Bài 6:

Lý thuyết mẫu

1. Nhập, xử lý, xuất dữ liệu

1.1. Nhập dữ liệu

a) Working directory

Trước khi nhập dữ liệu ta nên thiết lập thư mục làm việc. Đây là thư mục chứa những thứ mà ta muốn tương tác với R (như file dữ liệu, code script R, hình ảnh, đồ thị, package,…).
- Để thiết lập thư mục làm việc, ví dụ ở ổ D, thư mục Works

```
> setwd('D:/Works')
```

hoặc vào `File -> Change dir`
- Xem thư mục hiện hành

```
> getwd()
```

- Liệt kê tất cả file trong thư mục làm việc

`> list.files()` hoặc `> dir()`

b) Workspace

Từ khi mở R (cửa sổ R console xuất hiện) cho đến khi tắt R là một phiên làm việc. Những đối tượng ta tạo ra trong một phiên làm việc được R lưu trong Workspace. Ta có thể lưu lại mọi thứ trong Workspace này để tiếp tục công việc đang làm của ta ở một thời điểm khác.
- Lưu Workspace

```
> save.image('ten_file.rda')
```

- Tải Workspace đã lưu

```
> load('ten_file.rda')
```

- Lưu biến đang làm việc, chẳng hạn biến x

```
> save(x, file='ten_file.rda')
```

- Khôi phục biến x

`> load('ten_file.rda')`, `ten_file.rda` là file chứa biến x vừa lưu ở trên.
- Xóa 1 biến ra khỏi Workspace

```
> rm(x)
```

- Xóa tất cả

```
> rm(list=ls())
```

- Liệt kê tất cả những biến trong Workspace

```
> ls()
```

- Xem thông tin của 1 biến

```
> str(x)
```

- Xem thông tin của tất cả biến đang làm việc

```
> ls.str()
```

c) Nhập dữ liệu

- Nhập trực tiếp từ R, dùng lệnh `edit(data.frame())`

```
> frame <- edit(data.frame())
```

Sẽ mở 1 cửa sổ nhập trực tiếp, ta nhập các biến cần thiết thuộc frame vào.
- Nhập từ file .txt, dùng lệnh `read.table`

```
> data <- read.table('D:/Đường dẫn/solieu.txt ', header=TRUE,
sep= "")
```

Đọc file solieu.txt rồi gán cho biến data. Nếu file để trong thư mục làm việc thì không cần chỉ ra đường dẫn.
- Nhập từ file excel (file excel phải save dưới dạng .csv)

```
> data <- read.csv('solieu.csv', header= FALSE)
```

1.2. Xử lý dữ liệu

0/ Đưa một data frame vào workspace để xử lý: attach(dataframe)
1/ Tách dữ liệu: Lệnh subset(*bien_goc*, *dieu_kien*)
2/ Nhập 2 dataframe thành một: dùng lệnh merge(*frame_1*, *frame_2*, *by=)*
3/ Biến đổi số liệu: từ biến dạng numeric sang biến phân loại, sử dụng các phép toán logic hoặc dùng lệnh replace( ).
Dùng lệnh factor( ): chuyển từ biến dạng numeric sang nhân tố.
4/ Phân nhóm số liệu, dùng hàm cut2 (trong thư viện Hmisc).
> library(Hmisc)
> cut2(*bien_goc*, g= *so_nhom*)

1.3. Xuất dữ liệu

a) Định dạng R (.rda)

Lệnh: `save()`
VD: `save(data frame, file= "data.rda")`

b) Định dạng excel (.xls, xlsx)

- Sao chép 1 vector trong R vào clipboard sau đó dán vào Excel.
`writeClipboard(x)`, trong đó `x` là 1 vector dạng ký tự (character)
Ví dụ:

```
writeClipboard(as.character(factor.name))
```

Ví dụ:

```
writeClipboard(as.character(numeric.variable))
```

- Sao chép 1 data frame trong R vào clipboard sau đó dán vào Excel.

```
write.table(data,"clipboard",sep="\t",col.names=NA)
```

- Lưu trực tiếp một data frame trong R thành 1 file Excel
Hàm `write.xlsx()` trong package `xlsx` có thể được dùng để lưu một data frame R thành một workbook Excel 2007.

```
library(xlsx)
write.xlsx(x , outfile , col.Names=TRUE, row.names=TRUE,
sheetName="Sheet 1", append=FALSE)
```

Ví dụ:

```
library(xlsx)
write.xlsx(mydata, "mydata.xlsx")
```

xuất data frame `mydata` vào một worksheet (Sheet 1 theo mặc định) trong một work-book Excel được đặt tên là `mydata.xlsx` trong thư mục làm việc hiện tại. Mặc định, các tên biến trong dataset được dùng để tạo tiêu đề cột trong spreadsheet (bảng tính) và các tên hàng được đặt trong cột đầu tiên của bảng tính. Nếu `mydata.xlsx` đã có, nó sẽ bị ghi đè. Nếu cần lưu ở một thư mục khác thì cho đường dẫn vào. Ví dụ,
`write.xlsx(mydata, "D:/mydata.xlsx")` hoặc `write.xlsx(mydata, "D:\\mydata.xlsx")`

c) Định dạng text (.txt)

Hàm `write.table()` dùng để lưu một đối tượng trong R vào 1 file text
`write.table(x, outfile , sep=delimiter, quote=TRUE, na="NA")`
trong đó
`x` là đối tượng cần lưu lại
`outfile` là file chứa đối tượng.
Ví dụ:
`write.table(mydata, "mydata.txt", sep=",")`
sẽ lưu dataset mydata vào 1 file được phân cách bởi dấu phẩy đặt tên là mydata.txt trong thứ mực làm việc hiện tại. Chỉ rõ đường dẫn (ví dụ, "`d:/myprojects/mydata.txt`") để lưu file output ở nơi khác. Thay `sep=","` bằng `sep="\t"` sẽ lưu dữ liệu trong một file được phân cách bởi dấu tab. Mặc định, các chuỗi được đặt trong dấu nháy kép ("") và các giá trị khuyết được viết là NA.

2. Một số hàm về vec-tơ: cho vec-tơ x

`max(x), min(x)` : giá trị lớn nhất, bé nhất của x.
`sum(x)`: tổng các giá trị trong x
`mean(x)`: trung bình của x
`median(x)`: trung vị của x
`range(x)`: bằng max(x) – min(x)
`var(x)`: phương sai của x
`sort(x)`: sắp xếp x, mặc định theo thứ tự tăng dần
`order(x)`: trả về các vị trí của x khi đã sắp theo thứ tự tăng dần
`quantile(x)`: tính các phân vị của x
`cumsum(x)`: tổng tích lũy
`cumprod(x)`: tích tích lũy

3. Vẽ đồ thị một số phân phối thông dụng: (nhị thức, poisson, đều, mũ, chuẩn)

Dùng hàm plot( );
Ví dụ: vẽ đồ thị hàm số
$y = e^x$ và $y = \ln(x)$

```
> x <- seq(0,10,0.1)
> y <- exp(x)
> plot(y~x, type='l ')
> y <- log(x)
> plot(y~x, type='l')
```

Phân phối nhị thức:

```
> x <- 0:50
> y <- dbinom(x,50,0.25)
> plot(x,y, 'S ')  hoặc > plot(x,y, 'h ')
```

Phân phối chuẩn:

```
> sample <- rnorm(50)
> hist(sample, prob=T)
>  mu <- mean(sample)
> sigma <- sd(sample)
> x <- seq(-4,4,length=500)
> y <- dnorm(x,mu,sigma)
> lines(x,y)
```

4. Bài tập:

1 Tạo vec-tơ: x =[1,2,5,7,-3,0,5,1,5,6] và y=[2,2,0,-5,7,8,11,9,3,2]

a. Tính x+y, x*y,x-y.
b. Tao z =[Những phần tử chẵn của x], t =[Những phần tử lẻ của y]
c. Trích những phần tử lớn hơn 0 của x và y.
d. Tính trung bình, độ lệch tiêu chuẩn, sai số chuẩn của x và y.
e. Tìm phần tử lớn nhất, bé nhất của x, y.
f. Sắp xếp x tăng dần, y giảm dần.
g. Lưu x và y.

2 Nhập số liệu từ file data01.xls bằng lệnh read.csv( ) (chuyển file .xls -> .csv) gán vào frame data1. Thực hiện:

a. Tính trung bình, phương sai, trung vị của các biến FPSA và TPSA.
b. Vẽ biểu đồ dạng đường, boxplot cho FPSA và TPSA.
c. Tách những giá trị của biến FPSA có K=0 và K=1.
d. Đọc số liệu từ file data02.csv gán vào frame data2, merge 2 frame này theo biến K.
e. Tạo biến mới tPSA theo yêu cầu sau: Nếu tuổi <= 30, tPSA=0; nếu 30< tuổi <=50, tPSA=1; nếu tuổi>50, tPSA =2. Tạo bảng thống kê cho tPSA.

3 Bảng sau là điểm một bài kiểm tra gồm 3 câu hỏi của 10 SV

Sinh viên	Câu hỏi 1	Câu hỏi 2	Câu hỏi 3
1	3	5	1
2	3	3	3
3	3	5	1
4	4	5	1
5	3	2	1
6	4	2	3
7	3	5	1
8	4	5	1
9	3	4	1
10	4	2	1

a. Nhập các số liệu sau và gán vào biến tương ứng sử dụng 3 cách: Dùng lệnh c( ); dùng lệnh scan( ); lệnh read.table( ) (Tạo file .txt) , edit(data.frame()).

b. Tạo bảng kết quả riêng cho câu hỏi 1 và câu hỏi 2.
c. Vẽ biểu đồ bar cho 3 câu hỏi.
d. Vẽ biểu đồ bar dạng nằm ngang cho câu hỏi 2 và 3. (Gợi ý: dùng đối số horiz = T trong lệnh barplot).

4

a. Tạo ngẫu nhiên 100 giá trị có phân phối nhị thức, với n = 60 và xác suất thành công mỗi lần 0.4. Vẽ biểu đồ tổ chức tần số.
b. Tạo ngẫu nhiên 100 giá trị có phân phối Poisson với lambda = 4, vẽ biểu đồ tổ chức tần số.
c. Tạo ngẫu nhiên 100 giá trị có phân phối chuẩn có trung bình là 50 và độ lệch tiêu chuẩn 4. Vẽ hàm phân phối, hàm mật độ.
d. Tạo ngẫu nhiên 100 giá trị có phân phối mũ với lambda=1/25. Vẽ hàm phân phối, hàm mật độ.

5 File *diesel_engine.dat* và *diesel_time.xls* chứa số liệu về hoạt động của các động cơ chạy bằng dầu diesel. Thực hiện:

a. Đọc số liệu từ hai file này, gán và hai dataframe, đặt tên hai dataframe cùng tên với file.

b. Liệt kê tên các biến có trong hai dataframe vừa nhập.

c. Xác định có bao nhiêu dữ liệu bị khuyết (missing data) trong *diesel_engine*. Thay thế các giá trị khuyết trong biến *speed* bằng 1500, biến *load* bằng 20.

d. Tính: trung bình, phương sai, độ lệch tiêu chuẩn, giá trị lớn nhất, nhỏ nhất của biến *alcohol* trong dataframe *diesel_engine*.

e. Ghép hai dataframe *diesel_engine* và *diesel_time* lại thành một frame có tên là *diesel*.

f. Trích giá trị của biến *run* (số thứ tự các động cơ) mà có thời gian trễ (biến *delay*) dưới 1.000.

g. Đếm xem có bao nhiêu động cơ có *timing* bằng 30.

h. Vẽ biểu đồ boxplot cho các biến *speed, timing* và *delay*.

i. Vẽ biểu đồ phân tán cho các cặp biến (*timing, speed*), (*temp, press*).

j. Chuyển biến *load* sang biến nhân tố.

k. Chia phạm vi giá trị của biến *delay* thành 4 đoạn đều nhau và đếm số giá trị nằm trong các đoạn đó. Tạo bảng thống kê và vẽ biểu đồ cột.

l. Chia phạm vi giá trị của biến *delay* thành 4 đoạn như sau: (0.283, 0.7], (0.7, 0.95], (0.95, 1.2], (1.2, 1.56]. Tạo bảng thống kê và vẽ biểu đồ cột.

6 Cho số liệu sau:

year	*snow.cover*
1970	6.5
1971	12.0
1972	14.9
1973	10.0
1974	10.7

1975	7.9
1976	21.9
1977	12.5
1978	14.5
1979	9.2

a. Nhập số liệu trên vào R.
b. Vẽ *snow.cover* theo *year*.
c. Vẽ biểu đồ histogram cho *snow.cover*.
d. Lặp lại câu b. và c. sau khi lấy logarit của biến *snow.cover*.

7 Cho số liệu sau:

Temperature (F)	*Erosion incidents*	*Blowby incidents*	*Total incidents*
53	3	2	5
57	1	0	1
63	1	0	1
70	1	0	1
70	1	0	1
75	0	2	1

Nhập số liệu trên vào một dataframe, vẽ đồ thị biểu diễn tổng số *incidents* theo *temperature*.

8 Thống kê số liệu tỉ lệ lạm phát tại 4 nước trong giai đoạn 1960-1980 được thu thập trong 2 bảng số liệu sau (Đvt: %)

Nam	**US**	**Anh**
1960	**1.5**	**1**
1961	**1.1**	**3.4**
1962	**1.1**	**4.5**
1963	**1.2**	**2.5**
1964	**1.4**	**3.9**
1965	**1.6**	**4.6**
1966	**2.8**	**3.7**
1967	**2.8**	**2.4**
1968	**4.2**	**4.8**
1969	**5**	**5.2**
1970	**5.9**	**6.5**
1971	**4.3**	**9.5**
1972	**3.6**	**6.8**
1973	**6.2**	**8.4**
1974	**10.9**	**16**
1975	**9.2**	**24.2**
1976	**5.8**	**16.5**
1977	**6.4**	**15.9**
1978	**7.6**	**8.3**
1979	**11.4**	**13.4**

Nam	**Nhat**	**Duc**
1960	**3.6**	**1.5**
1961	**5.4**	**2.3**
1962	**6.7**	**4.5**
1963	**7.7**	**3**
1964	**3.9**	**2.3**
1965	**6.5**	**3.4**
1966	**6**	**3.5**
1967	**4**	**1.5**
1968	**5.5**	**18**
1969	**5.1**	**2.6**
1970	**7.6**	**3.7**
1971	**6.3**	**5.3**
1972	**4.9**	**5.4**
1973	**12**	**7**
1974	**24.6**	**7**
1975	**11.7**	**5.9**
1976	**9.3**	**4.5**
1977	**8.1**	**3.7**
1978	**3.8**	**2.7**
1979	**3.6**	**4.1**

1980	**13.6**	**18**		**1980**	**8**	**5.5**

a. Nhập dữ liệu trên vào 2 data.frame *lamphat1* và *lamphat2* trong R bằng 3 cách.
b. Trộn 2 data.frame trên vào 1 data.frame duy nhất là *lamphat* theo Nam.
c. Đếm số năm các nước US, Anh, Nhật, Đức có tỉ lệ lạm phát trên 5%.
d. Vẽ đồ thị phân tán về tỉ lệ lạm phát cho mỗi quốc gia theo thời gian. Cho nhận xét tổng quát về lạm phát của 4 nước?
e. Tính trung bình, trung vị, Max, Min, độ lệch chuẩn, sai số chuẩn của từng nước?
f. Để xác định lạm phát nước nào biến thiên nhiều hơn, ta cần dựa vào tham số thống kê nào? Kết luận?
g. Tạo một data.frame mới *lamphat1* với số biến như trong data.frame *lamphat* nhưng không chứa dữ liệu của năm 1980.
h. Ta biết rằng hệ số của phương trình hồi quy tuyến tính $\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 \hat{X}_i$ được xác định như sau:

$$\hat{\beta}_2 = \frac{\sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y}}{\sum_{i=1}^{n} X_i^2 - n(\bar{X})^2}$$

$$\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X}$$

Xác định các hệ số này trong mô hình hồi quy: lạm phát theo thời gian cho US bằng cách sử dụng data.frame *lamphat1*. Vẽ đồ thị phương trình hồi quy này?
i. Sử dụng phương trình hồi quy trong câu h) hãy xác định tỉ lệ lạm phát trong năm 1980 của US. So sánh với số liệu thực tế?