Basic Inferential Data Analysis

Vijay Ramanujam

January 5, 2018

Overview

Now in the second portion of the project, we're going to analyze the ToothGrowth data in the R datasets package. The dataset contains the effect of vitamin C on tooth growth in guinea pigs. 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).

Format

A data frame with 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

Load dependent libraries and dataset

```
library(dplyr)
library(knitr)
library(printr)
library(datasets)

data(ToothGrowth)
```

Structure of the dataset

```
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 ...
```

Top rows

```
head(ToothGrowth)
```

len	supp	dose
4.2	VC	0.5
11.5	VC	0.5
7.3	VC	0.5
5.8	VC	0.5
6.4	VC	0.5
10.0	VC	0.5

Observations by supplement type/dosage

```r
table(ToothGrowth$supp, as.factor(ToothGrowth$dose))
```

/	0.5	1	2
OJ	10	10	10
VC	10	10	10

Basic Summary Statistics

```r
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	NA	Median :1.000
Mean :18.81	NA	Mean :1.167
3rd Qu.:25.27	NA	3rd Qu.:2.000
Max. :33.90	NA	Max. :2.000

Summary by supplement type

```r
kable(as.data.frame(summarise(
                        group_by(ToothGrowth, supp)
                        , count= n()
                        , mean=mean(len)
                        , median=median(len)
                        , "sd" = sd(len)
                    )
    ), digits=3, align='c'
)
```

supp	count	mean	median	sd
OJ	30	20.663	22.7	6.606
VC	30	16.963	16.5	8.266

Summary by supplement type and dosage

```r
kable(as.data.frame(summarise(
                        group_by(ToothGrowth, supp, dose)
                        , count= n()
                        , mean=mean(len)
                        , median=median(len)
                        , "sd" = sd(len)
                    )
    ), digits=3, align='c'
)
```

supp	dose	count	mean	median	sd
OJ	0.5	10	13.23	12.25	4.460
OJ	1.0	10	22.70	23.45	3.911
OJ	2.0	10	26.06	25.95	2.655
VC	0.5	10	7.98	7.15	2.747
VC	1.0	10	16.77	16.50	2.515
VC	2.0	10	26.14	25.95	4.798

Average tooth growth by supplement type/dosage

```
aggregate(len ~ supp + dose, FUN=mean, data=ToothGrowth)
```

supp	dose	len
OJ	0.5	13.23
VC	0.5	7.98
OJ	1.0	22.70
VC	1.0	16.77
OJ	2.0	26.06
VC	2.0	26.14

As shown above, the average tooth growth increases as the dosages increase with both supplement types. Also, the tooth growth is larger with orange juice at dosage levels 0.5mg/day and 1mg/day, and it is almost the same between the supplements at dosage level 2mg/day.

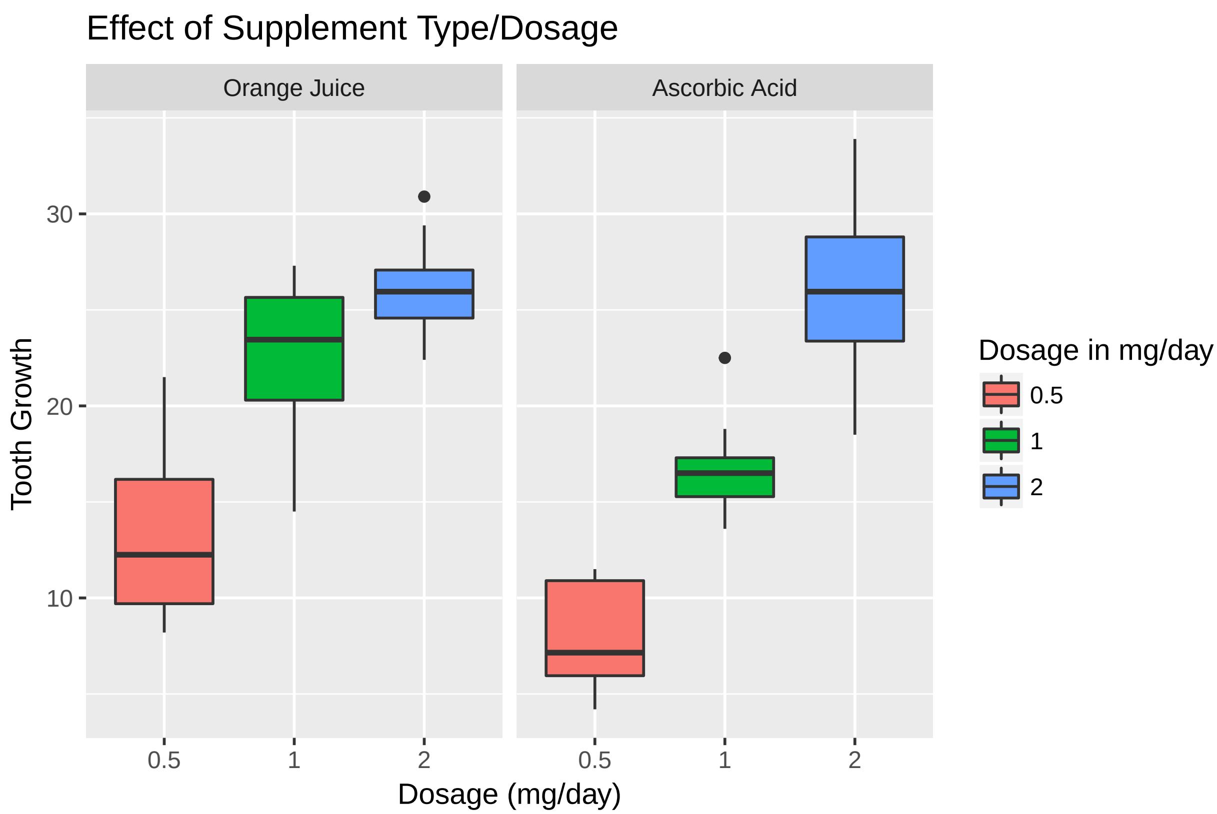
Plot the dataset

```
library(ggplot2)

plot <- ggplot(
    ToothGrowth
    , aes(x=factor(dose), y=len, fill=factor(dose))
)

plot <- plot +
    geom_boxplot(notch=F) +
    facet_grid(.~supp, labeller = as_labeller(
        c("OJ" = "Orange Juice", "VC" = "Ascorbic Acid"))) +
    scale_x_discrete("Dosage (mg/day)") +
    scale_y_continuous("Tooth Growth") +
    ggtitle("\nEffect of Supplement Type/Dosage") +
    scale_fill_discrete(name = "Dosage in mg/day")

plot
```



Hypothesis Testing

Testing by supplement type

```
ttest <- t.test(len ~ supp
                , data=ToothGrowth
                , var.equal = FALSE
                , paired=FALSE
                , conf.level = .95)

kable(data.frame(
        "t_statistic"  = ttest$statistic
        , "DF" = ttest$parameter
        , "p_value"  = ttest$p.value
        , "CL_Lower" = ttest$conf.int[1]
        , "CL_Upper" = ttest$conf.int[2]
        , "OJ_Mean" = ttest$estimate[1]
        , "VC_Mean" = ttest$estimate[2]
        , row.names = "OJ versus VC "
), digits = 3, align = 'c'
    , caption = '<br/><br/><br/>Summary of two sample t-test for tooth growth by supplement type')
```

Table 7: Summary of two sample t-test for tooth growth by supplement type

	t_statistic	DF	p_value	CL_Lower	CL_Upper	OJ_Mean	VC_Mean
OJ versus VC	1.915	55.309	0.061	-0.171	7.571	20.663	16.963

From the above summary, with 95% confidence (alpha=0.05), we are unable to reject the null hypothesis i.e. there is no significant difference in tooth growth between suppplement types in guniea pigs since the p-value (0.061) is greater than alpha (0.05). It means, we do not have sufficient statistical evidence to show that there is a significant difference in the average tooth growth between supplement types. This can also be verified by the lower and upper levels of confidence interval which contains 0 (no difference).

Testing by supplement type and dosage levels

```
ttest05 <- t.test(len ~ supp
                  , data=filter(ToothGrowth, dose==0.5)
                  , var.equal = FALSE
                  , paired=FALSE
                  , conf.level = .95)


ttest10 <- t.test(len ~ supp
                  , data=filter(ToothGrowth, dose==1)
                  , var.equal = FALSE
                  , paired=FALSE
                  , conf.level = .95)

ttest20 <- t.test(len ~ supp
                  , data=filter(ToothGrowth, dose==2)
                  , var.equal = FALSE
                  , paired=FALSE
                  , conf.level = .95)


kable(data.frame(
        "t_statistic" = c(ttest05$statistic, ttest10$statistic, ttest20$statistic)
        , "DF" = c(ttest05$parameter, ttest10$parameter, ttest20$parameter)
        , "p_value" = c(ttest05$p.value, ttest10$p.value, ttest20$p.value)
        , "CL_Lower" = c(ttest05$conf.int[1], ttest10$conf.int[1], ttest20$conf.int[1])
        , "CL_Upper" = c(ttest05$conf.int[2], ttest10$conf.int[2], ttest20$conf.int[2])
        , "OJ_Mean" = c(ttest05$estimate[1], ttest10$estimate[1], ttest20$estimate[1])
        , "VC_Mean" = c(ttest05$estimate[2], ttest10$estimate[2], ttest20$estimate[2])
        , row.names = c(
          "OJ versus VC @ 0.5mg/day"
          , "OJ versus VC @ 1mg/day"
          , "OJ versus VC @ 2mg/day")
), digits = 3, align = 'c'
    , caption = 'Summary of two sample t-test for tooth growth by supplement type and dosage')
```

Table 8: Summary of two sample t-test for tooth growth by supplement type and dosage

	t_statistic	DF	p_value	CL_Lower	CL_Upper	OJ_Mean	VC_Mean
OJ versus VC @ 0.5mg/day	3.170	14.969	0.006	1.719	8.781	13.23	7.98
OJ versus VC @ 1mg/day	4.033	15.358	0.001	2.802	9.058	22.70	16.77
OJ versus VC @ 2mg/day	-0.046	14.040	0.964	-3.798	3.638	26.06	26.14

From the above summary, with 95% confidence, there is statistical evidence to believe that the average tooth growth of guinea pigs differ between supplement types at dosage levels 0.5mg/day and 1mg/day whereas, there is no such evidence when it comes to 2mg/day since the p-value (0.964) is greater than alpha (0.05) and the confidence interval contains 0.

Assumptions

```
* Random samples of guinea pigs, but guinea pigs are similar as a population
* Variances between different samples are not equivalent
* The observations are not paired; groups of guinea pigs that are
given OJ and VC are different samples
* The null hypothesis is: there is no significant difference of
average tooth growth in guinea pigs samples that are given OJ and VC
* Confidence level is 95% (alpha = 0.05)
```

Conclusion

Based on the analyses performed in the above sections, we can conclude that the average tooth growth is larger with orange juice at dosage levels 0.5mg/day and 1mg/day. But, the difference in average tooth growth is uncertain at dosage level of 2mg/day between OJ and VC.