# Analyze the ToothGrowth data in the R datasets package

By: Oleg Rozenshtain

## Overview

This report is written as a solution to the second part of the "statistical inference" course project, as part of the "data science" specialization by Johns Hopkins University on Coursera.

The data was taken out of a study by E. W. Crampton[1], which tried to check if obtaining vitamin C in a chemically pure way, can replace obtaining the vitamin via a biological way. The problem of the assay of this vitamin was of particular concern to the Canadian Government during the Second World War because of the difficulty of providing natural sources of vitamin C to the armed forces for a considerable portion of the year.

The data consists of length measurements of the odontoblast cells harvested from the incisor teeth of a population of 60 guinea pigs. The vitamin C was given in three dose levels: 0.5, 1, and 2 mg (doses less than 0.5 mg shown frank scurvy, doses more than 2 mg didn't have increased response). The vitamin was delivered either by an ascorbic acid (chemical way, denoted by VC) or by orange juice (biological way, denoted by OJ). The guinea pigs were divided into 6 groups of 10, with equal numbers of each sex on each dose level, and consistently fed a diet with one of 6 vitamin C supplement regimes for a period of 42 days.

## Data description

First load ToothGrowth data set from the R datasets package.

```
> library(datasets)
> data(ToothGrowth)
```

Now let's observe the structure of the data:

```
> 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 ...

> table(ToothGrowth$supp, ToothGrowth$dose)

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

The data has 60 observations, each one for a different guinea pig with a specific vitamin C supplement regime. The data considers three variables:

1. len: a numeric variable indicates the length in microns of the odontoblast cells harvested from the incisor teeth of the guinea pigs after 42 days.

[1] E. W. Crampton 1946, The growth of the odontoblasts of the incisor tooth as a criterion of the vitamin C intake of the guinea pig. The Journal of Nutrition.

2. supp: a factorial variable with 2 levels (OJ, VC), indicates vitamin C supplement method.
3. dose: a numeric variable indicates vitamin C delivery dosage, 0.5, 1, or 2 milligrams.

The *table()* function tells me that the 60 guinea pigs are divided into 6 vitamin C supplement even groups of 10 guinea pigs each, as stated in the overview.

**Data summary and exploratory analyses**

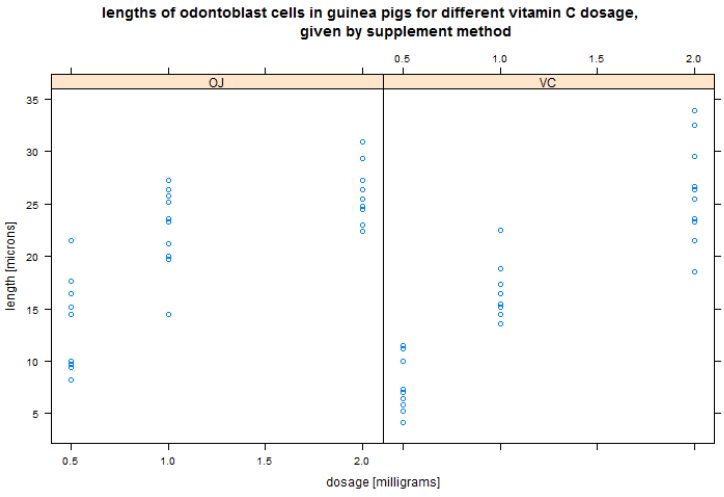
Now I can look at the data summary:

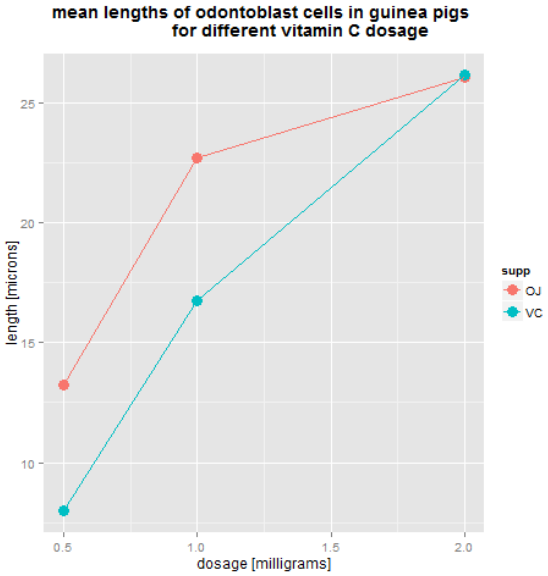
*> aggregate(len ~ supp + dose, data = ToothGrowth, summary)*

| | supp | dose | len.Min. | len.1st Qu. | len.Median | len.Mean | len.3rd Qu. | len.Max. |
|---|---|---|---|---|---|---|---|---|
| 1 | OJ | 0.5 | 8.20 | 9.70 | 12.25 | 13.23 | 16.18 | 21.50 |
| 2 | VC | 0.5 | 4.20 | 5.95 | 7.15 | 7.98 | 10.90 | 11.50 |
| 3 | OJ | 1.0 | 14.50 | 20.30 | 23.45 | 22.70 | 25.65 | 27.30 |
| 4 | VC | 1.0 | 13.60 | 15.27 | 16.50 | 16.77 | 17.30 | 22.50 |
| 5 | OJ | 2.0 | 22.40 | 24.58 | 25.95 | 26.06 | 27.08 | 30.90 |
| 6 | VC | 2.0 | 18.50 | 23.38 | 25.95 | 26.14 | 28.80 | 33.90 |

The median and the mean are pretty close which means that the length is not skewed and that the data doesn't contain outliers (can also be deducted out of the min and max values). I also notice a positive correlation between the dosage and the length. Generally the OJ supplement type leads to a bigger dontoblast cells length than VC supplement.

These initial conclusions can be inspected visually:



**Analysis and conclusion**

In order to check whether chemically obtaining vitamin C has the same effect as biologically, I will use t-test comparing the two supplement ways for each dose level. The null hypothesis is set to $H_0: \mu_{OJ} = \mu_{VC}$. For t-test usage I assume that the original population from which the samples were drawn is normally distributed. The samples histograms (appendix A) don't show any kind of mound-shaped distribution, but I assume it is because of the relatively small sample size. Furthermore, I assume the population variance is equal. It is a fair assumption because it is a randomized trial in which subjects were divided randomly into groups, and a lot of efforts were invested in keeping all parameters the same except vitamin C supplement (including parents and subjects diet). I left the default value of parameter *paired = FALSE* because the samples include different guinea pigs, and I kept the *conf.level = 0.95*. I used a two sided test because I wanted to find out any difference in the mean, either upwards or downwards because both ways may cause negative effect, so the alternative hypothesis is $H_a: \mu_{OJ} \neq \mu_{VC}$. I used *saplly()* to perform a *t.test()* on each one of the dose levels, and extracted the confidence interval and the p-value:

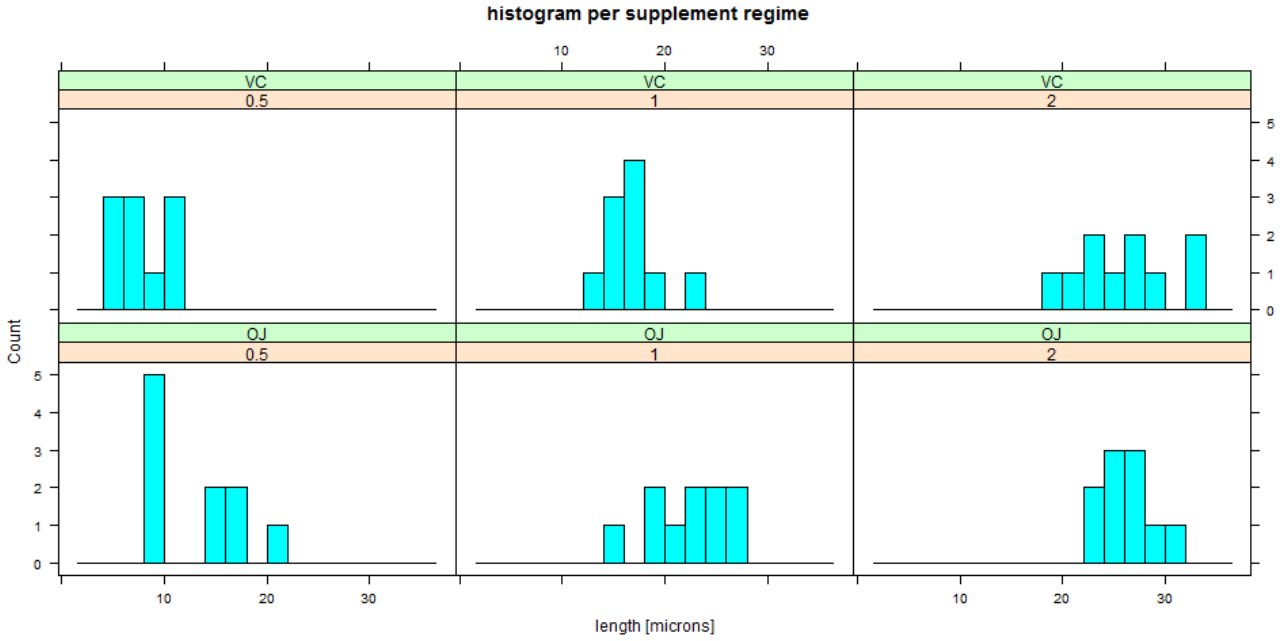
```
> t(sapply(c(dose0.5=0.5, dose1.0=1, dose2.0=2),
+          function(d) with(ToothGrowth,
+                           t.test(len[(supp == "OJ") & (dose == d)],
+                                  len[(supp == "VC") & (dose == d)],
+                                  var.equal = TRUE))$conf))
             [,1]     [,2]
dose0.5  1.770262 8.729738
dose1.0  2.840692 9.019308
dose2.0 -3.722999 3.562999

> sapply(c(dose0.5=0.5, dose1.0=1, dose2.0=2),
+          function(d) with(ToothGrowth,
+                           t.test(len[(supp == "OJ") & (dose == d)],
+                                  len[(supp == "VC") & (dose == d)],
+                                  var.equal = TRUE))$p.value))
          dose0.5      dose1.0     dose2.0
[1,] 0.005303661 0.0007807262 0.9637098
```

My conclusions from the analysis are that for doses 0.5 and 1 mg $H_0$ is rejected and $\mu_{OJ} \neq \mu_{VC}$. This can be deducted from the fact that zero isn't in the confidence interval, and p-value<$\alpha(= 0.05)$. For 2 mg dose $H_0$ is failed to be rejected (insufficient evidence to reject it) because zero is within the confidence interval and p-value is very high, p-value>$\alpha$.

The bottom line is that for high vitamin C dosage consumption chemical supplement such as ascorbic acid can replace a biological delivery. Though, for low dosages of vitamin C chemical supplement has different effect from a biological one.

# Appendix A


histogram per supplement regime
VC
0.5
VC
1
VC
2
OJ
0.5
OJ
1
OJ
2
Count
length [microns]

# Appendix B

```
investigateToothGrowth<-function()
{
  library(datasets)
  data(ToothGrowth)

  print("Data description:")
  cat("\n")
  print(str(ToothGrowth))
  # count number of observations for each (supp,dose) combination
  print(table(ToothGrowth$supp, ToothGrowth$dose))
  cat("\n")

  print("Data summary:")
  cat("\n")
  # get summary of len for each (supp,dose) combination
  print(aggregate(len ~ supp + dose, data = ToothGrowth, summary))
  cat("\n")

  library(lattice)
  png("lengths_of_odontoblast_cells.png", width = 720)
  print(xyplot(len ~ dose | supp, data = ToothGrowth,
               main = "lengths of odontoblast cells in guinea pigs for different vitamin C dosage,
               given by supplement method" ,
                 xlab = "dosage [milligrams]", ylab = "length [microns]"))
  dev.off()

  library(ggplot2)
  png("means_of_lengths.png")
  print(ggplot(aggregate(len ~ supp + dose, data = ToothGrowth, mean),
                   aes(x = dose, y = len, colour = supp, group = supp)) +
        geom_point(size = 4) + geom_line() +
        theme(plot.title = element_text(face = "bold", vjust = 2)) +
        labs(title = "mean lengths of odontoblast cells in guinea pigs
              for different vitamin C dosage",
              x = "dosage [milligrams]", y = "length [microns]"))
  dev.off()

  png("lengths_histogram.png", width = 960)
  print(histogram(~len | factor(dose)*supp, data = ToothGrowth,
                       layout = c(3,2), breaks = seq(2,36,2), type = "count",
                       xlab = "length [microns]",
                       main = "histogram per supplement regime"))
  dev.off()
```

```
  # perform t-test to compare OJ mean with VC mean over all dosage leveles.
  # extract first the confidence intervals, and second the p-value.

  cat("\n")
  print("95 percent confidence interval:")
  print(t(sapply(c(dose0.5=0.5, dose1.0=1, dose2.0=2),
                 function(d) with(ToothGrowth,
                                  t.test(len[(supp == "OJ") & (dose == d)],
                                         len[(supp == "VC") & (dose == d)],
                                         var.equal = TRUE))$conf)))
  cat("\n")
  print("p-values:")
  print(sapply(c(dose0.5=0.5, dose1.0=1, dose2.0=2),
               function(d) with(ToothGrowth,
                                t.test(len[(supp == "OJ") & (dose == d)],
                                       len[(supp == "VC") & (dose == d)],
                                       var.equal = TRUE))$p.value))
}
```