Central Limit Therem Demonstration using Exponetial Distribution

Statistical Inference Course Project 1

Xinchun Tian

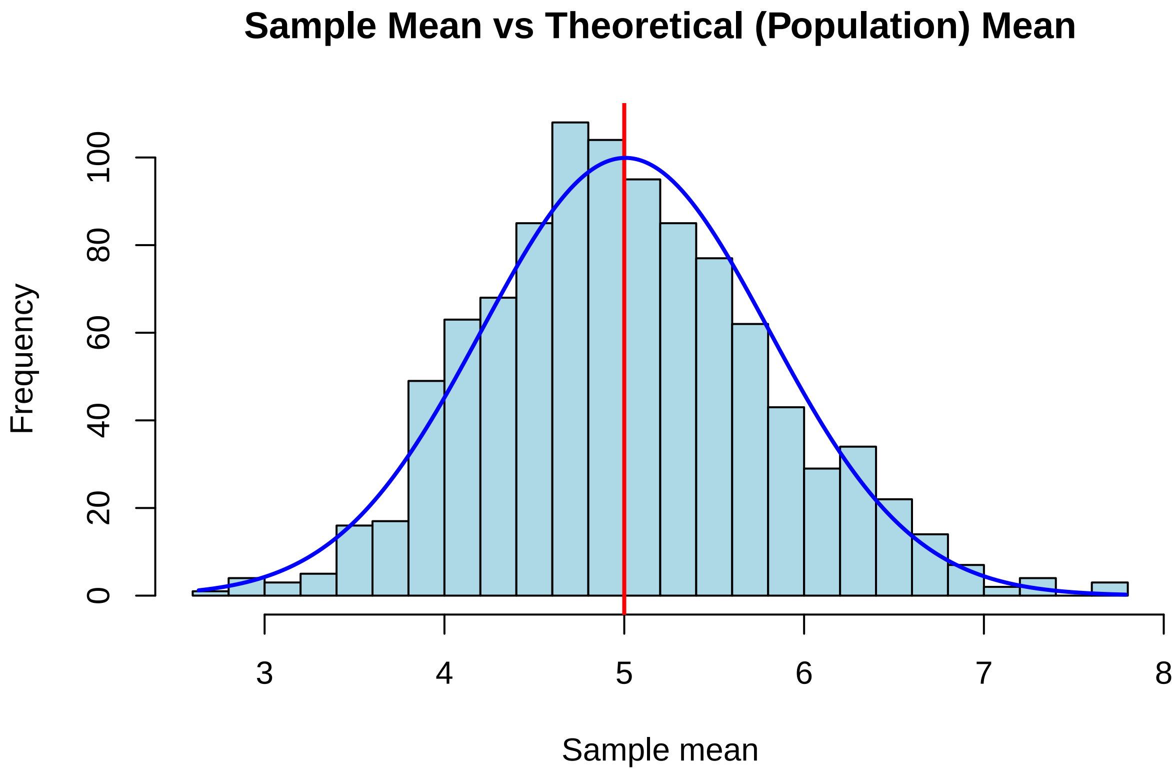
Overview

According to the Central Limit Theorem (CLT), given sufficiently large number of sample statistics with independent random variables, both the sample mean and variance, will be approximately normally distributedn. The sample mean is the estimate of the population mean, and the sample variance is the estimate of the popluation variance. The distribution used to demonstrate the CLT in this document is the exponential distribution: $y = e^{-x}$.

Sample Mean vs Theoretical Mean

Following the example code from the course project page, the following code will simulate 40 exponential processes with $\lambda = 0.2$, take their sample mean and repeat this process 1,000 times.

```r
set.seed(20150810)
lambda <- 0.2
mns    <- NULL
for (i in 1:1000) mns = c(mns, mean(rexp(40, rate = lambda)))
h <- hist(mns, col="lightblue", breaks=20, xlab="Sample mean",
          main="Sample Mean vs Theoretical (Population) Mean")
abline(v=1/lambda, col="red", lwd=2)
xfit <- seq(min(mns), max(mns), length=1000)
yfit <- dnorm(xfit, mean=mean(mns), sd=sd(mns))
yfit <- yfit*diff(h$mids[1:2])*length(mns)
lines(xfit, yfit, col="blue", lwd=2)
```

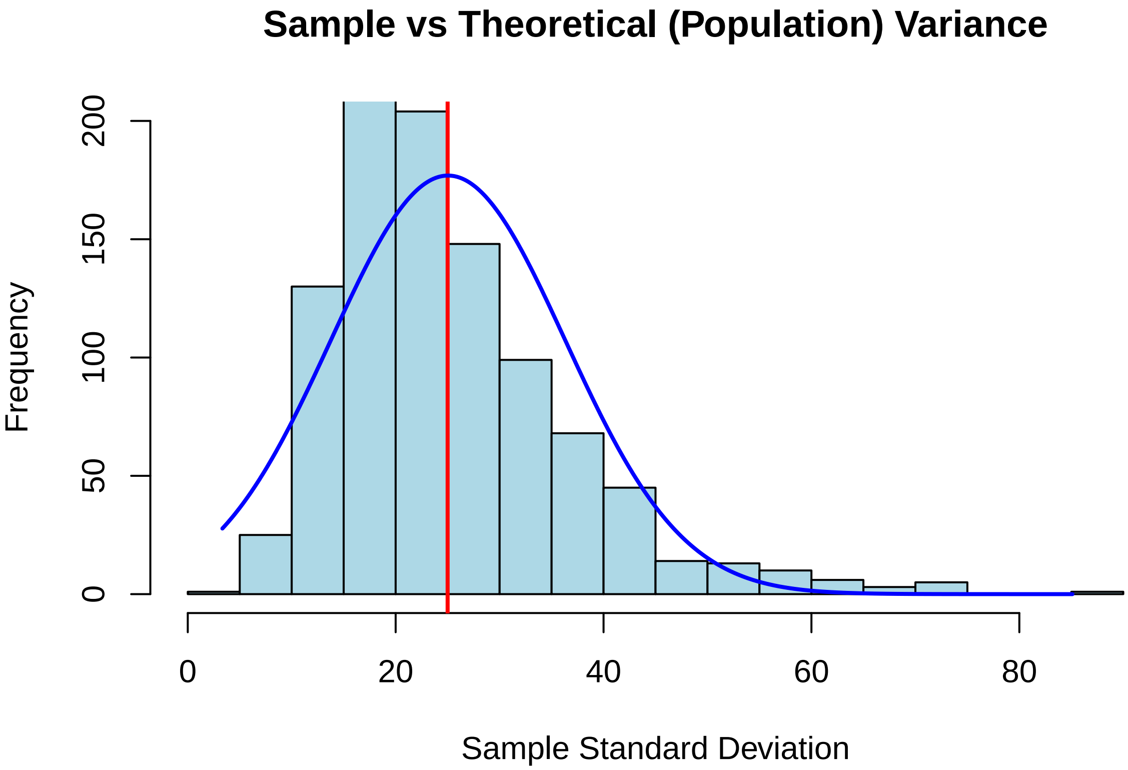


The above figure shows the sample mean (5.006) in light blue histogram vs the theoretical mean (μ=5) which is the vertical red line. The sample mean has a bell shape which is consistent with the Gaussian density curve in blue. With sufficient sample statistics, the sample mean is a good estimate of the population mean or theoretical mean.

Sample Standard Deviation vs Theoretical Standard Deviation

Similarly, the following code will simulate 40 exponential processes with $\lambda = 0.2$, take their sample standard deviation and repeat this process 1,000 times.

```
set.seed(20150710)
lambda <- 0.2
n      <- 40
nosim  <- 1000
exp_df <- data.frame(matrix(rexp(n*nosim, rate=lambda), nrow=nosim, ncol=n))
vars   <- apply(exp_df, 1, var)
h <- hist(vars, col="lightblue", breaks=20, xlab="Sample Standard Deviation",
          main="Sample vs Theoretical (Population) Variance", ylim=c(0,200))
abline(v=(1/lambda)^2, col="red", lwd=2)
xfit <- seq(min(vars), max(vars), length=1000)
yfit <- dnorm(xfit, mean=mean(vars), sd=sd(vars))
yfit <- yfit*diff(h$mids[1:2])*length(vars)
lines(xfit, yfit, col="blue", lwd=2)
```



The above figure shows the sample variance (25.039) in light blue histogram vs the theoretical variance (σ^2=25) which is the vertical red line. The sample variance has a bell shape which is consistent with the Gaussian density curve in blue. With sufficient sample statistics, the sample variance is a good estimate of the population variance or theoretical variance.