Statistical Inference Course Project

Manu

20/07/2020

1. Instructions

In this project you 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. Set lambda = 0.2 for all of the simulations. You will investigate the distribution of averages of 40 exponentials. Note that you will need to do a thousand simulations.

1.1 Sample mean compared to theoretical mean

Show the sample mean and compare it to the theoretical mean of the distribution.

```
library(ggplot2)

set.seed(1234)
lambda <- .2
n <- 40
nSim <- 1000

exponentialDistribSimFun <- function(n, lambda){
        mean(rexp(n,lambda))
}

SampleMeanDf <- data.frame(ncol=2, nrow=1000) ## A data frame is created to store the results
names(SampleMeanDf) <- c("Index", "sampleMean")

for (i in 1:nSim){
        SampleMeanDf[i,1] <- i
        SampleMeanDf[i,2] <- exponentialDistribSimFun(n, lambda)
}
```

The mean of 1000 simulations is the following:

```
sampleMean <- mean(SampleMeanDf$sampleMean)
sampleMean
```

```
## [1] 4.974239
```

The theoretical mean is the following:

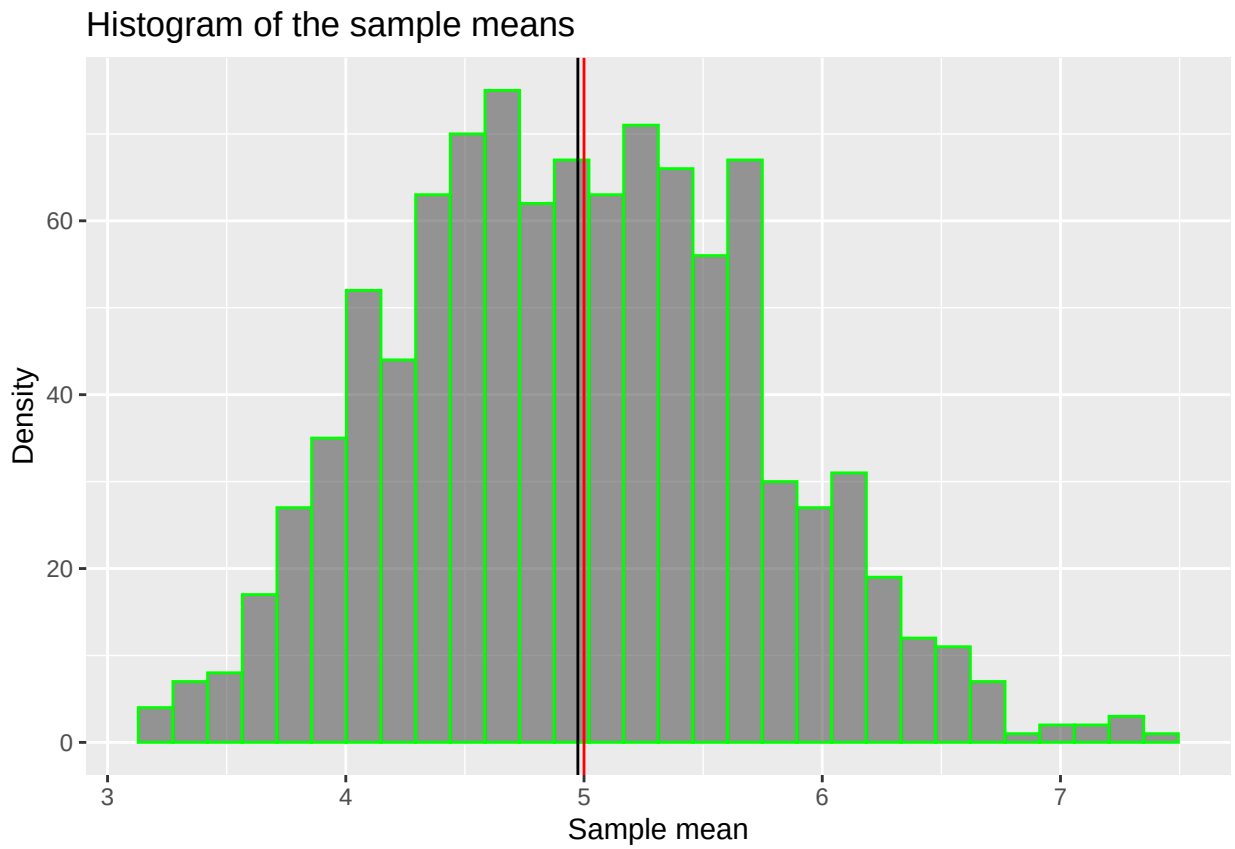
```
theoreticalMean <- 1/lambda
theoreticalMean
```

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

Conclusion: The simulation mean (4.97) is very close to the theoretical mean of 5. The following histogram will demonstrates this:

```
ggplot(SampleMeanDf,
       aes(sampleMean)) +
  geom_histogram(
    alpha=.6,
    position="identity",
    col="green") +
  geom_vline(
    xintercept = theoreticalMean,
    colour = "red",
    show.legend = TRUE) +
  geom_vline(
    xintercept = sampleMean,
    colour = "black",
    show.legend = TRUE)+
  ggtitle ("Histogram of the sample means ")+
  xlab("Sample mean")+
  ylab("Density")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



1.2 Sample variance compared to theoretical variance

Show how variable the sample is (via variance) and compare it to the theoretical variance of the distribution.

```r
sampleVariance <- var(SampleMeanDf$sampleMean)
sampleVariance
```

```
## [1] 0.5706551
```

```r
theoreticalVariance <- (1/lambda)^2 / n
theoreticalVariance
```

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

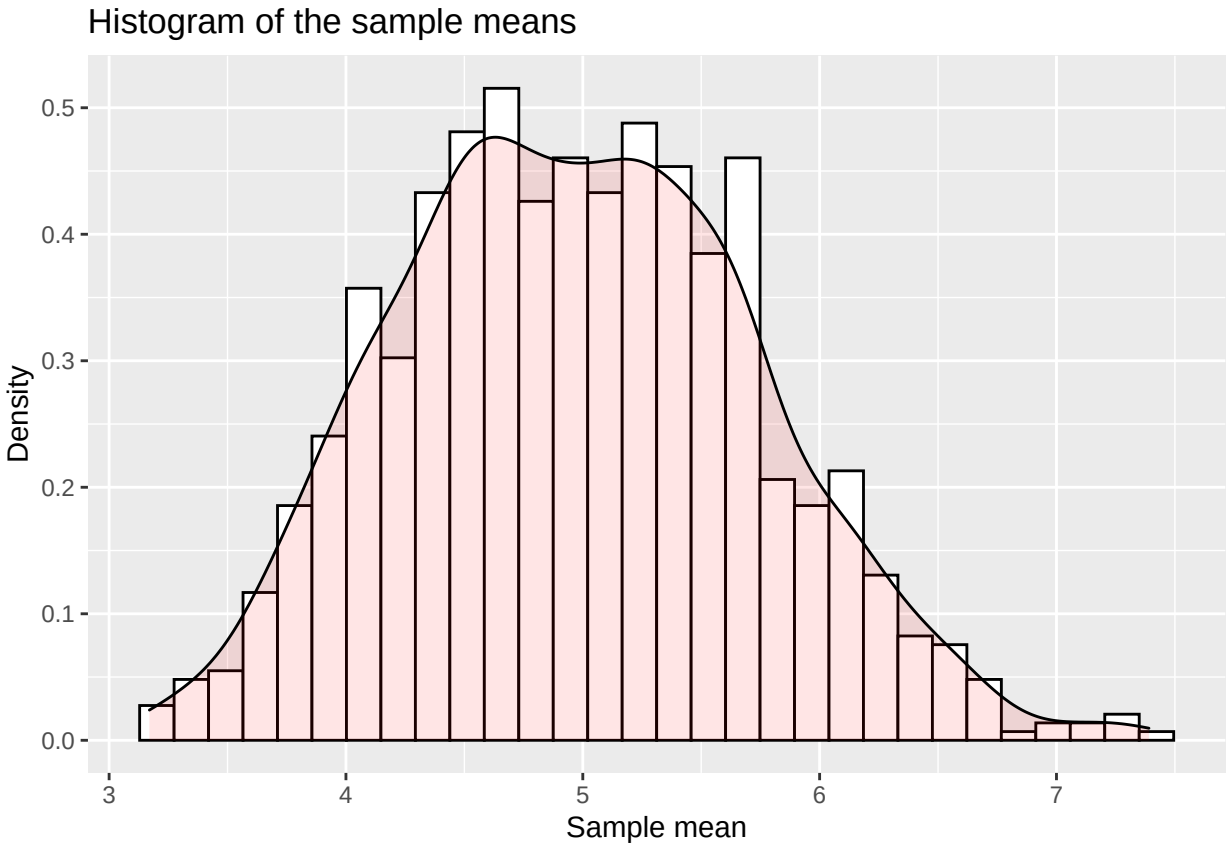
Conclusion The sample and theoretical variances are close.

1.3 Normality of the distribution

Show that the distribution is approximately normal:

```r
ggplot(SampleMeanDf,
       aes(x = sampleMean)) +
  geom_histogram(
    aes(y=..density..),
    col="black",
    fill="white") +
  geom_density(
    alpha = .1,
    fill="red"
  ) +
  ggtitle ("Histogram of the sample means ")+
  xlab("Sample mean")+
  ylab("Density")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Conclusion The central limit theorem states that with a sufficiently large number of random samples taken from the population with replacement, the distribution of sample means will approximate a normal distribution.

The red density layer on the upper histogram demonstrates this. Its gaussian distribution with a bell-shaped curved can be identified.