Statistical inference project part 1: simulation

Nick Plummer

26/07/2018

Overview

In this project we will investigate the exponential distribution in R and compare it with the Central Limit Theorem (CLT).

The exponential distribution can be simulated in R with the `rexp(n, lambda)` function where λ represents the rate parameter. The mean of an exponential distribution $\mu = \frac{1}{\lambda}$ and the standard deviation $\sigma = \frac{1}{\lambda}$.

According to the CLT if a sample is "large" (i.e. consisting of at least 30 independent observations) and the data are not strongly skewed, then the distribution of the sample mean is well approximated by a normal model ($\bar{x}_n \sim N(\mu, \frac{\sigma}{\sqrt{n}})$).

A sampling distribution represents the distribution of the point estimates based on samples of a fixed size from a certain population. We will demonstrate therefore that the sampling distribution of the mean of an exponential distribution $\lambda = 0.2$ and with $n = 40$ observations is approximately normally distributed ($N(\frac{1}{0.2}, \frac{\frac{1}{0.2}}{\sqrt{40}})$).

Question 1: Show where the distribution is centered at and compare it to the theoretical center of the distribution

Draw 1000 samples of size 40 from the $Exp(\frac{1}{0.2}, \frac{1}{0.2})$ distribution and calculate the mean of each sample.

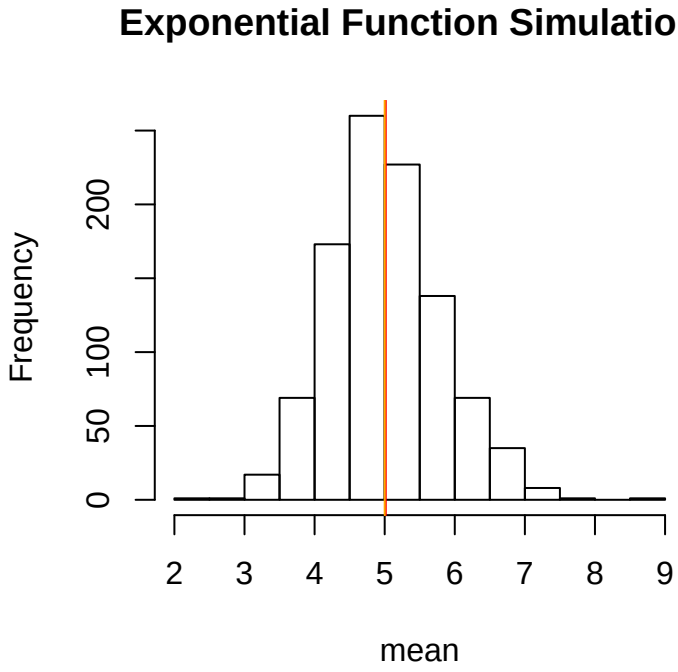
```
lambda <- 0.2  # set lambda to 0.2
n <- 40        # 40 samples
sims <- 1000   # 1000 simulations

sim_exps <- replicate(sims, rexp(n, lambda))
means_exps <- apply(sim_exps, 2, mean)
```

According to the CLT we would expect that the mean of the simulated samples is approximately $\frac{1}{\lambda} = \frac{1}{0.2} = 5$, meaning we expect the mean of 1000 sampled means we to be very close to 5.

```
analytical_mean <- mean(means_exps)
theoretical_mean <- 1/lambda
```

The analytical mean $\bar{x}$ in our case is 5.015 which is very close to the theoretical mean 5. We can demonstrate this by plotting the simulated means as a histogram, and the analytical mean (red) and theoretical mean (orange) almost overlie.

```
hist(means_exps, xlab = "mean", main = "Exponential Function Simulations")
abline(v = analytical_mean, col = "red")
abline(v = theoretical_mean, col = "orange")
```



Question 2: Show how variable it is and compare it to the theoretical variance of the distribution.

According to the CLT we would expect that the variance of the sample of the 1000 means is approximately $\frac{1}{\frac{0.2^2}{40}} = 0.625$.

```
analytical_sd <- sd(means_exps)
analytical_variance <- analytical_sd^2

theoretical_sd <- (1/lambda)/sqrt(n)
theoretical_variance <- ((1/lambda)*(1/sqrt(n)))^2
```

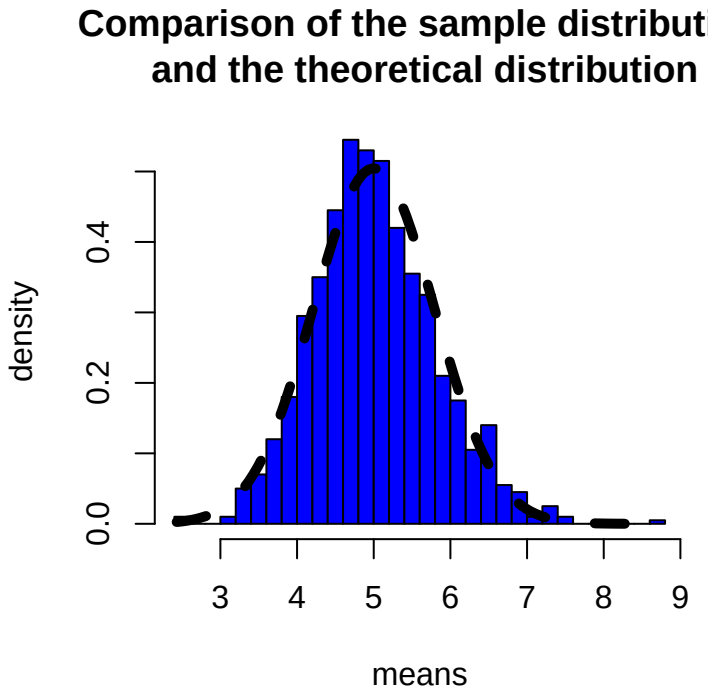
s^2 in our case is 0.6372728 which is close to the theoretical variance 0.625.

Question 3 - Show that the distribution is approximately normal

The sample is sufficiently large that following the CLT the distribution of averages of 40 exponentials is very close to a normal distribution.

In order to demonstrate that the sample distribution of the sampled means is approximately normal we will plot the these as a histogram (blue) and overlay it with the density function from the theoretical sampling distribution (dashed line).

```
xfit <- seq(min(means_exps), max(means_exps), length=100)
yfit <- dnorm(xfit, mean=1/lambda, sd=(1/lambda/sqrt(n)))
hist(means_exps,breaks=n,prob=T,col="blue",xlab = "means",main="Comparison of the sample distribution\n
lines(xfit, yfit, pch=22, col="black", lty=20, lwd=5)
```



We can also show via a Q-Q plot that the distribution of averages of 40 exponentials is very close to a normal distribution.

```
qqnorm(means_exps)
qqline(means_exps, col = 2)
```