

Statistical Inference Class Project

Cathy Snell

October 10, 2018

Part 2: Basic Inferential Data Analysis

For part 2, we will analyze the ToothGrowth data in the R datasets package. This data set shows the effect of Vitamin C on tooth growth in guinea pigs.

Exploritory Data Analysis

First we load the ToothGrowth data and perform some basic exploratory data analyses.

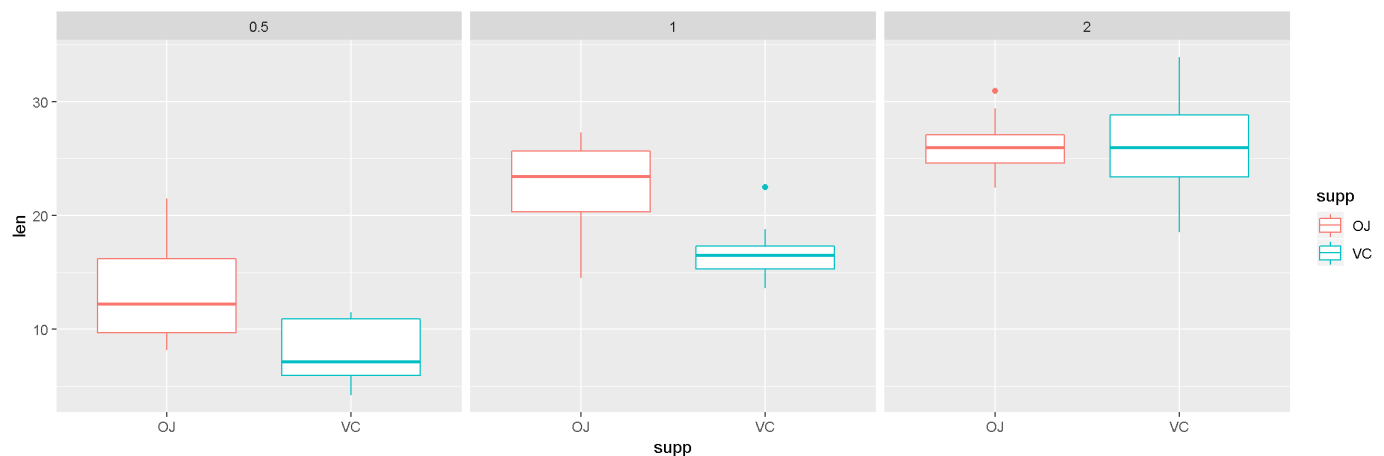
```
## 'data.frame': 60 obs. of 3 variables:
## $ len : num 4.2 11.5 7.3 5.8 6.4 10 11.2 11.2 5.2 7 ...
## $ supp: Factor w/ 2 levels "OJ","VC": 2 2 2 2 2 2 2 2 2 2 ...
## $ dose: num 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 ...
```

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

We see that there are 60 observations of 3 variables. There are 2 levels of supp and 3 distinct values of dose. There are 10 observations for each combination of supp and dose.

Compare Tooth Growth

We will use confidence intervals and hypothesis testing (p-values) to compare tooth growth by supp and dose. Let's start with a visual of the data. We will do a hypothesis test to determine if there is significant difference between the two supp types at each dose level.



The null hypothesis (H_0) will be that $\text{mean}(\text{OJ}) = \text{mean}(\text{VC})$ for each dose.

The alternative hypothesis (H_a) will be that $\text{mean}(\text{OJ}) > \text{mean}(\text{VC})$ for each dose.

Using the confidence interval calculation for comparing groups with unequal variances, means we will need to calculate the pooled standard error and the degrees of freedom. We'll compare the manual calculation with the R `t.test` command.

```
## [1] "Manual Calculation Confidence Interval: 2.34604034665748 , 8.15395965334252"
```

```
##
## Welch Two Sample t-test
##
## data:  ToothGrowth %>% filter(supp == "OJ", dose == "0.5") %>% pull(len) and ToothGrowth %>% filter(supp
== "VC", dose == "0.5") %>% pull(len)
## t = 3.1697, df = 14.969, p-value = 0.003179
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  2.34604      Inf
## sample estimates:
## mean of x mean of y
##      13.23      7.98
```

Let's quickly look at the confidence intervals for the other doses.

```
## [1] "Dose = 1 Confidence Interval: c(3.35615763634793, Inf) / p-value: 0.000519187936149939"
```

```
## [1] "Dose = 2 Confidence Interval: c(-3.13349957439956, Inf) / p-value: 0.518074205638314"
```

Conclusion

At dose = .5, there is a significant difference between OJ and VC. We know this because the 95% confidence interval does not include 0 (meaning it's less than 5% likely to have the means be equal), and the p-value is .003, which is less than .05.

At dose = 1, there is a significant difference between OJ and VC. Again, the confidence interval does not include 0, and the p-value is .0005 (well under .05).

At dose = 2, we can not reject the null hypothesis. The confidence interval includes 0, and the p-value is > .05.

These calculations are evidence of what we visually see in the plot of the differences between OJ and VC.

Appendix

Code

```

library(tidyverse)
library(ggplot2)
library(gridExtra)

# Part 2: Basic Inferential Data Analysis

## Exploratory Data Analysis

# load data
data(ToothGrowth)
str(ToothGrowth)
table(ToothGrowth$dose, ToothGrowth$supp)

## Compare Tooth Growth

ggplot(ToothGrowth, aes(supp, len)) +
  geom_boxplot(aes(colour=supp)) +
  facet_grid(.~dose)

mean.OJ.0.5 <- mean(ToothGrowth%>%filter(supp=="OJ",dose=="0.5")%>%pull(len))
mean.VC.0.5 <- mean(ToothGrowth%>%filter(supp=="VC",dose=="0.5")%>%pull(len))
sd.OJ.0.5 <- sd(ToothGrowth%>%filter(supp=="OJ",dose=="0.5")%>%pull(len))
sd.VC.0.5 <- sd(ToothGrowth%>%filter(supp=="VC",dose=="0.5")%>%pull(len))

sp = sqrt((sd.OJ.0.5^2)/10 + (sd.VC.0.5^2)/10) # pooled standard error
sx = (sd.OJ.0.5^2)/10
sy = (sd.VC.0.5^2)/10
df = (sx+sy)^2/(sx^2/9 + sy^2/9) # degrees of freedom

paste("Manual Calculation Confidence Interval:",
      (mean.OJ.0.5-mean.VC.0.5)+-1*qt(.95,df)*sp, # confidence interval for unequal variances
      ", ",
      (mean.OJ.0.5-mean.VC.0.5)+qt(.95,df)*sp)

t.test(ToothGrowth%>%filter(supp=="OJ", dose=="0.5")%>%pull(len),
       ToothGrowth%>%filter(supp=="VC", dose=="0.5")%>%pull(len),
       paired=FALSE, var.equal=FALSE, alternative="greater", conf.level=.95) # t test for unequal variances

paste("Dose = 1 Confidence Interval:",
      t.test(ToothGrowth%>%filter(supp=="OJ", dose=="1")%>%pull(len),
             ToothGrowth%>%filter(supp=="VC", dose=="1")%>%pull(len),
             paired=FALSE, var.equal=FALSE, alternative="greater", conf.level=.95)[4],
      " / p-value:",
      t.test(ToothGrowth%>%filter(supp=="OJ", dose=="1")%>%pull(len),
             ToothGrowth%>%filter(supp=="VC", dose=="1")%>%pull(len),
             paired=FALSE, var.equal=FALSE, alternative="greater", conf.level=.95)[3]) # t test for dose = 1

paste("Dose = 2 Confidence Interval:",
      t.test(ToothGrowth%>%filter(supp=="OJ", dose=="2")%>%pull(len),
             ToothGrowth%>%filter(supp=="VC", dose=="2")%>%pull(len),
             paired=FALSE, var.equal=FALSE, alternative="greater", conf.level=.95)[4],
      " / p-value:",
      t.test(ToothGrowth%>%filter(supp=="OJ", dose=="2")%>%pull(len),
             ToothGrowth%>%filter(supp=="VC", dose=="2")%>%pull(len),
             paired=FALSE, var.equal=FALSE, alternative="greater", conf.level=.95)[3]) # t test for dose = 2

## Conclusion

```