Statistical Inference Course Project Part 1

Melody Wolk

June 8, 2015

Synopsis

In this project we investigate the exponential distribution in R and compare it with the Central Limit Theorem. The exponential distribution is simulated in R using 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. In the following we set lambda = 0.2 for all of the simulations.

We investigate the distribution of averages of 40 exponentials using a thousand of simulations.

Setting the parameters

```
# set lambda to 0.2
lambda <- 0.2
# 40 samples
n <- 40
# 1000 simulations
simulations <- 1000
# set seed for reproducability
set.seed(1)
```

Performing the simulations and extracting the mean for each

```
simu <- replicate(simulations, rexp(n, lambda))
mean_exp <- apply(simu, 2, mean)
```

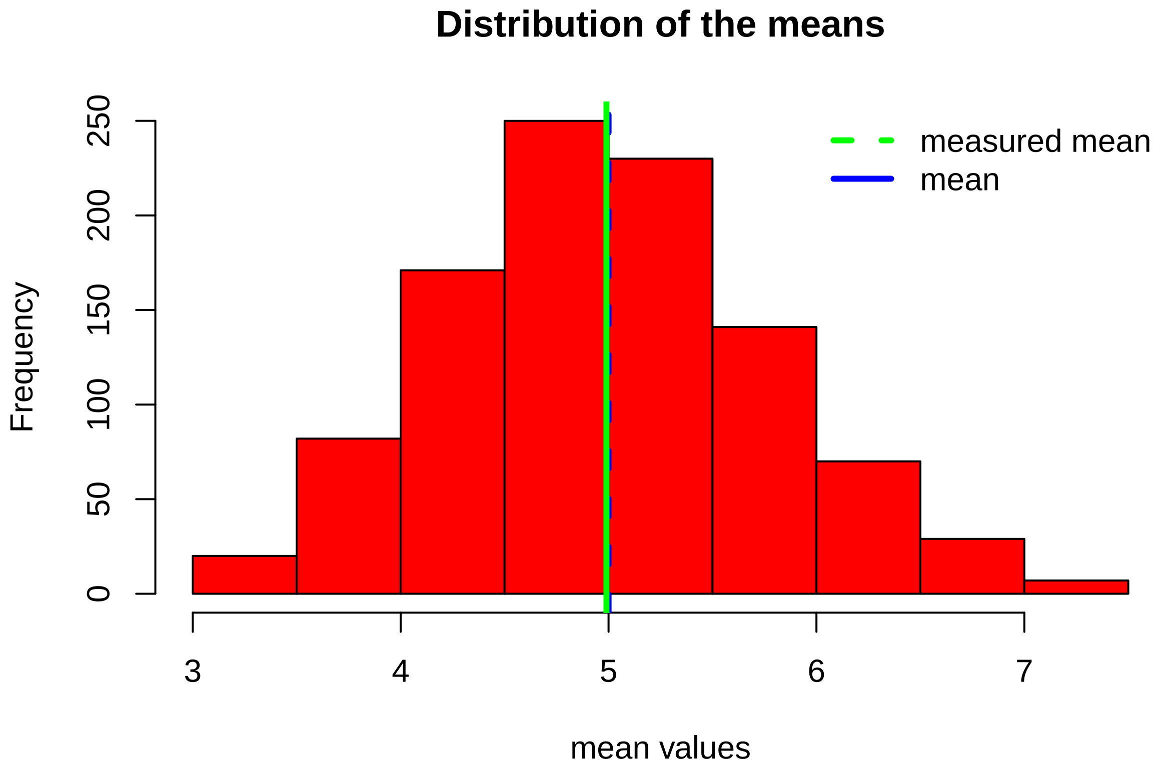
Comparison of the sample and theoretical means

Let's calculate our sample mean and the theoretical mean:

```
true_mean <- 1./lambda
sample_mean <- mean(mean_exp)
```

and plot them:

```
hist(mean_exp, breaks=10, col="red", xlab="mean values", main="Distribution of the means")
abline(v=true_mean, col="blue", lwd = 3, lty = 2)
abline(v=sample_mean, col="green", lwd = 3)
legend('topright',c("measured mean", "mean"), col=c("green","blue"), lty=c(2,1), bty="n", lwd=c(3,3))
```



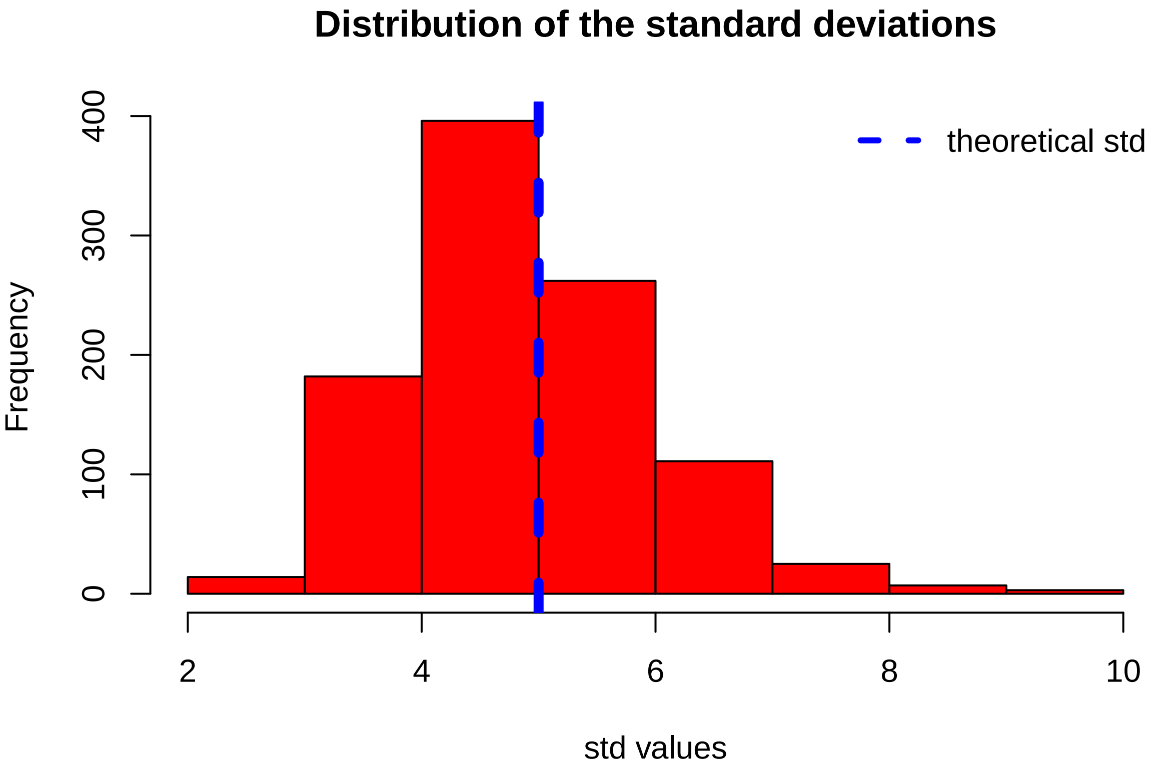
```
diff <- 100*abs(true_mean-sample_mean)/true_mean
```

We see that the distribution of the means peak at the theoretical value. The theoretical and sample means differ by 0.199496 %.

How variable is the sample mean?

Let's look at the variance of our simulations

```
std_exp <- apply(simu, 2, sd)
hist(std_exp, breaks=10, col="red", xlab="std values", main="Distribution of the standard deviations")
abline(v=(1./lambda), col="blue", lwd = 5, lty = 2)
legend('topright',"theoretical std", col="blue", lty=2, bty="n", lwd=3)
```



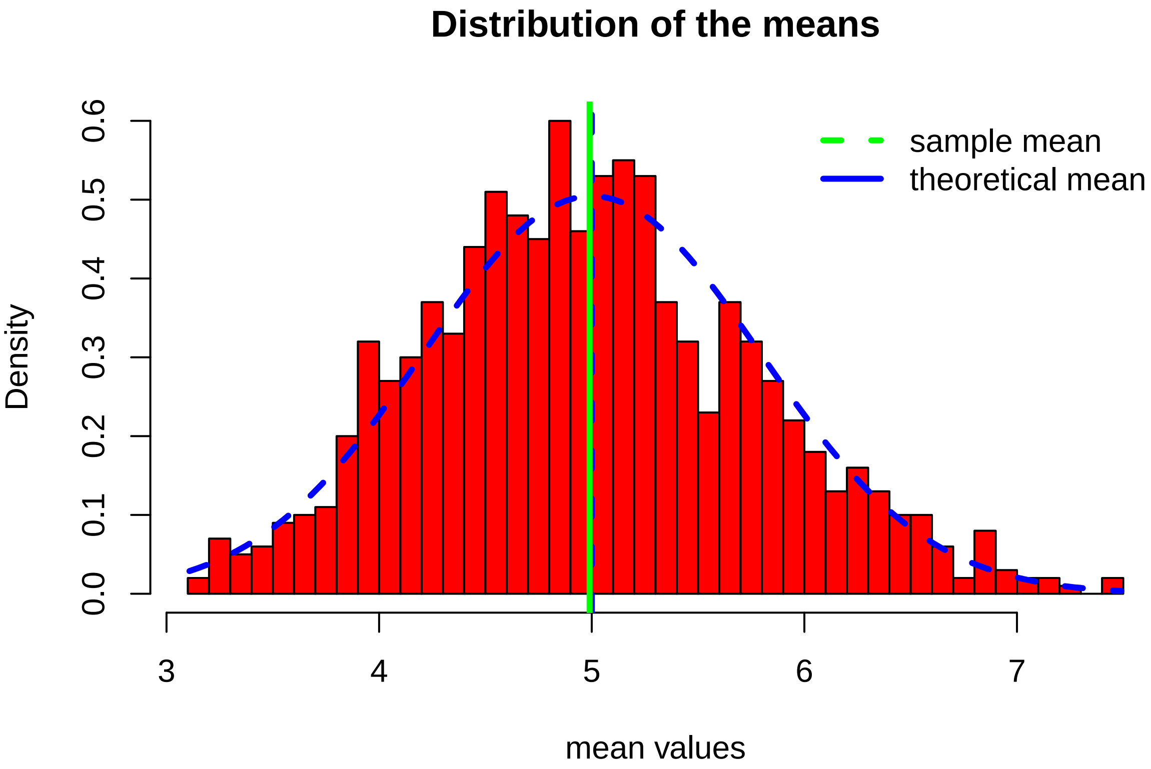
The distribution of the standard deviations (the square root of the variance) peaks around the theoretical value.

What about the variance of the means?

```
sample_std <- sd(mean_exp)
true_std <- (1./lambda)/sqrt(n)
diff_std <- 100*abs(true_std-sample_std)/true_std
diff_var <- 100*abs((true_std)^2-(sample_std)^2)/(true_std)^2
```

The theoretical and sample standard deviationa differ by 1.1169202 % or by 2.2213654 % for the variances. Let's plot these quantities:

```
x <- seq(min(mean_exp), max(mean_exp), length = 100)
gauss <- dnorm(x, mean=true_mean, sd=true_std)
hist(mean_exp, breaks=n, col="red", xlab="mean values", main="Distribution of the means", prob=T)
lines(x, gauss, lwd=3, col="blue", lty=2)
abline(v=true_mean, col="blue", lwd = 3, lty = 2)
abline(v=sample_mean, col="green", lwd = 3)
legend('topright',c("sample mean", "theoretical mean"), col=c("green","blue"), lty=c(2,1), bty="n", lwd=
```

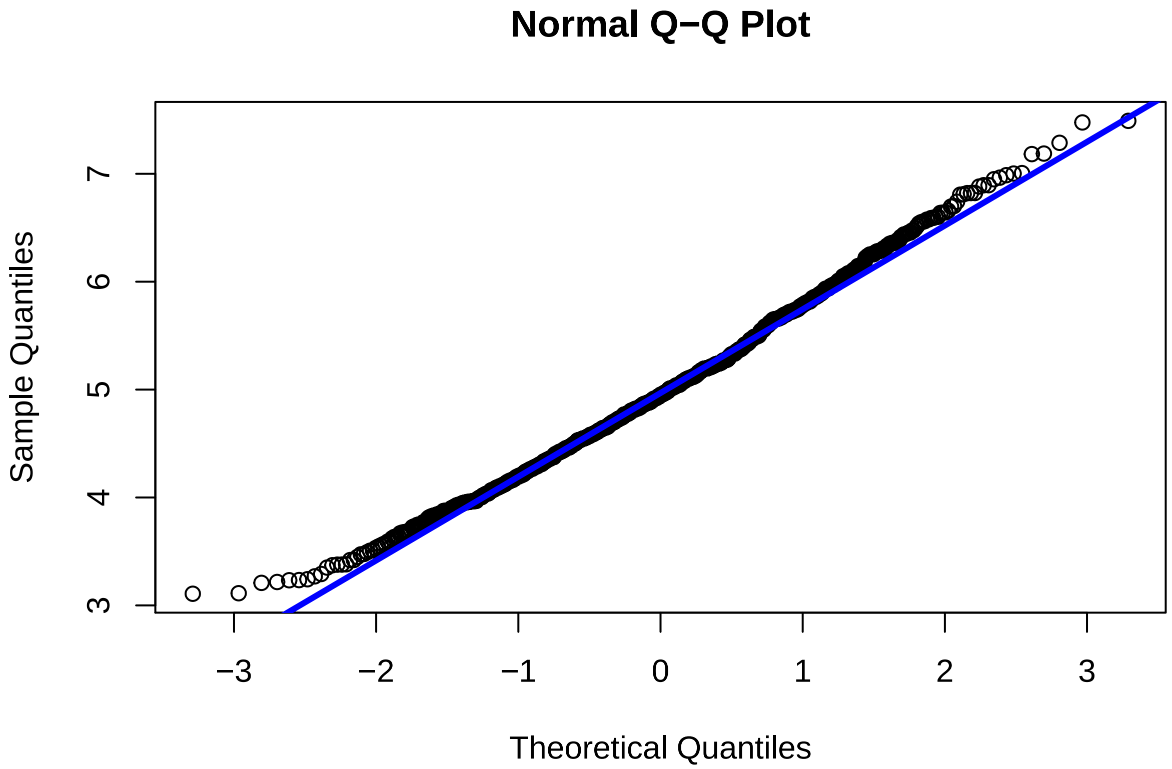


The dotted-line shows a Gaussian distribution with mean equals to the theoretical mean and the standard deviation equals to the theoretical standard deviation divided by the square root of the number of simulations. According to the central limit theorem, we see by eye that this distribution approximates well the distribution of the means.

Is this distribution truly Gaussian?

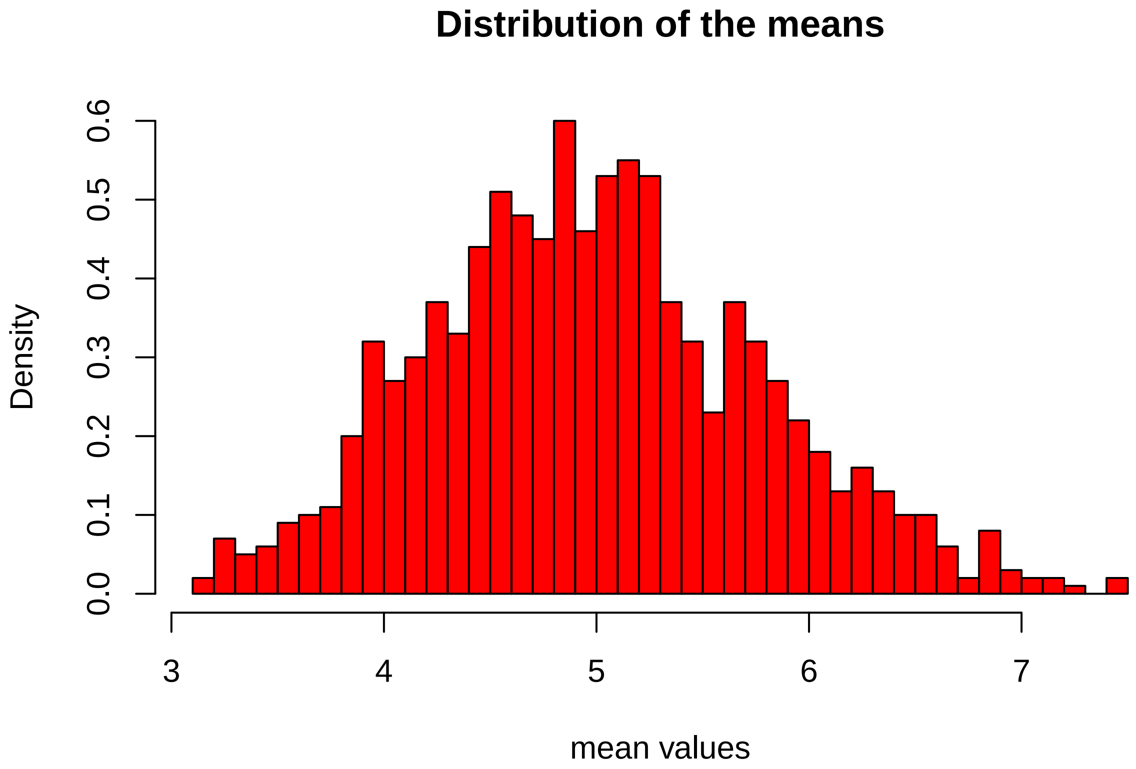
We can perform a Q-Q plot ("Q" stands for quantile) which is agraphical method for comparing two probability distributions by plotting their quantiles against each other. If the two distributions being compared are similar, the points in the QQ plot will approximately lie on the line y = x. Here we compare the distribution of the means with a normal distribution using qqnorm() in R:

```r
qqnorm(mean_exp)
qqline(mean_exp, col = "blue", lwd=3)
```



We do see that the two distributions are very similar however we can notice departures from normality at the low- and high-ends. Another test that we can perform is the “Shapiro-Wilk normality test”. The null-hypothesis of this test is that the population is normally distributed. Thus if the p-value is less than the chosen alpha level, then the null hypothesis is rejected and there is evidence that the data tested are not drawn from a normally distributed population. In other words, the data are not normal. On the contrary, if the p-value is greater than the chosen alpha level, then the null hypothesis that the data came from a normally distributed population cannot be rejected. We choose here alpha = 5%

```
y <- hist(mean_exp, breaks=n, col="red", xlab="mean values", main="Distribution of the means", prob=T)
```



```
shapiro.test(y$density)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  y$density
## W = 0.9121, p-value = 0.002627
```

We see that the p-value is less than our alpha level thus we reject the null-hypothesis that the distribution is gaussian. Increasing the number of simulations will make the distribution of the means more Gaussian.