Statistical Inference Course Project, Part 1: Simulation exercise

Javier Nieto

Monday, January 12, 2015

Exploring Exponential Distribution

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 $1/\lambda$. In this simulation, we investigate the distribution of averages of 40 numbers sampled from exponential distribution with $\lambda = 0.2$.

Setting up required environment in R

```
# Load libraries
library(ggplot2)
library(knitr)

# Changing locale time to English
Sys.setlocale("LC_TIME", "english")
```

Simulations

Let's do a thousand simulated averages of 40 exponentials.

We generate random numbers from an exponential distribution with $\lambda = 0.20$. For each simulation, we draw 40 samples. We do this 1000 times, taking the average of the values each time.

```
# Lambda value
lambda  <- 0.2

# Number of simulations
num_sim <- 1000

# Sample numbers
n       <- 40

# Simulations
simulaciones <- do.call(rbind, replicate(num_sim, rexp(n, lambda), simplify = F))

# Average simulations
promedios <- data.frame(x = rowMeans(simulaciones))
```

Sample Measures versus theoretical measures

To show proximity between sample and theoretical averages we calculate it as following

```
avg_sim <- mean(promedios$x)
avg_teo <- 1/lambda
```

To show the variability of the simulation results, we solve for both the standard deviaion and variance statistics for both simulated and theoretical data, respectively, and then compare. The code chunk below allows us to do just this.

```
sd_sim  <- round(sd(promedios$x), 3)
var_sim <- round(sd_sim^2, 3)
sd_teo  <- round((1/lambda)*(1/sqrt(n)), 3)
var_teo <- round(sd_teo^2, 3)
```

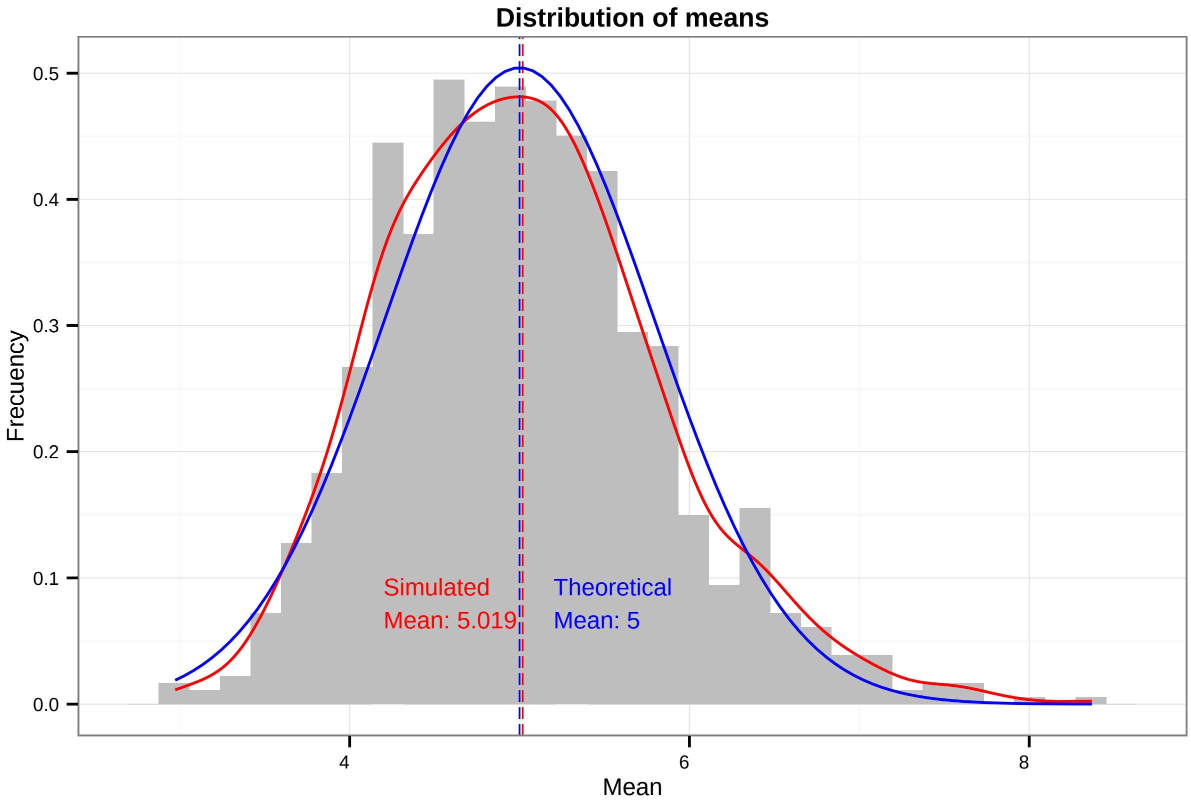
In the next table we show a summary from computed values previously

```
statistics <- data.frame(Mean = c(avg_sim, avg_teo),
                         `Standard deviation` = c(sd_sim, sd_teo),
                         Variance = c(var_sim, var_teo)
                         )

row.names(statistics) <- c("Simulated", "Theoretical")

kable(statistics, format = "pandoc", caption = "Sample and theoretical measures", dig=3)
```

```
Table: Sample and theoretical measures

                 Mean   Standard.deviation   Variance
------------   ------   ------------------   ---------
Simulated       5.019                0.813      0.661
Theoretical     5.000                0.791      0.626
```

Distribution plots

```
ggplot(promedios, aes(x=x)) +
       geom_histogram(binwidth = diff(range(promedios$x))/30,
                      fill = "gray", aes(y=..density..)) +
       ggtitle("Distribution of means") +
       xlab("Mean") +
       ylab("Frecuency") +
       theme_bw() +
       theme(plot.title   = element_text(lineheight=.8, size = 10, face  = "bold"),
             axis.text.x  = element_text(lineheight=.8, size =  7, hjust = 1),
             axis.text.y  = element_text(lineheight=.8, size =  7),
             axis.title.x = element_text(lineheight=.8, size =  9),
             axis.title.y = element_text(lineheight=.8, size =  9)
             ) +
       geom_line(stat="density", col="red") +
       geom_vline(xintercept=avg_sim, lwd=.3, col="red", linetype = "longdash") +
       annotate("text", x = 4.2, y = 0.08,
             label = paste("Simulated\nMean:", round(avg_sim,3), " "), cex=3,
                           col="red", hjust=0 ) +
    stat_function(fun = dnorm, arg = list(mean = avg_teo, sd = sd_teo),
                  color="blue") +
    geom_vline(xintercept=avg_teo, lwd=.3, col="blue", linetype = "longdash") +
    annotate("text", x = 5.2, y = 0.08,
             label = paste("Theoretical\nMean:", round(avg_teo,3), " "), cex=3,
                           col="blue", hjust=0 )
```


Distribution of means
Frecuency
Simulated
Mean: 5.019
Theoretical
Mean: 5
Mean