Statistical Inference Course Project - Part 2

Juan Agustín Morello

20/8/2020

Overview

This is the second part (basic inferential data analysis) of the Statistical Inference Course Project from Coursera.

The `ToothGrowth` data set is found in the datasets R package and registers the effecto of vitamin C on tooth growth in guinea pigs.

The data consist in the variability on 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.

The data frame contains 60 observations on 3 variables.

- `len`: (numeric) Tooth length
- `supp`: (factor) Supplement type. VC (ascorbic acid) or OJ (orange juice).
- `dose`: (numeric) Dose in milligrams/day

ToothGrowth data set Exploratory Analysis

Load the data set

```
data("ToothGrowth")

# I'll change the supplement names for clarity
ToothGrowth$supp <- factor(ToothGrowth$supp, levels = c("OJ", "VC"),
                           labels = c("Orange Juice", "Vitamin C"))
```

Display summary of the data set

```
summary(ToothGrowth)
```

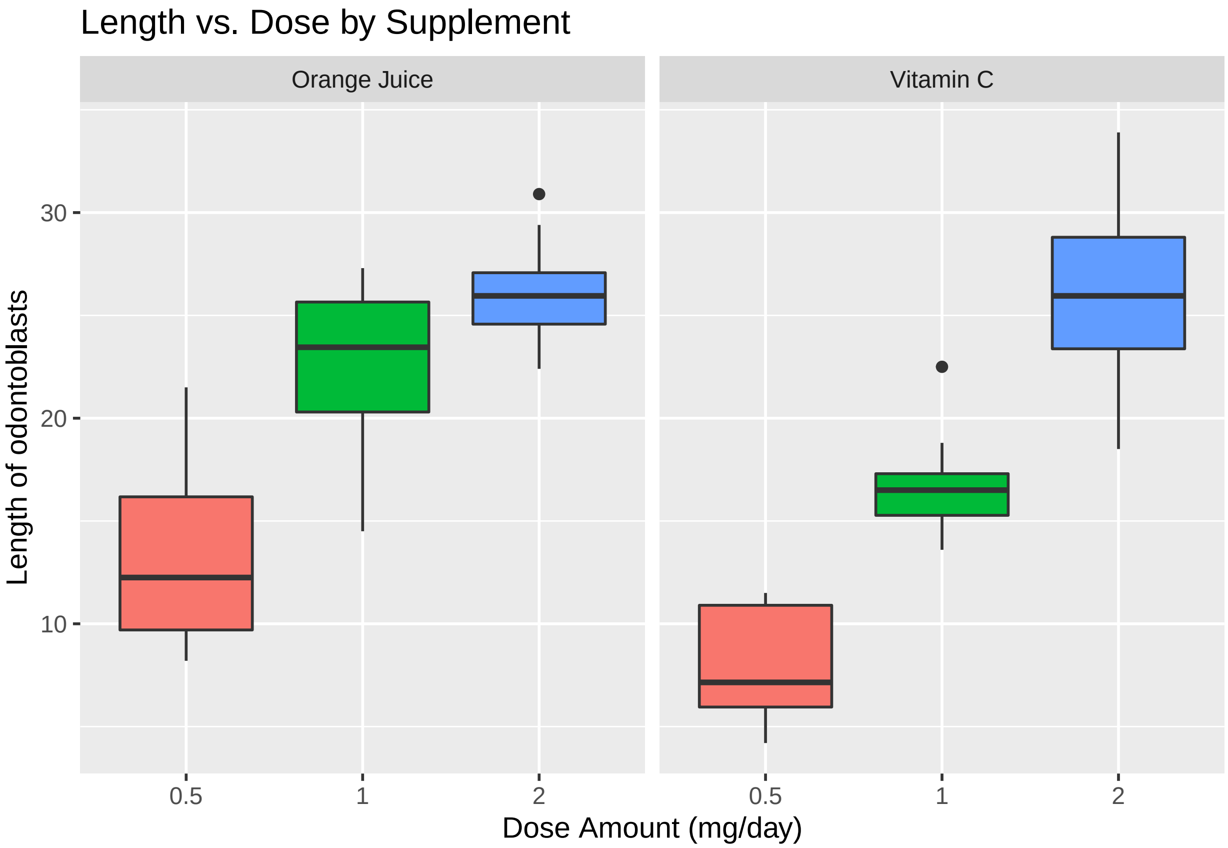
```
##       len                    supp         dose      
##  Min.   : 4.20   Orange Juice:30   Min.   :0.500  
##  1st Qu.:13.07   Vitamin C   :30   1st Qu.:0.500  
##  Median :19.25                     Median :1.000  
##  Mean   :18.81                     Mean   :1.167  
##  3rd Qu.:25.27       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

From the summary, we know that 30 observations are related to Orange Juice and other 30 are with Ascorbic Acid. Since the dose levels are fixed at 0.5, or 1, or 2 mg level, the median and mean or even quantile values do not carry much meaning in our analysis, we should focus on the length observation in the dataset, which the summary shows the values are between 4.20 and 33.90 with mean value at 28.81 and median at 19.25.

Plot some graphs to understand the data

I´ll use boxplot to provide a quick visual on the impact of dosage and supplement on the tooth growth

```
library(ggplot2)
ggplot(ToothGrowth, aes(as.factor(dose), len))+
  geom_boxplot(aes(fill=as.factor(dose)))+
  facet_grid(.~supp,)+
  labs(x="Dose Amount (mg/day)",
       y="Length of odontoblasts",
       title="Length vs. Dose by Supplement")+
  theme(legend.position = "none")
```



Based on the graph, we can know that the higher the dosage the longer the tooth grows. it appears that when the dosage is high at 2 mg/day, the median value of tooth growth appears to be similar between OJ and VC, however, when the dosage is 0.5 mg/day or 1 mg/day, the chart definitely shows that OJ has a obvious positive impact on tooth growth compared to VC.

Hypothesis Test

The objective now is to evaluate the impact of control variables `supp` and `dose` on the target variable `len`.

- Hypothesis #1 - For impact of control variable `supp` only on target variable `len`, **`OJ` has a higher impact on target variable `len`**:

```r
t.test(len ~ supp, paired = FALSE, var.equal=FALSE, 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 Orange Juice    mean in group Vitamin C
##                   20.66333                   16.96333
```

P-Value of 0.06063 is greater than alpha = 0.05 (alpha for confidence interval of 95%). Confidence interval (-0.171, 7.571) for the difference of the means of each group spans 0. Hence *null hypothesis is Failed to Reject*, hence **Hypothesis #1 is Rejected**.

- Hypothesis #2 - For impact of control variable `dose` only on target variable `len`, **Higher the `dose`, higher is the impact**:

2a : dose 1 mg/day has higher impact than dose 0.5 mg/day

```r
t.test(len ~ dose, paired = FALSE, var.equal = FALSE,
       data = ToothGrowth[ToothGrowth$dose %in% c(0.5, 1),])
```

```
##
##  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
```

P-Value of 1.268×10ˆ-7 is less than alpha=0.05 (alpha for confidence interval of 95%). Confidence interval (-11.9838, -6.2762) for the difference of the means of each group does not span 0. Hence *null hypothesis is Rejected*, hence **Hypothesis #2a is Failed to Reject**.

2b : dose 2 mg/day has higher impact than dose 1 mg/day

```r
t.test(len ~ dose, paired = FALSE, var.equal = FALSE,
       data = ToothGrowth[ToothGrowth$dose %in% c(1, 2),])
```

```
##
##  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
```

P-Value of 1.906×10ˆ-5 is less than alpha=0.05 (alpha for confidence interval of 95%). Confidence interval (-8.9965, -3.7335) for the difference of the means of each group does not span 0. Hence *null hypothesis is Rejected*, hence **Hypothesis #2b is Failed to Reject**.

Hypothesis #2 is Failed to Reject, based on above two evaluations.

- Hypothesis #3 - For combined impact of control variable `supp` and `dose`, **`OJ` has higher impact on target variable `len` for `dose` 0.5 mg/day & 1 mg/day**:

3a : `OJ` has higher impact for dose 0.5

```r
t.test(len ~ supp, paired = FALSE, var.equal = FALSE, data = ToothGrowth[ToothGrowth$dose == 0.5,])
```

```
##
##  Welch Two Sample t-test
##
## data:  len by supp
## t = 3.1697, df = 14.969, p-value = 0.006359
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.719057 8.780943
## sample estimates:
## mean in group Orange Juice    mean in group Vitamin C
##                      13.23                       7.98
```

P-Value of 0.0064 is less than alpha=0.05 (alpha for confidence interval of 95%). Confidence interval (1.7191, 8.7809) for the difference of the means of each group does not span 0. Hence *null hypothesis is Rejected*, hence **Hypothesis #3a is Failed to Reject**.

3b : `OJ` has higher impact for dose 1 mg/day

```r
t.test(len ~ supp, paired = FALSE, var.equal = FALSE, data = ToothGrowth[ToothGrowth$dose == 1,])
```

```
##
##  Welch Two Sample t-test
##
## data:  len by supp
## t = 4.0328, df = 15.358, p-value = 0.001038
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  2.802148 9.057852
## sample estimates:
## mean in group Orange Juice    mean in group Vitamin C
##                      22.70                      16.77
```

P-Value of 0.001 is less than alpha=0.05 (alpha for confidence interval of 95%). Confidence interval (2.8021, 9.0579) for the difference of the means of each group does not span 0. Hence *null hypothesis is Rejected*, hence **Hypothesis #3b is Failed to Reject**.

Hypothesis #3 is Failed to Reject, based on above two evaluations.

- Hypothesis #4 - For combined impact of control variables `supp` and `dose`, **`OJ` and `VC` have same impact on target variable `len` for `dose` 2 mg/day**:

```
t.test(len ~ supp, paired = FALSE, var.equal = FALSE, data = ToothGrowth[ToothGrowth$dose == 2,])
```

```
## 
##  Welch Two Sample t-test
## 
## data:  len by supp
## t = -0.046136, df = 14.04, p-value = 0.9639
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3.79807  3.63807
## sample estimates:
## mean in group Orange Juice    mean in group Vitamin C
##                      26.06                      26.14
```

P-Value of 0.9639 is greater than alpha=0.05 (alpha for confidence interval of 95%). Confidence interval (-3.7981, 3.6381) for the difference of the means of each group spans 0. Hence *null hypothesis is Failed to Reject*, hence **Hypothesis #4 is Failed to Reject**.

Assumptions and Conclusions

Assumptions:

1. Populations are random, independent, and comprised similar guinea pigs
2. Variance between populations are unequal
3. Double blind research methos were used.
4. A higher value of `len` indicates a higher impact.
5. Higher value of `dose` indicates increased dosages.

Conclusions: Based on the above evaluation of the four hypothesis, following are the conclusions:

1. For impact of control variable `supp` only, there is no significant difference on target variable `len` for different values of `supp`.
2. For impact of control variable `dose` only, higher the dose, higher is the impact on target variable `len`.
3. For combined impact of control variables, there is significant difference on target variable `len` for different values of `supp` for `dose` 0.5 mg/day and 1 mg/day. There is no significant difference for different values of `supp` for `dose` 2 mg/day".

Assuming the supplements were independently and identically distributed among the subjects, initially it appeared that the delivery method had no significant impact on tooth growth, but when controlling for dose level we discovered that there was a significant difference at 0.5 mg/day and 1 mg/day when applying orange juice supplements, but not at 2.0 mg/day (both OJ and CV looks similar).