Statistical Inference Course Project (Coursera)

Yevgeny V.Yorkhov

04/25/2015

Overview

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.

Illustrate via simulation and associated explanatory text the properties of the distribution of the mean of 40 exponentials. You should 1. Show the sample mean and compare it to the theoretical mean of the distribution. 2. Show how variable the sample is (via variance) and compare it to the theoretical variance of the distribution. 3. Show that the distribution is approximately normal.

In point 3, focus on the difference between the distribution of a large collection of random exponentials and the distribution of a large collection of averages of 40 exponentials.

Simulations

The task is: to generate a vector which consist of 1000 averages of exponential distribution (each of exponentials consist of 40 observations). Calculate mean, variance and standard deviation of the simulated and theoretical distribution of averages. Compare them.

Sample Mean versus Theoretical Mean

At first I generate sample distribution, after that I compare expected (theoretical) mean of distribution and simulated mean.

```r
sim <- 1000 # simulations
n <- 40 # samples
lambda <- .2

# Set some randomness
set.seed(as.numeric(Sys.time()));

# Simulate exponential distribution
mns = NULL
for ( i in 1:sim ) {
  mns = c(mns, mean(rexp(n, lambda)))
}

# Here we have averages (means)
head(mns)
```

```
## [1] 4.549430 5.465942 4.507545 5.892741 5.705227 4.112257
```

Theoretical mean of exponential distribution is 1/lambda. Let's compare theoretical mean with simulation mean and show that they are pretty close to each other.

```
# expected (theoretical) mean and simulated mean
c(1/lambda, mean(mns))
```

```
## [1] 5.000000 4.991193
```

Conclusion: Simulated mean is close to expected (theoretical) mean of a normal distribution.

Sample Variance versus Theoretical Variance

Theoretical variance of a random sample for exponential distribution with n observations is

$$Var(\bar{X}) = \frac{\sigma^2}{n} = \frac{1}{\lambda^2} * \frac{1}{n} = \frac{1}{\lambda^2 * n}$$

Let's compare theoretical variance with simulation variance and show that they are pretty close to each other as well.

```
tsd <- sqrt(1/(lambda^2*n)) # expected (theoretical) standard deviation

# Here we have theoretical and simulated standard deviations
c(tsd, sd(mns))
```

```
## [1] 0.7905694 0.7903798
```

```
# Here we have theoretical and simulated variances
c(round(1/(lambda^2*n),1), round(var(mns),1))
```

```
## [1] 0.6 0.6
```

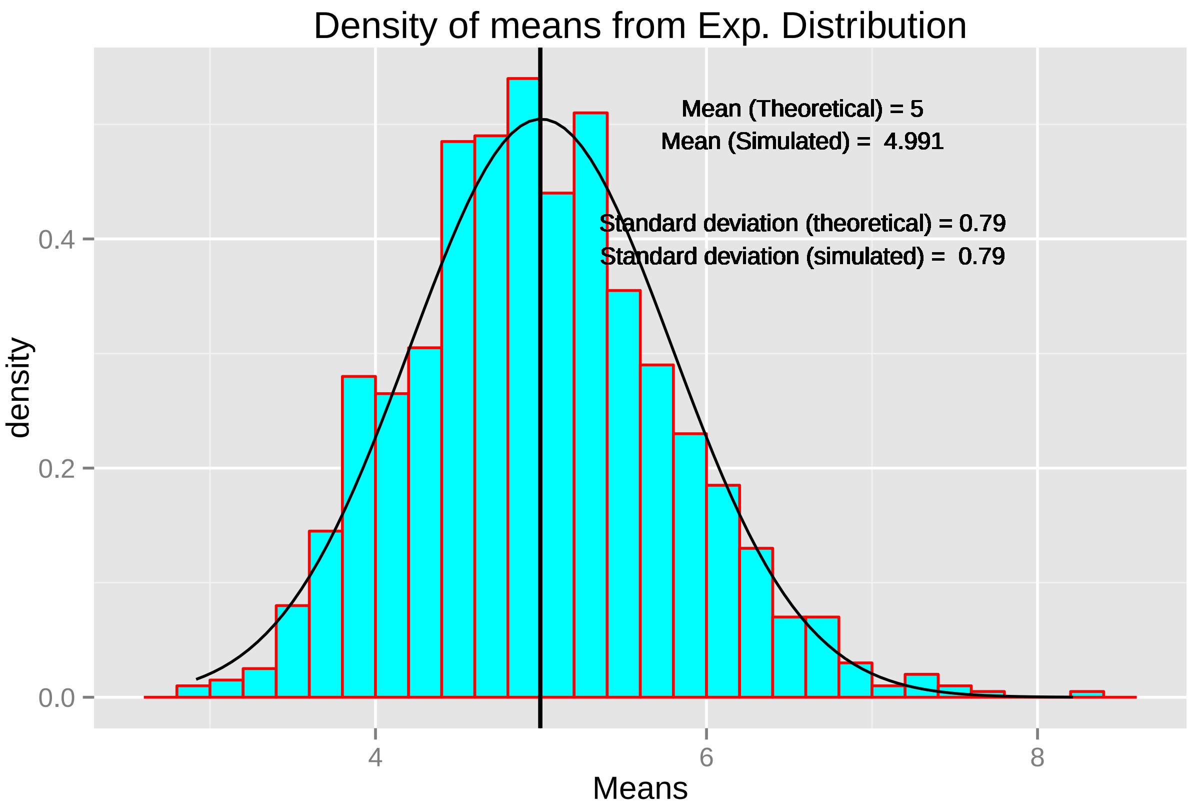
Conclusion: Simulated variance is close to expected (theoretical) variance of a normal distribution.

Distribution

Here I make a plot which shows that the distribution is bell-shaped, it's mean and standard deviation are pretty close to theoretical vlalues of normal distribution.

```
library(ggplot2)

xdf <- as.data.frame(mns)
ggplot(data = xdf, aes(x = mns)) + geom_histogram(aes(y = ..density..), fill = I("cyan"),
    binwidth = 0.2, color = I("red")) + stat_function(fun = dnorm, arg = list(mean = 1/lambda,
    sd = 1/(lambda*sqrt(n))   )) + geom_vline(xintercept = 1/lambda,
      lwd=.5) + geom_vline(xintercept = mean(mns),
      lwd=.5) + geom_text(aes(x=1/lambda + 2*tsd,
        label=paste("Mean (Theoretical) = 5\nMean (Simulated) = ", round(mean(mns),3)),
        y=0.5), size=3) + geom_text(aes(x=1/lambda + 2*tsd,
        label=paste("Standard deviation (theoretical) = 0.79\nStandard deviation (simulated) = ",
                    round(sd(mns), 2)),y=0.4),size=3) + labs(title="Density of means from Exp. Distribu
                                                                     x="Means")
```



According the CLT, the sample mean ***X*** is approximately normal with mean ***mu*** and ***sd=sigma/sqrt(n)*** .

With this level of sample size, we can assess how well the data matches a 95% confidence interval to the population.

The ***95% confidence interval*** for ***mu*** is calculated as follows:

$$\bar{X} \pm 1.96 * StandardError = \bar{X} \pm 1.96 * \frac{\sigma}{\sqrt{(n)}}$$

Taking the sample mean and +/- 1.96 * stderror we can then determine the number of times this result includes the population mean. This will be expected to be approximately 95% of the time.

Now find out how sample means match 95% confidence interval.

```
mdiff <- abs(mns-1/lambda) - 1.96* (1/lambda) / (sqrt(n))
proportion <- 100 * length(mdiff[mdiff<=0])/sim
proportion # the result of matching
```

```
## [1] 95.2
```

Conclusion: As we can see the simulated means match 95% confidence interval pretty well.

Conclusion

Simulated mean, variance and standard deviation of simulated data are pretty close to expected (theoretical) values of normal distribution. The plot of the density of simulated distribution is bell-shaped. Population means matches 95% confidence interval of normal distribution. So we can say that the distribution is approximately normal.