# ToothGrowth Dataset, Basic Inferential Data Analysis

*emeko*

*January 25, 2016*

## Overview

The *ToothGrowth* data set from the R *datasets* package is a dataset of the response in 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).The data set is provided as a data frame of 60 observations across three variables: length (*len*), supplement (*supp*; a factor that is either 'OJ' for orange juice, or 'VC' for vitamin C (ascorbic acid)), and dose in milligrams (*dose*).

### Basic Exploratory Analysis of the Dataset

The *summary()* function can provide basic information. We know from the description of the data that the tooth length is a reponse (dependent) variable.

```
data('ToothGrowth')

summary( ToothGrowth )
```

```
##       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
```

```
table( ToothGrowth$supp, ToothGrowth$dose )
```

```
##     
##      0.5  1  2
##   OJ  10 10 10
##   VC  10 10 10
```

There are **only 10 data points** for each combination of dose and supplement. For this reason, we will use Student's t-statistic when considering the data.

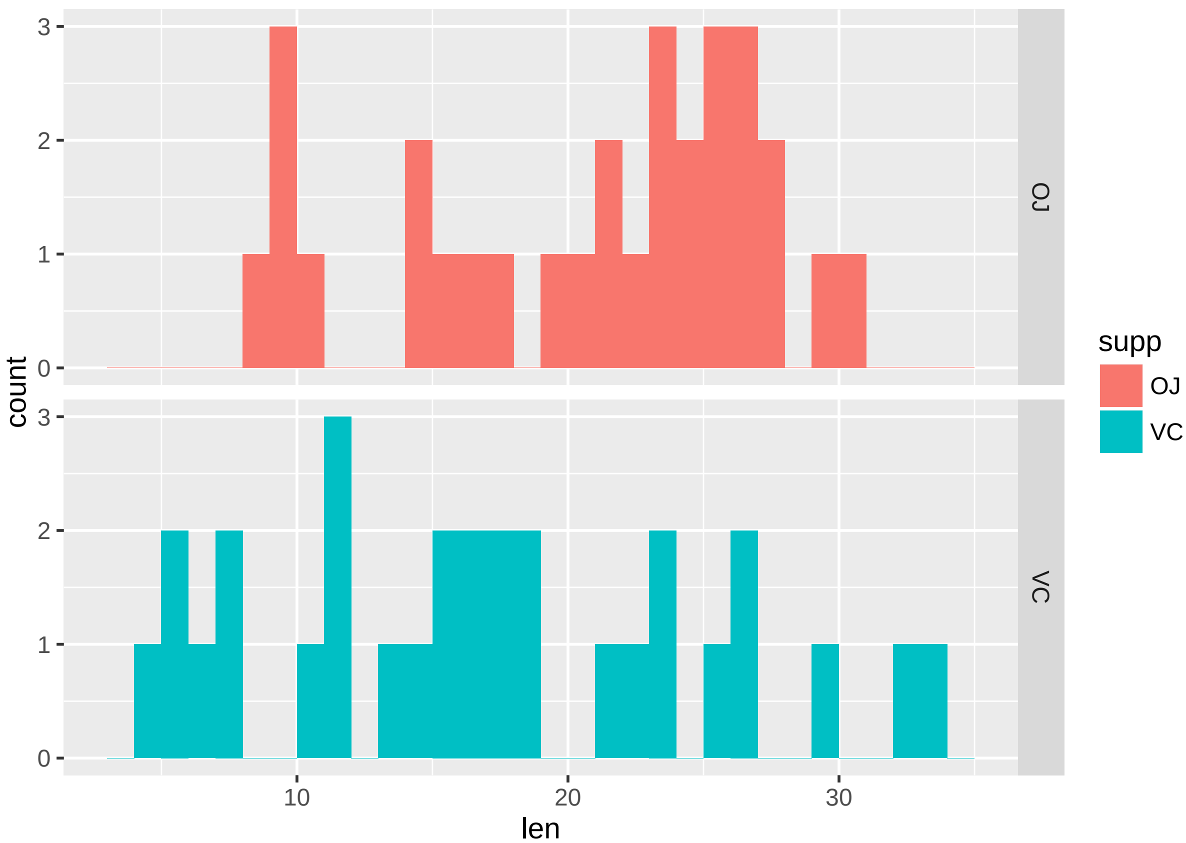
Since there is nothing in the description of the data that suggests that the observations are paired in any way, we **assume that the obeservations are not paired**.

## Supplement Dependence

*QUESTION*: Does the observed amount of tooth growth depend on the type of supplement provided?

Lets consider the data faceted by the supplement:

```r
ggplot(ToothGrowth, aes(x=len, fill=supp)) +
    geom_histogram( binwidth = 1 ) +
    facet_grid(supp ~ .)
```




From the above plot, the distributions look similar save that the ascorbic acid results appear more spread out. The difference in spread suggests that the variance for VC-treated cases is greater than for the OJ cases:

```r
aggregate( len ~ supp, ToothGrowth, var)
```

```
##   supp      len
## 1   OJ 43.63344
## 2   VC 68.32723
```

There are large differences in the length variance between the treatments at each dose, which suggests, in the absence of any prior knowledge that would suggest otherwise, that we should **not treat the variances as equal**.

The question "Does the observed amount of tooth growth depend on the type of supplement provided?" suggests a statistical test, where the null-hypothesis ($H_0$) is that the amount of tooth growth is the same for VC and OJ treated animals ($\mu_{VC} = \mu_{OJ}$) with 95% confidence. Note that there are multiple dose levels, so we'd like to test the hypothesis at each dosage. The alternate hypothesis is $H_a : \mu_{VC} \neq \mu_{OJ}$.

The length means by dosage and supplement:

```r
aggregate( len ~ supp + dose, ToothGrowth, mean)
```

```
##   supp dose   len
```

```
## 1   OJ  0.5 13.23
## 2   VC  0.5  7.98
## 3   OJ  1.0 22.70
## 4   VC  1.0 16.77
## 5   OJ  2.0 26.06
## 6   VC  2.0 26.14
```

For $H_0 : \mu_{VC} = \mu_{OJ}$ at each dose:

```
test.results <- by(
    ToothGrowth,
    ToothGrowth$dose,
    function( data ) {
        t.test( len ~ supp, data, paired = FALSE, var.equal = FALSE )
    }
)

test.results
```

```
## ToothGrowth$dose: 0.5
##
##  Welch Two Sample t-test
##
## data:  len by supp
## t = 3.1697, df = 14.969, p-value = 0.006359
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.719057 8.780943
## sample estimates:
## mean in group OJ mean in group VC
##            13.23             7.98
##
## --------------------------------------------------------
## ToothGrowth$dose: 1
##
##  Welch Two Sample t-test
##
## data:  len by supp
## t = 4.0328, df = 15.358, p-value = 0.001038
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  2.802148 9.057852
## sample estimates:
## mean in group OJ mean in group VC
##            22.70            16.77
##
## --------------------------------------------------------
## ToothGrowth$dose: 2
##
##  Welch Two Sample t-test
##
## data:  len by supp
## t = -0.046136, df = 14.04, p-value = 0.9639
## alternative hypothesis: true difference in means is not equal to 0
```

```
## 95 percent confidence interval:
##  -3.79807  3.63807
## sample estimates:
## mean in group OJ mean in group VC
##            26.06            26.14
```

This suggests that at lower doses, we reject the null-hypothesis that true means are equal, but at the largest does, we cannot reject the null-hypothesis.

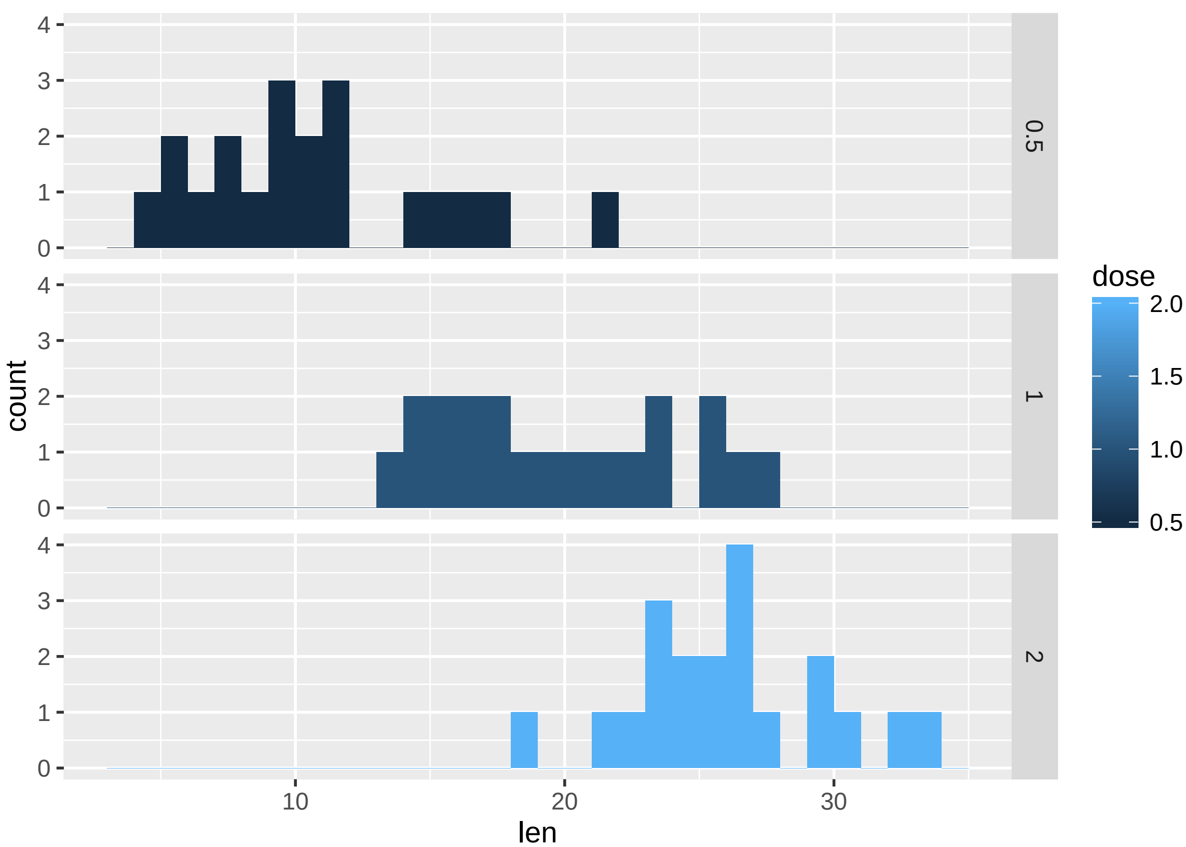
Conclusion: at doses of 0.5 and 1.0 milligrams, Orange Juice is better and ascorbic acid at stimulating tooth growth (with P-values of 0.0063586 and 0.0010384 repsectively). At 2.0 milligrams, the effect is about the same.

## Supplement Dose

*QUESTION*: Is the observed amount of tooth growth dose dependent?

Lets consider the data faceted by the dose:

```
ggplot(ToothGrowth, aes(x=len, fill=dose)) +
    geom_histogram( binwidth = 1 ) +
    facet_grid(dose ~ .)
```




Visually, the distributions of dose versus response seem to have very different means and similar variance. There isn't a reason to believe that there exists any difference in variance by dose in the parent population, so we'll treat the observations as having the same variance.

```r
aggregate( len ~ dose, ToothGrowth, var)
```

```
##   dose      len
## 1  0.5 20.24787
## 2  1.0 19.49608
## 3  2.0 14.24421
```

That being the case, we postulate the null-hypothesis ($H_0$) that there's no difference in the means between doses. The alternative hypothesis is $H_a : \mu_x \neq \mu_y \; \forall \; (x, y) \in [(0.5, 1), (0.5, 2), (1, 2)]$.

```r
t.test( len ~ dose, subset(ToothGrowth, dose %in% c(0.5, 1)), paired = FALSE)
```

```
##
##  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
```

```r
t.test( len ~ dose, subset(ToothGrowth, dose %in% c(0.5, 2)), paired = FALSE)
```

```
##
##  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
```

```r
t.test( len ~ dose, subset(ToothGrowth, dose %in% c(1, 2)), paired = FALSE)
```

```
##
##  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
```

In each of the cases, the means of lengths landed far outside the indicated confidence intervals, and attained a significance level (P-value) of $<1e^{-6}$. Therefore, we can reject the null-hypothesis that the means are equal for the alternative hypothesis that they are not equal.

Conclusion: tooth growth in guinea pigs has a dose-dependent response to vitamin C.

## Appendix

1. R libraries used - datasets, ggplot2
2. ToothGrowth R documentation - ?ToothGrowth
3. TeX is required to create the PDF output - You should install a recommended TeX distribution for your platform
4. Source code - https://github.com/emeko/StatInfProject