Basic Inferential Data Analysis

Writen 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 are len = Tooth length, supp = Supplement type (Vitamin C or Orange Juice), dose = Dose in milligrams

Summary Analysis

The data appears to resemble a normal distribution as confimred by shapiro test p-value above .05. With 30 samples per supp, we will use t-tests for our hypothesis testing.

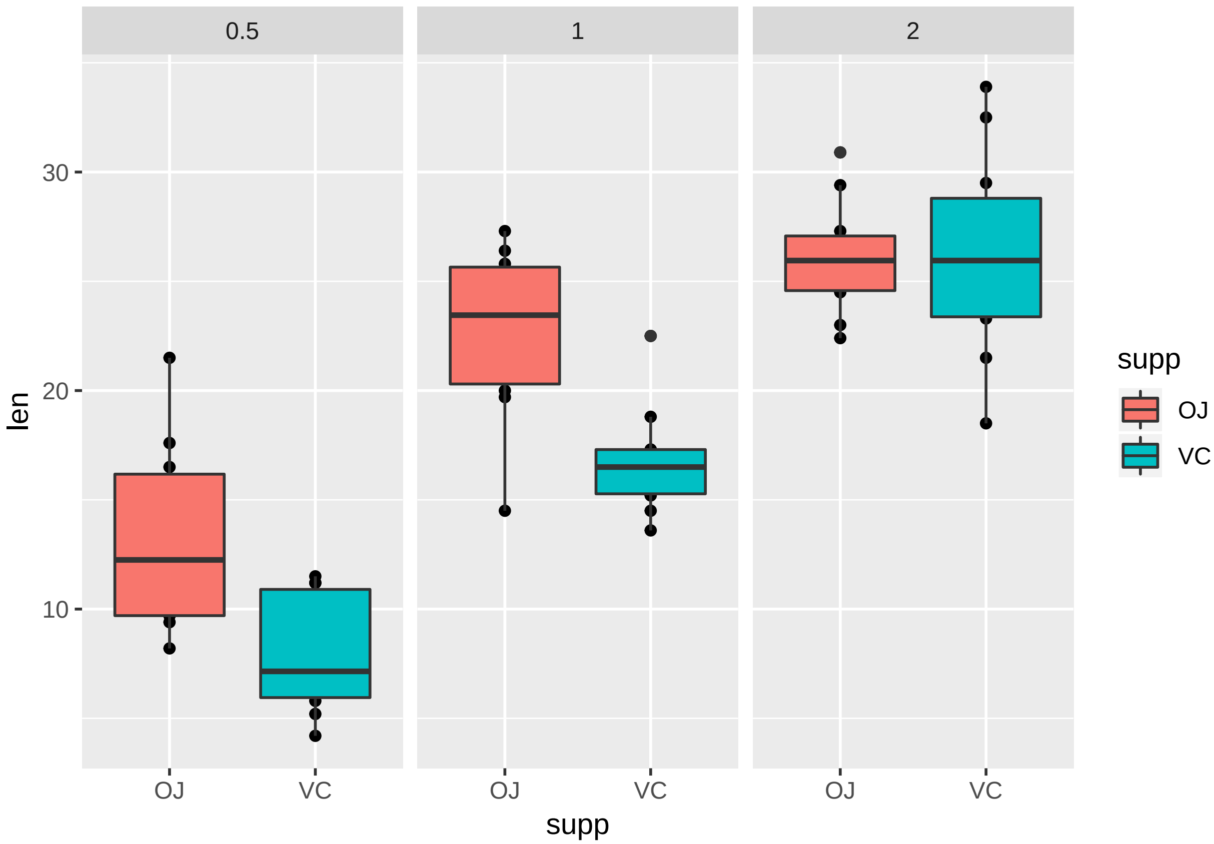
```
library(ggplot2)
library(dplyr)
data("ToothGrowth")
summary(ToothGrowth)
```

```
##       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
```

```
shapiro.test(ToothGrowth$len)
```

```
## 
##  Shapiro-Wilk normality test
## 
## data:  ToothGrowth$len
## W = 0.96743, p-value = 0.1091
```

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



It appears tooth length increases with higher dosage of both VC and OJ. It's unclear from plotting how tooth length is affected by supplement type (VC or OJ). We will test both below.

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
tstat
```

```
## [1] 1.915268
```

```
#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
pv
```

```
## [1] 0.06572358
```

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 significant difference in tooth growth by suppliment type at a 95% confidence interval.

Part 2: Tooth Length and Dosage

For the second portion, we will look at impact of dosage on tooth length. 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
tstat
```

```
## [1] -11.79905
```

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

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

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

```
## [1] 3.30998e-10
```

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 tooth length for a double dose is higher than tooth length at half a dose for guinea pigs.