Statistical Inference - Course Project - Part I

Woldetsadick Selam Getachew

Wednesday, October 22, 2014

Part I: Simulation Exercises

The following analysis was made using the R statistic software version R version 3.1.1 (2014-07-10) on a computer using Windows OS version x86_64-w64-mingw32/x64 (64-bit) and completed on Thu, Oct 23 2014, 2:50:35 AM.

Generating the Data

The exponential distribution can be simulated in R with repx(n,λ) where λ represents the rate parameter of the exponential distribution. It can also be demonstrated that the mean of exponential distribution is $\mu = 1/\lambda$ and the standard deviation θ is also $\theta = 1/\lambda$. In this simulation, λ is set to equal 0.2, and n to 40. Hence are generated 1000 simulated averages of 40 exponential of parameter equal to .2.

```
# 1000 simulations of 40 exponentials of parameter .2
set.seed(123)
n <- 40
lambda <- .2
nosim <- 1000
data <- matrix(rexp(nosim * n, rate = lambda), nosim, n)
Dim <- dim(data)
```

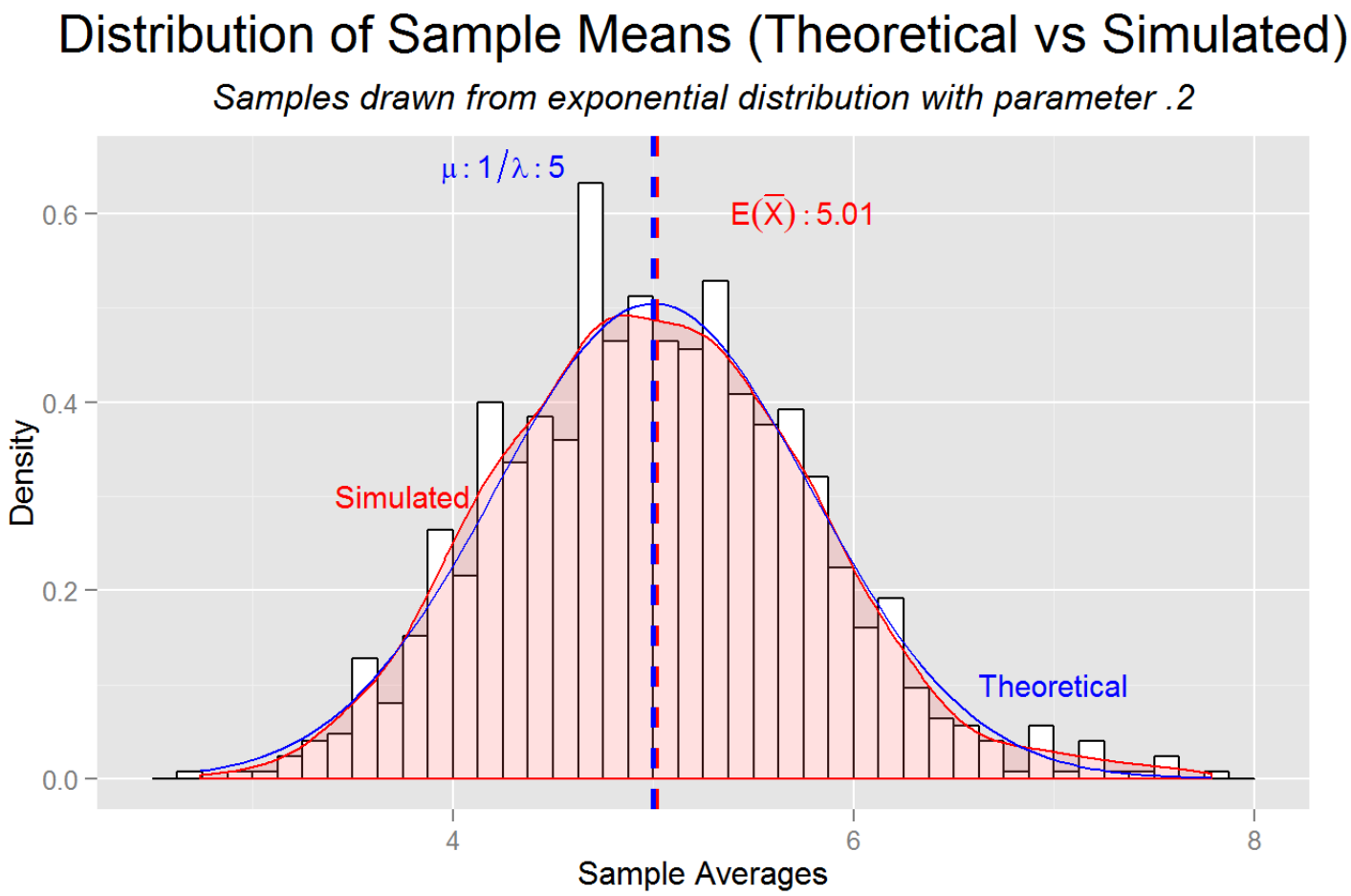
In the end, a data matrix of dimension `1000, 40` is generated where each row i represents the i^{th} simulation out of the thousand simulated 40 exponential.

1. Central Tendency

```
#Calculating averages by row, or for the 1000 different simulations over the 40 sample drawn
xbar <- round(mean(apply(data, 1, mean)), digits = 2)
#Calculating theoretical mean
mu <- 1/lambda
```

Let's note $E(\bar{X}_n)$ the expected value of the averages of the n simulated exponential of parameter λ, and μ its true (theoretical) mean. Their values for n = 40 are respectively $E(\bar{X}_{40})$ = `5.01` and μ = `5`. Observe the near equality of these two values. This near equality is guaranteed by the LAW of LARGE NUMBERS (LLN).
The plot below shows a histogram of the 1000 averages of 40 samples drawn from the exponential distribution with parameter λ = .2. Its corresponding density is shown in red and an-noted "Simulated". A vertical line with equation x = `5.01` in red dashed is introduced to show the center of mass of the simulated exponential.

Moreover, for the sake of comparison a normal density of parameters $\mu_N = \mu$ and $\theta_N = \theta/\sqrt{n} = (1/\lambda)/\sqrt{n}$ is shown in blue, while a dashed line of the same color and of equation x = `5` marks not only the center mass of such density, but also the theoretical mean of simulated distribution.



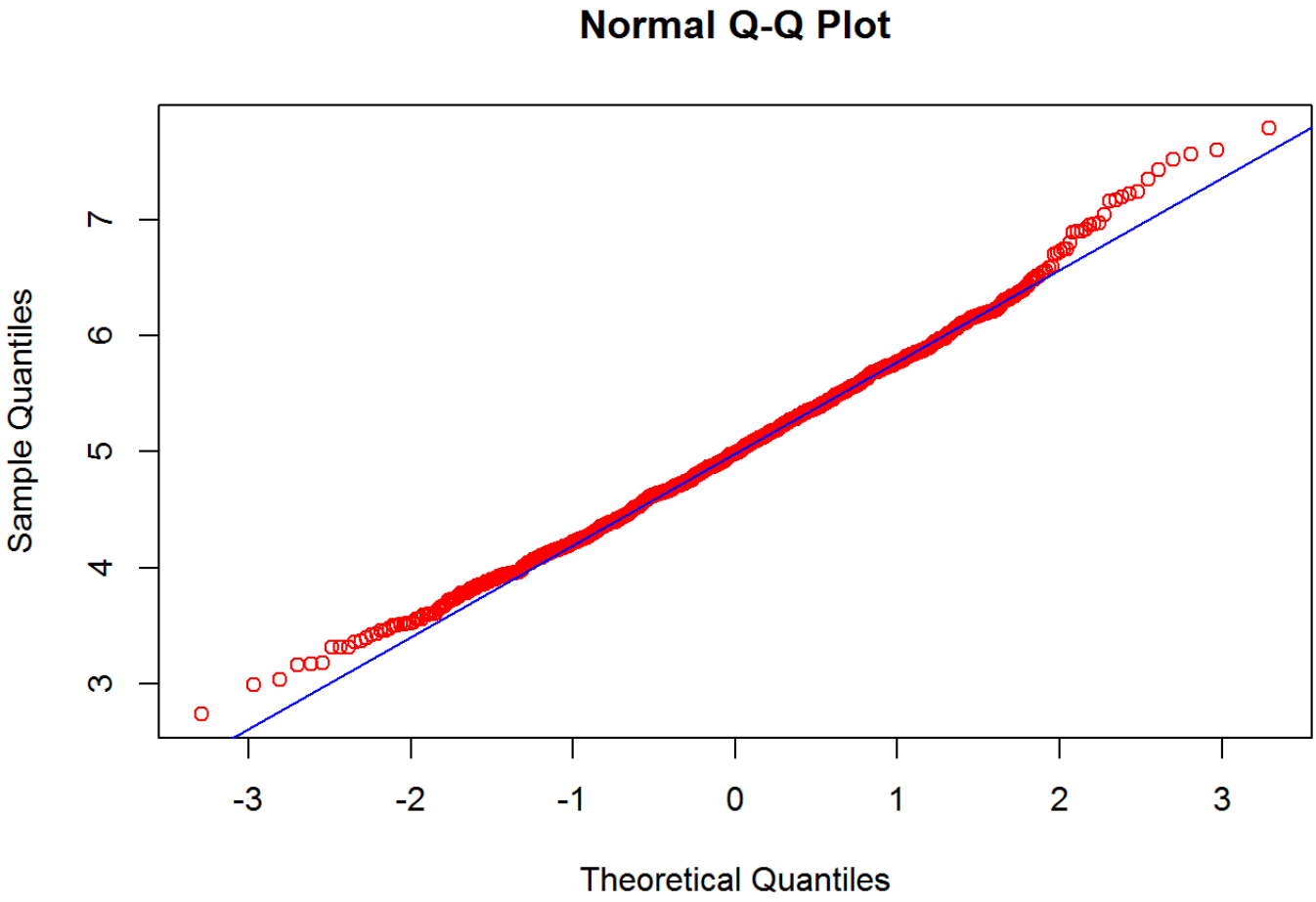
Observe how closely the distribution of sample averages of the simulated data follows the normal distribution centred at the theoretical mean of the sample averages. This closeness is again guaranteed by the Central Limit Theorem CLT.

2. Dispersion

```
#Calculating sd of the 1000 averages
Ssqr <- round((sd(apply(data, 1, mean))), digits = 2)
#Calculating theoretical standard deviation
theta <- round(((1/lambda)/sqrt(n)), digits = 2)
```

Let's note *S* the standard error of the averages of the n simulated exponential of parameter λ, and θ its true (theoretical) standard deviation. Their values for n = 40 are respectively *S* = `0.78` and θ = `0.79`. Observe the near equality of these two values. This near equality is guaranteed by the CLT.
The theoretical & the empiric variability of the averages of the n simulated exponential around the mean is close. One can then conclude that that the expected value of sample averages is an unbiased estimator of the mean of the exponential ditribution from which the sample are drawn from.

3. Central Limit Theorem

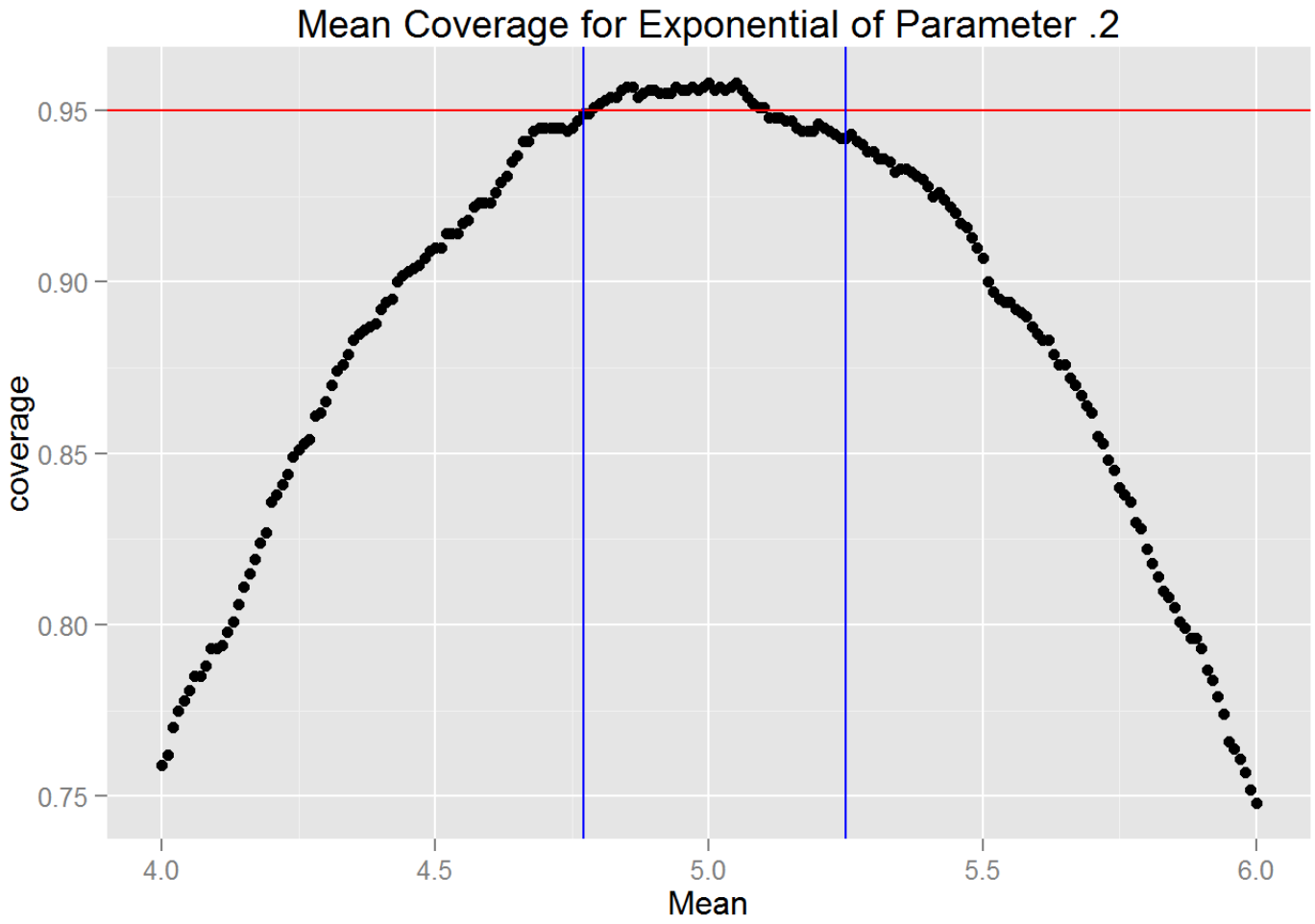


It was observed in the first plot up that the distribution of sample averages follows closely a normal distribution. It was also observed that this closeness was guaranteed by the CLT as long as the draws are idependent and identical, the distribution that they are drawn from is finite and n is large enough.
As can be seen in the plot up, the QQplot is not perfectly linear. The higher and lower tail don't fall exactly on the 45° line but the very large majority of observations does so one can conclude that the distribution is indeed approximately normal.

4. Confidence Interval Coverage

```
interval <- mean(Rmeans) + c(-1,1)*qnorm(.975)*(sd(Rmeans)/sqrt(n))
n <- 40
nosim <- 1000
lambda <- .2

avgvals <- seq(4, 6, by = 0.01)
coverage <- sapply(avgvals, function(avg){
         set.seed(123)
         datas <- matrix(rexp(n * nosim, rate = lambda), nosim, n)
         xbars <- apply(datas, 1, mean)
         ll <- xbars - qnorm(.975) * sqrt(1/lambda**2/n)
         ul <- xbars + qnorm(.975) * sqrt(1/lambda**2/n)
         mean(ll < avg & ul > avg)
})
```



So if one assumes that the samples in this simulation are an IID draw of an exponential distribution of whose parameter λ is unknown, one can conclude that 95% of the time, the true value of $\frac{1}{\lambda}$ will be in the interval [`4.7701, 5.2537`] Observe how the theoretical mean $\frac{1}{\lambda} = 5$ is covered by the 95% confidence interval.