Solutions

September 13, 2024

1 Revision on probability

Exercise 1

a) The marginal distribution (pmf) of Z is

$Z=-1$	$Z=5$
$0.1+0.2$	$0.5+0.2$

Then $\mathbb{E}\left(Z\right) = -1\times 0.3 + 0.7\times 5 = 3.2$, $\mathbb{E}\left(Z^2\right) = 1\times 0.3 + 0.7\times 25 = 17.8$, $\mathbb{V}\left(Z\right) = 17.8 - 3.2^2 = 7.56$.

b)$\mathbb{E}\left(Y\right) = -4\times 0.6 + 0.4\times 6 = 0$, $\mathbb{E}\left(Y^2\right) = 16\times 0.6 + 0.4\times 36 = 24$, $\mathbb{V}\left(Y\right) = 24$.

c) Covariance? The cross moment is

$$\begin{aligned}\mathbb{E}\left(YZ\right) &= 0.1\times 4 - 0.5\times 20 - 6\times 0.2 + 30\times 0.2\\ &= -4.8\end{aligned}$$

Then $\sigma_{Z,Y} = -4.8 - 0\times 3.2 = -4.8$ and $\rho_{Z,Y} = \frac{-4.8}{\sqrt{7.56}\sqrt{24}} = -0.3563$.

Exercise 2

a) k is such that $\int_{-1}^{1} k(1+x^2) = 1$ then $k=\frac{3}{8}$.

b) $\mathbb{E}(X) = k\int_{-1}^{1}(x+x^3)dx = 0$. $\mathbb{V}\left(X\right) = \mathbb{E}(X^2) = k\int_{-1}^{1}(x^2+x^4)dx = k\frac{16}{15} = \frac{2}{5}$.

c) quantile q of order α is such that $\int_{-1}^{q_\alpha} k(1+x^2)dx = \alpha$ but

$$\begin{aligned}\int_{-1}^{q_\alpha} k(1+x^2)dx &= \frac{k}{3}(q^3+3q+4)\\ &= \frac{1}{8}(q^3+3q+4)\end{aligned}$$

Then q is the solution of $\frac{1}{8}(q^3+3q+4) = \alpha$ inside the interval $[-1,1]$.

d) The pdf of Y is

$$\begin{aligned}
f_Y(y) &= \frac{1}{3}f_X(\frac{y-2}{3}) = \frac{k}{3}(1+\left(\frac{y-2}{3}\right)^2) \\
&= \frac{1}{8}(1+\frac{1}{9}\left(y^2-4y+4\right)) \\
&= \frac{1}{8}+\frac{1}{72}\left(y^2-4y+4\right) \\
&= \frac{1}{72}\left(y^2-4y+13\right)
\end{aligned}$$

and the support is $\in [-1,5]$.

e) In theory

$$\mathbb{E}\left(XY\right) \quad = \quad \int_{-1}^{1}\int_{-1}^{5} xy\, f(x,y)\, dy\, dx$$

where $f(x,y)$ is the joint density of (X,Y), which is here unknown. But we don't need it since Y is a linear function of X. Then the correlation $\rho_{X,Y}$ is $+1$ (Notice that if $Y := -3X+2$ the correlation would be -1). The covariance is then

$$\sigma_{X,Y} = \rho_{X,Y}\sigma_X\sigma_Y$$

where $\sigma_X = \sqrt{\frac{2}{5}}$. The expectation of Y is $\mathbb{E}\left(Y\right) = 3\mathbb{E}\left(X\right)+2 = 2$ and its variance is

$$\mathbb{V}\left(Y\right) = 9\mathbb{V}\left(X\right) = \frac{18}{5}$$

then $\sigma_Y = \sqrt{\frac{18}{5}}$ and $\sigma_{X,Y} = \sqrt{\frac{2}{5}}\sqrt{\frac{18}{5}} = \frac{6}{5}$.

Remark: the joint density $f(u,v)$ is the function such that

$$P(X \le x\,,\, Y \le y) \quad = \quad \int_{-\infty}^{x}\int_{-\infty}^{y} f(u,v)dv\, du$$

where $x \in [-1,1]$ and $y \in [-1,5]$ but here

$$\begin{aligned}
P(X \le x\,,\, Y \le y) =& P(X \le x\,,\, 3X+2 \le y) \\
=& P(X \le x\,,\, X \le \frac{y-2}{3}) \\
=& P(X \le x \cap X \le \frac{y-2}{3}) \\
=& P\left(X \le \min\left(x\,,\, \frac{y-2}{3}\right)\right)
\end{aligned}$$

but we do not need this result.

Exercise 3

a) Y is a binomial random variable with $n = 1000$ and $p = 1 - 0.18 = 0.82$. Then

$$\begin{aligned} \mathbb{E}(Y) &= np = 820 \\ \mathbb{V}(Y) &= np(1-p) = 147.6 \end{aligned}$$

b) According to the central limit theorem $Y \approx \mathcal{N}(\mu, \sigma^2)$ where $\mu = 820$ and $\sigma = \sqrt{147.6}$. Using standardization:

$$P(Y \geq 840) = P(Z \geq \frac{840 - 820}{\sqrt{147.6}}) = P(Z \geq 1.645)$$

where $Z \sim \mathcal{N}(0, 1)$. From tables we infer that $P(Y \geq 840) = 0.05$. In a similar way, using symmetry:

$$\begin{aligned} P(Y \leq 815) &= P(Z \leq \frac{815 - 820}{\sqrt{147.6}}) = P(Z \leq -0.411) \\ &= P(Z \geq 0.411) = 0.3409 \end{aligned}$$

Exercise 4

a) False. The covariance (and correlation) are measure of linear dependence. A null covariance simply means that X and Y do not display any linear relationship. But they may be dependent at higher order.

b) True. This is an exception and the only case in which a null covariance (or correlation) implies independence and vice-versa.

c) False. The covariance is equal to

$$\sigma_{X,Y} = \mathbb{E}\left(X^3\right) - \mathbb{E}(X)\,\mathbb{E}\left(X^2\right)$$

and the correlation is

$$\rho_{X,Y} = \frac{\mathbb{E}\left(X^3\right) - \mathbb{E}(X)\,\mathbb{E}\left(X^2\right)}{\sqrt{\mathbb{E}(X^2) - \mathbb{E}(X)^2}\sqrt{\mathbb{E}(X^4) - \mathbb{E}(X^2)^2}}$$

which is different from 1 in nearly all cases.

d) False. Only if X and Y have a null covariance.

Exercise 5

a) $Z \sim N(\mu_Z, \sigma_Z)$ where

$$\mu_Z = 2\mathbb{E}(X) + 3\mathbb{E}(Y) - 1 = 4 + 9 - 1 = 12$$

The covariance between X and Z is

$$\begin{aligned}\mathbb{C}(X,Y) &= \rho_{X,Y}\sqrt{\mathbb{V}\,(X)}\sqrt{\mathbb{V}\,(Y)}\\ &= 0.63\times 0.5\times 0.8\\ &= 0.252\,.\end{aligned}$$

Therefore

$$\begin{aligned}\sigma_Z^2 &= 2^2\mathbb{V}\,(X)+3^2\mathbb{V}(Y)+2\times 2\times 3\mathbb{C}(X,Y)\\ &= 4\times 0.5^2+9\times 0.8^2+12\times 0.252\\ &= 9.784\end{aligned}$$

or $\sigma_Z = 3.128$. Using standardization and the notation $W \sim \mathcal{N}(0,1)$

$$\begin{aligned}P(Z\geq 14) &= P(W\geq \frac{14-12}{3.128})\\ &= P(W\geq 0.64)\\ &= 0.2611\end{aligned}$$

Exercise 6

a) Notation: H=highly contaminated , M=averagely contaminated and L=healthy.

$$\begin{aligned}P(H) &= P(X\geq 3)\\ &= \int_3^\infty xe^{-x}dx\end{aligned}$$

But the gamma function is $\Gamma(z)=\int_0^\infty x^{z-1}e^{-x}dx$ and the incomplete gamma function is $\Gamma(z,a)=\int_a^\infty x^{z-1}e^{-x}dx$. Therefore, $P(H)=\Gamma(2,3)=0.1991$ (numerical calculation). Remark : if you do not wish to use the incomplete gamma function, you can as well solve the integral using the technique of integration by parts.

$$\begin{aligned}P(L) &= P(X\leq 1)\\ &= 1-P(X\geq 1)\\ &= 1-\Gamma(2,1)\\ &= 0.2642\end{aligned}$$

Finally,

$$\begin{aligned}P(M) &= P(X\leq 3)-P(X\leq 1)\\ &= 1-\Gamma(2,3)-\Gamma(2,1)\\ &= 0.5367\end{aligned}$$

b) We use a binomial random variable, Y, with $n=25$ trials and the probability of "success" is $p=P(H)=0.1991$. Using tables available on Moodle:

$$P(Y\leq 3) \quad = 0.234$$

c) Assumption $Y \sim \mathcal{N}\,(np = 4.9775\,,\, np(1-p) = 3.9865)$. Using standardization and symmetry, we infer from tables that

$$\begin{aligned} P(Y \leq 3) &= P\left(Z \leq \frac{3 - 4.9775}{\sqrt{3.9865}}\right) \\ &= P\,(Z \leq -0.9904) \\ &= P\,(Z \geq 0.9904) \\ &= 0.1611\,, \end{aligned}$$

which is quite far from the real probability of 0.234. In fact, the size of the sample set is too small. See slides (central limit theorem): the criterion $n > 9\frac{max(p,1-p)}{min(p,1-p)} = 36.20$ is not fulfilled. We need more experiment to approximate the binomial r.v. by a normal one.

Exercise 7

a) Let us denote by $t_{df,\alpha}$ the α-quantile of a Student with df degrees of freedom. From the table, we directly have $t_{20,0.975} = 2.086$ and by symmetry, we have $t_{20,0.025} = -t_{20,0.975} = -2.086$.
b) Using a similar notation to before, we obtain from the table $\chi^2_{20,0.025} = 9.59083$ and $\chi^2_{20,0.975} = 34.1696$.
c) Let us denote by $F_{df_1,df_2;\alpha}$ the α-quantile of a Fisher distribution with df_1 degrees of freedom on the numerator and df_2 on the denominator. From the table, we directly retrieve that $F_{4,10;0.975} = 4.47$ and $F_{4,10;0.025} = 0.11$.
d) Let $X \sim N(\mu = 10, \sigma^2 = 9)$, we need to find $x_{0.025}$ and $x_{0.975}$, its quantiles of level 0.025 and 0.975 respectively. Using standardization, we have $Z = \frac{X-10}{3} \sim N(0,1)$. Besides, $X = 3Z + 10$. Hence,

$$\begin{aligned} x_{0.025} &= 3z_{0.025} + 10 \\ &= 3 \times (-1.96) + 10 \\ &= 4.12 \end{aligned}$$

where -1.96 is obtained using the symmetry of the normal. Besides,

$$\begin{aligned} x_{0.975} &= 3z_{0.975} + 10 \\ &= 3 \times 1.96 + 10 \\ &= 15.88 \end{aligned}$$

Exercise 8

a) We perform a change of variable to calculate the expectation of X:

$$\begin{aligned} \mathbb{E}\,(X) &= \int_0^\infty \frac{x}{\beta} e^{-\frac{x}{\beta}}\,dx \\ &= \beta \int_0^\infty y e^{-y} dy \quad (y = \frac{x}{\beta}) \\ &= \beta\,\Gamma(2) \end{aligned}$$

where $\Gamma(.)$ is the gamma function. Since $\Gamma(n) = (n-1)!$ for $n \in \mathbb{N}$ then $\Gamma(2) = 1$. For the variance, we need

$$\begin{aligned}
\mathbb{E}\left(X^2\right) &= \int_0^\infty \frac{x^2}{\beta} e^{-\frac{x}{\beta}} dx \\
&= \beta^2 \int_0^\infty y^2 e^{-y} dy \quad (y = \frac{x}{\beta}) \\
&= \beta^2\,\Gamma(3) \\
&= \beta^2\,2!
\end{aligned}$$

Therefore $\mathbb{V}(X) = \mathbb{E}\left(X^2\right) - \mathbb{E}(X)^2 = 2\beta^2 - \beta^2 = \beta^2$. You can also obtain these results without relying on the incomplete gamma function and solving the integrals with the technique of integration by parts.

b) Recall that $t \in]-\infty, \frac{1}{\beta}[$. The mgf is:

$$\begin{aligned}
m_X(t) := \mathbb{E}\left(e^{tX}\right) &= \int_0^\infty e^{tx} \frac{1}{\beta} e^{-\frac{x}{\beta}} dx \\
&= \frac{1}{\beta} \int_0^\infty e^{-\left(\frac{1}{\beta}-t\right)x} dx \\
&= \frac{1}{\beta} \left[-\left(\frac{1}{\beta} - t\right)^{-1} e^{-\left(\frac{1}{\beta}-t\right)x} \right]_{x=0}^{x=\infty} \\
&= \frac{1}{\beta} \left(\frac{1-\beta t}{\beta}\right)^{-1} = (1-\beta t)^{-1}
\end{aligned}$$

c) The support of $Y = -5X$ is $(-\infty, 0]$ and its pdf is

$$\begin{aligned}
f_Y(y) &= \frac{1}{5} f_X\left(-\frac{y}{5}\right) \\
&= \frac{1}{5\beta} e^{\frac{y}{5\beta}}
\end{aligned}$$

This is like an exponential random variable with a parameter $\beta' = 5\beta$ but defined on $\mathbb{R}^-$. By symmetry, we immediately infer that $\mathbb{E}(Y) = -5\beta$ and the variance is $\mathbb{V}(Y) = 25\beta^2$.

d) The X_i for $i = 1, ..., n$ are independent therefore

$$\begin{aligned}
\mathbb{E}\left(e^{t\,\sum_{i=1}^n X_i}\right) &= \prod_{i=1}^n \mathbb{E}\left(e^{t\,X_i}\right) \\
&= \left(m_X(t)\right)^n \\
&= (1-\beta t)^{-n}
\end{aligned}$$

for $t < \frac{1}{\beta}$. We recognize the mgf of a Gamma random variable $Gamma(n, \beta)$. Z is therefore a gamma random variable. and

$$\mathbb{E}(Z) = n\beta \quad \mathbb{V}(X) = n\beta^2$$

Exercise 9

a) Since X and Y are independent, the mgf of Z is the product of mgf:

$$\begin{aligned}\mathbb{E}\left(e^{tZ}\right) &= \mathbb{E}\left(e^{tX}\right)\mathbb{E}\left(e^{tY}\right)\\ &= \exp\left(\lambda_1\left(e^t-1\right)\right)\exp\left(\lambda_2\left(e^t-1\right)\right)\\ &= \exp\left(\left(\lambda_1+\lambda_2\right)\left(e^t-1\right)\right)\end{aligned}$$

and we recognize the mgf of a Poisson random variable with parameter $\lambda_z = \lambda_1 + \lambda_2$. Z is then a Poisson r.v.

Exercise 10

a) Since X and Y are independent, the mgf of Z is the product of mgf:

$$\begin{aligned}\mathbb{E}\left(e^{tZ}\right) &= \mathbb{E}\left(e^{tX}\right)\mathbb{E}\left(e^{tY}\right)\\ &= \left(1-\beta t\right)^{-\alpha}\left(1-\beta t\right)^{-\delta}\\ &= \left(1-\beta t\right)^{-(\alpha+\delta)}\end{aligned}$$

and we recognize the mgf of a Gamma random variable with parameters $\alpha + \delta$ and β. Z is then a Gamma r.v.

Exercise 11

a) Firstly, we are going to determine the probability density function of X.

In theory:

$$f_X(x) = \frac{1}{2\pi\sqrt{1-a^2}}\int_{-\infty}^{+\infty} e^{-\frac{x^2-2axy+y^2}{2(1-a^2)}}\,\mathrm{dy}$$

Let us notice that: $x^2 - 2axy + y^2 = (y-ax)^2 + x^2(1-a^2)$

Hence:

$$f_X(x) = \frac{1}{2\pi\sqrt{1-a^2}}e^{-\frac{x^2}{2}}\int_{-\infty}^{+\infty} e^{-\frac{(y-ax)^2}{2(1-a^2)}}\,\mathrm{dy}$$

We make the following change of variable $u = \frac{y-ax}{\sqrt{2(1-a^2)}}$ to reach:

$$f_X(x) = \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}\int_{-\infty}^{+\infty} e^{-u^2}\,\mathrm{du}$$

Starting from the Gauss integral $\int_{-\infty}^{+\infty} e^{-u^2}\mathrm{du} = \sqrt{\pi}$, we find the probability density function of X :

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

We recognize directly a normal distribution with a zero mean and a variance equal to 1, therefore $X \sim N(0,1)$.
We notice that the law of probability of X does not depend on the parameter a !

b) As like in the question a):

$$f_Y(y) = \frac{1}{2\pi\sqrt{1-a^2}} \int_{-\infty}^{+\infty} e^{-\frac{x^2-2axy+y^2}{2(1-a^2)}} \mathrm{dx}$$

Starting from $x^2 - 2axy + y^2 = (x-ay)^2 + y^2(1-a^2)$, by making the following change of variable $u = \frac{x-ay}{\sqrt{2(1-a^2)}}$, and by using again Gauss integral, we reach:

$$f_Y(y) = \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}}$$

We notice again that $Y \sim N(0,1)$ and does not depend on the parameter a.

c) In the case of $a = 0$, the joint density function is

$$f_{(X,Y)}(x,y) = \frac{1}{2\pi} e^{-\frac{x^2+y^2}{2}}$$

Then: $f_{(X,Y)}(x,y) = f_X(x) f_Y(y)$

Which leads directly to deduce that X and Y are independent.

Therefore: $\sigma_{X,Y} = 0$ and $\rho_{X,Y} = 0$.

Exercise 12

The aim of the first four questions is to exercise the symmetry of the normal distribution and the use of the table.

1.

$$P(X \geq 2.5) = P\left(\frac{X-3}{\sqrt{1.25}} \geq \frac{2.5-3}{\sqrt{1.25}}\right) = P(Z \geq -0.45) = 1 - P(Z \geq 0.45) = 1 - 0.3264.$$

2.

$$P(X \leq 1) = P\left(\frac{X-3}{\sqrt{1.25}} \leq \frac{1-3}{\sqrt{1.25}}\right) = P(Z \leq -1.79) = P(Z \geq 1.79) = 0.0367.$$

3.

$$P(X \leq 4) = P\left(\frac{X-3}{\sqrt{1.25}} \leq \frac{4-3}{\sqrt{1.25}}\right) = P(Z \leq 0.89) = 1 - P(Z \geq 0.89) = 1 - 0.1867.$$

4. $$P(X \geq 3.5) = P\left(\frac{X-3}{\sqrt{1.25}} \geq \frac{3.5-3}{\sqrt{1.25}}\right) = P(Z \geq 0.45) = 0.3264.$$

5. Use commands

   ```
   import numpy as np
   import scipy.stats as sp
   1-sp.norm.cdf(2.5, loc=3, scale=np.sqrt(1.25))
   sp.norm.cdf(1, loc=3, scale=np.sqrt(1.25))
   sp.norm.cdf(4, loc=3, scale=np.sqrt(1.25))
   1-sp.norm.cdf(3.5, loc=3, scale=np.sqrt(1.25))
   ```

6. We have

   $$P(X \geq a) = P\left(\frac{X-3}{\sqrt{1.25}} \geq \frac{a-3}{\sqrt{1.25}}\right) = P\left(Z \geq \frac{a-3}{\sqrt{1.25}}\right) = 0.674.$$

   This suggests that if we find $z_{0.674}$ which is such that $P(Z \geq z_{0.674}) = 0.674$ we can find a. However we can not find directly $z_{0.674}$ from the table, we need to use the symmetry of the Normal distribution according to which $z_{0.674} = -z_{1-0.674} = -z_{0.326}$ with $P(Z \geq z_{0.326}) = 0.326$. Form the table we find $z_{0.326} = 0.45$. Hence $\frac{a-3}{\sqrt{1.25}} = z_{0.674} = -z_{0.326} = -0.45$ and we get $a = 2.496$.

   Next we have

   $$\P(X \leq b) = \P\left(\frac{X-3}{\sqrt{1.25}} \leq \frac{b-3}{\sqrt{1.25}}\right) = \P\left(Z \leq \frac{b-3}{\sqrt{1.25}}\right) = 0.037.$$

   This suggests that if we find $z_{0.037}$ which is such that $P(Z \leq z_{0.037}) = 0.037$ we can find b. However we can not find directly $z_{0.037}, P(Z \leq z_{0.037}) = 0.037$ from the table, we need to use the symmetry of the Normal distribution according to which $z_{0.037}, P(Z \leq z_{0.037}) = 0.037$ equals $-z_{0.037}, P(Z \geq z_{0.037}) = 0.037$. Form the table we find $z_{0.037} = 1.79, P(Z \geq z_{0.037}) = 0.037$. Hence $\frac{b-3}{\sqrt{1.25}} = -1.79$ and we get $b = 1$.

   Next we have

   $$P(X \leq c) = P\left(\frac{X-3}{\sqrt{1.25}} \leq \frac{c-3}{\sqrt{1.25}}\right) = P\left(Z \leq \frac{c-3}{\sqrt{1.25}}\right) = 0.81.$$

   This suggests that if we find $z_{1-0.81}$ which is such that $P(Z \geq z_{0.19}) = 0.19$ we can find c. Form the table we find $z_{0.19} = 0.87$. Hence $\frac{c-3}{\sqrt{1.25}} = 0.87$ and we get $c = 3.97$.

   Next we have

   $$P(X \geq d) = P\left(\frac{X-3}{\sqrt{1.25}} \geq \frac{d-3}{\sqrt{1.25}}\right) = P\left(Z \geq \frac{d-3}{\sqrt{1.25}}\right) = 0.3264.$$

This suggests that if we find $z_{0.3264}$ which is such that $P(Z \geq z_{0.3264}) = 0.3264$ we can find d. Form the table we find $z_{0.3264} = 0.45$. Hence $\frac{d-3}{\sqrt{1.25}} = 0.45$ and we get $d = 3.50$.

We have obtained the same results as in the first four questions.

7. By 'standardization' of a random variable we mean that the standardized variable has mean 0 and variance 1. We get this by subtracting the mean and dividing by the standard deviation of the variable. We standardize a normal or a t-distributed random variable in order to bring it to a unified scale, a unit-free scale (imagine one variable measured in cm and another in inches then the standardization of each of them is unit free). After standardization we can use the standard normal/t distribution and the tables of it.

8. We need the expectation and the variance in order to standardize $aX+b \sim N(a\mu + b, a^2\sigma^2)$ hence standardization

$$\frac{aX + b - (a\mu + b)}{\sqrt{a^2\sigma^2}} = (X - \mu)/\sigma$$

$$\frac{\hat{\beta}_i - \beta_i}{\sigma\sqrt{\{(A^T A)^{-1}\}_{ii}}}$$

9. We have $X - Y \sim N(0, 2 \times 1.25) = N(0, 2.5)$.

$$\P(X - Y \leq 2.5) = \P\left(\frac{X-Y}{\sqrt{2.5}} \leq \frac{2.5}{\sqrt{2.5}}\right) = \P(Z \leq 1.58) = 1 - \P(Z \geq 1.58) = 1 - 0.0571.$$

10. A weighted sum of normal variables is normal again with mean

$$\mathbb{E}\Big(\sum_{i=1}^n a_i X_i\Big) = \sum_{i=1}^n \mathbb{E}(a_i X_i) = \sum_{i=1}^n a_i \mathbb{E}(X_i) = \mu \sum_{i=1}^n a_i$$

and variance

$$Var\Big(\sum_{i=1}^n a_i X_i\Big) = \sum_{i=1}^n Var(a_i X_i) = \sum_{i=1}^n a_i^2 Var(X_i) = \sigma^2 \sum_{i=1}^n a_i^2$$

The standardization of $\sum_{i=1}^n a_i X_i$ is

$$\frac{\sum_{i=1}^n a_i(X_i - \mu)}{\sigma\sqrt{\sum_{i=1}^n a_i^2}}.$$

$$P\Big(\sum_{i=1}^5 a_i X_i \leq 25\Big) = P\left(\frac{\sum_{i=1}^5 a_i(X_i - 3)}{\sqrt{1.25\sum_{i=1}^5 a_i^2}} \leq \frac{25 - 3\sum_{i=1}^5 a_i}{\sqrt{1.25\sum_{i=1}^5 a_i^2}}\right)$$
$$= P(Z \leq -2.41) = \P(Z \geq 2.41) = 0.008.$$

11. We can use CLT according to which $(\bar{X}-3)/\sqrt{1.25/50} \approx N(0,1)$ to approximate this probability.

$$
\begin{aligned}
P\Big(\sum_{i=1}^{n} X_i \leq 150\Big) &= P\Big(\frac{1}{50}\sum_{i=1}^{50} X_i \leq 150/50\Big)\\
&\approx P\left(\frac{(\bar{X}-3)}{\sqrt{1.25/50}} \leq \frac{3-3}{\sqrt{1.25/50}}\right)\\
&= P\left(Z \leq 0\right) = 0.5
\end{aligned}
$$

2 Parameters estimation

Exercise 1

a) Positive integer random variable, counting events. Then we use a Poisson random variable, N. The expectation is $\mathbb{E}N = \lambda$. Using the method of moments, we would estimate $\widehat{\lambda} = \bar{x} = 2340$. For a Poisson r.v. $\mathbb{V}(N) = \lambda$, if $\widehat{\lambda} = 2340$, the standard deviation is $\sqrt{2340} = 48.37$ that is close to the observed one, 48.91. This validates our choice.

b) If the standard deviation is 21.01<48.37, the Poisson law is not adapted. An alternative consists to work with a Binomial r.v. or with a Normal approximation (in order to fit the variance).

Exercise 2

a)

$$
\begin{aligned}
\mathbb{E}\left(X\right) &= \frac{1}{3\theta}\int_0^{\theta} x\;dx + \frac{2}{3\left(21-\theta\right)}\int_{\theta}^{21} x\;dx\\
&= \frac{1}{3\theta}\left(\frac{\theta^2}{2}\right) + \frac{2}{3\left(21-\theta\right)}\left(\frac{21^2}{2}-\frac{\theta^2}{2}\right)\\
&= \frac{\theta}{6} + \frac{1}{3}\left(21+\theta\right)\\
&= \frac{\theta}{2} + 7
\end{aligned}
$$

Variance (not asked in the question)

$$
\begin{aligned}
\mathbb{E}\left(X^2\right) &= \frac{1}{3\theta}\int_0^{\theta} x^2\;dx + \frac{2}{3\left(21-\theta\right)}\int_{\theta}^{21} x^2\;dx\\
&= \frac{\theta^2}{9} + \frac{1}{9}\left(2\theta^2 + 42\theta + 882\right)\\
&= \frac{\theta^2}{3} + \frac{14}{3}\theta + 98
\end{aligned}
$$

Then

$$\begin{aligned}\mathbb{V}(X) &= \frac{\theta^2}{3}+\frac{14}{3}\theta+98-\left(\frac{\theta}{2}+7\right)^2\\ &= \frac{\theta^2}{3}+\frac{14}{3}\theta+98-\frac{\theta^2}{4}-7\theta-49\\ &= \frac{\theta^2}{12}-\frac{7}{3}\theta+49\end{aligned}$$

Using the moments method we have that

$$\hat{\theta} = 2\left(\bar{X}-7\right)$$

b) Method of moments:

$$\begin{aligned}\mathbb{E}\left(\hat{\theta}\right) &= 2\left(\mathbb{E}\left(\bar{X}\right)-7\right)\\ &= 2\left(\frac{n}{n}\mathbb{E}(X)-7\right)\\ &= 2\left(\frac{\theta}{2}+7-7\right)\\ &= \theta\end{aligned}$$

therefore, the bias is null.

c) Let $n_1(\theta)$ and $n_2(\theta)$ be respectively the number of observations in $[0,\theta]$ and $(\theta,21]$. The likelihood is then:

$$L(\theta) = \left(\frac{1}{3\theta}\right)^{n_1(\theta)}\left(\frac{2}{3(21-\theta)}\right)^{n_2(\theta)},$$

whereas the log-likelihood is

$$\ln L(\theta) = -n_1(\theta)\ln(3\theta)-n_2(\theta)\ln\left(\frac{3}{2}(21-\theta)\right) \tag{1}$$

The functions $n_1(\theta)$ and $n_2(\theta)$ are not smooth and therefore we cannot derive the log-likelihood to obtain a closed form expression of $\hat{\theta}$. Nevertheless, we can still find numerically the $\hat{\theta}$ maximizing equation (1).

Exercise 3

a)

$$\begin{aligned}\mathbb{E}(X) &= \int_0^1(\alpha+1)x^{\alpha+1}dx=\frac{\alpha+1}{\alpha+2}\\ \mathbb{E}\left(X^2\right) &= \int_0^1(\alpha+1)x^{\alpha+2}dx=\frac{\alpha+1}{\alpha+3}\end{aligned}$$

then $\mathbb{V}\left(X\right)=\frac{\alpha+1}{\alpha+3}-\left(\frac{\alpha+1}{\alpha+2}\right)^2$. Using the moments method we have that

$$\frac{\hat{\alpha}+1}{\hat{\alpha}+2} = \bar{X}$$

or

$$\begin{aligned}\hat{\alpha}-\hat{\alpha}\bar{X} &= 2\bar{X}-1\\ \hat{\alpha} &= \frac{2\bar{X}-1}{1-\bar{X}}\end{aligned}$$

Here, we cannot estimate easily the bias because $\mathbb{E}\left(\frac{2\bar{X}-1}{1-\bar{X}}\right)$ =???

c) The likelihood is equal to

$$L(\alpha) = (\alpha+1)^n\prod_{i=1}^{n}x_i^{\alpha}$$

and

$$\ln L(\alpha) = n\ln\left(\alpha+1\right)+\sum_{i=1}^{n}\alpha\ln x_i\,.$$

Deriving this last expression with respect to θ gives us:

$$\frac{\partial L(\alpha)}{\partial\alpha} = \frac{n}{1+\alpha}+\sum_{i=1}^{n}\ln x_i\,.$$

We infer from this last equation that the LL estimator is

$$\hat{\alpha}=-\left(1+\frac{n}{\sum_{i=1}^{n}\ln X_i}\right)$$

c) First,

$$\begin{aligned}P(Y\leq y) &= P(X\geq e^{-y})\\ &= \int_{e^{-y}}^{1}\left(\alpha+1\right)x^{\alpha}dx\end{aligned}$$

The pdf of Y is therefore

$$f_Y(y) = \frac{\partial P(Y\leq y)}{\partial y}=\left(\alpha+1\right)e^{-(\alpha+1)y}\quad y\geq 0$$

and we recognize an exponential random variable of parameter $\beta=\frac{1}{\alpha+1}$. Therefore $-\sum_{i=1}^{n}\ln X_i$ is a Gamma random variable $U\sim Gamma(n,\beta=\frac{1}{\alpha+1})$. The expectation of $\hat{\alpha}$ is therefore:

$$\mathbb{E}\left(\hat{\alpha}\right) = -1+n\,\mathbb{E}\left(\frac{1}{U}\right)$$

Using the definition of a gamma pdf and $\Gamma(n) = (n-1)\Gamma(n-1)$ we get that

$$\begin{aligned}
\mathbb{E}\left(\frac{1}{U}\right) &= \int_0^\infty \frac{1}{u}\left(\frac{1}{\Gamma(n)}\frac{u^{n-1}}{\beta^n}e^{-\frac{u}{\beta}}\right)du \\
&= \frac{1}{(n-1)\,\beta}\int_0^\infty \underbrace{\frac{1}{\Gamma(n-1)}\frac{u^{n-2}}{\beta^{n-1}}e^{-\frac{u}{\beta}}}_{\text{pdf of a Gamma(n-1,}\beta\text{)}}\,du \\
&= \frac{1}{(n-1)\,\beta} = \frac{\alpha+1}{n-1}\,.
\end{aligned}$$

Therefore

$$\mathbb{E}\left(\hat{\alpha}\right) = -1 + n\,\frac{\alpha+1}{n-1} = \frac{1+\alpha n}{n-1}$$

and the bias is

$$\begin{aligned}
bias &= \mathbb{E}\left(\hat{\alpha}\right) - \alpha \\
&= \frac{1+\alpha n}{n-1} - \alpha
\end{aligned}$$

Exercise 4

a-b) The likelihood of the normal sample is the following

$$L = \left(\sigma\sqrt{2\pi}\right)^{-n}\exp\left(-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i-\mu}{\sigma}\right)^2\right)$$

and its logarithm is

$$\ln L = -n\ln\left(\sigma\right) - n\ln\left(\sqrt{2\pi}\right) - \frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i-\mu}{\sigma}\right)^2$$

The estimators of μ and σ are solution of the following system of equations:

$$\begin{aligned}
\frac{\partial\ln L}{\partial\mu} &= \sum_{i=1}^{n}\frac{x_i-\mu}{\sigma^2} = 0 \\
\frac{\partial\ln L}{\partial\sigma} &= -\,n\frac{1}{\sigma} + \sum_{i=1}^{n}\frac{\left(x_i-\mu\right)^2}{\sigma^3} = 0
\end{aligned}$$

From the first equation we infer that $\hat{\mu} = \bar{X}$. Given that $\mathbb{E}\left(\bar{X}\right) = \mu$, this estimator is unbiased. The second equation leads to

$$\hat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2\,.$$

Is this estimator biased (i.e. $\mathbb{E}\left(\hat{\sigma}^2\right) \neq \sigma^2$) ? Yes and this is a major drawback of this estimator! You can refer to the properties of $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$ seen during the lecture to infer that

$$\begin{aligned}\mathbb{E}\left(\hat{\sigma}^2\right) &= \mathbb{E}\left(\frac{(n-1)}{n}S^2\right)\\ &= \frac{(n-1)}{n}\sigma^2\end{aligned}$$

Therefore, the bias $b = \mathbb{E}\left(\hat{\sigma}^2\right) - \sigma^2$ is not null! This explain why we use S^2 instead of $\hat{\sigma}$ for estimating the variance.

c) As a sum of independent normal random variables is also normal, we infer that $\hat{\mu} = \bar{X}$ is $\mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$ (see also lecture 3). The MSE of $\hat{\mu}$ is the sum of the bias (null for this estimator) and of its variance. Therefore, the MSE of $\hat{\mu}$ is $\frac{\sigma^2}{n}$: i.e. the higher is n, the better is the accuracy of the estimation (since the MSE is inversely proportional to n).

Exercise 5

a) The likelihood of a sample of discrete random variables is the product of pmf. We denote by n_I, n_S and n_D the number of observations in each of the three categories. The likelihood is then:

$$\begin{aligned}L &= p_I^{n_I} p_S^{n_S} p_D^{n_D}\\ \ln L &= n_I \ln p_I + n_S \ln p_S + n_D \ln p_D\end{aligned}$$

but $p_D = 1 - p_I - p_S$:

$$\ln L = n_I \ln p_I + n_S \ln p_S + n_D \ln\left(1 - p_I - p_S\right)$$

We cancel the first order derivative of $\ln L$ to find estimators:

$$\begin{aligned}\frac{\partial L}{\partial p_I} &= \frac{n_I}{p_I} - \frac{n_D}{1 - p_I - p_S} = 0\\ \frac{\partial L}{\partial p_S} &= \frac{n_S}{p_S} - \frac{n_D}{1 - p_I - p_S} = 0\end{aligned}$$

Solving this system (and remember that $n = n_I + n_S + n_D$) leads to the quite intuitive estimators:

$$\hat{p}_I = \frac{N_I}{n} \quad \hat{p}_S = \frac{N_S}{n} \quad \hat{p}_D = 1 - \hat{p}_I - \hat{p}_S\,.$$

where N_I and N_S are the random variables counting the number of observations respectively in category I and S.

b) The bias is

$$bias = \mathbb{E}\left(\frac{N_I}{n}\right) - p_I\,.$$

By definition, the N_I is a binomial random variable with n trials and a probability of success equal to p_I. Therefore $\mathbb{E}\left(N_I\right) = \frac{np_I}{n}$ and the bias is null.

Exercise 6

a) This random variable is distributed as a $Gamma(2, \theta)$ and $\mathbb{E}(X) = 2\theta$. If you want to prove it, let us recall that $\Gamma(\alpha) = \int_0^\infty y^{\alpha-1}e^{-y}dy$ and $\Gamma(n) = (n-1)!$ if $n \in \mathbb{N}$. After a change of variable, the expectation is

$$\begin{aligned}\mathbb{E}(X) &= \frac{1}{\Gamma(2)}\int_0^\infty \frac{x^2}{\theta^2}e^{-x/\theta}dx \quad (y = x/\theta)\\ &= \theta\frac{1}{\Gamma(2)}\underbrace{\int_0^\infty y^{3-1}e^{-y}dy}_{\Gamma(3)} = 2\theta\end{aligned}$$

By the method of moments, an estimator of θ is $\hat{\theta} = \frac{\bar{X}}{2}$.

b) The likelihood and log-likelihood are respectively:

$$L(\theta) = \left(\frac{1}{\Gamma(2)}\right)^n \prod_{i=1}^n \frac{x_i}{\theta^2}e^{-x_i/\theta}$$

$$\ln L(\theta) = -n\ln\Gamma(2) + \sum_{i=1}^n \ln(x_i) - 2n\ln(\theta) - \frac{1}{\theta}\sum_{i=1}^n x_i$$

The estimator solves then:

$$\frac{\partial \ln L(\theta)}{\partial\theta} = -\frac{2n}{\theta} + \frac{1}{\theta^2}\sum_{i=1}^n x_i = 0$$

whose the solution is $\hat{\theta} = \frac{\sum_{i=1}^n X_i}{2n} = \frac{\bar{X}}{2}$. Moments and log-likelihood estimators are the same.

c) The sum of a $Gamma(\alpha, \beta)$ and $Gamma(\delta, \beta)$ random variables is $Gamma(\alpha + \delta, \beta)$ distributed (see exercise 10, first session). As $X_i \sim Gamma(2, \theta)$, then $\sum_{i=1}^n X_i \sim Gamma(2n, \theta)$, and

$$\begin{aligned}\mathbb{E}\left(\sum_{i=1}^n X_i\right) &= 2n\theta\\ \mathbb{V}\left(\sum_{i=1}^n X_i\right) &= 2n\theta^2\end{aligned}$$

and given that $\mathbb{E}(aY) = a\mathbb{E}(Y)$ and $\mathbb{V}(aY) = a^2\mathbb{V}(Y)$

$$\begin{aligned}\mathbb{E}\left(\hat{\theta}\right) = \mathbb{E}\left(\frac{\sum_{i=1}^n X_i}{2n}\right) &= \theta\\ \mathbb{V}\left(\hat{\theta}\right) = \mathbb{V}\left(\frac{\sum_{i=1}^n X_i}{2n}\right) &= \frac{2n}{4n^2}\theta^2 = \frac{\theta^2}{2n}\end{aligned}$$

d) For the normal approximation, we approximate $\hat{\theta}$ by a $\mathcal{N}\left(\theta, \frac{\theta^2}{2n}\right)$. Then $Z = \frac{\hat{\theta}-\theta}{\sqrt{\frac{\theta^2}{2n}}}$ is $\mathcal{N}(0,1)$ and $Z_{97.5\%} = 1.96$. The interval for $\hat{\theta}$ is then such that

$$P\left(-1.96 \leq \frac{\hat{\theta}-\theta}{\sqrt{\frac{\theta^2}{2n}}} \leq 1.96\right) = 95\% \tag{2}$$

or $\hat{\theta} \in \left[\theta - 1.96\sqrt{\frac{\theta^2}{2n}}\,;\, \theta + 1.96\sqrt{\frac{\theta^2}{2n}}\right]$ with a 95% probability. This interval cannot be computed in practice since θ is unknown.

e) In practice, θ is unknown and we rather seek for a confidence interval for θ rather than for $\hat{\theta}$. Recall that

$$\begin{aligned} 0.95 &= P\left(-1.96 \leq \frac{\hat{\theta}-\theta}{\sqrt{\frac{\theta^2}{2n}}} \leq 1.96\right) \\ &= P\left(\frac{\hat{\theta}}{1+1.96/\sqrt{2n}} \leq \theta \leq \frac{\hat{\theta}}{1-1.96/\sqrt{2n}}\right) \end{aligned}$$

Hence, a 0.95 confidence interval for θ is given by

$$IC_{0.95}[\theta] = \left[\frac{\hat{\theta}}{1+1.96/\sqrt{2n}}; \frac{\hat{\theta}}{1-1.96/\sqrt{2n}}\right]$$

Exercise 7

a) See exercise 9, first session.

b) We have d observations of N, noted $n_i = n_i^1 + n_i^2$ for $i = 1, ..., d$. The likelihood is the product of pmf and the log-likelihood the sum of log-pmf:

$$\begin{aligned} L(\lambda) &= \prod_{i=1}^{d} \frac{\lambda^{n_i}}{n_i!} e^{-\lambda}\,, \\ \ln L(\lambda) &= \sum_{i=1}^{d} n_i \ln \lambda - d\,\lambda - \sum_{i=1}^{d} \ln n_i!\,. \end{aligned}$$

The estimator is solution of the following equation:

$$\frac{\partial \ln L(\lambda)}{\partial \theta \lambda} = \frac{1}{\lambda}\sum_{i=1}^{d} n_i - d = 0\,,$$

which suggests the following estimator: $\hat{\lambda} = \frac{1}{d}\sum_{i=1}^{d} N_i = \bar{N}$ where N_i are iid Poison random variables with parameter λ.

c) A sum of Poisson r.v. is also Poisson distributed. Therefore, $\sum_{i=1}^{d} N_i \sim Po(d\,\lambda)$ (discrete r.v.). Furthermore,

$$\begin{aligned} P\left(\hat{\lambda} \leq \lambda_0\right) &= P\left(\frac{1}{d}\sum_{i=1}^{d} N_i \leq \lambda_0\right) \\ &= P\left(\sum_{i=1}^{d} N_i \leq \lambda_0 d\right) \\ &= \sum_{k=0}^{\lfloor \lambda_0 d \rfloor} P\left(\sum_{i=1}^{d} N_i = k\right) \\ &= \sum_{k=0}^{\lfloor \lambda_0 d \rfloor} \frac{(d\,\lambda)^k}{k!} e^{-d\,\lambda}\,, \end{aligned}$$

where $\lfloor \lambda_0 d \rfloor$ is the nearest lower integer than $\lambda_0 d$ (because $\sum_{i=1}^{d} N_i$ takes only integer values).

d) For the normal approximation, we need the expectation and variance of $\hat{\lambda}$. Since $N_i \sim Po(\lambda)$, $\mathbb{E}(N_i) = \lambda$ and $\mathbb{V}(N_i) = \lambda$. Therefore

$$\begin{aligned} \mathbb{E}\left(\hat{\lambda}\right) &= \frac{1}{d}\sum_{i=1}^{d} \mathbb{E}\left(N_i\right) = \frac{d}{d}\lambda = \lambda\,, \\ \mathbb{V}\left(\hat{\lambda}\right) &= \frac{1}{d^2}\sum_{i=1}^{d} \mathbb{V}\left(N_i\right) = \frac{d}{d^2}\lambda = \frac{\lambda}{d}\,, \end{aligned}$$

and we approximate $\hat{\lambda}$ by a $\mathcal{N}\left(\lambda, \frac{\lambda}{d}\right)$. Then $Z = \frac{\hat{\lambda}-\lambda}{\sqrt{\frac{\lambda}{d}}}$ is $\mathcal{N}(0,1)$ and $Z_{97.5\%} = 1.96$. The confidence interval of $\hat{\lambda}$ is then such that

$$P\left(-1.96 \leq \frac{\hat{\lambda}-\lambda}{\sqrt{\frac{\lambda}{d}}} \leq 1.96\right) = 95\% \tag{3}$$

or $\hat{\lambda} \in \left[\lambda - 1.96\sqrt{\frac{\lambda}{d}}\,;\,\lambda + 1.96\sqrt{\frac{\lambda}{d}}\right]$ with a 95% probability.

e) In practice, λ is unknown and we rather seek for a confidence interval for λ rather than for $\hat{\lambda}$. Contrary to Exercise 5, a direct expression for this confidence interval does not exist (since we cannot isolate λ in the probability displayed in (3).

First way out: simply replace λ in the denominator in (3). We obtain

$$\frac{\hat{\lambda}-\lambda}{\sqrt{\frac{\hat{\lambda}}{d}}} \approx N(0,1)$$

Notice that the asymptotic normality is retained even though we replaced λ by $\hat{\lambda}$. Formally, this is a consequence of Slutsky's theorem but this is beyond the scope of this course. Hence,

$$\begin{aligned} 0.95 &= P\left(-1.96 \leq \frac{\hat{\lambda}-\lambda}{\sqrt{\frac{\hat{\lambda}}{d}}} \leq 1.96\right) \\ &= P\left(\hat{\lambda} - 1.96\sqrt{\frac{\hat{\lambda}}{d}} \leq \lambda \leq \hat{\lambda} + 1.96\sqrt{\frac{\hat{\lambda}}{d}}\right) \end{aligned}$$

A 95% confidence interval for λ is then given by

$$IC_{0.95}[\lambda] = \left[\hat{\lambda} \pm 1.96\sqrt{\frac{\hat{\lambda}}{d}}\right]$$

Second way out: The denominator in (3) corresponds to the standard deviation of the estimator. We can relate it to the variance of the data, i.e. $\mathbb{V}(N_i)$, and approximate the latter by its empirical estimator (the sample variance) $S^2 = \frac{1}{d-1}\sum_{k=1}^{d-1}(n_i - \bar{n})^2$ ($\bar{n}$ is here the average of n_i). Indeed, we know that

$$\mathbb{V}(N_i) = \lambda$$

can be well estimated by S^2. So we simply replace $\sqrt{\lambda/d}$ by $\sqrt{S^2/d}$. We obtain

$$\frac{\hat{\lambda}-\lambda}{\sqrt{\frac{S^2}{d}}} \approx N(0,1)$$

Notice that the asymptotic normality is retained even though we replaced the true variance by the empirical variance. Formally, this is a consequence of Slutsky's theorem but this is beyond the scope of this course. Hence, a 95% confidence interval for λ is then given by

$$IC_{0.95}[\lambda] = \left[\hat{\lambda} \pm 1.96\sqrt{\frac{S^2}{d}}\right]$$

Exercise 8

We define four random variables (N_1, N_2, N_3, N_4) according to the following table:

	Polluting	Not Polluting
Worn	N_1	N_2
Not Worn	N_3	N_4

The vector (N_1, N_2, N_3, N_4) is a multinomial distribution with parameters

$$\begin{aligned} p_1 &= p_A p_B \\ p_2 &= p_A(1-p_B) \\ p_3 &= (1-p_A)p_B \\ p_4 &= (1-p_A)(1-p_B) \end{aligned}$$

The likelihood function is therefore:

$$\begin{aligned} L(p_A, p_B) = & \frac{n!}{N_1!N_2!N_3!N_4!} \times (p_A p_B)^{N_1} \\ & \times [p_A(1-p_B)]^{N_2} \times [p_B(1-p_A)]^{N_3} \times [(1-p_A)(1-p_B)]^{N_4} \end{aligned}$$

The log-likelihood is then

$$\begin{aligned} \ln L(p_A, p_B) = \ln\left(\frac{n!}{N_1!N_2!N_3!N_4!}\right) & + N_1 \left(\ln\left(p_A\right) + \ln\left(p_B\right)\right) + N_2 \left(\ln\left(p_A\right) + \ln\left(1-p_B\right)\right) \\ & + N_3 \left(\ln\left(1-p_A\right) + \ln\left(p_B\right)\right) + N_4 \left(\ln\left(1-p_A\right) + \ln\left(1-p_B\right)\right) . \end{aligned}$$

The log-likelihood estimators are then solution of the system

$$\begin{aligned} \frac{N_1+N_2}{p_A} - \frac{N_3+N_4}{1-p_A} &= 0 \\ \frac{N_1+N_3}{p_B} - \frac{N_2+N_4}{1-p_B} - &= 0 \end{aligned}$$

Recalling that $n = N_1 + N_2 + N_3 + N_4$, we infer that

$$\hat{p}_A = \frac{N_1+N_2}{n} \qquad \hat{p}_B = \frac{N_1+N_3}{n} .$$

By construction, $N_1 + N_2$ is a $Bi(n, p_A)$ and therefore

$$\begin{aligned} \mathbb{E}\left(\hat{p}_A\right) &= \frac{\mathbb{E}\left(N_1+N_2\right)}{n} = p_A \\ \mathbb{V}\left(\hat{p}_A\right) &= \frac{\mathbb{V}\left(N_1+N_2\right)}{n^2} = \frac{p_A\left(1-p_A\right)}{n} \end{aligned}$$

In the same manner, we prove that

$$\mathbb{E}\left(\hat{p}_B\right) = p_B \quad , \quad \mathbb{V}\left(\hat{p}_B\right) = \frac{p_B\left(1-p_B\right)}{n}$$

Exercise 9

a) Let X be a uniform r.v. Its pdf is:

$$f_X(x) = \begin{cases} 1/\theta & x \in [0, \theta] \\ 0 & otherwise \end{cases} .$$

For a sample of n observations $(x_1, ..., x_n)$,

$$\begin{aligned} L(\theta) &= 1/\theta^n \quad and\, x_i \in [0,\theta]\, for\, i = 1, ..., n \\ &= 0 \quad if\, \exists\, x_i \notin [0,\theta] \end{aligned}$$

The log-likelihood estimator is such that $\hat{\theta} = \arg\max_{\theta\in\mathbb{R}^+} L(\theta)$. Clearly we must have $\theta \geq x_i$ for $1, ..., n$ otherwise the likelihood is null. For any $\theta \geq x_i$ for $1, ..., n$, the higher is θ, the lower is the product $1/\theta^n$. Therefore,

$$\hat{\theta} = \max(X_1, ..., X_n)$$

The expectation of X is $\mathbb{E}(X) = \frac{1}{\theta}\int_0^\theta x\, dx = \frac{\theta^2}{2\theta} = \frac{\theta}{2}$ and therefore $\tilde{\theta} = 2\bar{X}$.

b) We need the distribution of the maximum $Y = \max(X_1, ..., X_n)$.

$$\begin{aligned} P(Y \leq y) &= P(\max(X_1, ..., X_n) \leq y) \\ &= P(X_1 \leq y, X_2 \leq y, ..., X_n \leq y) \\ &= P(X_1 \leq y) \times ... \times P(X_n \leq y) \\ &= (y/\theta)^n \end{aligned}$$

for $y \leq \theta$ and $P(Y \leq y) = 1$ if $y \geq \theta$. Therefore$f_Y(y) =$

$$f_Y(y) = \begin{cases} n\frac{y^{n-1}}{\theta^n} & y \in [0,\theta] \\ 0 & otherwise \end{cases}.$$

The expectation and variance of the log-likelihood estimator are then

$$\mathbb{E}\left(\hat{\theta}\right) = \mathbb{E}(Y) = n\int_0^\theta \frac{y^n}{\theta^n}dy = \frac{n}{(n+1)}\theta$$
$$\mathbb{V}\left(\hat{\theta}\right) = \mathbb{V}(Y) = \frac{n}{(n+2)(n+1)^2}\theta^2.$$

We see that the log-likelihood estimator is biased! For the the moment estimator, we have

$$\begin{aligned} \mathbb{E}\left(\tilde{\theta}\right) = 2\mathbb{E}\left(\bar{X}\right) &= 2\frac{n}{n}\mathbb{E}(X) = \theta, \\ \mathbb{V}\left(\tilde{\theta}\right) = 2\mathbb{V}\left(\bar{X}\right) &= 2\frac{n}{n^2}\mathbb{V}(X) = \frac{\theta^2}{3n} \end{aligned}$$

an then the moment estimator is unbiased but its variance is higher than the one of the log-likelihood estimator.

c) Let us compute the MSE of both estimators.

$$MSE[\tilde{\theta}] = \mathbb{V}\left(\tilde{\theta}\right) = \frac{\theta^2}{3n}$$
$$MSE[\hat{\theta}] = \mathbb{V}\left(\hat{\theta}\right) + [\text{Biais}(\hat{\theta})]^2 = \ldots = \frac{2\theta^2}{(n+2)(n+1)}$$

The method of moments estimator is considered better if its MSE is lower than that of the log-likelihood estimator. This happens when

$$\begin{aligned} 6n &> n^2 + 3n + 2 \\ \Leftrightarrow n^2 &- 3n + 2 < 0 \end{aligned}$$

We can check that this inequality is never satisfied for $n \geq 1$. Besides, the gap between the two MSE's increases with sample size. Hence, the log-likelihood estimator is better. The reason is that even though it is biased, the log-likelihood estimator has a smaller variance and, hence, converges quicker to its expectation $\frac{n}{(n+1)}\theta$. In order to remove the bias, I would simply estimate θ by $\frac{n+1}{n}\hat{\theta}$ since

$$\mathbb{E}\left(\frac{n+1}{n}\hat{\theta}\right) = \frac{n+1}{n}\frac{n}{(n+1)}\theta\,.$$

However, you can check that this would increase the MSE.

Exercise 10

a) (This is a en exponential random variable). The likelihood and log-likelihood are equal to:

$$\begin{aligned} L(\lambda) &= \prod_{i=1}^{n} \frac{1}{\lambda}\exp(-\frac{x_i}{\lambda})\,, \\ \ln L(\lambda) &= -n\ln\lambda - \frac{\sum_{i=1}^{n} x_i}{\lambda}\,. \end{aligned}$$

Then

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

b) The estimator is unbiased since X is exponential and $\mathbb{E}\left(X\right) = \lambda$. The variance is:

$$\mathbb{V}\left(\hat{\lambda}\right) \quad = \quad \frac{\sum_{i=1}^{n}\mathbb{V}\left(X_i\right)}{n^2} = \frac{\mathbb{V}(X)}{n} = \frac{\lambda^2}{n}\,.$$

The mean squared error is therefore

$$MSE \quad = \quad \underbrace{0}_{bias} + \frac{\lambda^2}{n}\,.$$

c) We can approach $\hat{\lambda} \sim \mathcal{N}\left(\lambda, \frac{\lambda^2}{n}\right)$. Hence, we have

$$\begin{aligned} 0.95 &= P\left(-1.96 \leq \frac{\hat{\lambda}-\lambda}{\sqrt{\frac{\lambda^2}{n}}} \leq 1.96\right) \\ &= P\left(\frac{\hat{\lambda}}{1+\frac{1.96}{\sqrt{n}}} \leq \lambda \leq \frac{\hat{\lambda}}{1-\frac{1.96}{\sqrt{n}}}\right) \end{aligned}$$

Hence, a 95% confidence interval for λ is given by

$$IC_{0.95}[\lambda] = \left[\frac{\hat{\lambda}}{1+\frac{1.96}{\sqrt{n}}}; \frac{\hat{\lambda}}{1-\frac{1.96}{\sqrt{n}}}\right]$$

Exercise 11

a) It is well clear that $\mathbb{E}(\overline{Y}) = \mathbb{E}(Y)$, moreover we can notice that $(Y_k)_{1\leq k\leq n} \sim Bernoulli(P(X=0))$ by construction. Then:

$$\mathbb{E}(\overline{Y}) = P(X=0) = e^{-\lambda} = \theta$$

So the bias of the estimator $\overline{Y}$ is null.

It is also clear that $\mathbb{V}(\overline{Y}) = \frac{\mathbb{V}(Y)}{n}$, hence:

$$\mathbb{V}(\overline{Y}) = \frac{\theta(1-\theta)}{n}.$$

b) Let $j \in \mathbb{N}$. Starting from the definition of the conditional probability, we have that

$$\phi(j) = \frac{P[X_1 = 0, \sum_{k=1}^{n} X_k = j]}{P[\sum_{k=1}^{n} X_k = j]}$$

which is equivalent to

$$\phi(j) = \frac{P[X_1 = 0, \sum_{k=2}^{n} X_k = j]}{P[\sum_{k=1}^{n} X_k = j]}$$

Since $(X_k)_{1\leq k\leq n}$ are iid, we have first that

$$P[X_1 = 0, \sum_{k=2}^{n} X_k = j] = P[X_1 = 0]P\left[\sum_{k=2}^{n} X_k = j\right]$$

. Besides, using the result of the exercise 9 of TP1 we have

$$\sum_{k=2}^{n} X_k \sim Poi((n-1)\lambda)$$

$$\sum_{k=1}^{n} X_k \sim Poi(n\lambda)$$

Hence,

$$\phi(j) = \frac{e^{-\lambda} \times e^{-(n-1)\lambda} \times \frac{((n-1)\lambda)^j}{j!}}{e^{-n\lambda}\frac{(n\lambda)^j}{j!}}$$

So

$$\phi(j) = (\frac{n-1}{n})^j.$$

c) By definition $T = \left(\frac{n-1}{n}\right)^S$ and since $S \sim Poisson(n\lambda)$ (result of the exercise 9 of the first TP), we have

$$\mathbb{E}(T) = \sum_{k=0}^{+\infty} (\frac{n-1}{n})^k e^{-n\lambda} \frac{(n\lambda)^k}{k!} = e^{-\lambda} \sum_{k=0}^{+\infty} \frac{((n-1)\lambda)^k}{k!} e^{-(n-1)\lambda}$$

The last sum is equal to 1 since the term inside of the sum correspond to the probability mass function of a Poisson with parameter $(n-1)\lambda$. Therefore:

$$\mathbb{E}(T) = e^{-n\lambda} e^{(n-1)\lambda} = \theta$$

The estimator T has no bias.

Let us now move to the computation of the variance.

$$\mathbb{E}(T^2) = \sum_{k=0}^{+\infty} (\frac{n-1}{n})^{2k} e^{-n\lambda} \frac{(n\lambda)^k}{k!} = e^{-n\lambda} e^{\frac{(n-1)^2}{n}\lambda} \sum_{k=0}^{+\infty} \frac{\left(\frac{(n-1)^2}{n}\lambda\right)^k}{k!} e^{-\frac{(n-1)^2}{n}\lambda}$$

Therefore:

$$\mathbb{E}(T^2) = e^{\lambda(\frac{1}{n}-2)}$$

Since $\mathbb{V}(T) = \mathbb{E}(T^2) - \mathbb{E}(T)^2$ and after an algebraic calculation we reach

$$\mathbb{V}(T) = e^{-2\lambda}(e^{\frac{\lambda}{n}} - 1) = \theta^2(\theta^{-\frac{1}{n}} - 1).$$

d) T and $\overline{Y}$ are unbiased estimators, so $MSE[T] = \mathbb{V}(T)$ and $MSE[\overline{Y}] = \mathbb{V}(\overline{Y})$. Let us then compare between $MSE[T]$ and $MSE[\overline{Y}]$

$$MSE[T] - MSE[\overline{Y}] = \frac{\theta^2}{n} f(\lambda)$$

where f is the following function:

$$\begin{array}{rccl} f: & \mathbb{R} & \to & \mathbb{R} \\ & t & \mapsto & ne^{\frac{t}{n}} - e^t - n + 1 \end{array}$$

As $t \to -\infty$, the function reaches $-(n-1)$. The function then starts increasing until $t = 0$, where it reaches 0. After, the function decreases infinitely. This implies that: $\forall t \in \mathbb{R}\ f(t) \leq 0$ and $f(t) = 0 \Leftrightarrow t = 0$.

Consequently $MSE[T] < MSE[\overline{Y}]$, and therefore the estimator T is better than $\overline{Y}$.

Exercise 12

1. In both integrals we use integration by parts.

$$\begin{aligned} E(X) &= \int_\eta^\infty \frac{x}{\theta} e^{-(x-\eta)/\theta} dx = [-xe^{-(x-\eta)/\theta}]_\eta^\infty + \int_\eta^\infty e^{-(x-\eta)/\theta} dx \\ &= \eta + [-\theta e^{-(x-\eta)/\theta}]_\eta^\infty = \eta + \theta. \end{aligned}$$

$$\begin{aligned} E(X^2) &= \int_\eta^\infty \frac{x^2}{\theta} e^{-(x-\eta)/\theta} dx = [-x^2 e^{-(x-\eta)/\theta}]_\eta^\infty + \int_\eta^\infty 2xe^{-(x-\eta)/\theta} dx \\ &= \eta^2 + [-2x\theta e^{-(x-\eta)/\theta}]_\eta^\infty + \int_\eta^\infty 2\theta e^{-(x-\eta)/\theta} dx \\ &= \eta^2 + 2\theta\eta + [-2\theta^2 e^{-(x-\eta)/\theta}]_\eta^\infty = \eta^2 + 2\theta\eta + 2\theta^2 \\ Var(X) &= E(X^2) - (EX)^2 = \eta^2 + 2\theta\eta + 2\theta^2 - (\eta+\theta)^2 = \theta^2 \end{aligned}$$

$$\begin{aligned} L(\theta, \eta; x_1, \dots, x_n) &= \prod_{i=1}^n f(\theta, \eta; x_i) = \prod_{i=1}^n \frac{1}{\theta} e^{-(x_i-\eta)/\theta} \mathbb{1}(x_i \geq \eta) \\ &= \frac{1}{\theta^n} e^{-\sum_{i=1}^n (x_i-\eta)/\theta} \mathbb{1}(x_1 \geq \eta, \dots, x_n \geq \eta) \\ &= \frac{1}{\theta^n} e^{-\sum_{i=1}^n (x_i-\eta)/\theta} \mathbb{1}(\min(x_1, \dots, x_n) \geq \eta). \end{aligned}$$

If all x_i's are larger than η the minimum of them is larger too. Hence the event $\{x_1 \geq \eta, \dots, x_n \geq \eta\}$ is equivalent to $\{\min(x_1, \dots, x_n) \geq \eta\}$.

$$\begin{aligned} l(\theta, \eta; x_1, \dots, x_n) &= \Big(-n \ln\theta - \sum_{i=1}^n (x_i - \eta)/\theta \Big) \mathbb{1}(\min(x_1, \dots, x_n) \geq \eta) \\ &= \Big(-n \ln\theta - \sum_{i=1}^n x_i/\theta + n\eta/\theta \Big) \mathbb{1}(\min(x_1, \dots, x_n) \geq \eta) \end{aligned}$$

2. About η we see that $l(\theta, \eta; x_1, \dots, x_n) \neq 0$ it is increasing in η, hence the larger value of η, the larger will be the value of the log-likelihood. However

if η is larger then the minimum of the observations $\min(x_1, \dots, x_n)$ the value of the log-likelihood becomes zero. Hence the maximum of l is attained at $\eta = \min(x_1, \dots, x_n)$. Hence $\min(x_1, \dots, x_n)$ is the MLE of η.

For θ we have

$$\frac{\partial l(\theta, \eta; x_1, \dots, x_n)}{\partial \theta} = -n/\theta + n\bar{x}/\theta^2 - n\eta/\theta^2 = 0$$

Solving for θ the equation above we obtain $\theta = \bar{x} - \eta$ hence, the estimator $\hat{\theta}^{MLE} = \bar{x} - \hat{\eta}^{MLE} = \bar{x} - \min(x_1, \dots, x_n)$.

3. For $x \geq \eta$ we have

$$P(X \leq x) = \int_\eta^x \frac{1}{\theta} e^{-(s-\eta)/\theta} ds = [-e^{-(s-\eta)/\theta}]_\eta^x = 1 - e^{-(x-\eta)/\theta}$$

4. The α-percentile is given by q_α which solves $P(X \leq q_\alpha) = \alpha$. Hence using the previous answer we get

   Solving the last equation with respect to q_α we obtain $q_\alpha = \eta - \theta \ln(1 - \alpha)$. To find an estimator of q_α it is natural to think to use the MLEs $\hat{\eta}^{MLE}, \hat{\theta}^{MLE}$:

$$\hat{q_\alpha} = \hat{\eta}^{MLE} - \hat{\theta}^{MLE}(1 - \alpha).$$

5.

$$P(\min(X_1, \dots, X_n) \leq x) = 1 - P(\min(X_1, \dots, X_n) \geq x)$$
$$= 1 - P(X_1 \geq x, \dots, X_n \geq x) = 1 - \prod_{i=1}^n P(X_i \geq x)$$

   Next we use

$$P(X \geq x) = 1 - P(X \leq x) = 1 - (1 - e^{-(x-\eta)/\theta}) = e^{-(x-\eta)/\theta}$$

   We obtain

$$P(\min(X_1, \dots, X_n) \leq x) = 1 - e^{-n(x-\eta)/\theta}$$

   for $x \geq \eta$. We recognize that this is the distribution of the two parameter Exponential again, with location parameter η and scale θ/n. The pdf is given by

$$f_{\min}(x; \eta, \theta/n) = \frac{n}{\theta} e^{-n(x-\eta)/\theta}$$

6. We need to compute $E(\hat{\eta}^{MLE}) = E(\min(X_1, \dots, X_n))$. We could proceed with calculating the corresponding integral and using the pdf found in the previous question. However from the first question we have established that if X has bivariate Exponential distribution with location parameter η and scale θ, then $E(X) = \eta + \theta$ and $E(X^2) = \eta^2 + 2\eta\theta + 2\theta^2$. This means

that $E(\hat{\eta}^{MLE}) = E(\min(X_1, \dots, X_n)) = \eta + \theta/n$. The last expression goes to η for $n \to \infty$ so indeed $\hat{\eta}^{MLE}$ is asymptotically unbiased for η. Similarly we have $E(\min(X_1, \dots, X_n)^2) = \eta^2 + 2\eta(\theta/n) + 2(\theta/n)^2$. Then for the variance

$$\begin{aligned} Var(\min(X_1, \dots, X_n)) &= E(\min(X_1, \dots, X_n)^2) - (E \min(X_1, \dots, X_n))^2 \\ &= \eta^2 + 2\eta(\theta/n) + 2(\theta/n)^2 - \eta^2 \end{aligned}$$

Indeed the variance tends to zero as $n \to \infty$.

7.

$$E(\hat{\theta}^{MLE}) = E(\bar{X} - \min(X_1, \dots, X_n)) = E(\bar{X}) - E(\min(X_1, \dots, X_n)) = \eta + \theta - (\eta + \theta/n)$$

The last expression goes to θ as $n \to \infty$, so indeed $\hat{\theta}^{MLE}$ is asymptotically unbiased for θ.

8. The first moment of X around zero is $E(X)$ and the second moment around zero is $E(X^2)$. To find the MMEs of η and θ we need to solve the system with respect to η and θ:

$$\begin{aligned} \bar{X} &= \eta + \theta \\ \overline{X^2} &= \eta^2 + 2\eta\theta + 2\theta^2 \end{aligned}$$

where $\bar{X} = (1/n) \sum_{i=1}^n X_i$ and $\overline{X^2} = (1/n) \sum_{i=1}^n X_i^2$. This leads to the solution

$$\begin{aligned} \theta &= \sqrt{\overline{X^2} - \bar{X}^2} \\ \eta &= \bar{X} - \sqrt{\overline{X^2} - \bar{X}^2}. \end{aligned}$$

Hence the corresponding estimators

$$\begin{aligned} \hat{\theta}^{MME} &= \sqrt{\overline{X^2} - \bar{X}^2} \\ \hat{\eta}^{MME} &= \bar{X} - \sqrt{\overline{X^2} - \bar{X}^2}. \end{aligned}$$

9. As for the sample variance we have $S^2 = \frac{n}{n-1}(\overline{X_i^2} - \bar{X}^2)$ we have $S^2 = \frac{n}{n-1}(\hat{\theta}^{MME})^2$ and hence $\hat{\theta}^{MME} = \sqrt{\frac{n}{n-1}} S$. Then $\hat{\eta}^{MME} = \bar{X} - \sqrt{\frac{n}{n-1}} S$.

3 Properties of $\bar{X}$ and S^2

Exercise 1

According to the CLT, whatever the distribution, the empirical mean tends to a normal, i.e.

$$\frac{\bar{X} - \mu}{\sqrt{55.0564/100}} \sim N(0, 1)$$

Since $Z_{95\%} = 1.645$ then

$$P\left(-1.645 \leq \frac{\bar{X}-\mu}{\sqrt{55.0564/100}} \leq 1.645\right) = 90\%$$

and the confidence interval for μ is then

$$\left[176 - 1.645\sqrt{55.0564/100}\,;\, 176 + 1.645\sqrt{55.0564/100}\right] = [174.7794; 177.2206]$$

Exercise 2

a) Under the assumption that X is normal, $\frac{\bar{X}-\mu}{\sqrt{S^2/n}}$is a Student's T random variable with 99 degrees of freedom

$$\frac{\bar{X}-\mu}{\sqrt{136.89/100}} \sim t_{99}$$

Since $t_{99\,,\,95\%} = 1.645$ (using the table, we use $df = 100$ since $df = 99$ is not displayed)

$$P\left(-1.645 \leq \frac{\bar{X}-\mu}{\sqrt{136.89/100}} \leq 1.645\right) = 90\%$$

and the confidence interval for μ is

$$\left[174 - 1.645\sqrt{136.89/100}\,;\, 174 + 1.645\sqrt{136.89/100}\right]$$

b) Under the assumption that X is normal, $(n-1)\,\frac{S^2}{\sigma^2}$ is a chi-square random variable with 99 degrees of freedom

$$99\frac{136.89}{\sigma^2} \sim \chi^2_{99}$$

Since $\chi^2_{99\,,\,95\%} = 124.32$ and $\chi^2_{99\,,\,5\%} = 77.9295$ (again, we take $df = 100$) then

$$P\left(77.9295 \leq 99\frac{136.89}{\sigma^2} \leq 124.342\right) = 90\%$$
$$P\left(\sigma^2 \leq 99\frac{136.89}{77.9295}\,;\, 99\frac{136.89}{124.342} \leq \sigma^2\right) = 90\%$$

and the confidence interval for σ^2 is

$$[108.9906\,; 173.9022]$$

Exercise 3

a) If the two populations have the same variance $\sigma_m^2 = \sigma_f^2 = \sigma^2$, an unbiased "pooled" estimator of this variance is

$$\begin{aligned} S^2_{pool} &= \frac{(n_m - 1)\, S_m^2 + (n_f - 1) S_f^2}{n_m + n_f - 2} \\ &= \frac{99\,(9.3)^2 + 99\,(8.7)^2}{198} = 81.09\,. \end{aligned}$$

In this case,

$$\frac{\left(\bar{X_m} - \bar{X_f}\right) - (\mu_m - \mu_f)}{S_{pool}\sqrt{\frac{1}{n_m} + \frac{1}{n_f}}} \sim t_{n_m + n_f - 2}$$

Given that $t_{198,97.5\%} = 1.96$,

$$P\left(-1.96 \leq \frac{(177 - 173) - (\mu_m - \mu_f)}{\sqrt{81.9}\sqrt{\frac{1}{100} + \frac{1}{100}}} \leq 1.96\right) = 95\%$$

the confidence interval for $\mu_m - \mu_f$ is

$$\left[\underbrace{4 - 1.96 \times 1.27}_{1.5108} ; \underbrace{4 + 1.96 \times 1.27}_{6.4892}\right] .$$

b) $(n_m + n_f - 2)\, \frac{S^2_{pool}}{\sigma^2}$ is a chi-square random variable with 198 degrees of freedom

$$198 \frac{81.9}{\sigma^2} \sim \chi^2_{198}$$

Since $\chi^2_{198\,,\,97.5\%} = 238.86$ and $\chi^2_{198\,,\,2.5\%} = 160.92$ (we invite you to compute these values using Python since the table is not precise enough) then

$$\begin{aligned} P\left(160.92 \leq 198 \frac{81.9}{\sigma^2} \leq 238.86\right) &= 95\% \\ P\left(\sigma^2 \leq \frac{16216.2}{160.92} \,;\, \frac{16216.2}{238.86} \leq \sigma^2\right) &= 95\% \end{aligned}$$

and the confidence interval for σ^2 is

$$[67.88\,;\,100.77]\ .$$

Exercise 4

a) First of all, we are going to workout S^2_{Obs} and $\overline{X}_{Obs}$. We recall that

$$\overline{X} = \frac{1}{n}\sum_{k=1}^{n} X_k$$

and

$$S^2 = \frac{1}{n-1}\sum_{k=1}^{n}(X_k - \overline{X})^2$$

Therefore: $S^2_{Obs} = (2,683)^2$ and $\overline{X}_{Obs} = 55,083$.
Since $(X_i)_{1\leq i\leq n}$ are iid $\sim N(\mu, \sigma)$, we have: $\frac{\overline{X}-\mu}{\sqrt{\frac{S^2}{n}}} \sim t_{35}$.
Therefore, the confidence interval for μ at 2% is:

$$[\overline{X}_{Obs} - t_{35,99\%}\sqrt{\frac{S^2_{Obs}}{n}}; \overline{X}_{Obs} + t_{35,99\%}\sqrt{\frac{S^2_{Obs}}{n}}] = [53,99; 56,17]$$

(Use Python to find the value of $t_{35,99\%} = 2.437$).

b) Under the assumption of $(X_i)_{1\leq i\leq n}$ are iid $\sim N(\mu, \sigma)$, we obtain:

$$35\frac{S^2}{\sigma^2} \sim \chi^2_{35}$$

Therefore, the confidence interval of σ^2 at 2% is:

$$[\frac{35S^2_{Obs}}{\chi^2_{35,99\%}}; \frac{35S^2_{Obs}}{\chi^2_{35,1\%}}] = [4,39; 13,62]$$

(Use python to find $\chi^2_{35,99\%} = 57,34$ and $\chi^2_{35,1\%} = 18,5$).

4 Hypothesis testing

Exercise 1

a) We test the following hypothesis

$$\begin{cases} H_0 & \mu = 60 \\ H_1 & \mu \neq 60 \end{cases}$$

with the Student's t statistic $\frac{\bar{X}-\mu}{S/\sqrt{n}}$. Given that $t_{obs} = 1.625$ and $t_{8,0.975} = -2.31$, $t_{8,0.975} = 2.31$ we do not reject H_0 (since $t_{8,0.025} \leq t_{obs} \leq t_{8,0.975}$).

Since the distribution of the statistics of test is symmetric around 0, the probability of observing t_{obs} or $-t_{obs}$ are the same. Given that extremely low or high realizations of the statistics disqualify H_0, the p-value is

$$\begin{aligned} p_{val} &= P(t_8 \leq -t_{obs}) + P(t_8 \geq t_{obs}) \\ &= 2P(t_8 \leq -t_{obs}) = 14.28\% \end{aligned}$$

As $p_{val} \geq 5\%$, we do not reject H_0.

b)We test the following hypothesis

$$\begin{cases} H_0 & \sigma^2 = 0.81 \\ H_1 & \sigma^2 > 0.81 \end{cases}$$

with the chi-square statistics $\frac{(n-1)S^2}{\sigma^2}$. We have: $\chi^2_{obs} = 12.94$ and $\chi^2_{8,0.95} = 15.51$. Since $\chi^2_{obs} \leq \chi^2_{8,0.95}$ we do not reject this assumption.

Given that extremely high realizations of this statistics disqualify H_0, the p-value is

$$\begin{aligned} p_{val} &= P(\chi^2_8 \geq \chi^2_{obs}) \\ &= 11.39\% \end{aligned}$$

Given that $p_{val} \geq 5\%$, we do not reject H_0.

c) The confidence interval for μ is obtained with the Student's t statistic $\frac{\bar{X}-\mu}{S/\sqrt{n}}$ as the variance is unknown. We have then

$$P\left(\bar{X} - t_{n-1,1-\frac{\alpha}{2}} \frac{S}{\sqrt{N}} \leq \mu \leq \bar{X} + t_{n-1,1-\frac{\alpha}{2}} \frac{S}{\sqrt{N}}\right) = 1 - \alpha$$

i.e. [59.74; 61.50]. Notice that 60 is well in the confidence interval. This confirms the assumption tested in question a).

d) A confidence interval for σ^2 is given by

$$P\left(\frac{(n-1)S^2}{\chi^2_{n-1\,;\,1-\frac{\alpha}{2}}} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{n-1\,;\,\frac{\alpha}{2}}}\right) = 1 - \alpha$$

Exercise 2

First we find

A	B
$n_1 = 8$	$n_2 = 6$
$\bar{x}_1 = 61$	$\bar{x}_2 = 49$
$s_1^2 = 221.4$	$s_2^2 = 138.0$

a) We test the following hypothesis

$$\begin{cases} H_0 & \sigma_1^2 = \sigma_2^2 \\ H_1 & \sigma_1^2 \neq \sigma_2^2 \end{cases}$$

with the Fisher statistic $\frac{S_1^2}{S_2^2} \sim F_{n_1-1,n_2-1}$. In our case, we obtain $F_{7,5,obs} = \frac{221.4}{138.0} = 1.604$. Given that $F_{7,5,1\%} = 0.1340$ and $F_{7,5,99\%} = 10.46$, we do not reject this assumption (as $F_{7,5,1\%}\leq F_{7,5,obs}\leq F_{7,5,99\%}$).

b)We test the following hypothesis

$$\begin{cases} H_0 & \mu_1-\mu_2=5 \\ H_1 & \mu_1-\mu_2>5 \end{cases}.$$

As we have not rejected the equality of variances but this variance being unknown, we use the statistics

$$\frac{\bar{X}_1-\bar{X}_2-5}{S_{pool}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\sim t_{n_1+n_2-2}$$

where

$$\begin{aligned} S_{pool}^2 &= \frac{(n_1-1)\,S_1^2+(n_2-1)S_2^2}{n_1+n_2-2} \\ &= 186.65\,. \end{aligned}$$

Given that $t_{obs}=0.9487\leq t_{12,95\%}=1.78$, we do not reject H_0.

High values of the test statistics disqualify H_0, the p-value is then (you can compute it easily using Python)

$$\begin{aligned} p_{val} &= P(t_{12}\geq t_{obs}) \\ &= 18.14\% \end{aligned}$$

As $p_{val}\geq 5\%$, we do not reject H_0.

c)We check that

$$\begin{cases} H_0 & \mu_1=\mu_2 \\ H_1 & \mu_1\neq\mu_2 \end{cases}.$$

We a test statistic similar to that of question b) (you just need to withdraw -5). We find

$$t_{12,2.5\%}=-2.18\leq t_{obs}=1.626\leq t_{12,97.5\%}=2.18$$

and then do not reject the assumption of equality of averages.

High and low values of the test statistics disqualify H_0. Since the distribution of the statistics of test is symmetric around 0, the probability of observing t_{obs} or $-t_{obs}$ are the same. The p-value is hence

$$\begin{aligned} p_{val} &= P(t_{12}\leq -t_{obs}=-1.626)+P(t_{12}\geq t_{obs}=1.626) \\ &= 12.99\% \end{aligned}$$

As $p_{val} \geq 5\%$, we do not reject H_0.

d) A confidence interval for $\mu_1 - \mu_2$ is given by:

$$\left[\bar{X}_1 - \bar{X}_2 - t_{12,97.5\%} S_{pool}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\,; \bar{X}_1 - \bar{X}_2 + t_{12,97.5\%} S_{pool}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right]$$

which is $[-4.077356\ ;\ 28.077356]$. Notice that 0 is well in this interval. This confirms the conclusion obtained in question c).

Exercise 3

a) To apply a 2 populations test, the assumption of independence must be fulfilled. Here, we have one single population (the 10 samples of ore) on which we perform two experiments. Therefore, we will study the differences between measures in order to come back to a one dimension random variable. This difference is noted D and the 10 realizations of D are:

Sample	1	2	3	4	5	6	7	8	9	10
D	3.2	-3.9	-2.3	-0.6	1.1	-2.9	-1.8	-0.5	-2.3	-0.3

The sample mean and variance are $\bar{d} = -1.03$ and $s^2 = 4.33$. We test

$$\begin{cases} H_0 & \mu_D = 0 \\ H_1 & \mu_D \neq 0 \end{cases}.$$

We use the statistics $\frac{\bar{D}-\mu_D}{S_D/\sqrt{10}} \sim t_9$. Since $t_{9,2.5\%} = -2.26 < t_{obs} = -1.565 < t_{9,97.5\%} = 2.26$, we cannot conclude that both methods lead to different measures.

Remark: if the methods were tested separately on two different samples, we would have applied a 2 population test.

b) The two-sided p-value (computed with Python) is

$$\begin{aligned} p_{val} &= P(t_9 \leq -t_{obs}) + P(t_9 \geq t_{obs}) \\ &= 15.20\%\,. \end{aligned}$$

As $p_{val} \geq 5\%$, we do not reject H_0.

Exercise 4

We want to test

$$\begin{cases} H_0 & \mu = 120 \\ H_1 & \mu = 140 \end{cases}.$$

a) This is the probability of an error of type II

$$P(accept\, A\,|\,B\, is\, true) = error\, of\, type\, II = P(\bar{X} \leq 135)$$

where $X_i \sim \mathcal{N}(\mu = 140\,;\, \sigma = 18)$. Then $\bar{X} \sim \mathcal{N}(\mu = 140\,;\, \frac{\sigma}{\sqrt{25}} = 3.6)$ and $P(\bar{X} \leq 135) = 8.24\%$ (you can compute it using Python).

b)This is the probability of an error of type I

$$P(accept\, B \,|\, A\, is\, true) = error\, of\, type\, I = P(\bar{X} \geq 135)$$

where $X_i \sim \mathcal{N}(\mu = 120\,;\, \sigma = 18)$. Then $\bar{X} \sim \mathcal{N}(\mu = 120\,;\, \frac{\sigma}{\sqrt{25}} = 3.6)$ and $P(\bar{X} \geq 135) = 0.001\%$ (you can compute it using Python).

Exercise 5

a) We test the following hypothesis

$$\begin{cases} H_0: & p = 0.10 \\ H_1: & p > 0.10 \end{cases}$$

We haven't seen during the lecture how to test a proportion but the method for solving this is similar to the one applied for testing the mean of a random sample. Firstly, we assume that H_0 holds ($p = 0.10$) and calculate the probability $P(X \geq 4)$, that is the p-value. If this p-value is smaller than 5%, we reject H_0.

$$\begin{aligned} P(X \geq 4) &= 1 - P(X \leq 3) \\ &= 1 - 0.764 = 23.6\%\,. \end{aligned}$$

We do not reject H_0 based on this test.

b) We test the following hypothesis

$$\begin{cases} H_0: & p = 0.10 \quad \Leftrightarrow \mu = 100\,;\, \sigma^2 = 90 \\ H_1: & p \neq 0.10 \end{cases}$$

Under H_0, $N \sim \mathcal{N}(100, 90)$ and then the statistic $Z = \frac{N-100}{\sqrt{90}} \sim \mathcal{N}(0,1)$ and symmetric around zero and $Z_{obs} = -2.95$. You reject H_0 because:

$$Z_{2.5\%} = -1.96 \geq -2.95\,.$$

The distribution of the statistics being symmetric around zero, the events $Z_{obs} = 2.95$ and $Z_{obs} = -2.95$ have the same probability of being observed. The p-value is then (you can compute it precisely using Python; you can approximate it using the tables)

$$\begin{aligned} p\,value &= P(Z \leq -2.95) + P(Z \geq 2.95) \\ &= 0.00317 \end{aligned}$$

which is well lower than $\alpha = 0.05$. Therefore we reject H_0.

Exercise 6

a) $\bar{x} = 460.8/8 = 57.6$ and $s_x^2 = 4.12$. The statistics $\frac{\bar{X}-\mu_x}{S_x/\sqrt{n_x}} \sim t_{n_x-1}$ therefore

$$P\left(-t_{7,1-\frac{\alpha}{2}} \leq \frac{\bar{X}-\mu_x}{S_x/\sqrt{n_x}} \leq t_{7,1-\frac{\alpha}{2}}\right) = 1-\alpha$$
$$P\left(-2.36 \leq \frac{\bar{X}-\mu_x}{S_x/\sqrt{n_x}} \leq 2.36\right) = 1-\alpha$$

Therefore

$$\mu \in \left[\bar{X} - 2.36 S_x/\sqrt{n_x}\,;\, \bar{X} + 2.36\, S_x/\sqrt{n_x}\right] = [55.90279\,;\, 59.29721]$$

b) We perform the following test:

$$\begin{cases} H_0 & \mu_Y - \mu_X = 4 \\ H_1 & \mu_Y - \mu_X > 4 \end{cases}.$$

Under the assumption that variances are equal, we use the statistics

$$\frac{\bar{Y} - \bar{X} - 4}{S_{pool}\sqrt{\frac{1}{n_x} + \frac{1}{n_y}}} \sim t_{n_x+n_y-2}$$

where $\bar{x} = 460.8/8 = 57.6$ and $\bar{y} = 762/12 = 63.5$

$$\begin{aligned} S_{pool}^2 &= \frac{(n_x-1)\,S_x^2 + (n_y-1)S_y^2}{18} \\ &= \frac{7 \times 4.12 + 11 \times 2.56}{18} \\ &= 3.167 \end{aligned}$$

Given that $t_{obs} = 2.34 > t_{18,95\%} = 1.734$, we reject this assumption.

High values of the test statistics disqualify H_0, the p-value is then

$$\begin{aligned} p_{val} &= P(t_{18} \geq t_{obs}) \\ &= 1.55\% \end{aligned}$$

As $p_{val} \geq 5\%$, we reject H_0.

c) we must check the equality of variances. If they are not equal, results obtained in b do not hold. We test the following hypothesis

$$\begin{cases} H_0 & \sigma_x^2 = \sigma_y^2 \\ H_1 & \sigma_x^2 \neq \sigma_y^2 \end{cases}$$

with the Fisher statistic $\frac{S_x^2}{S_y^2} \sim F_{n_1-1,n_2-1}$. In our case, we obtain $F_{7,11,obs} = 1.61$. Given that $F_{7,11,2.5\%} = 0.212$ (computed with Python... With the tables, you'll only be able to state that is between 0.21 and 0.214) and $F_{7,11,97.5\%} = 3.76$, we do not reject this assumption. Therefore, we have no reason not to trust the results obtained in b). If, instead, we had rejected the assumption of equality of variances, the results would no longer be valid. Another test would then need to be used to test the equality of means (for example Welch's t-test, not seen in the lecture).

Exercise 7

1. A *statistic* is a function of the data $X_1, X_2, \ldots, X_n$ that does not contain any unknown parameter or quantity.

2. $$E(\bar{X}) = E\Big(\frac{1}{n}\sum_{i=1}^{n} X_i\Big) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n} nE(X) = E(X) = \mu$$

   $$Var(\bar{X}) = Var\Big(\frac{1}{n}\sum_{i=1}^{n} X_i\Big) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) = \frac{Var(X)}{n} = \sigma^2/n$$

   The second equation follows from the fact that since the sample is random, the observations are independent from each other and hence the covariance terms are zero.

   Yes $\bar{X}$ is a statistic as it does not contain any unknown quantity and it is a function of the data.

3. The sample variance denoted by S^2 or a hat over the parameter, i.e., $\hat{\sigma}^2$ is

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

   It also equals

   $$\begin{aligned} S^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i^2 - 2X_i\bar{X} + \bar{X}^2) \\ &= \frac{1}{n-1}\Big(\sum_{i=1}^{n} X_i^2\Big) - \frac{2n\bar{X}^2}{n-1} + \frac{n\bar{X}^2}{n-1} = \frac{1}{n-1}\Big(\sum_{i=1}^{n} X_i^2\Big) - \frac{n\bar{X}^2}{n-1} \\ &= \frac{n}{n-1}\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \frac{n}{n-1}\bar{X}^2 = \frac{n}{n-1}(\overline{X^2} - \bar{X}^2). \end{aligned}$$

   In the third equality use $\sum_{i=1}^{n}(2X_i\bar{X}) = 2\bar{X}\sum_{i=1}^{n} X_i = 2\bar{X}n\bar{X}$. Yes, it is a statistic as it does not contain any unknown quantity and it is a function of the data.

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

Now we use that $E(X^2) = Var(X) + (E(X))^2 = \sigma^2 + \mu^2$ and $E(\bar{X}^2) = Var(\bar{X}) + (E\bar{X})^2 = \sigma^2/n + \mu^2$

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

Hence the sample variance is unbiased for the population variance.

4. According to the theorem that linear combination of normal variables is again Normal, it follows that the distribution of $\bar{X}$ is Normal with parameters μ and σ^2/n.

5. We have $\bar{x} = -0.734, s = 4.181$.

$$\mu \in [\bar{x} - t_{4,0.975}s/\sqrt{n}, \quad \bar{x} + t_{4,0.975}s/\sqrt{n}]$$
$$= [-0.734 - 2.776 \times 4.181/\sqrt{5}, \quad -0.734 + 2.776 \times 4.181/\sqrt{5}] = [-5.92, 4.46]$$

The quantity $t_{k,p}$ is such that $P(T_k \leq t_{k,p}) = p$ if T_k has t-distribution with k degrees of freedom.

$$\sigma^2 \in \Big[\frac{(n-1)s^2}{\chi^2_{4,0.975}}, \frac{(n-1)s^2}{\chi^2_{4,0.025}}\Big] = \Big[\frac{4 \times 4.181^2}{11.14}, \frac{4 \times 4.181^2}{0.4844}\Big] = [6.28, 144.35]$$

The quantity $\chi^2_{k,p}$ is such that $P(H_k^2 \leq \chi^2_{k,p}) = p$ where H_k^2 is a χ^2-distributed variable with k degrees of freedom.

6.

$$\mu \in [\bar{x} - z_{0.975}\sigma/\sqrt{n}, \quad \bar{x} + z_{0.975}\sigma/\sqrt{n}]$$
$$= [-0.734 - 1.96 \times 3/\sqrt{5}, \quad -0.734 + 1.96 \times 3/\sqrt{5}] = [-3.36, 1.9]$$

The quantity z_p is such that $P(Z \leq z_p) = p$ if $Z \sim N(0,1)$.

Knowing the population variance reduces the length of the confidence interval for the mean, the effect is larger because the sample is really small, only 5 observations.

7. No, the sample size is too small, if it was larger we could use the CLT to motivate the use of the confidence interval derived in the previous question.

8. We will use the information $\sigma = 3$. We have $H_0 : \mu \leq 5, H_a : \mu > 5$. The value of the test statistic is $(\bar{x} - 5)/(3/\sqrt{5}) = -4.27$ compared to a critical value of $z_{0.05} = 1.64$ we see that it is less than the critical value, hence we can not reject the null hypothesis.

9. We will use the information $\sigma = 3$. We have $H_0 : \mu \geq 5, H_a < 5$. The value of the test statistic is $(\bar{x} - 5)/(3/\sqrt{5}) = -4.27$ compared to a critical value of $z_{0.05} = -1.64$ we see that it is less than the critical value, hence we reject the null hypothesis.

Exercise 8

a) Let p_1 and p_2 be respectively the proportion of students who passed the exam in group 1 and group 2.
Let $\hat{p}_1$ and $\hat{p}_2$ be respectively the (empirical) proportion of students who passed the exam in group 1 and group 2.
According to the CLT, we can approximate $\hat{p}_1 \approx N(p_1, \frac{p_1(1-p_1)}{n_1})$ and $\hat{p}_2 \approx N(p_2, \frac{p_2(1-p_2)}{n_2})$.
Since the two groups are independent, we obtain:

$$\frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \approx N(0,1)$$

We wish to test the following hypothesis:

$$\begin{cases} H_0: \ p_1 = p_2 = p \\ H_1: \ p_1 < p_2 \end{cases}$$

.

At this point, we need to find a test statistic. Examine how the quantity above reduces under H_0. On the numerator, the unknown parameters disappear since $p_1 - p_2 = 0$. On the denominator, this won't be the case. Indeed, the denominator becomes

$$\sqrt{p(1-p)(\frac{1}{n_1} + \frac{1}{n_2})}$$

As the hint indicates, we can replace p by a suitable estimator and keep the asymptotic normality.[1] One obvious estimator is $\hat{p} = \frac{n_1\hat{p}_1 + n_1\hat{p}_1}{n_1+n_2}$. Hence, our test statistic finally is

$$\frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}}$$

and is approximately $N(0,1)$ under H_0. Since $\hat{p_1} = \frac{126}{180}$ and $\hat{p_2} = \frac{129}{150}$, we find that $\hat{p} = 0,77$ and then we calculate z_{Obs} to find:

$$z_{Obs} = -3,44$$

[1] This is beyond the scope of this course but it uses the so-called Slutsky's Theorem and this requires that we find a *consistent* estimator for p.

Since $z_{Obs} < z_{5\%} = -1,645$, we reject H_0.
Therefore, the success is better with more hours of tutorials.

b) The p-value is:

$$p_{value} = P(Z < -3,44) = 0.00029 < \alpha$$

(Use python to workout the accurate value)
Hence we reject H_0.

5 Linear regression

Exercise 1

a) $\widehat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = 0.89$ and $\widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x} = 68.26$. The goodness of fit is measured by $r^2 = \frac{SSR}{SST} = 96.64\%$ which is very good. This tends to confirm the linearity.

b) $\hat{\sigma} = \sqrt{\frac{SSE}{n-2}} = 0.73$. As $(n-2)\frac{\widehat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-2}$, we have

$$P\left(\chi^2_{n-2\,2.5\%} \leq (n-2)\frac{\widehat{\sigma}^2}{\sigma^2} \leq \chi^2_{n-2\,97.5\%}\right) = 95\%\,,$$

or

$$\sigma \in \left[\sqrt{(n-2)\frac{\widehat{\sigma}^2}{\chi^2_{n-2\,97.5\%}}}\,;\,\sqrt{(n-2)\frac{\widehat{\sigma}^2}{\chi^2_{n-2\,2.5\%}}}\right]$$

here $\sigma \in [0.53, 1.21]$.

c) $\hat{\sigma}_{\beta_0} = \sqrt{\frac{\widehat{\sigma}^2\,\overline{x^2}}{S_{xx}}} = 0.35$ and $\hat{\sigma}_{\beta_1} = \sqrt{\frac{\widehat{\sigma}^2}{S_{xx}}} = 0.05$ which are low with respect to estimates.

d) The linear regression is highly significant as $p_{val} << 5\%$:

	d.f.	Value	Mean of ...	F^*	p-value
SSR	1	185.76	185.76	346.16	0
SSE	12	6.44	6.44		
SST	13	192.20			

e) We use the statistics $\frac{\widehat{\beta}_j - \beta_{j,0}}{\widehat{\sigma}\sqrt{c_{jj}}} \sim t_{n-2}$ where $c_{0,0} = \frac{\overline{x^2}}{S_{xx}}$ and $c_{1,1} = \frac{1}{S_{xx}}$. $t_{obs} = 193.66$ and $t_{n-2\,,\alpha/2} = -2.18$, $t_{n-2\,,1-\alpha/2} = 2.18$. Then we reject the hypothesis. The confidence interval is:

$$\beta_0 \in \left[\widehat{\beta}_0 - t_{n-k-1\;1-\alpha/2}\,\widehat{\sigma}\sqrt{\frac{\overline{x^2}}{S_{xx}}}\,;\,\widehat{\beta}_0 + t_{n-k-1\;1-\alpha/2}\,\widehat{\sigma}\sqrt{\frac{\overline{x^2}}{S_{xx}}}\right].$$

Here $\beta_0 \in [67.49 \,:\, 69.03]$ which is quite accurate.

f) The prediction is

$$\widehat{y}_0 = \widehat{\beta}_0 + \widehat{\beta}_1 x_0 \ = 81.5755$$

We need S^2_{pred} to evaluate the prediction interval:

$$S^2_{pred} = \widehat{\sigma}^2 \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right) \quad = \quad 0.753\,.$$

Finally, the prediction interval is:

$$\left[\widehat{y}_0 - S_{pred}\, t_{n-2\,;\,1-\frac{\alpha}{2}} \;;\; \widehat{y}_0 + S_{pred}\, t_{n-2\,;\,1-\frac{\alpha}{2}}\right] = [79.68\,;\, 83.47]$$

Exercise 2

a) According to the properties of expectation,

$$\mathbb{E}\left(Y_0 - \widehat{\beta}_0 - \widehat{\beta}_1 x_0\right) \quad = \quad \mathbb{E}\left(Y_0\right) - \mathbb{E}\left(\widehat{\beta}_0\right) - x_0 \mathbb{E}\left(\widehat{\beta}_1\right)\,.$$

But $\mathbb{E}\left(Y_0\right) = \beta_0 + \beta_1 x_0 + \mathbb{E}\left(\epsilon\right) = \beta_0 + \beta_1 x_0$. Given that $\mathbb{E}\left(\widehat{\beta}_0\right) = \beta_0$ and $\mathbb{E}\left(\widehat{\beta}_1\right) = \beta_1$, we infer that $\mathbb{E}\left(Y_0 - \widehat{\beta}_0 - \widehat{\beta}_1 x_0\right) = 0$.

b) According to the properties of variance (variances of a sum or a difference sum up),

$$\begin{aligned}\mathbb{V}\left(Y_0 - \widehat{\beta}_0 - \widehat{\beta}_1 x_0\right) &= \mathbb{V}\left(Y_0\right) + \mathbb{V}\left(\widehat{\beta}_0 + \widehat{\beta}_1 x_0\right) \\ &= \sigma^2 + \left(\mathbb{V}\left(\widehat{\beta}_0\right) + x_0^2 \mathbb{V}\left(\widehat{\beta}_1\right) + 2x_0 \mathbb{C}\left(\widehat{\beta}_0, \widehat{\beta}_1\right)\right)\end{aligned}$$

where (see lectures), $\widehat{\beta}_{j=0,1}$ are normal with

$$\mathbb{V}\left(\widehat{\beta}_0\right) = \frac{\sigma^2\, \overline{x^2}}{S_{xx}} \qquad \mathbb{V}\left(\widehat{\beta}_1\right) = \frac{\sigma^2}{S_{xx}} \quad \mathbb{C}\left(\widehat{\beta}_0, \widehat{\beta}_1\right) = \frac{-\sigma^2 \bar{x}}{S_{xx}}\,.$$

The variance is then

$$\mathbb{V}\left(Y_0 - \widehat{\beta}_0 - \widehat{\beta}_1 x_0\right) = \sigma^2 \left(1 + \frac{\overline{x^2} + x_0^2 - 2x_0\bar{x}}{S_{xx}}\right)\,.$$

If we remember that

$$\begin{aligned}S_{xx} &= \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} 2x_i\bar{x} + n\bar{x}^2 \\ &= \sum_{i=1}^{n} x_i^2 - 2n\bar{x}^2 + n\bar{x}^2 = n\overline{x^2} - n\bar{x}^2\end{aligned}$$

We finally infer that

$$\mathbb{V}\left(Y_0 - \widehat{\beta}_0 - \widehat{\beta}_1 x_0\right) = \sigma^2 \left(1 + \frac{1}{n} + \frac{\left(x_0 - \bar{x}\right)^2}{S_{xx}}\right) .$$

As Y_0, $\widehat{\beta}_0$ and $\widehat{\beta}_1$ being normal random variables, their difference is also normal:

$$Y_0 - \widehat{\beta}_0 - \widehat{\beta}_1 x_0 \sim \mathcal{N}\left(0; \sigma^2 \left(1 + \frac{1}{n} + \frac{\left(x_0 - \bar{x}\right)^2}{S_{xx}}\right)\right) .$$

c) Let us recall that if $Z \sim \mathcal{N}(0,1)$ and Y are two independent r.v. then $\frac{Z}{\sqrt{Y/n}}$ is a Student's T random variable with n degrees of freedom. Given that $(n-2)\,\frac{\widehat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-2}$, we have that

$$\begin{aligned}
\frac{Y_0 - \widehat{\beta}_0 - \widehat{\beta}_1 x_0}{S_{pred}} &= \frac{Y_0 - \widehat{\beta}_0 - \widehat{\beta}_1 x_0}{\sqrt{\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}} \frac{1}{\widehat{\sigma}} \\
&= \underbrace{\frac{Y_0 - \widehat{\beta}_0 - \widehat{\beta}_1 x_0}{\sigma\sqrt{\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}}}_{\mathcal{N}(0,1)} \sqrt{\left(\frac{\underbrace{\frac{\widehat{\sigma}^2}{\sigma^2}(n-2)}_{\chi^2_{n-2}}}{(n-2)}\right)^{-1}} \\
&\sim t_{n-2} \, .
\end{aligned}$$

d) The confidence interval for Y_0 is

$$\left[\widehat{\beta}_0 + \widehat{\beta}_1 x_0 - S_{pred}\, t_{n-2\,;\,1-\frac{\alpha}{2}} \ ; \ \widehat{\beta}_0 + \widehat{\beta}_1 x_0 + S_{pred}\, t_{n-2\,;\,1-\frac{\alpha}{2}}\right]$$

Exercise 3

a) $\widehat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = -1.49$ and $\widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x} = 920.07$. The goodness of fit is measured by $r^2 = \frac{SSR}{SST} = 92.60\%$ which is very good. This tend to confirm the linearity.

b) $\hat{\sigma} = \sqrt{\frac{SSE}{n-2}} = 31.92$ and a confidence interval is given by

$$\begin{aligned}
IC_{0.95}(\sigma) &= \left[\sqrt{\frac{SSE}{\chi^2_{n-2;0.995}}}; \sqrt{\frac{SSE}{\chi^2_{n-2;0.005}}}\right] \\
&= \left[\sqrt{\frac{SSE}{29.8194}}; \sqrt{\frac{SSE}{3.565}}\right] \\
&= [21.076; 60.955]
\end{aligned}$$

c) $\hat{\sigma}_{\beta_0} = \sqrt{\frac{\widehat{\sigma}^2\,\overline{x^2}}{S_{xx}}} = 35.93$ and $\hat{\sigma}_{\beta_1} = \sqrt{\frac{\widehat{\sigma}^2}{S_{xx}}} = 0.12$ which are low with respect to estimates.

d) The linear regression is highly significant

	d.f.	Value	Mean of ...	F^*	p-value
SSR	1	165912.03	165912.03	162.83	0
SSE	13	13245.70	1018.90		
SST	14	179157.73			

e) $t_{obs} = -12.76$ and $t_{n-2,0.5\%} = -3.012$. Here $\beta_1 \in [-1.839 \, : \, -1.135]$ which is quite accurate.

f) The predictions are $\hat{y}_{410} = 310.26$ and $\hat{y}_{500} = 176.5$. The confidence intervals are $Y_{410} \in [203.71\,;\,416.81]$ and $Y_{500} \in [54.78\,;\,298.03]$ ($\alpha = 1\%$).

Exercise 4

a)The consumption is a random variable denoted by Y. We consider 3 binary variables, X_1, X_2 and X_3 that take the following values:

$$X_1 = \begin{cases} 1 & Vehicle\,A \\ 0 & otherwise \end{cases} \quad X_2 = \begin{cases} 1 & Vehicle\,B \\ 0 & otherwise \end{cases} \quad X_3 = \begin{cases} 1 & Vehicle\,C \\ 0 & otherwise \end{cases}.$$

The consumption is linked to explanatory variables by the linear model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$$

where $\epsilon \sim \mathcal{N}\left(0, \sigma^2\right)$. The vehicles A, B, C and D are respectively represented by the triplets $(1,0,0)$, $(0,1,0)$, $(0,0,1)$ and $(0,0,0)$. In this model, the expected consumption of vehicles A, B, C and D are respectively equal to: $\beta_0+\beta_1$, $\beta_0+\beta_2$, $\beta_0+\beta_3$ and β_0. This approach is a standard way of coding categorical variables (here the car model) in a linear regression model.

Remark: it would be wrong to consider a fourth variable

$$X_4 = \begin{cases} 1 & Vehicle\,D \\ 0 & otherwise \end{cases}$$

with

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \epsilon$$

because the model would be ill-posed, i.e. there is no unique set of parameters. Indeed, for any $(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4)$ and $\delta \in \mathbb{R}$, the parameters $\beta_0^{'} = \beta_0 - \delta$, $\beta_1^{'} = \beta_1 + \delta$, $\beta_2^{'} = \beta_2 + \delta$, $\beta_3^{'} = \beta_3 + \delta$ and $\beta_4^{'} = \beta_4 + \delta$ lead to the same Y. Furthermore, it is not possible to invert the matrix $\boldsymbol{X}^\top \boldsymbol{X}$ to find parameter estimates because the matrix $\boldsymbol{X}$ is ill-posed (null rank).

b) Testing this hypothesis is tantamount to testing the following assumptions:

$$\begin{cases} H_0 : & \beta_1 = \beta_2 = \beta_3 = 0\,, \\ H_1 : & \exists j \; such \; \beta_j \neq 0\,. \end{cases}$$

Under H_0, the expected consumption of a vehicle, whatever the model would be β_0. This test is performed with classic F-test:

	d.f.	Value	Mean of ...	F^*	p-value
SSR	3	9.4655	3.155	8.736	0.0006646
SSE	20	7.2233	0.3612		
SST	23	16.6888			

Based on these calculations, we reject H_0.

Exercise 5

a) $\widehat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = 1.49$ and $\widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x} = -336$.

b)

	d.f.	Value	Mean of ...	F^*	p-value
SSR	1	13912.90	13912.90	37944	2.98e-7
SSE	3	1.10	0.37		
SST	4	13914.00			

c) I would test $H_0 : \beta_0 = 0$ and (in the same test) $\beta_1 = 1$. There exists a testing procedure to conduct such a test but it is beyond the scope of this course.
d) The confidence interval is $[1.4676\,; 1.5164]$

Exercise 6

a) We consider 2 binary variables, X_1, X_2 that take the following values:

$$X_1 = \begin{cases} 1 & Technique\,1 \\ 0 & otherwise \end{cases} \quad X_2 = \begin{cases} 1 & Technique\,2 \\ 0 & otherwise \end{cases}$$

The tension is linked to explanatory variables by the linear model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

where $\epsilon \sim \mathcal{N}\left(0, \sigma^2\right)$. Using a software, we get

	d.f.	Value	Mean of ...	F^*	p-value
SSR	2	419.047	209.523	0.3710	0.7115
SSE	4	2258.66	564.666		
SST	6	2677.71			

Therefore we do not reject the assumption that $\mu_1 = \mu_2 = \mu_3$, i.e. all techniques yield bricks with the same resistance i.e. $Y = \beta_0 + \epsilon$.

b) An estimator of the variance is $S^2 = \frac{1}{6}\sum_{i=1}^{7}(y_i - \bar{y})^2 = 446.285$. A confidence interval for σ^2 is given by

$$P\left(\frac{6S^2}{\chi^2_{6\,;\,97.5\%}} \leq \sigma^2 \leq \frac{6S^2}{\chi^2_{6\,;\,2.5\%}}\right) = 1 - \alpha$$

where $\frac{6S^2}{\chi^2_{6\,;\,97.5\%}} = \frac{6\,446.3}{14.45} = 185.31$ and $\frac{6S^2}{\chi^2_{6\,;\,2.5\%}} = \frac{6\,446.3}{1.24} = 2177.07$. Then $\sigma^2 \in [185.3091\,;\,2164.6855]$.

Exercise 7

1. Random variables with distribution Chi-squared:
   - the sample variance of iid Normal variables

   $$(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$$

   - the sample pooled variance of two samples independent from each other, each of them containing iid Normal variables

   $$(n_X - n_Y - 2)S^2_{pool}/\sigma^2 \sim \chi^2_{n_X - n_Y - 2}$$

   - the sample variance of the error term in a linear regression model with k independent variables

   $$(n-k-1)\hat{\sigma}^2/\sigma^2 = SSE/\sigma^2 \sim \chi^2_{n-k-1}$$

   - the sum of squares total for a regression model

   $$SST/\sigma^2 \sim \chi^2_{n-1}$$

   - the sum of squares of the regression for a regression model with k independent variables

   $$SSR/\sigma^2 \sim \chi^2_k$$

2. Random variables with distribution Fisher (F):
   - the ratio of the sample variances of two samples independent from each other each of them containing iid Normal variables

   $$\frac{S^2_X \sigma^2_X}{S^2_Y \sigma^2_Y} \sim F_{n_X - 1, n_Y - 1}$$

- the ratio of the sum of squared error to the sum of squares of the regression for a linear regression model with k independent variables and iid normal error term

$$\frac{SSR/k}{SSE/(n-k-1)} \sim F_{k,n-k-1}$$

3. Random variables with distribution Student:
- the standardized sample mean with *sample standard deviation*, iid Normal variables

$$T(X) = \frac{\bar{X} - \mu_0}{s/\sqrt{n}} \sim t_{n-1}$$

- the standardized difference in sample means with the *pooled estimator of the standard deviation*, iid Normal samples independent from each other

$$T(X,Y) = \frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{s_{pool}/\sqrt{1/n_X + 1/n_Y}} \sim t_{n_X+n_Y-2}$$

- standardized estimates of the coefficient in a multiple regression with *sample standard deviation* of the error term, iid errors Normally distributed

$$T_j^* = \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}/\sqrt{c_{jj}}} \sim t_{n-k-1}$$

- standardized estimate of the intercept coefficient in a simple regression with *sample standard deviation* of the error term, iid errors Normally distributed

$$T_0^* = \frac{\hat{\beta}_0 - \beta_0}{\hat{\sigma}/\sqrt{\bar{x}^2 s_{XX}^{-1}}} \sim t_{n-2}$$

- standardized estimate of the slope coefficient in a simple regression with *sample standard deviation* of the error term, iid errors Normally distributed

$$T_1^* = \frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}/\sqrt{s_{XX}^{-1}}} \sim t_{n-2}$$

4. When we replace the sample standard deviation denoted by $\hat{\sigma}$ or s with the standard deviation of the population σ we obtain Normal distributions. For the sample mean we have

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1).$$

For the difference in the sample means we have

$$\frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{\sigma/\sqrt{1/n_X + 1/n_Y}} \sim N(0,1), \qquad \frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{\sqrt{\sigma_X^2/n_X + \sigma_Y^2/n_Y}} \sim N(0,1)$$

in case of equal or different variances respectively. For the estimator of a coefficient of a linear regression

$$T_j^* = \frac{\hat{\beta}_j - \beta_j}{\sigma/\sqrt{c_{jj}}} \sim N(0,1).$$

5.

$$y_i = \beta_0 + \beta_{male}\mathbb{1}(x_{sex,i} = male) + \beta_{school}x_{school,i} + \beta_{exper}x_{exper,i} + \varepsilon_i$$

where $x_{sex,i}$ is the sex of person i, $x_{school,i}$ is the number of years of person i, $x_{exper,i}$ is the number of years of experience of person i and y_i is the wage of person i.

6. We test $H_0 : \beta_{male} = \beta_{school} = \beta_{exper} = 0$ versus $H_a :$ at least one is different than zero. The value of the test statistic is given by 167.63, it is compared to $f_{3,3290;0.95} = 2.6$. Since it is larger than 2.6, we reject the null hypothesis in favor of the alternative hypothesis. The value $f_{k,m;p}$ is such that $P(F_{k,m} \leq f_{k,m;p}) = p$, where $F_{k,m}$ has Fisher distribution with k, m df in the numerator and the denominator respectively.

7. The coefficient on sex suggests that men earn more than woman; the coefficient on school suggests that one year additional schooling increases the salary by 0.64 keeping all other things constant and similar for the coefficient on experience: it increases the salary by 0.12 units. The results for school and experience make sense.

8.

$$\hat{y}_i = x_i^T\hat{\beta} = -3.38 + 1.34 \times 0 + 0.64 \times 17 + 0.12 \times 10 = 15.46$$

9. We test $H_0 : \beta_{male} = 0, H_a : \beta_{male} \neq 0$. The test statistic is computed in the table as t-ratio $\hat{\beta}_{male}/(\hat{\sigma}\sqrt{c_{male}}) = 12.18$ compared to the critical value of $t_{3294-3-1,0.975} = 1.96$ we see that it is larger than it, hence we reject the null hypothesis in favor of the alternative. Indeed the coefficient is different than zero. The quantity $t_{k,p}$ is such that $P(T_k \leq t_{k,p}) = p$ if T_k has t-distribution with parameter k.

   We test $H_0 : \beta_{male} = 1, H_a : \beta_{male} > 1$. The test statistic is $(\hat{\beta}_{male} - 1)/(\hat{\sigma}\sqrt{c_{male}}) = (1.34 - 1)/0.11 = 3.1$ compared to the critical value of $t_{3294-3-1,0.95} = 1.645$ we see that it is larger than it, hence we reject the null hypothesis in favor of the alternative. Indeed the coefficient is larger than one. Note that in the alternative hypothesis we put the statement that we want to test, that is, we put the statement $\beta_{male} > 1$. Then we need to check whether the value of the test statistic *exceeds* the critical value in order to decide whether to reject the Null hypothesis in favor of the alternative.

Another way of answering the same questions without doing tests is to derive confidence intervals and check if the values 0 or 1 are in it. So we have

$$\begin{aligned}\beta_{male} &\in [\hat{\beta}_{male} - t_{3290,0.975}\hat{\sigma}\sqrt{c_{male}}, \quad \hat{\beta}_{male} + t_{3290,0.975}\hat{\sigma}\sqrt{c_{male}}]\\ &= [1.34 - 1.96 \times 0.11, 1.34 + 1.96 \times 0.11]\\ &= [1.12, 1.56]\end{aligned}$$

Since 0 is not in the confidence interval, β_{male} is statistically different than 0. The confidence interval may also be used in order to check $\beta_{male} > 1$ but it would correspond to a test of level $\alpha = 0.025$. This stems from the fact that, in that situation, only the lower bound of the interval matters. Since 1 is not in the confidence interval, we can also conclude that $\beta_{male} > 1$.

10.

$$\begin{aligned}\beta_{exper} &\in [\hat{\beta}_{exper} - t_{3290,0.975}\hat{\sigma}\sqrt{c_{jj}}, \quad \hat{\beta}_{exper} + t_{3290,0.975}\hat{\sigma}\sqrt{c_{jj}}]\\ &= [0.12 - 1.96 \times 0.02, 0.12 + 1.96 \times 0.02]\\ &= [0.0808, 0.1592]\end{aligned}$$

Where $t_{k,p}$ is such that $P(T_k \leq t_{k,p}) = p$ if T_k has t- distribution with parameter k. Zero does not belong to the interval, which means that the effect of experience on wage is statistically different than zero.

11.

$$\begin{aligned}SSE &= (n-k-1)\hat{\sigma}^2 = (3294 - 3 - 1)3.0462^2 = 30529.01\\ SST &= SSE/(1-R^2) = 30529.01/(1-0.1326) = 35196\\ SSR &= SST - SSE = 4667\end{aligned}$$

Exercise 8

a) From the course:

$$\hat{\beta_1}^{LS} = \frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^n (x_i - \overline{x})^2}$$

$\Leftrightarrow$

$$\hat{\beta_1}^{LS} = \frac{\sum_{i=1}^n (x_i - \overline{x})y_i - \overline{y}(\sum_{i=1}^n x_i - n\overline{x})}{\sum_{i=1}^n (x_i - \overline{x})^2}$$

$\Leftrightarrow$

$$\hat{\beta_1}^{LS} = \frac{\sum_{i=1}^n (x_i - \overline{x})y_i - \overline{y}(n\overline{x} - n\overline{x})}{\sum_{i=1}^n (x_i - \overline{x})^2}$$

Therefore:

$$\hat{\beta_1}^{LS} = \sum_{i=1}^{n} \frac{(x_i - \overline{x})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} y_i$$

If we put $c_i := \frac{(x_i - \overline{x})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$, then:

$$\hat{\beta_1}^{LS} = \sum_{i=1}^{n} c_i y_i.$$

Hence, $\hat{\beta_1}^{LS}$ is linear at Y.

b) Let $\hat{\beta_1}$ be a linear unbiased estimator of β_1.

Which means that there exists $(a_i)_{1 \le i \le n} \in \mathbb{R}$ such as: $\hat{\beta_1} = \sum_{i=1}^{n} a_i y_i$ and $\mathbb{E}(\hat{\beta_1}) = \beta_1$.

Since $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, then:

$$\hat{\beta_1} = \beta_0 \sum_{i=1}^{n} a_i + \beta_1 \sum_{i=1}^{n} a_i x_i + \sum_{i=1}^{n} a_i \epsilon_i$$

Since β_1 is unbiased, we obtain the following equation:

$$\beta_1 = \beta_0 \sum_{i=1}^{n} a_i + \beta_1 \sum_{i=1}^{n} a_i x_i$$

This leads us to conclude that: $\sum_{i=1}^{n} a_i = 0$ and $\sum_{i=1}^{n} a_i x_i = 1$.

Now it remains to show that $\hat{\beta_1}$ is more volatile than $\hat{\beta_1}^{LS}$, i.e. $Var(\hat{\beta_1}) \ge Var(\hat{\beta_1}^{LS})$.

Obviously: $Var(\hat{\beta_1}) = Var(\hat{\beta_1} - \hat{\beta_1}^{LS}) + Var(\hat{\beta_1}^{LS}) + 2Cov(\hat{\beta_1} - \hat{\beta_1}^{LS}, \hat{\beta_1}^{LS})$.

Using the bi-linearity of the covariance and the fact that $(y_i)_{1 \le i \le n}$ are iid $\sim N(\beta_0 + \beta_1 x_i, \sigma^2)$, we have:

$$Cov(\hat{\beta_1} - \hat{\beta_1}^{LS}, \hat{\beta_1}^{LS}) = Cov(\hat{\beta_1}, \hat{\beta_1}^{LS}) - Var(\hat{\beta_1}^{LS}) = \sum_{i=1}^{n} c_i a_i \sigma^2 - Var(\hat{\beta_1}^{LS})$$

$\Leftrightarrow$

$$Cov(\hat{\beta_1} - \hat{\beta_1}^{LS}, \hat{\beta_1}^{LS}) = \sum_{i=1}^{n} \frac{(x_i - \overline{x})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} a_i \sigma^2 - \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

If we remind that $\sum_{i=1}^{n} a_i = 0$ and $\sum_{i=1}^{n} a_i x_i = 1$, we find directly that $Cov(\hat{\beta_1} - \hat{\beta_1}^{LS}, \hat{\beta_1}^{LS}) = 0$.

Therefore, $Var(\hat{\beta_1}) = Var(\hat{\beta_1} - \hat{\beta_1}^{LS}) + Var(\hat{\beta_1}^{LS})$, and then:

$$Var(\hat{\beta_1}) \ge Var(\hat{\beta_1}^{LS}).$$