# The Effect of Vitamin C on Tooth Growth in Guinea Pigs

*Derwin McGeary*

*09/20/2015*

## Introduction

This is a brief analysis of the `ToothGrowth` dataset which is distributed with the R programming language.

From the documentation for that dataset:

> 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).

We wish to know how the delivery method and dosage affect tooth length.

## Exploratory analysis

First we load the data into memory, check the structure, and do a preliminary plot to see what it looks like.

```r
data(ToothGrowth)
library(ggplot2)
library(knitr)
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 ...
```

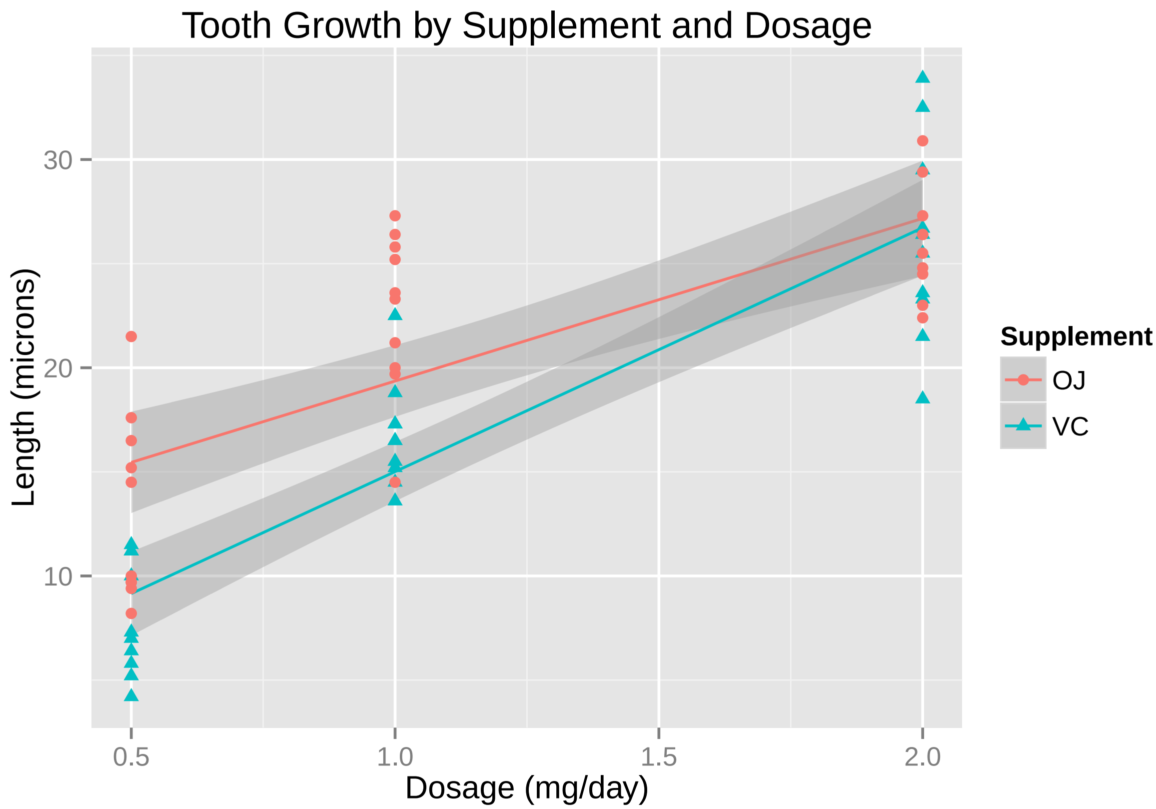
```r
kable(table(ToothGrowth$dose,ToothGrowth$supp))
```

| | OJ | VC |
|---|---|---|
| 0.5 | 10 | 10 |
| 1 | 10 | 10 |
| 2 | 10 | 10 |

```r
summary(ToothGrowth$len)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4.20   13.08   19.25   18.81   25.28   33.90
```

As expected, we have 60 observations (3 dosages, 0.5, 1, and 2; multiplied by 2 supplements, OJ and VC; and 10 guinea pigs per group). We know from the documentation that dosage is in `mg/day`, that `OJ` stands for Orange Juice and `VC` stands for Vitamin C, or ascorbic acid. The units of length are not specified there, but from Crampton 1947 they appear to be microns. For the purposes of this analysis, the units aren't critical. Now we can look at the data visually.




Looking at this graph it appears that there is a positive relationship between dosage and odontoblast length, and that orange juice results in more growth than plain vitamin C.

## Confidence Intervals and Hypothesis Testing

For all of these tests we will use a T-test, and the samples are assumed to be independent (not paired: these are different animals) and roughly normally distributed, and we will examine the p-value and 95% confidence interval.

### Dosage and Odontoblast Length

Here we will make three groups by dosage and take two comparisons: 0.5 mg/day versus 1 mg/day; and 1 mg/day versus 2 mg/day.

```
d1 <- subset(ToothGrowth, dose==0.5); d2 <- subset(ToothGrowth, dose==1.0); d3 <- subset(ToothGrowth, d
halfvsone <- t.test(d1$len,d2$len, paired=FALSE,var.equal=FALSE)
onevstwo <- t.test(d2$len,d3$len, paired=FALSE,var.equal=FALSE)
halfvsone$conf.int
```

```
## [1] -11.983781  -6.276219
## attr(,"conf.level")
## [1] 0.95
```

Our first set of tests compares 0.5 mg/day with 1 mg/day. Our p-value is $1.2683007 \times 10^{-7}$ which is very low. Our confidence interval is -11.9837813 to -6.2762187 which does **not** include zero, so we can say that **there is a significant increase in tooth size at the p=0.5 level**.

```
onevstwo$conf.int
```

```
## [1] -8.996481 -3.733519
## attr(,"conf.level")
## [1] 0.95
```

Our second set of tests compares 1 mg/day with 2 mg/day. Our p-value is $1.9064295 \times 10^{-5}$ which is also very low. Our confidence interval is -8.9964805 to -3.7335195 which does **not** include zero, so we can again say that **there is a significant increase in tooth size at the p=0.5 level**.

### Supplement and Odontoblast Length

Here we will combine all dosages together and group by which supplement was given.

```
s1 <- subset(ToothGrowth, supp=="OJ"); s2 <- subset(ToothGrowth, supp=="VC")
stest <-t.test(s1$len,s2$len,paired=FALSE, var.equal=FALSE)
stest$conf.int
```

```
## [1] -0.1710156  7.5710156
## attr(,"conf.level")
## [1] 0.95
```

This shows a 95% confidence interval which does include zero, and the p-value of 0.0606345 is greater than 0.05 so we can say that there is **no significant difference at the p=0.05 level** between the OJ and VC groups.

## Conclusion

There is a **definite positive correlation between increased dosage and larger odontoblasts at the p=0.5 level**. This is using the T-test which makes the assumption that the distributions are roughly Gaussian and independent (`iid`).

There appears to be a relationship between supplement type and growth, but it is **not statistically significant at the p=0.5 level**. From our preliminary graph, it may even be important only at lower doses, but that could also be cherry-picking our subsets, a temptation which I will resist. We can also reduce the p-value by getting more data.

## Appendices

### Output of `sessionInfo()`

```r
sessionInfo()
```

```
## R version 3.2.1 (2015-06-18)
## Platform: arm-unknown-linux-gnueabihf (32-bit)
## Running under: Ubuntu 14.10
##
## locale:
##  [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
##  [3] LC_TIME=en_US.UTF-8        LC_COLLATE=en_US.UTF-8
##  [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
##  [7] LC_PAPER=en_US.UTF-8       LC_NAME=C
##  [9] LC_ADDRESS=C               LC_TELEPHONE=C
## [11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] knitr_1.10.5  ggplot2_1.0.1
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_0.12.1      digest_0.6.4     MASS_7.3-33      grid_3.2.1
##  [5] plyr_1.8.1       gtable_0.1.2     formatR_1.2      evaluate_0.7
##  [9] scales_0.2.4     highr_0.5        reshape2_1.4     rmarkdown_0.7.3
## [13] labeling_0.2     proto_0.3-10     tools_3.2.1      stringr_0.6.2
## [17] munsell_0.4.2    yaml_2.1.13      colorspace_1.2-4 htmltools_0.2.6
```

### Code for the graph

```r
qplot(x = ToothGrowth$dose, y=ToothGrowth$len,geom = "smooth",
      method="lm",
      colour=ToothGrowth$supp,
      main="Tooth Growth by Supplement and Dosage",
      xlab="Dosage (mg/day)",
      ylab="Length (microns)") +
  scale_colour_discrete(name="Supplement") +
  scale_shape_discrete(name="Supplement") +
  geom_point(aes(shape=ToothGrowth$supp))
```

### Full output of the T-tests

This is included as a source for any numbers for which code was not visible in the main report, namely the p-values.

For 0.5 mg/day versus 1 mg/day:

```
halfvsone
```

```
##
##  Welch Two Sample t-test
##
## data:  d1$len and d2$len
## 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 of x mean of y
##    10.605    19.735
```

For 1 mg/day versus 2 mg/day:

```
onevstwo
```

```
##
##  Welch Two Sample t-test
##
## data:  d2$len and d3$len
## 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 of x mean of y
##    19.735    26.100
```

For OJ versus VC:

```
stest
```

```
##
##  Welch Two Sample t-test
##
## data:  s1$len and s2$len
## 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 of x mean of y
##  20.66333  16.96333
```