Statistical Inference Course Project: A Simulation Exercise

Anand Akella

August 21, 2015

Overview

The goal of this exercise is to 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.

On the other hand according to the central limit theorem distribution of $\bar{x}$ is approximately normal. The approximation can be poor if the sample size is small, but it improves with larger sample sizes. In absence of identified clear outliers, the distribution of the sample mean, is well approximated by a normal model $\bar{x}_n \sim N(\mu, \frac{\sigma}{\sqrt{n}})$.

In order to accomplish this exercise we take $\lambda = 0.2$ and the exercise will be conducted on thousand simulations.

1. Comparison of Sample and theoretical Means of the Distribution

In the following we will draw 1000 samples of size 40 from an $Exp(\frac{1}{0.2}, \frac{1}{0.2})$ distribution. For each of the 1000 samples we will calculate the mean. Theoretically, this the same as drawing a single sample of size 1000 from the corresponding sampling distribution with $N(\frac{1}{0.2}, \frac{\frac{1}{0.2}}{\sqrt{40}})$.

```
# load necessary libraries
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.2.1
```

```
# Set input parameters for simulation
lam <- 0.2 # lambda for rexp
ne <- 40 # number of exponentials
nSim <- 1000 # number of simulations

# set seed for reproducibility
set.seed(679)

#Run test to calculate mean.It will be an ne x nSim matrix
expDist <- matrix(data=rexp(ne*nSim,lam),nrow=nSim)
expDistMean <- data.frame(means=apply(expDist, 1, mean))
# Avearge Sample Mean of 1000 Simulations
meanDist <- mean(expDistMean$means)
```

Comparison of Sample and Theoretical Mean

Sample Mean:

```
meanDist
```

```
## [1] 5.004808
```

Theoretical Mean: $\mu = \frac{1}{\lambda}$

```
1/lam
```

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

Conclusion: As expected the sample mean from the simulations is very close to the Theoretical Mean as calculated above

2. Comparison of Sample Variance to the Theoretical Variance

The theoretical variance is calculated as square of the standard deviation or calculated as $\sigma = \frac{1/\lambda}{\sqrt{n}}$, $Var = \sigma^2$

```
sampVar <- var(expDistMean$means) # Sample variance
theoVar <- (1/lam)^2/ne # theoretical variance
sampVar
```

```
## [1] 0.6000051
```

```
theoVar
```

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

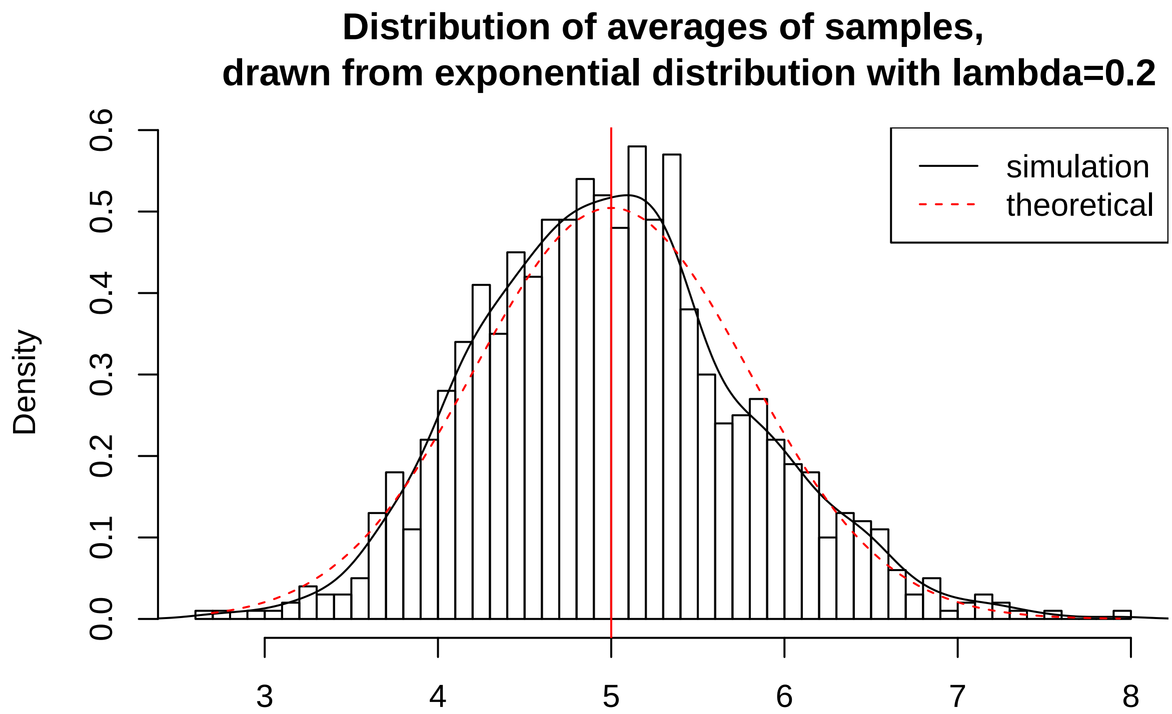
Conclusion: The calculated Sample Varicance and the theretical variance are very close. Minor differences are possible due to the fact that variance is square of standard deviation and hence small differences will be magnified.

3. Investigating Normality of the Distribution

In order to investigate the normality of the sample distribution with 1000 simulated sample means a histogram is plotted and overlayed with the density function from the theoretical sampling which is $N(\frac{1}{0.2}, \frac{\frac{1}{0.2}}{\sqrt{40}})$ distributed.

```
#plot the means
mu <- 1/lam
sd <- 1/lam/sqrt(ne)
# plot the histogram of averages
hist(expDistMean$means, breaks=50, prob=TRUE,
     main="Distribution of averages of samples,
     drawn from exponential distribution with lambda=0.2",
     xlab="")
# density of the averages of samples
lines(density(expDistMean$means))
# theoretical center of distribution
abline(v=1/lam, col="red")
```

```r
# theoretical density of the averages of samples
xfit <- seq(min(expDistMean$means), max(expDistMean$means), length=100)
yfit <- dnorm(xfit, mean=mu, sd = sd)
lines(xfit, yfit, pch=22, col="red", lty=2)
# add legend
legend('topright', c("simulation", "theoretical"), lty=c(1,2), col=c("black", "red"))
```



Conclusion: From the graph above it can be inferred that the distribution of means of random sampled exponential distributions overlaps very closely with the normal distribution of expected values based on the given λ. Hence it can be considered approximately normal.

The complete report can be found on github at https://github.com/anakella/Statistical-Inference.git as stat_inf_project_part1.Rmd and stat_inf_project_part1.pdf