

Coursera Data Science Project: Statistical Inference (Part 2)

Rob Rolleston

August 23, 2015

Introduction

This is the project for the statistical inference class. In it, I will use simulation to explore inference and do some simple inferential data analysis. The project consists of two parts:

1. A simulation exercise.
2. Basic inferential data analysis (this report)

Basic Inferential Data Analysis

tbd

Load and explore data

```
library(datasets)
data("ToothGrowth")
ToothGrowth_tbl <- tbl_df(ToothGrowth)
glimpse(ToothGrowth_tbl)
```

```
## Observations: 60
## Variables:
## $ len   (dbl) 4.2, 11.5, 7.3, 5.8, 6.4, 10.0, 11.2, 11.2, 5.2, 7.0, 16....
## $ supp  (fctr) VC, VC, VC, VC, VC, VC, VC, VC, VC, VC, VC, VC, VC, VC, ...
## $ dose  (dbl) 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 1.0, 1....
```

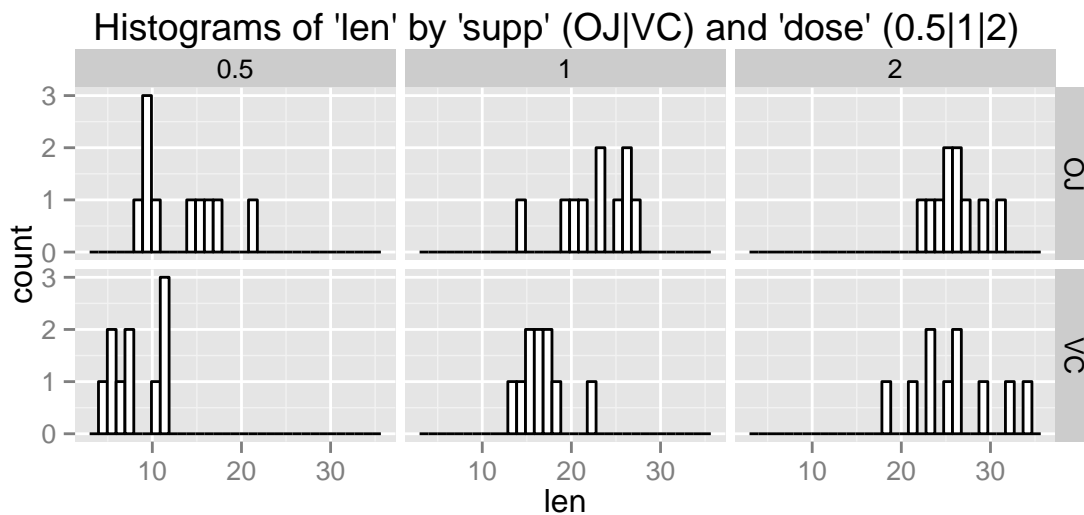
Some inspection of the ToothGrowth data indicates it has 60 observations of 3 values: len, supp, dose. The ‘supp’ value is a factor with only 2 levels: OJ, VC. The ‘dose’ value, although a number, actually has only 3 values: 0.5, 1, 2. For processing, this value was converted to a factor.

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

Data Summary

A faceted set of histograms

```
ggplot(ToothGrowth_tbl, aes(x=len)) +
  geom_histogram(color="black", fill="white") +
  facet_grid(supp ~ dose) +
  ggtitle("Histograms of 'len' by 'supp' (OJ|VC) and 'dose' (0.5|1|2) ")
```



The basic question is: “Is ToothGrowth ‘len’ affected by ‘supp’ or ‘dose’?”

The mean, sd, and count of values of ‘len’ are:

```
summaryLen <- ToothGrowth_tbl %>% group_by(supp, dose) %>%
  summarize (mean = mean(len), sd=sd(len), n=n())
print(summaryLen)
```

```
## Source: local data frame [6 x 5]
## Groups: supp
##
##   supp dose  mean      sd  n
## 1   OJ  0.5  13.23 4.459709 10
## 2   OJ   1  22.70 3.910953 10
## 3   OJ   2  26.06 2.655058 10
## 4   VC  0.5   7.98 2.746634 10
## 5   VC   1  16.77 2.515309 10
## 6   VC   2  26.14 4.797731 10
```

Compare tooth growth by supp and dose

Compare different doses of supp=="OJ"

```
OJ_by_dose <- tbl_df(data.frame(matrix(ncol=4, nrow=3)))
colnames(OJ_by_dose) <- c("supp", "dose1", "dose2", "pvalue")
OJ_by_dose[1,] = c("OJ", ".5", "1",
  t.test(len~dose, data=droplevels(filter(ToothGrowth_tbl, supp=="OJ" & dose!=2)))[3]$p.value)
OJ_by_dose[2,] = c("OJ", ".5", "2",
  t.test(len~dose, data=droplevels(filter(ToothGrowth_tbl, supp=="OJ" & dose!=1)))[3]$p.value)
OJ_by_dose[3,] = c("OJ", "1", "2",
  t.test(len~dose, data=droplevels(filter(ToothGrowth_tbl, supp=="OJ" & dose!=0.5)))[3]$p.value)
OJ_by_dose <- mutate(OJ_by_dose, H0 = ifelse(pvalue > 0.05, "Accept", "Reject"))
print(OJ_by_dose)
```

```
## Source: local data frame [3 x 5]
```

```
##
##      supp dose1 dose2                pvalue      H0
## 1    OJ   .5      1  8.7849190551615e-05 Accept
## 2    OJ   .5      2  1.32378387769724e-06 Accept
## 3    OJ    1      2   0.0391951420462442 Reject
```

Compare different doses of supp=="VC"

```
VC_by_dose <- tbl_df(data.frame(matrix(ncol=4, nrow=3)))
colnames(VC_by_dose) <- c("supp", "dose1", "dose2", "pvalue")
VC_by_dose[1,] = c("VC", ".5", "1",
  t.test(len~dose, data=droplevels(filter(ToothGrowth_tbl, supp=="VC" & dose!=2)))[3]$p.value)
VC_by_dose[2,] = c("VC", ".5", "2",
  t.test(len~dose, data=droplevels(filter(ToothGrowth_tbl, supp=="VC" & dose!=1)))[3]$p.value)
VC_by_dose[3,] = c("VC", "1", "2",
  t.test(len~dose, data=droplevels(filter(ToothGrowth_tbl, supp=="VC" & dose!=0.5)))[3]$p.value)
VC_by_dose <- mutate(VC_by_dose, H0 = ifelse(pvalue > 0.05, "Accept", "Reject"))
print(VC_by_dose)
```

```
## Source: local data frame [3 x 5]
##
##      supp dose1 dose2                pvalue      H0
## 1    VC   .5      1  6.81101770286506e-07 Accept
## 2    VC   .5      2  4.6815774144921e-08 Accept
## 3    VC    1      2  9.15560305663864e-05 Accept
```

Compare different supp by dose

```
dose_by_supp <- tbl_df(data.frame(matrix(ncol=4, nrow=3)))
colnames(dose_by_supp) <- c("dose", "supp1", "supp2", "pvalue")
dose_by_supp[1,] <- c(".5", "OJ", "VC",
  t.test(len~supp, data=droplevels(filter(ToothGrowth_tbl, dose==.5)))[3]$p.value)
dose_by_supp[2,] <- c("1", "OJ", "VC",
  t.test(len~supp, data=droplevels(filter(ToothGrowth_tbl, dose==1)))[3]$p.value)
dose_by_supp[3,] <- c("2", "OJ", "VC",
  t.test(len~supp, data=droplevels(filter(ToothGrowth_tbl, dose==2)))[3]$p.value)
dose_by_supp <- mutate(dose_by_supp, H0 = ifelse(pvalue > 0.05, "Accept", "Reject"))
print(dose_by_supp)
```

```
## Source: local data frame [3 x 5]
##
##      dose supp1 supp2                pvalue      H0
## 1   .5    OJ    VC  0.00635860676409681 Reject
## 2    1    OJ    VC  0.00103837587229988 Reject
## 3    2    OJ    VC   0.963851588723373 Accept
```

Conclusions

tbd