Exercises and solutions: Chapter 4 only

Sébastien Ott

December 2, 2025

Exercise 4.1. Consider the following situation. We play roulette 30 times (with possible results $\{0, 1, 2, \ldots, 36\}$). Denote $X_1, X_2, \ldots, X_{30}$ the sequence of random values representing the results of each game. Suppose that after playing, we obtained the numbers

$$12, 15, 19, 21, 13, 6, 34, 9, 18, 4, 34, 5, 5, 27, 14, 34, 8,$$
$$26, 7, 25, 2, 36, 28, 21, 27, 2, 6, 14, 35, 12.$$

The total sum is 519. Define the empirical average and the average number of results in $[10, 26]$:

$$\bar{X}_{30} = \frac{1}{30}\sum_{i=1}^{30} X_i, \quad Y = \frac{1}{30}\sum_{i=1}^{30} \mathbb{1}_{[10,26]}(X_i).$$

1. What are the sample and the sample size in this experiment?
2. Are $\bar{X}_{30}, Y$ statistics?
3. What is the realization of the sample?
4. Compare the values of $\bar{X}_{30}, Y$ in the obtained sample realization to their expected values in a fair game.

We run the same experiment 3 times in different casinos and find the numbers

- Casino 1:

  $$14, 10, 11, 34, 22, 24, 32, 18, 20, 31, 12, 10, 24, 24, 13, 22, 32,$$
  $$31, 35, 19, 18, 24, 32, 15, 28, 32, 20, 35, 27, 24,$$

  total sum: 693.

- Casino 2:

  $$7, 19, 20, 0, 20, 11, 12, 3, 25, 11, 0, 4, 9, 8, 19, 22, 20,$$
  $$1, 19, 16, 4, 23, 25, 8, 14, 8, 6, 17, 15, 16,$$

  total sum: 382.

- Casino 3:

  $$13, 17, 11, 10, 24, 15, 22, 19, 20, 13, 18, 19, 25, 23, 10, 14, 21,$$
  $$13, 10, 17, 23, 22, 25, 12, 16, 26, 17, 24, 13, 25,$$

  total sum: 537.

5. What can be said about the game being fair or not in each cases?

Solution 4.1.

1. The sample is $(X_1, X_2, \ldots, X_{30})$ and the sample size is 30.
2. Yes.

3. The realization is the sequence of numbers

$$12, 15, 19, 21, 13, 6, 34, 9, 18, 4, 34, 5, 5, 27, 14, 34, 8,$$
$$26, 7, 25, 2, 36, 28, 21, 27, 2, 6, 14, 35, 12.$$

4. In a fair game, we would have that $X_1, \dots, X_{30}$ are independent random variables with uniform distribution on $\{0, 1, \dots, 36\}$. In particular, using the mean value of a uniform, we have

$$E(\bar{X}_{30}) = \frac{1}{30}\sum_{i=1}^{30} E(X_i) = \frac{1}{30} \cdot 30 \cdot \frac{36}{2} = 18,$$
$$E(Y) = \frac{1}{30} \cdot 30 \cdot P(\mathrm{Uni}(\{0, 1, \dots, 36\} \in [10, 26])) = \frac{|\{10, 11, \dots, 26\}|}{37} = \frac{17}{37} \approx 0.459.$$

The values we obtain from the sample realization are

$$\bar{X}_{30} = \frac{519}{30} = 17.3, \quad Y = \frac{12}{30} = 0.4.$$

We can say that they are fairly close to the ones of a fair game.

5. We will compare the expected value of $\bar{X}_{30}, Y$ in a the case of a fair game (uniform) to the values obtained in the sample. By the LLN, if these values are far apart, there is a high chance that the distribution of the X_i was *not* uniform. Let us compute the obtained values in each casinos:

	$\bar{X}_{30}$	Y
Fair game expect.	18	0.459
Casino 1	$\frac{693}{30} = 23.1$	$\frac{19}{30} \approx 0.63$
Casino 2	$\frac{382}{30} \approx 12.7$	$\frac{18}{30} = 0.6$
Casino 3	$\frac{537}{30} = 17.9$	$\frac{30}{30} = 1$

We can see that the value of $\bar{X}_{30}$ is way too high in the first case, way too low in the second but close to the fair game one in the third casino. On the other hand, we can see that the value of Y is completely off in the third casino. We can reasonably assume that non of these casinos are playing fair (indeed, the numbers were generated using a $\mathrm{Uni}(\{10, \dots, 36\})$ in the first case, a $\mathrm{Uni}(\{0, \dots, 26\})$ in the second, and a $\mathrm{Uni}(\{10, \dots, 26\})$ in the third).

To idea of comparing what we obtained in the experiment to some a priori guess is the central idea of *Hypotheses testing*.

Exercise 4.2. Show that for any $a > 0$,

$$\frac{1}{1 + a^{-2}} \frac{e^{-a^2/2}}{a} \leq \int_a^\infty e^{-x^2/2} dx \leq \frac{e^{-a^2/2}}{a}$$

Hint: write $1 = \frac{x}{x}$ *and integrate by part.*

Solution 4.2. For $a > 0$, let

$$I(a) = \int_a^\infty e^{-x^2/2} dx \geq 0.$$

One has

$$I(a) = \int_a^\infty \frac{1}{x} x e^{-x^2/2} dx = \big[-\tfrac{1}{x} e^{-x^2/2} \big]_a^\infty + \int_a^\infty \frac{1}{x^2} e^{-x^2/2} dx = \frac{e^{-a^2/2}}{a} - \int_a^\infty \frac{1}{x^2} e^{-x^2/2} dx$$

where we used integration by part. As $\int_a^\infty \frac{1}{x^2} e^{-x^2/2} dx \geq 0$,

$$I(a) \leq \frac{e^{-a^2/2}}{a},$$

which is the wanted upper bound. For the lower bound, notice that $\int_a^\infty \frac{1}{x^2} e^{-x^2/2} dx \leq \frac{1}{a^2} I(a)$, which gives

$$I(a) \geq \frac{e^{-a^2/2}}{a} - \frac{1}{a^2} I(a).$$

Re-arranging, we obtain

$$I(a) \geq (1 + a^{-2})^{-1} \frac{e^{-a^2/2}}{a},$$

which is the wanted lower bound.

Exercise 4.3. In each of the following examples, give the parameter space Θ and the family of probability laws $(\mathbb{P}_\theta)_{\theta \in \Theta}$ that one is led to consider.

1. We make a sequence of (potentially biased) coin flips.
2. We roll a (potentially biased) 6-faces dice many times.
3. We measure the sequence of inter-arrival times between clients in a candy store.

Solution 4.3.

1. We look at a two-outcomes experiments, that we can label 0,1. We thus need to understand the probability of getting 1 (tails will be 1 minus that). This gives a parameter $\theta = p = P(\text{coin gives } 1) \in [0, 1]$, so $\Theta = [0, 1]$, and $(\mathbb{P}_\theta)_{\theta \in \Theta}$ is the collection of Bernoulli laws: $(\text{Bern}(p))_{p \in [0,1]}$.

2. Reasoning as i the first point, we can encode the result with 1,2,3,4,5 or 6. The parameters will be $p_i = P(\text{dice rolls } i)$, $i = 1, \ldots, 6$, the parameter space will be

$$\Theta = \big\{ (p_1, \ldots, p_6) \in [0, 1]^6 : \ p_1 + \cdots + p_6 = 1 \big\},$$

and the family of laws $(\mathbb{P}_\theta)_{\theta \in \Theta}$ is the family of probability measures on $\{1, 2, \ldots, 6\}$.

3. We saw that we can model inter-arrival times using exponential random variables. So, we can take the parameter space to be $\Theta = (0, +\infty)$, as exponential random variables are determined by a positive number λ. The family of laws is then $(\text{Exp}(\lambda))_{\lambda > 0}$.

Exercise 4.4. Let $X_1, \dots, X_n$ be a sample of law $\mathbb{P}_\theta = \text{Bern}(p)$ (so that the parameter θ is p). In which of the following cases is $\hat{f}$ an estimator of $f(\theta)$? When it is the case, is $\hat{f}$ biased or not?

1. Consider $f(\theta) = p$, and
$$\hat{f}(X_1, \dots, X_n) = X_1.$$
2. Consider $f(\theta) = p$, and
$$\hat{f}(X_1, \dots, X_n) = X_1 + X_3.$$
3. Consider $f(\theta) = \text{Var}(X_1)$, and
$$\hat{f}(X_1, \dots, X_n) = \frac{1}{n}\sum_{i=1}^{n}(X_i - p)^2.$$
4. Consider $f(\theta) = p^2$, and
$$\hat{f}(X_1, \dots, X_n) = \frac{p}{n}\sum_{i=1}^{n} X_i.$$
5. Consider $f(\theta) = p^2$, and
$$\hat{f}(X_1, \dots, X_n) = \Big(\frac{1}{n}\sum_{i=1}^{n} X_i\Big)^2.$$

Solution 4.4.

1. X_1 is an estimator of p: it depends only on $X_1, X_2, \dots$ and not on p. Moreover, it is unbiased as $E(X_1) = P(X_1 = 1) = p$.
2. $X_1 + X_3$ is again an estimator of p: it depends only on $X_1, X_2, \dots$ and not on p, but it is biased $E(X_1 + X_3) = 2p$.
3. $\frac{1}{n}\sum_{i=1}^{n}(X_i - p)^2$ is **not** an estimator of p as it depends on p.
4. Same as the previous point.
5. $\big(\frac{1}{n}\sum_{i=1}^{n} X_i\big)^2$ is an estimator of p^2 as its does not depend on p^2. It is a biased estimator as
$$\begin{aligned}
E\Big(\Big(\frac{1}{n}\sum_{i=1}^{n} X_i\Big)^2\Big) &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} E(X_iX_j) \\
&= \frac{1}{n^2}\Big(\sum_{i=1}^{n} E(X_i^2) + 2\sum_{1\le i<j\le n} E(X_i)E(X_j)\Big) \\
&= \frac{1}{n^2}\Big(\sum_{i=1}^{n} E(X_i) + 2\sum_{1\le i<j\le n} p^2\Big) \\
&= \frac{1}{n^2}\big(pn + (n^2 - n)p^2\big) = p^2 + \frac{p}{n} - \frac{p^2}{n}.
\end{aligned}$$

Exercise 4.5. 1. What is the median for $X \sim \mathcal{N}(\mu, \sigma^2)$?

2. What is the median for $X \sim \text{Uni}([0,1])$?

3. Give two different medians for $X \sim \text{Uni}(\{-1, 34, 45, 55\})$.

4. Suppose $X \overset{\text{Law}}{=} -X$. Give a median for X.

5. Let $X_1, X_2, \ldots, X_n$ be an n-sample of a continuous law with density f. Suppose that $f(x) = f(-x) > 0$ for all $x \in \mathbb{R}$. Show that the median for X_1 is 0 and that the empirical median of $X_1, \ldots, X_n$ is an un-biased estimator of the median of X_1 and of the mean of X_1.

Solution 4.5. 1. It is μ as

$$P(X \geq \mu) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{\mu}^{\infty} e^{-(x-\mu)^2/2\sigma^2} = -\frac{1}{\sqrt{2\pi\sigma^2}} \int_{\mu}^{-\infty} e^{-(y+\mu)^2/2\sigma^2}$$
$$= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\mu} e^{-(y+\mu)^2/2\sigma^2} = P(X \leq \mu)$$

where we used the change of variable $y = 2\mu - x$.

2. It is 0.5 as
$$P(X \leq 0.5) = \frac{1}{2} = P(X \geq 0.5).$$

3. Any number in $(34, 45)$ gives a median for X: indeed, if $34 < a < 45$,
$$P(X \leq a) = P(X \in \{-1, 34\}) = \frac{1}{2}, \quad P(X \geq a) = P(X \in \{45, 55\}) = \frac{1}{2}.$$
We can take 40 and 41.3345 as medians.

4. We can take 0 as a median: as $X \overset{\text{Law}}{=} -X$, we have,
$$P(X \leq 0) = P(-X \leq 0) = P(X \geq 0).$$

5. By the previous point, 0 is a median for X_1 (f even implies $X_1 \overset{\text{Law}}{=} -X_1$). We need to show that it is the only one. Let $a > 0$, we have, as $X_1 \overset{\text{Law}}{=} -X_1$ and 0 is a median for X_1
$$P(X_1 \leq a) - P(X_1 \geq a) = P(X_1 \leq 0) + P(X_1 \in (0, a]) - P(X_1 \geq 0) + P(X_1 \in [0, a))$$
$$= 2P(X_1 \in [0, a]) = 2\int_0^a f(x)dx > 0,$$
as $f > 0$, where we used that X_1 is continuous which implies $P(X_1 = 0) = P(X_1 = a) = 0$. In particular, a is not a median for X_1. Proceed in the same way for $a < 0$. Now, as $X_1 \overset{\text{Law}}{=} -X_1$, $E(X_1) = 0$ which is also the median of X_1. The empirical median is an estimator of $E(X_1)$ as it does not depend on $E(X_1)$, remains to show that it is un-biased (has expectation 0 in this case). Let $\tilde{X}_1, \ldots, \tilde{X}_n$ be the increasing re-arrangement of $X_1, X_2, \ldots, X_n$ defined in the

definition of the empirical median. Then, as $(X_1, \dots, X_n) \overset{\text{Law}}{=} (-X_1, \dots, -X_n)$, we have that

$$\left(\tilde{X}_1, \dots, \tilde{X}_n\right) \overset{\text{Law}}{=} (-\tilde{X}_n, \dots, -\tilde{X}_1)$$

as taking the negative of an ordered sequence reverses the order. In particular, we obtained that the empirical median has the same law as -1 times the empirical median, thus the empirical median has expected value 0.

Exercise 4.6. Show that the empirical variance

$$\hat{\sigma}_n^2 = \overline{X^2}_n - \overline{X}_n^2$$

is a biased estimator. Show that one can remedy this by considering the estimator

$$\frac{n}{n-1}\hat{\sigma}_n^2$$

instead. Show that this estimator is convergent when $E(X_1^2) < \infty$ (i.e.: $\mathbb{P}_\theta$ has a second moment).

Solution 4.6. We compute the expected value of the estimator.

$$\begin{aligned}
E(\hat{\sigma}_n^2) &= E(\overline{X^2}_n) - E(\overline{X}_n^2) \\
&= \frac{1}{n}\sum_{i=1}^n E(X_i^2) - \frac{1}{n^2}\sum_{i,j=1}^n E(X_i X_j) \\
&= E(X_1^2) - \frac{1}{n^2}\sum_{i=1}^n E(X_i^2) - \frac{1}{n^2}\sum_{i,j=1}^n \mathbb{1}_{i\neq j} E(X_i)E(X_j) \\
&= E(X_1^2) - \frac{1}{n}E(X_1^2) - \frac{1}{n^2}n(n-1)E(X_1)^2 = \frac{n-1}{n}\mathrm{Var}(X_1),
\end{aligned}$$

where we used that the X_i's are i.i.d., and that there are $n^2 - n = n(n-1)$ elements of $\{1, \dots, n\}^2$ with non-matching entries. The quantity we wanted to estimate is $\mathrm{Var}(X_1)$, so we have a bias

$$\mathrm{Bias}_\theta(\hat{\sigma}_n^2) = E(\hat{\sigma}_n^2) - \mathrm{Var}(X_1) = -\frac{1}{n}\mathrm{Var}(X_1) \neq 0.$$

From the previous computation, we also get that

$$\mathrm{Bias}_\theta\left(\tfrac{n}{n-1}\hat{\sigma}_n^2\right) = \frac{n}{n-1}E(\hat{\sigma}_n^2) - \mathrm{Var}(X_1) = 0.$$

As $E(X_1^2) < \infty$, the LLN (Theorem **??**), implies that $\overline{X}_n \xrightarrow{\text{a.s.}} E(X_1)$, and that $\overline{X^2}_n \xrightarrow{\text{a.s.}} E(X_1^2)$ as $n \to \infty$. So, $\hat{\sigma}_n^2$ converges towards $\mathrm{Var}(X_1)$ (almost surely and therefore in probability). As $\frac{n}{n-1} \to 1$, $\frac{n}{n-1}\hat{\sigma}_n^2$ also converges towards $\mathrm{Var}(X_1)$.

Exercise 4.7. Show that $\hat{\tau}_n$ is a biased estimator of the covariance. Show that $\frac{n}{n-1}\hat{\tau}_n$ is unbiased.

Solution 4.7. $\hat{\tau}_n$ is an estimator of the covariance as it only depend on the sample $(X_1, Y_1), \dots, (X_n, Y_n)$ and not on any other parameters. We then compute its expected value.

$$\begin{aligned}
E(\hat{\tau}_n) &= E\Big(\frac{1}{n}\sum_{i=1}^n X_i Y_i - \frac{1}{n}\Big(\sum_{i=1}^n X_i\Big)\frac{1}{n}\Big(\sum_{i=1}^n Y_i\Big)\Big) \\
&= \frac{1}{n}E\Big(\sum_{i=1}^n X_i Y_i\Big) - \frac{1}{n^2}E\Big(\Big(\sum_{i=1}^n X_i\Big)\Big(\sum_{i=1}^n Y_i\Big)\Big) \\
&= \frac{1}{n}\sum_{i=1}^n E(X_i Y_i) - \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n E(X_i Y_j) \\
&= \frac{1}{n}\sum_{i=1}^n E(X_i Y_i) - \frac{1}{n^2}\sum_{i=1}^n E(X_i Y_i) - \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n \mathbb{1}_{i\neq j} E(X_i)E(Y_j) \\
&= \frac{1}{n}\cdot n \cdot E(XY) - \frac{1}{n^2}\cdot n \cdot E(XY) - \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n \mathbb{1}_{i\neq j} E(X)E(Y) \\
&= E(X_1Y_1) - \frac{1}{n}\cdot E(XY) - \frac{1}{n^2}\cdot (n^2-n)\cdot E(X)E(Y) \\
&= \frac{n-1}{n}E(XY) - \frac{n-1}{n}\cdot E(X)E(Y) = \frac{n-1}{n}\mathrm{Cov}(X,Y),
\end{aligned}$$

where we used linearity of E in the first three lines, the fact that (X_i, Y_i) and (X_j, Y_j) are independent when $i \neq j$ in the fourth line, and the fact that $(X_1, Y_1), \dots, (X_n, Y_n)$ all have the same law which is the law of (X, Y) in the fifth line. In particular,

$$E(\hat{\tau}_n) - \mathrm{Cov}(X,Y) = -\tfrac{1}{n}\mathrm{Cov}(X,Y) \neq 0,$$

so $\hat{\tau}_n$ is a biased estimator of $\mathrm{Cov}(X, Y)$, and

$$E(\tfrac{n}{n-1}\hat{\tau}_n) - \mathrm{Cov}(X,Y) = 0,$$

so $\frac{n}{n-1}\hat{\tau}_n$ is an unbiased estimator of $\mathrm{Cov}(X, Y)$.

Exercise 4.8. Suppose that we do the following experiment: pick 10 first year students uniformly at random in EPFL, and ask them their coffee consumption in litres per week, and their grade in Analysis I. Suppose that we obtain the numbers

Quest. / Student	1	2	3	4	5	6	7	8	9	10
Coffee	1.12	0.3	0.95	0.76	0.04	0.88	1.1	1.1	0.1	0.73
Analysis I	5.5	3.5	4	5	3	4.5	5	5.5	4.5	4

Based on these numbers, answer the following using a suitable convergent estimator in each cases.

1. Estimate the average weekly coffee consumption (in litres) of a first year student.
2. Estimate the average Analysis I grade of a first year student.
3. Estimate the variance of the weekly coffee consumption amongst first year students.
4. Estimate the covariance between coffee consumption and Analysis I grades.

Solution 4.8.

1. We can use the empirical mean as estimator: this gives an estimated average coffee consumption of $\frac{7.08}{10} = 0.708$ litre per week.
2. We can use the empirical mean as estimator: this gives an estimated average grade of $\frac{41.5}{10} = 4.15$.
3. We can use the empirical variance to obtain $\frac{6.5634}{10} - 0.708^2 = 0.155076$, or its unbiased version to obtain $\frac{10}{9} \cdot 0.155076 = 0.172306\overline{}$.
4. We can use the empirical covariance to obtain $\frac{33.81}{10} - 0.708 \cdot 4.15 = 0.4428$, or its unbiased version to obtain $\frac{10}{9} \cdot 0.4428 = 0.429$.

Exercise 4.9. We look at $\mathbb{P}_\theta = \mathrm{Uni}([0,\theta])$, $\theta \in \Theta = (0, +\infty)$. Let $a \in \mathbb{N}^*$. Find an estimator of θ using the moment method with $g(x) = x^a$.

Solution 4.9. Let $X \sim \mathrm{Uni}([0,\theta])$. Start by computing $E(g(X)) = E(X^a)$. We have

$$E(X^a) = \frac{1}{\theta}\int_0^\theta x^a d\theta = \frac{1}{\theta}\left[\tfrac{x^{a+1}}{a+1}\right]_0^\theta = \frac{\theta^a}{a+1}.$$

We therefore have

$$\theta = \big((a+1)E(X^a)\big)^{1/a},$$

so we want to use the method of moments with $g(x) = x^a$, and $h(x) = ((a+1)x)^{1/a}$. This gives the family of estimators of θ: for $n \geq 1$

$$\hat{\theta}_n = h\Big(\frac{1}{n}\sum_{i=1}^n g(X_i)\Big) = \Big(\frac{(a+1)}{n}\sum_{i=1}^n X_i^a\Big)^{1/a}.$$

Exercise 4.10. A cow looks at the cars passing a mountain pass. As she is a modern cow, she has a watch, a computer, and an excel sheet to take notes. During a day, she observes the following times (in minutes) between each cars:

$$1.12, 5.67, 42.93, 5.56, 9.98, 2.13, 14.21, 19.51, 9.38, 59.44, 4.75, 1.07, 30.67, 6.37$$

Supposing that the inter-arrival times are independent, identically distributed, and follow an exponential distribution $\mathrm{Exp}(\lambda)$ for some $\lambda > 0$, use the moment's method to estimate λ.

Solution 4.10. We know that if $X \sim \mathrm{Exp}(\lambda)$, we have $E(X) = \frac{1}{\lambda}$. On our size 14 sample, we can take the moment method estimator of λ given by the choices $g(x) = x$, $h(x) = \frac{1}{x}$, which gives

$$\hat{\lambda}_{14} = \Big(\frac{1}{14}\sum_{i=1}^{14} X_i\Big)^{-1}.$$

On the given realisation of our sample, this estimator gives the value

$$\hat{\lambda}_{14}(1.12, 5.67, \ldots, 6.37) = \Big(\frac{1}{14}\cdot 212.79\Big)^{-1} \approx 0.06579.$$

Exercise 4.11. Let $\mu \in \mathbb{R}$, $\sigma > 0$, and $X \sim \mathcal{N}(\mu, \sigma^2)$. Define the random variable $Y = e^X$. The law of Y is called the *log-normal distribution*.

1. Using the moment generating function of X, show that

$$E(Y) = e^{\mu + \frac{\sigma^2}{2}}, \quad \mathrm{Var}(Y) = e^{2\mu+\sigma^2}(e^{\sigma^2} - 1).$$

2. If $Y_1, Y_2, \ldots, Y_n$ is a n-sample from the log-normal distribution, find estimators of $\psi = \mu + \frac{\sigma^2}{2}$, and for σ^2 using the method of moments.

Solution 4.11. 1. Recall that the moment generating function of a $\mathcal{N}(\mu, \sigma^2)$ is given by, for $t \in \mathbb{R}$,

$$M_X(t) = E(e^{tX}) = e^{\mu t + \sigma^2 t^2/2}.$$

In particular, as $Y = e^X$,

$$E(Y) = E(e^X) = M_X(1) = e^{\mu+\sigma^2/2}, \quad E(Y^2) = E(e^{2X}) = M_X(2) = e^{2\mu+2\sigma^2},$$

so

$$\mathrm{Var}(Y) = E(Y^2) - E(Y^2) = e^{2\mu+2\sigma^2} - e^{2\mu+\sigma^2} = e^{2\mu+\sigma^2}(e^{\sigma^2} - 1).$$

2. Start with estimating $\psi = \mu + \frac{\sigma^2}{2}$. By the first point, we have that

$$e^{\psi} = E(Y) \text{ so } \psi = \ln(E(Y)).$$

We can use the method of moments with $g(x) = x$, $h(x) = \ln(x)$ to obtain the estimator

$$\hat{\psi}_n(Y_1, \ldots, Y_n) = \ln\Big(\frac{1}{n}\sum_{i=1}^{n} Y_i\Big).$$

We can then notice that

$$\frac{E(Y^2)}{E(Y)^2} = e^{\sigma^2}.$$

Following the idea of the method of moments, we can then use the empirical mean to estimate $E(Y^2)$ and $E(Y)^2$, and find an estimator of σ^2 by taking the log of the ratio of those estimations:

$$\hat{\sigma}_n^2(Y_1, \dots, Y_n) = \ln\Big(\frac{\frac{1}{n}\sum_{i=1}^n Y_i^2}{\big(\frac{1}{n}\sum_{i=1}^n Y_i\big)^2}\Big).$$

Exercise 4.12. Find the maximum likelihood estimator in the following cases.

1. $X_1, \dots, X_n$ is a sample of law $\mathrm{Exp}(\lambda)$ (so $\theta = \lambda$).
2. $X_1, \dots, X_n$ is a sample of law $\mathrm{Bern}(p)$ (so $\theta = p$).
3. $X_1, \dots, X_n$ is a sample of law $\mathrm{Uni}([0, \theta])$, $\theta > 0$ (so $\theta = \theta$).

You observe cars passing on a small country road and note the time that passes between two cars (in minutes). You find the numbers

$$8.15,\ 2.77,\ 2.94,\ 19.20,\ 13.65,\ 4.99,\ 5.48,\ 17.37,\ 11.65,\ 4.54$$

Which law would you chose to approximate the times between two cars? Use the maximum likelihood estimator to estimate the parameter of the chosen law using the given data set.

Solution 4.12. 1. We are considering the case $(\mathbb{P}_\theta)_{\theta\in\Theta} = (\mathrm{Exp}(\lambda))_{\lambda>0}$. We are in the case of continuous random variables with density $f(x) = \lambda \mathbb{1}_{[0,+\infty)}(x)e^{-\lambda x}$, so, for $x_1, \dots, x_n \geq 0$, the likelihood function is given by

$$\mathcal{L}(x_1, \dots, x_n \,|\, \lambda) = \prod_{i=1}^n \lambda e^{-\lambda x_i} = \lambda^n e^{-\lambda \sum_{i=1}^n x_i}.$$

To maximize $\mathcal{L}(x_1, \dots, x_n \,|\, \lambda)$, we can equivalently minimize the negative log-likelihood function:

$$-\ln(\mathcal{L}(x_1, \dots, x_n \,|\, \lambda)) = -n\ln(\lambda) + \lambda \sum_{i=1}^n x_i.$$

We can notice that this function is strictly convex as its second derivative in λ is equal to $\frac{n}{\lambda^2} > 0$. It therefore has a unique minimum, and its minimum is found at the critical point which is the solution to

$$-\frac{n}{\lambda} + \sum_{i=1}^n x_i = 0 \iff \lambda = \frac{n}{\sum_{i=1}^n x_i}.$$

So,

$$\mathrm{MLE}(X_1, \dots, X_n) = \frac{n}{\sum_{i=1}^n X_i}.$$

2. We are considering the case $(\mathbb{P}_\theta)_{\theta\in\Theta} = (\mathrm{Bern}(p))_{p\in[0,1]}$. We are in the case of discrete random variables taking values in $\{0,1\}$, so, for $x_1,\dots,x_n \in \{0,1\}$, the likelihood function is given by

$$\mathcal{L}(x_1,\dots,x_n \,|\, p) = \prod_{i=1}^n P(X_i = x_i) = \prod_{i=1}^n p^{x_i}(1-p)^{1-x_i} = (1-p)^n \big(\tfrac{p}{1-p}\big)^{\sum_{i=1}^n x_i}.$$

It will be simpler to write $p = 1 - e^{-\beta}$ with $\beta \geq 0$. As in the first point, we can equivalently minimize the negative log-likelihood function:

$$-\ln\big(\mathcal{L}(x_1,\dots,x_n \,|\, 1-e^{-\beta})\big) = n\beta - \ln\big(e^\beta - 1\big)\sum_{i=1}^n x_i.$$

This function is then strictly convex in β (second derivative is $\frac{e^{-\beta}}{(1-e^{-\beta})^2} > 0$). We can again find the location of the minimum via the critical point equation:

$$n = \frac{e^\beta}{e^\beta - 1}\sum_{i=1}^n x_i = \frac{1}{p}\sum_{i=1}^n x_i.$$

So,

$$\mathrm{MLE}(X_1,\dots,X_n) = \frac{1}{n}\sum_{i=1}^n X_i.$$

3. We are considering the case $(\mathbb{P}_\theta)_{\theta\in\Theta} = (\mathrm{Uni}([0,\theta]))_{\theta>0}$. We are in the case of continuous random variables with density $f(x) = \frac{1}{\theta}\mathbb{1}_{[0,\theta]}(x)$, so, for $x_1,\dots,x_n \geq 0$, the likelihood function is given by

$$\mathcal{L}(x_1,\dots,x_n \,|\, \theta) = \prod_{i=1}^n \frac{1}{\theta}\mathbb{1}_{[0,\theta]}(x_i) = \theta^{-n}\prod_{i=1}^n \mathbb{1}_{[0,\theta]}(x_i).$$

Note then that this function is 0 if there is $i \in \{1,\dots,n\}$ such that $x_i > \theta$. Moreover, if $\theta \geq \max_{i=1,\dots,n} x_i$, $\mathcal{L}(x_1,\dots,x_n \,|\, \theta) = \theta^{-n}$ which is decreasing in θ. The value θ maximizing the likelihood is thus

$$\mathrm{MLE}(X_1,\dots,X_n) = \max_{i=1,\dots,n} X_i.$$

The time between cars is a continuous random variable, so we could model the situation using either uniform or the exponential laws. As we saw several times, the exponential is in general better suited for modelling waiting times, that is the choice we make here (but the choice of uniform is also correct). We thus have the estimation of λ given by

$$\mathrm{MLE}(8.15,\dots,4.54) = \frac{10}{90.74} = \frac{1}{9.074} \approx 0.11$$

Exercise 4.13. Let $X \sim \mathrm{Uni}([0,1])$, and $Y \sim \mathrm{Exp}(1)$.

1. Find the values of the 5-quantiles of X.
2. Same question for Y.

Solution 4.13. 1. We need to find the values $t_1, \dots, t_5$ such that

$$P(X \leq t_k) = \frac{k}{5},\ k = 1, \dots, 5.$$

We have

$$P(X \leq t_k) = \int_0^{t_k} dx = t_k.$$

So, the 5-quantiles of X are $\frac{1}{5}, \frac{2}{5}, \frac{3}{5}, \frac{4}{5}, 1$.

2. We need to find the values $t_1, \dots, t_5$ such that

$$P(Y \leq t_k) = \frac{k}{5},\ k = 1, \dots, 5.$$

We have

$$P(Y \leq t_k) = \int_0^{t_k} e^{-x} dx = 1 - e^{-t_k}.$$

So, t_k is the solution to $1 - e^{-t_k} = \frac{k}{5}$. The 5-quantiles of Y are thus $-\ln\big(1 - \frac{1}{5}\big)$, $-\ln\big(1 - \frac{2}{5}\big)$, $-\ln\big(1 - \frac{3}{5}\big)$, $-\ln\big(1 - \frac{4}{5}\big)$, $+\infty$.

Exercise 4.14. Let $X \sim \mathcal{N}(0, 1)$ and $Y \sim \mathcal{N}(2, 4)$. Answer the following questions using the table of Figure **??**.

1. For what value of t do we have $P(X \leq t) = 0.9$?
2. For what value of t do we have $P(X > t) = 0.05$?
3. For what value of t do we have $P(|X| > t) = 0.05$?
4. For what value of t do we have $P(Y \leq t) = 0.8$?

Solution 4.14. 1. $P(X \leq 0) = 0.5$ so $P(X \in [0, t]) = 0.4$, which gives $t = 1.28$.

2. $P(X > t) = 0.05$ so $P(X \in [0, t]) = 0.45$, which gives $t = 1.65$.

3. $P(|X| > t) = 0.05$ so $P(X > t) = 0.025$ so $P(X \in [0, t]) = 0.475$, which gives $t = 1.96$.

4. By the properties of Gaussian, see Lemma **??**,

$$P(Y \leq t) = P(Y - 2 \leq t - 2) = P\big(\tfrac{Y-2}{2} \leq \frac{t-2}{2}\big) = P\big(X \leq \frac{t-2}{2}\big).$$

So, we look for t such that $P\big(X \leq \frac{t-2}{2}\big) = 0.8$. First, let us see that we cannot find such t with $t \leq 2$. Indeed, in that case we have $P\big(X \leq \frac{t-2}{2}\big) \leq P(X \leq 0) = 0.5$. Now, for $t > 2$, $P\big(X \leq \frac{t-2}{2}\big) = 0.8$ is equivalent to $P\big(X \in \big[0, \frac{t-2}{2}\big]\big) = 0.3$, so, using Table **??**, $\frac{t-2}{2} = 0.84$, so $t = 3.68$.

Exercise 4.15. For $k \geq 1$ integer, the χ^2 square law with k degrees of freedom is the continuous probability measure on $\mathbb{R}$ with density

$$f_{\chi^2_k}(x) = \mathbb{1}_{[0,+\infty)}(x)\frac{x^{\frac{k}{2}-1}}{2^{\frac{k}{2}}\Gamma(\frac{k}{2})}e^{-x/2},$$

where $\Gamma : \mathbb{R}_+ \to \mathbb{R}$ is the Gamma function. We denote $X \sim \chi^2_1$ for "the random variable X follows a χ^2_1 law".

1. Using $\Gamma(\frac{1}{2}) = \sqrt{\pi}$, show that if $Z \sim \mathcal{N}(0,1)$, then $Z^2 \sim \chi^2_1$.
2. Let $X_1, X_2, \ldots$ be an independent family of uniform random variables on $\{-1, 1\}$. Using the CLT, show that

$$\frac{1}{n}\Big(\sum_{i=1}^n X_i\Big)^2 \xrightarrow{\text{Law}} \chi^2_1.$$

   Then, using Table **??**, find the value of

$$\lim_{n\to\infty} P\Big(\Big(\sum_{i=1}^n X_i\Big)^2 \in [2.26n, 2.5n]\Big).$$

3. Let $X, Y \sim \mathcal{N}(0,1)$ be two independent Gaussian random variables. Using $\Gamma(1) = 1$, show that $X^2 + Y^2 \sim \chi^2_2$.
4. For X, Y as in the previous point, use Table **??** to find $C \in \mathbb{R}$ such that $P(X^2 + Y^2 \leq C) = 0.893$. Also, find the value of $P(X^2 + Y^2 \in [3, 4])$.

Solution 4.15. 1. We will prove it by computing the CDF of Z^2. For $t \leq 0$, we have $F_{X^2}(t) = P(X^2 \leq t) = 0$, as for χ^2_1 random variables. For $t > 0$, we have

$$\begin{aligned}
F_{X^2}(t) &= P(X^2 \leq t) \\
&= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} \mathbb{1}_{(-\infty,t]}(x^2)e^{-x^2/2}dx \\
&= \frac{1}{\Gamma(2^{-1})\sqrt{2}}2\int_0^{+\infty} \mathbb{1}_{[0,t]}(x^2)e^{-x^2/2}dx \\
&= \frac{1}{\Gamma(2^{-1})\sqrt{2}}2\int_0^{+\infty} \mathbb{1}_{[0,t]}(y)\frac{1}{2\sqrt{y}}e^{-y/2}dy \\
&= \int_{-\infty}^{+\infty} \mathbb{1}_{(-\infty,t]}(y)f_{\chi^2_1}(y)dy
\end{aligned}$$

   which is the repartition function of the χ^2_1 law, where we used that $x \mapsto \mathbb{1}_{[0,t]}(x^2)e^{-x^2/2}$ is even, and did the change of variable $y = x^2$, $dx = \frac{1}{2\sqrt{y}}dy$.

2. By the first point, if $Z \sim \mathcal{N}(0,1)$, $Z^2 \sim \chi^2_1$. We need to show that for any $t \in \mathbb{R}$, the CDF of $\frac{1}{n}\Big(\sum_{i=1}^n X_i\Big)^2$ converges to the CDF of a χ^2_1, which is equivalent to

show that for any $t \in \mathbb{R}$,

$$P\Big(\frac{1}{n}\Big(\sum_{i=1}^{n} X_i\Big)^2 \leq t\Big) \xrightarrow{n\to\infty} P(Z^2 \leq t).$$

For $t < 0$, both sides are 0 for any n, so there is nothing to show. Let $t \geq 0$. Then,

$$P\Big(\frac{1}{n}\Big(\sum_{i=1}^{n} X_i\Big)^2 \leq t\Big) = P\Big(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i \in [-\sqrt{t}, \sqrt{t}]\Big) \xrightarrow{n\to\infty} P(Z \in [-\sqrt{t}, \sqrt{t}])$$

by the CLT as $\mathrm{Var}(X_1) = 1$, $E(X_1) = 0$. Finally, $P(Z \in [-\sqrt{t}, \sqrt{t}]) = P(Z^2 \geq t)$ which is what we wanted. Now, we have

$$\lim_{n\to\infty} P\Big(\Big(\sum_{i=1}^{n} X_i\Big)^2 \in [2.26n, 2.5n]\Big) = \lim_{n\to\infty} P\Big(\frac{1}{n}\Big(\sum_{i=1}^{n} X_i\Big)^2 \in [2.26, 2.5]\Big)$$
$$= P(Z^2 \in [2.26, 2.5]) = P(Z^2 \leq 2.5) - P(Z^2 \leq 2.26) = 0.886 - 0.867 = 0.019$$

by the previous convergence, the fact that the χ^2_1 law is continuous, and Table **??**.

3. We proceed similarly to the first point and compute the CDF of $X^2 + Y^2$. As in the first point, it is sufficient to consider $t \geq 0$. Then,

$$\begin{aligned} F_{X^2+Y^2}(t) &= P(X^2 + Y^2 \leq t) \\ &= \frac{1}{2\pi}\int_{-\infty}^{+\infty} dx \int_{-\infty}^{+\infty} dy \mathbb{1}_{[0,t]}(X^2 + Y^2) e^{-x^2/2} e^{-y^2/2} \\ &= \frac{1}{2\pi}\int_0^{2\pi} d\theta \int_0^{\infty} dr \mathbb{1}_{[0,t]}(r^2) r e^{-r^2/2} \end{aligned}$$

where we used polar coordinates, see Remark **??**:

$$x = r\cos(\theta),\ y = r\sin(\theta), \quad dxdy = rd\theta dr,$$

so

$$x^2 + y^2 = r^2(\cos^2(\theta) + \sin^2(\theta)) = r^2.$$

Now, as $\Gamma(1) = 1$, $\int_0^{2\pi} d\theta = 2\pi$, and using the change of variable $s = r^2$, $dr = \frac{1}{2\sqrt{s}} ds$, the last expression is equal to

$$\frac{1}{\Gamma(1)}\int_0^{\infty} ds \mathbb{1}_{[0,t]}(s) \frac{\sqrt{s}}{2\sqrt{s}} e^{-s/2} = \int_{-\infty}^{\infty} ds \mathbb{1}_{(-\infty,t]}(s) f_{\chi^2_2}(s)$$

which is indeed the CDF of a χ^2_2 evaluated at t.

4. By the previous point, $X^2+Y^2 \sim \chi^2_2$, so by Table **??**, $P(X^2+Y^2 \leq 4.47) = 0.893$, so $C = 4.47$. Then, still by Table **??**, and the fact that $X^2 + Y^2$ is a continuous random variable,

$$P(X^2+Y^2 \in [3,4]) = P(X^2+Y^2 \leq 4) - P(X^2+Y^2 \leq 3) = 0.865 - 0.777 = 0.088.$$

Exercise 4.16. During a test, we asked 190 people who was Niki Lauda. 19 of them answered "a songwriter" (wrong), and the others had the correct answer. Based on this experiment, answer the following.

1. Give an estimation of the percentage of the population who do not know who Niki Lauda is.

2. Given a 95% confidence interval for this proportion.

Solution 4.16.

1. We can estimate the percentage using the empirical mean: $\frac{171}{190} = \frac{9}{10} = 90\%$. More formally, we want to let X be a Bernoulli random variable that gives 1 if a uniformly sampled individual in the population answers the question correctly and 0 else. We want to estimate the parameter of this law, $p \in [0,1]$. We then write $X_i = 1$ if the ith person answers correctly to the question, and $X_i = 0$ else. $X_1, \dots, X_{190}$ is then a 190-sample of law $\mathrm{Bern}(p)$. As $p = E(X)$, we can use the empirical mean to estimate it. With the result of the experiment, this gives the estimation $\hat{p}(x_1, \dots, x_{190}) = \frac{171}{190} = \frac{9}{10}$.

2. We seek an interval $I(X_1, \dots, X_{190})$ such that the event $p \in I$ has probability $\frac{95}{100}$. We can use the CLT and Gaussian approximation (supposing that 190 is large enough) to say that

$$\frac{1}{\sqrt{190p(1-p)}} \sum_{i=1}^{190} (X_i - p) \approx \mathcal{N}(0,1).$$

We can take an interval of the form

$$I(X_1, \dots, X_{190}) = [\hat{p} - \delta, \hat{p} + \delta].$$

To find a suitable δ, we compute

$$\begin{aligned}
\sup_{p\in[0,1]} P\big(|\hat{p} - p| \geq \delta\big) &= \sup_{p\in[0,1]} P\Big(\Big|\frac{1}{190}\sum_{i=1}^{190}(X_i - p)\Big| \geq \delta\Big) \\
&= \sup_{p\in[0,1]} P\Big(\Big|\frac{1}{\sqrt{190p(1-p)}}\sum_{i=1}^{190}(X_i - p)\Big| \geq \frac{\sqrt{190}\delta}{\sqrt{p(1-p)}}\Big) \\
&\approx \sup_{p\in[0,1]} P\Big(|Z| \geq \frac{\sqrt{190}\delta}{\sqrt{p(1-p)}}\Big) \\
&= P\big(|Z| \geq 4\sqrt{190}\delta\big) \approx 2P(Z \geq 55.14\delta)
\end{aligned}$$

where $Z \sim \mathcal{N}(0,1)$. We then seek δ such that

$$2P(Z \geq 55.14\delta) = \frac{5}{100} \iff P(0 \leq Z \leq 55.14\delta) = 0.475.$$

Looking at the Gaussian table **??**, we find that $55.14\delta = 1.96$ solves this, so $\delta = 0.036$ does the job.

Exercise 4.17. We pick randomly 10 apricots from a tree. Their weight (in grams) are

$$30.2 \quad 40.7 \quad 35.1 \quad 36.2 \quad 38.8 \quad 29.9 \quad 42.1 \quad 45.2 \quad 39.5 \quad 37.1$$

Assume that the weight of an apricot is Gaussian with positive variance that is given by the empirical variance of our realisation of the sample.

1. Obtain a 90% confidence interval for the mean weight μ of an apricot.
2. Try to explain what "90% confidence interval" means in this example.

Solution 4.17. We assume that an apricot weight follows a Gaussian law $\mathcal{N}(\mu, \sigma^2)$. The variance is a priori unknown, but we assume it is given by the empirical variance: the empirical mean of our sample realisation is $374,8/10 = 37.48$, and the empirical variance is $215.036/10 = 21.5036 \approx 21.5$. We can thus make the assumption that the weight of an apricot is following a $\mathcal{N}(\mu, 21.5)$. As we know the weight of an apricot is non-negative, we can take as parameter space $\Theta = (0, +\infty)$, $\theta = \mu$. Using these assumptions, we can find a confidence interval for μ centred around the empirical mean: we look for $\delta > 0$ such that for all $\mu > 0$,

$$P(|\overline{X}_{10} - \mu| \geq \delta) = 10\%$$

where $X_1, \ldots, X_{10}$ are i.i.d. $\mathcal{N}(\mu, 21.5)$. By the properties of Gaussian, this is the same as finding $\delta > 0$ such that

$$P\big(|Z| \geq \tfrac{\delta}{\sqrt{2.15}}\big) = 10\% \iff P\big(Z \geq \tfrac{\delta}{\sqrt{2.15}}\big) = 5\%$$

where $Z \sim \mathcal{N}(0, 1)$, as $\overline{X}_{10} - \mu \sim \mathcal{N}(0, 2.15)$. For such a δ (that we can find using tables or R), $[\overline{X}_{10} - \delta, \overline{X}_{10} + \delta]$ is a 90% confidence interval.

A possible meaning of "$[\overline{X}_{10} - \delta, \overline{X}_{10} + \delta]$ is a 90% confidence interval for μ" is that if we were to perform our sampling (picking 10 apricots at random) several times, we will see that the empirical mean of the sample is δ-close to the real mean 90% of the time. So our assumption $\mu[\overline{X}_{10} - \delta, \overline{X}_{10} + \delta]$ is correct in 90% of our sampling repetitions.

Exercise 4.18. A physical constant μ is measured 30 times in a lab. We make the assumption that the measurements follow a normal $\mathcal{N}(\mu, \sigma^2)$ with some known variance σ^2. Based on the measurements, the 90% confidence interval for μ obtain using an interval centred around the empirical mean is $[6.04, 6.18]$. How many more measurements must be taken if we want:

1. to divide the length of the confidence interval for μ by 2?
2. to obtain a 95% confidence interval of the same length?

Solution 4.18. The setup is that we have $X_1, \ldots, X_n$ a sample of $\mathcal{N}(\mu, \sigma^2)$. We use the confidence interval

$$I(X_1, \ldots, X_n) = [\overline{X}_n - \delta, \overline{X}_n + \delta]$$

for some $\delta > 0$ as a confidence interval for μ. The confidence level of our test is

$$P(\mu \in I(X_1, \dots, X_n)) = P(|\overline{X}_n - \mu| \leq \delta) = P\big(|\mathcal{N}(0,1)| \leq \tfrac{\delta\sqrt{n}}{\sigma}\big).$$

In a realisation $x_1, \dots, x_{30}$ of our sample with $n = 30$, we find the interval $[6.04, 6.18]$. In particular, we have that $\overline{x}_{30} = 6.11$ (the midpoint) and $\delta = 0.7$. In particular, we know from the given information that

$$P\big(|\mathcal{N}(0,1)| \leq \tfrac{0.7\sqrt{30}}{\sigma}\big) = 0.9.$$

so

$$P\big(\mathcal{N}(0,1) \in \big[0, \tfrac{0.7\sqrt{30}}{\sigma}\big]\big) = 0.45. \tag{4.1}$$

Looking at Table **??**, this gives $\frac{0.7\sqrt{30}}{\sigma} = 1.64$, which is $\sigma = \frac{0.7\sqrt{30}}{1.64}$.

1. If we want to divide the length of the confidence interval, i.e.: to have $\delta = 0.35$ instead of $\delta = 0.7$, we need n such that

   $$P\big(|\mathcal{N}(0,1)| \leq \tfrac{0.35\sqrt{n}}{\sigma}\big) = P\big(|\mathcal{N}(0,1)| \leq \tfrac{0.7\sqrt{30}}{\sigma}\big) = 0.9,$$

   we can thus take a sample size n such that $\frac{0.35\sqrt{n}}{\sigma} = \frac{0.7\sqrt{30}}{\sigma}$ so, $n = 120$.

2. If we want to obtain a 95% confidence interval with the same value $\delta = 0.7$, we need to take n such that

   $$P\big(|\mathcal{N}(0,1)| \leq \tfrac{0.7\sqrt{n}}{\sigma}\big) = 0.95.$$

   using the value of σ we had, we need to solve

   $$P\big(|\mathcal{N}(0,1)| \leq \tfrac{1.64\cdot\sqrt{n}}{\sqrt{30}}\big) = 0.95,$$

   which is equivalent to

   $$P\big(\mathcal{N}(0,1) \in \big[0, \tfrac{1.64\cdot\sqrt{n}}{\sqrt{30}}\big]\big) = 0.475.$$

   Looking at Table **??**, we find that this is the same as asking $\frac{1.64\cdot\sqrt{n}}{\sqrt{30}} = 1.96$, so $n = 43$.

Exercise 4.19. We will try Hypotheses testing on a size-1 sample. Let $X \sim \text{Uni}([0, \theta])$ (the parameter space is $\Theta = (0, +\infty)$). We want to test the null Hypotheses H_0 "the mean of X is at least 10". Let $C \geq 0$. Consider the rejection region

$$D_C = \{x \in \mathbb{R} : \ x < C\}.$$

1. What is the parameter space region Θ_0 corresponding to the null Hypotheses H_0? and the parameter space region Θ_1 corresponding to the alternative hypotheses H_1?

2. If we sample X and find 23.45, for which values of C do we reject H_0 in the test procedure associated with the rejection region D_C?

3. Compute the risk level and the power of the test associated with the rejection region D_C.

Solution 4.19. 1. As $E(X) = \frac{\theta}{2}$, the condition "$E(X) \geq 10$" is equivalent to $\theta \geq 20$, so we can take $\Theta_0 = [20, +\infty)$, $\Theta_1 = \Theta \setminus \Theta_0 = (0, 20)$.

2. For any $C > 23.45$.

3. We first compute the risk level α.

$$\alpha = \sup_{\theta \in \Theta_0} P(X \in D_C) = \sup_{\theta \geq 20} P(X < C) = \sup_{\theta \geq 20} \min(\tfrac{C}{\theta}, 1) = \min(\tfrac{C}{20}, 1).$$

In particular, if $C \geq 20$, the confidence level of the test is 0. We then compute the power of the test $1 - \beta$:

$$\begin{aligned}\beta = \sup_{\theta \in \Theta_1} P(X \notin D_C) = \sup_{0<\theta<20} P(X \geq C) &= \sup_{0<\theta<20} \max(\tfrac{\theta - C}{\theta}, 0) \\ &= \sup_{0<\theta<20} \max(1 - \tfrac{C}{\theta}, 0) = \max(1 - \tfrac{C}{20}, 0).\end{aligned}$$

In particular, if $C \geq 20$, the power of the test is 1.

Exercise 4.20. Take the same setup as Example **??**. Based on what was done there, construct a test to test the Hypotheses H_0: "$|\mu_X - \mu_Y| \geq 10$" with confidence level at least 95/100. What is the power of that test?

Solution 4.20. We take the same test idea as in the Example: we used a rejection region of the form

$$\big\{|\overline{X}_{100} - \overline{Y}_{100}| \geq \delta\big\},$$

to test the null Hypotheses "$\mu_X - \mu_Y = 0$". This suggest a rejection region of the form

$$\big\{|\overline{X}_{100} - \overline{Y}_{100}| \geq 10 + \delta\big\}$$

to test our null Hypotheses H_0 "$|\mu_X - \mu_Y| \leq 10$" which corresponds to the parameter space $\Theta_0 = \{(\mu_X, \mu_Y) \in \mathbb{R}^2 : \ |\mu_X - \mu_Y| \leq 10\}$. We want a test with confidence 0.95 (i.e.: risk $\alpha = 0.05$). Now, let us find δ as function of the risk level α.

$$\begin{aligned}\alpha &= \sup_{\theta \in \Theta_0} P\big(|\overline{X}_{100} - \overline{Y}_{100}| \geq 10 + \delta\big) \\ &= \sup_{\substack{\mu_X, \mu_Y \in \mathbb{R}, \\ |\mu_X - \mu_Y| \leq 10}} P\big(|\overline{X}_{100} - \overline{Y}_{100}| \geq 10 + \delta\big) \\ &= \sup_{|\mu| \leq 10} P\big(|\sqrt{225.1}Z + \mu| \geq 10 + \delta\big) \\ &= \sup_{|\mu| \leq 10} P\Big(Z \notin \Big[\frac{\mu - 10 - \delta}{\sqrt{225.1}}, \frac{\mu + 10 + \delta}{\sqrt{225.1}}\Big]\Big) \\ &= P\Big(Z \notin \Big[\frac{-\delta}{\sqrt{225.1}}, \frac{20 + \delta}{\sqrt{225.1}}\Big]\Big)\end{aligned}$$

where $Z \sim \mathcal{N}(0, 1)$ and we used that, as in the Example, $\overline{X}_{100} - \overline{Y}_{100} \sim \mathcal{N}(\mu_X - \mu_Y, 225.1)$, and thus $\overline{X}_{100} - \overline{Y}_{100} \overset{\text{Law}}{=} \sqrt{225.1}Z + \mu_X - \mu_Y$. The last line uses that the

supremum is realized at $\mu \in \{-10, 10\}$ and that the two give the same probability as Z is symmetric. We then compute

$$P\Big(Z \notin \Big[\frac{-\delta}{\sqrt{225.1}}, \frac{20+\delta}{\sqrt{225.1}}\Big]\Big) = P\Big(Z > \frac{20+\delta}{\sqrt{225.1}}\Big) + P\Big(Z < \frac{-\delta}{\sqrt{225.1}}\Big)$$
$$= 0.5 - P\Big(0 \le Z \le \frac{20+\delta}{\sqrt{225.1}}\Big) + 0.5 - P\Big(0 \le Z \le \frac{\delta}{\sqrt{225.1}}\Big).$$

To get a risk level at most 0.05, we can ask that

$$P\Big(0 \le Z \le \frac{20+\delta}{\sqrt{225.1}}\Big) \ge 0.494, \quad P\Big(0 \le Z \le \frac{\delta}{\sqrt{225.1}}\Big) \ge 0.456.$$

Looking at Table **??**, we find that $\frac{20+\delta}{\sqrt{225.1}} \ge 2.49$ does the job for the first condition (so $\delta \ge 17.4$), and that $\frac{\delta}{\sqrt{225.1}} \ge 1.71$ (so $\delta \ge 25.7$) does the job for the second. Let's take $\delta = 26$ for simplicity.

Let us now compute the power $1 - \beta$ of the test by computing β, the worst case scenario for the probability of type-II error.

$$\begin{aligned}
\beta &= \sup_{\theta \in \Theta_1} P\big(|\overline{X}_{100} - \overline{Y}_{100}| \le 10 + \delta\big) \\
&= \sup_{\substack{\mu_X, \mu_Y \in \mathbb{R}, \\ |\mu_X - \mu_Y| > 5}} P\big(|\overline{X}_{100} - \overline{Y}_{100}| \le 36\big) \\
&= \sup_{|\mu| > 10} P\big(|\sqrt{225.1}Z + \mu| \le 36\big) \\
&= \sup_{|\mu| > 10} P\Big(Z \in \Big[\frac{\mu - 36}{\sqrt{225.1}}, \frac{\mu + 36}{\sqrt{225.1}}\Big]\Big) \\
&= P\Big(Z \in \Big[\frac{-26}{\sqrt{225.1}}, \frac{46}{\sqrt{225.1}}\Big]\Big) \le 2P\big(0 \le Z \le \tfrac{26}{\sqrt{225.1}}\big)
\end{aligned}$$

where $Z \sim \mathcal{N}(0,1)$, and we used the same properties as in the computation for finding the value of δ. Now, as $\frac{26}{\sqrt{225.1}} \approx 1.73$, we can use Table **??** to find that $P\big(0 \le Z \le \frac{26}{\sqrt{225.1}}\big) \approx 0.458$, and thus $\beta \le 0.916$, giving a power of 0.084, only a infinitesimal improvement on the power of the test performed in the example. Better tests will be seen in the course, such as *Two-samples t-tests*.

Exercise 4.21. Let $1 \leq n \leq N$ be integers. For $i_1, \ldots, i_n \in \{1, \ldots, N\}$, define

$$p_{N,n}(i_1, \ldots, i_n) = \frac{(N-n)!}{N!} \prod_{j=1}^{n} \prod_{l \in \{1,\ldots,n\} \setminus j} \mathbb{1}_{i_j \neq i_l},$$

$$q_{N,n}(i_1, \ldots, i_n) = \frac{1}{N^n},$$

$p_{N,n}$ is the mass function of the probability measure associated with putting N numbered balls in a bag, and drawing *without replacement* n of them. $q_{N,n}$ is the mass function of the probability measure associated with putting N numbered balls in a bag, and drawing *with replacement* n of them. Let $P_{N,n}$ be the probability measure on $\{1, \ldots, N\}^n$ with mass function $p_{N,n}$, and let $Q_{N,n}$ be the probability measure on $\{1, \ldots, N\}^n$ with mass function $q_{N,n}$.
Introduce the event

$$A_n = \big\{(i_1, \ldots, i_n) \in \{1, \ldots, N\}^n : \ \forall j \neq l,\, i_j \neq i_l\big\},$$

that all drawn numbers are different.

1. Show that for any event $B \subset \{1, \ldots, N\}^n$, $P_{N,n} = Q_{N,n}(B \,|\, A_n)$.
2. Show that if $n \leq N/2$, $Q_{N,n}(A_n) \geq 1 - \frac{n^2}{N}$. *Hint: show and use that for any* $x \in [0, 0.5]$, $1 - x \geq e^{-2x}$, *and that for any* $x \geq 0$, $e^{-x} \geq 1 - x$.

Solution 4.21. We first notice that

$$\mathbb{1}_{A_n}(i_1, \ldots, i_n) = \prod_{j,l \in \{1,\ldots,N\}:\, j \neq l} \mathbb{1}_{i_j \neq i_l} = \prod_{j=1}^{n} \prod_{l \in \{1,\ldots,N\} \setminus j} \mathbb{1}_{i_j \neq i_l}.$$

Then, note that the number of sequences in $\{1, \ldots, N\}^n$ with all their entries being different is $\frac{N!}{(N-n)!}$. Thus,

$$Q_{N,n}(A_n) = \frac{\frac{N!}{(N-n)!}}{N^n} = \frac{N!}{(N-n)! \cdot N^n}.$$

1. We have that for any event $B \subset \{1, \ldots, N\}^n$,

$$\begin{aligned}
Q_{N,n}(B \,|\, A_n) &= \frac{Q_{N,n}(B \cap A_n)}{Q_{N,n}(A_n)} \\
&= \frac{(N-n)! \cdot N^n}{N!} \sum_{(i_1,\ldots,i_n) \in B} \mathbb{1}_{A_n}(i_1, \ldots, i_n) q_{N,n}(i_1, \ldots, i_n) \\
&= \frac{(N-n)! \cdot N^n}{N!} \sum_{(i_1,\ldots,i_n) \in B} \Big(\prod_{j=1}^{n} \prod_{l \in \{1,\ldots,N\} \setminus j} \mathbb{1}_{i_j \neq i_l}\Big) \frac{1}{N^n} \\
&= \sum_{(i_1,\ldots,i_n) \in B} \frac{(N-n)!}{N!} \prod_{j=1}^{n} \prod_{l \in \{1,\ldots,N\} \setminus j} \mathbb{1}_{i_j \neq i_l} \\
&= \sum_{(i_1,\ldots,i_n) \in B} p_{N,n}(i_1, \ldots, i_n) = P_{N,n}(B).
\end{aligned}$$

2. First, for $x \in [0, 0.5]$,

$$1 - x = \exp(\ln(1-x)) = \exp\left(\int_0^x \frac{-1}{1-s} ds\right) \geq \exp\left(-2\int_0^x ds\right) = e^{-2x},$$

as $s \in [0, x] \subset [0, 0.5]$ implies $\frac{1}{1-s} \leq 2$. Then, for $x \geq 0$,

$$e^{-x} - 1 = -\int_0^x e^{-s} ds \geq -\int_0^x ds = -x$$

as $e^{-s} \leq 1$ for $s \geq 0$. So, for $x \geq 0$, $e^{-x} \geq 1 - x$. Now,

$$\begin{aligned} Q_{N,n}(A_n) = \frac{N!}{(N-n)! \cdot N^n} &= \prod_{k=0}^{n-1} \frac{N-k}{N} = \prod_{k=0}^{n-1} \left(1 - \frac{k}{N}\right) \geq \prod_{k=0}^{n-1} e^{-2k/N} \\ &= \exp\Big(-\frac{2}{N}\sum_{k=0}^{n-1} k\Big) = \exp\big(-\tfrac{1}{N}(n-1)n\big) \geq e^{-\frac{n^2}{N}} \geq 1 - \tfrac{n^2}{N} \end{aligned}$$

as $\frac{k}{N} < \frac{n}{N} \leq \frac{1}{2}$ by assumption.

Exercise 4.22. We have a sample $X_1, \ldots, X_{50}$ from a law $\mathbb{P}$ on $\mathbb{R}$, and we wonder if this law is a centred Gaussian. More precisely, we want to test the Hypotheses H_0 "$\mathbb{P} = \mathcal{N}(0, \sigma^2)$ for some $\sigma^2 > 0$". We want to do so with an χ^2 adequacy test. We partition $\mathbb{R}$ (which is the realization space of our experiment) as the disjoint union of $\mathbb{R}_- = (-\infty, 0)$ and $\mathbb{R}_+ = [0, +\infty)$.

1. Under the null Hypotheses, what is the probability for X_1 to fall into $\mathbb{R}_-$ and $\mathbb{R}_+$?
2. Give the test statistic and the degrees of freedom in this case if we want to perform a χ^2 adequacy test.
3. Give the rejection region corresponding to a level 0.9 confidence test ($\alpha = 0.1$). Use the approximation that 50 is close enough to infinity to be able to replace the test statistic by its limit in doing the computation of the rejection region.

We find the realisation

1.10, −0.33, −0.03, 0.90, 2.75, 1.55, −0.46, −1.12, 1.21, 1.65, −1.11, 0.04,
3.40, 3.37, 1.47, 0.69, 0.13, 1.07, 2.72, 0.63, −1.57, 2.72, 1.11, 0.30,
− 0.57, 6.37, 0.08, 3.25, 1.30, 1.09, 3.20, 2.57, 2.56, −3.30, 2.16, 1.99,
0.98, 3.39, 3.09, 2.16, 0.84, −0.05, −1.64, −0.57, −0.30, −0.70, 4.84, −0.54,
0.91, −2.12

4. Using the test setted up in the first points, do you reject the null Hypotheses?

Solution 4.22. 1. Under H_0, X_1 is a centred Gaussian, it is therefore a continuous symmetric random variable, so

$$P(X_1 < 0) = P(X_1 \leq 0) = P(X_1 \geq 0) = P(X_1 > 0) = 0.5.$$

2. The test statistic is

$$T(X_1,\dots,X_{50}) = 50\Big(\frac{\big(\frac{\sum_{i=1}^{50}\mathbb{1}_{(-\infty,0)}(X_i)}{50}-0.5\big)^2}{0.5}+\frac{\big(\frac{\sum_{i=1}^{50}\mathbb{1}_{[0,+\infty)}(X_i)}{50}-0.5\big)^2}{0.5}\Big)$$
$$= 100\Big(\Big(\frac{\sum_{i=1}^{50}\mathbb{1}_{(-\infty,0)}(X_i)}{50}-0.5\Big)^2+\Big(\frac{\sum_{i=1}^{50}\mathbb{1}_{[0,+\infty)}(X_i)}{50}-0.5\Big)^2\Big).$$

There are $2-1=1$ degrees of freedom as we partitioned the space into two classes, and there are therefore $2-1$ parameters being estimated.

3. By assumption that 50 is large enough, we can assume that T follows a χ^2_1 law to determine a rejection region. As we want a confidence level 0.9, we chose a rejection region of the form

$$\{T(X_1,\dots,X_{50})>\delta\},$$

with δ such that $P(Z\le\delta)=0.9$ with $Z\sim\chi^2_1$. Looking at Table **??**, we find that $\delta=2.7$ does the job.

4. In the given realisation of the sample, we have that there are 35 non-negative numbers and 15 negative numbers. So, $\sum_{i=1}^{50}\mathbb{1}_{(-\infty,0)}(x_i)=15$, and $\sum_{i=1}^{50}\mathbb{1}_{[0,+\infty)}(x_i)=35$, and thus

$$T(x_1,\dots,x_{50}) = 100\Big(\Big(\frac{15}{50}-0.5\Big)^2+\Big(\frac{35}{50}-0.5\Big)^2\Big) = 100\Big(\frac{1}{25}+\frac{1}{25}\Big) = 8$$

which is well above $\delta=2.7$, so we reject H_0 in this case.

Exercise 4.23. We have a group of 60 students who have to pass an exam. We want to test a new concentration enhancing drug on them. We assume that the variance of the test result is the same for all student, independently of whether they take the drug or not. Set a test procedure to test the null Hypotheses "the drug has an effect on the average performance of students" using a two-samples t-test. Explicitly state the assumptions you make along the way. Find a 0.9 confidence level test given your assumptions.
You now obtain the test results of 30 students who were given the drug: $x_1, \dots, x_{30}$ is equal to

63.9, 72.8, 78.1, 76.7, 50.7, 77.7, 77.9, 60.2, 57.9, 47.3, 94.6, 37.1, 66.7, 43.8, 50.4, 56.0, 51.5, 51.6, 49.5, 55.4, 37.6, 43.1, 83.0, 48.3, 49.8, 63.7, 57.5, 69.7, 61.8, 93.9

and the test results of 30 students who were not given the drug: $y_1, \dots, y_{30}$ is equal to

54.0, 22.3, 39.4, 43.7, 41.2, 61.1, 47.7, 26.0, 41.2, 38.4, 38.3, 34.3, 37.4, 33.6, 48.4, 22.0, 44.3, 28.8, 70.0, 22.1, 34.7, 47.6, 37.8, 58.6, 58.8, 57.5, 47.2, 31.7, 15.7, 16.6

Do you accept the null Hypotheses in this case? You can use that $P(0 \leq \text{Student}_\text{t}(58) \leq 1.67) \approx 0.45$, and that the above data satisfy

$$\sum_{i=1}^{30} x_i = 1828.2, \quad \sum_{i=1}^{30} y_i = 1200.4,$$
$$\sum_{i=1}^{30} (x_i - \overline{x}_n)^2 = 6880.032, \quad \sum_{i=1}^{30} (y_i - \overline{y}_n)^2 = 5426.395$$

Solution 4.23. We divide the student into two groups of equal sizes: 30 are given the drug, and 30 are given a placebo or nothing at all. Denote $X_1, \dots, X_{30}$ the sample of the test results of students with the drug, and $Y_1, \dots, Y_{30}$ the sample of the test results of student without the drug. We assume that all test results are independent and Gaussian, all with the same variance. In mathematical terms, we assume that there are $\mu_X, \mu_Y \in \mathbb{R}$, and $\sigma^2 > 0$ such that $X_1, \dots, X_{30}, Y_1, \dots, Y_{30}$ form an independent sequence and

$$X_i \sim \mathcal{N}(\mu_X, \sigma^2), \ i = 1, \dots, 30, \quad Y_i \sim \mathcal{N}(\mu_Y, \sigma^2), \ i = 1, \dots, 30.$$

We want to perform a two-sample t-test for testing the Hypotheses H_0 "$\mu_X = \mu_Y$" (so $\Theta = \mathbb{R}^2 \times (0, +\infty)$, $\theta = (\mu_X, \mu_Y, \sigma^2)$, and $\Theta_0 = \{(\mu, \mu, \sigma^2) : \ \mu \in \mathbb{R}, \sigma^2 > 0\}$). The test statistic is therefore

$$T(X_1, \dots, X_{30}, Y_1, \dots, Y_{30}) = \frac{\sqrt{30}(\overline{X}_{30} - \overline{Y}_{30})}{\sqrt{s_X^2 + s_Y^2}}$$

with

$$s_X^2 = \frac{1}{29} \sum_{i=1}^{30} (X_i - \overline{X}_{30})^2, \quad s_Y^2 = \frac{1}{29} \sum_{i=1}^{30} (Y_i - \overline{Y}_{30})^2.$$

Under the null Hypotheses H_0, $\mu_X = \mu_Y$ and therefore, by our assumptions on the Gaussian nature of the variables, $T \sim \text{Student}_\text{t}(2 \cdot 30 - 2) = \text{Student}_\text{t}(58)$ under H_0.

We take a rejection region of the form

$$\{|T(X_1,\dots,X_{30},Y_1,\dots,Y_{30})| \geq \delta\},$$

with δ such that $P(\text{Student}_t(58) \in [-\delta,\delta]) = 0.9$, as we want a confidence level 0.9. Using that the Student law is symmetric,

$$P(\text{Student}_t(58) \in [-\delta,\delta]) = 2P(\text{Student}_t(58) \in [0,\delta]).$$

Looking at the given numbers, we have that $\delta = 1.67$ does the job. We then look at the value of our test statistic on the given realisation of the samples.

$$T(x_1,\dots,x_{30},y_1,\dots,y_{30}) \approx \frac{\sqrt{30}(\frac{1828}{30}-\frac{1200}{30})}{\sqrt{\frac{6880}{29}+\frac{5426}{29}}} = \frac{628\cdot\sqrt{29}}{\sqrt{30}\cdot\sqrt{12306}} \approx 5.57$$

which is well above $\delta = 1.67$, so we reject the null Hypotheses "the drug has no influence on the test results".

Exercise 4.24. Let $n,m \geq 1$ be integers. Let $X \sim \mathrm{F}(n,m)$. Then,

1. Show that if $m > 2$,
$$E(X) = \frac{m}{m-2}.$$
2. Show that if $m > 4$,
$$\mathrm{Var}(X) = \frac{2m^2(m+n-2)}{n(m-2)^2(m-4)}.$$
3. Show that if $n = m > 4$,
$$P(|X-E(X)| \geq \epsilon) \leq \frac{12}{(n-4)\epsilon^2}.$$

Solution 4.24. 1. We compute using the density of the Fisher law: for $m > 2$,

$$\begin{aligned}
E(X) &= \int_{-\infty}^{+\infty} x f_X(x)dx \\
&= \frac{1}{\mathrm{B}(n/2,m/2)}\int_0^{+\infty}\Big(\frac{nx}{nx+m}\Big)^{n/2}\Big(1-\frac{nx}{nx+m}\Big)^{m/2}dx \\
&= \frac{m}{n}\frac{1}{\mathrm{B}(n/2,m/2)}\int_0^1 \underbrace{y^{\frac{n}{2}}}_{\downarrow}\underbrace{(1-y)^{\frac{m}{2}-2}}_{\uparrow}dy \\
&= \frac{mn}{n2(\frac{m}{2}-1)}\frac{1}{\mathrm{B}(n/2,m/2)}\int_0^1 y^{\frac{n}{2}-1}(1-y)^{\frac{m}{2}-1}dy \\
&= \frac{m}{m-2}\frac{1}{\mathrm{B}(n/2,m/2)}\mathrm{B}(n/2,m/2) = \frac{m}{m-2},
\end{aligned}$$

where we used the change of variable $y = \frac{nx}{nx+m}$, $x = \frac{my}{n(1-y)}$, $dx = \frac{m}{n(1-y)^2}dy$, in the third line, and integrated by part in the fourth.

2. We start by computing the second moment of X using the same change of variable as in the previous point. For $m > 4$,

$$\begin{aligned}
E(X^2) &= \int_{-\infty}^{+\infty} x^2 f_X(x) dx \\
&= \frac{1}{\mathrm{B}(n/2, m/2)} \int_0^{+\infty} x \Big(\frac{nx}{nx+m}\Big)^{n/2} \Big(1 - \frac{nx}{nx+m}\Big)^{m/2} dx \\
&= \frac{1}{\mathrm{B}(n/2, m/2)} \int_0^1 \frac{my}{n(1-y)} y^{\frac{n}{2}} (1-y)^{\frac{m}{2}} \frac{m}{n(1-y)^2} dy \\
&= \frac{m^2}{n^2} \frac{1}{\mathrm{B}(n/2, m/2)} \int_0^1 \underbrace{y^{\frac{n}{2}+1}}_{\downarrow} \underbrace{(1-y)^{\frac{m}{2}-3}}_{\uparrow} dy \\
&= \frac{m^2}{n^2} \frac{1}{\mathrm{B}(n/2, m/2)} \Big(\frac{n}{2} + 1\Big) \frac{1}{\frac{m}{2} - 2} \int_0^1 \underbrace{y^{\frac{n}{2}}}_{\downarrow} \underbrace{(1-y)^{\frac{m}{2}-2}}_{\uparrow} dy \\
&= \frac{m^2(n+2)}{n^2(m-4)} \frac{1}{\mathrm{B}(n/2, m/2)} \frac{n}{2} \frac{1}{\frac{m}{2} - 1} \int_0^1 y^{\frac{n}{2}-1} (1-y)^{\frac{m}{2}-1} dy \\
&= \frac{m^2(n+2)}{n(m-4)(m-2)}
\end{aligned}$$

where we used integration by part in the fifth and sixth lines. We then compute the variance:

$$\mathrm{Var}(X) = E(X^2) - E(X)^2 = \frac{m^2(n+2)}{n(m-4)(m-2)} - \frac{m^2}{(m-2)^2} = \frac{2m^2(m+n-2)}{n(m-2)^2(m-4)}.$$

3. We can use Chebychev's inequality and the previous point,

$$P(|X - E(X)| \geq \epsilon) \leq \frac{\mathrm{Var}(X)}{\epsilon^2} = \frac{2n^2(2n-2)}{n(n-2)^2(n-4)\epsilon^2} = \frac{4n(n-1)}{(n-2)^2(n-4)\epsilon^2} \leq \frac{12}{(n-4)\epsilon^2}$$

as $n(n-1) \leq 3(n-2)^2$ for $n > 4$.

Exercise 4.25. We measured the wine consumption of 74 former Swiss politicians (in litres per week). We found the following data.
For 12 former members of the "conseil fédéral": $x_1, \dots x_{12}$ are equal to

$$5.4,\ 5.2,\ 5.2,\ 4.7,\ 5.6,\ 5.6,\ 4.8,\ 5.4,\ 6.1,\ 4.9,\ 3.2,\ 4.8$$

For 26 former members of the "assemblée fédérale": $y_1, \dots y_{26}$ are equal to

$$2.9,\ 2.6,\ 3.4,\ 3.8,\ 5.5,\ 2.1,\ 4.3,\ 3.2,\ 2.6,\ 3.4,\ 1.5,\ 2.0,\ 2.4,\ 5.0,\ 4.9,\ 2.3,$$
$$3.0,\ 3.2,\ 2.3,\ 3.2,\ 4.9,\ 4.5,\ 3.0,\ 3.5,\ 4.7,\ 4.2$$

For 36 former members of "conseils d'états": $z_1, \dots z_{36}$ are equal to

$$3.4,\ 1.3,\ 3.8,\ 3.2,\ 2.2,\ 4.4,\ 0.9,\ 4.4,\ 2.8,\ 0.4,\ 0.8,\ 1.5,\ 2.8,\ 1.2,\ 2.6,\ 2.9,\ 3.7,\ 2.6,\ 3.6,$$
$$0.3,\ 1.9,\ 1.7,\ 3.4,\ 3.3,\ 3.2,\ 1.5,\ 2.4,\ 1.7,\ 2.7,\ 1.9,\ 1.3,\ 2.7,\ 3.9,\ 2.1,\ 2.0,\ 1.8$$

We want to test whether belonging to different political levels has an influence on the amount of wine drank per week using a one-way ANOVA. Setup the test procedure stating the assumptions made at each step. Set the parameters so that the test has a confidence level of 0.9. Do you reject the Hypotheses H_0 "all political levels have the same average wine consumption" based on the test procedure you created and the given data?
You can use

$$\sum_{i=1}^{12} x_i = 60.9, \quad \sum_{i=1}^{26} y_i = 88.4, \quad \sum_{i=1}^{36} z_i = 86.3,$$
$$\sum_{i=1}^{12} (x_i - \overline{x}_{12})^2 \approx 5.7, \quad \sum_{i=1}^{26} (y_i - \overline{y}_{26})^2 \approx 28.6, \quad \sum_{i=1}^{36} (z_i - \overline{z}_{36})^2 \approx 41.1,$$

and the fact that

$$P(\mathrm{F}(2, 71) \leq 2.379) = 0.9.$$

Solution 4.25. We have 74 politicians divided into three groups of sizes $12, 26, 36$. Denote $X_1, \dots, X_{12}$ a sample of the first group, $Y_1, \dots, Y_{26}$ a sample of the second group, and $Z_1, \dots, X_{36}$ a sample of the third group. We make the assumption that weekly wine consumption is normally distributed and that the variance of an individual consumption is not depending on the individual. Mathematically, we assume that there are $\mu_{X,1}, \dots, \mu_{X,12} \in \mathbb{R}$, $\mu_{Y,1}, \dots, \mu_{Y,26} \in \mathbb{R}$, $\mu_{Z,1}, \dots, \mu_{Z,36} \in \mathbb{R}$, and $\sigma^2 > 0$ such that

$$\begin{aligned} X_i &\sim \mathcal{N}(\mu_{X,i}, \sigma^2),\ i = 1, \dots, 12, \\ Y_i &\sim \mathcal{N}(\mu_{Y,i}, \sigma^2),\ i = 1, \dots, 26, \\ Z_i &\sim \mathcal{N}(\mu_{Z,i}, \sigma^2),\ i = 1, \dots, 36. \end{aligned}$$

We then use the test statistic

$$\begin{aligned} &T(X_1, \dots, X_{12}, Y_1, \dots, Y_{26}, Z_1, \dots, Z_{36}) \\ &\qquad = \frac{71(12(\overline{X}_{12} - \overline{\mu}_{74})^2 + 26(\overline{Y}_{26} - \overline{\mu}_{74})^2 + 36(\overline{Z}_{36} - \overline{\mu}_{74})^2)}{2(s_{X,12}^2 + s_{Y,26}^2 + s_{Z,36}^2)} \end{aligned}$$

where

$$\overline{X}_{12} = \frac{1}{12}\sum_{i=1}^{12} X_i, \quad \overline{Y}_{26} = \frac{1}{26}\sum_{i=1}^{26} Y_i, \quad \overline{Z}_{36} = \frac{1}{36}\sum_{i=1}^{36} Z_i,$$

$$\overline{\mu}_{74} = \frac{1}{74}\Big(\sum_{i=1}^{12} X_i + \sum_{j=1}^{26} Y_j + \sum_{k=1}^{36} Z_k\Big) = \frac{12\overline{X}_{12} + 26\overline{Y}_{26} + 36\overline{Z}_{36}}{74}$$

$$s^2_{X,12} = \sum_{i=1}^{12}(X_i - \overline{X}_{12})^2, \quad s^2_{Y,26} = \sum_{i=1}^{26}(Y_i - \overline{Y}_{26})^2, \quad s^2_{Z,36} = \sum_{i=1}^{36}(Z_i - \overline{Z}_{36})^2.$$

By assumptions on the sample, $T \sim \mathrm{F}(2, 71)$. As suggested in the description of the test, we take a rejection region of the form

$$\{T > \delta\}.$$

We now want a confidence level 0.9, which is achieved by picking δ such that $P(\mathrm{F}(2,71) \leq \delta) = 0.9$, which is given to be $\delta = 2.379 \approx 2.4$. We now perform the test on the given realisation. We have

$$\overline{x}_{12} = \tfrac{60.9}{12} \approx 5.1, \quad \overline{y}_{26} = \tfrac{88.4}{26} = 3.4, \quad \overline{z}_{36} = \tfrac{86.3}{36} \approx 2.4, \quad \overline{\mu}_{74} \approx \tfrac{12\cdot 5.1 + 26\cdot 3.4 + 36\cdot 2.4}{74} \approx 3.2,$$

and, from the given data,

$$s^2_{x,12} \approx 5.7, \quad s^2_{y,26} = 28.6, \quad s^2_{z,36} = 41.1.$$

So, on our realisation,

$$T(x_1, \ldots, x_{12}, y_1, \ldots, y_{26}, z_1, \ldots, z_{36}) \approx \frac{71(12(5.1-3.2)^2 + 26(3.4-3.2)^2 + 36(2.4-3.2)^2)}{2(5.7 + 28.6 + 41.1)} \approx 31.7$$

This is well above $\delta = 2.4$, so we reject the null Hypotheses.

Exercise 4.26. Let $X_1, \ldots, X_n$ be an n-sample of law $\mathrm{Uni}([0,\theta])$, $\theta > 0$ (so $\Theta = (0, +\infty)$).

1. Show that the estimator of θ $\mathrm{Est}_1(X_1, \ldots, X_n) = 2\overline{X}_n$ is an unbiased estimator of θ.
2. Show that the estimator $\mathrm{Est}_2(X_1, \ldots, X_n) = \mathrm{MLE}(X_1, \ldots, X_n)$ is a continuous random variable with density $f(x) = \frac{n}{\theta^n} x^{n-1} \mathbb{1}_{(0,\theta)}(x)$.
3. Show that Est_2 is a biased estimator of θ, but that it is asymptotically unbiased.
4. Compute $\mathrm{MSE}_\theta(\mathrm{Est}_1)$ and $\mathrm{MSE}_\theta(\mathrm{Est}_2)$.
5. What do you observe?

Solution 4.26. 1. $\mathrm{Est}_1(X_1, \ldots, X_n) = 2\overline{X}_n$ is an estimator of θ as it does not depend on θ. It is an un-biased estimator as

$$E(\mathrm{Est}_1(X_1, \ldots, X_n)) = \frac{2}{n}\sum_i^n E(X_i) = \frac{2}{n}\sum_i^n \frac{\theta}{2} = \theta.$$

2. Recall first that by Exercise **??**,

$$\mathrm{MLE}(X_1, \dots, X_n) = \max(X_1, \dots, X_n).$$

We first compute the CDF of $\mathrm{Est}_2 = \max(X_1, \dots, X_n)$. For $t \leq 0$, we obviously have $F_{\mathrm{Est}_2}(t) = 0$, and for $t \geq \theta$, $F_{\mathrm{Est}_2}(t) = 1$. Then, for $t \in [0, \theta]$,

$$F_{\mathrm{Est}_2}(t) = \prod_{i=1}^{n} F_{X_i}(t) = \frac{t^n}{\theta^n}$$

by Exercise **??**, and the fact that $P(X \leq t) = \frac{t}{\theta}$ for $t \in [0, \theta]$ and $X \sim \mathrm{Uni}([0, \theta])$. Then, we can recall that the repartition function is a primitive of the density function, so that we can find the density $f_{\mathrm{Est}_2}(x)$ as a constant c to be determined plus $F'_{\mathrm{Est}_2}(x) = \mathbb{1}_{(0,\theta)}(x)\frac{nx^{n-1}}{\theta^n}$. As

$$\int_0^\theta \frac{nx^{n-1}}{\theta^n} dx = 1,$$

we have that $c = 0$, and $f_{\mathrm{Est}_2}(x) = F'_{\mathrm{Est}_2}(x) = \mathbb{1}_{(0,\theta)}(x)\frac{nx^{n-1}}{\theta^n}$ as wanted.

3. We compute the expected value of Est_2:

$$\begin{aligned} E(\mathrm{Est}_2) = E\big(\max(X_1, \dots, X_n)\big) &= \int_0^\theta \frac{nx^{n-1}}{\theta^n} x dx \\ &= \frac{n}{\theta^n}\int_0^\theta x^n dx = \frac{n}{\theta^n}\big[\tfrac{1}{n+1}x^{n+1}\big]_0^\theta = \frac{n}{(n+1)}\theta, \end{aligned}$$

so $E(\mathrm{Est}_2) - \theta = -\frac{1}{n+1}\theta \xrightarrow{n\to\infty} 0$ so Est_2 is biased but asymptotically un-biased.

4. We first compute $\mathrm{MSE}_\theta(\mathrm{Est}_1)$:

$$\begin{aligned} \mathrm{MSE}_\theta(\mathrm{Est}_1) = E\big((2\overline{X}_n - \theta)^2\big) &= E\big((2\overline{X}_n - 2E(\overline{X}_n))^2\big) \\ &= 4\mathrm{Var}(\overline{X}_n) = \frac{4}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i) = \frac{4}{n^2}n\frac{\theta^2}{12} = \frac{\theta^2}{3n}. \end{aligned}$$

Then compute $\mathrm{MSE}_\theta(\mathrm{Est}_1)$:

$$\begin{aligned} \mathrm{MSE}_\theta(\mathrm{Est}_2) = E\big((\mathrm{Est}_2 - \theta)^2\big) &= \int_{-\infty}^{+\infty}(x-\theta)^2 f_{\mathrm{Est}_2}(x)dx \\ = \frac{1}{\theta^n}\int_0^\theta \underbrace{(x-\theta)^2}_{\downarrow}\underbrace{nx^{n-1}}_{\uparrow}dx &= \frac{1}{\theta^n}\big[(x-\theta)^2 x^n\big]_0^\theta - \frac{2}{\theta^n}\int_0^\theta (x-\theta)x^n dx \\ = -\frac{2}{\theta^n}\Big(\int_0^\theta x^{n+1}dx - \theta\int_0^\theta x^n dx\Big) &= -\frac{2}{\theta^n}\Big(\frac{\theta^{n+2}}{n+2} - \theta\frac{\theta^{n+1}}{n+1}\Big) = \frac{2\theta^2}{(n+1)(n+2)} \end{aligned}$$

where we integrated by part in the fourth equality.

5. If we look at the difference between the two risks, we find

$$\begin{aligned}\mathrm{MSE}_\theta(\mathrm{Est}_1) - \mathrm{MSE}_\theta(\mathrm{Est}_2) &= \frac{\theta^2}{3n} - \frac{2\theta^2}{(n+1)(n+2)} \\ &= \theta^2 \frac{(n+1)(n+2) - 6n}{3n(n+1)(n+2)} = \theta^2 \frac{n^2 - 3n + 2}{3n(n+1)(n+2)}.\end{aligned}$$

In particular, for any $n \geq 3$, the second estimator has a strictly smaller mean square error than the first, whilst being biased!