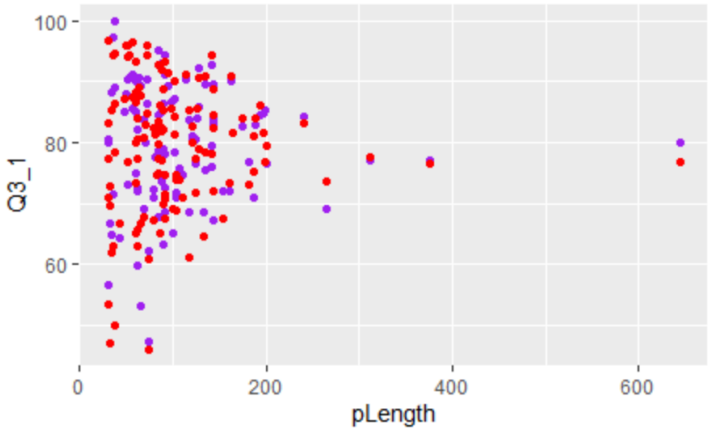
Question 1:

a) Red scatter plot is the Q3 accuracy vs. protein length in PCI,
Purple scatter plot is the Q3 accuracy vs. protein length in PSIPRED



b) I choose correlation coefficient to describe the correlation between Q3 accuracy and protein length for PCI and for PSIPRED

$$\rho_{X,Y} = \frac{E(XY) - E(X)E(Y)}{\sqrt{E(X^2) - E^2(X)}\,\sqrt{E(Y^2) - E^2(Y)}}.$$

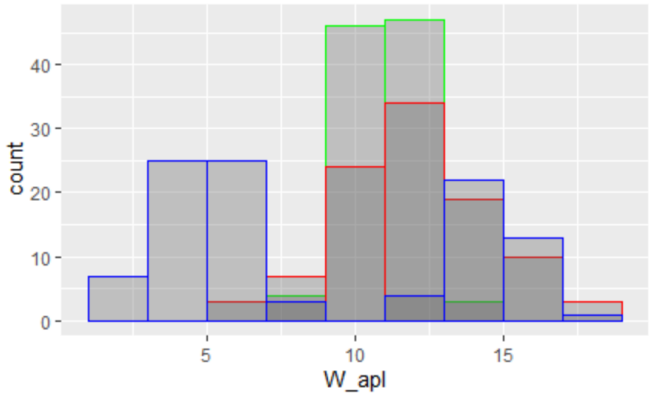
According to the formula:
Correlation coefficient in PCI is 0.02814
Correlation coefficient in PSIPRED is 0.01079
Because the correlation coefficients between Q3 accuracy and protein length for both PCI and PSIPRED are close to 0, which means little correlation of the objects.

c) 1)PCI Matthews' correlation coefficient:
Mean: 0.65604
Median: 0.669
Standard deviation: 0.17787
2)PSIPRED Matthews' correlation coefficient:
Mean: 0.65797
Median: 0.673
Standard deviation: 0.17572

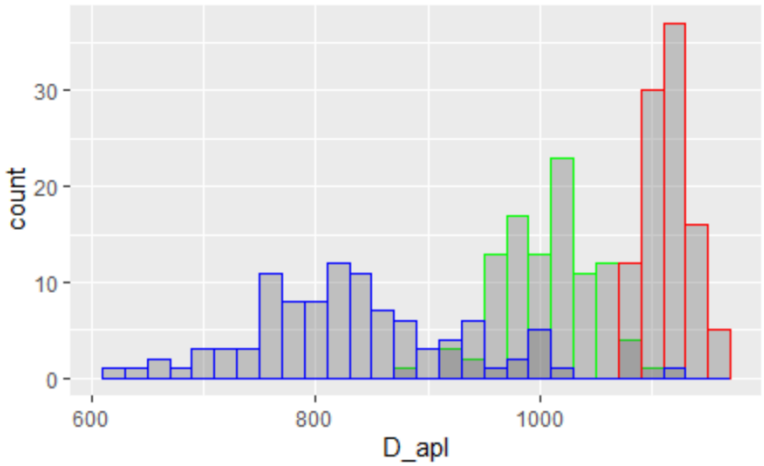
Question 2:

a) According to the maximum likelihood conception, if each class feather follows normal distribution, then the optimal selection of each class feather for their mean and variance would be their sample mean and variance.
   Estimates of mean:
   W_apl = 11.003
   W_orng = 11.945
   W_grp = 8.733
   D_apl = 1006.707
   D_orng = 1114.834
   D_grp = 832.546
   Estimates of variance:
   W_apl = 1.401
   W_orng = 6.807
   W_grp = 24.787
   D_apl = 1621.37
   D_orng = 8356.382
   D_grp = 383.406

b) The histogram for fruits weights is shown below:

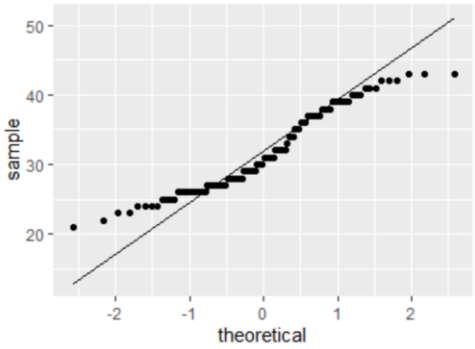


   The histogram for fruits diameters is shown below:



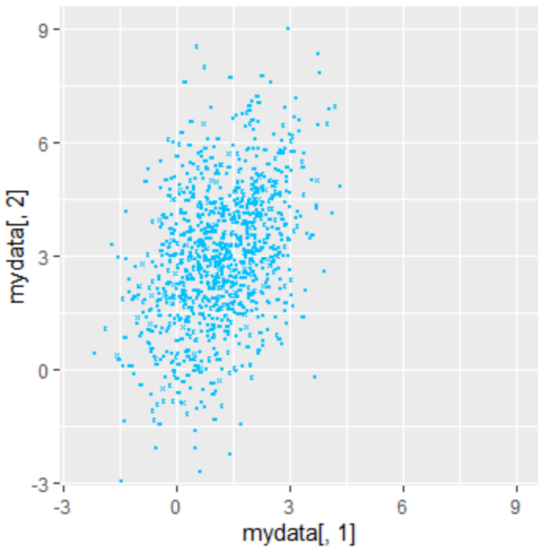
I prefer the diameter feather, because each different kind of fruit has quite different range of diameter, but not different range of weight.

c) Here is the normal QQ plot for all weight data. If the weight data is normally distributed, the scatter points will close to the straight line.
But in the figure below, we can see that, the distribution of scatter points is totally not consistent with the line.
Therefore, all weight data is not normally distributed.



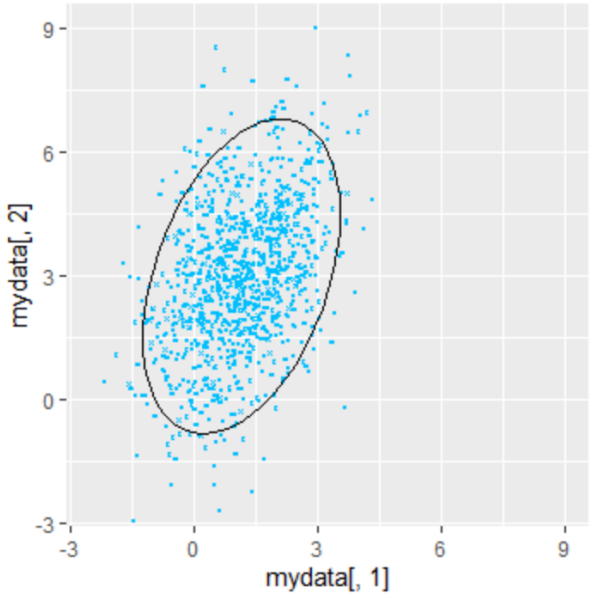
Question 3:

a) See the code
b)

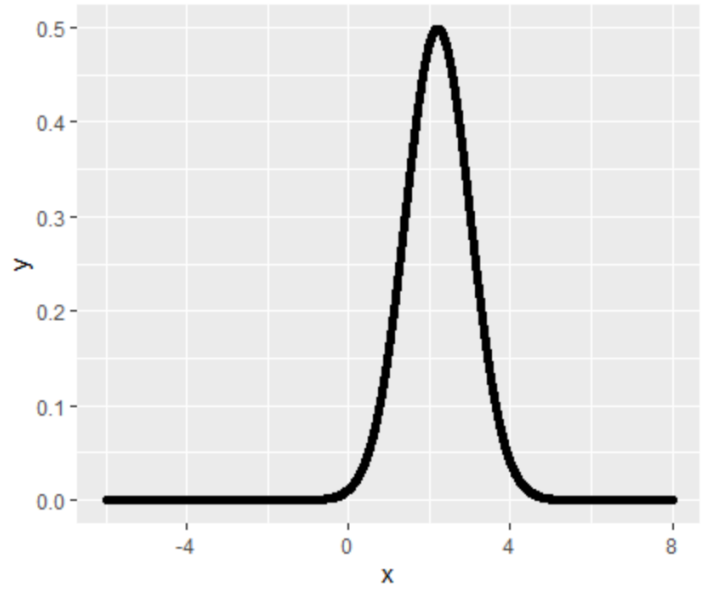


c) The determinant is 3.47, trace is 4.5, it is positive definite.

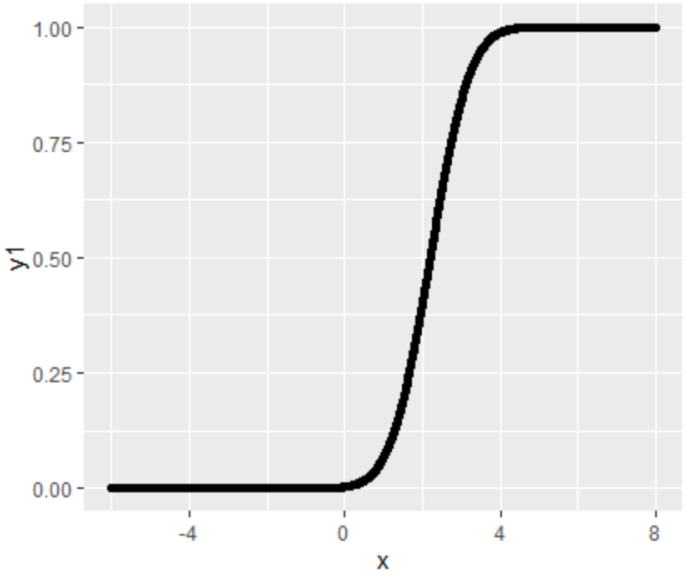
d) Eigenvalue is 3.512 and 0.988, eigenvectors are [0.2898, 0.9571] and [-0.9571, 0.2898].



e) PDF picture for 1D normal distribution.



CDF picture for 1D normal distribution.



Appendix:

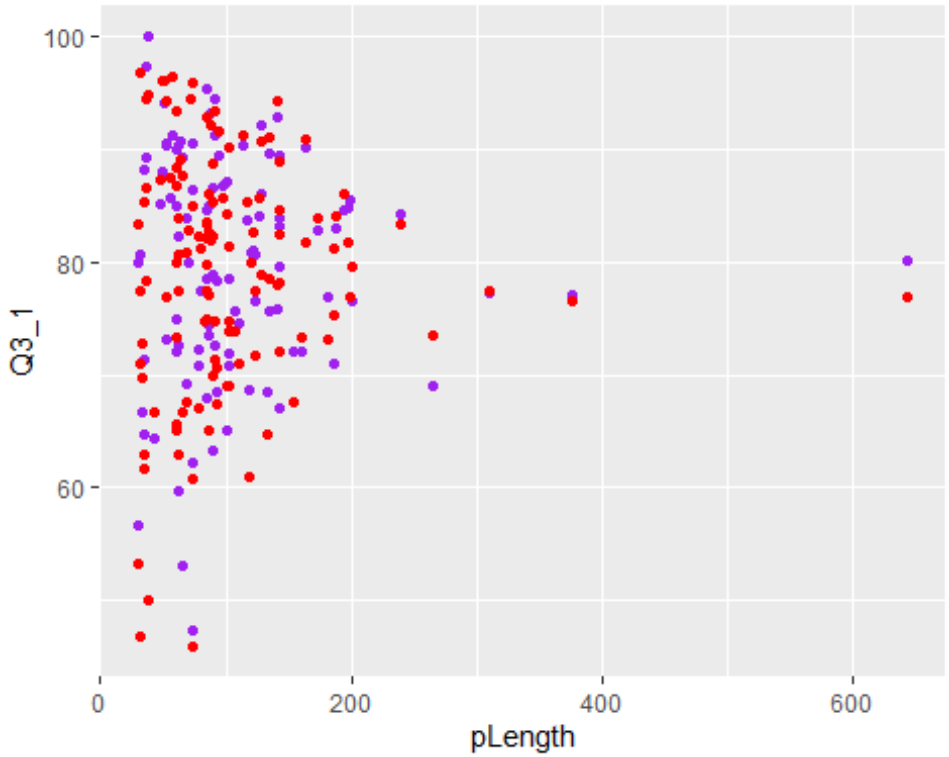
Question1.R

Yiyin Zhang

2019-09-25

```
library(readxl)
library(ggplot2)
```

```r
# read data
Pci_Data <- read_excel("assigData1.xls", sheet = 1)

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

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

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

#question 1.a
Q3_1 <- as.numeric(unlist(Pci_Data[,4]))
Q3_2 <- as.numeric(unlist(Psipred_Data[,4]))
pLength <- as.numeric(unlist(Pci_Data[,2]))

df1 <- data.frame(pLength, Q3_1, Q3_2)
p <- ggplot(df1) +
  geom_point(df1, mapping = aes(x = pLength, y = Q3_1), color = "purple") +
  geom_point(df1, mapping = aes(x = pLength, y = Q3_2), color = "red")
print(p)
```




```r
#question 1.b
#PCI Q3 and Length corelation
x1_mean <- mean(Q3_1)
y1_mean <- mean(pLength)
Mole1_x_y <- sum((Q3_1 - x1_mean)*(pLength - y1_mean))
Deno1_x_y <- sqrt(sum((Q3_1 - x1_mean)^2)*sum((pLength - y1_mean)^2))
```

```r
cor1_x_y <- Mole1_x_y/Deno1_x_y
#PSIPRED Q3 and Length corelation
x2_mean <- mean(Q3_2)
Mole2_x_y <- sum((Q3_2 - x2_mean)*(pLength - y1_mean))
Deno2_x_y <- sqrt(sum((Q3_2 - x2_mean)^2)*sum((pLength - y1_mean)^2))
cor2_x_y <- Mole2_x_y/Deno2_x_y

#question 1.c
#PCI CC
CC1 <- as.numeric(unlist(Pci_Data[,3]))
CC1.mean = mean(CC1)
CC1.median = median(CC1)
CC1.sd = sd(CC1)
#PSIPRED CC
CC2 <- as.numeric(unlist(Psipred_Data[,3]))
CC2.mean = mean(CC2)
CC2.median = median(CC2)
CC2.sd = sd(CC2)
```

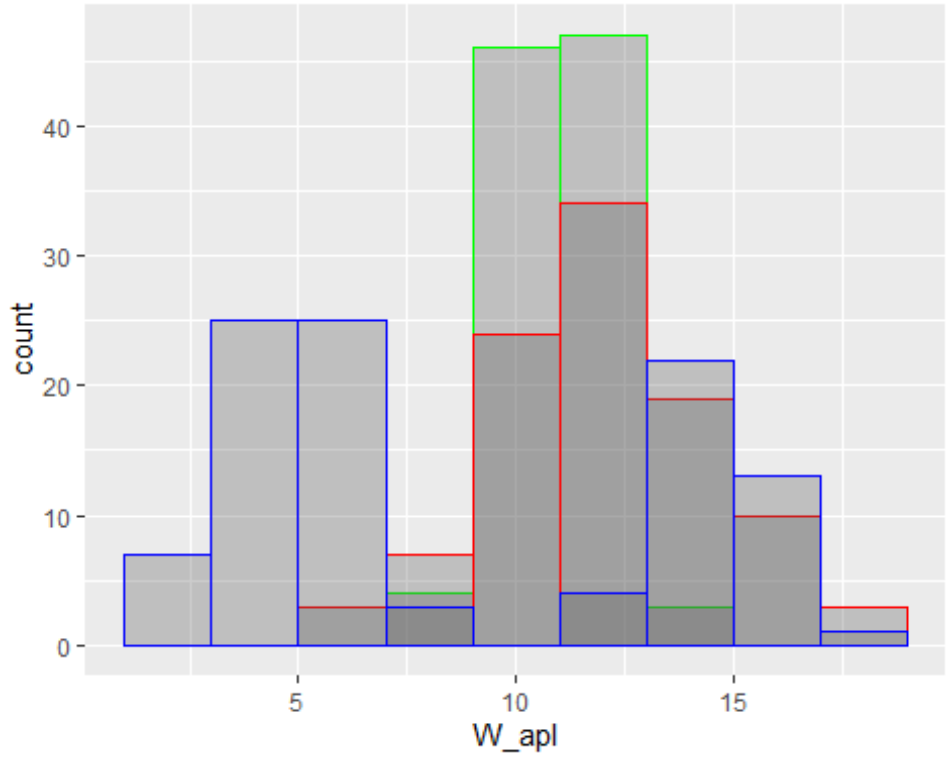
Question2.R

Yiyin Zhang

2019-09-25

```r
library(ggplot2)
library(grid)

#Question 2.a
fruit_data <- read.table(file = "assigData2.tsv")
W_apl <- fruit_data[,1]
W_orng <- fruit_data[,2]
W_grp <- fruit_data[,3]
D_apl <- fruit_data[,4]
D_orng <- fruit_data[,5]
D_grp <- fruit_data[,6]

W_apl.mean <- mean(W_apl)
W_orng.mean <- mean(W_orng)
W_grp.mean <- mean(W_grp)
D_apl.mean <- mean(D_apl)
D_orng.mean <- mean(D_orng)
D_grp.mean <- mean(D_grp)

W_apl.var <- var(W_apl)
W_orng.var <- var(W_orng)
W_grp.var <- var(W_grp)
D_apl.var <- var(D_apl)
```

```r
D_orng.var <- var(D_orng)
D_grp.var <- var(D_grp)

#question 2.b
df2 = data.frame(W_apl, W_orng, W_grp)
W <- ggplot(df2) +
  geom_histogram(aes(x = W_apl), col = "green", alpha = 0.3,binwidth = 2)+
  geom_histogram(aes(x = W_orng), col = "red",alpha = 0.3,binwidth = 2)+
  geom_histogram(aes(x = W_grp), col = "blue",alpha = 0.3,binwidth = 2)
print(W)
```



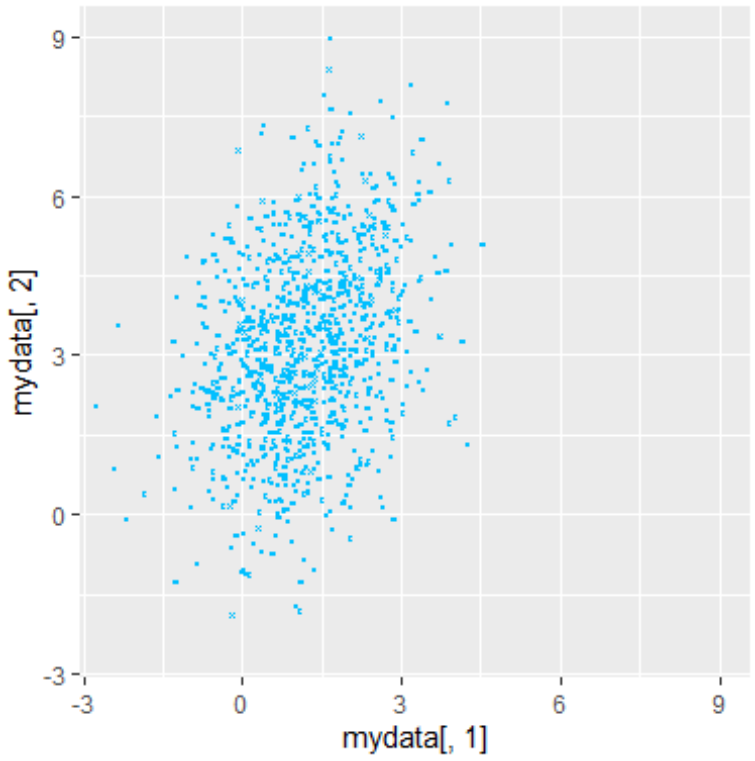

```r
df3 = data.frame(D_apl, D_orng, D_grp)
D <- ggplot(df2) +
  geom_histogram(aes(x = D_apl), col = "green", alpha = 0.3,binwidth = 20)+
  geom_histogram(aes(x = D_orng), col = "red",alpha = 0.3,binwidth = 20)+
  geom_histogram(aes(x = D_grp), col = "blue",alpha = 0.3,binwidth = 20)
print(D)
```

Question3.R

Yiyin Zhang

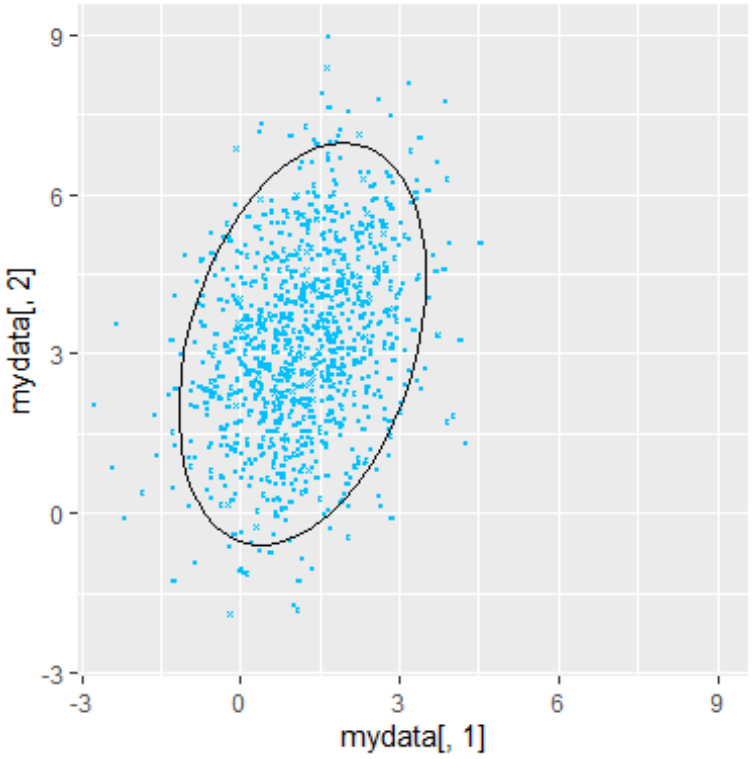
2019-09-25

```
library(MASS)
library(ggplot2)

#question 3.a
mean <- c(1.2, 3.1)
sigma <- matrix(c(1.2, 0.7, 0.7, 3.3), nrow = 2, ncol = 2)
mydata <- mvrnorm(1000, mean, sigma)
#question 3.b
df <- data.frame(mydata[,1], mydata[,2])
p <- ggplot(df, aes(mydata[,1], mydata[,2])) + geom_point(color = "#00bfff",
shape = 4, size = 0.1) +
  coord_fixed(ratio = 1, xlim = c(-2.5,9), ylim = c(-2.5,9))
print(p)
```




```
#question 3.c
detM <- det(sigma)
trace <- sum(diag(sigma))
#question 3.d
V_Lamda <- eigen(sigma)
eigenvalues <- V_Lamda$values
```

```r
eigenvectors <- V_Lamda$vectors
p + stat_ellipse(level = 0.95)
```




```r
#question 3.e
x <- seq(-6, 8, length = 1000)
y <- dnorm(x, 2.2, 0.8)
qplot(x,y)
```