A Simulation Exercise

NiaS

October 27, 2018

Goal

The exponential distribution can be simulated in R with `rexp(n, lambda)` where `lambda` λ is the rate parameter. The mean of exponential distribution is $1/\lambda$ and the standard deviation is also $1/\lambda$. For this simulation, we set $\lambda = 0.2$. In this simulation, we investigate the distribution of averages of 40 numbers sampled from exponential distribution with $\lambda = 0.2$.

Generated the Simulation Data

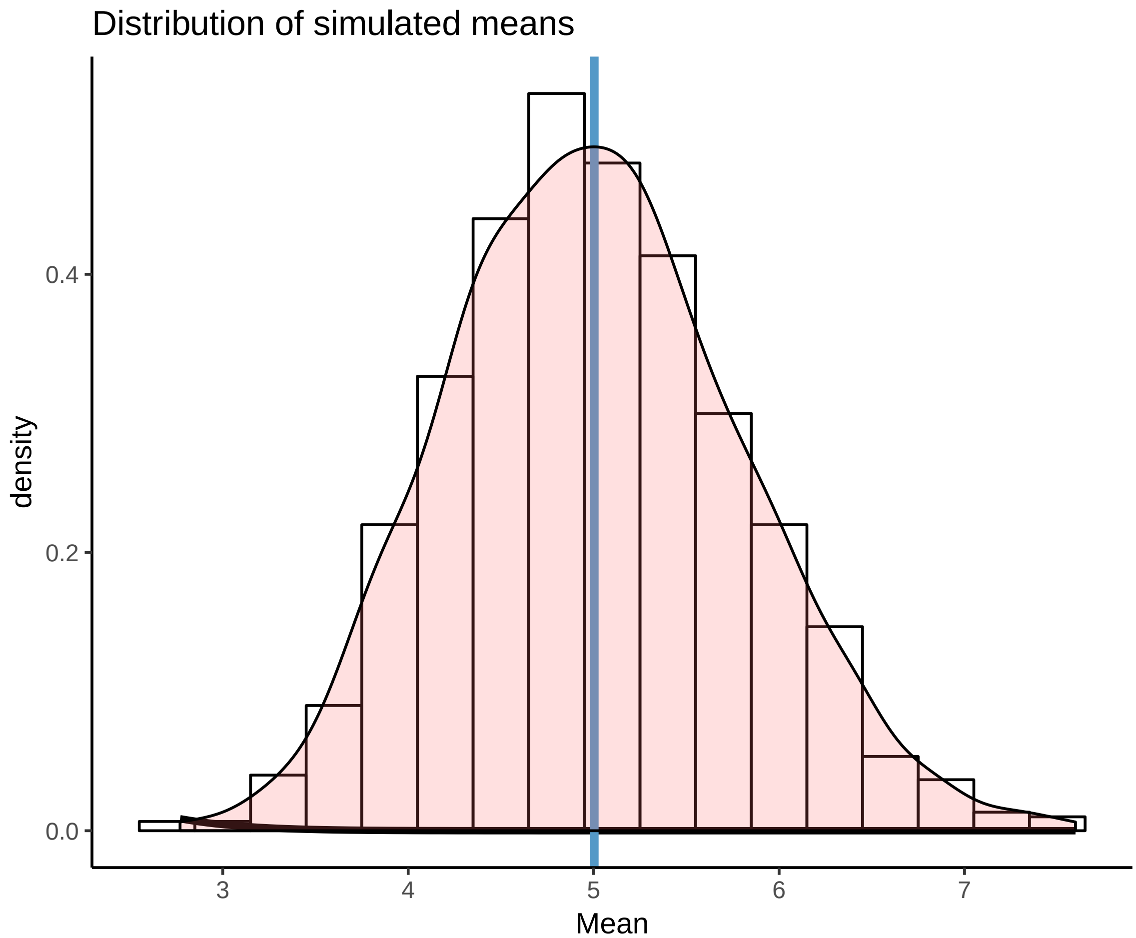
```
#The function parameter
lambda <- 0.2
s <- 40
B <- 1000

set.seed(1232434)

#construct a 1000 x 40 matrix
sim <- matrix(data=rexp(s*B, lambda), nrow=B)
means_of_row <- rowMeans(sim)
```

Sample mean vs theoritical mean

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
#calculate the mean for each row of the matrix
m <- data.frame(Mean = apply(sim, 1, mean))

#calculate the mean of the simulated mean
m.df <- m %>% summarize(sim.mean = mean(Mean)) %>% unlist()

m %>% ggplot(aes(x=Mean))+
    geom_histogram(alpha=0.6,  binwidth = 0.3, fill="white", color="black", mapping = aes(y = ..density
  stat_function(fun = dnorm, size = 1.3) +
  geom_vline(xintercept = m.df, color="#5499C7", size = 1.5) + geom_density(alpha=.2, fill="#FF6666")+
  ggtitle("Distribution of simulated means")+theme_classic()
```



From above plot, we can conclude that the distribution of mean is centered around the mean of our simulated distribution, i.e. indicate by the position of blue vertical line.

Sample Variance vs Theoritical Variance

```r
#Calculate the standard deviation of the sample
sample.sd <- m %>% select(Mean) %>% unlist() %>% sd()
#Calculate the variance using the value of standard deviation
sample.var <- sample.sd^2
sample.var
```

```
## [1] 0.6142007
```

```r
#Theoritical variance
((1/lambda) / sqrt(s)) ^2
```

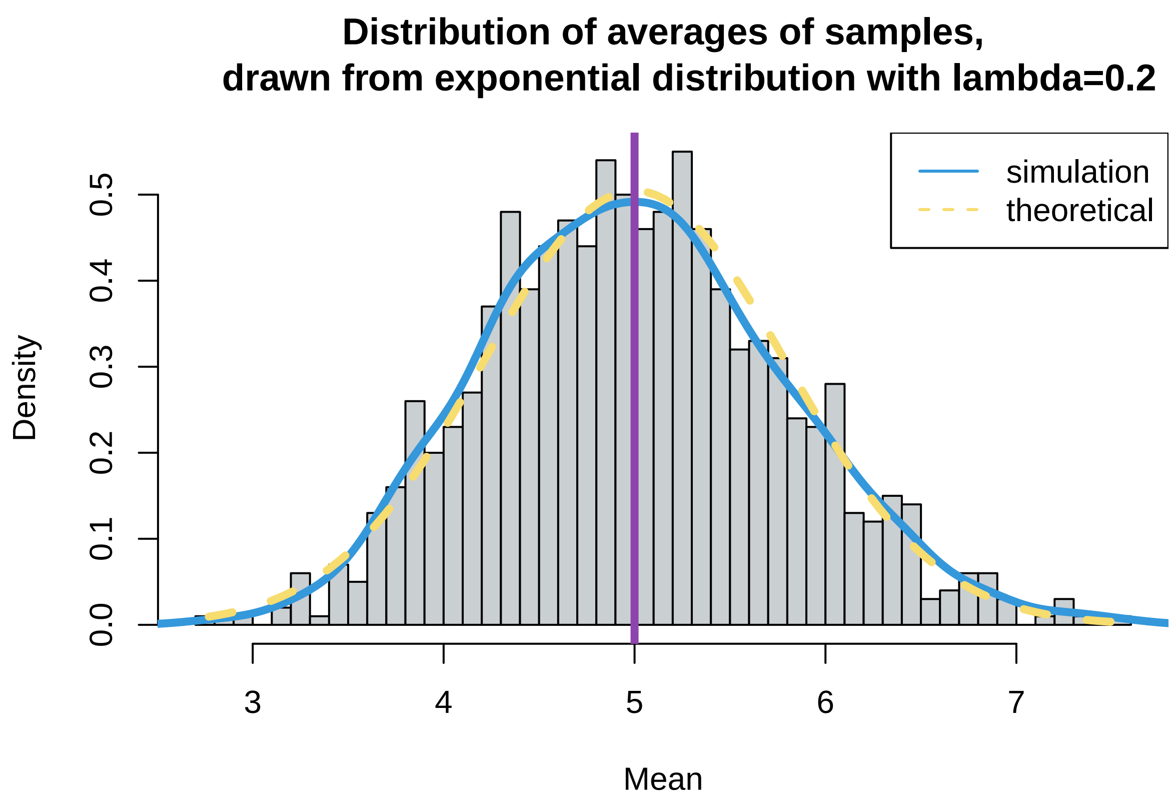
```
## [1] 0.625
```

We can observe their value are close, which are 0.625 and 0.6142007 respectively.

Normality of the Distribution

CLT says the distribution of averages of normalized variables becomes that of a standard normal distribution as the sample size increases.

```r
#plot the correspondent histogram and overaly with density function from theoritical sampling distribut
xfit <- seq(min(m$Mean), max(m$Mean), length=100)
yfit <- dnorm(xfit, mean=1/lambda, sd=(1/lambda/sqrt(s)))

hist(means_of_row, breaks=50, prob=TRUE,
     main="Distribution of averages of samples,
     drawn from exponential distribution with lambda=0.2",
     xlab="Mean", col = "#CACFD2")
# Density of the averages of samples
lines(density(m$Mean), col="#3498DB", lwd=4)
# Theoretical center of distribution
abline(v=1/lambda, col="#8E44AD", lwd=4)
# Theoretical density of the averages of samples

lines(xfit, yfit, pch=20, col="#F7DC6F", lty=2, lwd=4)
legend('topright', c("simulation", "theoretical"), lty=c(1,2), col=c("#3498DB", "#F7DC6F"), lwd=1.5)
```

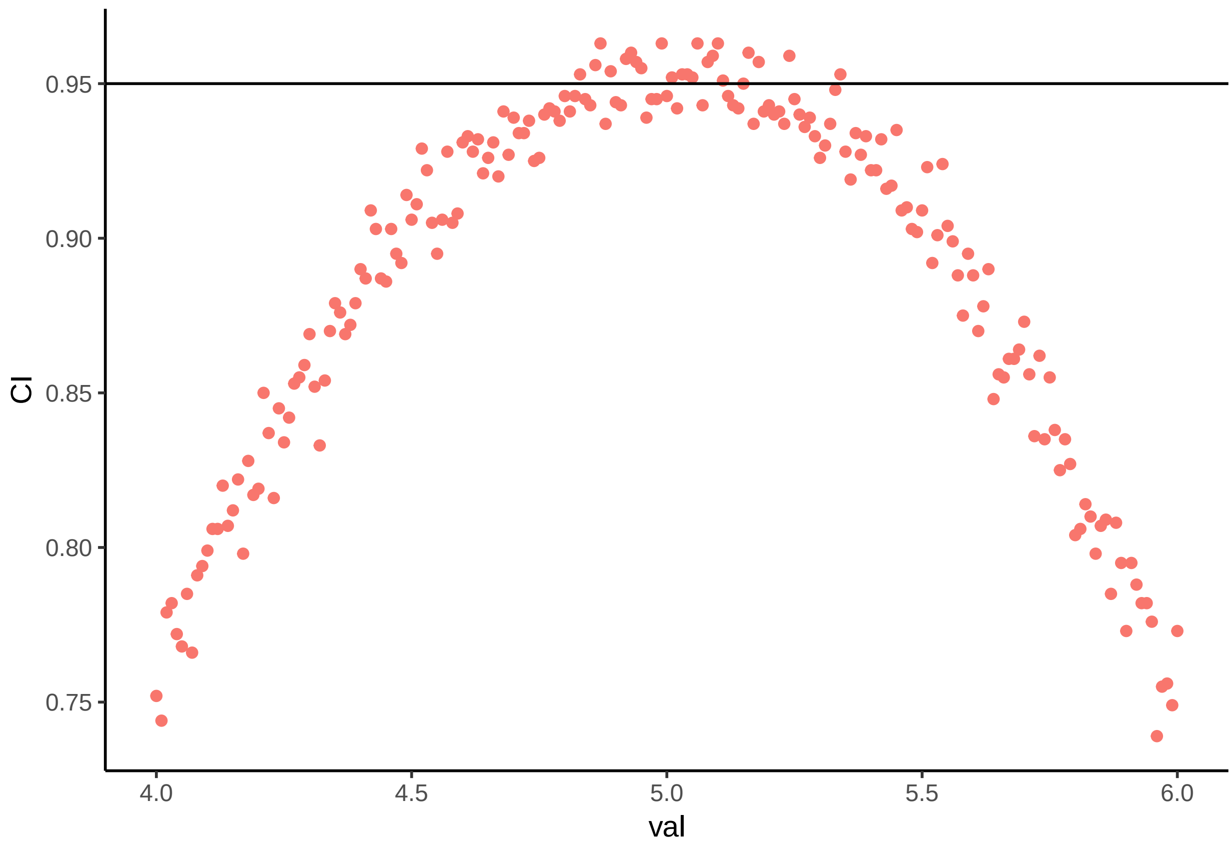


The distribution of sample means is centered at 5.0036654 and the theoretical center of the distribution is $\lambda^{-1} = 5$. The variance of sample means is 0.6142007 where the theoretical variance of the distribution is $\sigma^2/n = 1/(\lambda^2 n) = 1/(0.04 \times 40) = 0.625$.

Due to the Central Limit Theorem (CLT), the averages of samples follow normal distribution. The figure above also shows the density computed using the histogram and the normal density plotted with theoretical mean and variance values.

Let's evaluate the coverage for 95% CI, i.e. $1/\lambda = \bar{X} \pm 1.96 \frac{S}{\sqrt{n}}$

```r
library(ggplot2)

val <- seq(4, 6, by=0.01)
CI <- sapply(val, function(l) {
    mu_hats <- rowMeans(matrix(rexp(s*B, rate=0.2),
                               B, s))
    ll <- mu_hats - qnorm(0.975) * sqrt(1/lambda**2/s)
    ul <- mu_hats + qnorm(0.975) * sqrt(1/lambda**2/s)
    mean(ll < l & ul > l)
})

qplot(val, CI,color = "#E9967A", show.legend=FALSE) + geom_hline(yintercept=0.95)+theme_classic()
```



As can be seen from the plot above, for selection of $\hat{\lambda}$ around 5, the average of the sample mean falls within the confidence interval at least 95% of the time. Note that the true rate, λ is 5.

Conclusion

These analysis show that the sampling distribution of the mean of an exponential distribution with $n = 40$ observations and $\lambda = 0.2$ is approximately $N(\frac{1}{0.2}, \frac{\frac{1}{0.2}}{\sqrt{40}})$ distributed.