# Basic inferential data analysis

*Katica Ristic*

*July 2, 2017*

## Overview

Basic inferential data analysis will be prefomed on the dataset called **ToothGrowth** - The Effect of Vitamin C on Tooth Growth in Guinea Pigs. Data has measures of the response set as 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).

## Looking into data

ToothGrowth is a data frame with 60 observations on 3 variables, **len** - tooth length ( numeric ), **supp** - supplement type VC or OJ ( factor ) and **dose** - dose in milligrams/day ( numeric ).

**Table 1.**

```
##       len        supp         dose
##  Min.   : 4.20   OJ:30   Min.   :0.500
##  1st Qu.:13.07   VC:30   1st Qu.:0.500
##  Median :19.25           Median :1.000
##  Mean   :18.81           Mean   :1.167
##  3rd Qu.:25.27           3rd Qu.:2.000
##  Max.   :33.90           Max.   :2.000
```

From Table 1. we can see the range of the Tooth Length (`len`) . More sense of this variable is given by the histogram below, which is weighted by the black line representing the sample mean.

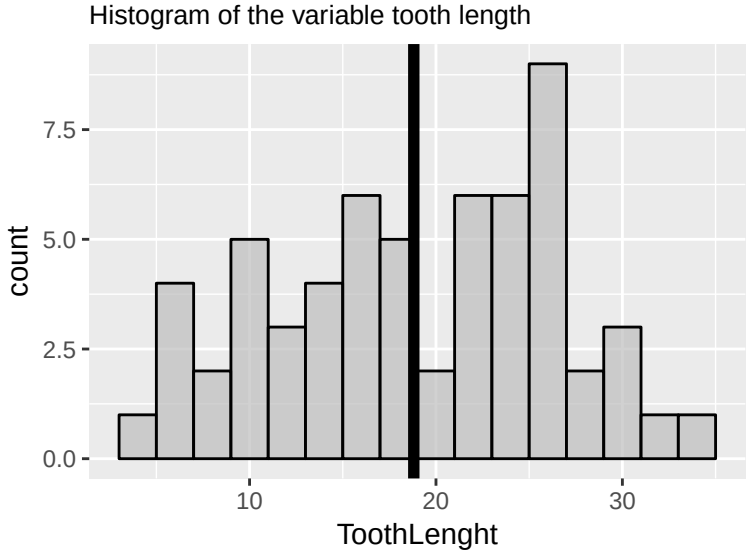
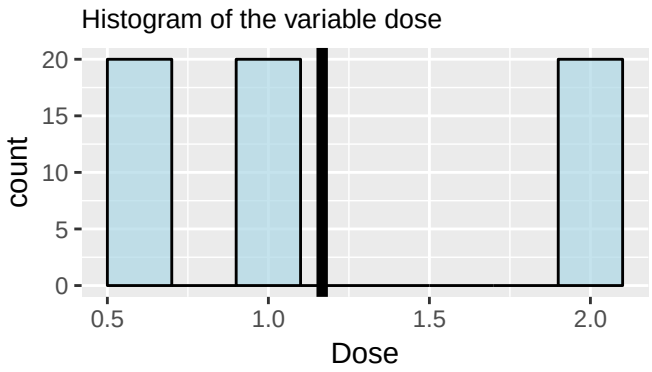



Table 1. also gives us information that the supplement type is the same between 30 guinea pigs. For the `dose` the next histogram shows us that there are three groups of 20 guinea pigs with the same dose. Histogram is also weighted by the black line representing the sample mean.




Let us now look at the means of the grouped data. The means are calculated using bootstrap ( $B = 10000$). The standard error and the 95% confidence intervals for the means are calculated as well. This helps us see how the data of the sample differs in the groups.

**Table 2.**

```
##                                mean    se   2.5%  97.5%
## Tooth Lenght - All           18.796 0.961 16.918 20.682
## Tooth Lenght - supplement VC 16.972 1.487 14.133 19.893
## Tooth Lenght - supplement OJ 20.652 1.198 18.227 22.947
## Tooth Lenght - dose 0.5      10.617 0.982  8.765 12.605
## Tooth Lenght - dose 1        19.746 0.979 17.845 21.645
## Tooth Lenght - dose 2        26.104 0.814 24.520 27.710
```

## Statistical inference

The differrence between the means of the gouped data 'Tooth Lenght - supplement VC' and 'Tooth Lenght - supplement OJ' is 3.6963 calculated from Table 2. Now let us do a t.test for this difference.

**Table 3.**

```
##    2.5%   97.5% p.value    mean
##  1.4087  5.9913  0.0025  3.7000
```

As you can see, the confidence interval doesn't even contain zero, so we will reject the null hypotisis $H_0 : \mu = 0$, and decide for the alternative hypotisis that the means of this two groups are diffrent. Also the p value is less then 0.05, so one more parameter help as reject the null hypotisis. Notice that the mean of the difference is prety close to the difference between the means got with bootstraping ( calculated in Table 2. ). All od these conclutions make us conclude one thing - **there is a significant influence of delivery methods on the length of odontoblasts** ( Tooth Lenght ).

As you can see in Table 2., the means of the groups made by dose differ a lot and thair confidence intervals do not intersect at all. Let us now do t.test comparing these differences.

**Table 4.**

```
##                            2.5%   97.5% p.value   mean
## dose 1 versus dose 0.5  6.3871 11.8729   0e+00  9.130
## dose 2 versus dose 1    3.4718  9.2582   2e-04  6.365
## dose 2 versus dose 0.5 12.6228 18.3672   0e+00 15.495
```

As we expected the differeces are very large and p values almost 0, so we can conclude that the **dosage significatly influence on the length of odontoblasts** ( Tooth Lenght ).

## Summary notes

- Basic inferential data analysis was prefomed on the dataset called ToothGrowth
- Tooth Length ( `len` ) was analysed trought all the groups, made by supplement and by dosage
- Using t.test we concluded that there is significant influence of delivery methods and dosage on the length of odontoblasts

## Appendix

Code in R for the first histogram:

```
library(UsingR)
library(ggplot2)
data("ToothGrowth")

d<-ToothGrowth
dat <- data.frame(ToothLenght = d$len)

g <- ggplot(dat, aes(x = ToothLenght))
     + geom_histogram(alpha = .7, binwidth= 2, fill= "grey", colour = "black")
g <- g + geom_vline(xintercept = 18.81, size = 2)
g <- g + labs(subtitle="Histogram of the variable tooth length",face="bold")
g
```

Code in R for the second histogram:

```
library(UsingR)
library(ggplot2)
data("ToothGrowth")

d<-ToothGrowth
dat <- data.frame(Dose = d$dose)

g <- ggplot(dat, aes(x = Dose))
     + geom_histogram(alpha = .7, binwidth= .2, fill= "lightblue", colour = "black")
g <- g + geom_vline(xintercept = 1.167, size = 2)
g <- g + labs(subtitle = "Histogram of the variable dose",face="bold")
g
```

Code in R for **Table 2.**

```
library(UsingR)
library(ggplot2)
data("ToothGrowth")
d<-ToothGrowth

resamples<-function(x,n){ m<- matrix(sample(x,10000*n,replace = TRUE),10000,n)
                          mn<-apply(m,1,mean)
                          c(mean(mn),sd(mn),quantile(mn,c(.025,.975))[[1]],
                          quantile(mn,c(.025,.975))[[2]])
                          }

l<-list(len=d$len,lenVC=d[1:30,1],lenOJ=d[31:60,1],
        "lendosehalf"=d[d$dose==.5,1],
        "lendoseone"=d[d$dose==1,1],
```

```
        "lendosetwo"=d[d$dose==2,1])

exit<-NULL
for(i in 1:6) exit<-rbind(exit,resamples(l[[i]],length(l[[i]])))

dimnames(exit)<-list(c("Tooth Lenght - All",
                       "Tooth Lenght - supplement VC",
                       "Tooth Lenght - supplement OJ",
                       "Tooth Lenght - dose 0.5",
                       "Tooth Lenght - dose 1",
                       "Tooth Lenght - dose 2"),
                       c("mean","se","2.5%","97.5%"))

round(exit,3)
```

Code in R for **Table 3.**

```
library(UsingR)
library(ggplot2)
data("ToothGrowth")
d<-ToothGrowth

l<-list(lenVC=d[1:30,1],lenOJ=d[31:60,1],
        "lendosehalf"=d[d$dose==.5,1],
        "lendoseone"=d[d$dose==1,1],
        "lendosetwo"=d[d$dose==2,1])

round(c("2.5%"=t.test(l[[2]]-l[[1]])$conf[1],
        "97.5%"=t.test(l[[2]]-l[[1]])$conf[2],
        "p.value"=t.test(l[[2]]-l[[1]])$p.value,
        "mean"= t.test(l[[2]]-l[[1]])$estimate[[1]]),4)
```