Stat_Inference_project_Part2

Patrick Close

November 22, 2015

Summary: Use the R data set "ToothGrowth" to run exploratory data analysis and test for significance of the two independent variables "dose" & "supp" on the dependent variable "len".

Step 1: perform exploratory data analysis to understand the dataset

```
library(datasets)
library(ggplot2)
data("ToothGrowth")
#
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 ...
```

```
ToothGrowth$dose <- as.factor(ToothGrowth$dose)
#overall summary:
summary(ToothGrowth)
```

```
##       len        supp     dose
##  Min.   : 4.20   OJ:30   0.5:20
##  1st Qu.:13.07   VC:30   1  :20
##  Median :19.25           2  :20
##  Mean   :18.81
##  3rd Qu.:25.27
##  Max.   :33.90
```

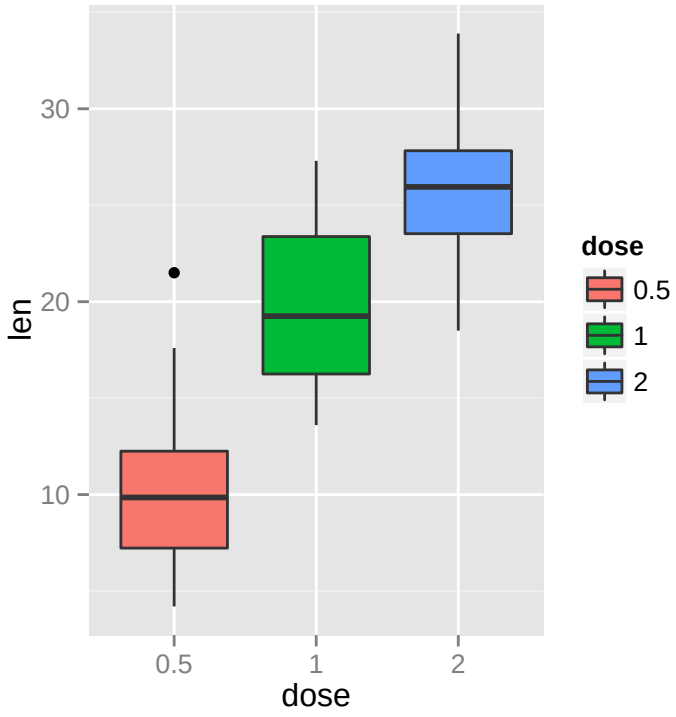
```
#summary grouped by supplement and dose
by(ToothGrowth$len, INDICES = list(ToothGrowth$supp, ToothGrowth$dose),summary)
```

```
## : OJ
## : 0.5
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.20    9.70   12.25   13.23   16.18   21.50
## ------------------------------------------------------------
## : VC
## : 0.5
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4.20    5.95    7.15    7.98   10.90   11.50
## ------------------------------------------------------------
## : OJ
## : 1
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   14.50   20.30   23.45   22.70   25.65   27.30
```

```
## ------------------------------------------------------------
## : VC
## : 1
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   13.60   15.27   16.50   16.77   17.30   22.50
## ------------------------------------------------------------
## : OJ
## : 2
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   22.40   24.58   25.95   26.06   27.08   30.90
## ------------------------------------------------------------
## : VC
## : 2
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18.50   23.38   25.95   26.14   28.80   33.90
```

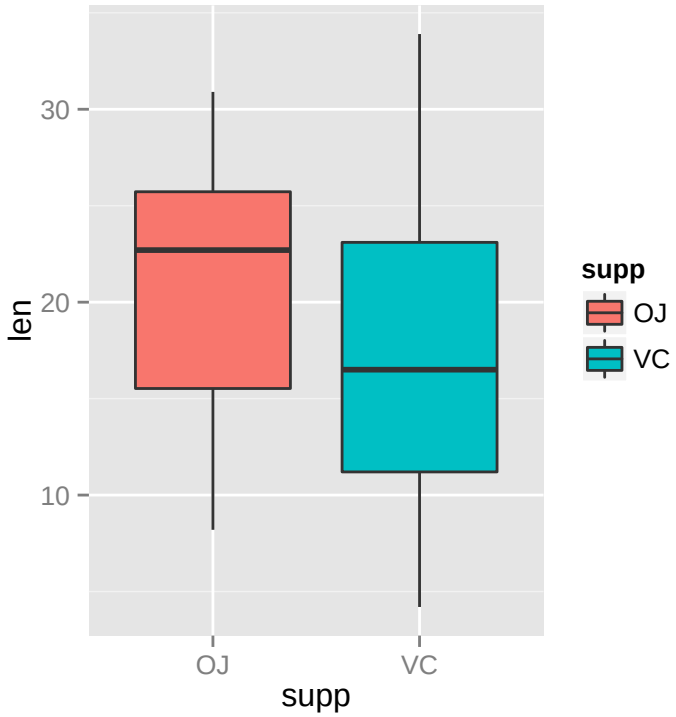
Plot the dependent variable "len" vs independent variable "dose"

```
g1 <- ggplot(aes(x=dose,y=len), data=ToothGrowth) + geom_boxplot(aes(fill=dose))
print(g1)
```



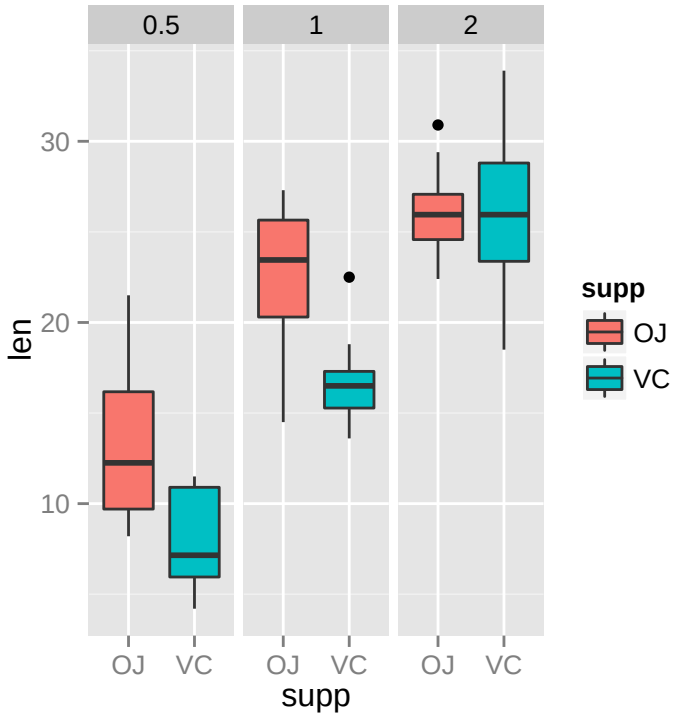
Plot the dependent variable "len" vs the independent variable "supp"

```
g2 <- ggplot(aes(x=supp, y=len), data = ToothGrowth) + geom_boxplot(aes(fill=supp))
print(g2)
```



Plot “supp” at the three dose levels

```
g3 <- ggplot(aes(x=supp, y=len), data = ToothGrowth) + geom_boxplot(aes(fill=supp))
g3 <- g3 + facet_wrap(~dose)
print(g3)
```



It appears there increasing dose has a directly positive relationship with len, but the relationship between dose and supp is less clear.

Step 2: use hypothesis testing to test for significance of the independent variables on dependent variable

```
t.test(len ~ supp, data = ToothGrowth)
```

```
##
##  Welch Two Sample t-test
##
## data:  len by supp
## 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:
##  -0.1710156  7.5710156
## sample estimates:
## mean in group OJ mean in group VC
##         20.66333         16.96333
```

with $p > 0.05$ we fail to reject the null hypothesis and cannot conclude the difference in response of len by supp is significantly different than zero (also note the CI contains zero)

Next create sub levels of dose to be tested

```
#create three dose sub-level pairs and test for dose significance
t_tests <- list()
doses <- c("0.5","1","2")
for (dL in doses){
     for(dH in doses){
```

```
          if (dL < dH){
               ToothGrowth_sub <- subset(ToothGrowth, dose %in% c(dL,dH))
               t <- t.test(len ~ dose, data = ToothGrowth_sub)
               test_doses <- paste(dL,"-",dH)
               t_tests <- rbind(t_tests, list(test_doses=test_doses, p.value = t$p.value))
          }
     }
}
print(t_tests)
```

```
##      test_doses p.value
## [1,] "0.5 - 1"  1.268301e-07
## [2,] "0.5 - 2"  4.397525e-14
## [3,] "1 - 2"    1.90643e-05
```

All three levels of dose yield a p-value < 0.05 and CIs do not contain zero, therefore we can reject the null hypothesis and conclude increasing dose levels have a significant effect on tooth length.

Test for significance between doses for each supplement

```
#create subsets of supp within dose and test for significance
t_tests <- NULL
supps <- c("OJ","VC")
for (dL in doses){
     for(dH in doses){
          if (dL < dH){
               ToothGrowth_sub <- subset(ToothGrowth, dose %in% c(dL,dH))
               for(s in supps){
                    ToothGrowth_sub_s <- subset(ToothGrowth_sub, supp %in% s)
                    t <- t.test(len ~ dose, data = ToothGrowth_sub_s)
                    test_doses <- paste(dL,"-",dH,s)
                    t_tests <- rbind(t_tests, list(test_doses=test_doses, p.value = t$p.value))
               }
          }
     }
}
print(t_tests)
```

```
##      test_doses   p.value
## [1,] "0.5 - 1 OJ" 8.784919e-05
## [2,] "0.5 - 1 VC" 6.811018e-07
## [3,] "0.5 - 2 OJ" 1.323784e-06
## [4,] "0.5 - 2 VC" 4.681577e-08
## [5,] "1 - 2 OJ"   0.03919514
## [6,] "1 - 2 VC"   9.155603e-05
```

When stratifying by supplement type the dose dependent responses maintain significance, implying the overall effect between dose and length is not driven by one supplement.

Conclusions:

1. Supplement does not have a significant effect on tooth length

2. Dose does have a significant effect on tooth length

Assumptions:

1. The experiment was randomized in its allocation of guinea pigs to different dose level categories and supplement.
2. The guinea pigs used in the experiment are representative of the entire population of guinea pigs, allowing inference about the population from this sample.
3. The two guinea pig groups studied are independent.
4. The variances are assumed to be different for the two groups being compared.