Statistical Inference Project: Simulations

Stephanie Denis

3/25/2017

Overview

The objective of this exercise is to compare the sampling distribution of an exponential distribution to the Central Limit Theorem. The results provide strong evidence that the sampling distribution of the mean and of the variance is centered around the theoretical mean and the theoretical variance. Additionally, the sample mean and the sample variance are approximately normally distributed, confirming the predictions of the Central Limit Theorem.

The Central Limit Theorem

The Central Limit Theorem requires that the sampled observations be independent and identically distributed and the sample size be $n \geq 30$. As the sample size increases, the Central Limit Theorem states that the sampling distribution becomes that of a standard normal. In other words, the mean and the variance of all samples from the same population will be approximately equal to the mean and the variance of the population.

Simulations

```r
#setup variables
n <- 40
lambda <- 0.2
nosim <- 1000
mu <- 1/lambda
sigma <- 1/lambda
var <- (sigma)^2
#simulations
set.seed(1010093)
mns <- replicate(nosim, mean(rexp(n, lambda)))
vars <- replicate(nosim, var(rexp(n, lambda)))
```

The data set consists of an exponential distribution of 40 observations with a mean, μ, and standard deviation, σ, both of which are equal to $1/\lambda$, where λ is equal to 0.2. The number of simulations is equal to 1000.

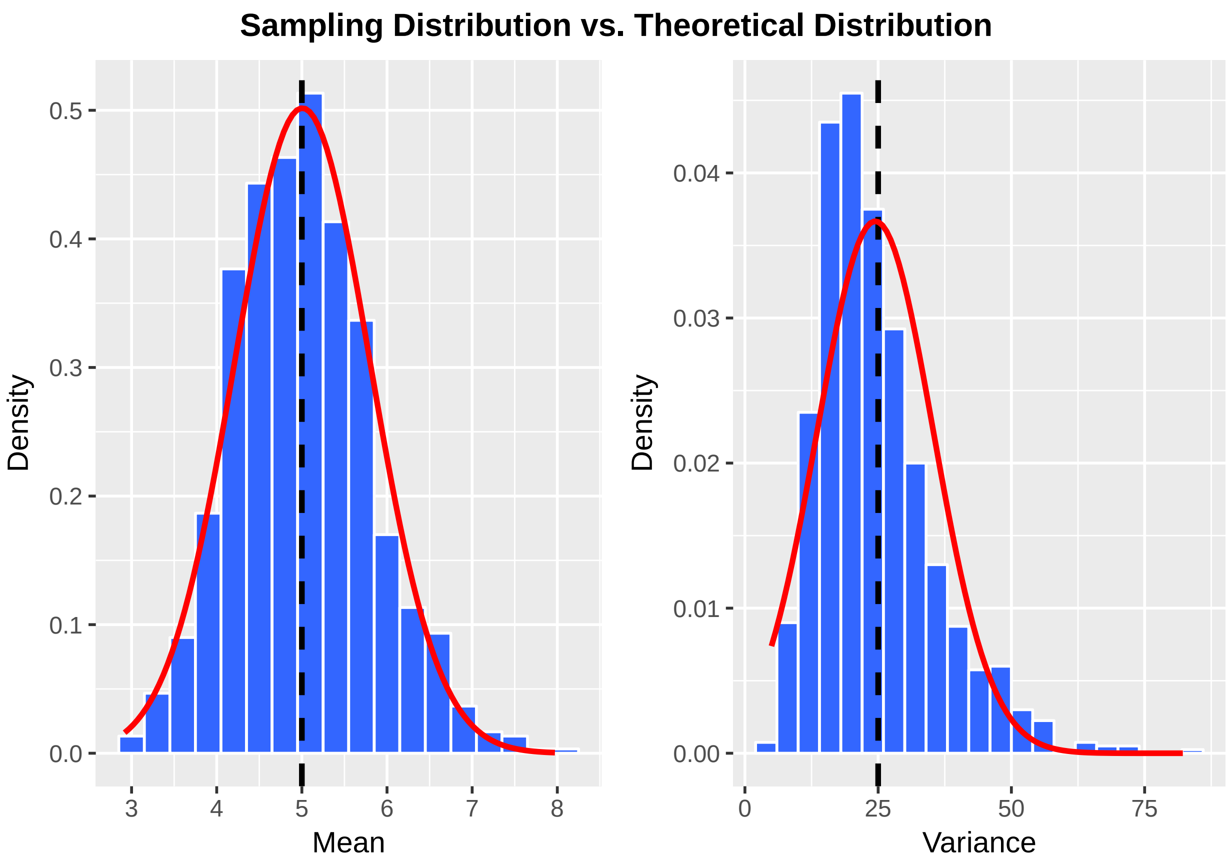
Distribution

1. Sampling Distribution vs. Theoretical Distribution

```r
sample_mn <- round(mean(mns), 2) #sample mean
sample_var<- round(mean(vars), 2) #sample variance
```

The sample mean (5.01) and the sample variance (24.45) are very close to the theoretical mean (5) and the theoretical variance (25).

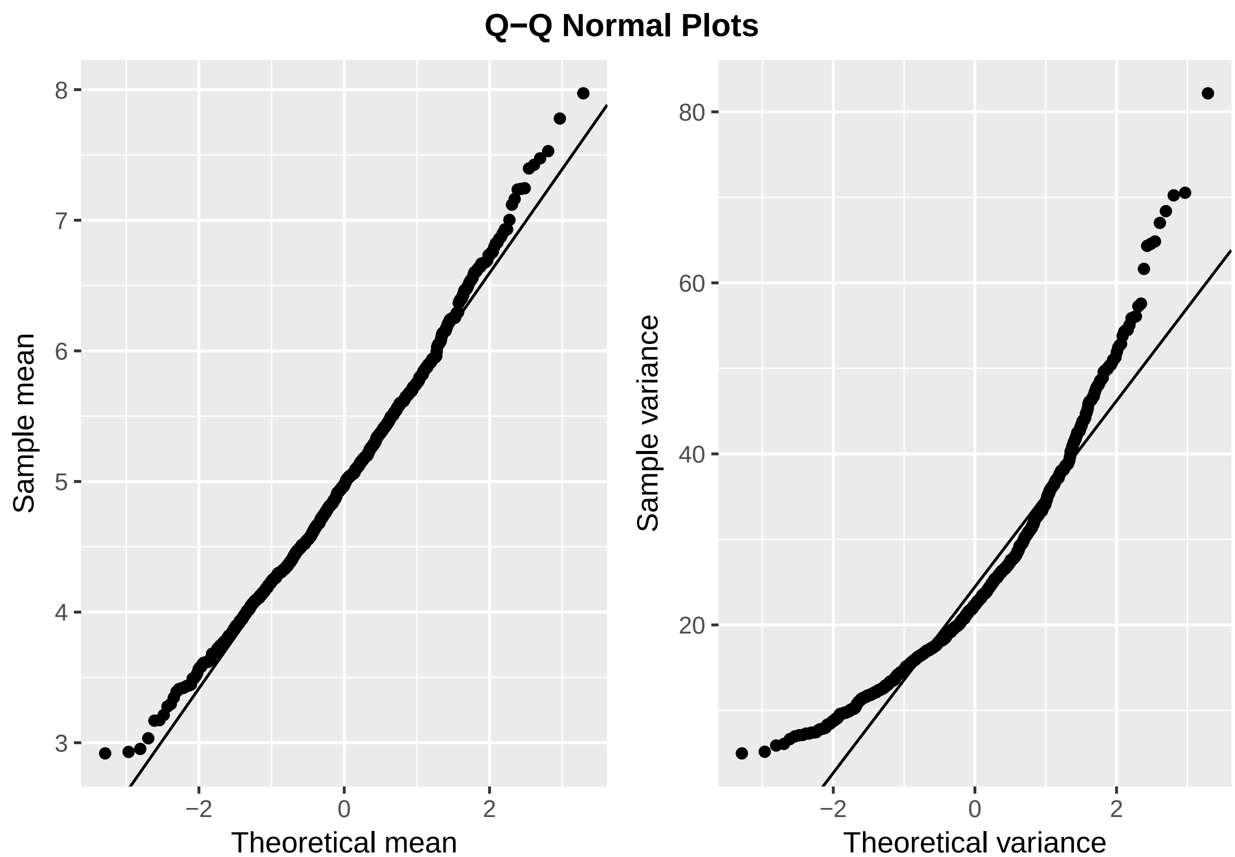
```r
#load packages
library(ggplot2); library(grid); library(gridExtra)
#histograms
p1 <- ggplot(data.frame(mns), aes(x = mns)) +
        geom_histogram(binwidth = .3, colour = "white",
                       fill = "#3366FF", aes(y = ..density..)) +
        geom_vline(xintercept = mu, aes(fill = "black"),
                   linetype = "dashed", size = 1) +
        stat_function(fun = dnorm, colour = "red", size = 1,
                      args = list(mean = mean(mns), sd = sd(mns))) +
        labs(x = "Mean", y = "Density")
p2 <- ggplot(data.frame(vars), aes(x = vars)) +
        geom_histogram(binwidth = 4, colour = "white",
                       fill = "#3366FF", aes(y = ..density..)) +
        geom_vline(xintercept = var, aes(fill = "black"),
                   linetype = "dashed", size = 1) +
        stat_function(fun = dnorm, colour = "red", size = 1,
                      args = list(mean = mean(vars), sd = sd(vars))) +
        labs(x = "Variance", y = "Density")
#panel1
title1=textGrob("Sampling Distribution vs. Theoretical Distribution",
              gp = gpar(fontsize = 12, fontface = "bold"))
grid.arrange(p1, p2, nrow = 1, ncol = 2, top = title1)
```



The histograms illustrate the sampling distrubtion of the mean and variance. The bell-shaped curve (red line) represents the theoretical density suggesting that the sampling distribution (blue bars) is approximately normally distributed and is centered around the theoretical mean and variance (vertical dashed black line).

2. Normality Test

```
#qq plots
p3 <- ggplot(data.frame(mns), aes(sample = mns)) +
        stat_qq() +
        geom_abline(intercept = mean(mns), slope = sd(mns)) +
        labs(x = "Theoretical mean", y = "Sample mean")
p4 <- ggplot(data.frame(vars), aes(sample = vars)) +
        stat_qq() +
        geom_abline(intercept = mean(vars), slope = sd(vars)) +
        labs(x = "Theoretical variance", y = "Sample variance")
#panel2
title2=textGrob("Q-Q Normal Plots",
                gp = gpar(fontsize = 12, fontface = "bold"))
grid.arrange(p3, p4, nrow = 1, ncol = 2, top = title2)
```



The quantile plots offer a more formal way to test for normality. The points falling along the straight theoretical line of the quantile plots suggest an approximately normally distributed sampling distribution.

Conclusion

The results provide strong evidence that the sampling distribution of the mean and the variance is centered around the theoretical mean and the theoretical variance. Additionally, the sample mean and the sample variance are approximately normally distributed confirming the predictions of the Central Limit Theorem.