Statistical Inference Course Project part2

Yao Bin

June 18, 2015

Overview

- In this study, we analyze the ToothGrowth data in the R datasets package.
- Exploratory analysis of basic features are performed.
- Then, confidence intervals and hypothesis tests are performed.
- At last, brief conclusions and assumptions are made.
- Code chunks for data analysis and plots are demonstrated in the appendix in the end of the document.

Exploratory data analysis

First, load the ToothGrowth data and check the structure and type of data (see Chunk1).

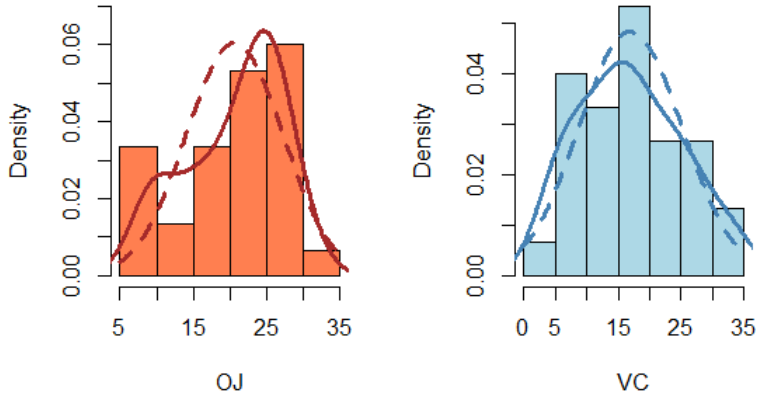
```
'data.frame':  60 obs. of  3 variables:
 $ len : num  4.2 11.5 7.3 5.8 6.4 10 11.2 11.2 5.2 7 ...
 $ supp: Factor w/ 2 levels "OJ","VC": 2 2 2 2 2 2 2 2 2 2 ...
 $ dose: num  0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 ...
```

We can tell that the data contains the tooth length of pigs fed with different doses of VC/OJ supplement. We reconstructure and summarize the data with "dplyr" package.(see chunk2)

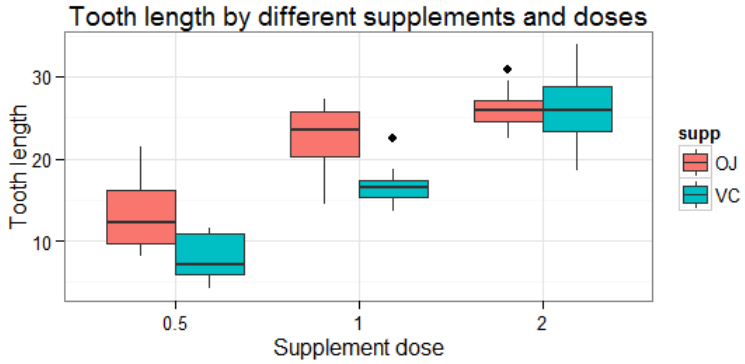
```
Source: local data frame [6 x 7]
Groups: supp

  supp dose  min median  max  mean     sd
1   OJ  0.5  8.2  12.25 21.5 13.23 4.4597
2   OJ  1.0 14.5  23.45 27.3 22.70 3.9110
3   OJ  2.0 22.4  25.95 30.9 26.06 2.6551
4   VC  0.5  4.2   7.15 11.5  7.98 2.7466
5   VC  1.0 13.6  16.50 22.5 16.77 2.5153
6   VC  2.0 18.5  25.95 33.9 26.14 4.7977
```

Then, we make some plots to get intuitive information. Check if the data is close to normal distribution, which most biomedical datasets are close to. The plot shows that the data is not so close to normal distribution (see chunk3), which determines the kind of statistical test we use.



Check the relationship between supplement type/dose and tooth length. (see Chunk4)



Confidence intervals and hypothesis tests

To compare tooth growth by supplement and dose, we perform confidence intervals and t-tests. For the data is not close to normal distribution, Welch t-test is used here. Significance level *a* is *0.05*, and confidential level is *0.95*. (see chunk3)

First, does tooth growth differ by supplement type?

Generally, the confidential interval is **[-0.17102, 7.57102]**. The p value is **0.06063**, which indicates that there is no significant difference between length by supplements.

Does tooth growth differ by different doses of two kinds of supplements?

```
     CI.L   CL.U   p.value
1  1.7191 8.7809 0.0063586
2  2.8021 9.0579 0.0010384
3 -3.7981 3.6381 0.9638516
```

At dose 0.5 mg, there is significant difference between length by the two supplements.
At dose 1.0 mg, there is significant difference between length by the two supplements.
At dose 2.0 mg, there is no significant difference between length by the two supplements.

Then, does tooth growth depend on dose for each supplement, respectly?

For VC:

```
          CI.L    CL.U    p.value
0.5~1.0 -12.659 -6.0612 0.0005573
0.5~2.0 -25.629 -11.651 0.00070937
1.0~2.0 -15.322 -1.0376 0.030569
```

Between 0.5 mg and 1.0 mg dose group, there is significant difference.
Between 0.5 mg and 2.0 mg dose group, there is significant difference.
Between 1.0 mg and 2.0 mg dose group, there is significant difference.

For OJ:

```
          CI.L    CL.U     p.value
0.5~1.0 -14.947 -5.6131  0.0012455
0.5~2.0 -16.943 -7.0967  0.00053965
1.0~2.0 -11.236 -0.16356 0.045095
```

Between 0.5 mg and 1.0 mg dose group, there is significant difference.
Between 0.5 mg and 2.0 mg dose group, there is significant difference.
Between 1.0 mg and 2.0 mg dose group, there is significant difference.

Conclusions and assumptions

1. At lower dose, 0.5 mg and 1.0 mg, teeth grow longer when pigs are fed with OJ than with VC; at higher dose, 2.0 mg, there are no significant differences between the two supplements.

2. For each supplement respectly, tooth length depends on the dose of supplement.

Appendix: Code chunks

Chunk1. Load necessary packages and data

```
library("datasets")
library("dplyr")
library("ggplot2")
tgdata<-ToothGrowth
```

Chunk2. Investigate structure of data

```
# head and tail of the data
datahead<-head(tgdata, 3)
datatail<-tail(tgdata, 3)

# structure of data
str(tgdata)

# summary of data
summary(tgdata)

#
group_by(tgdata, supp,dose) %>%
        summarize(min=min(len), median=median(len), max=max(len), mean=
mean(len), sd=sd(len))
```

Chunk3. Tooth length density curve versus normal distribution curve

```
len_oj<-tgdata[tgdata$supp=="OJ", "len"]
len_vc<-tgdata[tgdata$supp=="VC", "len"]
par(mfrow=c(1,2))
hist(len_oj, prob=T, col="lightblue", main="")
curve(dnorm(x, mean=mean(len_oj), sd=sd(len_oj)), add=T, lty=2, lwd=3,
col="steelblue")
lines(density(len_oj), col="steelblue", lwd=3)
hist(len_vc,prob=T, col="coral", main="")
curve(dnorm(x, mean=mean(len_vc), sd=sd(len_vc)), add=T, lty=2, lwd=3,
col="brown")
lines(density(len_vc), col="brown", lwd=3)
```

Chunk4. Tooth length by different supplements and doses

```
library("ggplot2")
ggplot(tgdata, aes(x=as.factor(dose), y=len))+
        geom_boxplot(aes(fill=supp))+
```

```
        labs(title="Tooth length by different supplements and doses",
             x="Supplement dose",
             y="Tooth length")+
        theme_bw()
```

Chunk5. Confidence intervals and hypothesis tests

```
# len by supp
t0<-with(tgdata, t.test(len~supp, paired=F, var.equal=F))
CI.L.0<-t0$conf.int[1]
CI.U.0<-t0$conf.int[2]
p0<-t0$p.value

t.supp<-data.frame()
dl<-unique(tgdata$dose) #dose.list
for(i in dl) {
        t.i<-t.test(len~supp, paired=F, var.equal=F, data=filter(tgdat
a, dose==i) )
        CI.L.i<-t.i$conf.int[1]
        CI.U.i<-t.i$conf.int[2]
        p.i<-t.i$p.value
        result.i<-cbind(CI.L.i, CI.U.i, p.i)
        t.supp<-rbind(t.supp, result.i)
}
names(t.supp)<-c("CI.L", "CL.U", "p.value")

# len by dose for each supp
tg.a<-filter(tgdata, dose==c(dl[1], dl[2]))
tg.b<-filter(tgdata, dose==c(dl[1], dl[3]))
tg.c<-filter(tgdata, dose==c(dl[2], dl[3]))

# len by dose for supp "VC"
vc.a<-t.test(len~dose, paired=F, var.equal=F, data=filter(tg.a, supp=="
VC") )
vc.b<-t.test(len~dose, paired=F, var.equal=F, data=filter(tg.b, supp=="
VC") )
vc.c<-t.test(len~dose, paired=F, var.equal=F, data=filter(tg.c, supp=="
VC") )
results.vc<-as.data.frame(rbind(vc.a, vc.b, vc.c))
CI.vc<-rbind(vc.a$conf.int, vc.b$conf.int, vc.c$conf.int)
p.value.vc<-results.vc[,"p.value"]
t.vc<-cbind(CI.vc, p.value.vc)
rownames(t.vc)<-c("0.5~1.0", "0.5~2.0", "1.0~2.0")
colnames(t.vc)<-c("CI.L", "CL.U", "p.value")

# len by dose for supp "OJ"
oj.a<-t.test(len~dose, paired=F, var.equal=F, data=filter(tg.a, supp=="
OJ") )
oj.b<-t.test(len~dose, paired=F, var.equal=F, data=filter(tg.b, supp=="
```

```
OJ") )
oj.c<-t.test(len~dose, paired=F, var.equal=F, data=filter(tg.c, supp=="
OJ") )
results.oj<-as.data.frame(rbind(oj.a, oj.b, oj.c))
CI.oj<-rbind(oj.a$conf.int, oj.b$conf.int, oj.c$conf.int)
p.value.oj<-results.oj[,"p.value"]
t.oj<-cbind(CI.oj, p.value.oj)
rownames(t.oj)<-c("0.5~1.0", "0.5~2.0", "1.0~2.0")
colnames(t.oj)<-c("CI.L", "CL.U", "p.value")

# output for hypothesis test
# this funcion put out a sentence judging if the data is significantly
different.
a<-0.05
output.test<-function(p.data){
        if(p.data<=a){
                paste("there is significant difference")
        }
                else{
                        paste("there is no significant difference")
                }
}
```