

# Practicising\_statistical\_testing.R

kriti

Wed Sep 19 11:53:25 2018

```
###Independent T test
```

```
#The independent t test is used to test if there is any statistically
#significant difference between two means.
#Use of an independent t test requires several assumptions to be satisfied.
#The assumptions are listed below
```

```
#The variables are continuous and independent
#The variables are normally distributed
#The variances in each group are equal
```

```
#When these assumptions are satisfied the results of the t test are valid.
#Otherwise they are invalid and you need to use a non-parametric test.
#When data is not normally
#distributed you can apply transformations to make it normally distributed.
```

```
##the data to be used for these tests are goin to be mtcars
```

```
summary(mtcars)
```

```
##      mpg          cyl          disp          hp
##  Min.   :10.40   Min.    :4.000   Min.    : 71.1   Min.    : 52.0
##  1st Qu.:15.43   1st Qu.:4.000   1st Qu.:120.8   1st Qu.: 96.5
##  Median :19.20   Median :6.000   Median :196.3   Median :123.0
##  Mean   :20.09   Mean    :6.188   Mean    :230.7   Mean    :146.7
##  3rd Qu.:22.80   3rd Qu.:8.000   3rd Qu.:326.0   3rd Qu.:180.0
##  Max.   :33.90   Max.    :8.000   Max.    :472.0   Max.    :335.0
##      drat          wt          qsec          vs
##  Min.   :2.760   Min.    :1.513   Min.    :14.50   Min.    :0.0000
##  1st Qu.:3.080   1st Qu.:2.581   1st Qu.:16.89   1st Qu.:0.0000
##  Median :3.695   Median :3.325   Median :17.71   Median :0.0000
##  Mean   :3.597   Mean    :3.217   Mean    :17.85   Mean    :0.4375
##  3rd Qu.:3.920   3rd Qu.:3.610   3rd Qu.:18.90   3rd Qu.:1.0000
##  Max.   :4.930   Max.    :5.424   Max.    :22.90   Max.    :1.0000
##      am          gear          carb
##  Min.   :0.0000   Min.    :3.000   Min.    :1.000
##  1st Qu.:0.0000   1st Qu.:3.000   1st Qu.:2.000
##  Median :0.0000   Median :4.000   Median :2.000
##  Mean   :0.4062   Mean    :3.688   Mean    :2.812
##  3rd Qu.:1.0000   3rd Qu.:4.000   3rd Qu.:4.000
##  Max.   :1.0000   Max.    :5.000   Max.    :8.000
```

```
head(mtcars)
```

```
##           mpg cyl disp  hp drat   wt  qsec vs am gear carb
## Mazda RX4      21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag  21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710      22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive  21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## Valiant         18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

```
colnames(mtcars)
```

```
## [1] "mpg"  "cyl"  "disp" "hp"   "drat" "wt"   "qsec" "vs"   "am"   "gear"
## [11] "carb"
```

```
##creating labels for am variable
```

```
mtcars$am_label<-factor(mtcars$am,levels=c(0,1),labels=c("Automatic","Manual"))
mtcars[,c('am','am_label')]
```

```
##          am  am_label
## Mazda RX4      1   Manual
## Mazda RX4 Wag  1   Manual
## Datsun 710      1   Manual
## Hornet 4 Drive  0 Automatic
## Hornet Sportabout 0 Automatic
## Valiant         0 Automatic
## Duster 360      0 Automatic
## Merc 240D       0 Automatic
## Merc 230        0 Automatic
## Merc 280        0 Automatic
## Merc 280C       0 Automatic
## Merc 450SE      0 Automatic
## Merc 450SL      0 Automatic
## Merc 450SLC     0 Automatic
## Cadillac Fleetwood 0 Automatic
## Lincoln Continental 0 Automatic
## Chrysler Imperial 0 Automatic
## Fiat 128        1   Manual
## Honda Civic     1   Manual
## Toyota Corolla  1   Manual
## Toyota Corona   0 Automatic
## Dodge Challenger 0 Automatic
## AMC Javelin     0 Automatic
## Camaro Z28      0 Automatic
## Pontiac Firebird 0 Automatic
## Fiat X1-9       1   Manual
## Porsche 914-2   1   Manual
## Lotus Europa    1   Manual
## Ford Pantera L  1   Manual
## Ferrari Dino    1   Manual
## Maserati Bora   1   Manual
## Volvo 142E      1   Manual
```

```
attach(mtcars)
```

```
##generating descriptive statistics for each group
```

```
#mean
```

```
aggregate(mtcars$mpg,by=list(mtcars$am_label),FUN=mean)
```

```
##      Group.1      x
## 1 Automatic 17.14737
## 2   Manual 24.39231
```

```
#range
```

```
aggregate(mtcars$mpg,by=list(mtcars$am_label),FUN=range)
```

```
##      Group.1  x.1  x.2
## 1 Automatic 10.4 24.4
## 2   Manual 15.0 33.9
```

```
##generating box plot for each group for each group to check the dist  
boxplot(mpg~am_label,main="Distribution of two groups",xlab="am label",ylab="mpg range")  
  
##to check the assumption that the two groups are normally dist  
  
##perform shapiro wilk normality test  
  
aggregate(mtcars$mpg,by=list(mtcars$am_label),FUN=function(x) shapiro.test(x))
```

```
## Warning in format.data.frame(x, digits = digits, na.encode = FALSE):  
## corrupt data frame: columns will be truncated or padded with NAs
```

```
##      Group.1      x  
## 1 Automatic 0.9767743  
## 2   Manual 0.9458037
```

```
##computing the p value for this test  
  
aggregate(mtcars$mpg,by=list(mtcars$am_label),FUN=function(x) shapiro.test(x)$p.value)
```

```
##      Group.1      x  
## 1 Automatic 0.8987358  
## 2   Manual 0.5362729
```

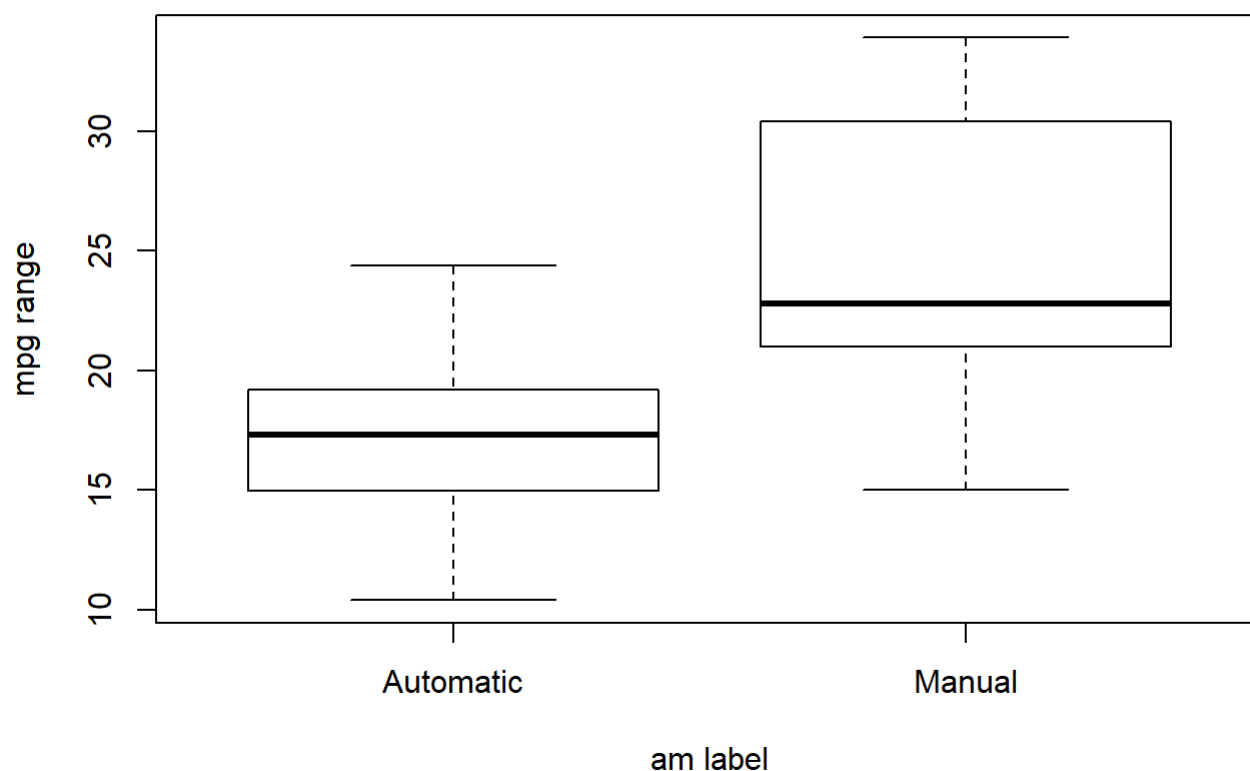
```
##in this test the null hypothesis is the dist is normal..since we see that  
##at 0.05 significance level the value is greater than 0.05 we accept the null  
##hypothesis and hence the two samples are normally dist
```

```
##now we check the assumption of variance
```

```
##performing Levene test
```

```
library(car)
```

## Distribution of two groups



```
leveneTest(mtcars$mpg~mtcars$am_label)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##      Df F value Pr(>F)
## group 1  4.1876 0.04957 *
##      30
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##the null hypothesis is the variance is equal
##since p value is less than 0.05 we accept the alternate hypothesis and hence
##the variance is not equal
```

```
##alternately we also perform var.test that computes f test
```

```
##inorder to stabilize the data variance we should take the log transformation
```

```
mtcars$log_mpg<-log(mpg)
leveneTest(mtcars$log_mpg~mtcars$am_label)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##      Df F value Pr(>F)
## group 1  0.3902 0.5369
##      30
```

*##now the variance is equal*

*##since all the assumptions are satisfied we can perform the t test*

```
t.test(mtcars$log_mpg~mtcars$am_label, var.equal=TRUE)
```

```
##
## Two Sample t-test
##
## data:  mtcars$log_mpg by mtcars$am_label
## t = -3.9087, df = 30, p-value = 0.0004905
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.5277597 -0.1655209
## sample estimates:
## mean in group Automatic    mean in group Manual
##           2.816692           3.163332
```

```
###We see that p-value is less than 0.05 so we reject the null hypothesis
##thus the means of the two samples are statistically different
```

```
#####
```

```
#Paired sample T testing
```

```
#The paired samples t test is used to check if there are any differences in
#the mean of the same sample at two different time points.
#For example a medical researcher collects data on the same patients
#before and after a therapy. A paired t test will show if the therapy
#improves patient outcomes.
```

```
#There are several assumptions that need to be satisfied so that
#results of a paired t test are valid. They are listed below
```

```
#The measured variable is continuous
#The differences between the two groups are approximately normally distributed
#We should not have any outliers in our data
#An adequate sample size is required
```

```
library(MASS)
```

```
head(anorexia)
```

```
##   Treat Prewt Postwt
## 1  Cont  80.7   80.2
## 2  Cont  89.4   80.1
## 3  Cont  91.8   86.4
## 4  Cont  74.0   86.3
## 5  Cont  78.1   76.1
## 6  Cont  88.3   78.1
```

```
library(psych)
```

```
##
## Attaching package: 'psych'
```

```
## The following object is masked from 'package:car':
##
##   logit
```

```
describe(anorexia)
```

```
##      vars  n mean   sd median trimmed  mad  min   max range  skew
## Treat*   1 72  1.83 0.79   2.00   1.79 1.48  1.0   3.0   2.0  0.29
## Prewt    2 72 82.41 5.18  82.30   82.47 5.49 70.0  94.9  24.9 -0.05
## Postwt   3 72 85.17 8.04  84.05   84.82 9.56 71.3 103.6  32.3  0.36
##      kurtosis   se
## Treat*    -1.35 0.09
## Prewt     -0.16 0.61
## Postwt    -0.81 0.95
```

```
##computing mean using apply function
```

```
apply(anorexia[,c(2,3)],2,mean)
```

```
##      Prewt   Postwt
## 82.40833 85.17222
```

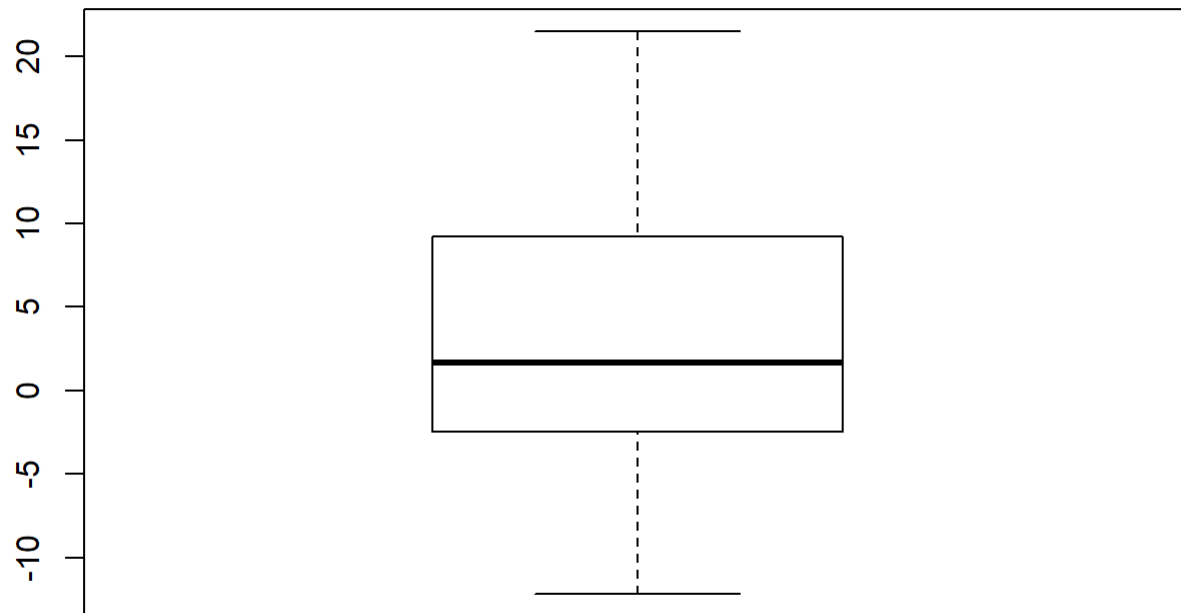
```
##creating a variable to check the diff
```

```
anorexia$diff=anorexia$Postwt-anorexia$Prewt
```

```
##creating a box plot
```

```
boxplot(anorexia$diff,xlab="The diff")
```



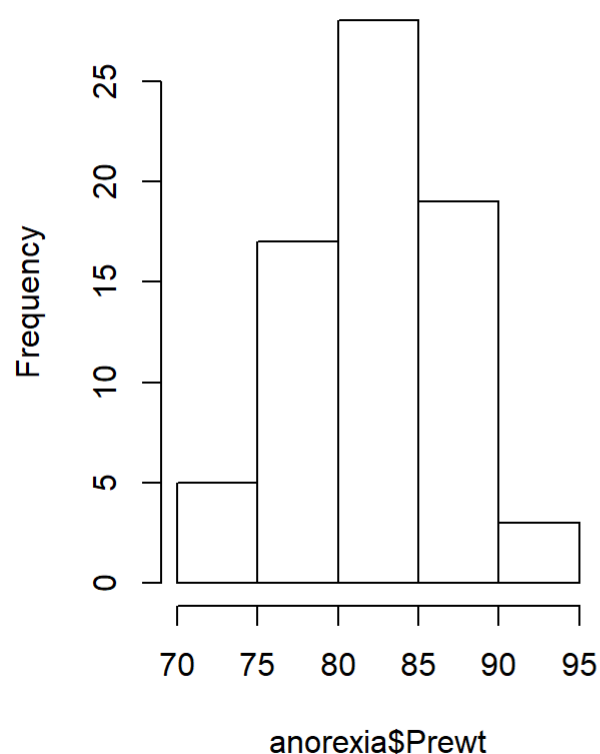
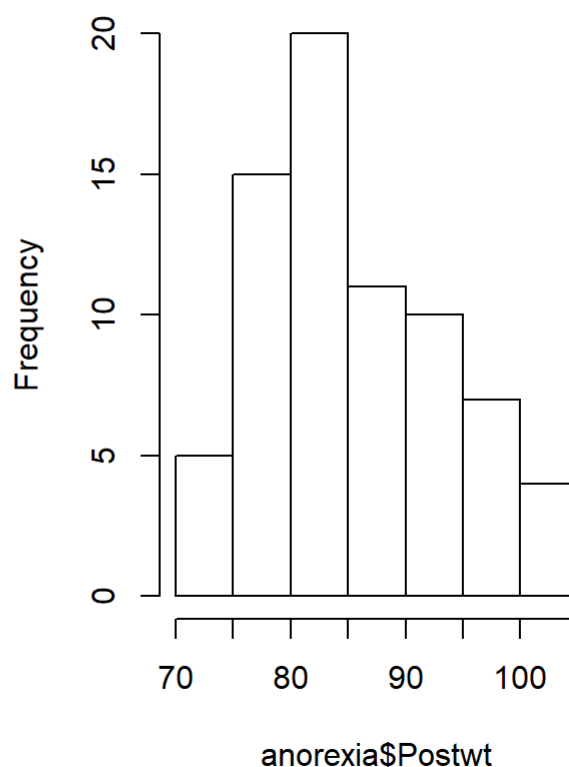


The diff

```
##checking the assumption to see if both the samples are normally dist  
##visually
```

```
par(mfrow=c(1,2))
```

```
hist(anorexia$Prewt)  
hist(anorexia$Postwt)
```

**Histogram of anorexia\$Prewt****Histogram of anorexia\$Postwt**

*#pre treatment is normally dist but post treatment is not normally dist*

*##checking through normality test*

```
shapiro.test(anorexia$Prewt)
```

```
##
## Shapiro-Wilk normality test
##
## data:  anorexia$Prewt
## W = 0.99248, p-value = 0.9484
```

```
shapiro.test(anorexia$Postwt)
```

```
##
## Shapiro-Wilk normality test
##
## data:  anorexia$Postwt
## W = 0.9673, p-value = 0.05781
```

```
shapiro.test(anorexia$diff)
```

```
##
## Shapiro-Wilk normality test
##
## data: anorexia$diff
## W = 0.97466, p-value = 0.1544
```

```
#the diff is normally distributed
```

```
#Perform a power analysis to check the sample size has adequate power to detect a difference if it exists
#install package pwr and load it
library(pwr)
```

```
## Warning: package 'pwr' was built under R version 3.4.4
```

```
pwr.t.test(n=72,d=0.5,sig.level = 0.05,type = c("paired"))
```

```
##
##      Paired t test power calculation
##
##              n = 72
##              d = 0.5
##      sig.level = 0.05
##      power = 0.9869471
##      alternative = two.sided
##
## NOTE: n is number of *pairs*
```

```
#we see the power of the test is 98% hence the sample size is appropriate
```

```
var.test(anorexia$Prewt,anorexia$Postwt)
```

```
##
## F test to compare two variances
##
## data: anorexia$Prewt and anorexia$Postwt
## F = 0.41599, num df = 71, denom df = 71, p-value = 0.000288
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.2602635 0.6648930
## sample estimates:
## ratio of variances
##      0.4159896
```

```
#performing paired t test
```

```
t.test(anorexia$Prewt,anorexia$Postwt,paired = TRUE)
```

```
##  
## Paired t-test  
##  
## data: anorexia$Prewt and anorexia$Postwt  
## t = -2.9376, df = 71, p-value = 0.004458  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -4.6399424 -0.8878354  
## sample estimates:  
## mean of the differences  
## -2.763889
```

```
##the two sample means are statistically different
```

```
#All assumptions required were satisfied
```

```
#There were no outliers, data was normally distributed and the t test had adequate power
```

```
#The difference in weight before and after treatment was statistically significant at 5% LOs.
```

```
###chisqaure test of independence
```

```
data("trees")  
head(trees)
```

```
## Girth Height Volume  
## 1 8.3 70 10.3  
## 2 8.6 65 10.3  
## 3 8.8 63 10.2  
## 4 10.5 72 16.4  
## 5 10.7 81 18.8  
## 6 10.8 83 19.7
```

```
##creating a two way frequency table
```

```
mytable<-table(trees$Height,trees$Volume)  
mytable
```

```
##
##      10.2 10.3 15.6 16.4 18.2 18.8 19.1 19.7 19.9 21 21.3 21.4 22.2 22.6
## 63    1    0    0    0    0    0    0    0    0    0    0    0    0
## 64    0    0    0    0    0    0    0    0    0    0    0    0    0
## 65    0    1    0    0    0    0    0    0    0    0    0    0    0
## 66    0    0    1    0    0    0    0    0    0    0    0    0    0
## 69    0    0    0    0    0    0    0    0    0    0    1    0    0
## 70    0    1    0    0    0    0    0    0    0    0    0    0    0
## 71    0    0    0    0    0    0    0    0    0    0    0    0    0
## 72    0    0    0    1    0    0    0    0    0    0    0    0    0
## 74    0    0    0    0    0    0    0    0    0    0    0    1    0
## 75    0    0    0    0    1    0    1    0    1    0    0    0    0
## 76    0    0    0    0    0    0    0    0    0    1    0    1    0
## 77    0    0    0    0    0    0    0    0    0    0    0    0    0
## 78    0    0    0    0    0    0    0    0    0    0    0    0    0
## 79    0    0    0    0    0    0    0    0    0    0    0    0    0
## 80    0    0    0    0    0    0    0    0    0    0    0    0    1
## 81    0    0    0    0    0    1    0    0    0    0    0    0    0
## 82    0    0    0    0    0    0    0    0    0    0    0    0    0
## 83    0    0    0    0    0    0    0    1    0    0    0    0    0
## 85    0    0    0    0    0    0    0    0    0    0    0    0    0
## 86    0    0    0    0    0    0    0    0    0    0    0    0    0
## 87    0    0    0    0    0    0    0    0    0    0    0    0    0
##
##      24.2 24.9 25.7 27.4 31.7 33.8 34.5 36.3 38.3 42.6 51 51.5 55.4 55.7
## 63    0    0    0    0    0    0    0    0    0    0    0    0    0
## 64    0    1    0    0    0    0    0    0    0    0    0    0    0
## 65    0    0    0    0    0    0    0    0    0    0    0    0    0
## 66    0    0    0    0    0    0    0    0    0    0    0    0    0
## 69    0    0    0    0    0    0    0    0    0    0    0    0    0
## 70    0    0    0    0    0    0    0    0    0    0    0    0    0
## 71    0    0    1    0    0    0    0    0    0    0    0    0    0
## 72    0    0    0    0    0    0    0    0    1    0    0    0    0
## 74    0    0    0    0    0    0    0    1    0    0    0    0    0
## 75    0    0    0    0    0    0    0    0    0    0    0    0    0
## 76    0    0    0    0    0    0    0    0    0    0    0    0    0
## 77    0    0    0    0    0    0    0    0    0    1    0    0    0
## 78    0    0    0    0    0    0    1    0    0    0    0    0    0
## 79    1    0    0    0    0    0    0    0    0    0    0    0    0
## 80    0    0    0    0    1    0    0    0    0    0    1    1    0
## 81    0    0    0    0    0    0    0    0    0    0    0    0    1
## 82    0    0    0    0    0    0    0    0    0    0    0    0    1
## 83    0    0    0    0    0    0    0    0    0    0    0    0    0
## 85    0    0    0    0    0    1    0    0    0    0    0    0    0
## 86    0    0    0    1    0    0    0    0    0    0    0    0    0
## 87    0    0    0    0    0    0    0    0    0    0    0    0    0
##
##      58.3 77
## 63    0    0
## 64    0    0
## 65    0    0
## 66    0    0
## 69    0    0
```

```
## 70 0 0
## 71 0 0
## 72 0 0
## 74 0 0
## 75 0 0
## 76 0 0
## 77 0 0
## 78 0 0
## 79 0 0
## 80 1 0
## 81 0 0
## 82 0 0
## 83 0 0
## 85 0 0
## 86 0 0
## 87 0 1
```

```
##generating individual freq using margin table
```

```
margin.table(mytable,1)
```

```
##
## 63 64 65 66 69 70 71 72 74 75 76 77 78 79 80 81 82 83 85 86 87
## 1 1 1 1 1 1 1 2 2 3 2 1 1 1 5 2 1 1 1 1 1
```

```
margin.table(mytable,2)
```

```
##
## 10.2 10.3 15.6 16.4 18.2 18.8 19.1 19.7 19.9 21 21.3 21.4 22.2 22.6 24.2
## 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1
## 24.9 25.7 27.4 31.7 33.8 34.5 36.3 38.3 42.6 51 51.5 55.4 55.7 58.3 77
## 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
```

```
#creating a new col called a
```

```
a=c(70, 65, 63, 72, 80, 83, 66, 75, 80, 75, 79, 76, 76, 69, 75, 74, 85, 8, 71,
    63, 78, 80, 74, 72, 77, 81, 82, 80, 86, 80, 87)
```

```
trees=cbind(trees,a)
length(trees$a)
```

```
## [1] 31
```

```
mytable2<-table(trees$Height,trees$a)
```

```
chisq.test(mytable)
```

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

```
##  
## Pearson's Chi-squared test  
##  
## data: mytable  
## X-squared = 589, df = 580, p-value = 0.3888
```

*##we observe that the p value is greater than 0.05 and hence we accept the null hypothesis that the columns are independent of each other*

```
chisq.test(mytable2)
```

```
## Warning in chisq.test(mytable2): Chi-squared approximation may be incorrect
```

```
##  
## Pearson's Chi-squared test  
##  
## data: mytable2  
## X-squared = 571.64, df = 400, p-value = 3.495e-08
```

*#we observe that the p value is much lesser than 0.05 and hence we reject the null hypothesis*