Statistical Inference Project Part 2

Abhishek Kumar

9 August 2020

Part 2: Basic Inferential Data Analysis

Overview

Now in the second portion of the project, we're going to analyze the ToothGrowth data in the R datasets package. This analysis includes some basic exploratory analysis specifically summary and hypothesis testing confidence intervals.

```
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

library(ggplot2)

## Warning: package 'ggplot2' was built under R version 4.0.2

library(scales)
```

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

```
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 ...

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

About the data

This data comes from a study which analyzed the effect of supplement (Vitamin C) on length of tooth in Guinea Pigs. This data has 60 observations and 3 variables (len, supp and dose). Each of 10 guinea pigs were given three Vitamin C dosage levels (0.5, 1, and 2 mg) with two delivery methods (orange juice or ascorbic acid). So, the data contains 60 observations of 3 variables.

- len : Tooth length
- supp : Supplement type (VC or OJ)
- dose : Dose in milligrams

Exploratory Data Analysis

2. Provide a basic summary of the data.

```
supp_summ <- ToothGrowth %>% group_by(Factor = supp) %>% summarise(mean =
mean(len),
                                                    sd = sd(len),
                                                    median = median(len),
                                                    minimum = min(len),
                                                    maximum = max(len),
                                                    IQR = IQR(len))

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

dose_summ <- ToothGrowth %>% group_by(Factor = as.factor(dose)) %>%
summarise(mean = mean(len),
                                                    sd = sd(len),
                                                    median = median(len),
                                                    minimum = min(len),
                                                    maximum = max(len),
                                                    IQR = IQR(len))

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

rbind(supp_summ, dose_summ)

## # A tibble: 5 x 7
##   Factor  mean    sd median minimum maximum   IQR
##   <fct>  <dbl> <dbl>  <dbl>   <dbl>   <dbl> <dbl>
## 1 OJ      20.7  6.61  22.7      8.2    30.9 10.2
## 2 VC      17.0  8.27  16.5      4.2    33.9 11.9
## 3 0.5     10.6  4.50   9.85     4.2    21.5  5.03
## 4 1       19.7  4.42  19.2     13.6    27.3  7.12
## 5 2       26.1  3.77  26.0     18.5    33.9  4.3
```

The basic summary indicates that the tooth length was ranged from 4.2 to 33.9 with a mean of 18.81 and Median of 19.25. Among groups the data is shown in above table.

3. Use confidence intervals and/or hypothesis tests to compare tooth growth by supp and dose.

i. Which supplementary method is more effective for tooth growth?

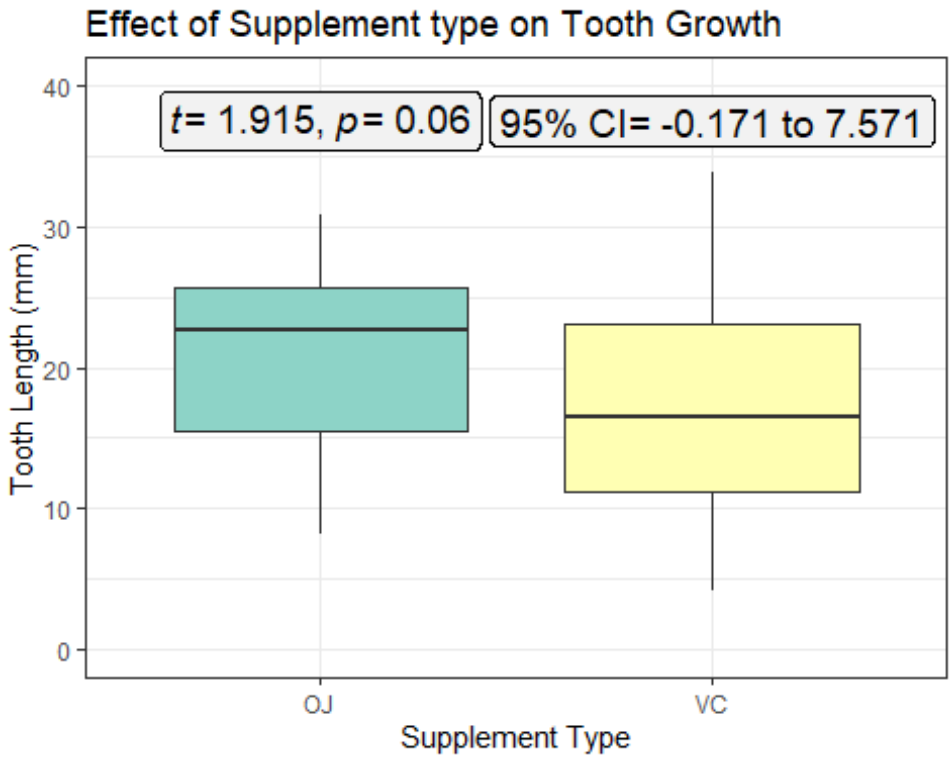
$H_0: \mu_{OJ} = \mu_{VJ}; \quad H_A: \mu_{OJ} \neq \mu_{VJ}$

```
OJ <- (ToothGrowth %>% filter(supp == "OJ"))$len
VC <- (ToothGrowth %>% filter(supp == "VC"))$len

t.test(OJ, VC, paired = FALSE, var.equal = FALSE, conf.level = 0.95)

##
##  Welch Two Sample t-test
##
## data:  OJ and 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

ggplot(ToothGrowth, aes(x = supp, y = len, fill = supp)) +
      geom_boxplot(show.legend = FALSE) +
      scale_fill_brewer(palette = "Set3") +
      theme_bw() + ylim(0,40) +
      labs(x = "Supplement Type", y = "Tooth Length (mm)",
           title = "Effect of Supplement type on Tooth Growth") +
      geom_label(x = 1, y = 37.5, fill = "gray95", size = 5,
                 label = "paste(italic(t), \"= 1.915\", \", \" , italic(p),
\"= 0.06\")",
                 parse = TRUE) +
      geom_label(x = 2, y = 37.5, fill = "gray95", size = 5,
                 label = "paste(\"95% CI\", \"= -0.171 to 7.571\")",
                 parse = TRUE)
```



So, the 95% confidence interval include zero suggesting that the null hypothesis (H_0) is true at this confidence interval and therefore, the alternative hypothesis (H_A) is rejected.

ii. Which dose is more effective for tooth growth?

Lets consider that the doses 0.5, 1, and 2 represents the levels as Low, Medium and high.

Hypotheses:

$H_0: \mu_{Low} = \mu_{Medium} = \mu_{High}; \quad \\ H_A: \mu_{Low} \ne \mu_{Medium} \ne \mu_{High}$

```
Low <- (ToothGrowth %>% filter(dose == "0.5"))$len
Medium <- (ToothGrowth %>% filter(dose == "1"))$len
High <- (ToothGrowth %>% filter(dose == "2"))$len

t.test(Low, Medium, paired = FALSE, var.equal = FALSE, conf.level = 0.95)

##
##  Welch Two Sample t-test
##
## data:  Low and Medium
## 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(Low, High, paired = FALSE, var.equal = FALSE, conf.level = 0.95)
```

```
##
##  Welch Two Sample t-test
##
## data:  Low and High
## 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
```

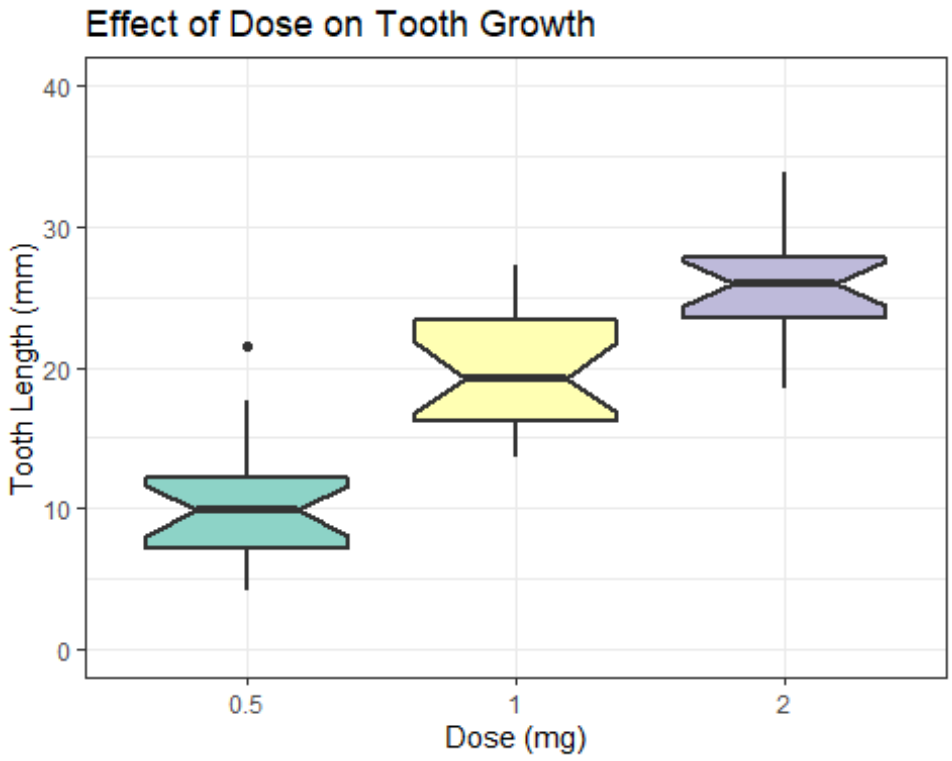
```
t.test(High, Medium, paired = FALSE, var.equal = FALSE, conf.level = 0.95)
```

```
##
##  Welch Two Sample t-test
##
## data:  High and Medium
## 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:
##  3.733519 8.996481
## sample estimates:
## mean of x mean of y
##    26.100    19.735
```

Data	t	df	p-value	CI	x_mean	y_mean
Low vs Medium	-6.477	37.986	<0.000	-11.98 to -6.276	10.60	19.73
Low vs High	-11.799	36.883	<0.000	-18.16 to -12.834	10.60	26.10
High vs Medium	4.901	37.101	<0.000	3.734 to 8.996	26.100	19.73

So, the 95% confidence interval did not included zero suggesting that the null hypothesis (H_0) is not true at this confidence interval and therefore, the alternative hypothesis (H_A) is accepted. Also, all the comparisons indicate a highly significant p-value showing the significant differences in the mean of the samples.

```
ggplot(ToothGrowth, aes(x = as.factor(dose), y = len, fill =
as.factor(dose))) +
      geom_boxplot(show.legend = FALSE, notch = TRUE, size = 0.75) +
      scale_fill_brewer(palette = "Set3") +
      theme_bw() + ylim(0,40) +
      labs(x = "Dose (mg)", y = "Tooth Length (mm)",
           title = "Effect of Dose on Tooth Growth")
```



4. State your conclusions and the assumptions needed for your conclusions.

Conclusions

The present analysis conclude that the tooth growth is not affected by supplement type (OJ or VC) at the 5% confidence intervals. However, the doses has the significant effects on the tooth growth, as the mean of doses was significantly differed among the dose concentrations.

Assumptions

The analysis used the t-test for hypothesis testing. So it was assumed that:

1. The population from which samples are drawn is normally distributed.
2. The tests also assumed that variances are equally distributed or homogenous.
3. The samples were randomly drawn and represents the population.
4. The samples were independent of each other.