STI-Project 2. Research on the effect of Vitamin C on Tooth Growth

Gerrit Versteeg

27 November 2016

1. Synopsis (Overview)

This report investigates the influence of Vitamin C on the growth of teeth of guinea pigs. The research compares the length of the teeth for the influence of the method of delivery (supp) as well as the dose applied daily (dose). The data is considered to be derived from iid samples, non-paired but with equal variances.

The results indicate that the method of delivery has no measurable influence on teeth length, **but the dose of vitamin C has a significant influence** (using a type 1 error level α of 0.05).

2. Loading and preparing the data

First step is to load the necessary libraries and the 'ToothGrowth' data from the library 'datasets'. We then put the data into a data frame (tibble).

```
library("dplyr", warn.conflicts=FALSE)      ## for data prep
library("pastecs", warn.conflicts=FALSE)    ## for data analysis
```

```
## Loading required package: boot
```

```
library("ggplot2", warn.conflicts=FALSE)    ## for the graphs
library("datasets", warn.conflicts=FALSE)   ## for the data iself
DF <- tbl_df(ToothGrowth)                   ## read data into tibble
attach(DF)                                  ## to reference column-names
```

The data consists of three variables for 60 individual pigs:

- len (numeric): showing the length of the incisor tooth
- supp (factor, 2 levels): the delivery method, OJ (Orange Juice) or VC (Vitamine C).
- dose (factor, 3 levels): reference to the dose (0.5, 1.0 or 2.0 mg/day)

2. Exploratory Data Analysis

Lets have a look at the measurement variable (len) and its properties.

```
stat.desc(len)     ## basic properties of variable len
```

```
##      nbr.val     nbr.null       nbr.na          min          max
##   60.0000000    0.0000000    0.0000000    4.2000000   33.9000000
##        range          sum       median         mean      SE.mean
##   29.7000000 1128.8000000   19.2500000   18.8133333    0.9875223
## CI.mean.0.95          var      std.dev     coef.var
##    1.9760276   58.5120226    7.6493152    0.4065901
```

There are no missing values (NULL, NA's), so nothing needs to be imputed via some imputing strategy. The mean of the data is 18.8, with a standard deviation of 7.65. The data varies from 4.2 up to 33.9. Let's have a look at the distribution of the data, using a histogram (see appendix, figure 1).

The graph shows a wide spread and a certain central tendency, but it may be difficult to call this normally distributed. So, lets check with a QQ-plot (see appendix, figure 2).

Looking at the QQ-plot, the distribution looks somewhat 'tailed' (lower Q's are up, higher Q's are down) and is therefore not convincingly 'normal'. To understand better, there is a variety of tests we can use to measure the 'normality' of our data. In cases of a limited number of observations (n=60), the one prefered is the 'Shapiro-Wilk'-test (see appendix, Figure 3). The resulting p-value is 0.11. Which is even higher than an alpha level of 0.1 (90%).

Based on the checks, we need to reject the hypothesis that these data have been drawn from a 'normal' distributed population.

3. Key Assumptions

Looking at the SW-test, it appears the data do not acceptably follow a 'normal' distribution. Therefore we will need to refer to t distributions for smaller sample sizes (T student's test). Also we do not have any 'prior to treatment' observations to compare the effects of the dose/supp-variables on the pigs. They only way to analyze the data is to compare tooth length between treatments (dose and/or delivery method). Because 60 individual pigs were measured once, there are **no paired** observations. A key assumption is that the pigs' tooth length measurements grouped by method of delivery and/or dose represent groups with **equal variances**. It is also assumed that the observations or **iid** (independent and identically distributed). We will be using the commonly used **error level**: $\alpha = 0.05$.

4. Hypothesis Tests

Testing the delivery method (supp)

The main null hypothesis is that the *delivery method* does **not** have any influence on the length of teeth:

$H_0 : \mu_{oj} = \mu_{vc}$

$H_a : \mu_{oj} \neq \mu_{vc}$ therefore $|TS| \geq t_{1-\alpha/2}$

The alternative hypothesis is that there is a significant difference between the average tooth length when varied by delivery method.

The test statistic and the pooled variance for grouped data to be used is:

$TS = \frac{X_{oj}-X_{vc}}{S_p\sqrt{1/n_{oj}+1/n_{vc}}}$

$S_p^2 = \frac{(n_{oj}-1)S_{oj}^2+(n_{vc}-1)S_{vc}^2}{n_{oj}+n_{vc}-1}$

```
g1 <- len[supp == "OJ"]                       ## let group 1 be pigs having OJ
m1 <- mean(g1); s1 <- var(g1); n1 <- length(g1)
g2 <- len[supp == "VC"]                       ## let group 2 be pigs having VC
m2 <- mean(g2); s2 <- var(g2); n2 <- length(g2)
a <- 0.05                                     ## alpha level = 0.05
Sp <- sqrt(((n1-1)*s1^2+(n2-1)*s2^2)/(n1+n2-2)) ## determine pooled variance
TS <- (m1 - m2)/(Sp*sqrt(1/n1 + 1/n2))        ## determine test statistic
t_dof <- qt((1-a/2), (n1+n2-2))               ## t-quantile: 1-alpha/2, DoF = 58
tCI <- m2-m1 + c(-1,1)*t_dof*Sp*sqrt(1/n1+1/n2) ## determine T conf. interval
print(c(TS, t_dof, tCI))                      ## print TS, t_dof and t Conf.Int
```

```
## [1]   0.2499755   2.0017175 -33.3283201  25.9283201
```

Because our test statistic TS (0.25) is smaller than the t-quantile (2.00) we fail to reject the null-hypothesis. The 95% T confidence interval ranges from -33.33 to 25.93 clearly containing the 0. So the case that the difference in tooth length based on the method of delivery is zero, is within the accepted probability. Therefore the change in means of 'len' is not significantly explained by the delivery method.

Testing the applied dose of vitamin C (dose)

The main null hypothesis is that the *dose* does **not** have any influence on the length of teeth. We will test the difference between the group of pigs with dose 0.5 mg/day versus the group of pigs with dose 2.0 mg/day, bacause this is the largest dose variation:

$H_0 : \mu_{0.5} = \mu_{2.0}$

$H_a : \mu_{0.5} \neq \mu_{2.0}$ -> $|TS| \geq t_{1-\alpha/2}$

The alternative hypothesis is that there is a (significant) difference between the avarage tooth length based on the dose of vitamins applied.

The test statistic for the regrouped data to be used is:

$TS = \frac{X_{0.5}-X_{2.0}}{S_p\sqrt{1/n_{0.5}+1/n_{2.0}}}$

$S_p = \frac{(n_{0.5}-1)S_{0.5}^2+(n_{2.0}-1)S_{2.0}^2}{n_{0.5}+n_{2.0}-1}$

```
g1 <- len[dose == 0.5]                    ## group 1 are pigs dosed 0.5 mg/day
m1 <- mean(g1); s1 <- var(g1); n1 <- length(g1)
g2 <- len[dose == 2.0]                    ## group 1 are pigs dosed 2.0 mg/day
m2 <- mean(g2); s2 <- var(g2); n2 <- length(g2)
a <- 0.05                                 ## alpha level = 0.05
Sp <- sqrt(((n1-1)*s1^2+(n2-1)*s2^2)/(n1+n2-2)) ## determine pooled variance
TS <- (m1 - m2)/(Sp*sqrt(1/n1 + 1/n2))    ## determine test statistic
t_dof <- qt((1-a/2), (n1+n2-2))           ## t-quantile: 1-alpha/2, DoF = 38
tCI <- m2-m1 + c(-1,1)*t_dof*Sp*sqrt(1/n1+1/n2) ## determine T conf. interval
print(c(TS, t_dof, tCI))                  ## print TS, t_dof and t Conf.Int
```

```
## [1] -2.799117  2.024394  4.288614 26.701386
```

Looking at the influence of the dose on the length of the teeth, we see that our new test statistic absolute TS (2.80) is larger than the t-quantile belonging to $1 - \alpha/2 = 0.975$ and 38 degrees of freedom (2.02). The 95% T confidence interval ranges from 4.29 to 26.70 and does NOT contain the 0. So the case that the difference in tooth length based on the dose of vitamin C is zero, is not within the accepted probability range. Therefore we now reject the null-hypothesis in favor of the alternative hypothesis. The change in means of 'len' is significantly explained by the dose.

5. Conclusions

Looking at the differences in the means of the length of teeth, the method of delivery (Orange Juice versus Vitamin C) does not have an effect. The dose however, has a signifant influence on the length of teeth, comparing the pigs with a dose of 0.5 mg/day versus the pigs with a dose of 2.0 mg/day and taking into account a normal error level ($\alpha = 0.05$).

APPENDIX

Figure 1. Histogram of len variable

```
hist(len)                               ## to look at the ditribution of 'len'
```

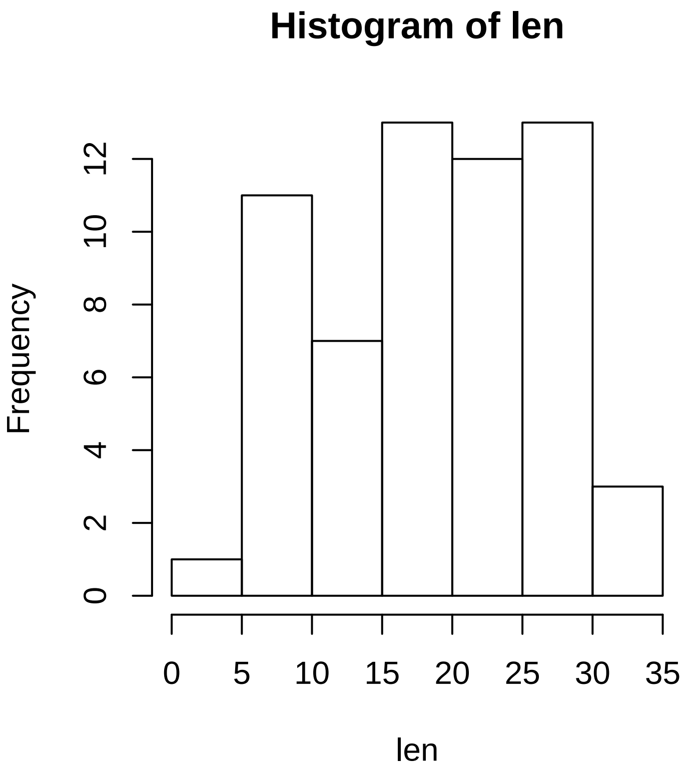

Figure 2. QQ-plot of distribution of len versus normal distribution

```
qqnorm(len); qqline(len, col = 2)       ## plot the quantiles
```

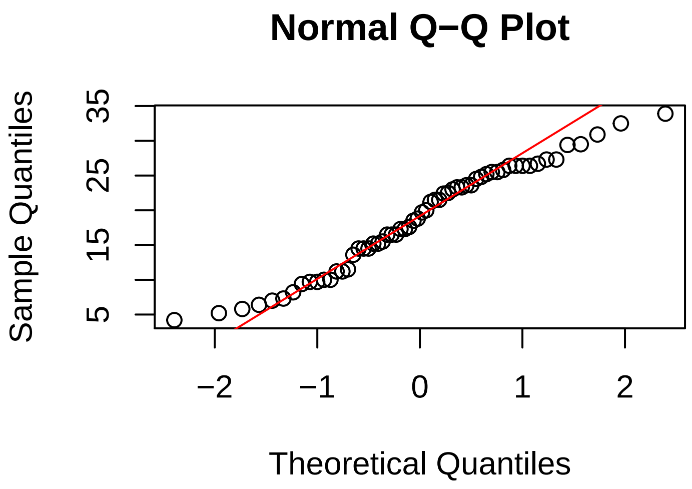

Figure 3. Shapiro-Wilk test for normal distribution comparison

```r
shapiro.test(len)      ## check for normal distribution by Shapiro-Wilk test
```

```
##
##  Shapiro-Wilk normality test
##
## data:  len
## W = 0.96743, p-value = 0.1091
```