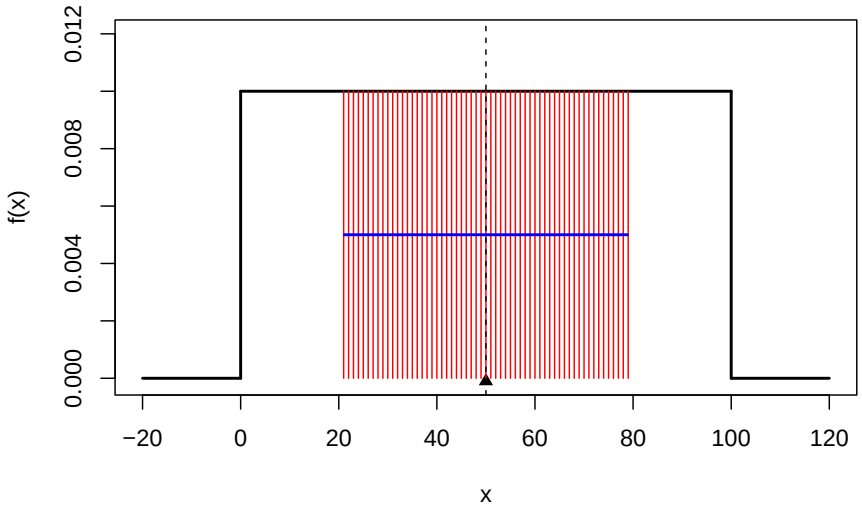
6.12 Exercises

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 (R:50)
- compute variance using $E(X^2) = V(X) + E(X)^2$ (R:100^2/12)
- compute $P(\mu - \sigma \leq X \leq \mu + \sigma)$ (R: 0.57)
- What are the first and third quartiles? (R: 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? (R: 1/15; 1/5)
- what is the mean of X? (R: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 (R: $1/\lambda$)
- Compute the expected value of X^2 (R: $2/\lambda^2$)
- Compute variance (R: $1/\lambda^2$)

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

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:

- $P(X > 0)$ (R:63/80)
- $E(X)$ (R:1.93)
- $P(X > 0 | X < 2)$ (R: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 thus 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 functions.

7.2 Probability mass function

Let us remenber 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 of of obtaining anything in 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 previous example 1, 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 previous example 2, 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 functions that depend on M: $f(x; M)$.

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

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

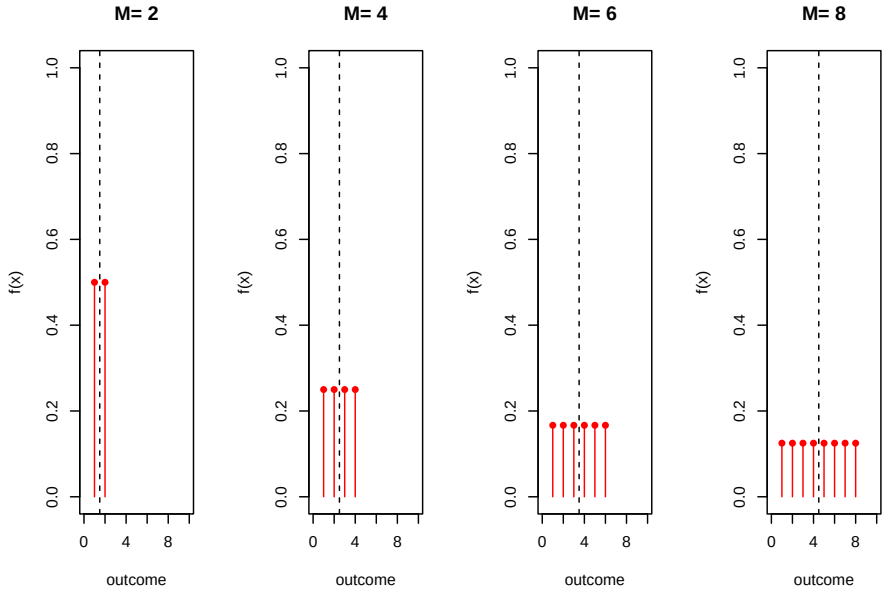
and

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

Note: $E(X)$ and $V(X)$ are also **parameters**. If we know any of them then we can fully determine the distribution. For instance:

$$f(x) = \frac{1}{2E(X) - 1}$$

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



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 $M = b - a + 1$.

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

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

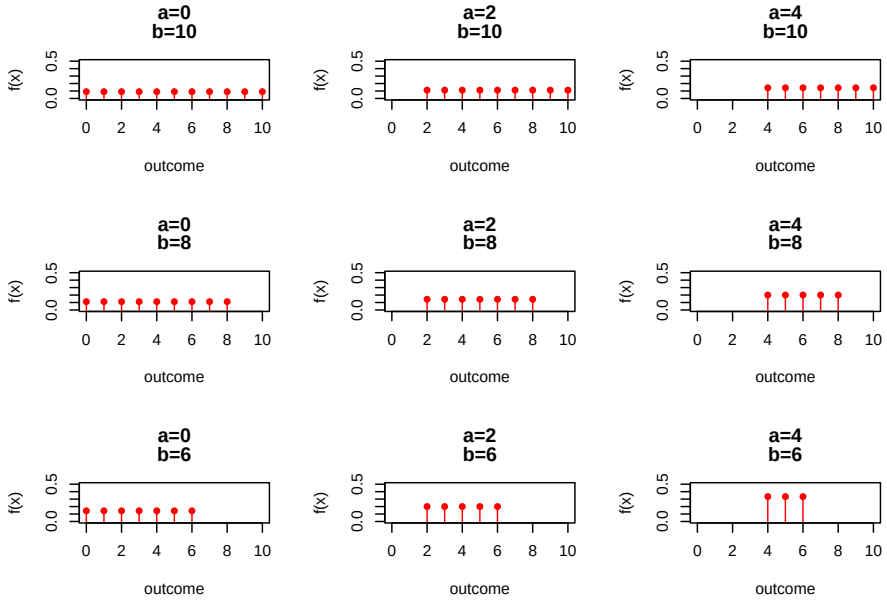
2) Its variance is

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

To prove this change variables $X = Y + a - 1$, $y \in \{1, ...M\}$.

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

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

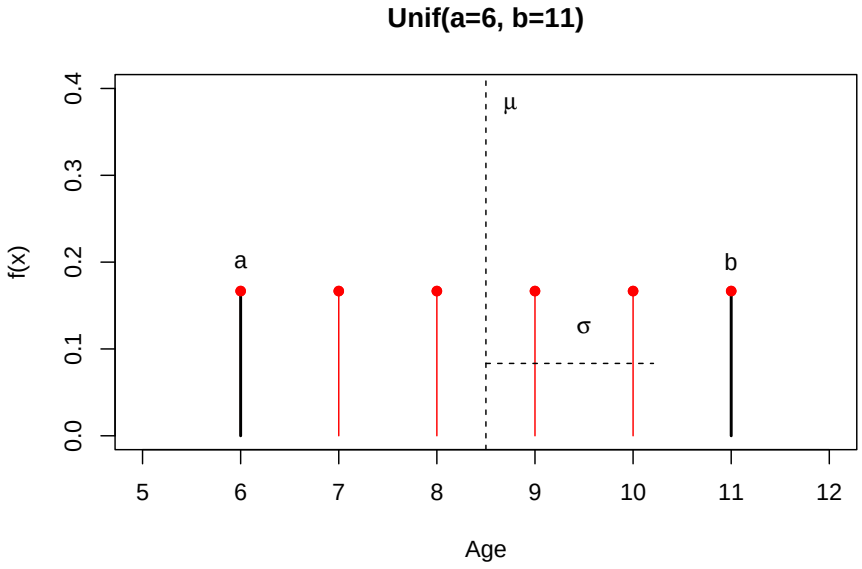
for $x \in \{6, 7, 8, 9, 10, 11\}$, and 0 otherwise.

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 value 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 B) that have **unequal** probabilities

Examples:

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

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:

$Outcome$	P_i
A	p
B	$1-p$

- $i \in \{A, B\}$
- outcome A (success): has probability p (parameter)
- outcome B (failure): has a probability $1-p$

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

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

3) As a concise probability mass function

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

for $k = (0, 1)$.

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

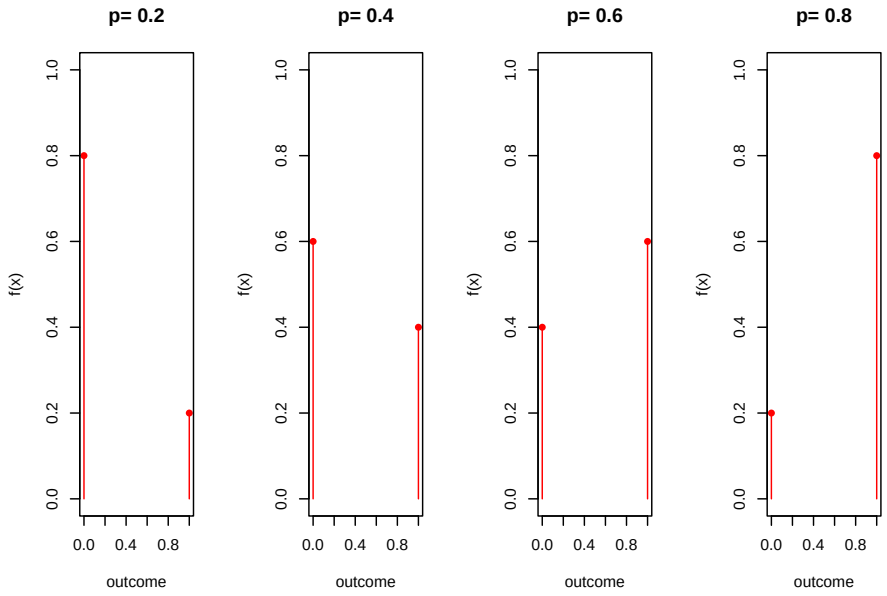
$$E(K) = p$$

2) its variance is

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

Note that the probability of the outcome A is the parameter p which is the same as its value at 1: $f(1) = P(X = 1)$. 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** when we know the parameter p of particular 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 a **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.

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 $B : failure$). 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 correctly ($A : yes$ and $B : no$). 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 that we obtain common or rare? what is its probability mass function $P(X = x) = f(x)$?

Example (transmission of pixels):

What is the probability of observing $X = x$ errors when transmitting $n = 4$ pixels, if the probability of an error is p?

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

2) Each

$$K_i \rightarrow Bernoulli(p)$$

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

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

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

$x \in \{0, 1, 2, 3, 4\}$. For example $X = 2$ for the outcome $(0, 1, 0, 1)$.

Now let's see the probabilities of some **error events** 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:

$$P(X = 4) = P(1, 1, 1, 1) = p * p * p * p = p^4$$

because K_i are **independent**.

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 probability of observing 3 errors in **different events**:

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

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

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

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

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

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

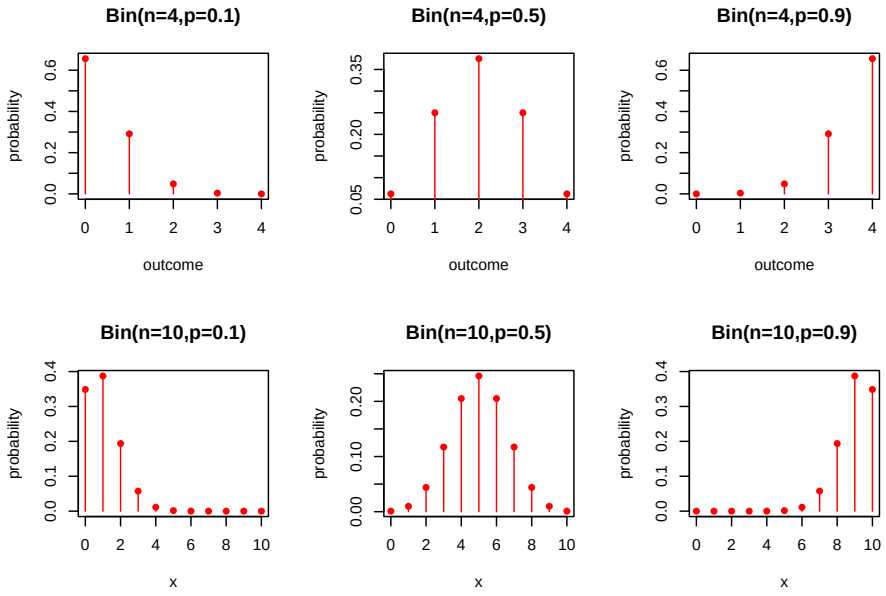
$\binom{n}{x} = \frac{n!}{x!(n-x)!}$ is called **the binomial coefficient** and gives the number of ways one can obtain x events of type A in a set of n.

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 $X \to Bin(n, p)$ 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

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

Example (transmission of pixels):

- The expected value for the number of errors in the transmission of 4 pixels is $np = 4 * 0.1 = 0.4$ 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 R dbinom(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 R dbinom(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

$$f(x) = \binom{10}{x} 0.9^x (0.1)^{4-x}$$

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

in R pbinom(8,10, 0.9)

7.10 Negative binomial probability function

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

Our random experiment is now: Repeat Bernoulli trials until we observe the outcome A appears r times.

The outcome of the experiment is the number of events $n_B = y$

We are interested in finding the probability of observing a particular number of events B, $P(Y = y)$, where $Y = N_B$ is the random variable.

Example (transmission of pixels):

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

Let's first find the probability of **one particular transmission event** with y number of correct pixels (B) 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 bit 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

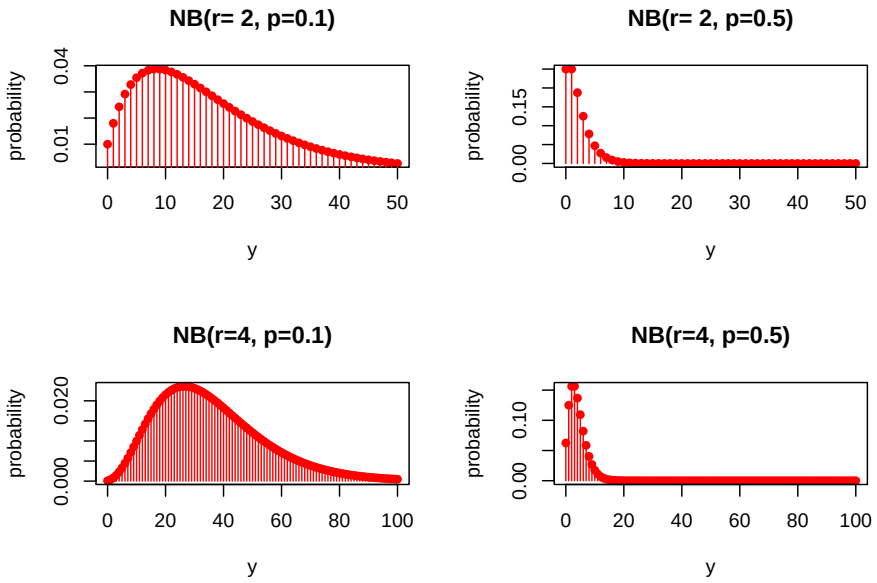
1) mean

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

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 B (successful requests) then

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

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 R this is computed with pnbinom(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 R dnbinom(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 R dnbinom(2, 4, 0.5)

7.11 Geometric distribution

We call **geometric distribution** to the **negative binomial** distribution with $r = 1$

The probability of observing B 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 B. 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

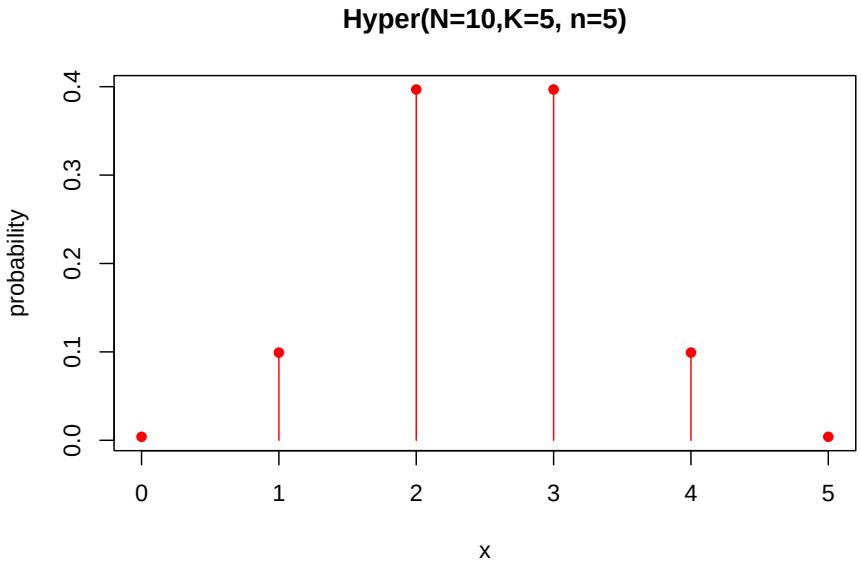
$$E(X) = n\frac{K}{N} = np$$

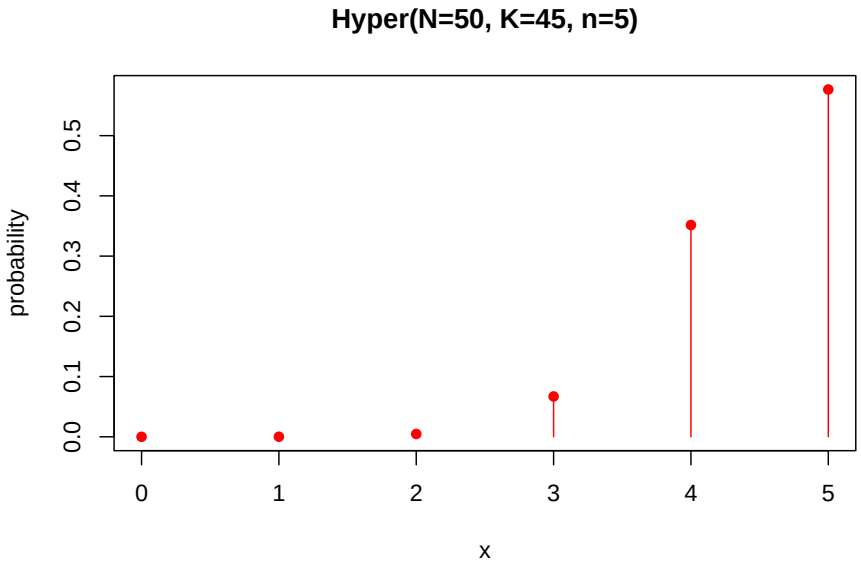
2) and variance

$$V(X) = np(1-p)\frac{N-n}{N-1}$$

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):

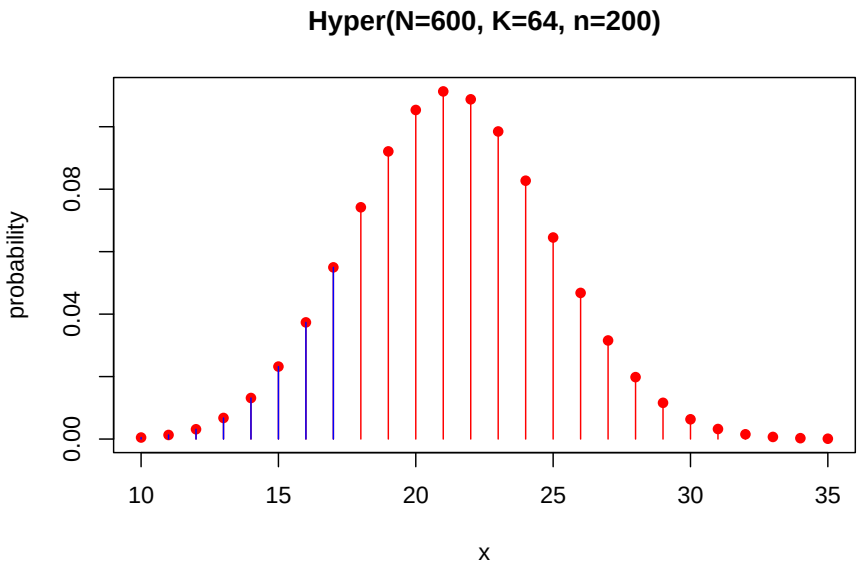
- what is the probability of infections less or equal than 17 in a sample of 200, drawn from a population of 600 where 64 are infected?

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 R phyper(17, 64, 600-64, 200)=0.140565

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:pbinom(x=49, n=100, p=0.55)=0.13; **b:**1-pbinom(x=49, n=100, p=0.55)=0.86; **c:**pbinom(x=51, n=100, p=0.45)=0.90; **d:**1-pbinom(x=51, n=100, p=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:dbinom(x=5, n=10, p=0.25)~ 0.05; **b:**pbinom(x=5, n=10, p=0.75)~ 0.07; **c:**dbinom(x=5, n=10, p=0.75)~ 0.05; **d:**1-pbinom(x=5, n=10, p=0.25)~ 0.02

7.14 Exercises

7.14.0.1 Exercise 1

If the 25% of screws produced by a machine are defective, determine the probability that, out of 5 randomly chosen screws:

- none of the screws is defective (R: 0.2373)
- 1 screw is defective (R: 0.3955)
- 2 screws are defective (R: 0.2636)
- at most 2 screws are defective (R: 0.8964)

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?(R: 0.240)
- What is the probability that a family has only one girl?(R: 0.259)
- What is the probability that a family has only one boy or only one girl?(R: 0.4999)
- What is the probability that the family has at least two boys?(R: 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?(R:$n = 3 > \log(0.25)/\log(0.51)$)

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?(R:0.1385651)
- What is the probability that the engine successfully completes 15 searches before the first failure?(R: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?(R: 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

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 $\downarrow$ **Bernoulli**: Introduction of a **parameter** p (family of models) $\downarrow$ **Binomial**: Introductio of the **Repetition** of a random experiment (n-times Bernoulli trials) $\downarrow$ **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 your 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 in one given centimeter of the wire (X).

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 specimen of one centimeter in particular?

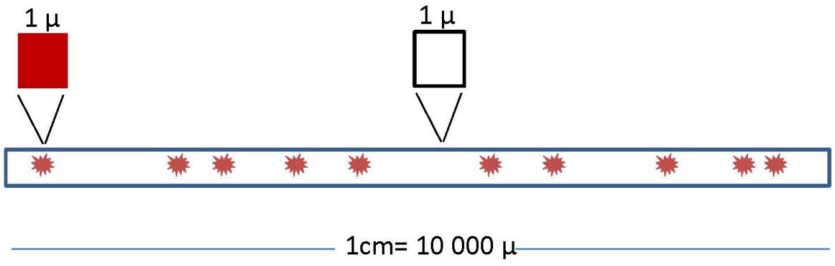
8.4 Poisson probability mass function

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 $\lambda = np$. 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:

$$f(x) = \frac{e^{-\lambda}\lambda^x}{x!}$$

Where λ is constant because 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) counts occur at random in the interval
3) the average number of counts 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 Bernoulli trials.

Then the random variable X that counts events across the interval is a **Poisson** variable with probability mass function

$$f(x) = \frac{e^{-\lambda}\lambda^x}{x!}, \lambda > 0$$

Properties: When $X \to Poiss(\lambda)$ it has

1) mean
$$E(X) = \lambda$$
2) and variance
$$V(X) = \lambda$$

Examples

1) What is the probability of receiving 4 emails in an hour, when the average number of emails in an hour is 1?

We have that the variable is Poisson: $X \to Poiss(\lambda)$ with $\lambda = 1$ and its probability mass function is:

$$f(x) = \frac{e^{-1}1^x}{x!}$$

Therefore the probability that the variable takes value 4 is $P(X = 4)$:

$f(4; \lambda = 1) = \frac{e^{-1}1^4}{4!} = 0.01532831$

in R dpois(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 R dpois(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;

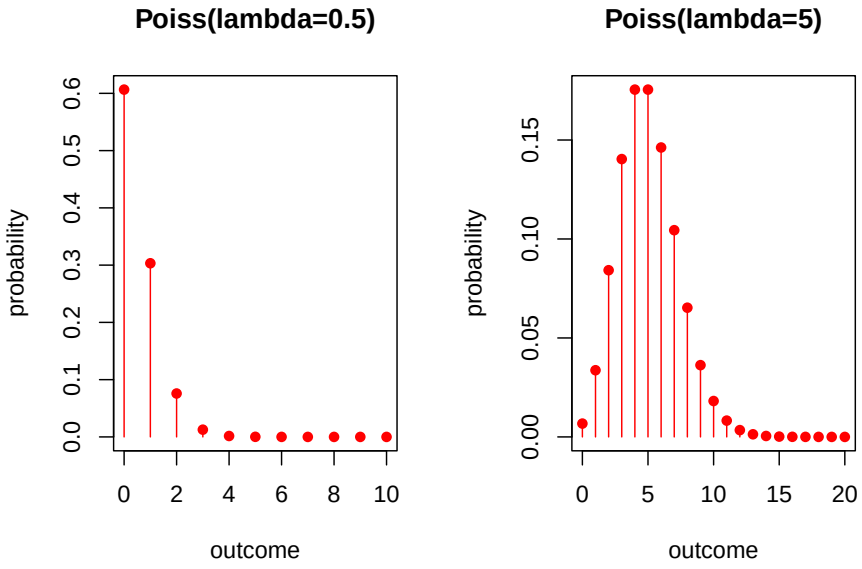
We have that the variable is Poisson: $X \to Poiss(\lambda)$ with $\lambda = 5$ and its probability mass function is:

$$f(x) = \frac{e^{-5}5^x}{x!}$$

$P(X \geq 10) = 1 - P(X < 10) = 1 - P(X \leq 9) = 1 - F(9; \lambda = 5) = 1 - \sum_{x=0,...10} f(x; \lambda = 5) = 0.03182806$

in R 1-ppois(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.

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 duration/time we should wait for the **first** count to occur.

We can ask about the probability that the first event occurs after duration/time X.

Therefore, since X is a **continuous** random variable, we look for its probaility density function $f(x)$.

8.7 Exponential probability density

The probability of 0 counts **if** an interval has unit x is

$$f(0|x) = \frac{e^{-x\lambda}x\lambda^0}{0!}$$

or

$$f(0|x) = e^{-x\lambda}$$

We can treat this as the conditional probability of 0 events in a distance x: $f(K = 0|X = x)$ and apply the Bayes theorem to reverse it:

$$f(x|0) = Cf(0|x) = Ce^{-x\lambda}$$

So we can calculate the **probability of observing a distance** a distance x with 0 counts. This is the distance between any two events or the distance until the first event.

Definition

In a Poisson process with parameter λ the probability of waiting a distance/time X between two counts is given by the **probability density**

$$f(x) = Ce^{-x\lambda}$$

- C is a constant that ensures: $\int_{-\infty}^{\infty} f(x)dx = 1$
- by integration $C = \lambda$

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 $X \to Exp(\lambda)$ then

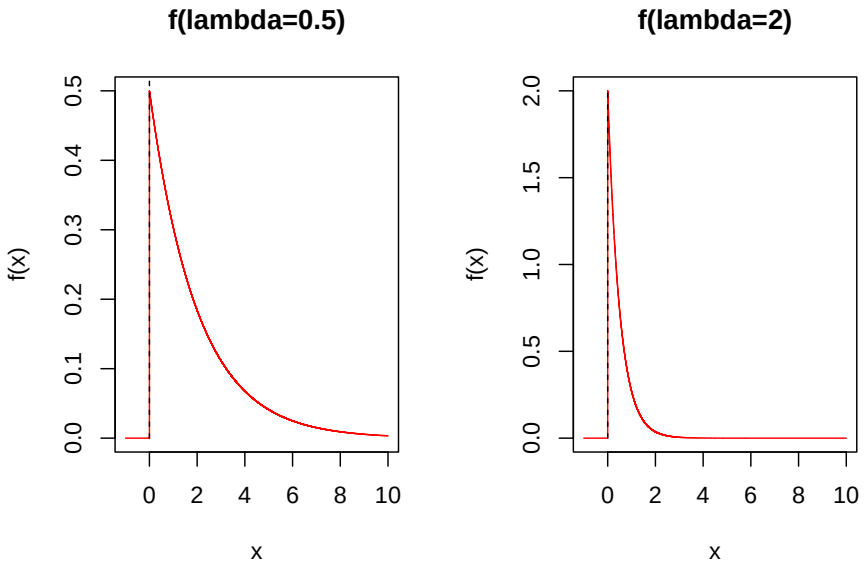
1) It has mean

$$E(X) = \frac{1}{\lambda}$$

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 count?

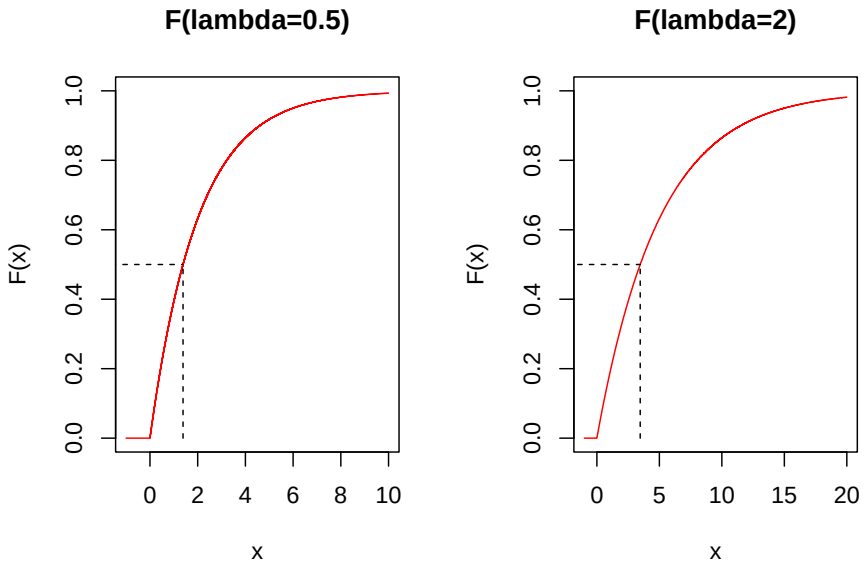
Remember that this probability $F(a) = P(X \leq a)$ is the probability density

$$F(a) = \lambda \int_{\infty}^{a} e^{-x\lambda} dx = 1 - e^{-a\lambda}$$

2) In a Poisson process ¿What is the probability of observing an interval **larger** than size a until the first event?

$$P(X > a) = 1 - P(X \leq a) = 1 - F(a) = e^{-a\lambda}$$

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?

$$P(X > 1) = 1 - P(X \leq 1) = 1 - F(1, \lambda = 2) = 0.1353353$$

in R 1-pexp(1,2)

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 R pexp(2,2)

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

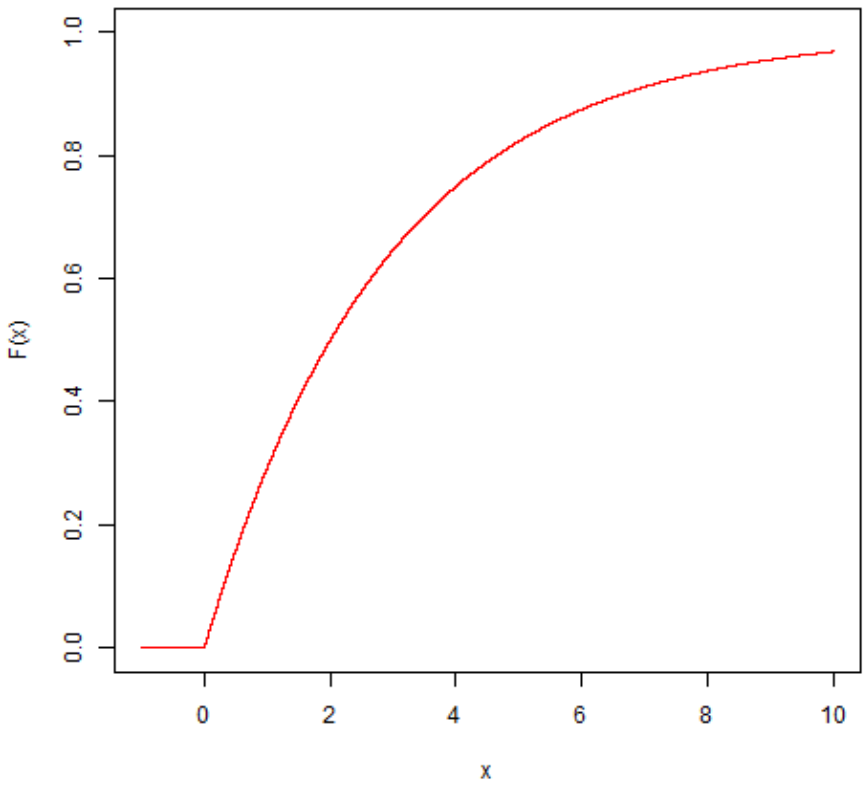
a:1-ppois(x=2, lambda=0.92); **b:**ppois(x=2, lambda=0.92); **c:**1-dpois(x=2, lambda=0.92); **d:**dpois(x=2, lambda=0.92)

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 minute? (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 data obtained for the position of the Ceres, a large asteroid between Mars and Jupiter.

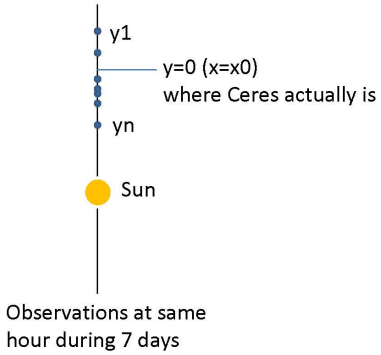
At the time people suspected it was a new planet, as it moved day to day against the fixed stars. In January, it could be seen at the horizon just before dawn. However, as days passed Ceres will rise latter and latter until it could not longer be seen because of the Sun rise.

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 a planet or a big asteroid.

Data was available only for the month of January. After which Ceres would disappear. He wanted to **predict** where astronomers should point their telescopes to find it six months after at dusk, once it had passed behind the Sun.

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) the error at a distance $-\epsilon$ from the value at the position of Ceres was equally likely as a distance ϵ

3) the most **likely** (that we believe) of Ceres at any given time in the sky was the **average** of multiple altitude measurements at that latitude.

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 actual position of Ceres.

2) and variance

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

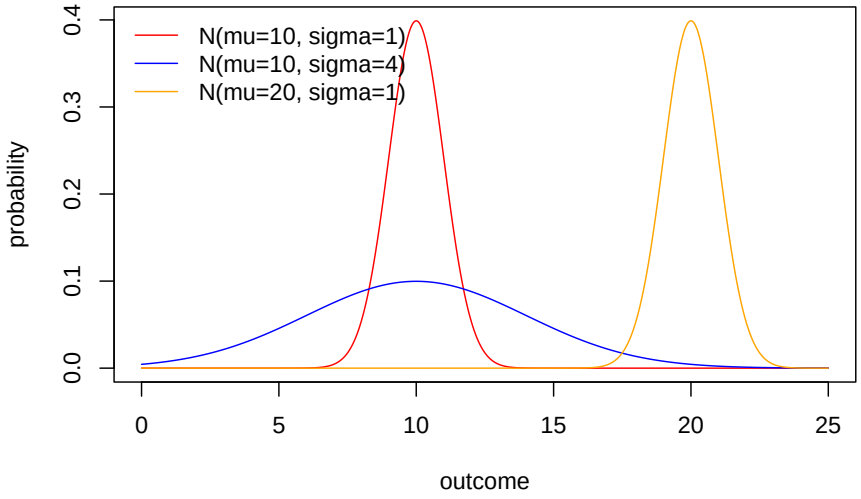
which represented the dispersion of the error in the observations, which depends on the quality of the telescope.

μ and σ are the **two parameters** that completely describe the normal density function and their **interpretation** depends on the random experiment.

When X follows a Normal density, that is, it is normally distributed, we write

$$X \to N(\mu, \sigma^2)$$

Let's look at some probability densities in the normal parametric model



9.5 Probability distribution

The probability distribution of the Normal density:

$$F_{normal}(a) = P(Z \leq a)$$

is the **error** function defined by the area under the curve from $-\infty$ to a

$$F_{normal}(a) = \int_{-\infty}^{a} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$

The function is found in most computer programs, and it does not have a closed form of known functions.

Example (women's 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 $8cm$?

$P(X \leq 150) = F(150, \mu = 165, \sigma = 8) = 0.03039636$

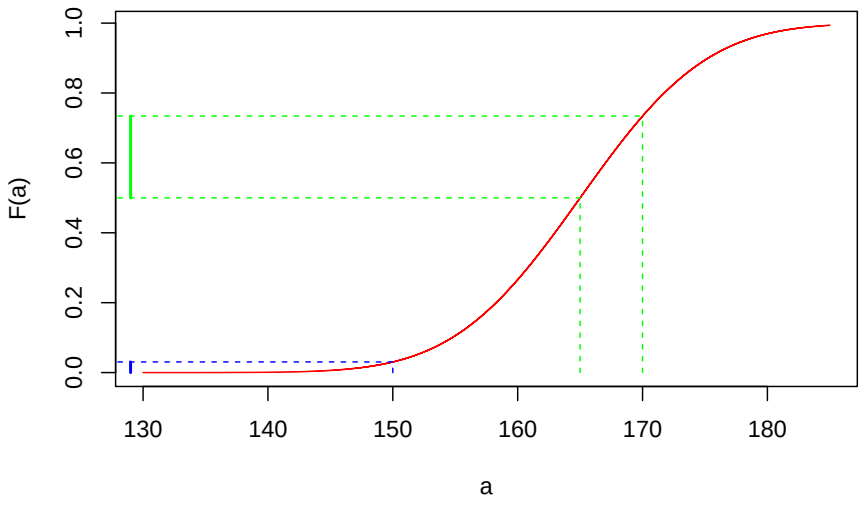
in R pnorm(150, 165, 8)

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 R pnorm(170, 165, 8)-pnorm(165, 165, 8)

Let's look at the probability distribution function



3) What is the first quartile for height in women?

The first quartile is defined as:

$F(x_{0.25}, \mu = 165, \sigma = 8) = 0.25$

or

$x_{0.25} = F^{-1}(0.25, = 165, \sigma = 8) = 159.6041$

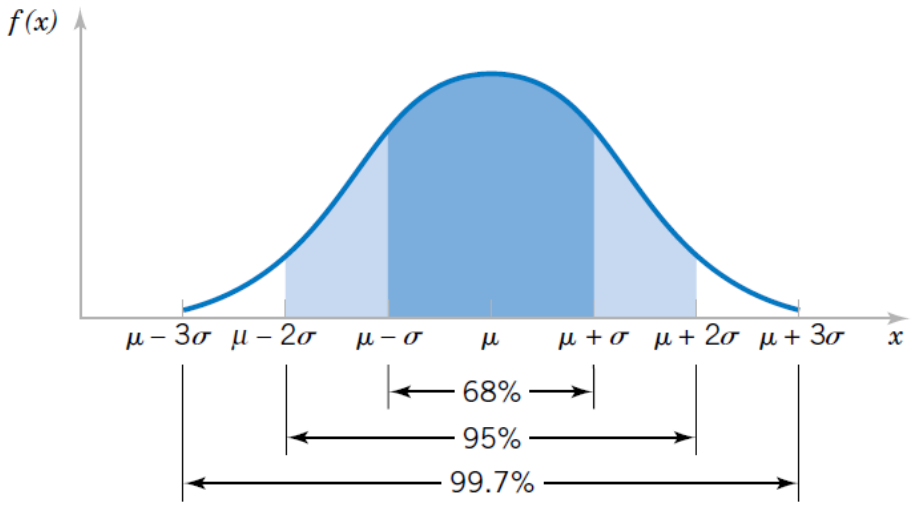
in R qnorm(0.25, 165, 8)

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** 5%

3) x values that fall farther than 3σ are considered **extremely rare** 0.2%



Example (women's height)

We can define the limits of **common observations** for the distribution of women's height in the population.

1) at a distance of one standard deviation from the mean, we find 68% of the population

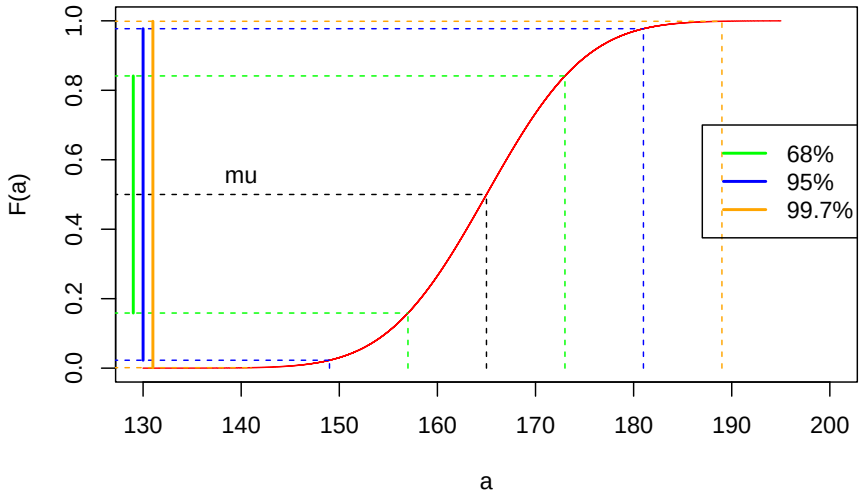
$$P(165 - 8 \leq X \leq 165 + 8) = P(157 \leq X \leq 173) = F(173) - F(157) = 0.68$$

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

$$P(165 - 3 \times 8 \leq X \leq 165 + 3 \times 8) = F(189) - F(141) = 0.997$$



9.6 Standard normal density

The standard normal density is the particular density from the normal family

$$f(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, x \in \mathbb{R}$$

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 it distributes normally and write

$$X \to N(0,1)$$

9.7 Standard distribution

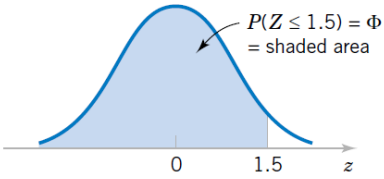
The probability distribution of the standard density:

$$\phi(a) = F_{N(0,1)}(a) = P(Z \leq a)$$

is the **error** function defined by

$$\phi(a) = \int_{-\infty}^{a} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz$$

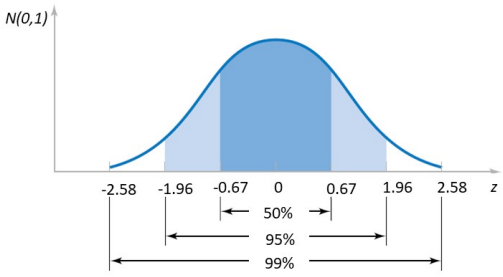
Because the standard distribution is special and will appear often then we use the letter ϕ for it.



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 R is pnorm(x) with the default parameters, 0 and 1.

We usually define the limits of the **most common observations** for the standard variable



1) The interquartile range

$$P(-0.67 \leq X \leq 0.67) = 0.50$$

in R: c(qnorm(0.25), qnorm(0.75))

2) 95% range

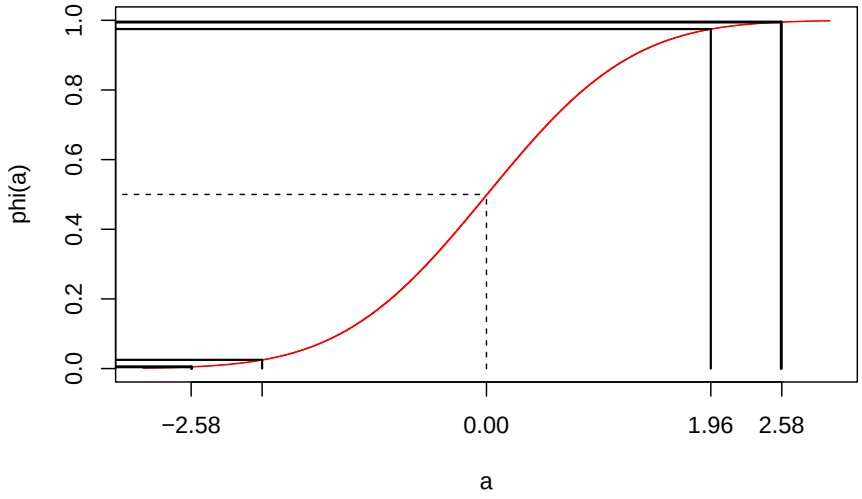
$$P(-1.96 \leq X \leq 1.96) = 0.95$$

in R: c(qnorm(0.025), qnorm(0.975))

3) 99% range

$$P(-2.58 \leq X \leq 2.58) = 0.99$$

in R: c(qnorm(0.005), qnorm(0.995))



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**

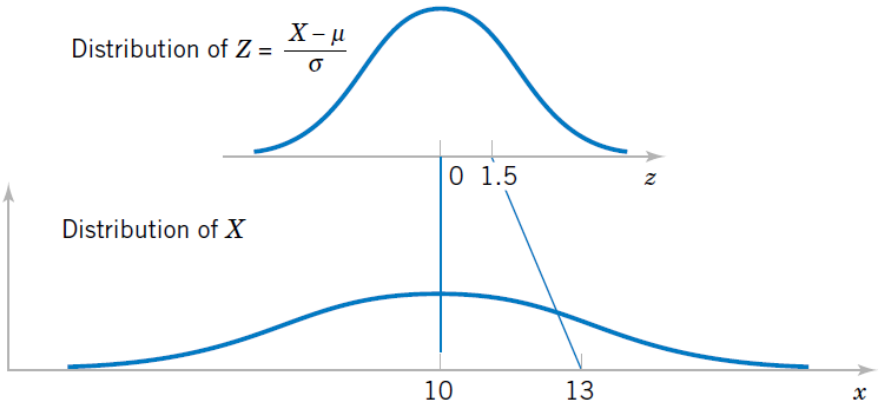
$$Z = \frac{X - \mu}{\sigma}$$

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 expression we have

$$P(x \leq X \leq x + dx) = P(z \leq Z \leq z + dz)$$

$$= \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$

$$= \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz$$

The probability of **any normal variable** $X \to N(\mu, \sigma^2)$ can be computed 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 can use the property of the probability distributions

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

$$= \phi(\frac{b - \mu}{\sigma}) - \phi(\frac{a - \mu}{\sigma})$$

9.9 Summary of probability models

Model	X	range of x	f(x)	E(X)	V(X)
Uniform	integer or real number	$[a, b]$	$\frac{1}{n}$	$\frac{b+a}{2}$	$\frac{(b-a+1)^2-1}{12}$
Bernoulli	event A	0,1	$(1-p)^{1-x}p^x$	p	$p(1-p)$
Binomial	# of A events in n repetitions of Bernoulli trials	0,1,...	$\binom{n}{x}(1-p)^{n-x}p^x$	np	$np(1-p)$
Negative Binomial for events	# of B events in Bernoulli repetitions before r As are observed	0,1,..	$\binom{x+r-1}{x}(1-p)^x p^r$	$\frac{r(1-p)}{p}$	$\frac{r(1-p)}{p^2}$
Hypergeometric	# A events in a sample n from population N with K As	$\max(0, n+K-N)$, ... $\min(K, n)$	$\frac{1}{\binom{N}{n}}\binom{K}{x}\binom{N-K}{n-x}$	$n * \frac{N}{K}$	$n\frac{N}{K}(1-\frac{N}{K})\frac{N-n}{N-1}$
Poisson	# of events A in an interval	0,1, ..	$\frac{e^{-\lambda}\lambda^x}{x!}$	λ	λ
Exponential	Interval between two events A	$[0, \infty)$	$\lambda e^{-\lambda x}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Normal	measurement 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 R functions of probability models

Model	R
Uniform (continuous)	dunif(x, a, b)
Binomial	dbinom(x,n,p)
Negative Binomial for events	dnbinom(x,r,p)

Model	R
Hypergeometric	dhyper(x, K, N-K, n)
Poisson	dpois(x, lambda)
Exponential	dexp(x, lambda)
Normal	dnorm(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: qnorm(1/3, 10, 2)=9.138545; **b:** qnorm(1-0.75, 10, 2)=8.65102 ; **c:** qnorm(1-1/3, 10, 2)=10.86145 ; **d:** qnorm(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 $z = 1.94$ (R:0.182)
- To the right of $z = -1.28$ (R:0.899)
- To the right of $z = 2.05$ or to the left of $z = -1.44$ (R: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? (R:0.841)
- What is the probability that a man's height is between 165cm and 185cm? (R:0.682)
- What is the height that defines the 5% of the smallest men? (R:158.55)

Chapter 10

Sampling distributions

10.1 Objective

In this chapter, we are going to 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 Aleatory sample

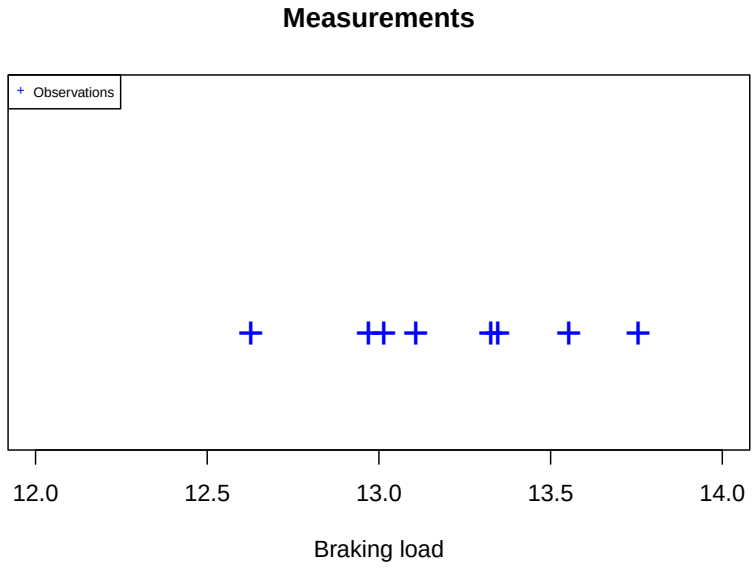
Example (Cables)

Imagine that a client asks your metallurgical company to sell cables for 8 that can transport up to 96 Tons; that's 12 Tons each. You must guarantee that none of them will break with this weight.

You **stock** a set of cables that might work, but you're not sure. So you take 8 Cables at random, and charge 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 **replication** of a random experiment n **independent** times.

- A random sample is an **random variable** n-dimensional

$$(X_1, X_2, ...X_n)$$

where X_i is the *i-th* iteration of the random experiment with common distribution $f(x;\theta)$ for any i

- **An observation** of a random sample is the set of n values obtained from the experiments

$$(x_1, x_2, ...x_n)$$

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
```

Example (Cables)

In the observed sample of the breaking load of the cables it was observed that

1) None of them 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 responsible for a cable break and has to pay a large fine?

To assure the customer that the cables will not 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 of 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 gravity of the **probabilities**, where $f(x_i)$ is the probability function. This definition was motivated by the center of gravity of the **observations**

$$\bar{x} = \frac{1}{n} \sum_{i}^{n} x_i = \sum_{i}^{m} x_i f_i$$

which we define 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** (by putting on a **hat**) by the relative frequencies when n is large, because $lim_{n\to\infty} f_i = f(x_i)$. Therefore, we should also have that the **mean** μ can be estimated by the **mean** $\bar{x}$

$$\hat{\mu} = \bar{x} = \sum_i^m x_i \hat{P}(X = x)$$

Thus, we may take the center of the probability function as the center of gravity of the data. Doing this we will make an error that we can assume, 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 $n \to \infty$

$$\hat{\sigma}^2 = s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

and we assume that the moment of inertia of the data is close to the moment of inertia of the probabilities.

Example (Cables)

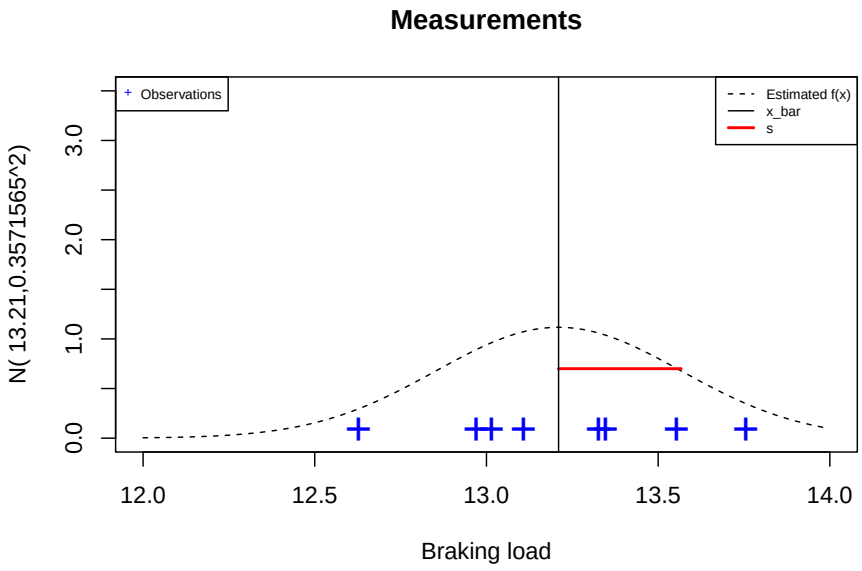
Assuming that the breaking load of our cable is a normal random variable

$$X \to N(x; \mu, \sigma^2)$$

we use the estimates $\bar{x}_{stock} = 13.21$ (mean(x)) and $s^2 = 0.3571565^2$ (sd (x)^2) 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 R pnorm(12,13.21, 0.3571565)= 0.000352188

Given the **observed** sample, there is an estimated probability of 0.03% that a single cable breaks in 12 Tons. We have a probabilistic argument for selling the cables.

10.5 Margin of error of estimates

When we estimate parameters using data, such as by taking the value of

$$\hat{\mu} = \bar{x}$$

for the value of μ; and the value of

$$\hat{\sigma}^2 = s^2$$

by the value of σ^2, we know that we are **making a mistake**. We know that if we take another sample of size 8 cables **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?

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 properties

1) is **unbiased**:
$$E(\bar{X}) = \mu$$
2) is **consistent**:
$$lim_{n\to\infty} V(\bar{X}) = 0$$

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 $V(\sum_{i=1}^{n} X_i) = nV(X)$.

Estimate of 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 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.

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 **error measure**. Therefore, we ask about the **probability of making an error** of a given size when estimating μ with $\bar{x}$.

When we calculate probabilities in 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 in estimating μ with $\bar{x}$.

To calculate probabilities we need

1. A probability model (probability function)
2. The parameters of the model (the values of the probability function)

What are 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 $\sigma^2 = 0.35^2$.

If 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 of 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 distribution

Theorem: If X follows a normal distribution

$$X \to N(\mu, \sigma^2)$$

then $\bar{X}$ is normal

$$\bar{X} \to N(\mu, \frac{\sigma^2}{n})$$

and $\bar{X}$ has

1) mean

$$E(\bar{X}) = \mu$$

We say that $\bar{X}$ is unbiased because its expected value is exaclty μ.

2) variance

$$V(\bar{X}) = \frac{\sigma^2}{n}$$

We say that $\bar{X}$ is consistent because is gets smaller with large n.

We call $se = \sqrt{V(\bar{X})}$ the **standard error** of the sample mean. The standard error is also written as $\sigma_{\bar{x}}$. Note that this is the error we expect when using $\bar{x}$ as the value of μ, and it is the bias we needed to correct for S_n^2.

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**:

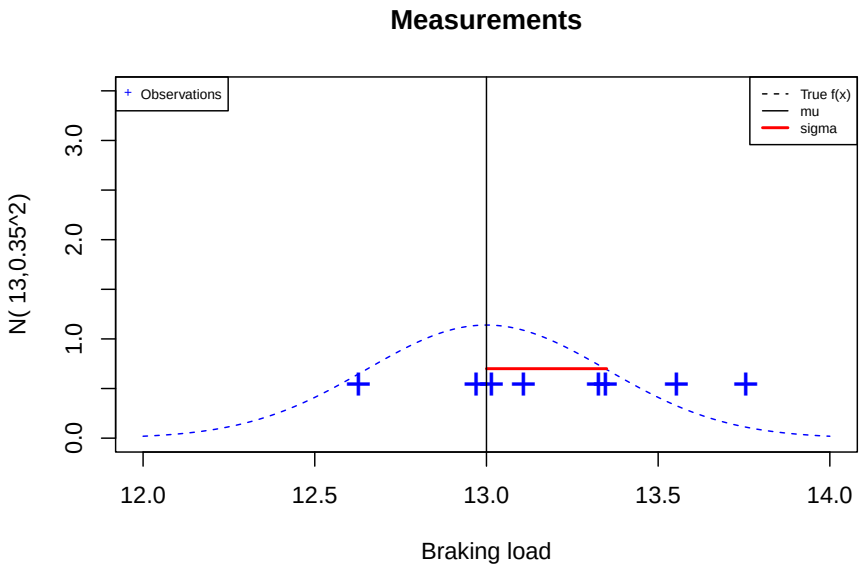
1. The probability function of X is also known as the **population** probability function
2. The probability function of $\bar{X}$ is a probability function of the **sample**.

Example (Cables)

Probability densities for X and $\bar{X}$

In our new problem, we now **know** μ and σ and the probability function of the **population**

$$X \to N(\mu = 13, \sigma^2 = 0.35^2)$$

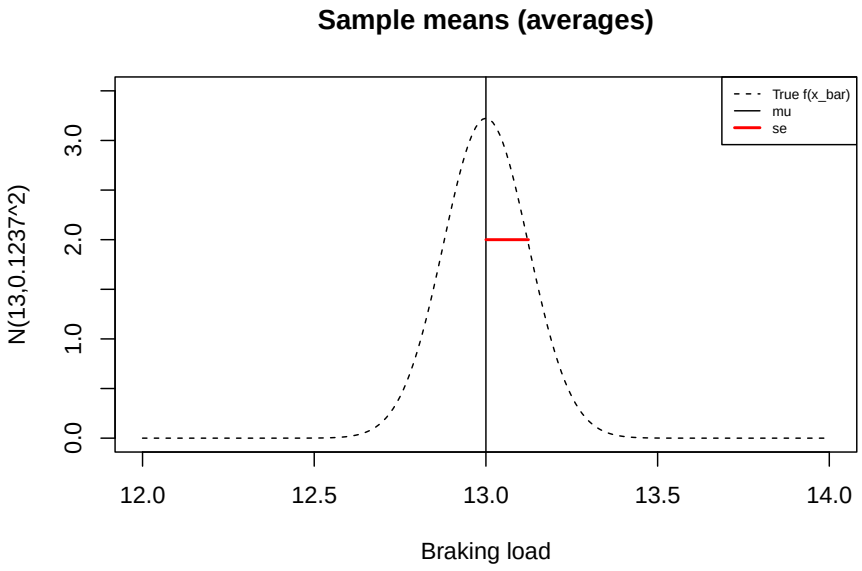


Since X is normal, then $\bar{X}$ is normal, and therefore we also know the probability function of the sample mean $\bar{X}$

$$\bar{X} \to N(13, \frac{0.35^2}{8})$$

which has mean and variance

1) $E(\bar{X}) = \mu = 13$
2) $V(\bar{X}) = \frac{\sigma^2}{n} = \frac{0.35^2}{8} = 0.01530169$



Finally we want to calculate **the probability** that our estimate has a margin of error of 0.25. That is a distance of 0.25 from the mean. That's

$$P(-0.25 \leq \bar{X} - 13 \leq 0.25) = P(12.75 \leq \bar{X} \leq 13.25)$$

$$= F(13.25; \mu, \sigma^2/n) - F(12.75; \mu, \sigma^2/n)$$

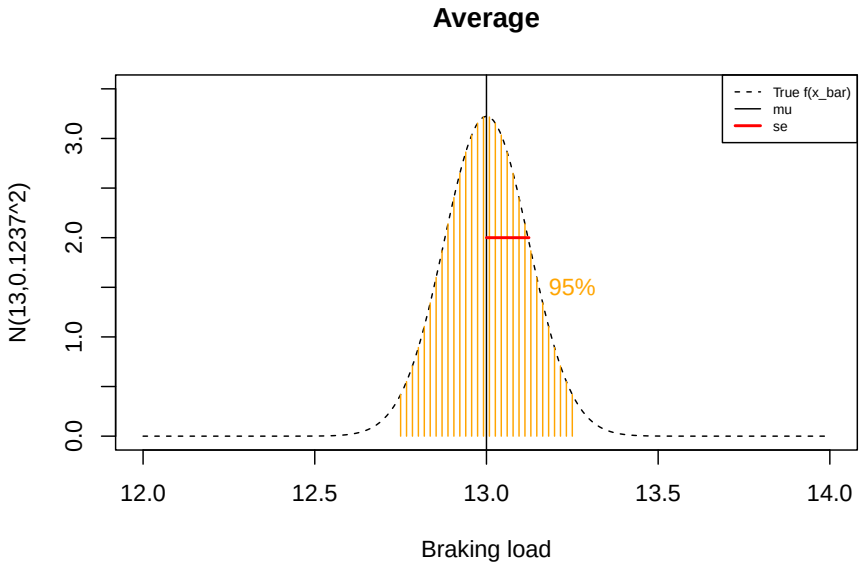
In R we can calculate it as:

pnorm(13.25, 13, 0.1237)-pnorm(12.75, 13, 0.1237)=0.956.

Remember: $se = \sigma_{\bar{x}} = \sqrt{0.01530169} = 0.1237$

Therefore 95.6% of the means $\bar{X}$ of random samples of size 8 are at a distance of 0.25 from the mean $\mu = 13$.

If we sell our process for building the cables, we can tell new manufacturers that when they follow our instructions, they can test the process by taking a sample of size 8 cables. In that case, they can expect the sample mean to fall between $(12.75, 13.25)$ about 95% of the times.



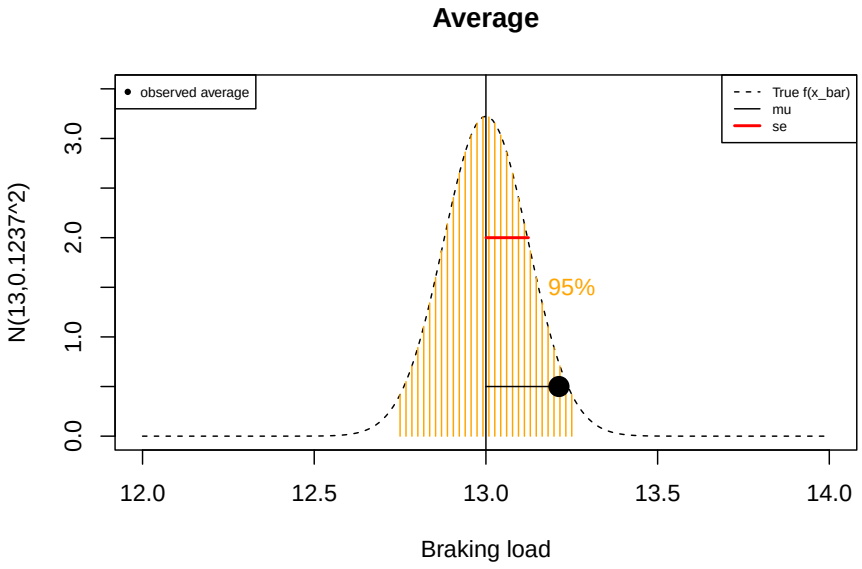
When we perform random sampling we observe:

```
## [1] 13.34642 13.32620 13.01459 13.10811 12.96999 13.55309 13.75557 12.62747
```

Assuming $\mu = 13$ then our **observed error** in estimating the mean is the difference

$$\bar{x}_{stock} - \mu = 13.21 - 13 = 0.21$$

This is within the margin of error of 95.6% and therefore, we must consider that the manufacturing process is working as expected.



10.7.1 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

$$X \to N(\mu, \sigma^2)$$

so Y is normal

$$Y \to N(n\mu, n\sigma^2)$$

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 R we can calculate it as:

pnorm(106, 104, 0.9899495)-pnorm(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 $8 * 0.25 = 2$ Tons from the total mean $n\mu = 104$.

10.8 Sample variance

By estimating the variance

$$s^2 = \hat{\sigma}$$

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: $E(S^2) = V(X) = \sigma^2$
2) consistent: $n \to \infty$, $V(S^2) \to 0$

and therefore S^2 consistently estimates σ^2

We can take a value of s^2 as an estimate for σ^2 or

$$s^2 = \hat{\sigma}^2$$

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 bias, writing equation 1 above as:

$$E(\frac{n}{n-1}S^2_n) = \sigma^2$$

We can define the **sample variance** (corrected)

$$S^2 = \frac{n}{n-1}S^2_n = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

which is an unbiased estimator of σ^2 because $E(S^2) = \sigma^2$.

We can also have inference problems when we are interested in the probability of the **sample variance** S^2.

Consider a quality control process that requires the leads to be 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: 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?

10.9 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

$$f(w) = C_n w^{\frac{n-3}{2}} e^{-\frac{w}{2}}$$

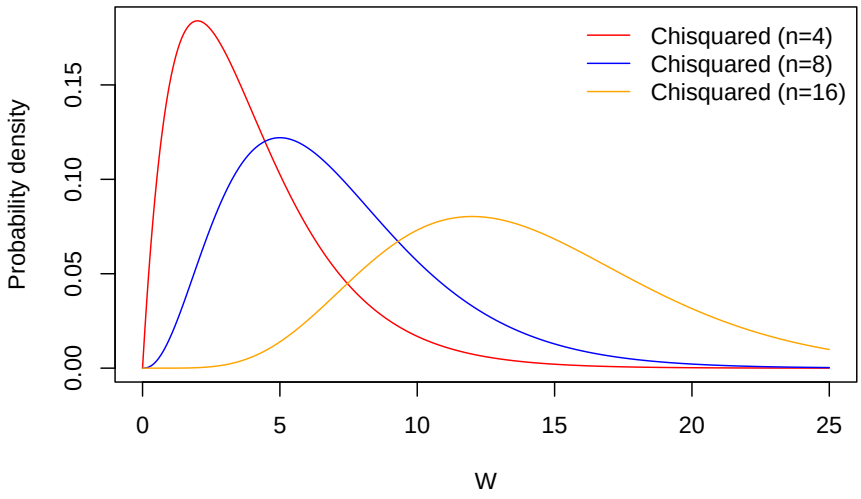
where:

1) $C_n = \frac{1}{2^{(n-1)/2}\sqrt{\pi(n-1)}}$ ensures \$ _{-∞}^{∞} 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.

10.10 χ^2-statistic

The probability density χ^2 has a parameter $df = n-1$, called 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

$$X \to N(\mu = 13, \sigma^2 = 0.35^2)$$

so

$$W = \frac{(n-1)S^2}{\sigma^2} = \frac{7S^2}{0.35^2} \to \chi^2(n-1)$$

we can calculate

$$P(S^2 > 0.3) = P(\frac{(n-1)S^2}{\sigma^2} > \frac{(n-1)0.3}{\sigma^2})$$

$= P(W > \frac{7*0.3}{0.35^2}) = P(W > 17.14286)$

$= 1 - P(W \leq 17.14286)$

$= 1 - F_{\chi^2, df=7}(17.14286) = 0.016$

in R 1-pchisq(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.

If we take a random sample and get a value of s^2 that is greater than 0.3, it will be a rare observation if all is well. We tend to believe that the observed values are common, not rare, so we may think that something is not right.

When we perform random sampling we observe:

```
## [1] 13.34642 13.32620 13.01459 13.10811 12.96999 13.55309 13.75557 12.62747
```

Therefore, our observed value was $s^2_{stock} = 0.1275608$

The sample is not very sparse because $s^2_{stock} < 0.3$ and we believe that all is well and production is under control.

10.11 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 implies:

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;

10.12 Exercises

10.12.0.1 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)

10.12.0.2 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 introduce the **margin of errors** when estimating the mean of the population distribution by 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.

The 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 m of $\bar{X}$ from μ that captures 95% of the estimations:

$$P(-m \leq \bar{X} - \mu \leq m) = P(\mu - m \leq \bar{X} \leq \mu + m) = 0.95$$

This means that 95% of the possible results of $\bar{X}$ are a distance m from μ

11.3 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, when we estimate μ by $\bar{x}$?

Calculating the maging of error of a normal variable

We want to know the number m in the equation

$$P(\mu - m \leq \bar{X} \leq \mu + m) = 0.95$$

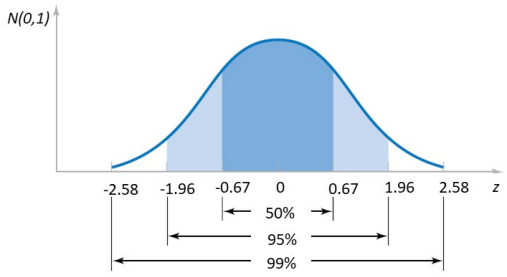
To solve this equation we need two steps. **First**, we need to know the **distribution** of $\bar{X}$.

1. When X is normal ($X \to N(\mu, \sigma^2)$) then

$$\bar{X} \to N(\mu, \frac{\sigma^2}{n})$$

We **then** need **standardize** $\bar{X}$. Remember that to standardize a normal variable we subtract its mean an divide it by its standard deviation.

$$Z = \frac{\bar{X} - E(\bar{X})}{\sqrt{V(\bar{X})}} = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \to N(0,1)$$



2. Substituting the mean of $\bar{X}$ and its standard deviation into the equation for the margin of error, we have:

$$P(\mu - m \leq \bar{X} \leq \mu + m) = P(-\frac{m}{\sigma/\sqrt{n}} \leq \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \leq \frac{m}{\sigma/\sqrt{n}})$$

$$= P(-\frac{m}{\sigma/\sqrt{n}} \leq Z \leq \frac{m}{\sigma/\sqrt{n}}) = 0.95$$

Compare it with the plot above: $\frac{m}{\sigma/\sqrt{n}}$ is the distance about 0 that captures 95% of the distribution standard normal variable. The distance leaves 2.5% probability at each tail of the distribution. For the upper tail, this is

$$\frac{m}{\sigma/\sqrt{n}} = \phi^{-1}(0.975) = 1.96$$

where ϕ^{-1} is the inverse of the standard normal distribution (qnorm(0.975)). Therefore

$$m = 1.96\frac{\sigma}{\sqrt{n}}$$

Example (cables)

The sample mean $\bar{X}$ of a sample of 8 cables follows a normal distribution with:

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 distribution

$$X \to f(x;\theta)$$

the standardized statistic

$$Z = \frac{\bar{X} - E(\bar{X})}{\sqrt{V(\bar{X})}}$$

approximates to a standard distribution

$$Z \rightarrow_d N(0,1)$$

when $n \rightarrow \infty$

Consequence: We can compute probabilities for $\bar{X}$ if n is large, using the normal distribution:

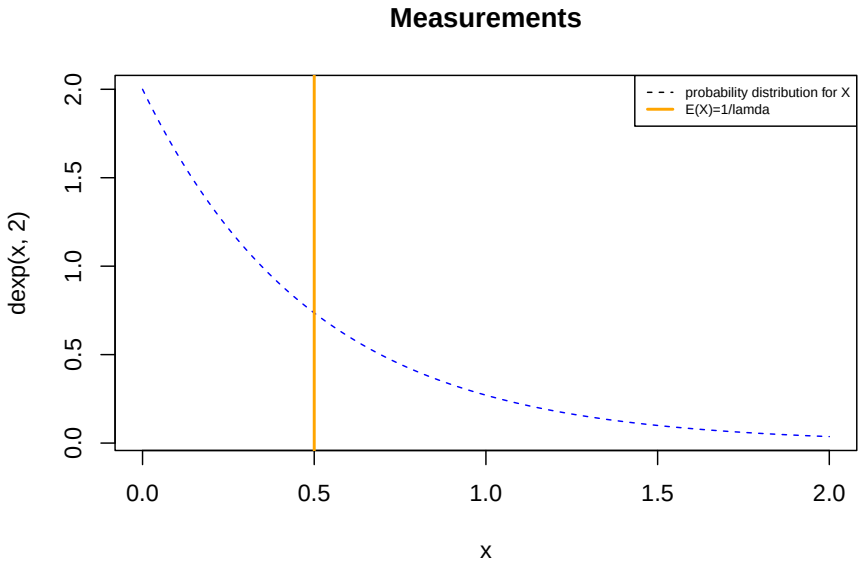
$$\bar{X} \sim_{aprox} N(E(X), \frac{V(X)}{n})$$

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

$$X \rightarrow exp(\lambda = 2)$$



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 $n \geq 30$

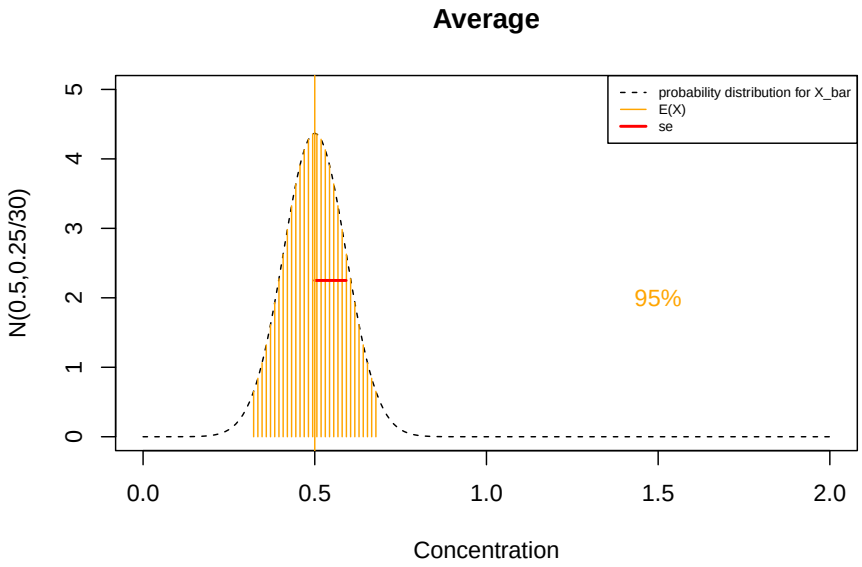
$$Z = \frac{\bar{X} - \lambda}{\sqrt{\frac{1}{n\lambda^2}}}$$

is a 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
$$V(Y) = nV(X) = n^2V(\bar{X})$$

The CLT tells us that for any random variable X with **unknown** (any type of) distribution

$$X \to f(x;\theta)$$

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

$$Z = \frac{n\bar{X} - nE(\bar{X})}{\sqrt{n^2V(\bar{X})}} = \frac{Y - E(Y)}{\sqrt{V(Y)}}$$

Consequence: We can compute probabilities for the sample sum $Y = n\bar{X}$ if n is large, using the normal distribution:

$$Y \sim_{aprox} N(nE(X), nV(X))$$

Example (Bernoulli trial)

For the Bernoulli trial $X \to Bernoulli(p)$, the average of a n sample of Bernoulli trials is $\bar{X} = \sum_i^n X_i$. Therefore

$$Z = \frac{\bar{X} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

is standard normal, because $E(\bar{X}) = p$ and $V(\bar{X}) = \frac{p(1-p)}{n}$.

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 $V(Y) = np(1-p)$. Since we can write

$$Z = \frac{Y - np}{\sqrt{np(1-p)}}$$

then using the CLT, we can proximate the Binomial probability mass function with the normal probability density when n is big

$$Y \to Binom(n, p) \sim_{aprox} N(np, np(1-p))$$

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:pnorm(0.9, 0.85, sqrt(0.85*0.15)/10); **b:**dnorm(0.85, 0.9, sqrt(0.9*0.1)/10); **c:**pnorm(0.85, 0.9, sqrt(0.9*0.1)/10); **d:**dnorm(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? (R:0.1863)
- How many components do we need such that the telescope is operational 2750 consecutive hours with at least 0.95 probability? (R: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? (R: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 $\sigma = 5$mg.

- for a random sample of size 50. What is the probability that the sample mean (average) is between 128 and 132gr? (R:0.995)
- what should be the size of the sample (n) such that the sample mean $\bar{X}$ is higher than 131gr with a probability less or equal than 0.025?(R: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? (R: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? (R:169)

Chapter 12

Maximum likelihood and Method of Moments

12.1 Objective

In this chapter we will discuss what an **estimator** is and give some examples. Then we will introduce two methods for obtaining **estimators** of the parameters of probability models.

These are the **maximum likelihood** and the **method of moments**.

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**

- mean: $\bar{X}$ for μ
- variance: S^2 for σ^2

3. **Inference** to say something about the parameters given the data

- mean: Z, T
- variance: χ^2

Remember: They are all random variables. Every time we take another sample they change their value.

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 population distribution as

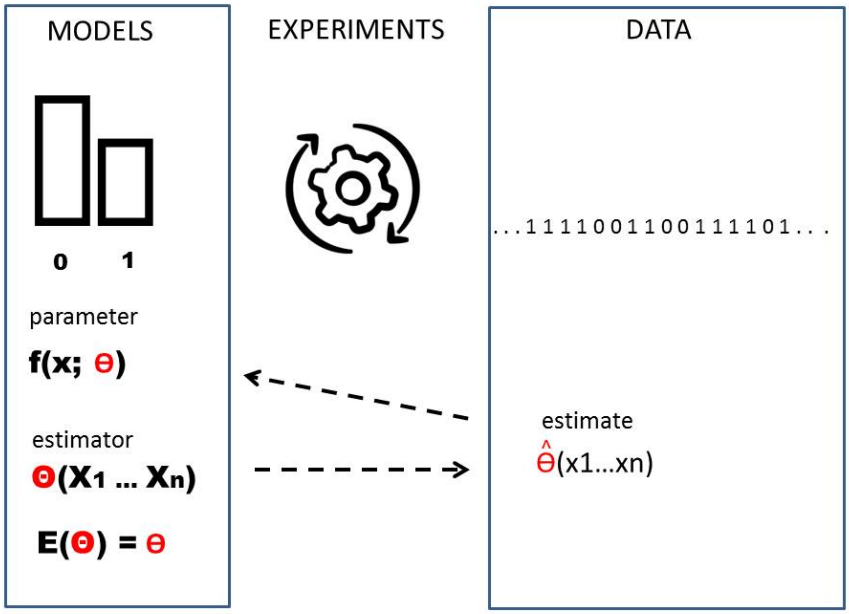
$$X \to f(x;\theta)$$

then θ is a parameter and Θ is a random variable whose observations $\hat{\theta}$ we take as estimations of θ

$$\hat{\theta} \sim \theta$$

Therefore 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. $\hat{\theta}$ is the **estimate** of θ: A realized value of Θ



Example (Sample mean)

When we have a normal random variable

$$X \to N(\mu, \sigma^2)$$

we identify the three different quantities:

1. μ is a **parameter** of the **population** distribution: distribution of X, $N(\mu, \sigma^2)$
2. $\bar{X}$ is an **estimator** of μ
3. $\bar{x} = \hat{\mu}$ 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 is **unbiased** if it 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 icreased

$$lim_{n\to\infty} V(\Theta) = 0$$

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

$$mse(\Theta) = se^2 + bias^2$$

where $se = \sqrt{V(\bar{X})}$ is the standard error.

12.4 Maximum likelihood

How can obtain **estimators** of the parameters of **any** probability model?

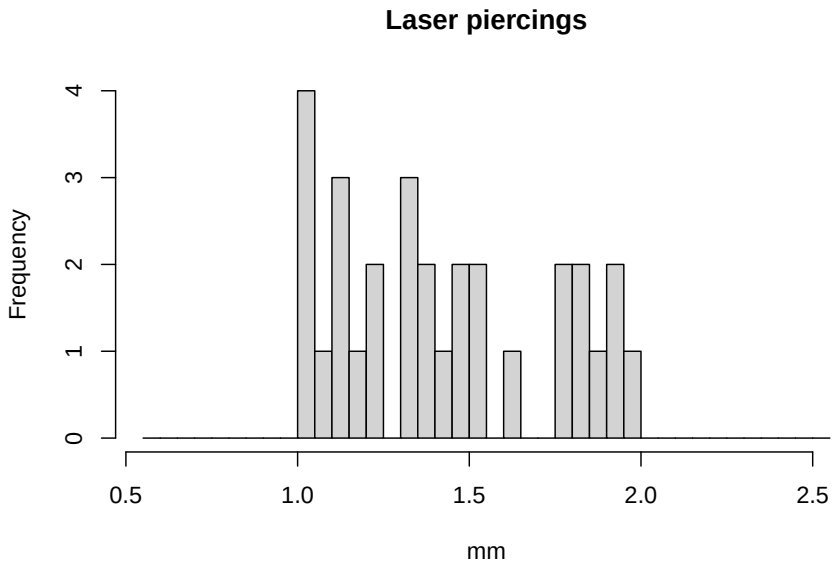
Example (Laser)

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.

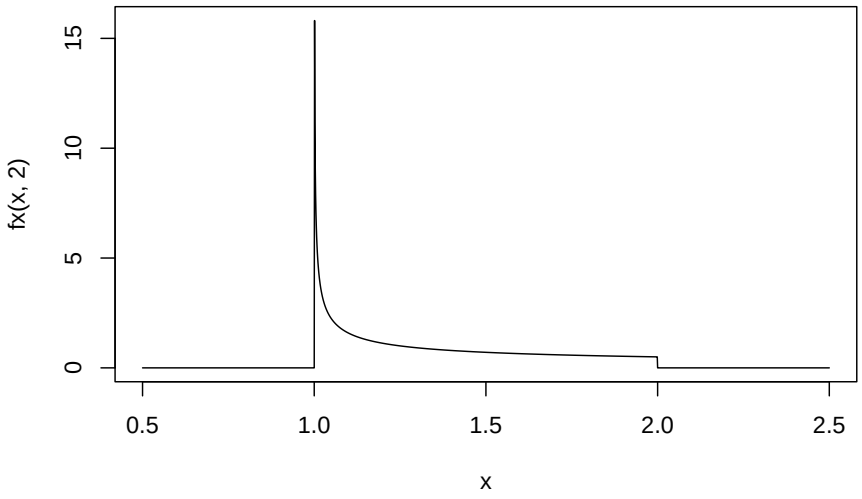
Example (Laser)

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

$$f(x) = \begin{cases} \frac{1}{\alpha}(x-1)^{\frac{1}{\alpha}-1}, & \text{if } x \in (1,2) \\ 0, & x \notin (1,2) \end{cases}$$

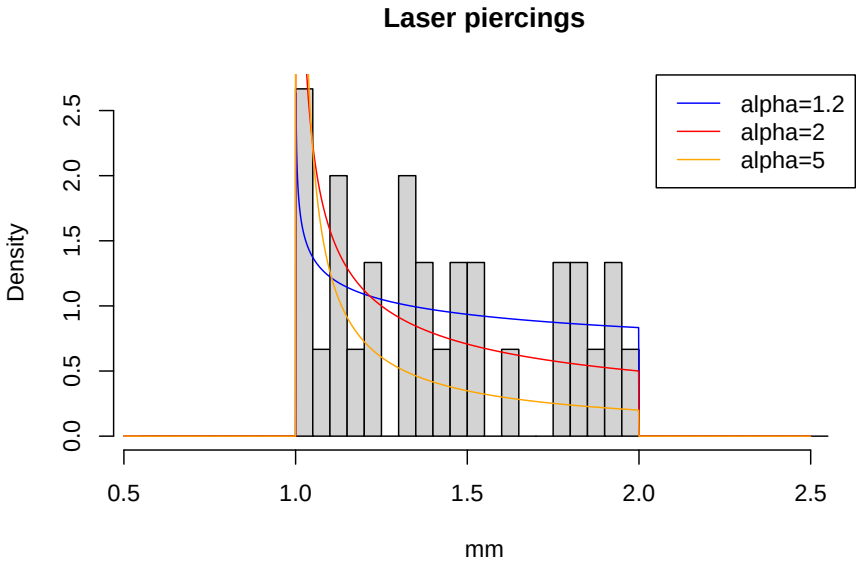
Where α is a parameter. This is a probability density because it integrates to one and it is positive. In particular, for $\alpha = 2$ we can plot it



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 that gives us the estimator for α

$$\hat{\alpha}_{ml}$$

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**.

Maximum likelihood step 1

We calculate the probability of having observed the n-sample: $x_1, ...x_n$. It is the product of probabilities because observations 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 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.

$$\frac{d \ln L(\alpha)}{d\alpha}|_{\hat{\alpha}} = 0$$

The value of the parameter at the maximum is called the **maximum likelihood estimate** for the parameter and its written with a hat $\hat{\alpha}$.

Example (Laser)

We derive the log-likelihood

$$\frac{d \ln L(\alpha)}{d\alpha} = -\frac{n}{\alpha} - \frac{1}{\alpha^2} \Sigma_{i=1...n} \ln(x_i)$$

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

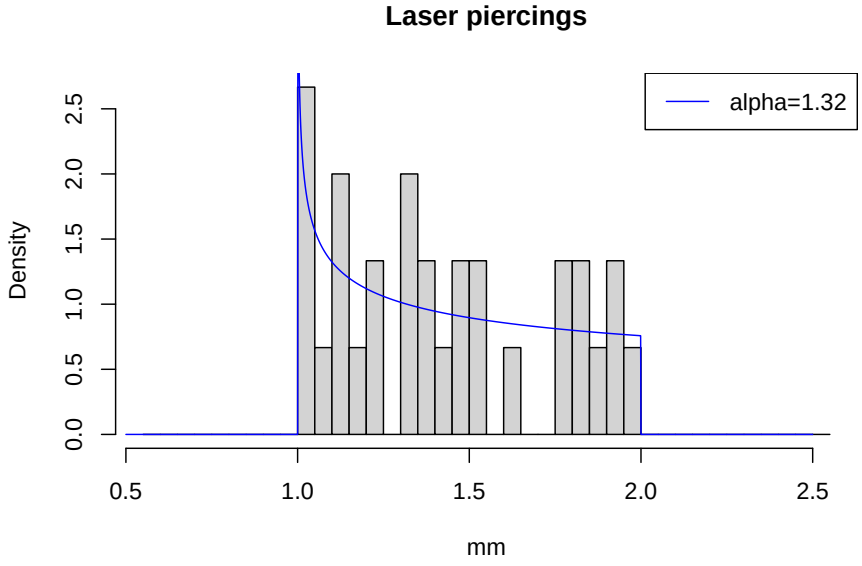
$$(X_1, X_2, ...X_n)$$

Estimating the parameters with the data

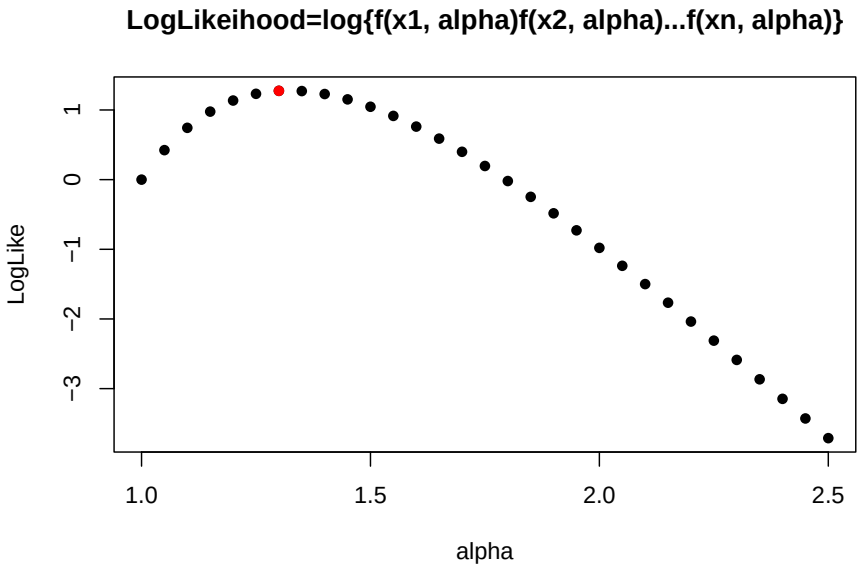
In our example, we then have the observation of the random sample as a set of 30 numbers $(x_1, x_2, ...x_{30})$, we therefore substitute the numbers in the estimator and this will give us its observed value.

$\hat{\alpha}_{ml} = -\frac{1}{n}\{\ln(1.11-1) + \ln(1.64-1) + ... \ln(1.04-1)\} = 1.320$

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.



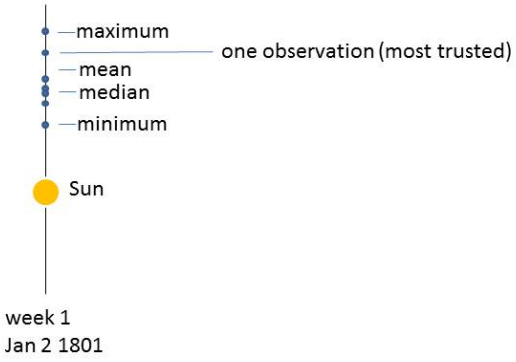
Maximum likelihood: History

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?

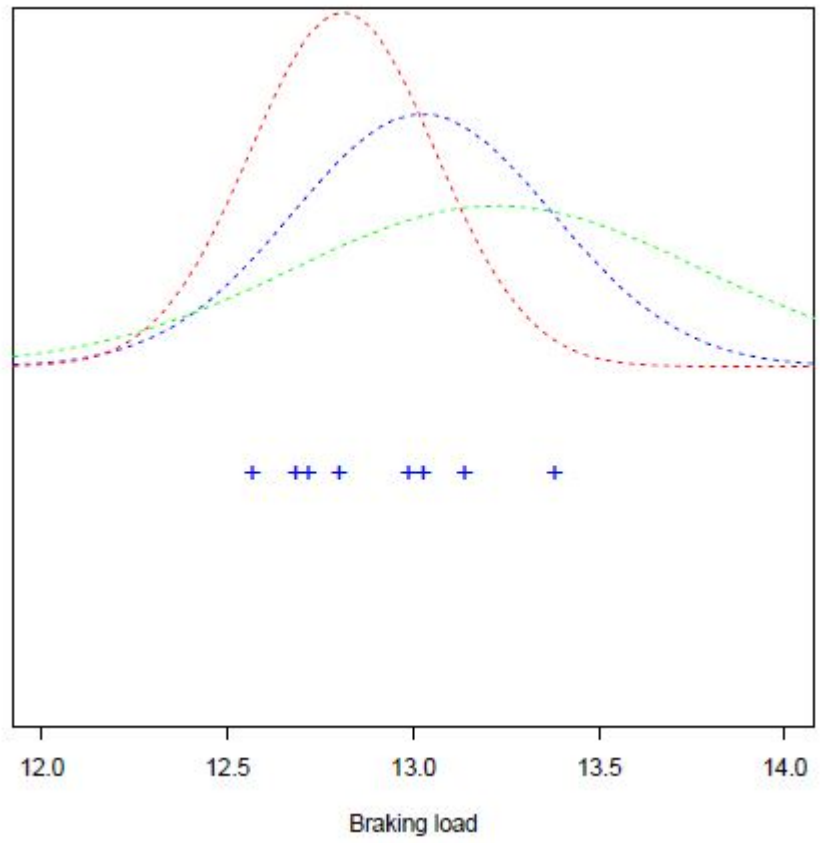
Maximum likelihood of the normal distribution

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 discover 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 Method of Moments

The method of maximum likelihood aims to produce the estimators of probability distributions from data. However, there is another way to produce those estimators, which is based on the main frequentist idea of probabilities.

Let's re-write the estimator $\hat{\mu} = \bar{x}$ for a normal randon variable in terms of the outcomes of X

For instance:

$$\hat{\mu} = \frac{1}{n}\sum_i x_i = \sum_x x\frac{n_x}{n}$$

and remember that in the limit $n \to \infty$ the frequentist interpretation requires $\frac{n_x}{n} \to P(X = x)$ and therefore in the limit

$$\hat{\mu} = \frac{1}{n}\sum_i x_i \to E(X) = \mu$$

The method of moments says that we can take the **observed** value of the average $\bar{X}$ as an estimator of $E(X) = \mu$

$$E(X) \sim \bar{x}$$

The average $\bar{X} = \frac{1}{n}\sum_i X_i$ is also called the first **sample moment**

If $X \to f(x, \theta)$ the estimator of the parameter θ is then obtained from the equation:

$$E(X; \hat{\theta}) = \bar{x}$$

Example (exponential)

If a random vartible follows an exponential distribution

$$X \hookrightarrow exp(\lambda)$$

then we can use the method of moments to estimate λ. The method consists on three steps:

1. Compute the expected value of variable
$$E(X; \lambda) = \mu$$
2. Write down the equation where the expected value is equal to the first sample moment
$$\frac{1}{\hat{\lambda}} = \bar{x}$$
3. Solve for the parameter

$$\hat{\lambda} = \frac{1}{\bar{x}}$$

In terms of the data this is $\hat{\lambda} = (\frac{1}{n}\sum_i x_i)^{-1}$.

Example (Bateries)

Suppose that we have several batteries (new and old) that we charge over the period of 1 hour. We measure the state of charge of the battery, being 1 a 100% charge.

The state of charge of a battery is a random variable that may have a uniform distribution, where we do not know the minimum value that x can take, but we know that the maximum is 1 (100% of charge)

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

What is the estimator of a (the minimum charge after one hour)?

If we run an experiment and obtain $x_1, ...x_n$, we ask how can we estimate a form the data?

We follow the three stepd of the method of moments:

1. We compute the expected value of the random variable

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

2. We obtain the equation for $\hat{a}$ where we make the expected value equal to the first sample moment

$$\frac{\hat{a}+1}{2} = \bar{x}$$

3. We solve for the estimator $\hat{a}$

$$\hat{a} = 2\bar{x} - 1$$

This is the estimator of the minimum charge we may observe.

Note that taking the minimum of the observations is clearly suboptimal. The method gave us a clever answer that can also be summarized by the following steps

a) We can compute $\bar{x}$ with increasing precision given by n
b) We know that no measurement surpasses $b = 1$
c) Then we compute the distance between $\bar{x}$ and b: $1 - \bar{x}$

d) Finally, we subtract it from $\bar{x}$:

$$\bar{x} - (1 - \bar{x}) = 2\bar{x} - 1$$

This should be our best guess for $\hat{a}$. As such we arrive at the same estimate given by the method of moments.

12.7 Method of Moments for several parameters

The method says that an estimator for the parameter θ of $f(x;\theta)$ can be found from the equation:

$$E(X) = \frac{1}{n}\sum_i x_i$$

If there are more parameters, we use the higher **sample moments**. Consider that the second sample moment is

$$\frac{1}{n}\sum_i X_i^2$$

Therefore, an observation of this moment is close to $E(X^2)$

$$E(X^2) \sim \frac{1}{n}\sum_i x_i^2$$

The method for two parameters says that an estimation for the the parameters θ_1 and θ_2 of $f(x;\theta_1,\theta_2)$ can be found from the equations:

a. $E(X) = \frac{1}{n}\sum_i x_i$

b. $E(X^2) = \frac{1}{n}\sum_i x_i^2$

We can have as many equations as parameters we need to compute, incrementing the degree of the moments.

Example (Normal distribution)

If X distributes normally, we have two parameters to estimate

$$X \to N(\mu, \sigma^2)$$

We follow the steps for the method of moments for two parameters:

1. We compute the mean and expected value of the second moment $E(X^2)$:

$$E(X) = \mu$$

and

$$E(X^2) = \sigma^2 + \mu^2$$

$E(X^2)$ follows from the property: $E(X^2) = V(X) + \mu^2$

2. We obtain the equations for the parameters where we make (a) the expected value of the variable equal to the first sample moment, and (b) the expected value of the second moment equal to the second sample moment

a. $E(X)$ is estimated by

$$\hat{\mu} = \frac{1}{n}\sum_i x_i$$

b. $E(X^2)$ is estimated by

$$\hat{\sigma}^2 - \hat{\mu}^2 = \frac{1}{n}\sum_i x_i^2$$

3. We solve for the parameters

The first equation gives the estimator for the mean μ.

$$\hat{\mu} = \frac{1}{n}\sum_i x_i$$

Which again if the average. From the second equation we obtain

$$\hat{\sigma}^2 = \frac{1}{n}\sum_i x_i^2 - \hat{\mu}^2$$

which can also be written as:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_i (x_i - \hat{\mu})^2$$

We find that the method of moments and the maximul likelihood estimates for the normal distribution are the same. However, this it not always the case.

Example (laser)

What is the estimator of parameter α for the laser cut given by the method of moments?

$$f(x;\alpha) = \begin{cases} \frac{1}{\alpha}(x-1)^{\frac{1}{\alpha}-1}, & \text{if } x \in (1,2) \\ 0, & x \notin (1,2) \end{cases}$$

Where α is a parameter.

The method says that an estimator for the parameter α of $f(x;\alpha)$ can be found from the equation:

$$E(X) = \frac{1}{n}\sum_i x_i$$

for $\hat{\alpha}$

1. We compute the expected value $E(X)$

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

Consider a change of variables $Z = X - 1$ then $E(X) = E(Z) + 1$ and

$E(Z) = \frac{1}{\alpha}\int_0^1 z z^{\frac{1-\alpha}{\alpha}} dz = \frac{1}{\alpha}\int_0^1 z^{1+\frac{1-\alpha}{\alpha}} dz$

$= \frac{1}{\alpha}\frac{z^{2+\frac{1-\alpha}{\alpha}}}{2+\frac{1-\alpha}{\alpha}}|_0^1 = \frac{1}{1+\alpha}$

Therefore,

$$E(X) = E(Z+1) = \frac{1}{1+\alpha} + 1$$

2. We obtain the equation for $\hat{\alpha}$ where we make the expected value equal to the first sample moment. Substituting for $\hat{\alpha}$, the method of moments gives us the equation

$$\frac{1}{1+\hat{\alpha}} + 1 = \bar{x}$$

3. We solve for $\hat{\alpha}$

$$\hat{\alpha}_m = \frac{1}{\bar{x} - 1} - 1$$

4. We compute the value for our data

$$\hat{\alpha}_m = 1.314$$

Note that this is an example for which the estimates by maximum likelihood and the method of moments are **different**.

The maximum likelihood estimate was:

$$\hat{\alpha}_{ml} = -\frac{1}{n}\sum_{i=1}^{n} \ln(x_i - 1) = 1.320$$

The method of moments estimate was:

$$\hat{\alpha}_m = \frac{1}{\bar{x} - 1} - 1 = 1.314$$

We need **simulation** studies, where **we know** the true value of the parameter α, to find which of these statistics have less error.

Note: the data for 30 laser piercings were simulated with $\alpha = 2$, therefore we should prefer the maximum likelihood estimate.

To obtain better estimates of α we need to increase the size of the sample.

12.8 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;

5) The first sample moment is

a: the mean; **b:** the variance; **c:** the expected value; **d:** the average;

12.9 Exercises

12.9.0.1 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?

12.9.0.2 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?

12.9.0.3 Exercise 3

Take a random variable with the following probability density function

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

- What is the maximum likelihood estimate for λ?
- If we take a 5-sample with observations $x_1 = 0.223$ $x_2 = 0.681$; $x_3 = 0.117$; $x_4 = 0.150$; $x_5 = 0.520$

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 $Y = \sum_1^n X_i$?

12.10 Method of moments

12.10.0.1 Exercise 1

What are the estimators of the following parametric models given by the method of moments?

Model	f(x)	E(X)
Bernoulli	$p^x(1-p)^{1-x}$	p

Model	f(x)	E(X)
Binomial	$\binom{n}{x}p^x(1-p)^{n-x}$	np
Shifted geometric	$p(1-p)^{x-1}$	$\frac{1}{p}$
Negative Binomial	$\binom{x+r-1}{x}p^r(1-p)^x$	$r\frac{1-p}{p}$
Poisson	$\frac{e^{-\lambda}\lambda^x}{x!}$	λ
Exponential	$\lambda e^{-\lambda x}$	$\frac{1}{\lambda}$
Normal	$\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	μ

12.10.0.2 Exercise 2

Take a random variable with the following probability density function

$$f(x) = \begin{cases} (1+\theta)x^\theta, & \text{if } x \in (0,1) \\ 0, & x \notin (0,1) \end{cases}$$

- Compute $E(X)$ as a function of θ
- What is the estimate for θ using the method of moments?
- 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?

12.10.0.3 Exercise 3

Consider a discrete random variable X that follows a negative binomial distribution with probability mass function:

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

Given that

- $E(X) = \dfrac{r(1-p)}{p}$
- $V(X) = \dfrac{r(1-p)}{p^2}$

compute:

- An estimate for the parameter r and an estimate for the parameter p obtained from a random sample of size n using the method of moments.
- The values of the estimates of r y p for the folowing random sample:

$x_1 = 27$; $x_2 = 8$; $x_3 = 22$; $x_4 = 29$; $x_5 = 19$; $x_5 = 32$