Exponential Distributions and the Central Limit Theorem

Peter Pih
Statistical Inference, Corsera(statinference-015)
June 17th, 2015

Overview

This exercise will look at the difference between a sample mean and it's related population mean and how the Central Limit Theorem works. The Central Limit Theorem (CLT) states that the larger a sample size, the closer the sample's mean and variance will reflect the true population mean and variance.

Simulations

We will first show an example of 40 draws from the population. Then we will show how the Central Limit Theorem comes into play with 1000 iterations of 40 draws. In both we will compare the extimated mean and observed variance.

Two sets of simulations will be covered:

- Create a sample of 40 draws, **Draw40** and compare the mean and variance
- Create 1000 iterations of 40 draws, **Draw1000**, and compare the observed means and variances

The two data sets created and used are:

- **sample40__** is the sample draw of 40
- **Draw100__** are the 1000 interations of 40 draws

true__ are the theoretical mean and variance for exp(lambda=0.2)

One of the **GIVENS** in this assignment is to use a **lambda value of 0.2**.

```
lambda <- 0.2   # this is a GIVEN in the assignment
```

This value of lambda implies that the theoretical **true mean and true variance** of the population is:

```
true_mean <- 1/lambda
true_var <- (1/lambda)^2
```

NOTE: We will use `set.seed(2)` before performing every sequence to ensure reproducibility.

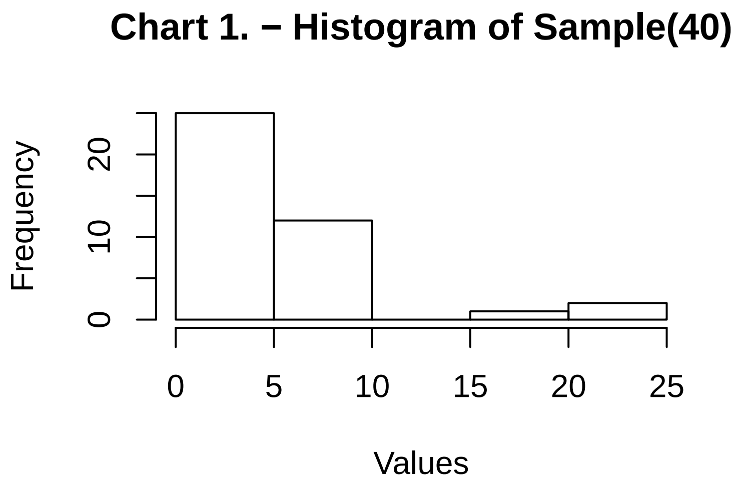
Simulation of 40 Draws

If we simulate 40 draws, what are the mean and variance of this sample?

```
set.seed(2)
sample40_data <- rexp(40, rate=lambda)
sample40_mean <- mean(sample40_data)
sample40_var <- var(sample40_data)
```

The graph of Sample40 looks like this:

```
hist(sample40_data, main="Chart 1. - Histogram of Sample(40)", xlab="Values")
```



The **true mean** and **true variance** of the population is **5** and **25**, how does this compare with the sample observations? The **sample40__mean** is 5.199 (shown in green on chart 2) and the sample variance is 27.8259. The sample variance is larger than the true variance and we will see later how this estimate decreases with a larger sample size.

To see if sample mean is meaningful, we calculate the true confidence interval from the given the true mean:

```
n <- 40
true_mean + c(-1,1) * qt(0.95, n-1) * sqrt(true_var/n)
```

```
## [1] 3.667989 6.332011
```

We can see that the observed sample mean 5.199 falls within this confidence interval, meaning that 95% of the time we will observe a sample's mean value within this interval as an estimation of the **true mean**.

Simulation of 1000 Draws of 40

This section will show how the Central Limit Theorem (CLT) works in estimating the **true mean** for many observations. The theorem states that as the size of the sample increases, the confidence interval decreases and there is less variation of the sample estimate around the theoretical mean. This is shown graphically in Chart 2.

Create a matrix by drawing 40 samples, 1000 times (1000 x 40) and taking the average of the mean and variance of each draw of 40.

```
set.seed(2)
Draw1000 <- matrix(rexp(40000, rate=lambda), 1000,40)

Draw1000_mean <- apply(Draw1000, 1, mean)
Draw1000_var <- apply(Draw1000, 1, var)
```

This shows an average Draw1000_mean of **5.0164** and average variance of **25.0024**.

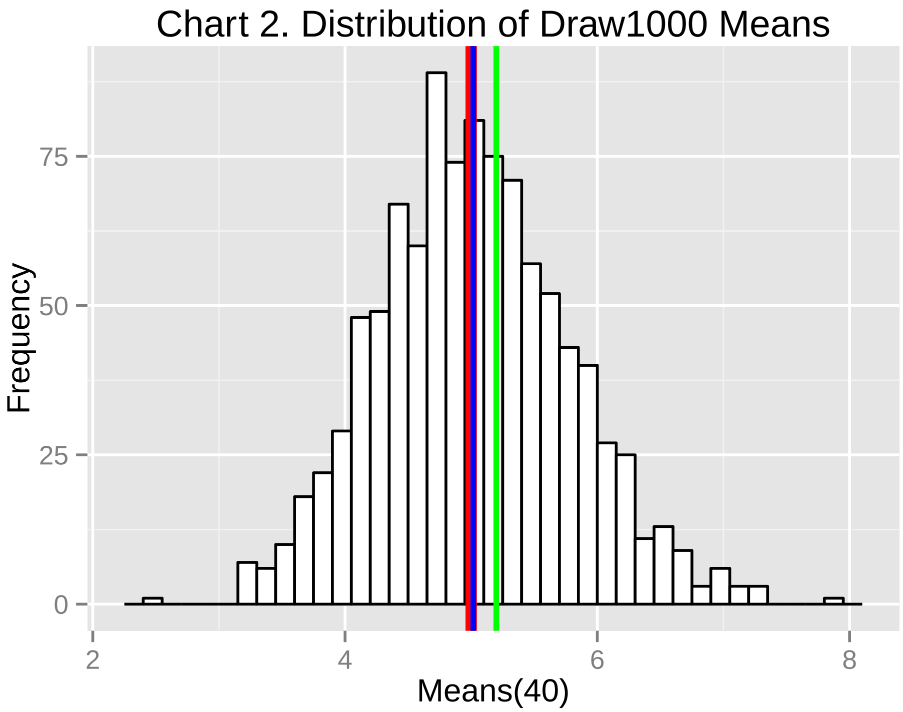
What is the confidence interval of the observed means?

```
n <- 40000
mean(Draw1000_mean) + c(-1,1) * qt(0.95, n-1) * sqrt(mean(Draw1000_var)/n)
```

```
## [1] 4.975232 5.057480
```

As predicted by the CLT, the confidence interval has shrunk by increasing the number of observations.

```
require("ggplot2", quietly=TRUE)
ggplot(as.data.frame(Draw1000_mean), breaks=40, aes(Draw1000_mean)) +
  geom_histogram(fill="white",colour="black", binwidth=0.15) +
  xlab("Means(40)") +
  ylab("Frequency") +
  ggtitle("Chart 2. Distribution of Draw1000 Means") +
  geom_vline(xintercept=true_mean, colour="red", size=2) +
  geom_vline(xintercept=mean(Draw1000_mean), colour="blue", size=1) +
  geom_vline(xintercept=mean(sample40_mean), colour="green", size=1)
```



The Central Limit Theorem states that as the sample size gets larger for mean estimates, the closer the sample mean gets the true mean (for iid processes, which this is). CLT also tells us that as we try to estimate the mean, the samples that we draw of the mean will be normally distributed around the true mean. This should not be confused with the actual distribution of the underlying population which is shown in the Appendix.

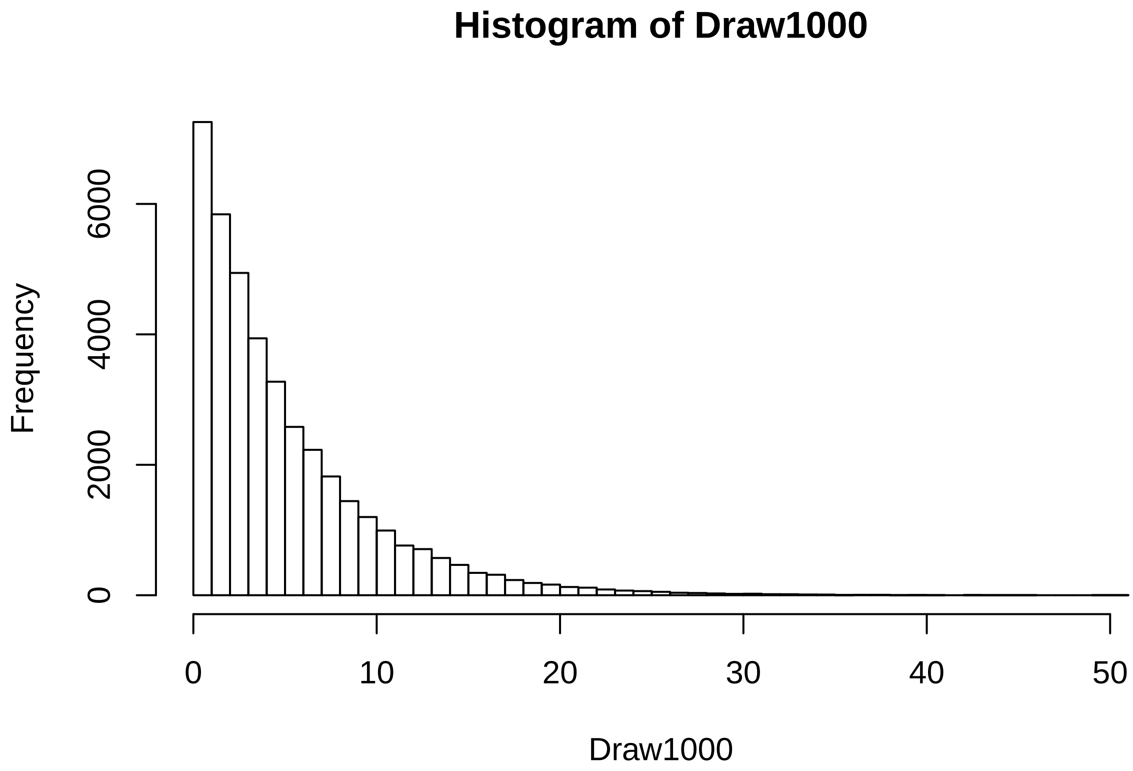
These properties are shown in the histogram. The **red** vertical line is the **true mean**, the **green** vertical line is the original mean estimate from Sample40, and the **blue** line is the mean estimate from Draw1000.

By increasing the sampling in Draw1000, the variance is reduced from 27.8259 in Sample40, to 25.0024 in Draw1000.

Appendix

Hstogram of Draw1000 Sample Data

```r
hist(Draw1000, breaks=40)
```


Histogram of Draw1000
Frequency
Draw1000