Statistical Inference - Course Project

Buva Ramamoorthy

September 19, 2014

Statistical Inference Course Project : Part#1 Simulation Exercise

This report is produced as part of the "Statistical Inference" course project by Coursera which is a part of specialization "Data Science" by John Hopkins University. In this first part, we perform simulation exercises using exponential distributions.

Simulations

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 exponential(0.2)s.

Lets start by doing a thousand simulated averages of 40 exponentials.

```
# Set seed
set.seed(3)
lambda <- 0.2

# We perform 1000 simulations with 40 samples
sample_size <- 40
simulations <- 1000

# Lets do 1000 simulations
simulated_exponentials <- matrix(rexp(simulations*sample_size, rate=lambda), simulations, sample_size)

# Averages of 40 exponentials
row_means <- rowMeans(simulated_exponentials)
```

Results

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

```
# mean of distribution of averages of 40 exponentials
mean(row_means)
```

```
## [1] 4.987
```

```
# mean from analytical expression
1/lambda
```

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

The distribution of sample means is shown below:

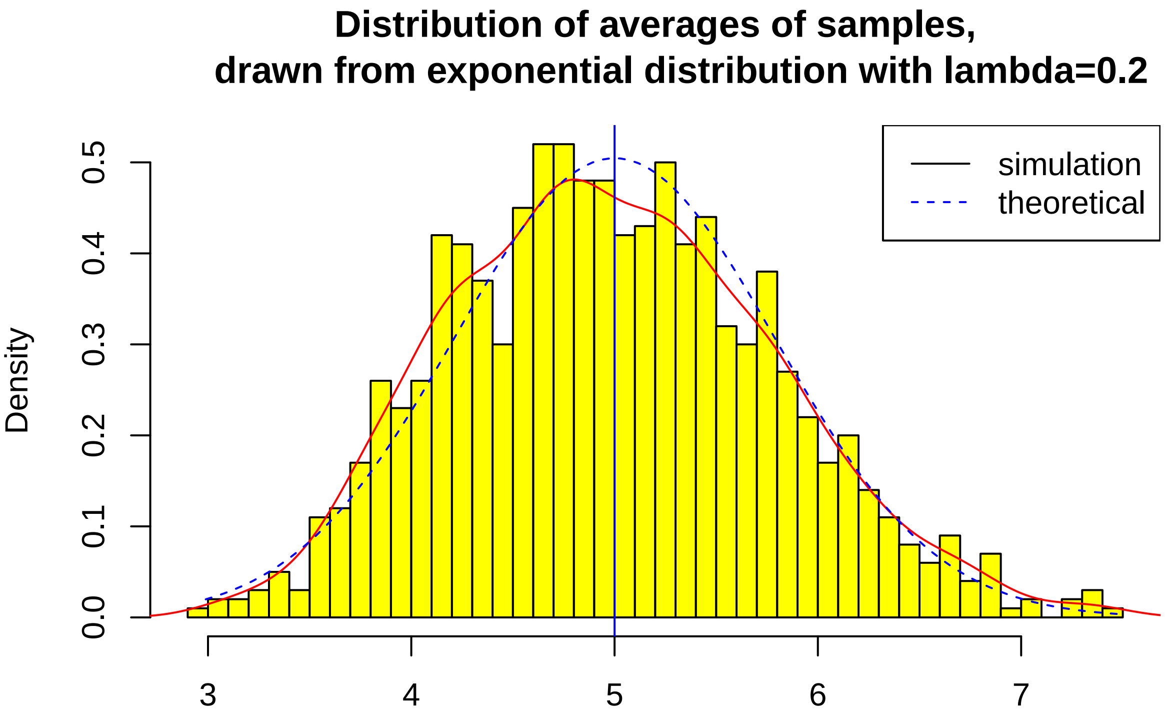
```
# plot the histogram of averages
hist(row_means, breaks=50, prob=TRUE,
     main="Distribution of averages of samples,
     drawn from exponential distribution with lambda=0.2",
     xlab="", col="yellow")

# density of the averages of samples
lines(density(row_means), col="red")

# theoretical center of distribution
abline(v=1/lambda, col="blue")

# theoretical density of the averages of samples
xfit <- seq(min(row_means), max(row_means), length=100)
yfit <- dnorm(xfit, mean=1/lambda, sd=(1/lambda/sqrt(sample_size)))
lines(xfit, yfit, pch=22, col="blue", lty=2)

# add legend
legend('topright', c("simulation", "theoretical"), lty=c(1,2), col=c("black", "blue"))
```



Therefore, the distribution of averages of 40 exponentials is centered at 4.9866 and the same is close to the theoretical center of the distribution, which is $\lambda^{-1} = 5$.

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

```
# standard deviation of distribution of averages of 40 exponentials
sd(row_means)
```

```
## [1] 0.791
```

```
# standard deviation from analytical expression
(1/lambda)/sqrt(sample_size)
```

```
## [1] 0.7906
```

```
# Variance of the sample mean
var(row_means)
```

```
## [1] 0.6258
```

```
# Theoritcal variance of the distribution
1/((0.2*0.2) * 40)
```

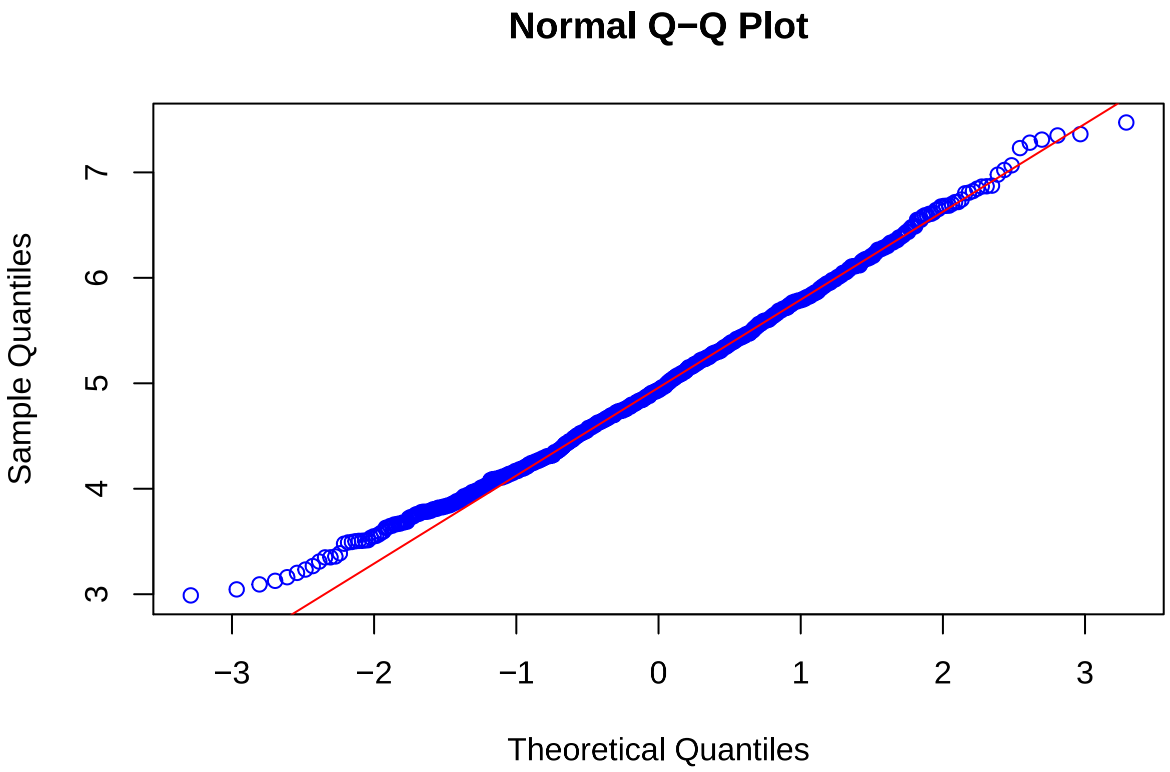
```
## [1] 0.625
```

Therefore, the variability in distribution of averages of 40 exponentials is close to the theoretical variance of the distribution.
The variance of sample means is 0.6258 where as the theoretical variance of the distribution is $\sigma^2/n = 1/(\lambda^2 n) = 1/(0.04 \times 40) = 0.625$.

3. Show that the distribution is approximately normal.

```
# use qqplot and qqline to compare the distribution of averages of 40 exponentials
# to a normal distribution
qqnorm(row_means, col="blue")

qqline(row_means, col = 2)
```



Due to the central limit theorem, 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. Also, the q-q plot suggests the distribution of averages of 40 exponentials is very close to a normal distribution.

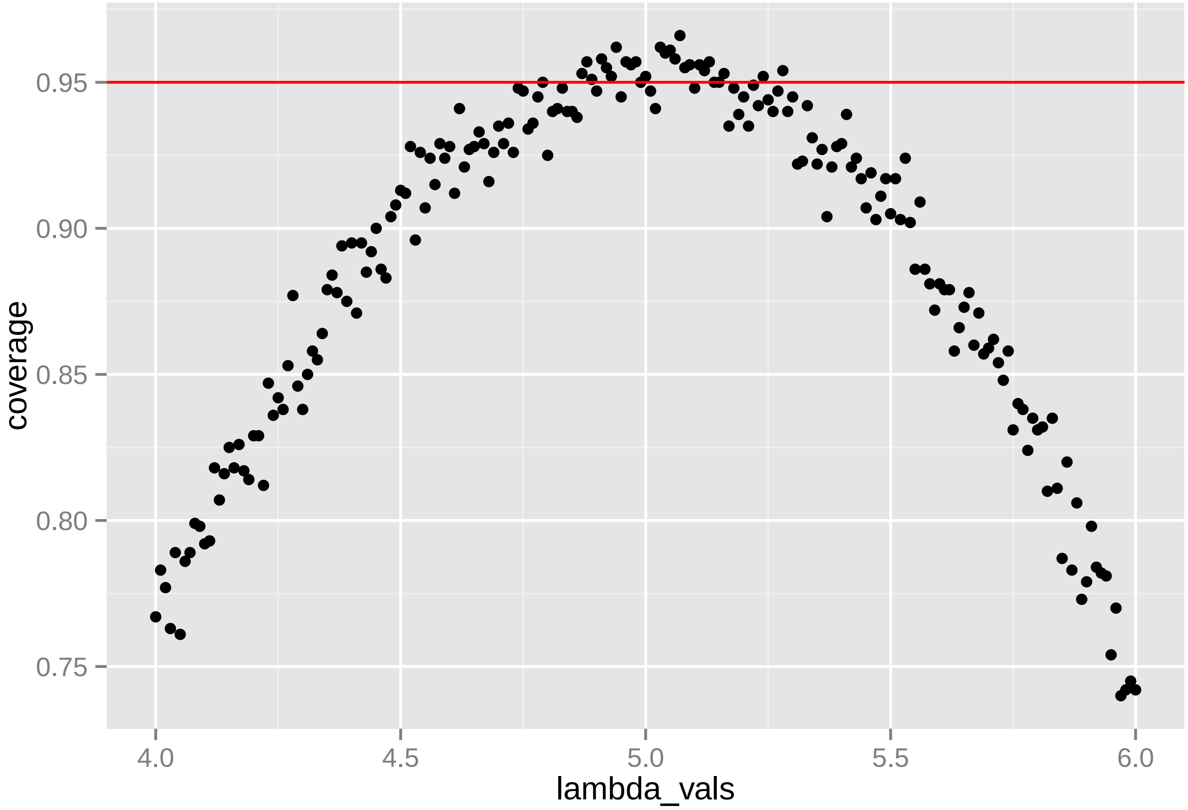
4. Evaluate the coverage of the confidence interval for $1/\lambda = \bar{X} \pm 1.96\frac{S}{\sqrt{n}}$.

```
# calculate the confidence interval
mean(row_means) + c(-1, 1) * 1.96 * sd(row_means)
```

```
## [1] 3.436 6.537
```

```
# Evaluate coverage of the confidence interval
lambda_vals <- seq(4, 6, by=0.01)
coverage <- sapply(lambda_vals, function(lamb) {
    mu_hats <- rowMeans(matrix(rexp(sample_size*simulations, rate=0.2),
                               simulations, sample_size))
    ll <- mu_hats - qnorm(0.975) * sqrt(1/lambda**2/sample_size)
    ul <- mu_hats + qnorm(0.975) * sqrt(1/lambda**2/sample_size)
    mean(ll < lamb & ul > lamb)
})

# Plot coverage
library(ggplot2)
qplot(lambda_vals, coverage) + geom_hline(yintercept=0.95, col=2)
```



The 95% confidence intervals for the rate parameter (λ) to be estimated ($\hat{\lambda}$) are $\hat{\lambda}_{low} = \hat{\lambda}(1 - \frac{1.96}{\sqrt{n}})$ and $\hat{\lambda}_{upp} = \hat{\lambda}(1 + \frac{1.96}{\sqrt{n}})$. 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.

Since, we consider the distribution of averages of exponentials, the standard deviation of this distribution already incorporates the $\sqrt{n}$ term i.e. it is the standard error.

The confidence interval is given by [3.460, 6.484].

Important Note: *Kindly visit the below link for the updated report. The plots are available in the figures folder.*
rpubs –> http://rpubs.com/bram/StatInf-Proj1-Simulation
github –> https://github.com/buva-datascience/Statistical-Inference