Tuần 7: LT Bài 6.pdf

- Điểm danh trên gg classroom của lớp.
- Mở R, thiết lập thư mục làm việc cho R.
- Mở file “Bai 6.pdf”: đọc và thực hành các bài tập.

Thực hành xác suất thống kê

Bài 6: LÝ THUYẾT MẪU

Bài tập

Bài 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.

```
# Bai 1
# Tao mau x,y
x <- c(1,2,5,7,-3,0,5,1,5,6)
y <- c(2,2,0,-5,7,8,11,9,3,2)
# a. tinh tong, tich, hieu
x+y;x*y; x-y
# b. Tach gia tri cua x chan gan vao z, gia tri cua y le cho t
z <- x[x%%2 == 0];z
t <- y[y%%2 == 1];t
```

```
# c. Tach gia tri duong cua x, y
x[x>0]; y[y>0]

# d.
# trung binh mau cua x
mean(x)
# do lech tieu chuan cua x
sd(x)
# sai so chuan cua x
sd(x)/sqrt(length(x))

# e. Tim phan tu be nhat va lon nhat cua x,y
min(x); max(x)
min(y); max(y)

# f. Sap xep x tang dan; y giam dan
sort(x)
sort(y,decreasing=T)

# g. Luu bien x,y
save(x,file='varx.rda');save(y,file='vary.rda')
```

Bài tập

Bài 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**, ghép 2 frame này theo biến **K**.

e. Tạo biến mới **tPSA** theo yêu cầu sau: Nếu tuổi <= 60, **tPSA**=0; nếu 60< tuổi<=70, **tPSA**=1; nếu tuổi>70, **tPSA** =2. Tạo bảng thống kê cho **tPSA**.

```
# Bai 2
# Doc du lieu tu file "data01.csv" vao R va gan vao bien data1
data1 <- read.csv("data01.csv", header = TRUE)
attach(data1); names(data1)
data1
#a. Trung binh
mean(FPSA); mean(TPSA)
#  Phuong sai
var(FPSA); var(TPSA)
# Trung vi
median(FPSA); median(TPSA)
```

```
#b. Ve bieu do dang duong
plot(FPSA, type = "l")
plot(TPSA, type = "l")
# Ve bieu do boxplot
boxplot(FPSA)
boxplot(TPSA)

#c. Tach nhung gia tri cua FPSA co K = 0
fpsa0 <- subset(FPSA,K==0)

# Tach nhung gia tri cua FPSA co K = 1
fpsa1 <- subset(FPSA,K==1)

#d. Doc du lieu tu file "data02.csv" vao R va gan vao bien data2
data2 <- read.csv("data02.csv",header = T)
names(data2)
attach(data2)
# Ghep 2 data lai theo K
dat <- data.frame(data1[ ,1:3],data2);dat

#e. Tao bien tPSA theo yeu cau de bai
tPSA <- Age
tPSA[Age <= 60] <- 0
tPSA[Age > 60 & Age <=70] <- 1
tPSA[Age >70] <- 2

# Tao bang thong ke cho tPSA
tab <- table(tPSA); tab
```

Bài tập

Bài 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`).

Bài tập

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

```
# Bai 3
sv <-1:10
ques1 <- c(3,3,3,4,3,4,3,4,3,4)
ques2 <- c(5,3,5,5,2,2,5,5,4,2)
ques3 <- c(1,3,1,1,1,3,1,1,1,1)
# a) Tao bang diem
Diem <- data.frame(sv,ques1,ques2,ques3)

# b) Tao bang ket qua rieng cho cau hoi 1,2
tab1 <- table(ques1);tab1
tab2 <- table(ques2);tab2
tab3 <- table(ques3);tab3

par(mfrow = c(1,3))
# c) Ve bieu do bar cho 3 cau hoi
barplot(tab1); barplot(tab2); barplot(tab3)
par(mfrow = c(2,1))
# d) Ve bieu do bar dang nam ngang cho cau hoi 2,3
barplot(tab2, horiz=T)
barplot(tab3, horiz=T)
```

Bài tập

Bài 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.

Bài tập

Bài 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.

```
# Bai 4
# a) Tao 100 gia tri co phan phoi nhi thuc B(60, 0.4)
x <- rbinom(100,60,0.4);x
hist(x, main='Mo phong phan phoi nhi thuc')

# b) Tao ngau nhien 100 gia tri co phan phoi Poisson voi lambda=4
y <- rpois(100,4);y
hist(y)
```

Bài tập

Bài 4.

c. Tạo ngẫu nhiên 100 giá trị có phân phối chuẩn có trung bình là 50 và độ lệch tiêu chuẩn 4. Vẽ hàm mật độ.

d. Tạo ngẫu nhiên 100 giá trị có phân phối mũ với lambda=1/25. Vẽ hàm mật độ.

```
# c) Tao ngau nhien 100 gia tri co phan phoi chuan moi trung binh
#bang 50 va do lech tieu chuan 4
z <- rnorm(100,50,4);z
# Ve ham mat do
plot(density(z),main='Bieu do ham mat do')

# d) Tao ngau nhien 100 gia tri co phan phoi mu voi lambda=1/2500
t <- rexp(100,1/2500);t
# Ve ham mat do
plot(density(t),main='Bieu do ham mat do')
```

Bài tập

Bài 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.

Bài tập

Bài 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*.

```
# Bai 5
# a) Doc du lieu
diesel.engine = read.table('diesel_engine.dat',header=T)
diesel.time = read.csv('diesel_time.csv',header=T)
attach(diesel.engine);attach(diesel.time)
# b) Liet ke ten cac bien trong hai dataframe
names(diesel.engine);names(diesel.time)
# c) Data frame diesel.engine
length(diesel.engine[diesel.engine=='NA'])
# Xac dinh so bien khuyet trong bien speed va thay doi gia tri
s = length(speed[speed=='NA'])
speed[is.na(speed)==T] = 1500

# Xac dinh so bien khuyet trong bien load va thay doi gia tri
l = length(load[load=='NA'])
load[is.na(load)==T]=20
s + l
speed;load
# d)Tinh trung binh, phuong sai, do lech chuan cua bien alcohol
mean(alcohol);var(alcohol);sd(alcohol); min(alcohol); max(alcohol)
```

Bài tập

Bài 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:

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*.

```
# e) Ghep hai dataframe diesel.engine va diesel.time thanh diesel
diesel = data.frame(diesel.engine,diesel.time)
diesel

# f) Trich gia tri cua bien run ma co delay < 1.000
run[delay<1.000]

# g) Dem so dong co co timing = 30
length(run[timing==30])

# h) Ve bieu do boxplot cho cac bien speed, timing, delay
par(mfrow = c(1,3))
boxplot(speed)
boxplot(timing)
boxplot(delay)
```

Bài tập

Bài 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:

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.

```
# i) Ve bieu do phan tan cho cac cap (timing,speed) va (temp,press)
plot(timing,speed)
plot(temp,press)

# j) Chuyen bien load sang bien nhan to
load = factor(load)
load

# k) Chia bien delay thanh 4 doan deu nhau
delay
new.delay = cut(delay,breaks=4)
new.delay

# So gia tri trong tung khoang
tab = table(new.delay)
tab
barplot(tab)
```

Bài tập

Bài 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:

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.

```
# l) Chia bien delay thanh cac doan: (0.283,0.7],(0.7,0.95],(0.95,1.2],(1.2,1.56]
cut.points = c(0.283,0.7,0.95,1.2,1.56)
new.delay1 = cut(delay,breaks=cut.points)
new.delay1

# So gia tri trong tung khoang
tab1 = table(new.delay1)
tab1
barplot(tab1)
```