ToothGrowth - Basic Inferential Data Analysis

Sohail Munir Khan

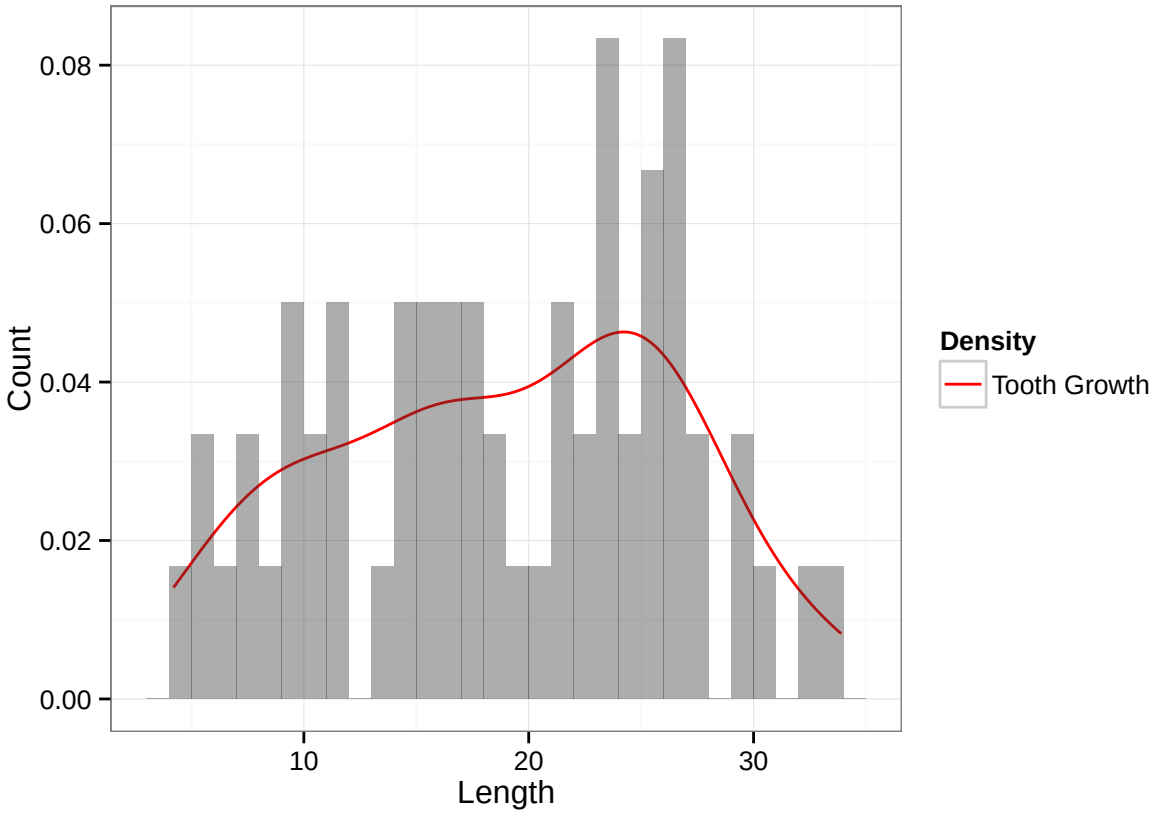
17 June 2015

Exploratory Data Analysis

ToothGrowth is a data.frame within R package {datasets}.
The response is the length of odontoblasts (teeth) in each of 10 guinea pigs at each of three dose levels of Vitamin C (0.5, 1, and 2 mg) with each of two delivery methods (orange juice or ascorbic acid).

```
library(datasets)
data(ToothGrowth)
summary(ToothGrowth) # Summarise ToothGrowth and its variables
```

```
##       len        supp         dose
##  Min.   : 4.20   OJ:30   Min.   :0.500
##  1st Qu.:13.07   VC:30   1st Qu.:0.500
##  Median :19.25           Median :1.000
##  Mean   :18.81           Mean   :1.167
##  3rd Qu.:25.27           3rd Qu.:2.000
##  Max.   :33.90           Max.   :2.000
```

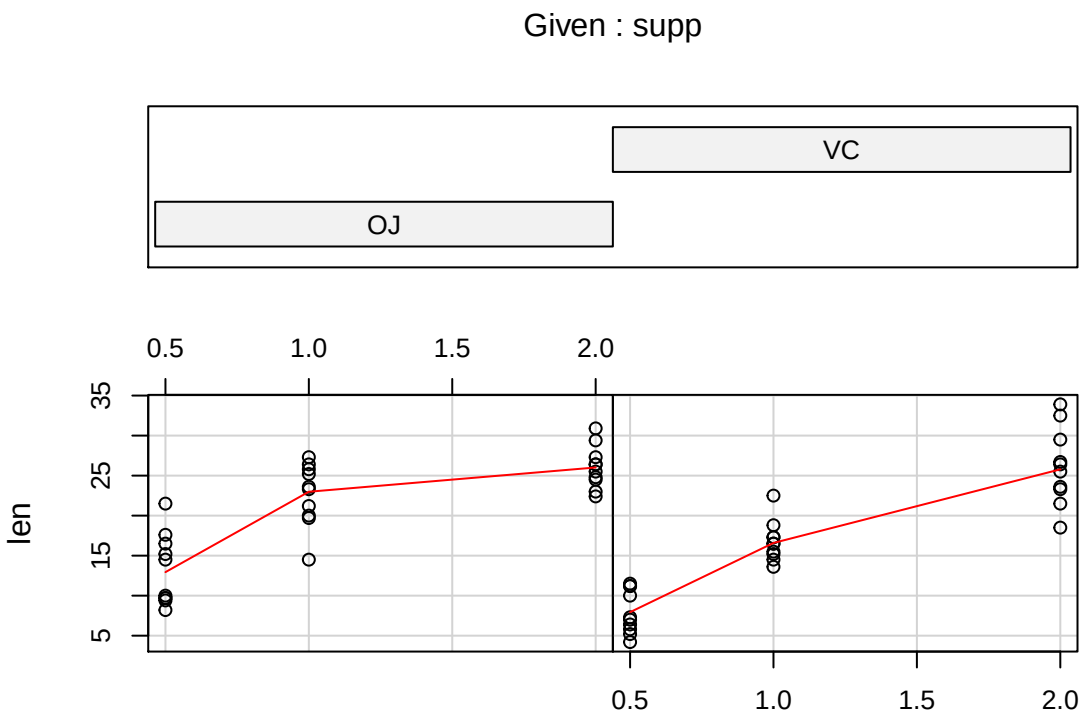


```r
mean(ToothGrowth$len) # mean for len of ToothGrowth
```

```
## [1] 18.81333
```

```r
sd(ToothGrowth$len) # SD for len of ToothGrowth
```

```
## [1] 7.649315
```



ToothGrowth data: length vs dose, given type of supplement

Data Summary

Full data has 60 Tooth Growth lengths (**len** variable). Mean is 18.81 and Standard Deviation is 7.65.
Column **supp** is a factor between the two delivery methods (ascorbic acid ("VC") vs orange juice ("OJ"))
Variable **dose** is essentially either a value of 0.5, 1 or 2 corresponding to the levell of Vitamin in mg
The density function for the len of Tooth is slightly skewed as you can see from the Tooth Growth Density figure on the previous page
Using the **coplot** function to look at Tooth Growth len, we can see that at lower dose levels (0.5 & 1.0) orange juice ("**OJ**") as a supplement does better than ascorbic acid ("**VC**") while at higher levels (2.0), both supplements do similarly although the gains look better for **VC**. We will try to prove this during our hypothesis testing phase

---

Confidence Intervals and / or Hypothesis Tests

```
vc_df <- subset(ToothGrowth, supp == "VC")
oj_df <- subset(ToothGrowth, supp == "OJ")
overall_result <- t.test(vc_df$len, oj_df$len, paired = TRUE)
overall_result
```

```
##
##  Paired t-test
##
## data:  vc_df$len and oj_df$len
## t = -3.3026, df = 29, p-value = 0.00255
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -5.991341 -1.408659
## sample estimates:
## mean of the differences
##                    -3.7
```

Overall between VC and OJ, and putting VC as the null hypothesis, OJ does better than VC. This can be seen from confidence interval -5.99, -1.41 not containing mean zero(0) as well as is way below zero(0) and 0.00255 not very close to or higher than significance level 0.05 (5%)

```
vc_0.5_len <- subset(vc_df, dose == 0.5)$len
oj_0.5_len <- subset(oj_df, dose == 0.5)$len
result_0.5 <- t.test(vc_0.5_len, oj_0.5_len, paired = TRUE)
result_0.5
```

```
##
##  Paired t-test
##
## data:  vc_0.5_len and oj_0.5_len
## t = -2.9791, df = 9, p-value = 0.01547
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -9.236542 -1.263458
## sample estimates:
## mean of the differences
##                   -5.25
```

Still looking at data for VC and OJ separately, but this time honing in on the 0.5 mg supplement, OJ still does better. This can be seen from confidence interval -9.24, -1.26 not containing mean zero(0) and is below zero and 0.01547 not very close to or higher than significance level 0.05 (5%)

```
oj_1_len <- subset(oj_df, dose == 1)$len
vc_1_len <- subset(vc_df, dose == 1)$len
result_1 <- t.test(vc_1_len, oj_1_len, paired = TRUE)
result_1
```

```
##
```

```
##  Paired t-test
##
## data:  vc_1_len and oj_1_len
## t = -3.3721, df = 9, p-value = 0.008229
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -9.908089 -1.951911
## sample estimates:
## mean of the differences
##                   -5.93
```

For 1 mg supplement and looking at VC and OJ data, OJ does better still. This can be seen from confidence interval -9.91, -1.95 not containing mean zero(0) as well as significtantly below zero(0) and 0.00823 not very close to or higher than significance level 0.05 (5%)

```
vc_2_len <- subset(vc_df, dose == 2)$len
oj_2_len <- subset(oj_df, dose == 2)$len
result_2 <- t.test(vc_2_len, oj_2_len, paired = TRUE)
result_2
```

```
##
##  Paired t-test
##
## data:  vc_2_len and oj_2_len
## t = 0.0426, df = 9, p-value = 0.967
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -4.168976  4.328976
## sample estimates:
## mean of the differences
##                    0.08
```

At 2 mg level of supplement, VC definitely sits comfortably as the winner. The null hypothesis interval (-4.17, 4.33) easily contains 0 as the mean with a p-value of 96.7%

Conclusion

Our initially reaction that compares the two plots together did provide us some hints on the type of delivery methods that might be the better than the other. After doing confidence intervals, we have concluded the same results with atleast 95% confidence.

Assumptions

- We assume that the distribution for **len** variable resembles a t-distribution even though it's slightly skewed.
- All tests are performed in a paired fashion because the same 10 guinea pigs were used for testing purposes. We also assume that appropriate time in between tests were provided to avoid any biased results