BỘ GIÁO DỤC VÀ ĐÀO TẠO
TRƯỜNG ĐẠI HỌC GIAO THÔNG VẬN TẢI
KHOA KHOA HỌC CƠ BẢN
BỘ MÔN: ĐẠI SỐ & XÁC SUẤT THỐNG KÊ

HƯỚNG DẪN THỰC HÀNH THỐNG KÊ

MỤC LỤC

BÀI 1: GIỚI THIỆU R

1.1 Giới thiệu về phần mềm R

R là một phần mềm phân tích thống kê và vẽ biểu đồ. Phần mềm R được hai nhà thống kê Ross Ihaka và Robert Gentleman đưa ra và được cộng đồng thống kê phát triển các gói lệnh (packages). R là phần mềm hoàn toàn miễn phí.

1.2 Cài đặt phần mềm R

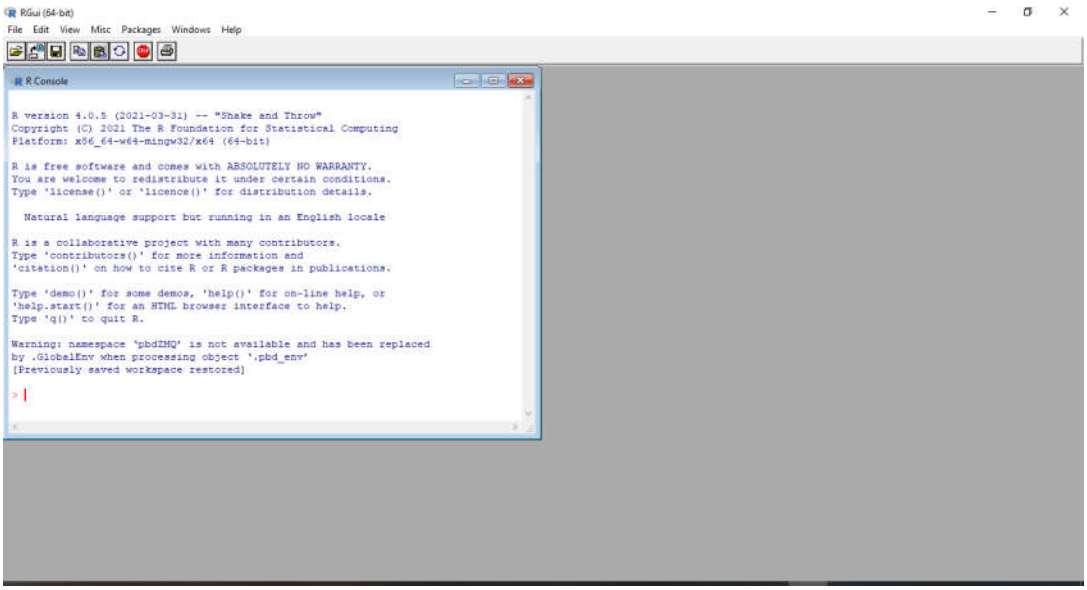
Để sử dụng phần mềm R, trước hết chúng ta truy cập trang web để tải phần mềm R:

https://cran.r-project.org/

phiên bản hiện tại của R là **`R-4.0.5 for Windows (32/64 bit).`** Phiên bản này của R có kích thước 85 MB và có địa chỉ cụ thể là **https://cran.r-project.org/bin/windows/base/R-4.0.5-win.exe**.

Sau khi cài đặt phần mềm R, truy cập vào phần mềm R trên màn hình Desktop

:



Chúng ta đã sẵn sàng sử dụng phần mềm R.

1.3 Cài đặt gói lệnh

Trong phần thục hành thống kê, chúng ta sử dụng một số gói lệch cơ bản trong thống kê và một số thư viện được dùng trong giảng dạy môn thống kê:

```
> install.packages("UsingR")
```

Để sử dụng thư viện UsingR ta dùng câu lệnh

```
> require("UsingR")
```

1.4 Các phép toán

Các phép toán cơ bản trên R gồm các phép toán số học: + (cộng) , - (trừ), * (nhân), / (chia), ^ lũy thừa. Trên R console thực hiện các phép tính sau:

```
> 1+2*(3+4)
```

Kết quả là:

```
[1] 15
```

```
> ((1+2)/(3+4))^4
```

```
[1] 0.03373594
```

Cấu trúc văn phạm của một hàm

Đối tượng ← tên hàm(tham số 1, tham số 2, …., tham số n)

Ví dụ trên R cung cấp tên hàm rnorm, trên R console ta thực hiện lệnh

```
> x<-rnorm(10)
```

Kết quả:

```
> x
 [1] -0.07092755  0.59279742  1.38188889  1.37016020 -1.54141683  0.19548770
 [7]  1.21982145  1.22848266 -0.30838562 -0.37820785
```

Trong đó **x** là một đối tượng (object), **rnorm** là một hàm, và **10** là một tham số của hàm. Với câu lệnh trên, thì x là một mẫu gồm 10 quan sát từ phân phối chuẩn tắc N(0,1).

1.5 Cách nhập dữ liệu

Cách 1: Nhập dữ liệu trực tiếp bằng lệnh c()

Ta có một bảng số liệu về lương (triệu đồng) của 10 kỹ sư ngành CNTT:

10, 11, 9, 8, 6, 7, 12, 15, 18, 20

Ta lưu trữ vào đối tượng luong như sau:

```
> luong<-c(10, 11, 9, 8, 6, 7, 12, 15, 18, 20)
> luong
 [1] 10 11  9  8  6  7 12 15 18 20
```

Truy cập đối tượng thứ nhất, luong[1], thứ 2, luong[2].

```
> luong[1]
[1] 10
> luong[2]
[1] 11
```

Lọc các đối tượng có mức lương trên 10 (triệu)

```
> luong[luong>10]
[1] 11 12 15 18 20
```

Tìm max và min của bảng lương

> max(luong)

[1] 20

> min(luong)

[1] 6

Độ dài của véc-tơ luong (dùng hàm length)

> length(luong)

[1] 10

Tính lương trung bình theo 2 cách như sau

> mean(luong)

[1] 11.6

hoặc

> sum(luong)/length(luong)

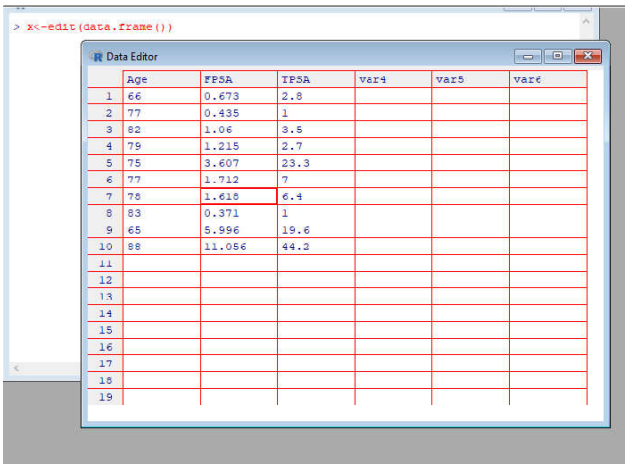
[1] 11.6

Cách 2: Nhập số liệu bằng lệnh edit(data.frame())

Ta có một bảng số liệu như sau:

Age	FPSA	TPSA
66	0.673	2.8
77	0.435	1
82	1.06	3.5
79	1.215	2.7
75	3.607	23.3
77	1.712	7
78	1.618	6.4
83	0.371	1
65	5.996	19.6
88	11.056	44.2

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



Sau đó ta nhập số liệu vào bảng, sau khi nhập ta bấm vào phím X để thoát phần nhập số liệu. Dữ liệu vừa nhập được nhập vào đối tượng x.

```
> x
   Age   FPSA TPSA
1   66  0.673  2.8
2   77  0.435  1.0
3   82  1.060  3.5
4   79  1.215  2.7
5   75  3.607 23.3
6   77  1.712  7.0
7   78  1.618  6.4
8   83  0.371  1.0
9   65  5.996 19.6
10  88 11.056 44.2
```

Ta truy cập dữ liệu như sau:

```
> x$Age
 [1] "66" "77" "82" "79" "75" "77" "78" "83" "65" "88"
> x$FPSA
 [1]  0.673  0.435  1.060  1.215  3.607  1.712  1.618  0.371  5.996 11.056
> x$TPSA
 [1]  2.8  1.0  3.5  2.7 23.3  7.0  6.4  1.0 19.6 44.2
```

Đối tượng x là một ma trận có kích thước 10x3, cỡ thực hiện bởi lệnh dim

```
> dim(x)
[1] 10  3
```

Truy cập phần tử của ma trận x bằng lệnh x[i,j], i=1,2,…,10; j=1,2,3.

```
> x[1,2]
[1] 0.673
```

Lưu dữ liệu thành file data1.rda bằng lệnh save

```
> save(x, file="data1.rda")
```

Để đọc dữ liệu từ file .rda ta dùng lệnh **load()**.

```
> data<-load(file="data1.rda")
> data
[1] "x"
> x
   Age   FPSA TPSA
1   66  0.673  2.8
2   77  0.435  1.0
3   82  1.060  3.5
4   79  1.215  2.7
5   75  3.607 23.3
6   77  1.712  7.0
7   78  1.618  6.4
8   83  0.371  1.0
9   65  5.996 19.6
10  88 11.056 44.2
```

Cách 3: Nhập dữ liệu từ một file bằng lệnh: **read.table**

Ta có một file **data1.txt** trong "C:/Users/xxx/Documents", ta có thể dùng lệnh getwd() để kiểm tra thư mục đang làm việc

```
> getwd()
[1] "C:/Users/xxx/Documents"
```

Nếu file ta muốn đọc nằm ở một đường dẫn khác thì ta dùng lệnh **setwd()**

```
> dulieu<-read.table("data1.txt",header=TRUE)
> dulieu
   Age   FPSA TPSA
1   66  0.673  2.8
2   77  0.435  1.0
3   82  1.060  3.5
4   79  1.215  2.7
5   75  3.607 23.3
6   77  1.712  7.0
7   78  1.618  6.4
8   83  0.371  1.0
9   65  5.996 19.6
10  88 11.056 44.2
> names(dulieu)
[1] "Age"  "FPSA" "TPSA"
```

Để đọc các file dữ liệu dạng csv ta dùng lệnh **read.csv()**

```
> data<-read.csv("data1.csv")
> data
   Age   FPSA TPSA
1   66  0.673  2.8
2   77  0.435  1.0
3   82  1.060  3.5
4   79  1.215  2.7
5   75  3.607 23.3
6   77  1.712  7.0
```

```
7   78  1.618  6.4
8   83  0.371  1.0
9   65  5.996 19.6
10  88 11.056 44.2
```

Để đọc các file excel thì chúng ta cần cài đặt thêm thư viện readxl bằng lệnh

```
> install.packages("readxl")
> library(readxl)
> data1<-read_excel('data1.xlsx')
> data1
# A tibble: 10 x 3
    Age   FPSA  TPSA
  <dbl>  <dbl> <dbl>
 1   66  0.673   2.8
 2   77  0.435   1
 3   82  1.06    3.5
 4   79  1.22    2.7
 5   75  3.61   23.3
 6   77  1.71    7
 7   78  1.62    6.4
 8   83  0.371   1
 9   65  6.00   19.6
10   88 11.1    44.2
```

Với các file excel có đuôi là xls hoặc xlsx ta dùng lệnh **read_excel()** như trên.

1.6 Bài tập thực hành

Bài 1: Thực hiện các phép toán sau đây trên R:

a) $\frac{1-2*3}{4+5*6}$ **b)** $\sqrt{10^3}$

Bài 2: Nhập dãy số liệu sau đây bằng lệnh **c()**

74 122 235 111 292 111 211 133 156 79

Sử dụng hàm length để tính số quan sát.

Bài 3: Cho file dữ liệu Bảng điểm có điểm của 10 SV. Sử dụng hàm **edit(data.frame())** để nhập dữ liệu trên và ghi lại file dữ liệu với tên bangdiem dưới dạng đuôi rda (bangdiem.rda). Sau đó áp dụng hàm load dữ liệu.

a) Tính điểm thi KTHP của các sinh viên trên với Điểm KTHP=0.5*DQT+0.5*THI.

b) Lọc các sinh viên có điểm thi lớn hơn hoặc bằng 4.

c) Lọc các sinh viên có điểm thi từ (>=) 5.5 đến 8.4 (<8.5).

STT	DQT	THI
1	8	2.5
2	8.5	4
3	6.5	0.5
4	7	1
5	4	0
6	1	1
7	5	4
9	10	5
10	8	6

Bài 4: Trong phần dữ liệu thực hành có file dữ liệu Boston.csv. Sử dụng hàm read.csv để đọc file dữ liệu đã cho.

a) Hãy cho biết kích thước của dữ liệu.

b) Sử dụng hàm names để liệt kê các thuộc tính của dữ liệu.

c) Sử dụng hàm save để ghi lại dữ liệu dưới dạng file Boston.rda.

BÀI 2: THỐNG KÊ MÔ TẢ

A. Mục tiêu

Sinh viên thực hành được trên phần mềm R các nội dung sau

- Lập bảng phân bố tần số, tần suất
- Vẽ đa giác tần số/tần suất
- Vẽ biểu đồ cột tần số/tần suất (còn gọi là biểu đồ thanh) cho dữ liệu một chiều và nhiều chiều
- Phân nhóm dữ liệu. Vẽ biểu đồ histogram tần số/tần suất của dữ liệu phân nhóm
- Vẽ biểu đồ hình tròn
- Tính toán các số đặc trưng của dữ liệu. Vẽ biểu đồ hộp và râu.

B. Một số hàm trong R

1. Bảng phân bố tần số, tần suất

- Trong R, để tính đươc tần số ta dùng hàm **table** với các tham số sau

table(x, exclude)

trong đó

x	Véc tơ dữ liệu cần tính tần số của các phần tử
exclude	Tham số chỉ những phần tử không tham gia vào quá trình tính tần số, mặc định **exclude = c(NA, NaN)**, tức là không tính tần số những dữ liệu trống **NA(Not available)** và những dữ liệu không phải là số **NaN (Not a number)**

Đơn giản nhất ta dùng hàm với cấu trúc table(x).

- Trong R, để tính đươc tần suất ta dùng hàm sau

prop.table(x, margin)

trong đó

x	Véc tơ dữ liệu cần tính tần suất của các phần tử
margin	Tham số chỉ cách tính tần suất trong bảng dữ liệu hai chiều. Nếu **margin=1** thì tính tần suất các phần tử trên mỗi hàng, nếu **margin=2** thì tính tần suất các phần tử trên mỗi cột. Mặc định margin =NULL, tức là tính tần suất trên tổng số phần tử trong bảng dữ liệu.

Với dữ liệu một chiều ta chỉ cần dùng hàm với cấu trúc prop.table(x).

2. Đa giác tần số/tần suất- Biểu đồ cột

- Trong R, để vẽ đa giác tần số/tần suất ta dùng hàm **plot** với một số tham số cơ bản của hàm như sau

plot(x, type)

trong đó

x	véc tơ dữ liệu dùng để vẽ đa giác tần số/tần suất
type	miêu tả kiểu vẽ. Với đa giác tần số/tần suất thường type= "b": dạng các điểm được nối bằng đoạn thẳng

Ngoài ra ta có thể thêm một số tham số khác nữa tạo hiệu ứng cho hình vẽ đẹp hơn như:

main	tiêu đề của hình vẽ
xlab, ylab	tên của trục nằm ngang và trục thẳng đứng
col	màu của đa giác

Đơn giản nhất ta dùng cấu trúc plot(x, type= "b").

• Để vẽ biểu đồ cột tần số/tần suất (còn gọi là biểu đồ thanh) trong R ta dùng hàm **barplot** với cấu trúc của hàm như sau

barplot(x)

trong đó **x** là véc tơ dữ liệu dùng để vẽ biểu đồ.
Ngoài ra ta có thể điều chỉnh thêm một số tham số cho hàm **barplot** để biểu đồ được đẹp hơn, chẳng hạn

col: màu của các cột
border: màu của đường biên các cột
main: tên của biểu đồ
xlab, ylab: tên trục x, y
xlim, y lim: giới hạn trên các trục

• **Trong trường hợp dữ liệu quan sát nhiều biến độc lập** ta sử dụng biểu đồ cột nhiều chiều với hàm **barplot** nhưng thường thêm một số tham số sau để chú thích các thông tin của dữ liệu trên biểu đồ

barplot(x, names.arg, legend.text, beside, horiz)

trong đó

x	Ma trận dữ liệu dùng để vẽ biểu đồ
names.arg	Tên viết dưới nhóm các thanh trong biểu đồ
legend.text	Véc tơ gồm các kí tự hoặc dạng logic dùng để ghi chú thích trong biểu đồ
beside	Dạng logic, nếu **beside = FALSE** thì các thanh của biểu đồ được vẽ chồng lên nhau, nếu **beside = TRUE** thì các thanh được vẽ cạnh nhau. Mặc định **beside = FALSE**
horiz	Dạng logic, nếu **horiz= FALSE** thì các thanh được vẽ vuông góc với trục nằm ngang với thanh đầu tiên ở bên trái, nếu **horiz= TRUE** thì các thanh được vẽ song song với trục nằm ngang với thanh đầu tiên nằm ở dưới cùng. Mặc định **horiz= FALSE**

Ngoài ra, tương tự như đối với biểu đồ cột một chiều, ta có thể điều chỉnh thêm một số tham số cho hàm **barplot** để biểu đồ được đẹp hơn như: **col, border, main, xlim, y lim**, ….

3. Phân nhóm dữ liệu- Biểu đồ histogram

- **Phân nhóm dữ liệu**
  - Đối với những tập dữ liệu có quá nhiều giá trị khác nhau, người ta tiến hành phân nhóm dữ liệu. Kĩ thuật này chỉ có thể áp dụng cho dữ liệu số.
  - Đầu tiên, chúng ta xác định các khoảng chia không lồng nhau, nhưng che phủ tất cả các giá trị quan sát. Sau đó đếm số giá trị nằm trong mỗi khoảng chia. Trong tài liệu này, khi phân nhóm dữ liệu, chúng ta áp dụng quy tắc **cận trái đúng**, tức là một giá trị của dữ liệu bằng với cận trái của một khoảng chia thì sẽ nằm trong khoảng đó.
  - Số khoảng chia có thể được cho trước hoặc xác định theo công thức, chẳng hạn công thức Sturge

$$\mathbf{K= 1+ 3,3}\log_{10} n \quad \textbf{hoặc K= 1+ } \log_2 n \ .$$

- Độ rộng của khoảng chia và các điểm chia cũng có thể được cho trước hoặc xác định theo công thức

độ rộng = (giá trị lớn nhất- giá trị nhỏ nhất)/số khoảng chia.

- Để chia khoảng dữ liệu trong R ta dùng hàm **cut** có cấu trúc như sau

cut(x, breaks, right)

trong đó

x	Là véc tơ dữ liệu dạng số cần được phân nhóm
breaks	Véc tơ số (ít nhất hai tọa độ) gồm các điểm chia hoặc là một số nguyên dương chỉ số khoảng chia (số nhóm)
right	dạng logic, nếu **right = TRUE** thì khoảng chia có dạng (a, b], nếu **right = FALSE** thì khoảng chia có dạng [a, b), mặc định **right= TRUE**

Chú ý rằng trong tài liệu này áp dụng quy tắc **cận trái đúng nên tham số right = FALSE** .

- **Biểu đồ histogram**

Biểu đồ histogram chính là biểu đồ phân phối tần số/tần suất của dữ liệu được phân nhóm. Để vẽ biểu đồ histogram ta dùng hàm **hist** với một số tham số cơ bản sau

hist(x, breaks, freq, right)

trong đó

x	là véc tơ dữ liệu dạng số cần được phân nhóm
freq	dạng logic, nếu **freq = TRUE** các cột của biểu đồ mô tả tần số, nếu **freq = FALSE** các cột của biểu đồ mô tả tần suất. Mặc định **freq = TRUE**
breaks	Véc tơ số (ít nhất hai tọa độ) gồm các điểm chia giữa các cột hoặc là một số nguyên dương chỉ số cột của biểu đồ
right	dạng logic, nếu **right = TRUE** thì các cột lấy các phần tử trong khoảng dạng (a, b], nếu **right = FALSE** thì khoảng dạng [a, b), mặc định **right= TRUE**

Chú ý:

- Vì chúng ta áp dụng quy tắc **cận trái đúng nên tham số right = FALSE** .
- Để điều chỉnh biểu đồ đẹp hơn ta có thể thêm một số tham số như bảng sau:

col	màu của các cột
border	màu đương biên của các cột
main, xlab, ylab	tên của biểu đồ, tên trục x, y
xlim, ylim	Giới hạn trên các trục
labels	dạng logic hoặc dạng kí tự điền tên trên đỉnh mỗi cột

….

4. Biểu đồ hình tròn

Để vẽ biểu đồ hình tròn trong R ta dùng hàm **pie** với cấu trúc đơn giản nhất là

pie(x)

Để biểu đồ đẹp hơn và thể hiện được các thông tin của dữ liệu trên biểu đồ chúng ta thường thêm một số tham số sau cho hàm **pie**

pie(x, labels, col, border, lty, main, sub)

trong đó

labels	Tên của những hình quạt trong biểu đồ
col	Màu của các hình quạt
border	Màu của đường ranh giới giữa các hình quạt
lty	Kiểu nét vẽ của đường ranh giới giữa các rẻ quạt: 1: liền nét, 2:nét, 3:chấm, 4:chấm nét, 5:nét dài, 6:hai nét
main, sub	Tiêu đề và tiêu đề phụ của biểu đồ

5. Tính toán các số đặc trưng của mẫu thực nghiệm

- Dưới đây là bảng giới thiệu một số hàm trong R để tính toán một số số đặc trưng của mẫu thực nghiệm.

Hàm	Chức năng
mean(x)	Tính trung bình cộng của các giá trị cho trong véc tơ x (chính là trung bình mẫu $\bar{x}$
median(x)	Tính trung vị của các giá trị cho trong véc tơ x
which(table(x) == max(table(x)))	Tìm các giá trị mode (các giá trị có tần số lớn nhất) của các giá trị cho trong véc tơ x và vị trí của các giá trị mode này trong table(x)
range(x)	Cho giá trị nhỏ nhất, giá trị lớn nhất của các giá trị cho trong véc tơ x
var(x)	Tính phương sai của các giá trị cho trong véc tơ x
sd(x)	Tính độ lệch chuẩn của các giá trị cho trong véc tơ x
summary(x)	Cho giá trị nhỏ nhất, giá trị lớn nhất, giá trị trung bình, các tứ phân vị của các giá trị cho trong véc tơ x

- **Biểu đồ hộp và râu**

Biểu đồ hộp và râu giúp ta minh họa các tham số: trung bình, tứ phân vị và các giá trị ngoại biên trên cùng hình vẽ. Trong R, để vẽ biểu đồ hộp và râu ta dùng hàm boxplot với cấu trúc đơn giản nhất là

boxplot(x)

Để điều chỉnh biểu đồ đẹp hơn ta có thể thêm một số tham số như:

boxplot(x, names, border, col, main, xlab, ylab, xlim, ylim)

names	Tham số ghi chú thích tên dưới mỗi biểu đồ
col	màu của hộp
border	màu của râu, đường biên của hộp và giá trị ngoại biên
main, xlab, ylab	tên của biểu đồ, tên trục x, y
xlim, ylim	Giới hạn trên các trục

C. Hướng dẫn thực hành qua một số ví dụ

Ví dụ 1: Dữ liệu định lượng một chiều

Dữ liệu sau đây về cường độ bê tông (đơn vị: ksi) được thu thập dựa vào phương pháp kiểm tra không phá hủy bằng sóng siêu âm tại một số vị trí của một công trình:
4.5, 4.2, 4.1, 4.5, 4.6, 4.2, 4.4, 4.9, 4.1, 4.6, 4.3, 4.5, 4.9, 4.8, 4.7, 4.4, 4.6, 4.5, 4.5, 4.7, 4.6, 4.8, 4.2, 4.4, 4.2, 4.6, 4.1, 4.9, 4.5, 4.5, 4.4, 4.2, 4.7, 4.8, 4.4, 4.6, 4.5, 4.2, 4.6, 4.8.
Thực hành trên R các nội dung sau:

- Lập bảng phân bố tần số, tần suất của dữ liệu.
- Vẽ đa giác tần số, biểu đồ cột tần số.

Hướng dẫn các bước làm:

• Bước 1: Nhập dữ liệu dưới dạng véc tơ hoặc nhập dữ liệu vào exel (lưu ý lưu tệp dưới dạng đuôi .csv) và đọc tệp này từ R bằng hàm read.csv. Nên lưu lại dữ liệu vào một thư mục nào đó.

```
> #Nhập dữ liệu dưới dạng véc tơ
> cuongdobetong <- c(4.5, 4.2, 4.1, 4.5, 4.6, 4.2, 4.4, 4.9, 4.1, 4.6, 4.3, 4.5, 4.9, 4.8, 4.7, 4.4, 4.6, 4.5, 4.5, 4.7, 4.6, 4.8, 4.2, 4.4, 4.2, 4.6, 4.1, 4.9, 4.5, 4.5, 4.4, 4.2, 4.7, 4.8, 4.4, 4.6, 4.5, 4.2, 4.6, 4.8)
> #Chọn thư mục để lưu dữ liệu
> setwd("F:/ThuchanhR")
> #Lưu dữ liệu vừa nhập vào thư mục này
> save(cuongdobetong, file = "cuongdobetong.rda")
```

Hoặc nhập dữ liệu vào exel lưu tên tệp là cuongdobetong.csv, chẳng hạn (lưu ý lưu tệp dưới dạng đuôi .csv) vào ổ F, thư mục ThuchanhR và đọc tệp này từ R bằng hàm read.csv.

```
> cuongdobetong <- read.csv("cuongdobetong.csv")
```

• Bước 2: Dùng hàm table để tính tần số của các giá trị

```
> #Tính tần số của các giá trị
> table(cuongdobetong)
cuongdobetong
4.1 4.2 4.3 4.4 4.5 4.6 4.7 4.8 4.9
  3   6   1   5   8   7   3   4   3
```

• Bước 3: Dùng hàm prop.table để tính tần suất của các giá trị

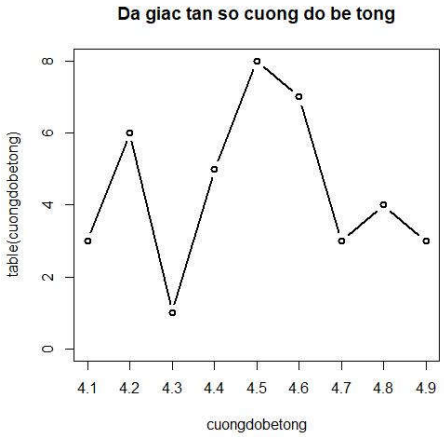
```
> #Tính tần suất các giá trị
> prop.table(table(cuongdobetong))
cuongdobetong
  4.1   4.2   4.3   4.4   4.5   4.6   4.7   4.8   4.9
0.075 0.150 0.025 0.125 0.200 0.175 0.075 0.100 0.075
```

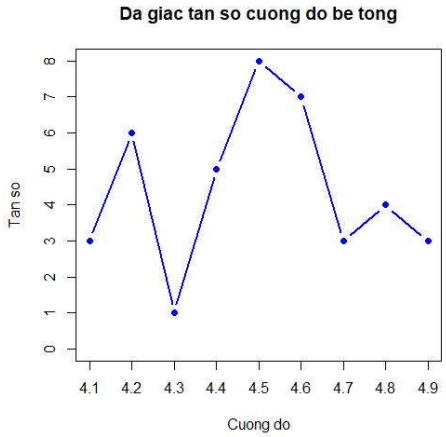
Chú ý rằng không dùng lệnh prop.table(cuongdobetong) (vì tính tần suất dựa vào bảng tần số)

• Bước 4: vẽ đa giác tần số

```
># vẽ đa giác tần số
> plot(table(cuongdobetong), type = "b", main = "Da giac tan so cuong do be tong")
># vẽ đa giác tần số có điều chỉnh
> plot(table(cuongdobetong), type = "b", main = "Da giac tan so cuong do be tong", col = "blue1", pch = 16, xlab = "Cuong do", ylab = "Tan so")
```

```
>#Chỉ rõ các điểm chia trên trục đứng
> axis(side = 2, c(0, 1, 2, 3, 4, 5, 6, 7, 8, 9))
```



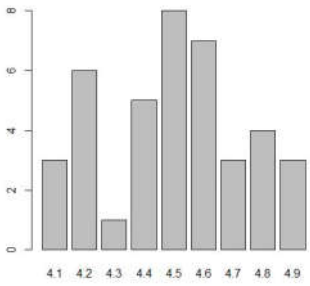


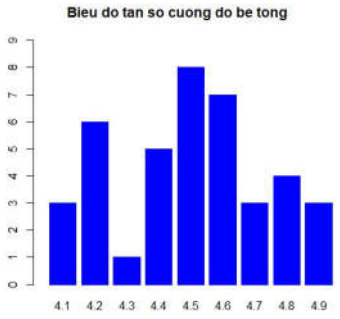
Hình 1.1: Đa giác tần số chưa điều chỉnh

Hình 1.2: Đa giác tần số có điều chỉnh

- Bước 5: vẽ biểu đồ cột

```
>#Sử dụng hàm barplot với cấu trúc đơn giản nhất
> barplot(table(cuongdobetong))
>#Điều chỉnh thêm một số tham số để biểu đồ đẹp hơn
> barplot(table(cuongdobetong), main = "Bieu do tan so cuong do be tong", col
= "blue1", border = "blue1", xlim = c(0,10), ylim = c(0, 9))
> axis(side = 2, c(0, 1, 2, 3, 4, 5, 6, 7, 8, 9))
```





Hình 1.3: Biểu đồ cột chưa điều chỉnh

Hình 1.4: Biểu đồ cột đã điều chỉ

Ví dụ 2: Dữ liệu định tính

Kết quả xếp loại học lực của 15 sinh viên lớp Kỹ sư tài năng chuyên ngành Cầu đường như sau

Khá, Giỏi, Khá, Khá, Trung bình, Xuất sắc, Khá, Giỏi, Trung bình, Xuất xắc, Khá, Khá, Trung bình, Giỏi, Giỏi.

Hãy lập bảng phân tần số, tần suất cho học lực của 15 sinh viên này.

Các bước làm

- Nhập dữ liệu dưới dạng véc tơ và lưu dữ liệu

```
> hocluc <- c("Kha", "Gioi", "Kha", "Kha", "Trung binh", "Xuat sac", "Kha", "Gioi",
"Trung binh", "Xuat sac", "Kha", "Kha", "Trung binh", "Gioi", "Gioi")
> #Chọn thư mục để lưu dữ liệu
```

```
> setwd("F:/ThuchanhR")
> #Lưu dữ liệu vừa nhập vào thư mục này
> save(hocluc, file = "hocluc.rda")
```

- Tính tần số

```
> table(hocluc)
hocluc
     Gioi       Kha  Trung binh  Xuat sac
        4         6           3         2
```

- Tính tần suất

```
> prop.table(table(hocluc))
hocluc
     Gioi       Kha  Trung binh  Xuat sac
0.2666667 0.4000000 0.2000000 0.1333333
```

Nếu muốn lấy số chữ số sau dấu phẩy theo ý muốn ta dùng hàm **round**

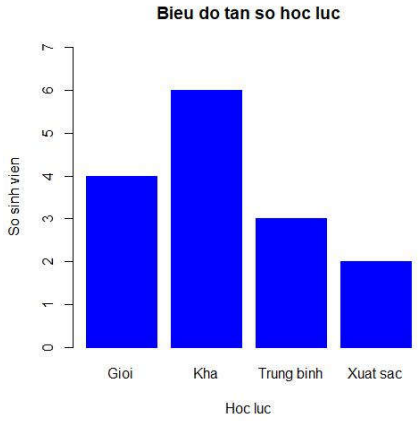
```
> #Tính tần suất, lấy 4 chữ số sau dấu phẩy
> round(prop.table(table(hocluc)), digits=4)
hocluc
     Gioi       Kha  Trung binh  Xuat sac
    0.2667    0.4000    0.2000    0.1333
```

- Vẽ biểu đồ tần số

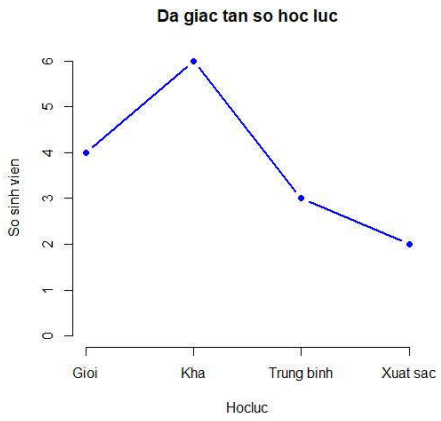
```
> barplot(table(hocluc), main = "Bieu do tan so hoc luc", col = "blue1", border = "blue1", ylim = c(0, 7), xlab = "Hoc luc", ylab= "So sinh vien")
```

- Vẽ đa giac tần số học lực

```
> plot(table(hocluc), type = "b", main = "Da giac tan so hoc luc", col = "blue1", pch = 16, xlab = "Hocluc", ylab = "So sinh vien")
```



Hình 2.1: Biểu đồ tần số học lực



Hình 2.2: Đa giác tần số học lực

Ví dụ 3: dữ liệu nhiều chiều

Hỏi ngẫu nhiên 10 người thuộc một số ngành nghề khác nhau về mức thu nhập hàng tháng và số năm kinh nghiệm được bảng số liệu sau

STT	Nghề nghiệp	Số năm kinh nghiệm	Mức thu nhập hàng tháng
1	LTV	2	18

2	KSCĐ	3	15
3	ĐD	2	10
4	NVVP	5	8
5	LTV	3	25
6	NVVP	2	7
7	ĐD	3	15
8	LTV	3	30
9	NVVP	5	10
10	LTV	4	50

(LTV: lập trình viên, KSCĐ: kĩ sư cầu đường, ĐD: điều dưỡng, NVVP: nhân viên văn phòng)

Lập bảng phân tần số, tần suất cho cột dữ liệu về nghề nghiệp, số năm kinh nghiệm.

Các bước làm

- Trước hết cần nhập dữ liệu dưới dạng bảng (**data.frame**) hoặc nhập dữ liệu vào exel (lưu ý lưu tệp dưới dạng đuôi **.csv**) và đọc tệp này từ R bằng hàm **read.csv**. Ở đây đã nhập dữ liệu vào exel và lưu tệp là **Mucthunhap.csv**

```
> #Truy cập vào thư mục chứa dữ liệu
> setwd("F:/ThuchanhR")
> #Đọc dữ liệu từ tệp Mucthunhap.csv và lưu vào đối tượng mucthunhap
>  mucthunhap <- read.csv("Mucthunhap.csv")
```

Điểm khác so với trường hợp dữ liệu một chiều xét ở hai mục trên là: **"Nghenghiep", "Kinhnghiem"** là các cột trong đối tượng dữ liệu **mucthunhap** chứ không phải là một đối tượng đang được làm việc trực tiếp. Do vậy, trước hết chúng ta phải dẫn cho R biết được chúng ta muốn xử lý cột **"Nghenghiep", "Kinhnghiem"** thuộc đối tượng dữ liệu nào. Muốn vậy ta dùng một số cách sau

- **Dùng hàm attach. Hàm attach giúp ta truy cập vào cột dữ liệu trong một bảng dữ liệu bằng tên cột.**

```
> #Dẫn cho R biết chúng ta muốn xử lý dữ liệu "mucthunhap"
> attach(mucthunhap)
The following objects are masked from mucthunhap (pos = 3):
    Kinhnghiem, Mucthunhap, Nghenghiep
```

- Hoặc có thể sử dụng toán tử $

```
> mucthunhap$Nghenghiep
 [1] LTV  KSCD DD   NVVP LTV  NVVP DD   LTV  NVVP LTV
Levels: DD KSCD LTV NVVP
```

- Tính tần số bằng hàm **table**

```
> #Tính tần số nghề nghiệp
```

```
> table(Nghenghiep)
Nghenghiep
 DD  KSCD  LTV  NVVP
  2     1    4     3
> #Tính tần số kinh nghiệm
> table(Kinhnghiem)
Kinhnghiem
2 3 4 5
3 4 1 2
```

- Tính tần suất bằng hàm **prop.table**

```
> prop.table(table(Nghenghiep))
Nghenghiep
 DD  KSCD  LTV  NVVP
0.2   0.1  0.4   0.3
> prop.table(table(Kinhnghiem))
Kinhnghiem
  2   3   4   5
0.3 0.4 0.1 0.2
```

<u>Tham khảo thêm</u>: với bảng nhiều chiều, ta có thể lập bảng tần số hai chiều (tính tần số chéo giữa hai cột dữ liệu)

```
> setwd("F:/ThuchanhR")
> mucthunhap = read.csv("Mucthunhap.csv")
> #Tính tần số cheo giữa kinh nghiệm và mức thu nhập
> attach(mucthunhap)
> table(Kinhnghiem, Mucthunhap)
           Mucthunhap
Kinhnghiem  7 8 10 15 18 25 30 50
         2  1 0  1  0  1  0  0  0
         3  0 0  0  2  0  1  1  0
         4  0 0  0  0  0  0  0  1
         5  0 1  1  0  0  0  0  0
```

```
>#Tính tần số chéo của mức thu nhập theo kinh nghiệm, làm tròn đến 2 chữ số thập phân

> round(prop.table(table(Kinhnghiem, Mucthunhap), margin = 1), digits = 2)

>#Có thể dùng cấu trúc ngắn gọn hơn: round(prop.table(table(Kinhnghiem, Mucthunhap), 1), dig = 2)
```

	Mucthunhap							
Kinhnghiem	7	8	10	15	1 8	25	30	50
2	0.33	0.00	0.33	0.00	0.33	0.00	0.00	0.00
3	0.00	0.00	0.00	0.50	0.00	0.25	0.25	0.00
4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00
5	0.00	0.50	0.50	0.00	0.00	0.00	0.00	0.00

Tức là, chẳng hạn, trong số những người có kinh nghiệm 3 năm thì có 50% số người có mức thu nhập 15 triệu, 25% số người có mức thu nhập 25 triệu và 25% số người có mức thu nhập 30 triệu.

```
> round(prop.table(table(Kinhnghiem, Mucthunhap), margin = 2), digits = 2)
```

	Mucthunhap							
Kinhnghiem	7	8	10	15	18	25	30	50
2	1.0	0.0	0.5	0.0	1.0	0.0	0.0	0.0
3	0.0	0.0	0.0	1.0	0.0	1.0	1.0	0.0
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
5	0.0	1.0	0.5	0.0	0.0	0.0	0.0	0.0

Có nghĩa là, chẳng hạn, trong số những người có mức thu nhập 10 triệu đồng thì có 50% số người có kinh nghiệm 2 năm, còn lại 50% là những người có kinh nghiệm 5 năm.

Nhắc lại rằng:

- Tham số `margin = 1` tức là tính tần suất các phần tử trên mỗi hàng; `margin =2` tức là tính tần suất các phần tử trên mỗi hàng.
- Hàm `round(x, digits = k)` : tính số làm tròn của các số trong véc tơ x đến k chữ số thập phân.

Ví dụ 4: Phân nhóm dữ liệu

Thời gian (tính bằng giây) cần thiết để công nhân hoàn thành một mối hàn trong một nhà máy lắp ráp ô tô được ghi lại dưới đây:

69	60	75	74	68	66	73	76	63	67
69	73	65	61	73	72	72	65	69	70

64	61	74	76	72	74	65	63	69	73
75	70	60	62	68	74	71	73	68	67

Thực hành trên R các nội dung sau:

- Xác định số khoảng chia K theo công thức Sturge.
- Xác định độ rộng của khoảng chia, các khoảng chia.
- Lập bảng phân bố tần số và tần suất theo kiểu chia khoảng cho dữ liệu này.
- Vẽ biểu đồ histogram tần số.

Lời giải:

- Tạo dữ liệu dạng véc tơ

```
> thoigianhan <- c(69, 60, 75, 74, 68, 66, 73, 76, 63, 67, 69, 73, 65, 61, 73, 72, 72, 65,
69, 70,64, 61, 74, 76, 72, 74, 65, 63, 69, 73, 75, 70, 60, 62, 68, 74, 71, 73, 68, 67)
```

- Xác định số khoảng chia

```
> #Xác định số quan sát
> n <- length(thoigianhan)
> n
[1] 40
>#Xác định số khoảng chia
> K <- 1+3.3*log(40, base=10)
> K
[1] 6.286798
>#Hoặc xác định K bởi công thức
> K <- 1+log(40, base=2)
> K
[1] 6.321928
```

Như vậy có thể chọn số khoảng chia là K=6.

- Xác định độ rộng của khoảng chia:

độ rộng = (giá trị lớn nhất- giá trị nhỏ nhất)/số khoảng chia.

```
> min(thoigianhan)
[1] 60
> max(thoigianhan)
[1] 76
> dorong <- (76-60)/6
> dorong
[1] 2.666667
```

Để thuận tiện ta chọn độ rộng của khoảng chia là 3. Dẫn đến, ta có thể chọn các khoảng chia như sau

[60, 63), [63, 66), [66, 69), [69, 72), [72, 75), [75, 78).

```
> thoigian <- cut(thoigianhan, breaks = c(60,63,66, 69, 72, 75, 78), right = FALSE)
```

● Tính tần số, tần suất các khoảng thời gian

```
>#Tính tần số các khoảng thời gian
> table(thoigian)
thoigian
[60,63) [63,66) [66,69) [69,72) [72,75) [75,78)
      5       6       6       7      12       4
>#Tính tần suất các khoảng thời gian
> prop.table(table(thoigian))
thoigian
[60,63) [63,66) [66,69) [69,72) [72,75) [75,78)
  0.125   0.150   0.150   0.175   0.300   0.100
```

Lưu ý rằng có thể để hàm cut tự chia khoảng khi biết trước số khoảng chia. Tuy nhiên, dùng hàm cut(x, breaks= k) các điểm chia thường không đẹp, ta nên để điểm chia cụ thể trong tham số breaks của hàm.

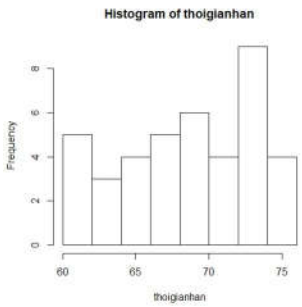
```
>#Không chọn khoảng chia trước mà để hàm cut tự chia khoảng khi biết trước số khoảng chia, sử dụng 6 khoảng chia.
 > table(cut(thoigianhan, breaks = 6, right= F))
 [60,62.7)  [62.7,65.3)  [65.3,68)  [68,70.7)  [70.7,73.3)  [73.3,76)
         5           6           3           9           9           8
```

● Vẽ biểu đồ histogram

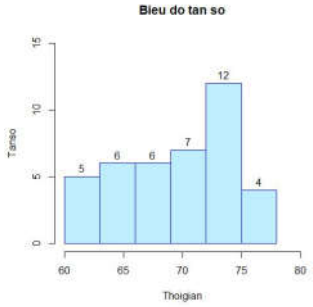
```
>#  Biểu đồ chưa điều chỉnh (Hình 4.1)
> hist(thoigianhan)
```

Ta có thể thêm các tham số để điều chỉnh biểu đồ tần số đẹp hơn, thể hiện được các thông tin số liệu trên biểu đồ như sau

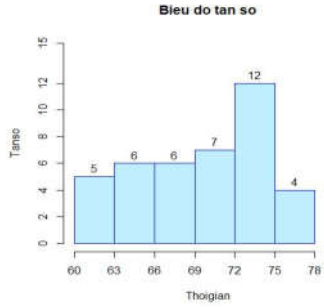
```
>#  Biểu đồ đã điều chỉnh (Hình 4.2)
>hist(thoigianhan, xlim  = c(60, 80),  ylim = c(0, 15), breaks = seq(60, 78, 3), right = F, xlab = "Thoigian", ylab = "Tanso", labels = T, main = "Bieu do tan so", col = "lightblue1", border = "blue1")
```



Hình 4.1



Hình 4.2



Hình 4.3

>#Thể hiện rõ các điểm chia trên các trục (Hình 4.3)

> hist(thoigianhan, xlim = c(60, 80), ylim = c(0, 15), breaks = seq(60, 78, 3), right = F, xlab = "Thoigian", ylab = "Tanso", xaxt = "n", yaxt = "n", labels = T, main = "Bieu do tan so", col = "lightblue1", border = "blue1")

> axis(side = 1, c(60, 63, 66, 69, 72, 75, 78))

> axis(side = 2, c(0, 2, 4, 6, 8, 10, 12, 15))

># Tham số xaxt = "n", yaxt = "n" tức là không vẽ trục x, trục y

Ví dụ 5: Biểu đồ cột nhiều chiều

Dữ liệu sau đây là về số lượng cầu được xây dựng trong các năm 1999, 2000, 2001 phân theo loại cầu

Loại cầu	**Số lương cầu được xây**		
	Năm 1999	**Năm 2000**	**Năm 2001**
Thép	5	10	12
Bê tông	10	6	7
Bê tông dự ứng lực	4	6	5
Tổng cộng	**19**	**22**	**24**

Hãy vẽ biểu đồ cột thể hiện dữ liệu trên.

Các bước làm:

>#Nhập dữ liệu dưới dạng matrix với 3 hàng

> Soluongcau = matrix(c(5, 10, 4, 10, 6, 6, 12, 7, 5), nrow = 3)

>#Biểu đồ cột chưa chú thích (Hình 5.1)

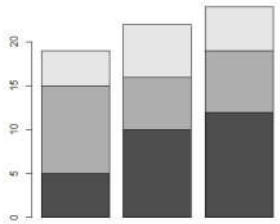
>barplot(Soluongcau)

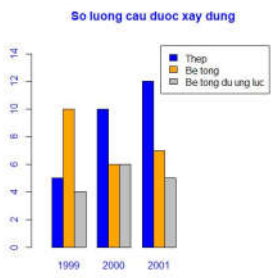
>#Biểu đồ cột đã điều chỉnh với các cột được vẽ cạnh nhau (Hình 5.2)

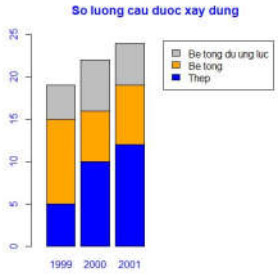
>barplot(Soluongcau, main = "So luong cau duoc xay dung", col = c("blue1", "orange", "gray"), names.arg = c(1999, 2000, 2001), beside = T, col.main = "blue", xlim = c(0,20), ylim = c(0,15), col.axis = "blue", legend.text = c("Thep", "Be tong", "Be tong du ung luc"))

>#Hoặc biểu đồ cột đã điều chỉnh với các cột được vẽ chồng lên nhau (Hình 5.3)

>barplot(Soluongcau, main = "So luong cau duoc xay dung", col = c("blue1", "orange", "gray"), names.arg = c(1999, 2000, 2001), beside = F, col.main = "blue", xlim = c(0,20), ylim = c(0,25), col.axis = "blue", legend.text = c("Thep", "Be tong", "Be tong du ung luc"))







Hình 5.1 **Hình 5.2** **Hình 5.3**

Ví dụ 6: Vẽ biểu đồ hình tròn

Thống kê số kilomet chiều dài đường bộ của Việt Nam tính đến năm 2013 như sau:

Loại đường	Số kilomet
Đường nhựa	108023
Đường đá	6509
Đường cấp phối	48555
Đường đất	48409

Vẽ biểu đồ hình tròn thể hiện chiều dài đường bộ theo phân loại đường.

Các bước làm:

>#Nhập dữ liệu dạng bảng

> chieudai <- c(108023, 6509, 48555, 48409)

> loaiduong <- c("Duongnhua", "Duongda", "Duongcapphoi", "Duongdat")

> Dulieuduong <- data.frame(chieudai, loaiduong)

>#Lưu dữ liệu vào ổ F thư mục ThuchanhR

> setwd("F:/ThuchanhR")

> save(Dulieuduong, file = "Dulieuduong.rda")

>#Tính t□ lệ phần trăm từng loại đường

> Tile <- round(prop.table(chieudai), 2)*100

> Tile

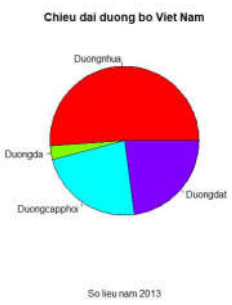
[1] 51 3 23 23

>#Vẽ biểu đồ hình tròn thể hiện t□ lệ các loai đường (Hình 6.1)

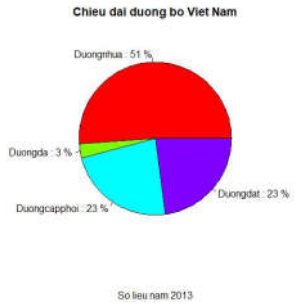
>pie(chieudai, labels = loaiduong, col = rainbow(4), lty = 1, main = "Chieu dai duong bo Viet Nam", sub = "So lieu nam 2013")

>#Thể hiện rõ số liệu t□ lệ phần trăm từng loại đường trên biểu đồ (Hình 6.2)

> pie(chieudai, labels = paste(loaiduong, ":", Tile, "%") , col = rainbow(4), lty = 1, main = "Chieu dai duong bo Viet Nam", sub = "So lieu nam 2013")



Hình 6.1



Hình 6.2

Ví dụ 7: Tính các số đặc trưng

Xét dữ liệu "**cuongdobetong**". Hãy tính các số đặc trưng của dữ liệu này.

```
>#Tính giá trị trung bình
> mean(cuongdo)
[1] 4.5
> #Tính trung vị
> median(cuongdo)
[1] 4.5
> #Tính mode
> which(table(cuongdo) == max(table(cuongdo)))
4.5
  5
>#Tìm giá trị nhỏ nhất, giá trị lớn nhất
> range(cuongdo)
[1] 4.1   4.9
>#Tính độ rộng của dữ liệu
> dorong <- 4.9 - 4.1
> dorong
[1] 0.8
```

```
>#Tính phương sai mẫu
> var(cuongdo)
[1] 0.05487179
>#Tính độ lệch chuẩn mẫu
> sd(cuongdo)
[1] 0.2342473
```

Như vậy ta tính được các số đặc trưng của dữ liệu như sau:

$$\bar{x} = 4,5; s^2 = 05487179, s = 0.2342473$$

Trung vị median = 4,5; mode= 4,5 và giá trị này ở vị trí thứ 5 trong bảng phân bố tần số.

min = 4,1; max= 4,9.

Ngoài ra, trong R ta có thể sử dụng hàm **summary(x)** để đồng thời tìm các giá trị min, max, median, và các tứ phân vị

```
>#Tìm các tứ phân vị, min, max, median
> summary(cuongdo)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  4.100   4.375   4.500   4.500   4.625   4.900
```

Tức là khoảng 50% vị trí có cường độ bê tông không quá 4,5 ksi, khoảng 25% vi trí có cường độ từ 4,625 ksi trở lên, khoảng 25% vị trí có cường độ không quá 4,375 ksi.

Hoặc dùng hàm **describe(x)** trong gói **psych**. Hàm **describe** cho rất nhiều số đặc trưng như kết quả dưới đây.

```
>library(psych)
> describe(cuongdo)
   vars  n mean  sd median trimmed  mad min max range skew kurtosis   se
X1    1 40  4.5 0.23    4.5     4.5 0.22 4.1 4.9   0.8 -0.07    -0.99 0,04
```

Nếu chỉ cần kết quả về giá trị trung bình, độ lệch chuẩn thì đặt các tham số **range = F, skew = F**

```
> describe(cuongdo, range = F, skew = F)
   vars  n mean  sd   se
X1    1 40  4.5 0.23 0.04
```

Ví dụ 8: Biểu đồ hộp và râu

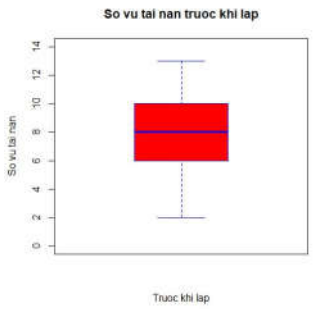
Số lượng các vụ tai nạn giao thông tại một điểm giao cắt quan sát trong khoảng thời gian 2 năm trước (A) và 2 năm sau (B) khi được lắp đặt các thiết bị kiểm soát giao thông được ghi lại dưới đây:

(A) 5, 2, 8 , 11, 7, 8, 5, 10, 6, 8, 9, 4, 6, 12, 7, 7, 10, 11, 6, 8, 13, 11, 7, 9
(B) 2, 0, 4, 3, 0, 1, 0, 4, 2, 1, 2, 2, 3, 0, 1, 5, 4, 2, 0, 2, 3, 1, 6, 1

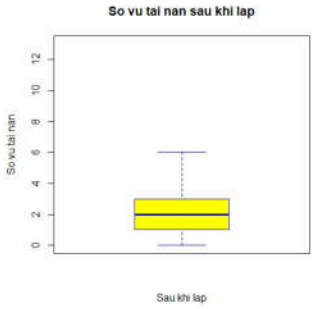
a) Xác định giá trị trung bình và các tứ phân vị của mỗi tập dữ liệu.
b) Vẽ biểu đồ hộp và râu cho hai tập dữ liệu trên.

Các bước làm:

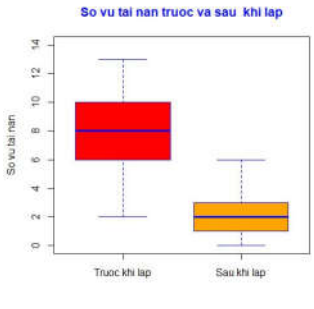
```
>#Nhập dữ liệu
> A = c(5, 2, 8 , 11, 7, 8, 5, 10, 6, 8, 9, 4, 6, 12, 7, 7, 10, 11, 6, 8, 13, 11, 7, 9)
> B = c(2, 0, 4, 3, 0, 1, 0, 4, 2, 1, 2, 2, 3, 0, 1, 5, 4, 2, 0, 2, 3, 1, 6, 1)
> Sovutainan <- data.frame(A, B)
>#Tìm trung bình, các tứ phân vị của dữ liệu A
> summary(A)
          Min.   1st Qu.  Median    Mean   3rd Qu.    Max.
          2.000   6.000   8.000     7.917   10.000   13.000
>#Tìm trung bình, các tứ phân vị của dữ liệu B
> summary(B)
           Min.   1st Qu.   Median    Mean   3rd Qu.    Max.
          0.000   1.000    2.000     2.042   3.000    6.000
>#Vẽ biểu đồ hộp râu của dữ liệu A (Hình 8.1)
>boxplot(A, main = "So vu tai nan truoc khi lap", border = "blue1", col = "red", horiz =
F, xlab = "Truoc khi lap", ylab = "So vu tai nan", ylim = c(0,13))

>#Vẽ biểu đồ hộp râu của dữ liệu B (Hình 8.2)

> boxplot(B, main = "So vu tai nan sau khi lap", border = "blue1", col = "yellow", horiz =
F, xlab = "Sau khi lap", ylab = "So vu tai nan", ylim = c(0,13))

>#Vẽ biểu đồ hộp râu của 2 tập dữ liệu trên cùng một biểu đồ để so sánh (Hình 8.3)
>boxplot(A, B, main = "So vu tai nan truoc va sau  khi lap", border = "blue1", col =
c("red", "orange"),horiz = F,  ylab = "So vu tai nan", ylim = c(0,14), names = c("Truoc khi
lap", "Sau khi lap"), col.main = "blue")
```



Hình 8.1



Hình 8.2



Hình 8.3

D. Bài tập thực hành ở nhà

Thực hành trên phần mềm R các bài tập sau

Bài 1: Một bảng mạch điện tử sẽ được sản xuất quy mô lớn với một quy trình sản xuất mới được đề xuất. Quy trình này được kiểm tra bằng cách sản xuất 20 bảng mạch, và số lỗi được

đếm trên mỗi bảng mạch. Số lỗi sau đây được quan sát trên những bảng mạch này: 0, 1, 3, 2, 1, 2, 2, 2, 3, 4, 1, 0, 0, 1, 0, 2, 0, 0, 2, 0.

a) Hãy lập bảng phân phối tần số và tần suất của số lỗi.
b) Vẽ biểu đồ cột tần số và đa giác tần số tương ứng với bảng phân phối tần số trên.
c) Vẽ biểu đồ cột cho tần suất của số lỗi.
d) Tính các số đặc trưng của dữ liệu.

Bài 2: Năng suất lao động (đơn vị: triệu đồng/người) được thống kê theo ngành kinh tế và theo năm được cho trong bảng dưới đây

Thành phần kinh tế	Năm 2005	Năm 2010	Năm 2015
Vận tải, kho bãi	21,7	43,8	71,9
Tài chính, ngân hàng và bảo hiểm	257,3	457,8	631,1
Kinh doanh bất động sản	3232,2	1300	1284,7
Giáo dục và đào tạo	21,4	30	72,1

Hãy vẽ biểu đồ cột biểu diễn năng suất lao động theo thành phần kinh tế và theo năm và cho nhận xét.

Bài 3: Dữ liệu sau đây là về chất thải rắn (đơn vị: triệu tấn) được thải ra môi trường hàng năm của một quốc gia:

Rác thải đô thị	150
Công nghiệp	350
Khai mỏ	1700
Nông nghiệp	2300

a) Hãy vẽ biểu đồ cột biểu diễn dữ liệu trên và cho nhận xét.

b) Hãy vẽ biểu đồ hình tròn biểu diễn dữ liệu trên và cho nhận xét.

c) So sánh thông tin có được khi dùng biểu đồ cột và biểu đồ hình tròn biểu diễn dữ liệu trên.

Bài 4: Một nhà máy xử lý nước cung cấp cho một khu vực dân cư được xây dựng với công suất thiết kế 17000 mét khối một ngày. Khi nhu cầu dùng nước vượt quá khả năng cung cấp, các hệ thống tưới tiêu công cộng sẽ bị dừng hoạt động. Người ta khảo sát nhu cầu nước (đơn vị: nghìn mét khối) trong một giai đoạn và kết quả được cho dưới đây:

8,7	12,1	12,6	13,7	14,8	15,1	15,4	15,9	16,2	16,5
16,8	16,8	17,1	17,2	17,3	17,4	17,6	17,7	17,7	17,9
18,0	18,1	18,2	18,2	18,4	18,5	18,6	18,6	18,6	18,7
18,9	19,1	19,1	19,1	19,5	19,5	19,5	20,2	21,0	16,6
17,9	18,9								

a) Xác định các tứ phân vị, các đặc trưng số về tâm, các đặc trưng số về sự phân tán của dữ liệu này. Giải thích ý nghĩa của các giá trị.

b) Vẽ biểu đồ hộp và râu để minh họa sự phân bố của tập dữ liệu.

c) Xác định số khoảng chia theo công thức Sturge, các điểm chia và vẽ biểu đồ histogram tần số. Tính tỉ lệ quan sát ở đó nhu cầu vượt quá khả năng cung cấp.

d) Vẽ biểu đồ histogram tần số, sử dụng 8 khoảng chia.

Bài 5: Để chọn một trong hai phương án đầu tư được đề xuất, một người thu thập dữ liệu về lợi nhuận của hai phương án đầu tư như dưới đây

Lợi nhuận của phương án đầu tư A					**Lợi nhuận của phương án đầu tư B**				
30,00	6,93	13,77	-8,55	-2,13	30,33	-34,75	30,31	24,30	-30,37
-13,24	22,42	-5,29	4,30	-18,95	54,19	6,06	-10,01	-5,61	44
34,40	-7,04	25	9,43	49,87	14,73	35,24	29	-20,23	36,13
-12,11	12,89	1,21	22,92	12,89	40,70	-26,01	4,16	1,53	22,18
-20,24	31,76	20,95	63	1,20	0,46	10,03	17,61	3,24	2,07
11,07	43,71	-19,27	-2,59	8,47	10,51	1,20	25,10	29,44	39,04
-12,83	-9,22	33,00	36,08	0,52	9,94	-24,24	11	24,76	-33,39
-17	14,26	-21,95	61	17,30	-38,47	-25,93	15,28	58,67	13,44
-15,83	10,33	-11,96	52	0,63	8,29	34,21	0,25	68	61
12,68	1,94	38	13,09	28,45	52	5,23	-20,44	-32,17	66

a) Tính giá trị trung bình, độ lệch chuẩn của mỗi tập dữ liệu và đưa ra nhận xét.
b) Hãy vẽ biểu đồ histogram tần số cho mỗi tập dữ liệu.
c) Phân tích biểu đồ và đưa ra kết luận về phương án đầu tư tốt hơn.

BÀI 3: ƯỚC LƯỢNG THAM SỐ

Trong bài học này, chúng ta sẽ làm quen với một số hàm trong R giúp tìm khoảng tin cậy của những tham số như: ***trung bình, tỷ lệ, phương sai.***

Chúng ta quan tâm đến các bài toán sau:

1. Khoảng tin cậy cho giá trị **trung bình**
   a. Khoảng tin cậy cho giá trị trung bình của phân phối chuẩn khi **biết phương sai**.
   b. Khoảng tin cậy cho giá trị trung bình của phân phối chuẩn khi **chưa biết phương sai**
      i. Khoảng tin cậy cho giá trị trung bình khi kích thước mẫu lớn $(n > 30)$
      ii. Khoảng tin cậy cho giá trị trung bình khi kích thước mẫu nhỏ $(n \le 30)$
2. Khoảng tin cậy cho giá trị **tỷ lệ**

❖ Một số gói lệnh cần cài đặt

Trong phần thực hành này, chúng ta cần cài đặt gói lệnh sau:

> install.packages("BSDA")

3.1 Khoảng tin cậy cho giá trị trung bình.

3.1.1 Khoảng tin cậy cho giá trị trung bình của phân phối chuẩn khi biết phương sai.

Bài toán 1: Giả sử X là biến ngẫu nhiên có phân phối chuẩn với giá trị trung bình là μ **chưa biết** và phương sai σ^2 **đã biết**. Hãy tìm khoảng tin cậy giá trị trung bình μ với độ tin cậy γ cho trước, biết một mẫu thực nghiệm của X là $(x_1, x_2, ..., x_n)$

Công thức: $\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ với $\gamma = 1 - \alpha$

- Khi phương sai đã biết, khoảng tin cậy cho trung bình được tìm qua hàm z.test với các tham số cần thiết được cho như sau

z.test (x, sigma x, conf.level)

trong đó

x	véc tơ dữ liệu mẫu.
sigma.x	độ lệch chuẩn.

Conf.level	số thuộc $[0,1]$ chỉ độ tin cậy của khoảng ước lượng, mặc định là 0.95.

Ví dụ 3.1.1a: Khối lượng của những chai nước của một dây chuyền đóng nước uống tinh khiết được giả sử là tuân theo phân phối chuẩn với độ lệch chuẩn là 10g. Để ước tính khối lượng của các chai nước trên dây chuyền, người ta chọn ngẫu nhiên ra 20 chai nước, đo khối lượng (gam) và được bảng dữ liệu sau

295	290	305	310	298
287	315	307	293	300
294	298	305	310	298
290	290	309	308	291

Tìm khoảng tin cậy 90% cho khối lượng trung bình của những chai nước sản xuất trên dây chuyền.

Lời giải:

- Các tham số cần thiết như sau:
  - x là dữ liệu về khối lượng các chai nước trong mẫu;
  - Độ lệch chuẩn của khối lượng các chai nước là 10g nên sigma.x = 10;
  - Độ tin cậy là 90% nên conf.level = 0.9.
- Thực hiện trên R:

```
> TrongLuong = scan ()
1: 295 290 305 310 298
6: 287 315 307 293 300
11: 294 298 305 310 298
16: 290 290 309 308 291
21:
Read 20 items
> z.test (TrongLuong, sigma.x=10, conf.level = 0.9)

        One-sample z-Test

data:  TrongLuong
z = 134.01, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
90 percent confidence interval:
 295.972 303.328
sample estimates:
mean of x
   299.65
```

- Khoảng tin cậy được tìm qua kết quả

90 percent confidence interval:
295.972 303.328

Như vậy ta có khoảng tin cậy 90% cho khối lượng trung bình của các chai nước sản xuất ra trên dây chuyền là [295.972, 303.328].

Bài toán 2: Giả sử X là biến ngẫu nhiên có phân phối chuẩn với giá trị trung bình là μ chưa biết và phương sai σ^2 đã biết. Hãy tìm khoảng tin cậy giá trị trung bình μ với độ tin cậy γ khi cho biết trung bình mẫu

Công thức: $\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ với $\gamma = 1 - \alpha$

- Khi phương sai đã biết, khoảng tin cậy cho trung bình được tìm qua hàm zsum.test với các tham số cần thiết được cho như sau

zsum.test (mean.x, sigma x, n.x, conf.level)

trong đó

mean.x	trung bình mẫu;
sigma.x	độ lệch chuẩn .
n.x	cỡ mẫu
Conf.level	số thuộc [0,1] chỉ độ tin cậy của khoảng ước lượng, mặc định là 0.95.

Ví dụ 3.1.1b: Mức lương tháng của trưởng phòng kinh doanh của các doanh nghiệp có qui mô trung bình tại thời điểm hiện tại được cho là tuân theo phân phối chuẩn với độ lệch chuẩn là 3.2 triệu. Người ta chọn ngẫu nhiên ra 100 trưởng phòng kinh doanh và thấy mức lương trung bình hàng tháng của nhóm là 16.5 triệu. Hãy xác định khoảng tin cậy 95% cho mức lương trung bình hàng tháng của các trưởng phòng kinh doanh.

Lời giải:

- Ta sử dụng các kết quả cho hàm zsum.test với các tham số cần thiết được cho cụ thể như sau:

- Lương của nhóm trưởng phòng trong mẫu là 16.5 triệu nên mean.x = 16.5;
- Độ lệch chuẩn của mức lương các trưởng phòng là 3.2 triệu nên sigma.x = 3.2;
- Mẫu điều tra 100 trưởng phòng nên n.x = 100
- Độ tin cậy là 95% nên conf.level = 0.95 hoặc ta không cần đưa tham số này vào (vì trùng với giá trị mặc định)

- Thực hiện trên R:

```
> library(BSDA)
> zsum.test(mean.x = 16.5, sigma.x = 3.2, n.x = 100, conf.level = 0.95)

        One-sample z-Test

data:  Summarized x
z = 51.562, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 15.87281 17.12719
sample estimates:
mean of x
     16.5
```

Theo kết quả đưa ra, ta có khoảng tin cậy 95% cho lương trung bình của các trưởng phòng kinh doanh là [15.87281, 17.12719]

3.1.2 Khoảng tin cậy cho giá trị trung bình của phân phối chuẩn khi chưa biết phương sai

Bài toán 1: Giả sử X là biến ngẫu nhiên có phân phối chuẩn với giá trị trung bình là μ chưa biết và phương sai σ^2 chưa biết. Hãy ước lượng giá trị trung bình μ với độ tin cậy γ cho trước, biết một mẫu thực nghiệm của X là $(x_1, x_2, ..., x_n)$

Khoảng tin cậy cho giá trị trung bình khi kích thước mẫu lớn $(n > 30)$

Công thức: $\bar{x} - z_{\alpha/2}\frac{s}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2}\frac{s}{\sqrt{n}}$ với $\gamma = 1 - \alpha$

- Khi phương sai chưa biết, khoảng tin cậy cho trung bình được tìm qua hàm t.test với các tham số cần thiết được cho như sau

t.test (x, conf.level)

trong đó:

x	véc tơ dữ liệu mẫu.
Conf.level	số thuộc $[0,1]$ chỉ độ tin cậy của khoảng ước lượng, mặc định là 0.95.

Ví dụ 3.1.2a: Kết quả kiểm tra độ bám dính trên 36 mẫu hợp kim U -700 có số liệu như sau

19.8 10.1 14.9 7.5 15.4 15.4 18.5 7.9 12.7 11.9 11.4 14.1 17.6 16.7 15.8 19.5 8.8 13.6 11.4 11.4 14.1 17.6 16.7 15.8 19.5 8.8 13.6 11.9 11.4 15.4 15.4 18.5 7.9 12.7 11.9 11.4

Với độ tin cậy 95%, hãy ước lượng độ bám dính trung bình của loại hợp kim trên. Biết rằng độ bám dính của hợp kim tuân theo phân phối chuẩn.

Lời giải:

- Ta dùng hàm t.test với các tham số cần thiết như sau:
  - x là véc tơ dữ liệu về độ bám dính trên 36 mẫu hợp kim U-700.
  - Độ tin cậy là 95% nên conf.level = 0.95
- Thực hiện trên R

```
> Dobamdinh = scan ()
1: 19.8 10.1 14.9 7.5 15.4 15.4 18.5 7.9 12.7 11.9 11.4 1
4.1 17.6 16.7 15.8 19.5 8.8 13.6
19: 11.4 11.4 14.1 17.6 16.7 15.8 19.5 8.8 13.6 11.9 11.4
 15.4 15.4 18.5 7.9 12.7 11.9 11.4
37:
Read 36 items
> t.test(Dobamdinh, conf.level = 0.95)

        One Sample t-test

data:  Dobamdinh
t = 23.754, df = 35, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 12.62566 14.98545
sample estimates:
mean of x
 13.80556
```

Theo kết quả đưa ra, ta có khoảng tin cậy 95% cho lương trung bình của các trưởng phòng kinh doanh là [12.6824, 14.9288]

Bài toán 2: Giả sử X là biến ngẫu nhiên có phân phối chuẩn với giá trị trung bình là μ chưa biết và phương sai σ^2 chưa biết. Hãy ước lượng giá trị trung bình μ với độ tin cậy γ cho trước, biết phương sai mẫu là s và trung bình mẫu là $\bar{x}$.

Công thức: $\bar{x} - z_{\alpha/2}\frac{s}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2}\frac{s}{\sqrt{n}}$ với $\gamma = 1 - \alpha$

- Khi phương sai chưa biết, khoảng tin cậy cho trung bình được tìm qua hàm tsum.test với các tham số cần thiết được cho như sau:

tsum.test (mean.x, s.x, n.x, conf.level)

trong đó

mean.x	trung bình mẫu
s.x	độ lệch chuẩn của mẫu
n.x	kích thước mẫu
Conf.level	số thuộc $[0,1]$ chỉ độ tin cậy của khoảng ước lượng, mặc định là 0.95.

Ví dụ 3.1.2b: Trong một cuộc khảo sát về năng khiếu học tập môn Toán của học sinh phổ thông ở một thành phố, người ta lấy một mẫu gồm 120 học sinh, cho trả lời các câu hỏi và tính được điểm trung bình của chúng là 501 điểm và độ lệch chuẩn của mẫu là 112. Hãy tìm khoảng tin cậy cho điểm năng khiếu môn Toán trung bình với độ tin cậy 95% của học sinh ở thành phố đó.

Lời giải:

- Với dữ liệu đã cho, ra sử dụng hàm tsum.test với các tham số cần thiết được cho cụ thể như sau:
  - Điểm trung bình là 501 nên mean.x = 501
  - Độ lệch chuẩn của mẫu là 112 nên s.x = 112
  - Mẫu thử nghiệm gồm 120 học sinh nên n.x = 120
  - Độ tin cậy là 95% nên conf.level = 0.95

- Thực hiện trên R

```
> library(BSDA)
> tsum.test(mean.x = 501, s.x = 112, n.x = 120, conf.leve
l = 0.95)

        One-sample t-Test

data:  Summarized x
t = 49.002, df = 119, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 480.7552 521.2448
sample estimates:
mean of x
      501
```

Theo kết quả đưa ra, ta có khoảng tin cậy 95% cho lương trung bình của các trưởng phòng kinh doanh là [480.7552, 521.2448]

Khoảng tin cậy cho giá trị trung bình khi kích thước mẫu nhỏ $(n \le 30)$

(Làm tương tự như khoảng tin cậy cho giá trị trung bình khi kích thước mẫu lớn (n>30))

3.2 Khoảng tin cậy cho giá trị tỷ lệ

Bài toán: Giả sử tỷ lệ (hay xác suất) gặp phần tử có dấu hiệu T nào đó trong tập chính. Khảo sát n phần tử từ tập chính, thấy có m phần tử có dấu hiệu T. Hãy ước lượng p với độ tin cậy γ cho trước.

Công thức:

$$f - z_{\alpha/2}\sqrt{\frac{f(1-f)}{n}} < p < f + z_{\alpha/2}\sqrt{\frac{f(1-f)}{n}}$$

- Ta sử dụng hàm prop.test với các tham số cụ thể như sau:

  Prop.test(x, n, conf.level, correct)

  Trong đó

  x số lần "thành công".

n	số lần thử nghiệm.
conf.level	số thuộc $[0,1]$ chỉ độ tin cậy của khoảng ước lượng, mặc định là 0.95.
correct	tham số dạng logic chỉ xem có hay không sự điều chỉnh liên tục Yate, mặc định là correct = TRUE (Tức là số liệu trong bài nếu thỏa mãn điều kiện $nf \geq 10, n(1-f) \geq 10$ thì là FALSE, không thỏa mãn là TRUE)

Ví dụ 3.2.1: Khảo sát một mẫu gồm 325 ổ trục quay động cơ ô tô, thấy có 74 ổ trục có bề mặt thô hơn so với thông số kỹ thuật cho phép. Hãy ước lượng khoảng tin cậy 95% cho tỷ lệ của ổ trục có bề mặt thô hơn thông số kỹ thuật cho loại động cơ ô tô này.

Lời giải:

- Khoảng tin cậy 95% cho tỉ lệ ổ trục có bề mặt thô hơn thông số kỹ thuật được ước tính qua hàm prop.test với các tham số cụ thể như sau:
- Có 74 ổ trục có bề mặt thô hơn nên x = 74
- Có 325 ổ trục trong mẫu điều tra nên n =325
- Do $n.f = 325.\frac{74}{325} = 74 \geq 10, n.(1-f) = 325.\left(1-\frac{74}{325}\right) = 251 \geq 10$ nên correct = F.
- Thực hiện trên R

```
> prop.test(x=74, n=325, correct = F, conf.level = 0.95)

        1-sample proportions test without continuity
        correction

data:  74 out of 325, null probability 0.5
X-squared = 96.397, df = 1, p-value < 2.2e-16
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.1854383 0.2763084
sample estimates:
        p
0.2276923
```

Vậy khoảng tin cậy 95% cho tỉ lệ tổng thể những ổ trục có bề mặt thô hơn thông số kỹ thuật là [0.1854383, 0.2763084]

BÀI TẬP

Bài 1: Độ bền kéo đứt (psi – pound per square inch) của sợi được sử dụng trong sản xuất vật liệu màn treo được yêu cầu tối thiểu 100. Giả sử lực kéo đứt sợi là một đại lượng ngẫu nhiên có phân phối chuẩn với độ lệch chuẩn bằng 2 psi. Kiểm tra ngẫu nhiên 9 màn treo ta thu được lực kéo đứt trung bình là 98 psi. Tìm khoảng tin cậy 95% cho lực kéo đứt trung bình.

Bài 2: Tuổi thọ (giờ) của một loại bóng đèn tuân theo quy luật phân phối chuẩn với độ lệch tiêu chuẩn 40 giờ. Chọn ngẫu nhiên 30 bóng đèn để thử nghiệm, thấy tuổi thọ trung bình mỗi bóng là 780 giờ. Hãy ước lượng tuổi thọ trung bình của loại bóng đèn trên với độ tin cậy 96%.

Bài 3: Bảy nhân viên giao hàng của một cửa hàng pizza được hỏi về số kilomet mà họ phải di chuyển trong một ngày làm việc. Kết quả được ghi lại dưới đây:

15.5 27.3 11.4 19.6 9.3 22.8 32.6

Hãy xác định khoảng tin cậy 95% cho quãng đường di chuyển trung bình trong ngày của nhân viên giao pizza biết quãng đường di chuyển là biến ngẫu nhiên tuân theo luật phân phối chuẩn.

Bài 4: Hao phí nguyên liệu cho một đơn vị sản phẩm là một đại lượng ngẫu nhiên X tuân theo luật phân phối chuẩn. Sản xuất thử 36 sản phẩm và thu được bảng số liệu

X	29-31	31-33	33-35	35-37	37-39
n_i	5	9	12	6	4

Bài 5: Người ta kiểm tra ngẫu nhiên 31 khoản vay thế chấp của một ngân hàng thương mại thấy số tiền vay trung bình là 24800 (đơn vị triệu đồng) với độ lệch tiêu chuẩn 650 triệu đồng. Tìm khoảng tin cậy cho số tiền thế chấp của ngân hàng trên với độ tin cậy 95% biết số tiền vay là biến ngẫu nhiên có luật phân phối chuẩn.

Bài 6: Kiểm tra ngẫu nhiên 400 người đi xe máy ở khu vực có 500.000 người đi xe máy thấy có 360 người có bằng lái. Với độ tin cậy 95%, hãy ước lượng số người đi xe máy có bằng lái trong khu vực.

BÀI 4. KIỂM ĐỊNH GIẢ THUYẾT THỐNG KÊ

Ngôn ngữ R cung cấp rất nhiều hàm kiểm định giúp người dùng giải quyết được các bài toán kiểm định giả thuyết.

Trong bài này, chúng ta quan tâm đến 5 bài toán kiểm định sau:

1. Kiểm định một mẫu
    1.1. Kiểm định giá trị trung bình tập chính
        a. Khi biết phương sai
        b. Khi chưa biết phương sai
    1.2. Kiểm định giá trị tỉ lệ tập chính
2. Kiểm định hai mẫu
    2.1. Kiểm định về hai giá trị trung bình tập chính ($\sigma_1 \neq \sigma_2$ chưa biết)
    2.2. Kiểm định về hai giá trị tỉ lệ tập chính

Quy trình làm một bài toán kiểm định giả thuyết với R trong bài này được tiến hành theo các bước sau:

- **Bước 1:** Tóm tắt bài toán
- **Bước 2:** Xác định hàm kiểm định và các tham số trong R
- **Bước 3:** Thực hiện kiểm định trên R
- **Bước 4:** Phân tích kết quả và kết luận.

Việc chấp nhận hay bác bỏ giả thuyết gốc H_0 với mức ý nghĩa α, có thể dựa vào trị số-p (p-value) với quy tắc:

- Nếu trị số-p < α thì bác bỏ H_0;
- Nếu trị số-p > α thì chưa bác bỏ H_0.

0. Một số gói lệnh cần cài đặt

Trong phần thực hành này, chúng ta cần cài đặt một vài gói lệnh sau:

```
> install.packages("BSDA")
> install.packages("readxl")
```

1. Một số hàm thường dùng

Để kiểm định một giả thuyết thống kê trong R, ta có thể sử dụng một số hàm được cho trong bảng sau:

Tham số	Hàm	Trường hợp	Gói
Giá trị trung bình	`z.test`	Phương sai đã biết	`BSDA`
	`t.test`	Phương sai chưa biết	
Tỷ lệ	`prop.test`		

1.1. Hàm `z.test`

Hàm `z.test` dựa trên phân phối chuẩn tắc, đưa ra khoảng tin cậy và kiểm định giả thuyết cho bài toán một mẫu và hai mẫu, ta có thể sử dụng nó với cú pháp như sau:

```
z.test(x, y, alternative=…, mu=…, sigma.x=…, sigma.y= …, conf.level=…)
```

với ý nghĩa của các tham số:

`x`	véctơ dữ liệu mẫu thứ nhất.
`y`	véctơ dữ liệu mẫu thứ hai, mặc định là `NULL` nếu chỉ có một mẫu.
`alternative`	chuỗi kí tự chỉ đối thuyết; là một trong ba chuỗi: **"`two.sided`"**, **"`less`"**, **"`greater`"**, tương ứng chỉ đối thuyết là hai phía, trái, phải; mặc định là **"`two.sided`"**. Ta có thể dùng kí tự đầu của chuỗi kí tự để thay cho chuỗi đó.
`mu`	hiệu chênh lệch của hai giá trị trung bình xác định theo giả thuyết gốc hoặc giá trị trung bình của giả thuyết gốc (một mẫu), mặc định là 0.
`sigma.x`	độ lệch chuẩn của tập chính thứ nhất
`sigma.y`	độ lệch chuẩn của tập chính thứ hai, mặc định là `NULL` nếu có một mẫu.
`conf.level`	độ tin cậy cho khoảng tin cậy được trả về, nằm trong khoảng (0; 1).

Để xem chi tiết hơn về hàm `z.test`, ta có thể dùng lệnh trợ giúp.

```
> library(BSDA)
> ?z.test
```

1.2. Hàm `t.test`

Hàm `t.test` có cú pháp như sau:

```
t.test(x, y, alternative=…, mu=…, paired=…, var.equal=…, conf.level=…)
```

với ý nghĩa của các tham số:

`x`	véctơ dữ liệu mẫu thứ nhất.
`y`	véctơ dữ liệu mẫu thứ hai, mặc định là `NULL` nếu chỉ có một mẫu.
`alternative`	chuỗi kí tự chỉ đối thuyết; là một trong ba chuỗi: **"`two.sided`"**, **"`less`"**, **"`greater`"**, tương ứng chỉ đối thuyết là hai phía, trái, phải; mặc định

	là "two.sided". Ta có thể dùng kí tự đầu của chuỗi kí tự để thay cho chuỗi đó.
mu	hiệu chênh lệch của hai giá trị trung bình xác định theo giả thuyết gốc hoặc giá trị trung bình của giả thuyết gốc (một mẫu), mặc định là 0.
paired	dạng logic (TRUE/FALSE) chỉ chọn mẫu theo đôi, mặc định là FALSE.
var.equal	dạng logic (TRUE/FALSE) chỉ phương sai hai tập chính bằng nhau, mặc định là FALSE.
conf.level	độ tin cậy cho khoảng tin cậy được trả về, nằm trong khoảng (0; 1).

Để xem chi tiết hơn về hàm z.test, ta có thể dùng lệnh trợ giúp.

```
> ?t.test
```

1.3. Hàm prop.test

Hàm prop.test dùng để kiểm định giá trị tỷ lệ tập chính hay kiểm định hai giá trị tỷ lệ tập chính, nó có cú pháp như sau:

```
prop.test(x, n, p=…, alternative=…, conf.level=…, correct=…)
```

với ý nghĩa các tham số:

x	véctơ chỉ số lần "thành công" trong mỗi mẫu.
n	véctơ chỉ số lần thử nghiệm trong mỗi mẫu.
p	véctơ chỉ xác suất thành công, có độ dài bằng số mẫu được chỉ định bởi x và các phần từ nằm trong khoảng từ 0 đến 1.
alternative	chuỗi kí tự chỉ đối thuyết; là một trong ba chuỗi: "two.sided", "less", "greater", tương ứng chỉ đối thuyết là hai phía, trái, phải; mặc định là "two.sided". Ta có thể dùng kí tự đầu của chuỗi kí tự để thay cho chuỗi đó.
conf.level	độ tin cậy cho khoảng tin cậy được trả về, nằm trong khoảng (0; 1).
correct	tham số logic (TRUE/FALSE) chỉ xem có hay không sự điều chỉnh liên tục Yate khi có thể, mặc định là TRUE.

Để xem chi tiết hơn về hàm z.test, ta có thể dùng lệnh trợ giúp.

```
> ?prop.test
```

2. Kiểm định một mẫu

2.1. Kiểm định giá trị trung bình tập chính

a. Khi biết phương sai

Ví dụ 1. Cho dữ liệu quan sát về cường độ (psi) của bê tông như sau:

4010; 3880; 3970; 3780; 3820.

Giả sử rằng cường độ của bê tông tuân theo luật phân phối chuẩn với độ lệch tiêu chuẩn 110 psi. Có thể kết luận cường độ trung bình của bê tông thấp hơn giá trị thiết kế là 4000 psi với mức ý nghĩa $\alpha = 5\%$ hay không?

- **Bước 1: Tóm tắt bài toán**

Xác định tham số	Ta quan tâm đến biến ngẫu nhiên X là cường độ của bê tông. Theo giả thiết, $X \sim N(\mu, \sigma^2)$ với $\sigma = 110$ psi. Ta kiểm định giá trị của cường độ trung bình μ.
Phát biểu giả thuyết	$H_0 : \mu = 4000$ $H_1 : \mu < 4000.$
Mức ý nghĩa	$\alpha = 0,05.$

- **Bước 2: Xác định hàm kiểm định và các tham số trong R**

Theo như phần trên, ở đây ta dùng hàm `z.test` với các tham số cần thiết như sau:

`x`	vécto dữ liệu mẫu là dữ liệu về cường độ
`y`	= `NULL` (do ở đây là một mẫu)
`alternative`	= "`less`" hoặc "`l`"
`mu`	= `4000`
`sigma.x`	= `110`
`sigma.y`	= `NULL`
`conf.level`	= $1-\alpha$ = `0.95`

- **Bước 3: Thực hiện kiểm định trên R**

```
> data1 <- c(4010, 3880, 3970, 3780, 3820)
> library(BSDA)
> z.test(data1, alternative="less", mu=4000, sigma.x=110, conf.level=0.95)
```

- **Bước 4. Phân tích kết quả và kết luận**

- Sau khi thực hiện các lệnh ở bước 3, ta thu được kết quả sau:

```
      One-sample z-Test

data:  data1
z = -2.1954, p-value = 0.01407
alternative hypothesis: true mean is less than 4000
95 percent confidence interval:
```

```
       NA 3972.916
sample estimates:
mean of x
     3892
```

- Kết quả trên cho ta một số thông tin sau:

+ Giá trị thống kê $z = \dfrac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} = -2.1954$;

+ Trị số-p của bài toán là `p-value = 0.01407`;

+ Cường độ bê tông trung bình trong mẫu $\bar{x} = 3892$.

- Kết luận: Vì p-value < α nên ta bác bỏ giả thuyết gốc H_0. Do đó, có cơ sở để nói cường độ trung bình của bê tông thấp hơn tiêu chuẩn thiết kế.

b. Khi chưa biết phương sai

Ví dụ 2. Một công ty công nghệ chuyên về nền tảng quảng cáo dựa trên chia sẻ hệ thống Wi-Fi miễn phí, quan tâm đến thời gian truy cập internet của khách hàng. Một mẫu ngẫu nhiên gồm 30 người dùng được chọn, dữ liệu về thời gian sử dụng mạng (phút) được cho dưới đây (dữ liệu có trong file **B4-KDGT-2.xlsx**):

5	15	14	8	8	6	2	10	11	12	4	7	19	22	17
8	9	3	9	12	13	8	7	16	11	23	15	5	6	14

Giả sử rằng thời gian truy cập mạng của người dùng tuân theo luật phân phối chuẩn. Có thể kết luận thời gian truy cập internet trung bình của người dùng bằng 10 phút hay không, với mức ý nghĩa 5% ?

- **Bước 1: Tóm tắt bài toán**

Xác định tham số	Ta quan tâm tới biến ngẫu nhiên X là thời gian truy cập internet của một khách hàng. Theo giả thiết $X \sim N(\mu, \sigma^2)$ với σ^2 chưa biết. Ta kiểm định giá trị của thời gian truy cập internet trung bình μ.
Phát biểu giả thuyết	$H_0 : \mu = 10$ (phút) $\quad$ $H_1 : \mu \neq 10$ (phút).
Mức ý nghĩa	$\alpha = 0,05$.

- **Bước 2: Xác định hàm kiểm định và các tham số trong R**

Ở đây, ta dùng hàm `t.test` với các tham số cần thiết như sau:

`x`	véctơ dữ liệu mẫu là dữ liệu về thời gian truy cập internet
`y`	`= NULL` (do ở đây là một mẫu)
`alternative`	`= "two.sided"` hoặc `"t"`

mu	= 10
paired	bỏ qua, mặc định là FALSE
var.equal	bỏ qua (do ở đây là một mẫu)
conf.level	= $1-\alpha$ = 0.95

- **Bước 3: Thực hiện kiểm định trên R**

```
> setwd("H:/GTVT/Thuc hanh R/Kiem dinh gia thuyet/dulieuthuchanh")
> library(readxl)
> data2 <- read_excel("B4-KDGT-2.xlsx")
> t.test(data2$ThoiGian, alternative="two.sided", mu=10, conf.level=0.95)
```

- **Bước 4: Phân tích kết quả và kết luận**

- Sau khi thực hiện các lệnh ở bước 3, ta thu được kết quả sau:

```
      One Sample t-test

data:  data2$ThoiGian
t = 0.64648, df = 29, p-value = 0.5231
alternative hypothesis: true mean is not equal to 10
95 percent confidence interval:
  8.629706 12.636961
sample estimates:
mean of x
 10.63333
```

- Kết quả trên cho ta một số thông tin sau:

+ Giá trị thống kê $t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}} = 0.64648;$

+ Bậc tự do (df: degree freedom) `df = n - 1 =29;`

+ Trị số-p của bài toán là `p-value = 0.5231;`

+ Thời gian truy cập trung bình trong mẫu $\bar{x} = 10.63333.$

- Kết luận: Vì p-value > α nên ta chưa bác bỏ giả thuyết gốc H_0. Do đó với mức ý nghĩa 5%, có cơ sở để nói rằng thời gian truy cập trung bình của người dùng là 10 phút.

2.2. Kiểm định giá trị tỷ lệ tập chính

Ví dụ 3. Một công ty tuyên bố rằng dịch vụ Internet của họ cung cấp cho 70% hộ gia đình của một khu vực dân cư. Kiểm tra ngẫu nhiên 200 hộ gia đình của khu vực trên thấy có 125 hộ sử dụng dịch vụ Internet của công ty đó. Với mức ý nghĩa 5%, có thể kết luận rằng tỉ lệ hộ gia đình sử dụng dịch vụ Internet của công ty trên thấp hơn mức tuyên bố 70% hay không?

- **Bước 1: Tóm tắt bài toán**

Xác định tham số Ta kiểm định giá trị của p, là tỉ lệ hộ gia đình trong khu vực dân cư sử dụng dịch vụ Internet của công ty được nói tới.

Phát biểu giả thuyết $H_0 : p = 0.7$ $H_1 : p < 0.7$

Mức ý nghĩa $\alpha = 0,05.$

- **Bước 2: Xác định hàm kiểm định và các tham số trong R**

Trong trường hợp này, ta sử dụng hàm kiểm định `prop.test` với các tham số:

`x` = `125` (số hộ gia đình sử dụng dịch vụ Internet)
`n` = `200` (tổng số hộ được điều tra)
`p` = `0.7`
`alternative` = `"less"`
`conf.level` = $1-\alpha$ = `0.95`
`correct` = `FALSE` (vì $n.\hat{p} \geq 5$ và $n(1-\hat{p}) \geq 5$)

- **Bước 3: Thực hiện kiểm định trên R**

```
> prop.test(125, 200, p=0.7, alternative="less", conf.level=0.95,
correct=FALSE)
```

- **Bước 4: Phân tích kết quả và kết luận**

- Sau khi thực hiện các lệnh ở bước 3, ta thu được kết quả sau:

```
      1-sample proportions test without continuity correction

data:  125 out of 200, null probability 0.7
X-squared = 5.3571, df = 1, p-value = 0.01032
alternative hypothesis: true p is less than 0.7
95 percent confidence interval:
 0.0000000 0.6792872
sample estimates:
    p
0.625
```

- Kết quả trên cho ta giá trị: `p-value = 0.01032.`

- Kết luận: Vì p-value < α nên ta bác bỏ giả thuyết gốc H_0. Do đó với mức ý nghĩa 5%, có cơ sở để nói rằng tỉ lệ hộ gia đình sử dụng dịch vụ Internet của công ty trên thấp hơn mức tuyên bố 70%.

3. Kiểm định hai mẫu

3.1. Kiểm định về hai giá trị trung bình tập chính ($\sigma_1 \neq \sigma_2$ chưa biết)

Ví dụ 4. Một người muốn lựa chọn một trong hai nhà cung cấp mạng Internet. Để quyết định, người đó dùng một ứng dụng để đo tốc độ đường truyền thực tế của hai nhà mạng trên cơ sở các gói cước tương đương nhau. Dữ liệu về tốc độ tải (đơn vị: Mbps) của hai nhà mạng (1) và (2) đo tại một số thời điểm được cho lần lượt trong file: **B4-KDGT-3.xlsx**. Với mức ý nghĩa 1%, có thể kết luận tốc độ đường truyền trung bình của nhà cung cấp (1) cao hơn nhà cung cấp (2) hay không?

- **Bước 1: Tóm tắt bài toán**

Xác định tham số: Ta quan tâm tới hai biến ngẫu nhiên X_1 và X_2 là tốc độ tải của hai nhà mạng (1) và (2). Ta kiểm định về hai giá trị trung bình μ_1 và μ_2 của chúng.

Phát biểu giả thuyết: $H_0 : \mu_1 = \mu_2$ $\quad$ $H_1 : \mu_1 > \mu_2$

Mức ý nghĩa: $\alpha = 0,01$.

- **Bước 2**: Xác định hàm kiểm định và các tham số trong R

Để kiểm định trung bình hai tập chính khi chưa biết phương sai của tập chính nhưng không được giả thiết bằng nhau, ta dùng hàm kiểm định `t.test` với các tham số:

`x`	vécto dữ liệu mẫu thứ nhất
`y`	vécto dữ liệu mẫu thứ hai
`alternative`	= "greater"
`mu`	= 0 (hiệu chênh lệch hai giá trị trung bình theo giả thuyết gốc)
`var.equal`	= FALSE (phương sai hai tập chính không bằng nhau)
`conf.level`	= $1-\alpha$ = 0.99

- **Bước 3: Thực hiện kiểm định trên R**

```
> setwd("H:/GTVT/Thuc hanh R/Kiem dinh gia thuyet/dulieuthuchanh")
> library(readxl)
> data3 <- read_excel("B4-KDGT-3.xlsx")
> t.test(data3$TocDo1, data3$TocDo2, alternative="greater", mu=0,
var.equal=FALSE, conf.level=0.99)
```

- **Bước 4: Phân tích kết quả và kết luận**

- Sau khi thực hiện các lệnh ở bước 3, ta thu được kết quả sau:

```
        Welch Two Sample t-test
```

```
data:  data3$TocDo1 and data3$TocDo2
t = 2.9247, df = 77.986, p-value = 0.002257
alternative hypothesis: true difference in means is greater than 0
99 percent confidence interval:
 0.2188485       Inf
sample estimates:
mean of x mean of y
 21.52829  20.36356
```

- Kết quả trên cho ta giá trị: `p-value = 0.002257.`

- Kết luận: Vì p-value $<\alpha$ nên ta bác bỏ giả thuyết gốc H_0. Do đó với mức ý nghĩa 5%, có cơ sở để nói rằng tốc độ đường truyền trung bình của nhà cung cấp (1) cao hơn nhà cung cấp (2).

3.2. Kiểm định về hai giá trị tỉ lệ tập chính

Ví dụ 5. Số trẻ em trong độ tuổi 6 – 15 tuổi tại Việt Nam mắc tật khúc xạ đang có xu hướng gia tăng. Có ý kiến cho rằng tỉ lệ trẻ em mắc tật khúc xạ sinh sống tại thành phố cao hơn tại nông thôn. Khảo sát ngẫu nhiên 560 trẻ em ta thu được dữ liệu:

Nhóm	Số trẻ em khảo sát	Số trẻ em mắc tật khúc xạ
Thành phố	320	80
Nông thôn	240	36

Với mức ý nghĩa 1%, hãy kết luận về ý kiến đã nêu.

- **Bước 1: Tóm tắt bài toán**

Xác định tham số Gọi p_1, p_2 là tỉ lệ trẻ em mắc tật khúc xạ ở thành phố và nông thôn.
Phát biểu giả thuyết $H_0 : p_1 = p_2$ $H_1 : p_1 > p_2$
Mức ý nghĩa $\alpha = 0,01.$

- **Bước 2: Xác định hàm kiểm định và các tham số trong R**

Để kiểm định những bài toán so sánh hai tỉ lệ với nhau ta vẫn dùng hàm kiểm định prop.test. Các tham số của hàm prop.test trong trường hợp này là:

`x` = `c(80,36)` (số trẻ em bị mắc tật khúc xạ ở thành phố và nông thôn)
`y` = `c(320,240)` (số trẻ em khảo sát ở thành phố và nông thôn)
`alternative` = "greater"
`conf.level` = $1-\alpha$ = `0.99`
`correct` = `FALSE`

- **Bước 3: Thực hiện kiểm định trên R**

```
> prop.test(c(80,36), c(320,240), alternative="greater", conf.level=0.99,
correct=FALSE)
```

- **Bước 4: Phân tích kết quả và kết luận**

- Sau khi thực hiện các lệnh ở bước 3, ta thu được kết quả sau:

```
      2-sample test for equality of proportions without continuity
      correction

data:  c(80, 36) out of c(320, 240)
X-squared = 8.3504, df = 1, p-value = 0.001928
alternative hypothesis: greater
99 percent confidence interval:
 0.02224332 1.00000000
sample estimates:
prop 1 prop 2
  0.25   0.15
```

- Kết quả trên cho ta giá trị: `p-value = 0.001928.`

- Kết luận: Vì p-value $<\alpha$ nên ta bác bỏ giả thuyết gốc H_0. Do đó với mức ý nghĩa 5%, có cơ sở để nói rằng tỉ lệ trẻ em mắc tật khúc xạ sinh sống tại thành phố cao hơn tại nông thôn.

BÀI TẬP THỰC HÀNH

Bài 1. Tuổi thọ của một loại bóng hình của máy vô tuyến truyền hình là một đại lượng ngẫu nhiên X tuân theo luật phân phối chuẩn với $EX = 3500$ giờ và độ lệch tiêu chuẩn là $\sigma = 20$ giờ. Nghi ngờ tuổi thọ bị thay đổi, người ta tiến hành theo dõi 25 bóng thấy tuổi thọ trung bình là 3422 giờ. Với mức ý nghĩa 5%, hãy kiểm định điều nghi ngờ trên.

Bài 2. Một nhà máy xử lý nước cung cấp cho một khu vực dân cư được xây dựng với công suất thiết kế 17000 mét khối một ngày. Gần đây, có thông tin cho rằng nhu cầu sử dụng nước cao hơn khả năng đáp ứng của nhà máy, do đó nhà máy cần phải nâng cấp. Người ta khảo sát nhu cầu sử dụng nước (đơn vị: nghìn mét khối) trong 40 ngày và kết quả được cho dưới đây:

8.7	12.1	12.6	13.7	14.8	15.1	15.4	15.9	16.2	16.5
16.8	16.8	17.1	17.2	17.3	17.4	17.6	17.7	17.7	17.9
18.0	18.1	18.2	18.2	18.4	18.5	18.6	18.6	18.6	18.7
18.9	18.9	19.1	19.1	19.1	19.5	19.5	19.5	20.2	21.0

Hãy kiểm định ý kiến trên với mức ý nghĩa 10%.

Bài 3. Mô hình thông tin xây dựng (BIM - Building Information Modeling) là một quy trình liên quan tới việc tạo lập và quản lý những đặc trưng kĩ thuật số trong các khâu thiết kế, thi công và vận hành các công trình. Để đánh giá về mức độ sử dụng BIM trong xây dựng công trình, người ta khảo sát 48 nhà thầu và kết quả cho thấy có 25 nhà thầu sử dụng BIM. Với mức ý nghĩa 1%, có thể kết luận tỉ lệ nhà thầu sử dụng BIM bằng 50% hay không?

Bài 4. Một loại vật liệu mới được nghiên cứu để sản xuất lốp xe ô tô và được so sánh với vật liệu đang được sử dụng. Các lốp xe sử dụng vật liệu cũ và mới được lắp vào các xe ô tô và cho chạy thử nghiệm trên quãng đường 60000 km, dưới các điều kiện giống nhau. Sau đó, độ mòn lốp xe (mm) được đo và dữ liệu được tổng hợp như dưới đây:

Loại vật liệu	Số lốp xe	Độ mòn trung bình	Độ lệch
Cũ	$n_1 = 40$	$\overline{x_1} = 3.68$	$s_1 = 0.74$
Mới	$n_2 = 40$	$\overline{x_2} = 3.19$	$s_2 = 0.83$

Với mức ý nghĩa 2%, có thể kết luận lốp xe sản xuất bằng vật liệu mới có độ mòn trung bình thấp hơn lốp xe sản xuất bằng vật liệu cũ hay không?

Bài 5. Hai nhà máy sử dụng hai công nghệ khác nhau để xử lý nước thải tại hai khu vực tương tự của một thành phố. Để đánh giá, người ta tiến hành thử nghiệm đối với hai nhà máy (1) và (2). Dữ liệu về số lần thử nghiệm và số kết quả cho thấy hàm lượng chất ô nhiễm bị giảm đáng kể của hai nhà máy được cho dưới đây:

$$n_1 = 90,\ m_1 = 33;\ n_2 = 100,\ m_2 = 44.$$

Với mức ý nghĩa 5%, hãy kiểm định xem có sự khác biệt về kết quả xử lý nước thải của hai nhà máy hay không?

TƯƠNG QUAN VÀ HỒI QUY

I. Thực hiện trên phần mềm R bài tập sau về tính hệ số tương quan và tính toán hàm hồi quy thực nghiệm:

Bài 1. *Một mẫu quan sát của đại lượng ngẫu nhiên hai chiều (*X; Y*) có giá trị như sau*
(2; 1; 4; 12); (2; 2; 4; 34); (2; 4; 4; 56); (2; 5; 4; 63)
(2; 25; 4; 38); (2; 45; 4; 75); (2; 16; 4; 4); (2; 34; 4; 62)

a) Hãy tính hệ số tương quan thực nghiệm của mẫu trên.
b) Hãy xây dựng hàm hồi quy tuyến tính của Y *theo* X.

Các thao tác cụ thể cần thực hiện với R:

1. Nhập biến X

```
> bienX <- c(2.1, 2.2, 2.4, 2.5, 2.25, 2.45, 2.16, 2.34)
```

2. Nhập biến Y

```
> bienY <- c(4.12, 4.34, 4.56, 4.63, 4.38, 4.75, 4.4, 4.62)
```

3. Tính hệ số tương quan theo cú pháp

```
> cor (bienX, bienY)
```

Xác nhận kết quả sau trên màn hình

```
[1] 0.9098077
>
```

4. Ước lượng các hệ số hồi quy theo cú pháp

```
> lm(bienY ~ bienX)
```

Xác nhận kết quả sau trên màn hình

```
lm(formula = bienY ~ bienX)

Coefficients:

(Intercept)       bienX

     1.544       1.274

>
```

Các thông tin cần nắm được khi thực hiện:

1. Cách nhập các biến X, Y.
2. Ký hiệu cor trong lệnh cor (bienX, bienY) nghĩa là *hệ số tương quan (coefficient of correlation)*. Công thức của hệ số này là
$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\ \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$
3. Giá trị nhận được từ R là
[1] 0.9098077
là giá trị tính được của hệ số tương quan r.
4. Ký hiệu lm trong lệnh lm(bienY ~ bienX) nghĩa là *mô hình tuyến tính (**l**inear **m**odel)*. Ký hiệu bienY ~ bienX có nghĩa là *mô tả bienY như một hàm số của bienX*. Công thức tính toán của mô hình là
$$y = \widehat{\beta_0} + \widehat{\beta_1}\, x$$
với $\widehat{\beta_0}, \widehat{\beta_1}$ là hai hệ số hồi quy thực nghiệm được ước lượng theo công thức
$$\widehat{\beta_1} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \qquad \widehat{\beta_0} = \bar{y} - \widehat{\beta_1}\bar{x}$$
5. Kết quả được xác định từ R là

Coefficients:

(Intercept)	bienX
1.544	1.274

có nghĩa là R tính ra được $\widehat{\beta_0} = 1.544$ *và* $\widehat{\beta_1} = 1.274$. Nói cách khác hàm hồi quy thực nghiệm được đưa ra là
$$y = 1{,}544 + 1{,}274\ x.$$

Trình bày lời giải của bài toán ra giấy sau các tính toán thực hành trên R (*áp dụng cho việc kiểm tra*):

Công thức tính hệ số tương quan là
$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\ \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$
Kết quả tính toán thực nghiệm là r =0,9098077 (sử dụng R)

Công thức tính các hệ số hồi quy tuyến tính là
$$\widehat{\beta_1} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \qquad \widehat{\beta_0} = \bar{y} - \widehat{\beta_1}\bar{x}$$

Kết quả tính toán thực nghiệm là $\widehat{\beta_0} = 1{,}544$ *và* $\widehat{\beta_1} = 1{,}274$ *(sử dụng R).*

Sinh viên có thể sử dụng các công thức tương đương để trình bày trong lời giải và cần giải thích được các ký hiệu $\bar{x}$, $\bar{y}$ trong lời giải trên.

Thực hành giải các bài tập sau bằng R

Bài 2. *Người ta lấy một mẫu thực nghiệm của đại lượng ngẫu nhiên hai chiều (X; Y) và thu được kết quả:*

X	*3, 6*	*3, 8*	*4, 3*	*4, 5*	*4, 9*	*5, 2*	*5, 4*
Y	*7, 1*	*7, 83*	*9, 62*	*10, 05*	*10, 7*	*11, 6*	*12,3*

a) Hãy tính hệ số tương quan thực nghiệm của mẫu trên.
b) Hãy xây dựng hàm hồi quy tuyến tính của Y theo X.

Kết quả đối chiếu

1. Kết quả tính hệ số tương quan thực nghiệm r
   [1] 0.9928191
2. Kết quả tính các hệ số hồi quy $\widehat{\beta_0}$, $\widehat{\beta_1}$
   lm(formula = bienY ~ bienX)
   Coefficients:
   (Intercept) bienX
   -2.590 2.755

Bài 3. *Để nghiên cứu về quan hệ giữa khối lượng bốc dỡ X (nghìn tấn) và thời gian bốc dỡ Y (giờ) người ta lấy một mẫu thực nghiệm và thu được kết quả:*

(10; 5, 5); (12; 6, 5); (11; 6, 3); (9; 4, 5);
(9, 5; 5, 3); (8; 4, 0); (12; 7, 0); (8, 5; 5, 0).

a) Hãy tính hệ số tương quan thực nghiệm của mẫu trên.
b) Hãy xây dựng hàm hồi quy tuyến tính của Y theo X.

Kết quả đối chiếu

1. Kết quả tính hệ số tương quan thực nghiệm r
   [1] 0.9619439
2. Kết quả tính các hệ số hồi quy $\widehat{\beta_0}$, $\widehat{\beta_1}$
   lm(formula = tgian ~ kluong)
   Coefficients:
   (Intercept) kluong
   -0.9420 0.6455

Bài 4. *Để nghiên cứu về quan hệ giữa khoảng cách X (km) từ nhà tới nơi làm việc và thời gian đi lại Y (phút), người ta lấy một mẫu thực nghiệm và có kết quả*

(10; 45); (12; 54); (11; 48); (9; 45);
(7; 30); (8; 32); (7, 5; 40); (8, 5; 42).

a) Hãy tính hệ số tương quan thực nghiệm của mẫu trên.
b) Hãy xây dựng hàm hồi quy tuyến tính của Y theo X.

Kết quả đối chiếu

1. Kết quả tính hệ số tương quan thực nghiệm r
   [1] 0.9012851
2. Kết quả tính các hệ số hồi quy $\widehat{\beta_0}$, $\widehat{\beta_1}$
   lm(formula = tgian1 ~ kcach)
   Coefficients:
   (Intercept) kcach
   4.433 4.117

II. Thực hiện trên phần mềm R bài tập về xác định hàm hồi quy và ước lượng giá trị dự báo:

Bài 5. *Số liệu về dân số (tính theo nghìn người) thành phố Hồ Chí Minh trong các năm gần đây được thống kê như sau:*

Năm	*2011*	*2012*	*2013*	*2014*	*2015*	*2016*
Số dân	*7498, 4*	*7660,3*	*7820, 0*	*7981,9*	*8146, 3*	*8320,1*

a) Hãy tìm hàm xu thế tuyến tính biểu thị dân số của thành phố Hồ Chí Minh.
b) Vẽ hình mô tả dữ liệu (biểu đồ phân tán) và đồ thị hàm hồi quy tuyến tính thực nghiệm.
c) Xác định sai số của dữ liệu được cung cấp và hàm hồi quy thực nghiệm tại các điểm quan sát.
d) Dự báo số dân năm 2017 của thành phố này và tìm khoảng tin cậy 98% cho giá trị đó.

Các thao tác cụ thể cần thực hiện với R:

1. Nhập biến thời gian

   > tgian <- c(2011, 2012, 2013, 2014, 2015, 2016)

2. Nhập biến dân số

   > danso <- c(7498.4, 7660.4, 7820, 7981.9, 8146.3, 8320.1)

3. Ước lượng các hệ số hồi quy theo cú pháp

   > lm(danso ~ tgian)

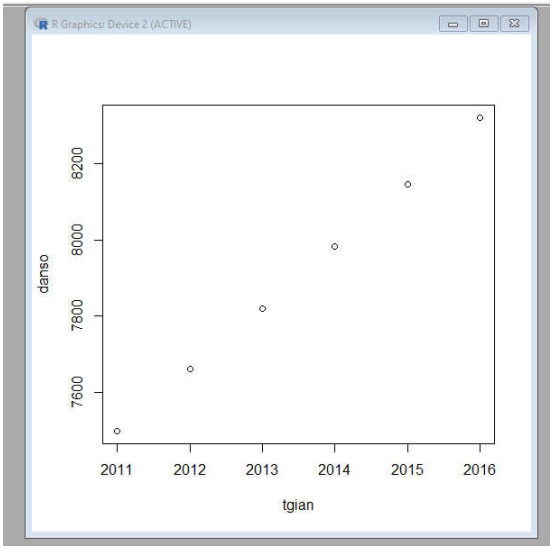
Xác nhận kết quả sau trên màn hình

Call:

```
lm(formula = danso ~ tgian)
Coefficients:
(Intercept)       tgian
  -321624.9       163.7
```

4. Vẽ biểu đồ miêu tả dữ liệu (biểu đồ phân tán) được cung cấp theo câu lệnh
> plot (tgian, danso)

Xác nhận hình ảnh được R đưa ra

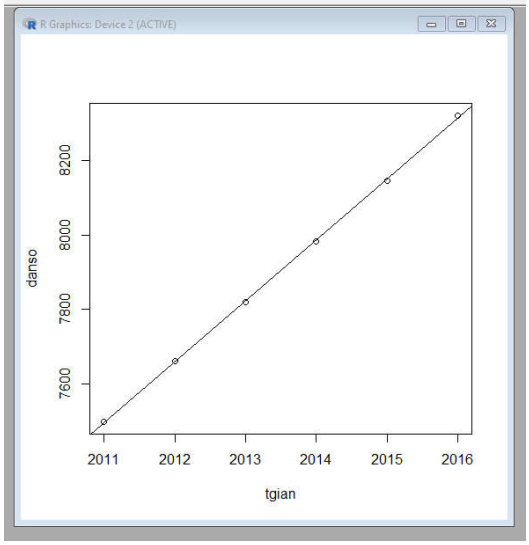


5. Tạo object chứa ***các thông tin về hồi quy*** trong R theo lệnh
> reg <- lm (danso ~ tgian)

6. Vẽ đường hồi quy thực nghiệm bằng R theo cú pháp

```
> abline(reg)
```

Xác nhận hình ảnh được R đưa ra



7. Tính sai số của dữ liệu được cung cấp và hàm hồi quy thực nghiệm tại các điểm quan sát theo lệnh

> residuals (reg)

hoặc câu lệnh thu gọn

> resid (reg)

Xác nhận kết quả R đưa ra

1	2	3	4	5	6
3.033333	1.373333	-2.686667	-4.446667	-3.706667	6.433333

8. Đưa ra công thức khoảng tin cậy sau để thực hiện tính toán theo yêu cầu d

$$\left(\widehat{y_0} - t_{(n-2,\frac{\alpha}{2})}\sqrt{s^2\left(\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right)}\ ;\ \widehat{y_0} + t_{(n-2,\frac{\alpha}{2})}\sqrt{s^2\left(\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right)}\right)$$

9. Nhập giá trị x_0 và tính $\widehat{y_0}$ theo các câu lệnh

> x0 <-2017

> beta0mu <- coef(reg)[1]

> beta1mu <- coef(reg)[2]

> y0mu <- beta0mu+beta1mu*x0

10. Đọc giá trị của $\widehat{y_0}$ từ R theo câu lệnh

>y0mu

Xác nhận kết quả được R đưa ra

(Intercept)

8477.327

11. Tính giá trị s^2 theo các câu lệnh

> n <- length(tgian)
> sbp <- sum(resid(reg)^2)/(n-2)

12. Đọc giá trị của s^2 từ R theo câu lệnh

>sbp

Xác nhận kết quả được R đưa ra

[1] **23.30133**

13. Tính $\bar{x}$ và S_{xx} theo các câu lệnh sau

> xtb <- mean(tgian)

> Sxx <-sum(tgian^2)-n*xtb^2

14. Đọc các giá trị $\bar{x}$ và S_{xx} và *xác nhận các kết quả từ R*

> xtb

[1] **2013.5**

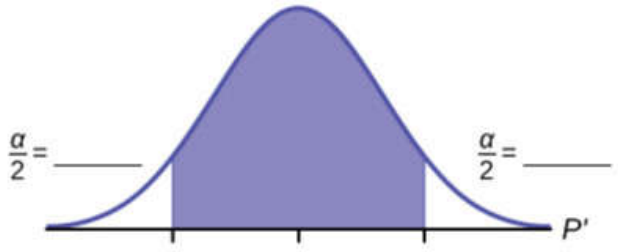
> Sxx

[1] **17.5**

15. Định nghĩa biến "phân vị" để ghi lại giá trị $t_{(n-2,\frac{\alpha}{2})}$

> phanvi <- qt(0.99,4)

Nhắc lại rằng *biến student với n-2 bậc tự do có đồ thị hàm mật độ là đường cong tạo với trục hoành một hình "quả chuông" có diện tích bằng 1.* ***Phân vị*** *(=vị trí phân chia) được xác định bởi hàm qt với 2 chỉ số. Chỉ số thứ nhất là diện tích mảnh chuông bên trái phân vị. Chỉ số thứ 2 là số bậc tự do.*



0,01 0,98 0,01

Do độ tin cậy là 98% nên ta lấy khoảng ước lượng theo mảnh chuông ở giữa có diện tích 0,98 và cắt bỏ 2 mẩu chuông cân đối 2 bên (mỗi mẩu diện tích 0,01). Điểm cắt bên phải là $t_{(n-2,\frac{\alpha}{2})}$*, điểm cắt bên trái là* $-t_{(n-2,\frac{\alpha}{2})}$*. Như vậy diện tích mẩu chuông bên phải phân vị* $t_{(n-2,\frac{\alpha}{2})}$ *là 0,01 và diện tích mảnh chuông bên trái phân vị* $t_{(n-2,\frac{\alpha}{2})}$ *là 1-0,01=0,99.* ***Đây là lý do ta đưa 0.99 vào chỉ số thứ nhất của hàm qt****.*

16. Đọc giá trị phân vị và *xác nhận các kết quả từ R*

> phanvi

[1] **3.746947**

17. Tính bán kính khoảng ước lượng theo các câu lệnh

bkinh <- phanvi*sqrt(sbp*(1/n+(x0-xtb)^2/Sxx))

và *xác nhận kết quả từ R*

> bkinh

[1] **16.83814**

18. Tính khoảng ước lượng theo câu lệnh

> y0mu+c(-1,1)*bkinh

và *xác nhận kết quả từ R*

[1] **8460.489 8494.165**

Trình bày lời giải các ý a và d của bài toán ra giấy sau các tính toán thực hành trên R (*áp dụng cho việc kiểm tra*):

a) Công thức tính các hệ số hồi quy tuyến tính là

$$\widehat{\beta_1} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \qquad \widehat{\beta_0} = \bar{y} - \widehat{\beta_1}\bar{x}$$

Kết quả tính toán thực nghiệm là $\widehat{\beta_0} = -321624{,}9$ *và* $\widehat{\beta_1} = 163{,}7$ *(sử dụng R)*

b) Thời điểm ước lượng dân số của TP Hồ Chí Minh là

$$x_0 = 2017$$

Điểm ước lượng cho dân số TP Hồ Chí Minh tại thời điểm được lựa chọn là

$$\widehat{y_0} = \widehat{\beta_0} + \widehat{\beta_1}\, x_0 = -321624{,}9 + 163{,}7 * 2017 = 8477{,}327$$

Công thức khoảng tin cậy cho dân số thành phố tại thời điểm được lựa chọn là

$$\left(\widehat{y_0} - t_{(n-2,\frac{\alpha}{2})}\sqrt{s^2\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}\ ;\ \widehat{y_0} + t_{(n-2,\frac{\alpha}{2})}\sqrt{s^2\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}\right)$$

Sử dụng R ta tính được các giá trị

$$\bar{x} = 2013{,}5;\ \ S_{xx} = 17{,}5;\ \ s^2 = 23{,}30133.$$

Phân vị của biến student là $t_{(n-2,\frac{\alpha}{2})} = 3{,}74694.$

Bán kính ước lượng là $\varepsilon = t_{(n-2,\frac{\alpha}{2})}\sqrt{s^2\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)} = 16{,}83814$

Kết quả tính toán khoảng tin cậy cho dân số thành phố tại thời điểm được lựa chọn là

(8460,489; 8494,165).

Thực hành giải các bài tập sau bằng R

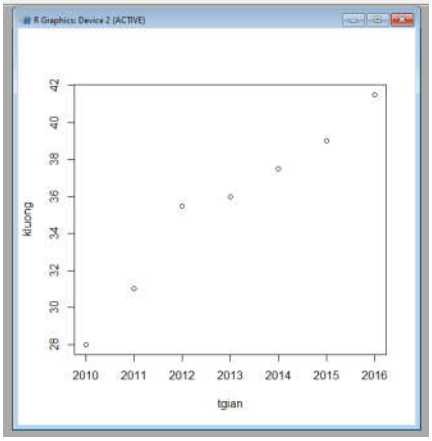
Bài 6. *Số liệu về lượng vận chuyển của một công ty vận tải trong các năm qua (tính theo triệu tấn) là như sau:*

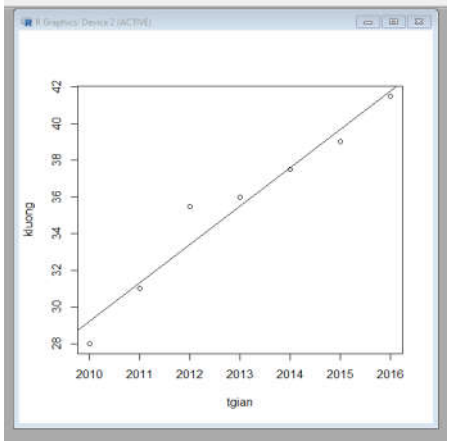
Năm	*2010*	*2011*	*2012*	*2013*	*2014*	*2015*	*2016*
Khối lượng	*28*	*31*	*35, 5*	*36*	*37, 5*	*39*	*41, 5*

a) Hãy tìm hàm xu thế tuyến tính biểu thị năng lực vận chuyển của công ty đó.
b) Vẽ hình mô tả dữ liệu và đồ thị hàm hồi quy tuyến tính thực nghiệm.
c) Xác định sai số của dữ liệu được cung cấp và hàm hồi quy thực nghiệm tại các điểm quan sát.
d) Dự báo khối lượng vận chuyển năm 2017 và tìm khoảng tin cậy 95% cho giá trị đó.

Kết quả đối chiếu

1. Kết quả tính các hệ số hồi quy $\widehat{\beta_0}, \widehat{\beta_1}$
   Call:
   lm(formula = kluong ~ tgian)
   Coefficients:
   (Intercept) tgian
   -4170.232 **2.089**
2. Hình ảnh của các mô tả trực quan





   Sai lệch giữa giá trị quan sát và hàm hồi quy
   > resid (reg)

1	2	3	4
-1.23214286	-0.32142857	2.08928571	0.50000000
5	6	7	
-0.08928571	-0.67857143	-0.26785714	

3. Điểm ước lượng cho lượng vận chuyển của công ty vận tải tại thời điểm được lựa chọn là
   > y0mu
   (Intercept)
   43.85714
4. Kết quả tính các biến trong công thức độ tin cậy
   > xtb

[1] **2013**
> Sxx
[1] **28**
> sbp
[1] **1.355357**

5. Phân vị và bán kính khoảng ước lượng
> phanvi
[1] **2.570582**
> bkinh
[1] **2.529265**

6. Khoảng tin cậy 95% cho lượng vận chuyển của công ty vận tải tại thời điểm được lựa chọn là

> y0mu+c(-1,1)*bkinh

[1] **41.32788 46.38641**

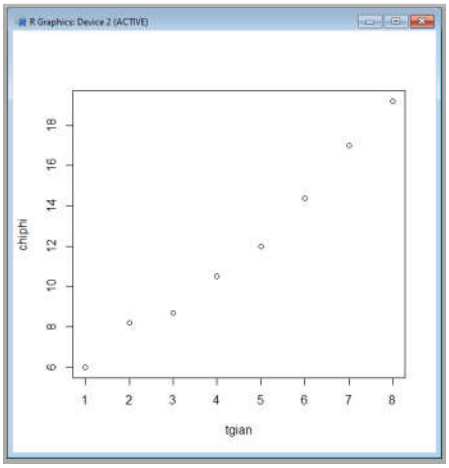
Bài 7. *Phân tích chi phí bảo dưỡng cho xe tải trong 8 năm sử dụng đầu tiên (tính theo triệu đồng) ta có kết quả:*

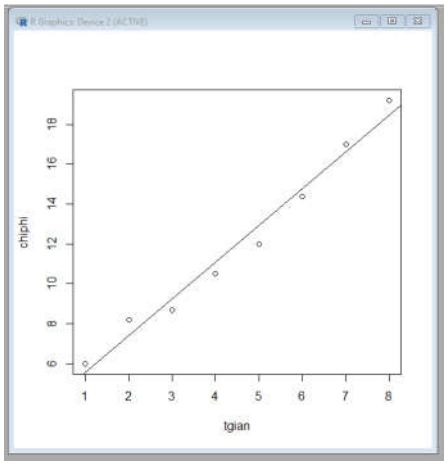
Năm thứ	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*
Chi phí TB	*6*	*8, 2*	*8, 7*	*10, 5*	*12*	*14, 4*	*17*	*19, 2*

a) Hãy tìm hàm xu thế tuyến tính biểu thị chi phí bảo dưỡng xe.
b) Vẽ hình mô tả dữ liệu và đồ thị hàm hồi quy tuyến tính thực nghiệm.
c) Xác định sai số của dữ liệu được cung cấp và hàm hồi quy thực nghiệm tại các điểm quan sát.
d) Dự báo chi phí bảo dưỡng trung bình cho xe trong năm sử dụng thứ 10 và tìm khoảng tin cậy 90% cho giá trị đó.

Kết quả đối chiếu

1. Kết quả tính các hệ số hồi quy $\widehat{\beta_0}, \widehat{\beta_1}$
Call:
lm(formula = chiphi ~ tgian)
Coefficients:
(Intercept) tgian
3.696 **1.845**
2. Hình ảnh của các mô tả trực quan





Sai lệch giữa giá trị quan sát và hàm hồi quy

> resid (reg)

1	2	3	4	5
0.4583333	0.8130952	-0.5321429	-0.5773810	-0.9226190
6	7	8		
-0.3678571	0.3869048	0.7416667		

3. Điểm ước lượng cho lượng vận chuyển của công ty vận tải tại thời điểm được lựa chọn là

> y0mu
(Intercept)
22.14881

4. Kết quả tính các biến trong công thức độ tin cậy

> xtb
[1] **4.5**
> Sxx
[1] **42**
> sbp
[1] **0.5290079**

5. Phân vị và bán kính khoảng ước lượng

> phanvi
[1] **1.94318**
> bkinh
[1] **1.299373**

6. Khoảng tin cậy 95% cho lượng vận chuyển của công ty vận tải tại thời điểm được lựa chọn là

> y0mu+c(-1,1)*bkinh

[1] **20.84944** **23.44818**

III. Bài tập làm thêm

Bài 8. Tóc độ xói mòn đất tại một công trường xây dựng được xem là hàm của độ dốc của khu vực địa hình đó. Dữ liệu về tốc độ xói mòn đất và độ dốc của một số điểm khảo sát được cho dưới đây:

Độ dốc (%)	1,2	1,6	2,4	3,2	3,6	4,1	4,9
Tốc độ xói mòn (tấn/ha/năm)	38	78	55	84	52	111	94

a) Vẽ biểu đồ phân tán của dữ liệu.
b) Hãy xác định đường hồi quy tuyến tính thực nghiệm biểu diễn tốc độ xói mòn theo độ dốc.

Bài 9. Tại một trường đại học, môn giải tích là điều kiện tiên quyết để sinh viên có thể học môn thống kê. Người ta lấy mẫu ngẫu nhiên 10 sinh viên đã hoàn thành cả hai môn học và ghi lại điểm của các sinh viên đó. Dữ liệu được cho dưới đây:

Giải tích	6,5	5,8	9,3	6,8	7,4	8,1	5,8	8,5	8,8	7,5
Thống kê	7,4	7,2	8,4	7,1	6,8	8,5	6,3	7,3	7,9	8,5

a) Tìm đường hồi quy tuyến tính biểu diễn điểm thống kê theo điểm giải tích.
b) Vẽ biểu đồ phân tán và đường hồi quy tuyến tính thực nghiệm. Dựa vào đồ thị để nhận xét về quan hệ giữa điểm của hai môn học.

Bài 10. Quảng cáo được xem là chìa khóa dẫn đến thành công. Để đánh giá hiệu quả của quảng cáo đến doanh thu, nhà quản lý của một chuỗi cửa hàng bán lẻ thu thập dữ liệu về doanh thu và chi phí dành cho quảng cáo (đơn vị: triệu đồng) từ các cửa hàng trong $n = 8$ tuần gần nhất. Dữ liệu được ghi lại trong bảng dưới đây.

Chi phí QC	3,0	7,0	6,5	3,5	4,5	7,0	7,5	8,5
Doanh thu	50	200	150	75	100	180	190	210

a) Hãy tính hệ số tương quan mẫu.
b) Hãy tìm hồi quy tuyến tính biểu diễn doanh thu qua chi phí quảng cáo.
c) Vẽ hình biểu đồ phân tán và đồ thị hàm hồi quy tuyến tính thực nghiệm.
d) Xác định sai số của dữ liệu được cung cấp và hàm hồi quy thực nghiệm tại các điểm quan sát.
e) Dự báo doanh thu đạt được trung bình ứng với chi phí quảng cáo 11 triệu và tìm khoảng tin cậy 95% cho giá trị đó.