Johns Hopkins Coursera - Statistical Inference - Project Part 2

shostiou

06/09/2020

## Part 02 - Basic Inferential Data Analysis

This second part of the project is dedicated to the analysis of the ToothGrowth dataset (embedded in the R package).

### Data upload & basic exploration

Let’s start by loading the dataset and by doing basic explorations.  
Refering to the help associated to the data, this dataset is describe as : **“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).”**

Content :  
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

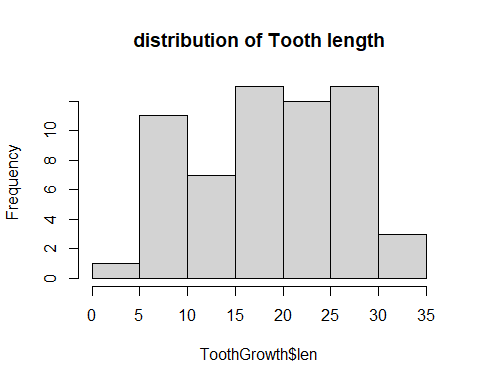
# Loading the dataset  
data(ToothGrowth)   
# Getting general overview of the data   
str(ToothGrowth)

## '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 ...

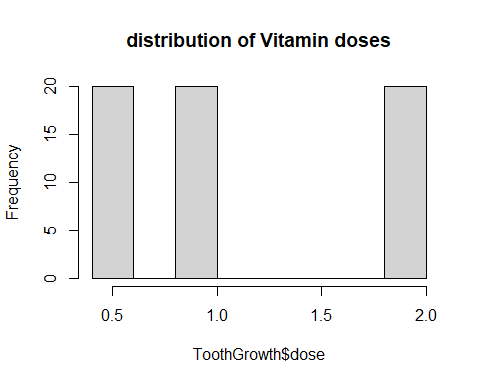
# Getting summary statistics.  
summary(ToothGrowth)

## 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

# Let's visualize the distribution of the numerical values  
hist(ToothGrowth$len, main = "distribution of Tooth length")



hist(ToothGrowth$dose, main = "distribution of Vitamin doses")



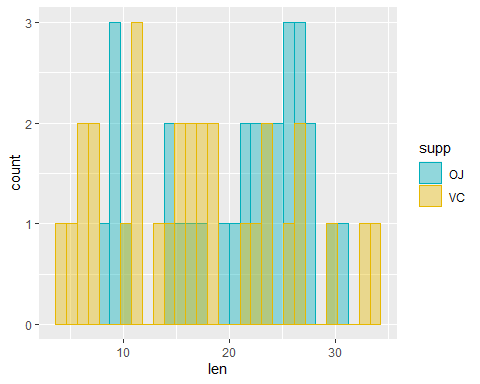
# Nb of rows  
df\_nb\_rows = as.numeric(nrow(ToothGrowth))

We can observe that the data contains 2 groups of observations. Each of those groups is made of 30 observation (Orange Juice / Ascorbic Acid).  
Even though the vitamin doses is numerical data, the values observed in the dataset is limited to 3 levels : 0.5; 1 ; 2. We can consider this variable as a discrete (not coninuous).

Let’s pursue the exploration by comparing the distribution of Tooth length based on the delivery method used (orange Juice - OJ ; ascorbid acid - VC)

# We call use ggplot2 to show the distributions   
library(ggplot2)  
# Building the histograms based on supp variable   
ggplot(ToothGrowth, aes(x = len)) +  
 geom\_histogram(aes(color = supp, fill = supp),   
 position = "identity", alpha = 0.4) +  
 scale\_color\_manual(values = c("#00AFBB", "#E7B800")) +  
 scale\_fill\_manual(values = c("#00AFBB", "#E7B800"))

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

 The observation of the 2 overlapped distributions doesn’t provide a key evidence between Tooth lengths & the delivery method used during the experiment.

### Comparing the means

Let’s compare the means of the tooth length distribution based on the delivery method.

# Let's call the dplyr package  
library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

# means & standard error computation   
ToothGrowth %>% group\_by(supp) %>% summarize(mean\_len = mean(len), std\_len = sd(len))

## `summarise()` ungrouping output (override with `.groups` argument)

## Warning: `...` is not empty.  
##   
## We detected these problematic arguments:  
## \* `needs\_dots`  
##   
## These dots only exist to allow future extensions and should be empty.  
## Did you misspecify an argument?

## # A tibble: 2 x 3  
## supp mean\_len std\_len  
## <fct> <dbl> <dbl>  
## 1 OJ 20.7 6.61  
## 2 VC 17.0 8.27

We can observe that mean values of the 2 distributions (tooth length with orange juice OJ / tooth length with ascorbic acis CV) differs.

### Hypothesis Test

Now to determine is this difference is statistically pertinent, we will use a Hypothesis test :  
H0 : xbar\_len\_OJ - xbar\_len\_CV = 0  
HA : xbar\_len\_OJ - xbar\_len\_CV != 0  
Note that the 2 groups will be considered as being unpaired.  
As a second assumption, we will consider a constant variance in the population.

Finally we will conclude the hypothesis test by computing the p-value associated to a confidence interval of 95%. The test to be used will be 2 sided test. Note : a two tail test will be used.  
Let’s use T distribution to calculate the p-value.

# Let's define specific dataframes for OJ and VC  
len\_OJ <- ToothGrowth %>% filter(supp == 'OJ') %>% select(len)  
len\_VC <- ToothGrowth %>% filter(supp == 'VC') %>% select(len)  
# Applying the T test using the R native command  
t.test(len\_OJ,len\_VC, alternative="two.sided", paired=FALSE, mu = 0, conf.level = 0.95)

##   
## Welch Two Sample t-test  
##   
## data: len\_OJ and len\_VC  
## t = 1.9153, df = 55.309, p-value = 0.06063  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -0.1710156 7.5710156  
## sample estimates:  
## mean of x mean of y   
## 20.66333 16.96333

**CONCLUSION :** as p-value > 0.05 we failed to reject the null hypothesis.  
The data doesn’t provide evidence of differences in means between the Orange Juice and Ascorbic Acid feeding methods.