StatsProject 2

George Liu

August 18, 2015

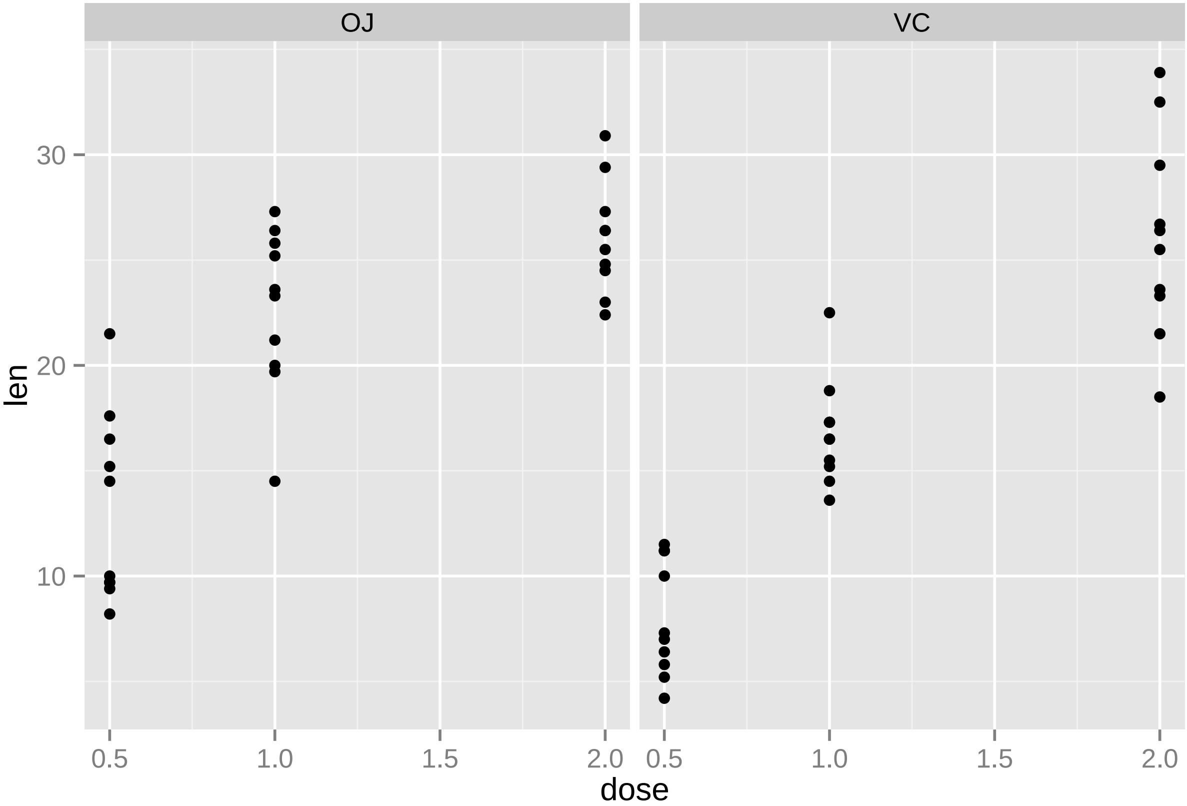
Data Loading and Exploratory Analyses

First, the data are loaded and some exploratory analyses are done. Based on the plot, it seems that's a positive correlation between tooth lengths and doses given. Also, the two panels present a possible difference in effect between the two delivery methods(orange juice and ascorbic acid).

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
##
## The following objects are masked from 'package:stats':
##
##     filter, lag
##
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
# load the data
data("ToothGrowth")
library(ggplot2)
# some exploratory data analyses
g <- ggplot(ToothGrowth, aes(dose, len))
gg <- g + geom_point() + facet_grid(.~supp)
gg
```



Summary of Data

The following code summarizes the data by calculating the number of observations, average length, variance and standard deviation for tooth lengths measured with both delivery methods.

```
ToothGrowth%>%
    group_by(supp, dose)%>%
    summarize(count = n(), avg.length = mean(len), variance = var(len), sd = sd(len))
```

```
## Source: local data frame [6 x 6]
## Groups: supp
##
##   supp dose count avg.length  variance       sd
## 1   OJ  0.5    10      13.23 19.889000 4.459709
## 2   OJ  1.0    10      22.70 15.295556 3.910953
## 3   OJ  2.0    10      26.06  7.049333 2.655058
## 4   VC  0.5    10       7.98  7.544000 2.746634
## 5   VC  1.0    10      16.77  6.326778 2.515309
## 6   VC  2.0    10      26.14 23.018222 4.797731
```

Confidence Interval

The data given are a sample from a possible large population. From the data, we can have some “point estimates”. With these estimates, we can make statistical inference - confidence interval is one of them. We can use the average tooth length with a certain dose using a delivery method as an estimate, then use it to generate the confidence interval for the population, i.e. what would the population's average tooth length when using the previously mentioned method and dose.

To construct the confidence interveal, we need parameters in the following formula: point estimate +/- Z*SE

For 95% confidence, we have a Z of 1.833113 (qt(.95, df = 9)). Below add two new columns for the upper and lower limits of the confidence intervals for each samples.

```r
z <- qt(.95, df = 9)
grouped <- group_by(ToothGrowth, supp, dose)
summary <- summarize(grouped, count = n(), avg.length = mean(len), variance = var(len), sd = sd(len))
summary.new <- mutate(summary, lower = avg.length - z * sd, upper = avg.length + z * sd)
summary.new
```

```
## Source: local data frame [6 x 8]
## Groups: supp
##
##   supp dose count avg.length  variance       sd     lower    upper
## 1   OJ  0.5    10      13.23 19.889000 4.459709  5.054851 21.40515
## 2   OJ  1.0    10      22.70 15.295556 3.910953 15.530781 29.86922
## 3   OJ  2.0    10      26.06  7.049333 2.655058 21.192979 30.92702
## 4   VC  0.5    10       7.98  7.544000 2.746634  2.945109 13.01489
## 5   VC  1.0    10      16.77  6.326778 2.515309 12.159155 21.38084
## 6   VC  2.0    10      26.14 23.018222 4.797731 17.345217 34.93478
```

Hypothesis Test

One interesting question we can ask based on the data is whether orange juice is more effective than ascorbic acid in terms of tooth growth. We can set it as a hypothesis, to test it, we have:

H0: mu.oj = mu.vc HA: mu.oj > mu.vc

To test this hypothesis, we need to run a t-test as below(Here we have 3 groups of data to test, i.e. the 3 different levels of dosage).

```r
# Get the lenth data for method OJ and dose 0.5
oj.5 <- filter(ToothGrowth, supp == "OJ", dose == 0.5)
oj.5 <- oj.5$len
# Get the lenth data for method VC and dose 0.5
vc.5 <- filter(ToothGrowth, supp == "VC", dose == 0.5)
vc.5 <- vc.5$len
# Get the lenth data for method OJ and dose 1
oj1 <- filter(ToothGrowth, supp == "OJ", dose == 1)
oj1 <- oj1$len
# Get the lenth data for method VC and dose 1
vc1 <- filter(ToothGrowth, supp == "VC", dose == 1)
vc1 <- vc1$len
# Get the lenth data for method OJ and dose 2
oj2 <- filter(ToothGrowth, supp == "OJ", dose == 2)
oj2 <- oj2$len
# Get the lenth data for method VC and dose 2
vc2 <- filter(ToothGrowth, supp == "VC", dose == 2)
vc2 <- vc2$len
```

Conclusion

Now we perform t-test for these three groups of data.

```
t.test(oj.5, vc.5, alternative = "greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  oj.5 and vc.5
## 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
```

Since p-value is less than .05, we should reject H0 hypothesis and it means for the population, OJ does seem to be more effective than VC at the 0.5 dose level.

Similiarly, for the other two levels, we can also run the t-test as follows:

```
t.test(oj1, vc1, alternative = "greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  oj1 and vc1
## t = 4.0328, df = 15.358, p-value = 0.0005192
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  3.356158      Inf
## sample estimates:
## mean of x mean of y
##     22.70     16.77
```

```
t.test(oj2, vc2, alternative = "greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  oj2 and vc2
## t = -0.046136, df = 14.04, p-value = 0.5181
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -3.1335     Inf
## sample estimates:
## mean of x mean of y
##     26.06     26.14
```

Therefore, for the dose 1 level, we have same conclusion, i.e. OJ is more effective than VC. But for the dose 2 level, since p-value is greater than .5, we don't have sufficient evidence to reject the null hypothesis, therefore won't conclude the whether VC is more effective in the population or not.

Assumptions

Here we assume the samples are independent.