# 5303hw3

*JinYao*

*2019/9/23*

## E5.5

```
library(cfcdae)
data("OrangePulpSilage")
oran <- lm(moisture ~ treatment, data = OrangePulpSilage)
compare.to.control(oran,treatment,control=2)
```

```
##                         difference     lower     upper
##   BeetPulp - Control           0.6 -7.600875  8.800875
##   NaCl - Control               2.1 -6.100875 10.300875
##   FormicAcid - Control         4.5 -3.700875 12.700875
```

```
# we should reject the null hypothesis because 0 is in the confidence interval, so the
# delay after exposure affects leaflet angle.
```

## P5.1

```
data("FruitFlyLifespan")
Fru <- lm(longevity~companions, data = FruitFlyLifespan)
sidelines(pairwise(Fru, companions, type = "hsd"))
```

```
##
## 8virgin   -18.72
## 8pregnant  -0.68 |
## none        5.92 |
## 1pregnant   6.12 |
## 1virgin     7.36 |
```

```
# It shows that Treatment 8virgin is lower than the other four treatments.
# Which is significant.This experiment have 5 treatments, and each treatment has
# 25 experimental untis. We should reject the null hypothesis and conclude that
# the reproductive affects longevity.
```

## P5.2

```
data("Autoclaving")
auto <- lm(modulus ~ treatment, data = Autoclaving)
pairwise(auto, treatment, type = "hsd")
```

```
##
## Pairwise comparisons ( hsd ) of treatment
##                     estimate signif diff        lower     upper
##   121_10 - 121_20   77.46667    177.3803  -99.9136199  254.8470
```

```
##   121_10 - 135_10  131.36667    177.3803  -46.0136199  308.7470
##   121_10 - 135_20  176.83333    177.3803   -0.5469532  354.2136
## * 121_10 - None   -317.96667    177.3803 -495.3469532 -140.5864
##   121_20 - 135_10   53.90000    177.3803 -123.4802866  231.2803
##   121_20 - 135_20   99.36667    177.3803  -78.0136199  276.7470
## * 121_20 - None   -395.43333    177.3803 -572.8136199 -218.0530
##   135_10 - 135_20   45.46667    177.3803 -131.9136199  222.8470
## * 135_10 - None   -449.33333    177.3803 -626.7136199 -271.9530
## * 135_20 - None   -494.80000    177.3803 -672.1802866 -317.4197
```

```
# Because the sample sizes are the same sizes, and we do all pairs to find if there is
# difference between each sample.
```

## P5.6

We want to make sure all locations are safe, so I think to control type II error, which means that I think false rejections are not so important. For per comparison error rate, the good news is that we can make sure every location is as safe as possible, because it allows some false rejections even the null hypothesis is true.But the bad news is that the pay is to big, because as the sample sizes increase sharply, the possibility of reject the null hypothesis will increase, so after several test, you will find that almost all the locations needs to be inspected, then the cost will increase incredibly. For FDR, the pros is that we are allowed more incorrect rejections as the number of true
rejections increase, then decrease the type II error, so that it is less likely to wrongly reject the null hypothesis, so that we can save more cost, but the type II error will increase. For the SFER, this is the most accurate, which means that it decreases the type I error to the least in comparison with the other two precedures, but the problem is that, controlling stronger error rates leads to less powerful tests, so it is more likely to have type II error, which is vital. In conclusion, I will choose the FDR, because it can control the type I error in a relatively low rate, but don't allow the type II error as big as the SFER.

## P5.9

As the sample size of the data increase, and the same hypothesis is tested multiple times, it has a trend to inflate the type I error, so there will be more false rejections, which means that it tends to conclude that there are significance, so this statistical issues are for ESP as the one sample z-test repeats.

## E6.5

(a) the first seems that it violates the independence, because as time goes by, there is a pattern that shows a serial correlation.
(b) it violates the non-constant variance, the pattern is a right-opening megaphone, with the increase of the fitted value, the residuals are more scattered.
(c) it is like a normal plot and it's good I think.
(d) it violates the normality, it looks like a long tail distribution.

## P6.1a

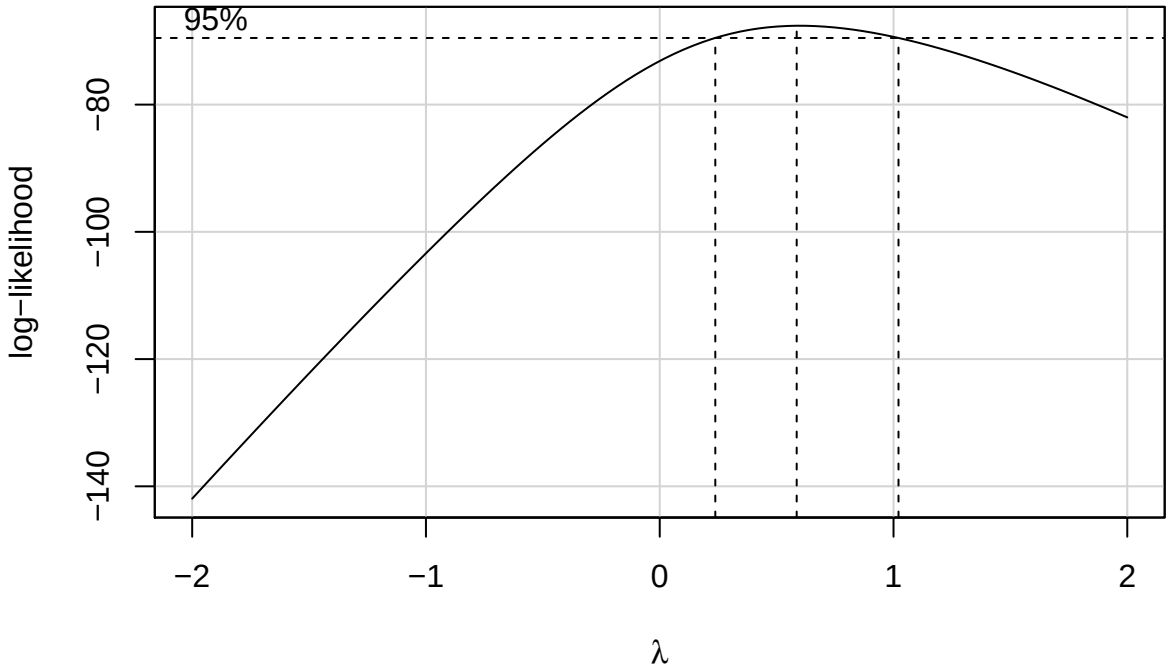
The largest value is around 100, and the smallest value is around 50, so the ratio is at about 2, so I think it is not necessary to take transformation to the data. While, I just found your hint. . . So it's a little bit tricky.

Because what we measured is the percent of the grass not the bluestem, so we need to use 100 to substract, then the range becomes really big:

```
res1 <- c(100-97, 100-96, 100-92, 100-95)
res2 <- c(100-83, 100-87, 100-78, 100-81)
res3 <- c(100-85, 100-84, 100-78, 100-79)
res4 <- c(100-64, 100-72, 100-63, 100-74)
res5 <- c(100-52, 100-56, 100-44, 100-50)
res6 <- c(100-48, 100-58, 100-49, 100-53)
res <- c(res1,res2,res3,res4,res5,res6)
treat <- c(rep("treat1",4),rep("treat2",4),rep("treat3",4),rep("treat4",4),rep("treat5",4),
           rep("treat6",4))
data <- data.frame(treat,res)
x <- lm(res~treat)
library(car)
```

```
## Loading required package: carData
```

```
#plot(x)
boxCox(x)
```




So from above, I can use the power to the square root to tast the large range between the largest and the smallest.

## P6.2

(a) The problem is that the ratio of the observed proportions of the largest and smallest are two big, more than 4.
(b) We need a transformation for the data.