Statistical Inference Course Project: Part 2

Cheng-Han Yu

July 17, 2015

Data analysis

We're going to analyze the ToothGrowth data in the R datasets package.

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

Solution

```
library(datasets)
data(ToothGrowth)
tooth <- ToothGrowth
str(tooth) # check structure
```

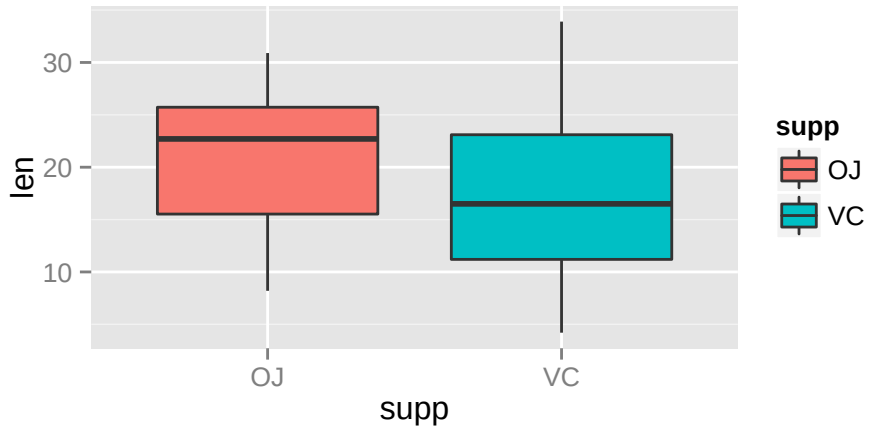
```
## 'data.frame':    60 obs. of  3 variables:
##  $ len : num  4.2 11.5 7.3 5.8 6.4 10 11.2 11.2 5.2 7 ...
##  $ supp: Factor w/ 2 levels "OJ","VC": 2 2 2 2 2 2 2 2 2 2 ...
##  $ dose: num  0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 ...
```

```
tooth$dose <- factor(tooth$dose) # convert class of dose into factor
head(tooth)
```

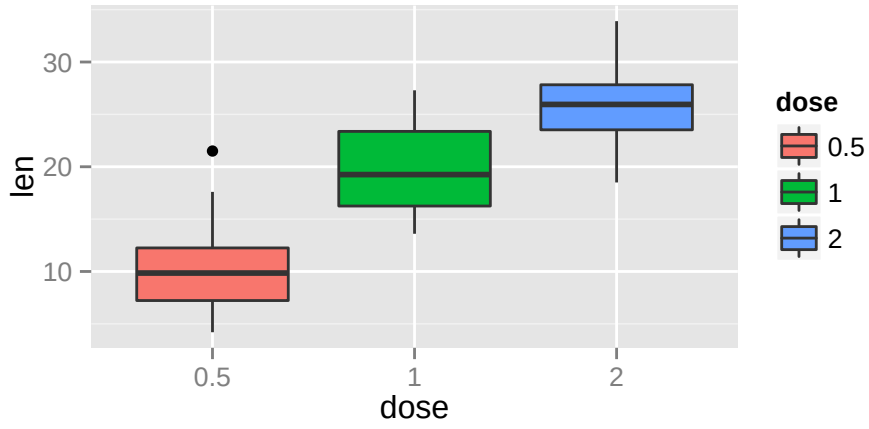
```
##    len supp dose
## 1  4.2   VC  0.5
## 2 11.5   VC  0.5
## 3  7.3   VC  0.5
## 4  5.8   VC  0.5
## 5  6.4   VC  0.5
## 6 10.0   VC  0.5
```

The dataset contains 60 observations and three variables where the response is the tooth length in each of 10 guinea pigs at each of three dose levels of Vitamin C (0.5, 1, and 2 mg) with each of two delivery methods (orange juice (OJ) or ascorbic acid (VC)). Some boxplots are shown below to explore the relationship between tooth length and supplement type and dose level before doing any inference. We can see that dose level affects tooth growth significantly no matter which type of supplement were used. Using orange juice tends to have longer teeth, especially when dose level is low. When dose level is 2mg, two types of supplement do not have much differnece for tooth growth.

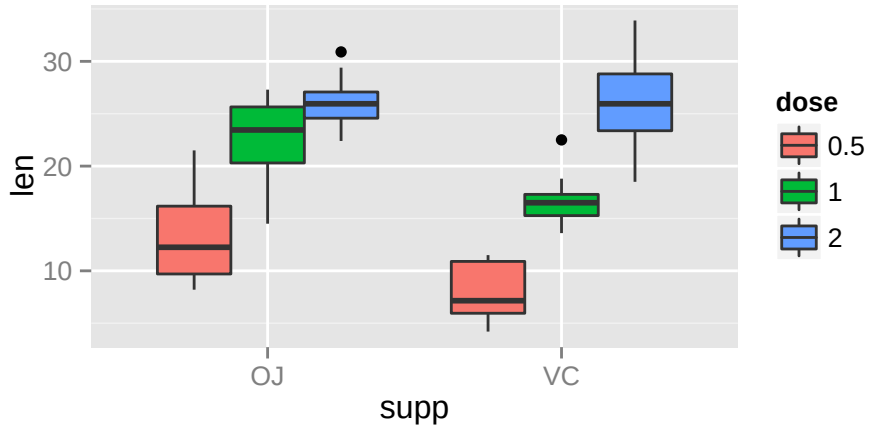
```
library(ggplot2)
ggplot(tooth, aes(supp, len, fill = supp)) + geom_boxplot()
```



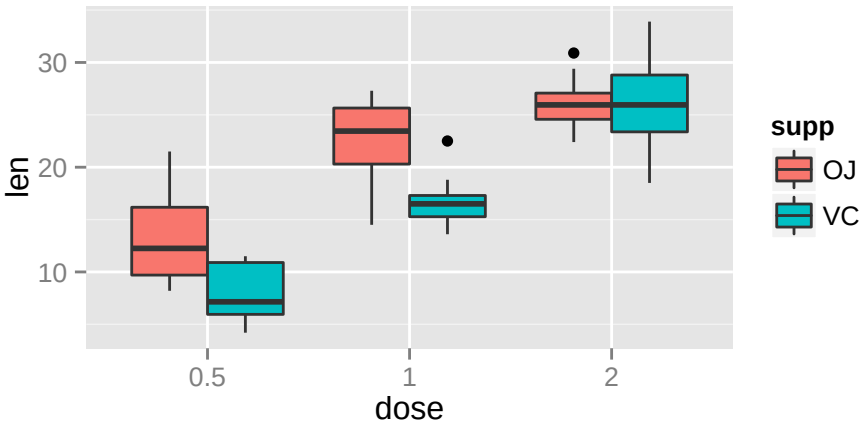

```
ggplot(tooth, aes(dose, len, fill = dose)) + geom_boxplot()
```




```
ggplot(aes(y = len, x = supp, fill = dose), data = tooth) + geom_boxplot()
```




```
ggplot(aes(y = len, x = dose, fill = supp), data = tooth) + geom_boxplot()
```



2. Provide a basic summary of the data.

Solution

Some basic decriptive satstistics summaries for the variable `len` are shown below.

```
library(psych)
describe(tooth$len)[-c(1, 6, 7)]
```

```
##    n  mean   sd median min  max range  skew kurtosis   se
## 1 60 18.81 7.65  19.25 4.2 33.9  29.7 -0.14    -1.04 0.99
```

```
describe(tooth[tooth$supp == "VC", ]$len)[-c(1, 6, 7)]
```

```
##    n  mean   sd median min  max range skew kurtosis   se
## 1 30 16.96 8.27   16.5 4.2 33.9  29.7 0.28    -0.93 1.51
```

```
describe(tooth[tooth$supp == "OJ", ]$len)[-c(1, 6, 7)]
```

```
##    n  mean   sd median min  max range  skew kurtosis   se
## 1 30 20.66 6.61   22.7 8.2 30.9  22.7 -0.52    -1.03 1.21
```

```
describe(tooth[tooth$dose == "0.5", ]$len)[-c(1, 6, 7)]
```

```
##    n  mean  sd median min  max range skew kurtosis   se
## 1 20 10.61 4.5   9.85 4.2 21.5  17.3 0.71    -0.31 1.01
```

```
describe(tooth[tooth$dose == "1", ]$len)[-c(1, 6, 7)]
```

```
##    n  mean   sd median  min  max range skew kurtosis   se
## 1 20 19.73 4.42  19.25 13.6 27.3  13.7 0.27    -1.44 0.99
```

```
describe(tooth[tooth$dose == "2", ]$len)[-c(1, 6, 7)]
```

```
##    n mean   sd median  min  max range skew kurtosis   se
## 1 20 26.1 3.77  25.95 18.5 33.9  15.4 0.25    -0.45 0.84
```

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

Solution

```
# len vs supp
t.test(tooth$len ~ tooth$supp, alternative = "two.sided",
       paired = FALSE, var.equal = FALSE, conf.level = 0.95) # Don not reject null
```

```
##
##  Welch Two Sample t-test
##
## data:  tooth$len by tooth$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 OJ mean in group VC
##         20.66333         16.96333
```

```
# len vs dose
t.test(x = tooth[tooth$dose == "0.5", ]$len,
       y = tooth[tooth$dose == "1", ]$len, alternative = "less",
       paired = FALSE, var.equal = FALSE, conf.level = 0.95) # reject null
```

```
##
##  Welch Two Sample t-test
##
## data:  tooth[tooth$dose == "0.5", ]$len and tooth[tooth$dose == "1", ]$len
## t = -6.4766, df = 37.986, p-value = 6.342e-08
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf -6.753323
## sample estimates:
## mean of x mean of y
##    10.605    19.735
```

```
t.test(x = tooth[tooth$dose == "0.5", ]$len,
       y = tooth[tooth$dose == "2", ]$len, alternative = "less",
       paired = FALSE, var.equal = FALSE, conf.level = 0.95) # reject null
```

```
##
##  Welch Two Sample t-test
##
## data:  tooth[tooth$dose == "0.5", ]$len and tooth[tooth$dose == "2", ]$len
## t = -11.799, df = 36.883, p-value = 2.199e-14
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf -13.27926
## sample estimates:
## mean of x mean of y
##    10.605    26.100
```

```r
t.test(x = tooth[tooth$dose == "1", ]$len,
       y = tooth[tooth$dose == "2", ]$len, alternative = "less",
       paired = FALSE, var.equal = FALSE, conf.level = 0.95) # reject null
```

```
##
##  Welch Two Sample t-test
##
## data:  tooth[tooth$dose == "1", ]$len and tooth[tooth$dose == "2", ]$len
## t = -4.9005, df = 37.101, p-value = 9.532e-06
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##      -Inf -4.17387
## sample estimates:
## mean of x mean of y
##    19.735    26.100
```

```r
# all 15 testings including bonferroni and FDR corrections
m <- choose(6, 2)
df <- data.frame(pvalue = 1:m, compare = 1:m, lower = 1:m, upper = 1:m)
supp_dose <- interaction(rep(c("VC", "OJ"), each = 3), rep(c("0.5", "1", "2")))
k <- 1
for (i in 1:5) {
    for (j in (i+1):6) {
        result <- t.test(x = tooth[(10*(i-1) + (1:10)), ]$len,
                         y = tooth[(10*(j-1) + (1:10)), ]$len,
                         alternative = "two.sided",
                         paired = FALSE, var.equal = FALSE)
        df$pvalue[k] <- result$p.value
        df$lower[k] <- round(result$conf.int[1], 3)
        df$upper[k] <- round(result$conf.int[2], 3)
        df$compare[k] <- paste(supp_dose[i], "vs", supp_dose[j])
        k <- k + 1
    }
}
df$pBon <- p.adjust(df$pvalue, method = "bonferroni")
df$pBH <- p.adjust(df$pvalue, method = "BH")
df$sig <- as.numeric(df$pvalue < 0.05)
df$sigBon <- as.numeric(df$pBon < 0.05)
df$sigBH <- as.numeric(df$pBH < 0.05)
cbind(round(df[, -2], 4), df$compare)
```

```
##    pvalue   lower   upper   pBon    pBH sig sigBon sigBH       df$compare
## 1  0.0000 -11.266  -6.314 0.0000 0.0000   1      1     1   VC.0.5 vs VC.1
## 2  0.0000 -21.902 -14.418 0.0000 0.0000   1      1     1   VC.0.5 vs VC.2
## 3  0.0064  -8.781  -1.719 0.0954 0.0087   1      0     1 VC.0.5 vs OJ.0.5
## 4  0.0000 -17.921 -11.519 0.0000 0.0000   1      1     1   VC.0.5 vs OJ.1
## 5  0.0000 -20.618 -15.542 0.0000 0.0000   1      1     1   VC.0.5 vs OJ.2
## 6  0.0001 -13.054  -5.686 0.0014 0.0002   1      1     1     VC.1 vs VC.2
## 7  0.0460   0.072   7.008 0.6902 0.0531   1      0     0   VC.1 vs OJ.0.5
## 8  0.0010  -9.058  -2.802 0.0156 0.0016   1      1     1     VC.1 vs OJ.1
## 9  0.0000 -11.720  -6.860 0.0000 0.0000   1      1     1     VC.1 vs OJ.2
## 10 0.0000   8.556  17.264 0.0001 0.0000   1      1     1   VC.2 vs OJ.0.5
```

```
## 11 0.0965  -0.684   7.564 1.0000 0.1034   0      0     0     VC.2 vs OJ.1
## 12 0.9639  -3.638   3.798 1.0000 0.9639   0      0     0     VC.2 vs OJ.2
## 13 0.0001 -13.416  -5.524 0.0013 0.0002   1      1     1   OJ.0.5 vs OJ.1
## 14 0.0000 -16.335  -9.325 0.0000 0.0000   1      1     1   OJ.0.5 vs OJ.2
## 15 0.0392  -6.531  -0.189 0.5879 0.0490   1      0     1     OJ.1 vs OJ.2
```

I use two sample t test to compare tooth growth by supp and dose. From the exploratory analysis and descriptive statistics in problem 1 and 2, I assume any two samples are independent, and their variances are not equal. All testings are using significant level $\alpha = 0.05$.

I first test if the mean lengths of teeth under two supplement types are different. The result shows that the p-value $= 0.06 > 0.05$ and the confidence interval includes zero, and so under $\alpha = 0.05$ using OJ or VC does not have significant difference for teeth growth.

I then compare tooth growth by dose. Since we are interested in if a higher dose level leads to longer teeth, I use one-sided test instead. Results above show that teeth lengths are statistically different under different dose levels.

Finally, to get more information about how `len` are related to `supp` and `dose`. I test all 15 different comparisons by supp and dose. If we don't do any correction, variable `pvalue`, confidence interval `lower` and `upper` and variable `sig` show that all paired comparisons are significant except two cases: `VC.2 vs OJ.1` and `VC.2 vs OJ.2`, i.e., VC with dose level 2mg vs OJ with dose level 1mg and VC with dose level 2mg vs OJ with dose level 2mg. If we use bonferroni correction, which is the most conservative one, we have three more insignificant results: `VC.0.5 vs OJ.0.5`, `VC.1 vs OJ.0.5`, and `OJ.1 vs OJ.2` by checking variable `pBon` and `sigBon`. If we use FDR correction with `method = "BH"`, only one more comparison `VC.1 vs OJ.0.5` is insignificant, as shown in `sigBH`.

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

Solution

Both confidence interval and hypothesis tesing suggest that dose level affects tooth growth, regardless of supplement types. Also, in general, using OJ makes teeth grows faster than using VC, althogh these two effects are similar when dose level is 2mg. This analysis is under assumptions that

- Those guinea pigs are randomly assigned to different types and dose levels, and hence any two subsamples for comparision are independent.
- The variances of subpopulations are unequal.
- Any sampling distribution is Student-t.
- The significant level is set at 0.05.