Lab 1

Bangda Sun, bs2996

September 13, 2016

Instructions

Before you leave lab today make sure that you upload an RMarkdown file to the canvas page (this should have a .Rmd extension) as well as the HTML output after you have knitted the file (this will have a .html extension). Note that since you have already knitted this file, you should see both a **Lab1_UNI.html** and a **Lab1_UNI.Rmd** file in your GR5206 folder. Click on the **Files** tab to the right to see this. The files you upload to the Canvas page should be updated with commands you provide to answer each of the questions below. You can edit this file directly to produce your final solutions. Partners, groups?

Background: The Normal Distribution

Recall from your probability class that a random variable X is normally-distributed with mean μ and variance σ^2 (denoted $X \sim N(\mu, \sigma^2)$) if it has a probability density function, or *pdf*, equal to

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

In *R* we can simulate $N(\mu, \sigma^2)$ random variables using the `rnorm()` function. For example,

```
rnorm(n = 5, mean = 10, sd = 3)
```

```
## [1]  8.120639 10.550930  7.493114 14.785842 10.988523
```

outputs 5 normally-distributed random variables with mean equal to 10 and standard deviation (this is σ) equal to 3. If the second and third arguments are ommited the default rates are **mean = 0** and **sd = 1**, which is referred to as the "standard normal distribution".

Tasks

Sample means as sample size increases

1) Generate 100 random draws from the standard normal distribution and save them in a vector named **normal100**. Calculate the mean and standard deviation of **normal100**. In words explain why these values aren't exactly equal to 0 and 1.

```
normal100 <- rnorm(n = 100, mean = 0, sd = 1); normal100
```

```
##   [1] -0.820468384  0.487429052  0.738324705  0.575781352 -0.305388387
##   [6]  1.511781168  0.389843236 -0.621240581 -2.214699887  1.124930918
##  [11] -0.044933609 -0.016190263  0.943836211  0.821221195  0.593901321
##  [16]  0.918977372  0.782136301  0.074564983 -1.989351696  0.619825748
##  [21] -0.056128740 -0.155795507 -1.470752384 -0.478150055  0.417941560
##  [26]  1.358679552 -0.102787727  0.387671612 -0.053805041 -1.377059557
##  [31] -0.414994563 -0.394289954 -0.059313397  1.100025372  0.763175748
##  [36] -0.164523596 -0.253361680  0.696963375  0.556663199 -0.688755695
```

```
##  [41] -0.707495157  0.364581962  0.768532925 -0.112346212  0.881107726
##  [46]  0.398105880 -0.612026393  0.341119691 -1.129363096  1.433023702
##  [51]  1.980399899 -0.367221476 -1.044134626  0.569719627 -0.135054604
##  [56]  2.401617761 -0.039240003  0.689739362  0.028002159 -0.743273209
##  [61]  0.188792300 -1.804958629  1.465554862  0.153253338  2.172611670
##  [66]  0.475509529 -0.709946431  0.610726353 -0.934097632 -1.253633400
##  [71]  0.291446236 -0.443291873  0.001105352  0.074341324 -0.589520946
##  [76] -0.568668733 -0.135178615  1.178086997 -1.523566800  0.593946188
##  [81]  0.332950371  1.063099837 -0.304183924  0.370018810  0.267098791
##  [86] -0.542520031  1.207867806  1.160402616  0.700213650  1.586833455
##  [91]  0.558486426 -1.276592208 -0.573265414 -1.224612615 -0.473400636
##  [96] -0.620366677  0.042115873 -0.910921649  0.158028772 -0.654584644
```

```
mean_normal100 <- mean(normal100); mean_normal100
```

```
## [1] 0.08256659
```

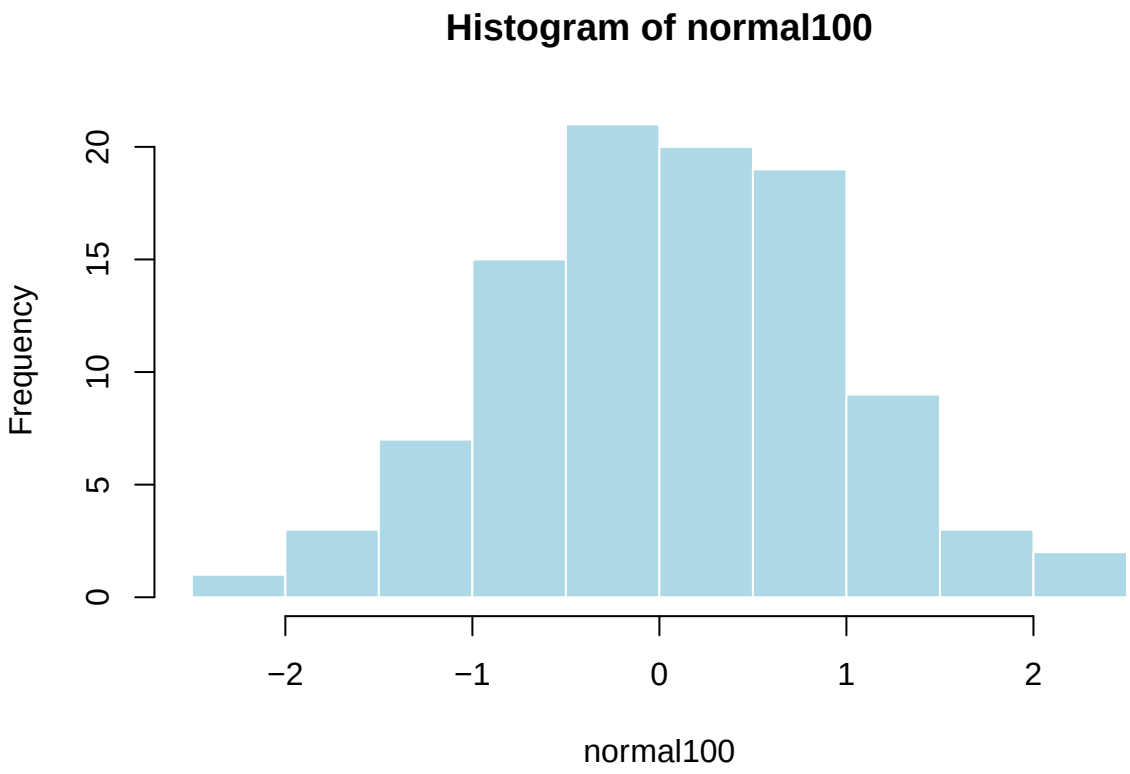
```
sd_normal100 <- sd(normal100); sd_normal100
```

```
## [1] 0.8891336
```

Because **normal100** is just one of the simple random samples of the population, 0 is the expectation of the sample mean, it is the unbiased estimates of the sample mean. But for one sample, it will not be exactly 0. So does the standard deviation.

2) The function **hist()** is a base *R* graphing function that plots a histogram of its input. Use **hist()** with your vector of standard normal random variables from question (1) to produce a histogram of the standard normal distribution. Remember that typing **?hist** in your console will provide help documents for the **hist()** function. If coded properly, these plots will be automatically embedded in your output file.

```
hist(normal100, col = "lightblue", border = "white")
```



3) Repeat question (1) except change the number of draws to 10, 1000, 10,000, and 100,000 storing the results in vectors called **normal10**, **normal1000**, **normal10000**, **normal100000**.

```
normal10 <- rnorm(n = 10, mean = 0, sd = 1)
normal1000 <- rnorm(n = 1000, mean = 0, sd = 1)
normal10000 <- rnorm(n = 10000, mean = 0, sd = 1)
normal100000 <- rnorm(n = 100000, mean = 0, sd = 1)
```

4) We want to compare the means of our four random draws. Create a vector called **sample_means** that has as its first element the mean of **normal10**, its second element the mean of **normal100**, its third element the mean of **normal1000**, its fourth element the mean of **normal10000**, and its fifth element the mean of **normal100000**. After you have created the **sample_means** vector, print the contents of the vector and use the **length()** function to find the length of this vector. (it should be five). There are, of course, multiple ways to create this vector. Finally, explain in words the pattern we are seeing with the means in the **sample_means** vector.

```
sample_means <- rep(0, 5)
sample_means[1] <- mean(normal10)
sample_means[2] <- mean(normal100)
sample_means[3] <- mean(normal1000)
sample_means[4] <- mean(normal10000)
sample_means[5] <- mean(normal100000)
print(sample_means)
```

```
## [1]  0.493735437  0.082566589 -0.026875723 -0.006719807 -0.001114476
```

```
length(sample_means)
```

```
## [1] 5
```

First we create an initial vector with 5 elements, than we assign the vector elements with its corresponding sample mean.

5) Let's push this a little farther. Generate 1 million random draws from a normal distribution with $\mu = 3$ and $\sigma^2 = 4$ and save them in a vector named **normal1mil**. Calculate the mean and standard deviation of **normal1mil**.

```
normal1mil <- rnorm(1000000, mean = 3, sd = 2)
mean_normal1mil <- mean(normal1mil); mean_normal1mil
```

```
## [1] 3.000241
```

```
sd_normal1mil <- sd(normal1mil); sd_normal1mil
```

```
## [1] 1.999961
```

6) Find the mean of all the entries in **normal1mil** that are greater than 3. You may want to generate a new vector first which identifies the elements that fit the criteria.

```
x <- which(normal1mil > 3) # identify the elements greater than 3
normal1mil_greatthan3 <- normal1mil[x] # elements greater than3
mean(normal1mil_greatthan3)
```

```
## [1] 4.596158
```

7) Create a matrix **normal1mil_mat** from the vector **normal1mil** that has 10,000 columns (and therefore should have 100 rows).

```
normal1mil_mat <- matrix(normal1mil, ncol = 10000)
```

8) Calculate the mean of the 1234^{th} column.

```
mean_col1234 <- mean(normal1mil_mat[, 1234]); mean_col1234
```

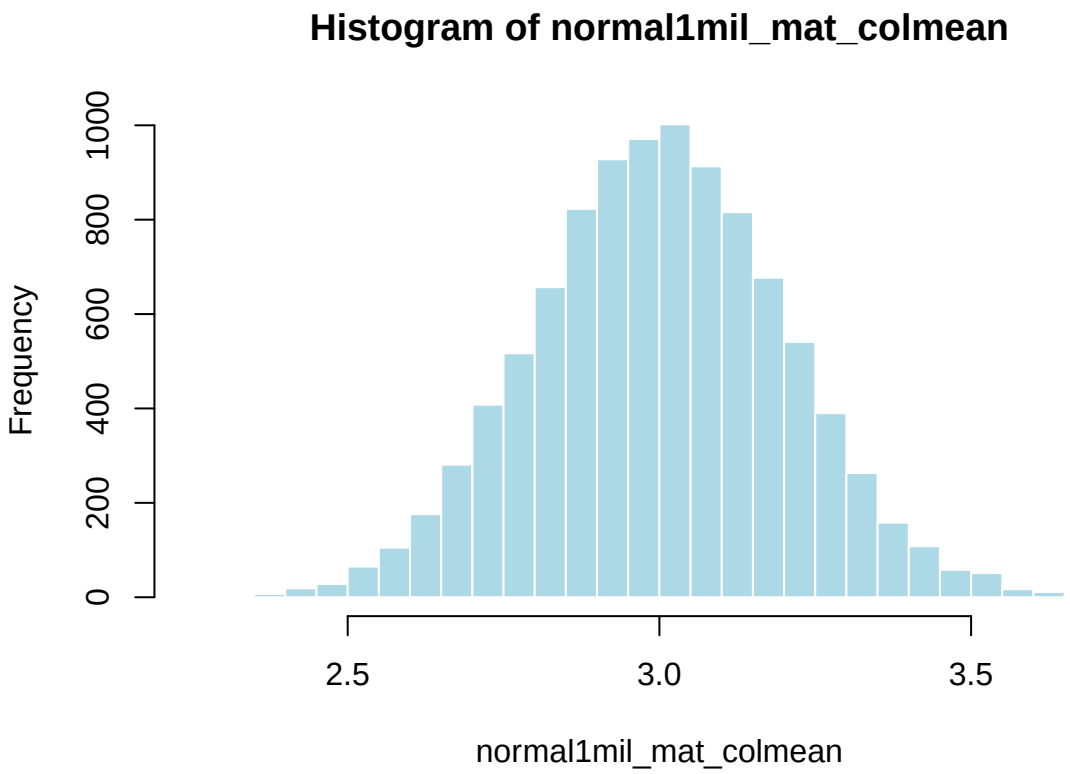
```
## [1] 3.056545
```

9) Use the **colSums()** functions to calculate the *means* of each column of **normal1mil_mat**. Remember, **?colSums** will give you help documents about this function. Save the vector of column means with an appropriate name as it will be used in the next task.

```
normal1mil_mat_colsum <- colSums(normal1mil_mat)
normal1mil_mat_colmean <- normal1mil_mat_colsum / dim(normal1mil_mat)[1]
```

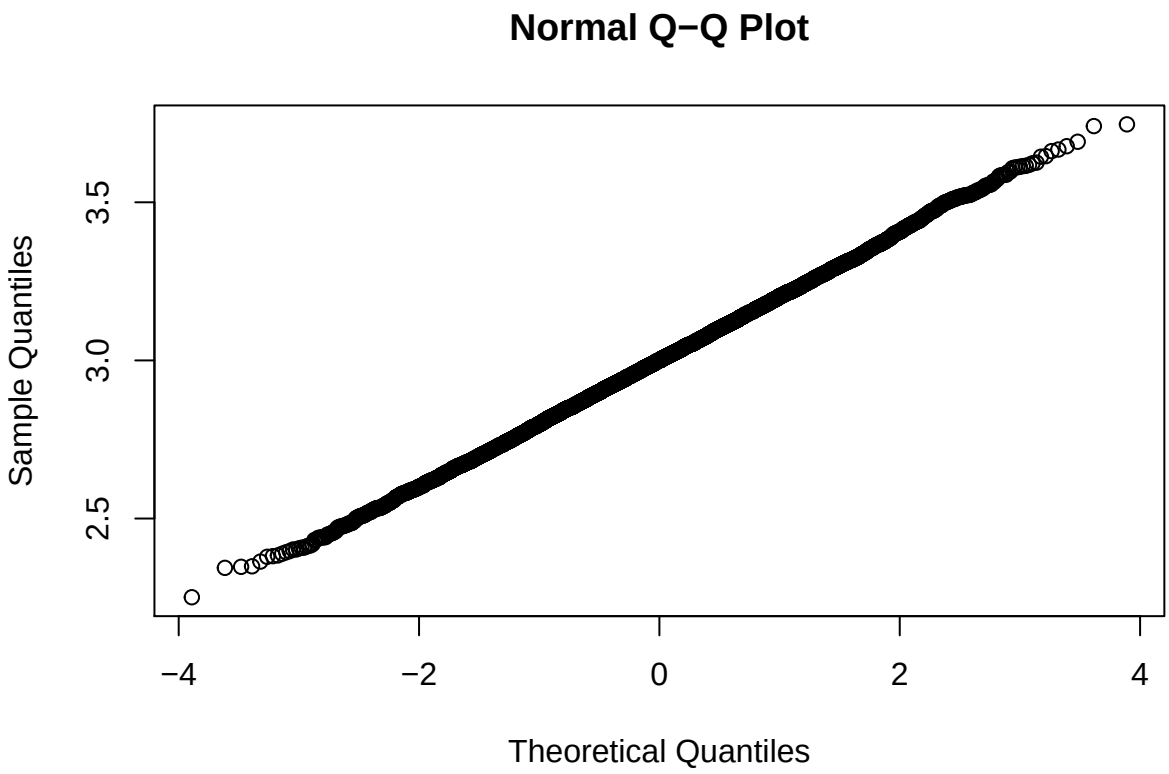
10) Finally, produce a histogram of the column means you calculated in task (9). What is the distribution that this histogram approximates (i.e. what is the distribution of the sample mean in this case)?

```
h <- hist(normal1mil_mat_colmean, breaks = 50, col = "lightblue", border = "white")
```



According to Central Limit Theorem, the distribution approximates to Normal distribution. The Q-Q plot of **normal1mil_mat_colmean** is almost a straight line :

```
qqnorm(normal1mil_mat_colmean)
```


Normal Q−Q Plot
Sample Quantiles
2.5
3.0
3.5
−4
−2
0
2
4
Theoretical Quantiles