Statistical inference on ToothGrowth dataset

Molina Rafidison

10 Mar 2016

Report

Preliminaries

Synopsis

Here we are going to analyze the ToothGrowth data, provided in the R `datasets` package. We want to investigate about how Vitamin C is related to tooth growth Guinea pig. The documentation is available here.

Packages

Load needed packages.

```
packages <- c("datasets", "dplyr", "ggplot2")
sapply(packages, require, character.only = TRUE, quietly = TRUE)
```

Getting and cleaning data

Load the dataset from the `datasets` package.

```
data(ToothGrowth); 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 ...
```

The documentation says that there are 3 dose levels of Vitamin C (0.5, 1.0 and 2.0) so let us convert it into a factor.

```
ToothGrowth$dose <- as.factor(ToothGrowth$dose)
```

Exploratory data analysis

Length

Look at how the odontoblasts respond in length to Vitamin C doses and types *(cf. Appendix Fig. 1)*.

The length increases along with the dosage. It seems like the orange juice is more efficient than the ascorbic acid (VC) until the dose level attains 2 mg/day.

Summary

Here is a basic summary of the data.

```r
summary(ToothGrowth)
```

```
##       len        supp     dose   
##  Min.   : 4.20   OJ:30   0.5:20  
##  1st Qu.:13.07   VC:30   1  :20  
##  Median :19.25           2  :20  
##  Mean   :18.81                   
##  3rd Qu.:25.27                   
##  Max.   :33.90                   
```

Then we provide a summary of the data giving the length mean for each dose level and delivery method for 60 tested Guinea pigs.

```r
by(ToothGrowth$len, INDICES = list(ToothGrowth$supp, ToothGrowth$dose), summary)
```

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

Statistical inference

Given the small sample of 60 Guinea pigs, we will use the t-test in the case of unequal variances for independent samples. It will be run with the independent, categorical variables: `supp` and `dose`.

Orange juice vs. ascorbic acid

The summary shows a greater length mean of odontoblasts with orange juice than VC.
Null hypothesis: Orange juice and VC have equal length mean of odontoblasts

Compare the length according to the supplement type *(cf. Appendix Test 1)*.

The p value is statistically not significant given a 5% significance level. Plus, 0 is in the confidence interval. We fail to reject the Null hypothesis and cannot conclude that orange juice is more or less efficient than ascrobic acid.

High dose vs. low dose levels

The summary shows that the higher the dose, the bigger the cells in mean. We will not compare the dose levels by supplement type as the test above does not valid any conclusion about one being more efficient than the other.
Null hypothesis: Low and high doses have equal length mean of odontoblasts

Compare the length according to the dose levels of Vitamin C 0.5 mg/day and 2 mg/day *(cf. Appendix Test 2)*, then 0.5 mg/day and 1 mg/day *(cf. Appendix Test 3)* and eventually 1 mg/day and 2 mg/day *(cf. Appendix Test 4)*.

Let us summarize the results we obtained *(cf. Appendix, Test summary)* .

```
##   doseL doseH       pValue lowerBoundCI upperBoundCI doseLMean doseHMean
## 1   0.5     1 4.397525e-14   -18.156167   -12.833833    10.605    26.100
## 2   0.5     2 1.268301e-07   -11.983781    -6.276219    10.605    19.735
## 3   1.0     2 1.906430e-05    -8.996481    -3.733519    19.735    26.100
```

The p value is also statistically significant given a 5% significance level. Plus, 0 is not in the confidence interval.
We reject the Null hypothesis and we turn in favor of the alternative hypothesis that suggests that low dose levels has different impacts on length mean than high ones.

Results

Conclusions

The hypothesis and tests allow two conclusions:
- The dose level of Vitamin C has an impact on the length of the odontoblasts: the higher the dose, the higher the mean;
- There is no evidence that one delivery method - or supplement type - is more efficient than the other: orange juice cannot be considered as being more powerful than ascorbic acid (VC).

Assumptions

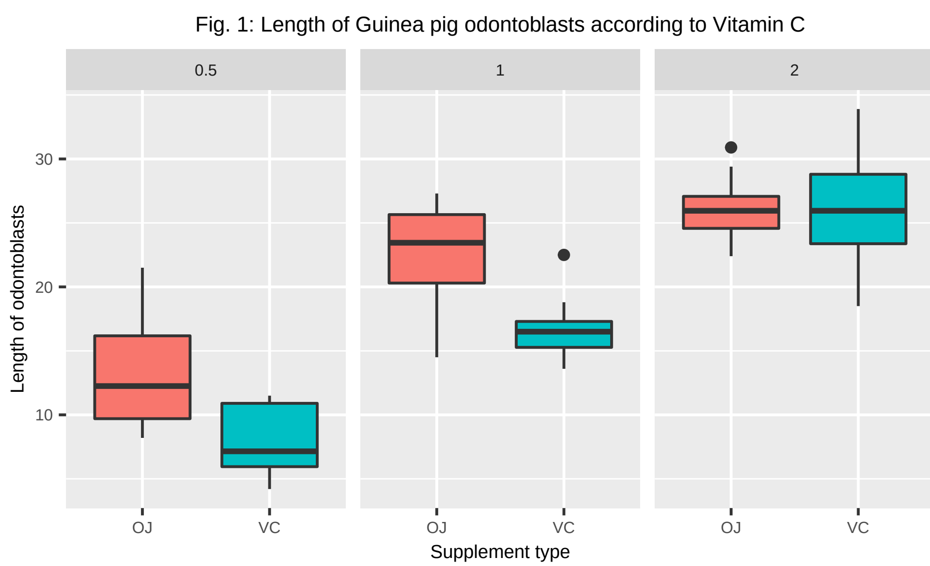
For this conclusions to be consistent with the population, we need to make sure:
- The sample of these 60 Guinea pigs is representative of the Guinea pig population;
- They were randomly selected;
- The difference in the length of odontoblasts due to the delivery of Vitamin C is significant compared to the natural variation;
- The delivery method was made in good conditions.

Appendix

Find below the supporting code and figures related to the Report part.

Figure

```r
ggplot(ToothGrowth, aes(x = supp, y = len)) +
    geom_boxplot(aes(fill = supp)) +
    facet_wrap( ~ dose) +
    labs(title = "Fig. 1: Length of Guinea pig odontoblasts according to Vitamin C",
         x = "Supplement type",
         y = "Length of odontoblasts") +
    scale_fill_discrete(guide = FALSE) +
    theme(text = element_text(size = 7))
```



Supporting code

Test 1

```r
## Test 1
t.test(len ~ supp, data = ToothGrowth)
```

```
##
##  Welch Two Sample t-test
##
## data:  len by supp
## 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 in group OJ mean in group VC
##         20.66333         16.96333
```

Test 2

```r
## Only keep the dose levels .5 and 2 mg/day
subTG1 <- ToothGrowth %>% filter(dose != 1)

## Test 2
T1 <- t.test(len ~ dose, data = subTG1)
print(T1)
```

```
##
##  Welch Two Sample t-test
##
## data:  len by dose
## 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 in group 0.5   mean in group 2
##            10.605            26.100
```

Test 3

```r
## Only keep the dose levels .5 and 1 mg/day
subTG2 <- ToothGrowth %>% filter(dose != 2)

## Test 3
T2 <- t.test(len ~ dose, data = subTG2)
print(T2)
```

```
##
##  Welch Two Sample t-test
##
## data:  len by dose
## 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 in group 0.5   mean in group 1
##            10.605            19.735
```

Test 4

```r
## Only keep the dose levels 1 and 2 mg/day
subTG3 <- ToothGrowth %>% filter(dose != .5)

## Test 4
T3 <- t.test(len ~ dose, data = subTG3)
print(T3)
```

```
##
##  Welch Two Sample t-test
##
## data:  len by dose
## 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 in group 1 mean in group 2
##          19.735          26.100
```

Test summary

```r
summaryTG <- data.frame(doseL = c(0.5, 0.5, 1),
                        doseH = c(1, 2, 2),
                        pValue = c(T1$p.value, T2$p.value, T3$p.value),
                        lowerBoundCI = c(T1$conf.int[1], T2$conf.int[1], T3$conf.int[1]),
                        upperBoundCI = c(T1$conf.int[2], T2$conf.int[2], T3$conf.int[2]),
                        doseLMean = c(T1$estimate[1], T2$estimate[1], T3$estimate[1]),
                        doseHMean = c(T1$estimate[2], T2$estimate[2], T3$estimate[2]))
print(summaryTG)
```