Statistical Inference - Course Project 1

Ivana Peretin

July, 2015

Overview

The purpose of this project is to use simulation to explore inference. We will investigate exponential distribution and compare it with the Central Limit Theorem. Via simulation and associated explanatory text we will illustrate the properties of the distribution of the mean of 40 exponentials through the following points:

1. Compare the sample mean to the theoretical mean of the distribution.
2. Compare the sample variance to the theoretical variance of the distribution.
3. Show that the distribution is approximately normal.

Simulations

The exponential distribution will be simulated in R with *rexp(n, lambda)* where lambda is the rate parameter. The mean of exponential distribution is 1/lambda and the standard deviation is also 1/lambda. For all of the simulations we will set lambda = 0.2. We will create a thousand simulations of 40 exponentials and store it all in a matrix with 1000 rows and 40 columns.

```
set.seed(12345)

lambda <- 0.2
n_exp <- 40
n_sim <- 1000

simulations <- matrix(rexp(n_exp*n_sim, lambda), n_sim, n_exp)
```

Table 1: Summary example for a sample of 40 exponentials

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.006411	1.332	3.768	4.735	5.986	20.51

Sample mean vs. Theoretical mean

Our next step will be to calculate a mean value for each sample of 40 exponentials. Calculated mean values will be shown in a histogram and compared to the theoretical mean of the sampling distribution which is equal to 5.

```
obs_means <- rowMeans(simulations)
```

Table 2: Simulated sample means summary

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
3.032	4.424	4.938	4.972	5.492	8.38

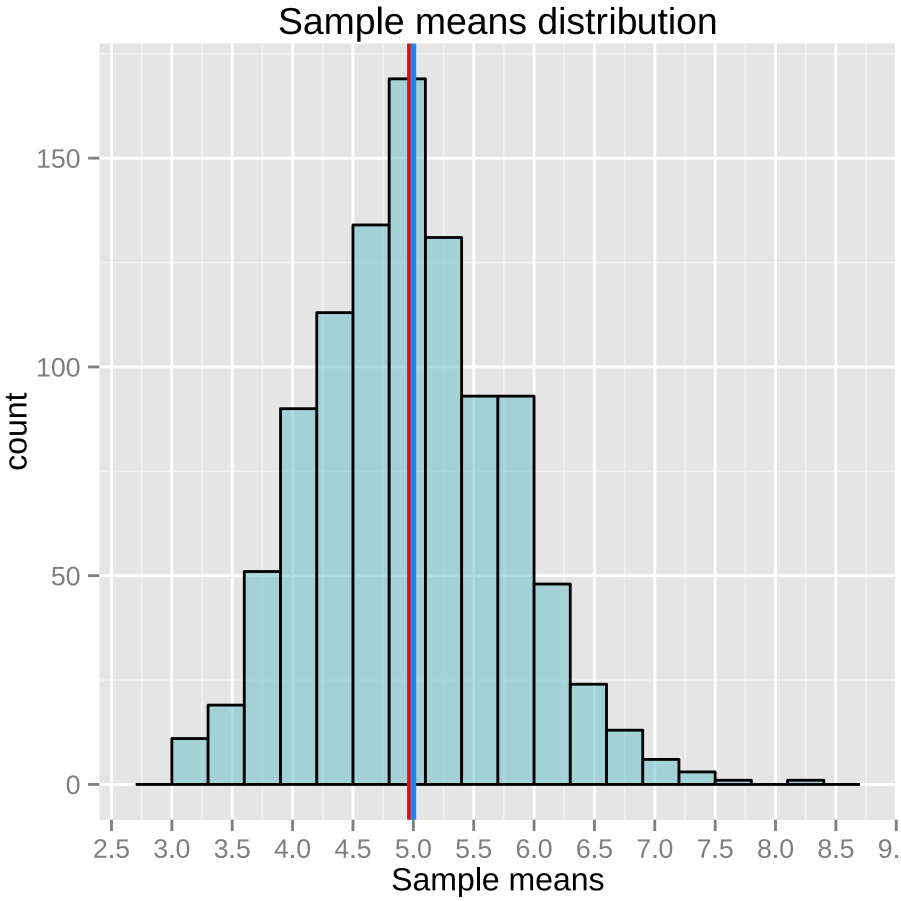

Figure 1: Comparison of sample mean (red line) and theoretical mean (blue line

Figure 1 shows the distribution of simulated sample means. Blue vertical line marks the theoretical mean and red line marks the observed mean (of sample means) of the sampling distribution. Observed mean is equal to **4.97** which is very close to the theoretical mean **5**.

Sample variance vs. Theoretical variance

It was stated in lecture notes that the variance of the sample mean is the population variance divided by the sample size. Therefore, our first step will be to calculate variance for each sample of 40 exponentials. Our sampling distribution is Exp(0.2) and we would expect that mean value of the 1000 sample variances will be near the theoretical value of **25**. As we can see from the table below, mean observed variance is equal to **24.57** which is very close to 25.

```
obs_vars <- apply(simulations, 1, var)
```

Table 3: Summary of simulated sample variances

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
6.382	17.29	22.79	24.57	29.47	96.98

Taking all this into account, the theoretical variance of sample mean is $25/40 = \mathbf{0.625}$. Now, lets divide the calculated sample variances with sample size. From the table below we can see that the mean variance of the sample mean is **0.614** which is near the theoretical value 0.625.

```r
obs_vars_mn <- apply(simulations, 1, function(x){var(x)/n_exp})
```

Table 4: Summary of simulated sample mean variances

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.1596	0.4324	0.5699	0.6142	0.7366	2.424

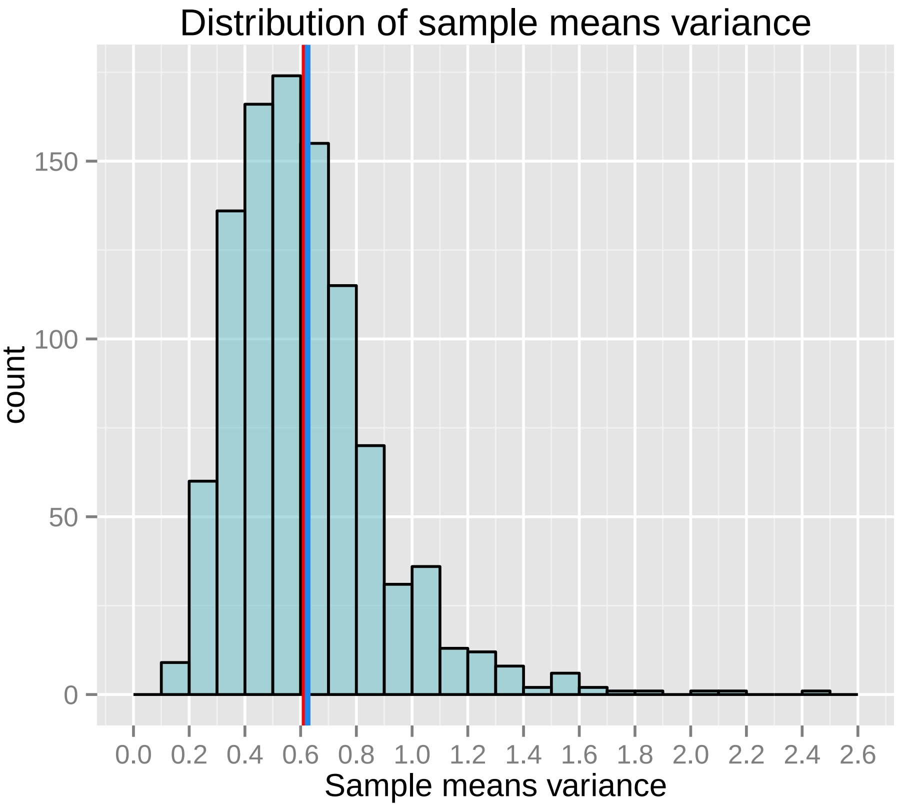

Figure 2: Comparison of variance of sample mean (red line) and theoretical variance (blue line)

Figure 2 shows the distribution of variances of sample means. Blue vertical line marks the theoretical variance and red line marks the mean observed variance (of the sample variances) of the sampling distrbution. Note that distribution is skewed to the right and not centered around what it's estimating (population variance). This could be fixed by increasing the sample size.

Comparison with Normal distribution

Finally, lets compare distribution of sample means to normal distribution.

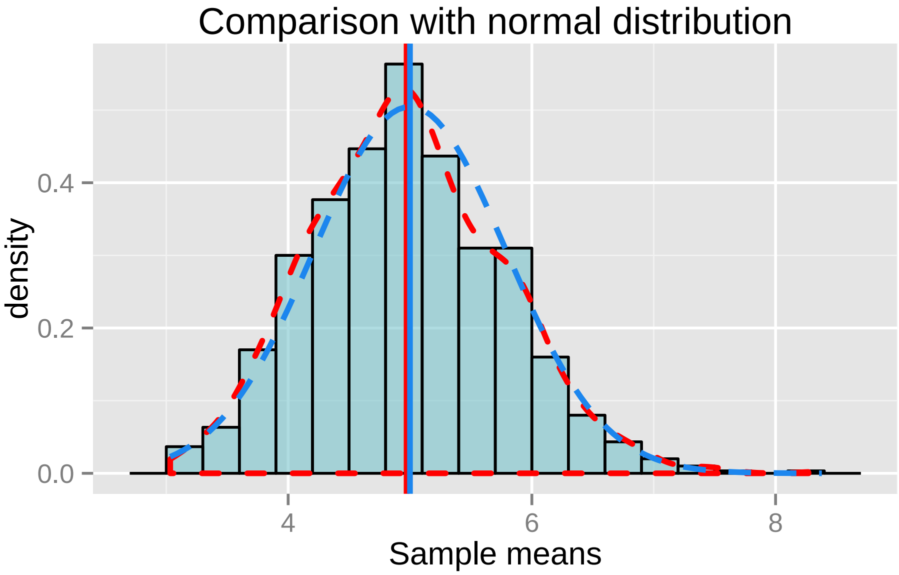

Figure 3: Comparison of simulated sample means distribution (red) and normal distribution (blue)

The figure above shows that the density of sample means (red dashed line) fits closely to a normal distribution (blue dashed line) plotted with theoretical mean and variance. The qqplot below also suggest normality. Therefore, we can conclude that due to the Central Limit Theorem the distribution of sample means is approximately N(5, 0.625). Source code for this report can be found on GitHub.

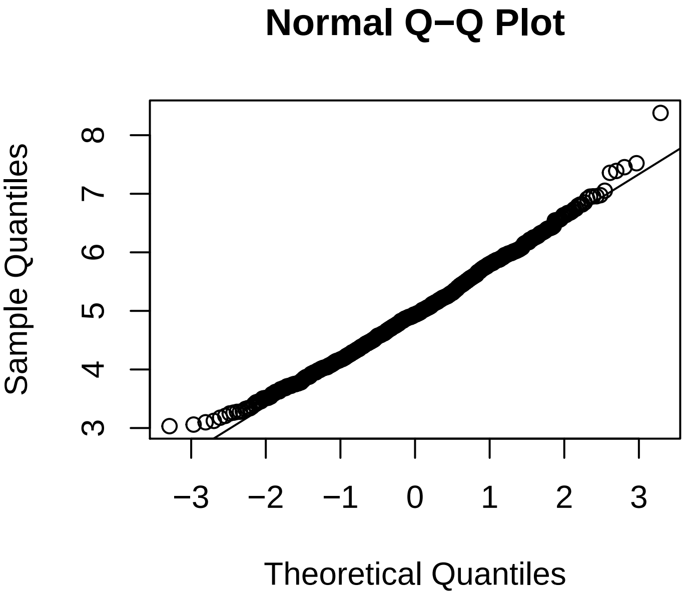

Figure 4: Comparison of sample quantiles and normal quantiles