Coursera Statistical Inference Project

Jiajun Zhang

October 12, 2019

```
library(ggplot2)
library(dplyr)
```

Part 2: Inferential Data Analysis

Data Loading and Basic Exploratory Data analysis

- First, we will load the `ToothGrowth` data directly since it is stored inside the R library.
- The dataset is about the effect of vitamin C on tooth growth in guinea pig. This dataset has 60 observations on 3 variables.

```
data(ToothGrowth)
dim(ToothGrowth)
```

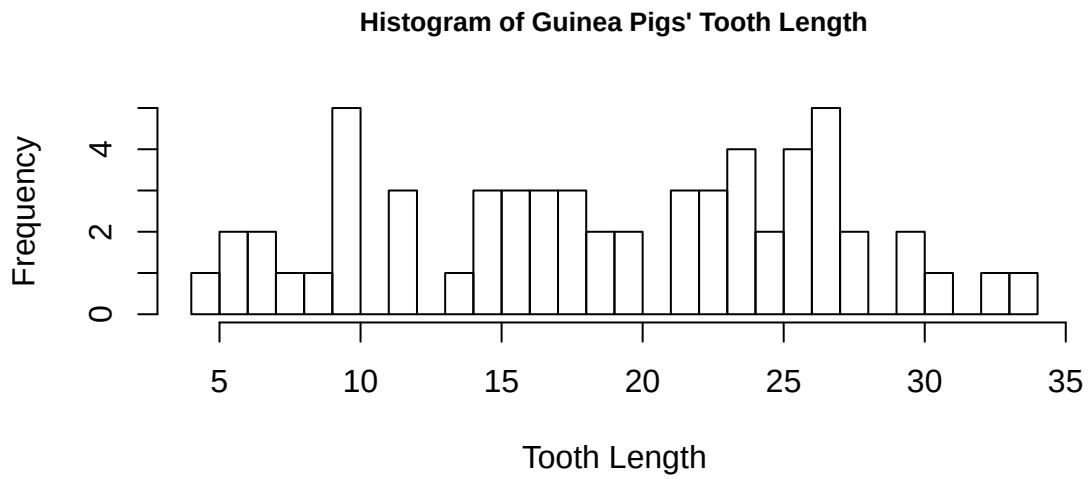
```
## [1] 60  3
```

```
head(ToothGrowth)
```

```
##    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
```

- Our response is the `len` variable, and we can make a histogram to see how they distributed.

```
hist(ToothGrowth$len, breaks=30, xlab="Tooth Length",
     main="Histogram of Guinea Pigs' Tooth Length", cex.main=0.8)
```



Basic Summary of the 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
```

```
sd(ToothGrowth$len)
```

```
## [1] 7.649315
```

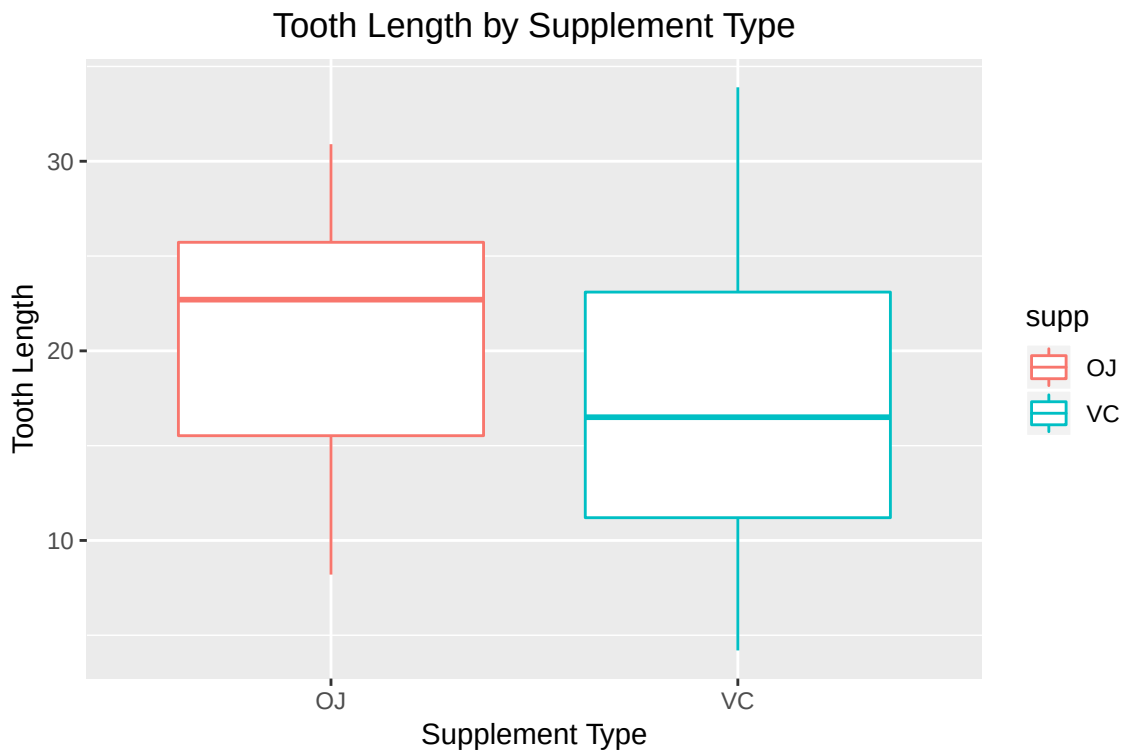
```
table(ToothGrowth$dose)
```

```
##
## 0.5   1   2
##  20  20  20
```

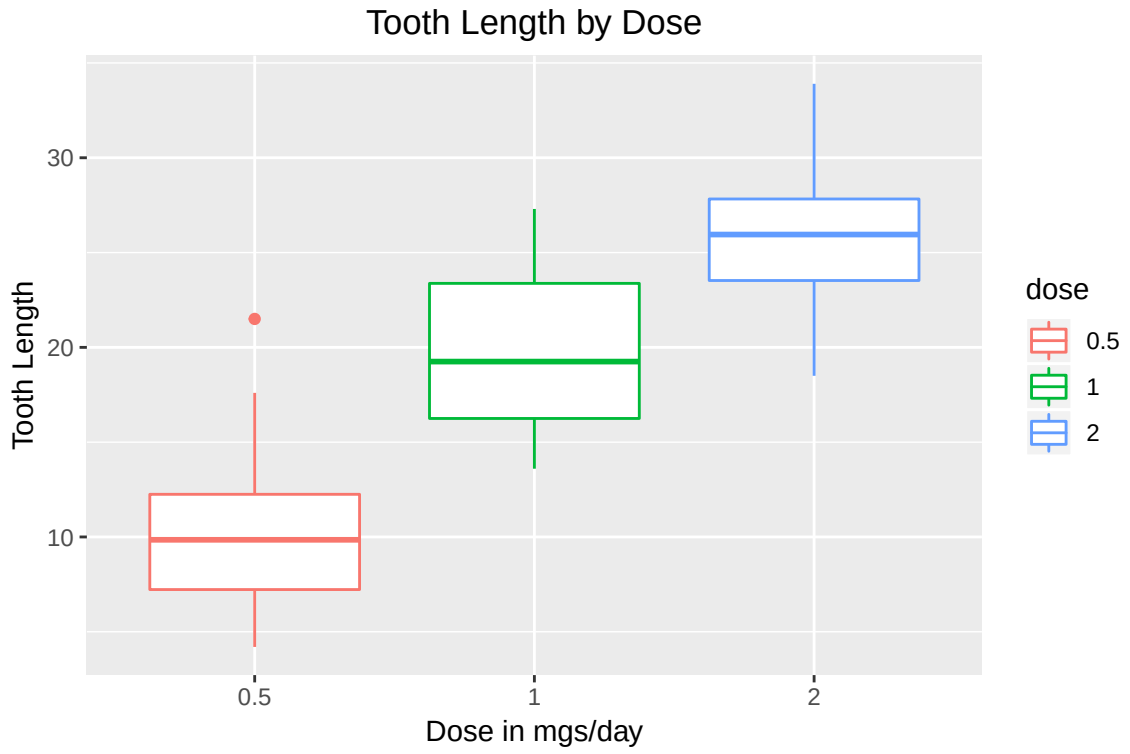
```
ToothGrowth$dose <- as.factor(ToothGrowth$dose)
```

- The variables in this dataset are `len`, `supp`, and `dose`, which stand for tooth length, supplement type, and dose in milligrams per day. The variables `len` and `dose` are numeric, and `supp` is a factor with 2 levels, VC or OJ.
- From the summary of the data, we can see that the mean of our response `len` is 18.81 and standard deviation is 7.65.
- Also, we shall see that the variable `dose` has only 3 unique values: 0.5, 1, and 2. So, we can convert it to a factor.
- Then, we can produce some graphs of the response `len` based on `supp` and `dose`.

```
ggplot(ToothGrowth, aes(supp, len, color=supp)) + geom_boxplot() +
      labs(x="Supplement Type", y="Tooth Length",
           title="Tooth Length by Supplement Type") +
      theme(plot.title=element_text(hjust=0.5))
```




```
ggplot(ToothGrowth, aes(dose, len, color=dose)) + geom_boxplot() +
      labs(x="Dose in mgs/day", y="Tooth Length",
           title="Tooth Length by Dose") +
      theme(plot.title=element_text(hjust=0.5))
```



Hypothesis Tests to Compare Tooth Growth By Supp and Dose

- From the previous plots we can sort of see some differences in tooth length between doses and a slightly difference in tooth length between 2 supplement types.
- Now, we can perform two sample t-tests on supplements and doses by stating our hypotheses as:
  H_0: difference in means equal to 0
  H_a: difference in means is not equal to 0

```
VC <- filter(ToothGrowth, supp=="VC")
OJ <- filter(ToothGrowth, supp=="OJ")
t.test(VC$len, OJ$len)
```

```
##
##  Welch Two Sample t-test
##
## data:  VC$len and OJ$len
## 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:
##  -7.5710156  0.1710156
## sample estimates:
## mean of x mean of y
##  16.96333  20.66333
```

- The result suggests a p-value of 0.06 which is slightly greater than 0.05. Therefore, we can say that there is a slightly significant evidence to indicate that the means between two supplement types are the same.

```
Dose0.5 <- filter(ToothGrowth, dose==0.5)
Dose1.0 <- filter(ToothGrowth, dose==1)
t.test(Dose0.5$len, Dose1.0$len)
```

```
##
##  Welch Two Sample t-test
##
## data:  Dose0.5$len and Dose1.0$len
## 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 of x mean of y
##    10.605    19.735
```

```
Dose2.0 <- filter(ToothGrowth, dose==2)
t.test(Dose0.5$len, Dose2.0$len)
```

```
##
##  Welch Two Sample t-test
##
## data:  Dose0.5$len and Dose2.0$len
## 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 of x mean of y
```

```
##    10.605    26.100
```

```
t.test(Dose1.0$len, Dose2.0$len)
```

```
##
##  Welch Two Sample t-test
##
## data:  Dose1.0$len and Dose2.0$len
## 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 of x mean of y
##    19.735    26.100
```

- Since the p-values from above are all smaller than the significant value 0.05, we will reject H_0 and conclude that there are significant evidences to indicate that the means between dose 0.5 and dose 1, dose 0.5 and dose 2, and dose 1 and dose 2 are different.

Conclusions

- By doing hypothesis tests, we can conclude that there is a significant evidence to say that the means of guinea pigs' tooth length between 2 supplement types of guinea pigs are the same. However, this evidence is not obvious since the p-value we obtained is 0.06 which is slightly greater than 0.05.
- On the other hand, the tooth length differences in means between 3 unique doses are significant.
- However, some assumptions we need to state before getting this conclusion and they are independence and normality. Independence states that our sample data must be randomly selected with biases. And normality assumes that the population distribution must be normal.