Statistical Interference - Project

Introduction

The exponential distribution can be simulated in R with `rexp(n, lambda)` where `lambda` is the rate parameter (λ). The mean of the exponential distribution is $\frac{1}{\lambda}$ and the standard deviation is also $\frac{1}{\lambda}$. In this simulation, we will investigate the distribution of averages of 40 exponential random values with λ 0.2. We will do a thousand simulated averages of 40 exponentials.

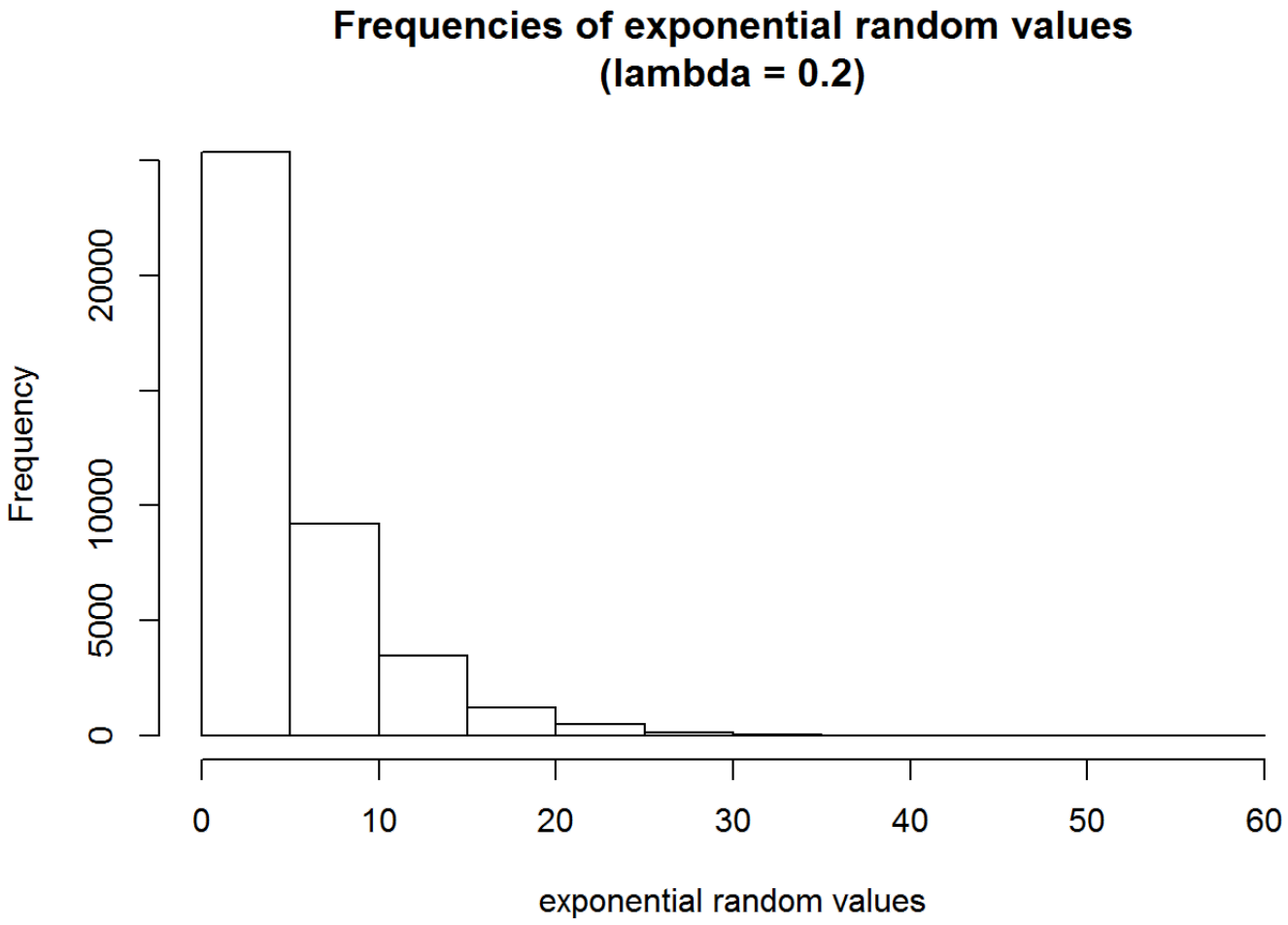
This report illustrates via simulation and associated explanatory text the properties of the distribution of the mean of 40 exponential random values with λ = 0.2. It shows
1. where the distribution is centered at and compares this to the theoretical center of the distribution
2. how variable it is and compares this to the theoretical variance of the distribution
3. that the distribution is approximately normal
4. the coverage of the confidence interval for $\frac{1}{\lambda}$ (i.e. the population mean): $\bar{X} \pm 1.96 \frac{S}{\sqrt{n}}$, where $\bar{X}$ is the sample mean and S the sample sd

The used R code is included in this report so that every part of it is reproducible.

Characteristics of the distribution

For the simulation we generate 1000 samples of size 40, each sample consists of 40 exponential random values. The distribution of the exponential random values is shown in the histogram below. Then we calculate the mean of each sample which results in a “set” of 1000 means. We consider the average of these means a (new) random variable, the **sample mean**. Using the sd function we can calculate the standard deviation of our “set” of 1000 means, and the square of it is the **variance of the sample mean**.

```
set.seed(1)
nosim <- 1000  # nr of simulations
n <- 40   ## sample size
lambda <- 0.2  ## rate parameter
rv <- rexp(n * nosim, lambda)  ## generate random values based on the exponential distribution
hist(rv, main="Frequencies of exponential random values\n(lambda = 0.2)", xlab = "exponential random
values") ## plot frequencies
```




```
# Define a matrix with n random values per row (this represents the sample), each row representing on
e simulation
m <- matrix(rv, nosim)

# Calculate the sample means of each row and its mean and standard deviation
means <- apply(m, 1, mean)
mean_means <- mean(means)
sd_means <- sd(means)
```

Comparing the sample distribution with the theoretical distribution

For the simulated distribution (for lambda 0.2) we get a **mean** of **4.99** and a **variance** of **0.6177**.
The **theoretical mean** (for λ = 0.2) is $\frac{1}{\lambda}$, which is **5** and the **theoretical variance** is the population variance (i.e. the square of the theoretical standard deviation) divided by the sample size (n), which results in $\frac{(\frac{1}{\lambda})^2}{n}$ = **0.625**. The table below summarizes the sample and the theoretical values.

```
tab <- matrix(c(mean_means, sd_means^2, 1/lambda, (1/lambda)^2/n), 2)  ## create table with comparabl
e values
colnames(tab) = c("sample", "theoretical")
rownames(tab) = c("mean", "variance")
tab
```

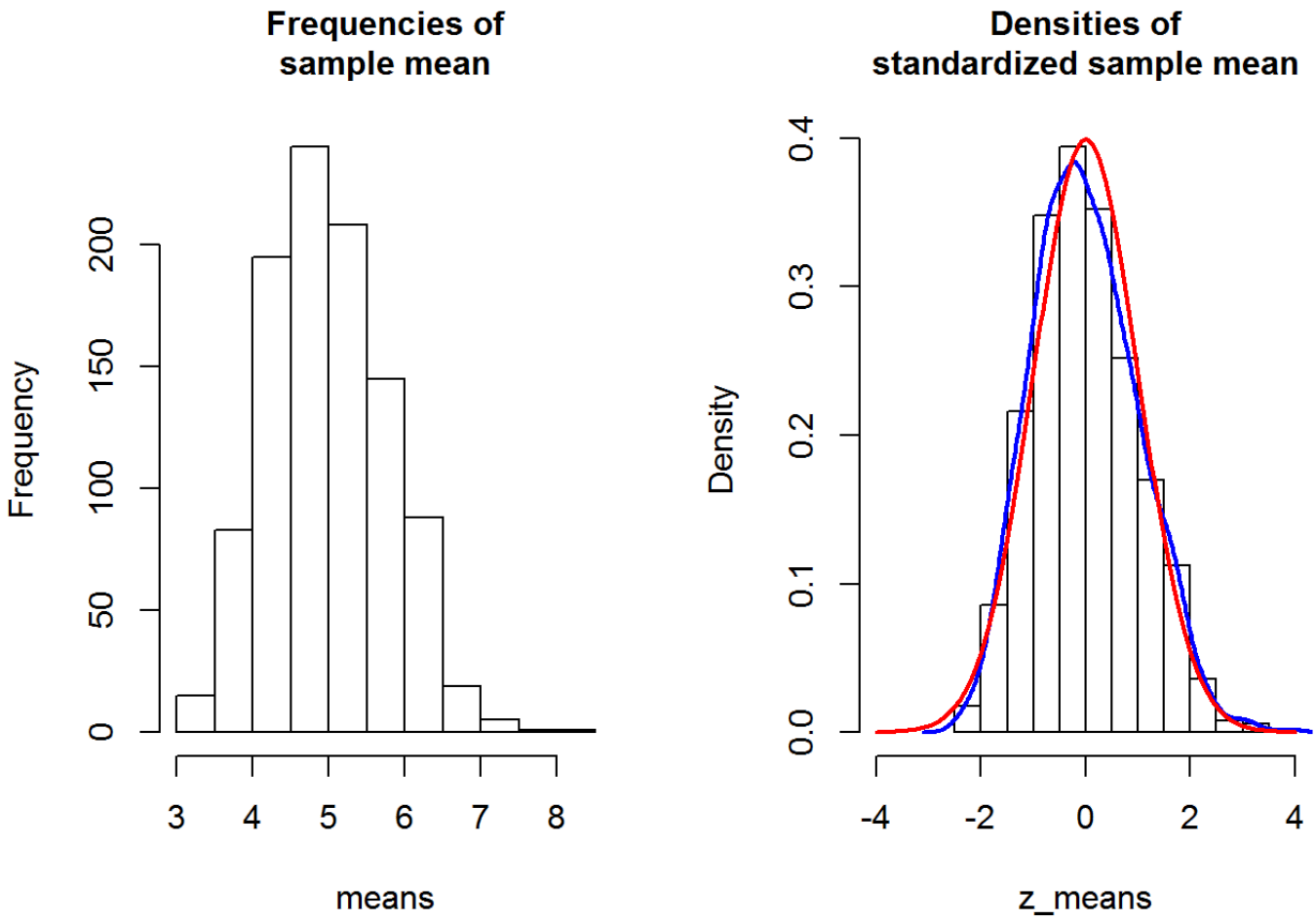
```
##          sample theoretical
## mean     4.9900       5.000
## variance 0.6177       0.625
```

Shape of the distribution density

The histograms below show the distribution of the sample mean (left) and the standardized probabilities (right) with the overlayed density curves of the sample mean (blue) and the standard normal distribution (red):

```
par(mfrow = c(1, 2))
hist(means, main="")  ## histogram plot shows the frequencies of a continuous variable
title(main="Frequencies of\nsample mean", cex.main=1)
z_means <- (means - mean_means) / sd_means  ## standardized values so that the mean is 0 and sd is 1
hist(z_means, prob=TRUE, xlim=c(-4,4), main="")  ## prob=TRUE shows probabilities not frequencies
title(main="Densities of\nstandardized sample mean", cex.main=1)
lines(density(z_means), lwd=2, col="blue")  ## add a density curve

# Plot a standard normal distribution density (for visual comparision)
x=seq(-4,4,length=200)
y=1/sqrt(2*pi)*exp(-x^2/2)
lines(x,y,type="l",lwd=2,col="red")
```



The density plot above on the right proves that the sample mean is approximately normal with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$: $\bar{X}_n \sim N(\mu, \frac{\sigma^2}{n})$.

Note: The plot actually shows that the standardized sample mean is approximately standard normal: $\bar{z}_n \sim N(0, 1)$.

Coverage of the 95% confidence interval for $\frac{1}{\lambda}$

The probability of the sample mean being at a max. distance of $\pm 1.96 \frac{S}{\sqrt{n}}$ from the population mean is 95% (1.96 is the 97.5th percentile, S the sample sd, and n the sample size).

```
ci <- (mean_means + c(-1, 1) * qnorm(0.975) * sd_means/sqrt(n))
```

The **95% confidence interval** for the population mean, $\frac{1}{\lambda}$, is **(4.7465, 5.2336)**.

```
sds <- apply(m, 1, sd)
ll <- means - qnorm(0.975) * sds/sqrt(n)
ul <- means + qnorm(0.975) * sds/sqrt(n)
coverage <- mean(ll < 1/lambda & ul > 1/lambda)  ## 1/lambda is the true population mean
```

In our simulation, the **coverage** of the true population mean, $\frac{1}{\lambda}$, is **0.932**, i.e. about **93.2%** of the intervals obtained in the simulation contain the true population mean, $\frac{1}{\lambda}$.