Simulations and Inferential Data Analysis, Statistical Inference: Inferential Testing

Scott Zuehlke

December 3, 2016

This report was produced as part of the Coursera Data Science Specialization, Statistical Inference course. All code was produced on a Mac running OS X 10.11.6, with RStudio version 0.99.902. This report was produced as part of the Coursera Data Science Specialization, Statistical Inference course. The purpose of this report is to twofold: Part one is to run simulations against the exponential distribution, therefore confirming the bootstrap method works, by checking sample distribution mean and spread, obtained by simulation, against the theoreteical mean and spread. Part two is to run basic inferential analysis by testing the effects of dosage and supplmenet on Guinea Pig tooth growth.

Part 2: Basic Inferential Data Analysis

Part 2: Basic Inferential Data Analysis Instructionsless 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.

Start by loading the ToothGrowth data set from the datasets library

```
library(datasets)
data(ToothGrowth)
```

First, take a look at the ToothGrowth data structures.

```
toothGrowthData <- ToothGrowth
str(toothGrowthData)
```

```
## '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 ...
```

Convert dose to a factor.

```
toothGrowthData$dose <- as.factor(toothGrowthData$dose)
str(toothGrowthData)
```

```
## '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: Factor w/ 3 levels "0.5","1","2": 1 1 1 1 1 1 1 1 1 1 ...
```

```r
summary(toothGrowthData)
```

```
##       len        supp     dose
##  Min.   : 4.20   OJ:30   0.5:20
##  1st Qu.:13.07   VC:30   1  :20
##  Median :19.25           2  :20
##  Mean   :18.81
##  3rd Qu.:25.27
##  Max.   :33.90
```

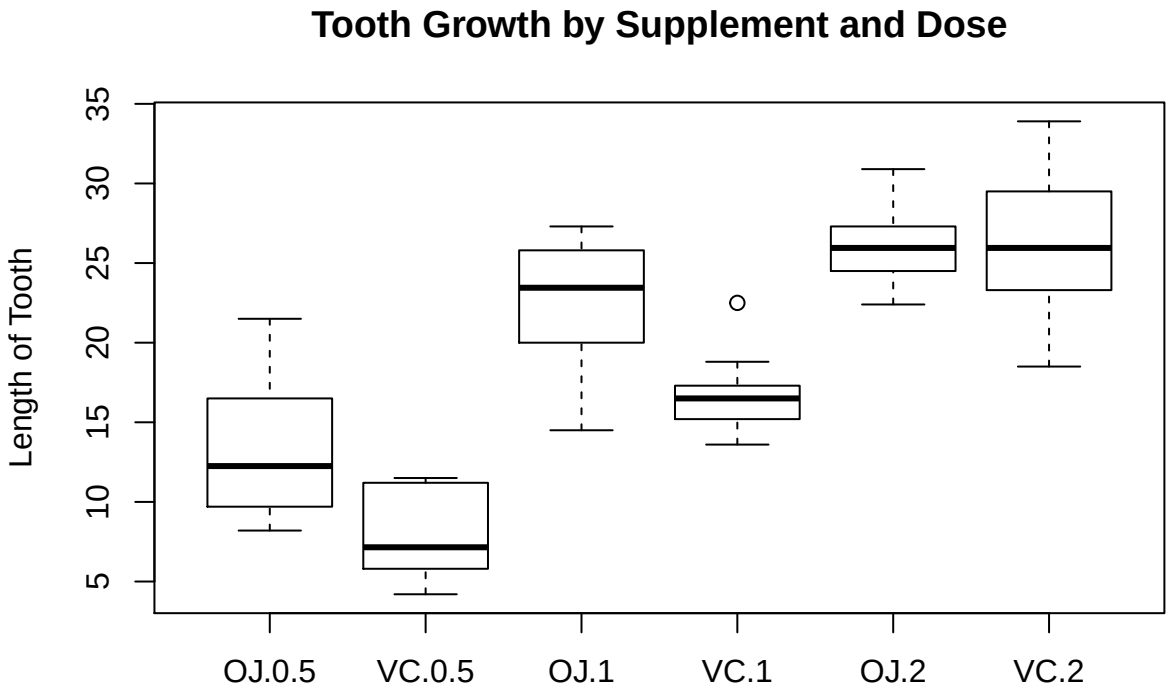
```r
head(toothGrowthData,n = 10)
```

```
##     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
## 7  11.2   VC  0.5
## 8  11.2   VC  0.5
## 9   5.2   VC  0.5
## 10  7.0   VC  0.5
```

```r
table(toothGrowthData$supp,toothGrowthData$dose)
```

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

```r
boxplot(len ~ supp * dose, toothGrowthData,
        ylab="Length of Tooth",
        main = "Tooth Growth by Supplement and Dose")
```



In looking at the boxplots, it looks like there could be a relationship between dose and tooth growth, but not as much a relationship in the supplement differences. To test these assumptions, we'll run hypothesis tests

Checking the sample size,

```
nrow(toothGrowthData)
```

```
## [1] 60
```

it makes sense to use a t test. Use the R function t.test.

First: tooth growth by supp. Check to see if different supp causes a change in tooth growth. Our test will be that H0: mu1 = mu2, and Ha: mu1 =/= mu2.

```
t.test(len ~ supp, data = toothGrowthData)
```

```
##
##  Welch Two Sample t-test
##
## data:  len by supp
## 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 in group OJ mean in group VC
##         20.66333         16.96333
```

The returned p value is .06, which confirms the initial assumption that the means aren't statistsically different. Looking at the 95% confidence interval, (-.17,7.57), it contains 0. This is a different way of confirming that the difference in lengths of teeth by changing supplements is not statistically different.

To test tooth growth by dose, first break the different dose levels out so they can be tested separately.

```r
toothGrowthdoses_0501 <- subset(toothGrowthData, dose %in% c(0.5, 1.0))
toothGrowthdoses_0502 <- subset(toothGrowthData, dose %in% c(0.5, 2.0))
toothGrowthdoses_0102 <- subset(toothGrowthData, dose %in% c(1.0, 2.0))
```

We're testing the hypothesis that there's a difference in lengths between a dose of .5 and 1. So, the hypotheses are H0:mu.5 = mu1, and Ha: mu.5 =/= mu1.

```r
t.test(len ~ dose, toothGrowthdoses_0501)
```

```
##
##  Welch Two Sample t-test
##
## data:  len by dose
## t = -6.4766, df = 37.986, p-value = 1.268e-07
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.983781  -6.276219
## sample estimates:
## mean in group 0.5   mean in group 1
##            10.605            19.735
```

The p-value is essentially 0, which confirms there is a statistical difference in tooth length between doses of .5 and 1. The confidence interval is (-11.98,-6.23), which illustrates again, but in a different way, there is a difference.

Here, we're testing the hypothesis that there's a difference in lengths between a dose of .5 and 2. So, the hypotheses are H0:mu.5 = mu2, and Ha: mu.5 =/= mu2.

```r
t.test(len ~ dose, toothGrowthdoses_0502)
```

```
##
##  Welch Two Sample t-test
##
## data:  len by dose
## t = -11.799, df = 36.883, p-value = 4.398e-14
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -18.15617 -12.83383
## sample estimates:
## mean in group 0.5   mean in group 2
##            10.605            26.100
```

Once again, the p-value is essentially 0, which confirms there is a statistical difference in tooth length between doses of .5 and 2. The confidence interval is (-18.16,-12.83), which illustrates again, but in a different way, there is a difference.

Finally, we're testing the hypothesis that there's a difference in lengths between a dose of 1 and 2. So, the hypotheses are H0:mu1 = mu2, and Ha: mu1 =/= mu2.

```r
t.test(len ~ dose, toothGrowthdoses_0102)
```

```
##
##  Welch Two Sample t-test
##
## data:  len by dose
## t = -4.9005, df = 37.101, p-value = 1.906e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -8.996481 -3.733519
## sample estimates:
## mean in group 1 mean in group 2
##          19.735          26.100
```

In this last test, the p-value is essentially 0, which confirms there is a statistical difference in tooth length between doses of .5 and 2. The confidence interval is (-8.99,-3.73), which illustrates again, but in a different way, there is a difference.

Conclusions: In the analysis in part 2, we confirmed there were no statistical differences in the length of teeth when changing the given supplements, however there are statistical differences between all three different levels of doses, confirming that a change in the doses of supplements does impact tooth length, even if the individual supplements do not.

Assumptions

1. We're assuming the experiment was done with a random assignment of guinea pigs. This random dispersement would include random samples of both supplement type and dosage.

2. We're also assuming that each sample of guinea pigs is representative of the true population.

3. Finally, in the t-tests, given the boxplot, we assume the variances are unequal.