Assignment 1 basic inferential analysis

Joris Schut

Saturday, February 28, 2015

Introduction

In this basic inferential analysis differnt levels in the supp and dose variables from the ToothGrowth dataset are compared to each other to see if they have similar means. This is done by using t-test. The process of loading the data, preparing the data, performing the t-tests and deriving conclusions from these test is described in this document.

1. Load the ToothGrowth data and perform some basic exploratory data analyses

First the libraries used for this study are loaded into R.

```
#Load libraries
library(datasets)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
##
## The following object is masked from 'package:stats':
##
##     filter
##
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

Second, the dataset is stored in the data variable.

```
#Read the data
data<-tbl_df(ToothGrowth)
```

Third, some basic exploratory data analysis is performed

```
#Exploratory data analysis
print(names(data))
```

```
## [1] "len"  "supp" "dose"
```

```
print(dim(data))
```

```
## [1] 60  3
```

```
print(head(data))
```

```
## Source: local data frame [6 x 3]
##
##    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
```

```
print(class(data$len))
```

```
## [1] "numeric"
```

```
print(class(data$supp))
```

```
## [1] "factor"
```

```
print(class(data$dose))
```

```
## [1] "numeric"
```

```
print(unique(data$supp))
```

```
## [1] VC OJ
## Levels: OJ VC
```

```
print(unique(data$dose))
```

```
## [1] 0.5 1.0 2.0
```

2. Provide a basic summary of the data

Summary data is provided by using the summary() function.

```
summary (data)
```

```
##       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
```

3. Use confidence intervals and/or hypothesis tests to compare tooth growth by supp and dose. (Only use the techniques from class, even if there's other approaches worth considering)

T-tests are used to determine if there is a statistacally significant difference between the different groups with in the supp variable (OJ & VC) and dose variable (0.5, 1.0 & 2.0) exists.

First, we will look at the supp variable. To prepare for the t-test the differnt levels of this variable are filtered by. Then the len column is selected and the variables are bound together by colomn.

```
#Create 2 data sets that contain the len entries filtered by the values of the sub variable
OJ <- filter(data, supp=="OJ") %>%
      select(len)
VC <- filter(data, supp=="VC") %>%
      select(len)

#Combine the two in a single variable
x1 <- cbind(OJ, VC)
```

After this is done a 2-sided t-test is carried out using the following hypotheses: - H0: Xbar1 equals Xbar2 - Ha: Xbar1 does not equals Xbar2

With X1 being the data related to OJ values in the supp variable and X2 being the data related to the VC values in the supp variable.

```
#Perform a two-sided t-test
t.test(x1[,1], x1[,2], alternative="two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  x1[, 1] and x1[, 2]
## 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
```

Second, we will look at the dose variable. To prepare for the t-test the differnt levels of this variable are filtered by. Then the len column is selected and the variables are bound together by colomn.

```
#Create 3 data sets that contain the len entries filtered by the values of the dose variable
dose0.5 <- filter(data, dose==0.5)%>%
            select(len)
dose1.0 <- filter(data, dose==1.0)%>%
            select(len)
dose2.0 <- filter(data, dose==2.0)%>%
            select(len)

#Combine the three in a single variable
x2 <- cbind(dose0.5, dose1.0, dose2.0)
```

After this is done three 2-sided t-test are carried out using the following hypotheses: - H0: Xbar1 equals Xbar2 H0: Xbar1 equals Xbar3 H0: Xbar2 equals Xbar3 - Ha: Xbar1 does not equals Xbar2 H0: Xbar1 does not equals Xbar3 H0: Xbar2 does not equals Xbar3

With X1 being the data related to 0.5 values in the dose variable, X2 being the data related to 1.0 values in the dose variable and X3 being the data related to 2.0 values in the dose variable.

```
#Perform a two-sided t-test
t.test(x2[,1], x2[,2], alternative="two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  x2[, 1] and x2[, 2]
## t = -6.4766, df = 37.986, p-value = 1.268e-07
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.983781  -6.276219
## sample estimates:
## mean of x mean of y
##    10.605    19.735
```

```
t.test(x2[,1], x2[,3], alternative="two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  x2[, 1] and x2[, 3]
## t = -11.799, df = 36.883, p-value = 4.398e-14
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -18.15617 -12.83383
## sample estimates:
## mean of x mean of y
##    10.605    26.100
```

```
t.test(x2[,2], x2[,3], alternative="two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  x2[, 2] and x2[, 3]
## t = -4.9005, df = 37.101, p-value = 1.906e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -8.996481 -3.733519
## sample estimates:
## mean of x mean of y
##    19.735    26.100
```

4 State your conclusions and the assumptions needed for your conclusions

With regard to the questions wherer the means of the observations of the 2 levels present in the subb variable are similar, it can be concluded that we reject H0 in favor of Ha as the p-value of the t-test is smaller than 0.05

(2.2e-16). The same conclusion can be drawn from the 3 t-test from the dose variable (p-values: 1.268e-07, 4.398e-14 and 1.906e-05).

Four assumptions have been made to arrive at these conlusions: the assumption the sample are indepent and identical distributed, the samples are taken from comparable groups, the samples are not paired and the variances were unequal. Given the data is from the same study, the first two assumptions can be considered reasonable. This is also true for the third assumption as data represent differnt persons. The fourth assumption was not tested but can easily be verified through the following block of code.

```
#Test if variances are equal
print(var(x1[,1])==var(x1[,2]))
```

```
## [1] FALSE
```

```
print(var(x2[,1])==var(x2[,2]))
```

```
## [1] FALSE
```

```
print(var(x2[,1])==var(x2[,3]))
```

```
## [1] FALSE
```

```
print(var(x2[,2])==var(x2[,3]))
```

```
## [1] FALSE
```

As these test all return FALSE it can be assumed variances are not equal.