Statistical Inference Project

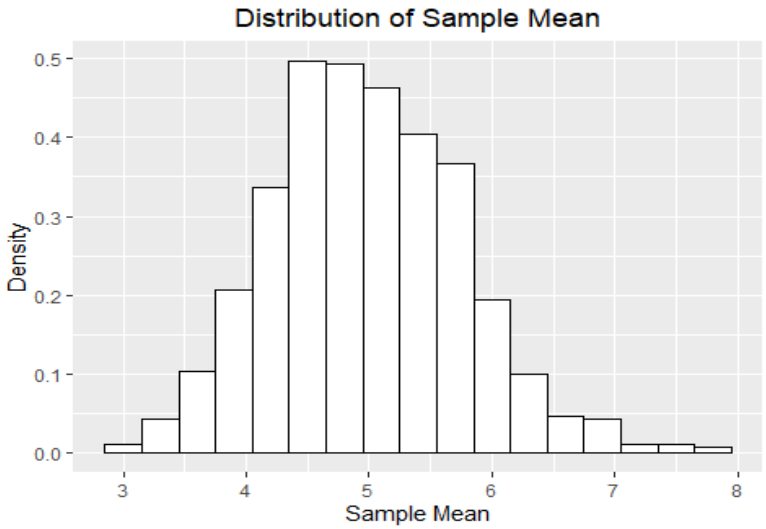
Shivam Mishra

07/08/2020

Overview

With this report we explore the relationship between the sample mean and the population mean from exponential distribution. Parameters and statistics of interest are the mean, the variance and, in a broader sense, the shape of the distribution. We run 1000 simulations of 40 observations each time.
Simulations We create a vector of size 1000 and fill it with the sample means of 1000 samples, each of which has size 40 observations. The distribution from which we draw the samples is exponential with rate parameter L = 0.2.

```
mns <- apply(matrix(rexp(40000, 0.2), 1000), 1, mean)
mns <- data.frame(mns)
## histogram of the distribution
library(ggplot2)
g <- ggplot(mns, aes(x = mns))
g <- g + geom_histogram(aes(y = ..density..), color = "black", fill = "white", binwidth = 0.3) +
  labs(x = "Sample Mean", y = "Density") + labs(title = "Distribution of Sample Mean") +
  theme(plot.title = element_text(hjust = 0.5))
print(g)
```



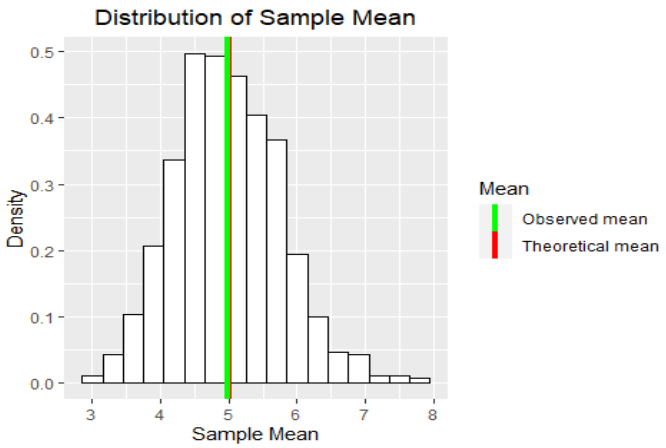
Calculating the Constants

```
obs_mean <- mean(mns[, 1]) ## observed mean
theo_mean <- 1/0.2         ## theoretical mean
Obs_var <- var(mns[, 1])   ## observed variance
theo_var <- 1/(0.2*0.2*40) ## theoretical variance
```

Sample Mean Vs Theoretical Mean

```
colors <- c("Theoretical mean" = "red", "Observed mean" = "green")
g <- ggplot(mns, aes(x = mns))
g <- g + geom_histogram(aes(y = ..density..), color = "black", fill = "white", binwidth = 0.3) +
  geom_vline(aes(xintercept = theo_mean, color = "Theoretical mean"), size = 1.5) +
  geom_vline(aes(xintercept = obs_mean, color = "Observed mean"), size = 1.5) +
  scale_color_manual(values = colors) + labs(x = "Sample Mean", y = "Density") + labs(title = "Distribution o
f Sample Mean") +
```

```
  theme(plot.title = element_text(hjust = 0.5)) + labs(colour = "Mean")
print(g)
```

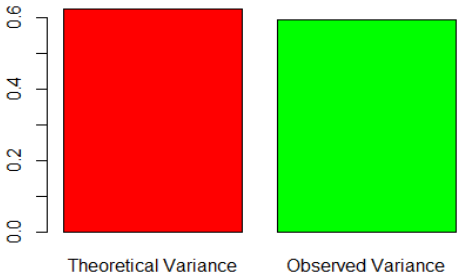



```
print(paste("The Sample Mean is ", obs_mean, sep = " "))

## [1] "The Sample Mean is  4.97918141591748"

print(paste("The Theoretical Mean is ", theo_mean, sep = " "))

## [1] "The Theoretical Mean is  5"
```

It can be seen that both Sample Mean and Theoretical Mean are approximately equal.

Sample Variance Vs Theoretical Variance

```
barplot(c(theo_var, Obs_var), col = c("red", "green"), names.arg = c("Theoretical Variance", "Observed Varian
ce"))
```



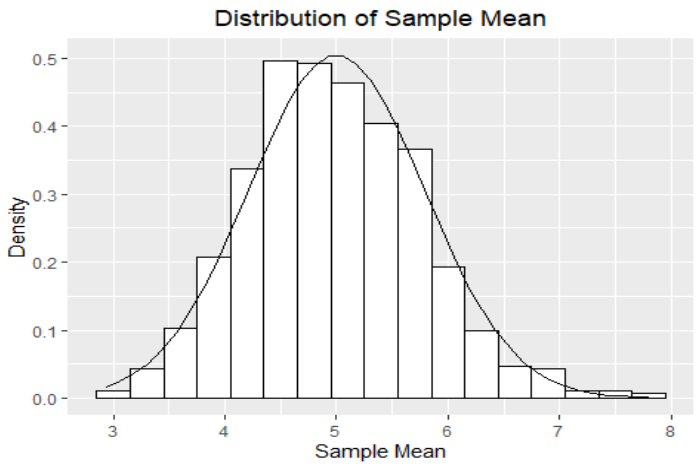
Just like with the expected values of the population and the sampling distributions, there is theoretically known link between the variances. The variance of the sample mean is equal to the population variance divided by the sample size. In mathematical notation, V(Xbar) = V(x)/n. The bigger the sample size, the smaller the uncertainty in the sample.

Distribution

In order to help you see that the sampling distribution of the sample mean is approximately normal - even when the underlying distribution is exponential (non-normal), we will overlay the histogram with a theoretical normal curve.

```
x <- seq(min(mns), max(mns), length.out = 40)
y <- dnorm(x, mean = theo_mean, sd = sqrt(theo_var))
d <- data.frame(x = x, y = y)
g <- ggplot(mns, aes(x = mns))
g <- g + geom_histogram(aes(y = ..density..), color = "black", fill = "white", binwidth = 0.3) +
  geom_line(data = d, aes(x = x, y = y)) +
  labs(x = "Sample Mean", y = "Density") + labs(title = "Distribution of Sample Mean") +
```

```
  theme(plot.title = element_text(hjust = 0.5))
g
```



The Central Limit Theorem proves that when we have a large size of the sample, if we sample a lot of samples - from whatever distribution - and construct their histogram, we will observe approximately normal distribution. The larger the sample, the closer to the normal look. Some textbook authors say that a sample size of just over 30 is enough to achieve this nice result. Here we had samples of 40. No less important was that we sampled a lot of samples - 1000. This aspect of the simulation also adds to the closeness to normality of the empirical distribution.