Statistical Inference Class Project

Cathy Snell

October 10, 2018

Part 1: Simulation Exercise

Overview

We will investigate the exponential distribution in R and compare it with the Central Limit Theorem (CLT). The CLT states that “the distribution of averages of iid random variables becomes that of a standard normal as the sample size increases.” (from 07 02 Asymptotics and the CLT lecture transcript). We will show that the mean of the sample means approaches the theoretical mean of the exponentials, is distributed as a normal curve, and that the variance decreases as the sample size increases.

Simulations

We have been directed to investigate the distribution of averages of 40 exponentials. Some parameters have been specified for us:

- Simulate the exponential distribution with “rexp(n, lambda)”
- Use lambda = 0.2 for all simulations
- The mean of the exponential distribution is 1/lambda
- The standard deviation of the exponential distribution is 1/lambda
- Do 1000 simulations (n=1000) of 40 exponentials

First, we generate a population of exponentials. We do this using the R code provided for exponentials, and the values provided for lambda and n.

```
pop <- rexp(1000, .2)
```

Next we generate samples of 40 exponentials, 1000 times, and take the mean of each. This is our simulation of multiple samples of the larger population.

```
sample.means <- NULL
for(i in 1:1000) sample.means = c(sample.means, mean(rexp(40, .2)))
```

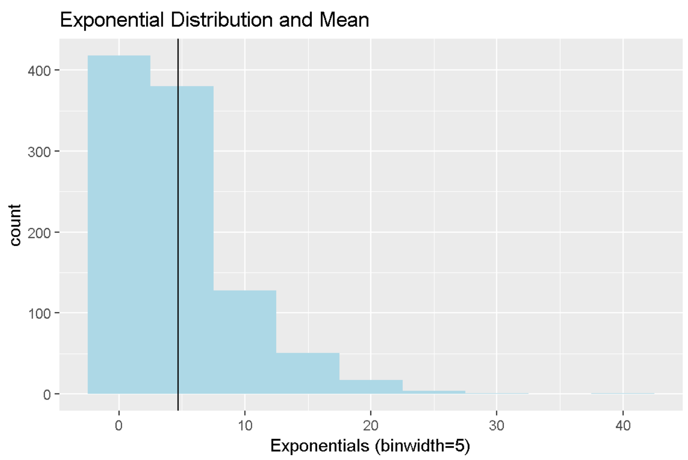
Sample Mean versus Theoretical Mean

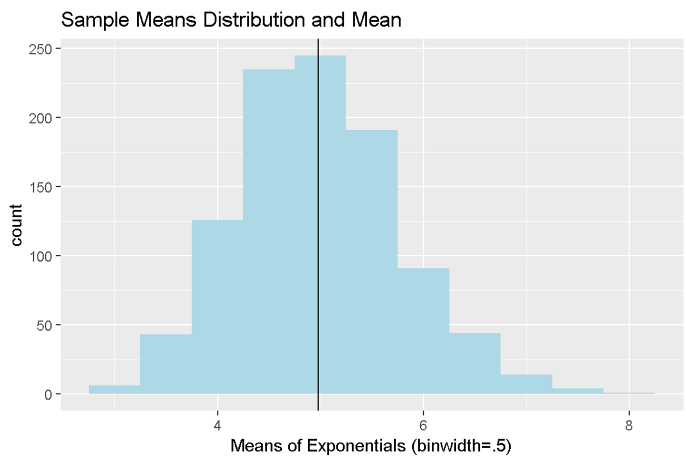
Per the CLT, the distribution of averages/means of the samples will approach normal, with a center/mean of the population mean.

We were given the theoretical mean of 1/lamba, which works out to 1/.2 or 5. Let’s see how this compares to the mean of the populations we simulated, and the mean of the sample means.

```
##  Population (1000) mean =  4.68896230358042
##  Means of 1000 Samples of 40 =  4.97912941889599
##  Theoretical mean =  5
```

Let’s visualize the distributions with their means.





We can see that both the means are very close to the theoretical mean of 5.

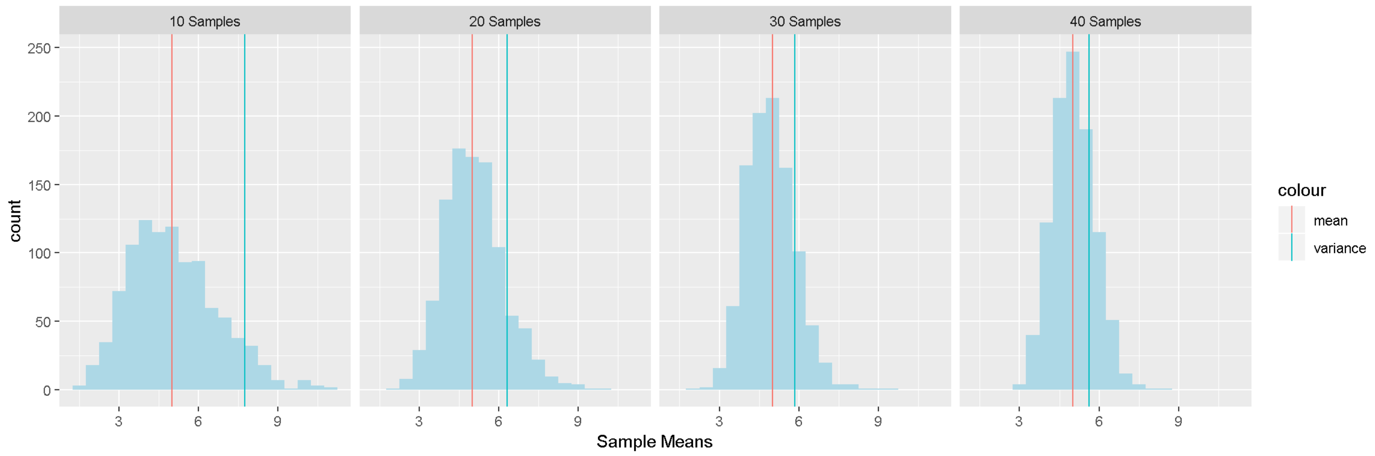
Sample Variance versus Theoretical Variance

We know that the standard deviation = sqrt(variance) and the theoretical sd = 1/lambda = 5. Thus, the theoretical variance is 25.

The CLT states that the variance of the sample means will approach the theoretical variance / n. With our sample size of 40, this gives 25/40 = .625.

If we consider 4 different sets of sample means which vary by the number of samples in the mean (n), we see the variance (which is the sd squared), decline from the theoretical variance of 25.

```
##  10 Samples Variance:  2.74426940254696
##  20 Samples Variance:  1.32465556728193
##  30 Samples Variance:  0.843721507094723
##  40 Samples Variance:  0.631044600086259
```

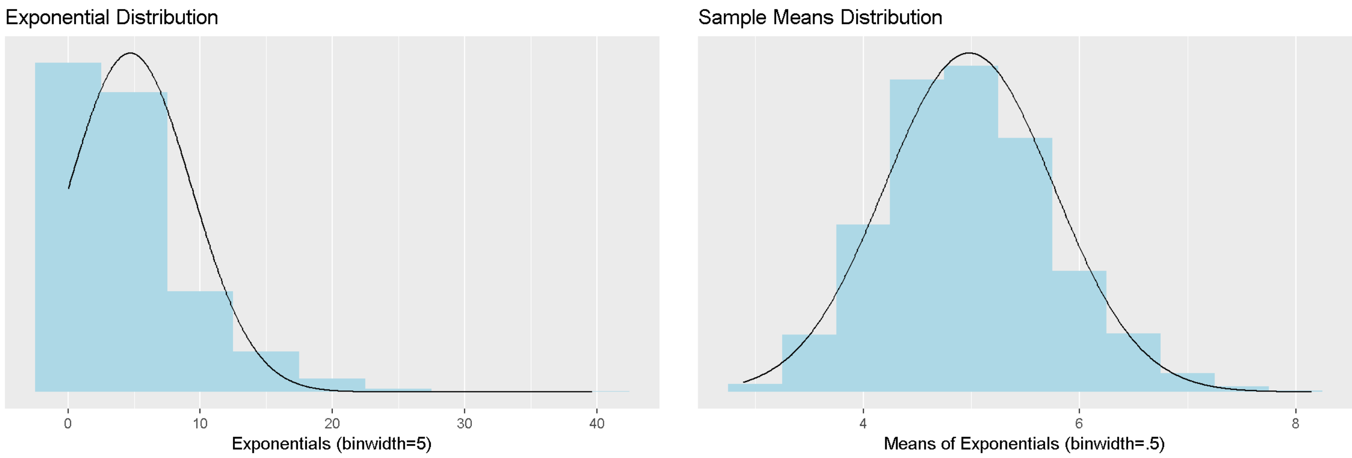


We have printed out the variance per sample size and visualized in a plot. With increasing sample size, we see the spread (variance) of the normal distribution decrease as the distribution clusters more tightly around the mean.

The acutal variance of our 40 samples is 0.631 which is close to the theoretical variance of 25/40 or .625.

Distribution

We will compare the simulated populations of exponentions and the sample means to a normal curve by plotting a normal on top of each. There are other methods that could be used, such as the Shaprio-Wilk test, which are out of the scope of this class.



For the exponentials, there is no left tail and we see a distinct left skew. For the sample means, we see a clearly defined hump in the middle with tails on either side, indicative of a normal curve.

Appendix

Code

```
library(tidyverse)
library(ggplot2)
library(gridExtra)

# Part 1: Simulation Exercise

## Overview

## Simulations

pop <- rexp(1000, .2)
```

```
lg.pop <- rexp(2000, .2)

     # combine into a data frame for plotting in ggplot2
df.pop <- rbind(as.data.frame(cbind(Exponential=pop, Group="1,000 Sample")),
                   as.data.frame(cbind(Exponential=lg.pop, Group="2,000 Sample")))

ggplot(df.pop, aes(as.numeric(as.character(Exponential)))) +
  geom_histogram(binwidth = 5) +
  facet_grid(.~Group) +
  xlab("Exponentials (binwidth=5)")

sample.means <- NULL
for(i in 1:1000) sample.means = c(sample.means, mean(rexp(40, .2)))

### Sample Mean versus Theoretical Mean

cat(paste(" Population (1000) mean = ", mean(pop),"\n",
          "Means of 1000 Samples of 40 = ", mean(sample.means),"\n",
          "Theoretical mean = ", 1/.2))

p1 <- ggplot(as.data.frame(pop), aes(pop)) +
  geom_histogram(binwidth = 5, fill="lightblue") +
  xlab("Exponentials (binwidth=5)") +
  ggtitle("Exponential Distribution and Mean") +
  geom_vline(xintercept = mean(pop))

p2 <- ggplot(as.data.frame(sample.means), aes(sample.means)) +
  geom_histogram(binwidth = .5, fill="lightblue") +
  xlab("Means of Exponentials (binwidth=.5)") +
  ggtitle("Sample Means Distribution and Mean") +
  geom_vline(xintercept = mean(sample.means))

grid.arrange(p1, p2, ncol=2)

### Sample Variance versus Theoretical Variance

sample.means2 <- NULL
c(for(i in 1:1000) sample.means2 = c(sample.means2, mean(rexp(10, .2))),
  for(i in 1:1000) sample.means2 = c(sample.means2, mean(rexp(20, .2))),
  for(i in 1:1000) sample.means2 = c(sample.means2, mean(rexp(30, .2))),
  for(i in 1:1000) sample.means2 = c(sample.means2, mean(rexp(40, .2))))

   # combine into a data frame for plotting in ggplot2
df.sample.means <- rbind(as.data.frame(cbind(Sample=sample.means2[1:1000], Group="10 Samples")),
                         as.data.frame(cbind(Sample=sample.means2[1001:2000], Group="20 Samples")),
                         as.data.frame(cbind(Sample=sample.means2[2001:3000], Group="30 Samples")),
                         as.data.frame(cbind(Sample=sample.means2[3001:4000], Group="40 Samples")))

df.intercepts <- as.data.frame(cbind(Variance= c(var(sample.means2[1:1000]),
                                                var(sample.means2[1001:2000]),
                                                var(sample.means2[2001:3000]),
                                                var(sample.means2[3001:4000])),
                               Group=c("10 Samples", "20 Samples", "30 Samples", "40 Samples")))

cat(paste(" 10 Samples Variance: ",
      df.intercepts[1,1], "\n",
      "20 Samples Variance: ",
      df.intercepts[2,1], "\n",
      "30 Samples Variance: ",
      df.intercepts[3,1], "\n",
      "40 Samples Variance: ",
      df.intercepts[4,1]))

ggplot(df.sample.means, aes(x=as.numeric(as.character(Sample)))) +
  geom_histogram(binwidth = .5, fill="lightblue") +
  geom_vline(data = df.intercepts, aes(xintercept = 5+as.numeric(as.character(Variance)), colour="variance")
) +
  geom_vline(aes(xintercept = 5, colour="mean")) +
  facet_grid(.~Group) +
  xlab("Sample Means")

### Distribution
```

```r
p1 <- ggplot(as.data.frame(pop), aes(pop)) +
  geom_histogram(binwidth = 5, fill="lightblue", aes(y=..density..)) +
  xlab("Exponentials (binwidth=5)") +
  ggtitle("Exponential Distribution") +
  stat_function(fun = dnorm, n = 1000, args = list(mean = mean(pop), sd = sd(pop))) +
  ylab("") +
  scale_y_continuous(breaks = NULL)

p2 <- ggplot(as.data.frame(sample.means), aes(sample.means)) +
  geom_histogram(binwidth = .5, fill="lightblue", aes(y=..density..)) +
  xlab("Means of Exponentials (binwidth=.5)") +
  ggtitle("Sample Means Distribution") +
  stat_function(fun = dnorm, n = 1000, args = list(mean = mean(sample.means), sd = sd(sample.means))) +
  ylab("") +
  scale_y_continuous(breaks = NULL)

grid.arrange(p1, p2, ncol=2)
```