École Polytechnique

Bachelor of Science

Introduction to Statistics

MAA204 - Homework

Authors:
Yubo Cai

Supervisor:
Jean-Marc Bardet

January 15, 2023

Abstract

This is the homework for MAA204-Introduction to Statistics of the Bachelor of Science Program of Ecole Polytechnique, finished by Yubo Cai.

Here is the source code in R markdown [Link] in my Github Repository, and you can find the output HTML Notebook of the graphs as well as the PDF format of the code.

Link: [https://github.com/yubocai-poly/Introduction-to-Statistics]

Contents

1 Exercise 1

Definition

We recall that an exponential random variable X with parameter $\lambda = \frac{1}{m} > 0$ has a density function given by:

$$f_X(x) = \frac{1}{m} e^{-\frac{x}{m}}, \forall x \geq 0$$

Question 1

Compute the mean and the variance of an exponential random variable X with general parameter $\lambda = \frac{1}{m} > 0$.

We have the parameter $\lambda = \frac{1}{m}$, the we have $f_X(x) = \lambda e^{-\lambda x}, \forall x \geq 0$. Then we compute the mean value which is the expectation of X. We have

$$\mathbb{E}[X] = \int_0^\infty x\lambda e^{-\lambda x} dx$$

We substitue $y = \lambda x$ we have

$$\begin{aligned}
\mathbb{E}[X] &= \frac{1}{\lambda} \int_0^\infty y e^{-y} dy \\
&= \frac{1}{\lambda} \left[-e^{-y} - y e^{-y} \right]_0^\infty \\
&= \frac{1}{\lambda}
\end{aligned}$$

Then we can compute the variance of X

$$\begin{aligned}
\mathbb{E}[X^2] &= \int_0^\infty x^2 \lambda e^{-\lambda x} dx \\
&= \frac{1}{\lambda^2} \int_0^\infty y^2 e^{-y} dy \\
&= \frac{1}{\lambda^2} \left[-2e^{-y} - 2ye^{-y} - y^2 e^{-y} \right]_0^\infty \\
&= \frac{2}{\lambda^2}
\end{aligned}$$

Therefore, we obtain

$$Var[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}$$

Then we have $\mathbb{E}[X] = \frac{1}{\lambda} = m$ and $Var[X] = \frac{1}{\lambda^2} = m^2$

Definition

The dataset contained in the file **exponential.csv** is represented by a sample of size $N = 500$ of independent observations $(x_i)_{1\leq i\leq N}$ drawn according to exponential distribution of parameter $\frac{1}{m} \in]0, +\infty[$. As we ignore the value of the parameter m, our goal is to estimate it. For that purpose, we shall study 2 estimators

1. Estimation with the Maximum Likelihood (ML).

2. the estimator $Y_n = n \min (X_1, X_2, \cdots, X_n)$.

Question 2

Maximum Likelihood estimation: compute the likelihood function that we denote $L(\lambda, x_1, x_2, \cdots, x_N)$ and deduce the maximum likelihood parameter $\hat{\lambda} = \frac{1}{\hat{m}_{\text{ML}}}$. What do you notice ?

To find the maximum likelihood estimate of $\lambda = \frac{1}{m}$ given a sample of exponential random variables $x_1, x_2, ..., x_N$, we can compute the likelihood function as follows:

$$L_\lambda(x_1, \ldots, x_N) = \prod_{i=1}^{N} f_\lambda(x_i) = \prod_{i=1}^{N} \lambda e^{-\lambda x_i} = \lambda^N exp(-\lambda \sum_{i=1}^{N} x_i)$$

Then we takie the natural log of the likelihood function

$$\ln(L_\lambda(X)) = \ln(L_\lambda(x_1, \ldots, x_N)) = N\ln(\lambda) - \lambda \sum_{i=1}^{N} x_i$$

Then we take the partial differentiating with respect to λ gives us:

$$\frac{\partial \ln L_\lambda(X)}{\partial\lambda} = \frac{\partial(N\ln(\lambda) - \lambda\sum_{i=1}^{N} x_i)}{\partial\lambda} = \frac{N}{\lambda} - \sum_{i=1}^{N} x_i$$

Since $\hat{\lambda} \in \underset{\lambda}{\operatorname{argmax}} \log(L_\lambda(X_1, \ldots, X_N)) = \underset{\lambda}{\operatorname{argmax}} L_\lambda(X_1, \ldots, X_N)$. In order to find the parameter $\hat{\lambda}$, we have

$$\frac{\partial \ln L_\lambda(X)}{\partial\lambda} = \frac{N}{\lambda} - \sum_{i=1}^{N} x_i = 0 \Rightarrow \hat{\lambda} = \frac{N}{\sum_{i=1}^{N} x_i}$$

From **Question 1** we got $\mathbb{E}[X] = \frac{1}{\lambda} = m$. Then we have that

$$\hat{\lambda} = \frac{1}{\hat{m}_{\text{ML}}}$$

We still need to prove that $\hat{\lambda} = \frac{1}{\hat{m}_{\text{ML}}} = \frac{N}{\sum_{i=1}^{N} x_i}$ is the maximum likelihood parameter, we take the second-order partial derivative

$$\frac{\partial^2 \ln L_\lambda(X)}{\partial\lambda^2} = -\frac{N}{\lambda^2} < 0$$

Then we finish the proof that $\hat{\lambda} = \frac{1}{\hat{m}_{\text{ML}}} = \frac{N}{\sum_{i=1}^{N} x_i}$ is a local and global maximum of the likelihood.

Also, we have $\hat{m}_{\text{ML}} = \frac{\sum_{i=1}^{N} x_i}{N} = \bar{x}_N$ which is the empirical mean of the observation. **Remark:** The result corresponds to the tuition that if N is big enough, $\bar{x}_N$ will be very close to $\mathbb{E}[X_1] = m$

Question 3

Is the estimator $\hat{m}_{\text{ML}}$ biased ? compute its quadratic risk.

From **Question 2** we have $\hat{m}_{\text{ML}} = \bar{x}_n$ where $\bar{x}_n$ is the mean that $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$. Then we have

$$B(\hat{m}_{\text{ML}}) = \mathbb{E}[\hat{m}_{\text{ML}}] - m = \mathbb{E}[\bar{x}_n] - m = \mathbb{E}[x_1] - m = 0$$

Then we prove that the estimator $\hat{m}_{\text{ML}}$ is unbiased. Now we try ot compute quadratic risk of $\hat{m}_{\text{ML}}$, we have

$$R(\hat{m}_{\text{ML}}) = \mathbb{E}[(\hat{m}_{\text{ML}} - m)^2] = B^2(\hat{m}_{\text{ML}}) + Var(\hat{m}_{\text{ML}}) = Var(\hat{m}_{\text{ML}})$$

Since $(X_i)_{1\leq i\leq N}$ are i.i.d.r.v and the estimator $\hat{m}_{\text{ML}}$ is unbiased, we have

$$R(\hat{m}_{\text{ML}}) = Var(\hat{m}_{\text{ML}}) = Var(\bar{x}_n) = \frac{Var(x_1)}{N} = \frac{1}{\lambda^2 N} = \frac{m^2}{N}$$

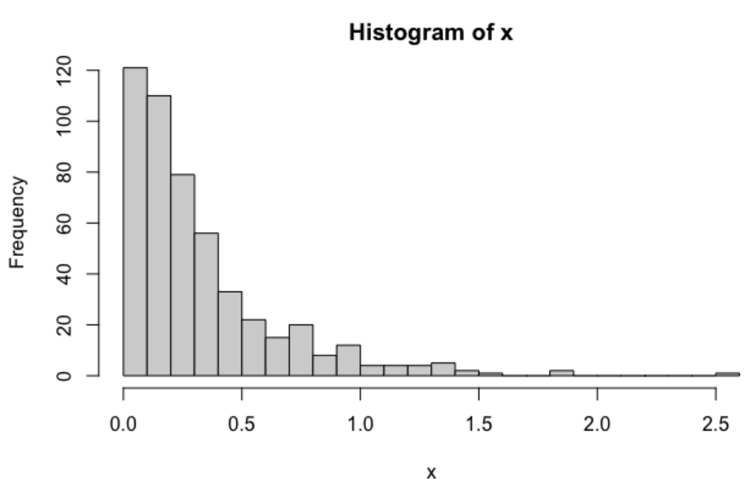
Therefore, the quadratic risk of $\hat{m}_{\text{ML}}$ is $\frac{m^2}{N}$

Question 4

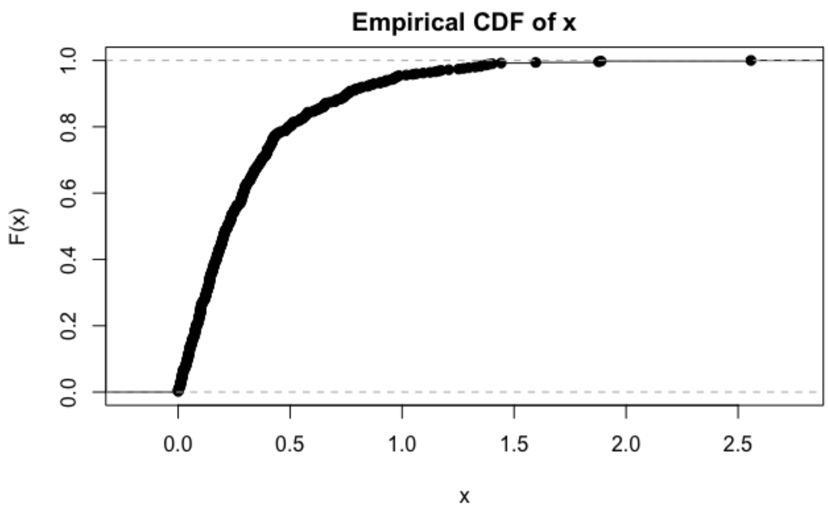
On R, read the file **exponential.csv**. Then compare through a plot the distribution of the dataset observations and the distribution of the exponential random variable with the estimated parameter $\hat{m}_{\text{ML}}$

From the dataset **exponential.csv**, we only have 500 values of observations. So we try to plot the histogram of the frequency for each interval in order to see the distribution since the observation value is only in 1-dimension.

```r
# import exponential.csv
data <- read.csv("exponential.csv")
summary(data)
hist(data$x[2:501], breaks=30, main="Histogram of x", xlab="x", ylab="Frequency")
plot(ecdf(data$x[2:501]), main="Empirical CDF of x", xlab="x", ylab="F(x)")
```



Also we have the Empirical CDF of x with the following graph



And we can use the following R code in order to compare. However, since we can't plot the exact PDF of the dataset, so we compare the CDF of the empirical (dataset) with the theoritical.

```
# m is the mean of the dataset
m <- 0.3256294
# plot the exponential distribution with parameter lambda
x <- seq(0, 3, length.out=1000)
y <- dexp(x, rate = 1/m)
plot(x, y, type="l", main="Exponential Distribution", xlab="x", ylab="f(x)")
```

Here we have the PDF graph of exponential distribution with parameter $\hat{\lambda} = \frac{1}{\hat{m}_{\mathrm{ML}}}$

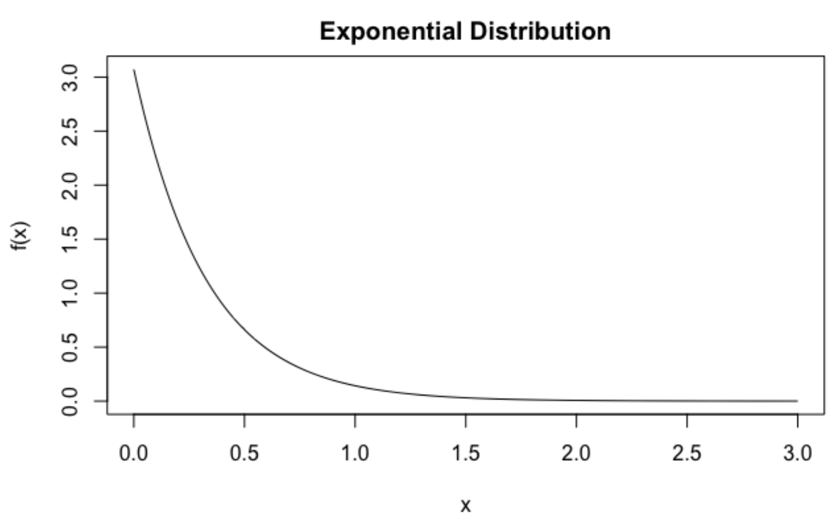
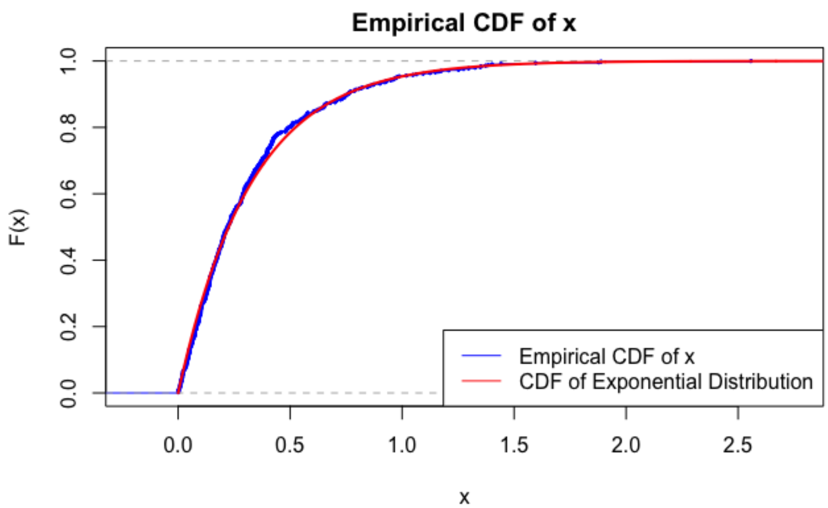


Fig. 1: The graph of the PDF for theoretical Exponential Distributinon with parameter $\hat{\lambda} = \frac{1}{\hat{m}_{\mathrm{ML}}}$

And we use the following code for comparing we have the graph

```
# plot the CDF of exponential distribution with parameter lambda
y1 <- pexp(x, rate = 1/m)
plot(x, y1, type="l", main="CDF of Exponential Distribution", xlab="x",
   ylab="F(x)")
plot(ecdf(data$x[2:501]), main="Empirical CDF of x", xlab="x", ylab="F(x)",
   col="blue", cex=0.3)
lines(x, y1, col="red", lwd=2)
legend("bottomright", legend=c("Empirical CDF of x", "CDF of Exponential
   Distribution"), col=c("blue", "red"), lty=1:1)
```



We can see that from the graph with the estimator is pretty close to the result from the dataset.

Question 5

We consider the second estimator Y_n. We set $M_n = \min(X_1, X_2, \cdots, X_n)$. What is the law of M_n ?

We set $M_n = \min(X_1, X_2, \cdots, X_n)$ and we observe that

$$\{\min\{X_1, \ldots, X_n\} \leq t\} = \left(\bigcap_{i=1}^{n} \{X_i > t\}\right)^c$$

and therefore,

$$\mathbb{P}[[\min\{X_1, \ldots, X_n\} \leq t]] = 1 - \prod_{i=1}^{n} \mathbb{P}[X_i \geq t]$$

The probabilities $\mathbb{P}[X_i \geq t]$ can be computed using point (i) and we obtain

$$\mathbb{P}[\{\min\{X_1, \ldots, X_n\} \leq t\}] = 1 - \exp\left(-\left(\sum_{i=1}^{n} \lambda_i\right) t\right)$$

Since the cumulative distribution function characterizes the law of a random variable we can conclude that $\min(X_1, X_2, \cdots, X_n) \sim Exp(\sum_{i=1}^{n} \lambda_i)$. Since we know for all $(X_i)_{1\leq i\leq N}$ are i.i.d.r.v, therefore we have $\min(X_1, X_2, \cdots, X_n) \sim Exp(n\lambda)$. And we have the law of M_n that

$$f_{M_n}(x) = \begin{cases} \frac{n}{m} e^{-\frac{n}{m}x} & (x > 0) \\ 0 & (x \leqslant 0) \end{cases}$$

Question 6

Is the estimator Y_n a biased estimator? compute its quadratic risk.

Since we have $Y_n = n \min(X_1, X_2, \cdots, X_n)$ and $\min(X_1, X_2, \cdots, X_n) \sim Exp(n\lambda)$. We have

$$B(Y_n) = \mathbb{E}[Y_n] - m = n\mathbb{E}[M_n] - m = \frac{n}{n\lambda} - m = 0$$

Therefore, we know that the estimator Y_n is an unbiased estimator. Then we compute the quadratic risk, we have

$$R(Y_n) = Var(Y_n) + B^2(Y_n) = Var(Y_n) = Var(nM_n) = n^2 Var(M_n) = \frac{n^2}{(n\lambda)^2} = \frac{1}{\lambda^2} = m^2$$

Therefore, we get the quadratic risk of Y_n is m^2

Question 7

Given these informations, what do you choose between these two estimators ? justify.

Since we have $R(Y_n) = m^2$ and $R(\hat{m}_{\text{ML}}) = \frac{m^2}{N}$. Since both are unbiased estimator of m, then we compare the quadratic risk of them. Since

$$R(\hat{m}_{\text{ML}}) = \frac{m^2}{N} < m^2 = R(Y_n)$$

Therefore, estimator $\hat{m}_{\text{ML}}$ is better since the quadratic risk (minimal square error) is smaller which means this estimator can better evaluate the value of m.

Definition

In what follows, we work with the Maximum likelihood estimator $\hat{m}_{\mathrm{ML}}$ and we want to analyze how close is the value of $\hat{m}_{\mathrm{ML}}$ to the real value of the parameter m in terms of probability.

We consider the empirical variance $\bar{V}_n = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}_n\right)^2$ where $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ is the empirical mean. For the remaining part of this exercice, we recall some properties that will be useful:

- **P.1** From Slutsky's theorem: Let $n \in \mathbb{N}$ and X_n, Y_n be random variables. if X_n converges in law to a random variable X (we use the notation $X_n \xrightarrow[n\to+\infty]{\mathcal{L}} X$) and Y_n converges in probability to a constant $c \in \mathbb{R}$ (we use the notation $Y_n \overset{\mathbb{P}}{\underset{n\to+\infty}{\to}} c$) then $X_n Y_n \underset{\substack{\mathcal{L} \\ n\to+\infty}}{\to} cX$ and $\frac{X_n}{Y_n} \overset{\mathcal{L}}{\underset{n\to+\infty}{\to}} \frac{X}{c}$ if $c \neq 0$.
- **P.2** Property from continuous mapping theorem: Let X_n, X be random variables defined on a metric space S and assume that $f : S \to S'$ (S' another metric space) such that g is continuous then
  - if $X_n \xrightarrow[n\to+\infty]{\mathcal{L}} X$ then $g\left(X_n\right) \xrightarrow[n\to+\infty]{\mathcal{L}} g(X)$
  - if $X_n \overset{\mathbb{P}}{\underset{n\to+\infty}{\to}} X$ then $g\left(X_n\right) \overset{\mathbb{P}}{\underset{n\to+\infty}{\to}} g(X)$.

Question 8

Show that $\bar{V}_n$ converges in probability to the variance of X_1, $\operatorname{Var}\left(X_1\right)$.

we have

$$\begin{aligned}
\bar{V}_n &= \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}_n\right)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(X_i^2 - 2X_i\bar{X}_n\right) + \bar{X}_n^2 \\
&= \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \frac{2}{n}\bar{X}_n\sum_{i=1}^{n} X_i + \bar{X}_n^2
\end{aligned}$$

Since X_i are i.i.d, then we have X_i^2 are also i.i.d. Also, we already know $\mathbb{E}[X_i^2] < +\infty$. We apply the law of large numbers we have

$$\frac{1}{n}\sum_{i=1}^{n} X_i^2 \to \mathbb{E}[X_1^2] = Var(X_1) + \mathbb{E}[X_1]^2$$

By the law of large number and $\mathbb{E}[|X_1|] < +\infty$ we have $\bar{X}_n \xrightarrow[n\to+\infty]{\mathcal{P}} \mathbb{E}[X_1]$, we got

$$\frac{2}{n}\bar{X}_n\sum_{i=1}^{n} X_i \xrightarrow[n\to+\infty]{\mathcal{P}} 2\mathbb{E}[X_1]^2$$

From $\bar{X}_n \xrightarrow[n\to+\infty]{\mathcal{P}} \mathbb{E}[X_1]$ and $y = x^2$ is continuous, we have

$$\bar{X}_n^2 \xrightarrow[n\to+\infty]{\mathcal{P}} \mathbb{E}[X_1]^2$$

Therefore we have that

$$\bar{V}_n \xrightarrow[n\to+\infty]{\mathcal{P}} Var(X_1)$$

So we prove that $\bar{V}_n$ converges in probability to the $Var(X_1)$

Question 9

What is the asymptotic distribution of the random variable $Z_n = \frac{\sqrt{n}(\hat{m}_{\text{ML}}-m)}{\sqrt{\bar{V}_n}}$ as $n \to +\infty$?

From **Question 9** we have $\bar{V}_n$ converges in probability to the $Var(X_1)$. Since $(X_n)_{n\in\mathbb{N}}$ is a sequence of i.i.d.r.v such that $\mathbb{E}[|X_1|] = \mathbb{E}[X_1] = m < \infty$. Also from **Question 4** we have $\hat{m}_{\text{ML}} = \bar{x}_n$. Therefore we have

$$Z_n = \frac{\sqrt{n}\,(\hat{m}_{\text{ML}} - m)}{\sqrt{\bar{V}_n}} \xrightarrow[n\to+\infty]{} \frac{\sqrt{n}\,(\bar{x}_n - \mathbb{E}[X_1])}{\sqrt{Var(X_1)}}$$

Then we apply the Central Limit Theorem we have

$$\frac{\sqrt{n}\,(\bar{x}_n - \mathbb{E}[X_1])}{\sqrt{Var(X_1)}} \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0,1)$$

Therefore, the asymptotic distribution of the random variable $Z_n = \frac{\sqrt{n}(\hat{m}_{\text{ML}}-m)}{\sqrt{\bar{V}_n}}$ as $n \to +\infty$ is standard normal distribution.

Question 10

For large $n = 10^5$, simulate the random variable Z_n and compare it to its asymptotic distribution (you can choose any value of m)

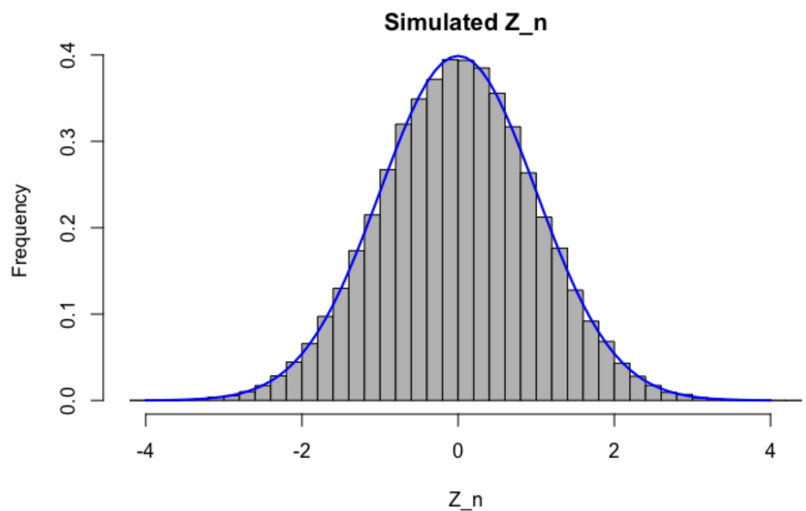
We set $n = 10^5$ and we apply the following code to simulate Z_n. Since Z_n is only one number, in order to see the distribution of Z_n is standard normal distribution, we need to execute the simulation of Z_n for multiple times.

```
# initialization of the parameters
n <- 100000
sqrt_n <- sqrt(n)
lis <- list()
for (i in 1:n){
  rand <- rexp(n,1)
  value <- sqrt_n * (mean(rand)-1)
  lis <- c(lis, value)
}
lis2 <- unlist(lis, use.names = FALSE)

# plot the graph with standard normal distribution
hist(lis2, breaks = 50, main = "Simulated Z_n", xlab = "Z_n", ylab =
    "Frequency",col = "gray", probability = TRUE)
```

```r
x <- seq(-4, 4, by = 0.1)
y <- dnorm(x)
lines(x, y, col = "blue", lwd = 2)
```

We have the graph of following



From the plot we can see that the distribution of the simulation result of Z_n is close to the standard normal distribution which is the blue line in the graph.

Question 11

Let $\alpha \in [0,1]$. By considering a large n, give the interval I such that:

$$P(m \in I) = 1 - \alpha$$

Since we know that the random variable $Z_n = \frac{\sqrt{n}(\hat{m}_{\mathrm{ML}} - m)}{m} = \sqrt{n}(\frac{\hat{m}_{\mathrm{ML}}}{n} - 1)$ as $n \to +\infty$ is standard normal distribution. For $\alpha > 0$, let $q_{1-\frac{\alpha}{2}}$ the quantile of order $(1 - \frac{\alpha}{2})$ of $\mathcal{N}(0,1)$. We have the asymptotic Confidence Interval for n is large enough

$$\begin{aligned}
1 - \alpha &\simeq \mathbb{P}\left(-q_{1-\alpha/2} \leq \sqrt{n}(\frac{\hat{m}_{\mathrm{ML}}}{n} - 1) \leq q_{1-\alpha/2}\right) \\
&\simeq \mathbb{P}(-\frac{q_{1-\alpha/2}}{\sqrt{n}} + 1 \leq \frac{\hat{m}_{\mathrm{ML}}}{m} \leq \frac{q_{1-\alpha/2}}{\sqrt{n}} + 1) \\
&\simeq \mathbb{P}(\frac{\sqrt{n}}{\sqrt{n} + q_{1-\alpha/2}}\hat{m}_{\mathrm{ML}} \leq m \leq \frac{\sqrt{n}}{\sqrt{n} - q_{1-\alpha/2}}\hat{m}_{\mathrm{ML}})
\end{aligned}$$

Then we find the asymptotic interval I that

$$I = [\frac{\sqrt{n}}{\sqrt{n} + q_{1-\alpha/2}}\hat{m}_{\mathrm{ML}}, \frac{\sqrt{n}}{\sqrt{n} - q_{1-\alpha/2}}\hat{m}_{\mathrm{ML}}]$$

2 Exercise 2

In this exercise, we will prove the Box Muller method for generating Gaussian random variables. Then, we will investigate the method in practice.

Question 12

Let U follow a uniform distribution $\mathcal{U}([0,1])$. Let $\phi(x) = \sqrt{-2\log(x)}$. Find the cumulative distribution function of the random variable $R = \phi(U)$. Deduce the density of R

From the definition of CDF, we have

$$\begin{aligned} F_R(x) &= F_{\phi(U)}(x) \\ &= \mathbb{P}(\sqrt{-2\log(U)} \leq x) \end{aligned}$$

For $\sqrt{-2\log(U)} \leq x$ we have

$$\begin{aligned} \sqrt{-2\log(U)} &\leq x \\ -2\log(U) &\leq x^2 \\ \log(U) &\geq -\frac{x^2}{2} \\ U &\geq e^{-\frac{x^2}{2}} \end{aligned}$$

Since U is a uniform distribution, therefore we have

$$F_R(x) = \mathbb{P}(U \geq e^{-\frac{x^2}{2}}) = 1 - exp(-\frac{x^2}{2})$$

$$F_R(x) = \begin{cases} 0, & x < 0 \\ 1 - exp(-\frac{x^2}{2}), & x \geq 0 \end{cases}.$$

And we take the derivative of $F_x(R)$ then we have the density

$$f_R(x) = \frac{d}{dx}F_R(x) = \begin{cases} 0, & x < 0 \\ xe^{-\frac{x^2}{2}}, & x \geq 0 \end{cases}.$$

Question 13

Now, let V be a second uniformly distributed random variable between 0 and 1 , independent from U. Give the density of the random variable $\Theta = 2\pi V$. Deduce the joint density of R, Θ (remember that U and V are independent).

Since the random variable $\Theta = 2\pi V$, we have for the CDF of Θ

$$F_\Theta(x) = \mathbb{P}[2\pi V \leq x]$$

Since we know that V is a uniformly distributed random variable, we got

$$F_\Theta(x) = \mathbb{P}[V \leq \frac{x}{2\pi}] = \begin{cases} 0, & x < 0 \\ \frac{x}{2\pi}, & 0 \leq x \leq 2\pi \\ 1, & x > 2\pi \end{cases}.$$

Then we take the derivative of $F_x(\Theta)$, we have

$$f_\Theta(x) = \begin{cases} \frac{1}{2\pi}, & 0 \le x \le 2\pi \\ 0, & otherwise \end{cases}.$$

Since the random variable U and V are independent, therefore Θ and R are independent, For joint density of R and Θ these two random variable we have for the joint CDF

$$F_{\Theta,R}(x,y) = \mathbb{P}(\Theta \le x, R \le y) = \mathbb{P}(\Theta \le x)\mathbb{P}(R \le y) = \begin{cases} 0, & x < 0 \text{ and } y < 0 \\ \frac{1}{2\pi}(x - xexp(-\frac{y^2}{2})), & 0 \le x \le 2\pi \text{ and } y \ge 0 \\ 1 - exp(-\frac{y^2}{2}), & x > 2\pi \text{ and } y \ge 0 \end{cases}.$$

Then for the joint density of continuous RV, we have the formula

$$f_{\Theta,R}(x,y) = \frac{\partial^2 F_{X,Y}(x,y)}{\partial x \partial y} = \begin{cases} 0, & x < 0 \text{ and } y < 0 \\ \frac{yexp(-\frac{y^2}{2})}{2\pi}, & 0 \le x \le 2\pi \text{ and } y \ge 0 \\ 0, & x > 2\pi \text{ and } y \ge 0 \end{cases}$$

In conclusion

$$f_{R,\Theta}(x,y) = \begin{cases} \frac{xexp(-\frac{x^2}{2})}{2\pi}, & 0 \le y \le 2\pi \text{ and } x \ge 0 \\ 0, & otherwise \end{cases}$$

We now have two independent random variables of known joint distribution. We may consider those two variables as the polar coordinates of a point in a plane. The following fact allows to compute the cartesian coordinates of the same point.

Question 14

Let $(a,b) = g(R,\Theta) = (R\cos(\Theta), R\sin(\Theta))$. Find a way to compute R given only a and b.

Since we know that $cos^2(\Theta)+sin^2(\Theta) = 1$, therefore we have $R^2 = R^2(cos^2(\Theta)+sin^2(\Theta)) = a^2+b^2$. Since $R \ge 0$, then we can deduce that

$$R = \sqrt{a^2 + b^2}$$

Definition

Let R, Θ be two random variables, and let $X, Y = g(R,\Theta) = (R\cos(\Theta), R\sin(\Theta))$. Then, $f_{X,Y}(a,b) = \frac{f_{R,\Theta}(g^{-1}(a,b))}{r}$.

Question 15

In the previous question, we only computed the first term of $g^{-1}(a,b)$. Why is the second term unnecessary? Find the analytical expression of $f_{X,Y}$.

In **Question 13**, we can see that the joint density function $f_{R,\Theta}(x,y)$ does not contain variable y, which means doesn't contain a variable of Θ. Therefore, in order to get $f_{X,Y}(a,b)$ we do not need to compute the second term Θ. For analytical expression of $f_{X,Y}$ we have

$$f_{X,Y}(a,b) = \frac{f_{R,\Theta}\left(g^{-1}(a,b)\right)}{r} = \frac{f_{R,\Theta}(r,\theta)}{r} = \frac{rexp(-\frac{r^2}{2})}{r2\pi} = \frac{exp(-\frac{r^2}{2})}{2\pi}$$

where $r = \sqrt{a^2+b^2}$. Then we got

$$f_{X,Y}(a,b) = \frac{exp(-\frac{a^2+b^2}{2})}{2\pi}$$

Question 16

Are X and Y independent? What are their respective law?

From **Question 15** we have

$$f_{X,Y}(a,b) = \frac{exp(-\frac{a^2+b^2}{2})}{2\pi} = \frac{1}{\sqrt{2\pi}}e^{-\frac{a^2}{2}} \cdot \frac{1}{\sqrt{2\pi}}e^{-\frac{b^2}{2}}$$

Since the joint variable is continuous RV, therefore we can compute the marginal PDF of X and Y, we have

$$\begin{aligned} f_X(x) &= \textstyle\int_{-\infty}^{\infty} f_{XY}(x,y)dy, \quad \text{for all } x,\\ f_Y(y) &= \textstyle\int_{-\infty}^{\infty} f_{XY}(x,y)dx, \quad \text{for all } y. \end{aligned}$$

Then we got

$$\begin{aligned} f_X(a) &= \int_{-\infty}^{\infty} f_{XY}(a,b)db \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}}e^{-\frac{a^2}{2}} \cdot \frac{1}{\sqrt{2\pi}}e^{-\frac{b^2}{2}} db \\ &= \frac{1}{2\pi}e^{-\frac{a^2}{2}} \int_{-\infty}^{\infty} e^{-\frac{b^2}{2}} db \end{aligned}$$

By Gaussian integral we got

$$f_X(a) = \frac{1}{2\pi}e^{-\frac{a^2}{2}} \int_{-\infty}^{\infty} e^{-\frac{b^2}{2}} db = \frac{1}{\sqrt{2\pi}}e^{-\frac{a^2}{2}}$$

With same method we have

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

Therefore we have

$$f_{X,Y}(a,b) = \frac{exp(-\frac{a^2+b^2}{2})}{2\pi} = \frac{1}{\sqrt{2\pi}}e^{-\frac{a^2}{2}} \cdot \frac{1}{\sqrt{2\pi}}e^{-\frac{b^2}{2}} = f_X(a)\cdot f_Y(b)$$

We prove that X and Y are independent, for respective law we got

$$\mathbb{P}[X=a] = f_X(a) = \frac{1}{\sqrt{2\pi}}e^{-\frac{a^2}{2}}$$

$$\mathbb{P}[Y=b] = f_Y(b) = \frac{1}{\sqrt{2\pi}}e^{-\frac{b^2}{2}}$$

Question 17

We recall that the Box-Muller method consists of generating normal distributions by observing samples of U and V two uniform random variables between 0 and 1, and applying either the transformation $X = \sqrt{-2\log(U)}\cos 2\pi V$ or $Y = \sqrt{-2\log(U)}\sin 2\pi V$. Conclude on the validity of the method.

From previous question we have $X = R\cos(\Theta) = \sqrt{-2\log(U)}\cos 2\pi V$. From last question we have

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

Since we know the density of normal distribution $\mathcal{N}(\mu, \sigma^2)$ is

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

Then for X we got $\mu = 0$ and $\sigma = 1$. Therefore we have $X \sim \mathcal{N}(0,1)$. For Y we have $f_Y(x) = \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$, due to the same reason we have $Y \sim \mathcal{N}(0,1)$. Therefore Box-Muller is a valid approach and we can use this method to simulate standard normal distribution $\mathcal{N}(0,1)$.

Question 18

We will now empirically test the Bow-Muller method. Generate 3000 samples of U and V i.i.d. following a uniform law in $[0,1]$, then X and Y according to the Box-Muller method. Plot the histograms of both functions and compute the covariance between X and Y. Generate 3000 samples using the rnorm function of R. Comment each results.

We first use the following R code to generate X and Y.

```
library(ggplot2)
# generate 3000 samples of U and V i.i.d. following a uniform law in [0,1]
U <- runif(3000)
V <- runif(3000)
# Then we apply the Box-Muller method to get X and Y
X <- sqrt(-2*log(U))*cos(2*pi*V)
Y <- sqrt(-2*log(U))*sin(2*pi*V)
# plot histograms of X and Y Seperately
ggplot() +
  geom_histogram(aes(x=X), bins=30, fill="blue", alpha=0.5) +
  ggtitle("Histograms of X (Box-Muller method)") +
  xlab("Value") +
  ylab("Frequency")
ggplot() +
  geom_histogram(aes(x=Y), bins=30, fill="red", alpha=0.5) +
  ggtitle("Histograms of Y (Box-Muller method)") +
  xlab("Value") +
  ylab("Frequency")

# compute the covariance between X and Y
cov(X, Y)
```

Then we got the histograms of X and Y with 3000 samples.

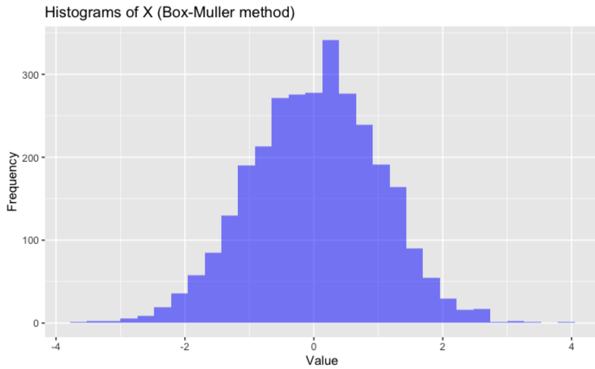


Fig. 2: Histogram of X by Box-Muller method

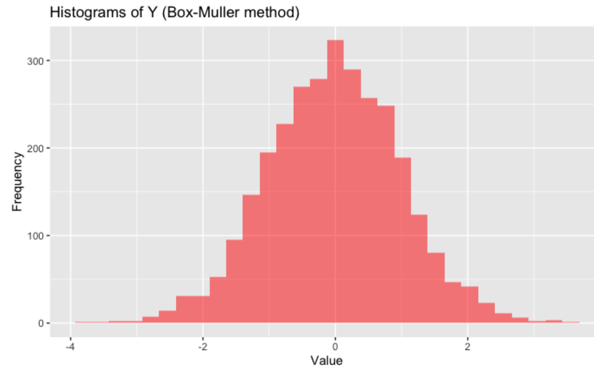


Fig. 3: Histogram of Y by Box-Muller method

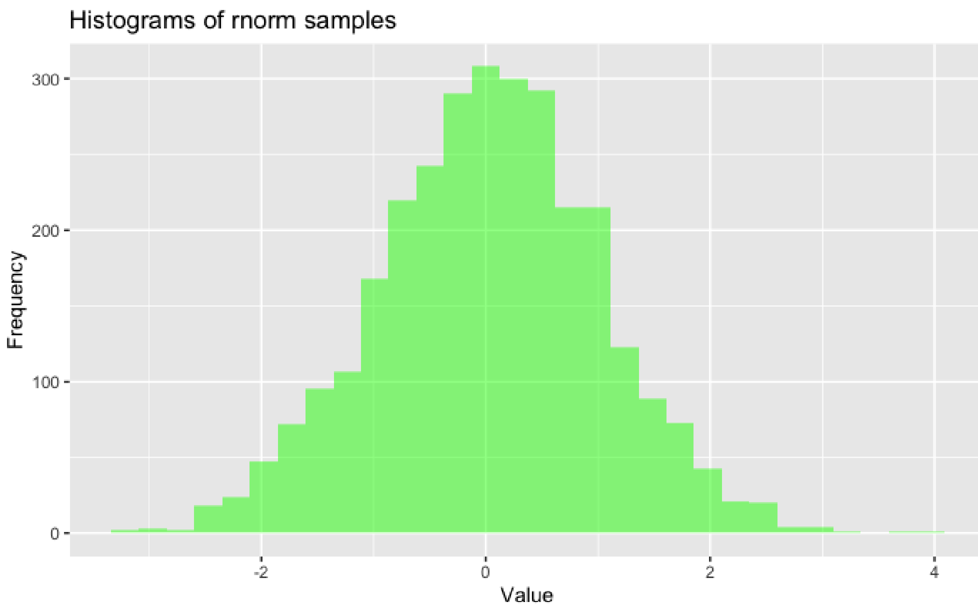
We can see that the histograms of X and Y is pretty similar to standard normal distribution. Also we can use the code to simulate 3000 samples of standard normal distribution, and compare the graph.

```
# generate 3000 samples using the rnorm function
rnorm_samples <- rnorm(3000)

# plot histogram of rnorm_samples
ggplot() +
  geom_histogram(aes(x=rnorm_samples), bins=30, fill="green", alpha=0.5) +
  ggtitle("Histograms of rnorm samples") +
  xlab("Value") +
  ylab("Frequency")

# Compare the graph
ggplot() +
  geom_histogram(aes(x=X, fill="X"), bins=30, alpha=0.5) +
  geom_histogram(aes(x=Y, fill="Y"), bins=30, alpha=0.5) +
  geom_density(aes(x=rnorm(3000), fill="Standard normal"), kernel = "gaussian",
     color = "black") +
  ggtitle("Histograms of X and Y with standard normal distribution") +
  xlab("Value") +
  ylab("Density") +
  scale_fill_discrete(name="Variable", labels=c("X","Y", "Standard normal"))
```

From the graph we can see that X, Y have a high degree of overlap with the images of our simulated standard normal distribution. Also, from the result of computation by R, we got $Cov(X, Y) = -0.003115452$. **Then we can conclude that X and Y are independent and $X, Y \sim \mathcal{N}(0, 1)$**



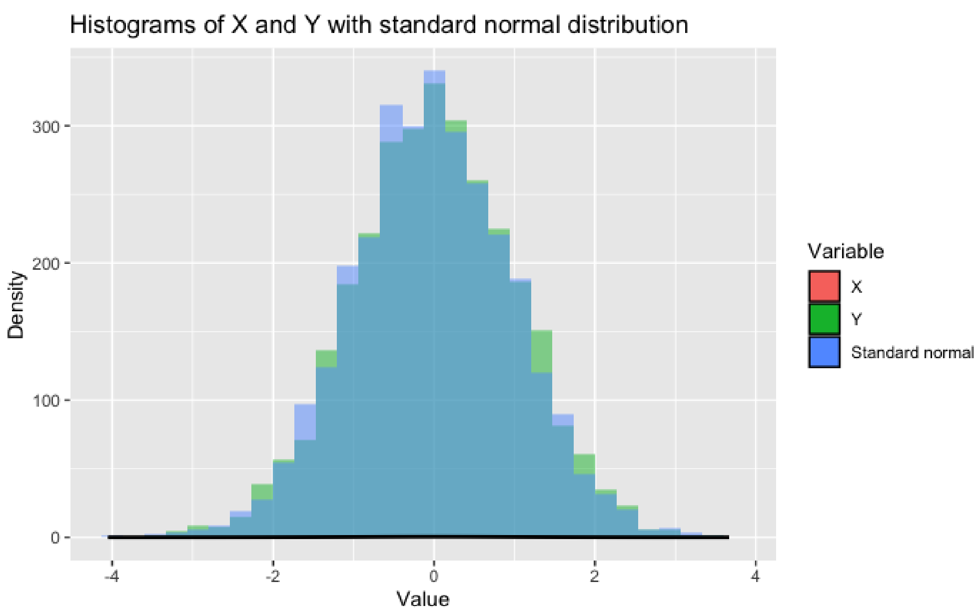


Fig. 4: Compare between X, Y, and standard normal distibution

Question 19

A common way to compare two distributions is through a quantile-quantile diagram. Generate $N1, N2$ two vectors of 3000 sample of a normal distribution using the funntion rnorm, and $E1, E2$ a vector of 3000 samples of an exponential distribution using the function rexp. Plot the quantile-quantile diagrams of $N1$ and $N2$, then of $E1$ and $E2$. What do you observe? Now plot the diagram of $N1$ and $E1$. How do you interpret the changes? (you may type help(qqplot) in the console to get information on the syntax and the procedure).

We can use the follwoing R code for the plot

```
# generate samples
N1 <- rnorm(3000, mean = 0, sd = 1)
N2 <- rnorm(3000, mean = 0, sd = 1)
E1 <- rexp(3000, rate = 1)
E2 <- rexp(3000, rate = 1)
# plot Q-Q diagrams for N1 and N2, E1 and E2
qqplot(N1, N2)
qqplot(E1, E2)
# plot Q-Q diagrams for N1 and E1
qqplot(N1, E1)
```

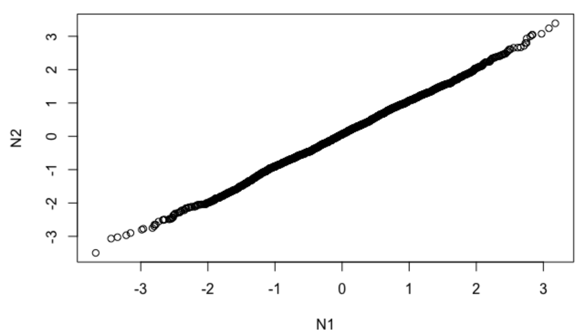


Fig. 5: QQ-plot of N_1 and N_2

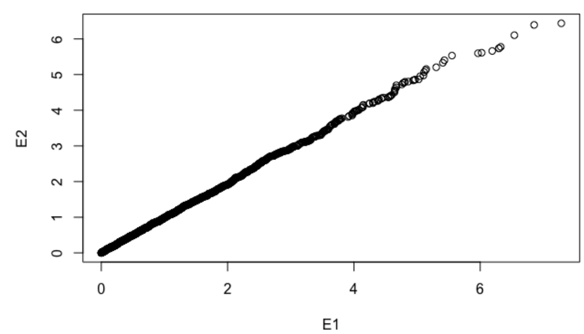


Fig. 6: QQ-plot of E_1 and E_2

From the graph we can see that the Q-Q diagram of $N1$ and $N2$, the points lie on a straight line, which indicates that the two normal distributions have the same quantiles, and the Q-Q plot is an identity line. The same reason apply to $E1$ and $E2$ since the graph is a straight line.

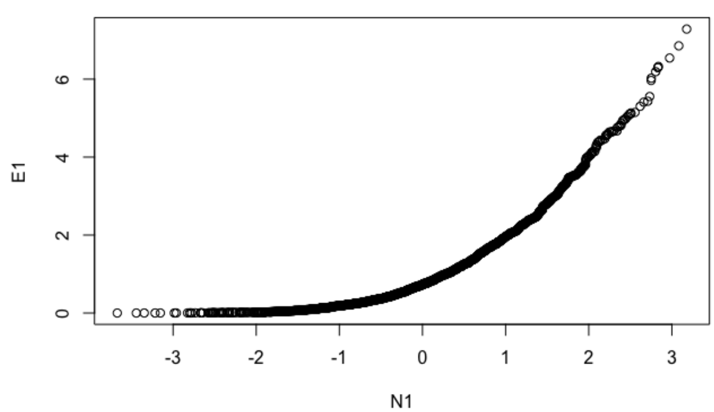


Fig. 7: QQ-plot of N_1 and E_1

We can observe that the points in the Q-Q plot is a curve with a 45-degree deviate line. The graph show that the distribution of E_1 and N_1 has different quantiles. This means that the Normal and Exponential distributions have different shapes and do not have the same quantiles.

Question 20

Draw the qqplot of the samples generated through the Box Muller method with the exponentially distributed samples first, then the normally distributed ones. Conclude.

We use the following R code for plot the graph

```
N <- 3000
# plot Q-Q plots for X and exponential distribution
qqplot(X, rexp(N, rate = 1))
# plot Q-Q plots for Y and normal distribution
qqplot(Y, rnorm(N))
```

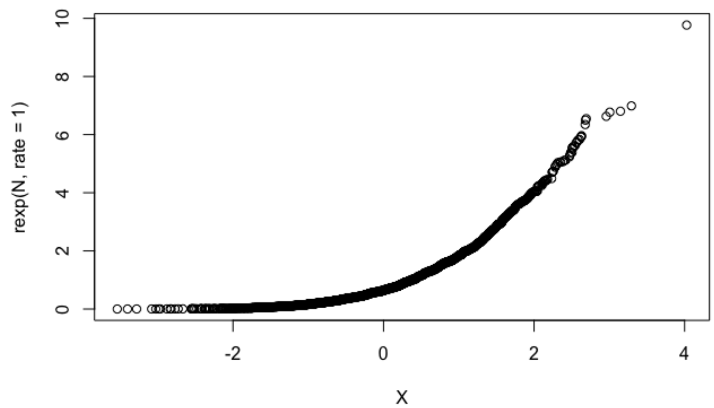


Fig. 8: QQ-plot of X and exponential distribution

We can find that the Q-Q plot of X and exponential distribution, the points deviate from the 45-degree line, which indicates that the two distributions are different and have different quantiles.

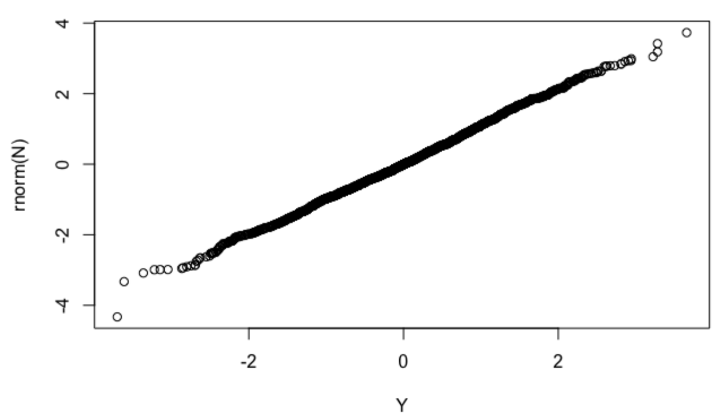


Fig. 9: QQ-plot of Y and Standard Normal Distribution

The Q-Q plot of Y and Standard Normal Distribution is a straight line, which means that Y (X)

has the same distribution with Standard Normal Distribution, this result is consistent with the results we proved before.