Coursera Statistical Inference Course Project

Sherman Wood

July 21, 2016

Introduction

This is my submission for the Coursera Statistical Inference Course Project.

The project consists of two parts:

1. A simulation exercise.
2. Basic inferential data analysis.

Simulation Exercise

We will investigate the exponential distribution in R and compare it with the Central Limit Theorem.

The Central Limit Theorem states that as n gets larger, the distribution of the difference between the sample average Sn and its limit mu, when multiplied by the factor sqrt(n) (that is sqrt(n)(Sn - mu)), approximates the normal distribution with mean 0 and variance sigma^2. For large enough n, the distribution of Sn is close to the normal distribution with mean mu and variance sigma^2/n.

The exponential distribution can be simulated in R with rexp(n, lambda) where lambda is the rate parameter. The theoretical mean of an exponential distribution is 1/lambda and the theoretical standard deviation is also 1/lambda.

For this investigation, we will evaluate the distribution of means of 1,000 samples of 40 random exponentials with lambda = 0.2.

For this exponential distribution, the Central Limit Theorem says that:

mu = 1/lambda = theoretical sample mean (Sn) = 5

Theoretical Sample variance = (1/lambda)^2 / n = 0.625

First let's get our sample exponential distribution.

```
for (i in 1 : simulations) means = c(means, mean(rexp(n,lambda)))
```

Sample mean (Sn) = 5.0156

Sample variance = 0.6065

Sample standard deviation = 0.7788

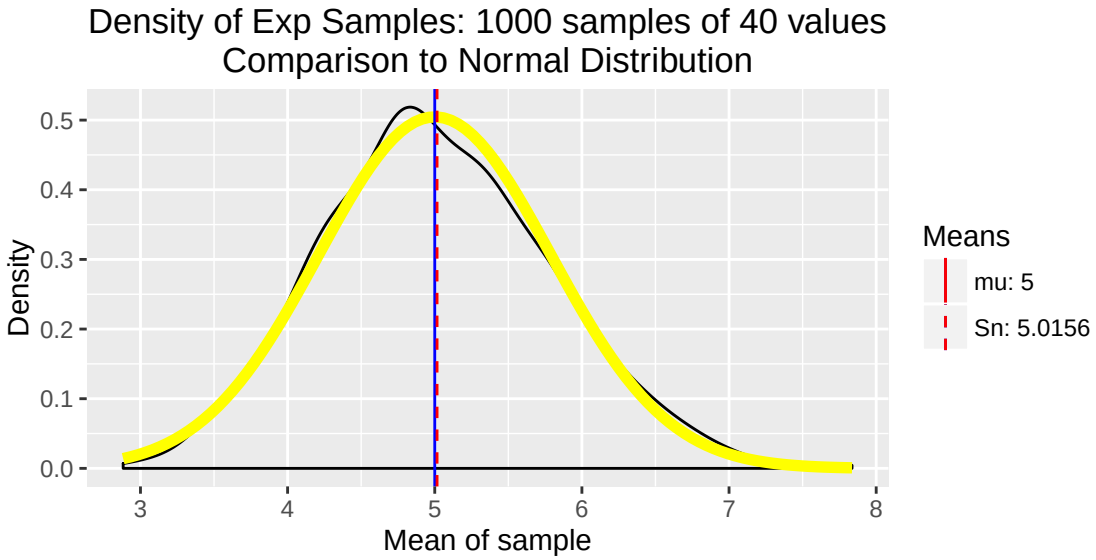
Compare the sample distribution statistics to the values expected from the Central Limit Theorem :

Theoretical mean mu - Sn = -0.0156

Theoretical - Sample variance = 0.0185

These differences are close to 0. The Central Limit Theorem assertions are thus supported.

A density plot allows us to compare the sample distribution to the normal distribution. The closeness of these curves highlights the Central Limit Theorem assertion that for large enough n, the distribution of Sn is close to the normal distribution with mean mu and variance sigma^2/n.



Basic Inferential Data Analysis

In this inferential analysis, we are going to explore the ToothGrowth data set that is part of the base R distribution. Let's start with an overview of the base ToothGrowth data set.

The documentation for the ToothGrowth data set says:

The response is the length of odontoblasts (cells responsible for tooth growth) in 60 guinea pigs. Each animal received one of three dose levels of vitamin C (0.5, 1, and 2 mg/day) by one of two delivery methods, (orange juice (coded as OJ) or ascorbic acid (a form of vitamin C and coded as VC).

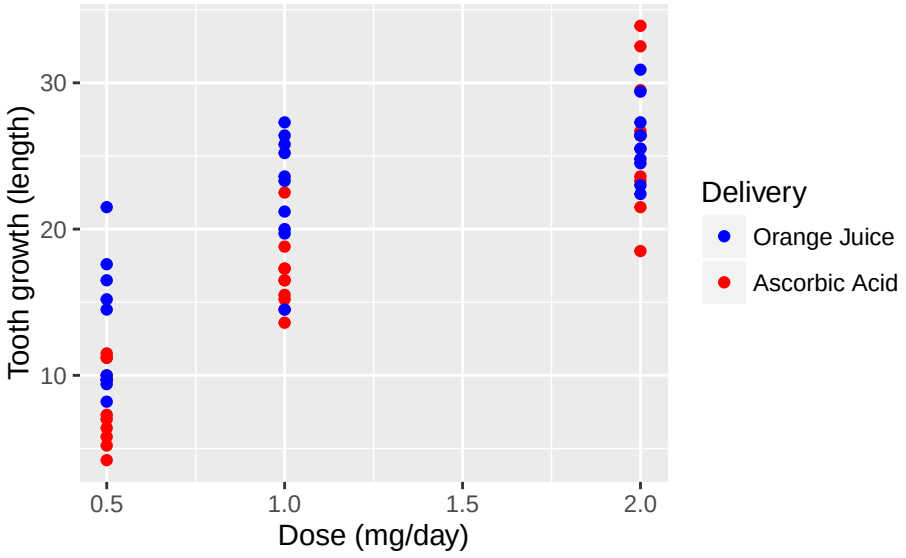
```
table(ToothGrowth$supp, ToothGrowth$dose)
```

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

We can see 10 observations for each dose/delivery method combination. The observations are independent.

We will assume that the length of time for the observations is the same, so tooth growth is comparable across the combinations.

ToothGrowth data profile



Comparison of tooth growth

Let us investigate whether the tooth growth was significantly differentiated by delivery method - orange juice versus ascorbic acid, regardless of dosage. This can be highlighted by a t-test of tooth growth across the samples partitioned by delivery method.

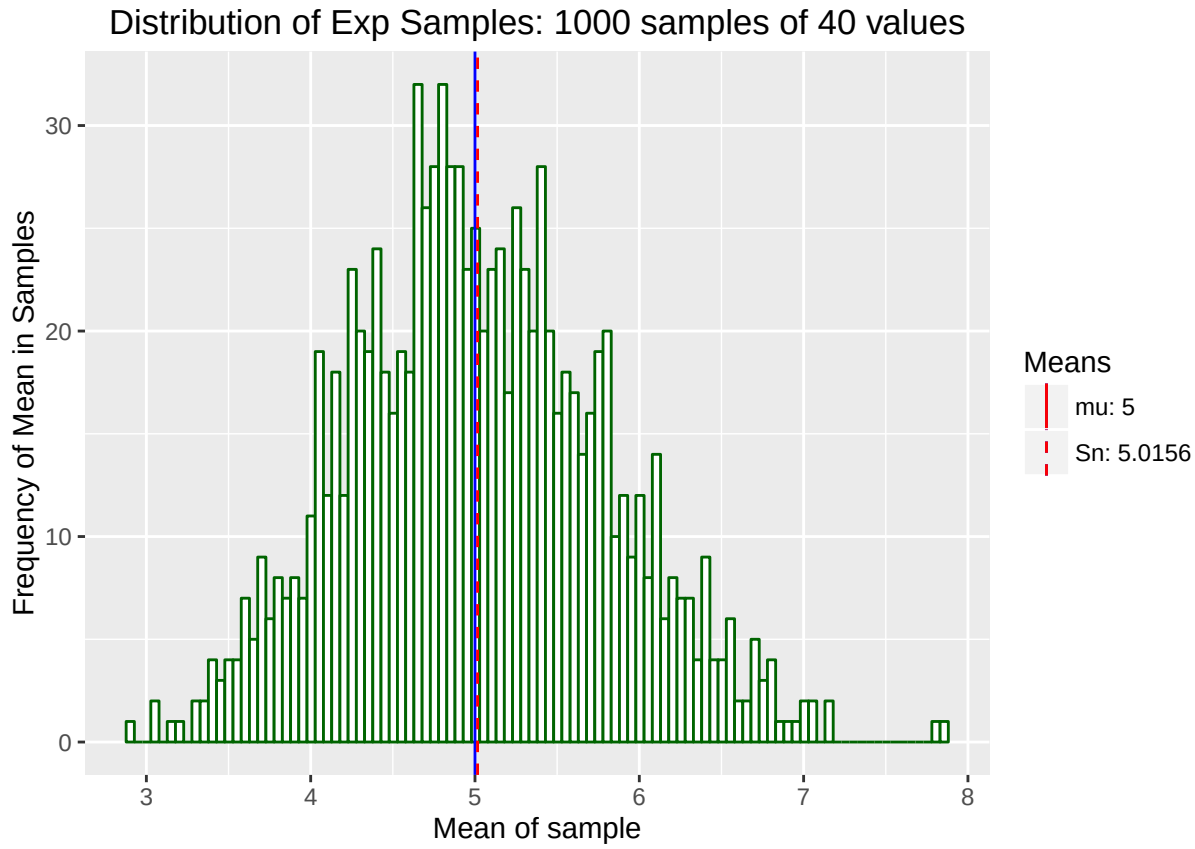
We will assume that the variances between the two groups is not equal.

```
##
##  Welch Two Sample t-test
##
## data:  oj$len and vc$len
## t = 1.9153, df = 55.309, p-value = 0.06063
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1710156  7.5710156
## sample estimates:
## mean of x mean of y
##  20.66333  16.96333
```

Conclusions

Given a t statistic of 1.9153 and a a low p-value of 0.0606, this test supports the hypothesis that orange juice provides significantly better tooth growth than asorbic acid. The 95% confidence interval of -0.171, 7.571, which is only just negative on the lower tail, also supports this conclusion.

Appendix



ToothGrowth data summaries

```
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
```

```
xtabs(len ~ dose, aggregate(len ~ dose,ToothGrowth,mean))
```

```
## dose
##    0.5      1      2
## 10.605 19.735 26.100
```

```
xtabs(len ~ supp, aggregate(len ~ supp,ToothGrowth,mean))
```

```
## supp
##       OJ       VC
## 20.66333 16.96333
```

```
xtabs(len ~ dose, aggregate(len ~ dose,ToothGrowth,quantile))
```

```
##
## dose      0%    25%    50%    75%   100%
##   0.5  4.200  7.225  9.850 12.250 21.500
##   1   13.600 16.250 19.250 23.375 27.300
##   2   18.500 23.525 25.950 27.825 33.900
```

```
xtabs(len ~ supp, aggregate(len ~ supp,ToothGrowth,quantile))
```

```
##
## supp     0%    25%    50%    75%   100%
##   OJ  8.200 15.525 22.700 25.725 30.900
##   VC  4.200 11.200 16.500 23.100 33.900
```

```
xtabs(len ~ dose + supp, aggregate(len ~ dose + supp,ToothGrowth,mean))
```

```
##      supp
## dose     OJ    VC
##   0.5 13.23  7.98
##   1   22.70 16.77
##   2   26.06 26.14
```