Hypothesis Case

Niranjan

20 August 2018

```
library(data.table)
```

```
## Warning: package 'data.table' was built under R version 3.4.4
```

```
options(scipen = 999)
```

Q1

Importing the data file

```
diet <- data.table::fread(file = "~/Case studies/Hypothesis Testing/dietstudy.csv",sep = ",",
 stringsAsFactors = F)
```

As the question is to find a comparision between pre and post test, a two sample t-test will be used

analyse the results of the test

Testing if there was any significant difference in the weights

```
t.test(diet$wgt0,diet$wgt1, paired = T)
```

```
##
##  Paired t-test
##
## data:  diet$wgt0 and diet$wgt1
## t = 3.6948, df = 15, p-value = 0.002162
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.9520306 3.5479694
## sample estimates:
## mean of the differences
##                    2.25
```

The first time weights after the new diet shows that there is a significant difference between the pre diet

```
t.test(diet$wgt0,diet$wgt2, paired = T)
```

```
##
##  Paired t-test
##
## data:  diet$wgt0 and diet$wgt2
## t = 9.4058, df = 15, p-value = 0.0000001112
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  3.286909 5.213091
## sample estimates:
## mean of the differences
##                   4.25
```

The Second time weights after the new diet shows that there is a significant difference between the pre diet

```
t.test(diet$wgt0,diet$wgt3, paired = T)
```

```
##
##  Paired t-test
##
## data:  diet$wgt0 and diet$wgt3
## t = 10.263, df = 15, p-value = 0.00000003546
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  4.952031 7.547969
## sample estimates:
## mean of the differences
##                   6.25
```

The Third time weights after the new diet shows that there is a significant difference between the pre diet

```
t.test(diet$wgt0,diet$wgt4, paired = T)
```

```
##
##  Paired t-test
##
## data:  diet$wgt0 and diet$wgt4
## t = 11.175, df = 15, p-value = 0.00000001138
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  6.524643 9.600357
## sample estimates:
## mean of the differences
##                  8.0625
```

The fourth time weights after the new diet shows that there is a significant difference between the pre diet

In all the above cases as the p-value>0.05 we can conclude that the new diet causes a significant difference

in the weights of the participants

Testing if there was any significant difference in the triglyceride levels

```
t.test(diet$tg0,diet$tg1,paired = T)
```

```
##
##  Paired t-test
##
## data:  diet$tg0 and diet$tg1
## t = 1.7083, df = 15, p-value = 0.1082
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3.436468 31.186468
## sample estimates:
## mean of the differences
##                  13.875
```

The first time measure of the triglyceride levels are not significantly different from before the diet

```
t.test(diet$tg0,diet$tg2,paired = T)
```

```
##
##  Paired t-test
##
## data:  diet$tg0 and diet$tg2
## t = 1.4653, df = 15, p-value = 0.1635
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -6.393188 34.518188
## sample estimates:
## mean of the differences
##                 14.0625
```

The second time measure of the triglyceride levels are not significantly different from before the diet

```
t.test(diet$tg0,diet$tg3,paired = T)
```

```
##
##  Paired t-test
##
## data:  diet$tg0 and diet$tg3
## t = 1.646, df = 15, p-value = 0.1205
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -5.787471 45.037471
## sample estimates:
## mean of the differences
##                  19.625
```

The third time measure of the triglyceride levels are not significantly different from before the diet

```
t.test(diet$tg0,diet$tg4,paired = T)
```

```
##
##  Paired t-test
##
## data:  diet$tg0 and diet$tg4
## t = 1.2, df = 15, p-value = 0.2487
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -10.91541  39.04041
## sample estimates:
## mean of the differences
##                 14.0625
```

The fourth time measure of the triglyceride levels are not significantly different from before the diet

All the results from this test show that the diet caused no difference in the triglyceride levels of the participants

Thus, it can be concluded that the new diet causes a difference only in the weights, but not in the triglyceride levels for patients with a family history of heart disease.

Q2

Importing the file

```
credit <- fread("~/Case studies/Hypothesis Testing/creditpromo.csv", sep = ",", stringsAsFact
ors = F)
credit
```

```
##          id        insert  dollars
##   1:    148      Standard 2232.772
##   2:    572 New Promotion 1403.808
##   3:    973      Standard 2327.092
##   4:   1096      Standard 1280.031
##   5:   1541 New Promotion 1513.563
##  ---
## 496: 130163 New Promotion 1513.060
## 497: 130204 New Promotion 1020.758
## 498: 130255      Standard 1919.856
## 499: 130583      Standard 1863.015
## 500: 130682      Standard 1295.012
```

This question requires comparision between two groups. Thus a grouped t test needs to be conducted

```
t.test(dollars~insert,data = credit,var.equal=T)
```

```
##
##  Two Sample t-test
##
## data:  dollars by insert
## t = 2.2604, df = 498, p-value = 0.02423
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##    9.301956 132.919948
## sample estimates:
## mean in group New Promotion      mean in group Standard
##                    1637.500                    1566.389
```

```
t.test(dollars~insert,data = credit)
```

```
##
##  Welch Two Sample t-test
##
## data:  dollars by insert
## t = 2.2604, df = 497.6, p-value = 0.02423
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##    9.301833 132.920071
## sample estimates:
## mean in group New Promotion      mean in group Standard
##                    1637.500                    1566.389
```

As the t value for both the assumptions are same, we can conclude that there is no significant difference in variance between the groups

From the test we observe that the p-value is <0.05, thus it can be concluded here that both the groups have different means. So, H0 is rejected.

This shows that there is a significant increase in sales from the new promotion as its mean is significantly different from the mean of the group that was given standard discount

Q3

Importing the data

```
pollination <- fread("~/Case studies/Hypothesis Testing/pollination.csv", sep= ",", stringsAs
Factors = F)
```

a

Here we need to find if the overall yield is significantly different from 200 or not

This would require a one sample t test

```
t.test(pollination$Seed_Yield_Plant, mu = 200)
```

```
##
##  One Sample t-test
##
## data:  pollination$Seed_Yield_Plant
## t = -2.3009, df = 19, p-value = 0.03289
## alternative hypothesis: true mean is not equal to 200
## 95 percent confidence interval:
##  163.3414 198.2656
## sample estimates:
## mean of x
##  180.8035
```

As the p-value is less than 0.05, we can conclude that the seed yield/plant is not equal to 200

b

This calls for comparision of two groups, hence a grouped t-test needs to be done

1. Checking fruit weight

```
t.test(Fruit_Wt~Group, data= pollination, vars.equal = T)
```

```
##
##  Welch Two Sample t-test
##
## data:  Fruit_Wt by Group
## t = 17.67, df = 10.397, p-value = 0.000000004307
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.6279281 0.8080719
## sample estimates:
##    mean in group Hand mean in group Natural
##                 2.566                 1.848
```

```
t.test(Fruit_Wt~Group, data= pollination, vars.equal = F)
```

```
##
##  Welch Two Sample t-test
##
## data:  Fruit_Wt by Group
## t = 17.67, df = 10.397, p-value = 0.000000004307
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.6279281 0.8080719
## sample estimates:
##    mean in group Hand mean in group Natural
##                 2.566                 1.848
```

As the t values for both the assumptions are same, variance between the groups is considered to be same

The result shows that the p-value is <0.05. Thus, H0 is rejected

Hence, it can be concluded that the fruit weight is significantly different between the groups and is higher when hand pollinated

2. Checking seed yield per plant

```
t.test(pollination$Seed_Yield_Plant~pollination$Group, vars.equal= T)
```

```
##
##  Welch Two Sample t-test
##
## data:  pollination$Seed_Yield_Plant by pollination$Group
## t = 13.958, df = 17.771, p-value = 0.00000000005136
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  59.10515 80.07285
## sample estimates:
##    mean in group Hand mean in group Natural
##               215.598               146.009
```

```
t.test(pollination$Seed_Yield_Plant~pollination$Group, vars.equal= F)
```

```
##
##  Welch Two Sample t-test
##
## data:  pollination$Seed_Yield_Plant by pollination$Group
## t = 13.958, df = 17.771, p-value = 0.00000000005136
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  59.10515 80.07285
## sample estimates:
##    mean in group Hand mean in group Natural
##               215.598               146.009
```

As the t values for both the assumptions are same, variance between the groups is considered to be same

The result shows that the p-value is <0.05. Thus, H0 is rejected

Hence, it can be concluded that the seed yield per plant is significantly different between the groups and is higher when hand pollinated

3. Checking seedling length

```
t.test(pollination$Seedling_length ~pollination$Group, vars.equal= T)
```

```
##
##  Welch Two Sample t-test
##
## data:  pollination$Seedling_length by pollination$Group
## t = 2.5422, df = 16.43, p-value = 0.02143
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.1482489 1.6177511
## sample estimates:
##    mean in group Hand mean in group Natural
##                18.590                17.707
```

```
t.test(pollination$Seedling_length~pollination$Group, vars.equal= F)
```

```
##
##  Welch Two Sample t-test
##
## data:  pollination$Seedling_length by pollination$Group
## t = 2.5422, df = 16.43, p-value = 0.02143
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.1482489 1.6177511
## sample estimates:
##    mean in group Hand mean in group Natural
##                18.590                17.707
```

As the t values for both the assumptions are same, variance between the groups is considered to be same

The result shows that the p-value is <0.05. Thus, H0 is rejected

Hence, it can be concluded that the Seedling length is significantly different between the groups and is higher when hand pollinated

In conclusion from all these grouped t-test conducted, it can be said that there is a significant difference when hand pollinated and when naturally pollinated. Hand pollinated is highly effective as its means for all the test factors were high

Q4

Importing the file

```
dvd <- fread("~/Case studies/Hypothesis Testing/dvdplayer.csv", sep=",", stringsAsFactors = F
)
```

As there are a lot of groups, an ANOVA will be used to show if the customers from different groups rated it differently

```
ANOVA_dvd <- aov(dvdscore~agegroup, data = dvd)
summary(ANOVA_dvd)
```

```
##             Df Sum Sq Mean Sq F value    Pr(>F)
## agegroup     5   1294  258.90   6.993 0.0000309 ***
## Residuals   62   2296   37.02
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Here, we observe that the pr value is less than 0.05. Hence, H0 is rejected.

Thus, it can be said that there is a significant difference in how the age groups rated the dvd design

Q5

Importing the file

```
sampsurv <- fread("~/Case studies/Hypothesis Testing/sample_survey.csv", sep=",", stringsAsFa
ctors = F)
```

a

This question asks to find if there was a relationship between two categorical variable. This can be found using a chi-square test

```
tbl1 <- xtabs(~wrkstat+marital,data = sampsurv)
chisq.test(tbl1)
```

```
## Warning in chisq.test(tbl1): Chi-squared approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  tbl1
## X-squared = 729.24, df = 28, p-value < 0.00000000000000022
```

```
chisq.test(sampsurv$wrkstat,sampsurv$marital)
```

```
## Warning in chisq.test(sampsurv$wrkstat, sampsurv$marital): Chi-squared
## approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  sampsurv$wrkstat and sampsurv$marital
## X-squared = 729.24, df = 28, p-value < 0.00000000000000022
```

As the p-value is less than 0.05, H0 is rejected and we say that the expected value is not the same as observed value

This means that there is no relationship between the labour force status and marital status

b

This again a question of influence of one categorical variable on the other. Thus, we again use a chi square

```
tbl2 <- xtabs(~marital+educ,data = sampsurv)
chisq.test(tbl2)
```

```
## Warning in chisq.test(tbl2): Chi-squared approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  tbl2
## X-squared = 221.11, df = 76, p-value = 0.000000000000000431
```

As the p-value is less than 0.05, H0 is rejected and we say that the expected value is not the same as observed value

This means that the education qualification does not predict the marital status

c

This question also asks for a influence in categorical variable by other categorical variables

```
tbl3 <-  xtabs(~happy+income,data = sampsurv)
chisq.test(tbl3)
```

```
## Warning in chisq.test(tbl3): Chi-squared approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  tbl3
## X-squared = 178.95, df = 22, p-value < 0.00000000000000022
```

As the p-value is less than 0.05, H0 is rejected and we say that the expected value is not the same as observed value

This means that earnings does not predict the happiness

```
tbl4 <-  xtabs(~happy+marital,data = sampsurv)
chisq.test(tbl4)
```

```
##
##  Pearson's Chi-squared test
##
## data:  tbl4
## X-squared = 260.69, df = 8, p-value < 0.00000000000000022
```

As the p-value is less than 0.05, H0 is rejected and we say that the expected value is not the same as observed value

This means that marital status does not predict the happiness