Central Limit Therem Demonstration using Exponetial Distribution

Statistical Inference Course Project

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: $y = e^{-x}$.

Generate the data sample

```
nosim  <- 1000 # number of simulations
n      <- 40   # sample size
lambda <- 0.2  # lambda
set.seed(1)    # set the seed for the first simulation
matrix_exp <- matrix(rexp(n,lambda),1) # simulate the first sample (first row)
# simulate the next 999 samples and merge them as one data frame
for (i in 2:nosim) {
  set.seed(i)
  tmp  <- matrix(rexp(n,lambda),1)
  matrix_exp <- rbind(matrix_exp, tmp)
}
df_exp <- as.data.frame(matrix_exp)
```

The generated data set is a 1000×40 dimension data frame. Each row is one sample statistic with the size of 40.

```
dim(df_exp)
```

```
## [1] 1000   40
```

Here shows the first few rows of the simulated data set.

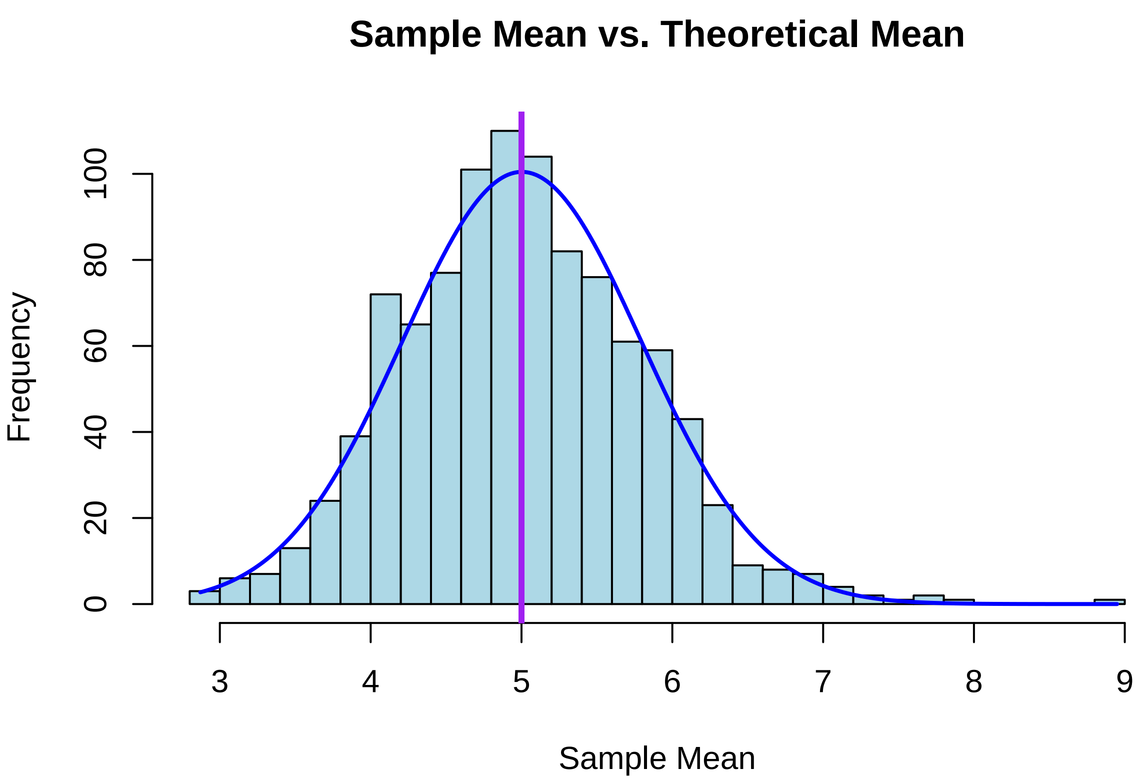
```
df_exp[1:2,]
```

```
##         V1       V2        V3        V4        V5        V6       V7
## 1 3.775909 5.908214 0.7285336 0.6989763 2.1803431 14.474843 6.147810
## 2 9.326762 2.023740 0.7332633 8.6535486 0.4476309  3.334488 5.371834
##         V8       V9       V10      V11      V12      V13      V14
```

```
## 1 2.698414 4.782837 0.7352300 6.953676 3.810149 6.188018 22.119671
## 2 7.558147 6.571380 0.7826514 3.725595 6.216727 3.368892  7.951529
##        V15      V16      V17       V18      V19      V20       V21
## 1 5.272716 5.176220 9.380176  3.273733 1.684667 2.942399 11.822576
## 2 5.411181 3.982986 7.226352 22.459712 8.517161 3.101921  1.782365
##        V22      V23      V24       V25       V26      V27       V28
## 1 3.209463 1.470602 2.829328 0.5303631 0.2971958 2.893562 19.794664
## 2 3.444290 4.163147 1.672256 7.9502619 0.3988510 2.754788  5.447504
##          V29      V30       V31       V32      V33      V34       V35
## 1 5.86656053 4.984065  7.176427 0.1863426 1.620051 6.602340 1.0175518
## 2 0.01969881 3.103973 24.311077 0.9544678 2.858789 2.127799 0.9021312
##        V36      V37      V38      V39       V40
## 1 5.113629 1.508705 3.626072 3.757713  1.175137
## 2 1.345207 3.694251 3.979756 4.673358 15.611028
```

Let's plot the sample statistics on the sample mean.

```r
x    <- apply(df_exp, 1, mean)
h    <- hist(x, breaks=40, col="lightblue", xlab="Sample Mean",
             main="Sample Mean vs. Theoretical Mean")
xfit <- seq(min(x), max(x), length=nosim)
yfit <- dnorm(xfit, mean=mean(x), sd=sd(x))
yfit <- yfit*diff(h$mids[1:2])*length(x)
lines(xfit, yfit, col="blue", lwd=2)
abline(v=1/lambda, col="purple", lwd=3)
```

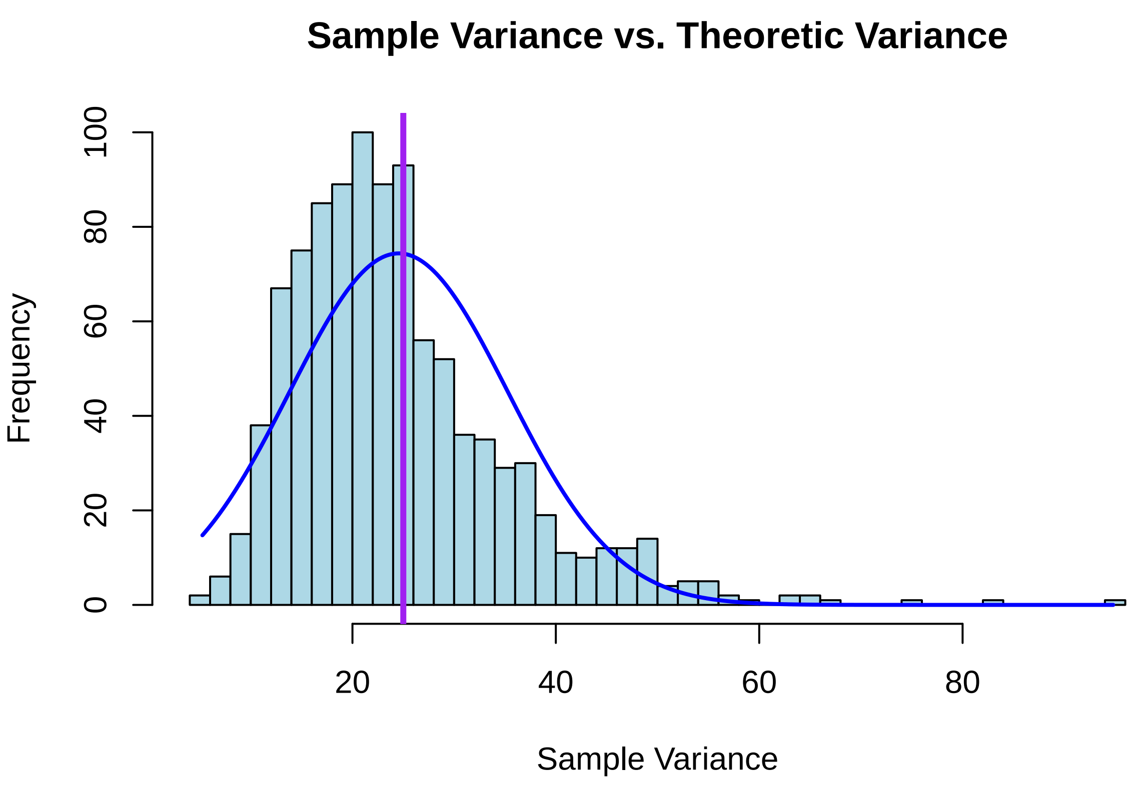


```
x_mean <- mean(x)
```

1. **The plot shows that the sample statistic has a bell shape which is consistent with the the gaussian density curve in blue with the mean (5) where the theoretical mean of the population distribution is 5.0 as the purple vertical line shows. With sufficient sample statistics, the sample mean is a good estimate of the population mean.**

Let's plot the sample statistics on the sample variance.

```
x    <- apply(df_exp, 1, var)
h    <- hist(x, breaks=40, col="lightblue", xlab="Sample Variance",
             main="Sample Variance vs. Theoretic Variance")
xfit <- seq(min(x), max(x), length=nosim)
yfit <- dnorm(xfit, mean=mean(x), sd=sd(x))
yfit <- yfit*diff(h$mids[1:2])*length(x)
lines(xfit, yfit, col="blue", lwd=2)
abline(v=(1/lambda)^2, col="purple", lwd=3)
```




```
# d <- density(x)
x_sd <- mean(x)
```

2. **The plot shows that the variance of the sample statistic has a skewed bell shape (the blue curve is the gaussian density distribution) with the mean (24.55) where the theoretical variance of the population distribution is 25.0 as the purple vertical line shows. More sample statistics may be needed to better estimate the population variance.**