Basic Inferential Data Analysis

Written by: Andrew Leonard

John Hopkins Data Science Specialization: Statistical Inference

Synopsis

We will use the Tooth Growth data set from the R datasets package to explore the question: "Do different suppliments and/or dosage affect tooth growth in guinea pigs?"

The data set looks at two suppliments: orange juice and vitamin C and different dosage levels of each.

Dataset variables: len: Tooth length supp: Supplement type (Vitamin C or Orange Juice) dose: Dose in milligrams

Summary Analysis

While the sampele size is low, the data appears resembles a normal distribution. With 30 samples per supp, we will use t-tests for our hypothesis testing.

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

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

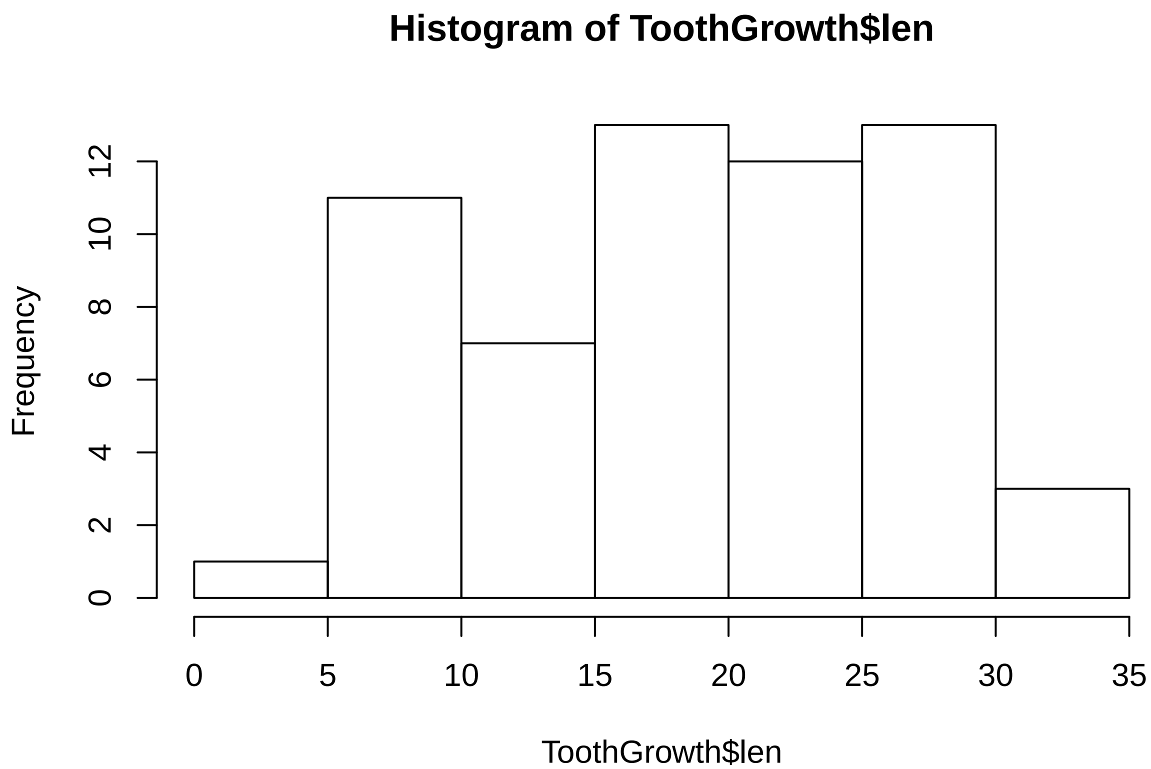
```
data("ToothGrowth")
summary(ToothGrowth)
```

```
##       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
```

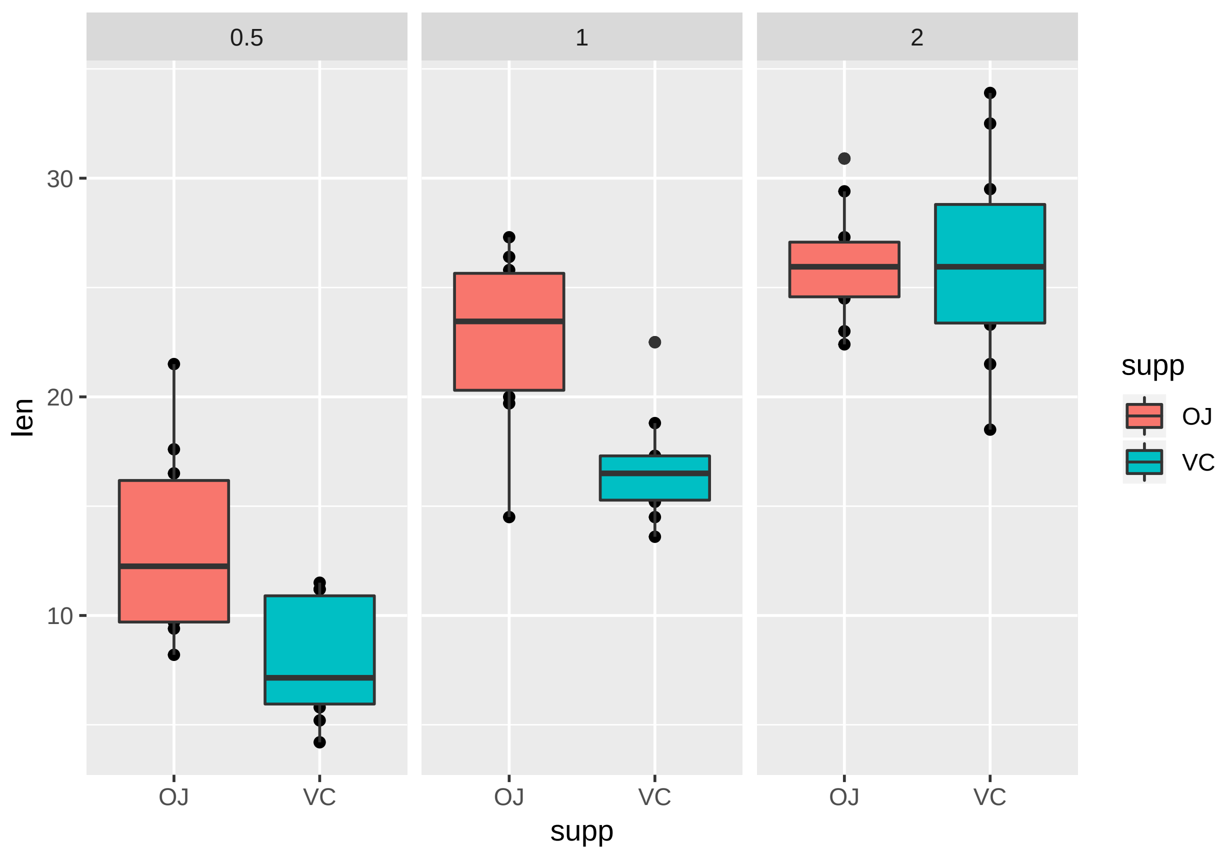
```
dim(ToothGrowth)
```

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

```
hist(ToothGrowth$len)
```




```r
qplot(supp, len, data = ToothGrowth, facets = ~dose)+
    geom_boxplot(aes(fill = supp))
```



Analysis Part 1: Tooth Length and Supplement Type

First, we will test if the supplement type has an impact on tooth growth.

For our test h0 is: OJ = VC, h1 is: OJ != VC

```
#group by suppliment
supp <- group_by(ToothGrowth, supp)
bysupp <- summarise(supp, mean(len), sd(len))
names(bysupp) <- c("supp", "mean", "sd")

muOJ <- bysupp$mean[1]
muVC <- bysupp$mean[2]
sdOJ <- bysupp$sd[1]
sdVC <- bysupp$sd[2]
n <- 30

se <- sqrt((sdOJ^2/n) + (sdVC^2/n))

tstat <- (muOJ - muVC)/se

#95% confidence interval
t <- qt(.975, 58)
(muOJ - muVC) + c(-1,1)*t*se
```

```
## [1] -0.1670064  7.5670064
```

```
#pvalue
pv <- pt(-abs(tstat), n-2)*2
```

Our t statistic is 1.9152683, which is lower than the 95% two tail confidence t statistic of 2.0017175. This indicates we should accept the null hypothesis. To confirm this is correct, we can look at our p-value. Our p-value is 0.0657236, which is higher than .05 at 95% confidence, confirming we cannot reject the null hypothesis. Thus, we cannot infer that there is a statistical difference in tooth growth dependent on suppliment type at a 95% confidence interval.

Part 2: Tooth Length and Dosage

For the second portion, we will look at dosage. We will compare a half dose to a double dose for this test.

For our test h0 is: halfdose = doubledose, h1 is:halfdose < doubledose

```
#group by suppliment
dose <- group_by(ToothGrowth, dose)
bydose <- summarise(dose, mean(len), sd(len))
names(bydose) <- c("supp", "mean", "sd")

muhalf <- bydose$mean[1]
mudouble <- bydose$mean[3]
sdhalf <- bydose$sd[1]
sddouble <- bydose$sd[3]
n <- 20

se <- sqrt((sdhalf^2/n) + (sddouble^2/n))

tstat <- (muhalf - mudouble)/se
```

```r
#95% confidence interval
t <- qt(.95, n-2)
(muhalf-mudouble)+(t*se)
```

```
## [1] -13.21776
```

```r
#pvalue
pv <- pt(-abs(tstat), n-2)
```

Our t statistic is -11.799046, which is below our 95% one tail confidence t statistic of 1.7340636. This indicates we should reject the null hypothesis. To confirm this is correct, we can look at our p-value. Our p-value is $3.3099796 \times 10^{-10}$, which is much lower than .05 at 95% confidence, confirming we can reject the null hypothesis and infer that a double dose has a larger impact on tooth growth than a half a dose.