Class Project in Two Parts: Central Limit Theorem & Inferential Data Analysis

Paul Clark

February 5, 2017

Contents

---

1 Central Limit Theorem (Simulation)

1.1 Overview

We investigate the *exponential* distribution in R and compare it with the Central Limit Theorem. The distribution is simulated via `rexp(n, lambda)`, and both mean and standard deviation are `1/lambda`. We set `lambda = 0.2` and investigate the distribution of 1000 samples of 40 exponentials each.

Review criteria

1. Did you show where the distribution of means is centered and compare it to the theoretical center of the population distribution?
2. Did you show how variable it is and compare it to the theoretical variance of the distribution?

We first load packages needed for the analysis.

```
library(ggplot2)
library(tidyr)
```

1.2 Simulations

We will show the properties of the distribution of the mean and variance of samples of 40 exponentials.

First, we simulate the data:

```
# We simulate S = 1000 samples of N = 40 exponentials each, all with lambda = 0.2
S <- 1000
N <- 40
lambda <- 0.2
set.seed(1234)
```

```r
sim_data <- matrix(rexp(n = S*N, rate = lambda), S, N)
means <- apply(sim_data, 1, mean)
variances <-  apply(sim_data, 1, function(x) {sd(x)^2}) # sample var, N-1 in denominator
sim <- data.frame(population = rexp(1000, rate = lambda), means, variances)
str(sim)
```

```
## 'data.frame':    1000 obs. of  3 variables:
##  $ population: num  4.166 2.465 0.782 2.598 3.058 ...
##  $ means     : num  4.6 6.02 5.46 4.18 7.14 ...
##  $ variances : num  11.8 36.4 35.9 16.8 56.7 ...
```

1.2.1 Sample Mean vs. Theoretical Mean

Next, we show our simulated distribution of sample means and compare it to the theoretical mean of the population.

```r
# We plot a histogram of sample means
g <- ggplot(sim, aes(x = means)) + geom_histogram()
g <- g + geom_vline(xintercept = mean(means), size = 1) +
     labs(title = paste0("The simulated mean lies just below the theoretical mean",
                         " of ", 1/lambda)) +
     theme(plot.title = element_text(face="italic", hjust=0.5))
print(g)
```

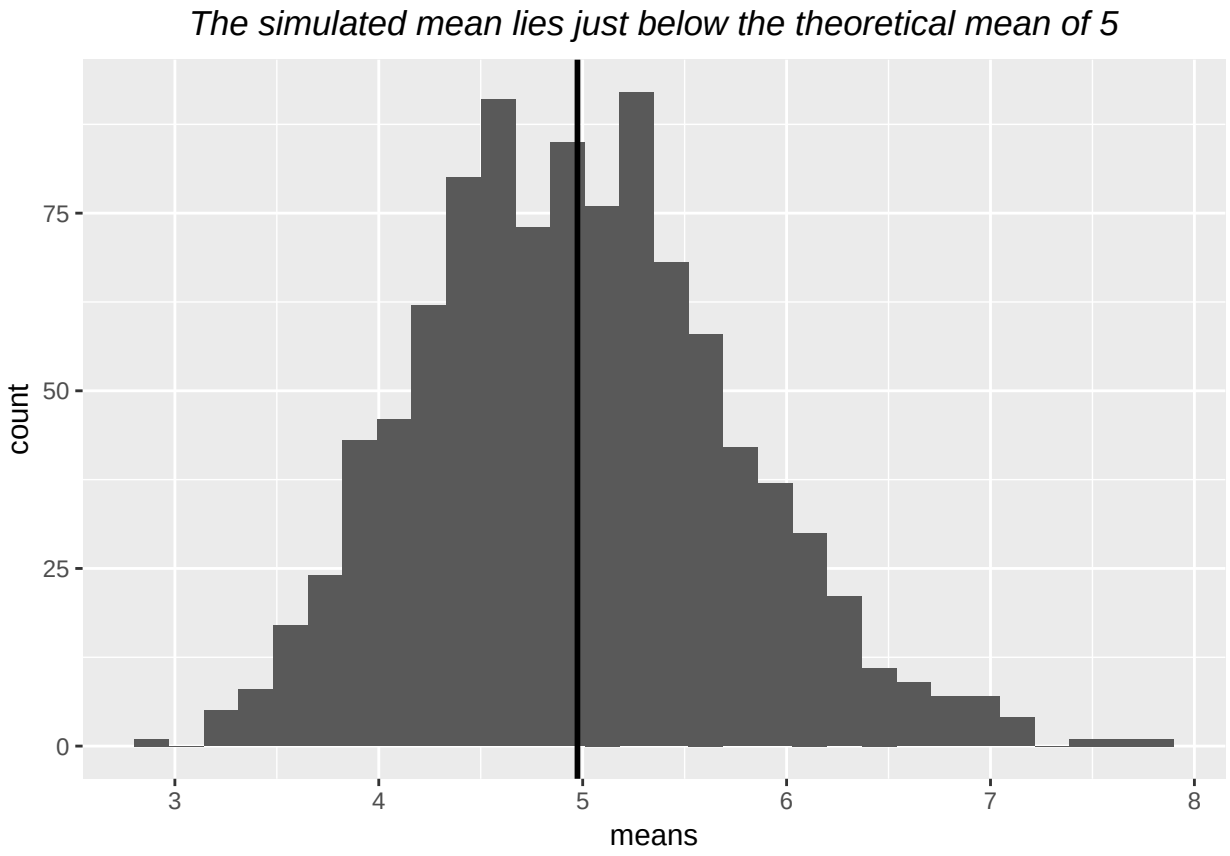

Figure 1: Histogram of means of 1000 samples of 40 exponentials with theoretical mean 5

The vertical line shows the position of the empirical mean. It almost exactly equals the theoretical value $1/\lambda$, or 5, for our lambda of 0.2. To be precise, the difference $\mu_{simulated} - \mu_{theory}$ is -0.03, or -0.52% from theoretical.

1.2.2 Sample Variance vs. Theoretical Variance

We show the sample variance and compare it to the theoretical variance of the population.

```
# Following slide 7 of lecture 5 of the statistical inference course, we plot a
# histogram of sample variances and show its mean
g <- ggplot(sim, aes(x = variances)) + geom_histogram()
g <- g + geom_vline(xintercept = mean(variances), size = 1) +
     labs(title=paste0("The mean of simulated variances is just below the ",
                       "theoretical value of ", 1/lambda^2)) +
     theme(plot.title = element_text(face="italic", hjust=0.5)) +
     scale_x_continuous(breaks = seq(from = 0, to = 90, by = 10))
print(g)
```

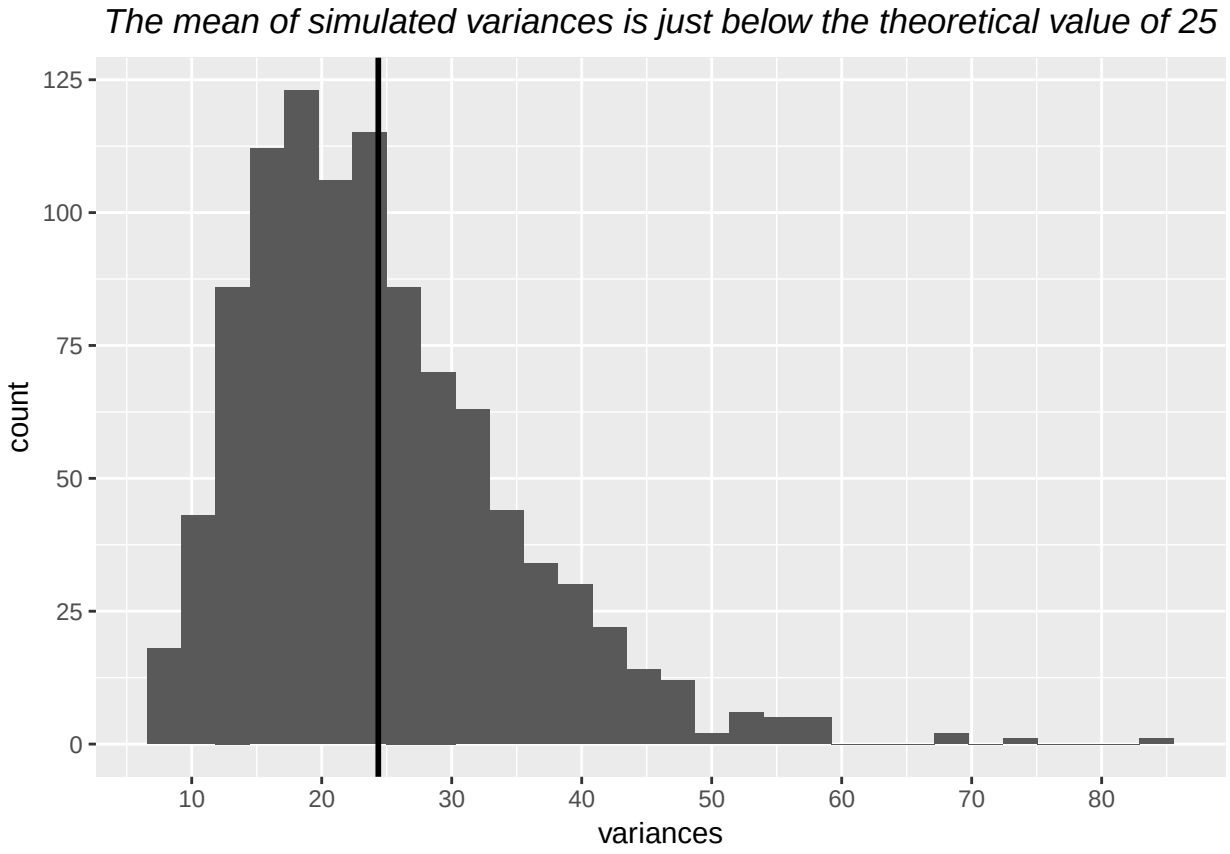

Figure 2: Histogram of estimated population variances of 1000 samples of 40 exponentials

The vertical line shows the position of the empirical mean of the distribution of estimated population variances. The mean is again close to the theoretical value of $1/\lambda^2$, or 25. To be precise, the difference $\sigma^2_{simulated} - \sigma^2_{theory}$ is -0.65, or -2.59% from theoretical.

We can also estimate population variance using the variance of the simulated means. By the properties of variances, $\sigma^2_{pop} = n\sigma^2_{mean}$. This estimate gives 23.80, which differs by -1.20 from the theoretical, or -4.80%.

1.2.3 Normality

We now show that the distribution of means is trending away from exponential toward normal. Our assessment will be primarily graphical.

```
# First, we re-format the data for use in ggplot
long_sim <- gather(sim, key = Distribution, value = Measure, population:variances)
# We construct a two-facet plot contrasting the randomly sampled distribution
# of means with a randomly sampled distribution of the population
long_sim_sub <- with(long_sim, long_sim[!Distribution %in% "variances",])
g <- ggplot(data = long_sim_sub, aes(x = Measure)) + geom_histogram() +
     facet_grid(Distribution~.) + labs(title =
     "The means trend toward normal, while the population values are far from it") +
     theme(plot.title = element_text(face="italic", hjust=0.5))
print(g)
```

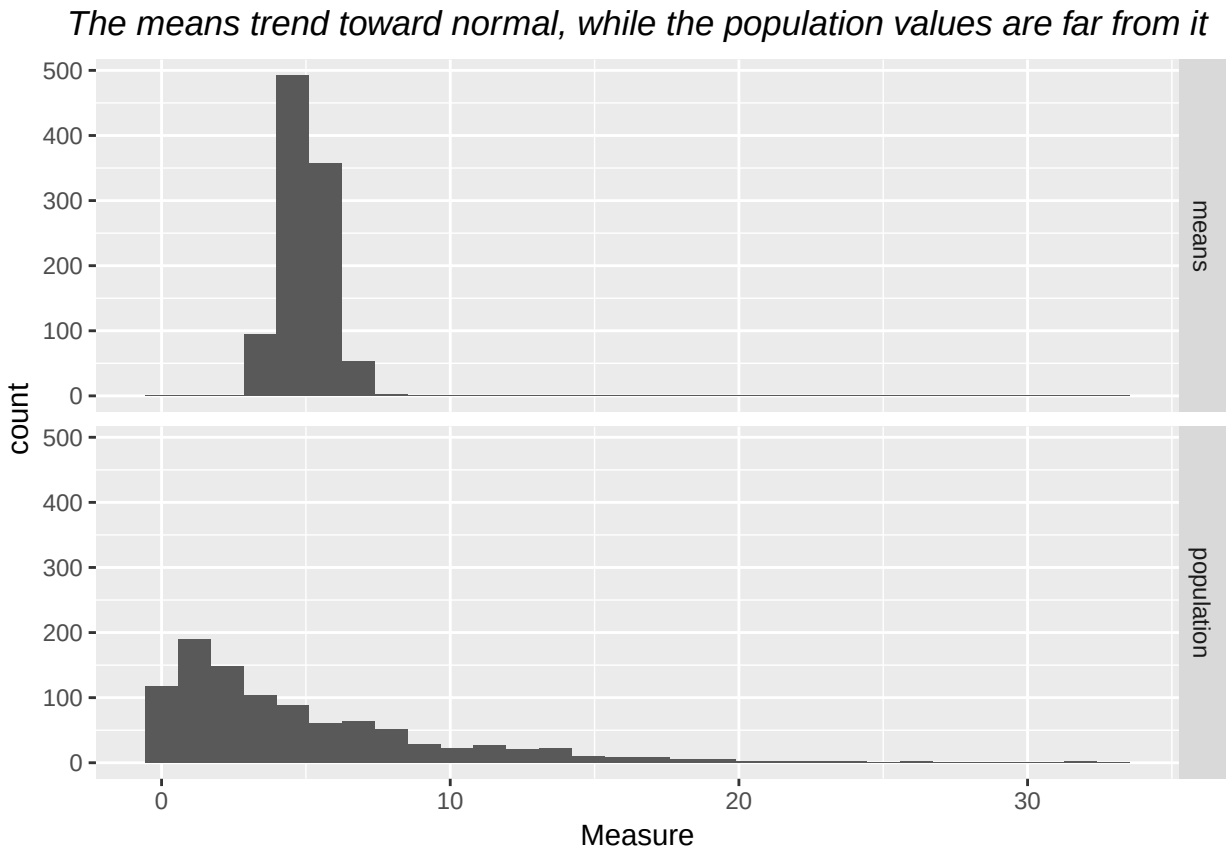

Figure 3: Histograms of 1000 Population Values and 1000 Means of random exponentials

By visual inspection, the population distibution is severely right-skewed and not normal. Therefore, we focus our evaluation of normality on the distribution of means.

1.2.3.1 Mean vs. Median

The Mean and Median of a normal distribution should be equal. This is approximately true for the distribution of means.

```
g <- ggplot(sim, aes(x = means)) + geom_histogram()
g <- g + geom_vline(xintercept = mean(means), size = 1, color = "black") +
         geom_vline(xintercept = median(means), size = 1, color = "red") +
     labs(title = "The mean and median are approximately equal") +
     theme(plot.title = element_text(face="italic", hjust=0.5))
print(g)
```

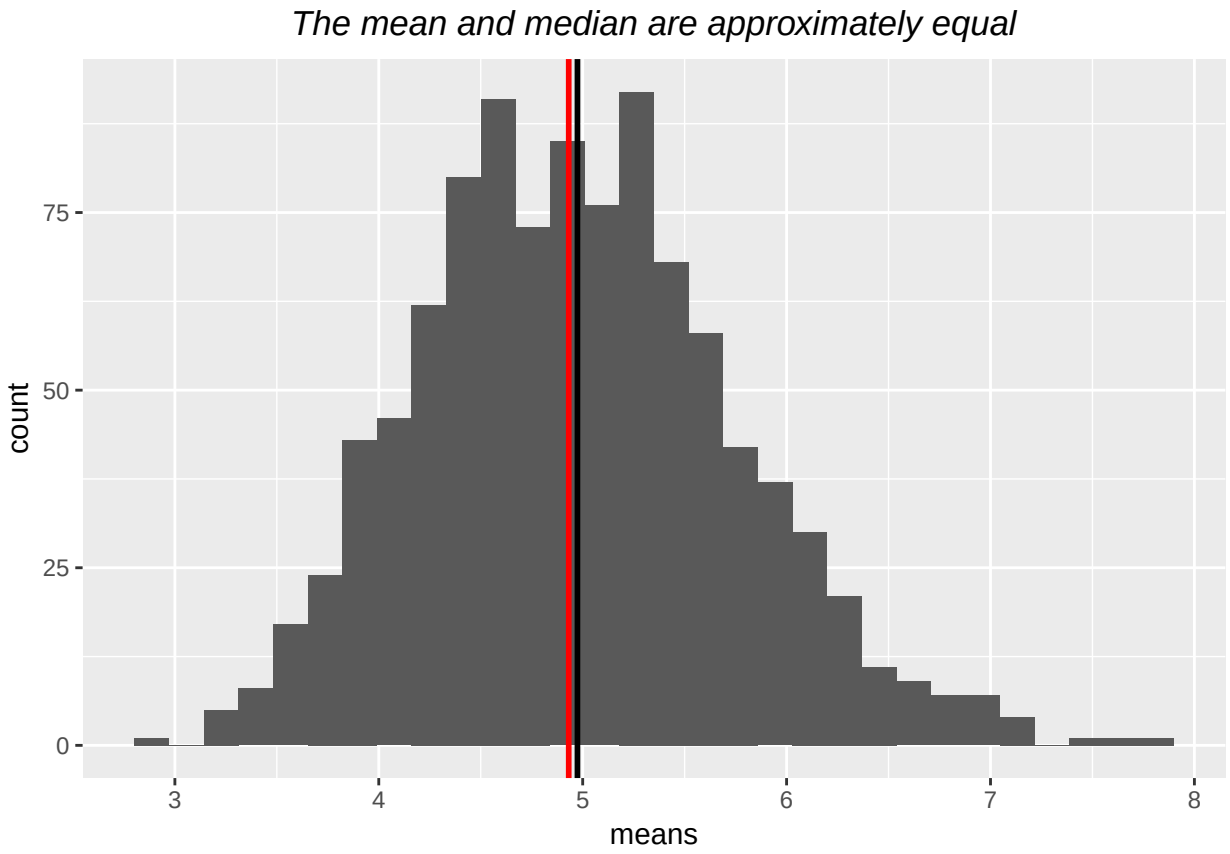

Figure 4: Histogram of means, with lines at mean and median

Above, as in **Figure 1**, the black line represents the mean, whereas the red line represents the median. The equality of the two is a characteristic of a normal distribution.

1.2.3.2 Quantiles of distribution vs. quantiles of standard normal

Here we compare the quantiles of the distribution of the mean with those of standard normal.

```
qqnorm(sim$means, main = "Normal sample with some left compression and right skew")
qqline(sim$means)
```

We see from the Q-Q plot above that as you move down from center (5) in the data distribution, data quantiles/values below 4 are traversed visibly slower than linearly with standard normal quantiles, indicating compression of the data's distributional mass on the left. On the other hand, as we move up from center, the data traverses its quantiles/values faster than (more than linear with) standard normal quantiles, indicating probablity mass that is spread out to the right. In summary, the distribution exhibits some left compression and rightward skew.

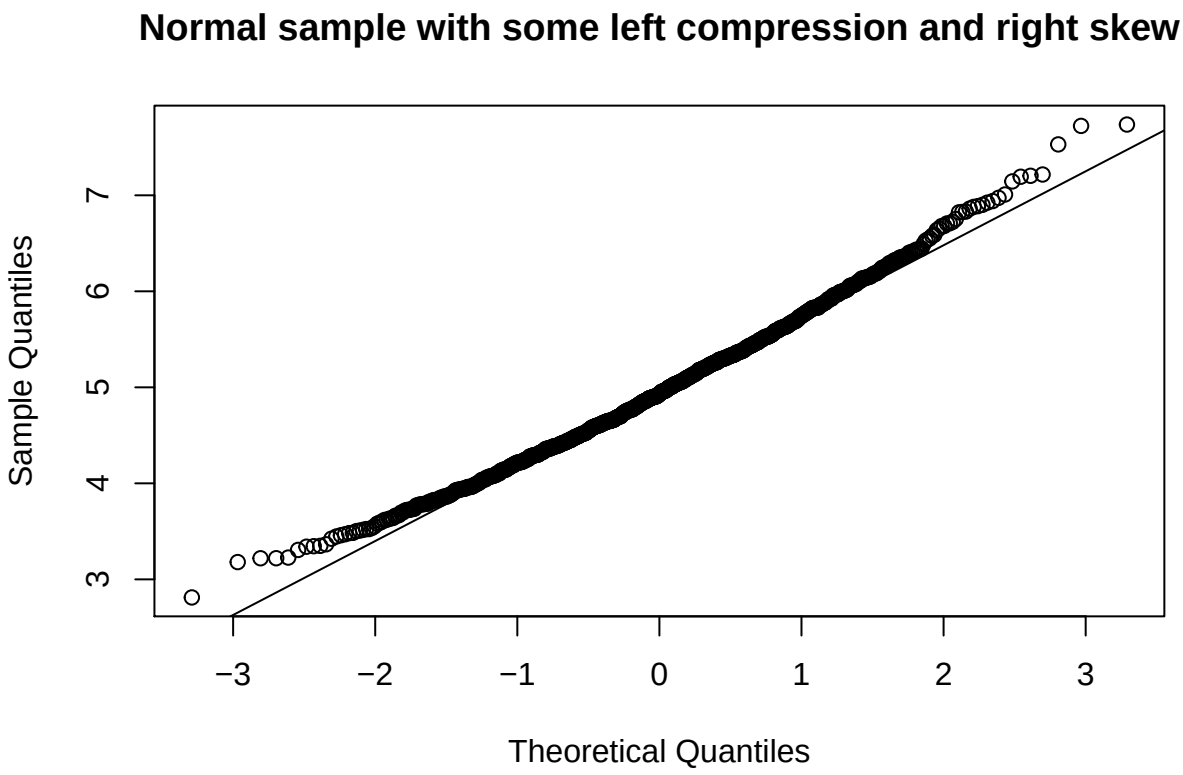

Figure 5: Plot of data against corresponding quantiles of standard normal

However, the linear part of the Q-Q plot between -1 and +1 normal quantiles attests to the fact that the data is relatively normally distributed. To show this, we compute the number of sample standard deviations between -1 and +1 standard normal quantiles from the center of the distribution, which should be 2 for a normal distribution.

```
data_quantiles_stdnorm <- quantile(sim$means, probs = c(pnorm(q=-1), pnorm(q=1)))
(med_plus_minus <- (data_quantiles_stdnorm - median(sim$means))/sqrt(var(sim$means)))
```

```
##  15.86553%  84.13447%
## -0.9296552  1.0526858
```

To be precise, the number of standard deviations within 1 standard normal quantile of center are just -0.88% from 2. Correspondingly, approximately 68% of the probability mass of the distribution lies within 1 standard deviation of the midpoint, an attribute of a normal distribution.

2 Basic Inferential Data Analysis

Review Criteria

3. Did you perform an exploratory data analysis of at least a single plot or table highlighting basic features of the data?
4. Did the student perform some relevant confidence intervals and/or tests?
5. Were the results of the tests and/or intervals interpreted in the context of the problem correctly?
6. Did the student describe the assumptions needed for their conclusions?

Now in the second portion of the project, we're going to analyze the ToothGrowth data in the R datasets package.

Load the ToothGrowth data and perform some basic exploratory data analyses

Provide a basic summary of the data.

Use confidence intervals and/or hypothesis tests to compare tooth growth by supp and dose. (Only use the techniques from class, even if there's other approaches worth considering)

State your conclusions and the assumptions needed for your conclusions.