Université Joseph Fourier, Grenoble I
Licence Sciences et Technologies 2e année
STA230 : Méthodes Statistiques pour la Biologie

Exercises in Statistics

http://ljk.imag.fr/membres/Bernard.Ycart/STA230/

Each theme begins with an abstract of the lecture and an exercise with detailed solution. The computations have been made with a software; due to rounding errors, there may be some minor differences with computations from statistical tables.

Contents

1 Data and models

1.1 Empirical distributions

Let $(x_1, \ldots, x_n)$ be a sample, *i.e.* a series of numerical values for a certain variable in a set of n individuals.

- The *modalities* are the different values.
- The *empirical mean* is $\overline{x} = \dfrac{1}{n}\sum_{i=1}^{n} x_i$.
- The *empirical variance* is $s_x^2 = \left(\dfrac{1}{n}\sum_{i=1}^{n} x_i^2\right) - \overline{x}^2$.
- The *empirical standard deviation* is the square root of the empirical variance.
- A sample is *centered and reduced* if its mean is 0 and its variance 1. In order to *center and reduce* a sample, substract the mean from each modality, then divide by the standard deviation.
- The *empirical frequency* of an interval is the ratio of the number of values in that interval, to the total number of individuals.
- The *median* is the smallest modality such that at least 50% of the values are smaller or equal.
- The *lower quartile* is the smallest modality such that at least 25% of the values are smaller or equal.
- The *upper quartile* is the smallest modality such that at least 75% of the values are smaller or equal.
- A statistical character is considered as *continuous* when (almost) all values are different. When for most modalities, several individuals have the same value, the character is *discrete*.

Exercise 1.1.1. Here are numbers by age of non-smoking mothers at delivery.

age	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35
number	7	8	9	10	12	3	2	5	4	5	2	4	2	0	1

1. What are the modalities ?

   The modalities are the whole numbers between 21 *and* 35.

2. Is this a discrete or a continuous variable ?

   Given the precision of the data, several individuals have the same modality (are considered as having the same age). Thus it is a discrete variable.

3. Find the empirical frequencies of the modalities.

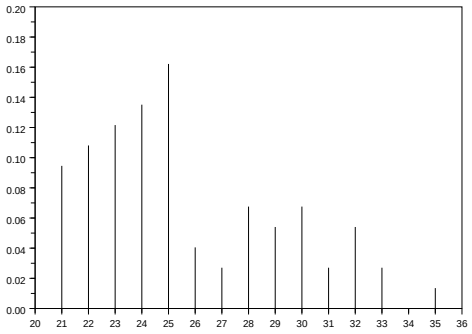
   To get the empirical frequencies, divide the numbers by the total number of individuals, which is 74.

age	21	22	23	24	25	26	27
frequency	$\frac{7}{74}$	$\frac{8}{74}$	$\frac{9}{74}$	$\frac{10}{74}$	$\frac{12}{74}$	$\frac{3}{74}$	$\frac{2}{74}$
rounded freq.	0.095	0.108	0.122	0.135	0.162	0.041	0.027

28	29	30	31	32	33	34	35
$\frac{5}{74}$	$\frac{4}{74}$	$\frac{5}{74}$	$\frac{2}{74}$	$\frac{4}{74}$	$\frac{2}{74}$	$\frac{0}{74}$	$\frac{1}{74}$
0.068	0.054	0.068	0.027	0.054	0.027	0	0.014

4. Represent the empirical frequencies on a bar chart.

   The bar chart consists of drawing a vertical segment above each modality, with height proportional to the number or to the empirical frequency.



5. Find the empirical mean, variance, and standard deviation of the sample.

   For the empirical mean:

$$\overline{x} = \frac{1}{74}\Big(7\times 21 + 8\times 22 + \cdots + 0\times 34 + 1\times 35\Big) = 25.662\ .$$

   The average age in this sample is approximately 25 *years and* 8 *months.*

   For the empirical variance:

$$s_x^2 = \frac{1}{74}\Big(7\times 21^2 + 8\times 22^2 + \cdots + 0\times 34^2 + 1\times 35^2\Big) - (25.662)^2 = 12.683\ .$$

The standard deviation is the square root of the variance:

$$s_x = \sqrt{12.683} = 3.561 \ ,$$

that is approximately 3 *years and* 7 *months.*

6. Find the values of the empirical distribution function.

   The values of the empirical distribution function are the cumulated sums of frequencies.

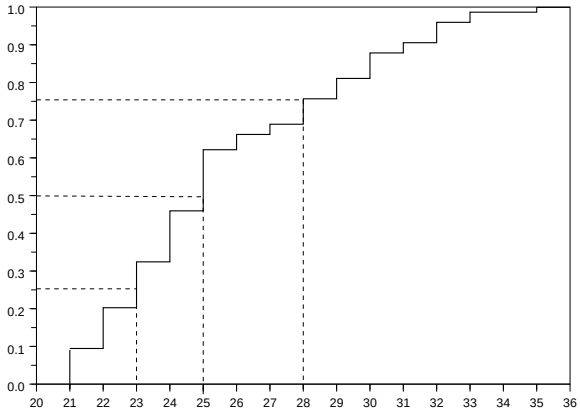
age	21	22	23	24	25	26	27
cum. freq.	$\frac{7}{74}$	$\frac{15}{74}$	$\frac{24}{74}$	$\frac{34}{74}$	$\frac{46}{74}$	$\frac{49}{74}$	$\frac{51}{74}$
rounded	0.095	0.203	0.324	0.459	0.622	0.662	0.689

28	29	30	31	32	33	34	35
$\frac{56}{74}$	$\frac{60}{74}$	$\frac{65}{74}$	$\frac{67}{74}$	$\frac{71}{74}$	$\frac{73}{74}$	$\frac{73}{74}$	$\frac{74}{74}$
0.757	0.811	0.878	0.905	0.959	0.986	0.986	1

7. What is the empirical frequency of the interval [22 ; 25]?

   It is the sum of empirical frequencies for the modalities 22, 23, 24, 25, *or else the increment of the empirical distribution function* $F(25) - F(21)$, *that is* $39/74 \simeq 0.527$. *More than half of the women in the sample are between* 22 *and* 25 *years old.*

8. Draw a graphical representation of the empirical distribution function. Determine from the graph the median and the quartiles of the sample.



The median is 25 *years; the first quartile is* 23 *years, the last quartile is* 28 *years.*

9. Compare the mean with the median, then the standard deviation with the distances between the median and the quartiles.

   The mean is larger than the median, which is normal for a distribution skewed to the right. For the same reason, the gap between the last quartile and the median is larger than that between the median and the first quartile. Both are lower than the standard deviation: this is the case for most distributions, whether they are symmetrical or skewed.

Exercise 1.1.2. Here are numbers by age of smoking birth mothers at delivery.

age	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35
number	5	5	4	3	3	5	1	4	3	2	3	2	1	1	1

1. What are the modalities ?
2. Is this a discrete or a continuous variable?
3. Find the empirical frequencies of the modalities.
4. Represent the empirical frequencies on a bar chart.
5. Find the empirical mean, variance, and standard deviation of the sample.
6. Find the values of the empirical distribution function.
7. What is the empirical frequency of the interval $[22\ ;\ 25]$?
8. Draw a graphical representation of the empirical distribution function. Determine from the graph the median and the quartiles of the sample.
9. Compare the mean with the median, then the standard deviation with the distances between the median and the quartiles.

Exercise 1.1.3. Consider the sample $(1, 0, 2, 1, 1, 0, 1, 0, 0)$.

1. What is its empirical mean?
2. What is its empirical variance?
3. Center and reduce this sample.
4. If you had to propose a model for these data: would you choose a discrete or a continuous model?

Exercise 1.1.4. Consider the sample

$$(1.2, 0.2, 1.6, 1.1, 0.9, 0.3, 0.7, 0.1, 0.4)\ .$$

1. What is its empirical mean?
2. What is its empirical variance?
3. Center and reduce this sample.
4. If you had to propose a model for these data: would you choose a discrete or a continuous model?

1.2 Probabilities and conditional probabilities

- The *probability of an event* in a population is the proportion of individuals for which the event is true.
- The *conditional probability of A knowing B*, is the proportion of individuals for which A is true *among those for which B is also true.* It is the ratio of the probability of "A and B" to the probability of B:

$$\mathbb{P}[A \,|\, B] = \frac{\mathbb{P}[A \text{ and } B]}{\mathbb{P}[B]} \ .$$

- The *total probability formula* gives the probability of an event A as a function of the conditional probabilities knowing another event B and its contrary $\overline{B}$:

$$\mathbb{P}[A] = \mathbb{P}[A \,|\, B]\,\mathbb{P}[B] + \mathbb{P}[A \,|\, \overline{B}]\,\mathbb{P}[\overline{B}] \ .$$

- The *Bayes formula* exchanges the order of conditional probabilities:

$$\mathbb{P}[B \,|\, A] = \frac{\mathbb{P}[A \,|\, B]\,\mathbb{P}[B]}{\mathbb{P}[A \,|\, B]\,\mathbb{P}[B] + \mathbb{P}[A \,|\, \overline{B}]\,\mathbb{P}[\overline{B}]} \ .$$

Exercise 1.2.1. In a sheep breeding farm, an estimated 30% of the sheep suffer from a certain disease. A test for this disease is available. If a sheep is not ill, it has 9 chances out of 10 to react negatively to the test; if it is ill, it has 8 chances out of 10 to have a positive reaction. All the sheep in the farm are submitted to the test.

Throughout the exercise, the event "the sheep is ill" will be denoted by M, and the event "the sheep has a positive reaction" by T. The text gives:

$$\mathbb{P}[M] = 0.3 \ , \quad \mathbb{P}[\overline{T} \,|\, \overline{M}] = 0.9 \ , \quad \mathbb{P}[T \,|\, M] = 0.8 \ .$$

1. What is the probability for a sheep in that farm not to be ill?

$$\mathbb{P}[\overline{M}] = 1 - \mathbb{P}[M] = 1 - 0.3 = 0.7 \ .$$

2. What is the conditional probability for a sheep to have a positive reaction knowing that it is not ill?

$$\mathbb{P}[T \,|\, \overline{M}] = 1 - \mathbb{P}[\overline{T} \,|\, \overline{M}] = 1 - 0.9 = 0.1 \; .$$

3. What is the probability for a sheep not to be ill and have a positive reaction?

$$\mathbb{P}[T \text{ and } \overline{M}] = \mathbb{P}[T \,|\, \overline{M}]\, \mathbb{P}[\overline{M}] = 0.1 \times 0.7 = 0.07 \; .$$

4. What proportion of the sheep have a positive reaction?

   Use the formula of total probabilities or compute it straight away, by distinguishing among those sheep reacting positively, those which are ill from those which are not.

$$\begin{aligned} \mathbb{P}[T] &= \mathbb{P}[T \text{ and } M] + \mathbb{P}[T \text{ and } \overline{M}] \\ &= \mathbb{P}[T \,|\, M]\, \mathbb{P}[M] + \mathbb{P}[T \,|\, \overline{M}]\, \mathbb{P}[\overline{M}] \\ &= 0.8 \times 0.3 + 0.1 \times 0.7 = 0.24 + 0.07 = 0.31 \; . \end{aligned}$$

5. What is the probability for a sheep to be ill, knowing that it has reacted positively?

   Use the Bayes formula or prove it again as follows.

$$\begin{aligned} \mathbb{P}[M \,|\, T] &= \frac{\mathbb{P}[T \text{ and } M]}{\mathbb{P}[T]} \\ &= \frac{\mathbb{P}[T \,|\, M]\, \mathbb{P}[M]}{\mathbb{P}[T \,|\, M]\, \mathbb{P}[M] + \mathbb{P}[T \,|\, \overline{M}]\, \mathbb{P}[\overline{M}]} \\ &= \frac{0.8 \times 0.3}{0.8 \times 0.3 + 0.1 \times 0.7} \simeq 0.774 \; . \end{aligned}$$

6. What is the probability for a sheep not to be ill, knowing it has reacted negatively?

   Use the Bayes formula or prove it again as follows.

$$\begin{aligned} \mathbb{P}[\overline{M} \,|\, \overline{T}] &= \frac{\mathbb{P}[\overline{T} \text{ and } \overline{M}]}{\mathbb{P}[\overline{T}]} \\ &= \frac{\mathbb{P}[\overline{T} \,|\, \overline{M}]\, \mathbb{P}[\overline{M}]}{\mathbb{P}[\overline{T} \,|\, \overline{M}]\, \mathbb{P}[\overline{M}] + \mathbb{P}[\overline{T} \,|\, M]\, \mathbb{P}[M]} \\ &= \frac{0.9 \times 0.7}{0.9 \times 0.7 + 0.2 \times 0.3} \simeq 0.913 \; . \end{aligned}$$

Exercise 1.2.2. There are three sorts of a given plant: early, normal, and late. It can also be either dwarf or tall. In a sample of plants grown from 1000 seeds, there are 600 dwarf, 200 late, 300 early dwarf, 250 normal tall, 100 late tall. Consider the plant grown from a seed taken at random.

1. What is the probability that it is early? normal? late? dwarf? tall?
2. A dwarf plant is observed. What is the probability that it is early? normal? late?
3. A tall plant is observed. What is the probability that it is early? normal? late?
4. A late plant is observed. What is the probability that it is dwarf? tall?

Exercise 1.2.3. In a batch of manufactured items, 5% are faulty. The items are checked, but the checking is not perfect. If the item is good, it is accepted with probability 0.96; if it is faulty, it is rejected with probability 0.98. An item is chosen at random, then checked.

1. What is the probability that this item is rejected?
2. What is the probability that it is good, knowing that it has been rejected?
3. What is the probability that it is faulty, knowing that it has been accepted?
4. What is the probability that there is an error in the checking (the item is good and rejected or bad and accepted)?

Exercise 1.2.4. Here are the percentages of the different blood types in France.

Group / Factor	O	A	B	AB
Rhesus +	37.0	38.1	6.2	2.8
Rhesus −	7.0	7.2	1.2	0.5

1. Determine the probability distribution of the four groups O, A, B, AB in the French population.
2. Determine the probability distribution of the four groups among the persons with positive rhesus
3. Determine the probability distribution of the four groups among the persons with negative rhesus
4. If a person of group O is chosen at random, what is the probability that he/she has a negative rhesus? Same question for a person of group B.

1.3 Binomial distribution

- When n *experiments are repeated independently*, the random variable X equal to the number of realizations of a given *event of probability* p, follows the *binomial* distribution with parameters n and p.
- The variable X may take all integer values between 0 and n.
- For any integer k between 0 and n, the variable X takes value k with probability:

$$\mathbb{P}[X = k] = \binom{n}{k} p^k (1-p)^{n-k} ,$$

where

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} = \frac{n \times (n-1) \cdots \times (n-k+1)}{k \times (k-1) \cdots \times 3 \times 2 \times 1}$$

is the number of ways to choose k objects among n.

- The expectation of X is np, its variance is $np(1-p)$.

Exercise 1.3.1. From past experience, it is known that a certain surgery has a 90% chance to succeed. This surgery is going to be performed on 5 patients. Let X be the random variable equal to the number of successes out of the 5 attempts.

1. What model do you propose for X?

   Assuming that the outcome (success or failure) of the 5 attempts are independent, the number of successes follows the binomial distribution with parameters 5 *and* 0.9. *The random variable X takes its values in the set* $\{0, 1, 2, 3, 4, 5\}$, *and for any integer k in that set:*

$$\mathbb{P}[X = k] = \binom{5}{k} 0.9^k \, 0.1^{5-k} .$$

2. What is the probability that the surgery will fail all 5 times?

$$\mathbb{P}[X = 0] = 0.1^5 = 0.00001 .$$

3. What is the probability for the surgery to fail exactly 3 times?

$$\mathbb{P}[X = 2] = \binom{5}{2} 0.9^2 \, 0.1^3 = 0.0081 .$$

4. What is the probability for the surgery to succeed at least 3 times?

$$\begin{aligned}\mathbb{P}[X \geqslant 3] &= \mathbb{P}[X=3] + \mathbb{P}[X=4] + \mathbb{P}[X=5] \\ &= \binom{5}{3} 0.9^3\, 0.1^2 + \binom{5}{4} 0.9^4\, 0.1^1 + \binom{5}{5} 0.9^5\, 0.1^0 \\ &= 0.0729 + 0.32805 + 0.59049 = 0.99144\ .\end{aligned}$$

Exercise 1.3.2. When a hunter aims at a helpless rabbit, he has 1 chance out of 10 to hit it.

1. Two hunters aim independently at the same rabbit. Find the probability that:
   (a) neither of them hit;
   (b) only one of them hits;
   (c) both hunters hit.
2. Four hunters aim independently at the same rabbit.
   (a) What is the probability distribution of the number of shots suffered by the poor animal? Give the expectation and variance of that distribution.
   (b) What is the probability that the rabbit is hit at most twice?
   (c) What is the probability that the rabbit is hit at least twice?
3. Ten hunters aim independently at the same rabbit.
   (a) What is the probability for the rabbit not to be hit?
   (b) What is the probability that the rabbit becomes inedible (if it has received at least 5 shots).

Exercise 1.3.3. At an identification session, 6 witnesses are asked to identify a murderer among 4 suspects, including yourself.

1. If each one of the 6 witnesses chooses at random, what are your chances:
   (a) of not being pointed out?
   (b) of being pointed out exactly once?
   (c) of being pointed out twice or more?
2. It turns out that 2 of the 6 witnesses have identified you as the murderer. Referring to 1 (c), do you expect that the judge will think that this may be due to chance?
3. What if 4 of the 6 witnesses have identified you?

1.4 Hypergeometric distribution

- In a set of N elements, among which m have been marked, n distinct elements are selected at random. The random variable X, equal to the number of marked elements among the selected n, follows the *hypergeometric distribution with parameters* N, m, n.
- In the case where $n \leq m$ and $n \leq N - m$, X may take all integer values between 0 and n.
- For any integer k between 0 and n, X takes value k with probability:

$$\mathbb{P}[X = k] = \frac{\binom{m}{k}\binom{N-m}{n-k}}{\binom{N}{n}} .$$

- The expectation of X is nm/N.

Exercise 1.4.1. There are 18 girls and 11 boys in a certain group of students. A sample of 5 persons is chosen at random in that group. Let X be the random variable equal to the number of girls in that sample.

1. What model do you propose for X?

   The distribution of X is the hypergeometric distribution with parameters $N = 29$ (total number of persons), $m = 18$ (the "marked" individuals are the girls), and $n = 5$ (the size of the sample). The values are the integers between 0 and 5. For any integer $k = 0, 1 \ldots, 5$:

$$\mathbb{P}[X = k] = \frac{\binom{18}{k}\binom{11}{5-k}}{\binom{29}{5}} .$$

2. Give the expectation of X.

   The expectation of X is $5 \times 18/29 \simeq 3.1$. It is the size of the sample, multiplied by the proportion of girls in the group.

3. Find the probability of having only girls in the sample.

$$\mathbb{P}[X = 5] = \frac{\binom{18}{5}}{\binom{29}{5}} \simeq 0.072 .$$

4. Find the probability of having at least one girl in the sample.

   The value of $\mathbb{P}[X \geqslant 1]$ must be calculated. It could be done as $\mathbb{P}[X = 1] + \mathbb{P}[X =$

$2] + \mathbb{P}[X = 3] + \mathbb{P}[X = 4] + \mathbb{P}[X = 5]$, *but it is quicker to calculate* $1 - \mathbb{P}[X = 0]$, *which is the same:*

$$\mathbb{P}[X \geqslant 1] = 1 - \mathbb{P}[X = 0] = 1 - \frac{\binom{11}{5}}{\binom{29}{5}} \simeq 0.996 \, .$$

5. Find the probability for the sample to have exactly 3 girls.

$$\mathbb{P}[X = 3] = \frac{\binom{18}{3}\binom{11}{2}}{\binom{29}{5}} \simeq 0.378 \, .$$

Exercise 1.4.2. In each of the following situations, give the probability distribution of the random variable X and its expectation. Find the probability for X to be 0, then the probability for X to be 2 or more.

1. At a card table, 8 cards are delt to each of the 4 players, out of a deck of 32. Let X be the number of aces received by a given player.

2. At a belote table, the four players make teams of two. Let X be the number of diamonds of a given team.

3. At a bridge table, thirteen cards are handed out to each of the four players. Let X be the number of figures (jack, queen, or king) of a given player.

4. On a loto card, you ticked 6 numbers out of an array of 49. Let X be the number of good numbers ticked on your card.

1.5 Normal distribution

- If no software is available, the following data for the normal distribution with mean 0 and variance 1, denoted by $\mathcal{N}(0, 1)$, are given in the tables:
  - ⋆ the values of the distribution function F: for a value of x, the table gives the probability $p = P[X \leqslant x] = F(x)$.
  - ⋆ the values of the quantile function $F^{-1}(p)$: for a probability p the table gives the value $x = F^{-1}(p)$ such that $p = \mathbb{P}[X \leqslant x]$.
- The density of the $\mathcal{N}(0, 1)$ distribution is symmetric:

$$\mathbb{P}[X \leqslant -x] = \mathbb{P}[X \geqslant x] \, .$$

- If a random variable X follows the $\mathcal{N}(\mu, \sigma^2)$ distribution, then $(X-\mu)/\sqrt{\sigma^2}$ follows the $\mathcal{N}(0,1)$ distribution. Thus:

$$\begin{aligned}\mathbb{P}[a \leqslant X \leqslant b] &= P\left[\frac{a-\mu}{\sqrt{\sigma^2}} \leqslant \frac{X-\mu}{\sqrt{\sigma^2}} \leqslant \frac{b-\mu}{\sqrt{\sigma^2}}\right] \\ &= F\left(\frac{b-\mu}{\sqrt{\sigma^2}}\right) - F\left(\frac{a-\mu}{\sqrt{\sigma^2}}\right) ,\end{aligned}$$

  where F is the distribution function of the $\mathcal{N}(0,1)$.

- If X and Y are two independent random variables, with respective distributions $\mathcal{N}(\mu_x, \sigma_x^2)$ and $\mathcal{N}(\mu_y, \sigma_y^2)$, then $X+Y$ follows the $\mathcal{N}(\mu_x+\mu_y, \sigma_x^2+\sigma_y^2)$ and $X-Y$ follows the $\mathcal{N}(\mu_x-\mu_y, \sigma_x^2+\sigma_y^2)$.

Exercise 1.5.1. The height X of men in France is modeled by a normal distribution $\mathcal{N}(172, 196)$ (unit: cm).

1. What proportion of French men are less than 160 cm tall?

   $$\mathbb{P}[X<160] = \mathbb{P}\left[\frac{X-172}{\sqrt{196}} < \frac{160-172}{\sqrt{196}}\right] = F(-0.857) = 1-F(0.857) = 0.1957 ,$$

   where F denotes the distribution function of the $\mathcal{N}(0,1)$ distribution.

2. What proportion of French men are more than two meters tall?

   $$\mathbb{P}[X>200] = \mathbb{P}\left[\frac{X-172}{\sqrt{196}} > \frac{200-172}{\sqrt{196}}\right] = 1-F(2) = 0.02275 .$$

3. What proportion of French men are between 165 and 185 centimeters tall?

   $$\begin{aligned}\mathbb{P}[165<X<185] &= \mathbb{P}\left[\frac{165-172}{\sqrt{196}} < \frac{X-172}{\sqrt{196}} < \frac{185-172}{\sqrt{196}}\right] \\ &= F(0.928) - F(-0.5) = 0.8234 - 0.3085 = 0.5149 .\end{aligned}$$

4. If ten thousand French men chosen at random were ranked by increasing size, how tall would be the 9000-th?

   The question amounts to finding the size such that 90% *of the French are smaller, i.e. the* 90*-th quantile of the ninth decile. Let x be that size.*

   $$\mathbb{P}[X<x] = \mathbb{P}\left[\frac{X-172}{\sqrt{196}} < \frac{x-172}{\sqrt{196}}\right] = 0.9$$

   Thus $\frac{x-172}{\sqrt{196}}$ is the value of the quantile function of the $\mathcal{N}(0,1)$ distribution for $p=0.9$, that is 1.2816*. Therefore:*

   $$x = 172 + 1.2816 \times \sqrt{196} \simeq 190 \text{ cm.}$$

5. The height of French women is modeled by a normal distribution $\mathcal{N}(162, 144)$ (in centimeters). What is the probability for a French man chosen at random to be taller than a French woman chosen at random?

   Let X denote the size of the man and Y that of the woman, and suppose they are independent. Then $X - Y$ follows the normal distribution $\mathcal{N}(10, 340)$. The probability for X to be larger than Y is the probability for $X - Y$ to be positive:

$$\mathbb{P}\left[X - Y > 0\right] = \mathbb{P}\left[\frac{(X-Y)-10}{\sqrt{340}} > \frac{0-10}{\sqrt{340}}\right] = 1 - F(-0.5423) = 0.7062\ .$$

Exercise 1.5.2. Let X be a random variable with $\mathcal{N}(0, 1)$ distribution.

1. Express with the distribution function of X, then compute using the table the following probabilities.
   (a) $\mathbb{P}[X > 1.45]$
   (b) $\mathbb{P}[-1.65 \leqslant X \leqslant 1.34]$
   (c) $\mathbb{P}[|X| < 2.05]$
2. Find the value of u in the following cases.
   (a) $\mathbb{P}[X < u] = 0.63$
   (b) $\mathbb{P}[X \geqslant u] = 0.63$
   (c) $\mathbb{P}[|X| < u] = 0.63$

Exercise 1.5.3. Let X be a random variable with $\mathcal{N}(0, 1)$ distribution. Let $Y = 2X - 3$.

1. What is the distribution of Y?
2. Find $\mathbb{P}[Y < -4]$.
3. Find $\mathbb{P}[-2 < Y < 3]$.

Exercise 1.5.4. Let X be a random variable with $\mathcal{N}(3, 25)$ distribution.

1. Express with the distribution function of the $\mathcal{N}(0, 1)$ distribution, then compute using the table the following probabilities.
   (a) $\mathbb{P}[X < 6]$
   (b) $\mathbb{P}[X > -2]$
   (c) $\mathbb{P}[-1 \leqslant X \leqslant 1.5]$
2. Find the value of u in the following cases.

(a) $\mathbb{P}[X < u] = 0.63$

(b) $\mathbb{P}[X > u] = 0.63$

(c) $\mathbb{P}[|X - 3| \leqslant u] = 0.63$

Exercise 1.5.5. In a given country, the cholesterol concentration of a person taken at random is modeled by a normal distribution with mean 200 mg/100 mL and standard deviation 20 mg/100 mL.

1. What is the probability that a person taken at random in that country has a cholesterol rate below 160 mg/100 mL?

2. What proportion of the population has a cholesterol rate between 170 and 230 mg/100 mL?

3. In another country, the mean cholesterol rate is 190 mg/100 mL, for the same standard deviation as before. Answer the previous questions for this other country.

4. A person is taken at random in each country. What is the probability for the person from the first country to have a higher cholesterol rate than the person from the second country?

Exercise 1.5.6. The size of an ear of wheat in a field is modeled by a random variable X with normal distribution $\mathcal{N}(15, 36)$ (unit: cm).

1. What is the probability for an ear to be smaller than 16 cm?

2. There are about 15 million ears in the field. Give an estimate of the number of ears larger than 20 cm.

3. A sample of 10 ears is picked up in the field. What is the probability that all of them have sizes in the interval $[16\ ;\ 20]$?

4. In another field, the size of an ear of wheat is modeled by a random variable Y with normal distribution $\mathcal{N}(10, 16)$. What is the probability for an ear from the first field to be larger than an ear from the second field?

1.6 Approximation of a binomial to a normal distribution

- If n is large enough, the binomial distribution $\mathcal{B}(n, p)$ can be approximated to the normal distribution $\mathcal{N}(np, np(1-p))$, having the same expectation and variance.

- In that case, if X follows the $\mathcal{B}(n,p)$ distribution, one computes the probability for X to be in the interval $[a,b]$ by:

$$\begin{aligned}\mathbb{P}[a \leqslant X \leqslant b] &= P\left[\frac{a-np}{\sqrt{np(1-p)}} \leqslant \frac{X-np}{\sqrt{np(1-p)}} \leqslant \frac{b-np}{\sqrt{np(1-p)}}\right] \\ &\simeq F\left(\frac{b-np}{\sqrt{np(1-p)}}\right) - F\left(\frac{a-np}{\sqrt{np(1-p)}}\right) ,\end{aligned}$$

  where F is the distribution function of the $\mathcal{N}(0,1)$.

Exercise 1.6.1. From past experience, it is known that a certain surgery has a 90% chance to succeed. This surgery is performed by a certain clinic 400 times each year. Let N be the number of successes next year. The normal approximation will be used for N.

1. Find the expectation and variance of N.

   The expectation is $400 \times 0.9 = 360$*, the variance is* $400 \times 0.9 \times 0.1 = 36$*.*

2. Find the probability for the clinic to perform successfully the surgery at least 345 times.

$$\begin{aligned}\mathbb{P}[N \geqslant 345] &= \mathbb{P}\left[\frac{N-360}{\sqrt{36}} \geqslant \frac{345-360}{\sqrt{36}}\right] \\ &= 1 - F(-2.5) = F(2.5) = 0.9938 .\end{aligned}$$

3. Find the probability that the surgery fails in that clinic more than 28 times in the year.

$$\begin{aligned}\mathbb{P}[N \leqslant 372] &= \mathbb{P}\left[\frac{N-360}{\sqrt{36}} \leqslant \frac{372-360}{\sqrt{36}}\right] \\ &= F(2) = 0.9772 .\end{aligned}$$

4. The insurance accepts to cover a certain number of failed surgeries: that number has only a 1% chance to be exceeded. What number is it?

   Let n *be the number of failed surgeries that must be determined. The corresponding number of successes is* $400-n$*. Therefore* $\mathbb{P}[N \leqslant 400-n] = 0.01$*. Now:*

$$\begin{aligned}\mathbb{P}[N \leqslant 400-n] &= \mathbb{P}\left[\frac{N-360}{\sqrt{36}} \leqslant \frac{400-n-360}{\sqrt{36}}\right] \\ &= F\left(\frac{40-n}{\sqrt{36}}\right) = 0.01 .\end{aligned}$$

The number $\frac{40-n}{\sqrt{36}}$ *is the quantile at* 0.01 *of the* $\mathcal{N}(0,1)$*, that is* -2.3236*. Thus:*

$$\frac{40-n}{\sqrt{36}} = -2.3263 \implies n = 40 + 2.3263\sqrt{36} \simeq 54 \ .$$

The reasoning could also be applied to the number of failed surgeries $R = 400-N$*. It follows the binomial distribution* $\mathcal{B}(400, 0.1)$*, that can be approximated by the normal* $\mathcal{N}(40, 36)$*. The desired number is such that* $\mathbb{P}[R > n] = 0.01$*.*

$$\begin{aligned}
\mathbb{P}[R > n] &= \mathbb{P}\left[\frac{R-40}{\sqrt{36}} > \frac{n-40}{\sqrt{36}}\right] \\
&= 1 - F\left(\frac{n-40}{\sqrt{36}}\right) \\
&= F\left(\frac{40-n}{\sqrt{36}}\right) = 0.01 \ .
\end{aligned}$$

Of course the result is the same.

Exercise 1.6.2. Among people old enough to receive an injection against the flu, 40% of them ask for it. In a population of 150000 persons old enough to receive the injection, let N be the number of those that will ask for it.

1. What model would you propose for N?
2. If 60500 syringes are prepared, what is the probability that these will not suffice?
3. Find the number of syringes that should be prepared to ensure that there will be enough with 90% probability at least.

Exercise 1.6.3. A restaurant, serving only upon reservation, has 50 seats. The probability that a someone with a reservation does not show up is 1/5. Let N be the number of meals served on a given day. The normal approximation will be used for N.

1. If the chef accepts 50 reservations, what is the probability he will serve more than 45 meals?
2. If he accepts 55 reservations, what is the probability he will find himself in an embarassing situation?

Exercise 1.6.4. Suppose there is probability 0.1 of being controlled in the tramway. Mr A. makes 700 trips per year. The normal approximation will be used for the number of fraud checks.

1. Find the probability that Mr. A will be controled between 60 and 80 times in the year.

2. Mr A. always travels without paying. Knowing that the price of a ticket is 1 euro, what minimal fine should the transportation company charge if they wanted Mr. A to have, over a 1 year period, a probability 0.75 of spending more than he would were he honest.

Exercise 1.6.5. Between Grenoble and Valence TGV, two buses of 50 seats depart on fridays at 4:10 pm. The number of travellers showing up for the trip has a mean of 80 and a standard deviation of 10. The normal approximation will be used for that number.

1. Find the probability for the two buses to be full.

2. One of the buses departs from the station, the other from Victor Hugo square. The passengers choose one or the other at random, but they cannot change if the bus is full. Assume 90 passengers want to go from Grenoble to Valence. What is the probability that at least one of them cannot ?

3. With the same hypotheses as in the previous questions, what should the size of the buses be in order to ensure that the probability of turning down a passenger is lower than 0.05?

Exercise 1.6.6. On average, a passenger that has bought a plane ticket shows up at registration with probability 0.9. A given plane has two hundred seats.

1. If the airline company accepts 220 reservations, what is the probability it will have to turn passengers away?

2. How many reservations should it accept at most to make sure that the probability of turning down at least one passenger is no larger than 0.01?

2 Parametric estimation

2.1 Estimating a parameter

- For an unknown parameter, an estimator is a function of the data, taking values close to that parameter. It is *unbiased* if its expectation is equal to the parameter. It is *convergent* if the probability for it to take a value at distance up to ε from the parameter tends to 1 as the size of the sample tends to infinity.
- The *empirical frequency* of an event is an unbiased convergent estimator of the probability of that event.
- The *empirical mean* of a sample is an unbiased convergent estimator of the theoretical expectation of the variables.
- The *empirical variance* of a sample is a convergent estimator of the theoretical variance of the variables. Un unbiased estimator is obtained by multiplying the empirical variance by $n/(n-1)$, where n is the size of the sample.

Exercise 2.1.1. Consider the statistical sample $(1, 0, 2, 1, 1, 0, 1, 0, 0)$.

1. Find its empirical mean and variance.

   $$\overline{x} = \frac{6}{9} = \frac{2}{3} \quad \textit{and} \quad s_x^2 = \frac{4}{9}\,.$$

2. Supposing that the data are realizations of a variable with an unknown distribution, give unbiased estimates for the expectation and the variance of that distribution.

   The empirical mean ($2/3$) is an unbiased estimate of the expectation. An unbiased estimate of the variance is obtained, multiplying s_x^2 by $9/8$: this gives $1/2$.

3. The data of the sample are modeled by a binomial distribution $\mathcal{B}(2, p)$. Use the empirical mean to propose an estimate for p.

   The expectation of the $\mathcal{B}(2, p)$ distribution is $2p$. It is estimated by the empirical mean (here $2/3$). Thus p can be estimated by:

   $$\frac{2/3}{2} = \frac{1}{3}\,.$$

4. With the same model, use the empirical variance to propose another estimate for p.

   The variance of the $\mathcal{B}(2, p)$ distribution is $2p(1-p)$. It is estimated by $1/2$. The value of p can be estimated by solving the equation $2p(1-p) = 1/2$, giving $p = 1/2$.

5. The data of the sample are now modeled by a Poisson distribution $\mathcal{P}(\lambda)$, the expectation of which is λ. What estimate would you propose for λ?

   The parameter λ can be estimated by the empirical mean, $2/3$.

Exercise 2.1.2. Consider the statistical sample $(1, 3, 2, 3, 2, 2, 0, 2, 3, 1)$.

1. Supposing that the variables are realizations of a variable with unknown distribution, give unbiased estimates for the expectation and the variance of that distribution.

2. The data of that sample are modeled by a $\mathcal{B}(3, p)$ distribution. Use the empirical mean to propose an estimate for p.

Exercise 2.1.3. Consider the statistical sample $(1.2, 0.2, 1.6, 1.1, 0.9, 0.3, 0.7, 0.1, 0.4)$.

1. The data of that sample are modeled by a uniform distribution on the interval $[0, \theta]$. What estimate would you propose for θ?

2. The data of the sample are now modeled by a normal distribution $\mathcal{N}(\mu, \sigma^2)$. What estimates would you propose for μ and σ^2?

2.2 Confidence intervals for a Gaussian sample

A Gaussian sample is a n-tuple $(X_1, \ldots, X_n)$ of independent random variables with normal distribution $\mathcal{N}(\mu, \sigma^2)$. The empirical mean and variance of the sample are given by:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{et} \quad S^2 = \left(\frac{1}{n}\sum_{i=1}^{n} X_i^2\right) - \overline{X}^2 \ ,$$

- If the theoretical variance σ^2 is *known*, a confidence interval at level $1-\alpha$ for μ is obtained by:

$$\left[\overline{X} - u_\alpha \frac{\sqrt{\sigma^2}}{\sqrt{n}} \ ; \ \overline{X} + u_\alpha \frac{\sqrt{\sigma^2}}{\sqrt{n}}\right] \ ,$$

  where u_α is the quantile of order $1-\alpha/2$ for the normal distribution $\mathcal{N}(0, 1)$.

- If the theoretical variance σ^2 is *unknown*, a confidence interval at level $1-\alpha$ for μ is obtained by:

$$\left[\overline{X} - t_\alpha \frac{\sqrt{S^2}}{\sqrt{n-1}} \ ; \ \overline{X} + t_\alpha \frac{\sqrt{S^2}}{\sqrt{n-1}}\right] \ ,$$

  where t_α is the quantile of order $1-\alpha/2$ for the Student distribution with parameter $n-1$.

- If the theoretical variance σ^2 is *unknown*, a confidence interval at level $1-\alpha$ for σ^2 is obtained by:

$$\left[\frac{nS^2}{v_\alpha}\,;\,\frac{nS^2}{u_\alpha}\right]\,,$$

where u_α is the quantile of order $\alpha/2$ for the chi-squared distribution with parameter $n-1$, and v_α is its quantile of order $1-\alpha/2$.

Exercise 2.2.1. The compression force of a certain type of concrete is modeled by a Gaussian random variable with expectation μ and variance σ^2. The measurement unit is the *psi* (pound per square inch). In questions 1. to 4., it will be supposed that the variance σ^2 is known and equal to 1000. An empirical mean of 3250 psi has been observed from a sample of 12 measurements.

1. Give a 95% confidence interval for μ.

   Here, $\alpha = 0.05$ *and* $1-\alpha/2 = 0.975$. *The quantile of order* 0.975 *for the* $\mathcal{N}(0,1)$ *distribution is* 1.96. *The confidence interval is:*

$$\left[3250 - 1.96\frac{\sqrt{1000}}{\sqrt{12}}\,;\,3250 + 1.96\frac{\sqrt{1000}}{\sqrt{12}}\right] \;=\; [3232\,;\,3268]\,.$$

   There is no point in giving more digits than in the empirical mean. The lower bound is rounded to the left, and the upper bound to the right; thus the rounding can only enlarge the interval, ensuring that the confidence level remains higher than 0.95.

2. Give a 99% confidence interval for μ. Compare its width with that of the interval in the previous question.

   Here, $\alpha = 0.01$ *and* $1-\alpha/2 = 0.995$. *The quantile of order* 0.995 *for the* $\mathcal{N}(0,1)$ *distribution is* 2.5758. *The confidence interval is:*

$$\left[3250 - 2.5758\frac{\sqrt{1000}}{\sqrt{12}}\,;\,3250 + 2.5758\frac{\sqrt{1000}}{\sqrt{12}}\right] \;=\; [3226\,;\,3274]\,.$$

   *This interval is wider than the previous one. The higher the probability that the mean belong to the interval (*0.99 *instead of* 0.95*), the wider the interval must be. To get greater confidence, less precision must be accepted.*

3. If using the same sample, a confidence interval of width 30 psi were given, what would its confidence level be?

   The width of a confidence interval at level $1-\alpha$ *is:*

$$2u_\alpha\frac{\sqrt{1000}}{\sqrt{12}}\,.$$

If that width is 30*, then:*

$$u_\alpha = \frac{30\sqrt{12}}{2\sqrt{1000}} = 1.6432 \ .$$

This value is the quantile of order $0.9498 = 1-\alpha/2$ *for the* $\mathcal{N}(0,1)$*. Thus* $\alpha = 0.1003$ *and* $1-\alpha = 0.8997$*.*

4. What minimal number of trials would be necessary to estimate μ with a precision of ± 15 psi, at confidence level 0.95?

   For n *trials, the precision of the confidence interval at level* 0.95 *is:*

$$\pm 1.96 \frac{\sqrt{1000}}{\sqrt{n}} \ .$$

   If it is ± 15*, then:*

$$n = \left(\frac{1.96\sqrt{1000}}{15} \right)^2 = 17.07 \ .$$

   The sample size must be at least 18*.*

5. From now on the theoretical variance is supposed to be unknown. For the 12 trials mentioned above:

$$\sum_{i=1}^{12} x_i^2 = 126761700 \ .$$

   Give a 95% confidence interval for μ and compare it with that of question 1. Repeat the calculation for a 99% confidence interval and compare it with that of question 2.

   The estimated variance is:

$$s^2 = \frac{1}{12} \times 126761700 - (3250)^2 = 975 \ .$$

   The quantile of order 0.975 *for the Student* $\mathcal{T}(n-1)$ *distribution is* 2.201*, that of order* 0.995 *is* 3.106*. The* 95% *confidence interval is:*

$$\left[3250 - 2.201 \frac{\sqrt{975}}{\sqrt{11}} \ ; \ 3250 + 2.201 \frac{\sqrt{975}}{\sqrt{11}} \right] \ = \ [3229 \, ; \, 3271] \ .$$

   The 99% *confidence interval is:*

$$\left[3250 - 3.106 \frac{\sqrt{975}}{\sqrt{11}} \ ; \ 3250 + 3.106 \frac{\sqrt{975}}{\sqrt{11}} \right] \ = \ [3220 \, ; \, 3280] \ .$$

Even though the estimated variance is lower than the theoretical one in this case, the confidence intervals calculated with the Student distribution (unknown variance) are wider thus less precise than those computed with the normal distribution (known variance). This is due to the fact that Student distributions are more scattered than the $\mathcal{N}(0,1)$ *one: the interval containing* 95% *of values for the* $\mathcal{T}(11)$ *is* $[-2.201\ ;\ +2.201]$*, instead of* $[-1.96\ ;\ +1.96]$ *for the* $\mathcal{N}(0,1)$*. It is reasonable to expect less precision when less information is available on the model.*

6. Give a 95% confidence interval for the variance, and for the standard deviation.

   The quantile of order 0.025 *for the khi-squared distribution* $\mathcal{X}^2(11)$ *is* $u_\alpha = 3.816$*. The quantile of order* 0.975 *is* $v_\alpha = 21.92$*. The* 95% *confidence interval for the variance is:*

$$\left[\frac{12 \times 975}{21.92}\ ;\ \frac{12 \times 975}{3.816}\right] \ = \ [533\,;\,3067]\ .$$

   By taking the square root of both bounds, a confidence interval for the standard deviation is obtained:

$$\left[\sqrt{\frac{12 \times 975}{21.92}}\ ;\ \sqrt{\frac{12 \times 975}{3.816}}\right] \ = \ [23.1\,;\,55.4]\ .$$

   The confidence intervals for the variance or the standard deviations on small samples are usually very wide.

Exercise 2.2.2. The weight of grapes produced per vine has been measured on 10 vines selected at random in a vineyard. The results in kilograms are the following:

$$2.4 \quad 3.4 \quad 3.6 \quad 4.1 \quad 4.3 \quad 4.7 \quad 5.4 \quad 5.9 \quad 6.5 \quad 6.9\ .$$

The weight of grapes produced by each vine is modeled by a $\mathcal{N}(\mu, \sigma^2)$ distribution.

1. Find the empirical mean and variance of the sample.
2. Give a 95% confidence interval for μ.
3. Give a 95% confidence interval for σ^2.
4. From now on, the standard deviation of productions per plant is supposed to be known and equal to 1.4. Give a 95% confidence interval for μ.
5. Find the minimal number of plants that should be taken to estimate μ with 99% confidence with a precision of ± 500 grams.

Exercise 2.2.3. A study on coronary blood flow velocity has lead to the following results in 18 people:

$$75,\ 77,\ 78,\ 77,\ 77,\ 72,\ 72,\ 72,\ 70,\ 71,\ 69,\ 69,\ 68,\ 66,\ 64,\ 66,\ 62,\ 61.$$

The values of that sample are modeled by a random variable with normal distribution $\mathcal{N}(\mu, \sigma^2)$, where μ and σ^2 are unknown.

1. Find the mean and variance of the sample.
2. Find the intervals of confidence for μ at levels 0.95, 0.98, and 0.99.
3. Find the intervals of confidence for σ^2 at levels 0.95, 0.98, and 0.99.
4. What would the confidence intervals for μ be, if the variance σ^2 was known and equal to 26?

Exercise 2.2.4. A laboratory uses an optical device to measure the fluorescein concentration in solutions. The measurements are modeled by a Gaussian random variable, the expectation of which is equal to the true concentration of the solution, and the standard deviation, guaranteed by the company selling the device is known: $\sigma = 0.05$.

1. Nine measurements are made for the same solution. The empirical mean of the 9 results is 4.38 mg/l. Give a 99% confidence interval for the true concentration of the solution.
2. For that same sample, what is the confidence level of the interval [4.36 ; 4.40]?
3. What should the sample size be if the concentration of the solution had to be known at confidence level 0.99, with a precision of ± 0.01 mg/l?
4. On the same sample of 9 measurements, a standard deviation of 0.08 mg/l has been observed. Give a 99% confidence interval for the theoretical standard deviation. What do you think of the company's guarantee?
5. Answer the first question again, this time supposing that the theoretical standard deviation is unknown, and estimated by the empirical one.

Exercise 2.2.5. To study the rotting of potatoes, a researcher injects bacteria that induce decay, into 13 potatoes. He then measures the rotten area (in mm^2) in these 13 potatoes. He gets a mean of 7.84 mm^2 and an empirical variance of 14.13. The rotten area of a potato is modeled by a random variable with $\mathcal{N}(\mu, \sigma^2)$ distribution.

1. Find a confidence interval for μ at level 0.95, then 0.99.
2. Find a confidence interval for σ^2 at level 0.95, then 0.99.

Exercise 2.2.6. The production of a new kind of apple tree has to be estimated. The production of one tree is modeled by a Gaussian random variable with expectation μ and standard deviation σ, both unknown.

1. A sample of 15 apple trees, has produced a mean crop of 52 kg with standard deviation 5 kg. Find a confidence interval for the expected production of the apple trees of the new species, at level 0.95, then 0.99.
2. Find a confidence interval for the standard deviation σ, at level 0.95.

2.3 Confidence interval for the expectation on a large sample

For a large sample, a confidence interval at approximate level $1-\alpha$ for the expectation is obtained by:

$$\left[\overline{X} - u_\alpha \frac{\sqrt{S^2}}{\sqrt{n}} \; ; \; \overline{X} + u_\alpha \frac{\sqrt{S^2}}{\sqrt{n}}\right] ,$$

where u_α is the quantile of order $1-\alpha/2$ for the normal distribution $\mathcal{N}(0,1)$.

Exercise 2.3.1. The fluorescein concentration of a given solution has been measured 90 times. An empirical mean of 4.38 mg/l and a standard deviation of 0.08 mg/l have been observed. Give a confidence interval for the true concentration of the solution, at confidence levels 0.95 and 0.99.

The quantile of order 0.975 *for the* $\mathcal{N}(0,1)$ *distribution is* 1.96. *The* 95% *confidence interval is:*

$$\left[4.38 - 1.96\frac{0.08}{\sqrt{90}} \; ; \; 4.38 + 1.96\frac{0.08}{\sqrt{90}}\right] \; = \; [4.363\,;\, 4.397] \; .$$

The quantile of order 0.995 *for the* $\mathcal{N}(0,1)$ *distribution is* 2.5758. *The* 99% *confidence interval is:*

$$\left[4.38 - 2.5758\frac{0.08}{\sqrt{90}} \; ; \; 4.38 + 2.5758\frac{0.08}{\sqrt{90}}\right] \; = \; [4.358\,;\, 4.402] \; .$$

Exercise 2.3.2. The production of a new kind of apple trees has to be estimated. On a sample of 80 trees, a mean crop of 51.5 kg, with standard deviation 4.5 kg has been observed. Find a confidence interval for the expected production of apple trees of that species, at level 0.95, then 0.99.

Exercise 2.3.3. The lengths in millimeters of 152 cuckoo eggs have been measured. The empirical mean was found to be 40.8 mm, with an empirical variance of 14.7 mm^2. Find a confidence interval for the expected length of a cuckoo egg, at levels 0.95, 0.98, and 0.99.

Exercise 2.3.4. The lengths in millimeters of 150 walnut shells have been measured. The empirical mean is 27.6 mm, and the empirical standard deviation is 3.7 mm. Give a confidence interval for the expected length of a wallnut shell, at level 0.99, then 0.998.

Exercise 2.3.5. Sleeping pills have been given to two groups of patients A and B. The 100 patient of group A have received a new sleeping pill whereas the 50 patient of group B have received an old one. The patients of group A have slept 7.82 hours on average, with a standard deviation of 0.24 h; the patients of group B have slept 6.75 hours on average, with a standard deviation of 0.30 h.

1. Find a confidence interval for the average sleeping time of patients receiving the new pill, at levels 0.90, 0.95, 0.99.

2. Same question for patients taking the old pill.

3. Do you think that the new sleeping pill is more efficient than the old one?

2.4 Confidence interval of a probability for a large sample

For a large binary sample, a confidence interval at level $1-\alpha$ for the probability of the event is obtained by:

$$\left[\overline{X} - u_\alpha \frac{\sqrt{\overline{X}(1-\overline{X})}}{\sqrt{n}} \; ; \; \overline{X} + u_\alpha \frac{\sqrt{\overline{X}(1-\overline{X})}}{\sqrt{n}}\right] ,$$

where n is the sample size, $\overline{X}$ is the empirical frequency of the event and u_α is the quantile of order $1-\alpha/2$ of the normal distribution $\mathcal{N}(0,1)$.

Exercise 2.4.1. In order to study the influence of X-rays on the spermatogenesis of Bombyx mori, males have been exposed to radiation on the second day and on the fourth day of the larval stage. These males have been mated with non exposed females, and the number of fertile eggs laid by the females have been counted: out of a total of 5646 eggs laid, 4998 were fertile. In a control group of non exposed males and females, 5834 fertile eggs out of 6221 were obtained.

1. Find a 95% confidence interval for the proportion of fertile eggs after radiation exposure of males.

   The empirical frequency of fertile eggs after exposure of males is:

   $$F = \frac{4998}{5646} = 0.885 .$$

   The 95% *confidence interval is:*

   $$\left[0.885 - 1.96\frac{\sqrt{0.885(1-0.885)}}{\sqrt{5646}} \; ; \; 0.885 + 1.96\frac{\sqrt{0.885(1-0.885)}}{\sqrt{5646}}\right]$$

   $$= \; [0.876\,;\, 0.894] .$$

2. Find a 95% confidence interval for the proportion of fertile eggs of non exposed couples.

   The empirical frequency of fertile eggs for non exposed couples is:

   $$F = \frac{5834}{6221} = 0.938 .$$

The 95% confidence interval is:

$$\left[0.938 - 1.96\frac{\sqrt{0.938(1-0.938)}}{\sqrt{6221}}\ ;\ 0.938 + 1.96\frac{\sqrt{0.938(1-0.938)}}{\sqrt{6221}}\right]$$

$$= \quad [0.931\ ;\ 0.944]\ .$$

3. What do you think of the influence of radiation exposure on fertility?

   The two confidence intervals do not intersect; thus the proportion of fertile eggs is significantly lower for exposed males.

Exercise 2.4.2. On a sample of $n = 500$ teenagers, 210 were found to be overweight. Let p be the proportion of overweight teenagers. Find a confidence interval for p, at levels 0.95 and 0.99.

Exercise 2.4.3. A clinic has proposed a new surgery, and has had 40 failures out of 200 attempts. Let p be the probability of success of that new surgery.

1. What estimate would you propose for p?
2. Using the normal approximation, give a 95% confidence interval for p.
3. How many surgeries should the clinic perform to know the success probability with a precision of $\pm 1\%$ with 95% confidence level?

3 Statistical testing

3.1 Decision rule, threshold and p-value

- In a test, the *null hypothesis* $\mathcal{H}_0$ is the one for which the probability of rejecting it wrongly is controled. It is the most valuable one, the one it would be costly to reject wrongly.
- The *threshold* of the test, also called the *first kind risk* is the probability of rejecting $\mathcal{H}_0$ wrongly:
$$\mathbb{P}_{\mathcal{H}_0}[\text{ Reject } \mathcal{H}_0\,] = \alpha\ .$$
- The *test statistic* is a function of the data, for which we know the probability distribution under the null hypothesis $\mathcal{H}_0$.
- The *decision rule* specifies, as a function of the values taken by the test statistic, in which cases the hypothesis $\mathcal{H}_0$ should be rejected.
- A test may be:
  - ⋆ *two-tailed* if the decision rule is:
$$\text{Reject } \mathcal{H}_0 \iff T \notin [l, l']$$
(reject too small or too large values). Usually, l and l' are chosen such that $\mathbb{P}_{\mathcal{H}_0}[T < l] = \mathbb{P}_{\mathcal{H}_0}[T > l'] = \alpha/2$.
  - ⋆ *one-tailed* if the decision rule is:
$$\text{Reject } \mathcal{H}_0 \iff T < l$$
(reject too small values),
or else:
$$\text{Reject } \mathcal{H}_0 \iff T > l$$
(reject too large values).
- The *p-value* is the threshold for which the observed value of the test statistic would be on the boundary of the rejection region. It is the probability under $\mathcal{H}_0$ that the test statistic be beyond the value that has been observed.
- The *second kind risk* is the probability of accepting $\mathcal{H}_0$ wrongly, *i.e.* the probability of accepting $\mathcal{H}_0$ when the *alternative hypothesis* $\mathcal{H}_1$ is true:
$$\mathbb{P}_{\mathcal{H}_1}[\text{ accept } \mathcal{H}_0\,] = \beta\ .$$
The *power* of the test is $1-\beta$. It is the probability of being right rejecting $\mathcal{H}_0$.

Exercise 3.1.1. For an adult, the logarithm of the D-dimer concentration, denoted by X, is modeled by a normal random variable with expectation μ and variance σ^2. The variable X is an indicator for the risk of thrombosis: it is considered that for healthy individuals, μ is -1, whereas for individuals at risk μ is 0. In both cases, the value of σ^2 is the same: 0.09.

1. Dr. House does not want to worry his patients if there is no need to. What hypotheses $\mathcal{H}_0$ and $\mathcal{H}_1$ will he choose to test? Give the decision rule for his test, at threshold 1%, and at threshold 5%.

   If Dr. House does not want to worry a patient, the hypothesis he considers as dangerous to reject wrongly is that the patient is not at risk, thus that his value of X (the test statistic) has expectation -1. His hypothesis $\mathcal{H}_0$ is $\mu = -1$ (the patient is not at risk), that he will test against $\mathcal{H}_1$: $\mu = 0$ (the patient is at risk). He will choose to reject too high values for X. The decision rule will be:

   $$\textit{Reject } \mathcal{H}_0 \iff X > l \ ,$$

   where:

   $$\mathbb{P}_{\mathcal{H}_0}[X > l] = \alpha \ .$$

   According to the null hypothesis $\mathcal{H}_0$, the test statistic X follows the $\mathcal{N}(-1, 0.09)$ distribution, hence $\frac{X-(-1)}{\sqrt{0.09}}$ follows the $\mathcal{N}(0,1)$ distribution. An equivalent decision rule is:

   $$\textit{Reject } \mathcal{H}_0 \iff \frac{X-(-1)}{\sqrt{0.09}} > \frac{l-(-1)}{\sqrt{0.09}} \ .$$

   Therefore $\frac{l-(-1)}{\sqrt{0.09}}$ is the value which has probability α to be overpassed for a $\mathcal{N}(0,1)$ distribution: 1.6449 for $\alpha = 0.05$, 2.3263 for $\alpha = 0.01$. At threshold 0.05 the decision rule of the test is:

   $$\begin{aligned} \textit{Reject } \mathcal{H}_0 &\iff \frac{X-(-1)}{\sqrt{0.09}} > 1.6449 \\ &\iff X > 1.6449\sqrt{0.09} + (-1) = -0.5065 \ . \end{aligned}$$

   Dr. House declares the patient is at risk if the logarithm of his D-dimer concentration is higher than -0.5065.

   At threshold 0.01 the decision rule of the test is:

   $$\begin{aligned} \textit{Reject } \mathcal{H}_0 &\iff \frac{X-(-1)}{\sqrt{0.09}} > 2.3263 \\ &\iff X > 2.3263\sqrt{0.09} + (-1) = -0.3021 \ . \end{aligned}$$

   The lower the threshold, the less the decision rule rejects risky patients: what should happen to reject $\mu = -1$ at threshold 0.01 is more unusual than at threshold 0.05.

2. Find the second kind risk and the power of the tests of the previous question.

   The second kind risk is the probability of rejecting $\mathcal{H}_1$ wrongly. Under hypothesis $\mathcal{H}_1$, $\mu = 0$, and the variable X follows the $\mathcal{N}(0, 0.09)$ distribution.
   For the test at threshold 0.05, the probability of accepting $\mathcal{H}_0$ wrongly (i.e. of declaring wrongly that a patient is not at risk) is:

$$\beta = \mathbb{P}_{\mathcal{H}_1}[X \leqslant -0.5065] = \mathbb{P}_{\mathcal{H}_1}\left[\frac{X-0}{\sqrt{0.09}} \leqslant \frac{-0.5065-0}{\sqrt{0.09}}\right]$$

   Under hypothesis $\mathcal{H}_1$, $\frac{X-0}{\sqrt{0.09}}$ follows the $\mathcal{N}(0,1)$ distribtuion. Therefore we must calculate, for a variable with $\mathcal{N}(0,1)$ distribution, the probability to fall below $\frac{-0.5065-0}{\sqrt{0.09}} = -1.6885$: this is the value of the distribution function of the $\mathcal{N}(0,1)$ at -1.6885, that is 0.0457. The power is:

$$1 - \beta = 1 - 0.0457 = 0.9543\ .$$

   For the test at threshold 0.01, the reasoning is the same, replacing the bound -0.5065 by -0.3021. A second kind risk equal to 0.1570 is found, with a power equal to 0.8430.
   When the threshold is lower, the risk of wrongly rejecting $\mathcal{H}_0$ is lowered, but the risk of accepting it wrongly is higher and the power is lower. For the test at threshold 0.01, the probability that Dr. House wrongly declaring that a patient is not at risk is about 16%.

3. A patient has a value of X equal to -0.46. Find the p-value of Dr. House's test.

   The p-value is the threshold at which -0.46 would be the bound. Knowing the results of the first question, since -0.46 lies between -0.5065 and -0.3021, the p-value must be between 0.05 and 0.01. It is the probability under $\mathcal{H}_0$, that the variable X is higher than -0.46.

$$\mathbb{P}_{\mathcal{H}_0}[X > -0.46] = \mathbb{P}_{\mathcal{H}_0}\left[\frac{X-(-1)}{\sqrt{0.09}} > \frac{-0.46-(-1)}{\sqrt{0.09}}\right] = \mathbb{P}_{\mathcal{H}_0}\left[\frac{X-(-1)}{\sqrt{0.09}} > 1.8\right].$$

   Under $\mathcal{H}_0$, $\frac{X-(-1)}{\sqrt{0.09}}$ follows the $\mathcal{N}(0,1)$ distribution. The probability we are looking for is $1 - F(1.8)$, where F is the distribution function of the $\mathcal{N}(0,1)$, i.e. 0.0359.

4. Dr. Cuddy's point of view is that she'd rather worry a patient wrongly than not warn him of an actual risk. What hypotheses $\mathcal{H}_0'$ and $\mathcal{H}_1'$ will she choose to test? Give the decision rule for her test, at threshold 1%, and at threshold 5%.

   If Dr. Cuddy does not want to miss a patient at risk, the hypothesis she considers as dangerous to reject wrongly is that he is at risk, and that his variable X has expectation 0. Her hypothesis $\mathcal{H}_0'$ is $\mu = 0$ (the patient is at risk), that she will

test against $\mathcal{H}'_1$*:* $\mu = -1$ *(the patient is not at risk). She will choose to reject lower values of* X*. The decision rule will be:*

$$\textit{Reject } \mathcal{H}'_0 \iff X < l' \ ,$$

where:

$$\mathbb{P}_{\mathcal{H}'_0}[X < l'] = \alpha \ .$$

Under hypothesis $\mathcal{H}'_0$*, the test statistic* X *follows the* $\mathcal{N}(0, 0.09)$ *distribution, therefore* $\frac{X-0}{\sqrt{0.09}}$ *follows the* $\mathcal{N}(0,1)$*. An equivalent decision rule is:*

$$\textit{Reject } \mathcal{H}'_0 \iff \frac{X-0}{\sqrt{0.09}} < \frac{l'-0}{\sqrt{0.09}} \ .$$

Thus $\frac{l'-0}{\sqrt{0.09}}$ *is the value such that a variable with* $\mathcal{N}(0,1)$ *distribution falls below with probability* α*:* -1.6449 *for* $\alpha = 0.05$*,* -2.3263 *for* $\alpha = 0.01$*. At threshold* 0.05 *the decision rule is:*

$$\begin{aligned} \textit{Reject } \mathcal{H}_0 &\iff \frac{X-0}{\sqrt{0.09}} < -1.6449 \\ &\iff X < -1.6449 \times \sqrt{0.09} + 0 = -0.4935 \ . \end{aligned}$$

Dr Cuddy declares that the patient is not at risk when his D-dimer variable is below -0.4935*.*
At threshold 0.01 *the decision rule is:*

$$\begin{aligned} \textit{Reject } \mathcal{H}'_0 &\iff \frac{X-0}{\sqrt{0.09}} < -2.3263 \\ &\iff X < -2.3263 \times \sqrt{0.09} + 0 = -0.6980 \ . \end{aligned}$$

5. Depending on the threshold, for what values of X will Drs. House and Cuddy's diagnoses agree?

   If $X < \min\{l, l'\}$*, Dr. House accepts* $\mathcal{H}_0$*, Dr. Cuddy rejects* $\mathcal{H}'_0$*. In both cases, the conclusion for the patient is the same: he is not at risk. Conversely, if* $X > \max\{l, l'\}$ *Dr. House rejects* $\mathcal{H}_0$*, Dr. Cuddy accepts* $\mathcal{H}'_0$ *and the conclusion is the same: the patient is at risk.*

   The conclusions differ for a patient whose value of X *lies between* l *and* l'*. At threshold* 0.05 *the bounds are* $l = -0.5065$ *and* $l' - 0.4935$*. For a patient whose variable* X *is between* -0.5065 *and* -0.4935*, Dr. House declares him at risk (he rejects* $\mathcal{H}_0$*), Dr. Cuddy declares he is not at risk (she rejects* $\mathcal{H}'_0$*).*

   At threshold 0.01*, the bounds are* $l = -0.3021$ *and* $l' = -0.6980$*. For a patient whose variable* X *is between* -0.6980 *and* -0.3021*, Dr. House declares he is not at risk (he accepts* $\mathcal{H}_0$*), Dr. Cuddy declares he is at risk (she accepts* $\mathcal{H}'_0$*).*

6. Give the decision rule of the test for the null hypothesis $\mathcal{H}_0'' : \mu = -1$ against the alternative one $\mathcal{H}_1'' : \mu \neq -1$.

   This is a two-tailed test. The decision rule will be:

   $$\textit{Reject } \mathcal{H}_0'' \iff X \notin [l_1, l_2] \,,$$

   where:

   $$\mathbb{P}_{\mathcal{H}_0''}[\, X \notin [l_1, l_2]\,] = 0.05 \,.$$

   Under hypothesis $\mathcal{H}_0''$, the test statistic X follows the $\mathcal{N}(-1, 0.09)$ distribution, hence $\frac{X-(-1)}{\sqrt{0.09}}$ follows the $\mathcal{N}(0,1)$. An equivalent decision rule is:

   $$\textit{Reject } \mathcal{H}_0'' \iff \frac{X-(-1)}{\sqrt{0.09}} \notin \left[\frac{l_1-(-1)}{\sqrt{0.09}}\,;\,\frac{l_2-(-1)}{\sqrt{0.09}}\right] \,.$$

   The interval $\left[\frac{l_1-(-1)}{\sqrt{0.09}}\,;\,\frac{l_2-(-1)}{\sqrt{0.09}}\right]$ must contain 95% of the values taken by a variable with $\mathcal{N}(0,1)$ distribution. The interval centered at 0 is chosen: $[-1.96\,;\,+1.96]$. Hence:

   $$\frac{l_1-(-1)}{\sqrt{0.09}} = -1.96 \implies l_1 = (-1) - 1.96\sqrt{0.09} = -1.588 \,,$$

   and:

   $$\frac{l_2-(-1)}{\sqrt{0.09}} = +1.96 \implies l_2 = (-1) + 1.96\sqrt{0.09} = -0.412 \,,$$

   At threshold 0.05 the decision rule of the two-tailed test is:

   $$\textit{Reject } \mathcal{H}_0 \iff X \notin [-1.588\,;\,-0.412] \,.$$

   The patient is said to have logarithm of D-dimer concentration significantly different from -1 when his variable X is either lower than -1.588, or larger than -0.488.

7. A patient has a value of X equal to -0.46. Find the p-value for the test of the previous question.

   The p-value is the threshold for which the observed value would be a bound of the rejection region. That rejection region is centered at -1. The other bound should be $-1-(-0.46-(-1)) = -1.54$.

The p-value is the following probability.

$$\begin{aligned} & \mathbb{P}_{\mathcal{H}_0''}[X \notin [-1.54\,;\,-0.46]\,] \\ = \; & \mathbb{P}_{\mathcal{H}_0''}\left[\frac{X-(-1)}{\sqrt{0.09}} \notin \left[\frac{-1.54-(-1)}{\sqrt{0.09}}\,;\,\frac{-0.46-(-1)}{\sqrt{0.09}}\right]\right] \\ = \; & \mathbb{P}_{\mathcal{H}_0''}\left[\frac{X-(-1)}{\sqrt{0.09}} \notin [-1.8\,;\,+1.8]\right]\,. \end{aligned}$$

Under the hypothesis $\mathcal{H}_0''$, the variable $\frac{X-(-1)}{\sqrt{0.09}}$ follows the $\mathcal{N}(0,1)$ distribution: the desired probability is 0.0719*. The p-value is found to be twice that of the one-tailed test in question 3.*

Exercise 3.1.2. A packaging machine is supposed to produce packs of 1 kg. The actual weight of a pack is modeled by a random variable following a normal distribution, with a standard deviation of 20 g. However, it is possible to tune the mean weight of the packs.

1. The production manager decides not to distribute packs with weight too far away from the prescribed value of 1 kg. What hypotheses $\mathcal{H}_0$ and $\mathcal{H}_1$ should he test? Establish the decision rule for that test at thresholds 5% and 1%.

2. The company manager thinks that the packs going out on sale are too heavy, causing money loss. What hypotheses $\mathcal{H}_0$ and $\mathcal{H}_1$ should the production manager use to answer the criticism? Establish the decision rule for that test at thresholds 5% and 1%.

3. A pack has been weighed at 1018 grams. What is the p-value for the test of the previous question? What is the p-value for the test of the first question?

4. A consumers' association sues the company for selling packs that are too light. What hypotheses $\mathcal{H}_0$ and $\mathcal{H}_1$ should the production manager use to answer? Establish the decision rule for that test at thresholds 5% and 1%.

5. A pack has been weighed at 982 grams. What is the p-value for the test of the previous question? What is the p-value for the test of the first question?

Exercise 3.1.3. A paracetamol concentration of more than 150 mg per kilogram body weight is considered as dangerous. The mesurements of paracetamol in blood tests are modelled by a random variable with normal distribution $\mathcal{N}(\mu, \sigma^2)$. The standard-deviation, linked to the testing method, is supposed to be known and equal to 5 mg.

1. Give the hypotheses and the decision rule for the test deciding, at threshold 5%, whether a patient is at risk, from the results of one blood test (you are a cautious doctor).

2. A patient arrives at the hospital with signs of paracetamol poisoning. A blood test is made and a concentration of 140 mg is found. Give the p-value for the test of the previous question. Should that patient be considered at risk?

Exercise 3.1.4. Let X be the pollution index measured close to a chemical plant. It is modeled by a $\mathcal{N}(\mu, \sigma^2)$ distribution. The standard deviation is supposed to be known and equal to 4. The state regulations fix the maximal pollution index at 30.

1. The head manager wants to show that his plant complies with the regulations. What hypotheses $\mathcal{H}_0$ and $\mathcal{H}_1$ should he test? Establish the decision rule for that test at thresholds 5% and 1%.

2. The Green party wants to prove that the pollution is higher than prescribed. What hypotheses $\mathcal{H}'_0$ and $\mathcal{H}'_1$ should they test? Establish the decision rule for that test at thresholds 5% and 1%.

3.2 Tests on a sample

Denote by:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{and} \quad S^2 = \left(\frac{1}{n}\sum_{i=1}^{n} X_i^2\right) - \overline{X}^2$$

the mean and standard deviation of the sample. The expectation of the unknown distribution is μ, its variance is σ^2. The test statistics and their probability distribution under the null hypothesis $\mathcal{H}_0$ are the following.

- Testing values of expectation, Gaussian sample, σ^2 known.

$$\mathcal{H}_0 : \mu = \mu_0 \quad ; \quad T = \sqrt{n}\left(\frac{\overline{X} - \mu_0}{\sqrt{\sigma^2}}\right) \text{ follows the } \mathcal{N}(0,1) \ .$$

- Testing values of the expectation, Gaussian sample, σ^2 unknown.

$$\mathcal{H}_0 : \mu = \mu_0 \quad ; \quad T = \sqrt{n-1}\left(\frac{\overline{X} - \mu_0}{\sqrt{S^2}}\right) \text{ follows the } \mathcal{T}(n-1) \ .$$

- Testing values of the variance, Gaussian sample, σ^2 unknown.

$$\mathcal{H}_0 : \sigma^2 = \sigma_0^2 \quad ; \quad T = n\left(\frac{S^2}{\sigma_0^2}\right) \text{ follows the } \mathcal{X}^2(n-1) \ .$$

- Testing values of the expectation, large sample, σ^2 known or not.

$$\mathcal{H}_0 : \mu = \mu_0 \quad ; \quad T = \sqrt{n}\left(\frac{\overline{X} - \mu_0}{\sqrt{S^2}}\right) \text{ follows the } \mathcal{N}(0,1) \, .$$

- Testing values of a probability, large binary sample.

$$\mathcal{H}_0 : p = p_0 \quad ; \quad T = \sqrt{n}\left(\frac{\overline{X} - p_0}{\sqrt{p_0(1-p_0)}}\right) \text{ follows the } \mathcal{N}(0,1) \, .$$

Exercise 3.2.1. For an adult, the logarithm of the D-dimer concentration, denoted by X, is modeled by a normal random variable with expectation μ and variance σ^2. The variable X is an indicator for the risk of thrombosis: it is considered that for healthy individuals, μ is -1, whereas for individuals at risk μ is 0. The influence of olive oil on thrombosis risk must be evaluated.

1. A group of 13 patients, previously considered as being at risk, had an olive oil enriched diet. After the diet, their value of X was measured, and this gave an empirical mean of -0.15. The variance σ^2 is supposed to be known and equal to 0.09. Give the decision rule for the test of $\mathcal{H}_0 : \mu = 0$ against $\mathcal{H}_1 : \mu = -1$, at threshold 5%. What p-value corresponds to -0.15? What is your conclusion? Find the second kind risk and the power of the test.

   We have a Gaussian sample with known variance, and we build a test on the value of the expectation. The test statistic is:

$$T = \sqrt{13}\frac{\overline{X} - 0}{\sqrt{0.09}} \, .$$

   According to the null hypothesis $\mathcal{H}_0$, T follows the normal distribution $\mathcal{N}(0,1)$. The hypothesis $\mathcal{H}_0$ is rejected when T takes low values. At threshold 5%, the bound is -1.6449. The decision rule is:

$$\textit{Reject } \mathcal{H}_0 \iff T < -1.6449 \iff \overline{X} < -0.1369 \, .$$

   For $\overline{X} = -0.15$, the test statistic takes a value of -1.8028, the corresponding p-value is 0.0357. At threshold 5%, the hypothesis $\mathcal{H}_0$ is rejected, thus the decision is that there has been a significant improvement. But at any threshold smaller than 3.57%, $\mathcal{H}_0$ is not rejected: the olive oil has not made a significant improvement.

   Under hypothesis $\mathcal{H}_1$, $\sqrt{13}\frac{\overline{X}-(-1)}{\sqrt{0.09}}$ follows the $\mathcal{N}(0,1)$ distribution. The second

kind risk is the probability of accepting $\mathcal{H}_0$ *wrongly, i.e.:*

$$\begin{aligned}\beta &= \mathbb{P}_{\mathcal{H}_1}[\overline{X} > -0.1369] \\ &= \mathbb{P}_{\mathcal{H}_1}\left[\sqrt{13}\frac{\overline{X}-(-1)}{\sqrt{0.09}} > \sqrt{13}\frac{-0.1369-(-1)}{\sqrt{0.09}}\right] \\ &= \mathbb{P}_{\mathcal{H}_1}\left[\sqrt{13}\frac{\overline{X}-(-1)}{\sqrt{0.09}} > 10.3732\right] \\ &\simeq 0\,.\end{aligned}$$

The second kind risk is very close to 0 *(lower than* 10^{-20}*), and the power is very close to* 1*.*

2. For the same group of 13 patients, an empirical standard deviation of 0.37 has been observed. Give the decision rule for the test of $\mathcal{H}_0 : \sigma^2 = 0.09$, against $\mathcal{H}_1 : \sigma^2 \neq 0.09$, at threshold 5%. What is your conclusion?

   We must test a value of the variance for a Gaussian sample. The test statistic is:

$$T = 13\frac{S^2}{0.09}\,.$$

   Under hypothesis $\mathcal{H}_0$*, it follows the chi-squared distribution with parameter* 12*. We want a two-tailed test, hence a decision rule rejecting values euther too low or too high.*

$$\textit{Reject } \mathcal{H}_0 \iff T \notin [l, l']\,,$$

   where l *and* l' *are the quantiles of order* 0.025 *and* 0.975 *for the chi-squared distribution with parameter* 12*:* $l = 4.4038$ *and* $l' = 23.3367$*. Here the test statistic takes a value of* 19.7744*. It is a high value, but not high enough to reject the hypothesis that the theoretical variance is* 0.09*.*

3. Assuming that the variance is unknown and using the estimate of the previous question, give the decision rule for the test of $\mathcal{H}_0 : \mu = 0$, against $\mathcal{H}_1 : \mu < 0$, at threshold 5%. What is your conclusion?

   We have a Gaussian sample with unknown variance and we must build a test on the value of the expectation. The test statistic is:

$$T = \sqrt{12}\frac{\overline{X}-0}{\sqrt{S^2}}\,.$$

   Under hypothesis $\mathcal{H}_0$*,* T *follows the Student distribution* $\mathcal{T}(12)$*. The hypothesis* $\mathcal{H}_0$ *is rejected when* T *takes low values. At threshold* 5% *the bound is* -1.7823*. The decision rule is:*

$$\textit{Reject } \mathcal{H}_0 \iff T < -1.7823\,.$$

For $\overline{X} = -0.15$ and $\sqrt{S^2} = 0.37$, the test statistic T takes a value of -1.4044, thus $\mathcal{H}_0$ cannot be rejected (the corresponding p-value is 0.0928). It can be said that there has not been a significant improvement.

4. The same experiment is repeated on a group of 130 patients, for whom an empirical mean of -0.12 and a standard deviation of 0.32 are observed. Give the decision rule for the test of $\mathcal{H}_0 : \mu = 0$ against $\mathcal{H}_1 : \mu < 0$, at threshold 5%. What p-value corresponds to -0.12? What is your conclusion?

   Now we must test a value of the expectation for a large sample. The test statistic is:

$$T = \sqrt{130}\frac{\overline{X} - 0}{\sqrt{S^2}} \ .$$

   Under hypothesis $\mathcal{H}_0$, T follows the normal distribution $\mathcal{N}(0,1)$. The hypothesis $\mathcal{H}_0$ is rejected when the values of T are too low values. At threshold 5% the bound is -1.6449. The decision rule is:

$$\textit{Reject } \mathcal{H}_0 \iff T < -1.6449 \ .$$

   For $\overline{X} = -0.12$ and $\sqrt{S^2} = 0.32$, the test statistic takes a value of -4.2757, the corresponding p-value is close to 10^{-5}. Thus it can be concluded without doubt that the expected value of X is significantly lower than 0.

5. The D-dimer concentration of the 130 patients was measured before the diet. After the diet, the concentration was found to be lower for 78 patients, higher for 52 patients. Build a test to decide whether the olive oil regime has improved the condition for a significant proportion of the patients. With the observations you have, what is the p-value for that test, what is your conclusion?

   Let p be the probability of improvement (lower value of X). If the regime had no effect, fluctuations in measurements would be merely random and there would be as many improvements as worsenings: the proportion of improvements would be $1/2$. We must test, for a large binary sample, the hypothesis $\mathcal{H}_0 : p = 0.5$, against $\mathcal{H}_1 : p > 0.5$. The test statistic is:

$$T = \sqrt{n}\frac{\overline{X} - 0.5}{\sqrt{0.5(1-0.5)}} \ .$$

   Under hypothesis $\mathcal{H}_0$, the test statistic follows the normal distribution $\mathcal{N}(0,1)$. Here $\overline{X}$ is the observed proportion of improvements, that is $78/130$. The test statistic takes a value of 2.2804, the corresponding p-value (probability that a $\mathcal{N}(0,1)$ variable be higher than 2.2804) is 0.0113. It can be concluded that the improvement is significant at threshold 5%, but not quite at threshold 1%.

Exercise 3.2.2. A packaging machine is supposed to produce 1 kg packs. The actual weight of a pack is modeled by a random variable following a normal distribution with a standard deviation of 20 g. It is possible to tune the mean weight of the packs. In order to check that the tuning is correct, a sample of 10 packs is weighed.

1. Let $\mathcal{H}_0$ be the hypothesis: "the mean weight is 1 kg". Build a test at threshold 1%, of $\mathcal{H}_0$ against hypothesis $\mathcal{H}_1$: "the mean weight is different from 1 kg". Find the p-value of that test, for a sample of average weight 1011 grams.

2. Same question for hypothesis $\mathcal{H}_1$: "the mean weight is larger than 1 kg".

3. Answer again the two previous questions for a sample of 100 packs, with mean weight 1005 g.

4. On a sample of 10 packs, a mean weight of 1011 g has been observed, with an empirical standard deviation of 32 g. At threshold 1%, is this observation compatible with the value of 20 g for the theoretical standard deviation?

5. For the sample of the previous question, supposing the variance is unknown, can it be said that the packs are significantly too heavy on average at threshold 1%?

Exercise 3.2.3. A paracetamol concentration of more than 150 mg per kilogram body weight is considered as dangerous. The mesurements of paracetamol in blood tests are modelled by a random variable with normal distribution $\mathcal{N}(\mu, \sigma^2)$. The standard-deviation, linked to the testing method, is supposed to be known and equal to 5 mg. For better assessment, 4 blood tests are usually made. The results are assumed to be independent realisations of the same normal distribution $\mathcal{N}(\mu, \sigma^2)$.

1. Give the hypotheses and the decision rule for the test deciding, at threshold 5%, whether a patient is at risk, on view of 4 blood tests. (you are a cautious doctor).

2. On a given patient, the 4 blood tests gave concentrations of 140, 133, 148, 144. Give the p-value for the test of the previous question. Is he at risk?

3. From now on, the standard-deviation is supposed *unknown*. Give the test statistic and the decision rule for the test deciding, at threshold 5%, whether a patient is at risk, on view of 4 blood tests.

4. For the patient of question 2, give an interval containing the p-value for the test of the previous question. What is your conclusion?

Exercise 3.2.4. For a given population, the weight of newborn babies is modeled by a normal distribution. In the whole population, the standard deviation of newborn weights is 380 g. The average weight of a newborn whose mother does not smoke is 3400 g. In order to study the effect of tobacco, the babies of 30 mothers who smoked during pregnancy have been weighed, giving an empirical mean of 3240 g, with standard deviation 426 g.

1. Assuming that the standard deviation of the sample is known and equal to that of the whole population, calculate the p-value of the test deciding whether the newborns of the sample are significantly lighter on average. What is your conclusion, at threshold 5%?

2. Assuming that the standard deviation is unknown, give a test statistic and a decision rule, to test the same hypotheses as in the previous question. What is your conclusion?

3. Is the observed standard deviation significantly higher than that of the whole population?

4. Answer question 1. for a sample of 300 newborns, for which a mean weight of 3340 g has been observed.

Exercise 3.2.5. The 15 lengths of 15 cuckoo eggs (expressed in millimeters) are as follows:

19.8, 22.1, 21.5, 20.9, 22.0, 21.0, 22.3, 21.0, 20.3, 20.9, 22.0, 22.0, 20.8, 21.2, 21.0 .

The following values are given:

$$\sum x_i = 318.8 \quad \text{and} \quad \sum x_i^2 = 6782.78 \ .$$

The length of a cuckoo egg is modeled by a random variable with distribution $\mathcal{N}(\mu, \sigma^2)$.

1. Find the empirical mean and variance of the sample.

2. Test the hypothesis $\mathcal{H}_0 : \sigma^2 = 0.4$ against $\mathcal{H}_1 : \sigma^2 > 0.4$, at threshold 5%.

3. Test the hypothesis $\mathcal{H}_0 : \mu = 21$" against $\mathcal{H}_1 : \mu > 21$, at threshold 5%.

4. Give an interval containing the p-value of the test in the previous question.

Exercise 3.2.6. After a treatment of a certain species of rodents, a sample of 10 animals is selected and weighed. The weights in grams are the following

$$83 \ , \quad 81 \ , \quad 84 \ , \quad 80 \ , \quad 85 \ , \quad 87 \ , \quad 89 \ , \quad 84 \ , \quad 82 \ , \quad 80 \ .$$

The following values are given:

$$\sum x_i = 835 \quad \text{and} \quad \sum x_i^2 = 69801 \ .$$

It is known that untreated rodents have an average weight of 87.6 g. The weight of a rodent is modeled by a normal random variable.

1. At threshold 5%, test the hypothesis that "the treatment has no effect on the mean weight" against "the treatment decreases the mean weight".

2. Give an interval containing the p-value of the test in the previous question.

Exercise 3.2.7. A renting car company makes an experiment to decide between two types of tyres. Eleven cars are driven on a circuit with type A tyres. The tyres are then replaced by type B, and the cars are driven again on the same circuit. The consumption in liters per 100 km of these cars are modeled by Gaussian random variables. Here are the observations:

Car	1	2	3	4	5	6	7	8	9	10	11
Tyres A	4.2	4.7	6.6	7	6.7	4.5	5.7	6	7.4	4.9	6.1
Tyres B	4.1	4.9	6.2	6.9	6.8	4.4	5.7	5.8	6.9	4.9	6

1. Assuming that the observed differences follow a normal distribution, what test statistic would you propose?

2. What hypotheses would you test to decide whether the tyres have an effect on gas consumption?

3. What hypotheses would you test to decide whether type B tyres are significantly better on average?

4. At threshold 5% what are your conclusions?

Exercise 3.2.8. Nine patients with anxiety symptoms receive a sedative. The condition of the patient before and after treatment is evaluated by an index that the doctor calculates according to the answers to a series of questions. If the treatment is efficient, the index must decrease. The observed values of the index on the 9 patients are the following:

Before	1.83	0.5	1.62	2.48	1.68	1.88	1.55	3.06	1.3
After	0.88	0.65	0.59	2.05	1.06	1.29	1.06	3.14	1.29

1. What modeling assumption would you make and what hypotheses would you test?

2. Give an interval containing the p-value of the test deciding whether the sedative significantly improves the condition of the patients on average. What is your conclusion?

Exercise 3.2.9. A factory must deliver rods the length of which is modeled by a normal distribution with expectation 40 mm. The rods cannot be used if they are smaller than 39 mm or longer than 41 mm, and the factory guarantees that less than 1% must be discarded.

1. Assuming that the machine produces rods having the right length on average, what should the standard deviation be in order to ensure that only 1% of the rods should be discarded?

2. On a sample of 15 rods, an empirical mean of 40.3 mm has been observed, with a standard deviation of 0.6 mm. Is the observed standard deviation significantly higher than the theoretical one of the previous question?

3. Are the rods significantly too long on average?

4. A customer claims he has received 112 useless rods out of a batch of 10000. Is he right to complain?

Exercise 3.2.10. The percentage of 35-year-old women having wrinkles is 25%. Out of 200 women having followed an anti-wrinkle treatment, 40 of them still had wrinkles. At threshold 5%, can it be said that the treatment is efficient?

Exercise 3.2.11. For a certain disease, there exists a treatment that cures 70% of the cases. A laboratory proposes a new treatment claiming that it is better than the previous one. Out of 200 patients having received the new treatment, 148 of them have been cured. As the expert in charge of deciding whether the new treatment should be authorized, what are your conclusions?

Exercise 3.2.12. Here is the table of blood group frequencies in France:

Group / Factor	O	A	B	AB
Rhesus +	0.370	0.381	0.062	0.028
Rhesus –	0.070	0.072	0.012	0.005

The blood transfusion center in Pau has observed the following distribution on 5000 blood donors.

Group / Factor	O	A	B	AB
Rhesus +	2291	1631	282	79
Rhesus –	325	332	48	12

One wishes to answer statistically the questions below. For each question, the test statistic and the p-value should be calculated, and the conclusion given.

1. Is type O+ significantly more frequent in Pau?

2. Among people with positive rhesus, is the frequency of group O significantly different in Pau?

3. Among people with group O, is the frequency of the positive rhesus significantly higher in Pau?

3.3 Comparison of two independent samples

For the first sample:

$$\overline{X} = \frac{1}{n_x}\sum_{i=1}^{n_x} X_i \quad \text{and} \quad S_x^2 = \left(\frac{1}{n_x}\sum_{i=1}^{n_x} X_i^2\right) - \overline{X}^2 \ ,$$

the expectation of the unknown distribution is μ_x, its variance is σ_x^2.

For the second sample:

$$\overline{Y} = \frac{1}{n_y}\sum_{j=1}^{n_y} Y_j \quad \text{and} \quad S_Y^2 = \left(\frac{1}{n_y}\sum_{j=1}^{n_y} Y_j^2\right) - \overline{Y}^2 \ ,$$

the expectation of the unknown distribution is μ_y, its variance is σ_y^2.

The test statistics to be used and their distribution under the null hypothesis $\mathcal{H}_0$ are the following.

- Fisher test: comparison of variances, Gaussian sample.

  $\mathcal{H}_0 : \sigma_x^2 = \sigma_y^2 \quad ; \quad T = \dfrac{\frac{n_x}{n_x-1}S_x^2}{\frac{n_y}{n_y-1}S_y^2}$ follows the Fisher distribution $\mathcal{F}(n_x - 1, n_y - 1)$.

  If $T < 1$, swap X and Y (i.e. replace T by $1/T$) and compare to the quantile of order $1 - \alpha/2$ of the Fisher distribution $\mathcal{F}(n_y - 1, n_x - 1)$.

- Student test: comparison of expectations, Gaussian sample.

  $$\mathcal{H}_0 : \mu_x = \mu_y \quad ; \quad T = \frac{\sqrt{n_x + n_y - 2}}{\sqrt{\frac{1}{n_x} + \frac{1}{n_y}}} \frac{\overline{X} - \overline{Y}}{\sqrt{n_x S_x^2 + n_y S_y^2}} \ ,$$

  follows the Student distribution $\mathcal{T}(n_x + n_y - 2)$, si $\sigma_x = \sigma_y$.

- Comparison of expectations, large samples.

  $\mathcal{H}_0 : \mu_x = \mu_y \quad ; \quad T = \dfrac{\overline{X} - \overline{Y}}{\sqrt{\frac{S_x^2}{n_x} + \frac{S_y^2}{n_y}}}$ follows the normal distribution $\mathcal{N}(0, 1)$.

Exercise 3.3.1. The question whether cooking with olive oil reduces risk of thrombosis is studied. For this, the logarithm of the D-dimer concentration, modeled by a normal distribution, is considered. A sample of 9 people regularly using sunflower oil, gives a mean of -0.78, with standard deviation 0.27. A sample of 13 people regularly using olive oil, gives a mean of -0.97, with standard deviation 0.32.

1. Test the equality of variances at threshold 0.05.

   The Fisher test has to be applied to see if the difference between the observed variances of the two samples is significant or not. The statistic of the test is calculated, first putting the the lower variance on the numerator:

   $$T = \frac{\frac{9}{8}0.27^2}{\frac{13}{12}0.32^2} = 0.7393 \ .$$

   The hypothesis $\mathcal{H}_0 : \sigma_x = \sigma_y$ *must be tested against* $\mathcal{H}_1 : \sigma_x \neq \sigma_y$*. This is a two-tailed test, that rejects those values outside the interval* $[l, l']$*, where* l *and* l' *are the quantiles of order* 0.025 *and* 0.975 *of the distribution of* T *under* $\mathcal{H}_0$*, that is the Fisher distribution* $\mathcal{F}(8, 12)$*. However the quantile of order* 0.025 *for* $\mathcal{F}(8, 12)$ *is the inverse of the quantile of order* 0.975 *for the* $\mathcal{F}(12, 8)$*. Therefore it is simpler to swap* X *and* Y*, which amounts to computing* $1/T = 1.3526$*. This value must be compared to the quantile of order* 0.975 *for the Fisher distribution with parameters* 12 *and* 8 *(and not* 8 *and* 12 *since* X *and* Y *have been swapped). That bound is* 4.1997*. The observed value* 1.3526 *is lower, thus the hypothesis of equality of variances has to be accepted at threshold* 5%*.*

2. At threshold 0.05, what test would you propose to decide whether olive oil lowers significantly the risk of thrombosis? What is your conclusion? Give an interval containing the p-value.

   Having accepted the equality of variances, the application of the Student test for equality of expectations is justified. Denoting by X *the variable "logarithm of the D-dimer concentration for a person having sunflower oil", and* Y *the same variable for persons having olive oil, the following hypotheses must be tested:*

   $$\mathcal{H}_0 : \mu_x = \mu_y \quad \textit{against} \quad \mathcal{H}_1 : \mu_x > \mu_y \ .$$

   The test statistic is

   $$T = \frac{\sqrt{n_x + n_y - 2}}{\sqrt{\frac{1}{n_x} + \frac{1}{n_y}}} \frac{\overline{X} - \overline{Y}}{\sqrt{n_x S_x^2 + n_y S_y^2}} \ ,$$

   for which high values will be rejected.

   $$\textit{Reject } \mathcal{H}_0 \iff T > l \ .$$

   The bound l *is such that a variable following the Student distribution with parameter* $9 + 13 - 2 = 20$ *is larger with probability* 0.05*, hence* $l = 1.7247$*. Here the test statistic takes a value of* 1.3055*, therefore the equality of expectations cannot be rejected: the decrease on average that has been observed is not significant at threshold* 5%*. The p-value is the probability that a variable following the Student distribution* $\mathcal{T}(20)$ *be larger than* 1.3055*. In the table,* 1.3055 *lies between the quantiles of order* 0.8 *and* 0.9*, close to that of order* 0.9*. Hence the p-value lies between* 0.1 *and* 0.2*. The numerical value is* 0.1033*.*

3. In another study on 110 sunflower oil users, a mean of -0.82 has been observed, with standard deviation 0.29, while 130 olive oil users, had a mean of -0.93, with standard deviation 0.31. Find the p-value of the test deciding whether the improvement is significant. At threshold 0.05, what is your conclusion?

   This is a test of expectation comparison on large samples. The test statistic is:

$$\frac{\overline{X}-\overline{Y}}{\sqrt{\frac{S_x^2}{n_x}+\frac{S_y^2}{n_y}}},$$

   that follows the $\mathcal{N}(0,1)$ *distribution under the hypothesis* $\mathcal{H}_0$*. The calculated value is* 2.8366*. The p-value is the probability for a variable with* $\mathcal{N}(0,1)$ *distribution to be larger than* 2.8366*, that is* 0.0023*. At any threshold lower than* 0.23% *(and in particular of course at thresholds* 5% *and* 1%*), the null hypothesis* $\mathcal{H}_0$ *is rejected, and it can be concluded that olive oil significantly decreases the risk of thrombosis.*

Exercise 3.3.2. The activity of the seric PDE enzyme is studied, depending on different factors. The results are expressed in international unit per liter of serum. For two groups of women, pregnant or not, the following results were obtained:

non pregnant	1.5	1.6	1.4	2.9	2.2	1.8	2.7	1.9
pregnant	4.2	5.5	4.6	5.4	3.9	5.4	2.7	3.9

non pregnant	2.2	2.8	2.1	1.8	3.7	1.8	2.1
pregnant	4.1	4.1	4.6	3.9	3.5		

(Indications: $\sum x_i = 32.5$, $\sum x_i^2 = 75.83$, $\sum y_i = 55.8$, $\sum y_i^2 = 247.32$).

1. Describe the modeling hypotheses.
2. Test the equality of variances at threshold 5%.
3. Can it be claimed that the seric PDE enzyme activity is significantly different between pregnant and non pregnant women?
4. Can it be claimed that the seric PDE enzyme activity is significantly higher among pregnant women?

Exercise 3.3.3. The IQ's of 9 children in a district of a large city have empirical mean 107 and standard deviation 10. The IQ's of 12 children in another district have empirical mean 112 and standard deviation 9.

1. Describe the modeling hypotheses.
2. Test the equality of variances at threshold 5%.

3. Can it be said the children in the second district have significantly higher IQ's than those in the first district? Give an interval containing the p-value for the test.

Exercise 3.3.4. The maximal tensions of gastrocnemian muscles (expressed in g) for the frog vary, according whether the nerves have been removed or not. During an experiment made on 9 frogs, the following measures have been obtained:

With nerves	75	96	32	41	50	39	59	45	30
Without nerves	53	67	32	29	35	27	37	30	21

1. Describe the modeling hypotheses.
2. Test the equality of variances at threshold 5%.
3. At threshold 5%, can it be said that the mean maximal tension is different among the two groups? Give an interval containing the p-value of that test.

Exercise 3.3.5. During a study comparing different sampling methods for forest soil, the K_2O concentration has been measured, on the one hand for 20 samples of soil individually extracted, and on the other hand for 10 mixed samples, each obtained from 25 differents soils. For the individual samples, the following values have been found:

$$\sum x_i = 259.2 \quad \text{and} \quad \sum x_i^2 = 3662.08 \ ,$$

and for the mixed samples:

$$\sum y_i = 109.2 \quad \text{and} \quad \sum y_i^2 = 1200.8 \ .$$

It can be expected that the two sampling methods give very different variances. Justify intuitively why, and prove it using the Fisher test.

Exercise 3.3.6. To determine the average weights of ears of two kinds of ears of wheat, 9 ears of each kind are weighed. The empirical mean and variance for the two samples are:

$$\overline{x} = 170.7 \ ; \quad \overline{y} = 168.5 \ ; \quad s_x^2 = 432.90 \ ; \quad s_y^2 = 182.70 \ .$$

1. Describe the modeling hypotheses.
2. Test the equality of variances at threshold 5%.
3. Give an interval containing the p-value for the test deciding whether the two kinds are significantly different. What is your conclusion?

Exercise 3.3.7. In a farming cooperative, the effect of a fertilizer on wheat production is to be tested. For this, 2000 plots of land of the same size are chosen. Half of them are treated with the fertilizer, the other half are not. The crops in tons for the untreated plots give $\sum x_i = 61.6$, $\sum x_i^2 = 292.18$ and for the fertilized plots $\sum y_i = 66.8$, $\sum y_i^2 = 343.48$. Test the hypothesis "the fertilizer is not efficient", against "the fertilizer is efficient" at thresholds 0.01 and 0.05.

Exercise 3.3.8. In a city A, 36 people out of 300 smoke at least two packs a day. In city B, 8 out of 100 smoke at least two packs a day. The following hypotheses must be tested: $\mathcal{H}_0$: "there is no difference between the two cities" against $\mathcal{H}_1$: "more people smoke more than two packs a day in city A than in B".

1. Let p_A (respectively p_B) denote the proportion of people smoking more than two packs a day in A (respectively in B). What variables would you propose to model the problem? Give their expectations and variances as a function of p_A and p_B.
2. What test would you propose for $\mathcal{H}_0$ against $\mathcal{H}_1$?
3. Give the p-value for that test. What is your conclusion?

Exercise 3.3.9. Let p_A be the probability of curing a certain disease by treatment A. A group of 50 patients had that treatment and 28 were cured. Another treatment B cures that same disease with probability p_B. Out of 60 patients having had that new treatment, 38 were cured.

1. What test would you propose to decide whether treatment B is better than treatment A?
2. Give the p-value for that test. What is your conclusion?

3.4 The chi-squared adjustment test

Let r be the number of classes. For $i = 1, \ldots, r$, denote by n_i the *observed* number of class i, and np_i its *theoretical* number.

- The chi-squared test statistic is:

$$T = \sum_{i=1}^{r} \frac{(n_i - np_i)^2}{np_i} .$$

- Under the null hypothesis where the theoretical model is true, T follows the chi-squared distribution with parameter $d = r-1-k$:
  - $\star$ r is the number of classes,
  - $\star$ k is the number of parameters that have been estimated from the data to establish the theoretical model.
- The test applies to a large sample ($n \geqslant 50$). The theoretical numbers in each class must be large enough ($np_i \geqslant 8$). If necessary, classes can be grouped to fulfill the second condition.

Exercise 3.4.1. White-flowered sweet peas are crossed with red-flowered sweet peas. The colors of 600 plants from the second generation are distributed as follows:

Phenotype	Red	Pink	White
Number	141	325	134

With the 600 plants, 150 bunches of 4 are made, out of which the number of white-flowered plants is counted. The observed numbers were the following.

White flowers	0	1	2	3	4
Numbers	53	68	23	4	2

1. Give the theoretical proportions of the Mendelian distribution for the three colors. Find the chi-squared test statistic. Give an interval containing the p-value. What is your conclusion?

   Denote by R the allele inducing the red color and by B the allele inducing the white color. It is supposed that the phenotypes "red flowers", "pink flowers", and "white flowers" correspond respectively to genotypes RR, RB, and BB. If genotype RR is crossed with BB, all offsprings at the first generation are RB hybrids. At the second generation, crossing two hybrids should produce one fourth genotypes RR, half genotypes RB, one fourth genotypes BB; therefore, one fourth red flowered plants, half pink flowered and one fourth white flowered should be observed. The theoretical numbers would be 150, 300, 150.

   The chi-squared test statistic takes a value of:

$$T = \frac{(141-150)^2}{150} + \frac{(325-300)^2}{300} + \frac{(134-150)^2}{150} = 4.33\ .$$

   This value should be compared to the quantiles of the chi-squared distribution with parameter $3-1=2$. *The p-value is the probability for a variable following the* $\mathcal{X}^2(2)$ *distribution to be larger than* 4.33. *According to the table, the p-value lies between* 0.1 *and* 0.5. *The exact value is* 0.1147. *The hypothesis that the empirical distribution agrees with the theoretical one is accepted.*

2. What theoretical model would you propose for the number of white flowered plants in a bunch of 4? Make the appropriate grouping of classes. Find the chi-squared test statistic. Give an interval containing the p-value. What is your conclusion?

   If the bunches are formed at random, the distribution of the number of white flowered plants in a bunch of 4 is the binomial distribution with parameters 4 *(total number of plants) and* 1/4 *(theoretical proportion of white flowered plants). For* $i = 0, \ldots, 4$, *the theoretical number for the number of bunches with* i *white flowered plants is:*

$$np_i = 150 \binom{4}{k} \left(\frac{1}{4}\right)^k \left(\frac{3}{4}\right)^{4-k}\ .$$

White flowers	0	1	2	3	4
Observed number	53	68	23	4	2
Theoretical number	47.46	63.28	31.64	7.03	0.59

In order to reach a theoretical distribution with at least 8 *in each class, the last three classes may be grouped.*

White flowers	0	1	2, 3, 4
Observed number	53	68	29
Theoretical number	47.46	63.28	39.26

The chi-squared test statistic takes a value of 3.6786*. The p-value is the probability for a variable following the* $\mathcal{X}^2(2)$ *distribution, to be larger than* 3.6786*. According to the table, the p-value is between* 0.1 *and* 0.2*. The exact value is* 0.1589*. The hypothesis that the empirical distribution agrees with the theoretical one is accepted.*

3. Let $\widehat{p}$ be the observed proportion of white flowered plants. For 4-plant bunches, test the adjustment of the observed distribution with the binomial distribution $\mathcal{B}(4, \widehat{p})$: calculate the test statistic and give an interval containing the p-value.

   The total number of white flowered plants is 134*, their proportion is* $\widehat{p} = \frac{134}{600} \simeq 0.2233$*. The theoretical numbers are now calculated using the* $\mathcal{B}(4, \widehat{p})$ *distribution.*

White flowers	0	1	2, 3, 4
Observed number	53	68	29
Theoretical number	54.59	62.78	32.64

The chi-squared test statistic takes a value of 0.8855*. Since one parameter has been estimated to establish the theoretical distribution, the parameter of the chi-squared distribution is* $3 - 1 - 1 = 1$*. According to the table, the p-value is between* 0.3 *and* 0.4*, the exact value is* 0.3467*. The hypothesis that the empirical distribution agrees with the theoretical one is accepted.*

Exercise 3.4.2. Here is the frequency table of blood types in France:

Groupe / Factor	O	A	B	AB
Rhesus +	0.370	0.381	0.062	0.028
Rhesus –	0.070	0.072	0.012	0.005

The transfusion center of Pau has observed the following distribution out of 5000 blood donors.

Group / Factor	O	A	B	AB
Rhesus +	2291	1631	282	79
Rhesus –	325	332	48	12

The following questions need to be answered statistically. For each question, write down the table of observed and theoretical distributions, calculate the test statistic, give an interval containing the p-value, and specify your conclusion.

1. Is the distribution of the 8 group-rhesus types in Pau different from the overall French distribution?
2. Is the distribution of the two rhesuses in Pau different from the overall French distribution?
3. Among type O people, is the distribution of the two rhesuses in Pau different from the overall French distribution?
4. Among people with positive rhesus, is the distribution of the four groups in Pau different from the overall French distribution?
5. Among people with negative rhesus, is the distribution of the four groups in Pau different from the overall French distribution?

Exercise 3.4.3. A group of 162 students have been asked to evaluate the time they spend cooking per month:

Hours	$[0\,;\,5[$	$[5\,;\,10[$	$[10\,;\,15[$	$\geqslant 15$
Students	63	49	19	31

Previous studies in the overall population established the following distribution:

Hours	$[0\,;\,5[$	$[5\,;\,10[$	$[10\,;\,15[$	$\geqslant 15$
Proportion	40%	35%	15%	10%

Test the adjustment of the observed distribution among students to that of the general population. Give an interval containing the p-value. What is your conclusion?

Exercise 3.4.4. The sleeping time of twelve year-old children is studied. On a sample of size $n = 50$ the sleeping times (expressed in hours) have been recorded. The following values are given: $\sum x_i = 424$ and $\sum x_i^2 = 3828$, with the following distribution in classes.

Class	$\leqslant 8$	$]8\,;\,9]$	$]9\,;\,10]$	> 10
Number	19	12	9	10

1. It is generally considered that the sleeping time of a 12 year old child follows a $\mathcal{N}(9, 3)$ distribution. Apply the adjustment test of the observed distribution with the theoretical one. Give the value taken by the test statistic, an interval for the p-value and your conclusion.
2. Find the empirical mean $\overline{x}$ and variance s^2. Answer the previous question again, replacing the $\mathcal{N}(9, 3)$ distribution by $\mathcal{N}(\overline{x}, s^2)$.

Exercise 3.4.5. A biometry study on the length of cuckoo eggs has produced the following results. The following values are given: $n = 152$, $\sum x_i = 6200$, $\sum x_i^2 = 255200$, with the following distribution by classes:

class	< 32	$[32;34[$	$[34;36[$	$[36;38[$	$[38;40[$	$[40;42[$	$[42;44[$	$[44;46[$	$[46;48[$	$\geqslant 48$
number	2	7	6	18	25	40	23	20	6	5

1. Previous studies had shown that the length of a cuckoo egg can be modeled by a normal distribution with expectation 40 and standard deviation 4. Apply the adjustment test of the observed distribution with the theoretical one. Give the value taken by the test statistic, an interval for the p-value and your conclusion.

2. Find the empirical mean $\overline{x}$ and variance s^2. Answer the previous question again, replacing the $\mathcal{N}(40, 4^2)$ distribution by $\mathcal{N}(\overline{x}, s^2)$.

3.5 The chi-squared independence test

This is a particular case of the chi-squared adjustment test, that allows to test the mutual independence of two discrete variables.

- The *contingency table* presents the *joint numbers*. At line i, column j, the table gives n_{ij}, i.e. the number of individuals in class i for the first variable and class j for the second. If the number of modalities for the two variables are r and s, the table has r lines and s columns.

- The *marginal numbers* are the sums by line or column in the contingency table; $n_{i\bullet} = \sum_j n_{ij}$ is the total number of individuals in class i for the first variable; $n_{\bullet j} = \sum_i n_{ij}$ is the total number of individuals in class j for the second variable. The total number is $n = \sum_i n_{i\bullet} = \sum_j n_{\bullet j}$.

- The test statistic is:
$$T = n\left(-1 + \sum_{i=1}^{r}\sum_{j=1}^{s}\frac{n_{ij}^2}{n_{i\bullet}n_{\bullet j}}\right).$$

- Under the null hypothesis where both variables are independent, T follows the chi-squared distribution with parameter $d = (r-1)(s-1)$.

Exercise 3.5.1. The blood transfusion center in Pau has observed the following type distribution on 5000 blood donors.

Group / Factor	O	A	B	AB
Rhesus +	2291	1631	282	79
Rhesus –	325	332	48	12

1. Complete the contingency table with the marginal numbers.

   The table gives the joint numbers. The marginal numbers are obtained by summing over lines and columns.

Group / *Factor*	*O*	*A*	*B*	*AB*	*Total*
Rhesus +	2291	1631	282	79	4283
Rhesus −	325	332	48	12	717
Total	2616	1963	330	91	5000

2. Find the value of the chi-squared independence test statistic.

$$T = 5000\left(-1 + \frac{2291^2}{2616 \times 4283} + \cdots + \frac{12^2}{717 \times 91}\right) = 18.5104\ .$$

3. At threshold 1% what is your conclusion?

   Under the independence hypothesis, the test statistic follows the the chi-squared distribution with parameter $(4-1)(2-1) = 3$. *The quantile of order* 0.99 *for this distribution is* 11.3449. *Since* 18.5104 *is larger, the conclusion is that there is a dependence between the blood group and rhesus. The exact p-value is* 0.000345.

Exercise 3.5.2. Two treatments A and B are available for a certain disease. The observed results for the evolution of the disease with the two treatments on two groups of patients are as follows:

Effect / Treatment	Cured	Improved	Stationary
A	280	210	110
B	220	90	90

1. Complete the contingency table.

2. Find the value of the independence test statistic.

3. Give an interval containing the p-value for the chi-squared test of independence. Would you say that treatments A and B yield significantly different results?

Exercise 3.5.3. During 10 years, 240 persons have been followed. Among them:
- 110 had sunflower oil on a regular basis
- 25 had olive oil and cardio-vascular troubles
- 78 had sunflower oil and did not have cardio-vascular troubles.

1. Write the contingency table matching these observations.

2. Find the value of the independence test statistic.

3. Give an interval containing the p-value for the chi-squared test of independence. Would you say that cardio-vascular troubles are independent of the type of oil?

Exercise 3.5.4. The observations of a couple (X, Y) of physiological variables for 100 individuals in a population has led, after the choice of two classes for X and three for Y, to the following contingency table.

X \ Y	1	2	3	Total
1	4	11	7	22
2	16	39	23	78
Total	20	50	30	100

1. Find the value of the independence test statistic.

2. Give an interval containing the p-value for the test. What is your conclusion?

Exercise 3.5.5. After the treatment for a certain disease, 40 young patients out of 70, and 50 old patients out of 100 have shown improvements.

1. Write the contingency table matching these observations.

2. Find the value of the independence test statistic.

3. Give an interval containing the p-value for the test. Would you say that the effect of the treatment depends on the age of the patient?

Exercise 3.5.6. The following contingency table concerns 592 women grouped according to their eye and hair color.

Eyes \ Hair	Black	Brown	Red	Blond
Brown	68	119	26	7
Hazel	15	54	14	10
Green	5	29	14	16
Blue	20	84	17	94

1. Complete this contingency table.

2. Find the value of the independence test statistic.

3. Give an interval containing the p-value for the test. Would you say that the hair color is independent of the eye color?

4 Linear regression

4.1 Regression line and prediction

The data are n couples of real numbers. The first coordinate is a variable considered as *deterministic* and *explaining*. The second one is considered as *random* and *explained*. The following quantities are calculated:

- mean of explaining variable: $\overline{x} = \dfrac{1}{n}\sum_{i=1}^{n} x_i$
- mean of explained variable: $\overline{y} = \dfrac{1}{n}\sum_{i=1}^{n} y_i$
- variance of explaining variable: $s_x^2 = \dfrac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{x}^2$
- variance of explained variable: $s_y^2 = \dfrac{1}{n}\sum_{i=1}^{n} y_i^2 - \overline{y}^2$
- *covariance* of the two variables: $c_{xy} = \dfrac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{x}\,\overline{y}$
- *correlation coefficient*: $r_{xy} = \dfrac{c_{xy}}{\sqrt{s_x^2\, s_y^2}}$.
- *slope* of the regression line: $\widehat{a} = \dfrac{c_{xy}}{s_x^2}$
- *intercept with y-axis*: $\widehat{b} = \overline{y} - \widehat{a}\,\overline{x}$
- *estimated variance*: $\widehat{\sigma}^2 = \dfrac{n}{n-2} s_y^2 (1 - r_{xy}^2)$
- *prediction* of an ordinate for a given abscissa x_*: $y_* = \widehat{a}\, x_* + \widehat{b}$.

Exercise 4.1.1. In order to measure the dependence between age and thrombosis risk, 12 patients have been considered. Their age in years (variable X) and the logarithm of D-dimer concentration (variable Y) are known. The following quantities are given:

$$\sum x_i = 596 \; ; \; \sum x_i^2 = 32435 \; ; \; \sum y_i = -5.2 \; ; \; \sum y_i^2 = 4.3 \; ; \; \sum x_i y_i = -188.58 \; .$$

1. Find the correlation coefficient between X and Y.

$$\overline{x} = 49.667 \; ; \quad \overline{y} = -0.43333 \; ; \quad s_x^2 = 236.139 \; ; \quad s_y^2 = 0.17056 \; ;$$
$$c_{xy} = 5.8072 \; ; \quad r_{xy} = 0.91506 \; .$$

Note that r_{xy} is close to 1, indicating a strong correlation.

2. Find the equation for the regression line for Y onto X.

$$\widehat{a} = 0.02459 \; ; \quad \widehat{b} = -1.6548 \; .$$

The equation of the regression line is $y = 0.02459\,x - 1.6548$. *It is increasing* ($a > 0$) *because the correlation is positive: the D-dimer concentration logarithm tends to increase with the age.*

3. Find the estimated variance of the regression.

$$\widehat{\sigma}^2 = 0.0333 \; .$$

4. What value for Y would you predict for a 60 year old person?

The prediction for $x_* = 60$ *is* $y_* = 0.02459 \times 60 - 1.6548 = -0.1792$.

Exercise 4.1.2. The air pollution has been recorded in 41 American cities by the variable Y, measuring the volume of SO_2 in micrograms per m^3 of air, as well as the average annual temperature X in degrees Fahrenheit. The following numerical results are given:

$$\sum x_i = 2286 \; ; \; \sum y_i = 1232 \; ; \; \sum x_i^2 = 129549 \; ; \; \sum y_i^2 = 59050 \; ; \; \sum x_i y_i = 74598 \; .$$

1. Find the correlation coefficient of X and Y.
2. Give the equation of the regression line for Y onto X.
3. What value for Y do you predict for a city where the average annual temperature is 60°F?

Exercise 4.1.3. In a study of the duration of the vegetation period in mountains, weather stations have been installed at different altitudes. The average temperature (variable Y in degrees Celsius) and the altitude (variable X in meters) for each station are given in the table below.

altitude	1040	1230	1500	1600	1740	1950	2200	2530	2800	3100
temperature	7.4	6	4.5	3.8	2.9	1.9	1	−1.2	−1.5	−4.5

The following values are given:

$$\sum x_i = 19690; \; \sum y_i = 20.3; \; \sum x_i^2 = 42925500; \; \sum y_i^2 = 162.41; \; \sum x_i y_i = 17671 \, .$$

1. Find the correlation coefficient.
2. Find the estimates for the parameters a, b and σ^2 of the regression of Y onto X.
3. What average temperature do you predict at 1100 m?

Exercise 4.1.4. The gain in weight of a young sheep in one year (variable Y in kilograms) is thought to depend on the initial weight (variable X also in kilograms). For 10 sheep, the following results have been obtained:

$$\sum x_i = 406 \; ; \; \sum y_i = 423 \; ; \; \sum x_i^2 = 16570 \; ; \; \sum y_i^2 = 18057 \; ; \; \sum x_i y_i = 17280 \; .$$

1. Find the correlation coefficient.
2. Estimate the parameters a, b and σ^2 for the regression of Y onto X.
3. According to this model, how much weight should a sheep with an initial weight of 50 kg gain? Same question for a 30 kg sheep.

Exercise 4.1.5. The volume Y of exhaled air is a standard measurement for lung condition. In order to identify a population with weak lung condition, a model for the normal lung condition is needed. In order to do this, the volume Y in liters and the size X in centimeters for 12 boys between 10 and 15 years old were measured.

The following values were obtained:

$$\sum x_i = 1872 \; ; \; \sum y_i = 32.3 \; ; \; \sum x_i^2 = 294320 \; ; \; \sum y_i^2 = 93.11 \; ; \; \sum x_i y_i = 5156.20 \; .$$

1. Find the correlation coefficient.
2. Find the estimates for the coefficients of the regression line of Y onto X, and its variance.
3. What volume of air should a boy 1.60 m tall exhale?

Exercise 4.1.6. One wishes to predict the height H of a tree as a function of the diameter D of its trunk. To make a linear regression, the logarithms of the variables are used: $Y = \ln(H)$ and $X = \ln(D)$. Here are the results for 5 trees:

X	-1.61	-1.20	-0.97	-0.51	-0.42
Y	2.22	2.27	2.38	2.60	2.65

The following results are given:

$$\sum x_i = -4.71 \; ; \; \sum y_i = 12.12 \; ; \; \sum x_i^2 = 5.4095 \; ;$$

$$\sum y_i^2 = 29.5282 \; ; \; \sum x_i y_i = -11.0458 \; .$$

1. Find the correlation coefficient of X and Y.
2. Give the equation of the regression line of Y onto X.
3. Find the predicted height for a tree with a trunk diameter 0.7.

4.2 Confidence and prediction intervals

The intervals given in what follows have level $1-\alpha$, and t_α is the quantile of order $1-\alpha/2$ for the Student distribution $\mathcal{T}(n-2)$.

- Confidence interval for the slope a:

$$\left[\, \widehat{a} \pm t_\alpha \sqrt{\frac{\widehat{\sigma}^2}{n s_x^2}} \,\right] .$$

- Confidence interval for $ax_* + b$:

$$\left[\, \widehat{a}x_* + \widehat{b} \pm t_\alpha \sqrt{\frac{\widehat{\sigma}^2(s_x^2 + (x_* - \overline{x})^2)}{n s_x^2}} \,\right] .$$

- *Prediction* interval for $Y_* = ax_* + b + E$:

$$\left[\, \widehat{a}x_* + \widehat{b} \pm t_\alpha \sqrt{\frac{\widehat{\sigma}^2((n+1)s_x^2 + (x_* - \overline{x})^2)}{n s_x^2}} \,\right] .$$

Exercise 4.2.1. In order to measure the dependence between age and thrombosis risk, 12 patients have been observed. Their age in years (variable X) and the logarithm of D-dimer concentration (variable Y) are known. The following quantities are given:

$$\sum x_i = 596 \; ; \; \sum x_i^2 = 32435 \; ; \; \sum y_i = -5.2 \; ; \; \sum y_i^2 = 4.3 \; ; \; \sum x_i y_i = -188.58 \; .$$

1. Give a 99% confidence interval for the slope of the regression line.

   The quantile of order 0.995 *for the Student distribution with parameter* $12-2 = 10$ *is* 3.169. *The confidence interval is* $[0.0137\,;\,0.0355]$.

2. Give a 99% confidence interval for the intercept of the regression line.

   A confidence interval for b *is obtained by letting* $x_* = 0$ *in the formula giving the confidence interval for* $ax_* + b$. *The desired interval is* $[-2.2195\,;\,-1.0900]$.

3. Give a 99% confidence interval for the average value of Y among 60 year old people.

   We want a confidence interval for $ax_* + b$, *with* $x_* = 60$. *The interval is* $[-0.380\,;0.022\,]$.

4. Give a prediction interval at level 0.99 for the value of Y of one 60 year old person in particular.

   We want a prediction interval for $Y_* = ax_* + b + E$, *with* $x_* = 60$. *The interval* $[-0.791\,;0.433\,]$. *Take care not to confuse:*

- estimation *of the* mean *value of D-dimer concentration among* all 60 *year old persons*
- prediction *of the value of the D-dimer concentration for* one 60 *year old person in particular.*

In the second case, the interval is necessarily wider than in the first.

Exercise 4.2.2. The air pollution has been recorded in 41 American cities through the variable Y, measuring the volume of SO_2 in micro-grams per m^3 of air, as well as the average annual temperature X in degrees Fahrenheit. The following numerical results are given:

$$\sum x_i = 2286,\ \sum y_i = 1232,\ \sum x_i^2 = 129549,\ \sum y_i^2 = 59050,\ \sum x_i y_i = 74598\ .$$

1. Give a 95% confidence interval for the slope and intercept of the regression line.
2. Give a 95% confidence interval for the mean value of Y in the cities where the average temperature is 60°F.
3. Give a prediction interval at level 0.95 for the value of Y in a given city where the average temperature is 60°F.

Exercise 4.2.3. In a study of the duration of the vegetation period in mountains, weather stations have been installed at different altitudes. The average temperature (variable Y in degrees Celsius) and the altitude (variable X in meters) for each station are given in the table below.

altitude	1040	1230	1500	1600	1740	1950	2200	2530	2800	3100
temperature	7.4	6	4.5	3.8	2.9	1.9	1	−1.2	−1.5	−4.5

The following results are given:

$$\sum x_i = 19690;\ \sum y_i = 20.3;\ \sum x_i^2 = 42925500;\ \sum y_i^2 = 162.41;\ \sum x_i y_i = 17671\ .$$

1. Give a 95% confidence interval for the slope and the intercept of the regression line.
2. Give a 95% confidence interval for the mean temperature at 1100 m.
3. Give a prediction interval at level 0.95 for the temperature of a given place at 1100 m.

Exercise 4.2.4. The gain in weight of a young sheep in one year (variable Y in kilograms) is thought to depend on the initial weight (variable X also in kilograms). For 10 sheep, the following results are given:

$$\sum x_i = 406\ ;\ \sum y_i = 423\ ;\ \sum x_i^2 = 16570\ ;\ \sum y_i^2 = 18057\ ;\ \sum x_i y_i = 17280\ .$$

1. Give a 99% confidence interval for the two coefficients of the regression line.
2. Give a 99% confidence interval for the mean weight gain of sheep initially weighing 30 kg.
3. Give a prediction interval at level 0.99 for the weight gain of a given sheep initially weighing 30 kg.

Exercise 4.2.5. The volume Y of exhaled air is a standard measurement for lung condition. In order to identify a population with weak lung condition, a model for the normal lung condition is needed. In order to do this, the volume Y in liters and the size X in centimeters for 12 boys between 10 and 15 years old were measured.

The following numerical values were obtained:

$$\sum x_i = 1872 \; ; \; \sum y_i = 32.3 \; ; \; \sum x_i^2 = 294320 \; ; \; \sum y_i^2 = 93.11 \; ; \; \sum x_i y_i = 5156.20 \; .$$

1. Give a 99% confidence interval for the slope and intercept of the regression line.
2. Give a 99% confidence interval for the mean volume of air exhaled by boys measuring 1.60 m.
3. Give a prediction interval at level 0.99 for the volume of air exhaled by one given boy measuring 1.60 m.

Exercise 4.2.6. One wishes to predict the height H of a tree as a function of the diameter D of its trunk. To make a linear regression, the logarithms of the variables are used: $Y = \ln(H)$ and $X = \ln(D)$. Here are the results for 5 trees:

X	-1.61	-1.20	-0.97	-0.51	-0.42
Y	2.22	2.27	2.38	2.60	2.65

The following results are given:

$$\sum x_i = -4.71, \; \sum y_i = 12.12, \; \sum x_i^2 = 5.4095,$$

$$\sum y_i^2 = 29.5282, \; \sum x_i y_i = -11.0458.$$

1. Give a 95% confidence interval for the two coefficients of the regression line.
2. Give a 95% confidence interval for the mean height of trees with diameter 0.7.
3. Give a prediction interval at level 0.95 for the height of one given tree with diameter 0.7.

4.3 Tests on a regression

Under the hypothesis $\mathcal{H}_0$, the model is $Y = ax + b + E$, where E follows the normal distribution $\mathcal{N}(0, \sigma^2)$. The parameters a, b and σ^2 are unknown. They are estimated by $\widehat{a}$, $\widehat{b}$ and $\widehat{\sigma}^2$. To test particular values, the following results are used. They give the distribution of the test statistic under $\mathcal{H}_0$.

- $\sqrt{\dfrac{ns_x^2}{\widehat{\sigma}^2}}\left(\widehat{a} - a\right)$ follows the $\mathcal{T}(n-2)$.
- $\sqrt{\dfrac{ns_x^2}{\widehat{\sigma}^2(s_x^2 + (x_* - \overline{x})^2)}}\left(\widehat{a}x_* + \widehat{b} - ax_* - b\right)$ follows the $\mathcal{T}(n-2)$.
- $\sqrt{\dfrac{ns_x^2}{\widehat{\sigma}^2((n+1)s_x^2 + (x_* - \overline{x})^2)}}\left(Y_* - \widehat{a}x_* - \widehat{b}\right)$ follows the $\mathcal{T}(n-2)$.
- $(n-2)\dfrac{\widehat{\sigma}^2}{\sigma^2}$ follows the $\mathcal{X}^2(n-2)$.

The *pertinence* or validity test for the regression consists of testing $\mathcal{H}_0 : a = 0$ against $\mathcal{H}_1 : a \neq 0$, by using the first result. The regression is declared valid by rejecting $\mathcal{H}_0$.

Exercise 4.3.1. In order to measure the dependence between age and thrombosis risk, 12 patients have been observed. Their age in years (variable X) and the logarithm of D-dimer concentration (variable Y) are known. The following quantities are given:

$$\sum x_i = 596 \; ; \; \sum x_i^2 = 32435 \; ; \; \sum y_i = -5.2 \; ; \; \sum y_i^2 = 4.3 \; ; \; \sum x_i y_i = -188.58 \; .$$

1. Test the pertinence of the regression at threshold 1%.

   This is a two-tailed test of $\mathcal{H}_0 : a = 0$ *against* $\mathcal{H}_1 : a \neq 0$. *The test statistic is:*

   $$T = \sqrt{\frac{ns_x^2}{\widehat{\sigma}^2}}\,\widehat{a}\,.$$

   Under hypothesis $\mathcal{H}_0$, T *follows the Student distribution with parameter* 10. *The decision rule is:*

   $$\textit{Reject } \mathcal{H}_0 \implies T \notin [-t_\alpha \,;\, +t_\alpha]\,,$$

   where t_α *is the quantile of order* $1-\alpha/2$ *for the Student distribution* $\mathcal{T}(10)$, *that is* 3.169. *Here, the value taken by* T *is* 7.177. *The null hypothesis* $\mathcal{H}_0$ *is rejected, the pertinence of the regression is accepted.*

2. Previous studies had shown a linear dependence betwen age and logarithm of D-dimer concentration of the form $Y = 0.02x - 2$. At threshold 1%, test whether the values of a and b previously found can still be accepted.

We first test $\mathcal{H}_0 : a = 0.02$ *against* $\mathcal{H}_1 : a \neq 0.02$. *The test statistic is:*

$$T = \sqrt{\frac{ns_x^2}{\widehat{\sigma}^2}}\,(\widehat{a} - 0.02)\,.$$

It takes a value of 1.341 *which is in the interval* $[-3.169\,;\,+3.169]$. *Thus we accept* $\mathcal{H}_0$ *(declare that the estimated value of* a *is not significantly far from* 0.02*).*

We now test $\mathcal{H}_0 : b = -2$ *against* $\mathcal{H}_1 : b \neq -2$. *The test statistic is:*

$$T = \sqrt{\frac{ns_x^2}{\widehat{\sigma}^2(s_x^2 + (0-\overline{x})^2)}}(\widehat{b} - (-2))\,.$$

It takes a value of 1.935 *which is in the interval* $[-3.169\,;\,+3.169]$. *Thus we accept* $\mathcal{H}_0$ *(declare that the estimated value of* b *is not significantly far from* -2*).*

Finally, both tests accept the previously admitted values for a *and* b.

3. A 60 year old patient has a value of Y equal to 0.14: should he worry?

   We are testing here a value of $Y_* = ax_* + b + E$, *with* $x_* = 60$. *The test statistic is:*

$$T = \sqrt{\frac{ns_x^2}{\widehat{\sigma}^2((n+1)s_x^2 + (x_* - \overline{x})^2)}}(Y_* - \widehat{a}x_* - \widehat{b})\,.$$

   It takes a value of 5.028. *This value is higher than the quantile of order* 0.0005 *for the* $\mathcal{T}(10)$ *distribution, therefore it is unusually high (by reference to the available data). The patient should see a doctor.*

4. At threshold 1% test the hypothesis $\mathcal{H}_0 : \sigma^2 = 0.03$ against $\mathcal{H}_1 : \sigma^2 > 0.03$.

   The test statistic is:

$$10\frac{\widehat{\sigma}^2}{0.03}\,.$$

   Under hypothesis $\mathcal{H}_0$, T *follows the chi-squared distribution with parameter* 10. *At threshold* 1%, *the test rejects those values higher than the quantile of order* 0.99 *for the* $\mathcal{X}^2(10)$ *distribution, that is* 23.21. *Here,* T *takes a value of* 11.09, *therefore* $\mathcal{H}_0$ *is accepted.*

Exercise 4.3.2. The air pollution has been recorded in 41 American cities through the variable Y, measuring the volume of SO_2 in micro-grams per m^3 of air, as well as the average annual temperature X in degrees Fahrenheit. The following numerical results are given:

$$\sum x_i = 2286,\ \sum y_i = 1232,\ \sum x_i^2 = 129549,\ \sum y_i^2 = 59050,\ \sum x_i y_i = 74598\,.$$

1. Test the pertinence of the regression at threshold 5%.

2. Test $\mathcal{H}_0 : a = 3$ against $\mathcal{H}_1 : a < 3$, at threshold 5%.

3. You were asked to fix an upper bound for the pollution of a city the average temperature of which is 60°F. You want this upper bound to be overpassed only in 5% of the cases. What upper bound will you choose?

Exercise 4.3.3. During a study on the duration of the vegetation period in mountains, weather stations have been installed at different altitudes. The average temperature (variable Y in degrees Celsius) and the altitude (variable X in meters) for each station are given in the table below.

altitude	1040	1230	1500	1600	1740	1950	2200	2530	2800	3100
temperature	7.4	6	4.5	3.8	2.9	1.9	1	−1.2	−1.5	−4.5

The following results are given:

$$\sum x_i = 19690;\ \sum y_i = 20.3;\ \sum x_i^2 = 42925500;\ \sum y_i^2 = 162.41;\ \sum x_i y_i = 17671\,.$$

1. Test the pertinence of the regression at threshold 1%.

2. In one given place at 1100 meters, a mean temperature of 3.2 degrees has been observed. At threshold 1%, would you say that it is unusually cold up there?

Exercise 4.3.4. The gain in weight of a young sheep in one year (variable Y in kilograms) is thought to depend on the initial weight (variable X also in kilograms). For 10 sheep, the following results are given:

$$\sum x_i = 406\ ;\ \sum y_i = 423\ ;\ \sum x_i^2 = 16570\ ;\ \sum y_i^2 = 18057\ ;\ \sum x_i y_i = 17280\ .$$

1. Test the pertinence of the regression at threshold 1%.

2. An old saying states that the weight of a sheep should double in one year. At threshold 1%, can you confirm?

3. A sheep with initial weight 30 kg, has gained only 20 kg after one year. At threshold 1%, should the shepherd worry?

Exercise 4.3.5. The volume Y of exhaled air is a standard measurement for lung condition. In order to identify a population with weak lung condition, a model for the normal lung condition is needed. In order to do this, the volume Y in liters and the size X in centimeters for 12 boys between 10 and 15 years old were measured.

The following numerical results were obtained:

$$\sum x_i = 1872\ ;\ \sum y_i = 32.3\ ;\ \sum x_i^2 = 294320\ ;\ \sum y_i^2 = 93.11\ ;\ \sum x_i y_i = 5156.20\ .$$

1. Test the pertinence of the regression at threshold 1%.

2. A 1.60 m boy exhales 2.1 litres: should he worry?

Exercise 4.3.6. One wishes to predict the height H of a tree as a function of the diameter D of its trunk. To make a linear regression, the logarithms of the variables are used: $Y = \ln(H)$ and $X = \ln(D)$. Here are the results for 5 trees:

X	-1.61	-1.20	-0.97	-0.51	-0.42
Y	2.22	2.27	2.38	2.60	2.65

The following results are given:

$$\sum x_i = -4.71,\ \sum y_i = 12.12,\ \sum x_i^2 = 5.4095,$$

$$\sum y_i^2 = 29.5282,\ \sum x_i y_i = -11.0458.$$

1. Test the pertinence of the regression at threshold 5%.

2. A tree with trunk diameter 0.7 m was 20 m in height. Was it unusually tall?