Inferential Statistics Course Project

Wally Thornton

August 22, 2015

This course project consists of two sections, one a simulation exercise and the second an inferential data analysis.

Section 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.

```
knitr::opts_chunk$set(fig.width=9)
options(scipen=999)
setwd("~/Documents/DataScience/Inferential Statistics")
# ensurePkg tests whether the packages that run_analysis uses are installed and, if not, installs them.
ensurePkg <- function(x) {
    if (!require(x,character.only = TRUE)) {
        install.packages(x,dep=TRUE, repos="http://cran.r-project.org")
        if(!require(x,character.only = TRUE)) stop("Package not found")
    }
}
ensurePkg('scales')
ensurePkg('tidyr')
ensurePkg('ggplot2')
```

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 pull values from an exponential distribution with a mean of $1/\lambda$, or 5, and 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
```

Calculate and compare the sample variance to the theoretical variance

So we've demonstrated that the mean of the sample is very close to that of the distribution, but how does the variance compare? The standard deviation of an exponential distribution is $1/\lambda$, so the variance will be $1/\lambda^2$. Therefore, the theoretical variance of the distribution will be 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:

```
se <- 1/lambda^2/n
paste("Standard error of the mean: ", se)
```

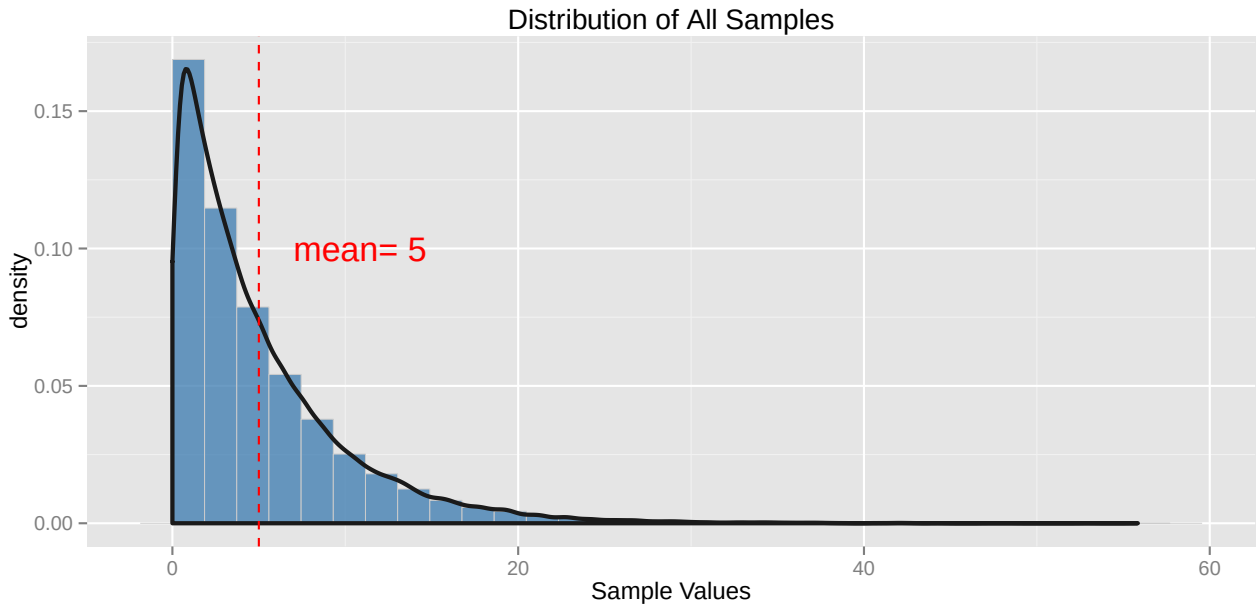
```
## [1] "Standard error of the mean:  0.625"
```

This is significantly less than the theoretical population variance. Why? This is because we are estimating how far the 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.

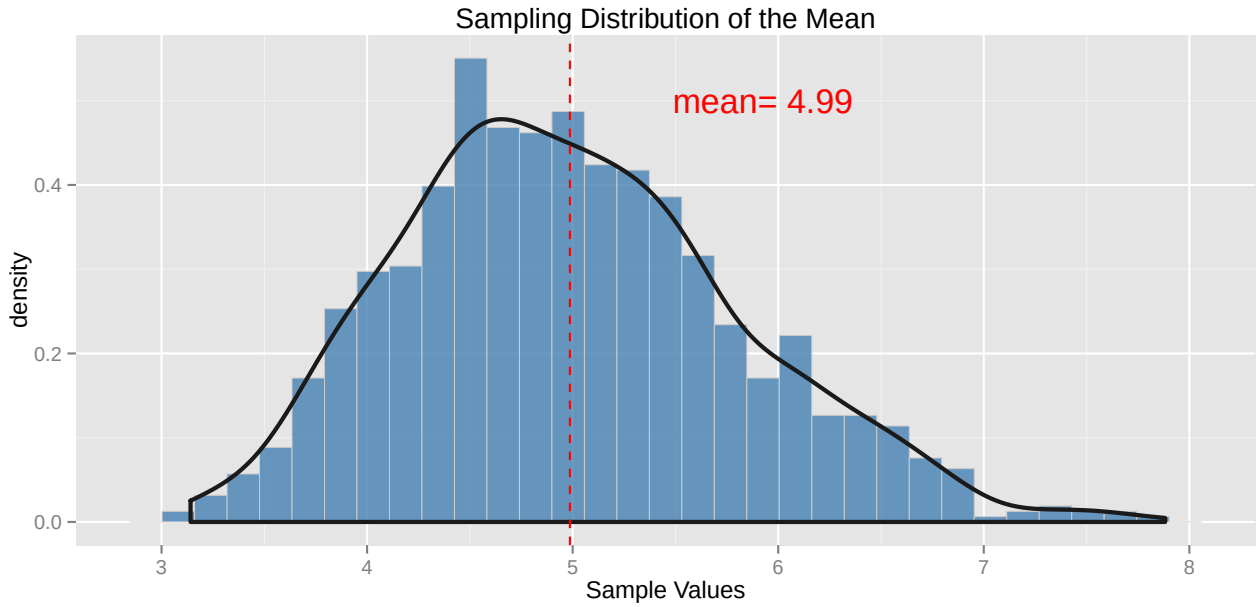
Analyze the distribution of the sample mean

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. 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 they are, with their respective density curves overlaid:

```
ggplot(data.frame(c(my_samples)), aes(x=c.my_samples., y=..density..)) +
    geom_histogram(fill="steelblue", color="grey80", size=.2, alpha=.8) +
    geom_density(color="grey10", size=1) +
    geom_vline(xintercept=1/lambda, linetype="dashed", color="red") +
    annotate("text", size=6, hjust=0, x = 1/lambda+2, y=.1, color="red"
             , label=paste("mean=",1/lambda)) +
    xlab("Sample Values") +
    ggtitle("Distribution of All Samples")
```




```
ggplot(data.frame(sample_means), aes(x=sample_means, 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=6, hjust=0, x = mean(sample_means)+.5, y=.5, color="red"
             , label=paste("mean=",round(mean(sample_means),2))) +
    xlab("Sample Values") +
    ggtitle("Sampling Distribution of the Mean")
```



Visually, the distribution of the sample means is much closer to normal than the distribution of the sample itself. To confirm our visual check with something more concrete, 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%"
```

We can therefore feel comfortable that the sampling distribution of the means is indeed normally distributed.

Section 2: Basic Inferential Data Analysis

Using the ToothGrowth data from the R datasets package, we will first perform some exploratory data analysis to get a feel for the data set and then provide a basic summary. We'll then compare tooth growth by supp and dose, using confidence intervals and hypothesis testing. Based on this analysis, we'll show that….

Exploratory Data Analysis

The first step is to load the packages we'll need, along with the dataset, and get a sense of the structure of ToothGrowth. R's documentation states that the ToothGrowth dataset is the effect of Vitamin C on tooth growth in guinea pigs. Ten guinea pigs were each given three dose levels of Vitamin C (0.5, 1 and 2 mg) with each of two delivery methods (orange juice and ascorbic acid), and their tooth growth measured after each test. Therefore, we'd expect to see 60 observations in the dataset.

```
ensurePkg("dplyr")
data("ToothGrowth")
str(ToothGrowth)
```

```
## 'data.frame':    60 obs. of  3 variables:
##  $ len : num  4.2 11.5 7.3 5.8 6.4 10 11.2 11.2 5.2 7 ...
##  $ supp: Factor w/ 2 levels "OJ","VC": 2 2 2 2 2 2 2 2 2 2 ...
##  $ dose: num  0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 ...
```

We see that there are 60 observations of three variables: * len, which is the length of observed tooth growth * supp, a variable with two values: "OJ" and "VC", which is the supplement type * dose, which is the dose in milligrams of orange juice or vitamin C given to each subject

Taking a look at a few of the rows, it appears that there might be some correlation between `dose` and `len`:

```
head(ToothGrowth)
```

```
##    len supp dose
## 1  4.2   VC  0.5
## 2 11.5   VC  0.5
## 3  7.3   VC  0.5
## 4  5.8   VC  0.5
## 5  6.4   VC  0.5
## 6 10.0   VC  0.5
```

```
tail(ToothGrowth)
```

```
##     len supp dose
## 55 24.8   OJ    2
## 56 30.9   OJ    2
## 57 26.4   OJ    2
## 58 27.3   OJ    2
## 59 29.4   OJ    2
## 60 23.0   OJ    2
```

Indeed, the correlation between the two variables is fairly high, but deeper analysis will show whether it's more than chance.

```
cor(ToothGrowth$len, ToothGrowth$dose)
```

```
## [1] 0.8026913
```

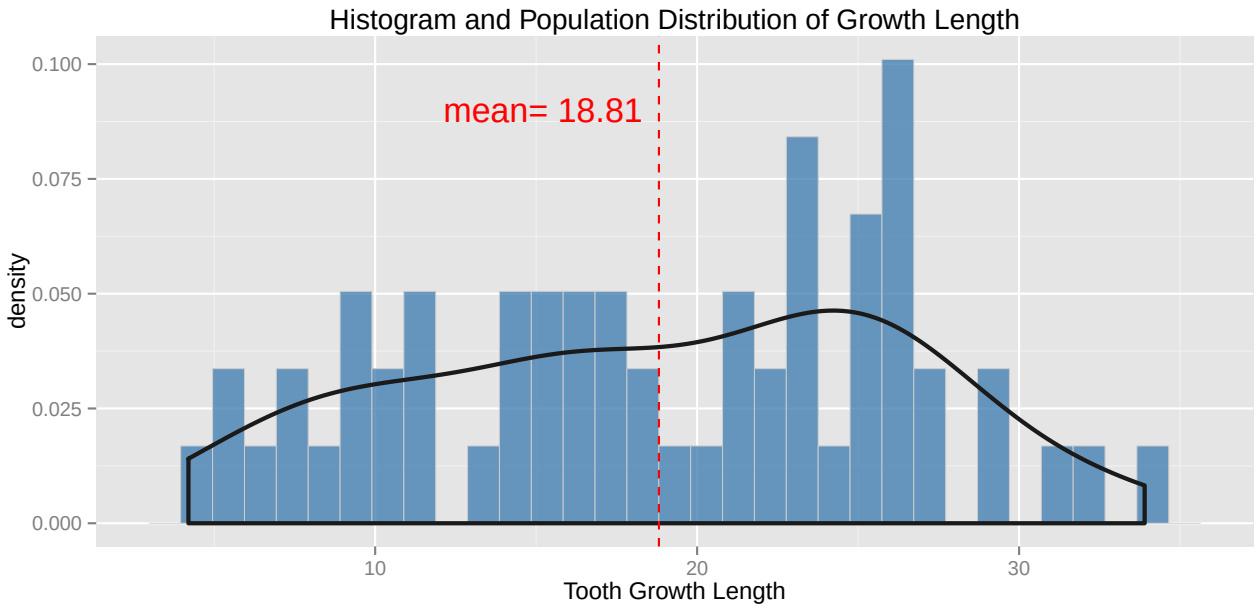
To get a sense of the range of values for `len`, we run a quick summary of the variable:

```
summary(ToothGrowth$len)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4.20   13.08   19.25   18.81   25.28   33.90
```

The results span quite a range, from 4.2 to 33.9. The mean and median are quite close, so let's plot a histogram of `len` to see if it is distributed normally.

```
ggplot(ToothGrowth, aes(x=ToothGrowth$len, y=..density..)) +
    geom_histogram(fill="steelblue", color="grey80", size=.2, alpha=.8) +
    geom_density(color="grey10", size=1) +
    geom_vline(xintercept=mean(ToothGrowth$len), linetype="dashed", color="red") +
    annotate("text", size=6, hjust=1, x = mean(ToothGrowth$len)-.5, y=.09, color="red"
             , label=paste("mean=",round(mean(ToothGrowth$len),2))) +
    xlab("Tooth Growth Length") +
    ggtitle("Histogram and Population Distribution of Growth Length")
```



With so few observations, it's difficult to tell the distribution visually but it is roughly normal so we check to see if 95% of the values are within 1.96 standard deviations of the mean:

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

```
## [1] "Percentage of tooth growth length within 1.96 standard deviations:  98.3%"
```

Not exactly 95% but also not dramatically distant, so it's again inconclusive and we cannot state definitively that the `len` variable follows a normal distribution.

Finally, what are the doses that were administered and can we confirm that they were given equally in the form of both orange juice and ascorbic acid?

```
unique(ToothGrowth$dose)
```

```
## [1] 0.5 1.0 2.0
```

```
with(ToothGrowth, table(supp, dose))
```

```
##     dose
## supp 0.5  1  2
##   OJ  10 10 10
##   VC  10 10 10
```

Summary of the Data

Before moving on to comparing tooth growth results by dose and source of ascorbic acid, let's summarize what we know:

- There are 60 observations in total, with six per guinea pig

- Each guinea pig was tested with three doses each of orange juice and ascorbic acid
- There is a wide range of tooth growth, so presumably some effect is taking place
- There appears to be a correlation between the length of tooth growth and the dose
- The measurements of tooth growth do not appear to adhere strictly to any particular distribution, although the sample size is not large
- The samples are paired, with equal numbers and types of treatments given to each subject

Comparison of Tooth Growth Results by 'supp' and 'dose'

For each guinea pig, we have six combinations of `supp` and `dose`. We can group the data for each of those six scenarios and compute the mean of each:

```
results <- ToothGrowth %>% group_by(supp, dose) %>% summarize(mean = mean(len))
print(results)
```

```
## Source: local data frame [6 x 3]
## Groups: supp
##
##   supp dose  mean
## 1   OJ  0.5 13.23
## 2   OJ  1.0 22.70
## 3   OJ  2.0 26.06
## 4   VC  0.5  7.98
## 5   VC  1.0 16.77
## 6   VC  2.0 26.14
```

The means of tooth growth do increase in response to greater dosages of Vitamin C, regardless of the source of the C, but how can we be sure that what we're seeing isn't just chance? One method is to compare the dosages. If Vitamin C supplements do not result in increased tooth growth, the means for 0.5mg and 2.0mg should not differ significantly. Therefore, our null hypothesis is:

$$H_0 : \mu_2 = \mu_{0.5} or,\ equivalently H_0 : \mu_2 - \mu_{0.5} = 0$$

On the other hand, if Vitamin C does enhance tooth growth, we would expect to see increased growth under the 2.0mg dose, so our alternative hypothesis is:

$$H_a : \mu_2 > \mu_{0.5} or,\ equivalently H_a : \mu_2 - \mu_{0.5} > 0$$

To determine whether or not to reject the null hypothesis above, we'll set the level of significance (α) to 0.05.

```
n <- 10
alpha <- 0.05
```

With a sample size of only 10, we will use a two-sample t-test but are the variances equal or not? We can run a quick F test for equality of variances and look at the P-Value. The null hypothesis of the F test is that the variances are equal, so if the resulting P-Value is greater than our α of 0.05, we fail to reject this null hypothesis.

```
var.test(ToothGrowth[ToothGrowth$dose == 0.5,]$len, ToothGrowth[ToothGrowth$dose == 2,]$len)$p.value
```

```
## [1] 0.4504979
```

The P-Value of 0.45 means that we can move forward with the two-sample t-test with equal variances.

Let's first subset by 0.5mg and 2.0mg dosages, and calculate μ_2 and $\mu_{0.5}$:

```
half_dose <- ToothGrowth[ToothGrowth$dose == 0.5, ]$len
two_dose <- ToothGrowth[ToothGrowth$dose == 2, ]$len
```

The formulas we're using to calculate the test statistic are:

$$t = \frac{\bar{X}_2 - \bar{X}_{0.5}}{s_p\sqrt{\frac{1}{n_2} + \frac{1}{n_{0.5}}}}$$

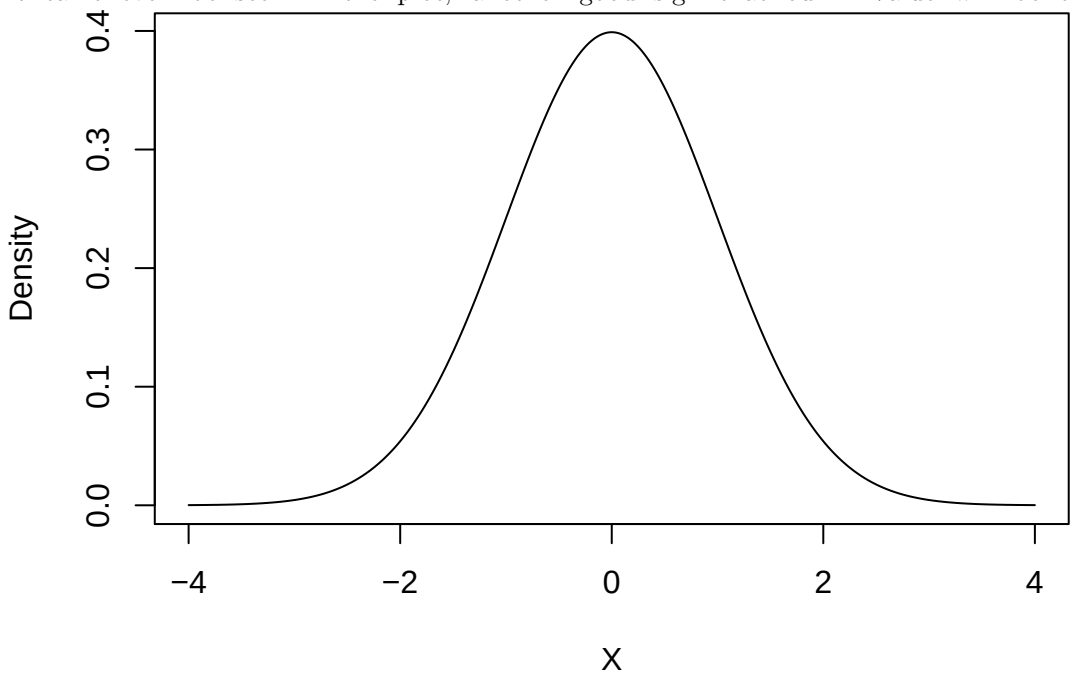
$$s_p = \sqrt{\frac{(n_2 - 1)s_2^2 + (n_{0.5} - 1)s_{0.5}^2}{n_2 + n_{0.5} - 2}}$$

We could use the R function `t.test()` but we'll solve for our t statistic formulaically. Note that since we have both orange juice and Vitamin C values for both doses, our sample size is 20.

```
n <- 20
x_two <- mean(two_dose)
x_half <- mean(half_dose)
s_two <- sd(two_dose)
s_half <- sd(half_dose)

sp <- sqrt( ((n-1)*s_two^2 + (n-1)*s_half^2)/(2*n-2) )
t <- (x_two-x_half)/(sp*sqrt(1/n + 1/n))
```

This results in a t statistic of 11.8, which is quite large and signals that we will have a very small P-Value. Since t distributions have the shape of a normal distribution, we can shade the portion of a normal curve that corresponds to our computed test statistic, but since it is so large, it can't even be seen in the plot, another good sign that our P-Value will be extremely small.



Even so, we calculate our P-Value as:

```r
t.test(two_dose, half_dose, alternative = "greater", var.equal = TRUE)$p.value
```

```
## [1] 0.00000000000001418777
```

Our P-Value $< \alpha$, so we reject the null hypothesis, in favor of the alternative hypothesis: there is evidence to suggest that Vitamin C in some form results in increased tooth growth. (It would require an additional analysis to determine if orange juice or Vitamin C is more effective.)

To double-check our work, we look at the confidence intervals, using the formula:

$$\bar{X}_2 - \bar{X}_{0.5} \pm t_{df} s_p \sqrt{\frac{1}{n_2} + \frac{1}{n_{0.5}}}$$

But we can pull the 95% confidence interval from the `t.test()` results:

```r
test <- t.test(two_dose, half_dose, var.equal = TRUE)
```

Which results in a lower bound of 12.84 and an upper bound 18.15, which comfortably include the difference between our two sample means of 15.495, so the difference in our sample means is solidly within the range of our 95% confidence interval.

Conclusions