Part2: Basi Inference

Haonan

October 26, 2016

Overview

```
library(datasets)
library(ggplot2)

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
```

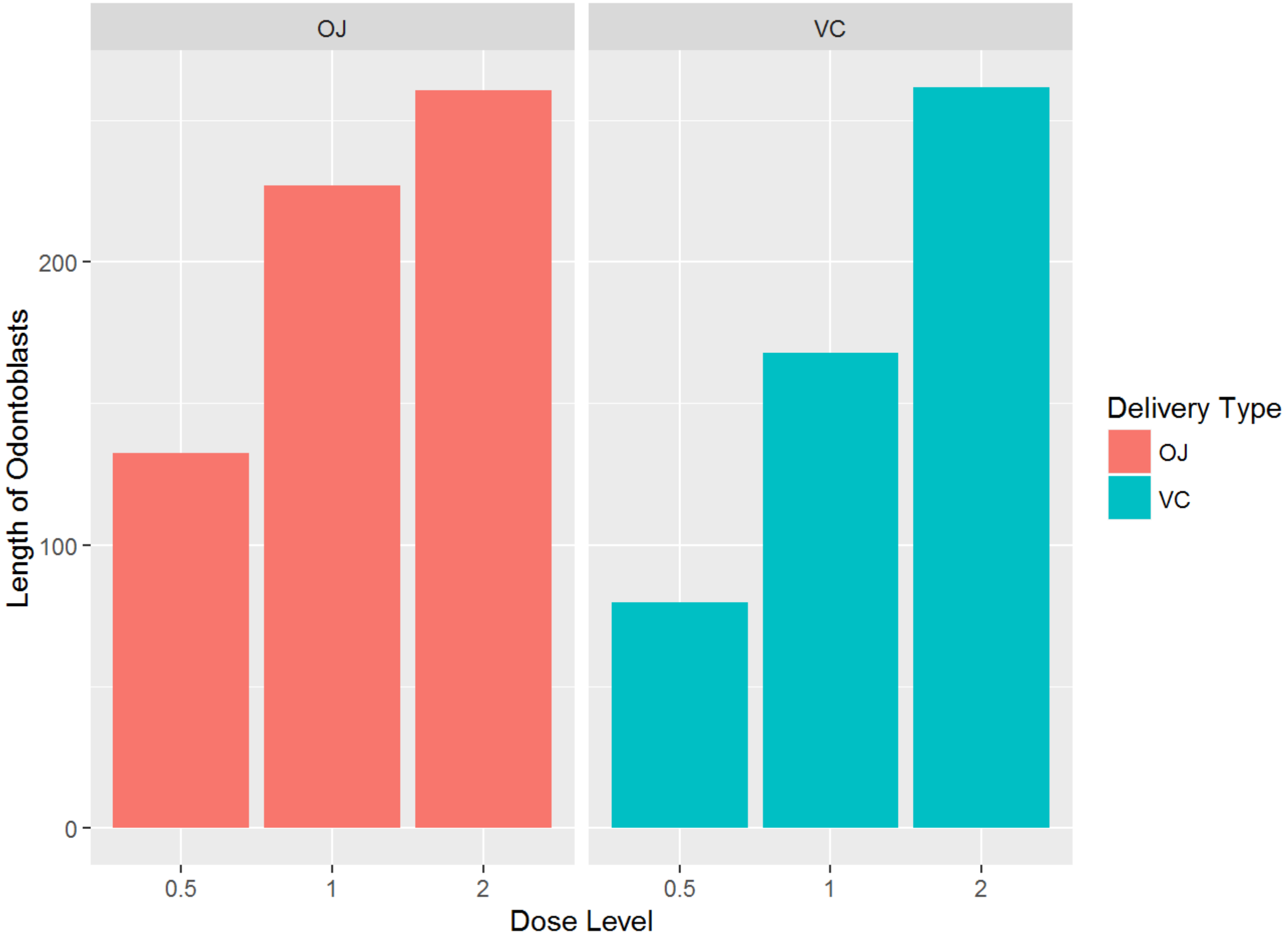
Description of the dataset:

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).

3 variables indicates the length of odontoblasts (len), type of delivery (supp) and dose levels (dose)

We plot the dataset using barplot, letting len as "y"" and dose as "x", group by supp:

```
gg <- ggplot(ToothGrowth,aes(x = as.factor(dose), y = len, fill = supp)) +
    geom_bar(stat = "identity") +
    facet_grid(.~supp) +
    guides(fill = guide_legend(title = "Delivery Type")) +
    labs(x = "Dose Level", y = "Length of Odontoblasts")
gg
```



We could identify that when the dose level is 0.5 and 1, there're a difference between OJ and VC. But when the dose level is 2, they look identical. We could make an assumption that there're difference at dose level 0.5 and 1, and we'll verify it later using t test

Take a look at its structure

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

```
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 ...
```

The length ranges from 4.20 to 33.90, with mean = 18.81. There're only 3 dose levels which are 0.5, 1 and 2. With each delivery type, there're 30 observations.

Analysis of Hypothesis Test

Comparison by "supp"

The first thing we would like to know is whether there's a significant difference between the different delivery types. We'll use the two sample t test to verify it

```
t.test(len~supp, data = ToothGrowth)
```

```
## 
##  Welch Two Sample t-test
## 
## data:  len by 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
```

Take a look at the result and we could find the mean difference is 3.7, the p-value is 0.06, very close to 0.05(which is the critical value to distinguish significance and insignificance). But statistically, we think the difference is not significant, even though it's nearly significant.

The next step is to make a comparison between different delivery by dose

Firstly, subset the data by delivery, and then conduct t test by each dose

```
oj <- subset(ToothGrowth, supp == "OJ")
vc <- subset(ToothGrowth, supp == "VC")
```

Dose == 0.5:

```
t.test(subset(oj, dose == .5)$len, subset(vc, dose == .5)$len)
```

```
##
##  Welch Two Sample t-test
##
## data:  subset(oj, dose == 0.5)$len and subset(vc, dose == 0.5)$len
## t = 3.1697, df = 14.969, p-value = 0.006359
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.719057 8.780943
## sample estimates:
## mean of x mean of y
##     13.23      7.98
```

Dose == 1

```
t.test(subset(oj, dose == 1)$len, subset(vc, dose == 1)$len)
```

```
##
##  Welch Two Sample t-test
##
## data:  subset(oj, dose == 1)$len and subset(vc, dose == 1)$len
## t = 4.0328, df = 15.358, p-value = 0.001038
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  2.802148 9.057852
## sample estimates:
## mean of x mean of y
##     22.70     16.77
```

Dose == 2

```
t.test(subset(oj, dose == 2)$len, subset(vc, dose == 2)$len)
```

```
##
##  Welch Two Sample t-test
##
## data:  subset(oj, dose == 2)$len and subset(vc, dose == 2)$len
## t = -0.046136, df = 14.04, p-value = 0.9639
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3.79807  3.63807
## sample estimates:
## mean of x mean of y
##     26.06     26.14
```

As we could see from the result, when the dose is 0.5 and 1, the difference of length is significant, but when the dose is 2, it is almost identical

Comparison by “dose”

From the plot we could see that there’s a difference in length between different dose. But to make the result more prudent, we’ll conduct a t test to verify it

Dose = 0.5 and 1:

```
t.test(subset(ToothGrowth, dose == .5)$len, subset(ToothGrowth, dose == 1)$len)
```

```
##
##  Welch Two Sample t-test
##
## data:  subset(ToothGrowth, dose == 0.5)$len and subset(ToothGrowth, dose == 1)$len
## 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
```

Dose = 0.5 and 2:

```
t.test(subset(ToothGrowth, dose == .5)$len, subset(ToothGrowth, dose == 2)$len)
```

```
##
##  Welch Two Sample t-test
##
## data:  subset(ToothGrowth, dose == 0.5)$len and subset(ToothGrowth, dose == 2)$len
## 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
```

Dose = 1 and 2:

```
t.test(subset(ToothGrowth, dose == 1)$len, subset(ToothGrowth, dose == 2)$len)
```

```
##
##  Welch Two Sample t-test
##
## data:  subset(ToothGrowth, dose == 1)$len and subset(ToothGrowth, dose == 2)$len
## 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:
##  -8.996481 -3.733519
## sample estimates:
## mean of x mean of y
##    19.735    26.100
```

The result is obvious, there’s a significant difference of length under different dose. We successfully verifed the assupmtion we talked before.

Conclusion

Given the reasonable assumption:

1: The data in the dataset is all correct

2: The sample is big enough to represent the population

From the basic analysis of the ToothGrowth data, we could make several conclusion as follows:

1: The more the dosage, the better then length

2: In general, type of delivery will not make a big difference, however, when the dosage is small enough(less than 1), it will cause great difference in length. Under this situation, OJ is better than VC

3: When the dosage equals 2, both VC and OJ will lead to the same length.