Statistical Inference Course Project - Inferential Data Analysis

Naman Gandhi

February 4, 2017

Overview

In this project we'll analyze the ToothGrowth data in the R datasets package.

We'll break it into 4 broad sections -
1) **Preparation and Data Summary**
2) **Exploratory Data Analysis**
3) **Inferential Analysis**
4) **Assumptions and Conclusion**

Let's dive in!!

Section 1: Preparation and Data Summary

This data analyses the effect of Vitamin C on Tooth Growth in Guinea Pigs

```
# Loading the required libraries and dataset
library(datasets)
library(ggplot2)
data("ToothGrowth")
# Understanding the structure and dispersion of data
str(ToothGrowth)
```

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

```
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
```

```
# Converting the dose to be factor to analyze it better
ToothGrowth$dose<-as.factor(ToothGrowth$dose)
```

Insights
Each of 60 animals 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).

Section 2: Exploratory Data Analysis

Figure 1 Explains how the length of the tooth is affected by supplements, OJ stands for Orange Juice and VC - (Vit C) Ascorbic Acid

```
ggplot(aes(x=supp, y=len), data=ToothGrowth) + geom_boxplot(aes(fill=supp))+ geom_point() +
    xlab("Supplement type") +ylab("Tooth Length") +
    ggtitle("Figure 1: Tooth Length summarized by supplement")
```

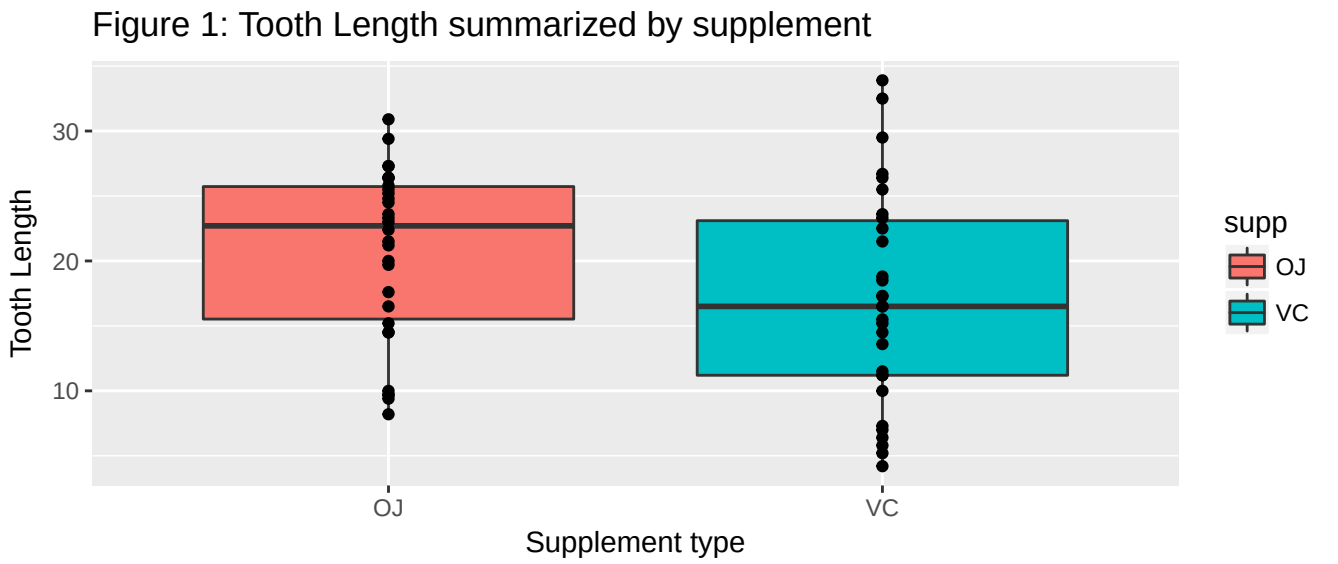
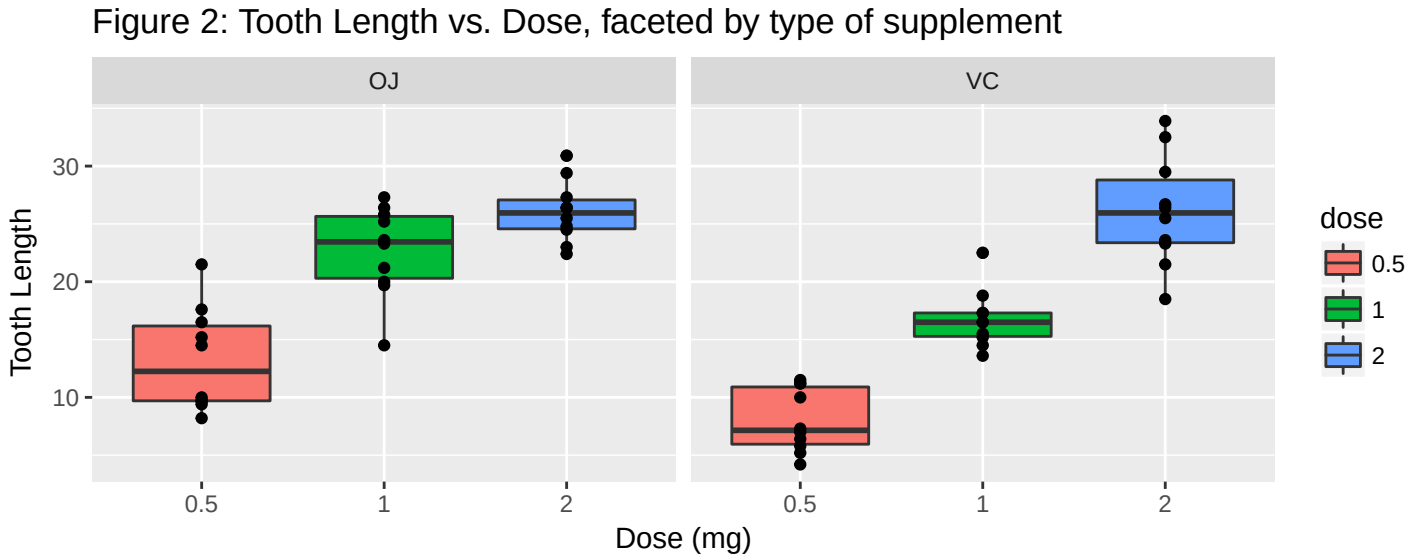


Figure 2 Explains how the length of the tooth grows as the dosage of the supplement increases, OJ stands for Orange Juice and VC - (Vit C) Ascorbic Acid

```
ggplot(data = ToothGrowth, aes(x = dose, y = len)) + geom_boxplot(aes(fill = dose)) +
    facet_grid(~ supp) + geom_point(aes(colour = I("black"))) + xlab("Dose (mg)") +
    ylab("Tooth Length") + ggtitle("Figure 2: Tooth Length vs. Dose, faceted by type of supplement")
```



Section 3: Inferential Analysis

We'll test the significance by performing hypothesis testing in two phases -

3A) Testing the significance of correlation between delivery method and tooth growth

```
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
```

Insights -
i) The p-value of this test is 0.06 which is greater than significance level of 0.05
ii) 95% onfidence interval contains zero
=> *Null Hypothesis Accepted - Supplement types seems to have no impact on Tooth growth*

3B) Testing the significance of correlation between dosage administered and tooth growth

```
# 3B1 - Dose amounts 0.5 and 1.0
t.test(len~dose,data=subset(ToothGrowth, ToothGrowth$dose %in% c(0.5,1.0)))
```

```
##
##  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
```

```
# 3B2 - Dose amounts 1.0 and 2.0
t.test(len~dose,data=subset(ToothGrowth, ToothGrowth$dose %in% c(1.0,2.0)))
```

```
##
##  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
```

```
# 3B3 - Dose amounts 0.5 and 2.0
t.test(len~dose,data=subset(ToothGrowth, ToothGrowth$dose %in% c(0.5,2.0)))
```

```
##
##  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
```

Insights -
i) The p-value of each of the three tests (3B1, 3B2 & 3B3) test is 0, which is less than significance level of 0.05
ii) 95% onfidence interval does not contains zero (-ve in all cases)
=> *Null Hypothesis Rejected - Tooth length increases with an inceasing dose*

Section 4: Assumptions and Conclusion

Assumptions
i) The distribution of the sample means follows the Central Limit Theorem
ii) The sample is representative of the population

Conclusion
Analyzing the results from the T-tests, we can conclude that supplement delivery method has no effect on tooth growth, however increased dosages do result in increased tooth length of Guinea pigs.