Statistical Inference Project - Part 2

Mariame M

September 11, 2014

In this project, we will analyze the ToothGrowth data from the R 'datasets' package. This data set contains 60 observations and the following variables:

- len (numeric): Tooth length - this represents the response variable;
- supp (factor): Supplement type (VC for ascorbic acid or OJ for orange juice) - this represents one explanatory variable;
- dose (numeric): Dose in milligrams - this represents one explanatory variable.

1. Load the ToothGrowth data and perform some basic exploratory data analyses:

We first load the 'datasets' library and display the dimensions of the ToothGrowth data, as well as the count of observations by supplement type and dosage level:

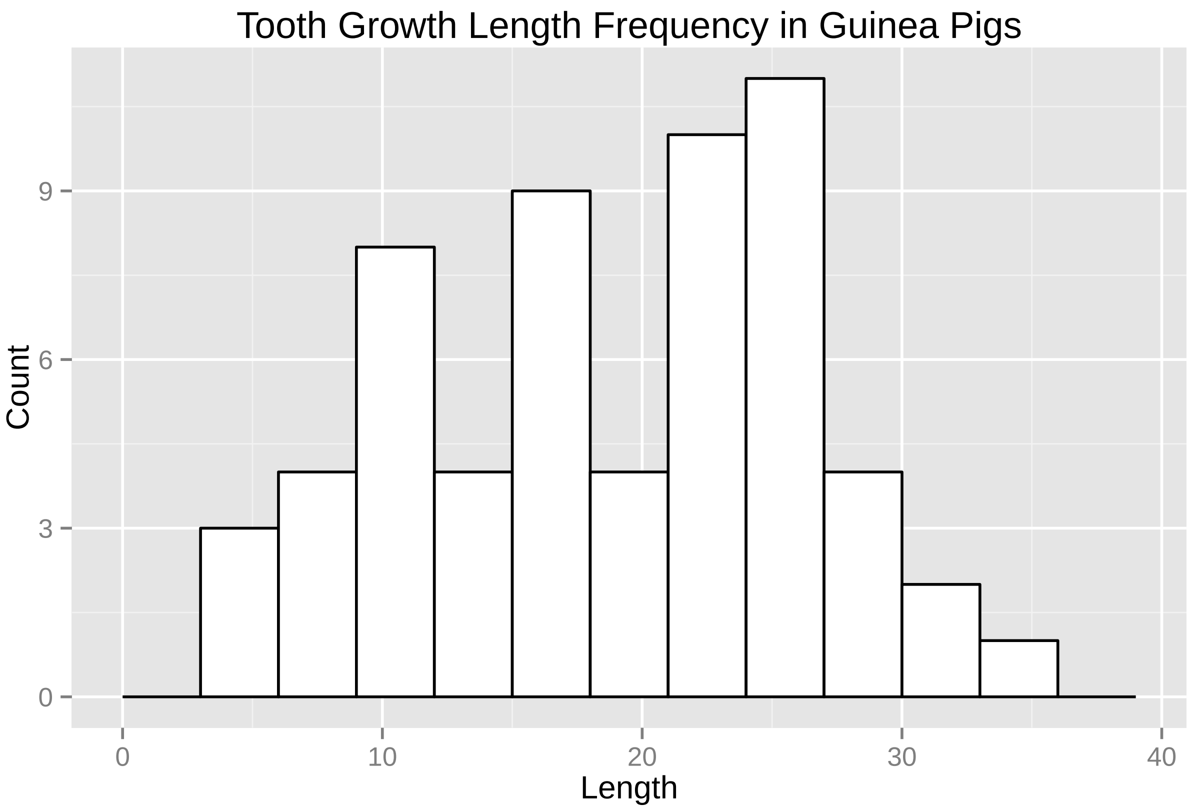
```
# load the required libraries
library(datasets)
library(ggplot2)

# show basic description of data frame
table(ToothGrowth$supp, ToothGrowth$dose)
```

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

We can then display a few exploratory data analysis plots to get a broad idea of the impact of supplement type and/or dosage level on tooth length in guinea pigs.

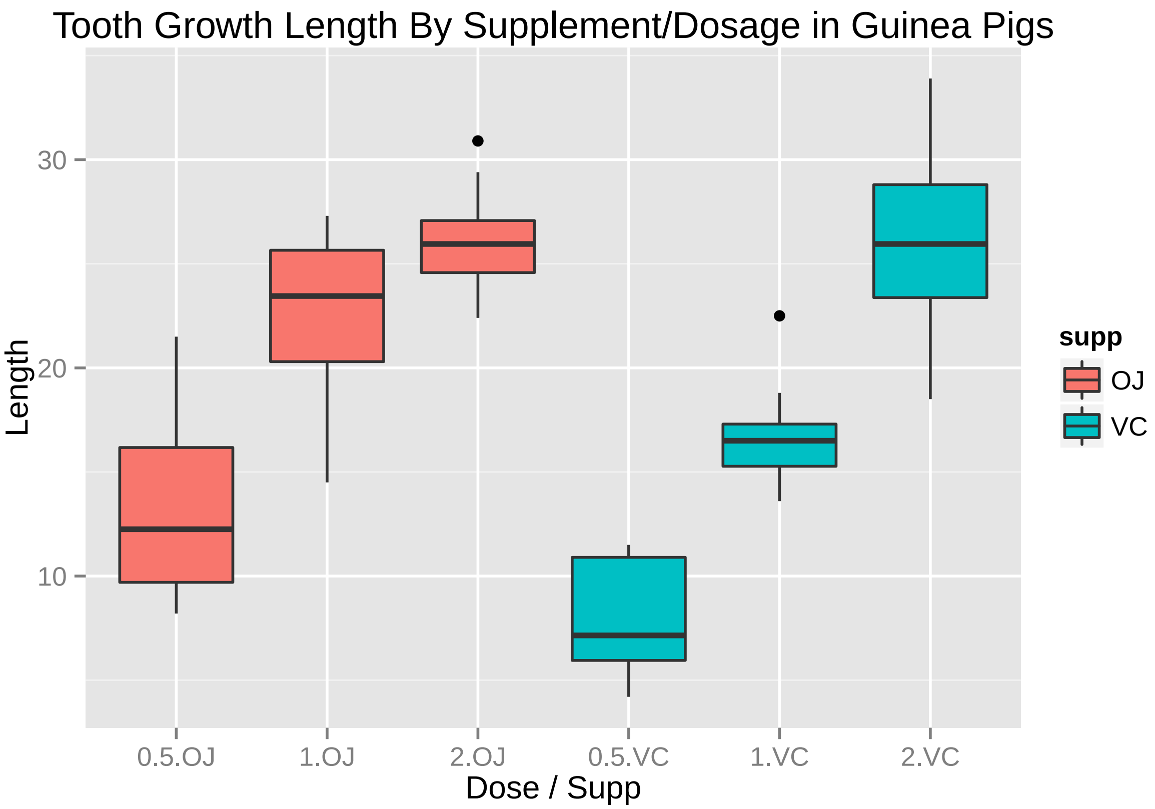
First, the histogram below shows the distribution of tooth length among all the guinea pigs in the experiement. We can see that our sample is not normally distributed, but we do not notice any skewness in our data. Assuming that guinea pigs were randomly drawn and assigned to the different groups in our experiment, we can use the Central Limit Theorem.

```
# histogram of the data
ggplot(ToothGrowth, aes(x=len)) +
    geom_histogram(binwidth=3, colour="black", fill="white") +
    ggtitle("Tooth Growth Length Frequency in Guinea Pigs") +
    xlab("Length") +
    ylab("Count")
```



We can then draw boxplots of the data by supplement type and dosage levels:

```r
# boxplots by dosage level and supplement type
ggplot(ToothGrowth, aes(x=interaction(dose,supp), y=len, fill=supp)) +
    geom_boxplot() +
    ggtitle("Tooth Growth Length By Supplement/Dosage in Guinea Pigs") +
    xlab("Dose / Supp") +
    ylab("Length") +
    scale_fill_discrete(breaks=c("OJ","VC"))
```



This plots gives us a first intuition that the supplement type OJ has a greater influence on tooth growth than the supplement type VC. However, it also seems that, as the dosage level gets higher, the difference between supplement types “slows down”.

2. Provide a basic summary of the data:

The below code provides a summary of the data:

```
summary(ToothGrowth)
```

```
##       len        supp         dose      
##  Min.   : 4.2   OJ:30   Min.   :0.50  
##  1st Qu.:13.1   VC:30   1st Qu.:0.50  
##  Median :19.2           Median :1.00  
##  Mean   :18.8           Mean   :1.17  
##  3rd Qu.:25.3           3rd Qu.:2.00  
##  Max.   :33.9           Max.   :2.00
```

While ‘dose’ is a numeric variable, the above summary is not meaningful for that variable, it can really only take one of three values: 0.5, 1.0 or 2.0. A workaround would be to convert that variable as a factor so that the summary would only show the possible values for that variable.

3. Use confidence intervals and hypothesis tests to compare tooth growth by supp and dose:

3.1. Test on Supplement Types Only:

First, ignoring the different dosages, we test the effect of the supplement type on tooth length.

Our null hypothesis H0 is that there is no difference in tooth growth between the group under VC supplement and the group under OJ supplement.
Our alternative hypothesis Ha is that there is indeed a difference between the group under VC supplement and the group under OJ supplement, or in other words the type of supplement has an incidence on tooth growth.

Even though our sample size is not small (> 30), we can use the t.test() function in R to compute the confidence interval for our hypothesis (as the T-distribution gets closer to a Z-distribution as the sample size increases):

```r
# test by supplement
t.test(len ~ supp, paired=FALSE, var.equal = FALSE, data = ToothGrowth)
```

```
##
##  Welch Two Sample t-test
##
## data:  len by supp
## t = 1.915, df = 55.31, p-value = 0.06063
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.171  7.571
## sample estimates:
## mean in group OJ mean in group VC
##            20.66            16.96
```

With a confidence interval of [-0.171, 7.571] for the difference in tooth growth average by supplement type (i.e. mean for the group in OJ minus mean for the group in VC), we *cannot* reject the null hypothesis that there is not a significant difference in tooth length between the two supplement types, as zero is included in the confidence interval. Therefore we conclude that the difference in tooth growth between the OJ guinea pigs and the VC guinea pigs can be due to chance.

3.2. Test on Dosage Levels Only:

In this second test, we ignore the type of supplement and only account for the dosage levels. To that purpose, we need to create three separate data sets to compare dosage levels two-by-two:

- one data set testing for the difference in tooth growth between the 0.5mg vs 1.0mg dosages;
- one data set testing for the difference in tooth growth between the 0.5mg vs 2.0mg dosages;
- one data set testing for the difference in tooth growth between the 1.0mg vs 2.0mg dosages.

For each sub-test above, our null hypothesis H0 is that there is no significant difference in the impact on tooth growth between the two dosages and, the alternative hypothesis Ha is that there is indeed a difference in the impact on tooth growth between the two dosages.

Below, we report the confidence intervals for the three sub-tests using T-intervals:

```r
# create 3 data frames
dose05_10 <- subset(ToothGrowth, dose %in% c(0.5, 1.0))
dose05_20 <- subset(ToothGrowth, dose %in% c(0.5, 2.0))
dose10_20 <- subset(ToothGrowth, dose %in% c(1.0, 2.0))

# display the confidence intervals for the different groups
rbind(
  t.test(len ~ dose, paired = FALSE, var.equal = FALSE, data = dose05_10)$conf,
  t.test(len ~ dose, paired = FALSE, var.equal = FALSE, data = dose05_20)$conf,
  t.test(len ~ dose, paired = FALSE, var.equal = FALSE, data = dose10_20)$conf
)
```

```
##         [,1]    [,2]
## [1,] -11.984  -6.276
## [2,] -18.156 -12.834
## [3,]  -8.996  -3.734
```

Based on the above results, we infer that:

- for 0.5mg vs 1.0mg dosages, with a confidence interval of [-11.98, -6.28], we can reject the null hypothesis in favor of the alternative hypothesis and, conclude that there is a significant difference in tooth growth between the two dosage levels.
- for 0.5mg vs 2.0mg dosages, with a confidence interval of [-18.16, -12.83], we can reject the null hypothesis in favor of the alternative hypothesis and, conclude that there is a significant difference in tooth growth between the two dosage levels.
- for 1.0mg vs 2.0mg dosages, with a confidence interval of [-9.00, -3.73], we can reject the null hypothesis in favor of the alternative hypothesis and, conclude that there is a significant difference in tooth growth between the two dosage levels.

3.3. Test on Supplement Types and Dosage Levels:

Finally, we test each dosage level separately to see if for a given dosage level, a significant difference in tooth growth between the two supplement types can be observed. This test allows us to compare supplement types while holding the dosage level constant.

Here again, we need to first create three separate data sets to compare supplement types across dosage levels:

- one data set containing data for VC and OJ supplement at the 0.5mg dosage level;
- one data set containing data for VC and OJ supplement at the 1.0mg dosage level;
- one data set containing data for VC and OJ supplement at the 2.0mg dosage level.

For each sub-test above, our null hypothesis H0 is that there is no significant difference in the impact on tooth growth between the two supplement types at the given dosage level and, the alternative hypothesis Ha is that there is indeed a difference in the impact on tooth growth between the two supplement types at the given dosage level.

Below, we report the confidence intervals for the three sub-tests using T-intervals:

```r
# create 3 data frames
dose05 <- subset(ToothGrowth, dose == 0.5)
dose10 <- subset(ToothGrowth, dose == 1.0)
dose20 <- subset(ToothGrowth, dose == 2.0)
```

```r
# display the confidence intervals for the different groups
rbind(
t.test(len ~ supp, paired = FALSE, var.equal = FALSE, data = dose05)$conf,
t.test(len ~ supp, paired = FALSE, var.equal = FALSE, data = dose10)$conf,
t.test(len ~ supp, paired = FALSE, var.equal = FALSE, data = dose20)$conf
)
```

```
##        [,1]  [,2]
## [1,]  1.719 8.781
## [2,]  2.802 9.058
## [3,] -3.798 3.638
```

Based on the above results, we infer that:

- at the dosage level 0.5mg, with a confidence interval of [1.72, 8,78], we can reject the null hypothesis in favor of the alternative hypothesis and, conclude that there is a significant difference in tooth length between the two supplement types.
- at the dosage level 1.0mg, with a confidence interval of [2.80, 9.06], we can reject the null hypothesis in favor of the alternative hypothesis and, conclude that there is a significant difference in tooth length between the two supplement types.
- at the dosage level 2.0mg, with a confidence interval of [-3.80, 3.64], we *cannot* reject the null hypothesis in favor of the alternative hypothesis and, conclude that there is *no* significant difference in tooth length between the two supplement types.

This means that for low dosage levels, we do see a difference in tooth growth average (i.e. mean for the group in OJ minus mean for the group in VC) by supplement type, and that as the dosage level gets higher that difference gets smaller.

4. State the assumptions and the conclusions:

Below are the assumptions we used in this study:

- each observation is independent from the other, meaning random assignment was used to assign the guinea pigs to each supplement type and to each dosage level - this is an important assumption as this allows us to conclude that any difference observed is due to one of the explanatory variable, and not genetic for example if we were using guinea pigs that are related;
- the sampling distribution of average differences in tooth growth between supplement types and/or dosage levels are nearly normal;
- the variances between the different groups (group VC vs group OJ and, group 0.5mg vs group 1.0mg vs gorup 2.0mg) are different (that is the reason why we used var.equal = FALSE in our tests).

Based on the above assumptions and the results observed in the experiment, we conclude that:

1. There is no significant difference in tooth growth between the VC and the OJ supplement types, when mixing the different dosage levels.
2. There is, however, a significant difference in tooth growth between the different dosage levels, with the tooth growth being higher as dosage levels increase.
3. When accounting for both supplement types and dosage levels, we can then see a significant difference for the lower dosage levels, with the OJ supplement type having a greater impact on tooth growth than the VC supplement type. When reaching a certain level in dosage (2.0mg in our experiment) however, that difference tends to disappear.

Therefore, we infer that below a certain level (< 2.0mg), supplement type is more important than dosage level when looking to increase the tooth length growth in guinea pigs. At (and most liely above) that level, it does not matter the type of supplement given to the guinea pigs.