Homework_sourcecode

2023-01-15

Introduction

This is the source code of MAA204-Introduction to Statistics at Ecole Polytechnique Bachelor of Science program, from Yubo Cai.

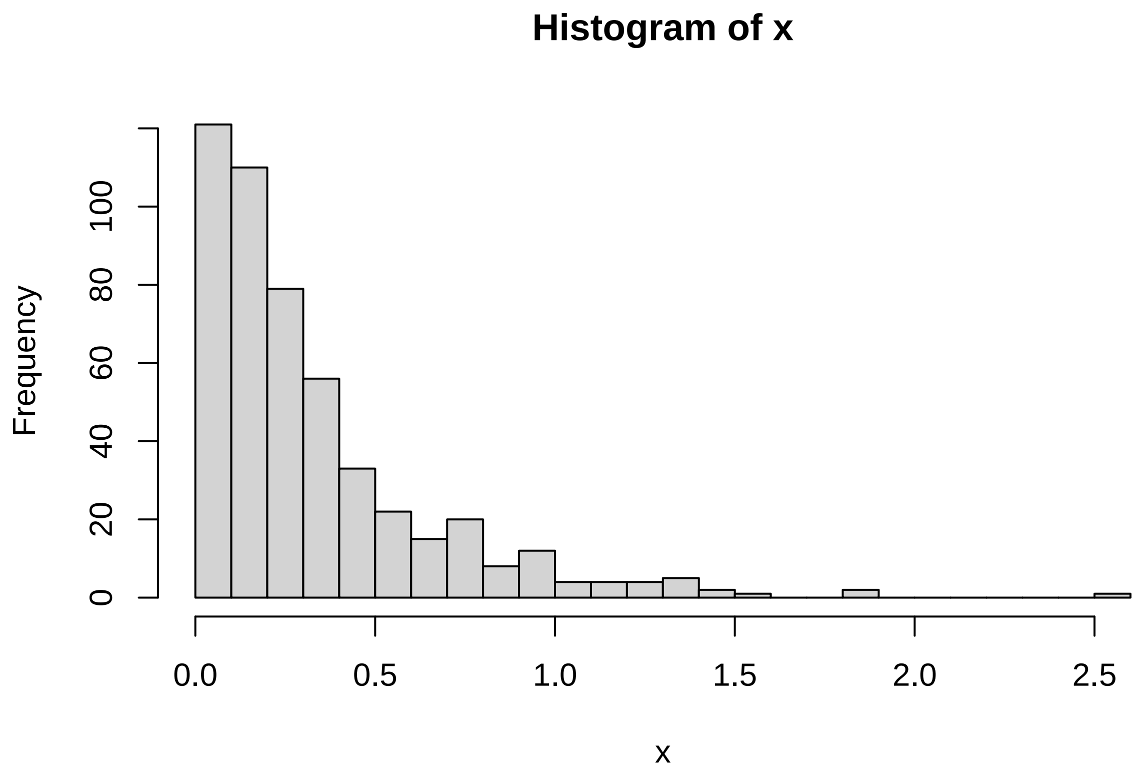
Question 4

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 $\hat{m}_{\mathrm{ML}}$

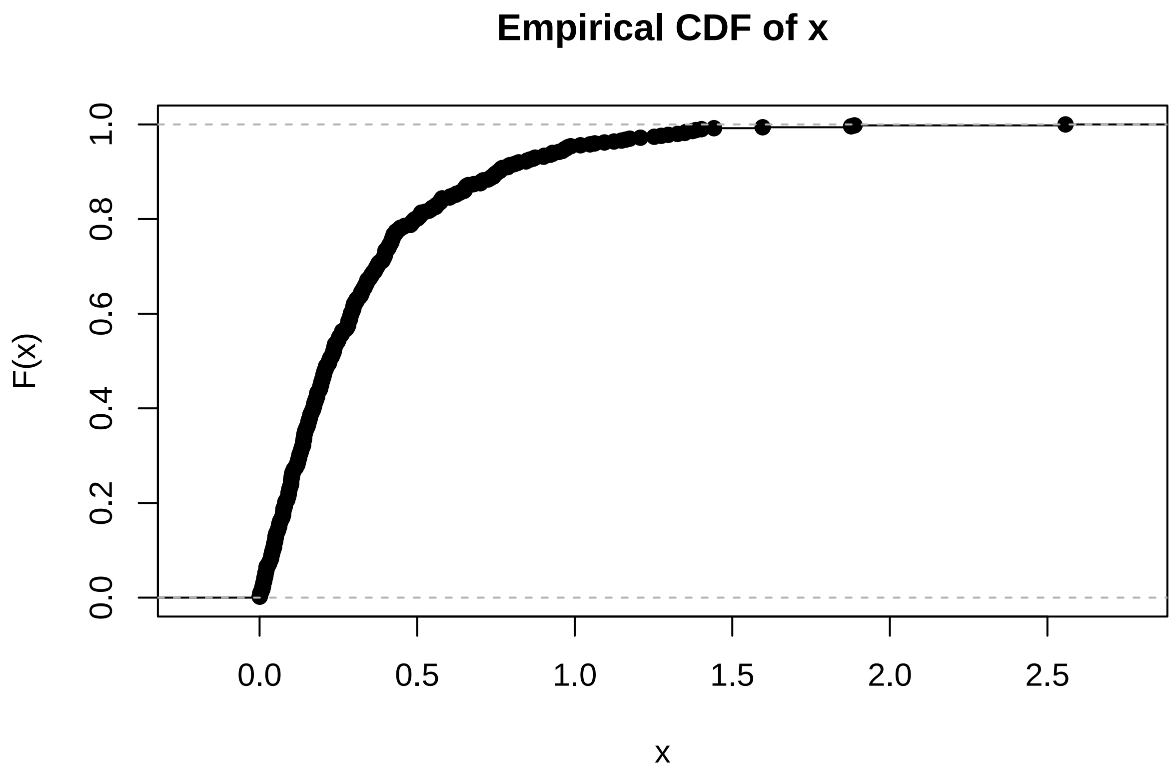
```
# import exponential.csv
data <- read.csv("exponential.csv")
summary(data)
```

```
##        x
##  Min.   :0.0004254
##  1st Qu.:0.1014974
##  Median :0.2222739
##  Mean   :0.3256294
##  3rd Qu.:0.4169742
##  Max.   :2.5573967
```

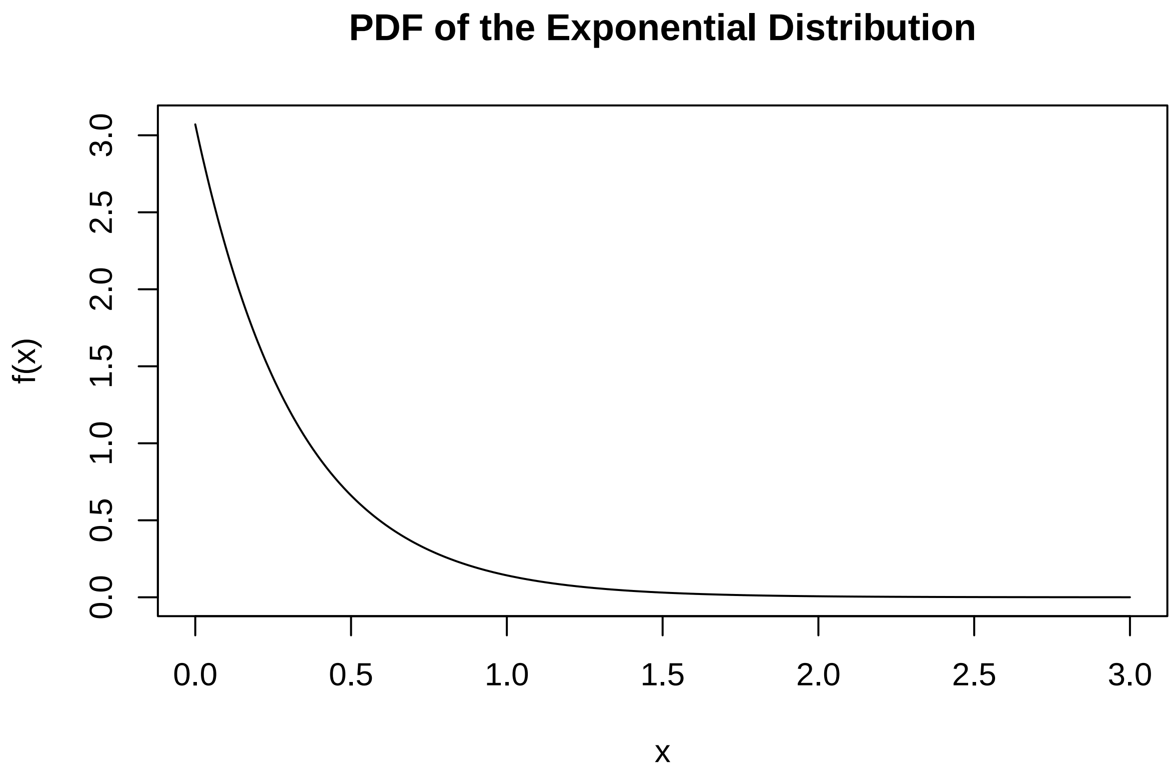
```
hist(data$x[2:501], breaks=30, main="Histogram of x", xlab="x", ylab="Frequency")
```



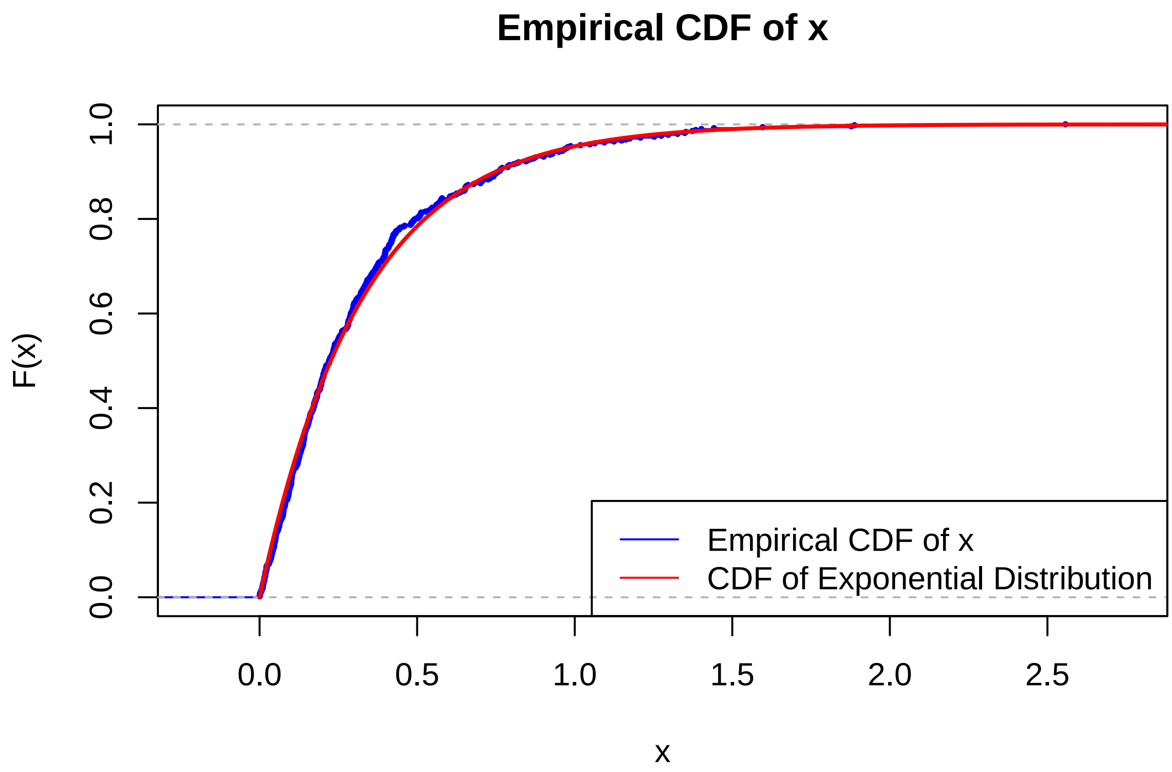

```
plot(ecdf(data$x[2:501]), main="Empirical CDF of x", xlab="x", ylab="F(x)")
```




```r
# 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="PDF of the Exponential Distribution", xlab="x", ylab="f(x)")
```



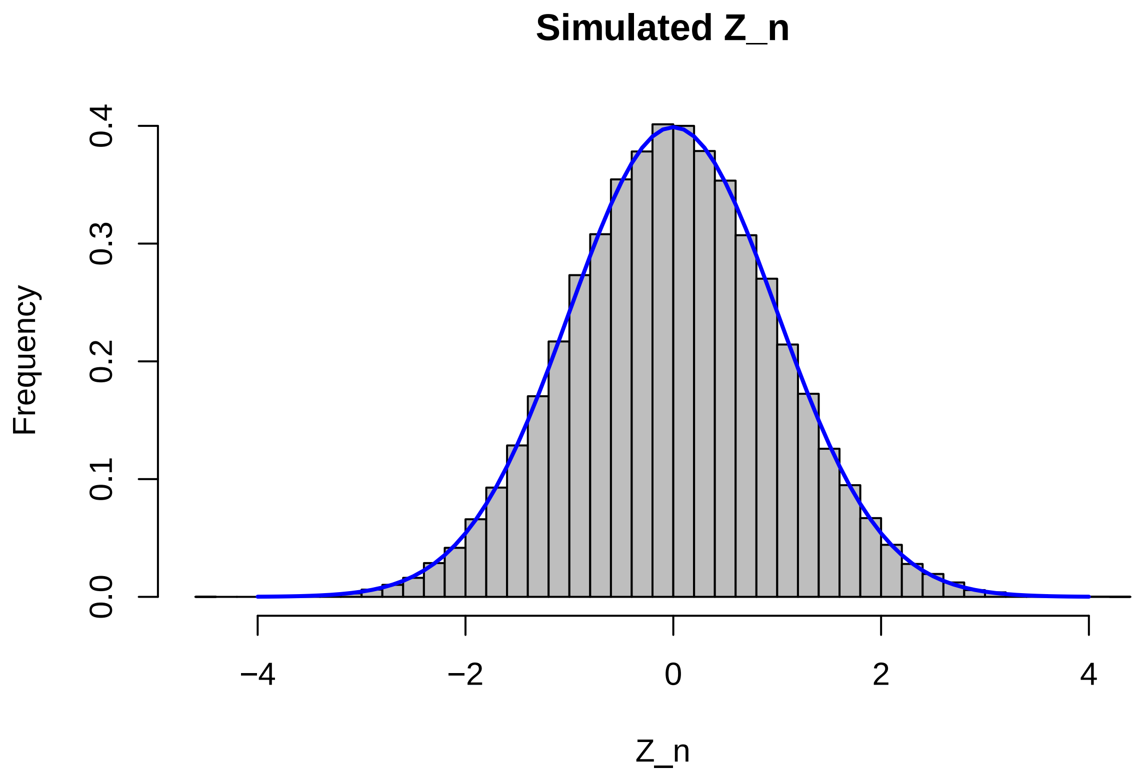

```
# plot the CDF of exponential distribution with parameter lambda
y1 <- pexp(x, rate = 1/m)
plot(ecdf(data$x[2:501]), main="Empirical CDF of x", xlab="x", ylab="F(x)", col="blue", cex=0.3)
lines(x, y1, col="red", lwd=2)
legend("bottomright", legend=c("Empirical CDF of x", "CDF of Exponential Distribution"),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



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)

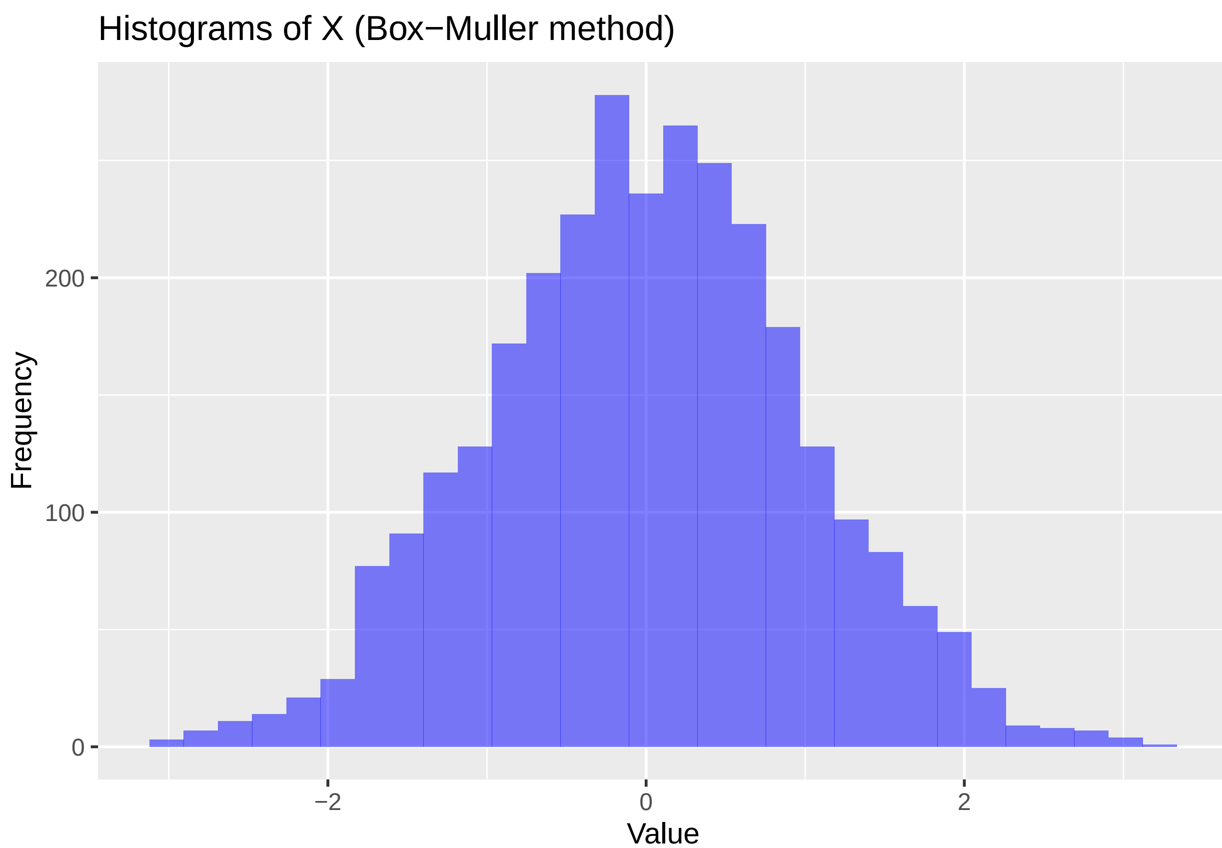
```
# initialization of the parameters
n <- 100000
sqrt_n <- sqrt(n)
lis <- list()
for (i in 1:n){
  rand <- rexp(n,1)
  value <- sqrt_n * (mean(rand)-1)
  lis <- c(lis, value)
}
lis2 <- unlist(lis, use.names = FALSE)

# plot the graph with standard normal distribution
hist(lis2, breaks = 50, main = "Simulated Z_n", xlab = "Z_n", ylab = "Frequency",col = "gray", probabil
x <- seq(-4, 4, by = 0.1)
y <- dnorm(x)
lines(x, y, col = "blue", lwd = 2)
```



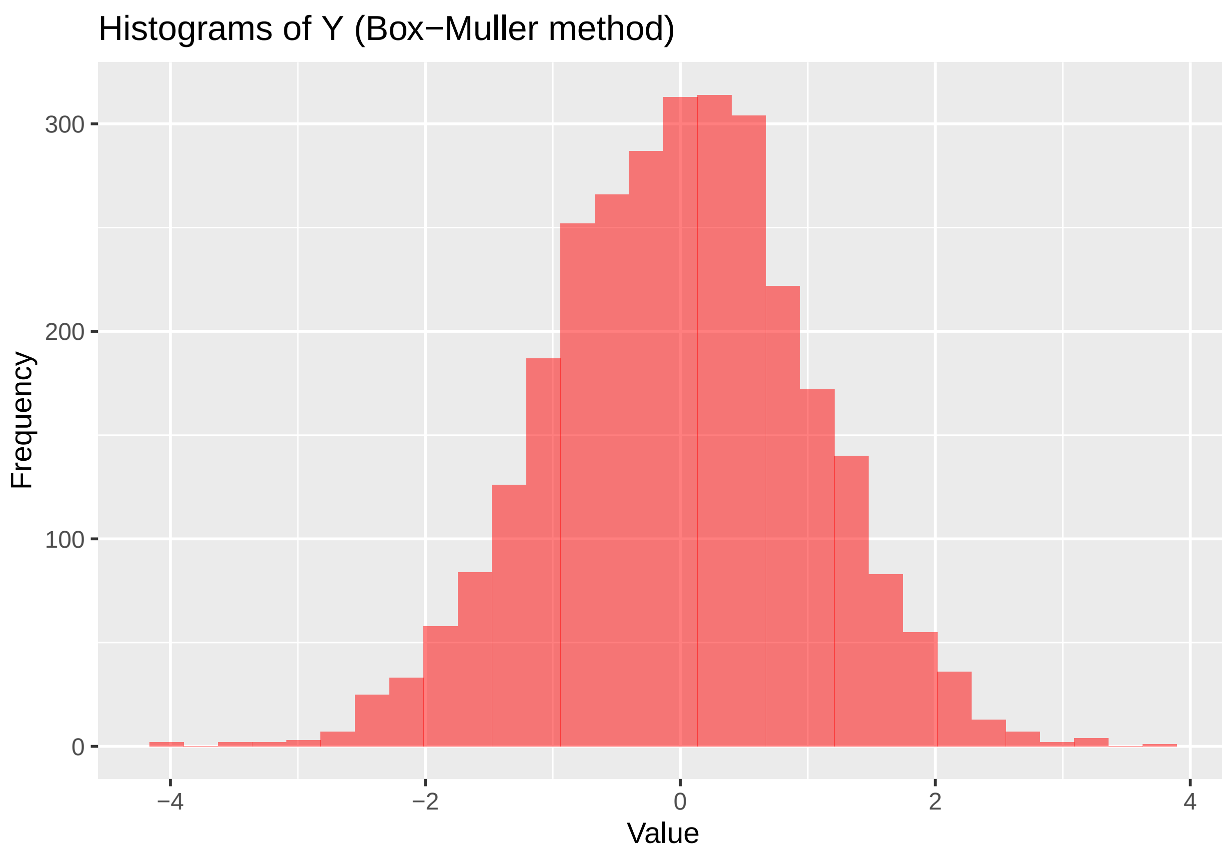
Question 18

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.

```
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")
```



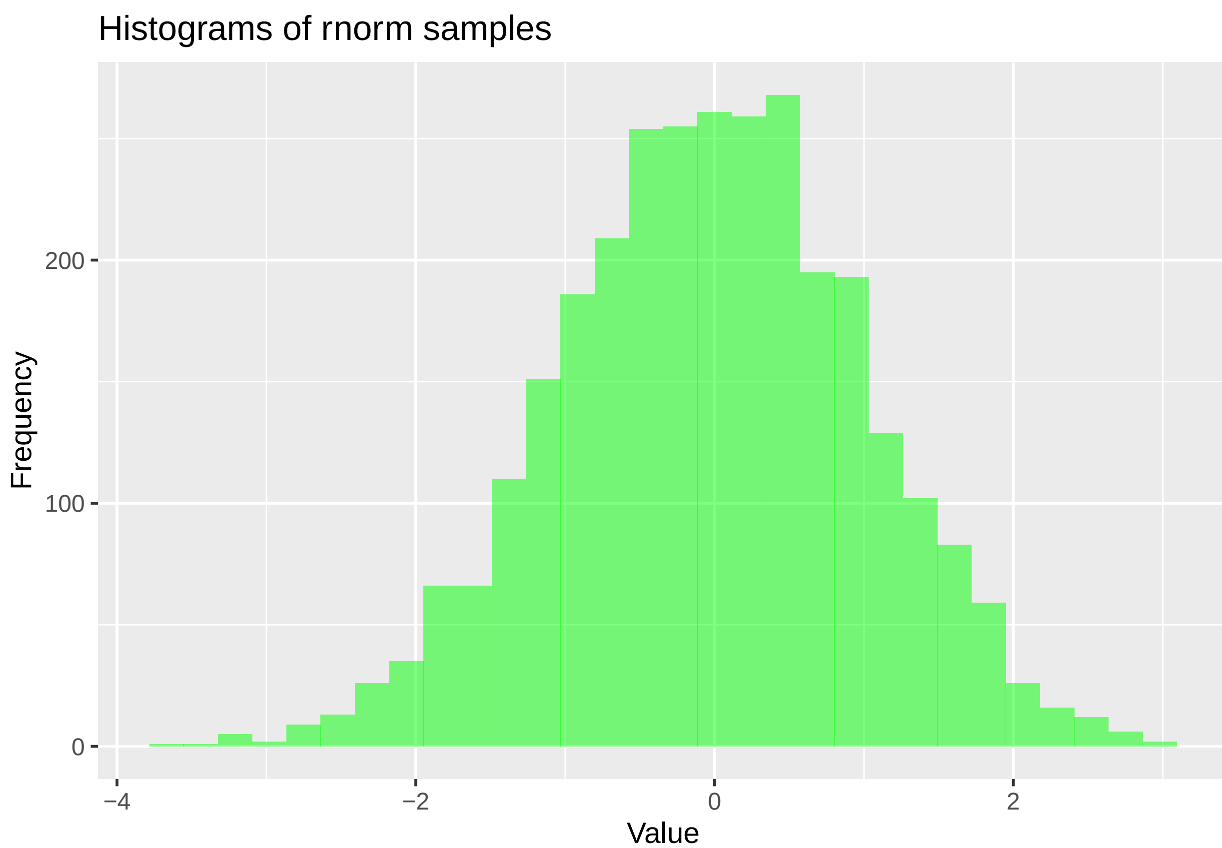

```
ggplot() +
  geom_histogram(aes(x=Y), bins=30, fill="red", alpha=0.5) +
  ggtitle("Histograms of Y (Box-Muller method)") +
  xlab("Value") +
  ylab("Frequency")
```



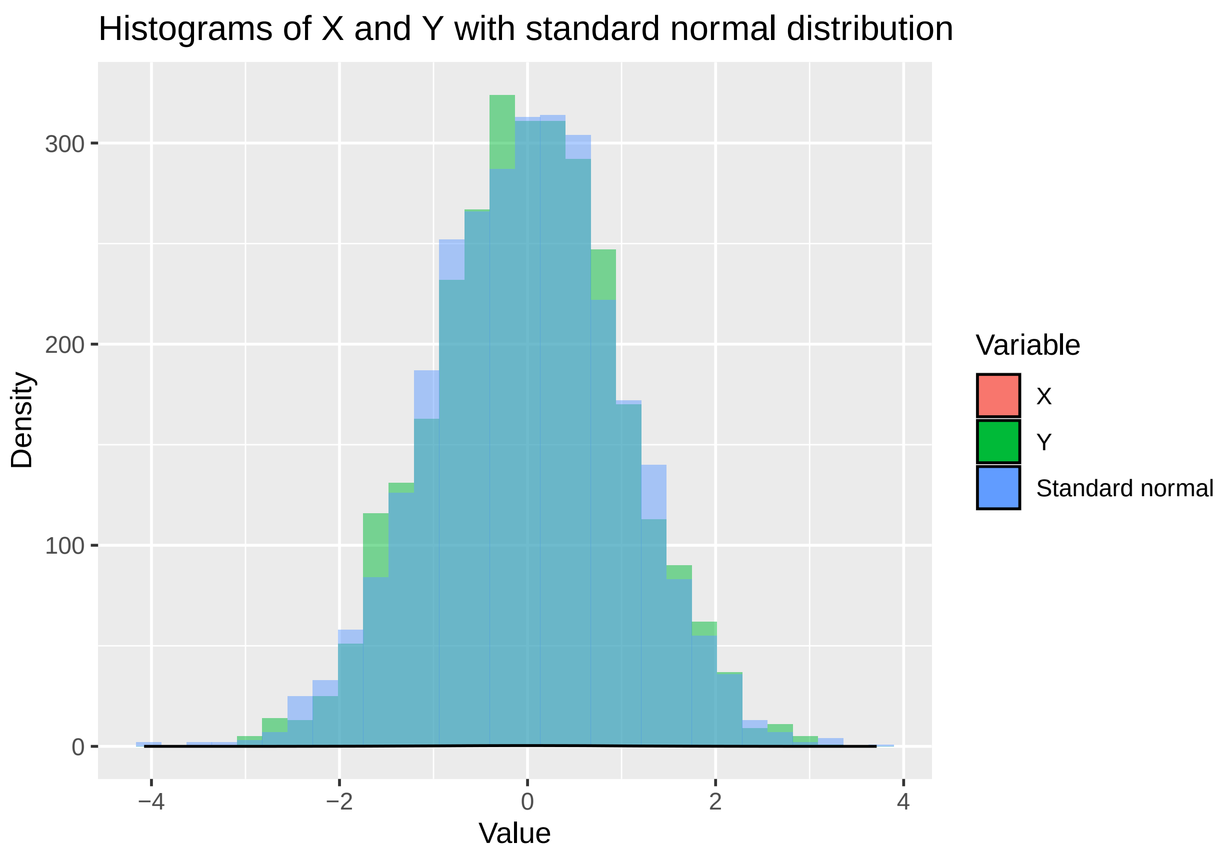

```r
# compute the covariance between X and Y
cov(X, Y)
```

```
## [1] 0.0115945
```

```r
# 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")
```



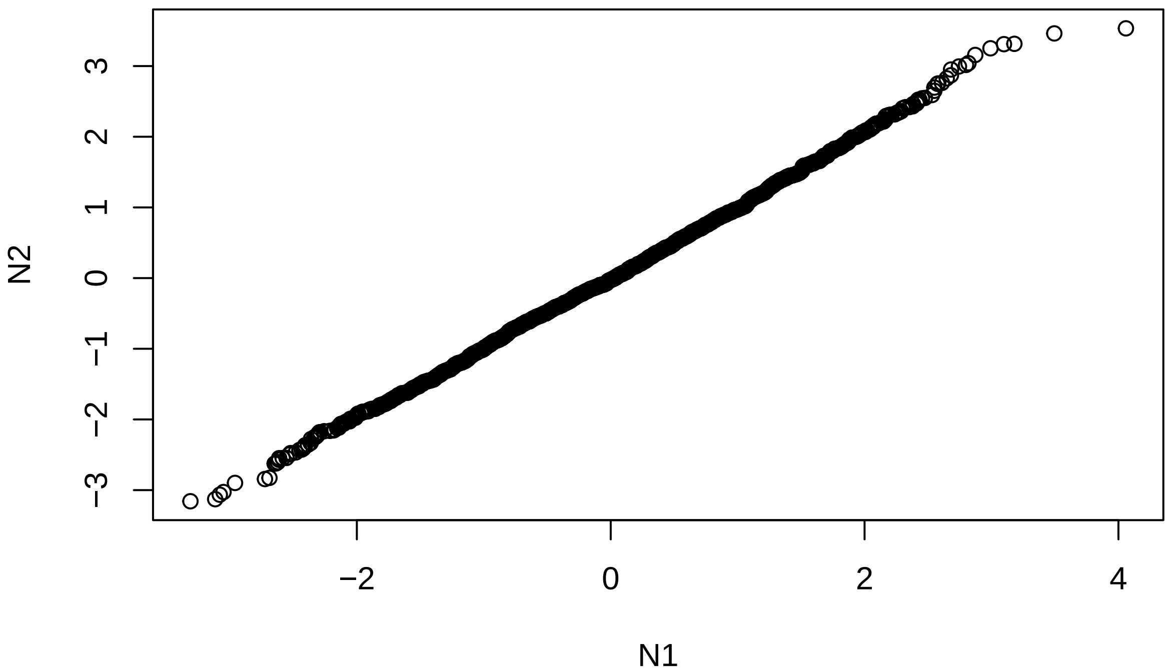

```
ggplot() +
  geom_histogram(aes(x=X, fill="X"), bins=30, alpha=0.5) +
  geom_histogram(aes(x=Y, fill="Y"), bins=30, alpha=0.5) +
  geom_density(aes(x=rnorm(3000), fill="Standard normal"), kernel = "gaussian", color = "black") +
  ggtitle("Histograms of X and Y with standard normal distribution")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



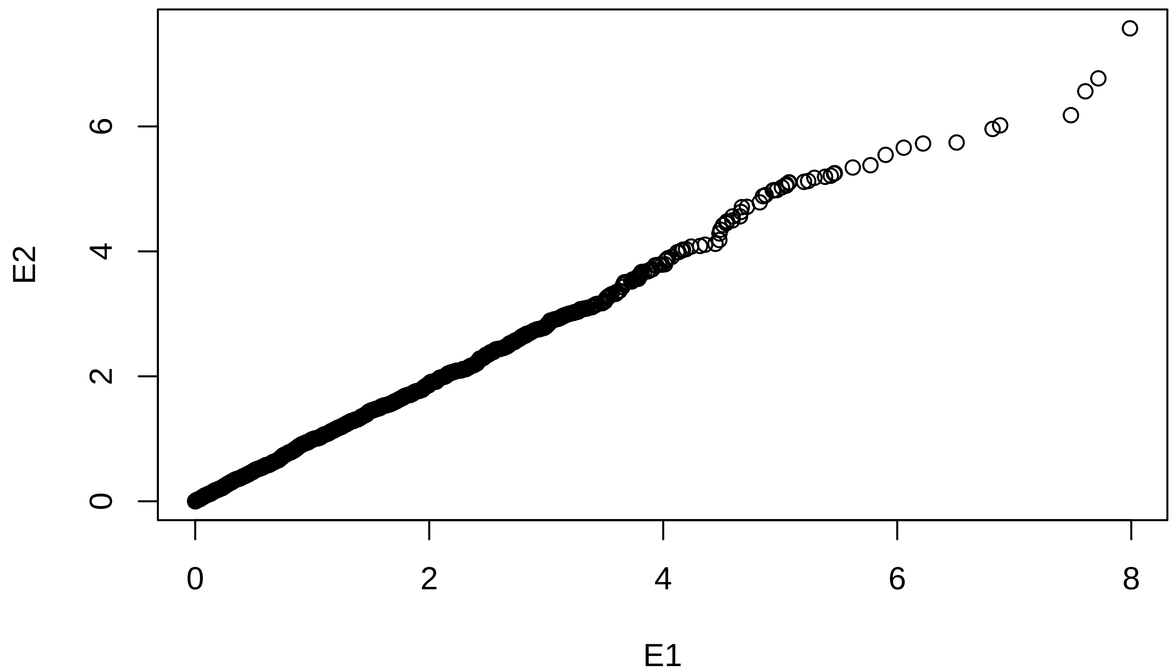
Question 19

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).

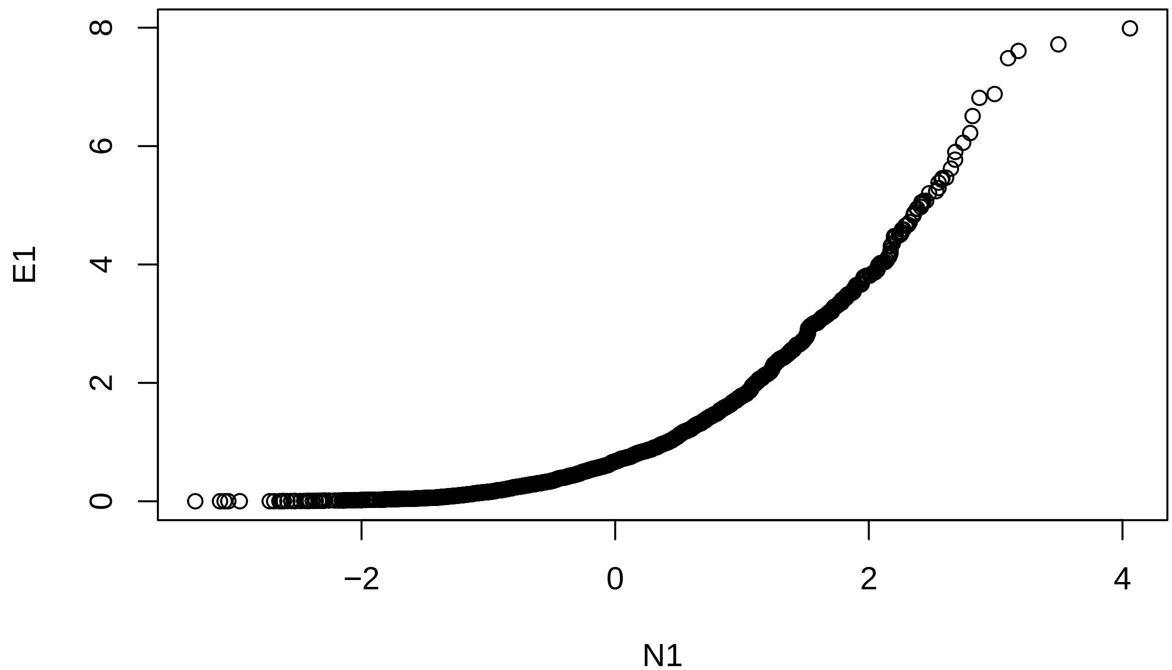
```
# generate samples
N1 <- rnorm(3000, mean = 0, sd = 1)
N2 <- rnorm(3000, mean = 0, sd = 1)
E1 <- rexp(3000, rate = 1)
E2 <- rexp(3000, rate = 1)

# plot Q-Q diagrams for N1 and N2, E1 and E2
qqplot(N1, N2)
```




```
qqplot(E1, E2)
```



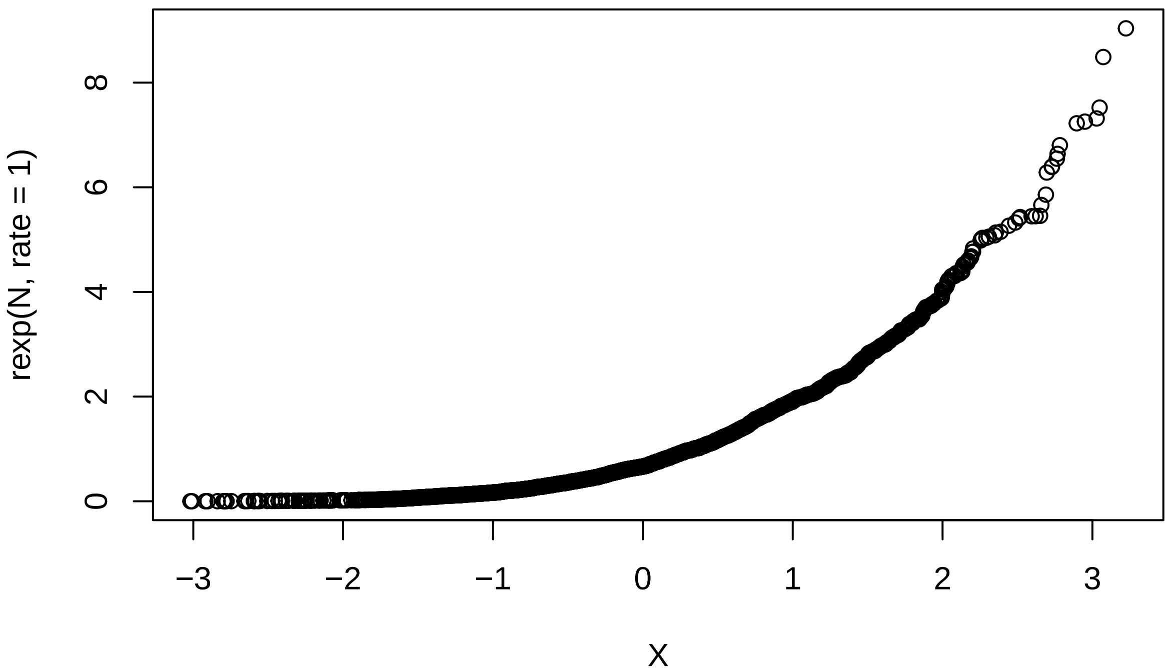

```
# plot Q-Q diagrams for N1 and E1
qqplot(N1, E1)
```



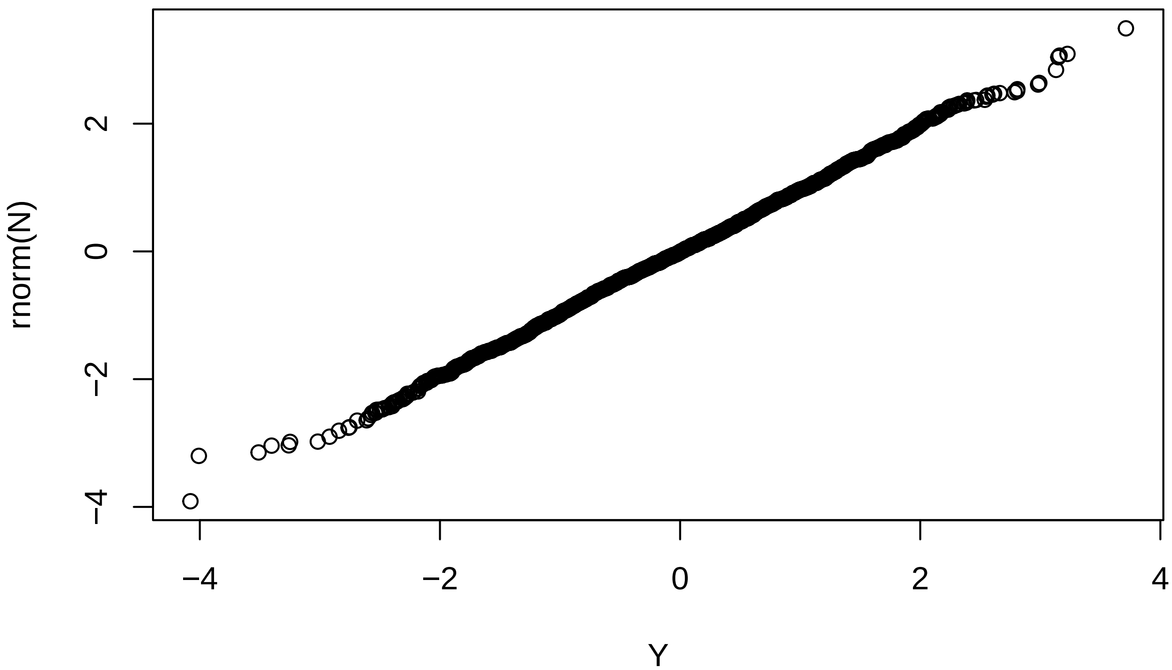
Question 20

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

```
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))
```


rnorm(N)
Y
−4
−2
0
2
4
−4
−2
0
2