Project Report 2

Yilun Zhang

Wednesday, October 22, 2014

Load data

Load data and show the first few lines of the data.

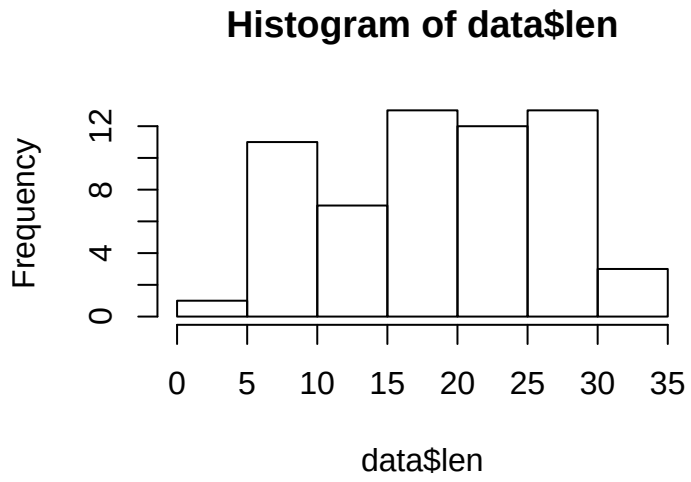
```
data(ToothGrowth)
data <- ToothGrowth
data$dose <- as.factor(data$dose)
head(data)
```

```
##    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
```

Exploratory analysis on ToothGrowth data ans basic summary

Histogram of len variable.

```
hist(data$len)
```



Mean and standard deviation of len by supp.

```
tapply(data$len,data$supp,mean)
```

```
##       OJ       VC
## 20.66333 16.96333
```

```
tapply(data$len,data$supp,sd)
```

```
##       OJ       VC
## 6.605561 8.266029
```

This shows that OJ has an average len of 20.66333 and VC has an average len of 16.96333 The standard deviations of len for the two catogories are 6.605561 for OJ and 8.266029 for VC.

Mean and standard deviation of len by supp.

```
tapply(data$len,data$dose,mean)
```

```
##    0.5      1      2
## 10.605 19.735 26.100
```

```
tapply(data$len,data$dose,sd)
```

```
##      0.5        1        2
## 4.499763 4.415436 3.774150
```

This shows that the mean len are 10.605 for 0.5 dose group, 19.735 for 1 dose group and 26.100 for 2 dose group.
The standard deviations of len are 4.499763 for 0.5 dose group, 4.41536 for 1 dose group and 3.774150 for 2 dose group. Basic summary of data.

```
summary(data)
```

```
##       len        supp     dose
##  Min.   : 4.20   OJ:30   0.5:20
##  1st Qu.:13.07   VC:30   1  :20
##  Median :19.25           2  :20
##  Mean   :18.81
##  3rd Qu.:25.27
##  Max.   :33.90
```

Confidence interval and hypothesis test on supp and dose groups

In the exploratory analysis part, we have already calculated the mean and standard deviations for supp subgroups and dose subgroups, and we also know from the summary that there are 30 each for OJ and VC for supp group and 20 each for each dose level.

supp group Confidence interval and hypothesis test for supp group. For the hypothesis test, we are assuming that the null hypothesis is: the two supp groups don't cause difference in len values.

```
# OJ group
tapply(data$len,data$supp,mean)[1] +
  c(-1,1)*qt(0.975,30-1)*tapply(data$len,data$dose,sd)[1]/sqrt(30)
```

```
## [1] 18.98309 22.34357
```

```r
# VC group
tapply(data$len,data$supp,mean)[2] +
  c(-1,1)*qt(0.975,30-1)*tapply(data$len,data$dose,sd)[2]/sqrt(30)
```

```
## [1] 15.31458 18.61208
```

```r
# t-test
t.test(data[data$supp == "OJ",]$len,data[data$supp == "VC",]$len,paied=F)
```

```
##
##  Welch Two Sample t-test
##
## data:  data[data$supp == "OJ", ]$len and data[data$supp == "VC", ]$len
## 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 of x mean of y
##  20.66333  16.96333
```

We have a p-value of 0.06063>0.05 for the t test, so we cannot reject the null hypothesis which is: the two supp groups don't cause difference in len values.

dose group Confidence interval and hypothesis test for dose group. For the hypothesis test, we are assuming that the null hypothesis is: the 3 dose groups don't cause difference in len values.

```r
# 0.5 dose group
tapply(data$len,data$dose,mean)[1] +
  c(-1,1)*qt(0.975,20-1)*tapply(data$len,data$dose,sd)[1]/sqrt(20)
```

```
## [1]  8.499046 12.710954
```

```r
# 1 dose group
tapply(data$len,data$dose,mean)[2] +
  c(-1,1)*qt(0.975,20-1)*tapply(data$len,data$dose,sd)[2]/sqrt(20)
```

```
## [1] 17.66851 21.80149
```

```r
# 2 dose group
tapply(data$len,data$dose,mean)[3] +
  c(-1,1)*qt(0.975,20-1)*tapply(data$len,data$dose,sd)[3]/sqrt(20)
```

```
## [1] 24.33364 27.86636
```

```r
# t-test: 0.5 vs 1
t.test(data[data$dose == "0.5",]$len,data[data$dose == "1",]$len,paied=F)
```

```
##
##  Welch Two Sample t-test
##
## data:  data[data$dose == "0.5", ]$len and data[data$dose == "1", ]$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
```

```
# t-test: 0.5 vs 2
t.test(data[data$dose == "0.5",]$len,data[data$dose == "2",]$len,paied=F)
```

```
##
##  Welch Two Sample t-test
##
## data:  data[data$dose == "0.5", ]$len and data[data$dose == "2", ]$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
```

```
# t-test: 1 vs 2
t.test(data[data$dose == "1",]$len,data[data$dose == "2",]$len,paied=F)
```

```
##
##  Welch Two Sample t-test
##
## data:  data[data$dose == "1", ]$len and data[data$dose == "2", ]$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
```

We have all the p-values being extremely less than 0,05 for all 3 t-tests, that means, we have sufficient evidence to conclude that dose is an important factor that affects len values.