Exploratory Data Analysis in Guinea Pig Tooth Growth

Kshitij Kulkarni

Wednesday, October 07, 2015

Prerequisite Knowledge

The Guinea pig is a species of rodent who have been used for scientific research since 17th century. One of many medical research purposes involving the Guinea Pig involves research on 'Scurvy', a disease caused due to deficiency of Vitamin C.

It is because Guinea pigs are one of the few animals which, like humans and other primates, cannot synthesize vitamin C, but must obtain it from their diet, they are ideal for researching scurvy. Started from the accidental discovery in 1907 that scurvy could be induced in guinea pigs, the guinea pig model time and again proved that it would play a crucial part in the vitamin C research.

About the dataset

This report focuses on a basic exploratory data analysis on the Tooth Growth dataset in the R datasets package. The ToothGrowth dataset contains the e???ect of Vitamin C on tooth growth in Guinea Pigs. According to the R documentation, The response is the length of odontoblasts (teeth) in each of 10 guinea pigs at each of three dose levels of Vitamin C (0.5, 1, and 2 mg) with each of two delivery methods (orange juice or ascorbic acid).

It contains a data frame with 60 observations on 3 variables : 1. Len : Numeric Tooth length 2. Supp : Factor Supplement type (VC or OJ) 3. Dose : Numeric Dose in milligrams

Aim

We try to answer two questions in this mini-project: 1. Whether the supplement type of Vitamin C effects the teeth lengths (to grow). 2. Whether the amount of dosage of Vitamin C effects the tooth growth. This can be useful for the medical experts to decide the dosage and supplement type o Vitamin C to be prescribed for their patients.

Data Exploration

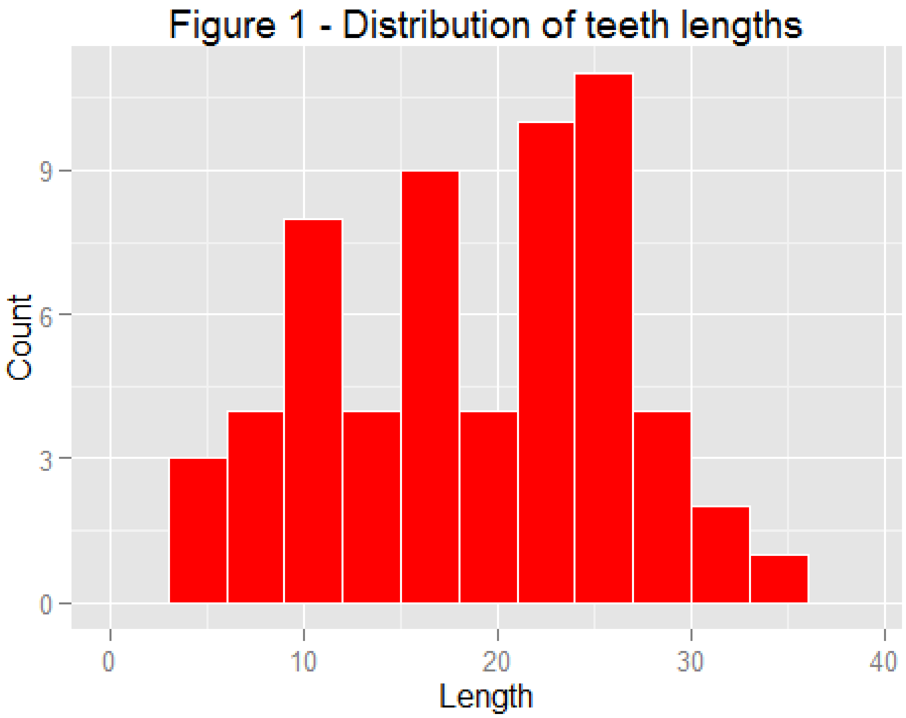
One of the ???rst steps in exploratory data analysis is to gain an understanding of the data.The summary of the data is as follows:

```
summary(ToothGrowth)
```

```
##       len        supp         dose      
##  Min.   : 4.20   OJ:30   Min.   :0.500  
##  1st Qu.:13.07   VC:30   1st Qu.:0.500  
##  Median :19.25           Median :1.000  
##  Mean   :18.81           Mean   :1.167
##  3rd Qu.:25.27           3rd Qu.:2.000
##  Max.   :33.90           Max.   :2.000
```

As there is variability in the len (length) variable, the next step is to visually explore the data through a histogram or distribution.As we can see in the figure 1 (below), the histogram appears to be tri-modal, or having 3 populations seen by the three peak,s around 10, 18 and 26.

```
library(ggplot2)
ggplot(ToothGrowth, aes(x = len)) +
  geom_histogram(colour = "white", fill = "red", binwidth = 3) +
  labs(title = "Figure 1 - Distribution of teeth lengths",
       y = "Count",
       x = "Length")
```



The supp and dose variables are categorical variables, representing the Supplement type and Dosage of vitamin C in milligrams. If the supplement type and dosage really have an e???ect on the length of teeth, di???erences may be seen in the mean length between groups. Speci???cally, the tri-modal distribution of the data in Figure 4 suggests that the di???erence is mainly due to a variable with three categories: Dosage.

Looking at the di???erence in teeth lengths accross supplement types, there is an average di???erence of 3.7 in the mean between Orange Juice and Ascorbic Acid, which does not appear signi???cant at face value

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

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

For dosages, there is an average difference in length of 9.13 between dosages 0.5mg and 1.0mg, and an average difference of 6.365 between dosages 1.0mg and 2.0mg. The mean lengths of teeth for each dosage confirms the tri-modal distribution in Figure 1.

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

##   dose    len
## 1  0.5 10.605
## 2  1.0 19.735
## 3  2.0 26.100
```

To investigate whethere there were significant differences in teeth lengths between supplement types and dosages, hypothesis tests were conducted.

Hypothesis Testing

The differences above, while relatively large, may not be real, or statistically significant differences.To test whether the difference is statistically significant, t-tests are run between groups. The differences shown above suggest that the higher the dosage, the longer the teeth lengths; a one-tail test is appropriate for the following tests.

T-test to check if the difference is statistically significant for the type of supplement

A t-test is run between the different supplement types, Orange Juice (OJ) and Ascorbic Acid (VC).

```
 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
```

The resulting p-value of 0.06 is greater than 0.05, meaning that at the 95% confidence level, the difference in mean teeth length is not statistically significant.

T-test to check if the difference is statistically significant for the amount of dosage given

T-tests can also be conducted against the three di???erent dosages: 0.5mg, 1.0mg and 2.0mg. As there are three categories, subsets of the data are extracted to conduct the t-tests. Also, as the preliminary investigation suggests that the higher the dosage, the longer the teeth lengths. Thus, one-sided t-tests were conducted.

Lets first observe the result from the t-test between dosages 0.5mg and 1.0mg.The p-value from this test is 0.00, less than 0.01. Thus at the 99% con???dence level, there is a statistically signi???cant di???erence between dosages of 0.5mg and 1.0mg

```
ToothGrowth2 <- ToothGrowth[ToothGrowth$dose == c(0.5, 1.0),]
t.test(len ~ dose, data = ToothGrowth2, alternative = "less")

##
##  Welch Two Sample t-test
##
## data:  len by dose
## t = -4.4725, df = 17.976, p-value = 0.0001476
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##      -Inf -6.01238
## sample estimates:
## mean in group 0.5   mean in group 1
##             10.63             20.45
```

Now let us examine the result from the t-test between dosages 1.0mg and 2.0mg.The p-value from this test is 0.00, less than 0.01. Thus at the 99% con???dence level, there is a statistically signi???cant di???erence between dosages of 1.0mg and 2.0mg

```
ToothGrowth3 <- ToothGrowth[ToothGrowth$dose == c(1.0, 2.0),]
t.test(len ~ dose, data = ToothGrowth3, alternative = "less")

##
##  Welch Two Sample t-test
##
## data:  len by dose
## t = -3.6827, df = 17.949, p-value = 0.0008549
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf -3.671655
## sample estimates:
## mean in group 1 mean in group 2
##           19.02           25.96
```

Since there is a signi???cant increase in teeth length between 0.5mg and 1.0, and also a signi???cant increase in teeth length between 1.0 and 2.0, it is expected that there is also a signi???cant increase in teeth length between 0.5mg and 2.0. The t-test below substantiates this argument.

```
ToothGrowth4 <- ToothGrowth[ToothGrowth$dose == c(0.5, 2.0),]
t.test(len ~ dose, data = ToothGrowth4, alternative = "less")
```

```
##
##  Welch Two Sample t-test
##
## data:  len by dose
## t = -7.3335, df = 17.635, p-value = 4.681e-07
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf -11.70102
## sample estimates:
## mean in group 0.5   mean in group 2
##             10.63             25.96
```

The p-value from the t-test is 0.00, confirming that there is a statistically significant difference in teeth lengths between vitamin C dosages of 0.5mg and 2.0mg.

Conclusion

From the T-tests conducted in the previous section: 1. There is no significant difference in teeth lengths between the supplement type for Vitamin C. 2. There is a significant difference in teeth lengths between all different dosages of Vitamin C. The higher the dosage, the longer are the teeth lengths.