Statistical Inference Course Project

T-StrawClown

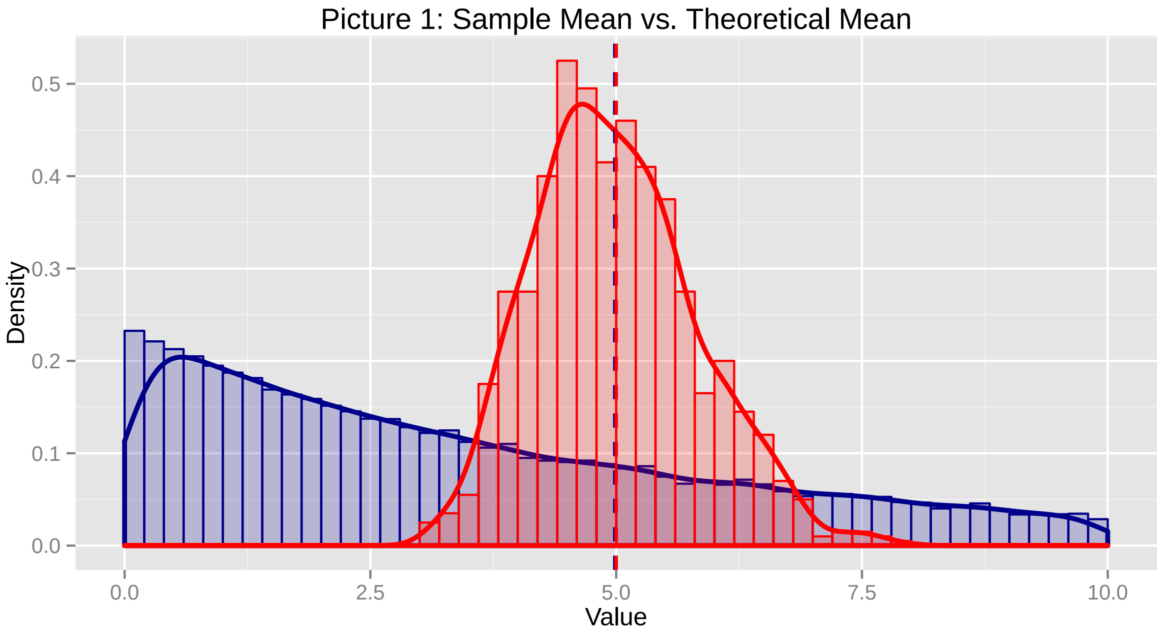
Overview

This exercise is done as part of Coursera Data Science Specialization, Statistical Interference course. We'll investigate the exponential distribution in R and compare it with the Central Limit Theorem. The exponential distribution is simulated in R with rexp(n, λ) where lambda is the rate parameter. The mean of exponential distribution is $\frac{1}{\lambda}$ and the standard deviation is also $\frac{1}{\lambda}$. We'll use $\lambda = 0.2$ for all of the simulations. We'll investigate the distribution of averages of 40 exponentials.

We're going to generate 1000x40 matrix with randomly generated data. Each row in matrix contains exponentially distributed sample data of 40 instances.

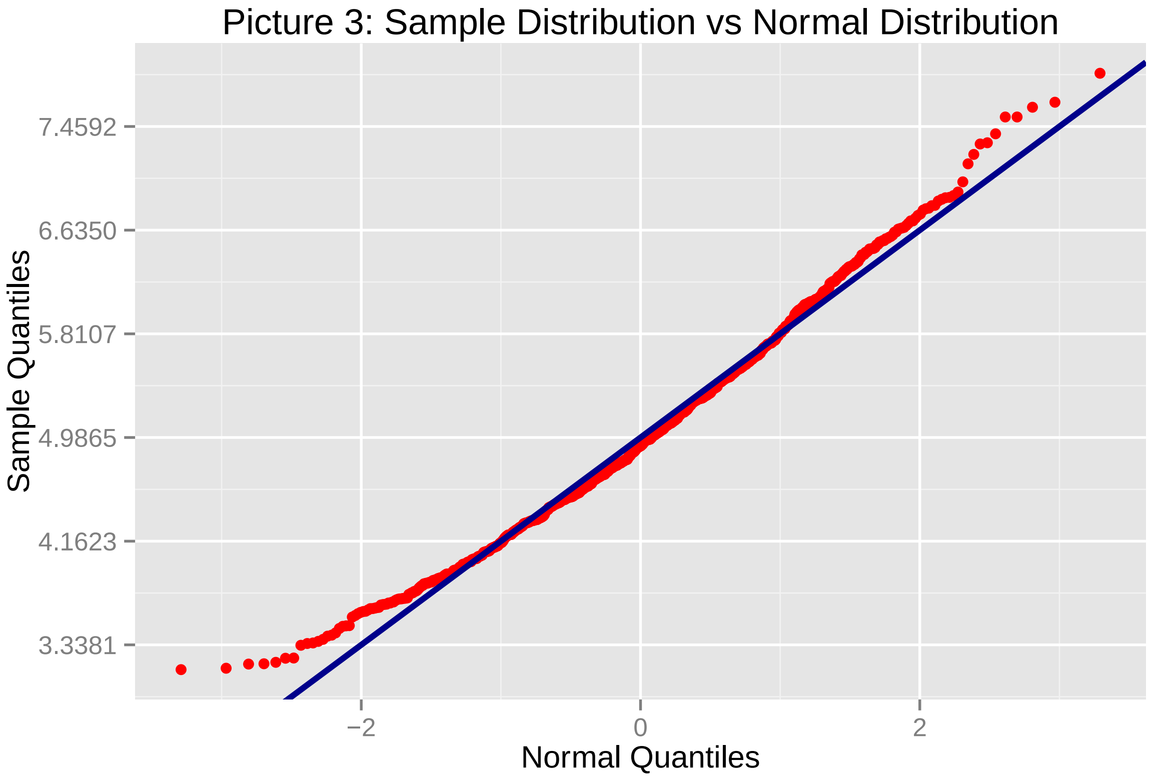
Sample versus Theoretical

Let's examine how exponentially distributed data from our simulations looks compared to distribution of means of 40 samples. The blue portion of the graph represents exponential distribution (the fragment of it to be honest, in reality it continues to the right, but we are saving space here) and the red portion is about our simulated means.



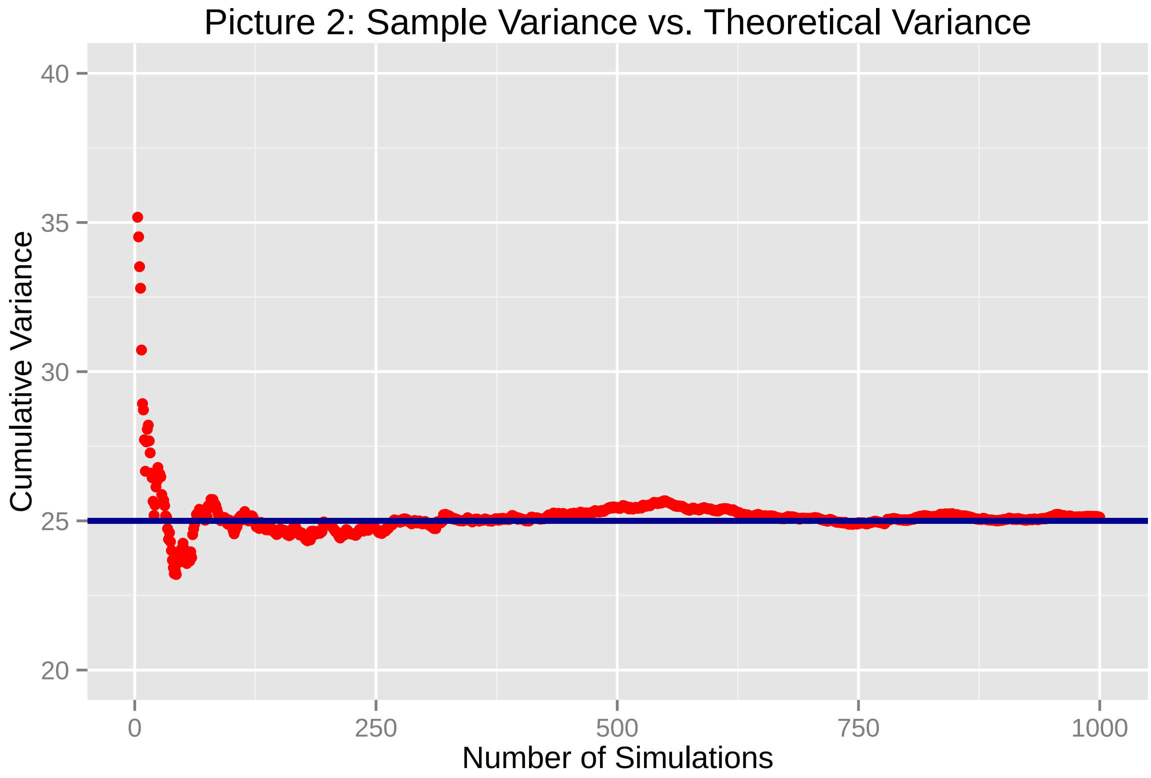
Although our sample size 40 is not that big, but is already hard to see that there are 2 vertical dashed lines, the blue one is theoretical mean of exponential distribution, which is 5, and the red one is actual mean of simulated data which is 4.9865. Note that red distribution looks almost normal, if we would increase the size of our sample to more than 40, it would look even more Gaussian. The CLT states that the distribution of averages of iid variables becomes that of a standard normal as the sample size increases.

Let's take a look at quantile to quantile comparison of our simulated data to normally distributed data. Red dots represent a single mean of 40 exponentially distributed samples in quantiles from the mean of all simulated data (mean of 1000 means) and blue line represents normal distribution in quantiles.



The values vertical axis is scaled in standard deviations of simulated data from the mean of simulated data 4.9865. Proximity of red dots to the blue line suggest that our simulation follows the CLT and distribution of means of randomly generated exponential data is very close to perfectly normal distribution, represented by the blue line. It would be even closer if we would increase our sample size to more that 40.

Let's get to variance now. The theoretical variance of exponential distribution is $\frac{1}{\lambda^2}$, which in our experiment is $\frac{1}{0.2^2} = 25$. Let's use the asymptotics and the Law of Large Numbers to investigate variance of our simulated data samples. Let's plot the cumulative variance and check were it is going.



As the picture suggests the variance of simulated samples (represented by red dots) approaches theoretical variance of exponential distribution (represented by the blue line) as the number of simulations increases.

Supplement 1: R code

```r
library(ggplot2)
set.seed(42)
lambda <- .2
sims <- 1000
n <- 40
# 1000 simulations
sim_data <- rexp(sims * n, lambda)
exp_data <- matrix(sim_data, sims, n)
means <- rowMeans(exp_data)
vars <- apply(exp_data, 1, var)
g1 <- ggplot() +
        scale_x_continuous(limits = c(0, 10)) +
        ggtitle("Picture 1: Sample Mean vs. Theoretical Mean") +
        xlab("Value") +
        ylab("Density") +
        geom_histogram(data = data.frame(x = sim_data),
                       aes(x = x, y = ..density..),
                       fill = "darkblue",
                       binwidth = .2,
                       color = "darkblue",
                       alpha = .2) +
        geom_density(data = data.frame(x = sim_data),
                     aes(x = x, y = ..density..),
                     fill = "transparent",
                     color = "darkblue",
                     size = 1.1) +
        geom_histogram(data = data.frame(x = means),
                       aes(x = x, y = ..density..),
                       binwidth = .2,
                       fill = "red",
                       color = "red",
                       alpha = .2) +
        geom_density(data = data.frame(x = means),
                     aes(x = x, y = ..density..),
                     fill = "transparent",
                     color = "red",
                     size = 1.1) +
        geom_vline(xintercept = mean(means),
                   size = 1,
                   color = "darkblue",
                   linetype = 2) +
        geom_vline(xintercept = 1 / lambda,
                   color = "red",
                   size = 1,
                   linetype = 2)
print(g1)

#sample variance approaching true variance
g2 <- ggplot(data = data.frame(y = cumsum(vars) / 1:sims),
            aes(y = y, x = 1:sims)) +
        scale_y_continuous(limits = c(20, 40)) +
        ggtitle("Picture 2: Sample Variance vs. Theoretical Variance") +
        xlab("Number of Tries") +
        ylab("Cumulative Variance") +
        geom_point(color = "red",
                   size = 2) +
        geom_hline(yintercept = 1 / lambda^2,
                   color = "darkblue",
                   size = 1.1)
print(g2)

#quantiles of means vs normal distribution quantiles
g3 <- ggplot(data = data.frame(x = means),
             aes(sample = x)) +
        ggtitle("Picture 3: Sample Distribution vs Normal Distribution") +
        xlab("Normal Quantiles") +
        ylab("Sample Quantiles") +
        scale_y_continuous(breaks = round(mean(means) +
                                    seq(-3, 3, by = 1) * sd(means), 4)) +
        stat_qq(color = "red") +
        geom_abline(slope = sd(means),
                    intercept = mean(means),
                    color = "darkblue",
                    size = 1.1)
print(g3)
```