Inferential Statistics Course Project: Simulation

Wally Thornton

August 23, 2015

Project 1: A Simulation Exercise

Using the simulation of an exponential distribution, we will explore the relationships between the sample and the theoretical population (including mean and variance), and demonstrate that the distribution of sample means adheres to a Gaussian (normal) distribution, even though the original distribution is exponential.

Run the simulations

The first step is to set up the simulation environment and load packages that might be needed (code loading packages not shown, to save space).

We then create an exponential distribution simulation with a sample size of 40 (n=40) and a rate of 0.2 (λ=0.2). We'll use the `rexp()` function in R, which randomly pulls values from an exponential distribution with a mean of $1/\lambda$, or 5, and then repeat this 1,000 times (r=1000) to get a nice, big matrix.

```
set.seed(42)
n <- 40
lambda <- 0.2
r <- 1000
my_samples <- matrix(rexp(n*r, lambda), r)
```

Calculate and compare the sample mean to the theoretical mean

Now that we have our simulation results, we'll calculate the mean for each 40-sample run and capture all 1,000.

```
sample_means <- apply(my_samples, 1, mean)
```

How do they compare to our theoretical mean of 5? While the mean of each sample ranges from 3.14 to 7.88, the mean of the sample means is 4.99, **very close to our theoretical mean of 5**.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.141   4.406   4.919   4.987   5.504   7.882
```

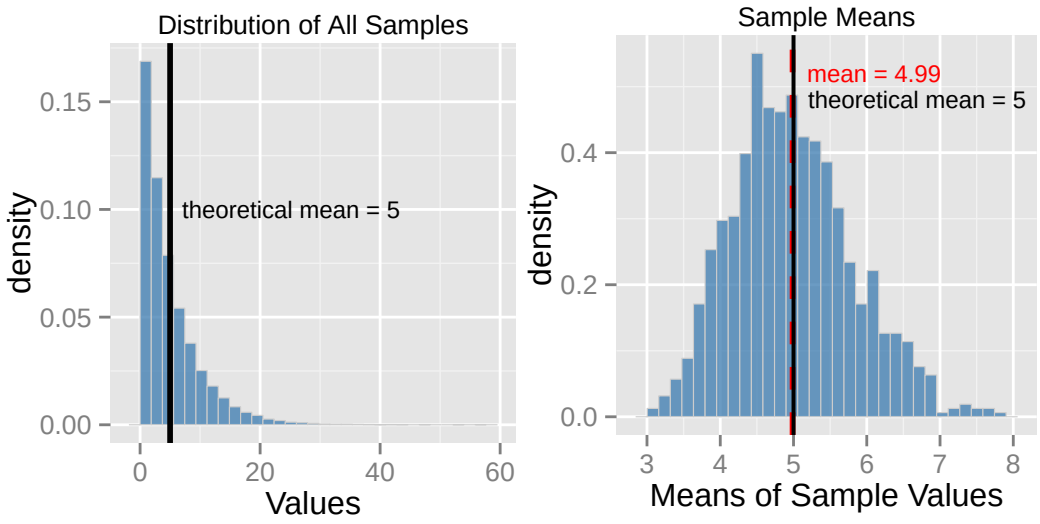
This is visually evident when we plot the distributions:

```
long_sample <- gather(data.frame(c(my_samples)), "sample", "x")
long_means <- gather(data.frame(sample_means), "sample", "x")
long_combined <- bind_rows(long_sample, long_means)

a <- ggplot(long_sample, aes(x=x, y=..density..)) +
    geom_histogram(fill="steelblue", color="grey80", size=.2, alpha=.8) +
```

```
    geom_vline(xintercept=1/lambda, linetype="solid", color="black", size = 1) +
    annotate("text", size=3, hjust=0, x = 1/lambda+2, y=.1, color="black"
             , label=paste("theoretical mean =",1/lambda)) +
    xlab("Values") +
    ggtitle("Distribution of All Samples") +
    theme(plot.title=element_text(size=10))

b <- ggplot(long_means, aes(x=x, y=..density..)) +
    geom_histogram(fill="steelblue", color="grey80", size=.2, alpha=.8) +
    geom_vline(xintercept=mean(sample_means), linetype="dashed", color="red", size=1) +
    geom_vline(xintercept=1/lambda, linetype="solid", color="black", size=.7) +
    annotate("text", size=3, hjust=0, x = mean(sample_means)+.2, y=.52, color="red"
             , label=paste("mean =",round(mean(sample_means),2))) +
    annotate("text", size=3, hjust=0, x = 1/lambda+.2, y=.48, color="black"
             , label=paste("theoretical mean =",round(1/lambda,2))) +
    xlab("Means of Sample Values") +
    ggtitle("Sample Means") +
    theme(plot.title=element_text(size=10))
```



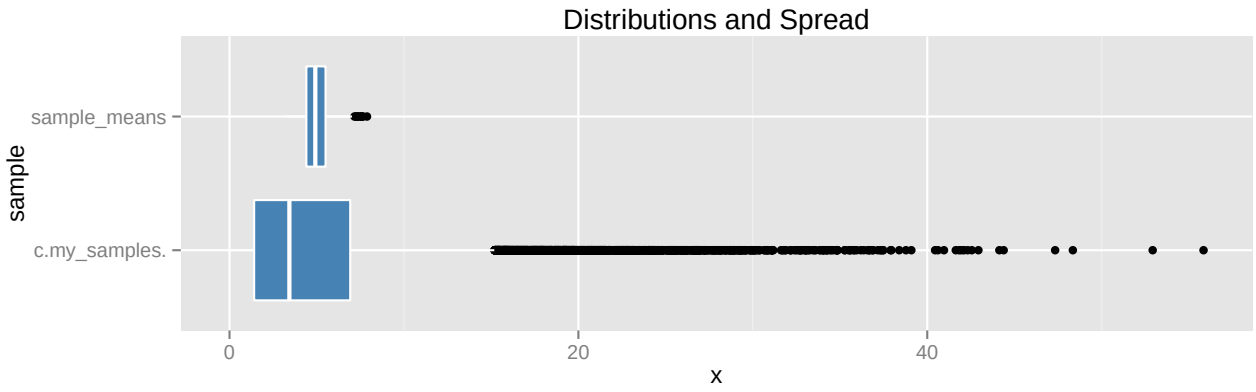
Calculate and compare the sample variance to the theoretical variance

The standard deviation of an exponential distribution is $1/\lambda$, so the variance is $(1/\lambda)^2$, which gives us the variance of the distribution of 25.

The variance of the sampling distribution of the mean (also known as the standard error of the mean) is defined as $\sigma_\mu^2 = \sigma^2/N$, that is, the population variance divided by the sample size. Plugging in our values results in a standard error of the mean of:

```
## [1] "Standard error of the mean:  0.625"
```

This is significantly less than the theoretical population variance. Why? This is because we are estimating how far each sample mean is likely to be from the population mean and as sample sizes get larger, the standard error will trend toward zero because the estimate improves. This difference in variance shows up in boxplots of the two distributions, with the sample means much more clustered near the median while the population is much more spread out.

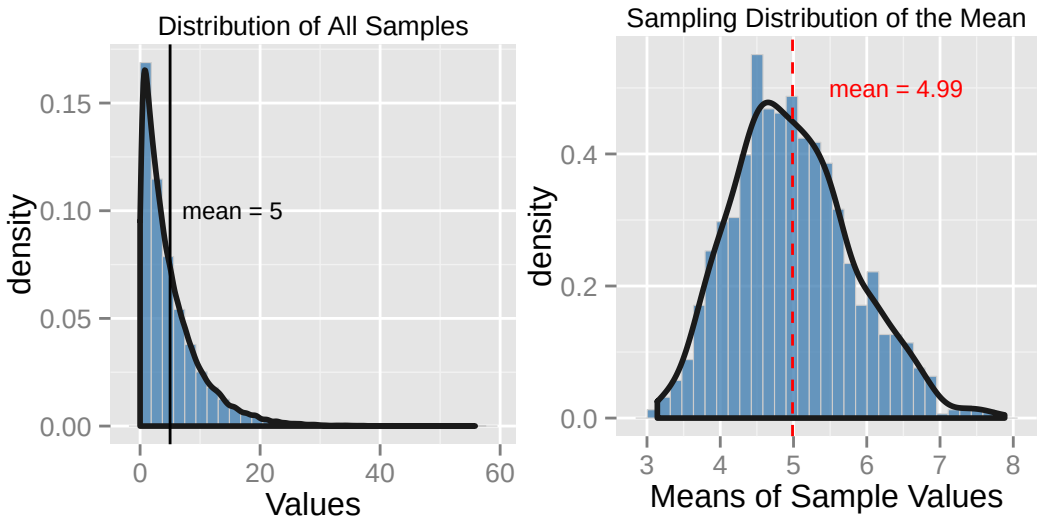


Analyze the distribution of the sample mean

Looking at the histogram of the 40,000 randomly generated exponents, it's obvious that the data are not normally distributed and actually follow an exponential distribution. On the other hand, if we plot the means of each 40-value observation, we find that they are not distributed exponentially, but rather follow a normal distribution. Here are the same plots as above, now with their respective density curves overlaid:

```
a <- ggplot(long_sample, aes(x=x, y=..density..)) +
    geom_histogram(fill="steelblue", color="grey80", size=.2, alpha=.8) +
    geom_density(color="grey10", size=1) +
    geom_vline(xintercept=1/lambda, linetype="solid", color="black") +
    annotate("text", size=3, hjust=0, x = 1/lambda+2, y=.1, color="black"
             , label=paste("mean =",1/lambda)) +
    xlab("Values") +
    ggtitle("Distribution of All Samples") +
    theme(plot.title=element_text(size=10))

b <- ggplot(long_means, aes(x=x, y=..density..)) +
    geom_histogram(fill="steelblue", color="grey80", size=.2, alpha=.8) +
    geom_density(color="grey10", size=1) +
    geom_vline(xintercept=mean(sample_means), linetype="dashed", color="red") +
    annotate("text", size=3, hjust=0, x = mean(sample_means)+.5, y=.5, color="red"
             , label=paste("mean =",round(mean(sample_means),2))) +
    xlab("Means of Sample Values") +
    ggtitle("Sampling Distribution of the Mean") +
    theme(plot.title=element_text(size=10))
```



Visually, the distribution of the sample means is much closer to normal than the distribution of the sample itself. If the distribution is normal, we expect the mean of the distribution to be equal to the median and about 95% of the results within 1.96 standard deviations, which we find to be approximately the case:

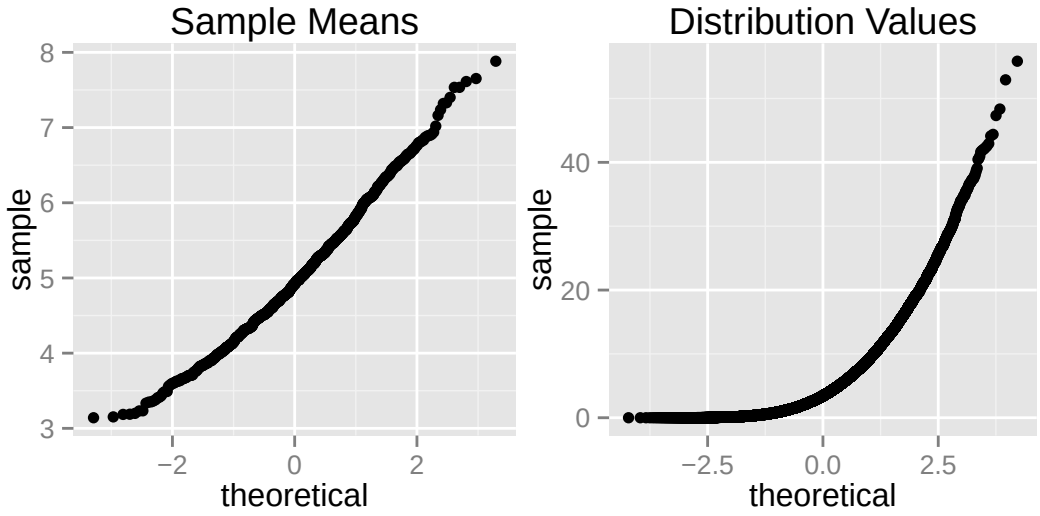
```
lower <- mean(sample_means) - 1.96*sd(sample_means)
upper <- mean(sample_means) + 1.96*sd(sample_means)
set_check <- sum(sample_means > lower & sample_means < upper)/length(sample_means)
```

```
## [1] "Mean:  4.99"
```

```
## [1] "Median:  4.92"
```

```
## [1] "Percentage of sampling means within 1.96 standard deviations:  95.4%"
```

To confirm these checks with something more concrete, we can run a quantile-quantile (Q-Q) plot. The Q-Q plot compares each point in the dataset to where they would be in a perfectly normal distribution with the same mean and standard deviation:



In a Q-Q plot, we look for general linearity as indicative of normality, which we roughly have with distribution of the sample means but very obviously do **not** have with the distribution of the values. We could check

specifically with a Shapiro-Wilk normality test, but with large datasets (like ours) this test can lead to conclusions that the data is not normal even though it is, in fact, quite normal.

Given the linearity of the Q-Q plot and what we see with the standard deviations, we can feel comfortable that the sampling distribution of the means is indeed normally distributed, consistent with the Central Limit Theorem.