Statistical Inference Project 2

Saturday, December 20, 2014

Project Overview

In this porject, the analysis of the **ToothGrowth** data in the **R datasets** package. The analysis starts with a exploratory plot for the data, and then proposes several hypothesis accordingly. The last part of the analysis tries to verify the hypothesis proposed before by statistical hypothesis test technique and makes a conclusion.

Exploratory Data Analysis

Description of The Source Data

- The data **ToothGrowth** describes the effect of Vitamin C on tooth growth in guinea pigs.
- The response is the length of odontoblasts (teeth) 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 or ascorbic acid).

See **Appendix** for the graph

Intuitive Sense on The Data

According to the graph, we can see that probably:
1. The supplement type OJ (orange juice) is more efficiet than VC (Vitamin C), especially for small dose levels of supplement.
2. The tooth growth of pigs is positively affected by the dose levels of supplement for both supplement types.

Hypothesis Testing

In order to verify our assumption, we need to do statistical hypothesis test. First of all, we need to make some reasonal assumptions for our analysis:
1. The experimental targets (pigs) are independently treated.
2. In the following analysis, all the compared groups are treated as independent groups with constant variance.

Hypothesis 1:

- Assumption: We construct two independent groups of obervations: (1) dose level 1, supplement OJ; (2) dose level 1, supplement VC. We assume that the observations within each group are i.i.d.
- Hypothesis Test:
  - **Null Hypothesis: For dose level = 1, the mean of Length with supplement OJ is not more than that of VC**

$$H_0 : \mu_{VC} \geq \mu_{OJ}$$

- **Alternative Hypothesis: For dose level = 1, the mean of Length with supplement OJ is greater than that of VC**

$$H_a : \mu_{VC} < \mu_{OJ}$$

The analysis is implemented in **R** by one-side t-test:

```
# For dose level = 1, the mean of Length with supplement OJ is
# higher than that of VC.
df1 <- ToothGrowth[ToothGrowth$dose==1,]
t.test(len~supp,data=df1,alternative="greater",mu=0,
       paired = FALSE, var.equal = FALSE,
       conf.level = 0.95)
```

```
##
##  Welch Two Sample t-test
##
## data:  len by supp
## t = 4.0328, df = 15.358, p-value = 0.0005192
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  3.356158      Inf
## sample estimates:
## mean in group OJ mean in group VC
##            22.70            16.77
```

The p-value for the t-test is 0.0005192, which is smaller than 5%. The t-statistics is 4.0328 lies in the 95% confidence interval $[3.356, \infty)$. So under the 95% confidence level, We can reject the NULL Hypothesis and accept the alternative one:
For dose level = 1, the mean of Length with supplement OJ is greater than that of VC

Hypothesis 2:

- Assumption: We construct two independent groups of obervations: (1) dose level 1, supplement VC; (2) dose level 0.5, supplement VC. We assume that the observations within each group are i.i.d.
- Hypothesis Test:
  - **Null Hypothesis: For supplement VC, the mean of Length with dose level 0.5 is not less than that with dose level 1.**

$$H_0 : \mu_{0.5} \geq \mu_1$$

  - **Alternative Hypothesis: For supplement VC, the mean of Length with dose level 0.5 is less than that with dose level 1.**

$$H_a : \mu_{0.5} < \mu_1$$

The analysis is implemented in **R** by one-side t-test:

```
# For supplement VC, the mean of Length with dose level 1 is
# higher than that with dose level 0.5.
df2 <- ToothGrowth[ToothGrowth$supp=="VC" & ToothGrowth$dose<2,]
t.test(len~dose,data=df2,alternative="less",mu=0,
       paired = FALSE, var.equal = FALSE,
       conf.level = 0.95)
```

```
##
##  Welch Two Sample t-test
##
## data:  len by dose
## t = -7.4634, df = 17.862, p-value = 3.406e-07
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf -6.746867
## sample estimates:
## mean in group 0.5   mean in group 1
##              7.98             16.77
```

The p-value for the t-test is 3.41e-7, which is smaller than 5%. The t-statistics is -7.4634 lies in the 95% confidence interval $[-\infty, -6.747]$. So under the 95% confidence level, We can reject the NULL Hypothesis and accept the alternative one:
Alternative Hypothesis: For supplement VC, the mean of Length with dose level 0.5 is less than that with dose level 1.

Appendix: Visualization of The Data

```
library(datasets)
data(ToothGrowth)
library(ggplot2)

ggplot(data=ToothGrowth,mapping=aes(x=factor(dose),y=len,fill=factor(dose))) +
  geom_boxplot() + geom_jitter() + facet_wrap(~supp) +
  labs(title="ToothGrowth data: \n Length vs Dose, given type of supplement",
       x="Dose",y="Length",fill="Dose") +
  theme(title=element_text(face = "bold"),
        strip.text=element_text(face = "bold",colour = "blue"))
```