Statistical Inference Course Project: Exploratory Analysis of ToothGrowth

equinaut

October 23, 2015

Overview

This report is to begin exploratory data analyses of the ToothGrowth data set.

Environment Set Up

```
# Libraries
library(ggplot2)
library(psych)
```

```
##
## Attaching package: 'psych'
##
## The following object is masked from 'package:ggplot2':
##
##     %+%
```

```
# Data Set
data("ToothGrowth")
```

Data Structure

Using R's `head` and `class` functions, below is an output of the data structure after converting dose to a factor.

Classes

```
##      [,1]      [,2]     [,3]
## [1,] "len"     "supp"   "dose"
## [2,] "numeric" "factor" "factor"
```

Head

```
##    len supp dose
## 1  4.2   VC  0.5
## 2 11.5   VC  0.5
## 3  7.3   VC  0.5
## 4  5.8   VC  0.5
## 5  6.4   VC  0.5
## 6 10.0   VC  0.5
```

```
length(ToothGrowth); nrow(ToothGrowth)
```

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

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

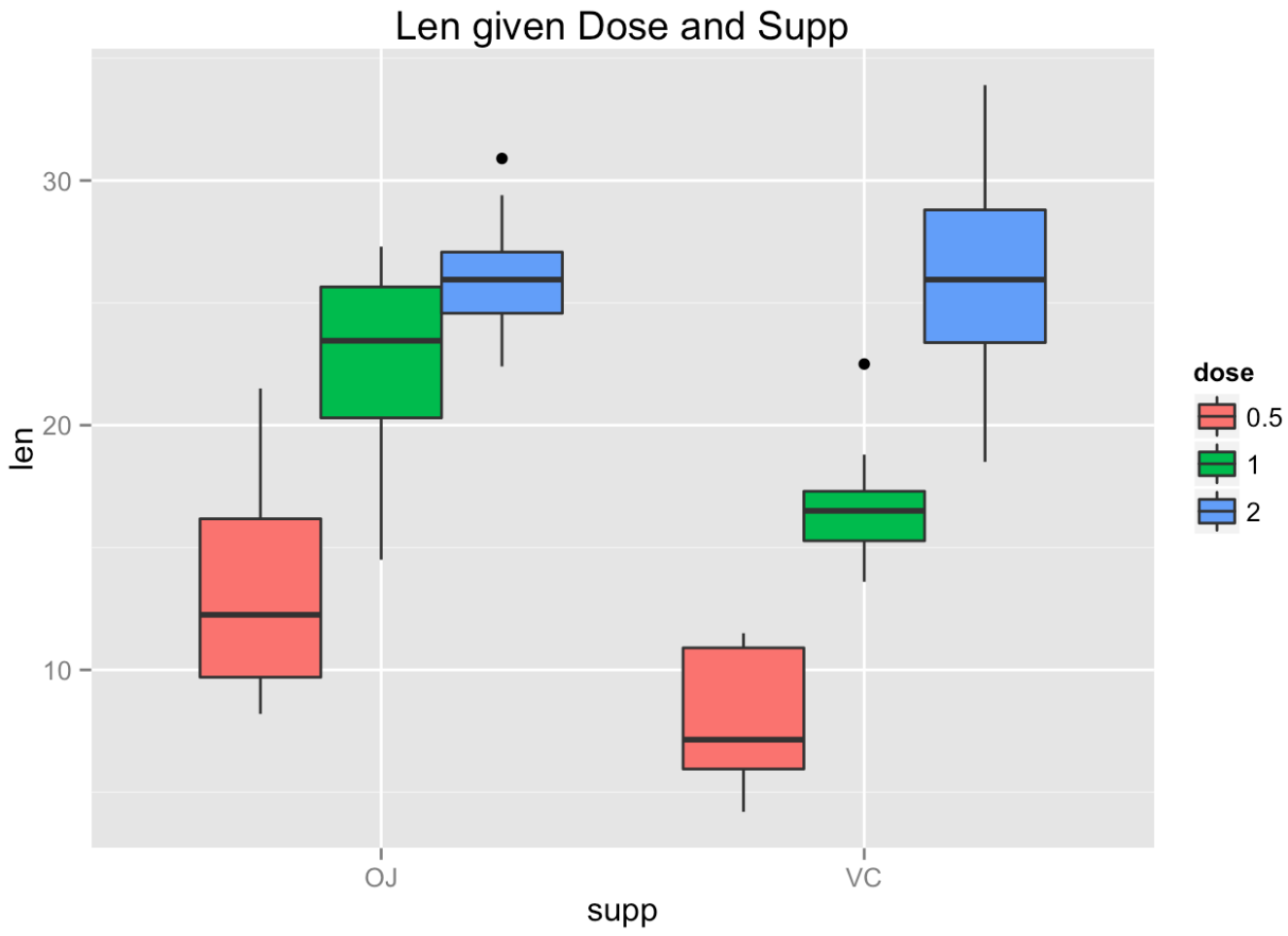
Data Summary

```
describe(ToothGrowth)
```

```
##       vars  n  mean   sd median trimmed  mad min  max range  skew kurtosis
## len      1 60 18.81 7.65  19.25   18.95 9.04 4.2 33.9  29.7 -0.14    -1.04
## supp*    2 60  1.50 0.50   1.50    1.50 0.74 1.0  2.0   1.0  0.00    -2.03
## dose*    3 60  2.00 0.82   2.00    2.00 1.48 1.0  3.0   2.0  0.00    -1.55
##         se
## len   0.99
## supp* 0.07
## dose* 0.11
```

`len` ranges from 4.20 to 33.90 with a mean of 18.81. `dose` ranges from 0.500 to 2.000 with a mean of 1.167. The boxplot below visually summarizes the above output.

```
ggplot(aes(x = supp, y = len), data = ToothGrowth) +
    geom_boxplot(aes(fill=dose)) +
    labs(title = "Len given Dose and Supp")
```



Confidence Intervals, Hypothesis Tests

Supp

```
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
```

At an alpha of .05 the p-value is not signifcant. Additionally, the confidence interval contains 0. These both indicate that the mean `len` difference between OJ and VC is not significantly different than zero.

Dose

To compare the doses in a similar manner as to that above, we must subset the data for each unique dose combination. This is so that we can compare the means and variances between doses.

The total number of combinations is 3:

- .5 vs 1
- .5 vs 2
- 1 vs 2

```
# Subset
set.0.5.VS.1.0 <- subset(ToothGrowth, ToothGrowth$dose %in% c(.5, 1))
set.0.5.VS.2.0 <- subset(ToothGrowth, ToothGrowth$dose %in% c(.5, 2))
set.1.0.VS.2.0 <- subset(ToothGrowth, ToothGrowth$dose %in% c(1, 2))

# 0.5 vs 1.0
t.test(len ~ dose, data = set.0.5.VS.1.0)
```

```
##
##  Welch Two Sample t-test
##
## data:  len by dose
## t = -6.4766, df = 37.986, p-value = 1.268e-07
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.983781  -6.276219
## sample estimates:
## mean in group 0.5   mean in group 1
##            10.605            19.735
```

```
# 0.5 vs 2.0
t.test(len ~ dose, data = set.0.5.VS.2.0)
```

```
##
##  Welch Two Sample t-test
##
## data:  len by dose
## t = -11.799, df = 36.883, p-value = 4.398e-14
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -18.15617 -12.83383
## sample estimates:
## mean in group 0.5   mean in group 2
##            10.605            26.100
```

```
# 1.0 vs 2.0
t.test(len ~ dose, data = set.1.0.VS.2.0)
```

```
##
##  Welch Two Sample t-test
##
## data:  len by dose
## t = -4.9005, df = 37.101, p-value = 1.906e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -8.996481 -3.733519
## sample estimates:
## mean in group 1 mean in group 2
##          19.735          26.100
```

For each combination we observe a signficant p-value indicating that all of the differences between the means are significantly different than 0.

Conclusions

We conclude that:

- Although the field `supp` may be useful in other analyses, these do not appear to have a direct effect on the field `len`
- The dosages appear to mean lengths that are statistically different from one another. This therefore would be an opportune area to focus the study on.

This is based off the assumptions that this data is randomly collected and representative of the population, and that the variances between the dose subsets are independent and unequal.