Universität Bielefeld
Tanja Pasurek
Ehsan Abedi WiSe 2023/24

Foundations of Statistics

Solutions to Homework 7

Exercise 1 (Order statistic, part I). Let $X_1, X_2, ..., X_n$ be i.i.d. real-valued random variables with CDF $x \mapsto F(x) \in [0,1]$. Let us consider their maxium and minimum:

$$Y := \max\{X_1, ..., X_n\}, \quad Z := \min\{X_1, ..., X_n\}.$$

The random variables Y and Z are called *largest order statistic* and *smallest order statistic*, respectively.

a Prove that the distribution function of Y is given by

$$F_Y(y) = F(y)^n \quad \forall y \in \mathbb{R}.$$

If, in particular, F has density function f, find the density function f_Y of the random variable Y.

Solution: Observe that

$$\begin{aligned} F_Y(y) &= \mathbb{P}(Y \leq y) \\ &= \mathbb{P}(X_1 \leq y, X_2 \leq y, \ldots, X_n \leq y) \\ &= \mathbb{P}(X_1 \leq y)\mathbb{P}(X_2 \leq y) \ldots \mathbb{P}(X_n \leq y) \qquad \text{(by independence)} \\ &= \boxed{F(y)^n}, \end{aligned}$$

and if F has density function f, then we obtain

$$\begin{aligned} f_Y(y) &= \frac{d}{dy} F_Y(y) \\ &= \frac{d}{dy} F(y)^n \\ &= nF(y)^{n-1} \frac{d}{dy} F(y) \\ &= \boxed{nF(y)^{n-1} f(y)}. \end{aligned}$$

b Prove that the distribution function of Z is given by

$$F_Z(z) = 1 - [1 - F(z)]^n \quad \forall z \in \mathbb{R}.$$

If, in particular, F has density function f, find the density function f_Z of the random variable Z.

Solution: Observe that

$$\begin{aligned}
F_Z(z) &= \mathbb{P}(Z \leq z) \\
&= 1 - \mathbb{P}(Z > z) \\
&= 1 - \mathbb{P}(X_1 > z, X_2 > z, \ldots, X_n > z) \\
&= 1 - \mathbb{P}(X_1 > z)\mathbb{P}(X_2 > z) \ldots \mathbb{P}(X_n > z) \qquad \text{(by independence)} \\
&= 1 - \big[1 - \mathbb{P}(X_1 \leq z)\big]\big[1 - \mathbb{P}(X_2 \leq z)\big] \ldots \big[1 - \mathbb{P}(X_n \leq z)\big] \\
&= \boxed{1 - [1 - F(z)]^n},
\end{aligned}$$

and if F has density function f, then we obtain

$$\begin{aligned}
f_Z(z) &= \frac{d}{dz} F_Z(z) \\
&= \frac{d}{dz}\Big(1 - [1 - F(z)]^n\Big) \\
&= n[1 - F(z)]^{n-1} \frac{d}{dz} F(z) \\
&= \boxed{n[1 - F(z)]^{n-1} f(z)}.
\end{aligned}$$

c Find the joint CDF of the random vector $\mathbf{U} := (Z, Y)^\top$.

Solution: We aim to find

$$F_U(z, y) := \mathbb{P}(Z \leq z, Y \leq y)$$

for different values of $y \in \mathbb{R}$ and $z \in \mathbb{R}$. Let us write

$$\mathbb{P}(Z \leq z, Y \leq y) = \mathbb{P}(\underbrace{\{Z \leq z\}}_{=:A} \cap \underbrace{\{Y \leq y\}}_{=:B}) \tag{1}$$

Recall that $A^C \cap B$ and $A \cap B$ are disjoint and their union forms the entire set B, that is $(A^C \cap B) \cup (A \cap B) = B$, and we also have

$$\mathbb{P}(A \cap B) = \mathbb{P}(B) - \mathbb{P}(A^C \cap B).$$

Applying this to (1), we obtain

$$\begin{aligned}\mathbb{P}(Z \le z, Y \le y) &= \mathbb{P}(Z \le z \cap Y \le y) \\ &= \mathbb{P}(Y \le y) - \mathbb{P}(Z > z \cap Y \le y) \\ &= \;\; F_Y(y) \quad - \mathbb{P}(Z > z \cap Y \le y) \\ &= \;\; F(y)^n \quad - \mathbb{P}(Z > z \cap Y \le y).\end{aligned}$$

It remains to find the second term. Note that we always have $Z \le Y$. Therefore if $z < y$, we have

$$\begin{aligned}\mathbb{P}(Z > z, Y \le y) &= \mathbb{P}(z < Z \le Y \le y) \\ &= \mathbb{P}(z < X_1, X_2, \dots, X_n \le y) \\ &= \mathbb{P}(z < X_1 \le y, z < X_2 \le y, \dots, z < X_n \le y) \\ &= \mathbb{P}(z < X_1 \le y)\mathbb{P}(z < X_2 \le y) \dots \mathbb{P}(z < X_n \le y) \\ &= [F(y) - F(z)]^n,\end{aligned}$$

and obviously if $z \ge y$, the quantity above is 0. All in all, we have

$$F_U(z, y) := \mathbb{P}(Z \le z, Y \le y) = \begin{cases} F(y)^n - [F(y) - F(z)]^n & \text{if } z < y \\ F(y)^n & \text{if } z \ge y \end{cases}.$$

d If, in particular, F has density function f, find the joint density function of $\mathbf{U}$. Are Z and Y independent?

Solution: To obtain the joint density function, we need to take derivative of the joint CDF:

$$\begin{aligned}f_U(z, y) &= \frac{\partial^2}{\partial z \partial y} F_U(z, y) \\ &= \begin{cases} n(n-1)[F(y) - F(z)]^{n-2} f(y) f(z) & \text{if } z < y \\ 0 & \text{if } z \ge y \end{cases}.\end{aligned}$$

We observe that

$$f_{Z,Y}(z, y) \ne f_Z(z) f_Y(y) = n^2 [1 - F(z)]^{n-1} F(y)^{n-1} f(y) f(z)$$

Therefore, we conclude that Z and Y are not independent (which is expected due to the relation $Y \ge Z$).

Exercise 2 (Order statistic, part I, examples).

a Let $U, V \sim \text{Unif}(0,1)$ be independent. Based on the previous exercise, find density function of $\max\{U,V\}$ and $\min\{U,V\}$. Compare your result with a simulation in R. Generate random samples from the uniform distribution, and for each pair, record both the maximum and minimum values. Finally, plot the histogram of these values.

Solution: Denoting by $Y := \max\{U,V\}$, we obtain

$$F_Y(y) = \begin{cases} 0 & y \le 0 \\ y^2 & 0 < y < 1, \\ 1 & y \ge 1. \end{cases} \quad \Rightarrow \quad f_Y(y) = \begin{cases} 0 & y \le 0 \\ 2y & 0 < y < 1, \\ 0 & y \ge 1. \end{cases}$$

Denoting by $Z := \min\{U,V\}$, we obtain

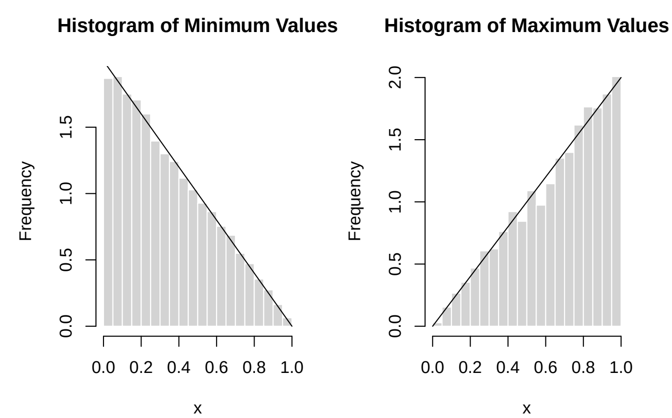
$$F_Z(z) = \begin{cases} 0 & z \le 0 \\ 1-(1-z)^2 & 0 < z < 1, \\ 1 & z \ge 1. \end{cases} \quad \Rightarrow \quad f_Z(z) = \begin{cases} 0 & z \le 0 \\ 2(1-z) & 0 < z < 1, \\ 0 & z \ge 1. \end{cases}$$

```
n <- 10^4
U <- runif(n, 0, 1)
V <- runif(n, 0, 1)
Y <- pmax(U, V) # parallel Maxima
Z <- pmin(U, V) # parallel Minima

par(mfrow=c(1,2))

hist(Z, freq = FALSE, main = 'Histogram of Minimum Values',
     xlab = 'x', ylab = 'Frequency', border = 'white')
curve(2*(1-x), from = 0, to = 1,add = TRUE)

hist(Y, freq = FALSE, main = 'Histogram of Maximum Values',
     xlab = 'x', ylab = 'Frequency', border = 'white')
curve(2*x, from = 0, to = 1,add = TRUE)
```



b Let $U, V \sim \text{Unif}(0,1)$ be independent and $p \in (0,1)$ be a constant. In HW4, Exercise 1, we studied indicator random variables and showed that $\mathbb{I}_{\{U \le p\}} \sim \text{Ber}(p)$. Now, use order statistic to find the distribution of the random variables $\mathbb{I}_{\{U \le p\}}\mathbb{I}_{\{V \le p\}}$ and $\mathbb{I}_{\{U \le p\}} + \mathbb{I}_{\{V \le p\}}$.

Solution:

- We have

$$\mathbb{I}_{\{U \le p\}}\mathbb{I}_{\{V \le p\}} = \mathbb{I}_{\{U \le p\} \cap \{V \le p\}} = \mathbb{I}_{\{\max\{U,V\} \le p\}},$$

which only takes two values $\{0,1\}$. By previous task (a), we have $\mathbb{P}(\max\{U,V\} \le p) = p^2$. Thus,

$$\mathbb{I}_{\{U \le p\}}\mathbb{I}_{\{V \le p\}} \sim \text{Ber}(p^2).$$

- The random variable $W := \mathbb{I}_{\{U \le p\}} + \mathbb{I}_{\{V \le p\}}$ is a sum of two independent Bernoulli random variables. Thus,

$$\mathbb{I}_{\{U \le p\}} + \mathbb{I}_{\{V \le p\}} \sim \text{Binom}(2, p).$$

c Let $X_1, \cdots X_n \sim \text{Unif}(a,b)$ be i.i.d random variables. Find CDF F_Y and F_Z of the largest and smallest order statistic Y and Z, respectively. Do random variables Y and Z converge in distribution as $n \to \infty$?

Solution: For uniform distribution Unif(a,b), the density function f and CDF F are given by

$$f(x) = \begin{cases} \frac{1}{b-a} & a \le x \le b, \\ 0 & \text{otherwise.} \end{cases} \qquad F(x) = \begin{cases} 0 & x \le a \\ \frac{x-a}{b-a} & a < x < b, \\ 1 & x \ge b. \end{cases}$$

respectively.

- Using part (a) for the largest order statistic Y, we get

$$F_Y^{(n)}(y) = \begin{cases} 0 & y \le a \\ F(y)^n = (\frac{y-a}{b-a})^n & a < y < b, \\ 1 & y \ge b. \end{cases}$$

Taking limit yields

$$\lim_{n \to \infty} F_Y^{(n)}(y) = \begin{cases} 0 & y \le a \\ 0 & a < y < b, \\ 1 & y \ge b, \end{cases} = \begin{cases} 0 & y < b, \\ 1 & y \ge b. \end{cases}$$

The limit function is continuous at all point except $y = b$.

The right-hand side is cumulative distribution function of the atomic measure δ_b. So we have convergence of CDFs and we conclude that

$$Y \xrightarrow{d} b \quad \Rightarrow \quad Y \xrightarrow{\mathbb{P}} b.$$

• Using part (b) for the smallest order statistic Z, we get

$$F_Z^{(n)}(z) = \begin{cases} 0 & z \leq a \\ 1 - [1 - F(z)]^n = 1 - (\frac{b-z}{b-a})^n & a < z < b, \\ 1 & z \geq b. \end{cases}$$

Taking limit yields

$$\lim_{n\to\infty} F_Z^{(n)}(z) = \begin{cases} 0 & z \leq a \\ 1 & a < z < b, \\ 1 & z \geq b. \end{cases} = \begin{cases} 0 & z \leq a \\ 1 & z > a \end{cases}$$

Note that the right-hand side is not cumulative distribution function of the atomic measure δ_a. However, ignoring discontinuity points (here $z = a$), we can say that $F_Z^{(n)}$ converges to cumulative distribution function of the atomic measure δ_a. We conclude

$$Z \xrightarrow{d} a \quad \Rightarrow \quad Z \xrightarrow{\mathbb{P}} a.$$

d Let $X_1, \cdots X_n \sim \exp(\lambda)$ be i.i.d random variables. Find CDF F_Y and F_Z of the largest and smallest order statistic Y and Z, respectively. Do random variables Y and Z converge in distribution as $n \to \infty$?

Solution: Recall that the support of Exponential distribution is $[0, \infty)$.
• Using part (a) for the largest order statistic Y, we get

$$F_Y^{(n)}(y) = \begin{cases} 0 & y < 0 \\ F(y)^n = (1 - \exp(-\lambda y))^n & y \geq 0 \end{cases}$$

Now observe that for fixed $y > 0$

$$\lim_{n\to\infty} F_Y^{(n)}(y) = \lim_{n\to\infty} (1 - \exp(-\lambda y))^n = 0.$$

So we have

$$\lim_{n\to\infty} F_Y^{(n)}(y) = \begin{cases} 0 & \text{if } y > 0 \\ 0 & \text{if } y = 0 \\ 0 & \text{if } y < 0 \end{cases}.$$

However, the right hand side is not any cumulative distribution function. Therefore, even though we have pointwise convergence of $F_Y^{(n)}(y)$, i.e., for any $y \in \mathbb{R}$, the random variable Y does not converge in distribution (or weakly). (*Technical remark: the family of distributions corresponding to $(F_Y^{(n)}(y))_{n\in\mathbb{N}}$ is not *tight*. They "escape to infinity.")

- Using part (b) for the smallest order statistic Z, we get for $z \geq 0$

$$\begin{aligned} F_Z^{(n)}(z) &= 1 - [1 - F(z)]^n \\ &= 1 - [1 - (1 - \exp(-\lambda z))]^n \\ &= 1 - [\exp(-\lambda z)]^n \\ &= 1 - \exp(-n\lambda z) \end{aligned}$$

and for $z < 0$, we obtain $F_Z^{(n)}(z) = 0$ becasue $F(z) = 0$. This shows that

$$Z \sim \exp(n\lambda).$$

Observe that

$$\lim_{n\to\infty} F_Z^{(n)}(z) = \begin{cases} \lim_{n\to\infty} 1 - \exp(-n\lambda z) = 1 & \text{if } z > 0 \\ 0 & \text{if } z = 0 \\ 0 & \text{if } z < 0 \end{cases}.$$

Note that the right-hand side is not cumulative distribution function of the atomic measure δ_0 (it is not right continuous). But again ignoring discontinuity points (here $z = 0$), we can say that

$$Z \xrightarrow{d} 0 \quad \Rightarrow \quad Z \xrightarrow{\mathbb{P}} 0.$$

Exercise 3 (Sample skewness and sample kurtosis).

a Show with Chebyshev's inequality that for any random variable (with finite non-zero variance) not more than about 11% of the data can be more than three standard deviations away from the mean.

Solution: Applying Chebyshev's inequality for a random variable X with $\mu := \mathbb{E}[X]$ and $\sigma^2 := \mathrm{Var}(X) \in (0, \infty)$, we have for any $a > 0$

$$\mathbb{P}(|X - \mu| \geq a) \leq \frac{\sigma^2}{a^2}.$$

Setting $a = k\sigma$, we obtain

$$\mathbb{P}(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}.$$

In particular for $k = 3$, we obtain

$$\mathbb{P}(|X - \mu| \geq 3\sigma) \leq \frac{1}{3^2} = \frac{1}{9} \approx 11.11\%.$$

b Show that for a $N(\mu, \sigma^2)$-distributed random variable the proportion calculated in a) is now about 0.27%.

Solution: We have

$$\begin{aligned}\mathbb{P}(|X - \mu| \geq 3\sigma) &= 1 - \mathbb{P}(|X - \mu| \leq 3\sigma) \\ &= 1 - \mathbb{P}(-3 \leq \frac{X - \mu}{\sigma} \leq 3) \\ &= 1 - \Big(\Phi_{0,1}(3) - \Phi_{0,1}(-3)\Big) \\ &\approx 0.27\%\end{aligned}$$

where this value can be found using `R`:

```
(1- (pnorm(3) - pnorm(-3)))*100
[1] 0.2699796
```

c For a sample $x_1, ..., x_n$ the z-score is defined by

$$z_i := \frac{1}{\tilde{s}}(x_i - \bar{x}), \quad i = 1, ..., n.$$

Here $\bar{x}$ and $\tilde{s}$ are the sample mean and standard deviation (with denominator n not $n - 1$), respectively. Explain what $z_i = 3$ means.

Solution: $z_i = 3$ simply means that i-th datapoint is 3 *sample standard deviation* $\tilde{s}$ above from the *sample mean* $\bar{x}$.

d Install the package `UsingR` with the commands `install.packages ("UsingR")` and `require("UsingR")`. The dataset `exec.pay` contains direct compensation data for 199 United States CEOs. Compare the mean, median and quantiles by using the function `summary(exec.pay)`. Draw the boxplot and determine the outliers.

Solution: Recall that a value x of the dataset is called an outlier if

$$x < x_{0.25} - 1.5 \times \mathrm{IQR} \quad \text{or} \quad x > x_{0.75} + 1.5 \times \mathrm{IQR},$$

where x_α is the α-quantiles and IQR is the interquartile range:

$$IQR := x_{0.75} - x_{0.25}$$

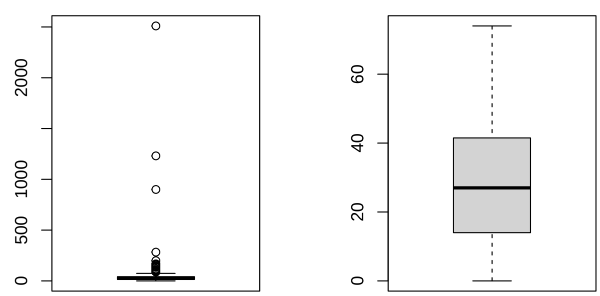
```
library(UsingR);
summary(exec.pay)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   14.00   27.00   59.89   41.50 2510.00
```

```
par(mfrow=c(1,2))

# box-plot of all data
boxplot(exec.pay)

# box-plot without outliers
boxplot(exec.pay, outline = FALSE)
```




```
# find the numbers with R function boxplot.stats
length(boxplot.stats(exec.pay)$out)
```

```
## [1] 24
```

```
# find outliers using direct computation
uq = quantile(exec.pay, p=0.75) + 1.5 * IQR(exec.pay)
uq
```

```
##   75%
## 82.75
```

```
lq = quantile(exec.pay, p=0.25) - 1.5 * IQR(exec.pay)
lq
```

```
##    25%
## -27.25
```

```r
# upper outliers
which(exec.pay > uq)
```

```
##  [1]   1  13  26  27  30  43  45  50  60  63  64  68  70  93  97  99 116 120 131
## [20] 136 149 185 189 190
```

```r
sum(exec.pay > uq)
```

```
## [1] 24
```

```r
# lower outliers
which(exec.pay < lq)
```

```
## integer(0)
```

```r
sum(exec.pay < lq)
```

```
## [1] 0
```

```r
# there are only upper outliers (all numbers bigger than 82.75)
```

In this dataset, the sample Mean is much higher than the sample Median. The distribution has positive skew: The right tail is longer.

e Calculate with R the z-score of the data to find out what proportion of the data are more than 3 standard deviations from the mean. Compare your result with the results in a) and b).

Solution:

```r
s_til <- sd(exec.pay)*sqrt ((n-1)/n)
z     <- (exec.pay - x_bar)/s_til
sum(abs(z)>3)
```

```
## [1] 3
```

```r
(sum(abs(z)>3) / n ) * 100
```

```
## [1] 1.507538
```

```r
# Only 1.5 % of this dataset is more than 3 standard deviations from the mean.
# This is certainly less than 11.11 % we obtained using Chebyshev's inequality.
# But it is more than 0.27 % corresponding to the normal distribution
```

f The *sample skewness* is defined by

$$\sqrt{n} \cdot \frac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{\left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^{3/2}}.$$

Show that this is equal to

$$\frac{1}{n}\sum_{i=1}^{n} z_i^3.$$

Solution: We start from the right-hand side and write:

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} z_i^3 &= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{\tilde{s}}\right)^3 \\
&= \frac{1}{n}\frac{1}{\tilde{s}^3}\sum_{i=1}^{n}(x_i-\bar{x})^3 \\
&= \frac{1}{n}\frac{1}{\left(\sqrt{\frac{1}{n}\sum_i (x_i-\bar{x})^2}\right)^3}\sum_{i=1}^{n}(x_i-\bar{x})^3 \\
&= \sqrt{n}\frac{\sum_{i=1}^{n}(x_i-\bar{x})^3}{\left(\sum_i (x_i-\bar{x})^2\right)^{3/2}}
\end{aligned}$$

g Calculate the sample skewness of the `exec.pay` dataset.

Solution: Continuing the code of task (e), we have

```
> mean (z**3)
[1] 9.651199
> # check with skewness() function in "moments"
> library(moments)
> skewness(exec.pay)
[1] 9.651199
```

h The *sample kurtosis* is the measure of the tails in a data set. Long tails will lead to larger values, while "normal" data will have kurtosis close to 0. It is defined by the formula

$$n \cdot \frac{\sum_{i=1}^{n}(x_i-\bar{x})^4}{\left[\sum_{i=1}^{n}(x_i-\bar{x})^2\right]^2} - 3.$$

Show that this is equal to

$$\frac{1}{n}\sum_{i=1}^{n} z_i^4 - 3.$$

Guess why we are taking out number 3 here.

Solution: We have

$$
\begin{aligned}
n \cdot \frac{\sum_{i=1}^{n}(x_i - \bar{x})^4}{\left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^2} &= \frac{\frac{1}{n}}{\frac{1}{n^2}} \frac{\sum_{i=1}^{n}(x_i - \bar{x})^4}{\left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^2} \\
&= \frac{1}{n} \frac{\sum_{i=1}^{n}(x_i - \bar{x})^4}{\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^2} \\
&= \frac{1}{n} \sum_{i=1}^{n} \frac{(x_i - \bar{x})^4}{\tilde{s}^4} \\
&= \frac{1}{n} \sum_{i=1}^{n} z_i^4
\end{aligned}
$$

If $X_1, X_2, \cdots \sim N(\mu, \sigma^2)$ are i.i.d random variables, by a law of large number,

$$
\frac{1}{n}\sum_{i=1}^{n}\left(\frac{X_i - \mu}{\sigma}\right)^4 \quad \xrightarrow{\mathbb{P}} \quad \mathbb{E}\left[\left(\frac{X - \mu}{\sigma}\right)^4\right] = 3
$$

Therefore, we subtract 3 becasue we want that the sample kurtosis of a sample from normal distribution to be close to 0, especially when n is large. Also, observe that:

```
> mean ((rnorm(10**7))**4)
[1] 3.002529
```

i Calculate the kurtosis of the `exec.pay` dataset.

Solution: Continuing the code of task (e), we have

```
> mean (z**4) -3
[1] 103.128
> # check with kurtosis() function in "moments"
> library(moments)
> kurtosis(exec.pay) - 3
[1] 103.128
```

Exercise 4. Suppose we have a computer program consisting of $n = 100$ pages of code. Let X_i be the number of errors on the i^{th} page of code. Suppose that the X_i's are Poisson with mean 1 and that they are independent. Let $Y = \sum_{i=1}^{n} X_i$ be the total number of errors. Use the Central Limit Theorem to approximate $\mathbb{P}(Y \leq 90)$. Check your answer with the exact value of $\mathbb{P}(Y \leq 90)$. (*Hint:* recall HW5, Exercise 4(c)).

Solution: For all $i = 1, \cdots, n = 100$, we have $X_i \sim Pois(\lambda = 1)$, whose mean and variance is given by $\mu = \lambda$ and $\sigma^2 = \lambda$. Let us define $\bar{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i$. Based on CLT for the standardized sum

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

Therefore, the distribution of Y can be approximated by normal distribution

$$Y = n\bar{X}_n \overset{\text{approx}}{\sim} N(n\mu, n\sigma^2) = N(n\lambda, n\lambda).$$

Recall HW 5 Ex.1 (b) that we showed that for $Z \sim N(\mu_*, \sigma_*^2)$ we have $\mathbb{P}(Z \leq b) = \Phi_{0,1}\left(\frac{b-\mu_*}{\sigma_*}\right)$, where $\Phi_{0,1}$ is CDF of standard normal distribution $N(0,1)$. Therefore, we get

$$\mathbb{P}(Y \leq 90) \approx \Phi_{0,1}\left(\frac{90 - n\lambda}{\sqrt{n\lambda}}\right) = \Phi_{0,1}(-1) = 0.15865.$$

Now, let us compute the exact value of $\mathbb{P}(Y \leq 90)$. In HW5, Exercise 4(c), we observed that the distribution of the sum of two independent Poisson random variables is also Poisson-distributed, with the rate being the sum of the individual rates. By induction, one obtains

$$Y \sim \text{Pois}(n\lambda).$$

Therefore, using the CDF in `R`, we obtain

```
ppois(90, lambda = 100)
[1] 0.1713851
```

$$\mathbb{P}(Y \leq 90) = 0.1713851.$$

Exercise 5. An accountant wants to simplify his bookkeeping by rounding amounts to the nearest integer, for example, rounding €99.53 and €100.46 both to €100. What is the cumulative effect of this if there are, say, 100 amounts? To study this we model the rounding errors by $n = 100$ independent Unif$(-0.5, 0.5)$ random variables $X_1, ..., X_{100}$.

a Compute the expectation and the variance of each X_i.

Solution: Recall the formulas for $X_i \sim \text{Unif}[a, b]$

$$\mathbb{E}[X_i] = \frac{a+b}{2} = \frac{-0.5+0.5}{2} = 0$$
$$\text{Var}(X_i) = \frac{(b-a)^2}{12} = \frac{(0.5-(-0.5))^2}{12} = \frac{1}{12}$$

b Use Chebyshev's inequality to compute an upper bound for the probability $\mathbb{P}(|X_1 + X_2 + ... + X_{100}| > 10)$ that the cumulative rounding error $X_1 + X_2 + \cdots + X_{100}$ exceeds €10.

Solution: Define
$$\bar{X} := X_1 + X_2 + ... + X_{100}.$$

By linearity of expectation

$$\mathbb{E}[\bar{X}] = 100\,\mathbb{E}[X_i] = 0.$$

By independence

$$\text{Var}(\bar{X}) = 100\,\text{Var}(X_i) = \frac{100}{12}.$$

Note that

$$\begin{aligned}
&\mathbb{P}\Big(|X_1 + X_2 + ... + X_{100}| > 10\Big) \\
&\qquad = \mathbb{P}\Big(|\bar{X}| > 10\Big) \\
&\qquad = \mathbb{P}\Big(|\bar{X} - \mathbb{E}[\bar{X}]| > 10\Big) \\
&\qquad \leq \mathbb{P}\Big(|\bar{X} - \mathbb{E}[\bar{X}]| \geq 10\Big) \\
&\qquad \leq \frac{\text{Var}(\bar{X})}{10^2} = \frac{1}{100}.\frac{100}{12} = \frac{1}{12}
\end{aligned}$$

c What can you say about the mean of the absolute error $\frac{1}{n}\sum_{i=1}^{n} |X_i|$, applying the Law of Large Numbers?

Solution: To apply LLN, we first need to find $\mathbb{E}(|X_i|)$. Here we present two approaches:

• 1st approach: We first claim that $X_i \overset{\text{i.i.d.}}{\sim} \text{Unif}[-a, a]$ implies $|X_i| \overset{\text{i.i.d.}}{\sim} \text{Unif}[0, a]$ for any $a > 0$. To see this, let $0 \leq x \leq a$ and write

$$\begin{aligned}\mathbb{P}(|X_i| \leq x) &= \mathbb{P}(-x \leq X_i \leq x) \\ &= \frac{2x}{2a} = \frac{x}{a},\end{aligned}$$

which shows that $|X_i| \overset{\text{i.i.d.}}{\sim} \text{Unif}[0, a]$. Then immediately

$$\mathbb{E}(|X_i|) = \frac{0.5 - 0}{2} = 1/4.$$

• 2nd approach: We can also check this analytically, using the PDF $f(x)$ of $X_i \overset{\text{i.i.d.}}{\sim} \text{Unif}[-0.5, 0.5]$:

$$\begin{aligned}\mathbb{E}(|X_i|) &:= \int_{-\infty}^{+\infty} |x| f(x)\, \mathrm{d}x = \int_{-0.5}^{+0.5} |x| \cdot 1\, \mathrm{d}x \\ &= 2\int_0^{+0.5} x\, \mathrm{d}x = 2\, \frac{x^2}{2}\bigg|_0^{0.5} = 2 \cdot \frac{1}{2} \cdot \frac{1}{4} = 1/4.\end{aligned}$$

Note that $f(x) = 1$ if $x \in [-0.5, 0.5]$ and $f(x) = 0$ otherwise.

• Final step: by the weak LLN:

$$\frac{1}{n} \sum\nolimits_{i=1}^{n} |X_i| \xrightarrow{\mathbb{P}} 1/4.$$

Remark: Notice that $\frac{1}{n} \sum_{i=1}^{n} X_i \xrightarrow{\mathbb{P}} 0$ which is not what has been asked in this exercise!