Statistical Inference Course Project Part 2 - Tooth Growth Analysis

Randall Helms

25 October 2016

Introduction

This document covers the second part of the Statistical Inference course project. This course is part of the Data Science Specialization offered by Johns Hopkins University via Coursera.

This second part uses the ToothGrowth data set from the `datasets` package to do some basic data visualization and inferential data analysis.

The ToothGrowth data set shows the results from clinical trials studying the effect of vitamin c on tooth growth in guinea pigs.

Here's the description from ?ToothGrowth:

> 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 or ascorbic acid (a form of vitamin C and coded as VC).

In order to analyze the results, we will do the following things:

1. Load the ToothGrowth data and perform some basic exploratory data analyses
2. Provide a basic summary of the data.
3. 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)
4. State your conclusions and the assumptions needed for your conclusions.

Part 1

Let's get started by loading the relevant libraries and the actual data set:

```
library(ggplot2)
library(psych)
```

```
##
## Attaching package: 'psych'
```

```
## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
```

```
library(RColorBrewer)

#load data

data("ToothGrowth")
ToothGrowth$dose<-as.factor(ToothGrowth$dose) #relevant because dose is a factor
```

Now that the data has been loaded into R, let's do some basic analyses:

```
head(ToothGrowth) #check the first six rows
```

```
##    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
```

```
str(ToothGrowth) #check the structure
```

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

Let's now do use the summary function to analyze the data set

```
summary(ToothGrowth) #basic summary
```

```
##       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
```

And a more detailed summary, from the `psych` package:

```
describe(ToothGrowth)
```

```
##       vars  n  mean   sd median trimmed  mad min  max range  skew kurtosis
## len      1 60 18.81 7.65  19.25   18.95 9.04 4.2 33.9  29.7 -0.14    -1.04
## supp*    2 60  1.50 0.50   1.50    1.50 0.74 1.0  2.0   1.0  0.00    -2.03
## dose*    3 60  2.00 0.82   2.00    2.00 1.48 1.0  3.0   2.0  0.00    -1.55
##         se
## len   0.99
## supp* 0.07
## dose* 0.11
```

Part 2

Now that we have loaded the data set up, let's use `ggplot2` and `stats` to provide a basic summary of the data.

Since this data set comprises one metric (tooth length) and two factors (supplement method and dose level), it makes sense to compare tooth length to each of the factors in turn.

Let's start by checking the mean tooth length by supplement method:

```
aggregate(len ~ supp, data=ToothGrowth,mean)
```

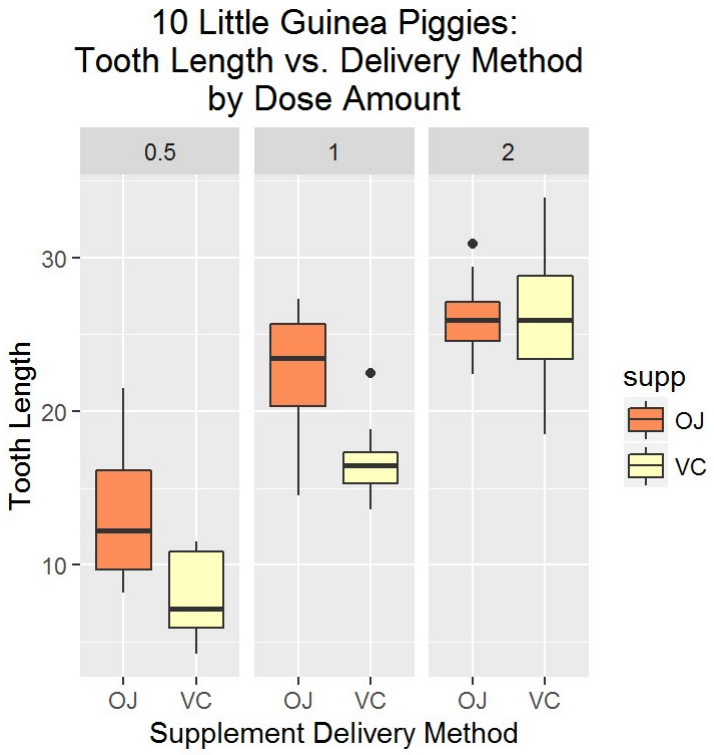
```
##   supp      len
## 1   OJ 20.66333
## 2   VC 16.96333
```

This would suggest that orange juice promotes greater tooth growth than ascorbic acid, but the extent to which we can statistically infer this will become clearer as we work through the other parts of the analysis.

Now let's plot out tooth growth by supplement, both overall and then seperated by dose level:

```
ggplot(ToothGrowth, aes(x = supp, y = len)) +
  geom_boxplot(aes(fill=supp)) +
  scale_fill_brewer(palette = "Spectral")+
  xlab("Supplement Delivery Method")+
  ylab("Tooth Length")+
  ggtitle("10 Little Guinea Piggies:\nTooth Length vs. Delivery Method")
```

```
ggplot(ToothGrowth, aes(x = supp, y = len)) +
  geom_boxplot(aes(fill=supp)) +
  facet_grid(~ dose)+
  scale_fill_brewer(palette = "Spectral")+
  xlab("Supplement Delivery Method")+
  ylab("Tooth Length")+
  ggtitle("10 Little Guinea Piggies:\nTooth Length vs. Delivery Method \nby Dose A
mount")
```



Now let's repeat that process, by checking out the mean tooth growth by dose level:

```
aggregate(len ~ dose, data=ToothGrowth,mean)
```

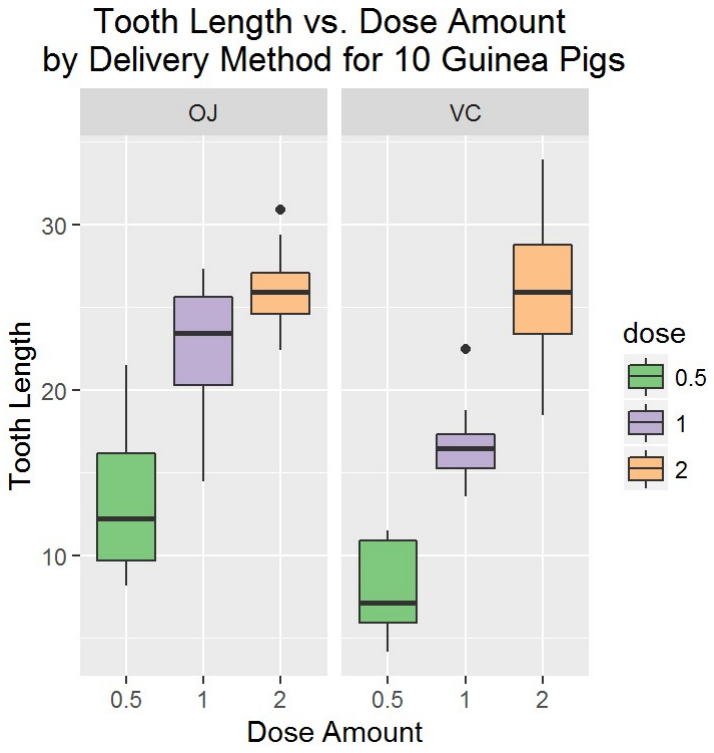
```
##   dose    len
## 1  0.5 10.605
## 2    1 19.735
## 3    2 26.100
```

Here we see what looks like a strong relationship between tooth growth and increasing dosages, however it is best to continue the analysis before coming to a final conclusion.

OK, let's graph tooth growth by dose level, both overall and by supplement method:

```
ggplot(ToothGrowth, aes(x = dose, y = len)) +
  geom_boxplot(aes(fill=dose)) +
  scale_fill_brewer(palette = "Accent")+
  xlab("Dose Amount")+
  ylab("Tooth Length")+
  ggtitle("Tooth Length vs. Dose Amount \n for 10 Guinea Pigs")
```


```
ggplot(ToothGrowth, aes(x = dose, y = len)) +
  geom_boxplot(aes(fill=dose)) +
  facet_grid(~ supp)+
  scale_fill_brewer(palette = "Accent")+
  xlab("Dose Amount")+
  ylab("Tooth Length")+
  ggtitle("Tooth Length vs. Dose Amount \nby Delivery Method for 10 Guinea Pigs")
```



Part 3

OK, now that we have a pretty good idea of the basic characteristics of tooth growth levels vis a vis the

different supplement types and dose levels, let's take the analysis one step further by conducting a variety of t-tests to test the relationship between tooth growth and the various supplement types and dosages.

First, let's run a t-test to compare the two supplement types and test the null hypothesis that both supply methods produce the same results:

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

```
##
##  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
```

With a p-value of 0.061 there is insufficient evidence to reject the null hypothesis.

Next let's run a series of t-tests to check the null hypothesis that all dosage levels produce the same results. To start, let's see what happens when we increase the dose from 0.5 to 1 mg per day:

```
tgTest1 <- subset(ToothGrowth,dose %in% c(0.5,1))
t.test(len~dose,data=tgTest1,paired = FALSE, var.equal = FALSE)
```

```
##
##  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 extremely low p-value resulting from this test would indicate that increasing the dosage results in a different mean value for tooth growth. Let's try this again by testing the dosage increase from 1 to 2 mg per day:

```
tgTest2 <- subset(ToothGrowth,dose %in% c(1,2))
t.test(len~dose,data=tgTest2,paired = FALSE, var.equal = FALSE)
```

```
##
##  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
```

This p-value isn't quite as low as for the previous test, however it is still well below the 0.05 threshold for us to reject the null hypothesis. For the last test, let's check what happens when you increase the dosage from 0.5 mg per day all the way up to 2 mg per day:

```
tgTest3 <- subset(ToothGrowth,dose %in% c(0.5,2))
t.test(len~dose,data=tgTest3,paired = FALSE, var.equal = FALSE)
```

```
##
##  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
```

This p-value is the lowest yet, again indicating that increasing the dosage results in increasing tooth growth. Taken together, this would show that there is sufficient evidence to reject the null hypothesis that tooth growth is equal across all dosage levels.

Part 4

Based on the results of the t-tests, we fail to reject the null hypothesis that supplement types don't affect tooth growth, and we reject the null hypothesis that dosage levels don't affect tooth growth.

To put that in more plain terms, we conclude from this analysis that mean tooth growth is positively increased by increasing Vitamin C dose levels, and unaffected by the choice of delivery method.