# Effects of Vitamin C doses on Guinea Pig Tooth Growth An analysis using R

*Kevin E. D'Elia*

*March 5, 2016*

## Synopsis of Analysis

In this paper, the techniques of Exploratory Data Analysis (EDA) will be applied to the dataset **ToothGrowth**. Among those techniques are the use of tabular (*summary()*, *str()*) as well as graphical (*histogram()*, *boxplot()*) representations of the data in order to glean patterns and structure within the data. The analysis will attempt to answer the following question: Is the data in this dataset normally distributed and, by extension, if not, then why not?

## Summary of the data

Usually the first step in any analysis is to acquire and load the data. In this case, however, the dataset is preloaded in **R** and is available via the **datasets** package. So, simply typing **ToothGrowth** at the console prompt will display the full dataset. The first few lines look like this:

```r
head(ToothGrowth)
```

```
##    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
```

A bit of background on this dataset: from the R documentation -

```
The response is the length of odontoblasts (cells responsible for tooth growth)
in 60 guinea pigs. Each animal received one of three dose levels of vitamin C
(0.5, 1, and 2 mg/day) by one of two delivery methods, (orange juice or ascorbic acid
(a form of vitamin C and coded as VC).
```

The structure of this dataset, as described by the documentation, is displayed using the following R command:

```r
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 ...
```

This table confirms that the dataset consists of 60 observations of 3 variables. Again, from the R documentation:

```
A data frame with 60 observations on 3 variables.

[,1]     len     numeric     Tooth length

[,2]     supp    factor  Supplement type (VC or OJ).

[,3]     dose    numeric     Dose in milligrams/day
```

Since the second variable is a *factor*, or *grouping*, variable, and the third variable has values in a fixed range from 0.5:2 (and thus not very interesting from an analytical perspective), the next step in the analysis will exclude those two columns and look only at changes in tooth length.

## Exploratory Data Analysis

To see a basic summary of the measures of central tendency for the ToothGrowth growth length variable, use the eponymous R command:

```r
summary(ToothGrowth["len"])
```

```
##       len
##  Min.   : 4.20
##  1st Qu.:13.07
##  Median :19.25
##  Mean   :18.81
##  3rd Qu.:25.27
##  Max.   :33.90
```

Other important descriptive statistics, which are absent from the base *summary()* function, are **skew** and **kurtosis**. Skew is an indication of which way the data is skewed, positively or negatively, while kurtosis describes the curvature of the shape of the distribution (more bell-shaped or less so). While functions to provide these statistics are available in several R packages, such as **pastecs** and **psych**, it is a relatively simple matter to write some R code which calculates these statistics. First, though, it is good practice to handle possible missing data (NA values) in the dataset.

```r
sum(is.na(ToothGrowth))
```

```
## [1] 0
```

The *is.na()* function returns a logical vector which is numerically represented as 0 for FALSE and 1 for TRUE. Taking the sum of that vector and getting a non-zero result indicates the presence of NA values. In this case, however, the result is 0, so the dataset is completely populated with meaningful values. Skew and kurtosis can now be computed.

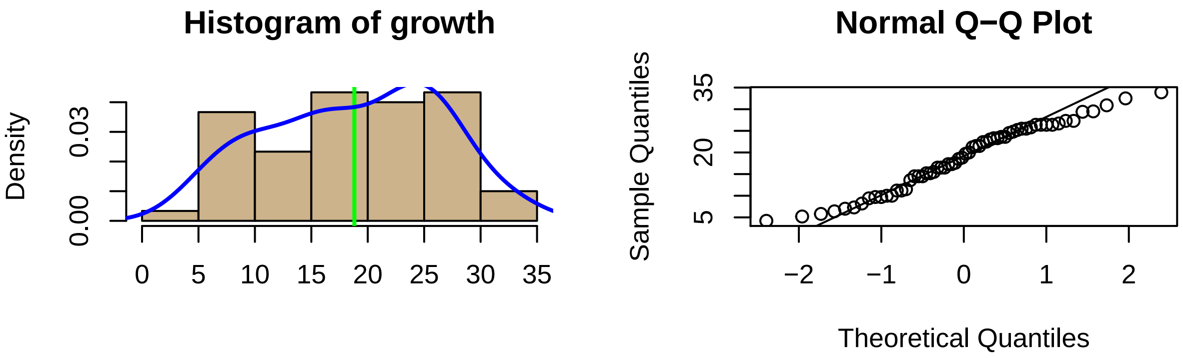
```r
other.stats <- function(x) {
  m <- mean(x);  n <- length(x);  s <- sd(x)
  skew <- round(sum( (x - m)^3 / s^3 ) / n, 3)
  kurtosis <- round(sum( (x - m)^4 / s^4 ) / n - 3, 3)
  return(c(skew=skew, kurtosis=kurtosis))
}
sapply(ToothGrowth["len"], other.stats)
```

```
##              len
## skew      -0.143
## kurtosis -1.043
```

What do these statistics tell us about the data? The mean is slightly lower than the median, so that indicates a small left-skewing of the data. The value for the skew is negative, which supports that assessment. The kurtosis is also slightly negative, indicating a curve that is minimally flatter than a standard bell curve, or relatively mesokurtic.

Tabular descriptive statistics are useful but, as the saying goes, "A picture is worth a thousand words", which leads now to the use of graphical descriptive techniques. Two of the more useful ones are **histograms**, which show either frequencies or probability densities, and **Q-Q plots**, which relate quantiles in the given data to standard quantiles:

```r
hist(growth, prob = TRUE, col = "navajowhite3", border = "black", xlab = "")
abline(v = mean(growth), col = "green", lwd = 2)
lines(density(growth), col="blue", lwd=2)
qqnorm(growth);qqline(growth)
par(original_par)
```



While the previous graphs are used in determining the normalcy of the data, other types of graphs are useful for displaying correlative information. A **coplot** (see *Appendix*) is one of a number of graphical functions that relates numerical data, such as the growth length in this case, to one or more grouping factor variables. Another popular choice is the **boxplot** (see *Appendix*), also known as a *tails and whiskers* plot.

## Observations on EDA artefacts

From an analysis of the graphs produced, the following observations can be made:

- Density curve on the histogram indicates a non-normal distribution; this is supported by the Q-Q plot.
- Data falls off sharply on both ends of the histogram.
- Skewing is slight due to fairly even distribution of data points in the 5-30 measurement range.
- The density curve, while not normal, is still relatively mesokurtic.
- There is a single outlier evident on the boxplot.
- The data appears most symmetric for the 2 milligrams/day dosage of the ascorbic acid (VC) supplement.

# Confidence intervals

Now that an estimate of the population mean has been calculated, how can its accuracy be quantified? Note that the variance for the population is unknown. The technique employed evaluates the margin of error and an interval estimate at the 95% confidence level. First, calculate the margin of error:

```
n <- length(growth)
s <- sd(growth)
SE <- s/sqrt(n); SE
```

```
## [1] 0.9875223
```

```
E <- qt(.975, n-1) * SE; round(E, 3)
```

```
## [1] 1.976
```

Next, using the margin of error, calculate the confidence interval:

```
x_bar <- mean(growth)
round(x_bar + c(-E, E), 3)
```

```
## [1] 16.837 20.789
```

Another approach to obtaining the confidence interval is by using the R function *t.test()* and examining the confidence interval (**conf.int**) parameter. One assumption for using **t.test()** is that the population from which the sample has been drawn should be normal, but minor departures from normality do not affect this test. This assumption is true for the tooth growth length data.

```
round(t.test(growth)$conf.int, 3)
```

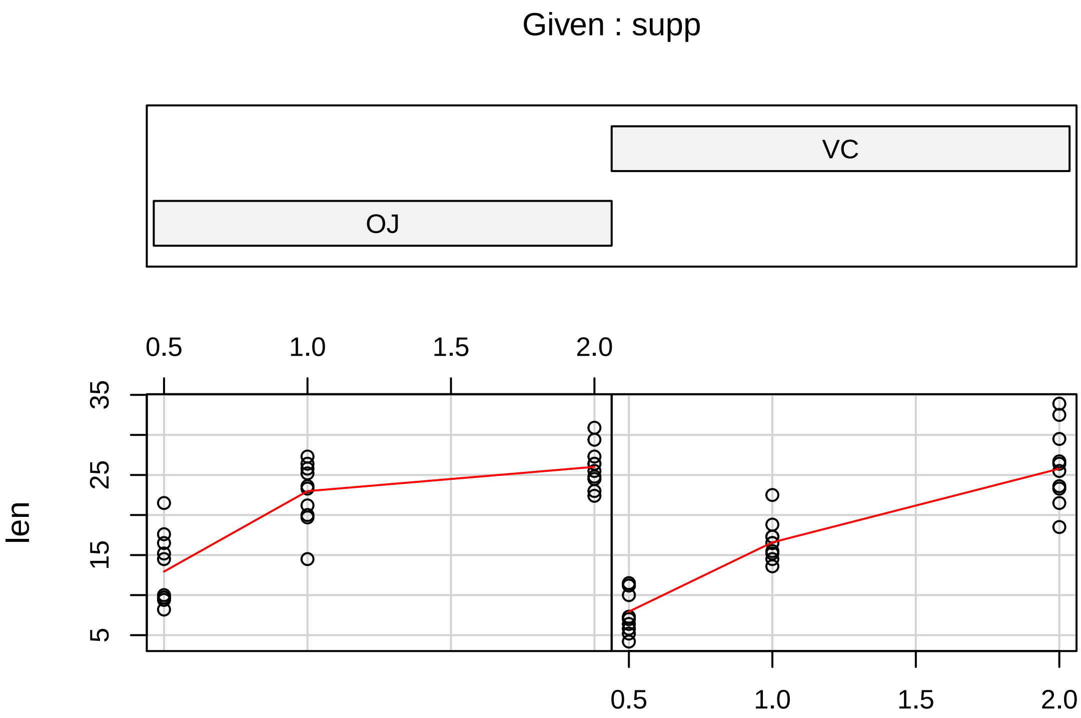
```
## [1] 16.837 20.789
## attr(,"conf.level")
## [1] 0.95
```

# Conclusions

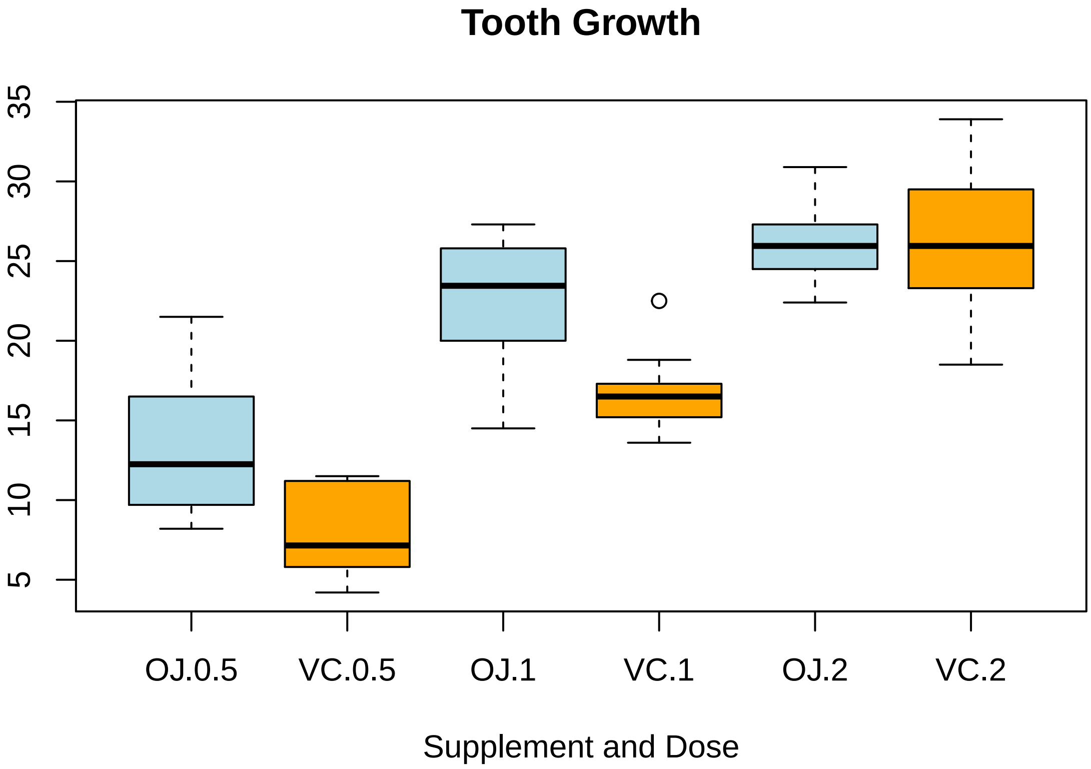
From the results of the exploratory and data analysis, the following conclusions can be drawn:

1. The distribution of growth length in odontoblasts is not nearly a normal distribution.
2. Without knowing the population standard deviation, the margin of error for tooth growth at a 95% confidence interval is **1.976** microns. The true mean has a 95% chance of being in the interval between **16.837, 20.789** microns.
3. There is a linear relationship between the amount of the respective dosages and the affects on tooth growth, irrespective of supplement type.
4. Ascorbic acid has the most effect on tooth growth at the highest dosage level; otherwise, Orange Juice provides a greater change overall.

## Appendix



ToothGrowth data: length vs. dose, given type of supplement


Tooth Growth
OJ.0.5
VC.0.5
OJ.1
VC.1
OJ.2
VC.2
Supplement and Dose