STAT 40001/STAT 50001 **Statistical Computing** **Fall 2024**

Homework 1- Solution

Name:

Instruction: Please submit your R code along with a brief write-up of the solutions (do not submit raw output). Some of the questions below can be answered with very little or no programming. However, write code that outputs the final answer and does not require any additional paper calculations. For example, suppose I ask for how many numbers are greater than 5 in the vector, x=c(1,9,2,8,10,12). Do not simply count the number of **TRUE***s by hand, instead let the R count the number of* **TRUE***s by coding* $\boldsymbol{sum(x > 5)}$ *or* $\boldsymbol{length(x[x > 5])}$.

Q.N. 1) Calculate the following numerical results to the three decimal places

a) $\mathbf{2\ln(3) + \sqrt{2}\sin(\pi)}$

b) $\mathbf{20 \times (25 + 3) - \sqrt{65} + 9^2}$

c) $\mathbf{\ln(5) - \exp(2) + 2^3}$

d) $\mathbf{(9 \div 2) \times 4 - \sqrt{10} + \ln 6 - exp(24)}$

e) $\mathbf{\log 14 + \ln 14 + (47 mod\ 5)}$

Solution: We used the following R codes to perform the calculations

```
a)
> round(2*log(3)+sqrt(2)*sin(pi),3)
[1] 2.197

b)
> round(20*(25+3)-sqrt(65)+9^2,3)
[1] 632.938
c)
> round((log(5)-exp(2)+2^3),3)
[1] 2.22
d)
> round(((9/2)*4-sqrt(10)+log(6)-exp(24)),3)
[1] -26489122113
e)
> round((log10(14)+log(14))+(47%%5),3)
[1] 5.785
```

Q.N. 2) Test scores of Fifteen students in Test 1 and Test 2 are presented below

a) How many students have their test 1 score greater than 80 ?

b) How many students have their test 2 score greater than 85 ?

c) Did all fifteen students take both tests?

d) How many students did better in the second test than the first test?

e) How many students have the same score in the first and second test?

S.N.	Test 1	Test2
1	56	86
2	78	67
3	87	78
4	89	89
5	95	87
6	98	67
7	NA	94
8	78	78
9	87	81
10	98	83
11	54	78
12	89	NA
13	78	93
14	98	98
15	97	100

Solution: We have used the following R code to answer (a)- (e)

```
> Test1
 [1] 56 78 87 89 95 98 NA 78 87 98 54 89 78 98 97
>  Test2=data$Test2
> Test2
 [1]  86  67  78  89  87  67  94  78  81  83  78  NA  93  98 100

 a)
>  sum(Test1>80, na.rm=T)
[1] 9
- 9 students have their Test 1 score higher than 80.
b)
>  sum(Test2>85,na.rm=T)
[1] 7
_ 7 students have their Test 2 score higher than 85.
c)
> any(is.na(Test1))
[1] TRUE
> any(is.na(Test2))
[1] TRUE
> which (is.na(Test1))
[1] 7
> which (is.na(Test2))
[1] 12
No, not all students took both tests.
d)
> sum(Test2>Test1, na.rm=T)
[1] 4
- 4 students did better in the second test than the first test.
```

```
e)
> sum(Test2==Test1, na.rm=T)
[1] 3
> which(Test2==Test1)
[1]  4  8 14
- 3 students have their Test 1 and Test 2 scores the same.
```

Q. N. 3) Create the following vectors using ***seq*** and ***rep*** function:

```
V1= 1, 2, 3, 4, 5, 1, 2, 3, 4, 5, 1, 2, 3, 4, 5
V2= 1, 1, 1, 1, 2, 2, 2, 2, 3, 3, 3, 3
V3=  MATH,  MATH,  STAT, STAT, STAT, STAT, STAT, ECE, ECE, ECE, BIO, BIO
V4=  B, B, B, B, C, C, C, C, C, D, D, D, D, D, D, E, E, E, E, E, E, E
```

Solution: We use the following R codes to generate the desired sample

```
> V1=rep(c(1,2,3,4,5), 3)
> noquote(paste(V1,collapse=","))
[1] 1,2,3,4,5,1,2,3,4,5,1,2,3,4,5

> V2=rep(c(1,2,3), each=4)
> noquote(paste(V2,collapse=","))
[1] 1,1,1,1,2,2,2,2,3,3,3,3

> V3=rep(c("MATH","STAT","ECE","BIO"),times=c(2,5,3,2))
> noquote(paste(V3,collapse=","))
[1] MATH,MATH,STAT,STAT,STAT,STAT,STAT,ECE,ECE,ECE,BIO,BIO

> V4=rep(c("B","C","D","E"),times=c(4,5,6,7))
> noquote(paste(V4,collapse=","))
[1] B,B,B,B,C,C,C,C,C,D,D,D,D,D,D,E,E,E,E,E,E,E
```

Q.N. 4) Create the following matrix (M) with the column and row names (Note that the numbers are in sequence from 1 to 20)

```
>M
             column-1 column-2 column-3 column-4 column-5
Experiment.1        1        5        9       13       17
Experiment.2        2        6       10       14       18
Experiment.3        3        7       11       15       19
Experiment.4        4        8       12       16       20
```

a) Determine the dimension of the matrix M
b) Select the first two row of the matrix M
c) Calculate the sum of all columns of the matrix M
d) Calculate the sum of all rows of the matrix M
e) Use "sample" to shuffle the elements of each row of the matrix M

Solution: We used the code below to create the desired matrix

```
> M<-matrix(1:20,nrow=4)
> rownames(M)<-rownames(M,do.NULL=FALSE,prefix="Experiment.")
> colnames(M)<-colnames(M,do.NULL=FALSE,prefix="Column-")
> M
             Column-1 Column-2 Column-3 Column-4 Column-5
Experiment.1        1        5        9       13       17
Experiment.2        2        6       10       14       18
Experiment.3        3        7       11       15       19
Experiment.4        4        8       12       16       20

a)
> dim(M)
[1] 4 5
b)
> M[c(1,2),]
             Column-1 Column-2 Column-3 Column-4 Column-5
Experiment.1        1        5        9       13       17
Experiment.2        2        6       10       14       18




c)
> colSums(M)
Column-1 Column-2 Column-3 Column-4 Column-5
      10       26       42       58       74
d)
> rowSums(M)
Experiment.1 Experiment.2 Experiment.3 Experiment.4
          45           50           55           60
e)

> t(apply(M,1,sample))
             [,1] [,2] [,3] [,4] [,5]
Experiment.1   17    1    5   13    9
Experiment.2   14   18    6    2   10
Experiment.3   19    3   11   15    7
Experiment.4    8   16    4   20   12
```

Q. N. 5) Use R to solve the following system of equations:

$$\begin{aligned} \mathbf{2}x_1 + x_2 + x_3 - \mathbf{3}x_4 + x_5 &= \mathbf{12} \\ x_1 - x_2 + \mathbf{2}x_3 + x_4 - x_5 &= \mathbf{1} \\ \mathbf{2}x_1 + x_2 - x_3 + \mathbf{2}x_4 + x_5 &= \mathbf{-2} \\ x_1 - \mathbf{3}x_2 + x_3 + \mathbf{2}x_4 - x_5 &= \mathbf{-9} \\ x_1 + \mathbf{2}x_2 - x_3 + \mathbf{3}x_4 - x_5 &= \mathbf{0} \end{aligned}$$

Solution:

```
> A=matrix(c(2,1,1,-3,1,1,-1,2,1,-1,2,1,-1,2,1,1,-3,1,2,-1,1,2,-1,3,-1),nrow=5,byrow=TRUE)
> A
     [,1] [,2] [,3] [,4] [,5]
[1,]    2    1    1   -3    1
[2,]    1   -1    2    1   -1
[3,]    2    1   -1    2    1
[4,]    1   -3    1    2   -1
[5,]    1    2   -1    3   -1

> Y=matrix(c(12,1,-2,-9,0),nrow=5)
> Y
     [,1]
[1,]   12
[2,]    1
[3,]   -2
[4,]   -9
[5,]    0
> X=solve(A,Y)
> X
     [,1]
[1,]    1
[2,]    3
[3,]    2
[4,]   -2
[5,]   -1
```

Hence, the solution of the given system of equations is $(x_1, x_2, x_3, x_4, x_5) = (1, 3, 2, -2, -1)$.

Q. N. 6) The Fibonacci sequence is a famous sequence in mathematics. The first two elements are defined as [1, 1]. Subsequent elements are defined as the sum of the preceding two elements. For example, the third element is 2 (= 1+1), the fourth element is 3 (= 1+2), the fifth element is 5 (= 2+3), and so on. Print first 20 Fibonacci numbers.
Hint: To obtain the first 10 Fibonacci numbers in R, we can use

```
> Fibonacci <- numeric(10)
> Fibonacci[1] <- Fibonacci[2] <- 1
> for (i in 3:10) Fibonacci[i] <- Fibonacci[i - 2] + Fibonacci[i - 1]
> Fibonacci
```

Solution: We use R code below to generate the first 50 Fibonacci numbers.

```
> Fibonacci <- numeric(20)
> Fibonacci[1] <- Fibonacci[2] <- 1
> for (i in 3:20) Fibonacci[i] <- Fibonacci[i - 2] + Fibonacci[i - 1]
> Fibonacci
 [1]    1    1    2    3    5    8   13   21   34   55   89  144  233  377  610
```

```
[16]  987 1597 2584 4181 6765
```

Q.N. 7) Stock and Watson (2007) provide several subsets created from March Current Population Surveys (CPS) with data on the relationship of earnings and education over several year. CPSS.XLS data provided with this assignment contains data for 2003 and 2004. Use R to extract for 2004 (2nd tab) and determine the dimension of the data.
Solution: In order to read ".xls" file first we need to install "readxl" package and use the R code below.

```
> library(readxl)
> file.choose()
[1] "G:\\My Drive\\Assignments\\CPSS.xls"
> data=read_excel("G:\\My Drive\\Assignments\\CPSS.xls", sheet=2)
> head(data)
  # A tibble: 6  5
   ...1 earnings degree     gender   age
  <dbl>    <dbl> <chr>      <chr>  <dbl>
1     1     34.6 bachelor   male      30
2     2     19.2 bachelor   female    30
3     3     13.7 highschool female    30
4     4     19.2 bachelor   female    30
5     5     19.2 bachelor   male      25
6     6     38.5 bachelor   female    32
> dim(data)
[1] 7986    5
```

There are 7986 observations with 5 variables.

Q.N. 8) An article entitled "What Does It Take to Heat a New Room? Estimating Utility Demand in a Home" by Robert Carver was published in the Journal of Statistics Education v.6, n.1 (1998). It has been observed that in a residential home, energy consumption is closely related to the outdoor temperature and size of the house. The date collected for this study are provided in the link below.
`http://jse.amstat.org/datasets/utility.dat.txt`

a)Import the data in R
b)How many variables are included in this dataset?
c)The missing values in this dataset are denoted by $*$. Please create a clean dataset by removing the missing values.

Solution:
a) We use R code below to import the data in R

```
> data=read.table("http://jse.amstat.org/datasets/utility.dat.txt")
```

b) Based on the R output there are 81 observations with13 variables included in the given dataset

```
> dim(data)
[1] 81 13
```

c) Since the missing values are denoted by $*$ We used R code below to convert them to "NA" and count the number of missing values.

```
> data=read.table("http://jse.amstat.org/datasets/utility.dat.txt", na.strings="*")
> sum(is.na(data))
[1] 30
```

Based on the above results there are 30 missing entries in the dataset. Note that a single individual may have more than one missing entry. We use R code below to create the Clean dataset by removing all missing values. The resulating clean data has 71 observations with 13 variables.

```
> Clean=na.omit(data)
> dim(Clean)
[1] 71 13
```