Statistical Inference Course Project1

doyougnu

July 20, 2015

Overview

Then following report will show the Central Limit Theorem (henceforth [CLT]) in action through the generation of a distribution of sample means, calculated from 1000 simulations of 40 random exponential deviates. The report will show that the sample mean distribution obeys the CLT by being normal-like distributed, while the population distribution is exponential.

Simulations

```r
#set seed
set.seed = 1234

#number of simulations to run
simnum <- 1000

#number of distributions to generate
exp_n <- 40

#exponential distribution parameters
exp_lambda <- 0.2
simMatrix <- matrix(rexp(exp_n * simnum, rate = exp_lambda), simnum, exp_n)
expMean <- rowMeans(simMatrix)
```

The simulation is done by generating 40,000 random exponential deviates with a lambda value = 0.2. Then populating a 1000 x 40 matrix with those values. We then call the rowMeans function to find the mean of each row (that is the mean of 40 simulation random exponential deviates). The rowMeans create our simulation distribution of means, called expMean.

Plots Comparing Simulation to Population

```r
#Compare Means
popMean <- 1 / exp_lambda
simMean <- mean(expMean)
```

The means are very close, the population mean is 5, the simulated mean is 5.0346682

```r
#Compare Variances
popVariance <- (1 / exp_lambda) ^ 2 / exp_n
simVariance <- var(expMean)
```

Similarly for the Variances, the pop variance is 0.625, the sim variance is 0.5810192

Normality test

```
#perform shapiro normality test
shapiro.test(expMean)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  expMean
## W = 0.99604, p-value = 0.01164
```

The shapiro wilk's normality test, test's the Null Hypothesis: That the samples come from a normal distribution, against the alternate Hypothesis: That the samples **do not** come from a normal distribution. As shown above, our p-value is < 0.05, so our samples do not form a normal distribution, but instead are normal-like.

Visual Comparison of Exponential Population vs. Samples

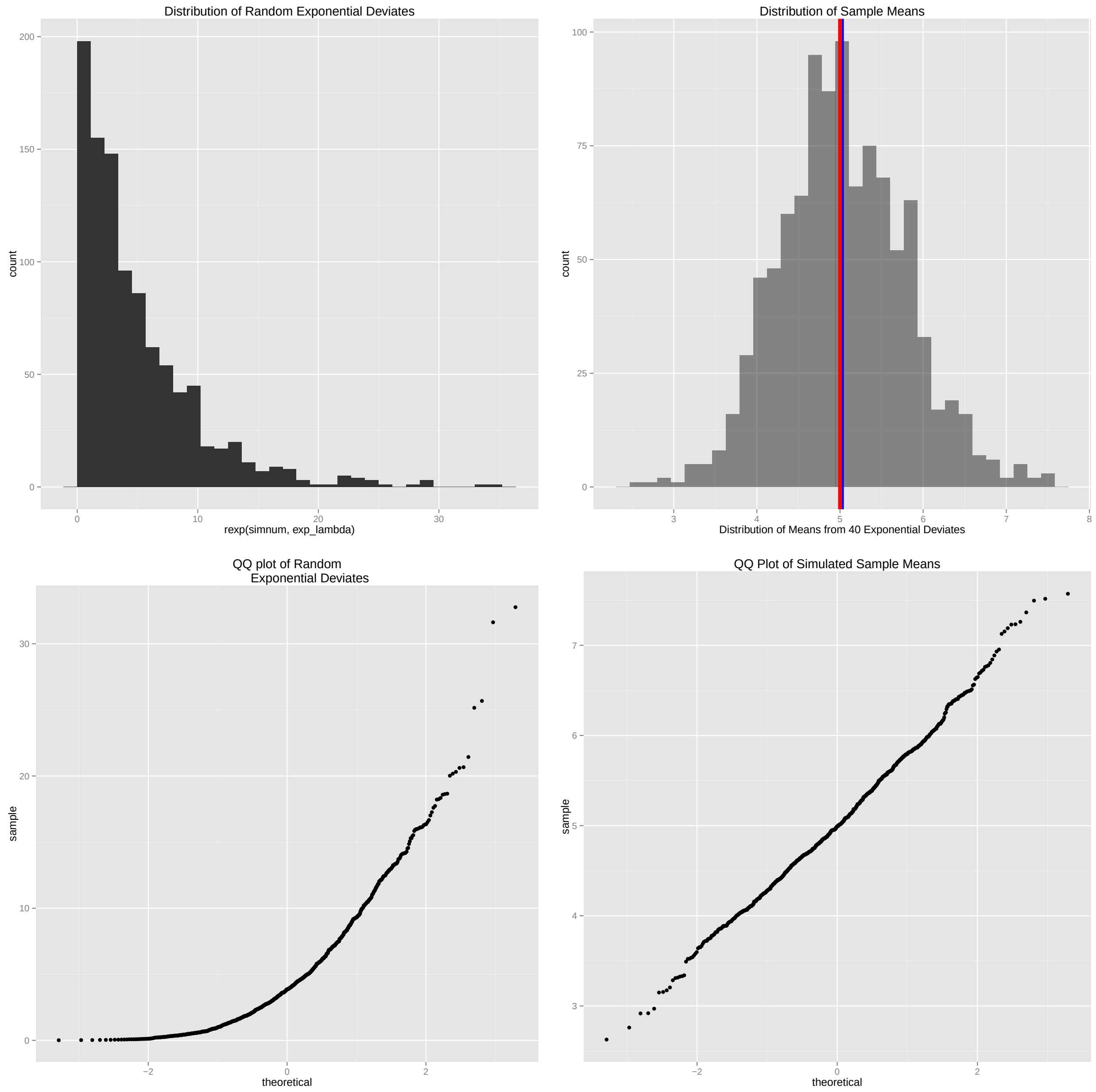
```
#load libs
library(ggplot2)
library(gridExtra)
library(gridBase)

#plot QQ plot
qQ <- qplot(sample = expMean, main = "QQ Plot of Simulated Sample Means")
qEQ <- qplot(sample = rexp(simnum, exp_lambda), main = "QQ plot of Random
             Exponential Deviates")

#plot rexp deviates
qE <- qplot(rexp(simnum, exp_lambda), main = "Distribution of Random Exponential Deviates")

#compare Variances visually
plt <- qplot(expMean, main = "Distribution of Sample Means", alpha = 0.3) +
    xlab("Distribution of Means from 40 Exponential Deviates") +
    geom_vline(size = 2, xintercept = popMean, colour = "red") +
    geom_vline(size = 1, xintercept = simMean, colour = "blue") +
    theme(legend.position = "none")

#compare plots in single panel
grid.arrange(qE, plt, qEQ, qQ)
```


Distribution of Random Exponential Deviates

Distribution of Sample Means

QQ plot of Random Exponential Deviates

QQ Plot of Simulated Sample Means

In the above plot we can see the CLT in action. Starting from the top left we have the distribution of 1000 random deviates; compare this with the sample of means of 40 random exponential deviates generated 1000 times. As you can see the top left plot is clearly exponential while the sample plot (top right) is more of a normal distribution (due to its bell shape). The red vertical line represents the Theoretical Mean, while the blue line represents the Simulation Mean. As you see the two are extremely close together (I altered the sizes of both lines so one could observe their proximity without overlap). The bottom left plot shows the Normal Quantile plot for the 1000 Random Exponential Deviates, it is obviously non-normal, compared to the plot on the lower right, which shows the Normal Quantile plot of the simulation samples; it is much more normal (Normal distributions are indicated by a linear line in a QQ plot, perfom a qplot(sample = rnorm(1000)) to see a normal distribution QQ plot) then the Exponential Deviate plot. This my friend, is the CLT observed.