Assignment 1 (Deadline: Sept. 27)

Question One

Suppose you track your commute times for two weeks (10 days) and you find the following times in minutes:

17 16 20 24 22 15 21 15 17 22

1. Enter this into R. Use the function **max** to find the longest commute time, the function **mean** to find the average and the function **min** to find the minimum.
2. If the 24 was a mistake. It should have been 18. How can you fix this? Do so, and then find the new average.
3. How many times was your commute 20 minutes or more? Answer this with R.
4. What percent of your commutes are less than 17 minutes? How can you answer this with R?

Answer

```r
# part one
comm_time <- c(17,16,20,24,22,15,21,15,17,22) #input the data
max(comm_time) # get the maximum element of the sequence = 24
mean(comm_time) # get the sequence's mean = 18.9
min(comm_time) # get the minimum element of the sequence = 15

# part two
comm_time[which(comm_time == 24)] <- 18
print(comm_time) # the result is  [1] 17 16 20 18 22 15 21 15 17 22
# This first command can disintegrate into two commands:
# substitute <- which(comm_time == 24)
# comm_time[substitute] <- 18

# part three
comm_time <- c(17,16,20,24,22,15,21,15,17,22)
comm_time[which(comm_time >= 20)] # the result is: [1] 20 24 22 21 22
# first loop is get the serial number of those who are larger than 20
# second step is to print them out

# part four
less_17_comm_time <- comm_time[which(comm_time < 17)]
sprintf("%.1f%%", length(less_17_comm_time) / length(comm_time) * 100)
# use `sprintf` can output a percentage mode number
```

Quetions Two

The volume of a sphere of radius r is given by $\frac{3}{4}\pi r^3$. For spheres having radius$3, 4, 5, \ldots, 20$ find the corresponding volumes and print the results out. Construct a data frame with columns **radius** and **volume**.

Answer

```
radius_Q2 <- c(seq(from = 3, to = 20, by = 1)) # store the radius data
volume <- c(3 / 4 * pi * (radius_Q2 ^ 3)) # calculate the volume of spheres
of radius_Q2
DF_1 <- data.frame(radius, volume) # create the data frame with columns
radius and volume
print(DF_1) # print out the data frame
```

Question Three

For the data frame `ais` in `DAAG` package，show the number of males and females for each different sport. In which sports is there a large imbalance (e.g. by a factor of more than 2:1) in the numbers of the two sexes? (Hint: you may need to use `install.packages("DAAG")` and `library(DAAG)`)

answer

```
install.packages("DAAG","lattice")
library(DAAG)
names(ais) # get the names of the columns
names(ais$sport) # we can get there are B_Ball Field Gym Netball Row Swim
T_400m T_Sprnt Tennis W_Polo
sport <-
c('B_Ball','Field','Gym','Netball','Row','Swim','T_400m','T_Sprnt','Tennis'
,'W_Polo')
i <- 1
for (i in seq(from = 1, to = length(sport))){
  print(paste("There are",length(which(ais$sex[which(ais$sport ==
sport[i])] == 'f')),"females in",sport[i]))
  i = i + 1
}
for (i in seq(from = 1, to = length(sport))){
  print(paste("There are",length(which(ais$sex[which(ais$sport ==
sport[i])] == 'm')),"males in",sport[i]))
  i = i + 1
}
```

Question Four

Create a matrix with 5 rows and 4 columns. Its elements are selected from $1:100$ randomly with replacement. Calculate the mean and standard deviation of each column.

Answer

```
matrix_Q4 <- matrix(sample(1:100,100,replace = T),nrow = 5, ncol = 4)
mean_col <- rep(0,4)
sd_col <- rep(0,4)

for (i in 1:4){
  mean_col[i] <- mean(matrix_Q4[1:5,i])
  print(paste("The mean of column",i,"is",mean_col[i]))
  sd_col[i] <- sd(matrix_Q4[1:5,i])
  print(paste("The  of standard deviation of column",i,"is",sd_col[i]))
  i = i + 1
}
```

Question Five

Suppose $\mathbf{A} = \begin{bmatrix} 1 & 1 & 3 \\ 5 & 2 & 6 \\ -2 & -1 & -3 \end{bmatrix}$

1. Check that $\mathbf{A}^3 = 0$ where 0 is a 3×3 matrix with every entry equal to 0
2. Replace the third column of $\mathbf{A}$ by the sum of the second and third columns.

Answer

```
#part one
matrix_Q5 <- matrix(c(1,5,-2,1,2,-1,3,6,-3),nrow = 3,ncol = 3)

# define a matrix power's function
"%^%" <- function(mat, pow) {
  if (!is.matrix(mat)) mat <- as.matrix(mat)
  stopifnot(!diff(dim(mat)))
  if (pow < 0) {
    pow <- -pow
    mat <- solve(mat)
  }
  pow <- round(pow)
  switch(pow + 1, return(diag(1, nrow(mat))), return(mat))
  get.exponents <- function(pow)
    if (pow == 0) NULL else c(k <- 2^floor(log2(pow)), get.exponents(pow -
k))
  ans <- diag(nrow(mat))
  dlog2exp <- rev(-diff(c(log2(get.exponents(pow)), 0)))
```

```
  for (j in 1:length(dlog2exp)) {
    if (dlog2exp[j]) for (i in 1:dlog2exp[j]) mat <- mat %*% mat
    ans <- ans %*% mat
  }
  ans
}
matrix_Q5 %^% 3
# the result is :
#       [,1] [,2] [,3]
# [1,]    0    0    0
# [2,]    0    0    0
# [3,]    0    0    0

# part two
matrix_Q5[1:3,3] <- matrix_Q5[1:3,2] + matrix_Q5[1:3,3]
matrix_Q5
# the result is:
#       [,1] [,2] [,3]
# [1,]    1    1    4
# [2,]    5    2    8
# [3,]   -2   -1   -4
```

Question Six

Solve the following system of linear equations in five unknowns: $\begin{align*} &x_1 +2x_2 + 3x_3 + 4x_4 + 5x_5 = 7 \ &2x_1 +x_2 + 2x_3 + 3x_4 + 4x_5 = -1 \ &3x_1 +2x_2 + x_3 + 2x_4 + 3x_5 = -3 \ &4x_1 +3x_2 + 2x_3 + x_4 + 2x_5 = 5 \ &5x_1 +4x_2 + 3x_3 + 2x_4 + x_5 = 17 \ \end{align*}$ by considering an appropriate matrix equation $\mathbf{A}x = y$

```
A <- matrix(c(1,2,3,4,5,2,1,2,3,4,3,2,1,2,3,4,3,2,1,2,5,4,3,2,1),nrow=5,
ncol=5)
y <- c(7,-1,-3,5,17)
solve(A,y)
#the solution is x = -2  3  5  2 -4
```

the solution is $\mathbf{x} = (-2, 3, 5, 2, -4)$

Question Seven

Create a 6×10 matrix of random integers chosen from $1, 2, \ldots, 10$by executing the following line of code: `aMat <- matrix( sample(1:10, size=60, replace=T), nr=6)`

1. Find the number of entries in each row which are greater than 4.
2. Which rows contain exactly two occurrences of the number seven?
3. Find those pairs of columns whose total (over both columns) is greater than 75. The answer should be a matrix with two columns; so, for example, the row $(1, 2)$ in the output matrix means that the sum of columns 1 and 2 in the original matrix is greater than 75.

Repeating a column is not permitted.

Answer

```
#part One
aMat <- matrix( sample(1:10, size=60, replace=T), nr=6)

i <- NULL
for ( i in 1:6){
  print(aMat[i,which(aMat[i,1:10] > 4)])
  i = i + 1
  }

#part two
```{r}
number_we_want <- readline()
frequence <- readline()
for ( i in 1:6){ # i represents the which line
  for ( j in 1:length(as.data.frame(table(aMat[i,1:10]))$Var1)){ # j
represents which line in the data.frame
    if(as.data.frame(table(aMat[i,1:10]))$Var1[j] == number_we_want ){
      if(as.data.frame(table(aMat[i,1:10]))$Freq[j] == frequence ){
        print(i)
      }
    }
  }
}

# According to the question we can let:
# number_we_want <- 7
# frequence <- 2

#The idea is that to change `table` function into a data.frame which can
make it
#clear that whether the numbers we need to find exists or not.
#Then, we  estimate the frequence of the number we want.
```

Question Eight

Physical fitness testing is an important aspect of athletic training. A common measure of the magnitude of cardiovascular fitness is the maximum volume of oxygen uptake during a strenuous exercise. A study was conducted on 24 middle-aged men to study the influence of the time that it takes to complete a 2-mile run. The oxygen uptake measure was accomplished with standard laboratory methods as the subjects performed on a motor driven treadmill. The
```

data are as follows:

Subject	Maximum Volume of $O_2(y)$	Time in Second(x)
1	42.33	918
2	53.10	805
3	42.08	892
4	50.06	962
5	42.45	968
6	42.46	907
7	47.82	770
8	49.92	743
9	36.23	1045
10	49.66	810
11	41.49	927
12	46.17	813
13	48.18	858
14	43.21	860
15	51.81	760
16	53.28	747
17	53.29	743
18	47.18	803
19	56.91	683
20	47.80	844
21	48.65	755
22	53.69	700
23	60.62	748
24	56.73	775

Write R codes for the following questions and show the R outputs. (Remarks: Write your own R codes based on the formula given in the class of regression analysis. **Don't use the existing regression analysis function `lm()` in R.)**

1. In put the data into R
2. Estimate the parameter(β_0, β_1 and σ^2) of a simple linear regression:
   $\mathbf{y_i = \beta_0 + \beta_1 x_i + \epsilon_i,\ i = 1, \ldots, 24}$
3. Calculate fitting value of y_i and residuals $\epsilon = y_i - \hat{y},\ i = 1, \ldots, 24$
4. Calculate 95%-confidence interval for the mean responses as each observed x_i by using the following formulas:

Answer:

```
# part one
x <-
c(918,805,892,962,968,907,770,743,1045,810,927,813,858,860,760,747,743,803,
683,844,755,700,748,775)
y <-
c(42.33,53.10,42.08,50.06,42.45,42.46,47.82,49.92,36.23,49.66,41.49,46.17,4
8.18,43.21,51.81,53.28,53.29,47.18,56.91,47.80,48.65,53.69,60.62,56.73)

# part two
Sx = sum(x)
```

```
Sy = sum(y)
Sxx = sum((x - mean(x))^2)
Syy = sum((y - mean(y))^2)
Sxy = sum((x - mean(x)) * (y - mean(y)))
beta_1 = Sxy/Sxx
alpha_1 = mean(y) - beta*mean(x)

# part three
residuals = y - (beta + alpha * x)

#part four
mean(x)-qt(0.975,49)*sd(x)/sqrt(50) # lower bound
mean(x)+qt(0.975,49)*sd(x)/sqrt(50) # upper bound
```

we can get $\beta_0 = 90.72172$ and $\beta_1 = -0.05102849$ the regression function is $Y = 90.72172 + -0.05102849x_1$

residuals is $(-83240.15,-72977.83,-80881.64,-87224.18,-87776.12,-82242.09,-69807.85
-67356.26,-94767.91,-73434.88,-84057.49,-73710.53,-77791.00,-77977.42
-68896.64,-67715.79,-67352.89,-72802.31,-61905.97,-76521.28,-68446.20
-63451.46,-67799.17,-70252.55)$

The confidence interval is $(800.1541, 852.8459)$