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:

```
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: **len** 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 **supp**: orange juice or ascorbic acid (a form of vitamin C and coded as VC).

The structure of this dataset is displayed using the following R command:

```
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 indicates that the dataset consists of 60 observations of 3 variables. 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 following **dplyr** command:

```
ToothGrowth %>% group_by(supp, dose) %>% summarise_each(funs(mean))
```

```
## Source: local data frame [6 x 3]
## Groups: supp [?]
##
##     supp  dose   len
##   (fctr) (dbl) (dbl)
## 1     OJ   0.5 13.23
## 2     OJ   1.0 22.70
## 3     OJ   2.0 26.06
## 4     VC   0.5  7.98
## 5     VC   1.0 16.77
## 6     VC   2.0 26.14
```

A quick analysis of this table hints that, at the two lower doses, Orange Juice results in greater tooth growth and that the results are equivalent for both supplements at the 2.0 mg/day dosage.

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.

```
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.

```
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)
```

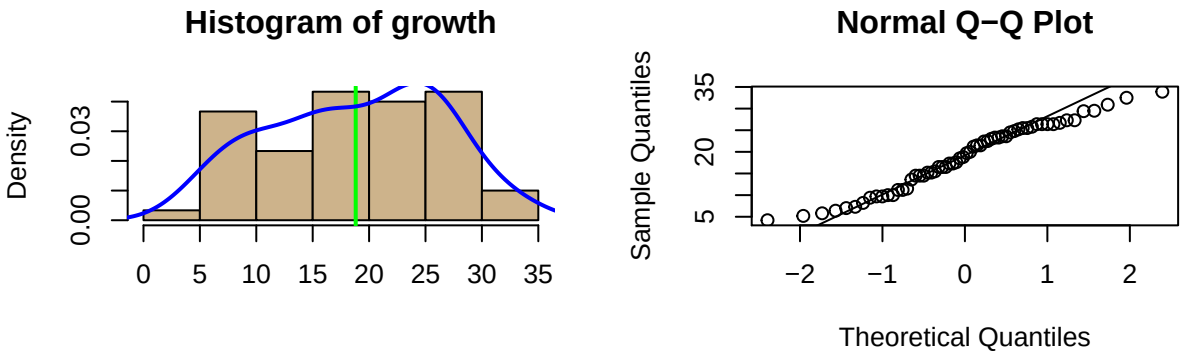
```
##             len
## skew     -0.143
## kurtosis -1.043
```

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:

```
growth <- ToothGrowth$len
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. One such popular choice is the **boxplot** (see *Appendix*), also known as a *box-and-whiskers* plot.

Observations on EDA artifacts

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 micron range.
- The density curve, while not normal, is still relatively mesokurtic.
- There is a single outlier evident on the boxplot for the 1.0 mg/day dosage of ascorbic acid.
- Ascorbic acid in the 1.0 and 2.0 mg/day dosages exhibit the most symmetry.

Hypothesis Testing and Confidence intervals

There are numerous hypotheses that can be developed from this dataset. This report will test Orange Juice versus Ascorbic Acid by Dose.

H_0: Tooth growth is the same for both delivery methods at the given dosage of Vitamin C

H_a: Tooth growth is probably impacted by the delivery method at the given dosage of Vitamin C

For the 0.5 Dosage

```
dose.0.5 <- t.test(len~supp,data=ToothGrowth[ToothGrowth$dose == 0.5, ], paired = F, var.equal = F)
dose.0.5$conf.int
```

```
## [1] 1.719057 8.780943
## attr(,"conf.level")
## [1] 0.95
```

```
dose.0.5$p.value
```

```
## [1] 0.006358607
```

Since the p-value is very much less than 0.05, H_0 is rejected at the 95% significance level. This indicates evidence that orange juice has a measurable impact on guinea pig tooth growth for a 0.5 mg dosage.

For the 1.0 Dosage

```
dose.1.0 <- t.test(len~supp,data=ToothGrowth[ToothGrowth$dose == 1.0, ], paired = F, var.equal = F)
dose.1.0$conf.int
```

```
## [1] 2.802148 9.057852
## attr(,"conf.level")
## [1] 0.95
```

```
dose.1.0$p.value
```

```
## [1] 0.001038376
```

This t.test gives the same result as for the 0.5 one.

For the 2.0 Dosage

```
dose.2.0 <- t.test(len~supp,data=ToothGrowth[ToothGrowth$dose == 2.0, ], paired = F, var.equal = F)
dose.2.0$conf.int
```

```
## [1] -3.79807  3.63807
## attr(,"conf.level")
## [1] 0.95
```

```
dose.2.0$p.value
```

```
## [1] 0.9638516
```

Since the p-value is very close to 1.0, H_0 is **not** rejected at the 95% significance level. This indicates evidence that orange juice has no significant impact on guinea pig tooth growth for a 2.0 mg dosage.

Assumptions

- The distribution of growth length in odontoblasts is not nearly a normal distribution. An explanation as to why this is so comes from the low contribution to growth rate from the 1.0 mg/day ascorbic acid supplement. This component of the data also contains the lone outlier.
- Guinea Pigs are similar as a population and this is a random population of Guinea Pigs.

Conclusions

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

- There is a linear relationship between the amount of the respective dosages and the affects on tooth growth, irrespective of supplement type.
- Ascorbic acid has the most effect on tooth growth at the highest dosage level. Orange juice provides a greater change at the 0.5 and 1.0 mg/day dosage levels.

Appendix