Statistical Inference - Course project (Author: Peter)

Part I

In this project I will investigate the exponential distribution in R and compare it with the Central Limit Theorem. The exponential distribution can 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. Let lambda be 0.2 for all of the simulations. I will investigate the distribution of averages of 40 exponentials. For this, I will need to do a thousand simulations.

Preliminaries

Let us first setting some variables and load some packages

```
library(ggplot2) # plot library
lambda = 0.2 # as required by the assignment
set.seed(157) # to reproduce the experiments
nSim = 1000
nExp = 40
```

Analysis of the distribution of averages of 40 exponentials

Let us simulate the data, by generating 1000 simulations for each of the 40 exponential distributions, and rearranging all values into a matrix $nSim \times nExp$, that is 1000×40. Then, using the `apply` function, we compute the actual mean of each sampled distribution (that is, each row in the matrix).

```
expDistr = matrix(data=rexp(nExp * nSim, lambda), nrow=nSim)
expDistrMeans = apply(expDistr, 1, mean)
```

Mean analysis

Theoretically, the mean of an exponential distribution should be $\frac{1}{\lambda}$, which yields

```
mu = 1 / lambda
mu
```

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

Let's calculate the sample mean of our 1000 simulations of 40 sampled exponential distributions

```
actualMean = mean(expDistrMeans)
actualMean
```

```
## [1] 5.038662
```

The results of the Central Limit Theorem yields true.

Variance analysis

Similarly to the mean, we compare the theoretical variance with the sampled one. The standard deviation of exponential distribution should be $\sigma = \frac{1/\lambda}{\sqrt{n}}$, which yields

```
theorySD = (1 / lambda) / sqrt(nExp)
theorySD
```

```
## [1] 0.7905694
```

and therefore the variance is σ^2, that is

```r
theoryVar = theorySD^2
theoryVar
```

```
## [1] 0.625
```

On the other hand, the variance computed over the sampled data is

```r
sd(expDistrMeans)
```

```
## [1] 0.8115078
```

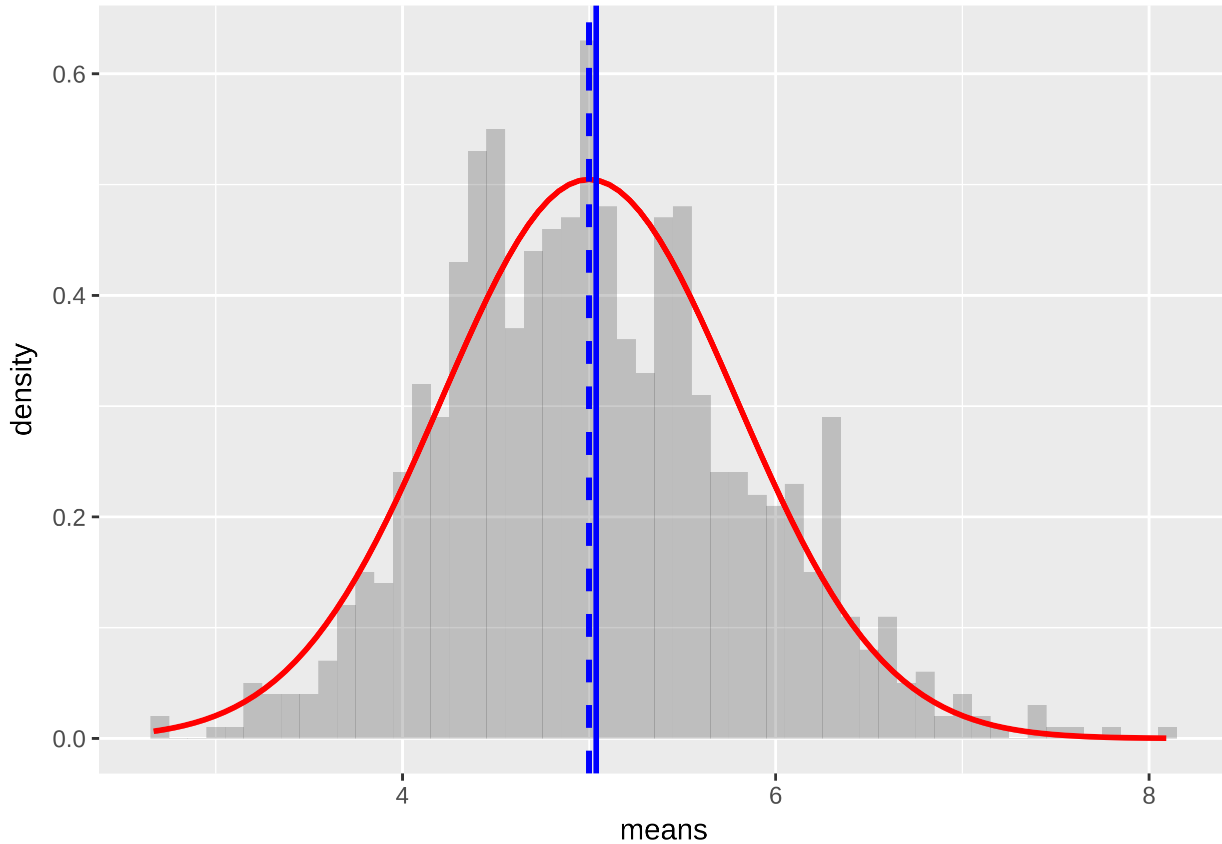
```r
var(expDistrMeans)
```

```
## [1] 0.6585449
```

that yields very close results to the theoretical ones.

Distribution visualization

```r
ggplot(
  data=data.frame(means=expDistrMeans),
  aes(x=means)
) +
  geom_histogram(binwidth=0.1, aes(y=..density..), alpha=0.3) +
  stat_function(fun=dnorm, args=list(mean=mu, sd=theorySD), colour="red", size=1) +
  geom_vline(aes(xintercept=mu), colour="blue", linetype="dashed", size=1) +
  geom_vline(aes(xintercept=actualMean), colour="blue", size=1)
```



The plot shows how close are the distribution of the actual data (histogram) and the theoretical normal distribution with the given mean and standard deviation (red solid line), as well as the theoretical mean value (dashed blue line) and the actual mean computed on the sampled data (solid blue line).