BMI713 Problem Set 1

Instructions:

Please submit this problem set before class on Tuesday, October 31. Problem sets may be submitted within a week past the due date at a 20% penalty; each person is allowed to submit one problem late (within a week) without penalty. Please comment your code, because it is part of the requirements of each exercise. Missing comments will not allow the full score.

If you have any questions, please post on the piazza site. This problem set was prepared by Tiziana Sanavia and Giorgio Melloni, so they will be most prepared to answer questions.

1. Random variables and distributions (points:30)

A. Assume that a die is fair, i.e. if the die is rolled once, the probability of getting each of the six numbers is 1/6. Calculate the probability of the following events.

- Rolling the die once, what is the probability of getting a number less than 4? (points: 5)

Solution: Let X be a random variable denoting the value of a die roll. The probability that one particular die roll satisfies some criteria is:

$$\frac{\text{number of die rolls that satisfy the criteria}}{\text{number of possible die rolls}}$$

We can easily count both of these values. X < 4 means either X $= 1$ or X $= 2$ or X $= 3$. So the number of die rolls satisfying X < 4 is 3. There are 6 possible die rolls. Hence:

$$P(X < 4) = 3/6 = 1/2$$

- Rolling the die twice, what is the probability that the sum of two rolling numbers is less than 4? (points: 7)

Solution: Let X and Y be random variables such that X is the value of the first die roll and Y is the value of the second. We wish to find the number of rolls where $X + Y < 4$. It is clear that $X + Y < 4$ occurs only in the pairs $X = 1, Y = 1$, $X = 2, Y = 1$ and $X = 1, Y = 2$. Therefore, three combinations of rolls satisfy the criteria. There are 6 possible values for both X and Y, hence there are $6x6 = 36$ possible rolls. It follows that the probability of rolling a sum less than 4 is:

$$\frac{\text{number of rolls with X + Y} < 4}{\text{total number of possible rolls}} = 3/36 = 1/12$$

B. Let p be the probability of obtaining a head when flipping a coin. Suppose that Jake flipped the coin n ($n \geq 1$) times. Let X be the total number of head he obtained.

- What distribution does the random variable X follow? Is X a discrete or continuous random variable? (points: 5)

Solution: X follows the binomial distribution with n trials and probability of success p. This is a discrete distribution.

- What is the probability of getting k heads when flipping the coin n times, i.e. what is $P(X = k)(0 \leq k \leq n)$? (Write down the mathematical formula for calculating this probability). (points: 5)

Solution: In general, the probability that a binomially distributed random variable X ~ Bin(n; p) is equal to some value k is:

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

- Suppose p = 0.2 and n = 20. Calculate the probabilities $P(X = 4)$ and $P(X \geq 4)$. (You may need the functions `dbinom` and `pbinom` in R to calculate these two probabilities. Use `?dbinom` and `?pbinom` to get help information of these two functions). (points 8)

Solution: Using the general formula above with n = 20 and p = 0.2:

$$P(X = 4) = \binom{20}{4} (0.2)^4 (0.8)^{16} \approx 0.218$$

Using R's `dbinom`:

```
dbinom(x=4,size=20,prob=0.2)
```

```
## [1] 0.2181994
```

We may compute $P(X \geq 4)$ either by summing the probabilities for $X = 4$, $X = 5, ..., X = 20$ or we could subtract the probabilities for $X = 0$, $X = 1$, $X = 2$ and $X = 3$ from 1:

$$P(X \geq k) = P(X = k) + P(X = k+1) + ... + P(X = n)$$

$$P(X \geq k) = 1 - P(X < k) = 1 - [P(X = 0) + P(X = 1) + ... + P(X = k-1)]$$

The latter method is more convenient since it requires fewer terms:

$$\begin{aligned} P(X \geq 4) &= 1 - P(X < 4) = 1 - [P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3)] \\ &= 1 - \left[\binom{20}{0}(0.2)^0(0.8)^{20}\right] - \left[\binom{20}{1}(0.2)^1(0.8)^{19}\right] - \left[\binom{20}{2}(0.2)^2(0.8)^{18}\right] - \left[\binom{20}{3}(0.2)^3(0.8)^{17}\right] \\ &\approx 1 - 0.01152922 - 0.05764608 - 0.1369094 - 0.2053641 \\ &\approx 0.5885512 \end{aligned}$$

Again, using R:

```
1 - pbinom(q=3, size=20, prob=0.2)
```

```
## [1] 0.5885511
```

2. Normal Distribution and Z-score

A. The so-called BMI (Body Mass Index) is a measure of the weight-height-relation, and it is defined as the weight (W) in kg divided by the squared height (H) in meters:

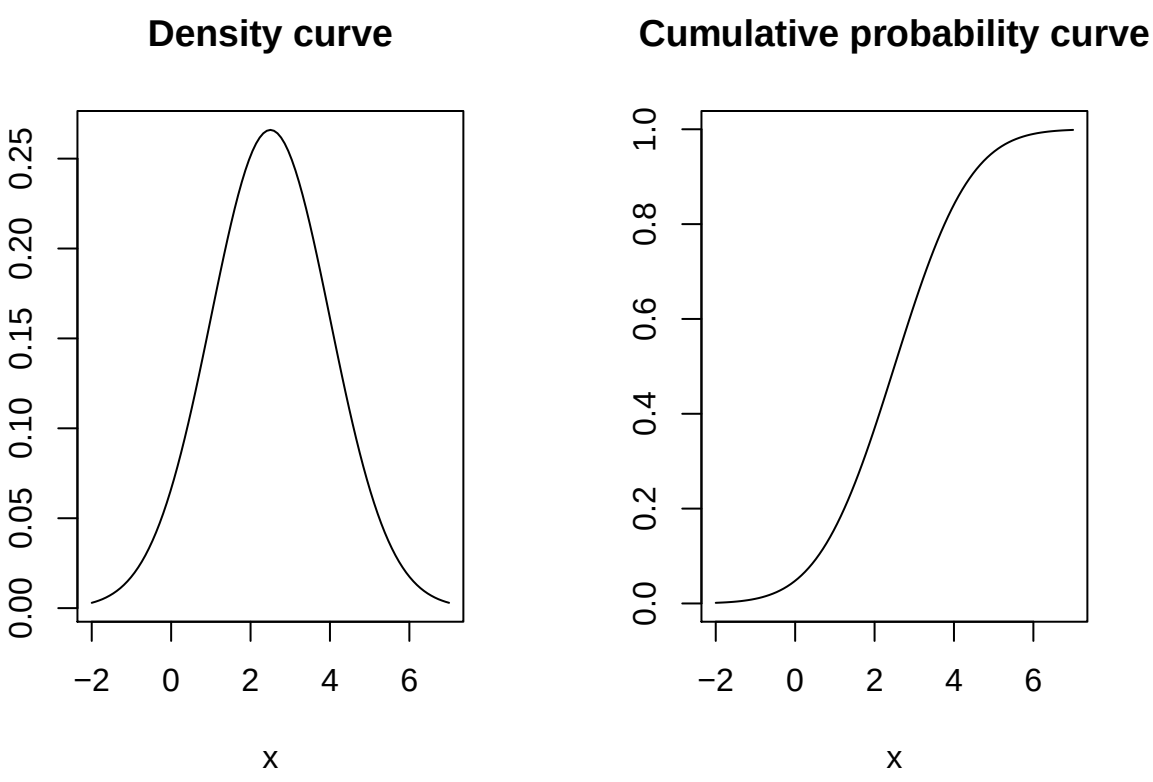
$$BMI = \frac{W}{H^2}$$

Assume that the population distribution of BMI is log-normal, therefore log(BMI) is a normal distribution with mean = 2.5 and variance = 2.25.

- Plot in R the density and the cumulative probability curves of log(BMI) as in the picture, using commands `dnorm` and `pnorm`: (points 4)

Solution:

```
par(mfrow=c(1,2))
curve(dnorm(x, mean = 2.5, sd = 1.5), from = 2.5 - 3 * 1.5, to = 2.5 + 3 * 1.5,
      ylab="", main= "Density curve") # sd = sqrt(2.25) = 1.5
curve(pnorm(x, mean = 2.5, sd = 1.5), from = 2.5 - 3 * 1.5, to = 2.5 + 3 * 1.5,
      ylab="", main= "Cumulative probability curve") # sd = sqrt(2.25) = 1.5
```



- Using the cumulative probability, calculate in R the area under the density curve between x=0.5 and x=4. (points: 6)

Solution:

We have to compute the probability $P(0.5 < x < 4) = P(x < 4) - (x < 0.5)$ with x $(log(BMI))$ following a normal distribution $N(2.5, 1.5)$. Let:

$$Z = \frac{x - 2.5}{1.5}$$

we have

$$P(x < 4) = P\left(Z < \frac{4 - 2.5}{1.5}\right) = P(Z < 1) \approx 0.8413447$$

$$P(x < 0.5) = P\left(Z < \frac{0.5 - 2.5}{1.5}\right) \approx P(Z < -1.33)$$

We can use symmetry to handle the negative value:

$$P(Z < -1.33) = P(Z > 1.33) = 1 - P(Z < 1.33) \approx 1 - 0.9087887 = 0.0912113$$

Finally,

$$P(0.5 < x < 4) = P(x < 4) - (x < 0.5) \approx 0.8413447 - 0.0912113 = 0.7501334$$

We may also calculate the probability using the following command in R:

```
pnorm(4, mean = 2.5, sd = 1.5) - pnorm(0.5, mean = 2.5, sd = 1.5)
```

```
## [1] 0.7501335
```

- A definition of "being obese" is a BMI-value of at least 30. How large a proportion of the population would then be obese? (points: 6)

Solution:

The distribution of BMI is log-normal and log(BMI) is normally distributed, therefore:

$$P(BMI \geq 30) = P(log(BMI) \geq log(30)) = P(Z \geq \frac{log(30) - 2.5}{1.5} = P(Z \geq 0.6) = 0.27425$$

In R we can subtract the area to the left of log(30) from 1. NOTE: here we have a continuous distribution, therefore $P(BMI \leq 30) = P(BMI < 30)$ (the probability that a continuous random variable will assume a particular value is zero).

```
1-pnorm((log(30)-2.5)/1.5)
```

```
## [1] 0.2739872
```

or

```
pnorm((log(30)-2.5)/1.5, lower.tail = FALSE)
```

```
## [1] 0.2739872
```

or

```r
pnorm(log(30), mean = 2.5, sd = 1.5, lower.tail=FALSE)
```

```
## [1] 0.2739872
```

- The 90th percentile of the BMI is the value such that 90% of the population has a BMI lower than this value. Find the 90th percentile for log(BMI) using `qnorm`. (points: 6)

Solution:

```r
qnorm(0.9,mean=2.5,sd=1.5)
```

```
## [1] 4.422327
```

B. Assume that blood-glucose levels in a population of adult women are normally distributed with mean 90 mg/dL and standard deviation 38 mg/dL. Answer the following questions:

- What percentage of women shows levels above or equal to 80.5 mg/dL?

Solution: We know that the blood-glucose levels follow a normal distribution with mean = 90 and standard deviation = 38. The solution we are searching for is then $P(X \geq 80.5)$. Let's standardize the normal variable and then determine the probability from the table of the cumulative distribution. Hence:

$$Z = (80.5 - 90)/38 = -0.25$$

Since the standard normal distribution is symmetric around 0, we know that $P(Z > -0.25) = P(Z < 0.25) \approx 0.5987 \approx 60\%$.

In R code:

```r
1-pnorm(80.5, mean=90, sd=38)
```

```
## [1] 0.5987063
```

or

```r
pnorm(80.5, mean=90, sd=38, lower.tail=F)
```

```
## [1] 0.5987063
```

or

```r
pnorm(-0.25, lower.tail=F)
```

```
## [1] 0.5987063
```

- Suppose that the "abnormal range" is defined to be glucose levels which are 1.5 standard deviations above the mean or 1.5 standard deviations below the mean. What percentage of women would be classified "abnormal"? (points: 6)

Solution: This is same as asking what percentage of standard normal distribution is below $Z = -1.5$ or above $Z = 1.5$. For standard normal the probability less than -1.5 is 0.0668. By the symmetry of the distribution the probability above 1.5 is the same. So answer is 2*0.0668=13.36%.

In R:

```
pnorm(1.5, lower.tail = F) + pnorm(-1.5)
```

```
## [1] 0.1336144
```

or

```
pnorm(-1.5) * 2
```

```
## [1] 0.1336144
```

or

```
pnorm(90 + 38*1.5, mean = 90, sd = 38, lower.tail = F) + pnorm(90 - 38*1.5, mean = 90, sd = 38)
```

```
## [1] 0.1336144
```

- Suppose now that we want to redefine the abnormal range to be more than 'c' standard deviations above the mean or less than 'c' standard deviations with 'c' chosen so that 4 % of the population will be classified as abnormal. What should 'c' be? (points: 6)

Solution: By symmetry there will be 2% below the mean minus 'c' standard deviations. For the standard normal, we can use `qnorm` to find the z-score of the 2nd percentile to obtain 'c'. Therefore:

```
abs(qnorm(0.04/2))
```

```
## [1] 2.053749
```

which is equivalent to:

```
(90-qnorm(0.04/2,90,38))/38
```

```
## [1] 2.053749
```

3. Simulation of distributions of random variables (points: 30)

Consider X a random variable from any distribution with mean μ and variance σ^2.

If we sample n values from that distribution, we can calculate the mean value $\bar{x_n}$ which is itself the realization of a random variable $\bar{X}_n$.

In this exercise we will evaluate some properties of the:

3.1 Normal Distribution (points: 15)

Using `rnorm` create a vector of 1000 values from a normal distribution with $\mu = 0$ and $\sigma = 1$. We call this vector $m0$. (points: 1)

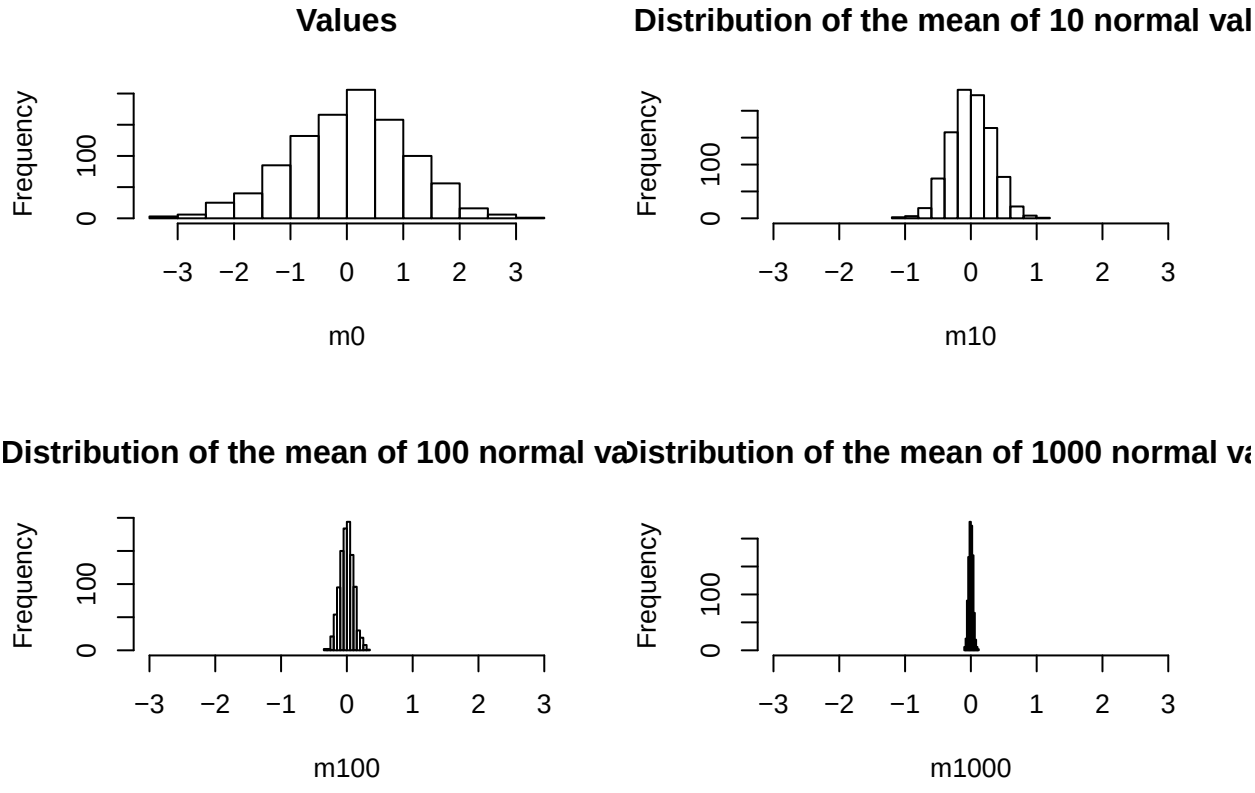
Using the same command, create a vector of $N = 1000$ mean values from a random sampling of $n = 10, 100$ and 1000 elements. (points: 1) We will call these vectors $m10$, $m100$, $m1000$.

Create a 4 panels plot (You can use an histogram or a density plot or both) showing the distributions of: (points: 2)

1) The 1000 values from the distribution ($m0$)
2) The 1000 means using n = 10 ($m10$)
3) The 1000 means using n = 100 ($m100$)
4) The 1000 means using n = 1000 ($m1000$)

Solution:

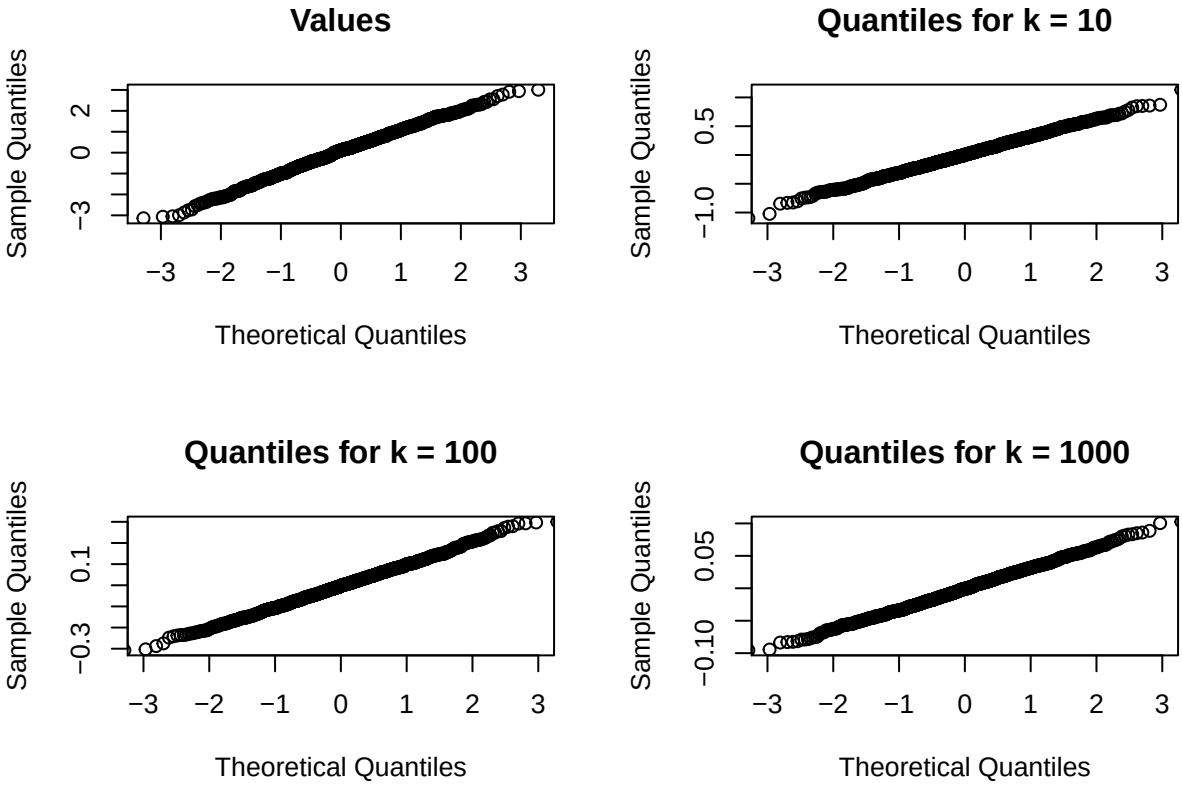
```
# Create the vectors
m0 <- rnorm(n = 1000 , mean = 0 , sd = 1)
m10 <- replicate(n = 1000 , expr = mean(rnorm(n = 10 , mean = 0 , sd = 1)))
m100 <- replicate(n = 1000 , expr = mean(rnorm(n = 100 , mean = 0 , sd = 1)))
m1000 <- replicate(n = 1000 , expr = mean(rnorm(n = 1000 , mean = 0 , sd = 1)))

# Plot their values
par(mfrow=c(2,2))
hist(m0 , main="Values")
hist(m10 , main="Distribution of the mean of 10 normal values" , xlim = c(-3 , 3))
hist(m100 , main="Distribution of the mean of 100 normal values" , xlim = c(-3 , 3))
hist(m1000 , main="Distribution of the mean of 1000 normal values" , xlim = c(-3 , 3))
```



Using the function `qqnorm`, compare theoretical and sample quantiles of a normal distribution. Do the distributions look normal? (points: 3)

Solution: All the sample distributions look normal.

```
par(mfrow=c(2,2))
qqnorm(m0 , main="Values")
qqnorm(m10 , main="Quantiles for k = 10" , xlim = c(-3 , 3))
qqnorm(m100 , main="Quantiles for k = 100" , xlim = c(-3 , 3))
qqnorm(m1000 , main="Quantiles for k = 1000" , xlim = c(-3 , 3))
```



Now evaluate the value of the mean and variance of each of the 4 vectors.

- Are the means substantially different from each other? (points: 2)

Solution: The means are not substantially different from each other. However, the increase of k allows a better estimation of the true mean (0).

- Are the variances different from each other? If yes, what is the ratio between $Var(m0)$ and the other variances? (e.g., $Var(m0)/Var(m10)$, $Var(m0)/Var(m100)$) (points: 3)

Solution: The variances have a 10 fold ratio between each other.

- If you see any pattern, can you derive a general formula to derive the Variance of any distribution of the means $\bar{X_n}$ for any given n (points: 3)

Solution: The standard deviation is in $1/sqrt(n)$ of the original SD and the Variance is:

$$Var(X_n) = Var(X)/n$$

3.2 Non-normal distribution (points: 15, evaluated as 3.1)

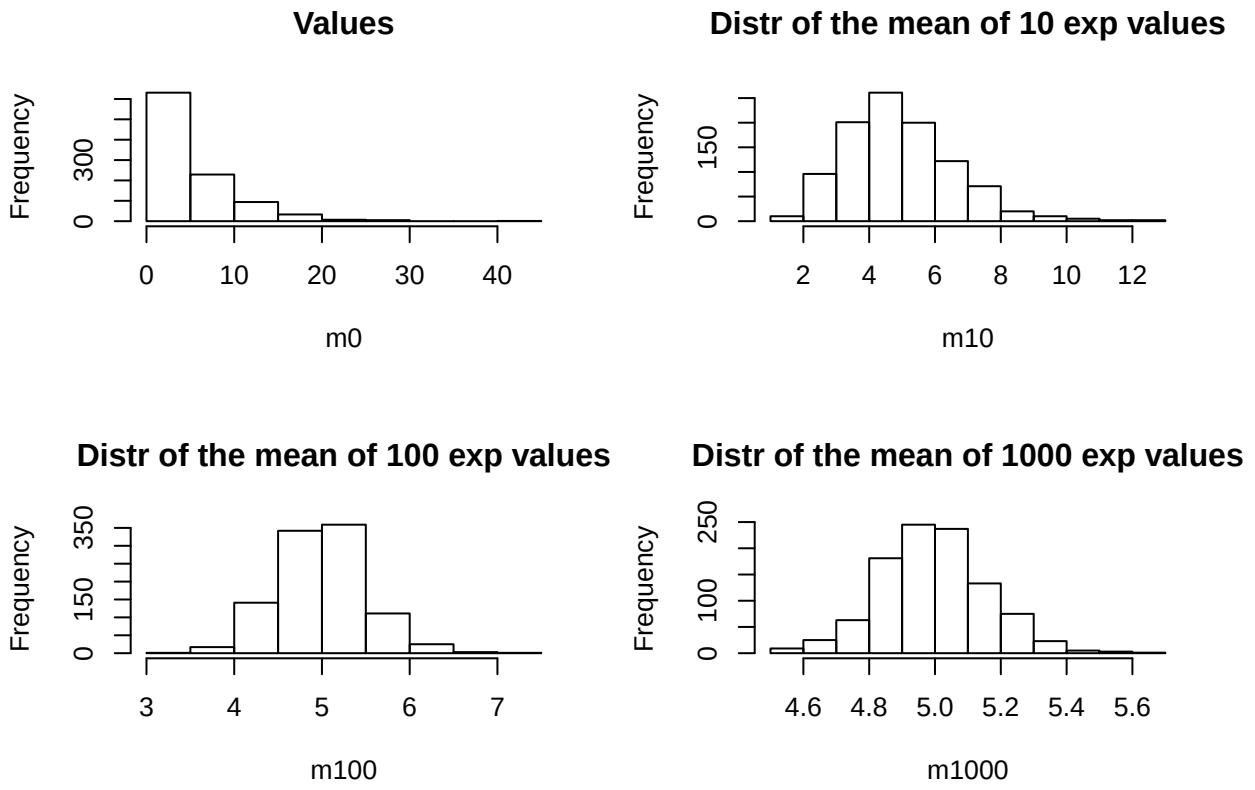
Repeat the exercise 3.1 but using a different random variable following the exponential distribution, $f(x) = \lambda e^{-\lambda x}$. To run this simulation use the function `rexp` with rate (i.e. λ) value **1**.

- Plot the distribution of the 4 vectors

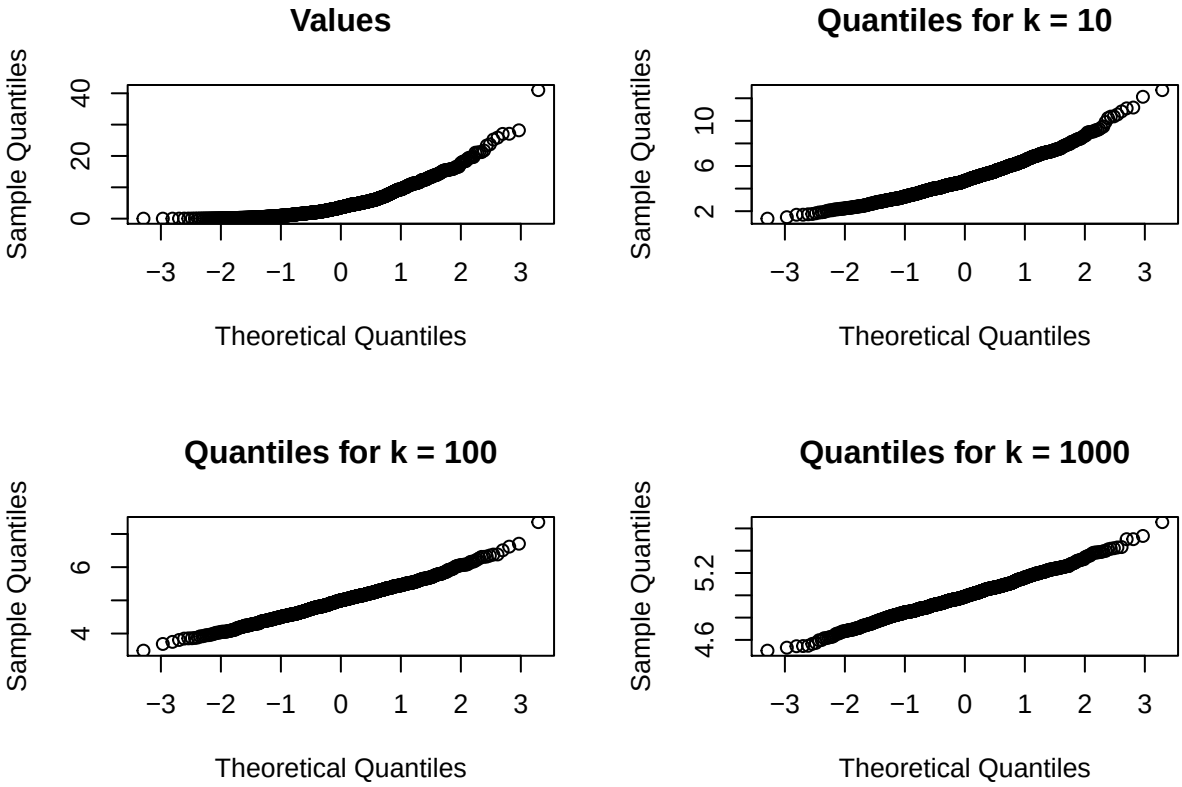
Solution:

```
m0 <- rexp(n = 1000 , rate =0.2)
m10 <- replicate(n = 1000 , expr = mean(rexp(n = 10 , rate =0.2)))
m100 <- replicate(n = 1000 , expr = mean(rexp(n = 100 , rate =0.2)))
m1000 <- replicate(n = 1000 , expr = mean(rexp(n = 1000 , rate =0.2)))
```

```
par(mfrow=c(2,2))
hist(m0 , main="Values")
hist(m10 , main="Distr of the mean of 10 exp values" )
hist(m100 , main="Distr of the mean of 100 exp values" )
hist(m1000 , main="Distr of the mean of 1000 exp values")
```




```
par(mfrow=c(2,2))
qqnorm(m0 , main="Values")
qqnorm(m10 , main="Quantiles for k = 10" )
qqnorm(m100 , main="Quantiles for k = 100")
qqnorm(m1000 , main="Quantiles for k = 1000")
```



- Using `qqnorm` like above, evaluate normality. Are the exponential values normally distributed? What about the means?

Solution: Exponential values are not normally distributed but the means are.

- Evaluate mean and variance as above

Solution: Although less clear than before, the ratio between Var(m0) and Var(m10) is ~10 and the formula is the same as above.