Stats theory (SDA)

Alejandro Caceres

2022-09-30

Contents

Chapter 1

About

The course is divided into **theory** and **practical** classes (Bootcamps). The classes on theory are subdivided into statistics (Stats), machine learning, and Bayesian inference. Here, are the times, schedules, and content for the statistics theory classes.

Stats theory classes comprise a total of 30 hours: 24 plenary lectures (24 hours) divided in

- Descriptive statistics and probability (4 days)
- Inference (4 days)

and 2 group work sessions (6 hours)

1.1 Schedule:

1.2 Recommended reading list

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

Chapter 2

Data description

2.1 Objective

- Data: discrete, continuous
- Summarizing data in tables and figures

2.2 Statistics

- Solve problems in a systematic way (science, engineering and technology)
- Modern humans use a general **method** historically developed for thousands of years! … and still under development.
- It has three main components: observation, logic, and generation of new knowledge

2.3 Scientific method

2.4 Outcome

Observation or *Realization*

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

… 1 0 0 1 0 **1** 0 1 1 … (the number in bold is an observation in a repetition of the experiment)

Outcome

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

1 is an outcome, **0** is the other outcome

2.5 Types of outcome

- **Categorical**: If the result of an experiment can only take discrete values (number of car pieces produced per hour, number of leukocytes in blood)
- **Continuous**: If the result of an experiment can only take continuous values (battery state of charge, engine temperature).

2.6 Random experiments

Definition:

A **random experiment** is an experiment that gives different outcomes when repeated in the same manner.

Examples:

- on the same object (person): temperature, sugar levels.
- on different objects but the same measurement: the weight of an animal.
- on events: a number of emails received in an hour.

2.7 Absolute frequencies

When we repeat a random experiment, we record a list of outcomes.

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

Absolute frequency:

$$n_i$$

is the number of times we observed the outcome i

2.8 Example

Random experiment: Extract a leukocyte from **one** donor and write down its type. Repeat experiment $N = 119$ times.

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

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

- For instance: $n_1 = 34$ is total number of T cells
- $N = \sum_i n_i = 119$

2.9 Relative frequencies

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

$$f_i = n_i/N$$

where N is the total number of observations

In our example there are recorded $n_1 = 34$ T cells, so we ask for the proportion of T cells from the total of 119.

2.10 Example

```
##      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
```

We have

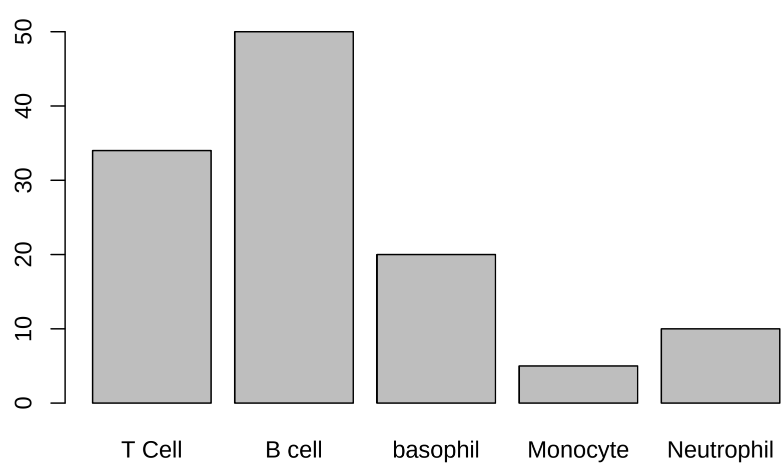
$\sum_{i=1..M} n_i = N$

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

where M is the number of outcomes.

2.11 Bar plot

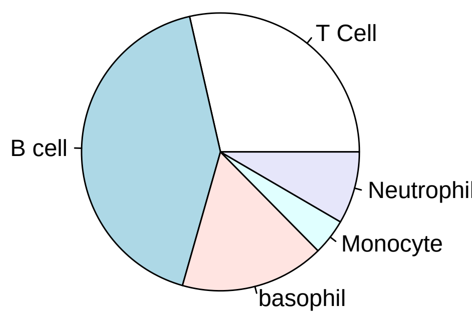
We can plot n_i Vs the outcomes, giving us a bar plot



2.12 Pie chart

We can visualize the relative frequencies with a pie chart

- Where the area of the circle represents 100% of observations (proportion = 1) and the sections the relative frequencies of all the outcomes.



2.13 Categorical and ordered variables

Cell types are not meaningfully ordered concerning the outcomes. However, sometimes **categorical** variables can be **ordered**.

Misophonia study:

- 123 patients were examined for misophonia: anxiety/anger produced by certain sounds
- They were categorized into 4 different groups according to severity.

2.14 Example

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
```

And its 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
```

2.15 Absolute and relative cumulative frequencies

Misophonia severity is **categorical** and **ordered**.

When outcomes can be ordered then it is useful to ask how many observations were obtained up to a given outcome we call this number the absolute cumulative frequency up to the outcome i:

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

It is also useful to compute the **proportion** of the observations that was obtained up to a given outcome

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

2.16 Frequency table

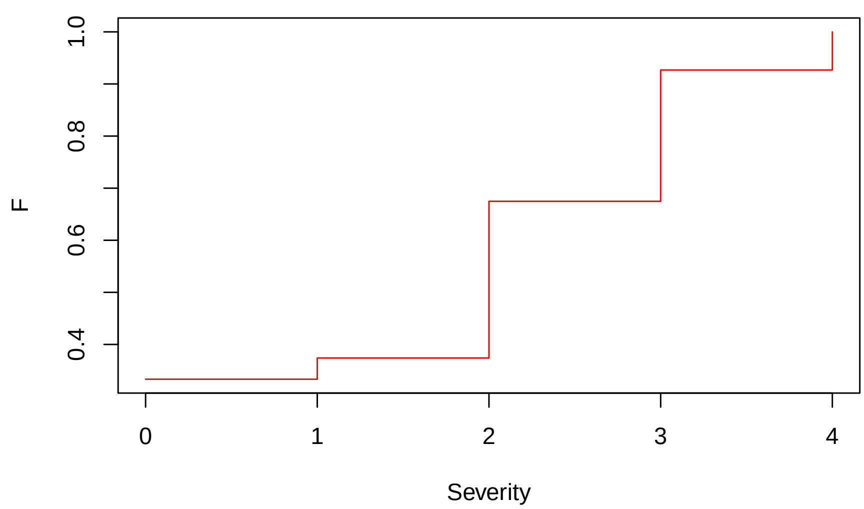
```
##   outcome ni         fi  Ni        Fi
## 0       0 41 0.33333333  41 0.3333333
## 1       1  5 0.04065041  46 0.3739837
```

```
## 2        2 37 0.30081301  83 0.6747967
## 3        3 31 0.25203252 114 0.9268293
## 4        4  9 0.07317073 123 1.0000000
```

- **67%** of patients had misophonia up to severity **2**
- **37%** of patients have severity less or equal than **1**

2.17 Cumulative frequency plot

We can also plot the cumulative frequency Vs the outcomes



2.18 Continuous variables

The result of a random experiment can also give continuous outcomes.

In the misophonia study, the researchers asked whether the convexity of the jaw would affect the misophonia severity (the scientific hypothesis is that the convexity angle of the jaw can influence the ear and its sensitivity). These are the results for the convexity of the jaw (degrees)

```
##   [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
```

2.19 Bins

Continuous outcomes cannot be counted!

We transform them into ordered categorical variables

- We cover the range of the observations into 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]"
```

2.20 Create a categorical variable from a continuous one

- We map each observation to its interval: creating an **ordered categorical** variable; 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]"
```

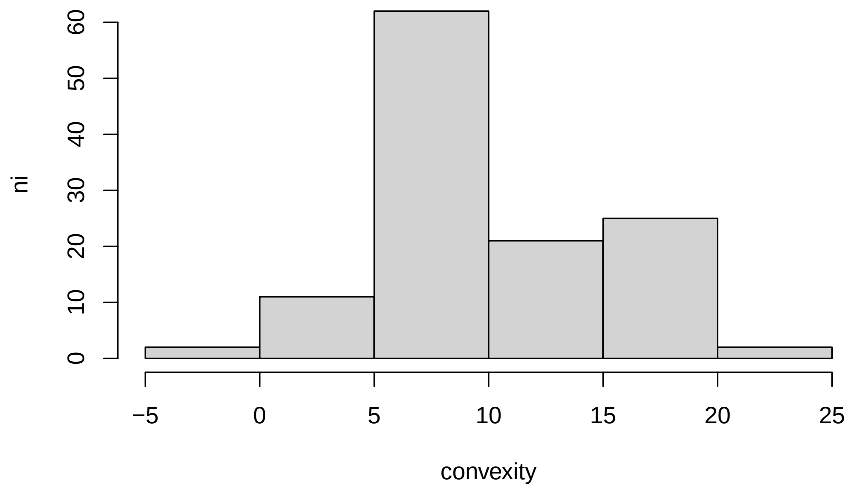
```
##  [41] "(3.46,7.92]"  "(3.46,7.92]"  "(3.46,7.92]"  "(7.92,12.4]"  "(12.4,16.8]"
##  [46] "(3.46,7.92]"  "(3.46,7.92]"  "(3.46,7.92]"  "(3.46,7.92]"  "(12.4,16.8]"
##  [51] "(7.92,12.4]"  "[-1.02,3.46]" "(3.46,7.92]"  "(3.46,7.92]"  "(3.46,7.92]"
##  [56] "(7.92,12.4]"  "(7.92,12.4]"  "(7.92,12.4]"  "(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]"
```

2.21 Frequency table for a continuous variable

```
##        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
```

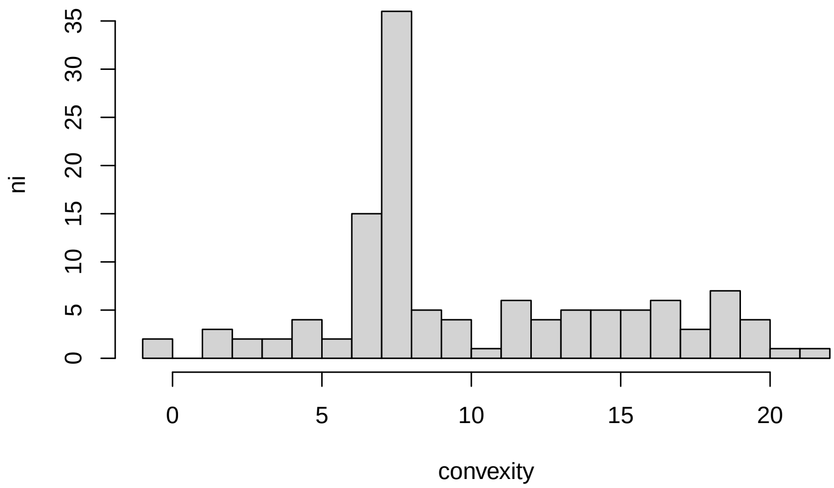
2.22 Histogram

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



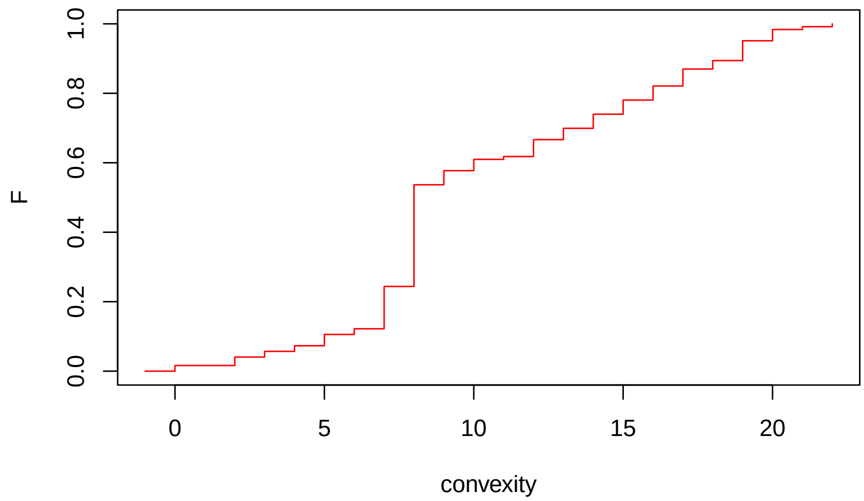
2.23 Histogram

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



2.24 Cumulative frequency plot: Continous variables

We can also plot the cumulative frequency Vs the outcomes



2.25 Summary statistics

The summary statistics are numbers computed from the data that tell us important features of numerical variables (categorical or continuous).

Limiting values:

- minimum: the minimum outcome observed
- maximum: the maximum outcome observed

Central value for the outcomes

- The average is defined as

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

where x_j is the **observation** j (convexity) from a total of N.

2.26 Average

The average convexity can be computed directly from the **observations**

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

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

2.27 Average (categorical ordered)

For **categorical ordered** variables we can use the frequency table to compute the average

```
##   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
```

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

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

$= 0 * f_0 + 1 * f_1 + 2 * f_2 + 3 * f_3 + 4 * f_4 = 1.691057$

(note the change from N to M in the second summation)

2.28 Average (categorical ordered)

In terms of the **outcomes** of categorical ordered variables, the **average** can be written as

$$\bar{x} = \sum_{i=1...M} x_i f_i$$

from a total of M possible outcomes (number of severity levels).

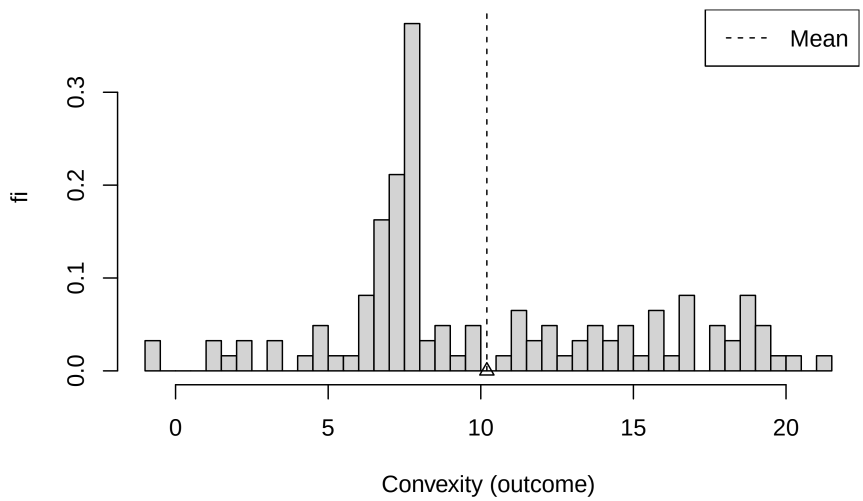
$\bar{x}$ is the **central value** or center of gravity of the outcomes. As if each outcome had a mass density given by f_i.

2.29 Average

- The average is not the result of one observation (random experiment).
- It is the result of a series of observations (sample).
- It describes the number where the observed values balance.

That is why we hear, for instance, that a patient with an infection can infect an average of 2.5 people.

2.30 Average



2.31 Median

Another measure of centrality is the median. The median $q_{0.5}$ is the value x_p

$$median(x) = q_{0.5} = x_p$$

below which we find half of the observations

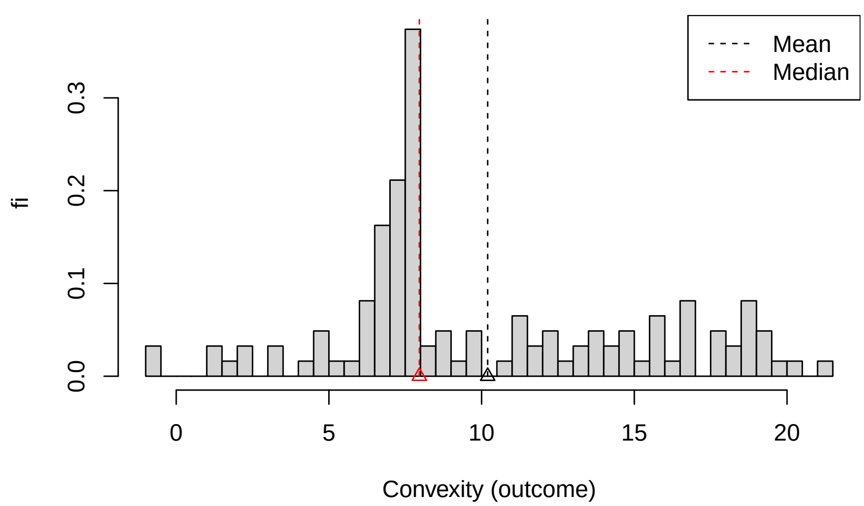
$$\sum_{x \le x_p} 1 = \frac{N}{2}$$

or in terms of the frequencies, is the value x_p that makes the cumulative frequency F_p equal to 0.5

$$q_{0.5} = \sum_{x \le x_p} f_x = F_p = 0.5$$

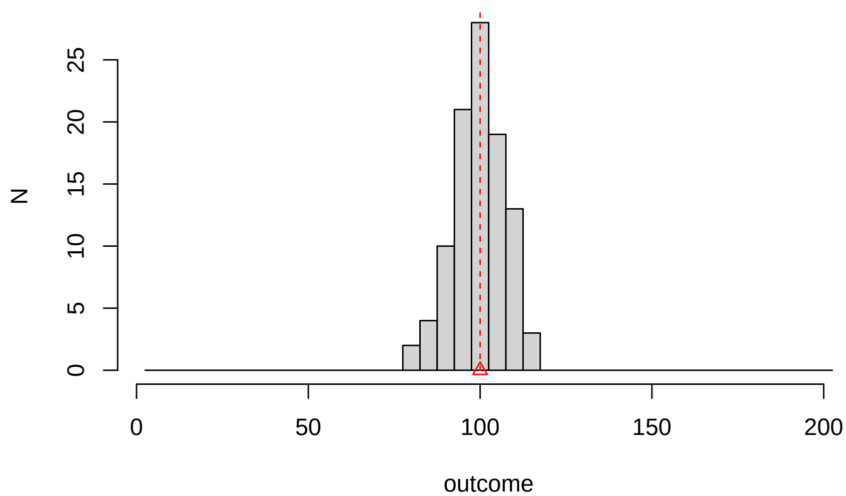
2.32 Median Vs Average

- Average: Center of mass (compensates distant values)
- Median: Half of the mass

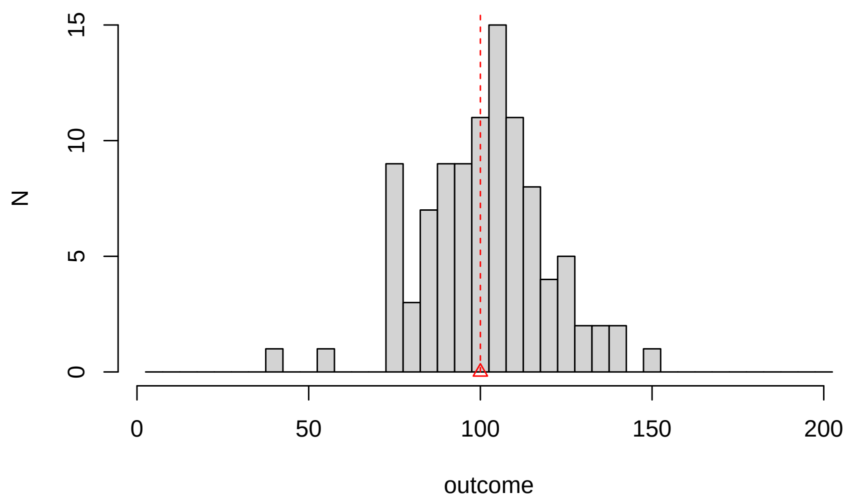


2.33 Dispersion

An important measure of the outcomes is their **dispersion**. Many experiments can share their mean but differ on how dispersed the values are.



2.34 Dispersion



2.35 Sample variance

Dispersion about the mean is measured with the

- The sample variance:

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

It measures the average square distance of the **observations** to the average. The reason for $N-1$ will be explained when we talk about inference.

2.36 Sample variance

- In terms of the frequencies of **categorical and ordered** variables

$$s^2 = \frac{N}{N-1} \sum_x (x - \bar{x})^2 f_x$$

s^2 can be thought of as the moment of inertia of the observations.

2.37 Standard deviation

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

The standard deviation of the convexity angle 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.

2.38 IQR

- Dispersion of data can also be measured with respect to the median by the **interquartile range**
- We define the **first** quartile as the value x_p that makes the cumulative frequency F_p equal to 0.25

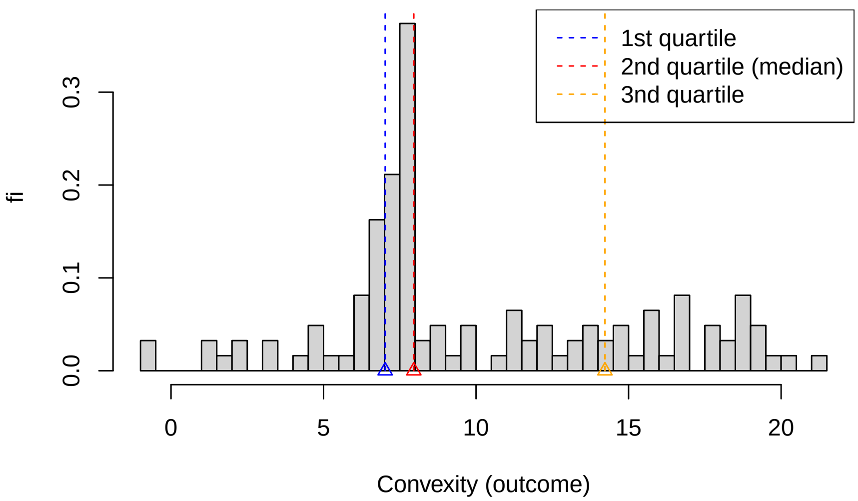
$$q_{0.25} = \sum_{x \leq x_p} f_x = F_p = 0.25$$

- We also define the **third** quartile as the value x_p that makes the cumulative frequency F_p equal to 0.75

$$q_{0.75} = \sum_{x \leq x_p} f_x = F_p = 0.75$$

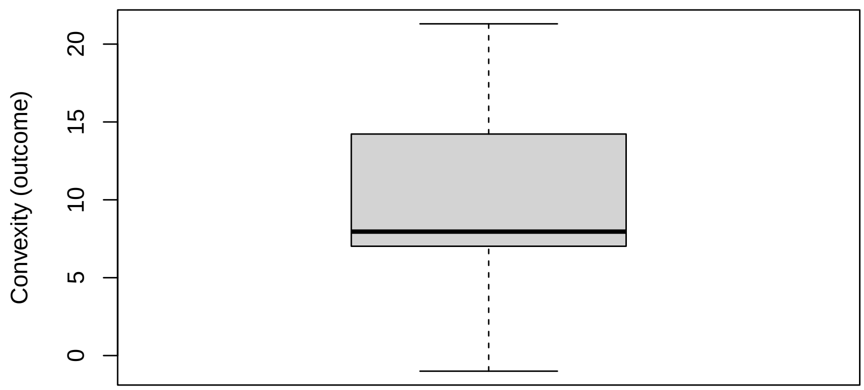
2.39 IQR

The distance between the third quartile and the first quartile is called the **interquartile range** (IQR) and captures the central 50% of the observations



2.40 Box plot

The interquartile range, the median, and the 5% and 95% of the data can be visualized in a **boxplot**, here the values of the outcomes are on the y-axis. The IQR is the box, the median is the line in the middle and the whiskers mark the 5% and 95% of the data.


Convexity (outcome)
0
5
10
15
20

Chapter 3

Probability

3.1 Objective

- Definition of probability
- Probability algebra
- Joint probability

3.2 Random experiments

Observation

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

Outcome

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

Random experiment

- An experiment that gives **different** outcomes when repeated in the same manner.

3.3 Probability

The **probability** of an outcome is a measure of how sure we are to observe that outcome when performing a random experiment.

- 0: We are sure that the observation will **not** happen.
- 1: We are sure that the observation will happen.

3.4 Example

- Consider the following observations of a random experiment:

1 5 1 2 2 1 2 2

- How sure we are to obtain 2 in the following observation?

3.5 Example

The frequency table is

```
##   outcome ni    fi
## 1       1  3 0.375
## 2       2  4 0.500
## 3       5  1 0.125
```

The **relative frequency** f_i

- is a number between 0 and 1.
- measures the proportion of total observations that we observed a particular outcome.
- seems a reasonable probability measure.

As $f_2 = 0.5$ then we would be half certain to obtain a 2 in the next repetition of the experiment.

3.6 Relative frequency

As a measure of certainty is f_i enough?

Say we repeated the experiment 12 times more:

1 5 1 2 2 1 2 2 **3 1 1 3 3 1 6 3 5 6 4 4**

The frequency table is now

```
##   outcome ni  fi
## 1       1  6 0.3
## 2       2  4 0.2
## 3       3  4 0.2
## 4       4  2 0.1
## 5       5  2 0.1
## 6       6  2 0.1
```

New outcomes appeared and f_2 is now 0.2, we are now a fifth certain of obtaining 2 in the next experiment… probability should not depend on N

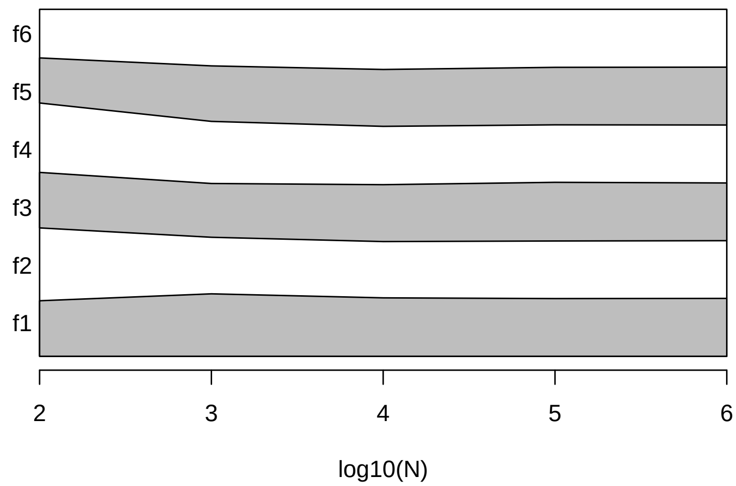
3.7 At infinity

Say we repeated the experiment 1000 times:

```
##   outcome  ni    fi
## 1       1 163 0.163
## 2       2 163 0.163
## 3       3 169 0.169
## 4       4 169 0.169
## 5       5 173 0.173
## 6       6 163 0.163
```

We find that f_i is converging to a constant value

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



3.8 Frequentist probability

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

Championed by Venn (1876)

The frequentist interpretation of probabilities is derived from data/experience (empirical).

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

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

3.9 Classical Probability

Whenever a random experiment has M possible outcomes that are all **equally likely**, the probability of each outcome is $\frac{1}{M}$.

Championed by Laplace (1814).

Since each outcome is **equally probable** we declare complete ignorance and the best we can do is to fairly distribute the same probability to each outcome.

What if I told you that our experiment was the throw of the dice? then

$P_2 = 1/6 = 0.166666$.

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

3.10 Classical and frequentist probabilities

3.11 Probability

Probability is a number between 0 and 1 that is assigned to each member E of a collection of **events** of a **sample space** (S) from a random experiment.

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

where $E \in S$

3.12 Sample space

We start by reasoning what are all the possible values (outcomes) that a random experiment could give.

Note that we do not have to observe them in a particular experiment: We are using **reason/logic** and not observation.

Definition:

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

3.13 Examples of sample spaces

- temperature 35 and 42 degrees Celcius
- sugar levels: 70-80mg/dL
- the size of one screw from a production line: 70mm-72mm
- number of emails received in an hour: 0-100
- a dice throw: 1, 2, 3, 4, 5, 6

3.14 Discrete and continuous sample spaces

- A sample space is discrete if it consists of a finite or countable infinite set of outcomes.

- A sample space is continuous if it contains an interval (either finite or infinite in length) of real numbers.

3.15 Event

Definition:

An **event** is a **subset** of the sample space of a random experiment. It is a **collection** of outcomes.

Examples of events:

- The event of a healthy temperature: temperature 37-38 degrees Celsius
- The event of producing a screw with a size: of 71.5mm
- The event of receiving more than 4 emails in an hour.
- The event of obtaining a number less than 3 in the throw of a dice

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

3.16 Event operations

For two events A and B, we can construct the following derived events:

- 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

3.17 Event operations example

Take

- Event $A : \{1, 2, 3\}$ a number less or equal to three in the throw of a dice
- Event $B : \{2, 4, 6\}$ an even number in the throw of a dice

New events:

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

3.18 Outcomes

Outcomes are events that are **mutually exclusive**

Definition:

Two events denoted as E_1 and E_2, such that

$$E_1 \cap E_2 = \emptyset$$

They cannot occur at the same time.

Example:

- The outcome of obtaining 1 **and** the outcome of obtaining 5 in the throw of one dice are mutually exclusive:
- The event of obtaining 1 and 5 is empty:

$$\{1\} \cap \{5\} = \emptyset$$

3.19 Probability definition

A probability is a number that is assigned to each possible event (E) of a sample space (S) of a random experiment that satisfies the following properties:

- $P(S) = 1$
- $0 \leq P(E) \leq 1$
- when $E_1 \cap E_2 = \emptyset$

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

Proposed by Kolmogorov's (1933)

3.20 Probability properties

Kolmogorov says that we can build a probability table (likewise the relative frequency table)

outcome	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

As $\{1, 2, 3, 4, 5, 6\}$ are mutually exclusive then

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

3.21 Addition Rule

When A and B are not mutually exclusive then:

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

Where $P(A)$ and $P(B)$ are called the **marginal probabilities**

3.22 Example Addition Rule

Take

- Event $A : \{1, 2, 3\}$ a number less or equal to three in the throw of a dice
- Event $B : \{2, 4, 6\}$ an even number in the throw of a dice

then:

- $P(A) : P(1) + P(2) + P(3) = 3/6$
- $P(B) : P(2) + P(4) + P(6) = 3/6$
- $P(A \cap B) : P(2) = 1/6$

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

Note: $P(2)$ appears in $P(A)$ and $P(B)$ that's why we subtract it with the intersection

---

3.23 Venn diagram

Note that can always break down the sample space in **mutually exclusive** sets involving the intersections:

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

Marginals:

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

3.24 Probability table

Let's look at the probability table

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

3.25 Example probability table

We also write $A \cap B$ as (A, B) and call it the **joint probability** of A and B

In our example:

outcome	Probability
(A, B)	$P(A, B) = 1/6$
(A, B')	$P(A, B') = 2/6$
(A', B)	$P(A', B) = 2/6$
(A', B')	$P(A', B') = 1/6$
sum	1

Note: each outcome has *two* values (one for the characteristic of type A and another for type B)

3.26 Contingency table

We can organize the probability of **joint outcomes** in a **contingency table**

	B	B'	sum
A	$P(A, B)$	$P(A, B')$	$P(A)$
A'	$P(A', B)$	$P(A', B')$	$P(A')$
sum	$P(B)$	$P(B')$	1

Marginals:

- $P(A) = P(A, B') + P(A, B)$
- $P(B) = P(A', B) + P(A, B)$

3.27 Example contingency table

- Event $A : \{1, 2, 3\}$ a number less or equal to three in the throw of a dice

- Event $B : \{2, 4, 6\}$ an even number in the throw of a dice

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

Three forms of the **addition rule**:

$P(A \cup B)$

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

3.28 Misophonia study

In the misophonia study, the patients were assessed for their misophonia severity **and** if they were depressed.

The outcome of one random experiment is to measure the misophonia severity **and** depression status of one patient. The repetition of the random experiment was to perform the same two measurements on another patient.

```
##     Misofonia.dic depresion.dic
## 1               4             1
## 2               2             0
## 3               0             0
## 4               3             0
## 5               0             0
```

```
## 6             0              0
## 7             2              0
## 8             3              0
## 9             0              1
## 10            3              0
## 11            0              0
## 12            2              0
## 13            2              1
## 14            0              0
## 15            2              0
## 16            0              0
## 17            0              0
## 18            3              0
## 19            3              0
## 20            0              0
## 21            3              0
## 22            3              0
## 23            2              0
## 24            0              0
## 25            0              0
## 26            0              0
## 27            4              1
## 28            2              0
## 29            2              0
## 30            0              0
## 31            2              0
## 32            0              0
## 33            0              0
## 34            0              0
## 35            3              0
## 36            0              0
## 37            2              0
## 38            3              1
## 39            2              0
## 40            2              0
## 41            0              0
## 42            2              0
## 43            3              0
## 44            0              0
## 45            0              0
## 46            2              0
## 47            2              0
## 48            3              0
## 49            3              0
## 50            0              0
## 51            0              0
```

```
## 52                  4             1
## 53                  3             0
## 54                  3             1
## 55                  2             1
## 56                  0             1
## 57                  2             0
## 58                  0             0
## 59                  0             0
## 60                  0             0
## 61                  2             0
## 62                  2             0
## 63                  0             0
## 64                  0             0
## 65                  2             0
## 66                  3             1
## 67                  0             0
## 68                  1             0
## 69                  3             0
## 70                  2             0
## 71                  4             1
## 72                  3             0
## 73                  2             1
## 74                  3             0
## 75                  0             1
## 76                  2             0
## 77                  3             0
## 78                  2             0
## 79                  4             1
## 80                  1             0
## 81                  2             0
## 82                  0             0
## 83                  2             0
## 84                  0             0
## 85                  2             0
## 86                  0             1
## 87                  2             0
## 88                  2             0
## 89                  4             1
## 90                  3             0
## 91                  0             1
## 92                  3             0
## 93                  0             0
## 94                  0             0
## 95                  0             0
## 96                  2             0
## 97                  2             0
```

```
## 98                 1              0
## 99                 3              0
## 100                0              0
## 101                0              0
## 102                3              1
## 103                2              0
## 104                1              0
## 105                3              0
## 106                0              0
## 107                4              1
## 108                4              1
## 109                2              0
## 110                3              0
## 111                3              0
## 112                3              1
## 113                0              0
## 114                3              0
## 115                2              0
## 116                1              0
## 117                2              0
## 118                3              1
## 119                3              0
## 120                4              1
## 121                2              0
## 122                3              0
## 123                2              0
```

3.29 Contingency table for frequencies

- For the number of observations $n_{i,j}$ of each outcome (x_i, y_i), misophonia: $x \in \{0, 1, 2, 3, 4\}$ and depression $y \in \{0, 1\}$ (no:0, yes:1)

```
##
##                Depression:0 Depression:1
##   Misophonia:4            0            9
##   Misophonia:3           25            6
##   Misophonia:2           34            3
##   Misophonia:1            5            0
##   Misophonia:0           36            5
```

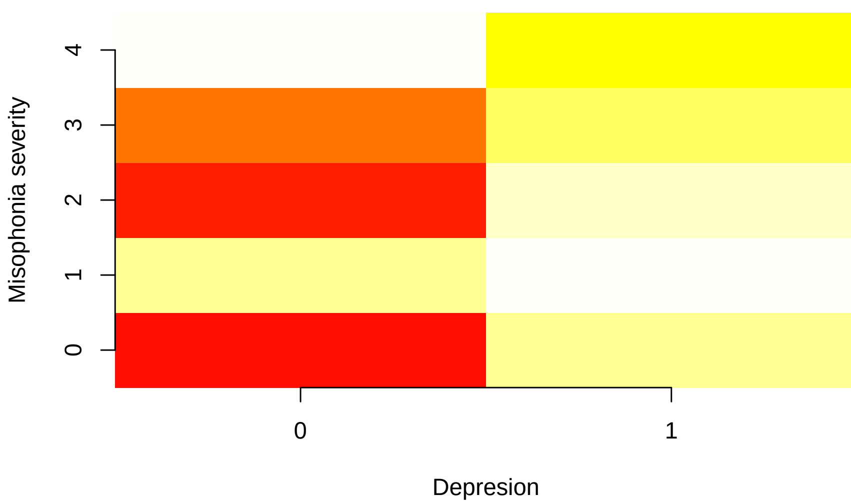
- For the relative frequencies $f_{i,j}$

```
##
```

```
##                 Depression:0 Depression:1
##   Misophonia:4   0.00000000   0.07317073
##   Misophonia:3   0.20325203   0.04878049
##   Misophonia:2   0.27642276   0.02439024
##   Misophonia:1   0.04065041   0.00000000
##   Misophonia:0   0.29268293   0.04065041
```

3.30 Heat map

The contingency table can be plotted as a **heat map**



3.31 Continous variables

In the misophonia study, the jaw protrusion was also measured as a possible cephalometric factor for de disease.

```
##     Angulo_convexidad protusion.mandibular
## 1                7.97                13.00
```

```
## 2                  18.23               -5.00
## 3                  12.27               11.50
## 4                   7.81               16.80
## 5                   9.81               33.00
## 6                  13.50                2.00
## 7                  19.30               -3.90
## 8                   7.70               16.80
## 9                  12.30                8.00
## 10                  7.90               28.80
## 11                 12.60                3.00
## 12                 19.00               -7.90
## 13                  7.27               28.30
## 14                 14.00                4.00
## 15                  5.40               22.20
## 16                  8.00                0.00
## 17                 11.20               15.00
## 18                  7.75               17.00
## 19                  7.94               49.00
## 20                 16.69                5.00
## 21                  7.62               42.00
## 22                  7.02               28.00
## 23                  7.00                9.40
## 24                 19.20              -13.20
## 25                  7.96               23.00
## 26                 14.70                2.30
## 27                  7.24               25.00
## 28                  7.80                4.90
## 29                  7.90               92.00
## 30                  4.70                6.00
## 31                  4.40               17.00
## 32                 14.00                3.30
## 33                 14.40               10.30
## 34                 16.00                6.30
## 35                  1.40               19.50
## 36                  9.76               22.00
## 37                  7.90                5.00
## 38                  7.90               78.00
## 39                  7.40                9.30
## 40                  6.30               50.60
## 41                  7.76               18.00
## 42                  7.30               18.00
## 43                  7.00               10.00
## 44                 11.23                4.00
## 45                 16.00               13.30
## 46                  7.90               48.00
## 47                  7.29               23.50
```

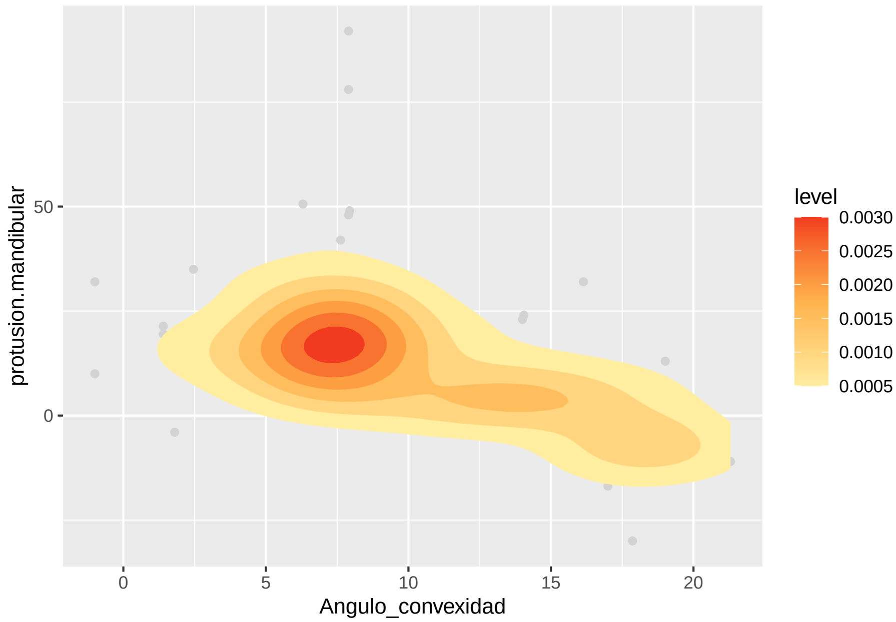
```
## 48             6.91             37.60
## 49             7.10             15.00
## 50            13.40              5.10
## 51            11.60             -2.20
## 52            -1.00             32.00
## 53             6.00             25.00
## 54             7.82             24.00
## 55             4.80             33.60
## 56            11.00              3.30
## 57             9.00             31.50
## 58            11.50             12.80
## 59            16.00              3.00
## 60            15.00              6.00
## 61             1.40             21.40
## 62            16.80            -10.00
## 63             7.70             19.00
## 64            16.14             32.00
## 65             7.12             15.00
## 66            -1.00             10.00
## 67            17.00            -16.90
## 68             9.26              2.00
## 69            18.70            -10.10
## 70             3.40             12.20
## 71            21.30            -11.00
## 72             7.50              5.20
## 73             6.03             16.00
## 74             7.50              5.80
## 75            19.00              5.20
## 76            19.01             13.00
## 77             8.10             13.60
## 78             7.80             16.10
## 79             6.10             33.20
## 80            15.26              4.00
## 81             7.95             12.00
## 82            18.00             -1.50
## 83             4.60             18.30
## 84            15.00              3.00
## 85             7.50             15.80
## 86             8.00             27.10
## 87            16.80            -10.00
## 88             8.54             25.00
## 89             7.00             27.10
## 90            18.30             -8.00
## 91             7.80             12.00
## 92            16.00             -8.00
## 93            14.00             23.00
```

```
## 94               12.30                   5.00
## 95               11.40                   1.00
## 96                8.50                  18.90
## 97                7.00                  15.00
## 98                7.96                  22.00
## 99               17.60                  -3.50
## 100              10.00                  20.00
## 101               3.50                  12.20
## 102               6.70                  14.70
## 103              17.00                  -5.00
## 104              20.26                  -4.15
## 105               6.64                  11.00
## 106               1.80                  -4.00
## 107               7.02                  25.00
## 108               2.46                  35.00
## 109              19.00                  -5.00
## 110              17.86                 -30.00
## 111               6.10                  12.20
## 112               6.64                  19.00
## 113              12.00                   1.60
## 114               6.60                  20.00
## 115               8.70                  17.10
## 116              14.05                  24.00
## 117               7.20                   7.10
## 118              19.70                 -11.00
## 119               7.70                  21.30
## 120               6.02                   5.00
## 121               2.50                  12.90
## 122              19.00                   5.90
## 123               6.80                   5.80
```

3.32 Heat map for continuous variables

- Two dimensional **histogram**.

- It illustrates the "continuous contingency" table for continuous variables

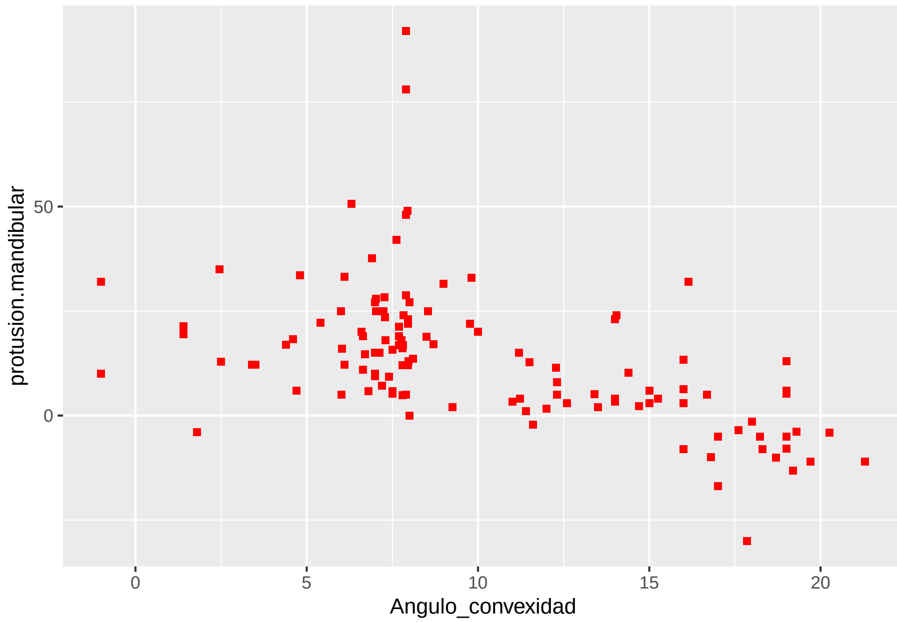


3.33 Scatter plot

- The **histogram** depends on the size of the bin (pixel).

- If the pixel is small enough to contain a single observation then the heat map results in a **scatter plot**

The scatter plot is the illustration of a "contingency table" for continuous variables when the bin (pixel) is small enough to contain one single observation (consisting of a pair of values).



protusion.mandibular
50
0
0
5
10
15
20
Angulo_convexidad

Chapter 4

Conditional Probability

4.1 Objective

- Conditional probability
- Independence
- Bayes' theorem

4.2 Joint Probability

The joint probability of two events A and B is

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

Let's imagine a random experiment that measures two different types of outcomes.

- height and weight of an individual: (h, w)
- time and place of an electric charge: (p, t)
- a throw of two dice: (n_1,n_2)
- cross two traffic lights in green: $(\bar{R_1}, \bar{R_2})$

In many cases, we are interested in finding out whether the values of one outcome **condition** the values of the other.

4.3 Diagnostics

Let's consider a **diagnostic tool**

We want to find the state of a system (s):

- inadequate (yes)
- adequate (no)

with a test (t):

- positive
- negative

We test a battery to find how long it can live. We stress a cable to find if it resists carrying a certain load. We perform a PCR to see if someone is infected.

4.4 Diagnostics Test

Let's consider diagnosing infection with a new test.

Infection status:

- yes (infected)
- no (not infected)

Test:

- positive
- negative

4.5 Observations

Each individual is a random experiment with two measurements: (Infection, Test)

Subject	Infection	Test
s_1	yes	positive
s_2	no	negative
s_3	yes	positive
...	...	...
s_i	no	positive*

Subject	Infection	Test
...	...	...
...	...	...
s_n	yes	negative*

4.6 Contingency tables

- For the number of observations of each outcome

	Infection: yes	Infection: no	sum
Test: positive	18	12	30
Test: negative	30	300	330
sum	48	312	360

- For the relative frequencies, if $N >> 0$ we will take $f_{i,j} = \hat{P}(x_i, y_j)$

	Infection: yes	Infection: no	sum
Test: positive	0.05	0.0333	0.0833
Test: negative	0.0833	0.833	0.9166
sum	0.133	0.866	1

4.7 Conditional probability

Let's think first in terms of those who are **infected**

Within those who are infected (**yes**), what is the probability of those who tested positive?

- Sensitivity (true positive rate)

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

$$= \frac{\frac{n_{positive,yes}}{N}}{\frac{n_{yes}}{N}} = \frac{f_{positive,yes}}{f_{yes}}$$

Therefore, in the limit, we expect to have a probability of the type

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

4.8 Conditional probability

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

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

- you can prove that the conditional probability satisfies the axioms of probability.
- it is the probability with the sampling space given by B: S_B.

4.9 Conditional contingency table

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

- True positive rate (Sensitivity): The probability of testing positive **if** having the disease $P(positive|yes)$
- True negative rate (Specificity): The probability of testing negative **if** not having the disease $P(negative|no)$
- False-positive rate: The probability of testing positive **if** not having the disease $P(positive|no)$
- False-negative rate: The probability of testing negative **if** having the disease $P(negative|yes)$

4.10 Example conditional contingency table

Taking the frequencies as estimates of the probabilities then

	Infection: Yes	Infection: No
Test: positive	18/48 = 0.375	12/312 = 0.038
Test: negative	30/48 = 0.625	300/312 =0.962
sum	1	1

Our diagnostic tool has low sensitivity (0.375) but high specificity (0.962).

4.11 Multiplication rule

Now let's imagine the real situation where we want to compute **joint** probabilities from conditional **probabilities**

- PCRs for coronavirus were (performed)[https://www.nejm.org/doi/full/10.1056/NEJMp2015897] in people in the hospital who we are sure to be infected. They have a sensitivity of 70%. They have also been tested in the lab in conditions of no infection with 96% specificity
- A prevalence study in Spain showed that $P(yes) = 0.05$, $P(no) = 0.95$ before summer.

With this data, what was the probability that a randomly selected person in the population tested positive **and** was infected: $P(yes \cap positive) = P(yes, positive)$?

4.12 Diagnostic performance

To study the performance of a new diagnostic test:

- you select specimens that are inadequate (disease: **yes**) and apply the test, trying to find its sensitivity: $P(positive|yes)$ (0.70 for PCRs)
- you select specimens that are adequate (disease: **no**) and apply the test, trying to find its specificity: $P(negative|no)$ (0.96 for PCRs)

	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

From this matrix, can we obtain $P(yes, positive)$?

4.13 Multiplication rule

How do you recover the joint probability from the conditional probability?

For two events A and B we have the multiplication rule

$$P(A, B) = P(A|B)P(B)$$

that follows from the definition of the conditional probability.

4.14 Contingency table in terms of conditional probabilities

	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

For instance the probability of testing *positive* and being infected *yes*:

- $P(positive, yes) = P(positive \cap yes) = P(positive|yes)P(yes)$

4.15 Conditional tree

4.16 Contingency table in terms of conditional probabilities

	Infection: yes	Infection: no	sum
Test: positive	0.035	0.057	0.092
Test: negative	0.015	0.893	0.908
sum	0.05	0.95	1

- $P(positive, yes) = 0.035$

But we also found the marginal of being positive:

- $P(positive) = 0.092$

4.17 Total probability rule

	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

When we write the unknown marginals in terms of their conditional probabilities we call it the **total probability rule**

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

4.18 Conditional tree

Total probability rule for the marginal of B: In how many ways I can obtain the outcome B?

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

4.19 Finding reverse probabilities

From the conditional contingency table

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

How can we calculate the probability of being infected if tested positive: $P(yes|positive)$?

4.20 Recover joint probabilities

1. We recover the contingency table for joint probabilities

	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

4.21 Reverse conditionals

2. We compute the conditional probabilities for the test:

$$P(infection|test) = \frac{P(test|infection)P(infection)}{P(test)}$$

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

For instance:

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

since we usually don't have $P(positive)$ we use the **total probability** rule in the denominator

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

4.22 Baye's theorem

The expression:

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

is called the **Bayes theorem**

Theorem

If $E1, E2, ..., Ek$ are k mutually exclusive and exhaustive events and B is any event,

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

It allows to reverse the conditionals:

$$P(B|A) \rightarrow P(A|B)$$

Or **design** a test B in controlled condition A and then use it to **infer** the probability of the condition when the test is positive.

4.23 Example: Bayes' theorem

Baye's theorem:

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

we know:

- $P(positive|yes) = 0.70$
- $P(positive|no) = 1 - P(negative|no) = 0.06$
- the probability of infection and not infection in the population: $P(yes) = 0.05$ and $P(no) = 1 - P(yes) = 0.95$.

Therefore:

$$P(yes|positive) = 0.47$$

Tests are not so good to **confirm** infections.

4.24 Example: Bayes' theorem

Let's now apply it to the probability of not being infected if the test is negative

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

Substitution of all the values gives

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

Tests are good to **rule out** infections.

4.25 Statistical independence

In many applications, we want to know if the knowledge of one event conditions the outcome of another event.

- there are cases where we want to know if the events are not conditioned

4.26 Statistical independence

Consider conductors for which we measure their surface flaws and if their conduction capacity is defective

The estimated **joint probabilities** are

	flaws (F)	no flaws (F')	sum
defective (D)	0.005	0.045	0.05
no defective (D')	0.095	0.855	0.95
sum	0.1	0.9	1

where, for instance, the joint probability of F and D is

- $P(D, F) = 0.005$

The marginal probabilities are

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

4.27 Statistical independence

What is the **conditional probability** of observing a defective conductor if they have a flaw?

	F	F'
D	P(D|F) = 0.05	P(D|F')=0.05
D'	P(D'|F)=0.95	P(D'|F')=0.95
sum	1	1

The marginals and the conditional probabilities are the same!

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

The probability of observing a defective conductor **does not** depend on having observed or not a flaw.

$$P(D) = P(D|F)$$

4.28 Statistical independence

Two events A and B are statistically independent if

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

and by the multiplication rule, their joint probability is

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

the multiplication of their marginal probabilities.

4.29 Products of marginals products

	F	F'	sum
D	0.005	0.045	0.05
D'	0.095	0.855	0.95
sum	0.1	0.9	1

Confirm that all the entries of the matrix are the product of the marginals.

For example:

- $P(F)P(D) = P(D \cap F)$
- $P(D')P(F') = P(D' \cap F')$

4.30 Example

Outcomes of throwing two coins: $S = (H, H), (H, T), (T, H), (T, T)$

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

- Obtaining a head in the first coin does not condition obtaining a tail in the result of the second coin $P(T|H) = P(T) = 1/2$
- the probability of obtaining a head and then a tail is the product of each independent outcome $P(H, T) = P(H) * P(T) = 1/4$

Chapter 5

Discrete Random Variables

5.1 Objective

- Random variables
- Probability mass function
- Mean and variance
- Probability distribution

5.2 How do we assign probability values to outcomes?

5.3 Random variable

Definition:

A **random variable** is a function that assigns a real **number** to each **outcome** in the sample space of a random experiment.

- Most commonly a random variable is the value of the **measurement** of interest that is made in a random experiment.

A random variable can be:

- Discrete (nominal, ordinal)

- Continuous (interval, ratio)

5.4 Random variable

A **value** (or **outcome**) of a random variable is one of the possible numbers that the variable can take in a random experiment.

We write the random variable in **capitals**.

Example:

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

Observation of a random variable

- An observation is the **acquisition** of the value of a random variable in a random experiment

Example:

1 0 0 1 0 **1** 0 1 1

The number in bold is an observation of X

5.5 Events of observing a random variable

- $X = 1$ 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 with value x (little x)
- Any two values of a random variable define two **mutually exclusive** events.

5.6 Probability of random variables

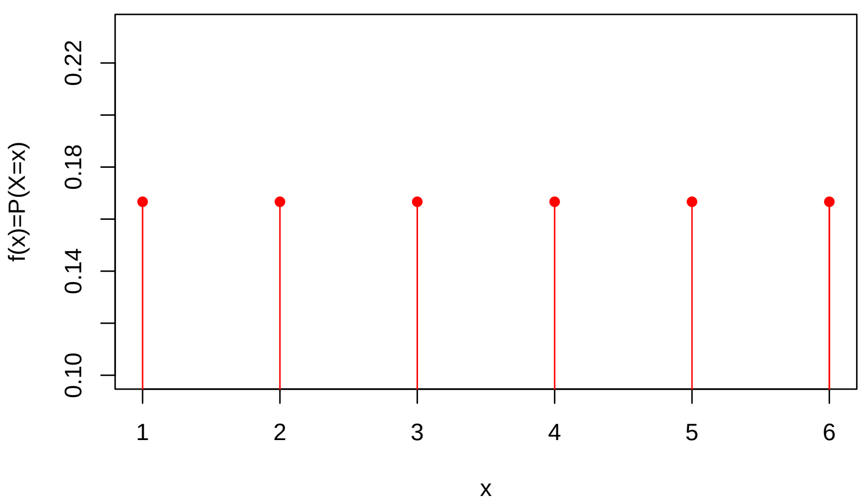
We are interested in assigning probabilities to the values of a random variable.

We have already done this for the dice: $X \in \{1, 2, 3, 4, 5, 6\}$ (classical interpretation of pribability)

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

5.7 Probability functions

- We can write the probability table
- plot it

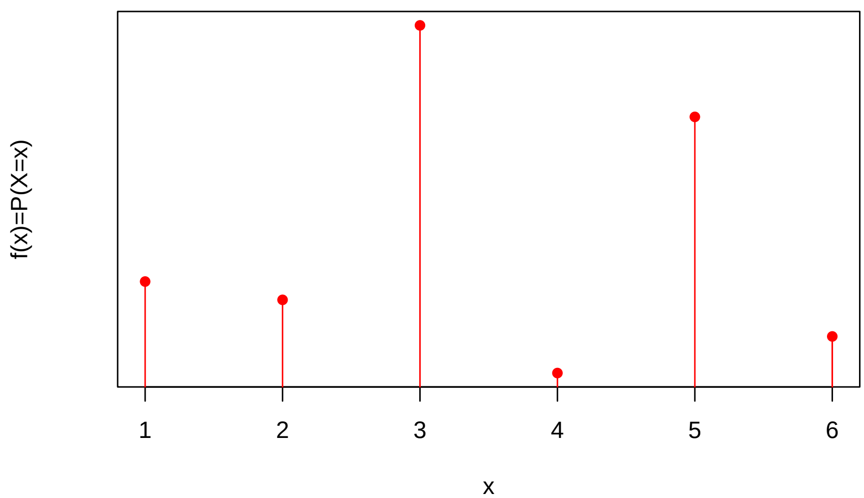


- or write as the function

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

5.8 Probability functions

We can **create** any type of probability function if we respect the probability rules:



5.9 Probability functions

For a discrete random variable $X \in \{x_1, x_2, .., x_M\}$, a **probability mass function**

is always positive

- $f(x_i) \geq 0$

is used to compute probabilities

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

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

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

5.10 Probability functions

- 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.
- 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.
- They have some **properties** that are derived exclusively from their definition.

5.11 Example: Probability mass function

Consider the following random variable X over the outcomes

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

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

5.12 Probability table for equally likely outcomes

outcome	Probability(outcome)
a	1/6
b	1/6
c	1/6
d	1/6
e	1/6
f	1/6

5.13 Probability table for X

X	$f(x) = P(X = x)$
0	$P(X = 0) = 2/6$
1.5	$P(X = 1.5) = 2/6$
2	$P(X = 2) = 1/3$
3	$P(X = 3) = 1/3$

We can compute, for instance, the following probabilities for events on the values of X

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

5.14 Example

Consider:

- we do not know what the primary events with equal probabilities are.
- we then **estimate** the probability mass function from the relative frequencies observed for a random variable

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

5.15 Example

Probability model:

These probabilities are consistent with the following experiment: 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

experiment: Take one ball and read the number.

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	$1/8 = 0.125$

5.16 Probabilities and frequencies

For computing the relative frequencies f_i you have to

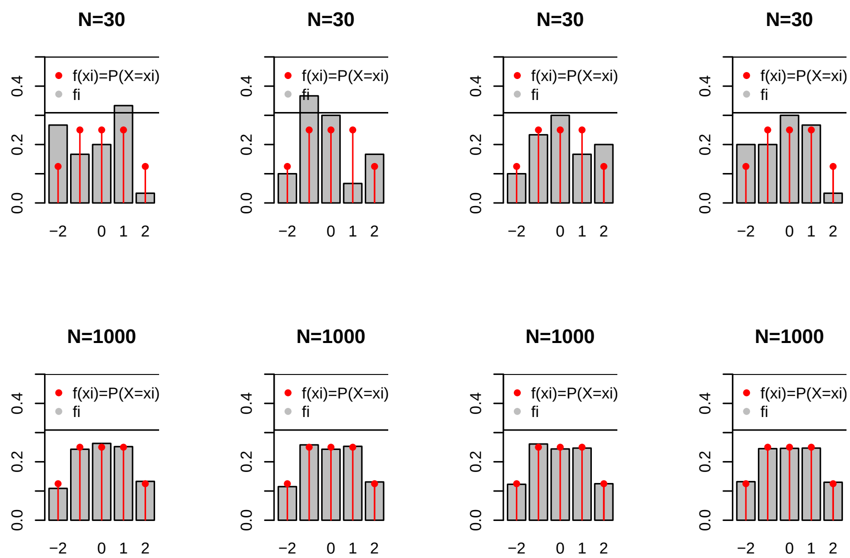
- **repeat** the experiment N times (you have to put the ball back in the urn each time) and at the end compute

$$f_i = n_i/N$$

We are assuming that:

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

5.17 Probabilities and relative frequencies



- In this example we **know** the probability **model** $f(x) = P(X = x)$ by design.
- We never observe $f(x)$
- We can use relative frequencies to estimate the probabilities

$$f_i = \hat{f}(x_i) = \hat{P}(X = x_i)$$

(f_i depends on N)

5.18 Mean and Variance

The probability mass functions $f(x)$ have two main properties

- its center
- its spread

We can ask,

- around which values of X the probability concentrated?

- How dispersed are the values of X in relation to their probabilities?

5.19 Mean and Variance

5.20 Mean

Remember that the **average** in terms of the relative frequencies of the values of x_i (categorical ordered outcomes) can be written 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

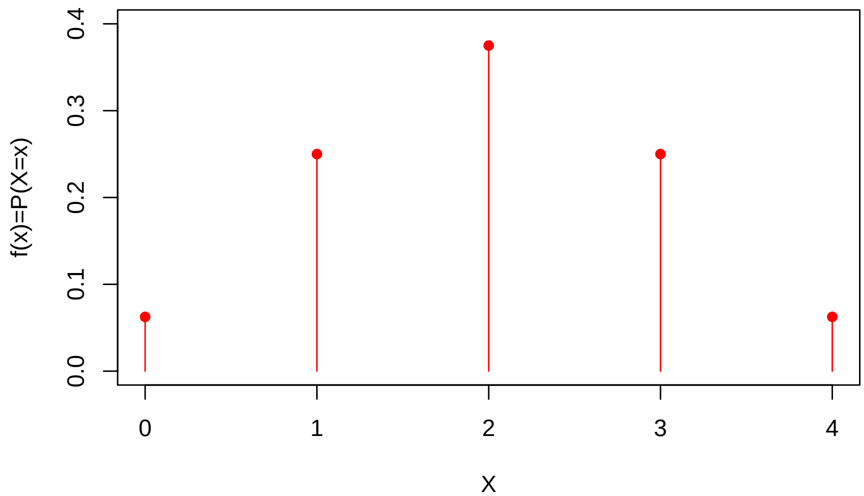
$$\mu = E(X) = \sum_{i=1}^{M} x_i f(x_i)$$

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

5.21 Example: Mean

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

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



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

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

5.22 Variance

In similar terms we define the mean squared distance from the mean:

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
- Think of it as the moment of inertia of probabilities about the mean.

5.23 Example: Variance

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

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

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

$V(X)$ =**(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

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

$$\sigma = 1$$

5.24 Functions of X

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 important properties of the median and variance:

- 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).

- The variance of a linear function of X is:

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

- The variance **about the origin** is the variance **about the mean** plus the mean squared:

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

5.25 Example: Variance about the origin

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

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

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

$E(X^2) =$**(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:

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

$5 = 1 + 2^2$

5.26 Probability distribution

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:

- $0 \leq F(x) \leq 1$
- If $x \leq y$, then $F(x) \leq F(y)$

5.27 Example: Probability distribution

For the probability mass function:

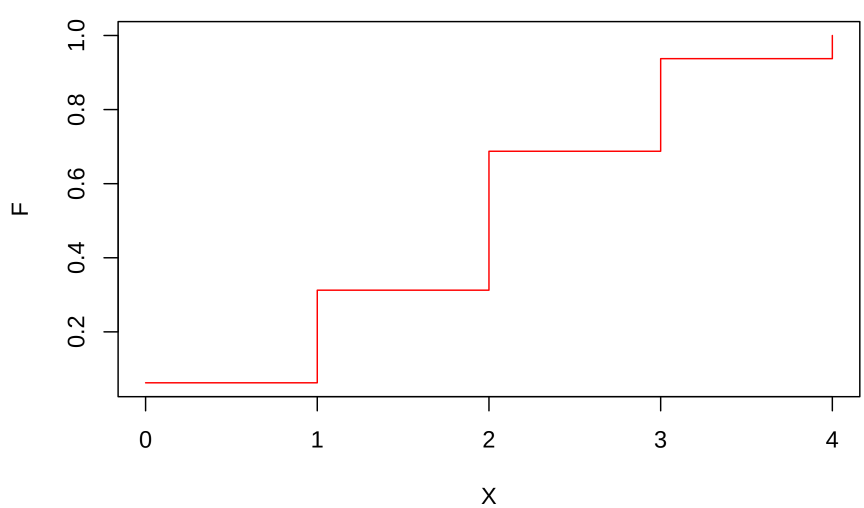
$f(0) = P(X = 0) = 1/16$ $f(1) = P(X = 1) = 4/16$ $f(2) = P(X = 2) = 6/16$ $f(3) = P(X = 3) = 4/16$ $f(4) = P(X = 4) = 1/16$

The probability distribution is:

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

For$X \in \mathbb{Z}$

5.28 Probability distribution



5.29 Probability function and Probability distribution

Compute the mass probability function of the following probability distribution:

$F(0) = 1/16$, $F(1) = 5/16$, $F(2) = 11/16$, $F(3) = 15/16$, $F(4) = 16/16$,

Let's work backward.

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

5.30 Probability function and Probability distribution

The Probability distribution is another way to specify the probability of a random variable

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

with

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

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

5.31 Quantiles

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 0.95-quantile is the value x_s such that $q = 0.95$

$$F(x_s) = 0.95$$

5.32 Summary

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

Note that:

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

Properties:

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

Chapter 6

Continous Random Variables

6.1 Objective

- Probability density function
- Mean and variance
- Probability distribution

6.2 Continuous random variable

What happens with continuous random variables?

Let's reconsider the convexity angle of misophonia patients (Section 2.21).

- We redefined the outcomes as little regular intervals (bins) and computed the relative frequency for each of them as we did in the discrete case.

```
##         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 Continuous random variable

Let's consider again that their relative frequencies are the probabilities when $N \to \infty$

$$f_i = \frac{n_i}{N} \to f(x_i) = P(X = x_i)$$

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

$P(X = x_i) \to 0$ when $\Delta x \to 0$, because $n_i \to 0$

```
##          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 Continuous random variable

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

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

6.5 Continuous random variable

Definition

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

The function is positive:

- $f(x) \geq 0$

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 probability of observing **any** value is 1:

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

6.6 Continuous random variable

- 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 ($\sum \to \int$).
- Assignment of probabilities to a random variable can be done with equiprobability (classical) arguments.
- Densities are mathematical quantities some will map to experiments some will not. *Which density will map best to my experiment?*

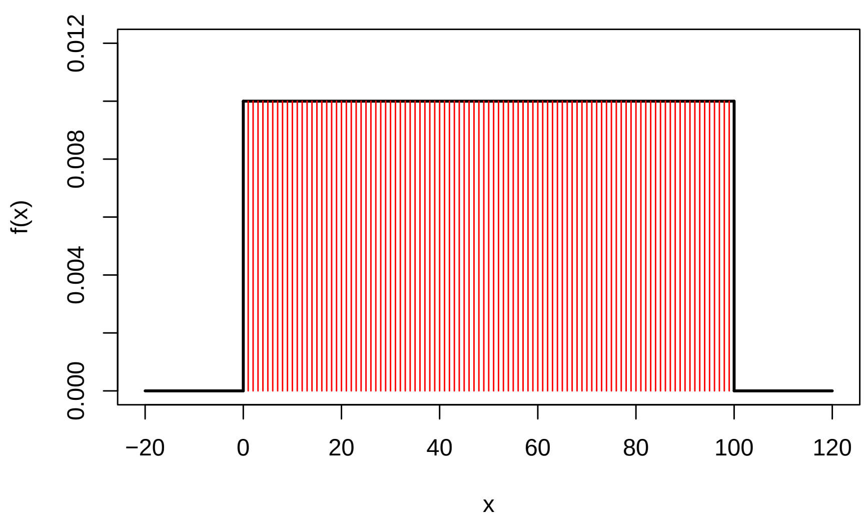
6.7 Total area under the curve

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

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

Then the probability of **any** observation is the total **area under the curve**

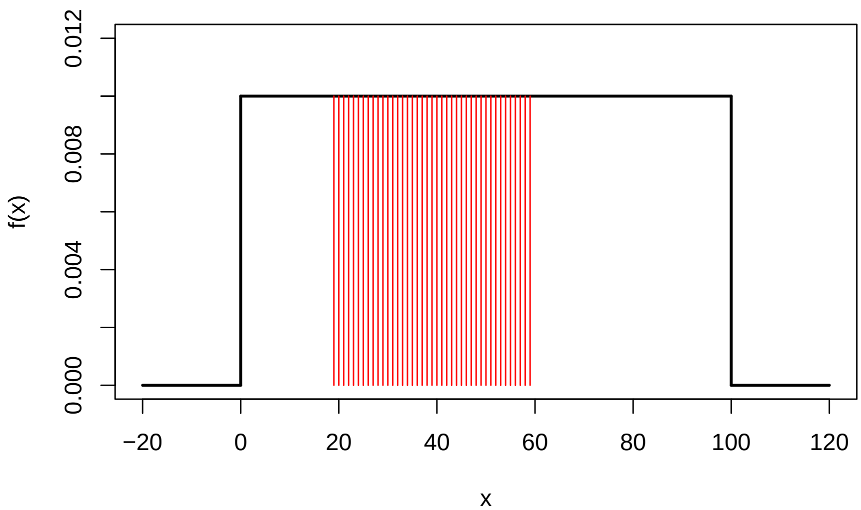
$P(-\infty \leq X \leq \infty) = \int_{-\infty}^{\infty} f(x)dx = 100 * 0.01 = 1$



6.8 Area under the curve

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

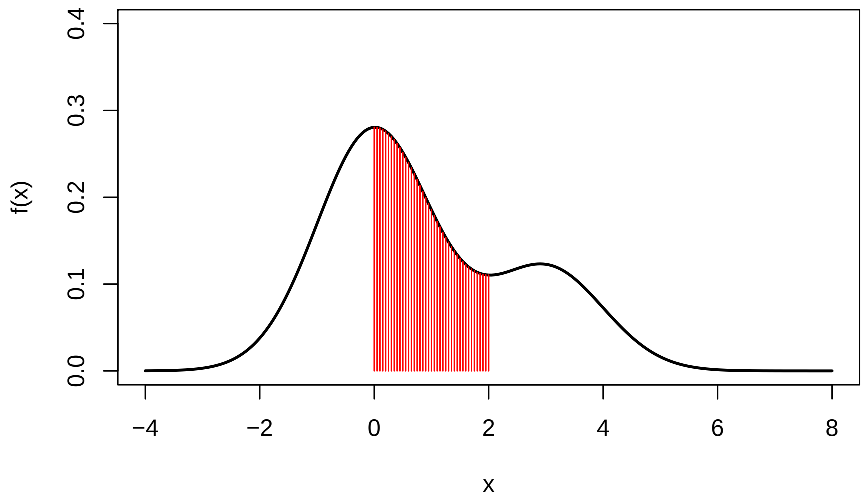
- $P(20 \leq X \leq 60) = \int_{20}^{60} f(x)dx = (60 - 20) * 0.01 = 0.4$



6.9 Area under the curve

In general $f(x)$ should satisfy:

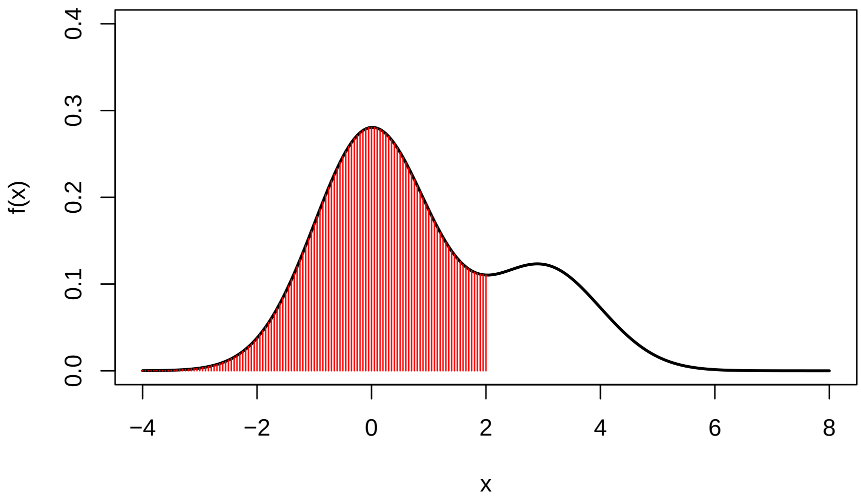
- $0 \leq P(a \leq X \leq b) = \int_a^b f(x)dx \leq 1$



6.10 Probability distribution

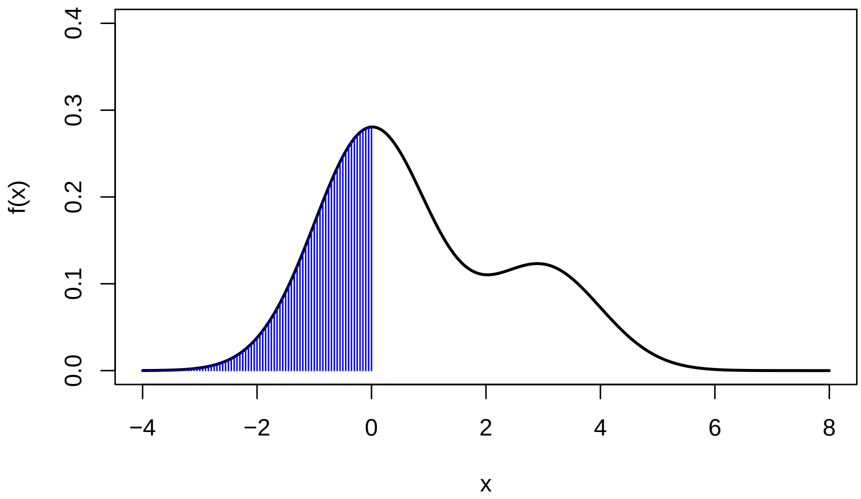
The probability accumulated up to b is defined by the probability distribution F

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



The probability accumulated up to a is

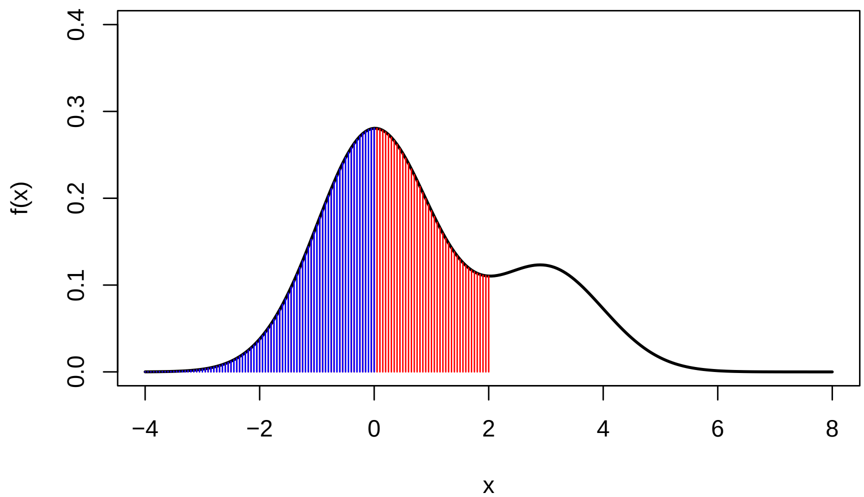
- $F(a) = P(X \leq a)$



6.11 Probability distribution

The probability between a and b is defined by the probability distribution F

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



6.12 Probability distribution

The probability distribution of a continuous random variable is defined as $F(a) = P(X \leq a) = \int_{-\infty}^a f(x)dx$

with the properties that:

It is between 0 and 1:

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

It always increases:

- if $a \leq b$ then $F(a) \leq F(b)$

It can be used to compute probabilities:

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

It recovers the probability density:

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

We use **probability distributions** to **compute probabilities** of a random variable with intervals

6.13 Probability distribution

For the uniform density function:

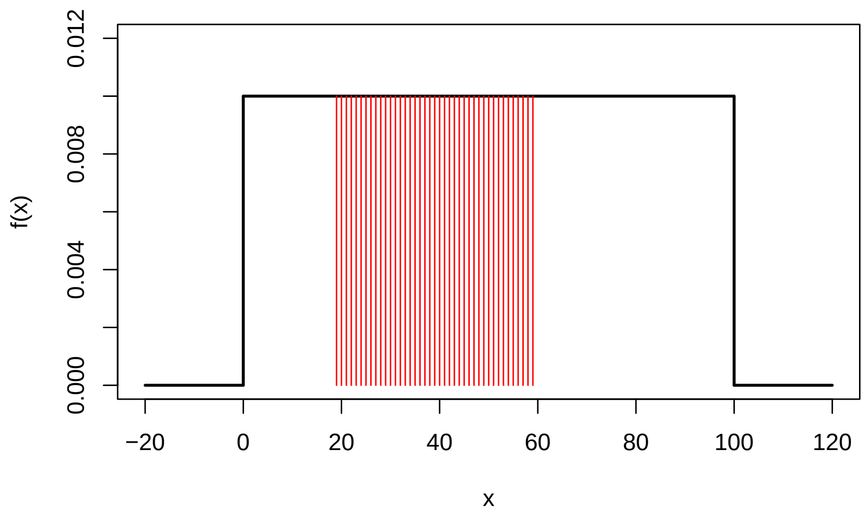
$$f(x) = \begin{cases} \frac{1}{100}, & \text{if } x \in (0, 100) \\ 0, & otherwise \end{cases}$$

The probability distribution is

$$F(a) = \begin{cases} 0, & a \leq 0 \\ \frac{a}{100}, & \text{if } a \in (0, 100) \\ 1, & 10 \leq a \\ \end{cases}$$

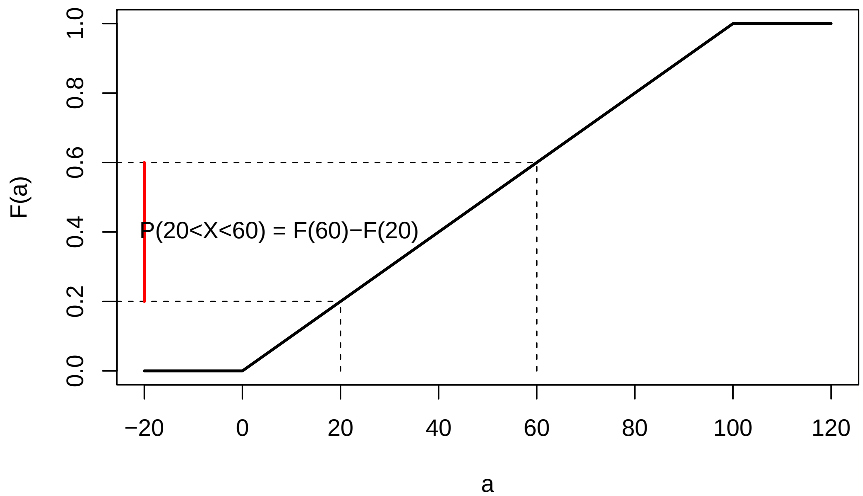
6.14 Probability graphics

The probability $P(20 < X < 60)$ is the *area* under the **density** curve



6.15 Probability graphics

The probability $P(20 < X < 60)$ is the *difference* in **distribution** values



6.16 Mean

As in the discrete case, the **mean** measures the center of the distribution

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

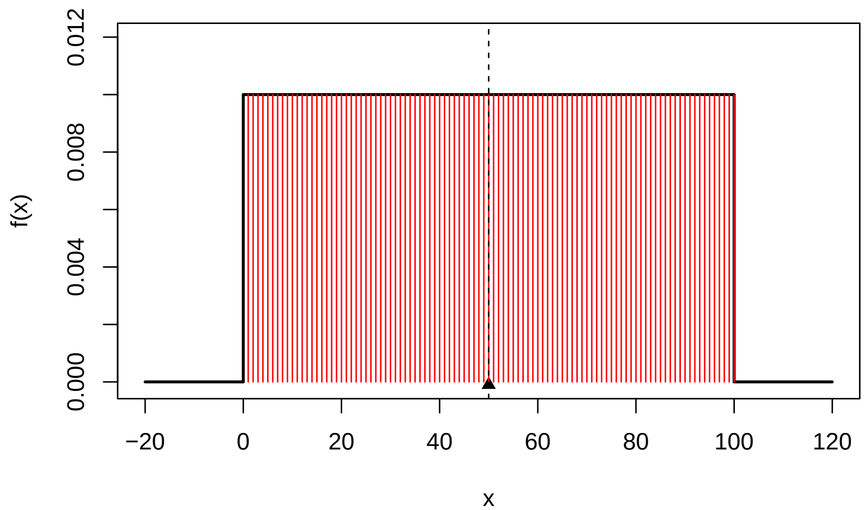
$$\mu = E(X) = \int_{-\infty}^{\infty} x f(x) dx$$

It is the continuous version of the center of mass.

6.17 Mean

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

$E(X) = 50$



6.18 Variance

As in the discrete case, the variance measures the dispersion 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

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

6.19 Functions of X

Definition

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

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

And we have the same properties as in the discrete case

- 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.

- The variance of a linear function of X is:

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

- The variance about the origin is the variance about the mean plus the mean squared:

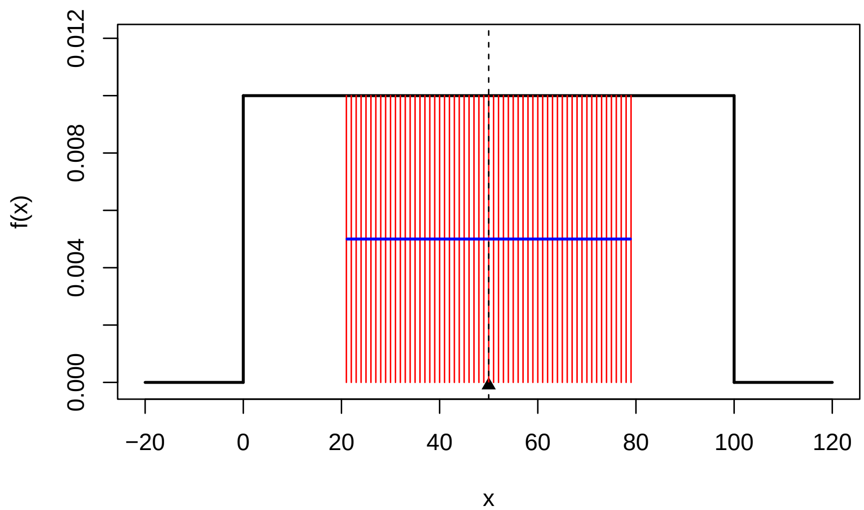
$$E(X^2) = V(X) + E(X)^2$$

6.20 Example

- for the probability density

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

- compute the mean
- compute variance using $E(X^2) = V(X) + E(X)^2$
- compute $P(\mu - \sigma \leq X \leq \mu + \sigma)$
- What are the first and third quartiles?


0.012
0.008
0.004
0.000
f(x)
−20
0
20
40
60
80
100
120
x

Chapter 7

Discrete Probability Models

7.1 Objective

Discrete probability models:

- Uniform and Bernoulli probability functions
- Binomial and negative binomial probability functions

7.2 Probability mass function

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

Positive:

- $f(x_i) \geq 0$

Allow us to compute probabilities:

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

The probability of any outcome is 1

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

Properties:

Central tendency:

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

Dispersion:

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

They are abstract objects with general properties that may or may not **describe** a natural or engineered process.

7.3 Probability model

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

Examples:

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

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 per label: $-1, 0, 1$ and one ball per label: $-2, 2$.

7.4 Parametric models

When we perform a random experiment and **do not** know the probabilities of the outcomes:

- We can always formulate the model given by the relative frequencies: $\hat{P}(X = x_i) = f_i$ (where $i = 1...M$).

We need to find M numbers each depending on N.

In many cases:

- We can formulate probability functions $f(x)$ that depend on **very few** numbers only.

Example:

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)

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

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

With mean and variance:

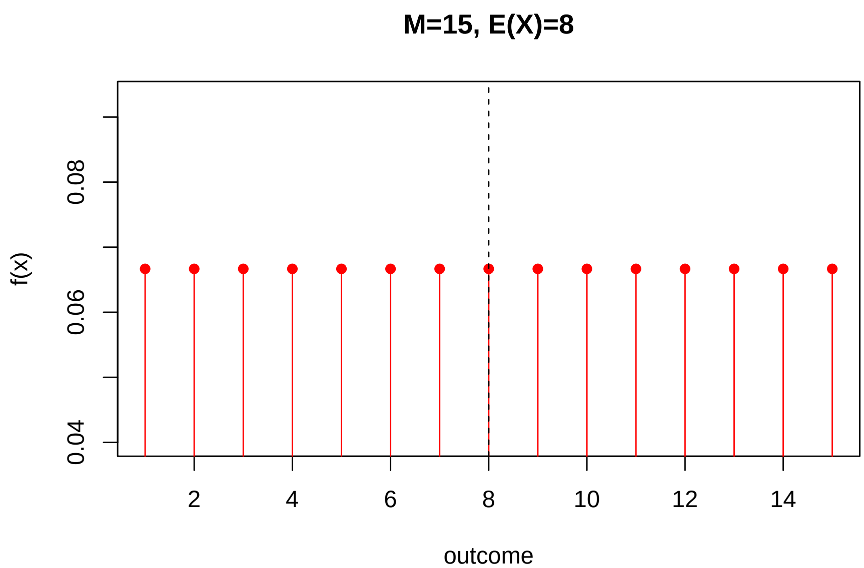
$E(X) = \frac{M+1}{2}$

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

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

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

7.6 Uniform distribution



7.7 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

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

7.8 Uniform distribution (two parameters)

Example:

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)?

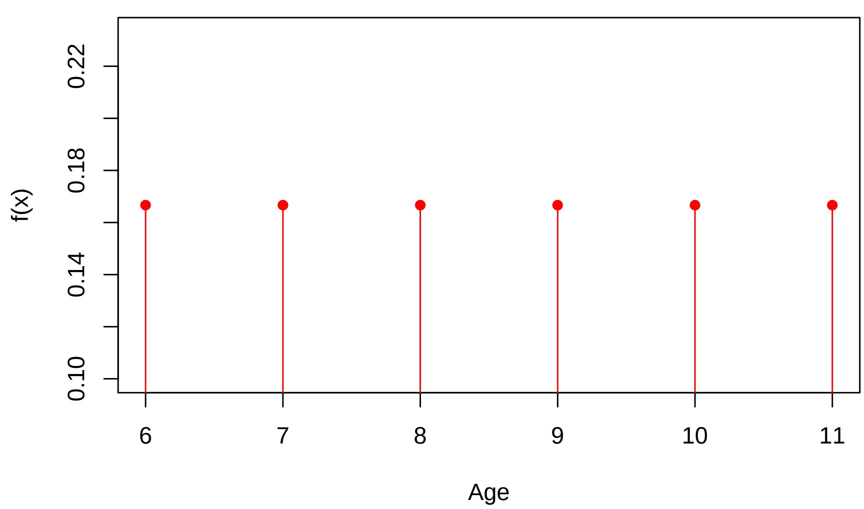
From the experiment we know: $a = 6$ and $b = 11$ then

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

that is

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

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



7.9 Uniform distribution

The probability model of a random variable X

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

for $x \in \{a, a+1, ...b\}$

has mean and variance:

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

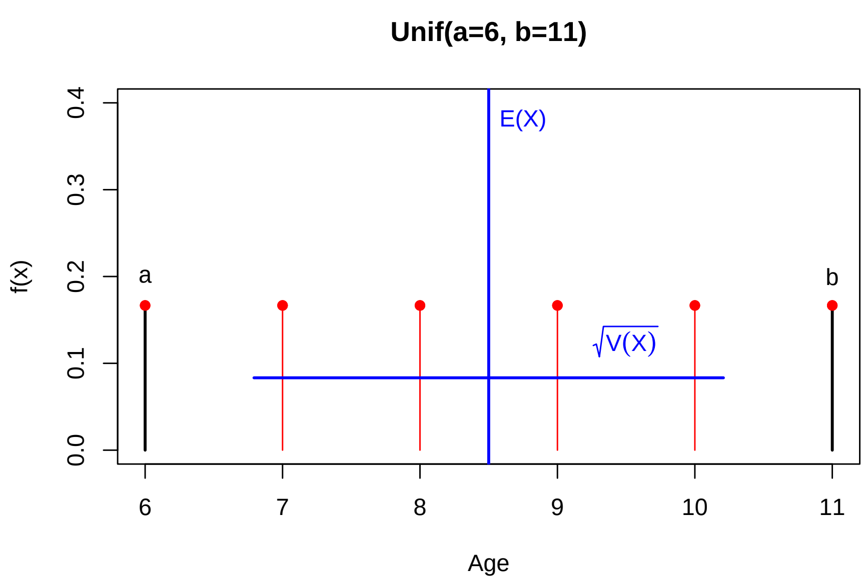
(Change variables $X = Y + a - 1$, $y \in \{1, ...M\}$)

We can either specify a and b or $E(X)$ and $V(X)$.

In our example:

- $E(X) = (11+6)/2 = 8.5$
- $V(X) = (6^2-1)/12 = 2.916667$

7.10 Uniform distribution (two-parameter)



7.11 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**

- Knowledge of $f(x)$ is reduced to the knowledge of the value of the parameters

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

Example:

Model: The data of our experiment is produced by a random process in which each age has the **same probability** of being observed.

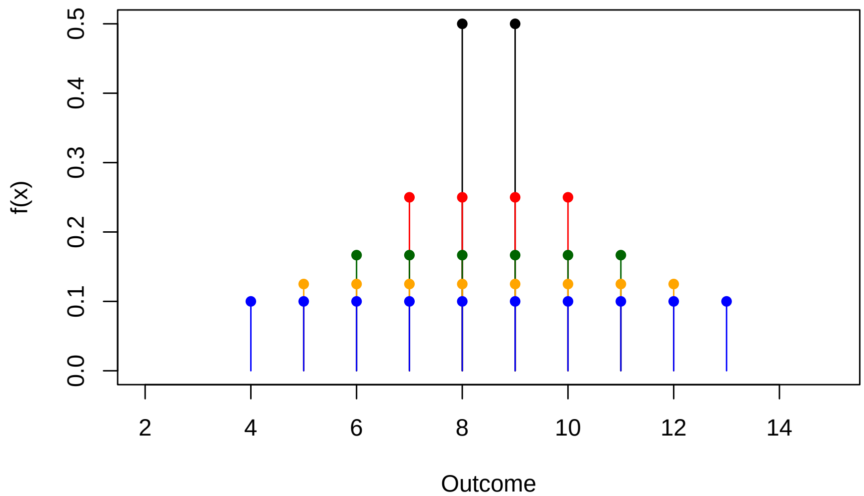
Parameters: a is the minimum age, $E(X)$ is the expected age ... they are **physical properties** of the experiment.

7.12 Parameters and Models

Example:

A **family** of models obtained from two-parameter uniform distributions changing the **variances** and keeping a constant mean ($E(X) = 8.5$). It results on **changing** both **minimum** and **maximum** outcomes.

- Note: Only one model makes sense for our experiment (only one model can represent the ages of children in a school).



- We can think of **families** that change only the **mean**, only the **minimum**, or only the **maximum**

7.13 Bernoulli trial

Let's try to advance from the equal probability case and suppose a model with two 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 $B : fine$).
- 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**.

7.14 Bernoulli trial

We will introduce the probability of an outcome (A) as the **parameter** of the model:

- outcome A (success): has probability p (parameter)
- outcome B (failure): has a probability $1 - p$

Or write, the probability mass function of K taking values $\{0, 1\}$ for A and B

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

or more shortly

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

for $k = (0, 1)$

We only need to know p.

7.15 Bernoulli trial

A Bernoulli variable K with outcomes $\{0, 1\}$ has a probability mass function

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

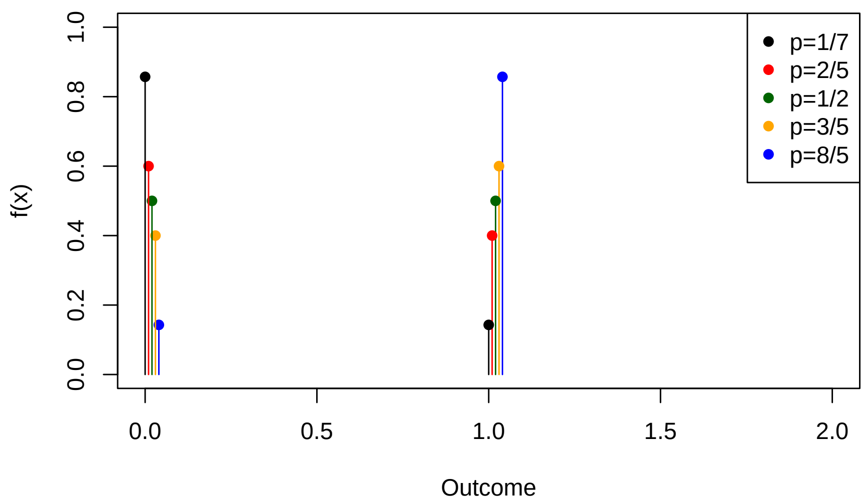
With mean and variance:

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

Note:

- The probability of the outcome A is the parameter p which is the same as $f(0) = P(X = 0)$.
- As p is usually **unknown** we typically estimated it by the relative frequency (more on this in the inference sections): $\hat{p} = f_A = \frac{n_A}{N}$

7.16 Bernoulli trial



7.17 Binomial distribution

When we are interested in learning about a particular Bernoulli trial

- We repeat the Bernoulli trial N times and count how many times we obtained A (n_A).
- We define a random variable $X = n_A$ taking values $x \in 0, 1, ...N$

We now ask for the probability of observing x events of type A in the repetition of n independent Bernoulli trials, when the probability of observing A is p.

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

7.18 Examples: Binomial distribution

- Writing down the sex of $n = 10$ patients who go into an emergency room of a hospital. What is the probability that $x = 6$ patients are men when $p = 0.9$?
- 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 $p = 0.1$?

7.19 Binomial distribution

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

Consider 4 random variables: K_1, K_2, K_3 and K_4 that record whether an error has been made in the 1^{st}, 2^{nd}, 3^{rd} and 4^{th} pixel.

Then

- k_i takes values $\{correct : 0; error : 1\}$
- $X = \sum_{i=1}^{4} K_i$ takes values $\{0, 1, 2, 3, 4\}$

Then the probability of observing 4 errors is:

- $P(X = 4) = P(1, 1, 1, 1) = p * p * p * p = p^4$ because K_i are independent.

The probability of 0 errors is:

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

The probability of 3 errors is:

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

7.20 Binomial distribution

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.

7.21 Binomial distribution: 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.

7.22 Binomial distribution: Mean and Variance

The mean and variance of $X \hookrightarrow Bin(n, p)$ are

- $E(X) = np$
- $V(X) = np(1 - p)$
- Since X is the sum of n independent Bernoulli variables

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

and

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

Example:

- 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 $n(1 - p)p = 0.36$

Remember: We can specify either the parameters n and p, or the parameters $E(X)$ and $V(X)$

7.23 Example 1

Now let's answer:

- What is the probability of observing 4 errors when transmitting 4 pixels, if the probability of an error is 0.1?

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

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

That is

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

7.24 Example 1

- We want to compute:

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

In R dbinom(4,4,0.1)

- We can also compute:

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

In R dbinom(2,4,0.1)

7.25 Example 2

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

For this case

$$X \rightarrow Bin(n = 8, p = 0.9)$$

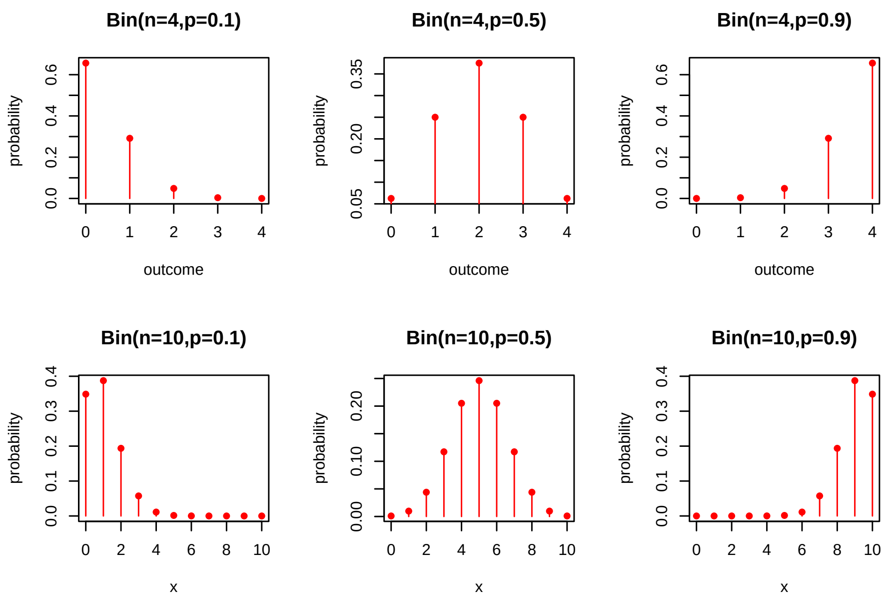
That is

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

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)

7.26 Binomial distribution



7.27 Negative binomial distribution

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.

- Experiment: Suppose performing Bernoulli trials until we observe the outcome A appears r times.
- Random variable: We count the number of events B
- Example: What is the probability of observing y well-transmitted (B) pixels before r errors (A)?

7.28 Negative binomial distribution

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

$(0, 0, 1, ., 0, 1, ...0, 1)$ (there are y zeros, and r ones)

We observe y correct pixels in a total of $y + r$ trials.

Then

- $P(0, 0, 1, ., 0, 1, ...0, 1) = p^r(1-p)^y$ (Remember: p is the probability of error)

How many transmissions can have y correct pixels before r errors?

Note:

- The last bit is fixed (marks the end of transmission)
- The total number of transmissions with y number of correct pixels (B) that we can obtain in $y + r - 1$ trials is: $\binom{y+r-1}{y}$

7.29 Negative binomial distribution

Therefore, the probability of observing y events of type B before r events of type A (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.

7.30 Mean and Variance

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

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

7.31 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

$$P(Y = y) = f(y) = p(1-p)^y$$

$$Y \to Geom(p)$$

with mean

- mean: $E(Y) = \frac{1-p}{p}$
- variance: $V(Y) = \frac{1-p}{p^2}$

7.32 Example

- 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?

7.33 Example

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

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

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

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

- Note that there are actually 6000 trials

7.34 Example

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

Recall the cumulative function distribution $F(y) = P(Y \leq 5)$

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

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

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

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

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

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

In R pnbinom(5,3,0.0005)

7.35 Examples

With the negative binomial probability function:

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

We can now answer questions like:

- What is the probability of observing 10 correct pixels before 2 errors, if the probability of an error is 0.1?

$f(10; r = 2, p = 0.1) = 0.03835463$

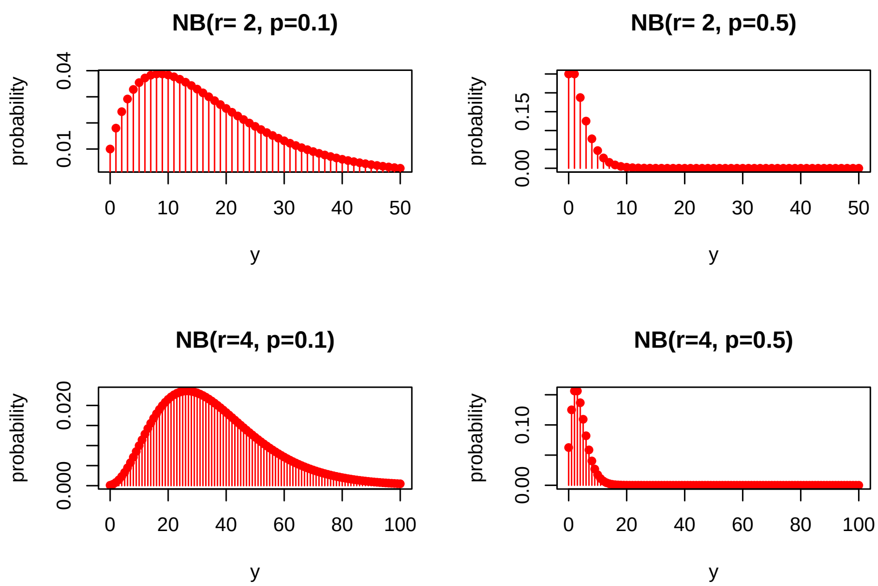
in R dnbinom(10, 2, 0.1)

- 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)

7.36 Negative binomial distribution



Chapter 8

Poisson and Exponential Models

8.1 Objective

Discrete probability model:

- Poisson

Continuous probability model:

- Exponential

8.2 Discrete probability models

We are building up more complex models from simple ones:

Uniform: Classical interpretation of probability ↓ **Bernoulli**: Introduction of a **parameter** p (family of models) ↓ **Binomial**: **Repetition** of a random experiment (n-times Bernoulli trials) ↓ **Poisson**: Repetition of random experiment within a continuous interval, having **no control** on when/where the Bernouilli trial occurs.

8.3 Counting events

Imagine that we are observing events that **depend** on time or distance **intervals**.

- 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.

8.4 Counting events

What is the probability of observing X events in an interval's unit (time or distance)?

Imagine that some impurities in a copper wire deposit randomly along a wire

- at each centimeter, you would count an average of $\lambda = 10/cm$.
- divide the centimeter into micrometers ($0.0001cm$)

8.5 Poisson distribution

micrometers are small enough so

- either there is or there is not an impurity in each micrometer
- each micrometer can be considered a **Bernoulli trial**

8.6 Poisson distribution

The probability of observing X impurities in $n = 10,000\mu$ (1cm) approximately follows a binomial distribution

$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$

where p is the probability of finding an impurity in a micrometer.

Remember that $E(X) = np$ so for $\lambda = np$ (average number of impurities per 1cm), we can write

$$P(X = x) = \binom{n}{x}(\frac{\lambda}{n})^x(1 - \frac{\lambda}{n})^{n-x}$$

- There **could** still be two impurities in a micrometer so we need to increase the partition of the wire and $n \to \infty$.

Then in the limit:

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

Where λ is constant because it is the density of impurities per centimeter, a **physical property** of the system.

8.7 Poisson distribution: Derivation details

For $P(X = x) = \binom{n}{x}(\frac{\lambda}{n})^x(1 - \frac{\lambda}{n})^{n-x}$

in the limit $(n \to \infty)$

- $\frac{1}{n^x}\binom{n}{x} = \frac{1}{n^x}\frac{n!}{x!(n-x)!} = \frac{(n-x)!(n-x+1)\ldots(n-1)n}{n^x x!(n-x)!} = \frac{n(n-1)..(n-x+1)}{n^x x!} \to \frac{1}{x!}$
- $(1 - \frac{\lambda}{n})^n \to e^{-\lambda}$ (definition of exponential)
- $(1 - \frac{\lambda}{n})^{-x} \to 1$

Therefore $P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}$

8.8 Poisson distribution

Definition

Given

- an interval in the real numbers
- counts occur at random in the interval
- the average number of counts on the interval is known (λ)
- if one can find a small regular partition of the interval such that each of them can be considered Bernoulli trials

Then…

8.9 Poisson distribution

Definition

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:

- mean $E(X) = \lambda$
- variance $V(X) = \lambda$

8.10 Poisson distribution

With the Poisson probability function:

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

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

We can now answer questions like:

- What is the probability of receiving 4 emails in an hour, when the average number of emails in **two** hours is 1?

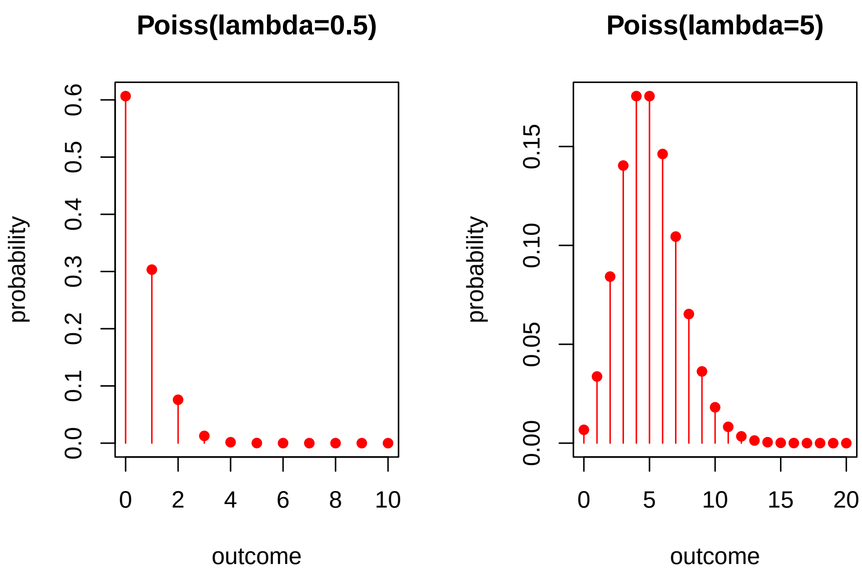
$f(4; \lambda = 0.5) = 0.001579507$

in R dpois(2,0.5)

- What is the probability of counting at least 10 cars arriving at a road toll in a minute, when the average number of cars that arrive at the toll in a minute is 5; $P(X \leq 10) = F(10; \lambda = 5) = 0.9863047$?

in R ppois(10,5)

8.11 Poisson distribution



8.12 Continuous probability models

Continuous probability models are probability density functions $f(x)$ of a continuous random variables that we **believe** describe real random experiments.

Definition:

Positive:

- $f(x) \geq 0$

Allows us to compute probabilities using the area under the curve:

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

The probability of any value is 1:

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

8.13 Exponential density

Let's go back to the Poisson probability for the number of events (k) in an interval

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

- Let's now consider only two consecutive (length/time) events
- the distance between them is a **continuous** random variable.

We can ask for the probability that the first event is at distance X.

8.14 Exponential 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).

8.15 Exponential density

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

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

8.16 Exponential density

An exponential random variable X has a probability density

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

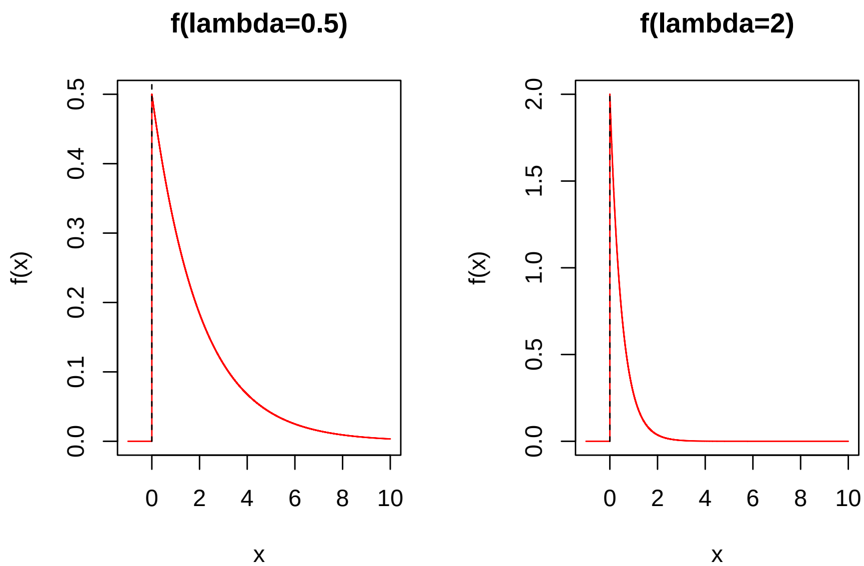
Properties:

- Mean: $E(X) = \frac{1}{\lambda}$
- Variance: $V(Y) = \frac{1}{\lambda^2}$

Where λ is its single parameter, known as a **decay rate**.

Note: The exponential model is a general model. It can describe the time/length until the first count in a Poisson process of the size of a whole made by a drill.

8.17 Exponential density



8.18 Exponential Distribution

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

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

- ¿What is the probability of observing an interval **larger** than size a until the first event?

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

8.19 Exponential Distribution

With the exponential density function:

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

We can answer questions like:

- 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?

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

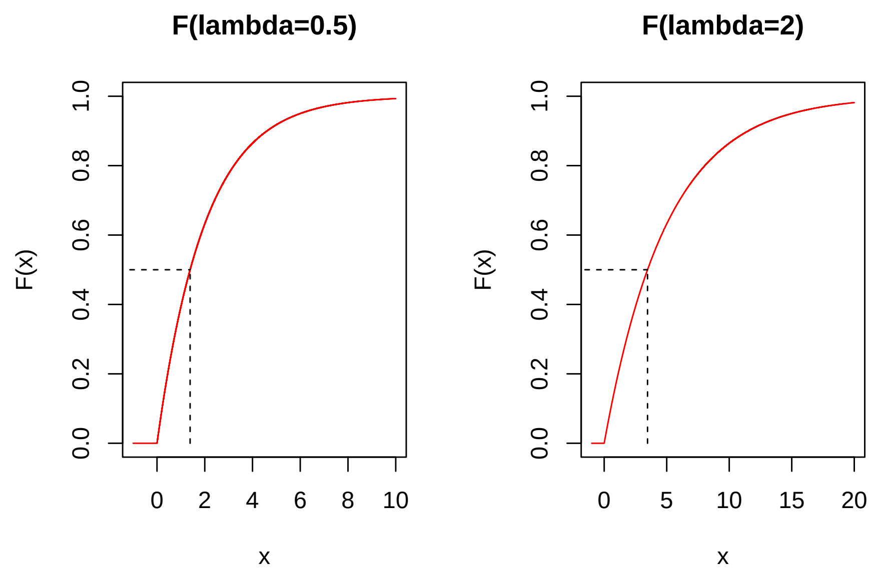
in R 1-pexp(1,2)

- 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)

8.20 Exponential Distribution



The median x_m is such that $F(x_m) = 0.5$. That is $x_m = \frac{\log(2)}{\lambda}$

Chapter 9

Normal Distribution

9.1 Objective

Continuous probability model:

- Normal distribution

9.2 Continuous probability models

Continuous probability models are probability density functions $f(x)$ of a continuous random variables that we **believe** describe real random experiments.

Definition:

Positive:

- $f(x) \geq 0$

Allows us to compute probabilities using the area under the curve:

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

The probability of any value is 1:

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

9.3 Normal density

In 1801 Gauss analyzed the orbit of Ceres (large asteroid between Mars and Jupiter).

- People suspected it was a new planet.
- The measurements had errors.
- He was interested in finding how the observations were distributed so he could find the most probable orbit.
- He wanted to predict where astronomers should point their telescopes to find it a few months after it had passed behind the Sun.

9.4 Normal density

Errors due to measurement.

9.5 Normal density

He assumed that

- small errors were more likely than large errors
- error at a distance $-\epsilon$ or ϵ from the most likely measurement were equally likely
- the most **likely** altitude of Ceres at a given time in the sky was the **average** of multiple altitude measurements at that latitude.

9.6 Normal density

That was enough to show that the random deviations y **from the orbit** distributed like

$$f(y) = \frac{h}{\sqrt{\pi}} e^{-h^2 y^2}$$

*The evolution of the Normal distribution, Saul Stahl, Mathematics Magazine, 2006.

9.7 Normal density

Let's write the distribution of errors

$$f(y) = \frac{h}{\sqrt{\pi}} e^{-h^2 y^2}$$

for the errors of measurements from the horizon X then $y = x - x_0$

$$f(x) = \frac{h}{\sqrt{\pi}} e^{-h^2 (x-x_0)^2}$$

- The **mean** of this probability density is:

$E(X) = \mu = x_0$, that represents the **true** position of Ceres from the horizon (property of the physical system).

- The **variance** is:

$V(X) = \sigma^2 = \frac{1}{2h^2}$, that represents the dispersion of the error in the observations (property of the measurement system).

9.8 Definition

A random variable X defined in the real numbers has a **Normal** density if it takes the form

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

with mean and variance:

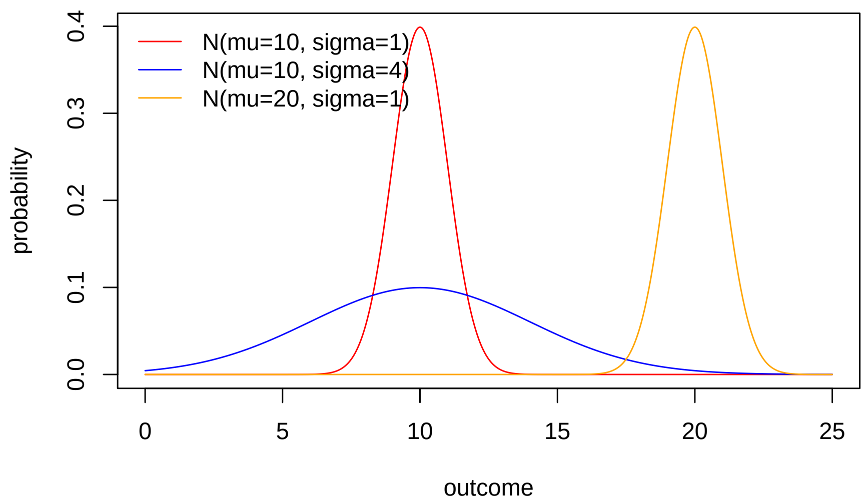
- $E(X) = \mu$
- $V(X) = \sigma^2$

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

When X follows a Normal density, i.e. distributes normally, we write

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

9.9 Normal probability density (Gaussian)



9.10 Normal 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.

9.11 Normal distribution

When

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

We can ask questions like:

- 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 $8cm$?

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

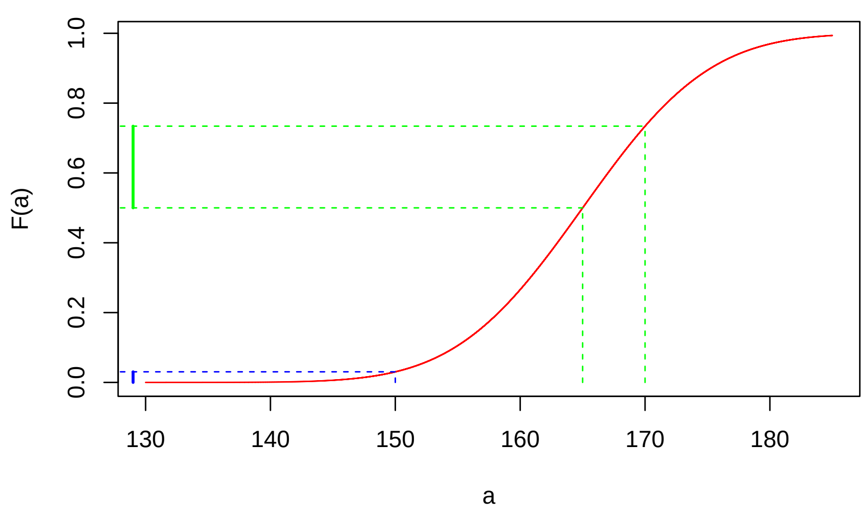
in R pnorm(150, 165, 8)

- 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)

9.12 Normal distribution



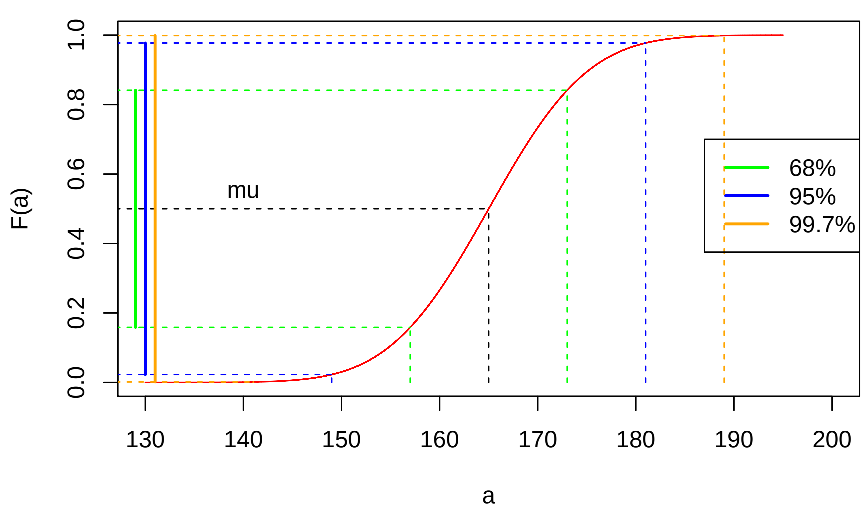
9.13 Normal distribution

- the mean μ is also the median as it splits the measurements in two
- x values that fall farther than 2σ are considered **rare** 5%
- x values that fall farther than 3σ are considered **extremely rare** 0.2%

9.14 Normal distribution

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

- $P(165 - 8 \leq X \leq 165 - 8) = P(157 \leq X \leq 173) = 0.68$
- $P(165 - 2 \times 8 \leq X \leq 165 - 2 \times 8) = P(149 \leq X \leq 181) = 0.95$
- $P(165 - 3 \times 8 \leq X \leq 165 - 3 \times 8) = P(141 \leq X \leq 189) = 0.997$



9.15 Standard normal density

Let's change variables to a **standardized variable**

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

in the density

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

replacing $x = \sigma z + \mu$ and $dx = \sigma dz$ in the probability expression we have

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

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

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

we obtain the **standardized** form of the normal density.

9.16 Standard normal density

Definition

A random variable Z defined in the real numbers has a **standard** density if it takes the form

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz, z \in \mathbb{R}$$

with mean and variance

- $E(X) = 0$
- $V(X) = 1.$

9.17 Standard normal density

The standard density:

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz, z \in \mathbb{R}$$

- is the normal density $N(\mu = 0, \sigma^2 = 1)$
- any normally distributed variable X can be transformed to a variable Z

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

that follows a standard distribution:

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

9.18 Normal distribution

All normal densities can be obtained from the standard density with the values of μ and σ

9.19 Standard distribution

The probability distribution of the standard density:

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

is the **error** function defined by

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

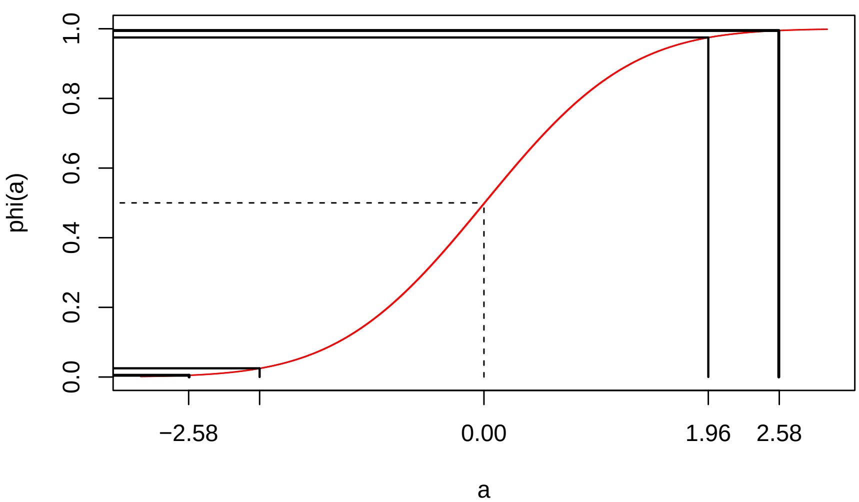
You can find it in most computer programs

9.20 Standard normal density

9.21 Standard normal density

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

- $P(-0.67 \leq X \leq 0.67) = 0.50$
- $P(-1.96 \leq X \leq 1.96) = 0.95$
- $P(-2.58 \leq X \leq 2.58) = 0.99$



9.22 Normal and standard distributions

For any normally distributed variable X, such that

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

its distribution $F(a) = P(X \leq a)$ can be computed from

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

9.23 Normal distribution

For computing $P(a \leq X \leq b)$, we use the property of the probability distributions

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

Let's standardize
$= P(\frac{X-\mu}{\sigma} \leq \frac{a-\mu}{\sigma}) - P(\frac{X-\mu}{\sigma} \leq \frac{b-\mu}{\sigma})$
$= P(Z \leq \frac{b-\mu}{\sigma}) - P(Z \leq \frac{a-\mu}{\sigma})$
$= \phi(\frac{b-\mu}{\sigma}) - \phi(\frac{a-\mu}{\sigma})$
Then

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

The probabilities of **any normal variable** can be obtained from the **standard distribution**, after standardization (subtract the mean and divide by the standard deviation).

9.24 Summary of probability models

Model	X	range of x	f(x)	E(X)	V(X)	R
Uniform	integer or real number	$[a, b]$	$\frac{1}{n}$	$\frac{b+a}{2}$	$\frac{(b-a+1)^2-1}{12}$	rep(1/n, n), dunif(x, a, b)
Bernoulli	event A	0,1	$p^x(1-p)^{1-x}$	p	$p(1-p)$	c(1-p,p)

Model	X	range of x	f(x)	E(X)	V(X)	R
Binomial	# of A events in n repetitions of Bernoulli trials	0,1,...	$\binom{n}{x}p^x(1-p)^{n-x}$	np	$np(1-p)$	dbimon(x,n,p)
Negative Binomial for events	# of B events in Bernoulli repetitions before r As are observed	0,1,..	$\binom{x+r-1}{x}p^r(1-p)^x$	$\frac{r(1-p)}{p}$	$\frac{r(1-p)}{p^2}$	dnbinom(x,r,p)
Poisson	# of events A in an interval	0,1, ..	$\frac{e^{-\lambda}\lambda^x}{x!}$	λ	λ	dpoiss(x, lambda)
Exponential	Interval between two events A	$[0,\infty)$	$\lambda e^{-\lambda x}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$	dexp(x, lambda)
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	dnomr(x, mu, sigma)

Chapter 10

Sampling Distributions

10.1 Objective

Distributions for

- Sample mean
- sample sum
- Sample variance

10.2 Normal distribution

When we have a normal random variable

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

How do we estimate μ and σ^2?

- we need to take a **random sample**
- we need to **estimate** each parameter

10.3 Example

Imagine a client asking your metallurgical company to sell them 8 cables that can carry up to 96 Tons; that is 12 Tons each.

- You have in **stock** a set of cables that could do the job.

Can you use the cables in stock or do you need to produce new ones?

10.4 Example

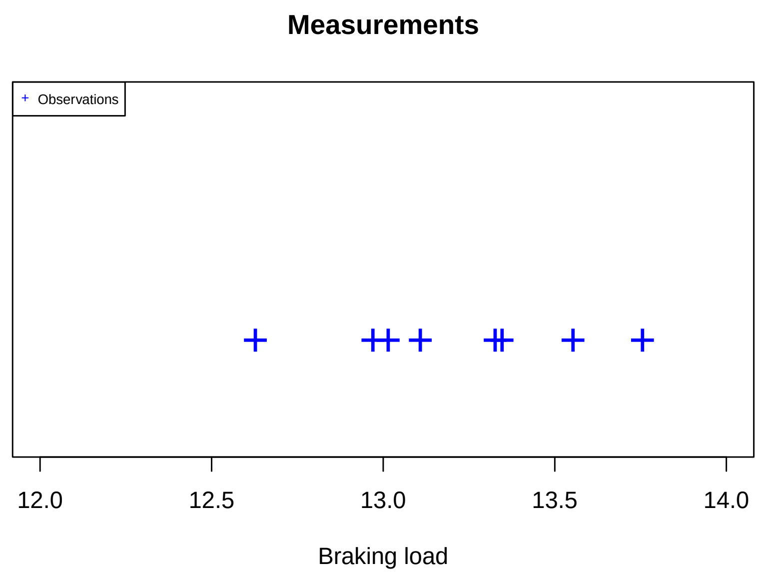
you take a sample of 8 random experiments, each of which consists of loading a cable until it breaks and recording the breaking load.

These are the results: The observation of a **sample** of size 8

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

- None of them broke at 12 Tons.
- There was one that broke at 12.62747 Tons.

Do you take the risk and sell a random sample of 8 cables from your stock?



10.5 Random sample

A **random sample** of size n is the **repetition** of a random experiment n **independent** times.

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

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

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

- **One 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 8 cables was

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

10.6 Example

We would like to compute $P(X < 12)$.

We are going to **assume** that the braking point is **normally distributed**.

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

- For computing $P(X < 12)$ we need the parameters μ and σ^2.
- How do we estimate the parameters from the observed sample?

10.7 Average or sample mean

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 8-size sample took the value

$$\bar{x}_{stock} = 13.21$$

10.8 Average as estimator

This number can be used to **estimate** the unkown parameter μ because:

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

(since each random experiment in the sample is independent)

as

- $n \rightarrow \infty$, $V(\bar{X}) \rightarrow 0$

then

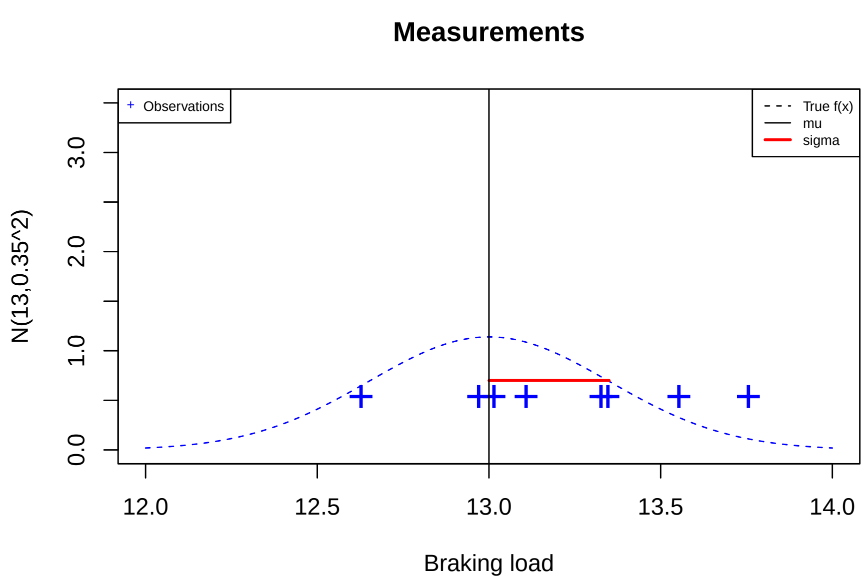
- $\bar{x}$ **concentrates closer and closer** to μ as n increases.

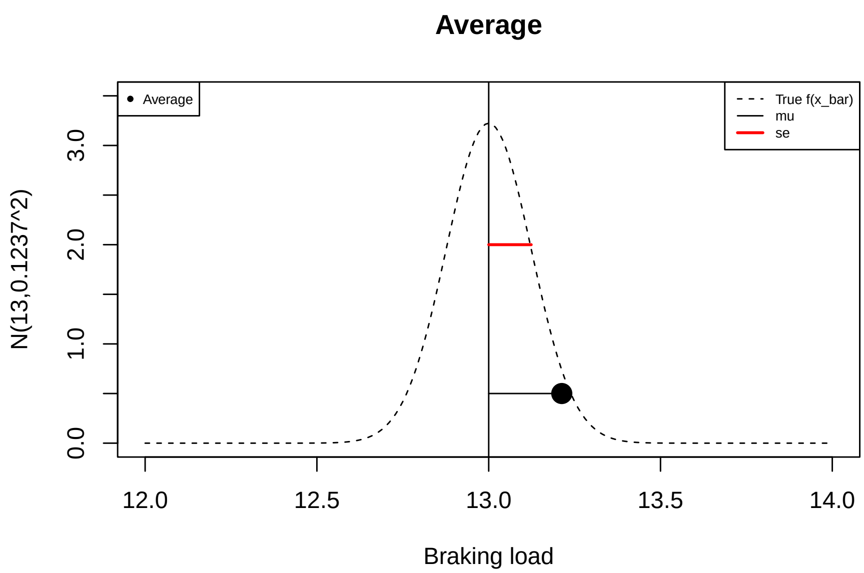
We can take one value of $\bar{x}$ as estimation for μ or

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

10.9 Outcome probability density and probability density of the average

If we **knew** that the **true** parameters were $\mu = 13$ and $\sigma = 0.35$ this is what we would see





10.10 Sample variance

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 about $\bar{X}$. In our 8-size sample S^2 took the value

$$s^2_{stock} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = 0.1275608$$

The expected value of S^2 is

- $E(S^2) = V(X) = \sigma^2$

and therefore S^2 is

- an estimator of $V(X)$

- it also concentrates around σ^2 because as $n \to \infty$, $V(\bar{S^2}) \to 0$

We can take one value of s^2 as estimation for σ^2 or

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

10.11 Sample variance

S^2 aims to estimate the dispersion of the outcomes about μ (the variance)

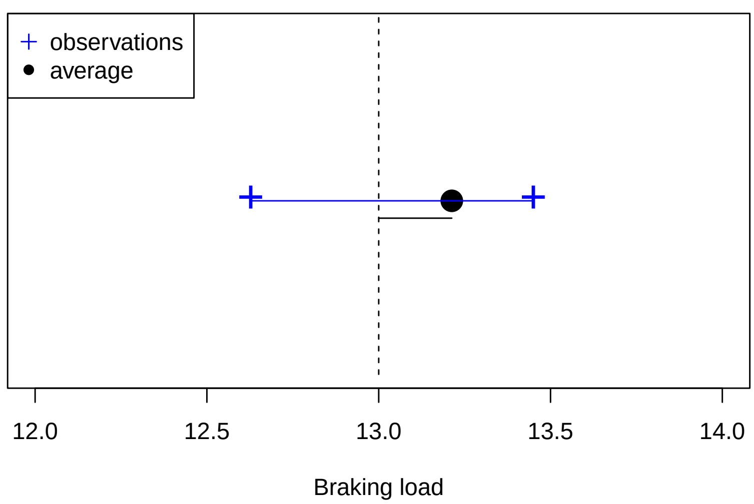
If we use $\bar{X}$ as an estimator of μ we need to correct for its dispersion (i.e. mean squared error of $\bar{X}$).

The correction is achieved by dividing by $n-1$ and not n in the definition of S^2

For:

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

$E(S_n^2) = \sigma^2 - \frac{\sigma^2}{n} \neq \sigma^2$ (we say that S_n^2 is **biased**)



10.12 Fitting a model

We **fit a model** when we

- **estimate** the parameters of the model

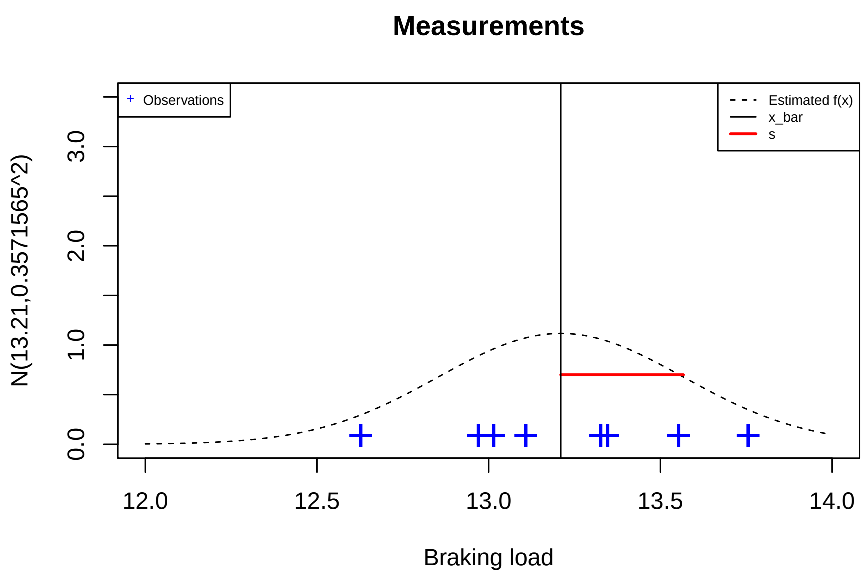
We also say we **train** a model (machine learning)

Assuming that

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

Since we do not know the parameters, we **plugin** the estimates $\bar{x}$ and s^2 as the values of μ and σ^2

$$X \to N(x; \mu = 13.21, \sigma^2 = 0.3571565^2)$$



10.13 Prediction

We **predict** the value of an **outcome** when we compute its **probability**

What is the probability that the cable breaks at 12 Tons?

If we assumed the random variable

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

We plug in the estimates $\bar{x}$ and s^2 into the probability distribution

$$P(X \leq 12) = F_{normal}(12; \mu = 13.21, \sigma^2 = 0.1275608)$$

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 will brake at 12 Tons.

10.14 Inference

We **infer** the value of an **estimator** when we compute its **probability**

Example:

- Imagine that our cables are certified to break with at a mean load of $\mu = 13$ Tons with variance $\sigma^2 = 0.35^2$.
- Can we claim that we actually produce stronger cables because we obtained $\bar{x} = 13.21$ in our 8-sample average?

We need to compute probabilities of $\bar{X}$.

When we make inferences, we usually ask the question:

How **confident** are we that the value of the estimator **is close** to the **true parameter**?

10.15 Sample mean distribution

When X follows a normal distribution $X \to N(\mu, \sigma^2)$

$\bar{X}$ is normal:

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

Then, if we **know** μ and σ we can compute the true **probabilities of** $\bar{X}$ using the normal distribution.

The mean and variance of $\bar{X}$ are

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

The errors in estimation are

- bias: $E(X) - E(\bar{X}) = 0$
- standard error: $se = \frac{\sigma}{\sqrt{n}}$

10.16 Inference on the average

Example:

If we **know** that for our cables trully distribute as

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

then

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

- $E(\bar{X}) = 13$
- $V(X) = \frac{0.35^2}{8} = 0.01530169$; $se = \frac{0.35}{\sqrt{8}} = 0.1237$

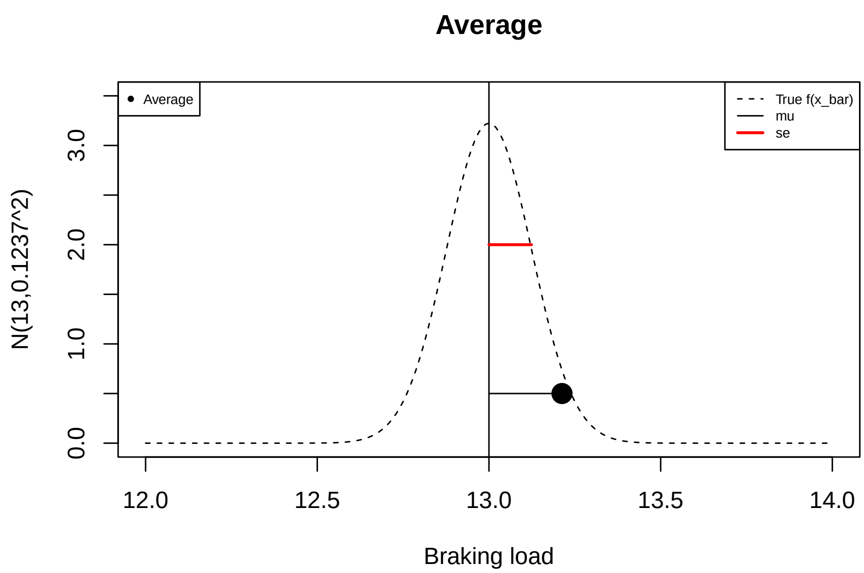
Our **observed error** in the estimation of the mean is the difference

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

We ask: Is this a **typical** error?

10.17 Outcome probability density and probability density of the average

If we **knew** that the **true** parameters were $\mu = 13$ and $\sigma = 0.35$ this is the error we would see



10.17.1 Probabilities of $\bar{X}$

If we **know** that the braking load of our cables **truly** distribute as

$$\bar{X} \to N(\mu = 13, \frac{\sigma^2}{n} = 0.1237^2)$$

What is the probability of observing an **error in estimation** of μ (distance between $\bar{X}$ and μ) smaller than 0.21?

We want to compute

$$P(-0.21 \leq \bar{X} - 13 \leq 0.21) = P(12.79 \leq \bar{X} \leq 13.21)$$

$= F_{normal}(13.21; \mu, se^2) - F_{normal}(12.79; \mu, se^2)$

In R we can compute it as:

pnorm(13.21, 13, 0.1237)-pnorm(12.79, 13, 0.1237)=0.9104.

91.0% of the errors are less than 0.21, therefore the **observed** error does not seem to be too typical (only 9% of the errors are higher). Maybe we have stronger cables than we thought.

10.17.2 Sample sum

If we are interested in using all the 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**:

$$Y = n\bar{X} = \sum_{i=1}^{n} X_i$$

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

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

With mean and variance:

- $E(Y) = n\mu$
- $V(Y) = n\sigma^2$

10.17.3 Inference on the sample sum

If we **know** that for our cables

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

then

$$Y \to N(n\mu = 104, n\sigma^2 = 8 \times 0.35^2)$$

- $E(Y) = 104$
- $V(Y) = 8 \times 0.35^2 = 0.98$

For our 8-sample, we observed

- $y_{stock} = 105.7014$

and, therefore, the **observed error** in the estimation of the mean of the **true** total braking load ($n\mu$) of 8 cables was

- $y_{stock} - n\mu = 1.7014$

Is this a **typical** error?

10.17.4 Probabilities of the sample sum: Propagation of error

What is the probability of observing a difference $Y - E(Y)$ smaller than 1.7014?

We want to compute the probability

$$P(-1.7014 \leq \bar{Y} - 104 \leq 1.7014) = P(102.2986 \leq Y \leq 105.7014)$$

$= F_{normal}(105.7014; n\mu, n\sigma^2) - F_{normal}(102.2986; n\mu, n\sigma^2)$

In R we can compute it as:

pnorm(105.7014, 104, sqrt(0.98)) - pnorm(102.2986, 104, sqrt(0.98))=0.914.

91.4% are smaller than 1.7014, a higher proportion than the proportion for individual cables because their individual errors accumulated.

10.18 Inference in the sample variance

Consider a quality control process that requires that the cables are produced close to the specified value μ.

If a sample of 8 cables is too dispersed ($S^2 > 0.3$), we stop production: the process is out of control.

What is the probability that the sample variance of a sample of 8 cables is greater than the required 0.3?

10.19 Probabilities of the sample variance

When X follows a normal distribution

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

The **statistic**:

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

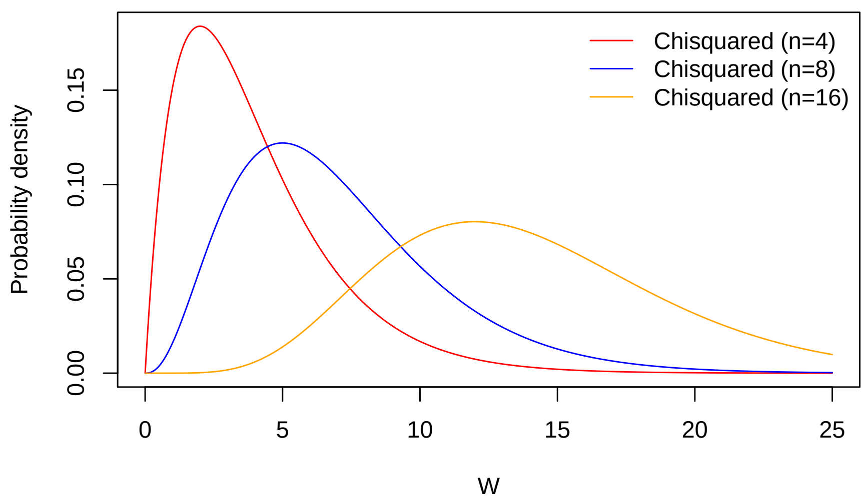
has a χ^2 (chi-squared) distribution with $df = n - 1$ degrees of freedom given by

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

where:

- $C_n = \frac{1}{2^{(n-1)/2}\sqrt{\pi(n-1)}}$ ensures $\int_{-\infty}^{\infty} f(t)dt = 1$
- $\Gamma(x)$ is Euler's factorial for real numbers
- If we **know** the true values of μ and σ we can compute probabilities of S^2 using the χ^2 distribution for W.

10.20 χ^2-statistic



10.21 χ^2-statistic

If we **know** that our cables trully distribute as

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

then we can compute

$$P(S^2 > 0.2) = P(\frac{(n-1)S^2}{\sigma^2} > \frac{(n-1)0.3}{\sigma^2})$$

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

$= 1 - P(W \leq \frac{(n-1)0.3}{\sigma^2}) = 1 - P(W \leq \frac{(8-1)0.3}{0.1225})$

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

In R 1-pchisq(17.14286, df=7)=0.016

There is only a probability of 1% of obtaining a value greater than $s^2 = 0.3$.

- $s^2 > 0.3$ seems to be a good criterion to stop production and revise the process.
- our observed value was $s^2_{stock} = 0.1275608$
- the sample is not too dispersed and we believe that the production is under control.

Chapter 11

Point Estimators

11.1 Objective

- Random sample
- Statistic
- Point estimators

11.2 Parameters

When we want to compute **probabilities** for the outcome of a random experiment **we need** a model and its parameter:

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

But we usually **don't know** θ

Let's look at a known example that will help us to introduce **terminology**

11.3 Bernoulli trial

Writing down the sex of a patient who goes into an emergency room of a hospital is a Bernoulli variable K with outcomes (0:female and 1:male), which has a probability mass function

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

The parameter p

- is the probability that one patient is male
- is usually **unknown**

What do we do?

11.4 Binomial distribution

We repeat the Bernoulli trial n times and count how many times we obtained A from the total number of repetitions:

$$f_A = \frac{n_A}{n}$$

- From $n = 100$ patients we count how many are men n_{man}.

f_A is the **observation** of the **average** over Bernoulli trials

$$\bar{K} = \frac{1}{n}\sum_{i}^{n} K_i$$

That is

$$f_{man} = \bar{k}$$

11.5 Binomial distribution

The average over n Bernoulli trials $\bar{K}$ is a random variable with mean and variance:

- $E(\bar{K}) = p$
- $V(\bar{K}) = \frac{p(1-p)}{n}$ (Remember: $V(aK) = a^2K$)

Therefore:

- as $n \to \infty$, $V(\bar{K}) \to 0$

and

- $\bar{K}$ **concentrates closer and closer** to p as n increases.

We can take one value of $\bar{k}$ as estimation for p or

$$\bar{k} = \hat{p}$$

Remember: $\bar{k} = f_{male}$ and $p = P(male)$, therefore $lim_{n\to\infty} f_{male} = P(male)$

11.6 Average

Example:

If we observe

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

after the repetition of n Bernoulli trials: Determining the sex of 10 patients (0:female, 1:male).

The **random variable** $\bar{K}$ takes the value $\bar{k} = 5/10 = 0.5$ and we use it

$$\hat{p} = 0.5$$

to **estimate** the unobserved probability (p) that one patient entering the emergency room is male (parameter of the Bernoulli trial).

11.7 Average

Situation 1:

If we wait for 10 other patients then

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

$\bar{K} = 7/10$ and

$$\hat{p} = 0.7$$

changes because $\bar{K}$ is random.

11.8 Average

Situation 2:

If we observe the event

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

with $N = 10000$ and $\bar{k} = 6675/10000$ then

$$\hat{p} = 0.6675$$

However, when we repeat the 10000-sampling

$$\hat{p} = 0.6698$$

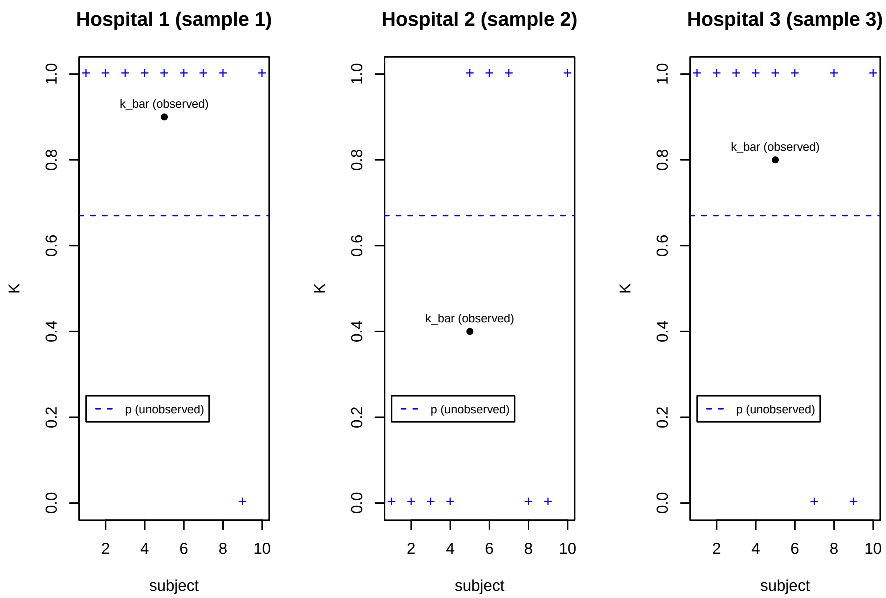
The two estimates get **closer and closer** because $V(\bar{K}) \to 0$

Solution: to estimate the probability that a man enters an emergency room, repeat the experiment many many times and take the average.

11.9 Average

Situation 1: small n

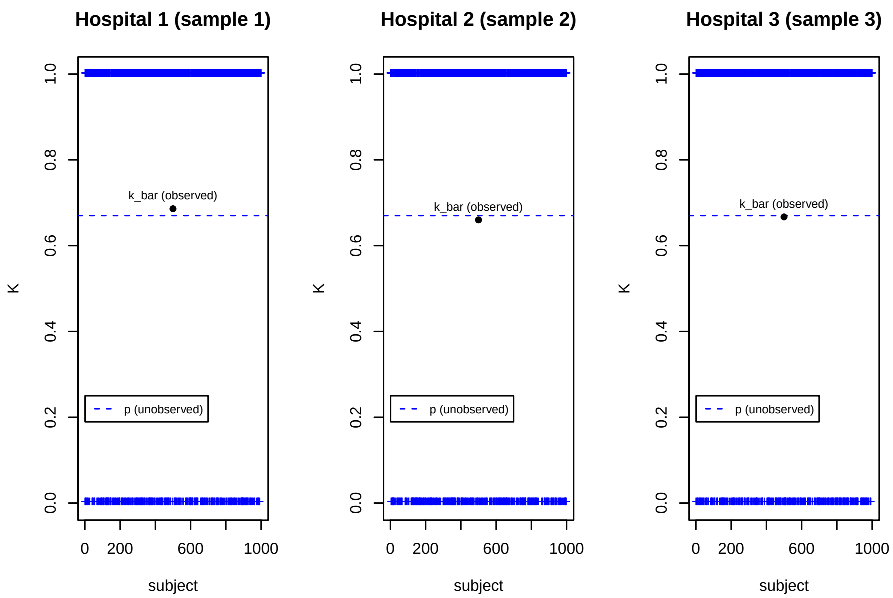
We record the sex of 10 patients going into an emergency room



11.10 Average

Situation 2: large n

We record the sex of 500 patients going into an emergency room



11.11 Random sample

A **random sample** of size n is the **repetition** of a random experiment n **independent** times.

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

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

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

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

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

11.12 Random sample

We repeat the random experiment n times to **learn from experience** and then we can

- Describe properties of the data and the underlying distribution model (Descriptive statistics)
- **Find** θ (Estimation)
- Make hypotheses on θ (Inference)

11.13 Statistic

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**
- The **probability distributions** of statistics are called **sampling distributions**

11.14 Statistics Examples 1

Statistics of **location** (center) of outcomes of random experiments

- **Average**:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

- **Median**:

$$Q_{0.5} = X_m$$

such that: $F_m = \sum_{i<m} \frac{n_i}{n} = 0.5$

- **Mode**:

$$\text{Mode} = X_m$$

such that: $\max_m\{\frac{n_i}{n}\}$

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

11.15 Statistics Examples 2

Statistics of **spread** of outcomes of random experiments

- **Sample variance**:

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

- **Inter quantile range**:

$$Q_{0.75} - Q_{0.25}$$

- **Range**:

$$\max\{X_i\} - \min\{X_i\}$$

11.16 Statistics Examples 3

statistics with important **distribution properties**:

- **Standard**:

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

- **t-statistics**:

$$T = \frac{\bar{X} - \mu}{S}$$

- χ^2**-statistics**:

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

11.17 Uses of Statistics

- **Description** of a sample's data
  - location: $\bar{X}$
  - Minimum: $\min\{X_i\}$
  - Maximum: $\max\{X_i\}$
- **Estimation** of a probability model's **parameters**
  - mean: $\bar{X}$ for μ
  - variance: S^2, for for σ^2
- **Inference** to say something about the parameters given the data
  - mean: Z, T
  - variance: χ^2

11.18 Estimation

We **assume** a probability model for the distribution of X,

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

where θ is a parameter

Main question:

- What is the value of θ, so we can compute the probability of an outcome?

Example: We observe n patients come in the emergency room, we take the relative frequency $\bar{k} = \frac{n_{men}}{n}$ and that is an estimation for the parameter p.

11.19 Point estimators

Point estimators are statistics that are used to **learn about the unknown parameters** of probability models:

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

Notation:

- A **parameter** of the distribution is a **number**

$$\theta$$

- A **point estimator** is a statistic (function of the random sample) and a **random variable**

$$\Theta$$

- An **estimate** is an **observed value** of the estimator

$$\hat{\theta}$$

11.20 Point estimators

For the Bernoulli trial

$$K \to Bernouilli(p)$$

Notation:

- The p is the **parameter** of the distribution
- A **point estimator** of p is the **random variable**

$$\bar{K} = \frac{1}{n}\sum_{i=1}^{n} K_i$$

- An **estimate** of p is an **observed value** of the estimator

$$\hat{p} = \bar{k}$$

11.21 Point estimators

11.22 Properties of estimators

As **random variables** estimators have their own probability functions:

$$\Theta \to f(\hat{\theta}; \beta)$$

and parameters β

They have their own mean and variance

- $E(\Theta)$
- $V(\Theta)$

11.23 Example:

For the Bernoulli trial if we take $\bar{K}$ as the estimator for p then

- $E(\bar{K}) = p$
- $V(\bar{K}) = \frac{p(1-p)}{n}$

Remember: The numbers we need to know to fully determine a probability function are call parameters, therefore p, n, are parameters for the probability function of $\bar{K}$

$$\bar{K} \to f(\bar{k}; n, p)$$

in particular, we know that

$$Y = n\bar{K} \to Binom(n, p)$$

11.24 Bias (Accuracy)

Some important properties of estimators are their **errors** when estimating:

Let's **assume that we know** the parameter θ that we want to estimate.

The **bias** of Θ is

$$bias = E(\Theta) - \theta$$

- how much the expectation of Θ differs from the parameter θ.
- *bias* is a property of Θ.
- Θ is **unbiased** if

$$E(\Theta) = \theta$$

Example:

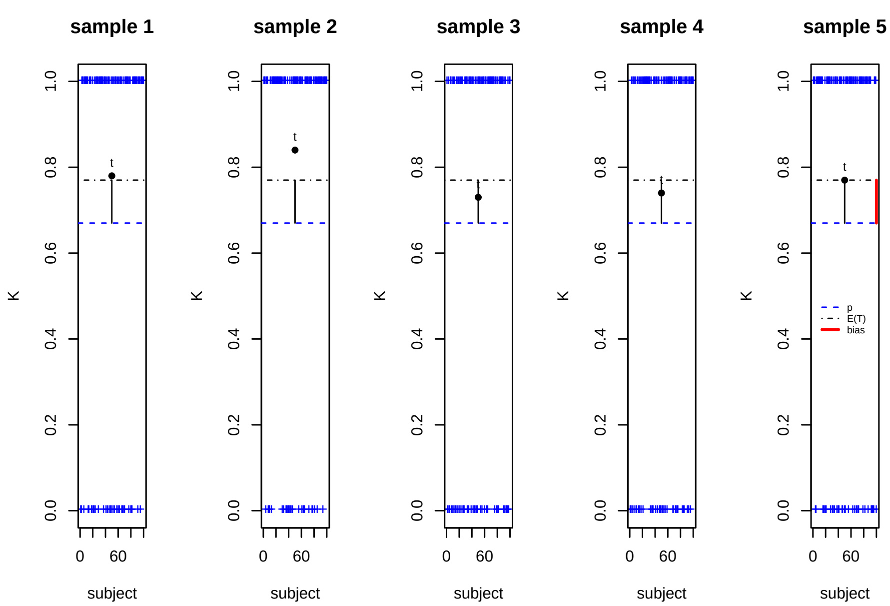
For the Bernoulli trial if we take $\bar{K}$ as the estimator for p then

- $E(\bar{K}) = p$ and therefore it is **unbiased**

11.25 A biased (inaccurate) estimator

Imagine another estimator for p that we call T

- p: bullseye (parameter)
- t: dart (estimate from a estatistic T)
- *bias*: error in accuracy $(E(T) - p)$



11.26 Standard Error (Precision)

The **standard error** se of an estimator Θ is its standard deviation

$$se = \sqrt{V(\Theta)}$$

- What Θ varies from its mean $E(\theta)$

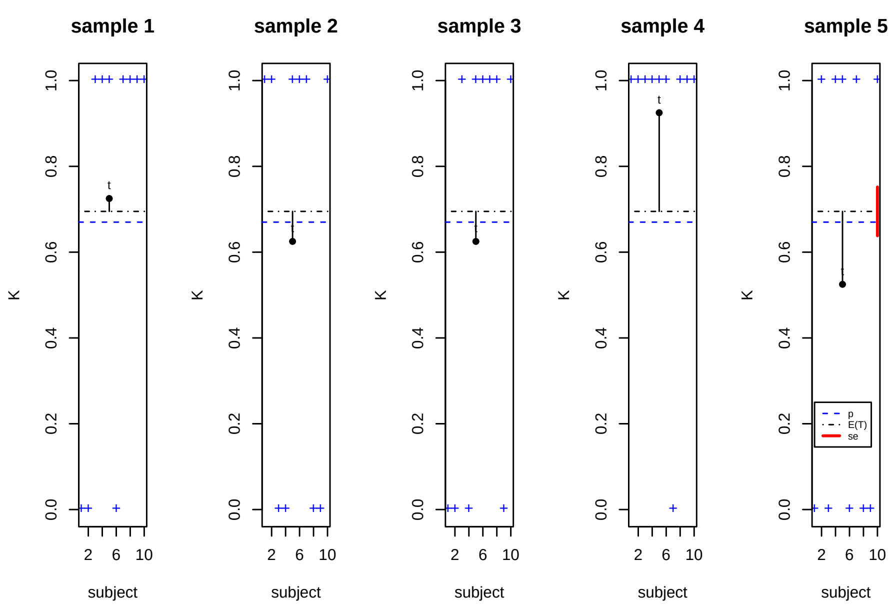
Example:

For the Bernoulli trial if we take $\bar{K}$ as the estimator for p then its standard error is

$$se = \sqrt{V(\bar{K})} = \sqrt{\frac{p(1-p)}{n}}$$

11.27 An unprecise estimator of p

- p: bullseye (parameter)
- t: dart (estimate from a estatistic T)
- se: error in precision ($\sqrt{V(T)}$)



11.28 Mean squared error

The *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

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

- what Θ varies from the parameter θ

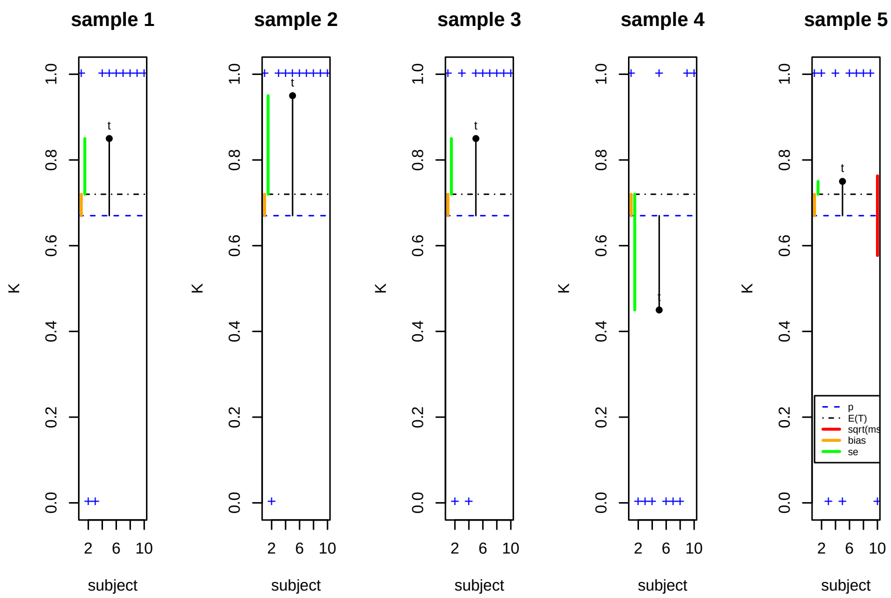
Example:

For the Bernoulli trial if we take $\bar{K}$ as an **unbiased** estimator for p then its *mse* is

$$mse = V(\bar{K}) = \frac{p(1-p)}{n}$$

11.29 An unprecise and inaccurate estimator of p

- p: bullseye (parameter)
- t: dart (estimate)
- mse: total error
- $bias$: error in accuracy
- se: error in precision


sample 1
sample 2
sample 3
sample 4
sample 5
K
subject
t
p
E(T)
sqrt(ms
bias
se

Chapter 12

Central limit theorem

12.1 Objective

- Margin of errors
- Central limit theorem
- t-statistic

12.2 Margin of error

When deciding whether an **observed error** is large or not we usually compare it with a **predefined** tolerance.

- The **margin of error** at 5% level is the distance m such that distribution of $\bar{X}$ captures 95% of the estimations:

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

- or that 95% of the values of $\bar{X}$ are a distance m from μ

12.3 Margin of error

Let's continue with the braking load example.

for the 8-sample

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

the **observed error** is the difference

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

Is this value below the margin of error at 5%?

12.4 Z-statistic

If we **know** that for our cables truly distribute as

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

then,

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

and the 5% margin of error for the average in our 8-sample can be computed from the **standardized statistic**:

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

12.5 Z-statistic

to compute the margin of error m at 5% level we standardize (subtract μ and divide by $\sigma/\sqrt{n}$)

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

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

(compare it with the plot) we have

$$m = z_{0.025}\frac{\sigma}{\sqrt{n}} = 1.96 \times se = 1.96\frac{0.35}{\sqrt{8}} = 0.24$$

where $z_{0.025} = 1.96$ is the value Z that leaves 2.5% at each side of standard normal density (0.025-quantile)

Our observed error 0.21

- is less than the margin of error 0.24 at level 5%.
- and, therefore, it is expected within the 95% of errors.

If an observation of $\bar{x}$ distance more than ~ 2 times the *se* we say that the error is **unusually** large.

12.6 Z-statistic

Definition

For a normal random variable X

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

with **known** σ

The Z statistic:

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

is a standard random variable whose $1 - \alpha/2$-quantiles ($z_{1-\alpha/2}$) give a measure of the margin of error of $\bar{X}$ at $1 - \alpha$ level

$$m = z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}$$

A common situation:

- What happens when X is not normally distributed?

12.7 Central Limit Theorem

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 $n \to \infty$

Therefore:

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

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

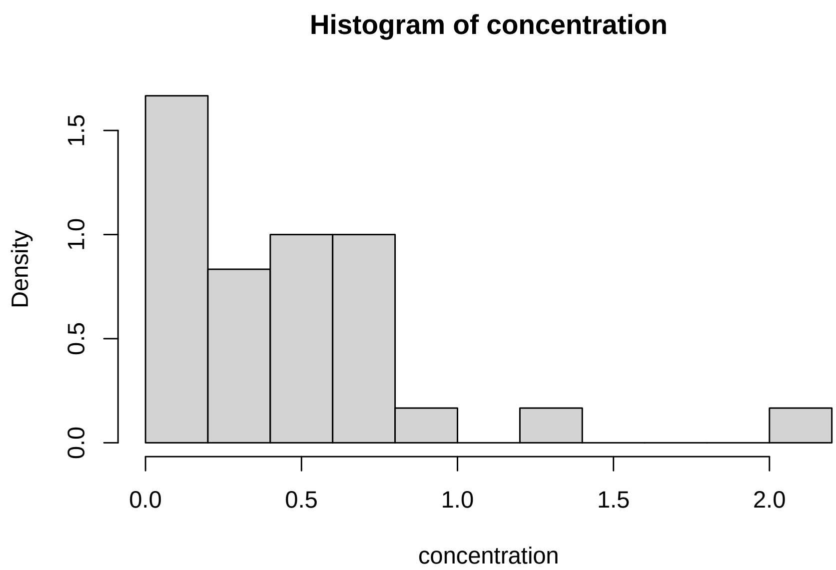
12.8 Central Limit Theorem

Example:

Consider an experiment where we measure the concentration in blood of a drug after 10-hour administration in 30 patients. We obtain the following results:

```
##  [1] 0.42172863 0.28830514 0.66452743 0.01578868 0.02810549 0.15825061
##  [7] 0.15711365 0.07263340 1.36311823 0.01457672 0.50241503 0.24010736
## [13] 0.14050681 0.18855892 0.09414202 0.42489306 0.78160177 0.23938021
## [19] 0.29546742 2.02050586 0.42157487 0.48293561 0.74263790 0.67402224
## [25] 0.58426449 0.80292617 0.74837143 0.78532627 0.01588387 0.29892485
```

- the average is $\bar{x} = 0.56$
- the histogram of the results is:



12.9 Central Limit Theorem

If we **know** that levels follow an exponential distribution

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

The mean and variance are:

- $E(X) = \frac{1}{\lambda} = 0.5$
- $V(X) = \frac{1}{\lambda^2} = 0.25$

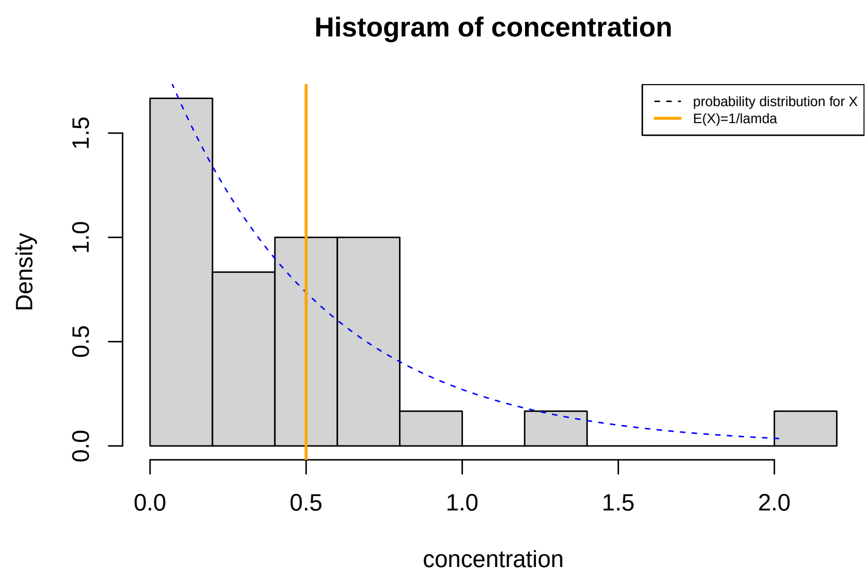
Therefore the mean and variance of $\bar{X}$ are:

- $E(\bar{X}) = \frac{1}{\lambda} = 0.5$
- $V(\bar{X}) = \frac{V(X)}{n} = \frac{1}{n\lambda^2} = 0.25/30$

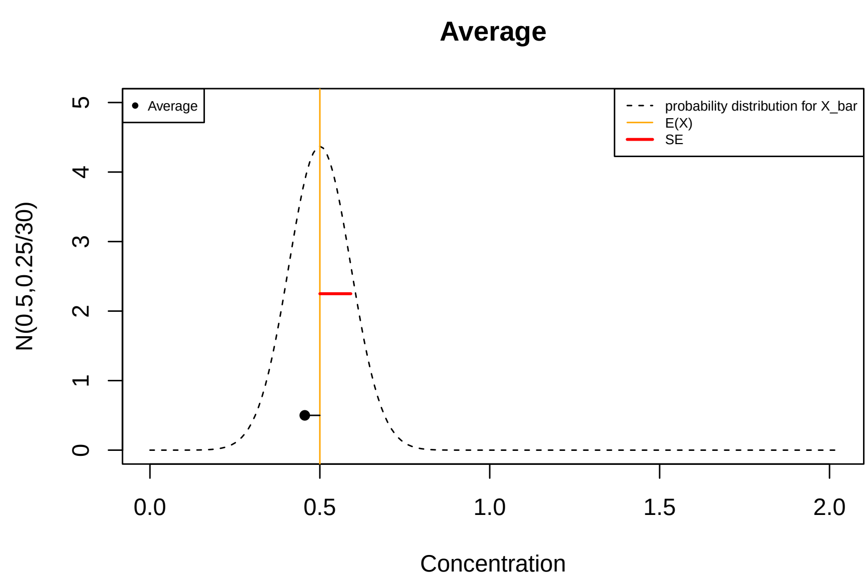
As $n \geq 30$

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

is a standard normal variable and: $\bar{X} \sim_{aprox} N(\lambda, \frac{1}{n\lambda^2})$



Histogram of concentration
probability distribution for X
E(X)=1/lamda
Density
0.0
0.5
1.0
1.5
2.0
concentration



Average
Average
probability distribution for X_bar
E(X)
SE
N(0.5,0.25/30)
0
1
2
3
4
5
0.0
0.5
1.0
1.5
2.0
Concentration

12.10 Margin of error with CLT

Since

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

The margin of error at 5% level

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

can be computed again with the standard distribution

$$m = z_{0.025}\sqrt{\frac{V(X)}{n}} = 1.96\sqrt{\frac{0.25}{30}} = 0.1789227$$

We **observed** $\bar{x} = 0.5638725$ therefore the **observed error** in estimation is

$$\bar{x} - E(X) = 0.5638725 - 0.5 = 0.063$$

which is within the margin of error.

The error that we observed is common and within the 95% of errors.

12.11 Sample sum and CLT

For any random variable X with **unknown** (any type of) distribution

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

the standardized statistic

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

approximates to a standard distribution

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

when $n \to \infty$

Therefore:

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

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

12.12 Unknown σ but large n

For any random variable X with **unknown** (any type of) distribution

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

with **unknown** variance $V(X)$, we can estimate the standard error ($se = \sqrt{V(X)/n}$) by the sample standard deviation

$$\hat{se} = \frac{s}{\sqrt{n}}$$

and write the standardized statistic

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

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

to recover the CLT when $n \to \infty$ (a good aproximation is when $n > 30$)

12.13 T-statistic

When

- σ is **unknown**

and

- n is small (cannot apply CLT)

However, if X is normal

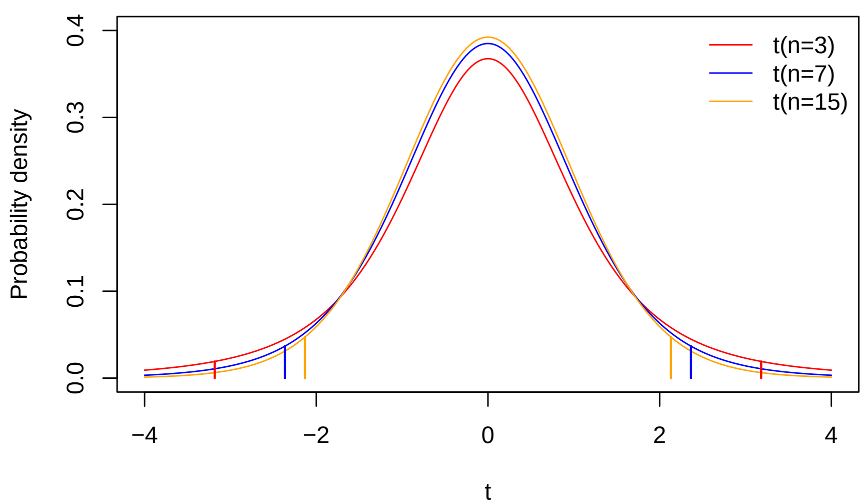
$$X \to N(\mu, \sigma^2)$$

then the standardized statistic

$$T = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}}$$

Follows a t-distribution with $n - 1$ degrees of freedom, and we can compute probabilities on $\bar{X}$.

12.14 T-statistic



12.15 T-statistic

To compute the margin of error m at 5% level when n is small, σ unknown but X normal

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

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

We use the t-distribution

$$m = t_{0.025,n-1}\frac{s}{\sqrt{n}}$$

where $t_{0.025,n-1}$ is the value T that leaves 2.5% at each side of t-distribution with $n-1$ degrees of freedom (0.025-quantile)

12.16 Example 1

Going back to the braking load example, we computed the margin of error with **known** $\sigma^2 = 0.35^2$.

$$m = z_{0.025}\frac{\sigma}{\sqrt{n}} = 1.96 \times se = 1.96\frac{0.35}{\sqrt{8}} = 0.24$$

- In most applications we **do not know** the parameters

If we only assumed that the braking load is a normal random variable

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

with **unknown** μ and σ^2 then from the data

- $s_{stock} = \sqrt{0.1275608}$

and the margin of error is

$$m = t_{0.025,n-1}\frac{s}{\sqrt{n}} = 2.36 \times \hat{se} = 2.36\frac{0.3571565}{\sqrt{8}} = 0.29$$

where $t_{0.025,n-1} = 2.36$

in R is qt(1-0.025, 7)

It increased from the value we obtained with **known** σ

12.17 Example 2

We can also ask for the probability of observing an error in the estimation of μ (distance between $\bar{X}$ and μ) smaller than the observed value 0.21?

We thus want to compute

$$P(-0.21 \leq \bar{X} - \mu \leq 0.21) = P(\frac{-0.21}{s/\sqrt{n}} \leq T \leq \frac{0.21}{s/\sqrt{n}})$$

$= P(\frac{-0.21}{0.3571565/\sqrt{8}} \leq T \leq \frac{0.21}{0.3571565/\sqrt{8}})$

$= F_{t,n-1}(0.21) - F_{t,n-1}(-0.21)$

In R we can compute it as:

pt(1.663052, 7)-pt(-1.663052, 7)=0.859.

85.9% of the errors are less than 0.21, therefore the **observed** error seems more typical than the 91% that we obtain with $\sigma^2 = 0.35^2$.

Note that in the calculations we have substituted $\sigma = 0.35$ by a higher estimate $s = 0.3571565$ **obtained from data**.

Chapter 13

Maximum likelihood

13.1 Objective

- Maximum likelihood
- Method of Moments

13.2 Statistic

Definition

Given a random sample $X_1, ...X_n$ a **statistic** is any real value function of the random variables that define the random sample: $f(X_1, ...X_n)$

- $\bar{X} = \frac{1}{N}\sum_{j=1..N} X_j$
- $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$
- $\max X_1, X_n$

are statistics

13.3 Estimator

Definition

An **estimator** is a statistic Θ whose values $\hat{\theta}$ are measures of a parameter θ of the population distribution on which the sample is defined: $E(\Theta) \sim \theta$

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

Then

- θ is a **parameter** of the population distribution $f(x;\theta)$
- Θ is an **estimator** of θ: A random variable
- $\hat{\theta}$ is the **estimate** of θ: A realized value of Θ

13.4 Estimator

13.5 Examples 1: Average (Sample mean)

When

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

For the mean:

- μ is a **parameter** of the **population** distribution: distribution of X, $N(\mu, \sigma^2)$
- $\bar{X}$ is an **estimator** of μ
- $\bar{x} = \hat{\mu} = 13.21\,Tons$ is the **estimate** of μ

13.6 Examples 2: Sample Variance

When

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

For the variance:

- σ^2 is a **parameter** of the population distribution $N(\mu, \sigma^2)$
- S^2 is an **estimator** of σ^2
- $s^2 = \widehat{\sigma^2} = 0.127\,Tons^2$ is the **estimate** of σ^2

13.7 Bias

An estimator is unbiased if $E(\Theta) = \theta$

- $\bar{X}$ is an **unbiased** estimator of μ because $E(\bar{X}) = \mu$
- S^2 is an **unbiased** estimator of σ^2 because $E(S^2) = \sigma^2$

13.8 Consistency

An estimator is consistent if $V(\Theta) \to 0$ when $n \to \infty$

- $\bar{X}$ is **consistent** because $V(\bar{X}) = \frac{\sigma^2}{n} \to 0$ when $n \to \infty$.
- S^2 is also **consistent** (we will not show).

13.9 Maximum likelihood

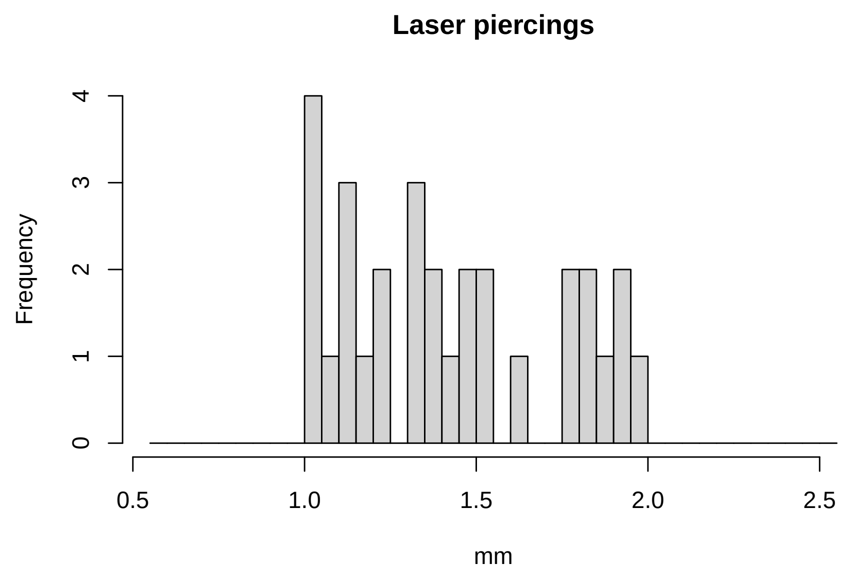
How can we **estimate** the parameter of **any** parametric model?

- 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

```
##  [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
```

13.10 Example

with histogram



13.11 Probability density

We consider 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.

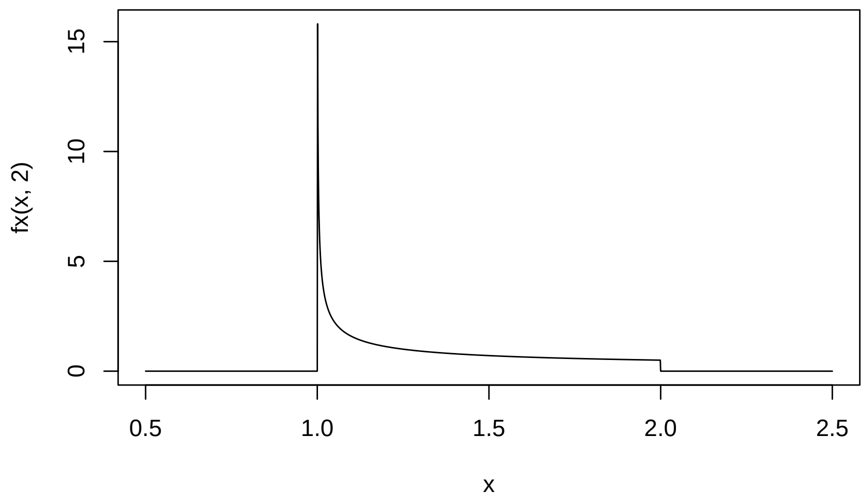
A suitable probability density distribution is

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

Where α is a parameter. This is a probability density.

13.12 Probability density

In particular, for $\alpha = 2$ we can plot it

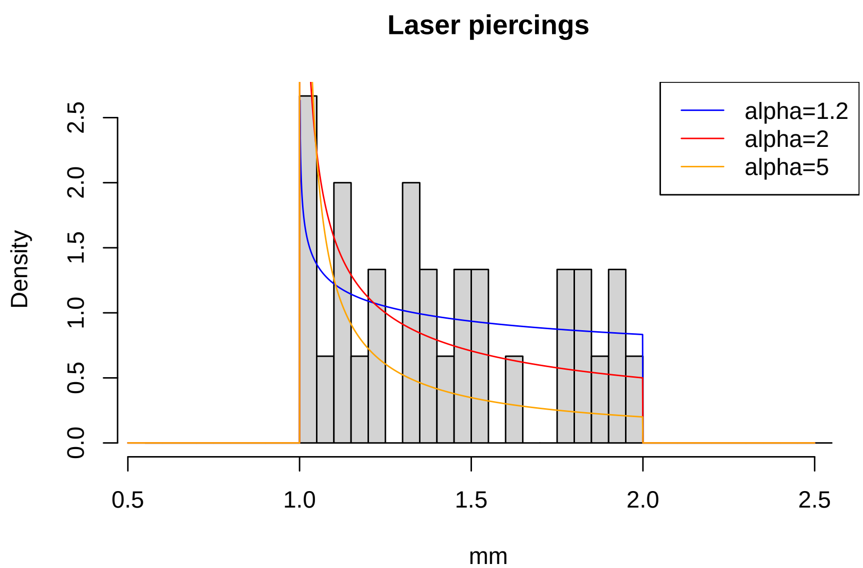


If we were to perform a n-sample: $X_1, ...X_n$, how should we combine the data for obtaining the best value of α?

- The **maximum likelihood** method **gives us the estimator** for α

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

13.13 Example: Maximum likelihood



13.14 Maximum likelihood

The objective is to find the value of the parameter that we **believe** can **best** represent the data.

We search for the parameter that makes the **observation** of the sample the most **probable**.

Note:

- **Probabilities** are assigned **to observations**.
- Probabilities are **not** assigned to **parameters** (we assign beliefs, and likelihoods).

Parameters are not supposed to change, they are properties of the system.

13.15 Method step 1

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:

$$P(M = x_1, ...x_n) = P(X_1 = x_1)P(X_2 = x_2)...P(X_n = x_n)$$
$$= f(x_1; \alpha)f(x_2; \alpha)...f(x_n; \alpha)$$

- Once the data is observed they are **fixed**.
- The unknown is α
- This probability as a function of the α we call it the **likelihood function**

$$L(\alpha) = \Pi_{i=1..n} f(x_i; \alpha)$$

then in our case

$$L(\alpha; x_1, ..x_n) = \frac{1}{\alpha^n}\Pi_{i=1..n}(x_i - 1)^{\frac{1-\alpha}{\alpha}} = \frac{1}{\alpha^n}\{(x_1 - 1)(x_2 - 1)...(x_n - 1)\}^{\frac{1-\alpha}{\alpha}}$$

13.16 Method step 2

We ask: what is the value of α that makes the observations the most probable? 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 $L(\alpha)$

2. Take the logarithm, obtain the **Log-likelihood**

$$\ln L(\alpha; x_1, ..x_n) = -n \ln(\alpha) + \frac{1-\alpha}{\alpha}\Sigma_{i=1...n} \ln(x_i - 1)$$

13.17 Method step 3

3. Maximize the log-likelihood with respect to the parameter

Therefore,

- we differentiate with respect to α

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

- The maximum is where the derivative is 0. This maximum is the value of our estimator $\hat{\alpha}_{ml}$.

$$\hat{\alpha}_{ml} = -\frac{1}{n}\Sigma_{i=1\ldots n} \ln(x_i - 1)$$

13.18 Method step 3

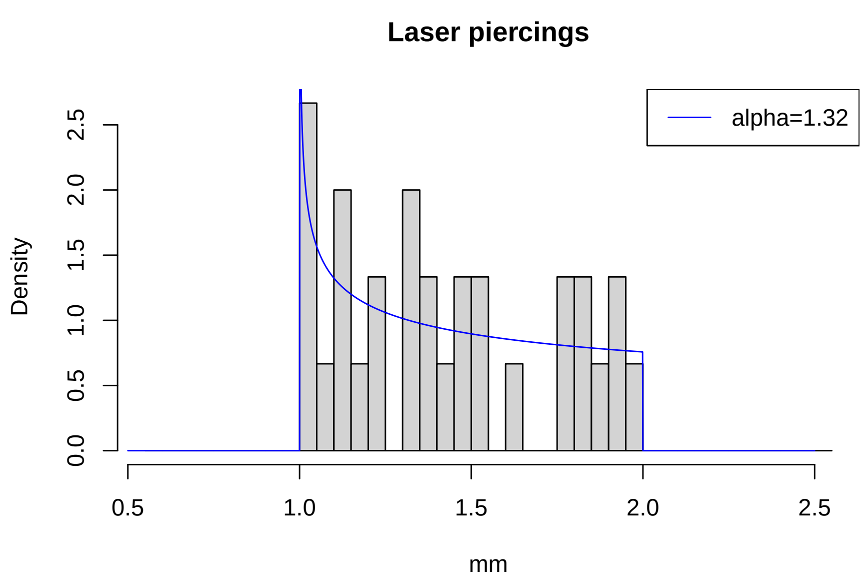
$\hat{\alpha}_{ml} = -\frac{1}{n}\Sigma_{i=1\ldots n} \ln(x_i - 1)$

is the **statistic** that estimates the parameter.

In our example we thus compute:

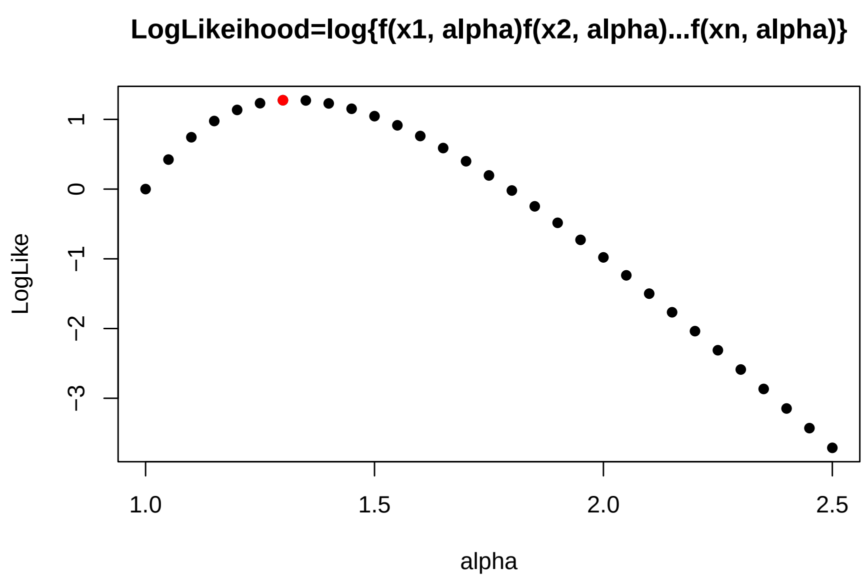
$\hat{\alpha}_{ml} = -\frac{1}{n}\{\ln(1.11 - 1) + \ln(1.64 - 1) + \ldots \ln(1.04 - 1)\} = 1.320$

13.19 Estimation



13.20 Estimation

Let's look at the log-likelihood for our 30 laser cuts. Remember, data is fixed by our experiment and α varies



Note: If we take another sample this function changes and so does its maximum.

13.21 Maximum likelihood: History

13.22 Maximum likelihood: History

- Ceres was thought to be a planet
- It disappeared behind the Sun

- Predictions were needed to know where in the sky to look for it after it passed behind the sun
- The trajectory (parallel to the planets) would determine if it was likely a planet
- With several observations with errors, what would be the best representative of the true position of Ceres at a given time?

13.23 Maximum likelihood: History

What is the statistic that best represents the true position of Ceres?

13.24 Maximum likelihood: History

Gauss proposed that at a **given** time

- the **true** position of Ceres was the mean μ
- the probabilities around the mean were symmetrical.

13.25 Maximum likelihood: History

Gauss discovered that if the average ($\bar{x}$) is the **most likely** value for the real position of Ceres (μ), then the probability density for the errors is

$$\frac{h}{\sqrt{\pi}}e^{-h^2(x-\mu)^2}$$

which we call the Gaussian and Pearson (1920) baptized it as the normal curve.

Note: We assume that the **true** position of Ceres exists μ.

Can we say the same about the height of men? is there a **true** mean height? (Galton)

13.26 Normal distribution

Imagine that we take a 8-sample for the breaking load of cables

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

and

- **assume** that
$$X \to N(\mu, \sigma^2)$$
.
- What are the estimators of μ and σ^2 that maximize the probability of the observed data?

13.27 Normal distribution

1. The likelihood function, the probability of having observed $(x_1, ...x_n)$ is

$L(\mu, \sigma^2) = \Pi_{i=1..n} N(x_i; \mu, \sigma)$

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

2. We can take the log of L, and compute the **log-likelihood**

$$\ln L(\mu, \sigma^2) = -n\ln(\sigma\sqrt{2\pi}) - \frac{1}{2\sigma^2}\Sigma_i(x_i - \mu)^2$$

13.28 Normal distribution

The estimates of μ, σ^2 are where the likelihood is maximum, and give the highest probability for the data.

3. we differentiate with respect to μ and σ^2

- $\frac{d\ln L(\mu,\sigma^2)}{d\mu} = \frac{1}{\sigma^2}\sum_i (x_i - \mu)$
- $\frac{d\ln L(\mu,\sigma^2)}{d\sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_i (x_i - \mu)^2$

13.29 Normal distribution

The derivatives are 0 at the maxima

- $\frac{1}{\hat{\sigma}^2}\sum_i(x_i - \hat{\mu}) = 0$
- $-\frac{n}{2\hat{\sigma}^2} + \frac{1}{2\hat{\sigma}^4}\sum_i(x_i - \hat{\mu})^2 = 0$

solving for the parameters we find

- $\hat{\mu}_{ml} = \frac{1}{n}\sum_i x_i = \bar{x}$ (the **average**)
- $\hat{\sigma^2}_{ml} = \frac{1}{n}\sum_i(x_i - \bar{x})^2$ (the **uncorrected** sample variance)

The maximum likelihood estimator of σ^2 is a **biased** estimator as:

$$E(\hat{\sigma}^2_{ml}) = \sigma^2 - \frac{\sigma^2}{n} \neq \sigma^2$$

13.30 Method of Moments

The method of maximum likelihood aims to produce the estimators of probability distributions from data.

- Is there another way to produce those estimators? would they be equal?

Let's look again at the maximum likelihood estimators for μ and σ^2 for a random variable that distributes normally:

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

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

13.31 Method of Moments

Let's re-write the estimators in terms of the values of X (outcomes), not of the observations

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

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

13.32 Method of Moments

The method os moments says that we can take the **observed** value of the average $\bar{X} = \frac{1}{n} \sum_i X_i$ as an estimator of $E(X) = \mu$

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

$\bar{X}$ is called the first **sample moment**

the estimator of the parameter θ is then obtained from the equation:

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

Example: If

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

then

$E(X; \mu, \sigma^2) = \mu$

Method of moments:

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

13.33 Method of Moments

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

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

What is the estimator of a?

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

13.34 Method of Moments

The distribution has one parameter. The method of moments gives us one equation

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

That is

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

That we solve for $\hat{a}$

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

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

13.35 Method of Moments

Note that taking the minimum of the measurements is clearly suboptimal.

The method gave us a clever answer:

- we can compute $\bar{x}$ with increasing precision given by n
- We know that no measurement surpasses $b = 1$
- Then compute the distance between $\bar{x}$ and b: $1 - \bar{x}$
- Subtract it from $\bar{x}$: $\bar{x} - (1 - \bar{x}) = 2\bar{x} - 1$

13.36 Method of Moments

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

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

If there are more parameters, we use the higher **sample moments**

- The second sample moment is

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

as such, an observation of this moment is

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

The method 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:

- $E(X;\hat{\theta}_1,\hat{\theta}_2) = \frac{1}{n}\sum_i x_i$
- $E(X^2;\hat{\theta}_1,\hat{\theta}_2) = \frac{1}{n}\sum_i x_i^2$

13.37 Normal distribution

If X distributes normally

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

then it has mean and variance:

$E(X;\mu,\sigma^2) = \mu$ and $V(X;\mu,\sigma^2) = \sigma^2$

Method of moments gives the equations:

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

13.38 Normal distribution

A substitution of $E(X)$ into the first equation gives the estimator for the mean μ.

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

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

then

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

which can also be written as: $\frac{1}{n}\sum_i (x_i - \hat{\mu})^2$

The method of moments and the maximum likelihood method give the same result for the normal distribution. Is this always the case?

13.39 Method of Moments

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.

13.40 Method of Moments

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

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

We need to compute the expected value $E(X)$

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

and equate it to the average $\bar{x}$

13.41 Method of Moments

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

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

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

Therefore, the method of moments gives us the equation

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

which solving for $\hat{\alpha}$ gives us the estimate

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

For our 30 lasers, this is

$\hat{\alpha}_m = 1.314$

13.42 Method of Moments

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

- $\hat{\alpha}_{ml} = -\frac{1}{n}\sum_{i=1}^{n} \ln(x_i - 1) = 1.320$
- $\hat{\alpha}_m = \frac{1-\bar{x}}{\bar{x}} = 1.314$

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

Chapter 14

Interval estimation

14.1 Objective

- Interval estimation for the mean and the proportion
- Interval estimation for the variance

14.2 Average or sample mean

Definition

The sample mean (or average) of a **random sample** of size n

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Because each random experiment is independent, the mean and variance of $\bar{X}$ are

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

$\bar{X}$ is therefore

- an **estimator** of $E(X)$, that is μ.
- a random variable

14.3 Inference on the average

Example:

You perform 8 random experiments: Load a cable until it breaks and record the breaking load. These are the results.

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

If we **know** that our cables truly distribute as

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

then

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

- $E(\bar{X}) = 13$
- $V(\bar{X}) = \frac{0.35^2}{8} = 0.01530169$; $se = \frac{0.35}{\sqrt{8}} = 0.1237$

then the **observed error** in the estimation is the difference

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

14.4 Margin of error

When deciding whether the **error** in estimation: $\bar{X} - \mu$ is large or not we usually compare it with a predefined tolerance.

- The **margin of error** at 5% level is the distance m such that distribution of $\bar{X}$ captures 95% of the estimations:

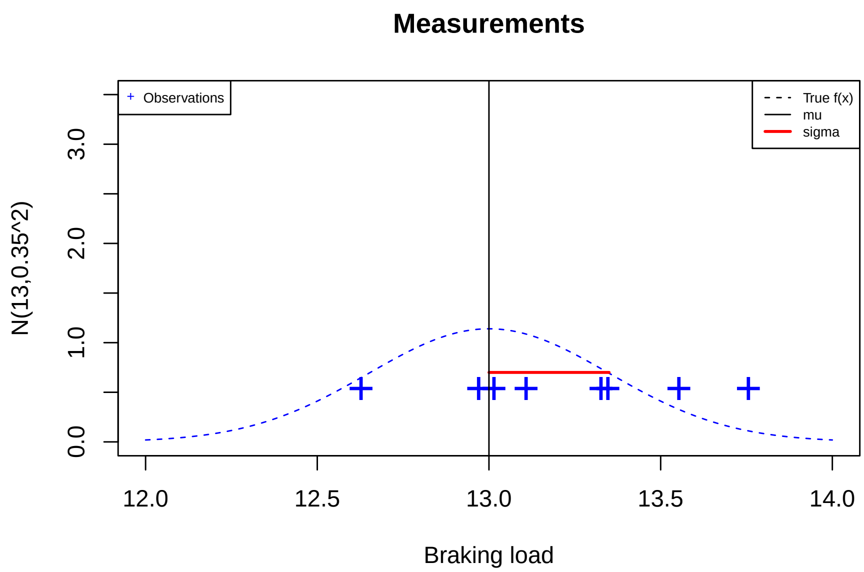
$$P(-m \leq \bar{X} - \mu \leq m) = P(\mu - m \leq \bar{X} \leq \mu + m) = 0.95$$

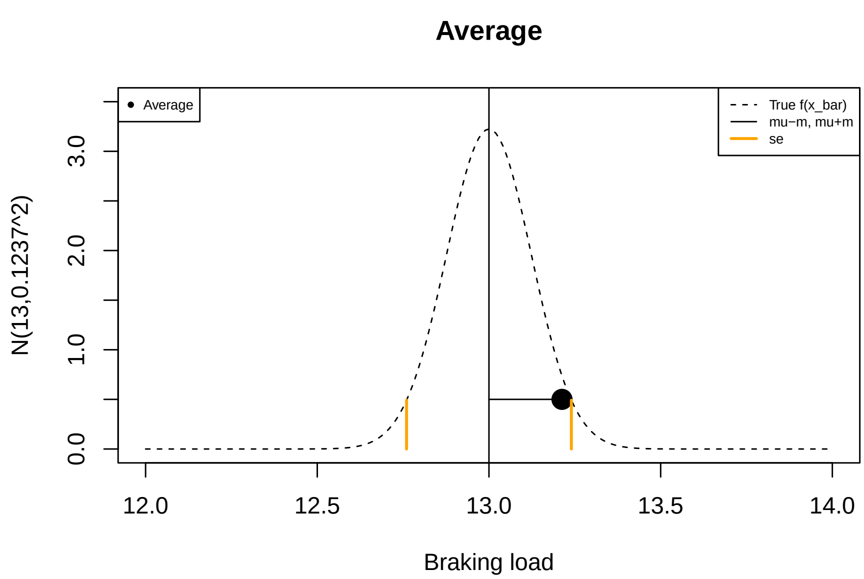
- or that 95% of the values of $\bar{X}$ are a distance m from μ.

In our example, we assume that $\bar{X}$ is normally distributed then

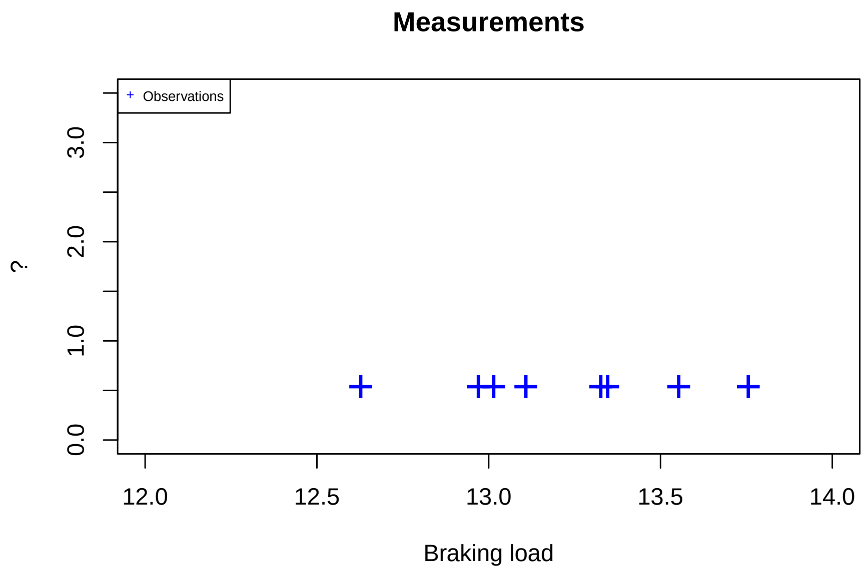
$$m = z_{0.025}\frac{\sigma}{\sqrt{n}} = 1.96 \times se = 1.96\frac{0.35}{\sqrt{8}} = 0.24$$

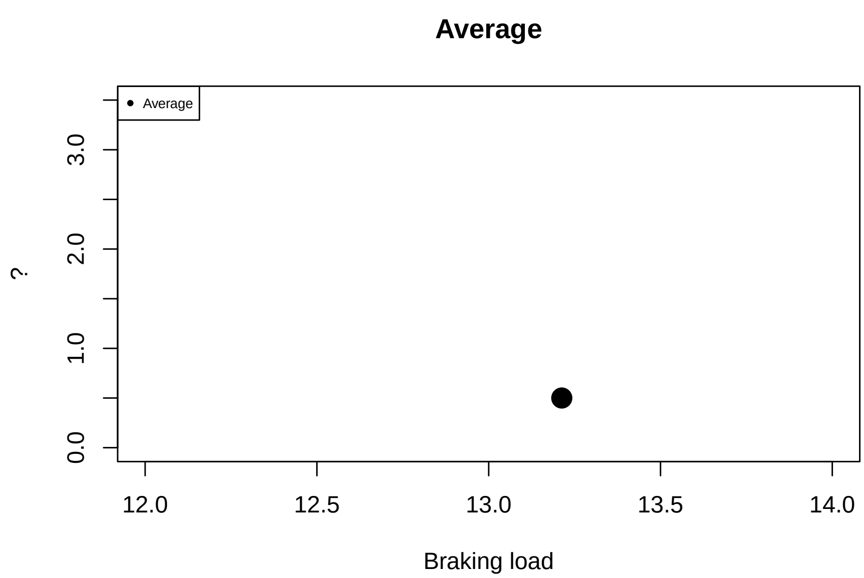
14.5 Outcome probability density Vs sample mean probability density





14.6 Real life





14.7 Interval estimation

We could estimate the margin of errors and errors because we "**knew**":

- A distribution for X
- with known μ
- and known σ^2

In real life we do not know any, but:

- We can assume a distribution of X
- We can estimate parameters

14.8 Interval estimation

From the margin of error equation:

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

let's solve for μ (the real unknown)

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

The left and right limits of the inequality are random variables which motivate the definition for the **random confidence interval at** 95%

- (L, U) such that $P(L \leq \mu \leq U) = 0.95$.

When the interval captures the **error**: $(\bar{X} - \mu)$ then

$$(L, U) = (\bar{X} - m, \bar{X} + m)$$

This interval is a **random variable** and it has by definition a probability of 0.95 to contain μ.

14.9 Interval estimation

When we perform n-random experiments (n-sample) we can calculate m if X is normal and we know σ^2.

- the interval that we obtain from the experiment is (script size)

$$(l, u) = (\bar{x} - m, \bar{x} + m)$$

- this interval either contains or does not the parameter μ: we will **never know**!
- We say that we have a confidence of 95% that the interval (l, u) will capture the true unknown parameter μ. Think of buying a lottery ticket for which you do not know the result.

14.10 Interval estimation

In our example, we assume that $\bar{X}$ is normally distributed then

$$m = z_{0.025} \frac{\sigma}{\sqrt{n}}$$

and the 95% confidence interval is

$$(l, u) = (\bar{x} - z_{0.025} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{0.025} \frac{\sigma}{\sqrt{n}}) = (12.97, 13.45)$$

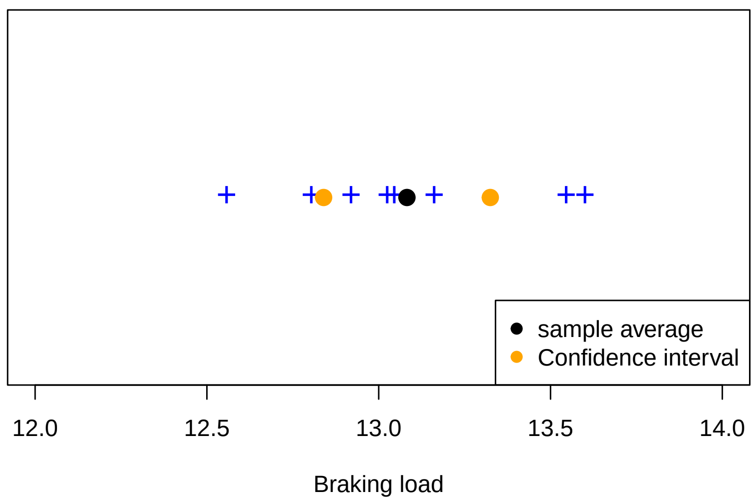
or

$$\hat{\mu} = 13.21 \pm 0.24$$

It also means that, in the estimation, we are confident about the units but not so much about the decimal places.

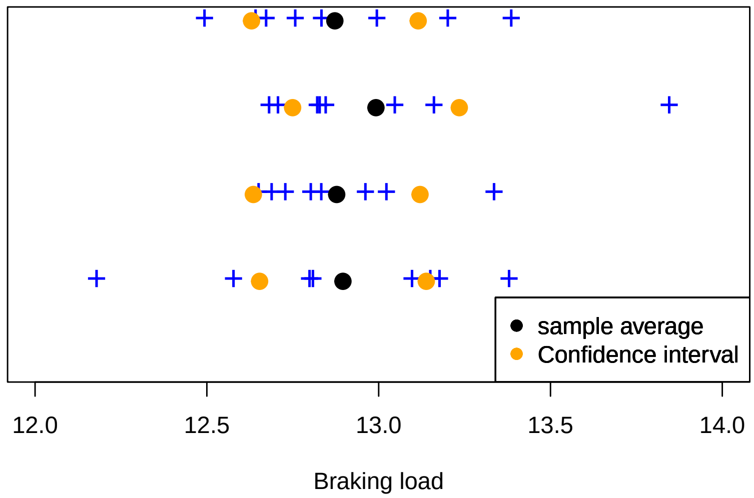
14.11 Interval estimation

For a sample of 8 observations, we have one estimate of the mean and one confidence interval



14.12 Interval estimation

Every time that we obtain a new sample then the estimates change. If we perform 100 samples then 95 of the confidence intervals will contain μ (we do not know which!)



14.13 Interval estimation

If the 95% confidence interval at confidence limit of $\alpha = 0.05$ (the amount of probability that is left out in the probability distribution) when $X \rightarrow N(\mu, \sigma)$ and known σ^2 is

$$(l, u) = (\bar{x} - z_{0.025}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{0.025}\frac{\sigma}{\sqrt{n}})$$

Likewise, the 99% confidence interval at confidence limit $\alpha = 0.01$ is

$(l, u) = (\bar{x} - z_{0.005}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{0.005}\frac{\sigma}{\sqrt{n}})$

$$= (\bar{x} - 2.58\frac{\sigma}{\sqrt{n}}, \bar{x} + 2.58\frac{\sigma}{\sqrt{n}})$$

or

$$\hat{\mu} = \bar{x} \pm 2.58\frac{\sigma}{\sqrt{n}}$$

If we want to be more confident then we need larger confidence intervals!

For our cables:

$$\hat{\mu} = 13.21 \pm 0.31$$

14.14 Interval estimation

14.15 Example

A metallic material is tested for impact to measure the energy required to cut it at a given temperature.

- Ten specimens of A238 steel were cut at 60ºC at the following impact energies (J)
- $64.1, 64.7, 64.5, 64.6, 64.5, 64.3, 64.6, 64.8, 64.2, 64.3$
- If we know that the impact energy is randomly distributed with $\sigma = 1J$ what is the 95% CI for the mean of these data?

14.16 Example

We know

- $x_i = \{64.1, 64.7, 64.5, 64.6, 64.5, 64.3, 64.6, 64.8, 64.2, 64.3\}$
- $X \to N(\mu, \sigma^2)$
- $\sigma = 1J$
- $\alpha = 0.05$

The confidence interval is then

$$CI = (\bar{x} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96\frac{\sigma}{\sqrt{n}})$$

$$= (64.46 - 1.96\frac{1}{\sqrt{10}}, 64.46 + 1.96\frac{1}{\sqrt{10}}) = (63.84, 65.08)$$

or

$$\hat{\mu} = 64.46 \pm 0.61$$

this tells us that we can be sure on the first digit (6), somewhat confident on the second (4), and unsure on the decimals (46).

What if σ^2 is **unknown**?

14.17 T-statistic

When

- σ is **unknown**

However, if X is normal

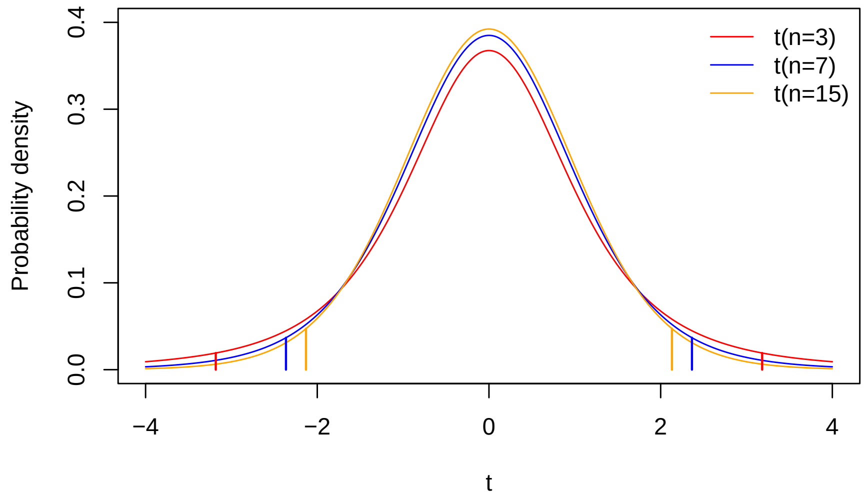
$$X \to N(\mu, \sigma^2)$$

then the standardized statistic

$$T = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}}$$

Follows a t-distribution with $n - 1$ degrees of freedom, and we can compute probabilities on $\bar{X}$.

14.18 T-statistic



14.19 T-statistic

To compute the margin of error m at 5% level when n is small, σ^2 unknown but X normal

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

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

We use the t-distribution

$$m = t_{0.025,n-1}\frac{s}{\sqrt{n}}$$

where $t_{0.025,n-1}$ is the value T that leaves 2.5% at right hand side of t-distribution with $n-1$ degrees of freedom (0.025-quantile)

the 95% confidence interval is then

$$(l, u) = (\bar{x} - t_{0.025,n-1}\frac{s}{\sqrt{n}}, \bar{x} + t_{0.025,n-1}\frac{s}{\sqrt{n}})$$

in R: $t_{0.025,n-1}$=qt(1-0.025, n-1)

14.20 Example

A metallic material is tested for impact to measure the energy required to cut it at a given temperature.

- Ten specimens of A238 steel were cut at 60ºC at the following impact energies (J)
- $64.1, 64.7, 64.5, 64.6, 64.5, 64.3, 64.6, 64.8, 64.2, 64.3$
- If we know that the impact energy is randomly distributed but we **do not know** the variance what is the 95% CI for the mean of these data?

14.21 Example

- $\bar{x} = 64.46$
- $s = 0.227$
- $\alpha = 0.05$
- $t_{0.025,9} = 2.26$ obtained from $P(T \leq t_{0.025,9}) = 0.975$; qt(1-0.025, 9)

The CI interval is then

$CI = (\bar{x} - t_{0.025,9}\frac{s}{\sqrt{n}}, \bar{x} + t_{0.025,9}\frac{s}{\sqrt{n}})$

$$= (64.46 - 2.26\frac{0.227}{\sqrt{10}}, 64.46 + 2.26\frac{0.227}{\sqrt{10}})$$

$$= (64.29, 64.62)$$

but $CI = (63.84, 65.08)$ when $\sigma = 1$. Data suggests $\sigma < 1$.

R: t.test(c(64.1,64.7,64.5,64.6,64.5,64.3,64.6,64.8,64.2,64.3))

14.22 IC with CLT

If we do not know how X distributes but take a large sample $n \geq 30$ then we can use the CLT to find the CI intervals.

the 95% confidence interval is then

$$(l, u) = (\bar{x} - z_{0.025}\frac{s}{\sqrt{n}}, \bar{x} + z_{0.025}\frac{s}{\sqrt{n}})$$

since $t_{0.025,n-1} \rightarrow z_{0.025}$ for $n \rightarrow \infty$ then it is also ok to use the T distribution for large n and unknown distribution of X.

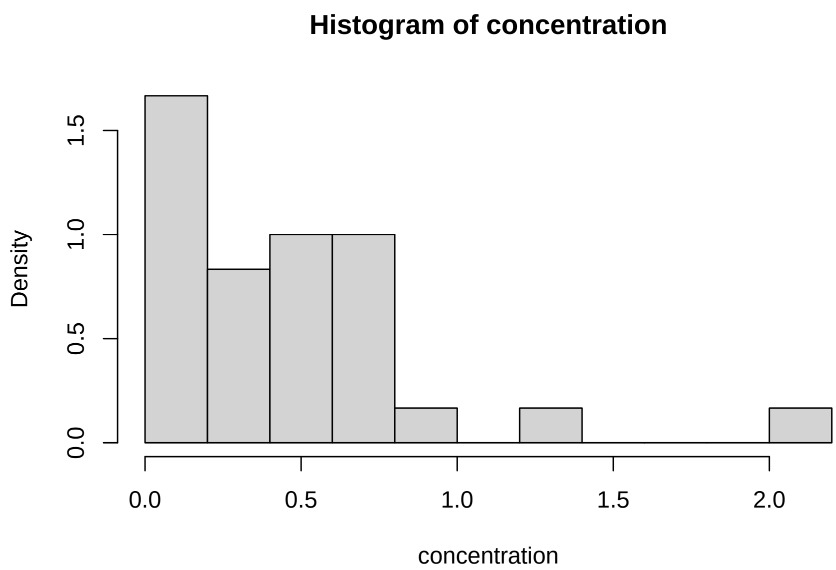
Note: This is why R only implements t.test and not z.test in the base functions to compute CI.

Example:

Consider an experiment where we measure the concentration in blood of a drug after 10-hour administration in 30 patients. We obtain the following results:

```
##  [1] 0.42172863 0.28830514 0.66452743 0.01578868 0.02810549 0.15825061
##  [7] 0.15711365 0.07263340 1.36311823 0.01457672 0.50241503 0.24010736
## [13] 0.14050681 0.18855892 0.09414202 0.42489306 0.78160177 0.23938021
## [19] 0.29546742 2.02050586 0.42157487 0.48293561 0.74263790 0.67402224
## [25] 0.58426449 0.80292617 0.74837143 0.78532627 0.01588387 0.29892485
```

- the average is $\bar{x} = 0.4556198$
- the standard deviation is $s = 0.4335571$
- the histogram of the results is:



14.23 Central Limit Theorem

We **assumed** that $X \to exp(\lambda = 2)$

With mean and variance:

- $E(X) = \frac{1}{\lambda} = 0.5$
- $V(X) = \frac{1}{\lambda^2} = 0.25$

The error was

$$\bar{x} - E(X) = 0.4556198 - 0.5 = -0.0443802$$

14.24 CI with CLT

What happens if we **do not know** the value of $E(X)$?

- We use a 95% CI to estimate it

Since $n \geq 30$ we can use the CLT

$$\bar{X} \sim_{aprox} N(\lambda, \frac{1}{n\lambda^2})$$

and the 95% confidence interval is then

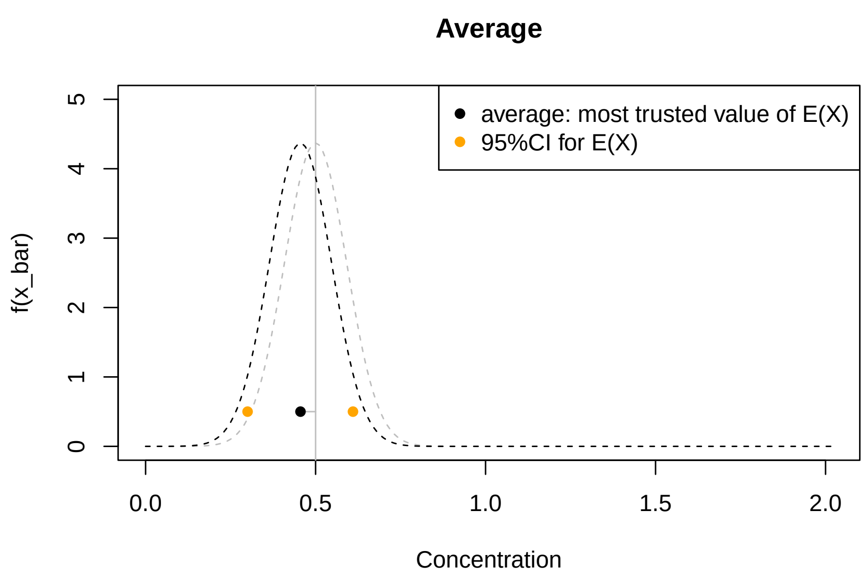
$$(l, u) = (\bar{x} - z_{0.025}\frac{s}{\sqrt{n}}, \bar{x} + z_{0.025}\frac{s}{\sqrt{n}})$$

$$(l, u) = (0.4556198 - 1.96\tfrac{0.4335571}{\sqrt{30}}, 0.4556198 + 1.96\tfrac{0.4335571}{\sqrt{30}})$$

$$= (0.300, 0.610)$$

or

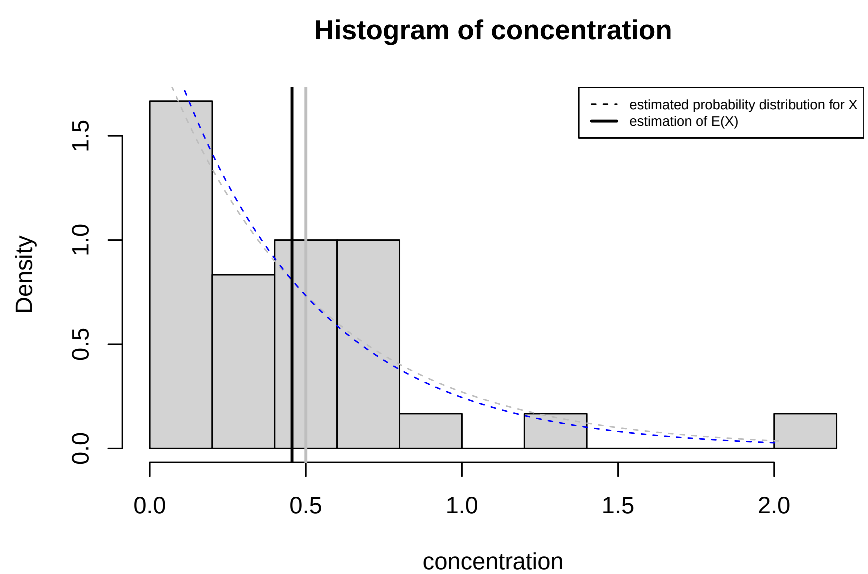
$$\hat{\mu} = 0.45 \pm 0.15$$



14.25 Parameter estimation

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

$$\hat{\lambda} = \frac{1}{\hat{\mu}} = 2.194812$$



or its 95% CI:

$$\hat{\lambda} = (1.66, 3.33)$$

14.26 Interval estimation for proportions

A random sample of 400 patients was selected for testing a new vaccine for the influenza virus, after 6 months of vaccination 136 were ill.

- What is the expected efficacy of the vaccine?

We have 136 failures in 400 trials, each trial is a Bernoulli trial

$$X \to Bernoulli(p)$$

with:

- the probability p of failure for one person ($x = 1$)
- mean $E(X) = p$
- variance $V(X) = p(1-p)$

We want to have a 95% CI for p.

14.27 Interval estimation for proportions

If the distribution of a random experiment is

$$X \to Bernoulli(p)$$

Then $\bar{X}$ has

- mean $E(\bar{X}) = E(X) = p$ (unbiased estimator of p)
- variance $V(\bar{X}) = \frac{V(X)}{n} = \frac{p(p-1)}{n}$ (consistent estimator of p)

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

14.28 Interval estimation for proportions

When $\hat{p}n > 5$ and $(\hat{p}-1)n > 5$

- The **standardized statistic** of $\bar{X}$ can be approximated by a standard distribution

$$Z = \frac{\bar{X} - E(\bar{X})}{\sqrt{V(\bar{X})}} = \frac{\bar{X} - p}{\left[\frac{p(1-p)}{n}\right]^{1/2}} \to N(0,1)$$

- The 95% CI interval of p is:

$$CI = (l, u) = (\bar{x} - z_{0.025}\left[\frac{\bar{x}(1-\bar{x})}{n}\right]^{1/2}, \bar{x} + z_{0.025}\left[\frac{\bar{x}(1-\bar{x})}{n}\right]^{1/2})$$

Where we estimate the Bernoulli variance $p(1-p)$ by $\bar{x}(1-\bar{x})$.

14.29 Interval estimation for proportions

In our case, we are counting failures on vaccinations 136 in 400 trials

we know

- $\bar{x} = 134/400 = 0.34$
- $z_{0.025} = 1.96$

$CI = (l, u) = (\bar{x} - 1.96[\frac{\bar{x}(1-\bar{x})}{n}]^{1/2}, \bar{x} + 1.96[\frac{\bar{x}(1-\bar{x})}{n}]^{1/2})$

$$= (0.29, 0.39)$$

The probability of failure of the vaccine is

$$\hat{p} = 0.34 \pm 0.05$$

Note: Polls for the intention to vote (Bernoulli trial) in a sample of n individuals report this type of estimate with its margin of error. It does not mean that the **true value** of p is within this interval with probability 95%.

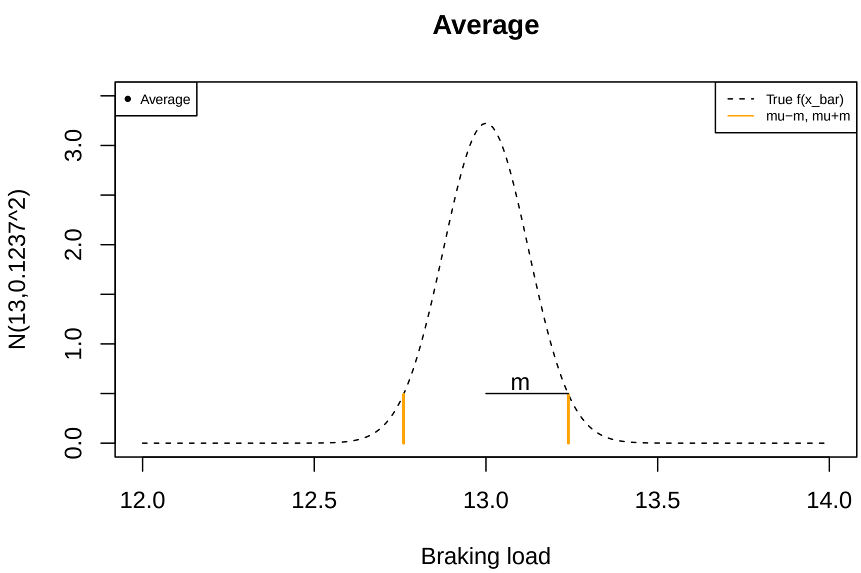
14.30 Probability Vs Confidence

There are two views of uncertainty:

- **Future**: probability on observations

When **we know** the probability distribution of our random experiment we ask:

What is the **probability** that a **new** value of $\bar{X}$ is close to μ ?



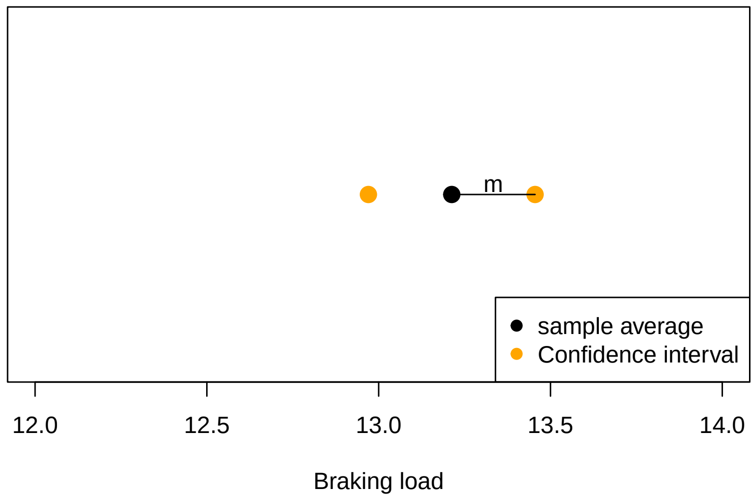
14.31 Probability Vs Confidence

- **Present**: confidence on parameters

When we have **observations** and do **not know** the parameter μ we ask:

What is the range of values of $\bar{X}$ where we **believe** that μ **is** with 95% confidence?

We use the **margin of error** to compute the CI, but it does not mean we have calculated an error (we don't know where μ is).



14.32 Interval estimation for the variance

A metallic material is tested for impact to measure the energy required to cut it at a given temperature.

- Ten specimens of A238 steel were cut at 60ºC at the following impact energies (J)
- $64.1, 64.7, 64.5, 64.6, 64.5, 64.3, 64.6, 64.8, 64.2, 64.3$

We know that the estimate for $s^2 = 0.227^2 = 0.051$, but what is its confidence interval?

14.33 Interval estimation for the variance

When $X \hookrightarrow N(\mu, \sigma^2)$.

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

Captures the proportion in the error of σ^2 and follows a χ^2 distribution with $n-1$ degrees of freedom

$$\frac{S^2}{\sigma^2}(n-1) \rightarrow \chi^2_{n-1}$$

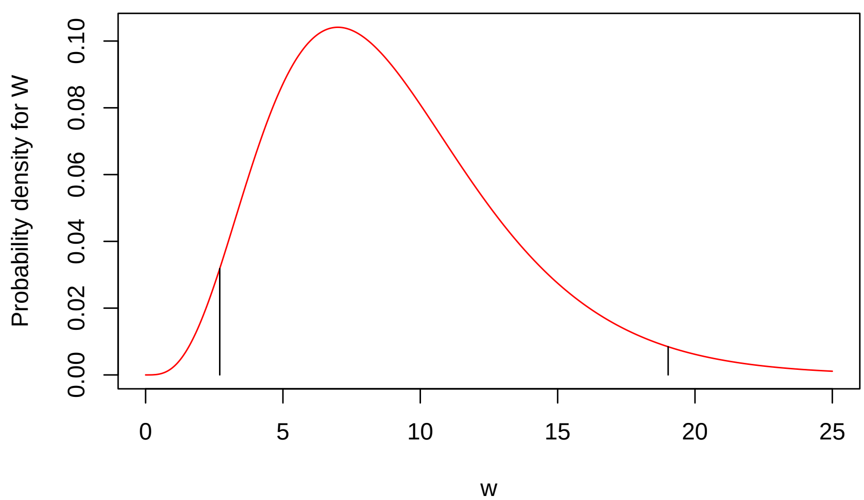
- We look for confidence interval of σ^2 at confidence 95% (L, U) such that

$$P(L \leq \sigma^2 \leq U) = 0.95$$

We can use the χ^2 to determine the 95% of the distribution about W

$$P(\chi^2_{0.975,n-1} \leq W \leq \chi^2_{0.025,n-1}) = 0.95$$

14.34 χ^2-statistic



14.35 Interval estimation for the variance

replacing the value of W

$$P(\chi^2_{0.975,n-1} \leq \frac{S^2}{\sigma^2}(n-1) \leq \chi^2_{0.025,n-1}) = 0.95$$

and solving for σ^2

$$P(\frac{S^2(n-1)}{\chi^2_{0.025,n-1}} \leq \sigma^2 \leq \frac{S^2(n-1)}{\chi^2_{0.975,n-1}}) = 0.95$$

The random interval at 95% confidence

$$(L, U) = (\frac{S^2(n-1)}{\chi^2_{0.025,n-1}}, \frac{S^2(n-1)}{\chi^2_{0.975,n-1}})$$

and the 95% confidence interval (script size)

$$(l, u) = (\frac{s^2(n-1)}{\chi^2_{0.025,n-1}}, \frac{s^2(n-1)}{\chi^2_{0.975,n-1}})$$

14.36 Interval estimation for the variance

$\chi^2_{0.975,n-1} = F^{-1}(0.025)$ for $n = 10$ or $df = n - 1 = 9$

```
chi0.975 <- qchisq(0.025, df=9)
chi0.975
```

```
[1] 2.700389
```

```
chi0.025 <- qchisq(0.975, df=9)
chi0.025
```

```
[1] 19.02277
```

14.37 Interval estimation

In our example

- $s = 0.227$
- $n = 10$

$\hat{\sigma}^2 = (l, u) = (\frac{s^2(n-1)}{\chi^2_{0.025,n-1}}, \frac{s^2(n-1)}{\chi^2_{0.975,n-1}})$

$$= (\frac{0.227^2(10-1)}{19.02277}, \frac{0.227^2(10-1)}{2.700389}) = (0.02, 0.17)$$

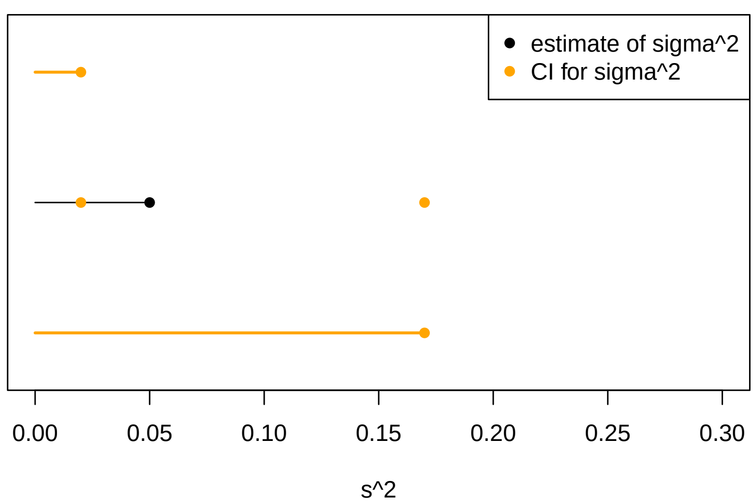
According to the data $\sigma^2 \neq 1$ at 95% confidence.

- Had we made an error considering $\sigma = 1$ when we calculated the first CI for this data?

in R: library(Ecfun); confint.var(0.05, 9)

14.38 Interval estimation

The interval for the variance is **not symmetric** and we cannot formulate it as an estimate $\pm$ margin of error.



Chapter 15

Hypothesis testing

15.1 Objective

- Hypothesis testing of means and proportions
- Hypothesis testing of variances
- Errors in hypothesis testing

15.2 Hypothesis

When we make inferences about our process, we often want to test if the process satisfies a desired condition/property

- Measurements and their inferences provide **evidence** for that condition.
- We can formulate the condition in terms of the values that some **parameters** of probability distributions can take.

15.3 Hypothesis

Examples:

- Tyre manufacturers want to know whether the half-life of the tires they produce is at least 20,000 km

- Fertilizer developers want to test whether their new product has a real effect on the growth of plants
- Pharmaceutical companies need to know if chemotherapy can cure 90% of cancer patients

These questions can be translated into statements of probability distributions

15.4 Hypothesis

- Tyre manufacturers want to know whether the half-life of the tires they produce is at least 20,000 km

Assuming that the life of tires follows a population probability distribution, we are interested in finding if the mean of the distribution is at least 20,000Km.

This can be done in two dichotomic statements

- The mean life of tires is **less** than 20,000km
- The mean life of tires is **greater** than 20,000km

15.5 Hypothesis

or being μ the mean of the population distribution

- $H_0 : \mu \leq 20,000km$
- $H_1 : \mu > 20,000km$

15.6 Hypothesis

Definition

In statistics, a statement (conjecture) about the distribution of a random variable is called a **hypothesis**.

The hypothesis is usually written in two dichotomous statements

- The **null** hypothesis: H_0 when the conjecture is False (usually refers to status quo)

- The **alternative** hypothesis: H_1 when the conjecture is True (usually refers to research hypothesis)

15.7 Null hypothesis

So what are the null and the alternative hypothesis for these situations?

- Tyre manufacturers want to know whether the half-life of the tires they produce is at least 20,000 km
- Fertilizer developers want to test whether their new product has a real effect on the growth of plants
- Pharmaceutical companies need to know if chemotherapy can cure 90% of cancer patients

15.8 Null hypothesis

- Fertilizer developers want to test whether their new product has a real effect on the growth of plants

Being μ_0 the mean growth of the plants **without** fertilizer (known) and μ the mean growth of the plants with the fertilizer (unknown)

- $H_0 : \mu \leq \mu_0$ (The fertilizer does nothing: status quo)
- $H_1 : \mu > \mu_0$ (The fertilizer has the desired effect: research interest)

What could be a suitable distribution of μ?

Example:

You perform 8 random experiments: Load a cable until it breaks and record the breaking load. These are the results.

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

- The average of the data is $\bar{x} = 13.21$
- The standard deviation is $s = 0.3571565$
- We may want to use this data to show that our cables break **on average** at more than 13 Tons.

Example:

If we **hypothesize** that our cables truly distribute as

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

then

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

We ask:

- Are the measurements consistent with the null hypothesis $H_0 : \mu = 13$?

15.9 Hypothesis test with acceptance/rejection zones

- $H_0 : \mu = 13$ (cables break as usual: status quo)
- $H_1 : \mu \neq 13$ (cables do not brake as usual: research interest)

To test the hypothesis contrast the standardized **observed error** with the standardized **margin of error** from the null hypothesis.

Since

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

Then the **standardized error** from the null hypothesis follows a standard distribution

$$Z = \frac{\bar{X} - 13}{\frac{0.35}{\sqrt{8}}} \to N(0, 1)$$

15.10 standardized margin of errors

- The **standardized margin of errors** are the quantiles of Z

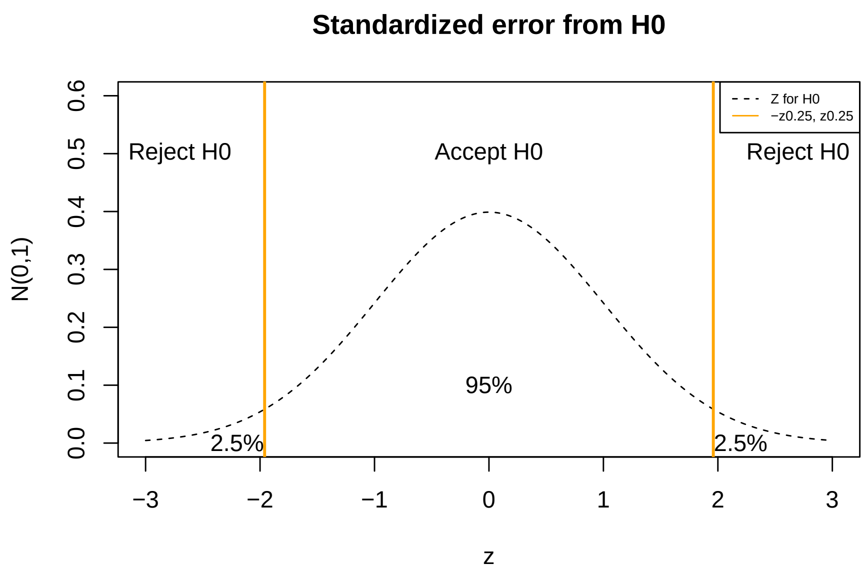
$$P(-z_{0.025} \leq Z \leq z_{0.025}) = 0.95$$

The interval:

$$(-z_{0.025}, z_{0.025})$$

is called acceptance interval of H_0 at 95% confidence level.

- $\alpha = 0.05 = 2 \times 0.025 = 1 - 0.95$ is called the **significance** limit.



15.11 Standardized observed error

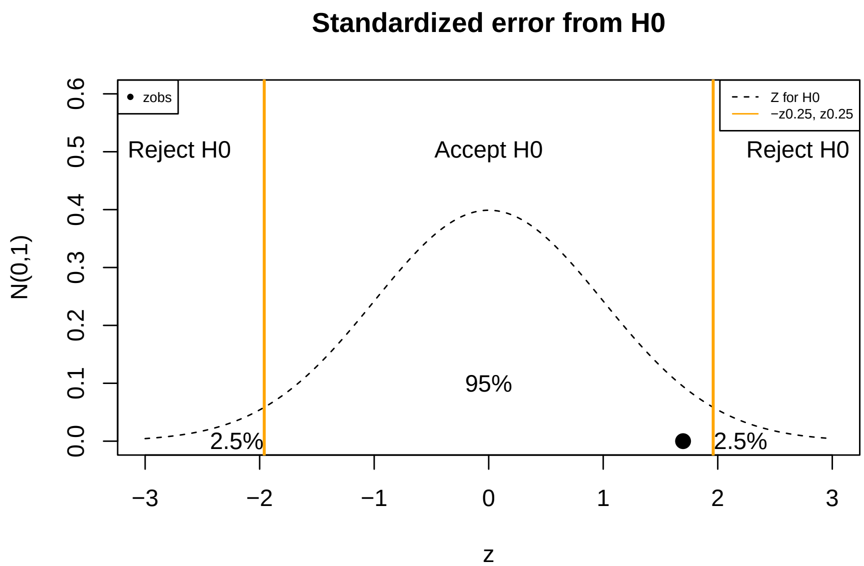
- The **standardized observed error** is

$$z_{obs} = \frac{\bar{x} - 13}{\frac{0.35}{\sqrt{8}}} = 1.697056 \in (-z_{0.025}, z_{0.025})$$

We conclude:

- Our observed error is consistent with 95% of the observations for the statistic Z when the null hypothesis is true

- We accept that the H_0 is true and give up on the idea that we have stronger cables than expected.

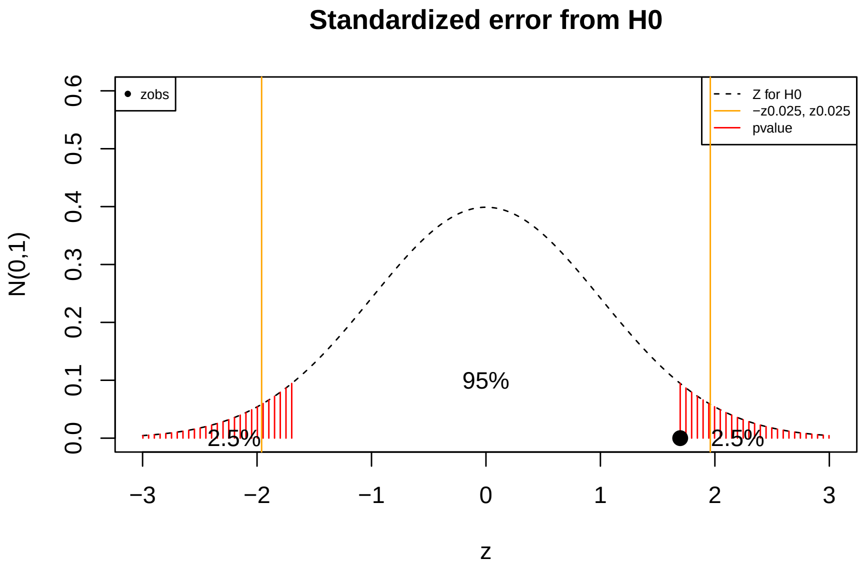


15.12 Hypothesis test with P-value

We can also contrast the hypothesis by calculating the probability that the average of another sample will be rarer than the average we just observed.

$$pvalue = P(Z \leq -z_{obs}) + P(z_{obs} \leq Z) = 2(1 - \phi(|z_{obs}|))$$

- We reject H_0 if $pvalue \leq \alpha = 0.05$



15.13 Standardized observed error

- The **pvalue** is

$$pvalue = 2(1 - \phi(1.697056)) = 0.089$$

R: 2*(1-pnorm(1.697056))

We conclude:

- If we performed a new sample is likely that we can get a more extreme result for the average at limit $\alpha = 0.05$ if the null hypothesis is true.
- We accept that H_0 could have produced our data and give up on the idea that we have stronger cables than expected.

15.14 Hypothesis test Confidence Interval

From the point of view of the estimation, we can also contrast the hypothesis.

- We trust that our estimation of μ is correct with 95% confidence

The CI is:

$$(l, u) = (\bar{x} - z_{0.025}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{0.025}\frac{\sigma}{\sqrt{n}}) = (12.97, 13.45)$$

- The CI tells us that we can be 95% confident that we have captured the true value of μ.

- We don't know the true value of μ but $H_0 : \mu = 13$ Tons could be it.

15.15 Hypothesis test Confidence Interval

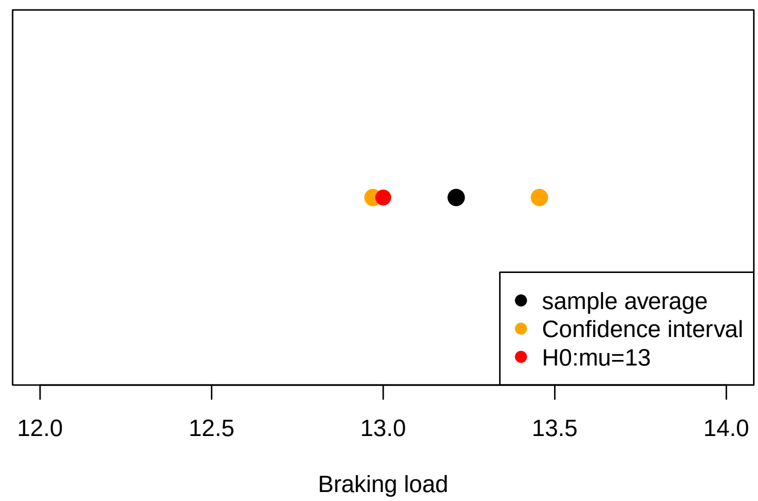
- Since

$$H_0 : \mu = 13 \in (12.97, 13.45)$$

We conclude:

- Our data is consistent with the fact that our estimate of μ is the null hypothesis.

- We accept that H_0 could have produced our interval and give up on the idea that we have stronger cables than expected.



15.16 Hypothesis test with unknown variance

It is common to **hypothesize** the values of the parameters we can contrast. Other nuisance parameters we may leave unknown.

We can **hypothesize** that our cables truly distribute as

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

then

$$\bar{X} \rightarrow N(13, \frac{\sigma^2}{8})$$

We ask again:

- Are the measurements consistent with the null hypothesis $H_0 : \mu = 13$?

15.17 Standardized error with unknown variance

If X is normal

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

then the **standardized errors** with respect to the **sample standard deviation** S

$$T = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}}$$

Follows a t-distribution with $n-1$ degrees of freedom.

15.18 Hypothesis testing with unknown variance

we accept H_0 because of any of following the equivalent contrasts:

1. The acceptance region for H_0 is:

$$(-t_{0.025,7}, t_{0.025,7}) = (-2.36, 2.36)$$

and the observed standardized error from H_0 is

$$t_{obs} = \frac{13.21268 - 13}{\frac{0.3571565}{\sqrt{8}}} = 1.6843$$

within the acceptance region.

15.19 Hypothesis testing with unknown variance

2. The

$$pvalue = 2(1 - F^{-1}_{t,7}(1.6843)) = 0.136$$

R: 2*(1-pt(1.6843,7))

is higher than $\alpha = 0.05$

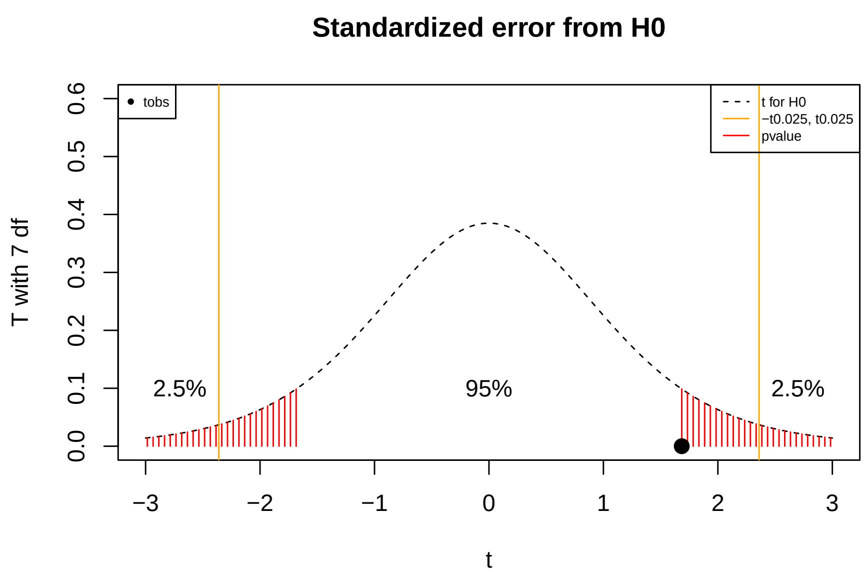
3. The confidence interval

$$(\bar{x} - t_{0.025,n-1}\frac{s}{\sqrt{n}}, \bar{x} + t_{0.025,n-1}\frac{s}{\sqrt{n}}) = (12.91409, 13.51127)$$

contains $H_0 : \mu = 13$.

---

in R: t.test(c(13.34642, 13.32620, 13.01459, 13.10811, 12.96999, 13.55309, 13.75557, 12.62747), mu=13)



15.20 One-tailed test

We may be interested in only testing for the fact that our estimate is higher than the null hypothesis (we do not care if it is lower)

Upper-tailed test:

- $H_0 : \mu \leq 13$ (at most cables break as usual)
- $H_1 : \mu > 13$ (cables break at a higher load)

We will test the higher tail of the distribution.

15.21 Hypothesis testing of the upper tail

In this example, we accept H_0 because of any of the following equivalent contrasts:

1. The acceptance region for H_0 is:

$$(-\infty, t_{0.05,7}) = (-\infty, 1.894579)$$

and the observed standardized error from H_0 is

$$t_{obs} = \frac{13.21268 - 13}{\frac{0.3571565}{\sqrt{8}}} = 1.6843$$

within the acceptance region.

15.22 Hypothesis testing with unknown variance

2. For the upper tail

$$pvalue = 1 - F^{-1}_{t,7}(1.6843) = 0.06799782$$

R: 1-pt(1.6843,7)

is higher than $\alpha = 0.05$

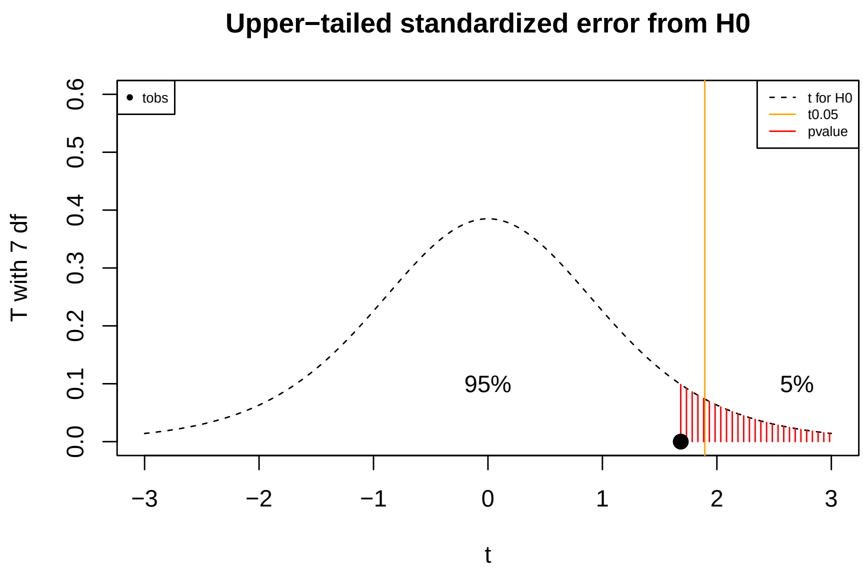
3. The **upper tailed** confidence interval

$$(\bar{x} - t_{0.05,n-1}\frac{s}{\sqrt{n}}, \infty) = (12.97344, \infty)$$

contains $H_0 : \mu = 13$.

---

in R: t.test(c(13.34642, 13.32620, 13.01459, 13.10811, 12.96999, 13.55309, 13.75557, 12.62747), mu=13, alternative="greater")



15.23 Example 1:

11.6g of NaCl is dissolved in 100g of water and has a molar concentration of 1.92mol/L

We design a process to remove salt from this concentration and obtain the following results

```
## [1] 1.716 1.889 1.783 1.849 1.891
```

- We want to test at 0.05 significant threshold if the process does remove salt from the concentration.

15.24 Example 1:

Two-tailed test:

- $H_0 : \mu = 1.92$; $H_1 : \mu \neq 1.92$

```
t.test(c(1.716, 1.901, 1.783, 1.849, 1.891),
       mu=1.92, alternative = "two.sided")
```

```
##
##  One Sample t-test
##
## data:  c(1.716, 1.901, 1.783, 1.849, 1.891)
## t = -2.6389, df = 4, p-value = 0.05764
## alternative hypothesis: true mean is not equal to 1.92
## 95 percent confidence interval:
##  1.731206 1.924794
## sample estimates:
## mean of x
##     1.828
```

Lower-tailed test:

- $H_0 : \mu \geq 1.92$; $H_1 : \mu < 1.92$

```
t.test(c(1.716, 1.901, 1.783, 1.849, 1.891),
       mu=1.92, alternative = "less")
```

```
##
##  One Sample t-test
##
## data:  c(1.716, 1.901, 1.783, 1.849, 1.891)
## t = -2.6389, df = 4, p-value = 0.02882
## alternative hypothesis: true mean is less than 1.92
## 95 percent confidence interval:
##      -Inf 1.902322
## sample estimates:
## mean of x
##     1.828
```

15.25 Example 2:

In some cases, we are not sure about the numerical value of the hypothesis to test, but we know that we want to improve the value of a parameter in two different conditions.

In the original paper of Gosset, he analyzed the effect of two soporific medicines.

- 10 individuals were given **soporific 1** and wrote down the additional hours slept under treatment, with a mean 0.75

```
##  [1]  0.7 -1.6 -0.2 -1.2 -0.1  3.4  3.7  0.8  0.0  2.0
```

- The same 10 individuals were given **soporific 2** and wrote down the additional hours slept under treatment, with a mean 2.33

```
##  [1]  1.9  0.8  1.1  0.1 -0.1  4.4  5.5  1.6  4.6  3.4
```

Scientific hypothesis: Soporific 2 is better than soporific 1

15.26 Example 2:

For each individual, Gosset made the difference between the treatments. Taking X as the difference between treatments, this was the sample observed for X

```
##  [1] 1.2 2.4 1.3 1.3 0.0 1.0 1.8 0.8 4.6 1.4
```

finding an average of treatment gain from soporific 2 with respect to soporific 1 of 1.58, and $s = 1.229995$

Upper-tailed paired t-test:

- $H_0 : \mu \leq 0$ (no treatment difference); $H_1 : \mu > 0$ (gain in treatment 2)

Where μ is the mean of the differences between treatments.

15.27 Example 2:

The **standardized error** is:

$$T = \frac{\bar{X}}{\frac{S}{\sqrt{n}}}$$

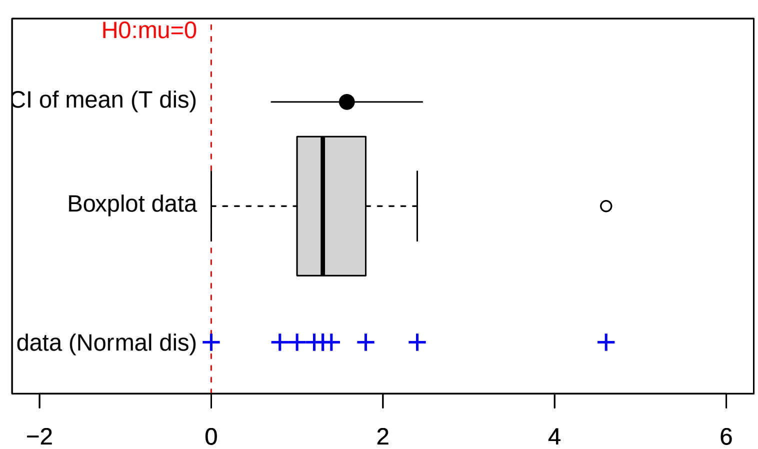
and its observation

$$t_{obs} = \frac{\bar{x}}{\frac{s}{\sqrt{n}}}$$

which is also known as the **signal** to **noise** ratio.

```
t.test(c(1.9,0.8,1.1,0.1,-0.1,4.4,5.5,1.6,4.6,3.4),
       c(0.7,-1.6,-0.2,-1.2,-0.1,3.4,3.7,0.8,0,2),
       paired = TRUE,
       alternative="greater")
```

```
##
##  Paired t-test
##
## data:  c(1.9, 0.8, 1.1, 0.1, -0.1, 4.4, 5.5, 1.6, 4.6, 3.4) and c(0.7, -1.6, -0.2,
## t = 4.0621, df = 9, p-value = 0.001416
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.8669947       Inf
## sample estimates:
## mean of the differences
##                    1.58
```



15.28 Hypothesis testing with large n and any distribution

On many occasions, X is not normally distributed but we can take large samples $n \geq 30$ then we can use the CLT:

Then the **standardized error** from the null hypothesis follows a standard distribution

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

and proceed as before, and if σ is unknown we replace it with its estimate s.

15.29 Hypothesis testing for proportions

Example:

We may be satisfied with a new process if 90% of the times we improve the previous process.

- If we run a sample of 200 new processes and find that 188 times we improved the previous process, can we be satisfied with the new process at 95% confidence?

15.30 Interval estimation for proportions

We hypothesize that the distribution of a random experiment is

$$X \to Bernoulli(p)$$

with an upper-tailed hypothesis contrast for p:

- $H_0 : p \leq 0.9$ (Not satisfactory)
- $H_1 : p > 0.9$ (Satisfactory)

Then if the null hypothesis is true $\bar{X}$ has

- mean $E(\bar{X}) = E(X) = p = 0.9$ (unbiased estimator of p)

- variance $V(\bar{X}) = \frac{V(X)}{n} = \frac{p(1-p)}{n} = 0.00045$ (consistent estimator of p)
- The observed $\bar{X}$ was $\bar{x} = 188/200 = 0.94$

15.31 Interval estimation for proportions

- By the CLT, the **standardized error** from the null hypothesis

$$Z = \frac{\bar{X} - E(\bar{X})}{\sqrt{V(\bar{X})}} = \frac{\bar{X} - p}{\left[\frac{p(1-p)}{n}\right]^{1/2}} \to N(0,1)$$

is a standard normal variable, when $pn > 5$ and $(p-1)n > 5$.

15.32 Interval estimation for proportions

In this example, we **reject** H_0 because of any of the following equivalent contrasts:

1. The acceptance region for H_0 is:

$$(-\infty, z_{0.05}) = (-\infty, 1.644854)$$

and the observed standardized error from H_0 is

$$z_{obs} = \frac{0.94 - 0.90}{\sqrt{0.00045}} = 1.885618$$

outside the acceptance region (inside the rejection zone).

15.33 Interval estimation for proportions

2. For the upper tail

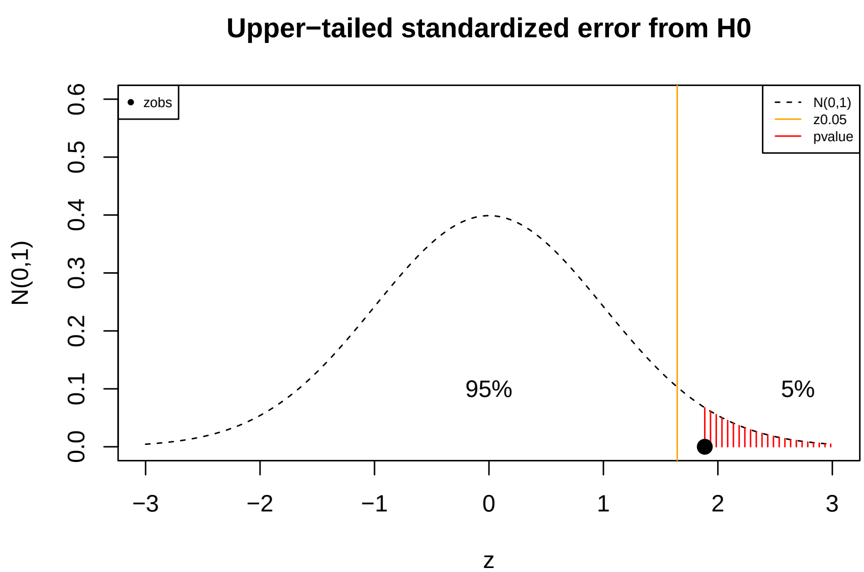
$$pvalue = 1 - \phi^{-1}(1.885618) = 0.02967323$$

R: 1-pnorm(1.885618)

is lower than $\alpha = 0.05$

15.34 Interval estimation for proportions

In R: prop.test(188, 200, p=0.9, alternative = "greater" , correct=FALSE)



15.35 Test for variances

In many cases, experiments are run to test specific values of the dispersion of data.

Such as

- for complying with strict design standards where measurements must be between certain values
- when relative measurements are taken such as the reaction of a treatment on an individual (insulin administration on an individual's sugar levels)

15.36 Test for variances

For a random sample $X_1, ...X_n$ with a normal population distribution ($X_i \to N(\mu, \sigma^2)$) the statistics defined by

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

Has a χ^2 (chi-squared) distribution with n-1 degrees of freedom given by

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

15.37 Test for variances

Suppose we want to test whether the variance of the population distribution is equal to a given value σ_0

- $H_0 : \sigma = \sigma_0$

Alternative hypothesis

- two tailed: $H_1 : \sigma \neq \sigma_0$
- upper tailed: $H_1 : \sigma > \sigma_0$
- lower tailed: $H_1 : \sigma < \sigma_0$

15.38 Test for variances

S^2 is an unbiased estimate of σ^2: $E(S^2) = \sigma^2$

The **standardized error ratio**

$$W = \frac{(n-1)S^2}{\sigma_0^2} \to \chi^2(n-1)$$

Follows a χ^2 distribution with $n-1$ degrees of freedom.

15.39 Example

- The production of a semiconductor chip is regulated by a process that requires that the thickness of a particular layer does not vary in more than $\sigma_0 = 0.6mm$, from its mean of $25mm$.
- To keep control of the process every so often a sample of 20 specimens is taken.
- If on one occasion the estimated standard deviation was $s = 0.8462188$ is the process out of control at 0.01 confidence and should be stopped?

This is the data:

```
##  [1] 24.51239 24.79975 26.35608 25.06134 25.11248 26.49211 25.40100 23.89940
##  [9] 24.40244 24.61227 26.06495 25.31304 25.34867 25.09629 24.51642 26.55461
## [17] 25.43313 23.28904 25.61018 24.58867
```

15.40 Test for variances

We want to contrast the hypotheses

- $H_0 : \sigma \leq 0.6$ (Process under control)
- $H_1 : \sigma > 0.6$ (Process out of control)
- Statistic: $W = \frac{(n-1)S^2}{\sigma_0^2} \rightarrow \chi^2(n-1)$
- Threshold limit $\alpha = 0.01$
- The acceptance region for H_0: $P(W \leq \chi^2_{0.01,19}) = 0.99$

$$(0, \chi^2_{0.01,19}) = (0, 36.19)$$

In R: $\chi^2_{0.01,19}$ =qchisq(0.99,19)= 36.19

15.41 Test for variances

In this example, we **reject** H_0 because of any of the following equivalent contrasts:

1. The observed **standardized error ratio** is:

$$w_{obs} = \frac{19(0.8462188)^2}{0.60^2} = 37.79344$$

That falls outside the acceptance region (inside the rejection zone)

2. For the upper tail

$$pvalue = 1 - F^{-1}_{\chi,19}(37.79344) = 0.006$$

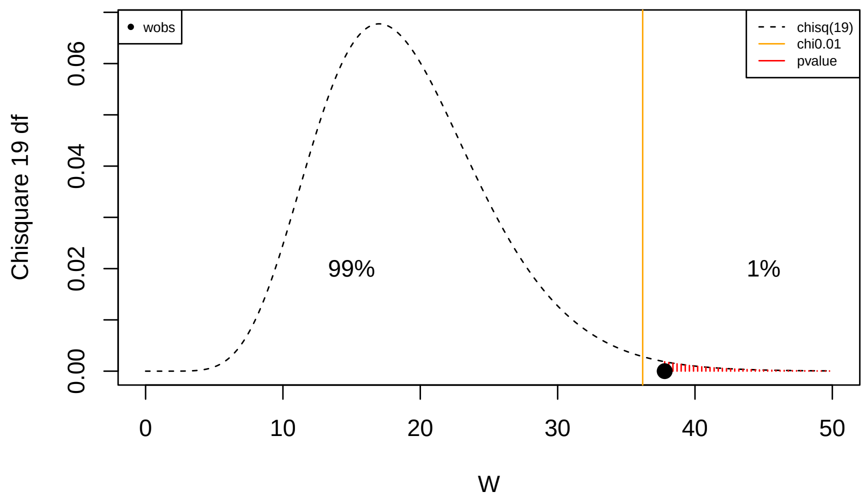
R: 1-pchisq(37.79344, 19)

is lower than $\alpha = 0.05$

Therefore we need to conclude that yes! the process is out of control.

15.42 χ^2-statistic

in R: library(EnvStats); varTest(thickness, sigma.squared = 0.6^2, alternative = “greater”)



15.43 Errors in hypothesis testing

When we infer a parameter with a statistic and then apply a criterion to decide on a hypothesis we have four possibilities

- H_0 unknown reality: **true** or **false**
- Result of H_0 testing: **negative**, **positive**

We usually aim to reject H_0 (prosecutor):

positive: expected research interest, rejection of H_0, or the status quo (rejecting innocence).

H_0	reality: true	reality: false
test: positive (negative result)	true negative	false negative
test: negative (positive result)	false positive	true positive

15.44 Errors in hypothesis testing

You take **one** PCR to test infection; H_0 you are **not infected**

- H_0 unknown reality is: **true**, **false**
- H_0 test is: **negative** (accept), **positive** (reject)

positive for infection.

H_0	not infected: true	not infected: false (infected: true)
negative (accept)	true negative: $P(accept \vert H_0 : true)$	false negative: $P(accept \vert H_0 : false)$
positive (reject)	false positive: $P(reject \vert H_0 : true)$	true positive: $P(reject \vert H_0 : false)$
sum	1	1

15.45 Errors in hypothesis testing

15.45.1 Errors are also known as

- Type I error: false positive (sending an innocent man to jail) $P(reject|H_0 : true)$
- Type II error: false negative (letting go of a criminal) $P(accept|H_0 : false)$

15.45.2 Correct contrasts are also known as

- Sensitivity: true positive (sending a criminal to jail) $P(reject|H_0 : false)$
- Specificity: true negative (letting go of an innocent man) $P(accept|H_0 : true)$

H_0	not infected: true	not infected: false (infected: true)
negative (accept)	Specificity: $P(accept \vert H_0 : true)$	Type II error: $P(accept \vert H_0 : false)$
positive (reject)	Type I error: $P(reject \vert H_0 : true)$	Specificity: $P(reject \vert H_0 : false)$
sum	1	1

15.46 Bayesian statistics

What happens if we apply the Bayes theorem to the previous table?

$$P(H_0|data) = \frac{P(data|H_0)P(H_0)}{P(data)}$$

We subvert the meaning of an event and apply it to a hypothesis.

Can we assign a probability to a hypothesis?

Bayesian interpretation of probability

- The probability is our **state of belief** on the veracity of a hypothesis given the data.

Chapter 16

Contingency tables

16.1 Objective

- χ^2 test
- Fisher exact test

16.2 Difference between proportions

For disease surveillance, we want to know if more hepatitis C patients are being observed in hospital A than in hospital B?

- We write down the status of hepatitis C of a patient who goes to **hospital A**. This is a Bernoulli variable K with outcomes (0:no hepatitis and 1:hepatitis) that has a probability mass function

$$K_A \to Bernoulli(p_A)$$

The parameter p_A is the probability of hepatitis at hospital A

- We also write down the hepatitis status of a patient who goes to **hospital B**.

$$K_B \to Bernoulli(p_B)$$

16.3 Difference between proportions

One random experiment has two outcomes: $(disease, hospital)$.

Categorical variables:

- $Disease \in \{no, yes\}$
- $Hospital \in \{A, B\}$

Repeating the experiment n times, the data for the first five repetitions look like

```
##   Hospital Disease
## 1        A     yes
## 2        A      no
## 3        B      no
## 4        A     yes
## 5        A      no
```

Question: Are $Disease$ and $Hopital$ statistically independent variables?

Let's formulate the null hypothesis.

16.4 Difference between proportions

Instead of taking random samples across hospitals, we take random samples **conditioned to** each hospital, for example:

- Hospital A included in the study total of $n_A = 200$ patients separately from hospital B that included $n_B = 400$.
- Hospital A observed 18 patients with hepatitis C, Hospital B observed 46

	Hospital: A	Hospital: B
Hepatitis (no)	$n_{no\|A} = 182$	$n_{no\|B} = 354$
Hepatitis (yes)	$n_{yes\|A} = 18$	$n_{yes\|B} = 46$
sum	$n_A = 200$	$n_B = 400$

16.5 Difference between proportions

For hospital A we have that $\bar{K_A}$ is an estimator of p_A

- $\bar{k}_A = \hat{p}_A = 18/200 = 0.09$

For hospital B we have that $\bar{K_B}$ is an estimator of p_B

- $\bar{k}_B = \hat{p}_B = 46/400 = 0.115$

These are the conditional frequencies:

	Hospital: A	Hospital: B
Hepatitis (no)	$f_{no\|A} = 0.91$	$f_{no\|B} = 0.885$
Hepatitis (yes)	$f_{yes\|A} = 0.09$	$f_{yes\|B} = 0.115$
sum	1	1

16.6 Difference between proportions

16.6.1 Null hypothesis:

- The null hypothesis (status quo) assumes that both hospitals have the same parameter (probability of hepatitis C) $H_0 : p = p_A = p_B$
- Therefore, the alternative hypothesis is that they are different $H_1 : p_A \neq p_B$

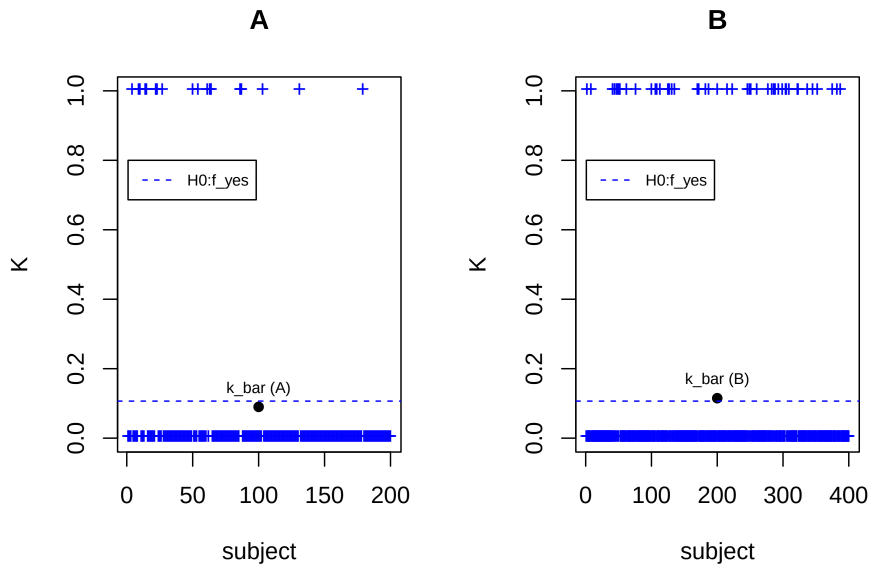
We don't know p, but we can take as the null hypothesis the value of p estimated from the two hospitals **taken together**, as if it was the same hospital:

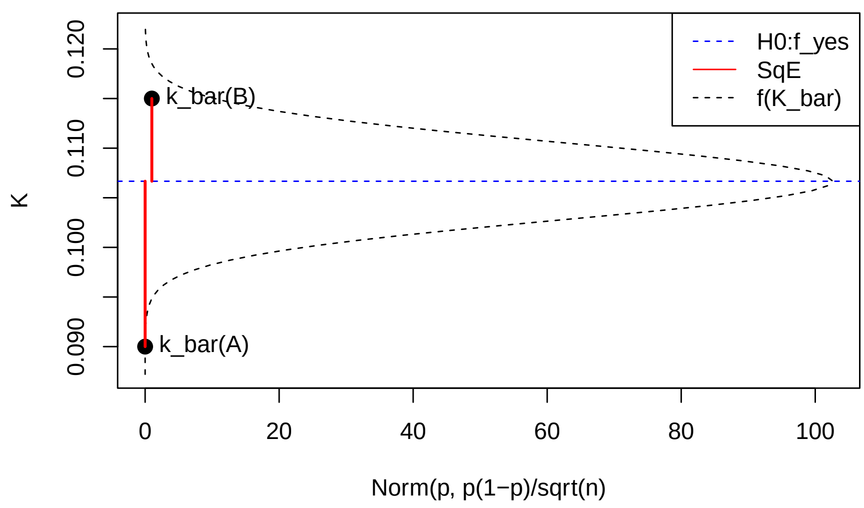
- $\hat{p} = \frac{n_{yes|A}+n_{yes|B}}{n_A+n_B} = \frac{18+46}{200+400} = f_{yes} = 0.152381$

16.7 χ^2 test

- By the CLT, the **standardized squared error** from the null hypothesis is

$$W = \frac{(\bar{K}_A - f_{yes})^2}{\frac{f_{yes}(1-f_{yes})}{n_A}} + \frac{(\bar{K}_B - f_{yes})^2}{\frac{f_{yes}(1-f_{yes})}{n_B}} \rightarrow \chi^2(1)$$


A
B
H0:f_yes
H0:f_yes
k_bar (A)
k_bar (B)
K
K
subject
subject
0.0
0.2
0.4
0.6
0.8
1.0
0
50
100
150
200
0
100
200
300
400


H0:f_yes
SqE
f(K_bar)
k_bar(B)
k_bar(A)
K
0.090
0.100
0.110
0.120
0
20
40
60
80
100
Norm(p, p(1−p)/sqrt(n))

16.8 χ^2 test

Under the **null hypothesis**, we consider that Hospital labeling, A or B, was really a random choice from the same hospital, therefore our data is rather the contingency table with overall disease marginals

	Hospital: A	Hospital: B	sum
Hepatitis (no)	$f_{no,A} = 182/600$	$f_{no,B} = 354/600$	$f_{no} = 536/600$
Hepatitis (yes)	$f_{yes,A} = 18/600$	$f_{yes,B} = 46/600$	$f_{yes} = 64/600$
sum	$f_A = 200/600$	$f_B = 400/600$	1

16.9 χ^2 test

In this context, the same **standardized squared error** of before can be re-written as:

$$W = \frac{(f_{no,A} - f_{no}f_A)^2}{f_{no}f_A} + \frac{(f_{no,B} - f_{no}f_B)^2}{f_{no}f_B} + \frac{(f_{yes,A} - f_{yes}f_A)^2}{f_{yes}f_A} + \frac{(f_{yes,B} - f_{yes}f_B)^2}{f_{yes}f_B}$$

Which are the squared differences between the

- **observed** frequencies $f_{disease,hospital}$ and
- the **expected** frequencies under statistical independence $f_{disease} * f_{hospital}$ (multiplication of marginals)

Therefore, our hypothesis test is now:

- $H_0 : p_{disease,hospital} = p_{disease} * p_{hospital}$ (Disease and Hospital are statistically **independent**)
- $H_1 : p_{disease,hospital} \neq p_{disease} * p_{hospital}$ (Disease and Hospital are statistically **dependent**)

16.10 χ^2 test

If the observed value for W is a rare error from the null hypothesis, for a χ^2 variable, we then reject the null hypothesis.

The observed value of W is

$$w_{obs} = 0.87453$$

And

$$pvalue = P(W \geq w_{obs}) = 0.3497$$

in R: chisq.test(matrix(c(182, 18, 354, 46), ncol=2), correct = FALSE)

Which is not lower than the significance level $\alpha = 0.05$ and therefore we **do not** reject H_0 and conclude:

- The frequencies of hepatitis C **are equal** between hospitals
- or, equivalently, that the frequency of hepatitis C **is independent** from hospital
- or, equivalently, that the frequency of hepatitis C **is not significantly associated** with the hospital.

16.11 Fisher's exact test

Another approach is **Fisher's exact test**

Take a ballot for each of the $N = 600$ patients of both studies into an urn:

From a population of N:

- There are $K = 64$ that have hepatitis C
- $N - K = 536$ do not have hepatitis C

Then, if we take a sample of $n = 200$ (similar in number to hospital A)

- what is the probability of observing more than 18 patients with hepatitis C, as observed in hospital A?

16.12 Fisher's exact test

- The null hypothesis (status quo) assumes that hospital A has the same parameter of both hospitals together $H_0 : p_A \geq 64/600$
- The alternative hypothesis is the parameter of hospital A is lower than the parameter for both hospitals together $H_1 : p_A < 64/600$

16.13 Hypergeometric distribution

The probability of obtaining x hepatitis C cases in a sample of n drawn from a population of N where K have hepatitis C is

$P(X = x) = P(one\, sample) \times (Number\, of\, ways\, of\, obtaining\, x)$

$$= \frac{1}{\binom{N}{n}} \binom{K}{x} \binom{N-K}{n-x}$$

where $k \in \{\max(0, n+K-N), ... \min(K, n)\}$

$$X \to Hypergeometric(N, K, n)$$

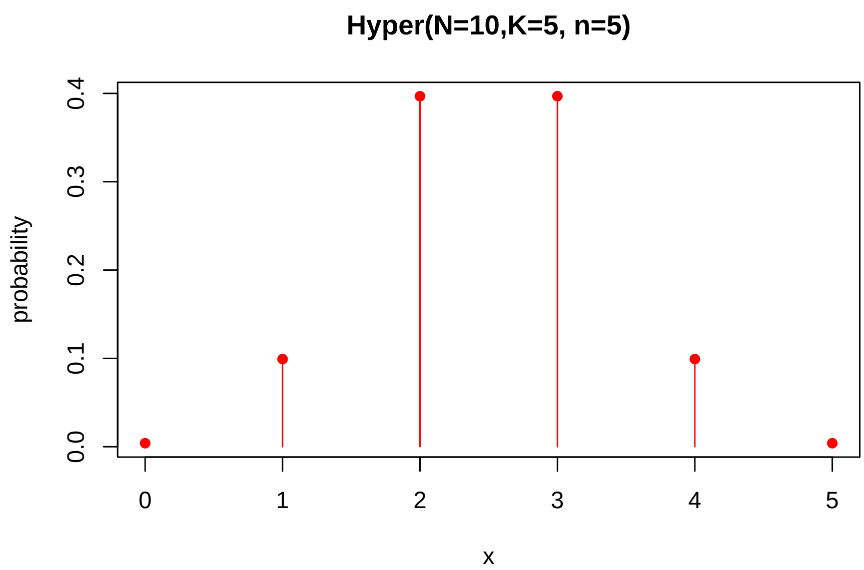
16.14 Hypergeometric distribution

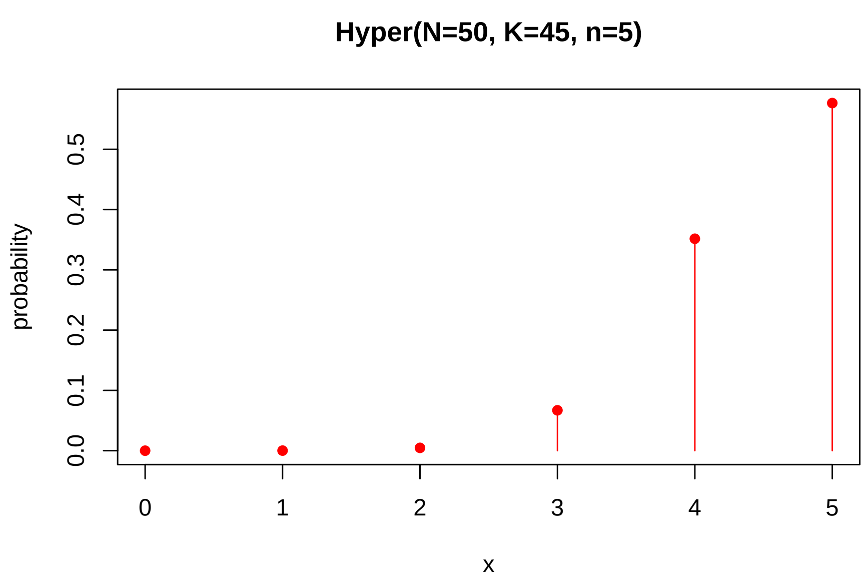
It has

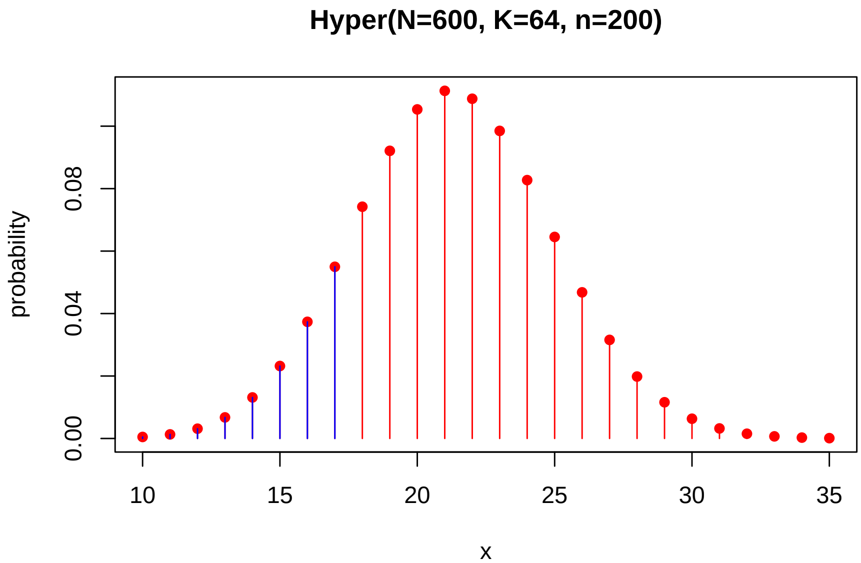
- mean: $E(X) = n\frac{K}{N} = np_0$
- variance: $V(X) = np_0(1-p_0)\frac{N-n}{N-1}$

when $p_0 = \frac{K}{N}$ is the proportion of hepatitis C in a population of size N.

16.15 Hypergeometric distribution







16.16 Fisher's exact test

If the observed value for $x = 18$ is a rare **observation** from the null hypothesis, for a hypergeometric variable, we then reject the null hypothesis.

The lower tail *pvalue* for an observation $X = 18$ is

$$pvalue = P_{hyper}(X \leq 18) = 0.2147683$$

In R: phyper(18, 64, 536, 200)

Which is not lower than the significance level $\alpha = 0.05$ and therefore we **do not** reject H_0 and conclude:

- that the frequency of hepatitis C **is not significantly associated** with the hospital.

16.17 Fisher's exact test

The odds ratio is defined as:

$$OR = \frac{f_{no,B}/f_{yes,B}}{f_{no,A}/f_{yes,A}} = 1.31$$

Gives the strength of the **observed association** between hospital and disease.

This is how we talk:

- There was **an increase** in 31% in the risk of hepatitis C for hospital B but it was not statistically significant.

This is how we compute it:

fisher.test(matrix(c(182, 18, 354, 46), ncol=2), alternative="greater")

16.18 Difference between several proportions

Now, we want to know if the frequency of hepatitis C is different across 5 difference hospitals.

	A	B	C	D	E	sum
Hepatitis (no)	182	354	375	85	90	90
Hepatitis (yes)	18	46	25	15	10	121
sum	200	400	400	100	100	1200

16.18.1 Null hypothesis:

- The null hypothesis (status quo) assumes that hospital ($i = \{A, B, C, D, E\}$) and disease ($j = \{yes, no\}$) are all independent $H_0 : p_i p_j = p_{i,j}$
- The alternative hypothesis is that **at least one** $p_i p_j \neq p_{i,j}$ is not independent.

16.19 Difference between several proportions

Writing the relative frequencies

	A	B	C	D	E	sum
Hepatitis (no)	0.1516667	0.29500000	0.31250000	0.07083333	0.075000000	0.905
Hepatitis (yes)	0.0150000	0.03833333	0.02083333	0.01250000	0.008333333	0.095
sum	0.16666667	0.33333333	0.33333333	0.08333333	0.08333333	1

We have that the **standardized squared error** from the null hypothesis can be written as:

$$W = \sum_{i=A,B,C,D,E} \sum_{j=yes,no} \frac{(f_{j,i} - f_j f_i)^2}{f_j f_i}$$

$= \frac{(0.1516667-0.16666667*0.905)^2}{0.16666667*0.905} + ... \rightarrow \chi^2(4)$

that follows a χ^2 distribution with $4 = 5 - 1$ degrees of freedom (number of hospitals -1).

16.20 Difference between several proportions

- The observed **standardized squared error** is

$$w_{obs} = 10.381$$

And

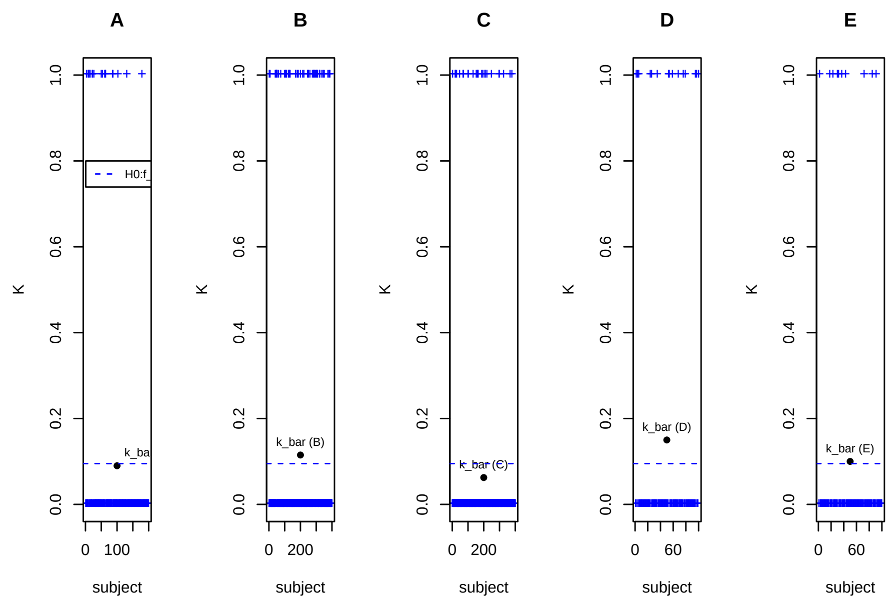
$$pvalue = P(W \geq w_{obs}) = 0.03448$$

in R: chisq.test(matrix(c(182, 18, 354, 46, 375, 25 , 85, 15, 90, 10), nrow=2))

Which is lower than the significance level $\alpha = 0.05$ and therefore we **reject** H_0 and conclude:

- that the frequency of hepatitis C **is significantly associated** with the hospital.

16.21 Difference between several proportions



Chapter 17

Mean differences between two samples

17.1 Objective

- large n: Z test
- small n with equal and unequal variances: t test

17.2 Difference between means

Let's consider an outcome of interest Y

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

we repeat the random experiment under two conditions A and B, to determine if the means between conditions change.

17.3 Difference between means

Leptin is an adipose tissue hormone that creates the sensation of satiety after eating. We want to study the serum leptin levels in obese children (PMID: 18755049) under different conditions, such as sex.

- We assume that the levels of leptin in girls have a probability density

$$Y_A \to N(\mu_A, \sigma_A^2)$$

- We assume a normal distribution of leptin in boys.

$$Y_B \to N(\mu_B, \sigma_B^2)$$

17.4 Difference between means

One random experiment has two outcomes: $(leptin, sex)$.

Continuous variable (outcome of interest)

- $leptin \in (0, 200)$

Categorical variable:

- $sex \in \{girl : A, boy : B\}$

Repeating the experiment n times, the data for the first five repetitions look like

```
##   leptin sex
## 1   37.8   B
## 2   40.1   B
## 3   48.6   A
## 4   39.0   A
## 5   43.9   A
```

Question: *Sex* and *Leptin* statistically independent variables?

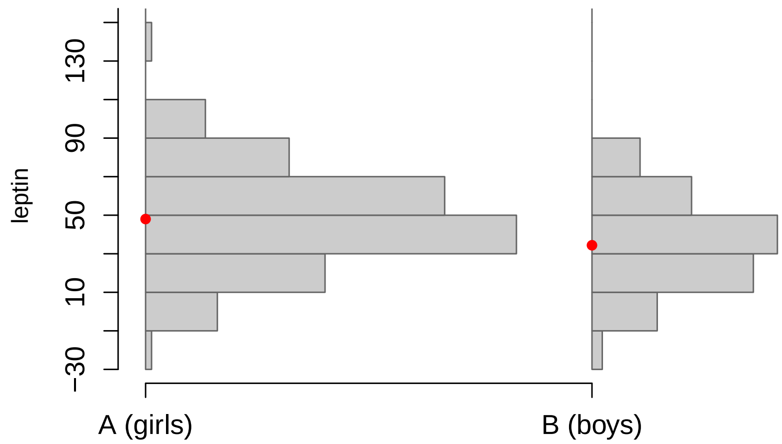
Let's formulate the null hypothesis.

17.5 Difference between means

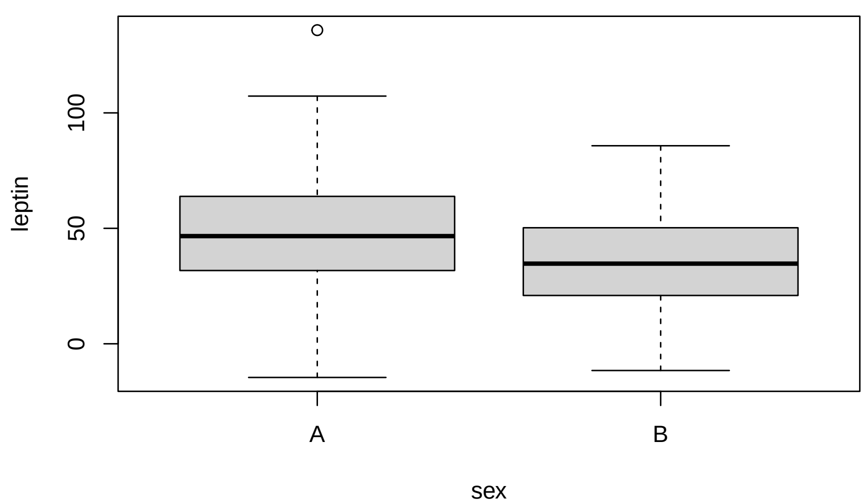
We take leptin levels **conditioned to** each sex, and observed:

- $n_A = 190$ girls had a mean of $\bar{y}_A = 48.0$ and $s = 27.1$
- $n_B = 166$ boys has a mean of $\bar{y}_B = 34.4$ and $s = 22.4$

Instead of a conditional table, we draw histograms of leptin for each condition



boxplots are also popular



17.6 Difference between means

17.6.1 Null hypothesis:

- The null hypothesis (status quo) assumes that both sexes have the same mean $H_0 : \mu_A = \mu_B$ or $H_0 : \delta = \mu_A - \mu_B = 0$
- Therefore, the alternative hypothesis is that they are different, that is $H_1 : \delta \neq 0$

We need an estimator of δ.

17.7 Estiamtor of the mean differnce

The statistic $D = \bar{Y}_A - \bar{Y}_B$ is an estimator of δ

- $E(D) = E(\bar{Y}_A - \bar{Y}_B) = \mu_A - \mu_B = \delta$ (unbiased estimator)
- $V(D) = V(\bar{Y}_A - \bar{Y}_B) = \frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}$ (consistent estimator)

More over if n_A and n_B are large (by CLT) then

$$Z = \frac{D - \delta}{\sqrt{V(D)}} = \frac{\bar{Y}_A - \bar{Y}_B - \delta}{\sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}} \rightarrow N(0, 1)$$

is a normal standard variable.

17.8 Standardized error

Then the **standardized error** from the null hypothesis ($\delta = 0$) follows a standard distribution

$$Z = \frac{\bar{Y}_A - \bar{Y}_B}{\sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}}$$

- Is our observed δ_{obs} within the acceptance region of the null hypothesis?

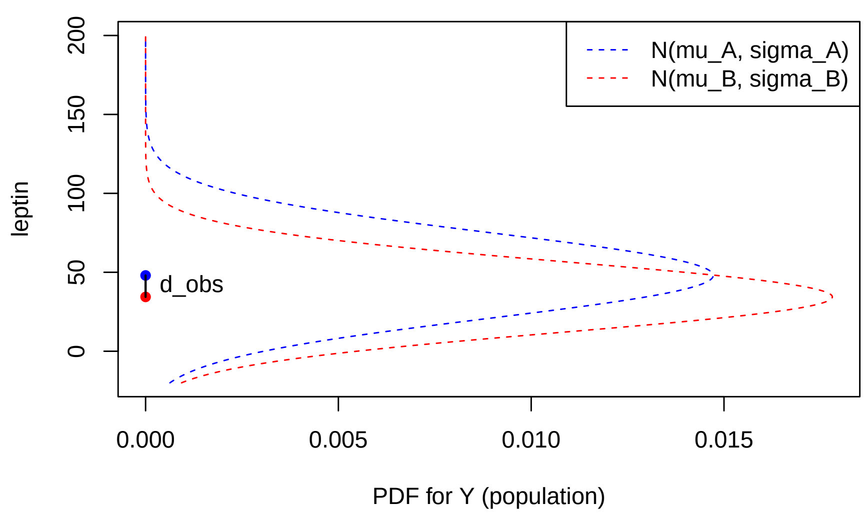
$$P(-z_{0.025} \leq Z \leq z_{0.025}) = P(-1.96 \leq Z \leq 1.96) = 0.95$$

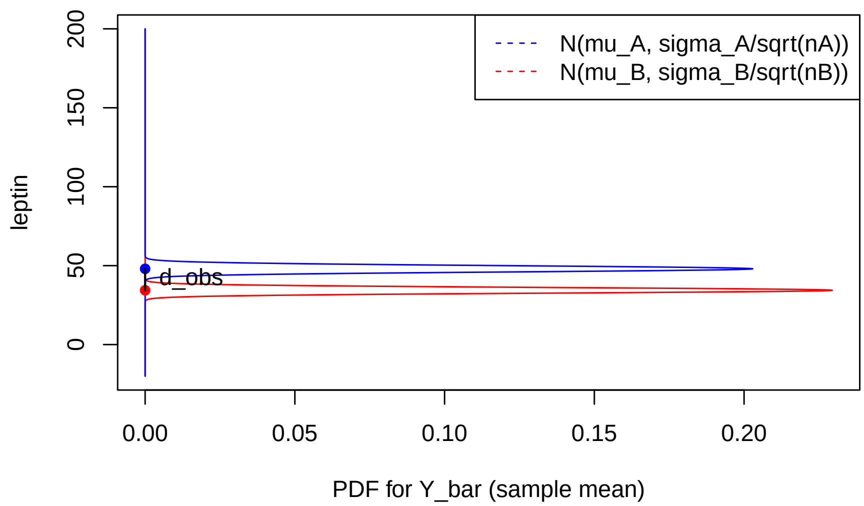
- Is the of z_{obs} of our experiment lower than $\alpha = 0.05$?

____

____

17.9 Mean comparison





17.10 Hypothesis testing

- The **observed mean difference**

$$z_{obs} = \frac{\bar{y}_A - \bar{y}_B}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}} = \frac{48 - 34.4}{\sqrt{\frac{27.1^2}{190} + \frac{22.4^2}{166}}} = 5.181952$$

is outside of the acceptance region.

- The two-tailed *pvalue*:

$$Pval = 2 * (1 - \phi(5.181952)) = 2.195757 \times 10^{-7}$$

is lower than α.

Therefore, we reject the null hypothesis that the leptin levels in obese children are equal between boys and girls.

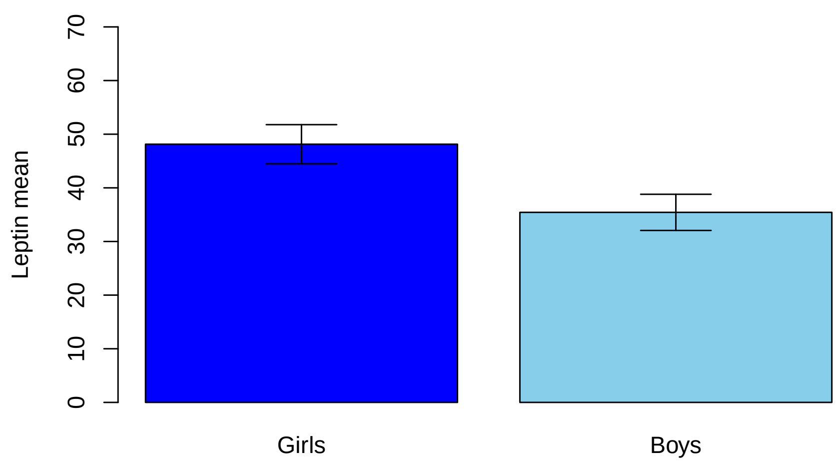
17.11 Reporting

17.11.1 Abstract

Obesity rates are different between boys and girls, suggesting that the physiopathology of the disease is different between the sexes.

In this study, we tested the hypothesis that the leptin levels in serum are different between boys and girls.

We analyzed data from 190 obese girls and 166 obese boys and found a significant difference in leptin between sexes (mean difference 13.6, $P = 2.195757 \times 10^{-7}$)



17.12 Mean difference small n

For performing the statistical test we computed

- The **observed mean difference**

$$z_{obs} = \frac{\bar{y}_A - \bar{y}_B}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}}$$

where we replaced the values of σ_A and σ_B for their estimated values s_A and s_B. The statistic D is approximately normal because $n \geq 30$ (CLT).

What happens when n is small?

17.13 Mean difference small n

In a study that wanted to test the effect of leptin in neurodevelopment, 7 male mice had their leptin gene knocked out. While 16 mice were left with normal leptin function (PMID: 30694175). An initial question was to test the effect of leptin on the body weight of the animals.

- We assume that the weight of the control animals has a probability density

$$Y_A \rightarrow N(\mu_A, \sigma_A^2)$$

- We assume a normal distribution weight for the mice with no leptin.

$$Y_B \rightarrow N(\mu_B, \sigma_B^2)$$

17.14 Difference between means

One random experiment has two outcomes: $(weight, leptin)$.

Continuous variable (outcome of interest)

- $weigth \in (20, 60)$

Categorical variable:

- $leptin \in \{control : A, knockout : B\}$

The data looks like

```
##    weight   group
## 1   27.67  Control
## 2   27.40  Control
## 3   25.77  Control
## 4   25.60  Control
## 5   25.03  Control
## 6   25.90  Control
## 7   26.67  Control
```

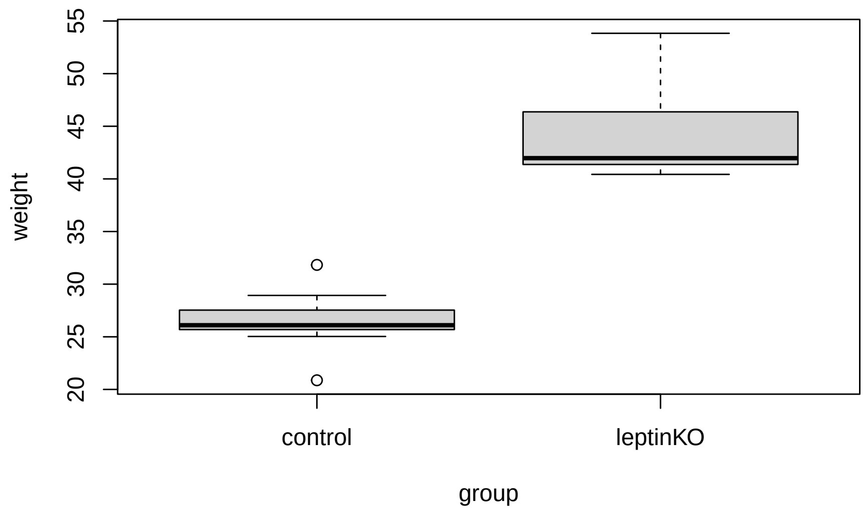
```
## 8   25.60  Control
## 9   28.93  Control
## 10  31.83  Control
## 11  25.90  Control
## 12  26.30  Control
## 13  27.90  Control
## 14  26.77  Control
## 15  25.83  Control
## 16  20.87  Control
## 17  46.57 leptinKO
## 18  40.43 leptinKO
## 19  41.97 leptinKO
## 20  41.17 leptinKO
## 21  41.57 leptinKO
## 22  46.17 leptinKO
## 23  53.83 leptinKO
```

17.15 Difference between means

We take weights **conditioned to** each lepting condition, and observed:

- $n_A = 16$ control mice had a weight mean of $\bar{y}_A = 26.49813$ and $s_A = 2.247577$
- $n_B = 7$ leptin KO mice had a weight mean of $\bar{y}_B = 44.53$ and $s_B = 4.774167$

We can draw boxplots per group



17.16 Difference between means

17.16.1 Null hypothesis:

- The null hypothesis (status quo) assumes that both mice (control, leptin KO) have the same mean $H_0 : \delta = \mu_A - \mu_B = 0$
- Therefore, the alternative hypothesis is that they are different, that is $H_1 : \delta \neq 0$

17.17 Estimator of the mean difference

The statistic $D = \bar{Y}_A - \bar{Y}_B$ is an estimator of δ

- $E(D) = \delta$ (unbiased estimator)

If Y_A and Y_B are normal variables with the same variance

$$\sigma^2 = \sigma_A^2 = \sigma_B^2$$

The **standardized error**

$$T = \frac{\bar{Y}_A - \bar{Y}_B - \delta}{\sqrt{\frac{s_p^2}{n_A} + \frac{s_p^2}{n_B}}} \to T(n_A + n_B - 2)$$

follows exactly a T-distribution with $n_A + n_B - 2$ degrees of freedom.

The **pooled variance** s_p^2, is an estimator of σ^2

$$\hat{\sigma}^2 = s_p^2 = \frac{(n_A - 1)s_A^2 + (n_B - 1)s_B^2}{n_A + n_B - 2}$$

17.18 Hypothesis testing

- Is our observed d_{obs} within the acceptance region of the null hypothesis?

$$P(-t_{0.025,21} \leq T \leq t_{0.025,21}) = P(-2.079614 \leq T \leq -2.079614) = 0.95$$

- Is the *pvalue* of the t_{obs} from our experiment lower than $\alpha = 0.05$?

17.19 Hypothesis testing

- The **observed mean difference**

$$t_{obs} = \frac{\bar{y}_A - \bar{y}_B}{\sqrt{\frac{s_p^2}{n_A} + \frac{s_p^2}{n_B}}} = \frac{26.49813 - 44.53}{\sqrt{\frac{3.18127^2}{16} + \frac{3.18127^2}{7}}} = -12.508$$

is outside of the acceptance region.

- The two-tailed *pvalue*:

$$pvalue = 2 * (1 - F_{t,21}^{-1}(12.508)) = 3.376854 \times 10^{-11}$$

is lower than α.

Therefore, the data shows a very significant increase in 18.03gr ($P = 3.376854 \times 10^{-11}$) in weight between the wild-type mice and leptin knockouts. "Absence of leptin signaling in early life alters the energy balance and predisposes the animals to obesity"

17.20 Hypothesis testing

in R

```
t.test(control, leptinKO, var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  control and leptinKO
## t = -12.508, df = 21, p-value = 3.377e-11
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -21.02992 -15.03383
## sample estimates:
## mean of x mean of y
##  26.49813  44.53000
```

17.21 Unequal variances

The boxplot suggests that the variances for each group are different.

The **standardized error**

$$T = \frac{\bar{Y}_A - \bar{Y}_B - \delta}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}} \rightarrow_{aprox} T(\nu)$$

approximately follows a t-distribution with

$$\nu = \frac{(\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B})^2}{\frac{(s_A^2/n_B)^2}{n_A - 1} + \frac{(s_B^2/n_B)^2}{n_B - 1}}$$

degrees of freedom

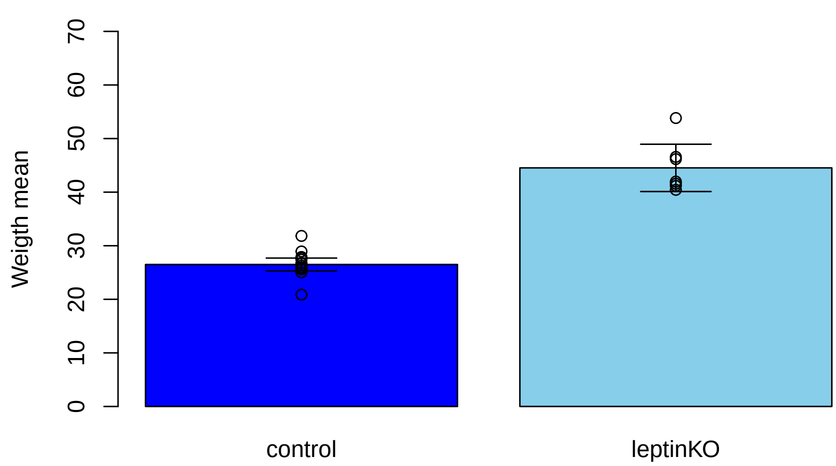
17.22 Hypothesis testing

in R

```
t.test(control, leptinKO, var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  control and leptinKO
## t = -9.541, df = 7.1929, p-value = 2.444e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -22.47665 -13.58710
## sample estimates:
## mean of x mean of y
##  26.49813  44.53000
```

Therefore, the data **still** shows a very significant increase in 18.03gr ($P = 2.444 \times 10^{-5}$) in weight between the wild-type mice and leptin knockouts (under a more appropriate test).



Chapter 18

Mean differences between several samples

18.1 Objective

- Two group ANOVA
- Several groups ANOVA
- Two-factor ANOVA
- Two-factor ANOVA with interaction

18.2 Revisiting letpin knockouts

Let's analyze the leptin experiment from a different perspective: Fisher's analysis of variance.

```
##    weight   group
## 1   27.67  Control
## 2   27.40  Control
## 3   25.77  Control
## 4   25.60  Control
## 5   25.03  Control
## 6   25.90  Control
## 7   26.67  Control
## 8   25.60  Control
## 9   28.93  Control
## 10  31.83  Control
```

```
## 11  25.90  Control
## 12  26.30  Control
## 13  27.90  Control
## 14  26.77  Control
## 15  25.83  Control
## 16  20.87  Control
## 17  46.57 leptinKO
## 18  40.43 leptinKO
## 19  41.97 leptinKO
## 20  41.17 leptinKO
## 21  41.57 leptinKO
## 22  46.17 leptinKO
## 23  53.83 leptinKO
```

18.3 Null hypothesis

If the **null hypothesis is true**, then mice with and without leptin are **identical** and then split into groups are two random samples of size A and B from the same population.

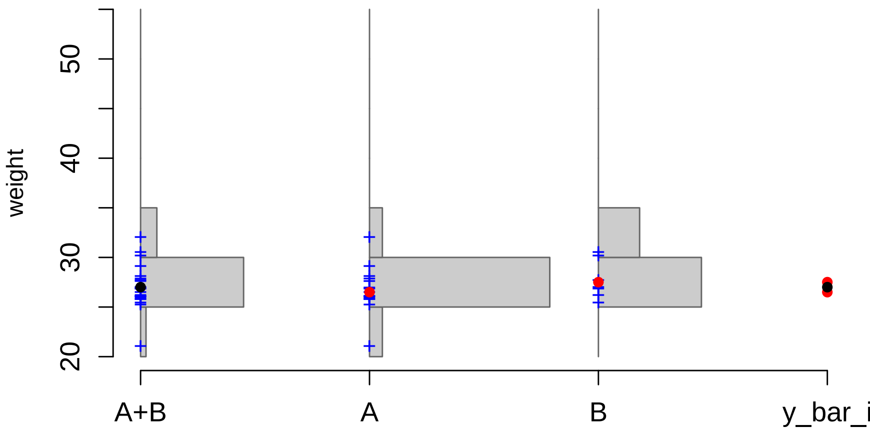
- The overall variance is equal to the **within group** variances

$$\sigma^2 = \sigma_A^2 = \sigma_B^2$$

- There is no difference between means

$$\delta = (\mu_A - \mu) - (\mu_B - \mu) = 0$$

This is how it would look like



18.4 Analisis of variance

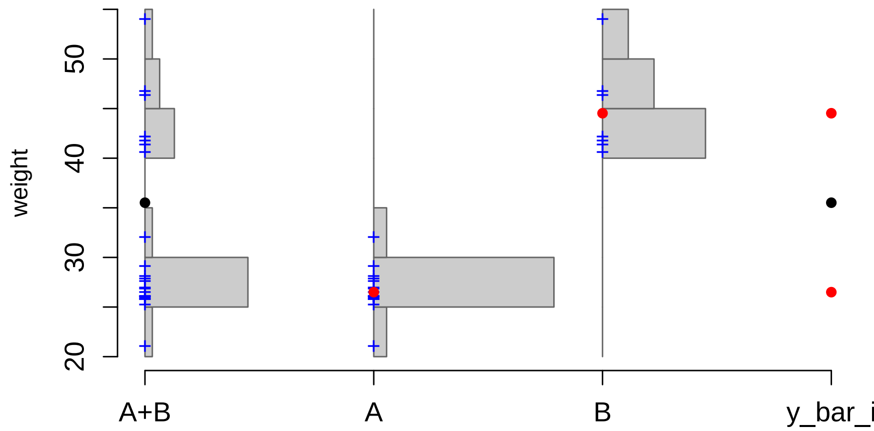
If the **null hypothesis is not true**, then the mean weight of the mice is **different** with and without leptin and the split into groups are two random samples of size A and B from **different** populations.

- The overall variance is greater than the within-group variances

$$\sigma^2 > \sigma_A^2 = \sigma_B^2$$

- The group means are different

$$\delta = (\mu_A - \mu) - (\mu_B - \mu) \neq 0$$



18.5 Linear model

Under the **alternative hypothesis**, let's consider the observations of mice weight conditioned to the leptin groups

Y_{Aj} for $i = 1...16$. For example: $Y_{A5} = 25.03$,

```
##    weight group
## 1   27.67     A
## 2   27.40     A
## 3   25.77     A
## 4   25.60     A
## 5   25.03     A
## 6   25.90     A
## 7   26.67     A
## 8   25.60     A
## 9   28.93     A
## 10  31.83     A
## 11  25.90     A
## 12  26.30     A
## 13  27.90     A
## 14  26.77     A
```

```
## 15  25.83     A
## 16  20.87     A
```

Y_{Bj} for $i = 1...7$. For example: $Y_{B2} = 40.43$,

```
##   weight group
## 1  46.57     B
## 2  40.43     B
## 3  41.97     B
## 4  41.17     B
## 5  41.57     B
## 6  46.17     B
## 7  53.83     B
```

18.6 Linear model

Let's assume that for all observations we can extract a **random error**

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$$

Fixed parameters:

- μ is the overall mean
- α_i is the deviation of group i to the overall mean: $i \in (A, B)$ and $\alpha_A = \mu_A - \mu$, $\alpha_B = \mu_B - \mu$.
- $j \in 1, ...n$ (all groups have the same number of observations $n_A = n_B = n$ for simplicity with no loss of generality)

Random error:

- ϵ_{ij} is a **random variable** with $E(\epsilon_{ij}) = 0$, $V(\epsilon_{ij}) = \sigma^2$

Then

- $E(Y_{ij}) =$
$$E(Y|i) = \mu + \alpha_i$$

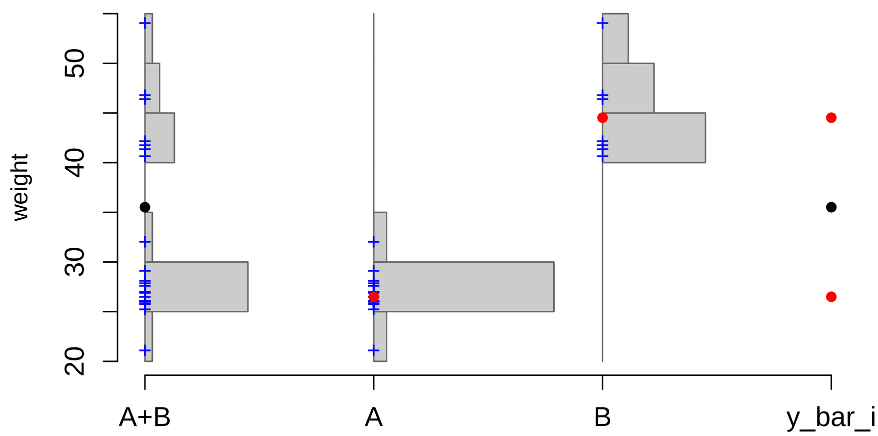
for instance: $E(Y|A) = \mu_A = \mu + \alpha_A$

- $V(Y_{ij}) = \sigma^2$

18.7 Variance components

The squared deviations of the observations to the overall average is

$$\sum_{i=A,B} \sum_{j=1}^{n} (Y_{ij} - \bar{Y})^2 = \sum_{i=A,B} \sum_{j=1}^{n} (Y_{ij} - \bar{Y}_i)^2 + n \sum_{i=A,B} (\bar{Y}_i - \bar{Y})^2$$



18.8 Variance components

Let's look at each term

$$\sum_{i=A,B} \sum_{j=1}^{n} (Y_{ij} - \bar{Y})^2 = \sum_{i=A,B} \sum_{j=1}^{n} (Y_{ij} - \bar{Y}_i)^2 + n \sum_{i=A,B} (\bar{Y}_i - \bar{Y})^2$$

- Sum of squares (total)

$$SS_T = \sum_{i=A,B} \sum_{j=1}^{n} (Y_{ij} - \bar{Y})^2$$

- Sum of squares (error)

$$SSE = \sum_{i=A,B} \sum_{j=1}^{n} (Y_{ij} - \bar{Y}_i)^2$$

- Sum of squares (treatment)

$$SS_{treatment} = n \sum_{i=A,B} (\bar{Y}_i - \bar{Y})^2$$

18.9 Variance components

- The **mean square error** (MSE)

$MSE = \frac{1}{2n-2} SSE$

$$= \frac{1}{2n-2} \sum_{i=A,B} \sum_{j=1}^{n} (Y_{ij} - \bar{Y}_i)^2 = \frac{(n-1)S_A^2 + (n-1)S_B^2}{2n-2}$$

is the **pooled variance estimator**

$$E(MSE) = \sigma^2$$

- The **mean square of treatments** (MST)

$MST = \frac{1}{2-1} SS_{treatment}$

$$= n \sum_{i=A,B} (\bar{Y}_i - \bar{Y})^2$$

is a **biased estimator** of the variance

$$E(MST) = \sigma^2 + n(\alpha_A^2 + \alpha_B^2)$$

18.10 Linear model

In the linear model for the weight of mice:

$$Y_{ij} = \mu + \alpha_i + E_{ij}$$

The null hypothesis is $H_0 = \mu_A = \mu_B = \mu$ therefore $H_0 : \alpha_i = 0$

- If the null hypothesis is true both MSE and MST are estimators of σ^2.
- If $Y_i \to N(\mu_i, \sigma_i^2)$ then $MSE \to \chi^2(2n-2)$ and $MST \to \chi^2(n-1)$

and the ratio of squares

$$\frac{MST}{MSE} \to F(2n-2, n-1)$$

follows a F distribution with $2n-2$ and $n-1$ degrees of freedom.

18.11 ANOVA

- H_0: observed values of $\frac{MST}{MSE}$ near 1 suggest that the means between groups **do not** differ.
- H_1: observed values of $\frac{MST}{MSE}$ far from 1 suggest that the means between groups **differ**.

$$\frac{MST}{MSE}_{obs} = \frac{(\bar{Y}_A - \bar{Y}_B)^2}{\frac{s_p^2}{n}} = t_{obs}^2$$

18.12 ANOVA

We have assumed the same number of observations in each group but when there are two groups the results above holds

$$f_{obs} = t_{obs}^2 = (-12.508)^2 = 156.45$$

The upper tailed pvalue for f_{obs} is

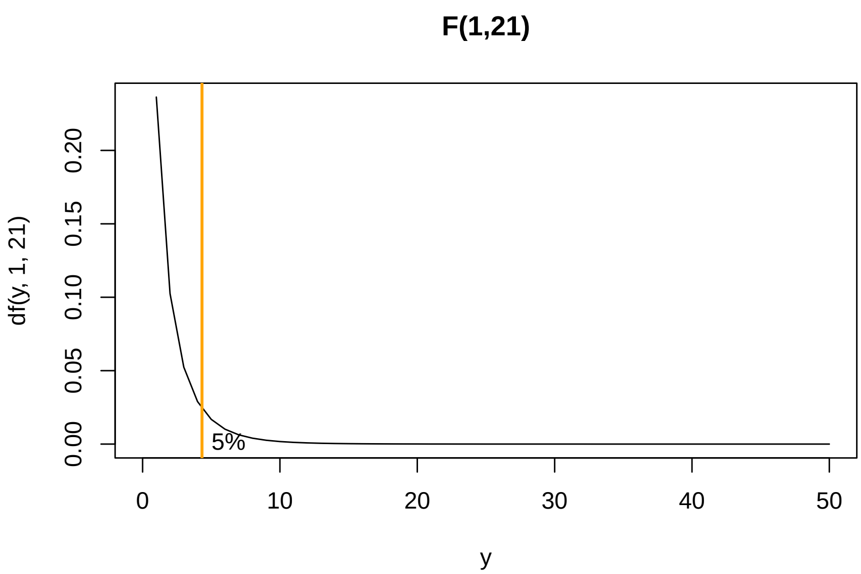
$$pvalue = 1 - F_{F,1,21}^{-1}(156.45) = 3.377 \times 10^{-11}$$

in R: 1-pf(156.45, 1,21)

Which is the same as the one obtained with the t-test with equal variances.

18.13 ANOVA

```
## Analysis of Variance Table
##
## Response: weight
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## group      1 1583.33 1583.33  156.45 3.377e-11 ***
## Residuals 21  212.53   10.12
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



18.14 ANOVA several groups

The ANOVA approach allows for the analysis of many groups.

Consider the **linear model**

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$$

with **Random error**:

- ϵ_{ij} is a **random variable** with $E(\epsilon_{ij}) = 0$, $V(\epsilon_{ij}) = \sigma^2$

and k groups.

- $\alpha_i, i \in \{1...k\}$ such that $\sum_i \alpha_i = 0$ are the deviations of the group means to the the overall mean.

$$E(Y|i) = \mu + \alpha_i$$

18.15 ANOVA several groups

- $H_0 : \alpha_1 = \alpha_2, ... = \alpha_k = 0$ There are no difference between group means

Then, observed values of $\frac{MST}{MSE}$ near 1 suggest that the means between groups **do not** differ.

- H_1 at least one α_i is different

Then, observed values of $\frac{MST}{MSE}$ far from 1 suggest that the means between groups **differ**.

$$\frac{MST}{MSE} \to F(k-1, k(n-1))$$

- where MSE is the estimated variance within groups

$$MSE = \frac{1}{k(n-1)} \sum_{i=1}^{k} \sum_{j=1}^{n} (Y_{ij} - \bar{Y}_i)^2$$

- MST is the estimated variance between groups

$$MST = \frac{n}{k-1} \sum_{i=A,B} (\bar{Y}_i - \bar{Y})^2$$

18.16 ANOVA several groups

In a study that wanted to test the effect of leptin in neurodevelopment, 7 male mice had their leptin gene knocked out. While 16 mice were left with normal leptin function. In a third group, 10 mice with knocked out leptin were injected leptin (PMID: 30694175). An initial question was to test the effect of the leptin group on the body weight of the animals.

- We assume that the weight of the control animals has a probability density

$$Y_A \rightarrow N(\mu_A, \sigma_A^2)$$

- We assume a normal distribution weight for the mice with no leptin.

$$Y_B \rightarrow N(\mu_B, \sigma_B^2)$$

- We assume a normal distribution weight for the mice with no leptin gene but were injected leptin.

$$Y_C \rightarrow N(\mu_C, \sigma_C^2)$$

18.17 Difference between means

One random experiment has two outcomes: $(weight, leptin)$.

Continuous variable (outcome of interest)

- $weigth \in (20, 60)$

Categorical variable:

- $leptin \in \{control : A, knockout : B, replace : C\}$

The data looks like

```
##    weight   group
## 1   27.67  Control
## 2   27.40  Control
## 3   25.77  Control
## 4   25.60  Control
## 5   25.03  Control
## 6   25.90  Control
## 7   26.67  Control
## 8   25.60  Control
## 9   28.93  Control
## 10  31.83  Control
## 11  25.90  Control
## 12  26.30  Control
## 13  27.90  Control
## 14  26.77  Control
```

```
## 15  25.83  Control
## 16  20.87  Control
## 17  46.57 leptinKO
## 18  40.43 leptinKO
## 19  41.97 leptinKO
## 20  41.17 leptinKO
## 21  41.57 leptinKO
## 22  46.17 leptinKO
## 23  53.83 leptinKO
## 24  24.33   KOplus
## 25  22.37   KOplus
## 26  26.10   KOplus
## 27  17.50   KOplus
## 28  35.17   KOplus
## 29  25.97   KOplus
## 30  27.67   KOplus
## 31  23.37   KOplus
## 32  31.83   KOplus
## 33  22.37   KOplus
```

18.18 Difference between means

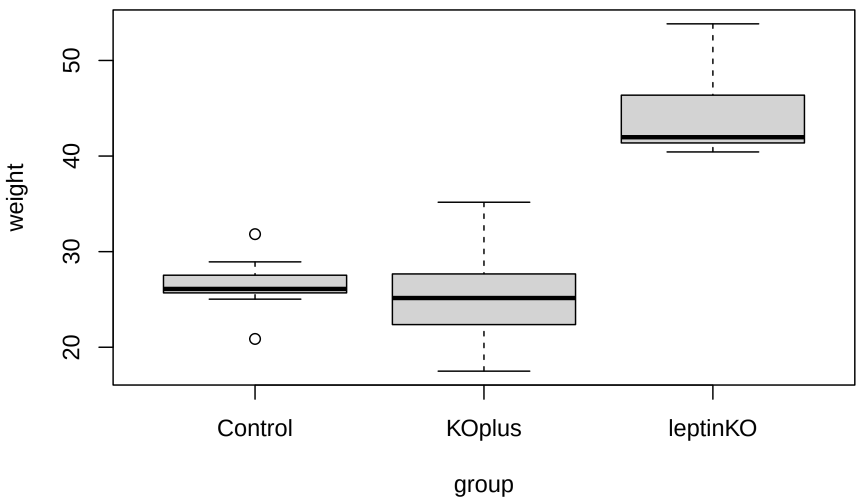
We take weights **conditioned to** each leptin condition, and observed:

- $n_A = 16$ control mice had a weight mean of $\bar{y}_A = 26.49813$ and $s_A = 2.247577$

- $n_B = 10$ leptin KO mice with leptin replacement had a weight mean of $\bar{y}_B = 25.668$ and $s_B = 5.034161$ We can draw boxplots per group

- $n_C = 7$ leptin KO mice had a weight mean of $\bar{y}_C = 44.53$ and $s_C = 4.774167$

We can see the differences in distributions with a boxplot



18.19 ANOVA several groups

The observed value of the statistics is

$$\frac{MST}{MSE} = 63.373$$

which is still a **rare** observation of an F distribution with 2 and $30 = (16+7+10)/3-1$ degrees of freedom

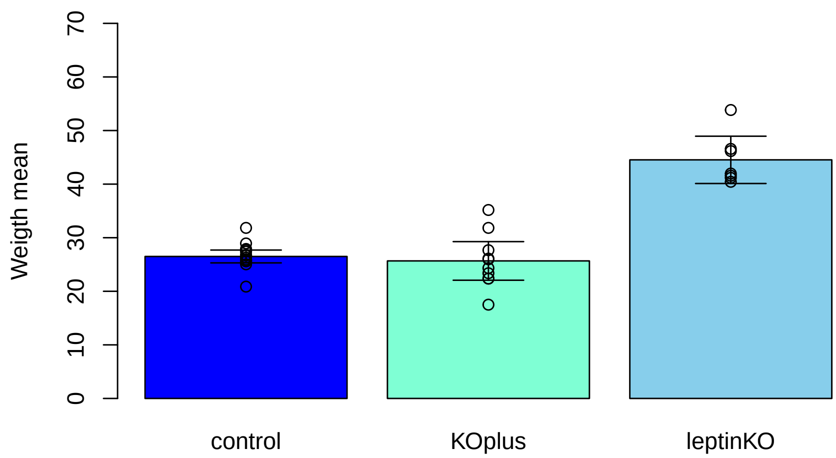
$$pvalue = 1 - F^{-1}_{F,2,30}(63.373) = 1.694 \times 10^{-11}$$

Suggesting significant differences in at least one group mean.

18.20 ANOVA several groups

```
## Analysis of Variance Table
```

```
##
## Response: weight
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## group      2 1861.55  930.77  63.373 1.694e-11 ***
## Residuals 30  440.62   14.69
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



We observed a significant difference between groups (ANOVA test $F(2, 30) = 63.373, P = 1.69 \times 10^{-11}$), due to the higher gain in weight of the knockout mice. Note that knocked-out mice with replacement recovered wild-type weight (t-test difference between means -0.83, $P = 0.63$)

18.21 ANOVA two factor

The ANOVA approach allows for the analysis of the joint effect of **two random** variables.

Let us include an additional sample of female mice in the leptin study and ask: Is the change in weight across different leptin groups that we observed in male mice the same in female mice?

- Is there an effect of **sex** on the weight of the mice?

- Is that effect different between leptin groups?

18.22 Two factor

One random experiment has three outcomes: $(weight, leptin, sex)$.

Continuous variable (outcome of interest)

- $weigth \in (20, 60)$

Categorical variable:

- $leptin \in \{control : A, knockout : B\}$

Categorical variable:

- $sex \in \{male : a, female : b\}$

The data looks like

```
##    weight    group sex
## 1   27.67  Control   M
## 2   27.40  Control   M
## 3   25.77  Control   M
## 4   25.60  Control   M
## 5   25.03  Control   M
## 6   25.90  Control   M
## 7   26.67  Control   M
## 8   25.60  Control   M
## 9   28.93  Control   M
## 10  31.83  Control   M
## 11  25.90  Control   M
## 12  26.30  Control   M
## 13  27.90  Control   M
## 14  26.77  Control   M
## 15  25.83  Control   M
## 16  20.87  Control   M
## 17  46.57 leptinKO   M
## 18  40.43 leptinKO   M
## 19  41.97 leptinKO   M
## 20  41.17 leptinKO   M
## 21  41.57 leptinKO   M
## 22  46.17 leptinKO   M
## 23  53.83 leptinKO   M
## 34  22.30  Control   F
## 35  23.30  Control   F
```

```
## 36  23.10  Control   F
## 37  22.20  Control   F
## 38  22.30  Control   F
## 39  19.90  Control   F
## 40  22.20  Control   F
## 41  20.60  Control   F
## 42  22.00  Control   F
## 43  20.40  Control   F
## 44  21.00  Control   F
## 45  22.00  Control   F
## 46  23.20  Control   F
## 47  22.00  Control   F
## 48  23.30  Control   F
## 49  20.60  Control   F
## 50  22.80  Control   F
## 51  19.50  Control   F
## 52  20.80  Control   F
## 53  20.20  Control   F
## 54  20.00  Control   F
## 55  20.80  Control   F
## 56  17.60  Control   F
## 57  16.30  Control   F
## 58  65.80 leptinKO   F
## 59  51.40 leptinKO   F
## 60  54.60 leptinKO   F
## 61  48.30 leptinKO   F
## 62  50.60 leptinKO   F
## 63  48.90 leptinKO   F
## 64  51.20 leptinKO   F
## 65  46.80 leptinKO   F
## 66  50.90 leptinKO   F
## 67  42.70 leptinKO   F
```

18.23 Difference between means

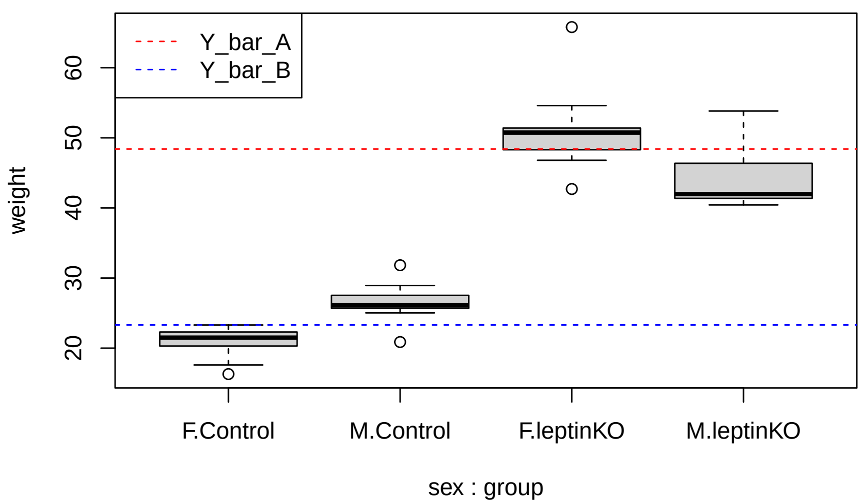
We take weights **conditioned to** each leptin by sex condition, and observed:

- $n_{Aa} = 16$ control **male** mice had a weight mean of $\bar{y}_{Aa} = 26.49813$ and $s_{Aa} = 2.247577$
- $n_{Ba} = 7$ leptin KO **male** mice had a weight mean of $\bar{y}_{Ba} = 44.53$ and $s_{Ba} = 4.774167$

- $n_{Ab} = 24$ control **female** mice had a weight mean of $\bar{y}_{Ab} = 21.18333$ and $s_{Ab} = 1.757386$
- $n_{Bb} = 10$ leptin KO **female** mice had a weight mean of $\bar{y}_{Bb} = 51.12$ and $s_{Bb} = 6.059483$

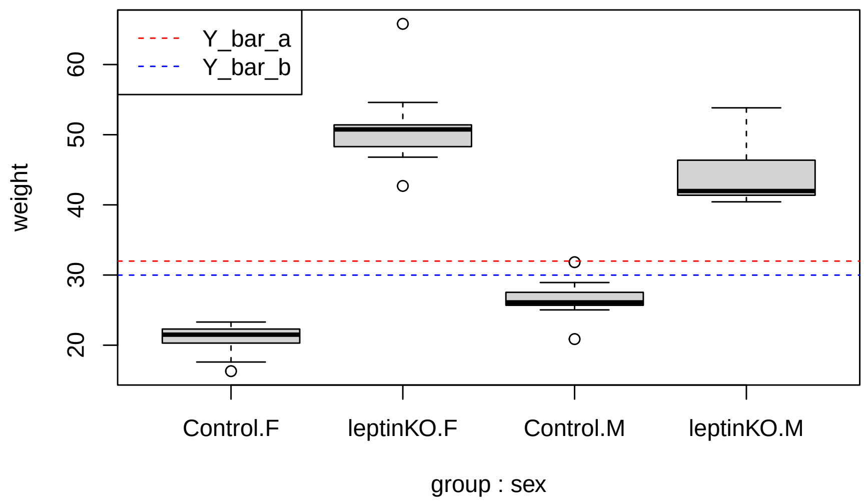
18.24 Difference between means

We draw the boxplot, grouping leptin conditions, and observe a strong overall effect of leptin



18.25 Difference between means

If we draw the boxplot grouping by sex, we observe that sex does not have such a strong effect on weight



18.26 ANOVA two factor

The ANOVA approach allows for the analysis of two factors (each with many groups-levels).

Consider the **linear model**

$$Y_{ijr} = \mu + \alpha_i + \beta_j + \epsilon_{ijr}$$

with **Random error**:

- ϵ_{ijr} is a **random variable** with $E(\epsilon_{ijr}) = 0$, $V(\epsilon_{ijr}) = \sigma^2$

and k groups for factor 1 (leptin).

- $\alpha_i, i \in \{1...k\}$ such that $\sum_i \alpha_i = 0$ are the deviations of the group means to the the overall mean.

$$E(Y|i) = \mu + \alpha_i$$

and m groups for factor 2 (sex).

$$E(Y|j) = \mu + \beta_j$$

Each experiment is defined by a given group in factor 1 and a given group in factor 2 (e.g $(control, male)$) and repeated n times (for simplicity but not loss of generality)

18.27 Variance components

The squared deviations of the observations to the overall average can be decomposed into their variations within each experiment (SSE) and the variations between factor 1 (SS_{Fac1}) and factor 2 (SS_{Fac2}).

$$SS_T = SSE + SS_{Fac1} + SS_{Fac2}$$

That defines F statistics

$$\frac{MS1}{MSE} \to F(k-1, (m-1)(nk-1))$$

and

$$\frac{MS2}{MSE} \to F(n-1, (m-1)(nk-1))$$

18.28 ANOVA several groups

ANOVA allows testing two null hypothesis

First

- $H_0 : \alpha_1 = \alpha_2, ... = \alpha_k = 0$ There are no difference between group means for the fist factor
- H_1 at least one α_i is different

Then, observed values of $\frac{MS1}{MSE}$ far from 1 suggest that the means between groups of the first factor **differ**.

Second

- $H_0 : \beta_1 = \beta_2, ... = \beta_k$ There are no difference between group means for the second factor

- H_1 at least one β_i is different

Then, observed values of $\frac{MS1}{MSE}$ far from 1 suggest that the means between groups of the second factor **differ**.

```
## Analysis of Variance Table
##
## Response: weight
##           Df Sum Sq Mean Sq  F value Pr(>F)
## group      1 7514.2  7514.2 396.8721 <2e-16 ***
## sex        1   41.6    41.6   2.1968 0.1441
## Residuals 54 1022.4    18.9
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As we observed from the boxplots, the statistical inference confirms that there are significant differences in weight between leptin conditions but no significant differences between sexes.

18.29 ANOVA interaction

From the **linear model**

$$Y_{ijr} = \mu + \alpha_i + \beta_j + \epsilon_{ijr}$$

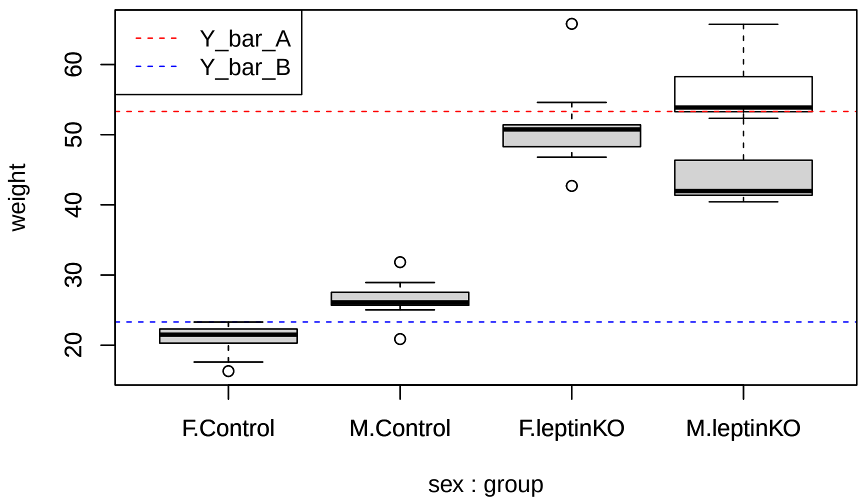
We have that the expected value of any observation

$$E(Y|i,j) = \mu + \alpha_i + \beta_j$$

is the sum of the overall mean and the means of each factor (the factors add together).

18.30 ANOVA interaction

This is only true if he had observed for the condition $(male, leptinKO)$ something like (white box)



18.31 ANOVA interaction

The apparent less-than-expected gain in weight of leptin KO males seems like a specific interaction between these two conditions.

We then formulate the **linear model** with an **interaction term**

$$Y_{ijr} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijr}$$

Such that each observation in each experiment have a specific contribution from the conditions in each factor

$$E(Y|ij) = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}$$

18.32 ANOVA interaction

The squared deviations of the observations to the overall average can be decomposed into their variations within each experiment (SSE) and the variations between factor 1 (SS_{Fac1}), factor 2 (SS_{Fac2}) and their interaction terms (SS_{Int}).

$$SS_T = SSE + SS_{Fac1} + SS_{Fac2} + SS_{Int}$$

That defines F statistics

$$\frac{MS1}{MSE} \rightarrow F(k-1, (m-1)(nk-1))$$

$$\frac{MS2}{MSE} \rightarrow F(n-1, (m-1)(nk-1))$$

and

$$\frac{MSI}{MSE} \rightarrow F((n-1)(k-1), (m-1)(nk-1))$$

18.33 ANOVA interaction

ANOVA allows testing three null hypothesis

First

- $H_0 : \alpha_1 = \alpha_2, ... = \alpha_k = 0$ There are no difference between group means for the fist factor

Second

- $H_0 : \beta_1 = \beta_2, ... = \beta_k = 0$ There are no difference between group means for the second factor

Third

- $H_0 : (\alpha\beta)_{ij} = 0$ There is no difference between group means for the second factor

And the alternatives are that at least one of the terms is different from 0

18.34 ANOVA interaction

```
## Analysis of Variance Table
##
## Response: weight
##           Df Sum Sq Mean Sq  F value    Pr(>F)
## group      2 7939.7  3969.8 299.5748 < 2.2e-16 ***
## sex        1   31.0    31.0   2.3422    0.1302
## group:sex  2  419.0   209.5  15.8095 1.975e-06 ***
## Residuals 73  967.4    13.3
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As we observed from the boxplots, the statistical inference confirms that there are significant differences in weight between leptin conditions, no significant differences between sexes, and significant interactions between sex and leptin condition. In particular, the effect of leptin knockout in weight is higher in females than males, opposite to what was observed in controls.

Chapter 19

Regression and correlation

19.1 Objective

- Bivariate normal distribution
- Correlation
- Regression
- Multiple regression

19.2 Regression

Leptin is a hormone produced by adipose tissue. We want to study the serum leptin levels in the adult population (PMID: 23628382, GEO:GSE45987) under a **continuous** condition, such as the amount in Kg of body fat.

- We assume that the levels of leptin have a probability density

$$Y \rightarrow N(\mu_y, \sigma_x^2)$$

19.3 Regression

One random experiment has two outcomes: $(leptin, fatmass)$.

Continuous variable (outcome of interest)

- $lepting \in (0, 5)$

Continuous variable:

- $fatmass \in (20, 80)$

Repeating the experiment n times, the data for the first five repetitions look like

```
##     leptin fatmass
## 1 3.355677  45.721
## 2 2.272126  43.895
## 3 1.071584  47.871
## 4 3.921082  65.801
## 5 1.536867  56.644
## 6 1.177115  56.355
```

Question: $fatmass$ and $leptin$ statistically independent variables?

19.4 Continous variation of the mean

Leptin levels are continuous

$$Y \rightarrow N(\mu_y, \sigma_y^2)$$

But fat mass is also a continuous variable

$$X \rightarrow N(\mu_x, \sigma_x^2)$$

To formulate the null hypothesis, we want to condition $leptin$ on $fatmass$.

19.5 Normal bivariate

A random 2D vector of two random variables (Y, X) follows a bivariate normal distribution if its probability density is

$$f(y,x) = \frac{1}{2\pi\sigma_y\sigma_x\sqrt{1-\rho^2}} e^{\frac{(y-\mu_y)^2}{\sigma_y^2} - \frac{2\rho(y-\mu_y)(x-\mu_x)}{\sigma_y\sigma_x} + \frac{(x-\mu_x)^2}{\sigma_x^2}}$$

With parameters $\mu_y, \mu_x, \sigma_y^2, \sigma_x^2, \rho$.

On the marginals:

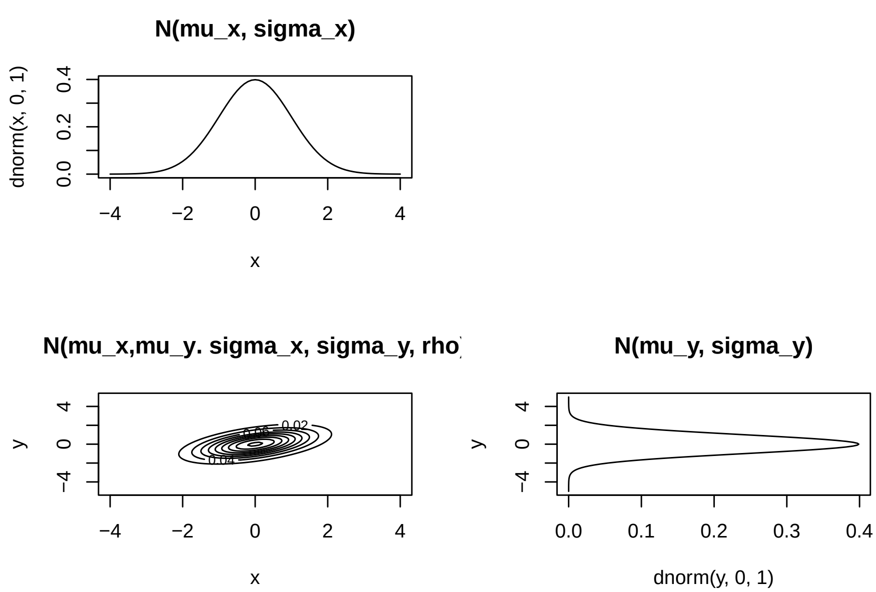
- $E(Y) = \mu_y$, $V(Y) = \sigma_y^2$
- $E(X) = \mu_x$, $V(X) = \sigma_x^2$

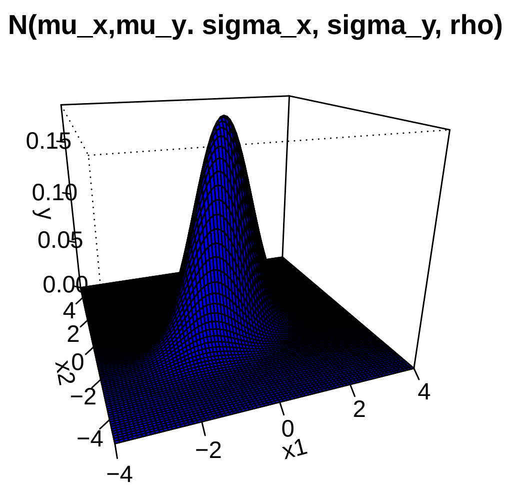
On the **correlation**:

- $\frac{E[(Y-\mu_y)(X-\mu_x)]}{\sigma_y \sigma_x} = \rho$

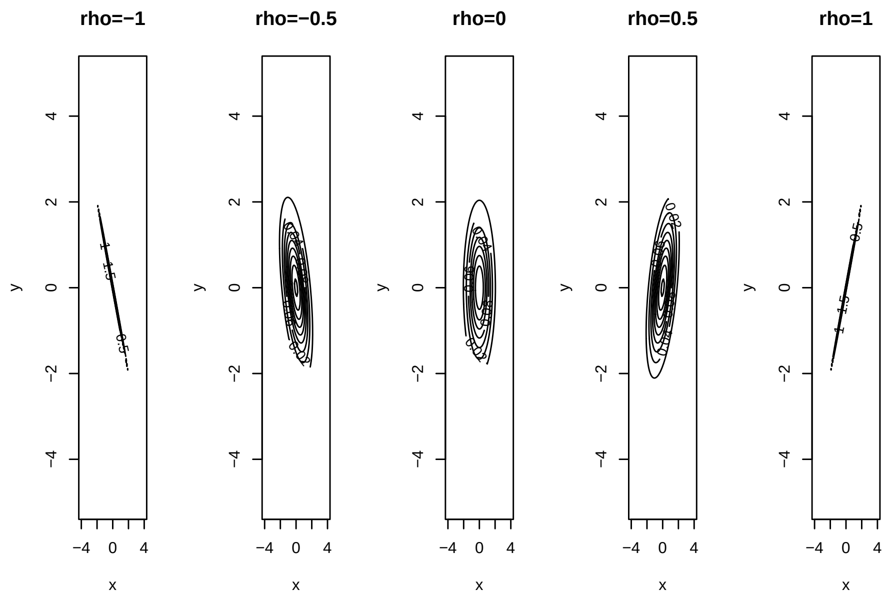
ρ is called the correlation coefficient

19.6 Normal bivariate




N(mu_x,mu_y. sigma_x, sigma_y, rho)
0.15
0.10
0.05
0.00
y
4
2
0
−2
−4
x2
−4
−2
0
2
4
x1

19.7 Normal bivariate



19.8 Estimators

If we formulate the likelihood function

$$L = \Pi_{i=1}^{n} f(y_i, x_i; \mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho)$$

Maximizing the function, we can obtain estimators for each of the parameters, resulting in

Estimators for

- μ_y: $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} y_i$
- μ_x: $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$
- σ_y^2: $S_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$
- σ_x^2: $S_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$
- ρ:

$$R = \frac{\sum_{i=0}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

19.9 Correlation coefficient

The transformation of R (Fisher's z transformation) has a distribution

$$\frac{1}{2}\ln(\frac{1+R}{1-R}) \rightarrow_{aprox} N(\frac{1}{2}\ln(\frac{1+\rho}{1-\rho}), \frac{1}{n-3})$$

- If R is 0 there is no direction in the relationship between y and x (the probability distribution is concentric)
- If R is near 1 most of the observations fall close to a line

19.10 Hypothesis

Null hypothesis:

- Y and X are statistically independent, therefore $f(y,x) = f(x)f(y)$ and $H_0 : \rho = 0$

Alternative hypothesis:

- Y and X are statistically dependent, therefore $H_1 : \rho \neq 0$

We, therefore, use the statistic R to test whether there is a dependency between y and x

19.11 Regression coefficient

The observed value of R is

$$r = \hat{\rho} = \frac{\sum_{i=0}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

For our data

- $r_{obs} = -0.2766492$
- The transformed value $ln((1 - 0.2766492)/(1 - 0.2766492))/2 = -0.2840499$ and it is rare under the distribution of R

$$pvalue = 2(1 - \phi(|r_{obs}|) = 0.0001214$$

Therefore, since it is lower than $\alpha = 0.05$, we reject the null hypothesis that leptin and fat mass are independent.

Leptin and fat mass are weakly correlated but highly significant.

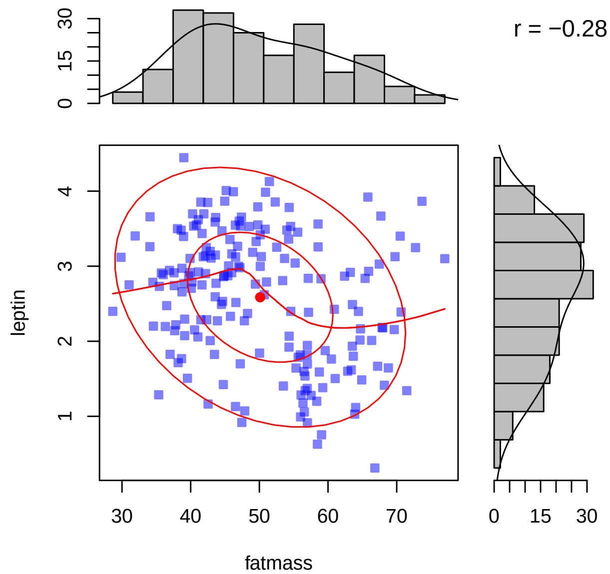
19.12 Correlation coefficient

```
##
##  Pearson's product-moment correlation
##
## data:  data$leptin and data$fatmass
## t = -3.9262, df = 186, p-value = 0.0001214
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.4037736 -0.1390439
## sample estimates:
##        cor
## -0.2766492
```

19.13 Correlation coefficient

We should take this result with care:

- We see that the **marginals** are not quite normal distributions.
- As we increase fat mass we should obtain more leptin, as it is released from adipose tissue.



19.14 Conditional distribution

We can rather ask: For a given value of fat mass, what is the probability density of leptin?

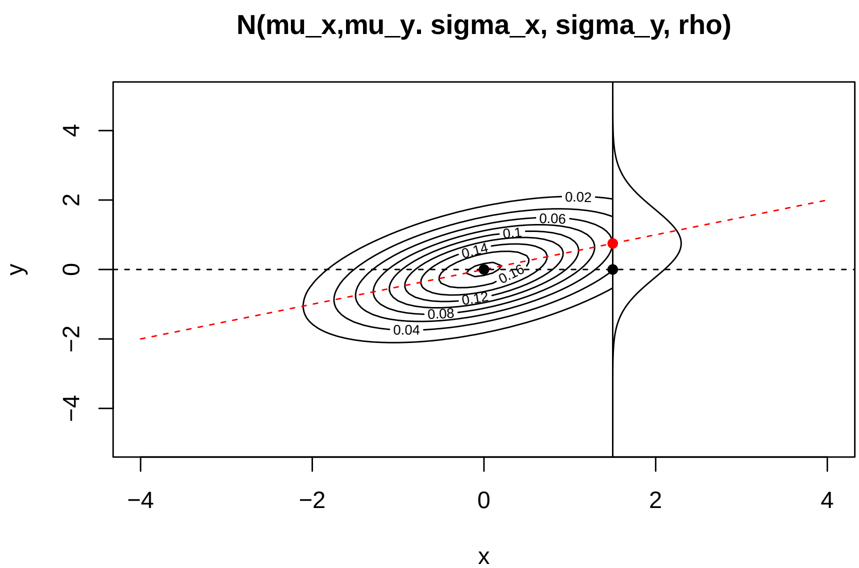
The conditional probability of Y (leptin) given X (fat mass) is

$f(y|x) = \frac{f(y,x)}{f(y)}$

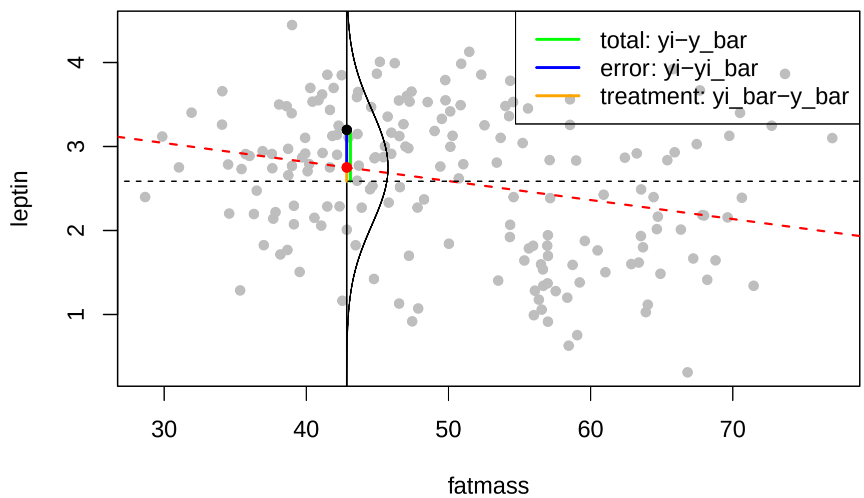
$$= N(\mu_{y|x}, \sigma^2_{y|x})$$

with

- mean: $E(Y|X) = \mu_{y|x} = \mu_y + \rho\frac{\sigma_y}{\sigma_x}(x - \mu_x)$
- variance: $V(Y|X) = \sigma^2_{y|x} = \sigma^2_y(1 - \rho^2)$


N(mu_x,mu_y. sigma_x, sigma_y, rho)
0.02
0.04
0.06
0.08
0.1
0.12
0.14
0.16
y
x
−4
−2
0
2
4

19.15 Sums of squares



- The total sum of squares for the conditional distributions are

$$SS_{tot} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}(Y_i - \hat{\mu}_y)^2$$

- The sum of squares for the error given x_i is

$$SSE = \sum_{i=1}^{n}(Y_i - \bar{Y}_i)^2 = \sum_{i=1}^{n}(Y_i - \hat{\mu}_{y|x_i})^2$$

- The sum of squares for the treatment (explained by variations of x) is

$$SS_{treatment} = \sum_{i=1}^{n}(\bar{Y}_i - \bar{Y})^2 = \sum_{i=1}^{n}(\hat{\mu}_{y|x} - \hat{\mu}_y)^2$$

19.16 Coefficient of determination

The coefficient of determination R^2 is the percentage explained of the total variance

$$R^2 = \frac{SS_{treatment}}{SS_{tot}}$$

It has mean:

$$E(R^2) = \rho^2 = \frac{\sigma_y^2 - \sigma_{y|x}^2}{\sigma_y^2}$$

- $\sigma_y^2 - \sigma_{y|x}^2$ is the variance associated to changes in x ($SS_{treatment} = SS_{tot} - SSE$).
- R^2 is the square of the correlation coefficient.
- If R^2 is near 1 most of the total variance is explained by the regression (the error is near zero)
- If R^2 is 0 no variance is explained (total variance is all error).
- In our data $r^2 = 0.07$ and therefore only 7 of the variation of x explained the variation on y

---

19.17 Linear model

Consider the **linear model**

$$Y_{x_i} = \alpha + \beta x_i + \epsilon_i$$

with **Random error**:

- ϵ_i is a **random variable** with $E(\epsilon_i) = 0$, $V(\epsilon_i) = \sigma_{y|x}^2$
- i is the index of the observation: $1...n$ (typically one for every x_i as x_i is continuous)

and

$$E(Y|x_i) = \alpha + \beta x_i$$

This is called a **regression line** of Y on x, it tells us how the mean of Y_x varies with x.

Note that:

$$\alpha = \mu_y - \beta\mu_x$$

$$\beta = \rho \frac{\sigma_y}{\sigma_x}$$

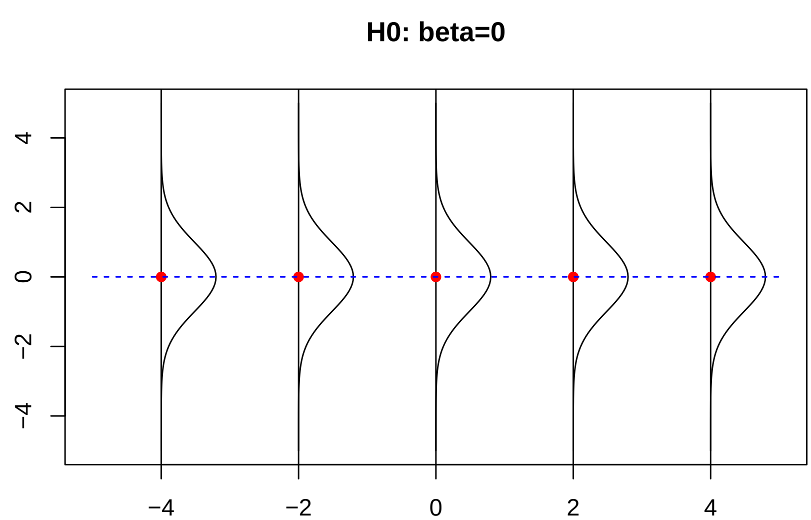
19.18 Hypothesis

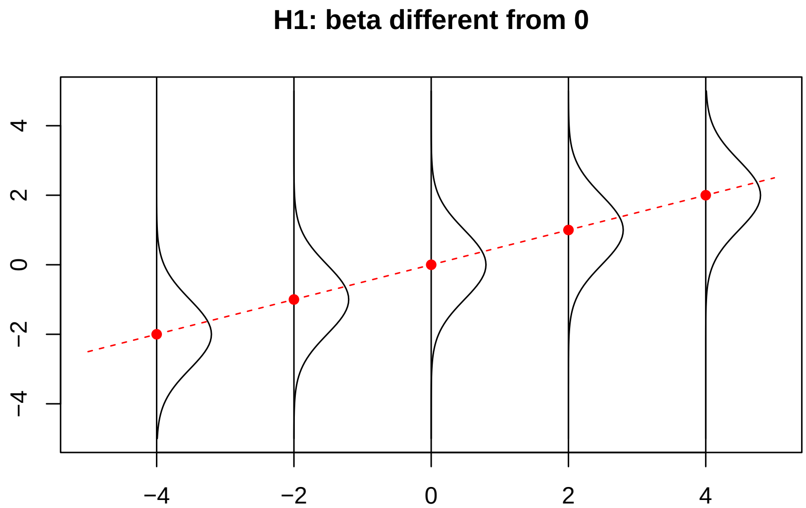
Null hypothesis:

- Y and X are statistically independent, therefore $f(y|x) = f(y)$ and $H_0 : \beta = 0$

Alternative hypothesis:

- Y and X are statistically dependent, therefore $H_1 : \beta \neq 0$





19.19 Estimators

- $\beta = \rho \frac{\sigma_y}{\sigma_x}$ suggest the estimator for β

$$B = \frac{\sum_{i=1}^{m}(x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- $\alpha = \mu_y - \beta\mu_x$ suggests the estimator for α

$$A = \bar{Y} - \hat{\beta}\bar{x}$$

19.20 Estimators

The estimators A and B for α and β can formally be derived from **minimizing the sum of squares** for the error given x_i

$$SSE = \sum_{i=1}^{n}(Y_i - \bar{Y}_i)^2 = \sum_{i=1}^{n}(Y_i - \alpha + \beta x_i)^2$$

with respect to α and β, leading to

$$B = \frac{\sum_{i=1}^{m}(x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- with mean

$$E(B) = \beta$$

- and distribution

$$B \rightarrow N(\beta, \frac{n\sigma_y^2}{(n-2)s_x^2})$$

The estimator A is for α is

$$A = \bar{Y} - \hat{\beta}\bar{x}$$

with mean $E(A) = \mu_y - \beta\mu_x$.

19.21 Hypothesis testing

Under the null hypothesis, $\beta = 0$ and the standardized error from the null hypothesis are

$$\frac{E(B)}{\sqrt{\frac{ns_y^2}{(n-2)s_x^2}}} \rightarrow T(n-2)$$

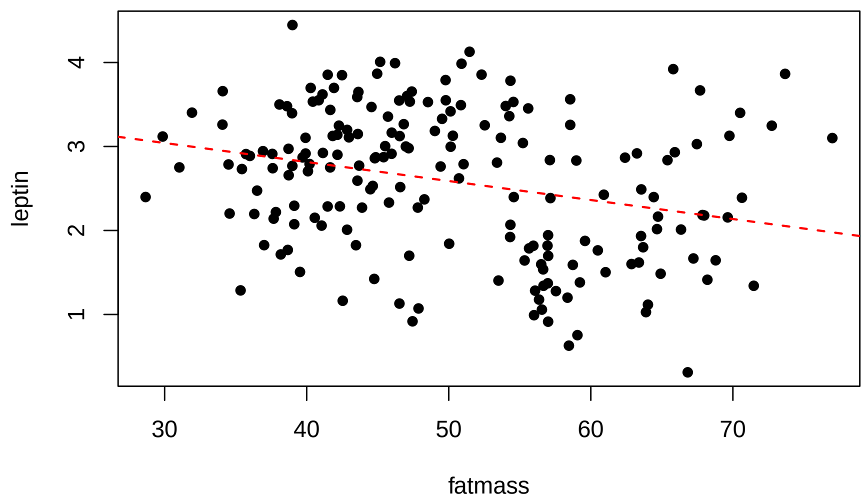
follows a t-distribution with $n-2$ degrees of freedom.

Is our t_{obs} a rare observation from this distribution?

19.22 Model fit

$$\beta_{obs} = \frac{\sum_{i=1}^{m}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = -0.02262$$

$$\alpha_{obs} = \bar{y} - \beta_{obs}\bar{x} = 3.72012$$



But leptin should increase with fat mass ...

19.23 Hypothesis test

```
##
## Call:
## lm(formula = leptin ~ fatmass, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.89723 -0.66914  0.04445  0.64555  1.81122
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.720119   0.295149  12.604  < 2e-16 ***
```

```
## fatmass     -0.022624   0.005762  -3.926 0.000121 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8338 on 186 degrees of freedom
## Multiple R-squared:  0.07653,    Adjusted R-squared:  0.07157
## F-statistic: 15.42 on 1 and 186 DF,  p-value: 0.0001214
```

19.24 Multiple Regression

We can include other conditions in the regression, such as sex or age.

Consider the **linear model**

$$Y_{ij} = \alpha + \beta x_i + \gamma z_j + \epsilon_{ij}$$

It is important to adjust for other factors that we believe are correlated with the outcome Y and the condition x.

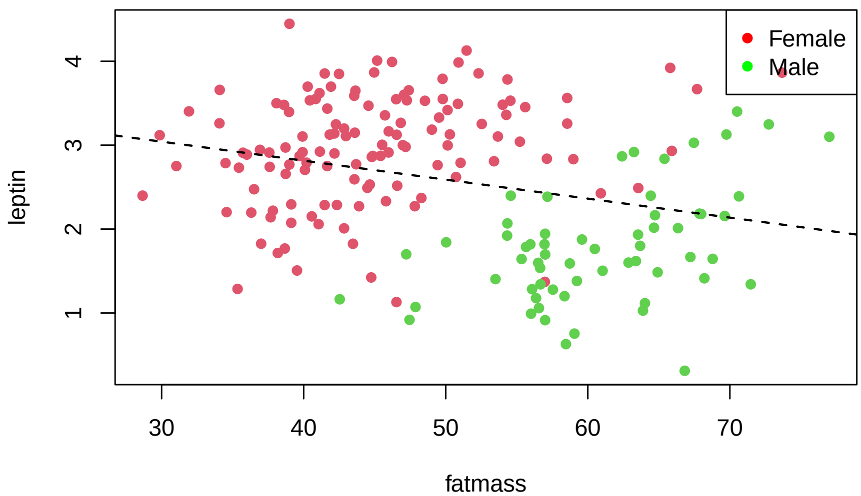
19.25 Multiple Regression

We can now have one outcome Y with multiple regressors

```
##     leptin fatmass sex age
## 1 3.355677  45.721   F  45
## 2 2.272126  43.895   F  77
## 3 1.071584  47.871   M  79
## 4 3.921082  65.801   F  58
## 5 1.536867  56.644   M  42
## 6 1.177115  56.355   M  75
```

19.26 Multiple Regression

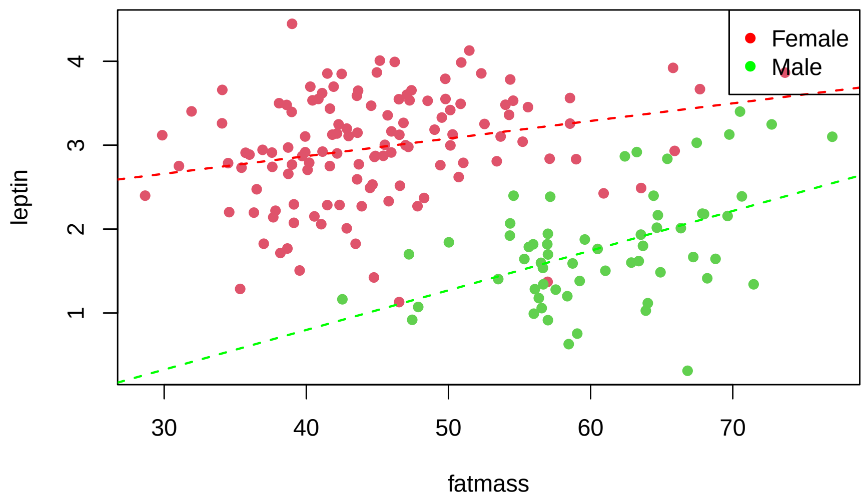
Consider the previous regression and color the points according to their sex



We find that the negative association between leptin and fat mass is given by the effect of sex.

19.27 Multiple Regression

If we run regression separating by sex we find a positive association between leptin and fat mass, as expected.



19.28 Multiple Regression

When we include other factors in the regression, we observe that there is a positive increase in leptin for females and a positive relationship between leptin and fat mass within each sex.

```
##
## Call:
## lm(formula = leptin ~ fatmass + sex + age, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.91314 -0.43523  0.07833  0.38451  1.49707
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.482557   0.304423   4.870 2.40e-06 ***
## fatmass      0.027585   0.006002   4.596 7.99e-06 ***
## sexM        -1.611965   0.135872 -11.864  < 2e-16 ***
## age          0.005353   0.002957   1.811   0.0718 .
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6226 on 184 degrees of freedom
## Multiple R-squared:  0.4907, Adjusted R-squared:  0.4824
## F-statistic: 59.09 on 3 and 184 DF,  p-value: < 2.2e-16
```

19.29 Multiple Regression interaction

We can include interactions between conditions in the regression.

Consider the **linear model**

$$Y_{ij} = \alpha + \beta x_i + \gamma z_j + \delta x_i z_j + \epsilon_{ij}$$

The parameter δ will add a contribution to β that is specific to the condition j

- if $z_i \in (0, 1)$ then when $z = 0$ the coefficient of x_i is β, when $z = 1$ the coefficient of x_i is $\beta + \gamma$

γ will test the differences between βs in males and females.

19.30 Multiple Regression interaction

Our data suggest that a steeper increase of leptin with body fat in males than in females (interaction: 0.028427, $pvalue = 0.03$)

```
##
## Call:
## lm(formula = leptin ~ fatmass * sex + age, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.82098 -0.38692  0.03192  0.42065  1.43851
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.800052   0.337279   5.337 2.77e-07 ***
## fatmass       0.020022   0.006949   2.881  0.00443 **
## sexM         -3.218004   0.775217  -4.151 5.07e-05 ***
## age           0.005846   0.002939   1.989  0.04815 *
```

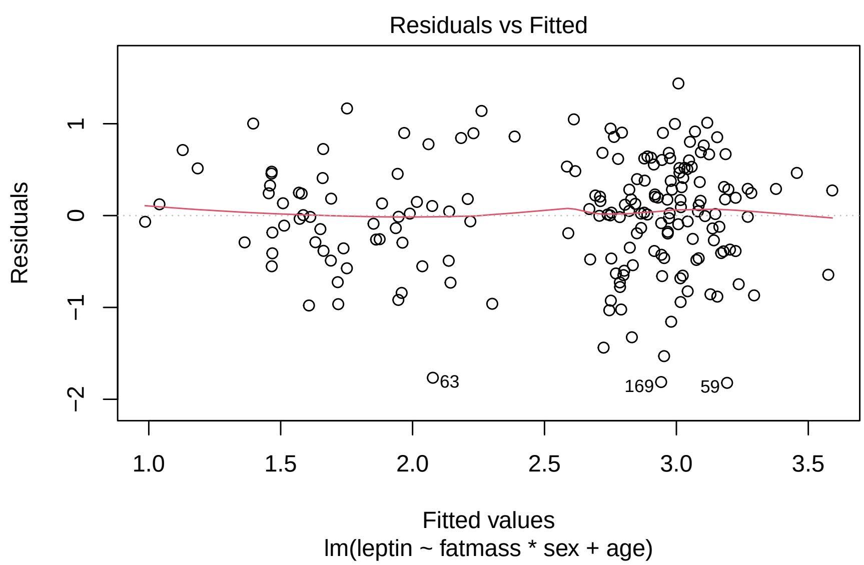
```
## fatmass:sexM  0.028427   0.013513   2.104  0.03677 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6169 on 183 degrees of freedom
## Multiple R-squared:  0.5027, Adjusted R-squared:  0.4918
## F-statistic: 46.25 on 4 and 183 DF,  p-value: < 2.2e-16
```

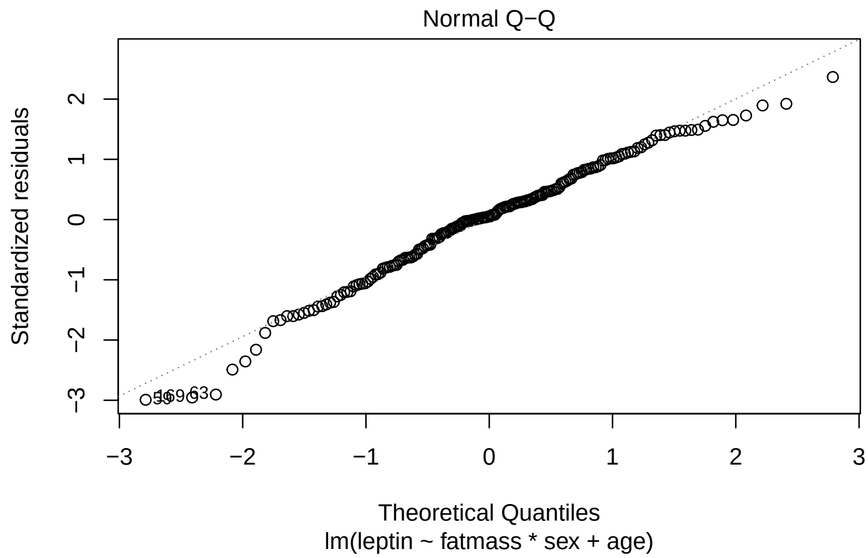
19.31 Model diagnostics

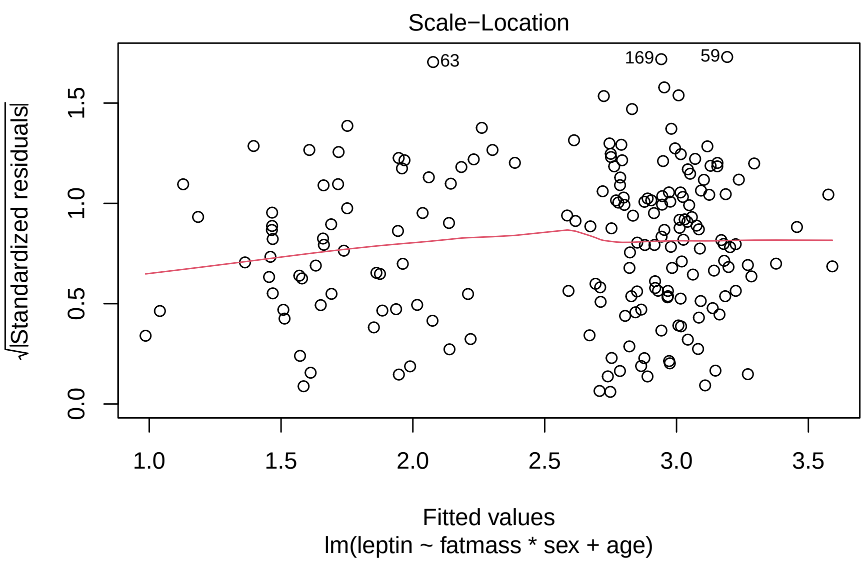
All linear models have been made on the supposition that

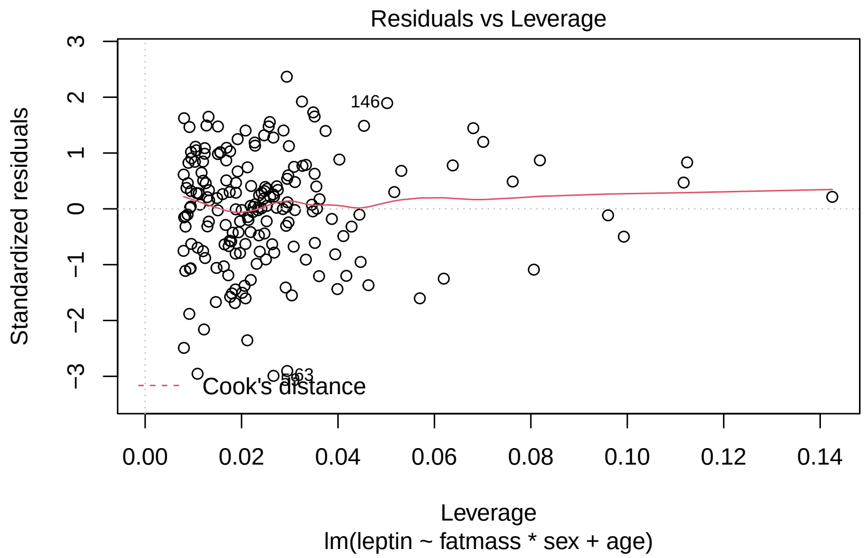
- Errors are distributed normally
- Errors have the same variance

There are a number of plots to check that at least the data is consistent with these suppositions




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(leptin ~ fatmass * sex + age)
59
169
63
−3
−2
−1
0
1
2
3


Scale−Location
√|Standardized residuals|
Fitted values
lm(leptin ~ fatmass * sex + age)
63
169
59
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5



19.32 Maximum likelihood

Let's look back at Gauss and study the maximum likelihood estimator of the regression.

- Gauss wanted to predict the position of Ceres in the summer of 1802 after it passed behind the sun. Depending on the position they could get decide whether Ceres was a new planet.

$$N(y_i; \mu_i = \alpha + \beta t_i, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(y_i - \mu_i)^2} = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(y_i - \alpha - \beta t_i)^2}$$

what are the maximum likelihood estimates for α and β?

19.33 Maximum likelihood

The likelihood function, the probability of having observed $(x_1, ...x_n)$ at $t_i, ...t_n$

$$L(\mu_i,\sigma) = \Pi_{i=1..n} N(\alpha+\beta t_i;\mu,\sigma)$$
$$= (\frac{1}{\sigma\sqrt{2\pi}})^n e^{-\frac{1}{2\sigma^2}\sum_i (y_i-\alpha-\beta t_i)^2}$$

The log-likelihood is

$\log(\Pi_{i=1..n} N(\alpha+\beta t_i;\mu,\sigma)) = -\frac{n}{2}\log(2\pi) - n\log(\sigma) - \frac{1}{2\sigma^2}\sum_i (y_i-\alpha-\beta t_i)^2$

that we differentiate with respect to α and β and equate to 0 to find the maxima.

After some algebra (exercise) we have

$$\hat{\beta} = \frac{\sum_i (t_i-\bar{t})(y_i-\bar{y})}{\sum_i (t_i-\bar{t})^2}$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{t}$$

19.34 Maximum likelihood

These are the values we obtained when we adjust a line to observations $(x_1, y_1)...(x_n, y_n)$ by minimum squares (when we had x instead o t).

$$\hat{\beta} = \frac{\sum_i (x_i-\bar{x})(y_i-\bar{y})}{\sum_i (x_i-\bar{x})^2}$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

β is the realization of the statistic

$$B = \frac{\sum_i (x_i-\bar{x})(Y_i-\bar{Y})}{\sum_i (x_i-\bar{x})^2}$$

That is the sum of the normal variables $Y_1, ...Y_n$, and therefore is normal.

19.35 Maximum likelihood

- We can then test the probability that $pval = P(B > 0)$ or that Ceres is moving in the sky.
- As we fix the values of α and β we can compute $E(Y_{t_n})$ as the most likely prediction of the position of Ceres at time t_n.

Gauss's story is one of the most important advancements in science. To predict where to find Ceres in the sky in 1802

- he discovered the **normal distribution**
- lay the foundations of **maximum likelihood** method and **regression** analysis.
- showed that the most likely value of Ceres was the **average**

Astronomers pointed their telescopes where Gauss told them, and there Ceres was!

Chapter 20

Group Work sessions

20.1 Objectives

- The objective of the work sessions is to work together with a student of a **different background** to perform a full analysis of the **misophonia dataset**.
- The analysis is **open**. You can formulate the analysis you consider interesting, trying to cover as much as possible the material we have seen in theory and bootcamps.
- **Justify** your analysis and **discuss** them.
- We will have **two sessions** to perform the report that will be done in colab and handed in through **google classroom**.
- *Work together, follow your interests and have fun!*

Next, we **describe** the data and show an **example** of the kind of analysis that can be performed in both group sessions.

20.2 Misophonia dataset

Misophonia is a recently described neurological condition whereby patients feel strong anxiety when hearing particular noises (someone blowing their nose, mobile ringing, trains passing, etc..). It is believed that 5% of the population suffers from this condition without knowing it, likely blaming their anxiety on other causes.

The misophonia dataset is from a recent (unpublished) study that aimed to describe the relationships between misophonia and anxiety, depression, and cephalometric measures (shape of the jaw).

```
##   Misofonia Misofonia.dic     Estado Estado.dic ansiedad.rasgo
## 1        si             4 divorciado          2             99
## 2        si             2     casado          1             75
## 3        no             0 divorciado          2             77
## 4        si             3     casado          1             95
## 5        no             0     casado          1             30
## 6        no             0     casado          1             30
##   ansiedad.rasgo.dic ansiedad.estado ansiedad.estado.dic ansiedad.medicada
## 1                  1              99                   1                no
## 2                  1              75                   1                no
## 3                  1              55                   0                no
## 4                  1              99                   1                no
## 5                  0              40                   0                no
## 6                  0              30                   0                no
##   ansiedad.medicada.dic depresion depresion.dic Sexo Edad CLASE
## 1                     0     33.65             1    M   44   III
## 2                     0     19.77             0    M   43    II
## 3                     0     29.57             0    M   24     I
## 4                     0      1.40             0    M   33   III
## 5                     0      5.98             0    H   41     I
## 6                     0     13.87             0    H   35     I
##   Angulo_convexidad protusion.mandibular Angulo_cuelloYtercio Subnasal_H
## 1              7.97                 13.0                 89.6        1.5
## 2             18.23                 -5.0                107.2        7.3
## 3             12.27                 11.5                101.4        5.0
## 4              7.81                 16.8                 75.3        2.7
## 5              9.81                 33.0                105.5        6.0
## 6             13.50                  2.0                105.0        7.0
##   cambio.autoconcepto Misofonia.post Misofonia.pre ansiedad.dif
## 1                   1             21            14            0
## 2                   0             14            13            0
## 3                  NA             NA            NA          -22
## 4                   1             NA            NA            4
## 5                  NA             NA            NA           10
## 6                  NA             NA            NA            0
```

Here is the description of the variables

[1] "Misofonia": Binary (si: misophinic, no: no misophinic)
[2] "Misofonia.dic": Categorical (0: no misophinic, 1: severity 1, 2: severity 2, 3: severity 3, 4: severity 4)
[3] "Estado": Marital status (casado: married, soltero: single, viuda: widow, divorciado:divorced)
[4] "Estado.dic": Numeric Marital status
[5] "ansiedad.rasgo": Score from 0-100 with anxiety personality trait
[6] "ansiedad.rasgo.dic": Binary score (0,1) of anxiety personality trait
[7] "ansiedad.estado": Score from 0-100 with current state of anxiety

[8] “ansiedad.estado.dic”: Binary score (0,1) with current state of anxiety
[9] “ansiedad.medicada”: Diagnosed with anxiety disorder (si, no)
[10] “ansiedad.medicada.dic”: Diagnosed with anxiety disorder (1, 0)
[11] “depresion”: Score from 0-50 with current state of depression
[12] “depresion.dic” : Binary score (0,1) with current state of depression
[13] “Sexo”: Male=H, Female:M
[14] “Edad”: Age
[15] “CLASE”: Type of jaw
[16] “Angulo_convexidad”: convexity angle
[17] “protusion.mandibular”: Projection of the jaw [18] “Angulo_cuelloYtercio”: angle between jaw and neck [19] “Subnasal_H”: Nasal angle
[20] “cambio.autoconcepto”: Whether people changed their self-concept after treatment.
[21] “Misofonia.post”: Misophionia diagnosed (A-MISO) after an educational program, where patients were made aware of a condition called misophonia.
[22] “Misofonia.pre”: Misophionia diagnosed (A-MISO) before an educational program, where patients were made aware of a condition called misophonia
[23] “ansiedad.dif”: Difference between anxiety state and anxiety trait scores

20.3 Group Work session 1: Data description

When reporting the results of a study, we first describe the variables of interest in tables and figures.

- We describe demographics (sex, age, marital status, etc..)
- We describe outcome variables (misophonia)
- We describe explanatory variables (cephalometric measures, anxiety, depression)

Example:

Imagine we want to study the anxiety of participants in the misophonia study

We load the data

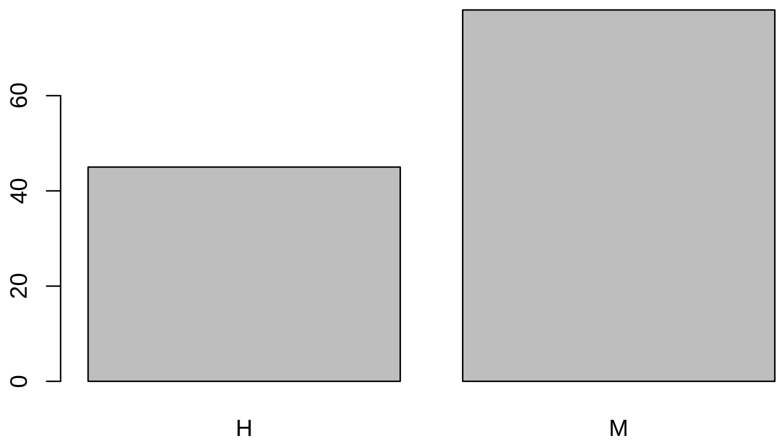
```
##   Misofonia Misofonia.dic     Estado Estado.dic ansiedad.rasgo
## 1        si             4 divorciado          2             99
## 2        si             2     casado          1             75
## 3        no             0 divorciado          2             77
## 4        si             3     casado          1             95
## 5        no             0     casado          1             30
## 6        no             0     casado          1             30
##   ansiedad.rasgo.dic ansiedad.estado ansiedad.estado.dic ansiedad.medicada
## 1                  1              99                   1                no
## 2                  1              75                   1                no
## 3                  1              55                   0                no
## 4                  1              99                   1                no
```

```
## 5                 0              40                  0                no
## 6                 0              30                  0                no
##   ansiedad.medicada.dic depresion depresion.dic Sexo Edad CLASE
## 1                     0     33.65             1    M   44   III
## 2                     0     19.77             0    M   43    II
## 3                     0     29.57             0    M   24     I
## 4                     0      1.40             0    M   33   III
## 5                     0      5.98             0    H   41     I
## 6                     0     13.87             0    H   35     I
##   Angulo_convexidad protusion.mandibular Angulo_cuelloYtercio Subnasal_H
## 1              7.97                 13.0                 89.6        1.5
## 2             18.23                 -5.0                107.2        7.3
## 3             12.27                 11.5                101.4        5.0
## 4              7.81                 16.8                 75.3        2.7
## 5              9.81                 33.0                105.5        6.0
## 6             13.50                  2.0                105.0        7.0
##   cambio.autoconcepto Misofonia.post Misofonia.pre ansiedad.dif
## 1                   1             21            14            0
## 2                   0             14            13            0
## 3                  NA             NA            NA          -22
## 4                   1             NA            NA            4
## 5                  NA             NA            NA           10
## 6                  NA             NA            NA            0
```

1. We describe the participants' sex, age, and marital status

a. Sex

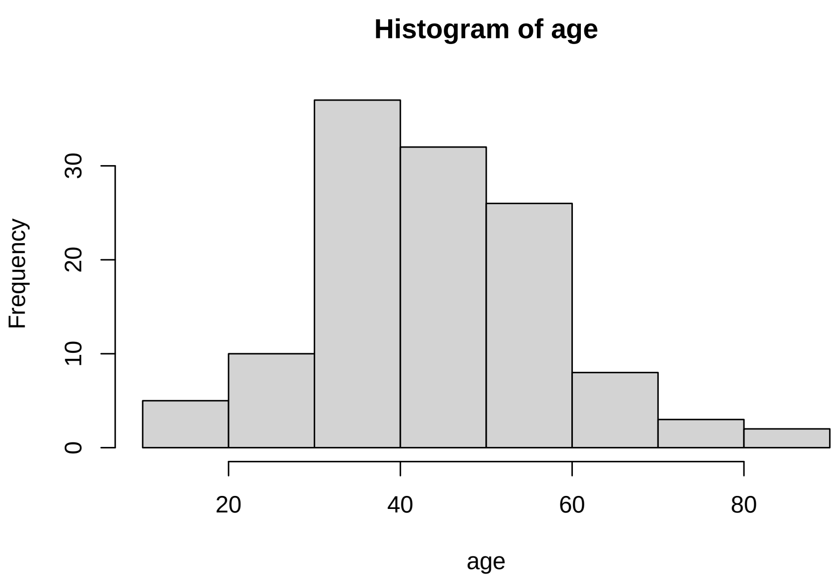
```
## sex
##         H         M
## 0.3658537 0.6341463
```

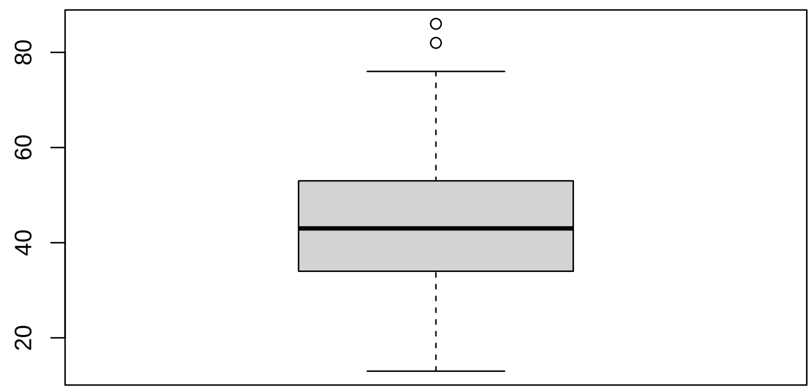


b. Age

```
## [1] 43.93496
```

```
## [1] 14.18654
```





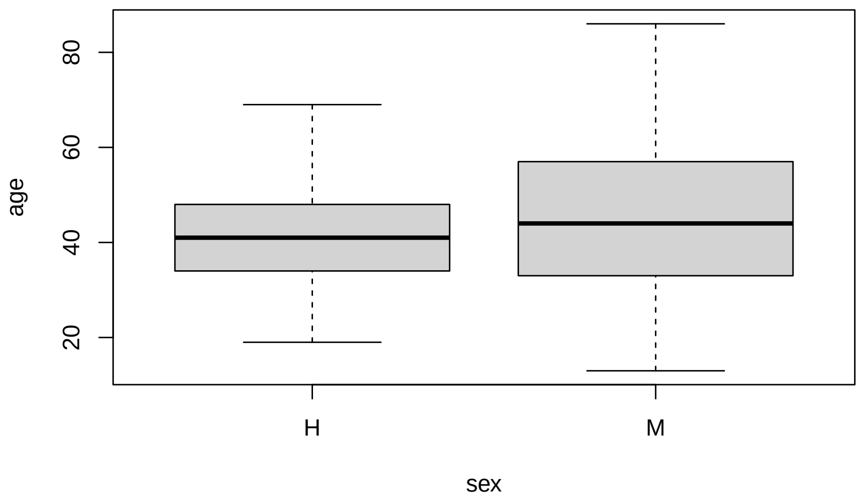
c. Age by sex

```
## [1] 40.64444
```

```
## [1] 10.75165
```

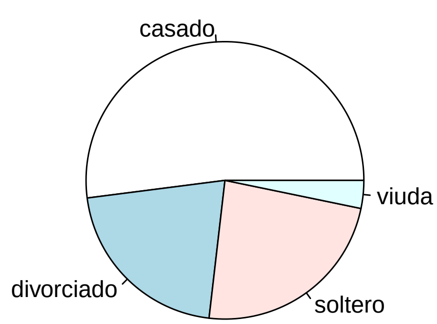
```
## [1] 45.83333
```

```
## [1] 15.58339
```



d. Marital status

```
## Mstate
##     casado divorciado    soltero      viuda
## 0.52032520 0.21138211 0.23577236 0.03252033
```



2. We describe the clinical outcome, for example, anxiety.

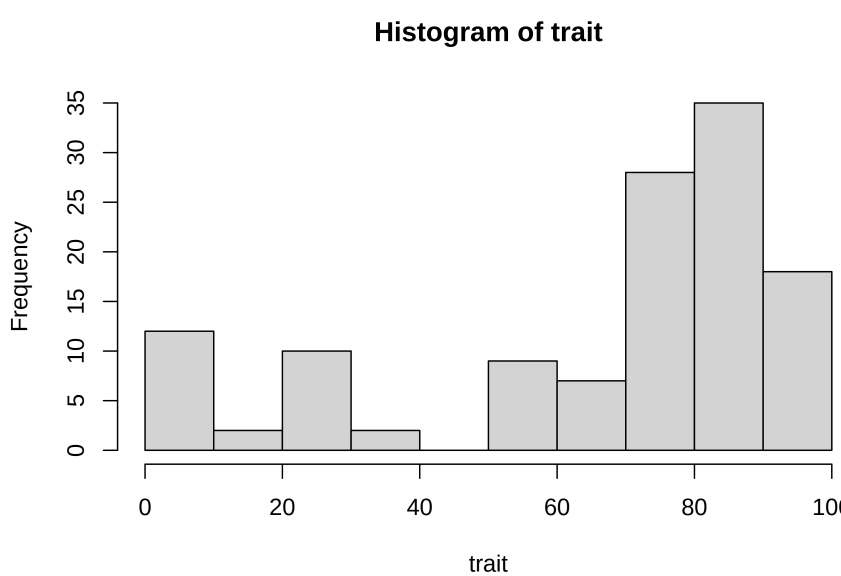
We have four measures of anxiety:

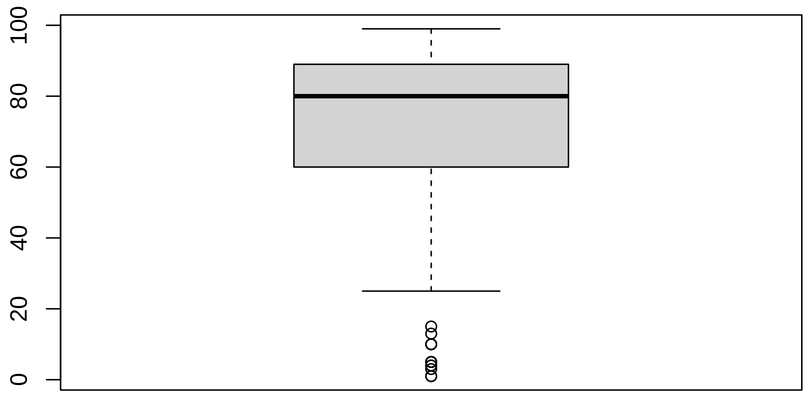
- Trait: ansiedad.rasgo (are you an anxious person?) continuous:0-100
- State: ansiedad.estado (are you currently feeling anxious?) continuous:0-100
- Diagnosed: ansiedad.medicada (have you been diagnosed with an anxiety disorder?) binary (si, no)
- Excess: ansiedad.dif (difference between State and Trait)

we describe these clinical outcomes

a. Trait (min, max, quantiles, median)

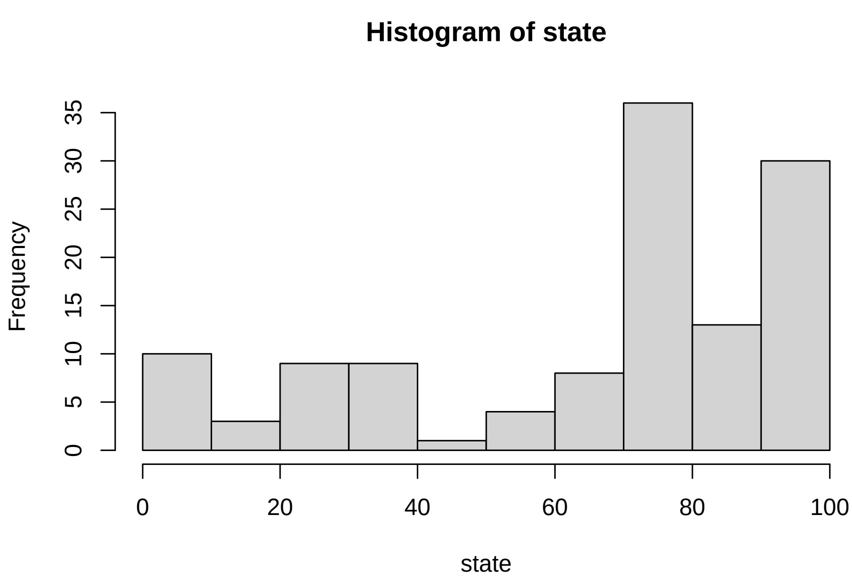
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    1.00   60.00   80.00   68.77   89.00   99.00      15
```

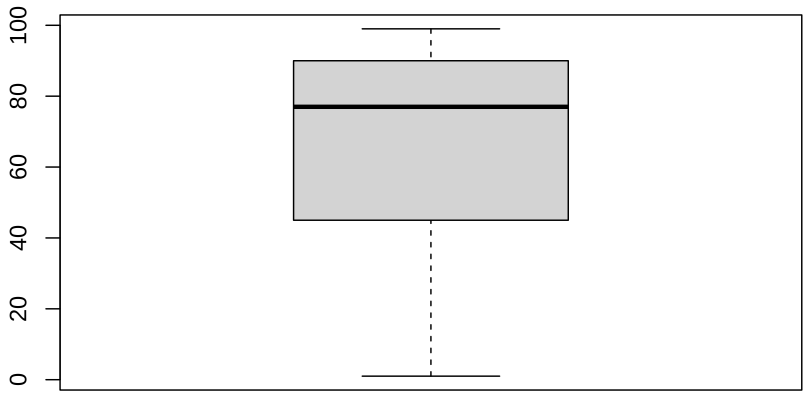




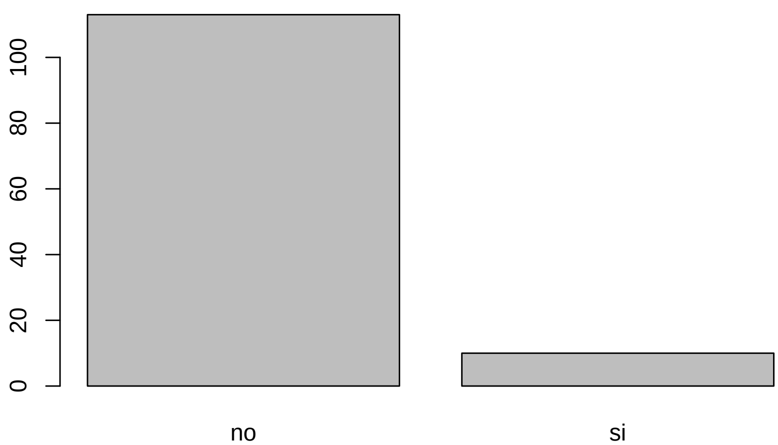
b. State (min, max, quantiles, median)

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    1.00   45.00   77.00   67.85   90.00   99.00      15
```



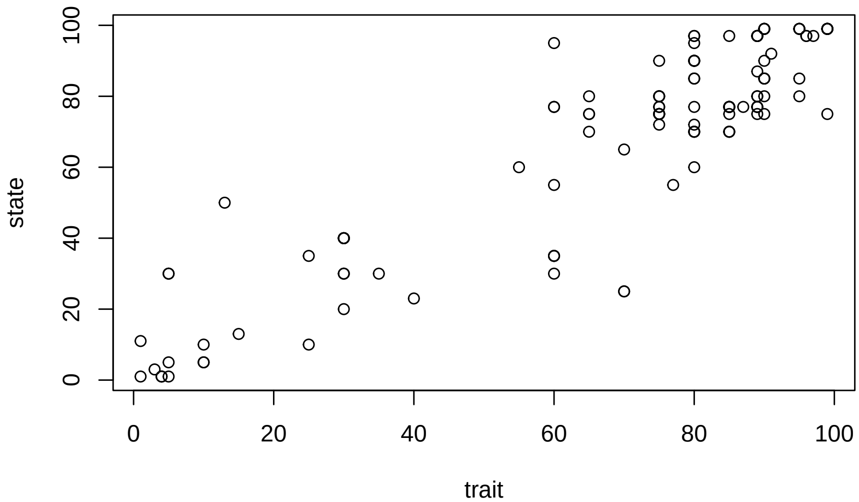


c. Diagnosed



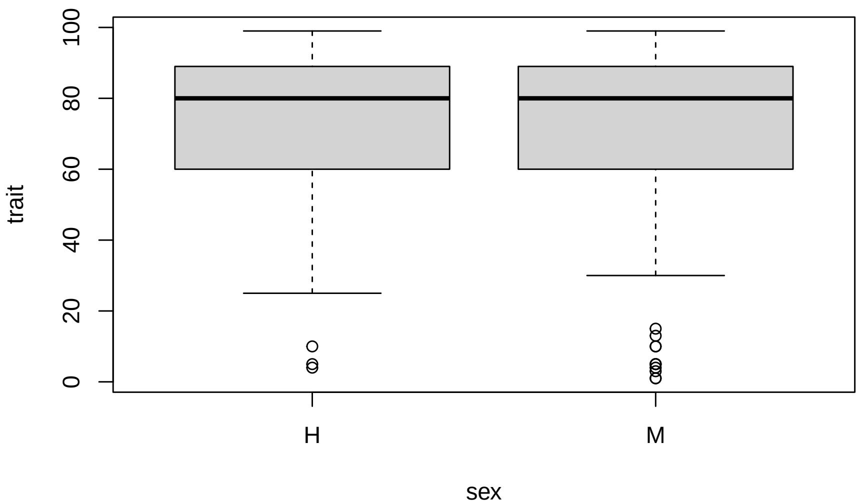
We can look at relationships between outcomes

d. Trait Vs Estate

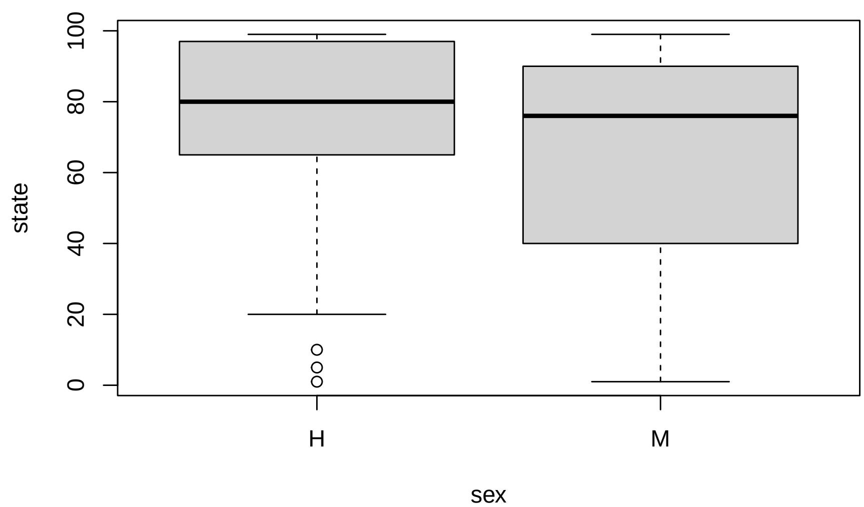


We can also look at the relationships between the clinical outcomes and the features of the participants

e. Trait by sex

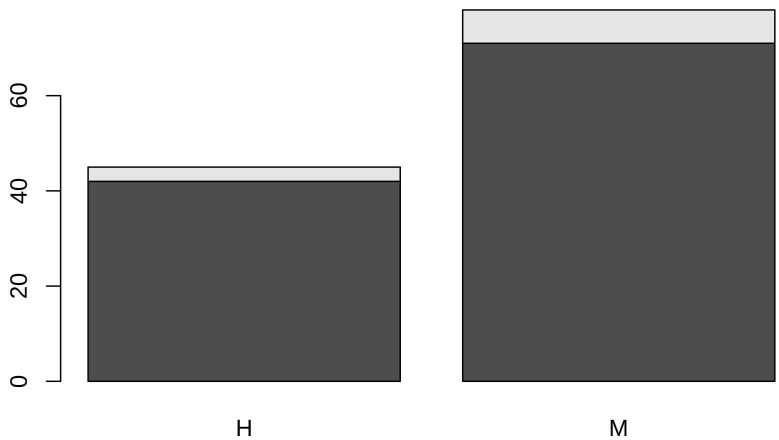


f. State by sex



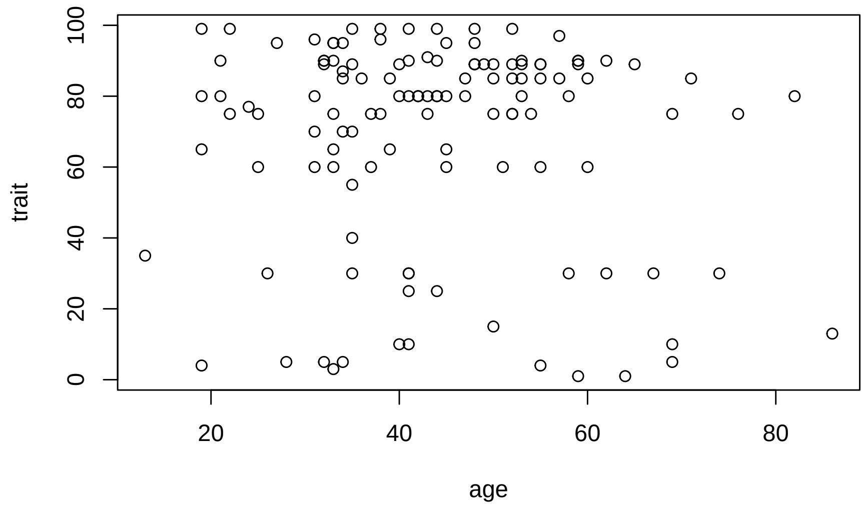
g. Diagnosed by sex

```
##          sex
## diagnosed  H  M
##        no 42 71
##        si  3  7
```

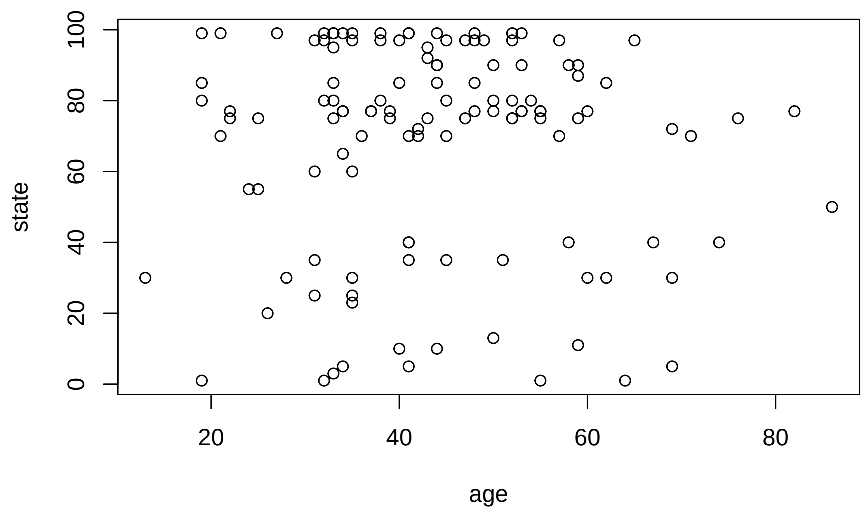



```
##          sex
## diagnosed          H          M
##        no 0.93333333 0.91025641
##        si 0.06666667 0.08974359
```

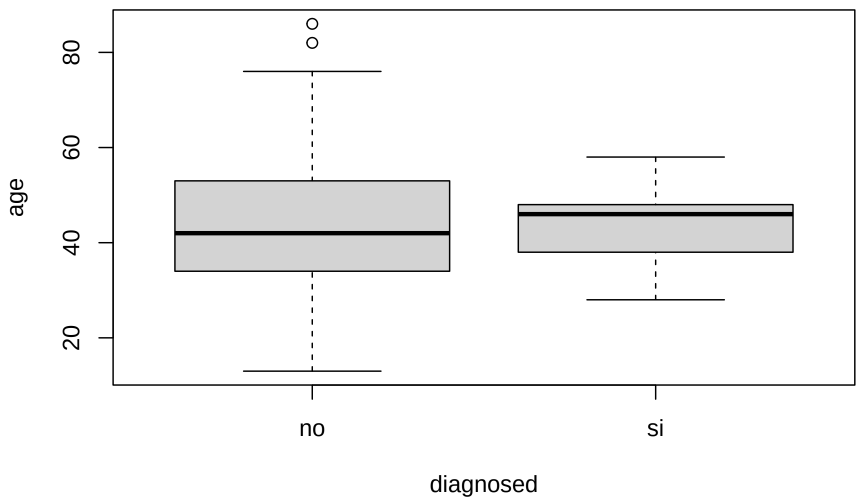
h. Trait Vs age



i. State Vs age



j. age by diagnosis



20.4 Group Work session 2: Inference

When reporting the results of a study, we first describe the variables of interest in tables and figures.

- We describe demographics (sex, age, marital status, etc..)
- We describe outcome variables (misophonia/axiety/depression/etc..)
- We describe explanatory variables (cephalometric measures, anxiety, depression)

We then test the main hypotheses of the study.

- We state the main relationships we want to study and formulate the statistical hypothesis (Introduction)

- We describe how the study was performed and the statistical methods to test the hypothesis (Methods)

- We describe the results of the hypothesis tests with statistics, and significance measures.

- We illustrate the results with figures.

Example:

Imagine we want to study the anxiety of participants in the misophonia study.

We formulate the following hypothesis:

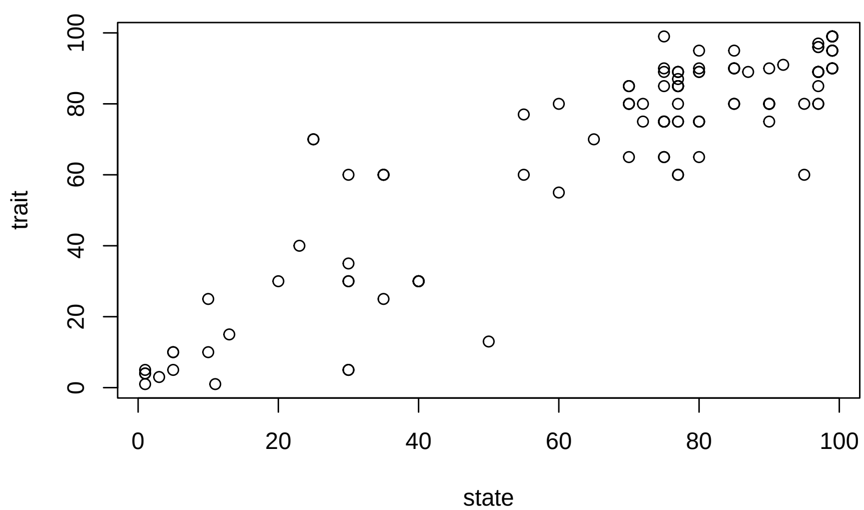
Participants who enrolled in the study had an increased level of anxiety from their baseline (trait) that is related to their:

- age
- sex
- misophonia state.

We are interested in the variable misofonia.dif, that is the observed **excess** of anxiety from the trait

$excess = state - trait$

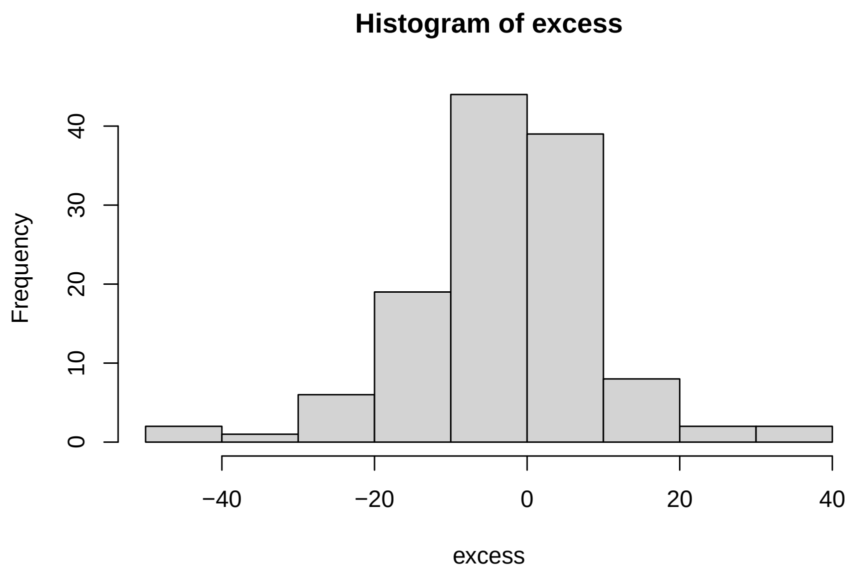
1. Are the state and trait of anxiety correlated?




```
##
##  Pearson's product-moment correlation
##
## data:  state and trait
## t = 23.282, df = 121, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.8656964 0.9320106
## sample estimates:
##       cor
## 0.9041609
```

2. Is excess in anxiety higher than 0?

a. We describe the Excess variable with summary statistics and figures (histogram)




```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's
## -45.0000  -8.0000   0.0000  -0.9187   8.0000  37.0000       15
```

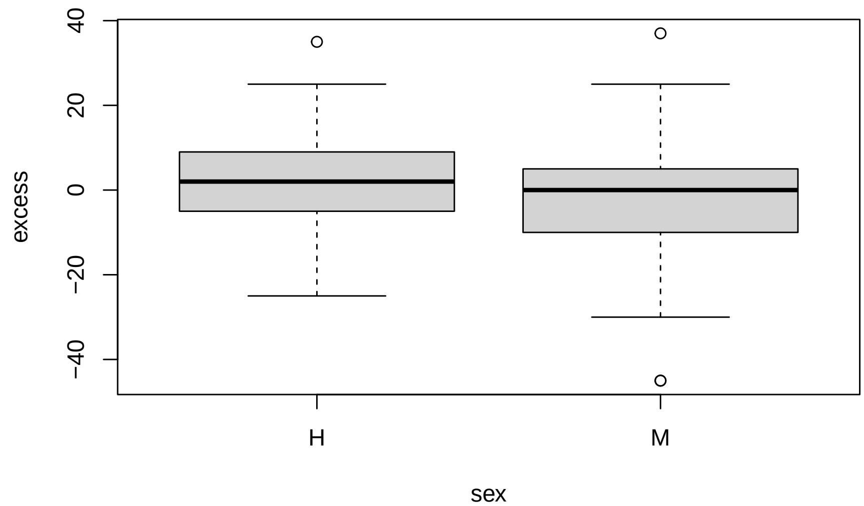
b. We then perform a hypothesis test for the mean of anxiety excess $H_0 : \mu = 0$ against $H_1 : \mu \neq 0$.

```
##
##  One Sample t-test
##
## data:  excess
## t = -0.79192, df = 122, p-value = 0.4299
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -3.215212  1.377814
## sample estimates:
##  mean of x
## -0.9186992
```

c. We conclude: We do not see significant large values of the difference in anxiety; Enrollment in the study does not seem to detect individuals with an excess of anxiety.

2. Is excess in anxiety higher than 0 for men and women separately?

a. We first describe the conditional distributions



b. We perform the hypothesis test for each sex separately

```
##
##  One Sample t-test
##
## data:  excess[sex == "M"]
## t = -1.6994, df = 77, p-value = 0.09328
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -5.5685793  0.4403741
## sample estimates:
## mean of x
## -2.564103
```

```
##
##  One Sample t-test
##
## data:  excess[sex == "H"]
## t = 1.1158, df = 44, p-value = 0.2706
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -1.558796  5.425462
## sample estimates:
## mean of x
##  1.933333
```

c. We conclude: We see that women (M) have a reduction in the excess of anxiety (almost significant), while men (H) had an increase (no significant). Why? perhaps because females tend to consult doctors before men do.

3. Is the excess of anxiety significantly different between the sexes?

a. We test the hypothesis $H_0 : \mu_{men} = \mu_{women}$ against $H_1 : \mu_{men} \neq \mu_{women}$ using a group t.test

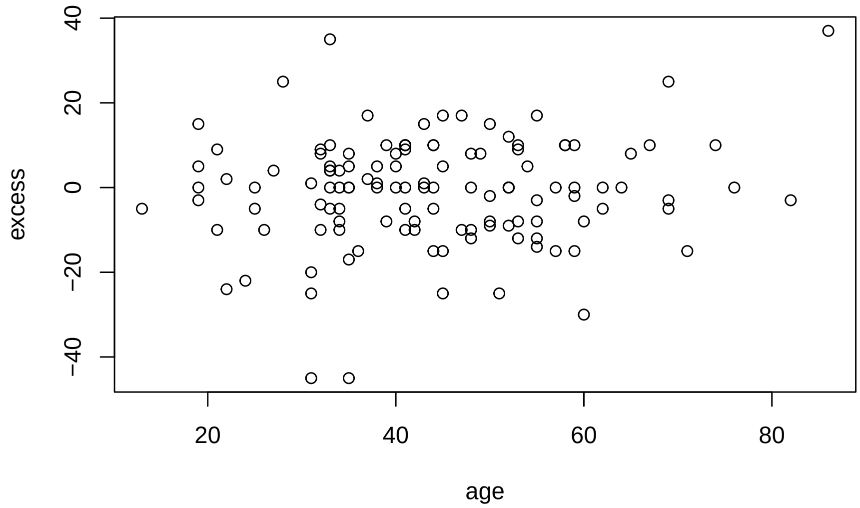
```
##
##  Welch Two Sample t-test
##
## data:  excess[sex == "M"] and excess[sex == "H"]
## t = -1.9574, df = 102.39, p-value = 0.05302
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -9.05452801  0.05965621
## sample estimates:
## mean of x mean of y
## -2.564103  1.933333
```

b. We conclude: we see that the difference between the group means is within the limit of significance with women having less excess anxiety than men.

```
##
## Call:
## lm(formula = excess ~ sex)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -42.436  -7.436   2.067   7.564  39.564
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.933      1.898   1.019   0.3105
## sexM          -4.497      2.384  -1.887   0.0616 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.73 on 121 degrees of freedom
##   (15 observations deleted due to missingness)
## Multiple R-squared:  0.02858,    Adjusted R-squared:  0.02055
## F-statistic:  3.56 on 1 and 121 DF,  p-value: 0.06158
```

4. Is excess in anxiety higher in older people?

a. We make a plot between anxiety and age

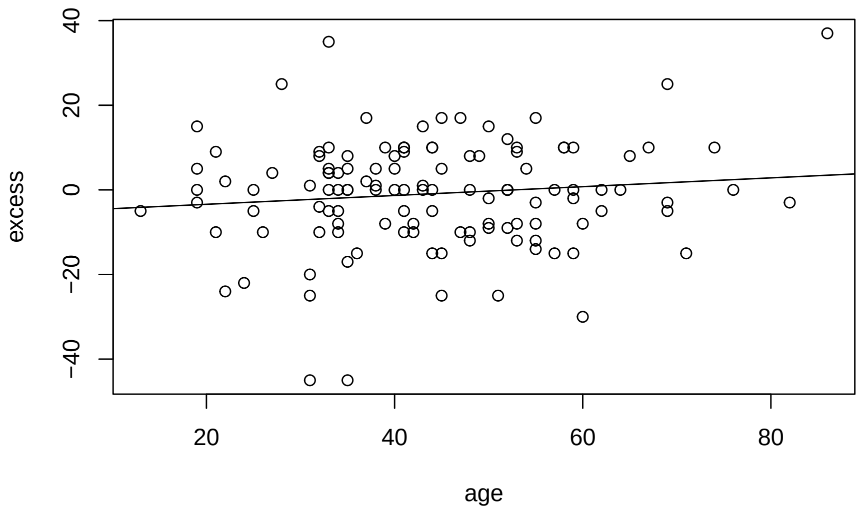


b. We fit the regression model

$$excess = \alpha + \beta * age + \epsilon$$

and test the hypothesis $H_0 : \beta = 0$ against $H_1 : \beta \neq 0$

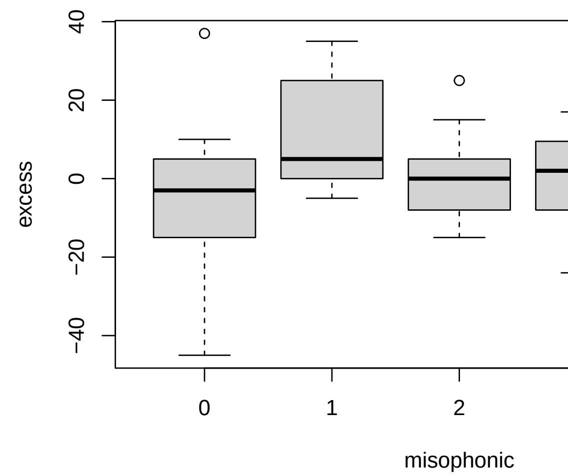
```
##
## Call:
## lm(formula = excess ~ age)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -43.151  -7.776   0.912   8.516  37.057
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -5.4917     3.7799  -1.453    0.149
## age           0.1041     0.0819   1.271    0.206
##
## Residual standard error: 12.83 on 121 degrees of freedom
##   (15 observations deleted due to missingness)
## Multiple R-squared:  0.01317,    Adjusted R-squared:  0.005016
## F-statistic: 1.615 on 1 and 121 DF,  p-value: 0.2062
```



c. We conclude: The association, while positive it is not significant. If we adjust by sex the association is a bit stronger but still not significant.

```
##
## Call:
## lm(formula = excess ~ age + sex)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -40.969  -6.849   0.781   8.019  34.124
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.57179    3.82807  -0.933   0.3527
## age          0.13545    0.08198   1.652   0.1011
## sexM        -5.20025    2.40467  -2.163   0.0326 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.64 on 120 degrees of freedom
##   (15 observations deleted due to missingness)
## Multiple R-squared:  0.05019,    Adjusted R-squared:  0.03436
## F-statistic:  3.17 on 2 and 120 DF,  p-value: 0.04553
```

5. Is excess in anxiety different between misophonic grades?



a. We plot the excess anxiety across groups (boxplot)

b. We test the hypothesizes $H_0 : \mu_0 = \mu_1 ... = \mu_4$ against H_1 : at least one of them is different. We fit an ANOVA model.

```
##
## Call:
## lm(formula = excess ~ misophonic)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -39.902  -8.257   1.243   7.152  42.098
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -5.098      1.944  -2.622  0.00988 **
## misophonic1   17.098      5.896   2.900  0.00445 **
## misophonic2    3.854      2.822   1.366  0.17464
## misophonic3    6.904      2.962   2.331  0.02148 *
## misophonic4    7.986      4.582   1.743  0.08391 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.45 on 118 degrees of freedom
##   (15 observations deleted due to missingness)
## Multiple R-squared:  0.09483,    Adjusted R-squared:  0.06414
## F-statistic:  3.09 on 4 and 118 DF,  p-value: 0.01847
```

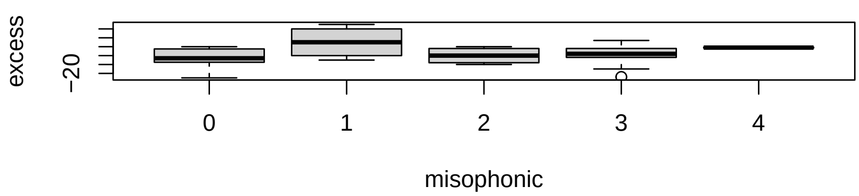
```
## Analysis of Variance Table
```

```
##
## Response: excess
##             Df Sum Sq Mean Sq F value  Pr(>F)
## misophonic   4   1915  478.76  3.0904 0.01847 *
## Residuals  118  18280  154.92
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

c. We conclude: We see that anxiety excess of misophonia grade 1 is significantly higher than misophonia grade 0 (no misophonia), as it is grade 3. The ANOVA table shows that we accept the alternative hypothesis, where the differences between groups are significantly higher than within groups.

6. Are the differences in excess anxiety between monophonic grades modulated by sex?

a. We plot excess anxiety for each misophonic group, for men and women separately



b. We perform an ANOVA test for the interaction

```
## Analysis of Variance Table
##
## Response: excess
##                 Df  Sum Sq Mean Sq F value  Pr(>F)
## misophonic       4  1915.0  478.76  3.0366 0.02026 *
## sex              1   179.7  179.74  1.1400 0.28792
```

```
## misophonic:sex   4   284.5   71.13  0.4512 0.77137
## Residuals      113 17815.9  157.66
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

c. We conclude: We do not see a significant interaction (modulation) of the effect of sex on the group differences. We cannot say that the profiles of anxiety excess across misophonia grades are different between sexes.

Chapter 21

Exercises

21.1 Data description

21.1.0.1 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:

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

21.1.0.2 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:

- Compute the relative frequencies of each outcome and draw the histogram
- Compute the cumulative frequencies of each outcome and sketch the cumulative plot.
- Sketch a boxplot.

21.2 Probability

21.2.0.1 Exercise 1

The outcome of one random experiment is to measure the misophonia severity **and** depression status of one patient.

- Misophonia severity: $x \in \{0, 1, 2, 3, 4\}$
- Depression: $y \in \{0, 1\}$ (no:0, yes:1)

```
##   Misofonia.dic depresion.dic
## 1             4             1
## 2             2             0
## 3             0             0
## 4             3             0
## 5             0             0
## 6             0             0
```

A large study on 123 patients showed the frequencies $n_{x,y}$ given in the contingency table:

```
##
##                Depression:0 Depression:1
##   Misophonia:4            0            9
##   Misophonia:3           25            6
##   Misophonia:2           34            3
##   Misophonia:1            5            0
##   Misophonia:0           36            5
```

Let's assume that $N >> 0$ and that the frequencies **estimate** the probabilities $f_{x,y} = \hat{P}(X, Y)$

```
##
##                Depression:0 Depression:1
##   Misophonia:4   0.00000000   0.07317073
##   Misophonia:3   0.20325203   0.04878049
##   Misophonia:2   0.27642276   0.02439024
##   Misophonia:1   0.04065041   0.00000000
##   Misophonia:0   0.29268293   0.04065041
```

- What is the marginal probability of misophonia severity 3?
- What is the probability of not being misophonic **and** not depressed?
- What is the probability of being misophonic **or** depressed?
- What is the probability of being misophonic **and** depressed?
- Describe in English the outcomes with probability 0.

21.2.0.2 Exercise 2

We have performed an experiment 10 times with 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 outcome (A, B)
- Create the contingency table for the relative frequency ($f_{i,j}$) of the outcomes
- What is the marginal frequency of being male?
- What is the marginal frequency of being alive?
- What is the frequency of being alive **or** female?

21.3 Conditional Probability

21.3.0.1 Exercise 1

A machine is tested for its performance to produce high-quality turning rods. These are the results of the testing

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

- What is the estimated probability that the machine produces a rod that does not satisfy any quality control?
- What is the estimated probability that the machine produces a rod that does not satisfy at least one quality control?
- What is the estimated probability that the machine produces rounded and smoothed surfaced rods?
- what is the estimated probability that the rod is rounded if the rod is smooth?
- what is the estimated probability that the rod is smooth if it is rounded?
- what is the estimated probability that the rod is neither smooth nor rounded if it does not satisfy at least one quality control?

- Are smoothness and roundness independent events?

21.3.0.2 Exercise 2

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

- What is the probability that one lake that tests positive is contaminated with bacteria?

21.3.0.3 Exercise 3

Two machines are tested for their performance to produce high-quality turning rods. These are the results of the testing

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 rod is rounded?
- What is the probability that the rod has been produced by machine 1?
- what is the probability that the rod is not smooth?
- What is the probability that the rod is smooth or rounded or produced by machine 1?
- What is the probability that the rod is rounded if it is smoothed and from machine 1?
- What is the probability that the rod is not rounded if it is not smoothed and is from machine 2?
- what is the probability that the rod has come from machine 1 if it it is smoothed and rounded?
- what is the probability that the rod has come from machine 2 if it does not pass at least one of the quality controls?

21.3.0.4 Exercise 4

We want to cross an avenue with two traffic lights. The probability of finding the first traffic light in red is 0.6. If we stopped at the first traffic light, the probability of stopping at the second one is 0.15. Whereas the probability of stopping on the second one if we do not stop on the first one is 0.25.

When we try to cross both traffic lights:

- what is the probability of having to stop at each traffic light?
- What is the probability of having to stop at at least one traffic light?
- What is the probability of having to stop at only one traffic light?
- If I stopped at the second traffic light, what is the probability that I had to stop at the first one?

- If I had to stop at any traffic light, what is the probability that I had to do it twice?
- Is stopping at the first traffic light an independent event from stopping at the second traffic light?

Now, we want to cross an avenue with three traffic lights. The probability of finding a traffic light in red only depends on the previous one. In particular, the probability of finding one traffic light in red given that the previous one was in red is 0.15. Whereas, the probability of finding one traffic right in red given that the previous one was in green is 0.25. Also, the probability of finding the first traffic light in red is 0.6.

- What is the probability of having to stop at each traffic light?
- What is the probability of having to stop at at least one traffic light?
- What is the probability of having to stop at only one traffic light?

hints:

- If the probability that one 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 traffic lights in red can be written as: $P(R_1, R_2, R_3) = P(R_3|R_2)P(R_2|R_1)P(R_1)$

21.3.0.5 Exercise 5

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

- Pass quality test: E, do no pass 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 a test is $P(E) = 0.893$ then

- what is the probability that a brick chosen at random is defective $P(D)$?

- What is the probability that a brick that has passed the test is really defective?
- The probability that a brick is not defective **and** that it does not pass the test
- Are D and $\bar{E}$ statistical independent?

21.4 Random variables

21.4.0.1 Exercise 1

Given the probability distribution for a discrete variable X

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

a) find $f(X)$
b) find $E(X)$ and $V(X)$
c) what is the expected value and variance of $Y = 2X + 3$
d) what is the median of X?

21.4.0.2 Exercise 2

We have a system of transmission of pixels that is totally noisy. We are testing the system and have designed an experiment to transmit 3 pixels.

- What is the probability of receiving 0, 1, 2, or 3 errors in the transmission of 3 pixels?
- Sketch the probability mass function
- What is the expected value of the error?
- What is its variance?
- Sketch the probability distribution
- What is the probability of transmitting at least 1 error?

hints:

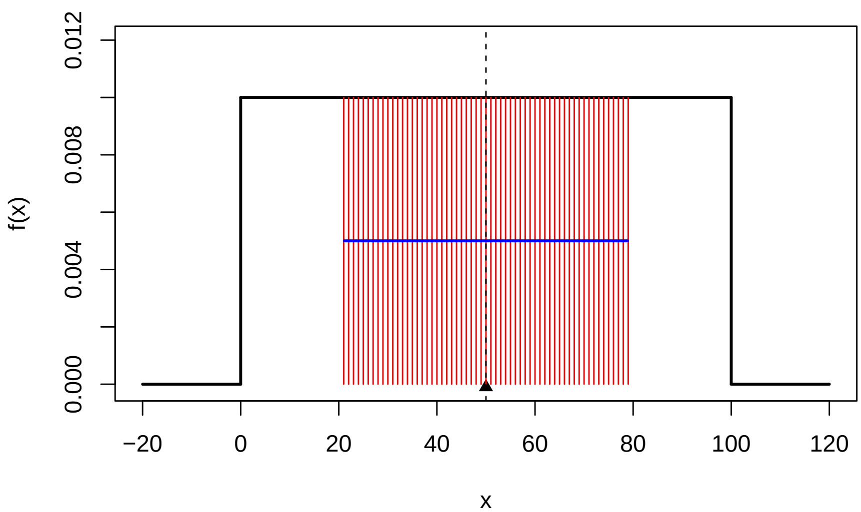
- Sample space: $\{(0,0,0), (1,0,0), (0,1,0), (0,0,1), (0,1,1), (1,0,1), (1,1,0), (1,1,1)\}$
- where, for example, the event $(0, 1, 1)$ is the event of receiving the first pixel with no error and the second and third pixels with errors.
- All events are equally probable.

21.4.0.3 Exercise 3

- for the probability density

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

- compute the mean
- compute variance using $E(X^2) = V(X) + E(X)^2$
- compute $P(\mu - \sigma \leq X \leq \mu + \sigma)$
- What are the first and third quartiles?



21.4.0.4 Exercise 4

For the probability density

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

- Confirm that this is a probability density
- Find the probability distribution $F(a)$
- Compute the mean
- Compute variance using $E(X^2) = V(X) + E(X)^2$

21.4.0.5 Exercise 5

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:

- $P(X > 0)$
- $E(X)$
- $P(X > 0 | X < 2)$

21.5 Probability Models

21.5.0.1 Exercise 1

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?
- What is the probability that the engine successfully completes 15 searches before the first failure?
- We consider that a search engine works sufficiently well when it is able to find information for 10 requests for every 2 failures. What is the probability that in a reliability trial our search engine is satisfactory?

21.5.0.2 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?
- What is the probability that a family has only one girl?
- What is the probability that a family has only one boy or only one girl?
- What is the probability that the family has at least two boys?
- What is the number of children that a family should have such that the probability of having at least a girl is more than 0.75?

21.5.0.3 Exercise 3

The average number of radioactive particles hitting a Geiger counter is 2.3 seconds.

- What is the probability of counting exactly 2 particles in a second?

- What is the probability of detecting exactly 10 particles in 5 seconds?
- What is the probability of at least one count in two seconds?
- What is the probability of having to wait 2.5 seconds after we switch on the detector?

21.5.0.4 Exercise 4

- 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?
- What is the probability that a man's height is between 165cm and 180cm.
- What is the height that defines the 5% of the smallest men?

21.6 Point Estimators

21.6.0.1 Exercise 1

Consider the probability model

$$f(x) = \begin{cases} 1/2 - a, & \text{if } x = -1 \\ 1/2, & \text{if } x = 0 \\ a, & 1\text{if } x = 1 \end{cases}$$

where a is a parameter.

Compute the mean and variance of the statistic:

$$T = \frac{\bar{X}}{2} + \frac{1}{4}$$

where $\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$

- is T a biased estimator of a?
- is T consistent? i.e. $V(T) \to 0$ when $N \to \infty$

21.6.0.2 Exercise 2

- Is $\bar{X}^2 = (\frac{1}{N}\sum_{i=1}^{N} X_i)^2$ an unbiased estimator of $E(X)^2$?

21.7 Sampling and Central Limit Theorem

21.7.0.1 Exercise 1

A battery model charges up to 75% of its capacity within an hour with a standard deviation of 15%.

- If we charge 25, what is the probability that the sample average is within a distance of 5% charge from the mean?
- If we charge 100, what is that probability?
- If, instead we only charge 9 batteries, what is the charge that is surpassed by the sample average with only 0.015 probability?

21.7.0.2 Exercise 2

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?
- How many components do we need such that the telescope is operational 2750 consecutive hours with 0.95 probability?

21.7.0.3 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?
- 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?

21.7.0.4 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?
- Which should be the sample size such that they have a probability of 0.025 that the sample mean is greater than 70 days?

21.7.0.5 Exercise 5

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?

hint: Use the CLT with a sample of 2000 Bernoulli trials. This is known as the normal approximation of the binomial distribution.

21.8 Maximum likelihood

21.8.0.1 Exercise 1

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?

21.8.0.2 Exercise 2

Take a random variable with the following probability density function

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

- 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?

21.8.0.3 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?

21.9 Method of moments

21.9.0.1 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
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

21.9.0.2 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?

21.9.0.3 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:

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

21.10 Confidence intervals

21.10.0.1 Exercise 1

In a scientific paper the authors report a 95% confidence interval of $(228, 232)$ for the natural frequency (Hz) of metallic beam. They useda sample of size 25 and considered that the measurements distributed normally.

- What is the mean and the standard deviation of the measurements?
- Compute the 99% confidence interval.

hints:

- in R $t_{0.025,24}$ = qt(0.975, 24)$\sim$ 2
- in R $t_{0.005,24}$ =qt(0.995, 24)$\sim$ 2.8

21.10.0.2 Exercise 2

compute 95% CI the mean of a normal variable with known variance $\sigma^2 = 9$ and $\bar{x} = 22$, using a sample of size 36.

21.10.0.3 Exercise 3

This year, 17 from 1000 of patients with influenza developed complications.

- Compute the 99% confidence interval for the proportion of complications.
- The previous year 2% showed complications. Can we say with 99% confidence that this year there is a significant drop in influenza complications?

21.11 Hypothesis testing

21.11.0.1 Exercise 1

Imagine we take a random sample of size $n = 41$ of a normal random variable X, and find that the sample average is 10 and the sample variance is 1.5.

- What is then the confidence interval for the mean of X at 95% confidence level?

Consider that $t_{0.025,40}$ = qt(0.975, 40) $\sim$ 2.

- Test the hypothesis that the mean of X is **different** than 10.5, using a 5% significance threshold.
- Write the code to calculate the P-value to test the hypothesis that the mean of μ is **lower** than 10.5, using a 5% significance threshold.

Consider that the code for the T probability distribution with $n-1$ degrees of freedom is pt(tobs, n-1).

21.11.0.2 Exercise 2

10 gas condensates showed the following concentrations of mercury (in ng/ml):

23.3, 22.5, 21.9, 21.5, 19.9, 21.3, 21.7, 23.8, 22.6, 24.7

Assuming that the mercury concentration is distributed normally a across gas condensates, test the hypothesis that a condensate does not surpass the toxicity limit established at $24ng/ml$.

21.11.0.3 Exercise 3

The manufacturer of gene expression microarrays guarantees that at least 97% of the microarrys they produce have high quality signals. A customer receives a batch of 200 pieces and finds that 8 unperformed.

Should the costumer return the lot due to poor quality?