Part2: Basic Inferential Data Analysis

Junyoung Kim

Feb. 25, 2018

1. Overview

The **"ToothGrowth"** data in R contains the experiment results of tooth cell growth by giving 60 guinea pigs different supplement types and dose levels of vitamin C. This study examines whether the experiment results differ statistically along with the types of vitamin C supplements provided: ascorbic acid(VC) and orange juice(OJ). As we can see in the data structure and the plot below, each supplement type is composed of three different dose levels(0.5, 1.0, 2.0 mg/day). Since each supplement/dose group consists of different 10 guinea pigs, each group has its own mean and variation. (**see Appendix 1 for codes**)

Data Structure:

```
## '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 ...
```

Distribution and Data Statistics:

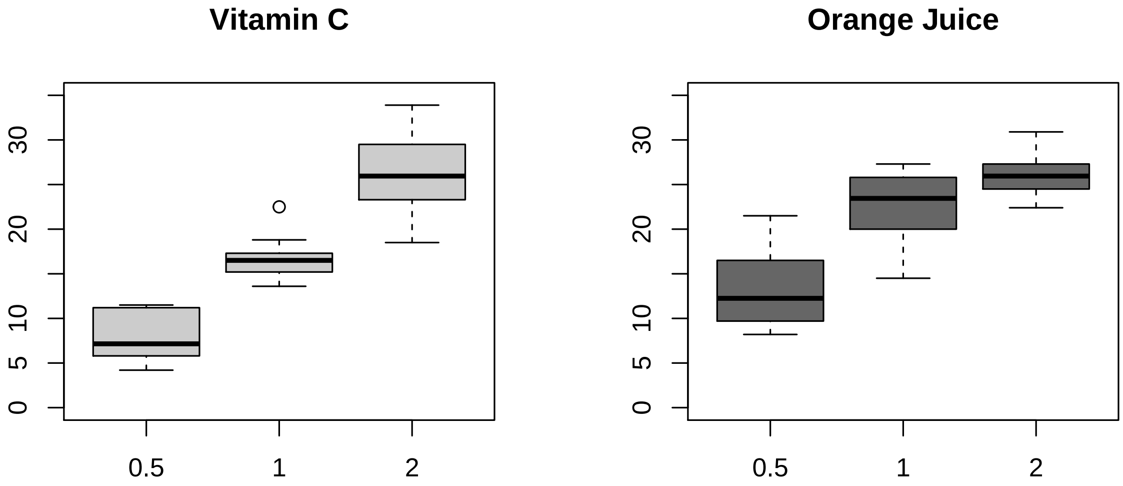
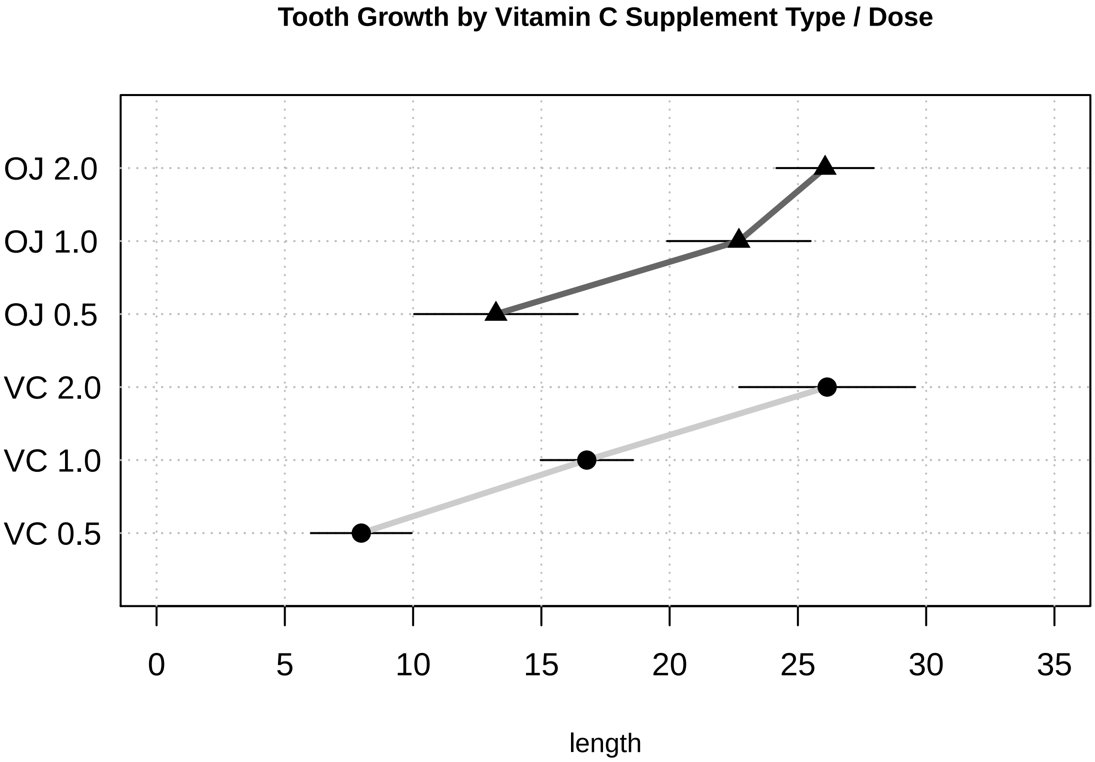


Table 1: Mean and Standard Deviation by Type and Dose

	VC 0.5	VC 1.0	VC 2.0	OJ 0.5	OJ 1.0	OJ 2.0
Mean	7.980	16.770	26.140	13.23	22.700	26.060
Standard Deviation	2.747	2.515	4.798	4.46	3.911	2.655

2. Hypothesis Test

Before testing difference between the effects in the two supplement groups, plotting means and confidence intervals is a good way to anticipate results. The following plot shows the means of each supplement group with different dosage levels, and the 95% confidence intervals of the means. (**see Appendix 2 for codes**)



What we should focus in this plot in particular is that the confidence interval of OJ 2.0 lies within that of VC 2.0 unlike other dosage levels. It implies that it is difficult to statistically distinguish the effects of those two groups. **Assuming that the effects are normally distributed, 95% C.I. two-tailed tests for each dosage level are performed below** to make this implication clear. The reason of running two-tailed test is not only because what we need to know is only whether the results of two supplement groups are statistically different at the same dosage level, but also because two-tailed test is more strict. As we can see the difference of the means in **Table 1**, we can find which one has greater effect at the same dosage level if the tests confirm the differences.

Hypotheses

- H0: Estimated mean of two supplement effects are equal → $E(OJ) - E(VC) = 0$
- H1: Estimated mean of two supplement effects are not equal → $E(OJ) - E(VC) \neq 0$

Table 2: 95% C.I. Hypothesis Test Result

	OJ mean	VC mean	difference	t-stat	df	p-value
0.5mg/day	13.23	7.98	5.25	3.170	14.969	0.006
1.0mg/day	22.70	16.77	5.93	4.033	15.358	0.001
2.0mg/day	26.06	26.14	-0.08	-0.046	14.040	0.964

3. Conclusion

The results in **Table 2** indicates that…

- at both dosage levels of 0.5 and 1.0 mg/day, two supplement types show notable mean effect differences($>$ 5) and their p-values(0.006, 0.001) are below the α-level of 0.05. Hence, the results reject their null hypotheses. The estimated means of two supplement effects are not equal, **therefore, at those two dosage levels, orange juice may have greater effects than ascorbic acid in tooth cell growth of guinea pigs**.
- at dosage level of 2.0 mg/day, two supplement types not only show very small difference(= -0.046) in mean effects, but also their p-value(0.964) is far larger than the α-level of 0.05. The result fails to reject the null hypothesis. Since the estimated means of two supplement effects are not statistically different, **at the dosage level of 2.0mg/day, giving ascorbic acid may have the same level of tooth growth effect on guniea pigs as giving orange juice, but its effect may vary more widely because its variance is larger.**
- Based on the two results above, **we can conclude that the difference of vitamin C supplement type can become meaningless, or at least their effect difference shrinks, if the dosage level is high enough.**

Appendix

1. Overview Codes

```r
data("ToothGrowth") # load data
str(ToothGrowth) # overview of data

# boxplot for overview
par(mfrow=c(1,2))
boxplot(len~dose, ToothGrowth[1:30,], main = "Vitamin C", ylim = c(0, 35),
        col="grey80")
boxplot(len~dose, ToothGrowth[31:60,], main = "Orange Juice", ylim=c(0, 35),
        col="grey40")
par(mfrow=c(1,1))

# Calculate mean and C.I. for each dose and supplement
tiny <- data.frame(ToothGrowth$len[1:10], ToothGrowth$len[11:20],
                   ToothGrowth$len[21:30], ToothGrowth$len[31:40],
                   ToothGrowth$len[41:50], ToothGrowth$len[51:60])

colnames(tiny) <- c("VC 0.5", "VC 1.0", "VC 2.0", "OJ 0.5", "OJ 1.0", "OJ 2.0")
means <- apply(tiny, 2, mean) # calculate means
sds <- apply(tiny, 2, sd) # calculate standard deviations

# calculate confidence intervals
ci.upper <- means + qt(.975, 9, lower.tail = TRUE) * sds/sqrt(10)
ci.lower <- means + qt(.025, 9, lower.tail = TRUE) * sds/sqrt(10)

# tabulate data
tab3 <- as.table(rbind(c(means[1],means[2],means[3],means[4],means[5],means[6]),
                       c(sds[1], sds[2], sds[3], sds[4], sds[5], sds[6])))
dimnames(tab3) <- list("Type/Dose" = c("Mean", "Standard Deviation"),
                       c("VC 0.5", "VC 1.0", "VC 2.0",
                         "OJ 0.5", "OJ 1.0", "OJ 2.0"))
```

```
knitr::kable(tab3, align = 'c', digits = 3,
             caption = "Mean and Standard Deviation by Type and Dose")
```

2. Hypothesis Test Codes

```
# plot means
dotchart(means, pch=c(rep(19, 3), rep(17,3)),
         xlim=c(0, 35), xlab="length", cex.main=0.8, cex.lab=0.8, cex.axis=0.8,
         main = "Tooth Growth by Vitamin C Supplement Type / Dose")
abline(h=1:6, v=(c(0:7)*5), col="gray", lty=3)
# add confidence intervals
for(i in 1:6){lines(x=c(ci.lower[i], ci.upper[i]), y=c(i,i))}
lines(means[1:3], y=c(1:3), col="grey80", lwd=3)
lines(means[4:6], y=c(4:6), col="grey40", lwd=3)
points(means, y=c(1:6), pch=c(rep(19, 3), rep(17,3)), cex=1.2)

# hypothesis test for each level of dose
## H0 : mean difference = 0
## H1 : mean difference != 0

# Test 1: 0.5 VC / OJ
mean.diff05 <- means[4] - means[1]
sd.diff05 <- sqrt(sds[1]^2/10 + sds[4]^2/10)
v.diff05 <- ((sds[1]^2/10)+(sds[4]^2/10))^2 / ((sds[1]^2/10)^2/9 + (sds[4]^2/10)^2/9)
p.value05 <- pt(mean.diff05/sd.diff05, v.diff05, lower.tail = FALSE) * 2

# Test 2: 1.0 VC / OJ
mean.diff10 <- means[5] - means[2]
sd.diff10 <- sqrt(sds[2]^2/10 + sds[5]^2/10)
v.diff10 <- ((sds[2]^2/10)+(sds[5]^2/10))^2 / ((sds[2]^2/10)^2/9 + (sds[5]^2/10)^2/9)
p.value10 <- pt(mean.diff10/sd.diff10, v.diff10, lower.tail = FALSE) * 2

# Test 3: 2.0 VC / OJ
mean.diff20 <- means[6] - means[3]
sd.diff20 <- sqrt(sds[3]^2/10 + sds[6]^2/10)
v.diff20 <- ((sds[3]^2/10)+(sds[6]^2/10))^2 / ((sds[3]^2/10)^2/9 + (sds[6]^2/10)^2/9)
p.value20 <- pt(mean.diff20/sd.diff20, v.diff20, lower.tail = TRUE) * 2

# tabulate
row1 <- c(means[4], means[1], mean.diff05, mean.diff05/sd.diff05, v.diff05, p.value05)
row2 <- c(means[5], means[2], mean.diff10, mean.diff10/sd.diff10, v.diff10, p.value10)
row3 <- c(means[6], means[3], mean.diff20, mean.diff20/sd.diff20, v.diff20, p.value20)
tab4 <- as.table(rbind(row1, row2, row3))
dimnames(tab4) <- list("dose"= c("0.5mg/day", "1.0mg/day", "2.0mg/day"),
                       c("OJ mean", "VC mean", "difference", "t-stat",
                         "df", "p-value"))
tab4 <- round(tab4, 3)
knitr::kable(tab4, align = 'c', digits = 3,
             caption = "95% C.I. Hypothesis Test Result")
```