An Exploration of ToothGrowth

Mark Bulkeley

April 2, 2016

Overview

This brief paper is intended to explore the data in the `R` `datasets` package called `ToothGrowth` and to make some appropriate statistical inferences about the data. Specifically, we will determine whether the Supplement (`supp`) and Dose (`dose`) have a significant impact on the Length (`len`).

Assumptions

We are assuming that the underlying population of Length is coming from a normal or gaussian distribution. Because we are dealing with small sample sizes (especially when segmented to do the statistical tests), we will be using a Student T distribution that is slightly more tail-heavy than the normal distribution, but provides a more accurate accounting at low sample size (small `n`). We will evaluate significance from an alpha of 95%, or from the perspective that our test will be wrong only 1 out of 20 times. In addition, we will run one-sided tests and structure them in a way that the alternate hypothesis will be that the second group promotes greater length than the first.

Brief Description of the Data

The ToothGrowth dataset consists of 60 rows with 3 columns. The three column names are `len`, `dose` and `supp`. `len` is a continuous variable with a range of 4.2, 33.9. `dose` has levels 0.5, 1, 2 (mg of dosage) and `supp` has levels of VC, OJ. Note that `VC` is short for Vitamin C and `OJ` is short for Orange Juice. The data is based on an original study by Crampton, E.W. "The growth of the odontoblasts of the incisor tooth as a criterion of the vitamin C intake of the guinea pig", published in *The Journal of Nutrition*, vol. 33, issue 5, May 1947, pp. 491-504. Details of the original study indicate that there were 60 different guinea pigs used, thus none of our tests will be paired (all independent samples). Please refer to Table 1 and Figure 1 in the Appendix to further understand and explore the data. Both include the code required to generate, for reproducibility.

Test of Supplement Significance

Because of the small data set, we are going to rely on Student T tests to determine significance levels. Given that there are only two levels of `supp`, this can be done simply with the `t.test` function provided with `R`. Note that the Null Hypothesis that will be tested is that the difference in means is zero. We would like to reject that to be able to say that the supplement type does matter.

```
lst.test <- t.test(x=tg[supp=="VC"]$len, y=tg[supp=='OJ']$len,
                   paired = FALSE, alternative = "greater")
print(lst.test)
```

```
##
##  Welch Two Sample t-test
##
## data:  tg[supp == "VC"]$len and tg[supp == "OJ"]$len
```

```
## t = -1.9153, df = 55.309, p-value = 0.9697
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -6.931731       Inf
## sample estimates:
## mean of x mean of y
##  16.96333  20.66333
```

Our sample difference in means is -3.7. Given that the p-value is 0.9696827, which is greater than 0.95, we will reject the Null Hypothesis that the means are the same. Thus, statistically, we can (barely) confidently distiguish between `OJ` and `VC` supplement types.

Test of Dose Significance

To test the significance of dosage, we are going to run two tests, the second dependent on the outcome of the first. Our first test will be between 0.5 and 1 mg. If that is not significant, the follow-on test will be between 0.5 and 2 mg. If it is significant, the second test will be between 1 and 2 mg. To do the tests, we will be using a Student T test with a Null Hypothesis that the distributions are the same and a significance level of p = 0.05 (the default for the `t.test` function). It will take the same form as our tests previously.

0.5mg Versus 1mg Test

```
lst.test <- t.test(x=tg[dose==0.5]$len, y=tg[dose==1]$len,
                   paired = FALSE, alternative = "greater")
print(lst.test)
```

```
##
##  Welch Two Sample t-test
##
## data:  tg[dose == 0.5]$len and tg[dose == 1]$len
## t = -6.4766, df = 37.986, p-value = 1
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -11.50668       Inf
## sample estimates:
## mean of x mean of y
##    10.605    19.735
```

Here, we have a very high p-value of 0.9999999 that supports the Null Hypothesis being rejected. Therefore, we can infer from our data that a 1mg dose has an impact on length, relative to a 0.5mg dose.

1mg Versus 2mg Test

```
lst.test <- t.test(x=tg[dose==1]$len, y=tg[dose==2]$len,
                   paired = FALSE, alternative="greater")
print(lst.test)
```

```
##
##  Welch Two Sample t-test
##
## data:  tg[dose == 1]$len and tg[dose == 2]$len
## t = -4.9005, df = 37.101, p-value = 1
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -8.55613      Inf
## sample estimates:
## mean of x mean of y
##    19.735    26.100
```

Here, as well, we have a just slightly less high p-value of 0.9999905 that supports the Null Hypothesis being rejected. Therefore, we can infer from our data that a 2mg dose has an impact on length, relative to a 1mg dose.

Conclusions

Based on our statistical inference tools, we conclude the following about the `ToothGrowth` dataset (for an alpha of 0.05):

1. We should accept the alternative hypothesis that the `OJ` has a greater impact on tooth growth than does `VC`.
2. We should infer that there is a very significant difference in dosage levels of 0.5mg, 1mg and 2mg.

Both of these conclusions make intuitive sense from our initial data exploration, particularly from a visual scan of the segmented boxplots in Figure 1. There are no further areas of study that we can recommend, given the classroom nature of this dataset.

Appendix

Code to Load ToothGrowth dataset

```
library(datasets)
library(data.table)
library(ggplot2)
library(gridExtra)
options(scipen = 999)
alpha <- 0.05

data(ToothGrowth)
tg <- data.table(ToothGrowth)
```

Table 1: ToothGrowth Data Exploration

```
grid.arrange(tableGrob(tg[, .(`Mean\nLength`=mean(len), `Std.Dev\nLength`=round(sd(len),4), n=.N),
                          by = list(Supplement =supp, Dose=dose)],
                       row = NULL, theme = ttheme_minimal()))
```

Supplement	Dose	Mean Length	Std.Dev Length	n
VC	0.5	7.98	2.7466	10
VC	1.0	16.77	2.5153	10
VC	2.0	26.14	4.7977	10
OJ	0.5	13.23	4.4597	10
OJ	1.0	22.70	3.9110	10
OJ	2.0	26.06	2.6551	10

Figure 1: ToothGrowth Data Exploration

```
ggplot(data = tg, mapping = aes(x = factor(dose), y = len, fill = supp)) +
    geom_boxplot() +
    scale_x_discrete(name = "Dose\n(in mg)") +
    scale_y_continuous(name = "Tooth Length (mm)") +
    scale_fill_discrete(name = "Supplement Type") +
    theme(legend.position = c(0.1, 0.8)) +
    ggtitle("Boxplot of ToothGrowth Data")
```