Question 1:

a) 1)Classical statistical test:
Using t-test to do the correction testing, according the Spearman correlation coefficient between Q3 and CC_AVG, we can calculate Sr, then according tdist function, we can calculate p-value, if p-value is less than 0.05, it means Q3 is not independent with CC_AVG
2)Permutation test:
Randomly shuffle the ranks of CC_AVG, compute Spearman Rank Correlation again with the new rank of CC_AVG and Q3.
Repeat previous 1000 trials,
If absolute Spearman Rank Correlation in 95% of the trials is less than the absolute original Spearman Rank Correlation. Then Q3 has a significant correlation with CC_AVG.

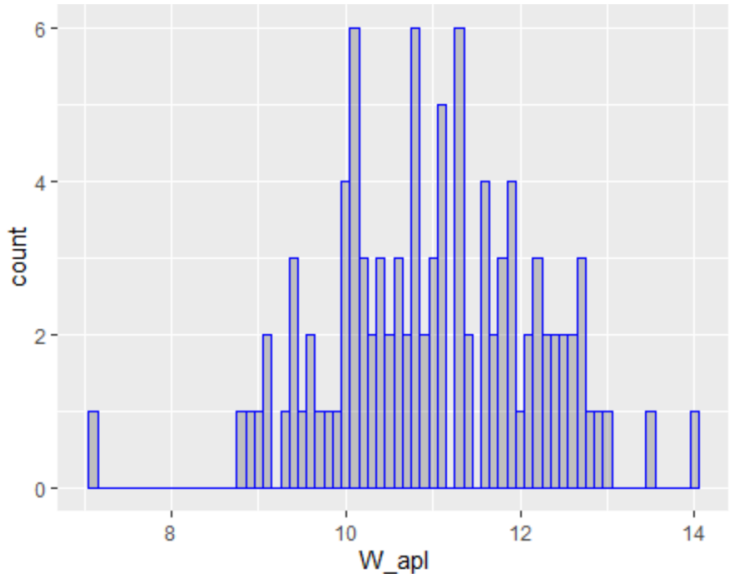
b) 1)Classical statistical test:
Assumption: Q3 and CC_AVG has a bivariate normal distribution.
2)Permutation test:
No assumption

c) Classical statistical test and permutation test:
Null hypothesis: Q3 is independent with CC_AVG

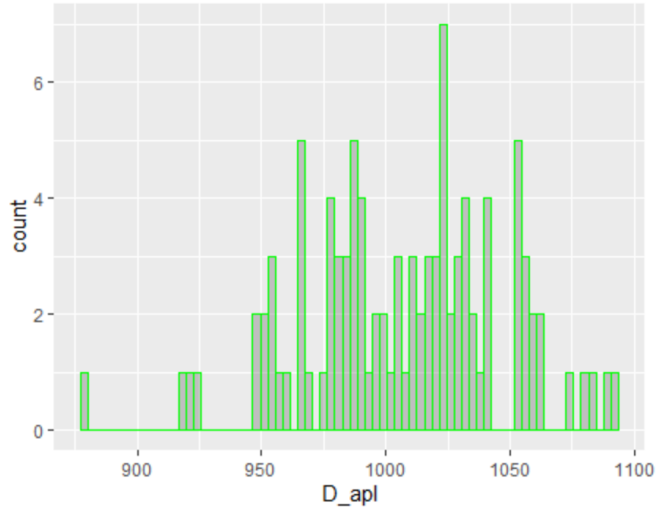
d) Classical statistical test:
Conclusion: Q3 is not independent with CC_AVG, and because the estimation of correlation between Q3 and CC_AVG is 0.8907, so Q3 has a “significant” correlation with CC_AVG
Permutation test:
Conclusion: Because in 100% trials Spearman Rank Correlation less than the original Spearman Rank Correlation, therefore, null hypothesis has been rejected, Q3 has a “significant” correlation with CC_AVG.

Question 2:

a) Apple’s weight feature and diameter feature are skewed.
I test them by compare the mean and median of these two features, if mean more than median and tails on the right, then the feature is right skewed, otherwise it is left skewed.
Apple’s weight is left skewed, its tail is on the left.



diameter is also left skewed, its tail is also on the left.



b) First calculate vectors' interquartile as IQR, 25% quartile as Q1, 75% quartile as Q3, and if there are any values is not in the range [Q1-IQR, Q3+IQR], then it will be defined as outliers.
None of the fruit weight vectors contains outliers.

c) For D_apl,
Min: 877.8837
Max: 1091.7998
Range: 213.9161
Inter-quartile: 50.07521

Question 3:

a) Null hypothesis:
There is no association between a person's propensity to leave the house and the degree to which they enjoyed the movie.

Expected contingency table according to the null hypothesis:

	0-5 stars	6-8 stars	9-10 stars	total
Never	3.36	5.88	4.76	14
Unlikely	2.88	5.04	4.08	12
Sometimes	2.64	4.62	3.74	11
Most weekends	3.12	5.46	4.42	13
total	12	21	17	50

Chi-Square is 8. Degrees of freedom is 6.
P-value is 0.26, so we cannot refuse the null hypothesis, so there is no association between a person's propensity to leave the house and the degree to which they enjoyed the movie.

b) i) Unsupervised method for feature selection:
Filter method:
LDA: Linear discriminant analysis is used to find a linear combination of features that characterizes or separate two or more classes of a categorical variable.

Supervised method for feature selection:
Wrapper method:
Forward selection: we start with have no feature in the model, in each iteration, we keep adding the features which can best improve our model.

ii)Wrapper method may lead to overfitting, because in each iteration it chooses the feature that has the highest correlation with dependent variable. This may cause that the features with high correlations been selected at the same time as the features has high correlation with the selected features can also have high correlation with the dependent variable.

iii) Using ten-fold cross-validation method

1)split the data into 10-fold.

2)choose 9 of the 10-fold data as training data and left one used as test data

3)Using filter method to do the feature selection, choose the feature which has less correlation

4)build the model

5)calculate the MSE and Adjusted R-squared error

6)do step 2 to 5 again, but choose another fold of data as test data until all the data been used as test data

7)calculate the mean of MSE and Adjusted R-squared error

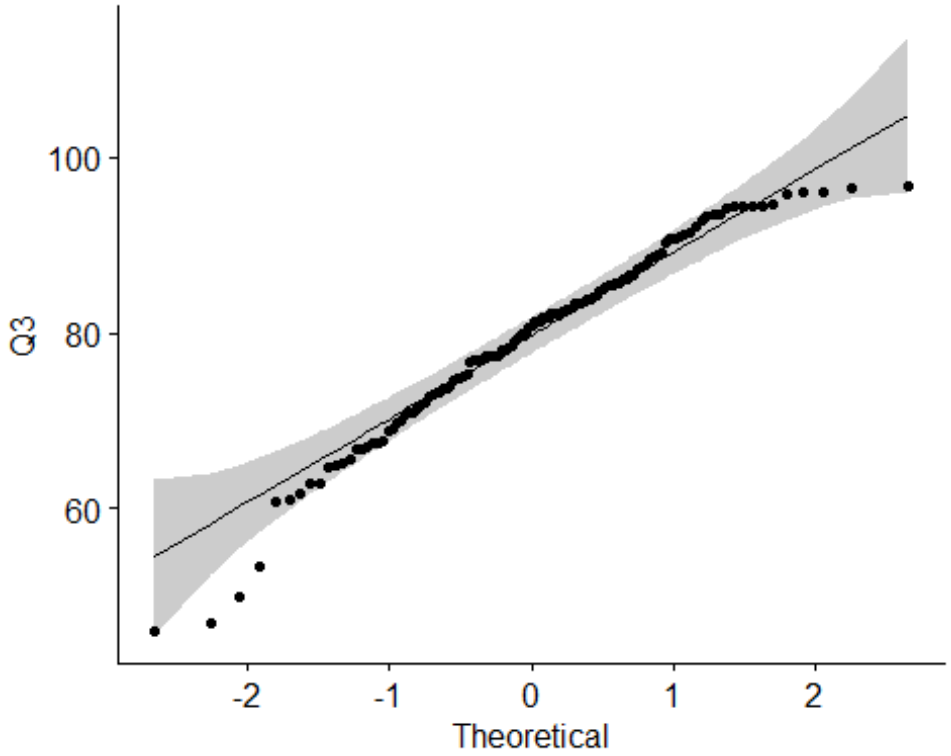
R code:

assignment2q1.R

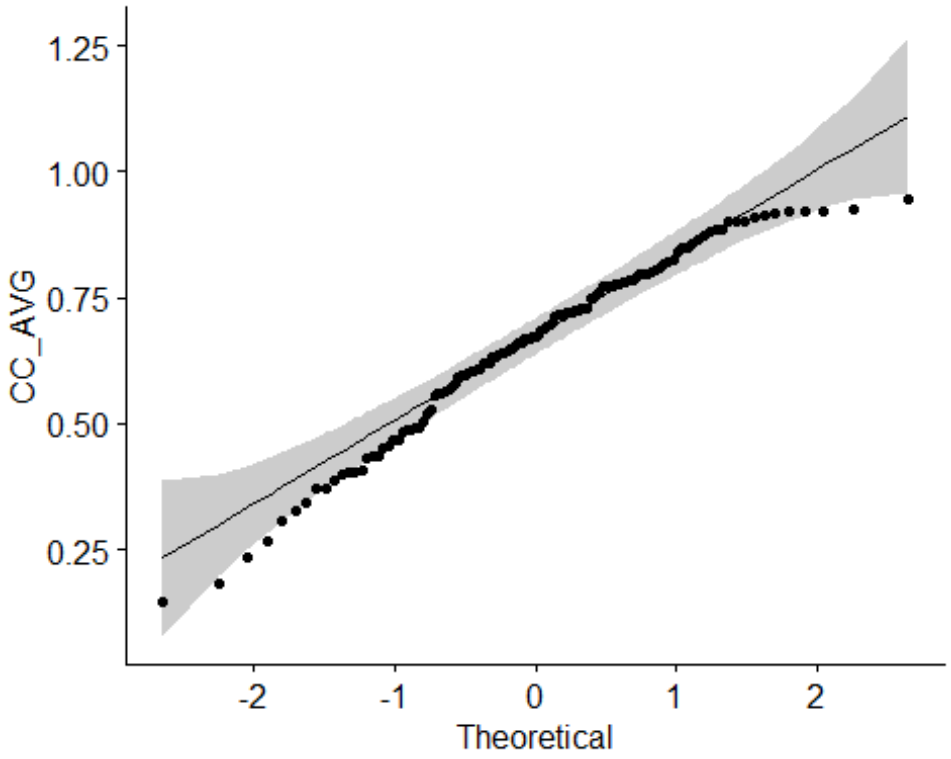
Yiyin Zhang

2019-10-05

```
library(readxl)
library(ggpubr)

## Loading required package: ggplot2

## Loading required package: magrittr

Psipred_data <- read_excel("assigData1.xls", sheet = 2)

## New names:
## * `` -> ...1

Q3 <- as.numeric(unlist(Psipred_data[,4]))
CC_AVG <- as.numeric(unlist(Psipred_data[,3]))
#Classical statistical test
n <- length(Q3)
ggqqplot(Q3, ylab = "Q3") #see if Q3 is close to normal distribution
```



```r
ggqqplot(CC_AVG, ylab = "CC_AVG") #see if CC_AVG is close to normal distribut
ion
```




```r
r <- cor(Q3, CC_AVG, method = "spearman")

print(r)

## [1] 0.8907183
Sr <- sqrt((1-r^2)/(n-2))
t <- r/Sr
p_value <- dt(abs(t),n-2)
print(p_value)

## [1] 1.383033e-43

#Permutation test
PT <- vector()
sum <- 0
for (i in 1:1000)
{
  CC_NEW <- sample(CC_AVG, length(CC_AVG))
  PT <- c(PT, cor(Q3, CC_NEW, method = "spearman"))
  if (PT[i] < r){
    sum  = sum + 1
  }
}
if (sum/1000 > 0.95){
  print("significant correlated")
```

```
  print(sum)
} else {
  print("not correlated")
  print(sum)
}
```

```
## [1] "significant correlated"
## [1] 1000
```

assignment2q2.R

Yiyin Zhang

2019-10-05

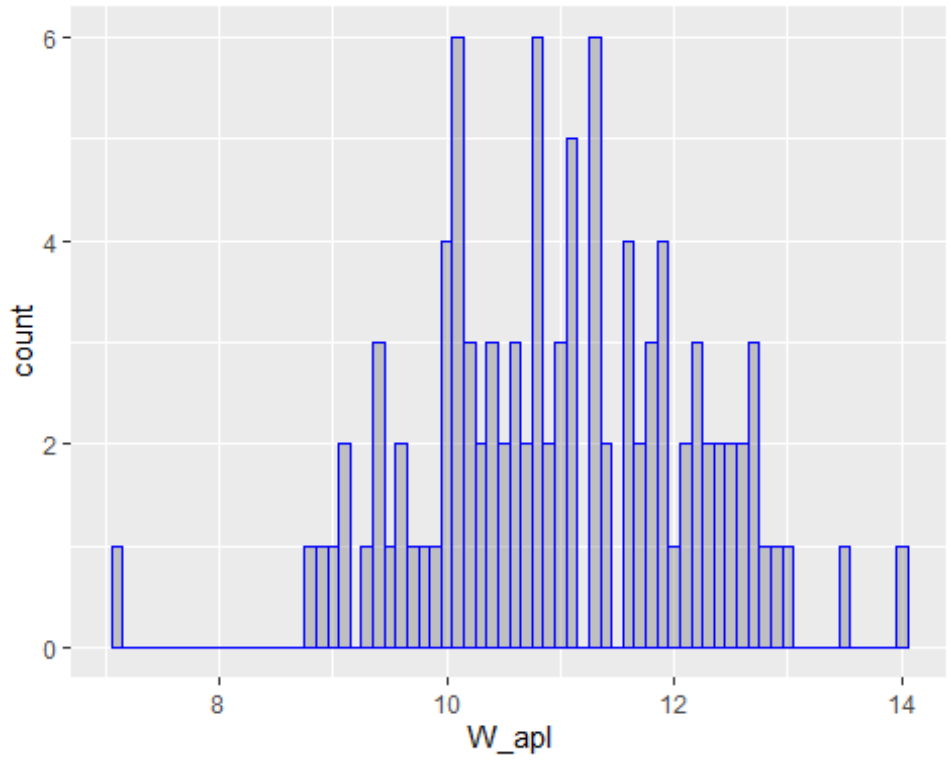
```
library(ggplot2)

fruit_data <- read.table(file = "assigData2.tsv")
#Question 2.a

W_apl <- fruit_data[,1]
W_orng <- fruit_data[,2]
W_grp <- fruit_data[,3]
D_apl <- fruit_data[,4]

skew_test <- function(A,B){
  if(mean(A) > median(A)){
    cat(B,"is right skewed")
  } else if(mean(A) < median(A)){
    cat(B,"is left skewed")
  } else {
    print("it is symmertical")
  }
}
skew_test(W_apl, "Apple's weight")
```

```
## Apple's weight is left skewed
```

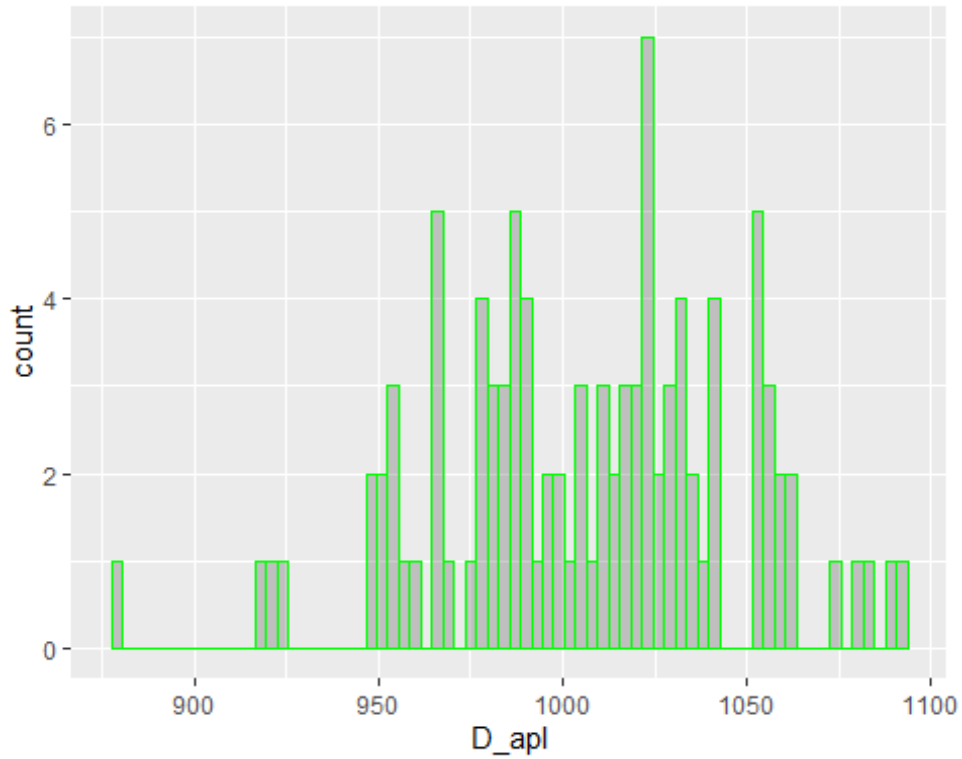
```
skew_test(D_apl,"Apple's diameter")
```

```
## Apple's diameter is left skewed
```

```
a = data.frame(W_apl,D_apl)
W <- ggplot(a)+
  geom_histogram(aes(x = W_apl),col = "blue", alpha = 0.3,binwidth = 0.1)
print(W)
```




```
D <- ggplot(a)+
  geom_histogram(aes(x = D_apl),col = "green", alpha = 0.3,binwidth = 3)
print(D)
```



```r
#Question 2.b
outlier_test <- function(A,B){
  test <- 0
  Q <- quantile(A, c(0.25, 0.75))
  IQR <- Q[2] - Q[1]
  for(i in length(A)){
    if(A[i] > Q[2] + IQR|| A[i] < Q[1]-IQR){
      test <- 1
      A[i] <- 0
    }
  }
  if(test == 1)
  {
    cat(B,"contains outliers")
  } else {
    cat(B, "does not have outliers")
  }
  A <- A[A!=0]
  return(A)
}

NEW_Wapl <- outlier_test(W_apl,"Apple's weight")

## Apple's weight does not have outliers

NEW_Worng <- outlier_test(W_orng,"orange's weight")

## orange's weight does not have outliers

NEW_Wgrp <- outlier_test(W_grp,"grape's weight")

## grape's weight does not have outliers

#Question 2.c
min(D_apl)

## [1] 877.8837

max(D_apl)

## [1] 1091.8

range(D_apl)

## [1]  877.8837 1091.7998

Q1 <- quantile(D_apl, c(0.25, 0.75))
IQR <- Q1[2]-Q1[1]
IQR

##      75%
## 50.07521
```

assignment2q3.R

Yiyin Zhang

2019-10-01

```
library(readxl)

#question 3.a
My_data <- read_excel("assigData4.xlsx")

## New names:
## * `` -> ...1

fix(My_data)

Frame_data <- data.frame(My_data)

as.matrix(Frame_data)

##   ...1            X0.5.Stars X6.8.Stars X9.10.Stars
## 1 "Never"         "3"        "4"        "7"
## 2 "Unlikely"      "2"        "6"        "4"
## 3 "Sometimes"     "1"        "6"        "4"
## 4 "Most weekends" "6"        "5"        "2"

chisq.test(Frame_data$X0.5,Frame_data$X6.8,
           Frame_data$X9.10)

## Warning in chisq.test(Frame_data$X0.5, Frame_data$X6.8, Frame_data$X9.10):
## Chi-squared approximation may be incorrect

##
##  Pearson's Chi-squared test
##
## data:  Frame_data$X0.5 and Frame_data$X6.8
## X-squared = 8, df = 6, p-value = 0.2381
```