Course Project 2: Baisc Inference Data Analysis

Xiaozhu Zhang

Jan. 15, 2021

In this part, I will do a basic inference data analysis on the dataset `ToothGrowth`. 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 supplement types, orange juice or ascorbic acid.

1. Data loading

```
library("datasets")
data("ToothGrowth")
dim(ToothGrowth)
```

```
## [1] 60  3
```

```
head(ToothGrowth)
```

```
##    len supp dose
## 1  4.2   VC  0.5
## 2 11.5   VC  0.5
## 3  7.3   VC  0.5
## 4  5.8   VC  0.5
## 5  6.4   VC  0.5
## 6 10.0   VC  0.5
```

We see that the dataset is a 60 by 3 matrix. Since the variable `supp` only takes 2 values, and `dose` only takes 3 values, we will set them as factors.

2. Basic EDA and data summary

```
dat <- ToothGrowth
dat$dose <- factor(dat$dose)
summary(dat)
```

```
##       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
```

From the output summary table, we know that 30 pigs were assigned to each supplement group respectively, while 20 pigs were assigned to each dose level.

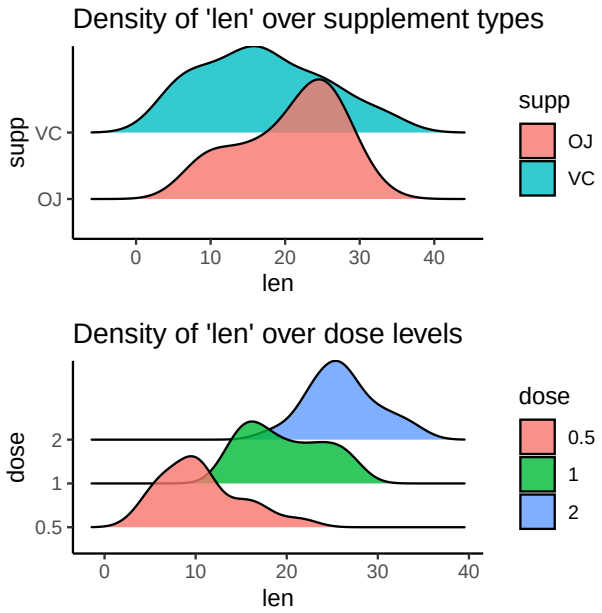

Figure 1: Density of different groups

The density plots over groups are shown in figure 1. We see that they all are bell-shaped though more or less skewed. The density plots of the two different supplement types are very close. However, different dose levels may lead to significantly different lengths.

3. Tests

Figures can imply but may never be exact. In order to give accurate results we use `t.test()` to perform hypothesis test and construct confidence interval.

```
# Tests of supp
len_OJ <- subset(dat, supp %in% "OJ")$len
len_VC <- subset(dat, supp %in% "VC")$len
t.test(len_OJ, len_VC, paired = FALSE, var.equal = FALSE, alt = "two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  len_OJ and len_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
```

For the hypothesis test, the null hypothesis is $H_0 : \mu_{\text{OJ}} = \mu_{\text{VC}}$. Since p-value is greater than $\alpha = 0.05$, we fail to reject the null hypothesis. In other words, there is no significant differences between the means of lengths over group "OJ" and group "VC". The 95% confidence interval is $[-0.171, 7.571]$ containing 0, which supports the same result.

```r
# Tests of dose
len_.5 <- subset(dat, dose %in% "0.5")$len
len_1 <- subset(dat, dose %in% "1")$len
len_2 <- subset(dat, dose %in% "2")$len
matrix <- cbind(len_.5, len_1, len_2)
library("combinat")
tab <- combn(3,2)
pValues <- c()
for(i in 1:3){
  a <- matrix[,tab[,i][1]]
  b <- matrix[,tab[,i][2]]
  t <- t.test(a, b, paired = FALSE, var.equal = FALSE, alt = "two.sided")$p.value
  pValues <- c(pValues, t)
}
pValues
```

```
## [1] 1.268301e-07 4.397525e-14 1.906430e-05
```

```r
p.adjust(pValues, method = "bonferroni")
```

```
## [1] 3.804902e-07 1.319257e-13 5.719289e-05
```

```r
p.adjust(pValues, method = "BH")
```

```
## [1] 1.902451e-07 1.319257e-13 1.906430e-05
```

Since there are 3 dose levels, in order to compare each pair, we need to perform $C(3,2) = 3$ tests. The null and hypotheses are: $H_{01} : \mu_{.05} = \mu_1$, $H_{02} : \mu_{.05} = \mu_2$, and $H_{03} : \mu_1 = \mu_2$. For multiple testing, we use "bonferroni" and "BH" methods to adjust p-values, controling the family-wise error rate and the discovery rate respectively. Regardless of the criteria, all p-values are far less than $\alpha = 0.05$, so we should reject all H_0's in the three tests. All three groups have significantly different means.

4. Assumptions and conclusions

The hypothesis tests and construction of confidence intervals require several **assumptions**:

- The population distribution of the tooth length variable from a certain group is approximately normal.
- The tooth length samples are independent and identically distributed.
- The sample size should not be too small if the population distribution is somewhat skewed.
- Confounders are randomly assigned. In other words, when we perform tests on supplement types over length, dose levels should be randomly distributed in the levels of VC and OJ. Similarly, when we perform tests on dose levels over length, supplement types should be randomly distributed in 0.5 mg/day, 1 mg/day, and 2 mg/day.
- The significant level is $\alpha = 0.5$.

Based on the test results, we have the following **conclusions**:

- Supplements types, namely orange juice or ascorbic acid, do not significantly affect the length of odontoblasts.
- Dose levels of vitmin C have an significant influence on the length of odontoblasts. In specific, pigs who recieve vitamin C 0.5 mg/day have smaller expected tooth length than those who recieve 1 mg/day; similarly, pigs who recieve 1 mg/day have smaller expected tooth length than those who recieve 2 mg/day. In a nutshell, vitamin C has a clear positive effect on tooth growth.

Appendix

```r
# Codes for Figure 1
library(ggplot2)
library(ggridges)
ggplot(dat, aes(x = len, y = factor(supp), fill = supp)) +
  geom_density_ridges(alpha = 0.8) +
  labs(title = "Density of 'len' over supplement types", x = "len", y = "supp") +
  theme_classic()

ggplot(dat, aes(x = len, y = factor(dose), fill = dose)) +
  geom_density_ridges(alpha = 0.8) +
  labs(title = "Density of 'len' over dose levels", x = "len", y = "dose") +
  theme_classic()
```