

KHOA KHOA HỌC CƠ BẢN
BỘ MÔN: ĐẠI SỐ & XÁC SUẤT THỐNG KÊ

HƯỚNG DẪN THỰC HÀNH THỐNG KÊ VỚI PHẦN MỀM R



Bài 1: Giới thiệu về R

1.1. Giới thiệu về phần mềm R

R là một phần mềm sử dụng cho phân tích thống kê và đồ thị. Về bản chất, R là ngôn ngữ máy tính đa năng, có thể sử dụng cho nhiều mục tiêu khác nhau, từ tính toán đơn giản, toán học giải trí (recreational mathematics), tính toán ma trận (matrix), đến các phân tích thống kê phức tạp. Vì là một ngôn ngữ, cho nên người ta có thể sử dụng R để phát triển thành các phần mềm chuyên môn cho một vấn đề tính toán cá biệt.

Hai người sáng tạo ra R là hai nhà thống kê học tên là Ross Ihaka và Robert Gentleman. Kể từ khi R ra đời, rất nhiều nhà nghiên cứu thống kê và toán học trên thế giới ủng hộ và tham gia vào việc phát triển R. Chủ trương của những người sáng tạo ra R là theo định hướng mở rộng (Open Access). Cũng một phần vì chủ trương này mà R hoàn toàn miễn phí và bất cứ ai ở bất cứ nơi nào trên thế giới đều có thể truy nhập và tải toàn bộ mã nguồn của R về máy tính của mình để sử dụng.

1.2. Cài đặt phần mềm R

Để sử dụng R, việc đầu tiên là chúng ta phải cài đặt R trong máy tính của mình. Để làm việc này, ta phải truy nhập vào mạng và vào website có tên là “Comprehensive R Archive Network” (CRAN) sau đây: <http://cran.R-project.org>. Tài liệu cần tải về, tùy theo phiên bản, nhưng thường có tên bắt đầu bằng mẫu tự R và số phiên bản (version).

R cung cấp cho chúng ta một “ngôn ngữ” máy tính và một số function để làm các phân tích căn bản và đơn giản. Nếu muốn làm những phân tích phức tạp hơn, chúng ta cần phải tải về máy tính một số package khác. Package là một phần mềm nhỏ được các nhà thống kê phát triển để giải quyết một vấn đề cụ thể, và có thể chạy trong hệ thống R. Chẳng hạn như để phân tích hồi qui tuyến tính, R có function `lm` để sử dụng cho mục đích này, nhưng để làm các phân tích sâu hơn và phức tạp hơn, chúng ta cần đến các package như `lme4`. Các package này cần phải được tải về máy tính và cài đặt. Địa chỉ để tải các package vẫn là: <http://cran.r-project.org>, rồi bấm vào phần “Packages” xuất hiện bên trái của mục.

Để thuận lợi hơn cho việc sử dụng ta có thể sử dụng phần mềm R Studio, bằng cách tải từ website: <https://www.rstudio.com/products/rstudio>.

Các câu lệnh trong bài giảng này sẽ được trình bày trong R Studio.

1.3. Ngôn ngữ của R

“Văn phạm” chung của R là một lệnh (command) hay function (hàm). Mà đã là hàm thì phải có thông số; cho nên theo sau hàm là những thông số mà chúng ta phải cung cấp.

Chẳng hạn như:

```
> reg <- lm(y ~ x)
```

thì *reg* là một object, còn *lm* là một hàm, và $y \sim x$ là thông số của hàm.

Hay

```
> setwd("c:/works/stats")
```

thì *setwd* là một hàm, còn “*c:/works/stats*” là thông số của hàm.

Một số kí hiệu hay dùng trong R

<code>x == 5</code>	x bằng 5
<code>x != 5</code>	x không bằng 5
<code>y < x</code> (<code>y > x</code>)	y nhỏ hơn (lớn hơn) x
<code>y <= x</code> (<code>y >= x</code>)	y nhỏ hơn hoặc bằng (lớn hơn hoặc bằng) x
<code>is.na(x)</code>	có phải x là biến số missing
<code>A & B</code>	A và B (AND)
<code>A B</code>	A hoặc B (OR)
<code>!</code>	không (NOT)

Với R, tất cả các câu chữ hay lệnh sau kí hiệu # đều không có hiệu ứng, đây là kí hiệu dành cho người sử dụng thêm vào các ghi chú, ví dụ

```
> #tạo mẫu gồm 10 giá trị quan sát từ phân phối chuẩn tắc  
N(0,1)
```

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

Đặt tên một đối tượng (object) hay một biến số (variable) trong R khá linh hoạt, vì R không có nhiều giới hạn như các phần mềm khác. Tên một object phải được viết liền nhau. Chẳng hạn như R chấp nhận *myobject* nhưng không chấp nhận *my object*.

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

```
> my object <- rnorm(10)
```

```
Error: syntax error in "my object"
```

Nhưng đôi khi tên *myobject* khó đọc, cho nên chúng ta nên tách rời bằng “.” Ví dụ như:

```
> my.object <- rnorm(10)
```

Một điều quan trọng cần lưu ý là R phân biệt mẫu tự viết hoa và viết thường.

1.4. Các phép toán đơn giản và ma trận

1.4.1. Các phép toán đơn giản

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

```
> 1+2*(3+4)-5^2/3
```

```
[1] 6.666667
```

Ngoài ra, ta có thể sử dụng các phép toán như: hàm căn bậc hai (`sqrt`), hàm mũ (`exp`), hàm lôgarit (`log`, `log(10)`), hàm lượng giác (`sin`, `cos`, ...)

R còn có công dụng tạo ra những dãy số rất tiện cho việc mô phỏng và thiết kế thí nghiệm. Những hàm thông thường cho dãy số là `seq` (sequence), `rep` (repetition). Ví dụ

- Hàm `seq`

```
> x<-(12:5)
```

```
> x
```

```
[1] 12 11 10 9 8 7 6 5
```

```
> seq(4,6,0.5) # Tạo vec to từ 4 đến 6 với khoảng cách 0.5
```

```
[1] 4.00 4.50 5.00 5.50 6.00
```

```
> seq(length=10, from =2, to =15) # Tạo một vec to 10 số,
với số nhỏ nhất là 2 và số lớn nhất là 15
```

```
[1] 2.000000 3.444444 4.888889 6.333333 7.777778
```

```
[6] 9.222222 10.666667 12.111111 13.555556 15.000000
```

- Hàm `rep(x, times, ...)` với, `x` là một biến số và `times` là số lần lặp lại.

```
> rep(10, 3) # Tạo ra số 10, 3 lần
```

```
[1] 10 10 10
```

```
> rep(c(1:4),3) # Tạo ra số 1 đến 4, 3 lần
```

```
[1] 1 2 3 4 1 2 3 4 1 2 3 4
```

1.4.2. Các phép toán với ma trận

R có một package *Matrix* chuyên thiết kế cho tính toán ma trận, ta có thể tải package xuống, cài vào máy, và sử dụng từ địa chỉ:

[CRAN - Package Matrix \(r-project.org\)](http://CRAN - Package Matrix (r-project.org))

1.5. Nhập dữ liệu trong R

1.5.1. Nhập dữ liệu trực tiếp bằng lệnh c()

Ví dụ 1.1: Chúng ta có số liệu về độ tuổi và insulin cho 10 bệnh nhân như sau, và muốn nhập vào R.

Tuổi	50	62	60	40	48	47	57	70	48	67
Insulin	16.5	10.8	32.3	19.3	14.2	11.3	15.5	15.8	16.2	11.2

Ta có thể sử dụng function `c` (concatenation – nghĩa là “móc nối”) để nhập dữ liệu như sau

```
> age <- c(50, 62, 60, 40, 48, 47, 57, 70, 48, 67)
```

```
> insulin <- c(16.5, 10.8, 32.3, 19.3, 14.2, 11.3, 15.5, 15.8, 16.2, 11.2)
```

R là một ngôn ngữ cấu trúc theo dạng đối tượng (object-oriented language), vì mỗi cột số liệu hay mỗi một *data.frame* là một đối tượng (object) đối với R. Vì thế, *age* và *insulin* là hai đối tượng riêng lẻ. Bây giờ, chúng ta cần phải nhập hai đối tượng này thành một *data.frame* để R có thể xử lý sau này. Để làm việc này chúng ta cần đến *function data.frame*:

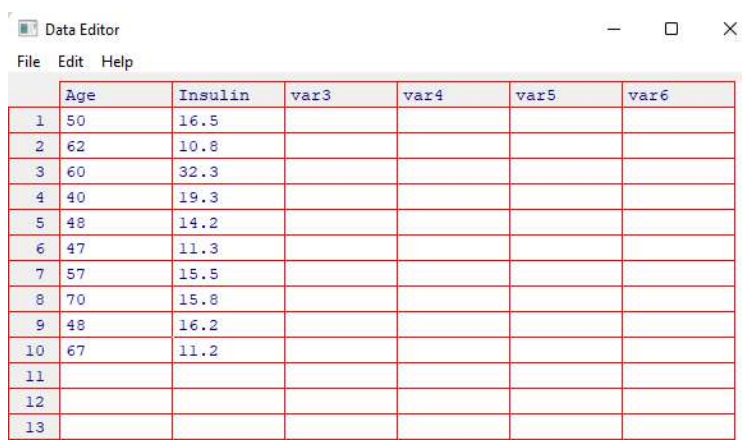
```
> VD1 <- data.frame(age, insulin)
```

Trong lệnh này, chúng ta muốn cho R biết rằng nhập hai cột (hay hai đối tượng) *age* và *insulin* vào một đối tượng có tên là *VD1*.

1.5.2. Nhập dữ liệu cho bảng

Chúng ta có thể nhập số liệu về độ tuổi và insulin cho 10 bệnh nhân trong Ví dụ 1 bằng một function rất có ích, đó là: `edit(data.frame())`. Với function này, R sẽ cung cấp cho chúng ta một window mới với một dãy cột và dòng giống như Excel, và chúng ta có thể nhập số liệu trong bảng đó. Ví dụ:

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



	Age	Insulin	var3	var4	var5	var6
1	50	16.5				
2	62	10.8				
3	60	32.3				
4	40	19.3				
5	48	14.2				
6	47	11.3				
7	57	15.5				
8	70	15.8				
9	48	16.2				
10	67	11.2				
11						
12						
13						

Sau đó ta nhập số liệu vào bảng, bấm nút chéo \times ở góc phải của spreadsheet, chúng ta sẽ có một data.frame tên `ins` với hai biến số `age` và `insulin`.

- Để xem lại dữ liệu trong `ins` ta dùng lệnh

```
> ins
```
- Để xem các biến trong `ins` ta dùng lệnh

```
> ins$age  
> ins$insulin
```
- Để xem kích thước của `ins` ta dùng lệnh

```
> dim(ins)
```
- Để truy cập các đối tượng của `ins` ta dùng lệnh

```
> ins[i,j] # truy cập đối tượng thuộc hàng i, cột j
```
- Lưu dữ liệu trong data.frame dưới dạng file.rda

```
> save(ins, file = "insulin.rda")
```
- Đọc dữ liệu của file.rda

```
> x <- load(file = "insulin.rda")
```

1.5.3. Nhập dữ liệu từ một file.csv hoặc file.xlsx

a. Đối với file.csv

Sử dụng hàm `read.csv` để đọc dữ liệu từ file.csv. Hàm này sẽ trả về một data frame.

Cú pháp:

```
> data <- read.csv("path/to/your/csv/file.csv")
```

Trong đó, `data` là tên biến mà bạn muốn lưu trữ dữ liệu và `"path/to/your/csv/file.csv"` là đường dẫn đến file csv trên máy tính của bạn.

b. Đối với file.xlsx

Để đọc dữ liệu từ file.xlsx, bạn cần cài đặt và sử dụng thư viện `readxl`. Bạn có thể cài đặt thư viện này bằng cách chạy lệnh sau trong R:

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

Sau khi cài đặt thành công, sử dụng hàm `read_excel()` để đọc dữ liệu từ file.xlsx. Hàm này cũng sẽ trả về một data frame.

```
> library(readxl)  
> data <- read_excel("path/to/your/excel/file.xlsx")
```

Trong đó, `data` là tên biến mà bạn muốn lưu trữ dữ liệu và `path/to/your/excel/file.xlsx` là đường dẫn đến `file.xlsx` trên máy tính của bạn.

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 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. Nội dung

2.1. Bảng phân phối tần số, tần suất

- Để lại lập *bảng tần số* ta dùng cấu trúc lệnh

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 <code>exclude = c(NA, NaN)</code> , 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)

Để đơn giản ta có thể sử dụng câu lệnh **table(x)**

- Để lại lập *bảng tần suất* ta dùng cấu trúc lệnh

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 <code>margin=1</code> thì tính tần suất các phần tử trên mỗi hàng, nếu <code>margin=2</code> 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)` .

Ví dụ 2.1: 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.

Hãy lập bảng phân phối tần số và tần suất của số lỗi.

```
> so.loi<-c(0,1,3,2,1,2,2,2,3,4,1,0,0,1,0,2,0,0,2,0)
```

```
> so.loi
```

```
[1] 0 1 3 2 1 2 2 2 3 4 1 0 0 1 0 2 0 0 2 0
```

```
> table(so.loi)
```

```
so.loi
```

```
0 1 2 3 4
```

```
7 4 6 2 1
```

```
> prop.table(table(so.loi))
```

```
so.loi
```

```
0 1 2 3 4
```

```
0.35 0.20 0.30 0.10 0.05
```

2.2. Đa giác tần số/ tần suất

Để vẽ đa giác tần số/tần suất ta dùng hàm

`plot(x, type)`

`x` véc tơ dữ liệu dùng để vẽ đa giác tần số/tần suất

`type` miêu tả kiểu vẽ, ta thường `type= "b"` (dạng các điểm được nối bằng đoạn thẳng)

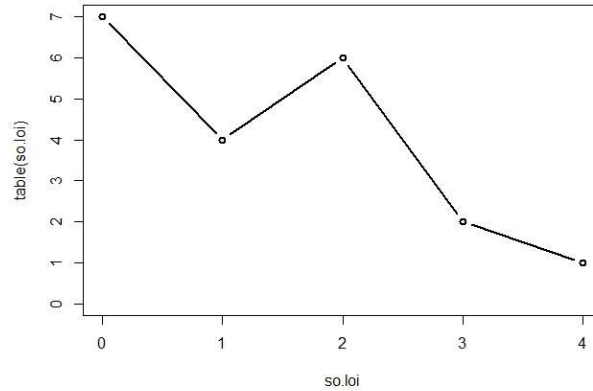
Để tạo hiệu ứng cho hình vẽ đẹp hơn ta có thể thêm vào các tham số:

`main` tiêu đề của hình vẽ

xlab, ylab tên của trục nằm ngang và trục thẳng đứng
col màu của đa giác

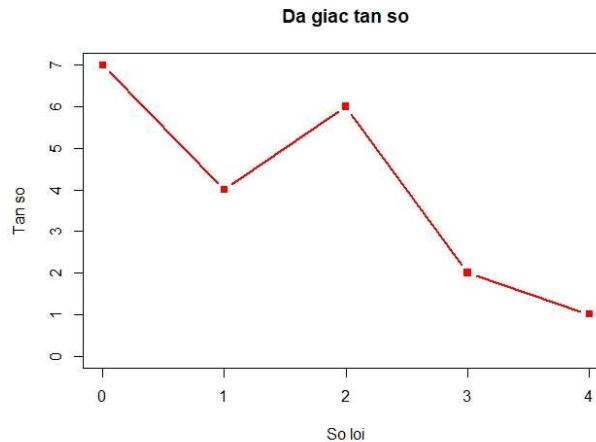
Ví dụ 2.2: Vẽ đa giác tần số cho số lỗi trong Ví dụ 2.1

```
> plot(table(so.loi), type="b")
```



Hình 1 (a) Biểu đồ về đa giác tần số

```
> plot(table(so.loi), type="b", main="Da giac tan so",  
col="red", pch=15, xlab="So loi", ylab="Tan so")
```



Hình 1 (b) Biểu đồ về đa giác tần số

2.3. Các dạng biểu đồ

2.3.1. Biểu đồ cột

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

barplot(x, col, border, main, xlab, ylab, xlim, ylim)

trong đó

col	màu của các cột
border	màu của đường biên các cột
main	tên của biểu đồ
xlab, ylab	tên trục x, y
xlim, ylim	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)

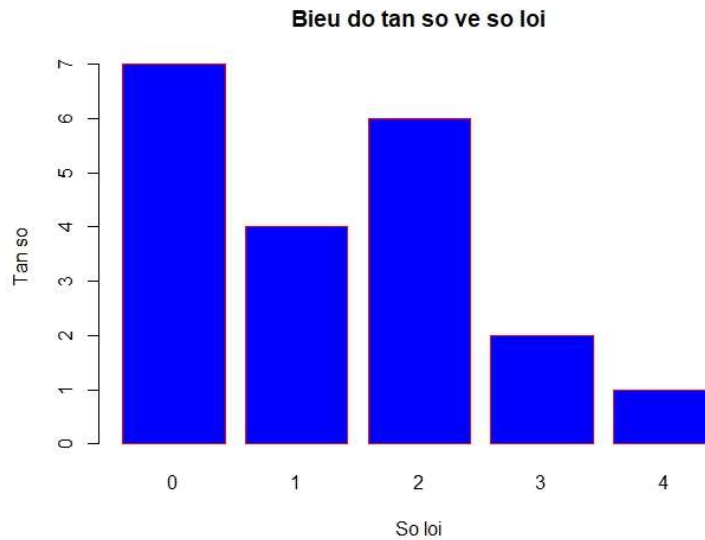
trong đó

x	Mã trận dữ liệu dùng để vẽ biểu đồ
names.arg	Tên viết dưới nhóm các thanh trong biểu đồ
legend.text	Ghi chú thích trong biểu đồ
beside	Dạng logic, nếu <code>beside = FALSE</code> (mặc định) thì các thanh của biểu đồ được vẽ chồng lên nhau, nếu <code>beside = TRUE</code> thì các thanh được vẽ cạnh nhau.
horiz	Dạng logic, nếu <code>horiz= FALSE</code> (mặc định) thì các thanh được vẽ vuông góc với trục nằm ngang với thanh đầu tiên ở bên trái, nếu <code>horiz= TRUE</code> thì các thanh được vẽ song song với trục nằm ngang với thanh đầu tiên nằm ở dưới cùng.

Ngoài ra, ta có thể điều chỉnh thêm một số tham số cho hàm `barplot` để biểu đồ được đẹp hơn như: `col, border, main, xlim, ylim, ...`.

Ví dụ 2.3: Vẽ biểu đồ tần số về số lỗi trong Ví dụ 2.1

```
> barplot(table(so.loi), col="blue", border="red", main =  
"Biểu đồ tần số về số lỗi", xlab="Số lỗi", ylab="Tần số")
```



Hình 1 Biểu đồ tần số

2.3.2. 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. Kỹ 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

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

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

$$\text{độ rộng} = (\text{giá trị lớn nhất} - \text{giá trị nhỏ nhất}) / \text{số khoảng chia}.$$

- Để chia khoảng dữ liệu, ta dùng hàm `cut` có cấu trúc như sau

`cut(x, breaks, right)`

trong đó

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

• 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

<code>x</code>	là véc tơ dữ liệu dạng số cần được phân nhóm
<code>freq</code>	dạng logic, nếu <code>freq = TRUE</code> (mặc định) các cột của biểu đồ mô tả tần số, nếu <code>freq = FALSE</code> các cột của biểu đồ mô tả tần suất.
<code>breaks</code>	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
<code>right</code>	dạng logic, nếu <code>right = TRUE</code> (mặc định) thì các cột lấy các phần tử trong khoảng dạng $(a, b]$, nếu <code>right = FALSE</code> thì khoảng dạng $[a, b)$.

Để đồ thị đẹp hơn, ta có thể thêm vào các tham số: `col`, `border`, `main`, `xlab`, `ylab`, `xlim`, `ylim`, `labels`, ...

Ví dụ 2.4: 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									

- Xác định số khoảng chia K theo công thức Sturge.
- Lập bảng phân bố tần số và tần suất theo kiểu chia khoảng cho dữ liệu này.

c) Vẽ biểu đồ histogram tần số.

Bước 1: Nhập dữ liệu

```
> 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)  
> thoigianhan  
[1] 69 60 75 74 68 66 73 76 63 67 69 73 65 61 73 72 72 65 69 70 64 61 74 76  
[25] 72 74 65 63 69 73 75 70 60 62 68 74 71 73 68 67
```

Bước 2: Tính số khoảng chia K theo công thức Sturge.

```
> n<-length(thoigianhan)  
> K <- 1+3.3*log(n, base=10)  
> K  
[1] 6.286798  
> do.rong=(max(thoigianhan)-min(thoigianhan))/K  
> do.rong  
[1] 2.545016
```

Vì độ rộng = 2,545016 nên để thuận tiện ta lấy độ rộng của mỗi khoảng là 3. Do đó, ta có thể chọn các khoảng chia: [60, 63), [63, 66), [66, 69), [69, 72), [72, 75), [75, 78).

Bước 3: Lập bảng tần số, tần suất

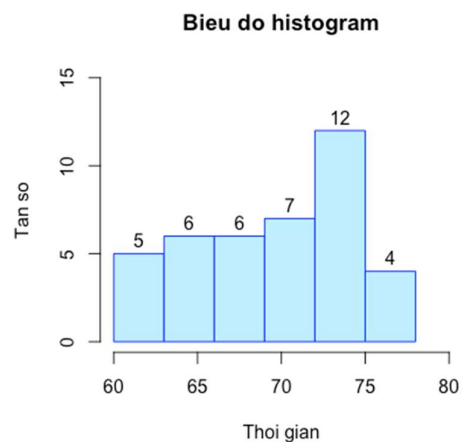
```
> thoigian <- cut(thoigianhan, breaks = c(60, 63, 66, 69, 72,  
75, 78), right = FALSE)  
> table(thoigian)  
thoigian  
[60, 63) [63, 66) [66, 69) [69, 72) [72, 75) [75, 78)  
          5         6         6         7        12         4  
> prop.table(table(thoigian))  
thoigian  
[60, 63) [63, 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.

```
> 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
```

Bước 4: Vẽ biểu đồ histogram tần số.

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



Hình 3. Biểu đồ histogram

Ví dụ 2.5: Dữ liệu về số lượng cầu được xây dựng trong các năm 1999, 2000, 2001 phân theo loại cầu

Loại cầu	Số lượng 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

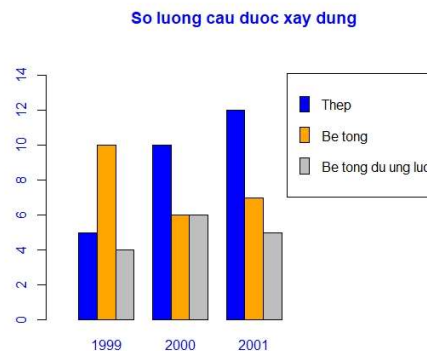
Hãy vẽ biểu đồ cột thể hiện dữ liệu trên.

Bước 1: Nhập dữ liệu

```
> so.luong.cau = matrix(c(5,10,4,10,6,6,12,7,5), nrow = 3)
> so.luong.cau
      [,1] [,2] [,3]
[1,]  5  10  12
[2,] 10   6   7
[3,]  4   6   5
```

Bước 2: Vẽ biểu đồ histogram

```
> barplot(so.luong.cau, 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("Thép", "Be
tong", "Be tong du ung luc"))
```



Hình 4. Biểu đồ histogram về số lượng cầu được xây dựng trong các năm 1999, 2000, 2001

Chú ý: Trong cấu trúc lệnh trên nếu ta thay `beside = T` bởi `beside = F` thì ta có biểu đồ với các cột thể hiện số lượng cầu xây dựng trong mỗi năm được xếp chồng lên nhau.

2.3.3. Biểu đồ hình tròn

Để vẽ biểu đồ hình tròn, ta dùng hàm `pie`

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

<code>x</code>	Là véc tơ dữ liệu dạng số cần được phân nhóm
<code>labels</code>	Tên của những hình quạt trong biểu đồ
<code>col</code>	Màu của các hình quạt
<code>border</code>	Màu của đường ranh giới giữa các hình quạt

lty Kiểu nét vẽ của đường ranh giới giữa các rổ 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

Ví dụ 2.6: Thống kê số kilomet chiều dài đường bộ của Việt Nam tính đến năm 2013

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.

Bước 1: Nhập dữ liệu

```
> chieudai <- c(108023, 6509, 48555, 48409)
> loaidualong <- c("Duongnhua", "Duongda", "Duongcapphoi",
"Duongdat")
> du.lieu.duong.bo <- data.frame(loaidualong, chieudai)
> du.lieu.duong.bo
  loaidualong  chieudai
1 Duongnhua  108023
2 Duongda    6509
3 Duongcapphoi 48555
4 Duongdat    48409
```

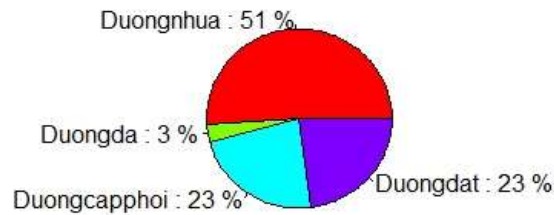
Bước 2: Tính tỉ lệ phần trăm từng loại đường

```
> Tile <- round(prop.table(chieudai), 2)*100
> Tile
[1] 51 3 23 23
```

Bước 3: Vẽ biểu đồ hình tròn

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

Chieu dai duong bo Viet Nam



So lieu nam 2013

Hình 5. Biểu đồ hình tròn

2.4. Các giá trị đặc trưng

- Để tính toán các giá trị đặc trưng của mẫu, ta có thể sử dụng các cấu trúc lệnh sau

Hàm	Chức năng
<code>mean(x)</code>	Tính trung bình mẫu
<code>median(x)</code>	Tính trung vị của các giá trị cho trong véc tơ <code>x</code>
<code>which(table(x) == max(table(x)))</code>	Tìm các giá trị mode và vị trí của các giá trị mode này trong <code>table(x)</code>
<code>range(x)</code>	Giá trị nhỏ nhất, giá trị lớn nhất của các giá trị trong véc tơ <code>x</code>
<code>var(x)</code>	Tính phương sai của các giá trị cho trong véc tơ <code>x</code>
<code>sd(x)</code>	Tính độ lệch chuẩn của các giá trị cho trong véc tơ <code>x</code>
<code>summary(x)</code>	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 <code>x</code>

- Để minh họa các tham số: trung bình, tứ phân vị và các giá trị ngoại biên trên cùng hình vẽ ta sử dụng biểu đồ hộp và râu bằng hàm `boxplot` với cấu trúc

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

Ví dụ 2.7: Cho dữ liệu về cường độ bê tông từ file `cuongdobetong.csv`. Tính các giá trị đặc trưng của mẫu quan sát.

Bước 1: Chuyển đường dẫn và đọc file `cuongdobetong.csv` và đặt biến mới

```
> setwd("D:/ThuchanhR/dulieuthuchanhR")
```

```
> cuongdobetong <- read.csv("cuongdobetong.csv")
> dim(cuongdobetong)
[1] 40 1
> cuongdo=cuongdobetong$cuong.do
> table(cuongdo)
cuongdo
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 2: Tính các giá trị đặc trưng của biến cuongdo

```
> mean(cuongdo)
[1] 4.5
> median(cuongdo)
[1] 4.5
> which(table(cuongdo) == max(table(cuongdo)))
4.5
  5
> range(cuongdo)
[1] 4.1 4.9
> var(cuongdo)
[1] 0.05487179
> sd(cuongdo)
[1] 0.2342473
```

Như vậy ta tính được các số đặc trưng của dữ liệu như sau: $\bar{x} = 4,5$; $s^2 \approx 0,05487$; $s \approx 0,23425$

Trung vị = 4,5; mode= 4,5 và giá trị này ở vị trí thứ 5 trong bảng tần số, min = 4,1; max= 4,9.

```
> 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% vị trí có cường độ từ 4,625 ksi trở lên, khoảng 25% vị trí có cường độ không quá 4,375 ksi.

Ta cũng có thể sử dụng hàm `describe(x)` trong gói **psych** để tính các giá trị đặc trưng.

Ví dụ 2.8: 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.

Bước 1: 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)
```

Bước 2: Tính giá trị trung bình và các tứ phân vị của mỗi tập dữ liệu

```
> summary(A)
```

```
Min. 1st Qu. Median Mean 3rd Qu. Max.
```

```
2.000 6.000 8.000 7.917 10.000 13.000
```

```
> summary(B)
```

```
Min. 1st Qu. Median Mean 3rd Qu. Max.
```

```
0.000 1.000 2.000 2.042 3.000 6.000
```

Bước 3: Vẽ biểu đồ hộp và râu

- Vẽ biểu đồ cho dữ liệu A (Hình 1.5(a))

```
> boxplot(A, main = "So vu tai nan truoc khi lap thiet bi  
kiem soat", border = "blue1", col = "red", horiz = F, xlab =  
"Truoc khi lap", ylab = "So vu tai nan", ylim = c(0,13))
```

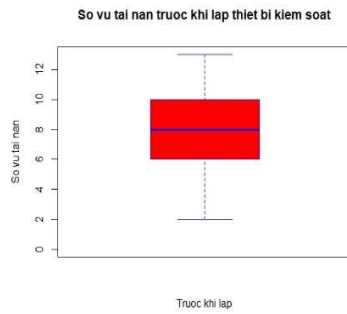
- Vẽ biểu đồ cho dữ liệu B (Hình 1.5(b))

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

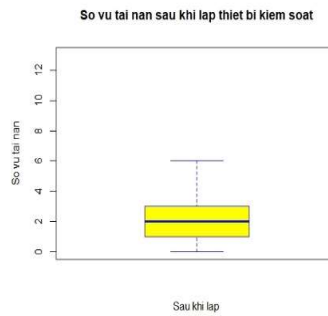
- Vẽ biểu đồ cho dữ liệu A và B (Hình 1.5(c))

```
> boxplot(A, B, main = "So vu tai nan truoc va sau khi lap  
thiet bi kiem soat", border = "blue1", col = c("red",
```

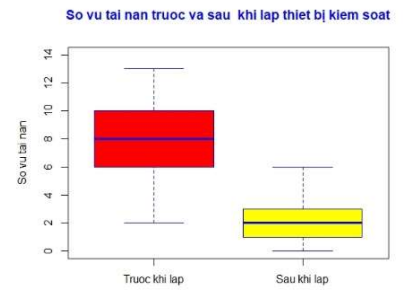
```
"yellow"),horiz = F, ylab = "So vu tai nan", ylim = c(0,14),
names = c("Truoc khi lap", "Sau khi lap"), col.main = "blue")
```



Hình 6a



Hình 6b



Hình 6c

Hình 6: Biểu đồ hộp và biểu đồ râu

C. Bài tập

Bài 1. Dữ liệu về số tai nạn mỗi tháng trên đoạn đường cao tốc được ghi nhận trong 24 tháng liên tiếp:

2, 0, 1, 0, 0, 4, 3, 4, 2, 1, 1, 1, 2, 0, 5, 1, 1, 1, 0, 2, 1, 3, 1, 3, 0

- Hãy nhập dữ liệu trên vào R.
- Lập bảng phân phối tần số và vẽ biểu đồ tần số dạng cột.
- Tính các giá trị đặc trưng của mẫu.

Bài 2. Số liệu về thời gian (tính theo phút) để hoàn thành một bài tập về nhà (gồm nhiều câu hỏi) của sinh viên được cho dưới theo bảng sau:

X (thời gian)	100	105	110	115	125	130	140
n_i (số sinh viên)	12	18	26	25	15	10	6

- Hãy nhập dữ liệu trên vào R.
- Vẽ đa giác tần số và vẽ biểu đồ tần số dạng cột.
- Tính các giá trị đặc trưng của mẫu.

Bài 3. Dữ liệu sau đây là tuổi thọ (đơn vị: năm) của 24 pin đặc biệt được sử dụng trong công nghiệp

1,8 5,1 4,2 6,3 3,3 5,8 4,4 4,8 3,0 4,3 4,7 5,1
4,3 4,2 1,8 4,9 5,8 4,4 4,4 3,0 4,3 5,8 5,1 5,8

- Hãy nhập dữ liệu trên vào R.
- Vẽ đa giác tần số và vẽ biểu đồ tần số dạng cột.
- Tính các giá trị đặc trưng của mẫu.

Bài 4. . Số liệu về thời gian (tính theo phút) để hoàn thành một bài tập về nhà (gồm nhiều câu hỏi) của sinh viên được cho dưới theo bảng sau:

X (thời gian)	85-90	90-95	95-100	100-105	105-110	110-115	115-120
n_i (số sinh viên)	25	30	35	30	27	19	8

- Hãy nhập dữ liệu trên vào R.
- Vẽ đa giác tần số và vẽ biểu đồ tần số dạng cột.
- Tính các giá trị đặc trưng của mẫu.

Bài 5. Số sản phẩm bị khách hàng trả lại của một nhà phân phối lớn được thống kê trong 25 ngày gần nhất là

21 8 17 22 19 18 19 14 17 24 6 21 11
 6 21 25 19 9 12 16 16 10 20 25 29

- Hãy nhập dữ liệu trên vào R.
- Vẽ đa giác tần số và vẽ biểu đồ tần số dạng cột.
- Tính các giá trị đặc trưng của mẫu.

Bài 6. Cho dữ liệu về cường độ bê tông (*ksi- kilopound per square inch*) được thu thập dựa vào phương pháp kiểm tra không phá hủy bằng sóng siêu âm tại một số vị trí

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 4,5 4,6 4,2
 4,3 4,2 4,3 4,6 4,4 4,1 4,5 4,6 4,9 4,9 4,4 4,3 4,5 4,4 4,4 4,1

- Hãy nhập dữ liệu trên vào R.
- Vẽ đa giác tần số và vẽ biểu đồ tần số dạng cột.
- Tính các giá trị đặc trưng của mẫu.

Bài 7. Chủ một cửa hàng bán thiết bị điện xem xét việc bán dây cắt sẵn theo chiều dài nhất định để giảm chi phí thuê nhân viên. Mẫu dưới đây là chiều dài (đơn vị: m) dây điện được bán trước đó.

3 7 4 2,5 3 20 5 5 15 3,5 3

- Hãy nhập dữ liệu trên vào R.
- Vẽ đa giác tần số và vẽ biểu đồ tần số dạng cột.
- Tính các giá trị đặc trưng của mẫu.

Bài 8. Độ bền chống trượt (đơn vị: tấn/ m^2) được xác định bằng thí nghiệm nén nổ hông tại khu vực miền núi phía bắc Việt Nam

1,31 2,30 3,94 4,05 4,27 5,03 5,14 5,47 5,47 5,58
 5,80 6,35 6,67 5,80 6,78 8,42 8,86 10,17 5,30 2,30

- Hãy nhập dữ liệu trên vào R.
- Vẽ đa giác tần số và vẽ biểu đồ tần số dạng cột.
- Tính các giá trị đặc trưng của mẫu.

Bài 9. Dữ liệu dưới đây thể hiện bề dày (đơn vị: m) tầng chứa nước trong trầm tích Holocene:

X (Độ dày)	22-29	29-36	36-43	43-50	50-57	57-64	64-71
n_i (số lượng)	7	8	17	17	18	3	6

- Hãy nhập dữ liệu trên vào R.

b) Vẽ đa giác tần số và vẽ biểu đồ tần số dạng cột.

c) Tính các giá trị đặc trưng của mẫu.

Bài 10. Theo dõi tốc độ (đơn vị: km/h) của một mẫu gồm 150 xe ô tô lưu thông trên đường cao tốc thường xuyên xảy ra tai nạn giao thông, ta thu được dữ liệu sau đây

Khoảng tốc độ	70-75	75-80	80-85	85-90	90-95	95-100
Số xe ô tô	20	45	34	17	24	10

a) Hãy nhập dữ liệu trên vào R.

b) Vẽ đa giác tần số và vẽ biểu đồ tần số dạng cột.

c) Tính các giá trị đặc trưng của mẫu.

Bài 11. Số liệu về tiếng ồn (đơn vị: dB -decibel) được đo từ các trạm như sau

Tiếng ồn	70-72	72-74	74-76	76-78	78-80
Số quan sát	5	6	14	8	3

a) Hãy nhập dữ liệu trên vào R.

b) Vẽ đa giác tần số và vẽ biểu đồ tần số dạng cột.

c) Tính các giá trị đặc trưng của mẫu.

Bài 12. Độ dày (đơn vị: mm) của một chi tiết kim loại trong một dụng cụ quang học được đo trên 100 sản phẩm như sau

Độ dày	3,20-3,25	3,25-3,30	3,30-3,35	3,35-3,40	3,40-3,45	3,45-3,50
Số sản phẩm	4	12	24	44	22	12

a) Hãy nhập dữ liệu trên vào R.

b) Vẽ đa giác tần số và vẽ biểu đồ tần số dạng cột.

c) Tính các giá trị đặc trưng của mẫu.

Bài 13. 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

a) Hãy nhập dữ liệu trên vào R.

b) Vẽ đa giác tần số và vẽ biểu đồ tần số dạng cột.

c) Tính các giá trị đặc trưng của mẫu.

Bài 14. 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 15. 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

- Hãy vẽ biểu đồ cột biểu diễn dữ liệu trên và cho nhận xét.
- Hãy vẽ biểu đồ hình tròn biểu diễn dữ liệu trên và cho nhận xét.
- So sánh thông tin có được khi dùng biểu đồ cột và biểu đồ hình tròn biểu diễn dữ liệu trên.

Bài 16. 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									

- 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.
- 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 17. 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	6.93	13.77	-8.55	-2.13	30.33	-34.75	30.31	24.30	-30.37
-13.24	22.42	-5.29	4.3	-18.95	54.19	6.06	-10.01	-5.61	44
34.40	-7.04	25	52	49.87	14.73	35.24	29	36.13	-20.23
-12.11	12.89	1.21	13.09	12.89	40.7	-26.01	4.16	1.53	22.18
-20.24	31.76	20.95	9.43	1.2	0.46	10.03	17.61	3.24	2.07
11.07	43.71	8.47	22.92	-19.27	10.51	1.2	25.1	29.44	39.04
-12.83	-9.22	33	63	0.52	9.94	-24.24	11	24.76	-33.39
-17	14.26	17.30	-2.59	-21.95	15.28	-38.47	58.67	13.44	-25.93
-15.83	10.33	0.63	36.08	-11.96	8.29	34.21	0.25	68	61
12.68	1.96	38	61	28.45	52	5.23	-20.44	66	-32.17

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: Lý thuyết ước lượng

A. Mục tiêu

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

- Tìm khoảng ước lượng cho giá trị trung bình của tập chính
- Tìm khoảng ước lượng cho giá trị tỉ lệ của tập chính

Để thực hành phần ước lượng, chúng ta cần cài đặt gói lệnh “BSDA” với cấu trúc

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

```
> library(BSDA)
```

Chú ý rằng, gói lệnh này chỉ thực hiện ở lần đầu tiên, trong các lần tiếp theo ta chỉ cần gọi thư viện “BSDA”.

B. Nội dung

Quy trình làm một bài toán tìm khoảng tin cậy với R trong bài này được tiến hành theo các bước sau:

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

3.1. Ước lượng khoảng cho giá trị trung bình

Bài toán: Cho X là biến ngẫu nhiên có phân phối chuẩn $N(\mu, \sigma^2)$. Tìm khoảng tin cậy cho giá trị trung bình μ với độ tin cậy γ từ mẫu quan sát (x_1, x_2, \dots, x_n) có kích thước n , trong hai trường hợp σ **đã biết** và σ **chưa biết**.

3.1.1. Trường hợp σ đã biết

- Thống kê ước lượng $T = \frac{(\bar{X} - \mu)\sqrt{n}}{\sigma}$
- Đặt $\alpha = 1 - \gamma$. Công thức khoảng ước lượng

$$\bar{x} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

- Cấu trúc hàm sử dụng trong R

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

trong đó

x

véc tơ dữ liệu mẫu

`sigma.x` độ lệch tiêu chuẩn mẫu
`conf.level` số thuộc $[0,1]$ chỉ độ tin cậy của khoảng ước lượng,
mặc định là 0.95

Ví dụ 3.1: 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.

Bước 1: Tóm tắt bài toán

- ✓ Gọi X là khối lượng chai nước của một dây chuyền đóng nước uống tinh khiết (đv: gam). Tìm khoảng tin cậy cho khối lượng trung bình μ của các chai nước với độ tin cậy $\gamma = 90\%$ và $\sigma = 10$ **đã biết**.

- ✓ Thống kê ước lượng: $T = \frac{(\bar{X} - \mu)\sqrt{n}}{\sigma}$

Bước 2: Xác định hàm ước lượng trong R là hàm `z.test` với các tham số x là véc tơ dữ liệu;

`sigma.x = 10; conf.level = 0.9.`

Thực hiện kiểm định trên R

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

Bước 3: Phân tích kết quả và kết luận

One-sample z-Test

```

data:  trong.luong
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

```

- Kết quả trên cho ta trọng lượng trung bình của 20 chai nước $\bar{x} = 299,65 \text{ g}$.

- Kết quả về khoảng tin cậy

```

90 percent confidence interval:
 295.972 303.328

```

Vậy ta có khoảng tin cậy 90% cho khối lượng trung bình của các chai nước sản xuất ra trên dây chuyền là $[295,972; 303,328]$.

Nhận xét 3.1: Nếu không có véc tơ dữ liệu mẫu mà chỉ có giá trị trung bình và kích thước của véc tơ dữ liệu mẫu, ta sử dụng hàm `zsum.test` với cấu trúc

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

trong đó

<code>mean.x</code>	trung bình mẫu quan sát
<code>sigma.x</code>	độ lệch chuẩn
<code>n.x</code>	kích thước mẫu quan sát
<code>conf.level</code>	độ tin cậy của khoảng ước lượng, mặc định 0,95

Ví dụ 3.2: 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.

Bước 1: Tóm tắt bài toán

- ✓ Gọi X là mức lương của trưởng phòng kinh doanh của các doanh nghiệp có quy mô trung bình tại thời điểm hiện tại (đv: triệu đồng). Tìm khoảng tin cậy cho mức lương trung bình μ của các trưởng phòng kinh doanh với độ tin cậy $\gamma = 95\%$ và $\sigma = 3.2$ đã biết.
- ✓ Thống kê ước lượng: $T = \frac{(\bar{X} - \mu)\sqrt{n}}{\sigma}$

Bước 2: Hàm ước lượng trong R là hàm `zsum.test` với các tham số `mean.x = 16.5`; `sigma.x = 3.2`; `n.x = 100`; `conf.level = 0.95` (hoặc `conf.level = NULL`).

Thực hiện kiểm định trên R

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

Bước 3: Phân tích kết quả và kết luận

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
```

Vậy ta có khoảng tin cậy 95% cho khối lượng trung bình của các chai nước sản xuất ra trên dây chuyền là $[15,87281; 17,12719]$.

3.1.2. Trường hợp σ chưa biết

- Thống kê ước lượng $T = \frac{(\bar{X} - \mu)\sqrt{n}}{s}$
- Đặt $\alpha = 1 - \gamma$. Công thức khoảng ước lượng

$$\bar{x} - t_{n-1;\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + t_{n-1;\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

- Cấu trúc hàm sử dụng trong R

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 3.3: 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 7.6 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.

Bước 1: Tóm tắt bài toán

✓ Gọi X là độ bám dính của hợp kim U - 700. Tìm khoảng tin cậy cho độ bám dính trung bình μ của hợp kim với độ tin cậy $\gamma = 95\%$ và σ **chưa biết**.

✓ Thống kê ước lượng: $T = \frac{(\bar{X}-\mu)\sqrt{n}}{s}$

Bước 2: Xác định hàm ước lượng trong R là hàm `t.test` với các tham số x là véc tơ dữ liệu;

`conf.level = 0,95`.

Thực hiện kiểm định trên R

```
> library(BSDA)
```

```
> do.bam.dinh = scan()
```

```
1: 19.8 10.1 14.9 7.5 15.4 15.4 18.5 7.9 12.7 11.9 11.4  
14.1
```

```
13: 17.6 16.7 15.8 19.5 8.8 13.6 11.4 11.4 14.1 7.6 16.7  
15.8
```

```
25: 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(do.bam.dinh, conf.level = 0.95)
```

Bước 3: Phân tích kết quả và kết luận

One Sample t-test

```
data: do.bam.dinh
t = 22.713, df = 35, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 12.31867 14.73688
sample estimates:
mean of x
 13.52778
```

- Kết quả trên cho ta độ bám dính trung bình của 36 hợp kim U-700 là $\bar{x} = 13,52778$.

- Kết quả về khoảng tin cậy

```
95 percent confidence interval:
 12.31867 14.73688
```

Vậy ta có khoảng tin cậy 95% cho độ bám dính trung bình của hợp kim U- 700 là [12,31867; 14,73688].

Nhận xét 3.2: Nếu không có véc tơ dữ liệu mẫu mà chỉ có giá trị trung bình, độ lệch tiêu chuẩn và kích thước của véc tơ dữ liệu mẫu, ta sử dụng hàm `tsum.test` với cấu trúc

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

trong đó

<code>mean.x</code>	trung bình mẫu quan sát
<code>s.x</code>	độ lệch chuẩn mẫu quan sát
<code>n.x</code>	kích thước mẫu quan sát
<code>conf.level</code>	độ tin cậy của khoảng ước lượng, mặc định 0,95

Ví dụ 3.4: 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 môn Toán trung bình với độ tin cậy 98% của học sinh ở thành phố đó.

Bước 1: Tóm tắt bài toán

- ✓ Gọi X là điểm môn Toán của học sinh. Tìm khoảng tin cậy cho điểm môn Toán trung bình μ của học sinh với độ tin cậy $\gamma = 98\%$ và σ chưa biết.

- ✓ Thống kê ước lượng: $T = \frac{(\bar{X}-\mu)\sqrt{n}}{s}$

Bước 2: Xác định hàm ước lượng trong R là `tsum.test` với các tham số `mean.x = 501`; `s.x = 112`; `n.x = 120`; `conf.level = 0,98`.

Thực hiện kiểm định trên R

```
> library(BSDA)
> tsum.test(mean.x = 501, s.x = 112, n.x = 120, conf.level
=0.98)
```

Bước 3: Phân tích kết quả và kết luận

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
98 percent confidence interval:
 476.8905 525.1095
sample estimates:
mean of x
      501
```

Vậy ta có khoảng tin cậy 98% cho điểm môn Toán trung bình của học sinh trong thành phố là $[476,8905; 525,1095]$.

3.2. Ước lượng khoảng cho giá trị tỷ lệ

Bài toán: Cho p là tỉ lệ cá thể có dấu hiệu T trong tập chính. Tìm khoảng tin cậy cho p khi biết trong mẫu quan sát có kích thước n có m cá thể có dấu hiệu T.

- Thống kê ước lượng $T = \frac{(\hat{p}-p)\sqrt{n}}{\sqrt{\hat{p}(1-\hat{p})}}$
- Đặt $\alpha = 1 - \gamma$. Công thức khoảng ước lượng

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

- Cấu trúc hàm sử dụng trong R

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

trong đó

x	số lần cá thể có dấu hiệu T trong mẫu quan sát
n	kích thước mẫu quan sát
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 Yates. Với $f = \frac{x}{n}$, nếu điều kiện $nf \geq 10; n(1-f) \geq 10$ thỏa mãn thì là <code>correct = FALSE</code> , trái lại <code>correct = TRUE</code> (mặc định <code>correct = TRUE</code>)

Ví dụ 3.5: 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.

Bước 1: Tóm tắt bài toán

- ✓ Gọi p là tỉ lệ ổ trục quay động cơ ô tô có bề mặt thô hơn so với thông số kỹ thuật cho phép. Tìm khoảng tin cậy cho tỉ lệ p với độ tin cậy $\gamma = 95\%$.
- ✓ Thống kê ước lượng: $T = \frac{(\hat{p}-p)\sqrt{n}}{\sqrt{\hat{p}(1-\hat{p})}}$

Bước 2: Xác định hàm ước lượng trong R là hàm `prop.test` với các tham số $n = 325$ là kích thước mẫu quan sát; $x = 74$ là số ổ trục có bề mặt thô hơn so với thông số kỹ thuật trong mẫu quan sát; `conf.level = 0,95`; `correct = FALSE` (vì $n.f = 74 > 10$; $n(1-f) = 251 > 10$ thỏa mãn).

Thực hiện kiểm định trên R

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

Bước 3: Phân tích kết quả và kết luận

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].

C. Bài tập

Bài 1. 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 2. Để thăm dò tình trạng sử dụng thuốc lá trong học đường, 996 thanh thiếu niên được hỏi về tình trạng sử dụng thuốc lá của bản thân và thu được kết quả sau

Yếu tố	Số người
Thường xuyên hút thuốc	269
Thi thoảng hút thuốc	385
Không hút thuốc	342

Hãy xây dựng khoảng tin cậy 90% cho tỉ lệ thanh thiếu niên không sử dụng thuốc lá.

Bài 3. Bảy nhân viên giao hàng của một cửa hàng pizza được hỏi về số kilomet mà họ phải đi 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. 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 5. 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 6. Gọi X là mức xăng tiêu thụ thực tế (lít) cho dòng xe ô tô con trên đoạn đường 100 km. Để ước lượng mức xăng hao phí trung bình, người ta lấy 36 chiếc và cho chạy thử thì tính được $\bar{x} = 9,45$ (lít). Biết rằng độ lệch tiêu chuẩn $\sigma = 3$, mức xăng tiêu thụ X là biến ngẫu nhiên có phân phối chuẩn. Tìm khoảng tin cậy 95% cho mức xăng hao phí trung bình.

Bài 7. Cho biết thời gian (tính theo phút) mà mỗi khách hàng truy cập và đọc tin tức trên một website điện tử là một biến ngẫu nhiên X có phân phối chuẩn. Người ta lấy một mẫu thực nghiệm và nhận được kết quả

X (thời gian)	0-10	10 - 20	20-30	30-40	40-50	50-60	60-80
n_i (số khách hàng)	42	63	75	52	39	24	13

Hãy tìm khoảng ước lượng của EX , với $\gamma = 0,95$.

Bài 8. Người ta khảo sát nhu cầu sử dụng nước (đơn vị: nghìn mét khối) trong 20 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

Tìm khoảng tin cậy 90% cho nhu cầu sử dụng nước trung bình.

Bài 9. Để đánh giá về bệnh ung thư phổi, người ta lấy ngẫu nhiên 1000 bệnh nhân mắc bệnh ung thư phổi và thấy có 823 người tử vong trong vòng 10 năm. Hãy ước lượng tỉ lệ người bị bệnh ung thư phổi có thể sống không quá 10 năm, với độ tin cậy 95%.

Bài 10. Cho biết thời gian di chuyển của một khách đi tàu hỏa (tính theo phút) từ nơi ở đến nhà ga là một biến ngẫu nhiên X có phân phối chuẩn. Người ta lấy một mẫu thực nghiệm và nhận được kết quả

X (thời gian)	0-10	10 - 15	15-20	20-25	25-30	30-40	40-50
n_i (số khách hàng)	11	46	62	35	32	22	14

Hãy tìm khoảng ước lượng của EX , với $\gamma = 0,99$.

Bài 11. Người ta tiến hành kiểm tra ngẫu nhiên 750 máy tính thì thấy rằng có 143 máy bị nhiễm mã độc. Với độ tin cậy $\gamma = 0,95$ hãy ước lượng tỷ lệ máy tính bị nhiễm mã độc.

Bài 12. Để đáp ứng số lượng và thể loại các chương trình dành cho trẻ em được phát sóng trên truyền hình, một đài truyền hình đã tiến hành cuộc khảo sát để ước tính số giờ trung bình trẻ em dành để xem tivi mỗi tuần. Dữ liệu lấy từ một mẫu gồm 24 trẻ em được cho như sau

39,7 21,5 40,6 15,5 43,9 33,0 21,0 15,8 27,1 23,8 18,3 23,4
 28,4 29,8 41,3 36,8 35,3 27,2 21,0 19,7 22,8 30,0 22,1 30

Tìm khoảng tin cậy 98% về số giờ trung bình trẻ em dành để xem tivi mỗi tuần biết rằng số giờ trẻ em xem ti vi mỗi tuần là biến ngẫu nhiên có phân phối chuẩn.

Bài 13. Để thăm dò ý kiến về yếu tố quan trọng nhất trong quyết định nơi mua sắm của phụ nữ, các nhà điều tra đã phát phiếu thăm dò cho 1200 phụ nữ ở một thành phố lớn và thu được kết quả như sau

Yếu tố	Số người
Giá	480
Chất lượng	360
Dịch vụ	180
Môi trường mua sắm	180

Hãy xây dựng khoảng tin cậy 95% cho tỉ lệ phụ nữ coi giá là yếu tố quan trọng nhất.

Bài 14. Số lượng xe ô tô đã qua sử dụng được bán ra bởi mỗi cửa hàng trong năm là biến ngẫu nhiên tuân theo luật phân phối chuẩn với độ lệch tiêu chuẩn là 15 xe. Kết quả điều tra số lượng xe ô tô đã qua sử dụng bán ra trong năm thông qua 15 cửa hàng như sau

79 43 58 66 101 63 79 33 58 71 60 101 74 55 88

Tìm khoảng tin cậy 90% cho số lượng trung bình xe ô tô đã qua sử dụng được bán ra.

Bài 15. Cho biết thời gian download một file dung lượng nhỏ (tính theo giây) của người dùng internet từ một trang chia sẻ tài liệu trực tuyến là một biến ngẫu nhiên X có phân phối chuẩn. Người ta lấy một mẫu thực nghiệm và nhận được kết quả

X (thời gian)	0-30	30 - 50	50-70	70-90	90-120	120-150	150-180
n_i (số lượt)	52	74	68	50	36	24	13

Hãy tìm khoảng ước lượng của EX , với $\gamma = 0,95$.

Bài 16. Một thí nghiệm kiểm tra tâm lý bằng cách đo thời gian phản ứng (đơn vị: giây) với một kích thích cụ thể. Kiểm tra ngẫu nhiên 52 đối tượng thấy thời gian phản ứng trung bình là 6,2 giây. Giả sử thời gian phản ứng là đại lượng ngẫu nhiên có luật phân phối chuẩn với độ lệch tiêu chuẩn $\sigma = 4$ giây. Tìm khoảng tin cậy 95% cho thời gian phản ứng trung bình.

Bài 4: Kiểm định giả thuyết thống kê

A. Mục tiêu

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

- Thực hiện về kiểm định giá trị trung bình
- Thực hiện về kiểm định giá trị tỉ lệ

Để thực hành phần kiểm định, chúng ta cần cài đặt gói lệnh “BSDA” và “readxl”

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

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

B. Nội dung

- Quy trình làm một bài toán kiểm định giả thuyết với R được tiến hành theo các bước:
 - ✓ **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 $p - value < \alpha$ thì bác bỏ H_0 ;
 - ✓ Nếu $p - value > \alpha$ thì chưa bác bỏ H_0 .

4.1. Bài toán kiểm định giá trị trung bình

- Xét bài toán kiểm định giá trị trung bình trong trường hợp 1 mẫu và 2 mẫu khi phương sai đã biết và phương sai chưa biết. Cụ thể, ta có các bài toán kiểm định

	Kiểm định 1 mẫu	Kiểm định hai mẫu
Giả thuyết gốc (H_0)	$\mu = \mu_0$	$\mu_1 = \mu_2$
Đối thuyết (H_1)	$\mu \neq \mu_0$	$\mu_1 \neq \mu_2$
	$\mu > \mu_0$	$\mu_1 > \mu_2$
	$\mu < \mu_0$	$\mu_1 < \mu_2$

4.1.1. Trường hợp phương sai đã biết

- Thống kê kiểm định

✓ Với bài toán kiểm định 1 mẫu, thống kê kiểm định $Z = \frac{(\bar{X} - \mu_0)\sqrt{n}}{\sigma}$.

✓ Với bài toán kiểm định 2 mẫu, thống kê kiểm định $Z = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$.

- Giá trị thống kê

✓ Bài toán 1 mẫu $z = \frac{(\bar{x} - \mu_0)\sqrt{n}}{\sigma}$

✓ Bài toán 2 mẫu $z = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$

- Cấu trúc hàm sử dụng trong R

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

trong đó

<code>x</code>	véc tơ dữ liệu thứ nhất
<code>y</code>	véc tơ dữ liệu thứ hai, mặc định là NULL nếu có 1 mẫu
<code>alternative</code>	chuỗi kí tự chỉ đối thuyết; là một trong ba chuỗi: “ <i>two.sided</i> ”, “ <i>less</i> ”, “ <i>greater</i> ”, 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
<code>mu</code>	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
<code>sigma.x</code>	độ lệch tiêu chuẩn của tập chính thứ nhất
<code>sigma.y</code>	độ lệch tiêu chuẩn của tập chính thứ hai, mặc định là NULL nếu 1 mẫu
<code>conf.level</code>	$= 1 - \alpha$, độ tin cậy cho khoảng tin cậy được trả về, thuộc (0; 1)

Ví dụ 4.1. 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

- ✓ Gọi X là cường độ của bê tông. Ta có, $X \sim N(\mu, \sigma^2)$ với $\sigma = 110$ **đã biết**. Ta kiểm định **một** giá trị trung bình μ với mức ý nghĩa $\alpha = 0,05$.
- ✓ Bài toán kiểm định: $H_0: \mu = 4000$ (psi); $H_1: \mu < 4000$ (psi)

Bước 2: Hàm kiểm định trong R là hàm `z.test` với các tham số $\mu = 4000$; $y = \text{NULL}$;

`sigma.x = 110; alternative = "less"; conf.level = 1 - α = 0.95.`

Bước 3: Thực hiện kiểm định trên R

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

Bước 4: Phân tích kết quả và kết luận

```
One-sample z-Test
data: cuongdo
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 $z = \frac{(\bar{x} - \mu)\sqrt{n}}{\sigma} = -2.1954$
- + 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 $p\text{-value} < \alpha$ 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.

Nhận xét 4.1: Nếu ta chỉ biết trung bình mẫu quan sát và kích thước mẫu quan sát thì ta sử dụng hàm `zsum.test` với cấu trúc

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

Ví dụ 4.2. Để đưa ra quyết định lựa chọn một trong hai quỹ đầu tư khác nhau, một nhà đầu tư quan sát lợi nhuận của quỹ đầu tư thứ nhất và thứ hai tương ứng trong $n_1 = 8$ năm và $n_2 = 11$ năm. Lợi nhuận trung bình của các quỹ đầu tư tính được là $\bar{x}_1 = 12,5\%$ và $\bar{x}_2 = 11,3\%$. Giả sử rằng lợi nhuận của các quỹ đầu tư trong 1 năm có phân phối chuẩn với độ lệch tiêu chuẩn $\sigma_1 = \sigma_2 = 5\%$. Một nhà môi giới tài chính nhận xét rằng quỹ đầu tư thứ nhất vượt trội hơn so với quỹ đầu tư thứ hai về lợi nhuận. Với mức ý nghĩa 2%, hãy kiểm định nhận xét trên.

Bước 1: Tóm tắt bài toán

- ✓ Gọi X_1, X_2 là lợi nhuận của quỹ đầu tư 1 và 2 trong một năm. Ta có, $X_1 \sim N(\mu_1, \sigma_1^2)$; $X_2 \sim N(\mu_2, \sigma_2^2)$ với $\sigma_1 = \sigma_2 = 5\%$ **đã biết**. Ta kiểm định **hai** giá trị trung bình μ_1, μ_2 với mức ý nghĩa $\alpha = 0,02$.
- ✓ Bài toán kiểm định: $H_0: \mu_1 = \mu_2$; $H_1: \mu_1 > \mu_2$

Bước 2: Hàm kiểm định trong R là hàm `zsum.test` với các tham số `mean.x = 12.5`; `mean.y = 11.3`; `n.x = 8`; `n.y = 11`; `sigma.x = sigma.y = 0.05`; `alternative = "greater"`; `conf.level = 0.98`

Bước 3: Thực hiện kiểm định trên R

```
>zsum.test(mean.x=0.125, mean.y=0.113, n.x=8, n.y=11, sigma.x=0.05, sigma.y=0.05, alternative = "greater", conf.level = 0.98)
```

Bước 4: Phân tích kết quả và kết luận

```
Two-sample z-Test
data: Summarized x and y
```

$z = 0.51651$, $p\text{-value} = 0.3028$

alternative hypothesis: true difference in means is greater than 0

98 percent confidence interval:

-0.03571477 NA

sample estimates:

mean of x mean of y

0.125 0.113

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

+ Giá trị thống kê

$$z = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = 0.51651;$$

+ Trị số-p của bài toán là

$$p\text{-value} = 0.0328;$$

- Kết luận: Vì $p\text{-value} > \alpha$ nên ta chưa bác bỏ giả thuyết gốc H_0 . Do đó, *chưa có cơ sở để nói quỹ đầu tư thứ nhất vượt trội hơn so với quỹ đầu tư thứ hai về lợi nhuận.*

4.1.2. Trường hợp phương sai chưa biết

- Thống kê kiểm định

✓ Với bài toán kiểm định 1 mẫu, thống kê kiểm định $T = \frac{(\bar{X} - \mu_0)\sqrt{n}}{s}$

✓ Với bài toán kiểm định 2 mẫu, thống kê kiểm định $T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$

- Giá trị thống kê

✓ Bài toán 1 mẫu $t = \frac{(\bar{x} - \mu_0)\sqrt{n}}{s}$

✓ Bài toán 2 mẫu $t = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$

- Cấu trúc hàm sử dụng trong R

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

trong đó

x véc tơ dữ liệu thứ nhất

y	véc tơ dữ liệu thứ hai, mặc định là NULL nếu có 1 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” .
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).

Ví dụ 4.3. 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

- ✓ Gọi X là thời gian truy cập internet của một khách hàng. Ta có $X \sim N(\mu, \sigma^2)$ với σ **chưa biết**. Ta kiểm định **một** giá trị trung bình μ , với mức ý nghĩa $\alpha = 0,05$
- ✓ Bài toán kiểm định: $H_0: \mu = 10$ (phút) $H_1: \mu \neq 10$ (phút)

Bước 2: Xác định hàm kiểm định trong R là hàm `t.test` với các tham số x là véc tơ tham số; $y = \text{NULL}$; `alternative = "two.side"`, `mu = 10`; `paired = FALSE` (mặc định);

`var.equal` (bỏ qua – vì 1 mẫu); `conf.level = 0.95`

Bước 3: Thực hiện kiểm định trên R

```
> setwd("D:/ThuchanhR/dulieuthuchanhR")
```

```
> 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

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 | $t = \frac{(\bar{x} - \mu_0)\sqrt{n}}{s} = 0.64648$ |
| + Bậc tự do df (<i>degree freedom</i>) | $df = n - 1 = 29;$ |
| + 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.

Nhận xét 4.2: Nếu ta chỉ biết trung bình mẫu quan sát và kích thước mẫu quan sát thì ta sử dụng hàm `tsum.test` với cấu trúc

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

Ví dụ 4.4. 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

- ✓ Gọi X_1, 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 khi σ_1, σ_2 **chưa biết** và mức ý nghĩa $\alpha = 0,01$.
- ✓ Bài toán kiểm định: $H_0: \mu_1 = \mu_2; \quad H_1: \mu_1 > \mu_2$

Bước 2: Xác định hàm kiểm định trong R là hàm `t.test` với các tham số `x, y` là hai véc tơ dữ liệu; `alternative = "greater"`; `mu = 0`; `var.equal = FALSE` (phương sai không

bằng nhau); `conf.level = 0.99`

Bước 3: Thực hiện kiểm định trên R

```
> setwd("D:/ThuchanhR/ dulieuthuchanhR")
> 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

```
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 . Vậy với $\alpha = 5\%$, có cơ sở để nói rằng tốc độ đường truyền trung bình của nhà cung cấp (1) cao hơn nhà cung cấp (2).

4.2. Bài toán kiểm định giá trị tỷ lệ

- Xét bài toán kiểm định giá trị tỉ lệ trong trường hợp 1 mẫu và 2 mẫu. Cụ thể, ta có các bài toán kiểm định

	Kiểm định 1 mẫu	Kiểm định hai mẫu
Giả thuyết gốc (H_0)	$p = p_0$	$p_1 = p_2$
Đối thuyết (H_1)	$p \neq p_0$	$p_1 \neq p_2$
	$p > p_0$	$p_1 > p_2$
	$p < p_0$	$p_1 < p_2$

- Thống kê kiểm định

✓ Với bài toán kiểm định 1 mẫu, thống kê kiểm định $T = \frac{(\hat{p}-p_0)\sqrt{n}}{\sqrt{p_0(1-p_0)}}$.

✓ Với bài toán kiểm định 2 mẫu, thống kê kiểm định $T = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$.

- Giá trị thống kê

✓ Bài toán 1 mẫu $t = \frac{(f-p_0)\sqrt{n}}{\sqrt{p_0(1-p_0)}}$ với $f = \frac{m}{n}$

✓ Bài toán 2 mẫu $t = \frac{f_1 - f_2}{\sqrt{f(1-f)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$ với $f_1 = \frac{m_1}{n_1}$; $f_2 = \frac{m_2}{n_2}$; $f = \frac{m_1 + m_2}{n_1 + n_2}$

- Điều kiện kích thước mẫu

✓ Bài toán 1 mẫu `correct = False` nếu $np_0 \geq 5$, $n(1-p_0) \geq 5$

✓ Bài toán 2 mẫu `correct = False` nếu

$$n_1 f_1 \geq 10, n_1(1-f_1) \geq 10; n_2 f_2 \geq 10; n_2(1-f_2) \geq 10$$

- Cấu trúc hàm sử dụng trong R

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

trong đó

x	véc tơ chỉ số lần “thành công” trong mỗi mẫu.
n	véc tơ chỉ số lần thử nghiệm trong mỗi mẫu
p	véc tơ chỉ xác suất thành công, có độ dài bằng số mẫu được chỉ định bởi x và các phần tử nằm trong khoảng từ 0 đến 1 ($p = p_0$ - ở 1 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.

`conf.level` độ tin cậy cho khoảng tin cậy được trả về, nằm trong khoảng (0; 1).

Ví dụ 4.5: 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

- ✓ Gọi p là tỉ lệ hộ gia đình trong khu vực dân cư sử dụng dịch vụ Internet của công ty được nói tới. Bài toán kiểm định **một** giá trị tỉ lệ p với mức ý nghĩa $\alpha = 0,05$.
- ✓ Bài toán kiểm định: $H_0: p = 0,7$ $H_1: p < 0,7$

Bước 2: Xác định hàm kiểm định trong R là hàm `prop.test` với các tham số $x = 125$;

$n = 200$; $p = 0,7$; `alternative = "less"`; `conf.level = 0,95`; `correct = False`

Bước 3: Thực hiện kiểm định trên R

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

Bước 4: Phân tích kết quả và kết luận

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 $p\text{-value} < \alpha$ nên ta bác bỏ giả thuyết gốc H_0 . Do đó với mức ý nghĩa $\alpha = 5\%$, 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%.

Ví dụ 4.6: 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

✓ Gọi p_1, p_2 là tỉ lệ trẻ em mắc tật khúc xạ ở thành phố và nông thôn. Ta kiểm định hai giá trị trung bình p_1, p_2 với mức ý nghĩa $\alpha = 0,01$.

✓ Bài toán kiểm định: $H_0: p_1 = p_2; \quad H_1: p_1 > p_2$

Bước 2: Xác định hàm kiểm định trong R là hàm `prop.test` với các tham số $x = (80, 36)$;

`n = c(320, 240); alternative = "greater"; conf.level = 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

```
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 $p\text{-value} < \alpha$ 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.

C. Bài tập

Bài 1. Các tua bin gió sử dụng tại Việt Nam được thiết kế phát điện ở tốc độ gió từ 5 m/s. Để đánh giá tính hiệu quả của một dự án xây dựng nhà máy điện gió tại Bạc Liêu, người ta khảo sát 20 điểm khác nhau xung quanh khu vực lập dự án. Giả sử tốc độ gió tại khu vực này tuân theo luật phân phối chuẩn với độ lệch tiêu chuẩn 2,2 m/s. Xét bài toán kiểm định để trả lời câu hỏi: Dự án có hiệu quả không? Nghĩa là có cơ sở để kết luận tốc độ gió trung bình của khu vực được xét có lớn hơn 5 m/s hay không? Mức ý nghĩa $\alpha = 10\%$.

Bài 2. Hai dây chuyền sản xuất cùng một loại linh kiện bán dẫn. Có ý kiến cho rằng tỉ lệ phế phẩm của hai dây chuyền bằng nhau. Kiểm tra một số sản phẩm của mỗi dây chuyền ta có bảng số liệu sau:

Dây chuyền	Số sản phẩm kiểm tra	Số phế phẩm
I	200	27
II	150	13

Với mức ý nghĩa 0,05 hãy đánh giá ý kiến trên.

Bài 3. Một kĩ sư đưa ra một quy trình sản xuất mới để giảm tỉ lệ phế phẩm của nhà máy.

Kiểm tra về hai quy trình sản xuất ta có bảng số liệu sau

Quy trình	Số sản phẩm kiểm tra	Số phế phẩm
Cũ	250	33
Mới	350	18

Với mức ý nghĩa 0,05 hãy đánh giá ý kiến của kĩ sư trên.

Bài 4. Điểm thi môn Xác suất thống kê của sinh viên có phân phối chuẩn. Kiểm tra điểm thi của sinh viên hai khối kinh tế và kĩ thuật, ta thu được bảng số liệu sau:

Khối	Kích thước mẫu	Trung bình mẫu	Phương sai mẫu
Kinh tế	$n_1 = 50$	$\bar{x}_1 = 7,24$	$s_1^2 = 1,65$
Kĩ thuật	$n_2 = 50$	$\bar{x}_2 = 6,78$	$s_2^2 = 0,94$

Có thể kết luận điểm thi trung bình môn xác suất thống kê của sinh viên khối kinh tế cao hơn của sinh viên khối kĩ thuật hay không? Mức ý nghĩa 10%.

Bài 5. Giảng viên môn Xác suất thống kê muốn chọn phần mềm thống kê để sử dụng để giới thiệu cho sinh viên. Sau khi xem xét, người đó phân vân giữa: (1) - phần mềm điều

khuyến bằng menu, và (2) - phần mềm bảng tính. Để quyết định, người đó yêu cầu hai nhóm sinh viên giải quyết một bài toán thống kê, mỗi nhóm chọn một phần mềm, kèm theo hướng dẫn sử dụng. Dữ liệu về thời gian trung bình (phút) và độ lệch tiêu chuẩn để hai nhóm sinh viên hoàn thành bài tập được cho sau đây:

$$n_1 = 54; \bar{x}_1 = 64,52; s_1 = 25,14; n_2 = 36; \bar{x}_2 = 56,32; s_2 = 11,53.$$

Có sự khác biệt về thời gian cần thiết để sinh viên học cách sử dụng hai phần mềm này hay không? Cho mức ý nghĩa $\alpha = 2\%$.

Bài 6. Cơ quan quản lý thị trường cho rằng có 10% mũ bảo hiểm xe máy không đạt tiêu chuẩn chất lượng. Người ta kiểm tra ngẫu nhiên 150 mũ bảo hiểm xe máy, thấy có 22 chiếc không đạt tiêu chuẩn. Với mức ý nghĩa 0,05 hãy kết luận về tính chính xác của ý kiến trên.

Bài 7. Kiểm tra mức tiêu thụ nhiên liệu của 22 xe máy cùng loại sử dụng bộ chế hòa khí thông thường, người ta thu được mức tiêu thụ nhiên liệu trung bình là $\bar{x}_1 = 31,7\text{km/l}$ với độ lệch tiêu chuẩn $s_1 = 2,7\text{ km/l}$. Trong các xe máy đó chọn ra ngẫu nhiên 15 xe và thay vào bộ chế hòa khí mới nghiên cứu chế tạo. Kiểm tra cho thấy mức tiêu thụ nhiên liệu trung bình của các xe này là $\bar{x}_2 = 31,7\text{ km/l}$ với độ lệch tiêu chuẩn $s_2 = 2,2\text{ km/l}$. Với mức ý nghĩa 10%, có thể kết luận bộ chế hòa khí mới tốt hơn hay không? Cho rằng mức tiêu thụ nhiên liệu của xe máy có phân phối chuẩn với phương sai bằng nhau.

Bài 8. Một công ty bảo hiểm xem xét đến yếu tố quãng đường di chuyển mỗi năm của lái xe khi xây dựng một sản phẩm bảo hiểm ô tô. Để xác định xem giới tính ảnh hưởng như thế nào đến yếu tố này, người ta khảo sát 50 lái xe nam và 50 lái xe nữ. Quãng đường di chuyển trung bình (nghìn kilomet) đối với tài xế nam và nữ tương ứng là: $\bar{x}_1 = 10,23$, $\bar{x}_2 = 9,66$. Công ty bảo hiểm tin rằng quãng đường di chuyển của lái xe nam cao hơn của lái xe nữ. Hãy kiểm định ý kiến trên với mức ý nghĩa 10%. Giả sử rằng quãng đường đi được mỗi năm của tài xế nam và nữ có phân phối chuẩn với độ lệch tiêu chuẩn tương ứng là $\sigma_1 = 3,6$; $\sigma_2 = 2,5$.

Bài 9. Bộ phận bảo trì tại một trạm bơm báo cáo rằng tỉ lệ số cảnh báo khẩn cấp hiện tại đã tăng lên so với trước năm 2019 và sử dụng nhận định đó làm cơ sở để đòi tăng thêm nhân sự. Nhà quản lý thu thập dữ liệu về các lần cảnh báo kể từ tháng 01 năm 2019 và thấy rằng có 59 lần trong số 150 lần cảnh báo là thực sự khẩn cấp. Giai đoạn trước 2019, trong 100 lần cảnh báo được theo dõi thì có 35 lần thực sự khẩn cấp và số nhân sự hiện giờ là đủ

để hoàn thành công việc. Với mức ý nghĩa 1%, có thể kết luận tuyên bố của bộ phận bảo trì là chính xác?

Bài 10. 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 kỹ 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 11. Nghiên cứu cho thấy, tai nạn giao thông tại những nút giao có đèn tín hiệu đếm lùi xảy ra nhiều hơn tại những nút giao có đèn tín hiệu không có số. Nguyên nhân là do người tham gia giao thông thường nhấn ga khi còn vài giây trước khi đèn xanh và cố gắng vượt qua khi gần đèn đỏ. Để minh chứng, người ta khảo sát hai giao lộ có lưu lượng giao thông tương đương (trong 1 giờ tại khung giờ cao điểm), một giao lộ gắn đèn tín hiệu đếm lùi (1) và một giao lộ gắn đèn tín hiệu không có số (2). Dữ liệu về số lượng người đi qua nút giao thông và số lượng người vượt đèn đỏ được cho dưới đây: $n_1 = 747$; $m_1 = 194$; $n_2 = 628$; $m_2 = 45$. Với mức ý nghĩa 5%, có thể kết luận tỉ lệ người vượt đèn đỏ tại nút giao có đếm số cao hơn tại nút giao không có đếm số hay không?

Bài 12. Chủ một cửa hàng muốn áp dụng phương thức thanh toán điện tử. Vì phải chi trả các khoản phí giao dịch nên phương pháp thanh toán mới chỉ hiệu quả nếu doanh số trung bình mỗi ngày lớn hơn 5 triệu đồng. Để đánh giá tính hiệu quả, chủ cửa hàng thống kê doanh số của 30 ngày gần nhất và tính được doanh số trung bình trong khoảng thời gian này là 5,7 triệu đồng một ngày. Biết rằng, doanh số mỗi ngày của cửa hàng có phân phối chuẩn với độ lệch tiêu chuẩn 2,1 triệu đồng. Với mức ý nghĩa 5%, hãy kiểm định tính hiệu quả của phương pháp thanh toán điện tử nếu áp dụng cho cửa hàng.

Bài 13. Nhiều lái xe buýt ở Hà Nội đồng ý rằng tình trạng giao thông ngày càng kém. Để đánh giá ý kiến đó, người ta đo thời gian chậm giờ của 20 chuyến xe buýt được chọn ngẫu nhiên trong giờ cao điểm và tính được thời gian chậm trễ trung bình là 45 phút. Các chuyên gia giao thông xác định rằng thời gian chậm trễ trung bình của xe buýt trong giờ cao điểm hai năm trước là 35 phút. Giả sử thời gian chậm giờ của xe buýt trong giờ cao điểm có phân phối chuẩn với độ lệch tiêu chuẩn 15 phút. Kiểm định xem tình trạng giao thông hiện tại có kém hơn hai năm trước hay không? Mức ý nghĩa 5%.

Bài 14. Một người đăng ký lắp đặt mạng Internet từ một nhà cung cấp. Gói cước anh ta lựa chọn được quảng cáo là có tốc độ 55Mbps. Để kiểm tra quảng cáo có chính xác hay không, người đó sử dụng ứng dụng Speedtest để đo tốc độ đường truyền (đơn vị: Mbps) tại một số thời điểm trong ngày. Dữ liệu thu được như sau:

48,7 43,4 46,6 36,6 35,2 41,0 44,9 30,9 46,4 30,7 43,5

Giả thiết rằng tốc độ đường truyền có phân phối chuẩn với độ lệch tiêu chuẩn 5,0 Mbps. Với mức ý nghĩa 5%, có thể kết luận tốc độ đường truyền trung bình trên thực tế thấp hơn 45 Mbps hay không?

Bài 15. Một máy sản xuất khi hoạt động bình thường thì khối lượng của sản phẩm có phân phối chuẩn với giá trị trung bình là 100(g), và độ lệch tiêu chuẩn bằng 2 (g). Sau một thời gian vận hành, người ta nghi ngờ khối lượng sản phẩm có xu hướng giảm xuống. Kiểm tra ngẫu nhiên 25 sản phẩm, dữ liệu về khối lượng sản phẩm thu được như sau:

Khối lượng	94-96	96-98	98-100	100-102	102-104
Số sản phẩm	2	5	7	8	3

Với mức ý nghĩa 5% hãy kết luận về điều nghi ngờ trên.

Bài 16. Một công ty công nghệ chuyên về nền tảng wifi-marketing: khai thác quảng cáo trên các hạ tầng wifi có sẵn. Doanh thu từ quảng cáo sẽ được chia sẻ giữa công ty và các đơn vị cung cấp hạ tầng internet. Theo đánh giá của các chuyên gia, công nghệ wifi-marketing sẽ có hiệu quả nếu có từ 10% khách hàng chạm vào quảng cáo khi sử dụng dịch vụ mạng internet miễn phí. Người ta khảo sát một mẫu gồm 325 khách hàng sử dụng wifi miễn phí, thấy có 49 khách chạm vào quảng cáo. Với mức ý nghĩa 10%, hãy kết luận về tính hiệu quả của công nghệ wifi-marketing do công ty cung cấp.

Bài 17. Cơ quan quản lý thị trường có vai trò kiểm soát các công ty sản xuất, đảm bảo sản phẩm của họ phải đóng gói đúng nội dung ghi trên bao bì. Gần đây, có nhiều khách hàng phản ánh về một sản phẩm nước đóng chai có dung tích thấp hơn 500 ml như ghi trên nhãn mác. Để kiểm tra, đơn vị quản lý thị trường lấy ngẫu nhiên 20 chai nước và dữ liệu đo dung tích nước được cho dưới đây:

475 465 500 486 503 479 501 466 488 476
485 479 473 502 505 489 474 485 492 486

Giả sử rằng dung tích nước đóng chai có phân phối chuẩn với độ lệch tiêu chuẩn 12 ml. Chọn mức ý nghĩa $\alpha = 10\%$.

Bài 18. AWING là một công ty công nghệ chuyên về nền tảng quảng cáo dựa trên chia sẻ hệ thống Wi-Fi miễn phí. Công ty 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:

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%?

Bài 19. Khảo sát ngẫu nhiên 18 sinh viên tại phòng tập gym nằm trong khuôn viên một trường đại học và hỏi về thời gian sử dụng phòng tập mỗi ngày. Dữ liệu cho thấy thời gian trung bình sử dụng phòng tập của nhóm sinh viên này là 43,56 phút và độ lệch tiêu chuẩn là 7,2 phút. Có thể kết luận thời gian sử dụng phòng tập gym trung bình của mỗi sinh viên là 45 phút/ngày hay không? Mức ý nghĩa $\alpha = 0.05$. Giả sử thời gian sử dụng phòng tập có phân phối chuẩn.

Bài 20. Độ nhớt (đơn vị: mPas) của hai loại dầu máy ô tô được đo đạc và cho kết quả sau

Loại dầu I	10.62	10.58	10.33	10.72	10.44	10.74
Loại dầu II	10.50	10.52	10.58	10.62	10.55	10.51

Hãy kiểm định giả thuyết rằng độ nhớt trung bình của hai loại dầu máy là như nhau. Cho biết độ nhớt của hai loại dầu máy có phân phối chuẩn với phương sai bằng nhau và mức ý nghĩa $\alpha = 1\%$.

Bài 5: Hồi quy và tương quan

A. Mục tiêu

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

- Xây dựng hàm hồi quy tuyến tính
- Vẽ biểu đồ phân tán về dữ liệu, vẽ đường hồi quy
- Tính toán và giải thích hệ số tương quan
- Đưa ra khoảng tin cậy cho giá trị dự báo

B. Nội dung

5.1. Biểu đồ phân tán và hệ số tương quan

Cho n điểm quan sát $(x_1, y_1); (x_2, y_2); \dots; (x_n, y_n)$ của hai biến x và y .

- *Hệ số tương quan* (coefficient of correlation), kí hiệu r , là đại lượng vô hướng dùng để đánh giá mối quan hệ tuyến tính giữa hai biến số, như giữa độ tuổi (x) và cholesterol (y). Hệ số tương quan có giá trị từ -1 đến 1 . Hệ số tương quan bằng 0 (hay gần 0) có nghĩa là hai biến số không có tuyến tính; ngược lại nếu hệ số gần bằng -1 hay 1 có nghĩa là hai biến số có một mối liên hệ tuyến tính.
- Công thức để đánh giá hệ số tương quan

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

- Trong R để tính hệ số tương quan ta dùng hàm

cor(x, y)

- Để vẽ biểu đồ phân tán, ta dùng hàm

plot(x, y) hoặc **plot(y~x)**

Ta có thể thêm các tham số `pch`, `color` để biểu đồ được đẹp hơn.

Ví dụ 5.1. Các nhà nghiên cứu đo lường độ cholesterol trong máu của 18 đối tượng nam.

Tỉ

trọng cơ thể (*body mass index - BMI*) cũng được ước tính cho mỗi đối tượng bằng công thức

lấy trọng lượng (kg) chia cho chiều cao (m) bình phương. Kết quả đo lường như sau:

Mã ID id	Độ tuổi age	BMI bmi	Cholesterol cholesterol
1	46	25.4	3.5
2	20	20.6	1.9
3	52	26.2	4.0
4	30	22.6	2.6
5	57	25.4	4.5
6	25	23.1	3.0
7	28	22.7	2.9
8	36	24.9	3.8
9	22	19.8	2.1
10	43	25.3	3.8
11	57	23.2	4.1
12	33	21.8	3.0
13	22	20.9	2.5
14	63	26.7	4.7
15	40	26.4	3.2
16	48	21.2	4.2
17	28	21.2	2.3
18	49	22.8	4.0

- Vẽ biểu đồ phân tán giữa độ tuổi và chỉ số bmi, độ tuổi và nồng độ cholesterol
- Tìm hệ số tương quan giữa độ tuổi và nồng độ cholesterol và hệ số tương quan giữa độ tuổi và chỉ số BMI.

Bước 1: Nhập dữ liệu

```
> age<-
c(46,20,52,30,57,25,28,36,22,43,57,33,22,63,40,48,28,49)
> bmi <-
c(25.4,20.6,26.2,22.6,25.4,23.1,22.7,24.9,19.8,25.3,
23.2, 21.8,20.9,26.7,26.4,21.2,21.2,22.8)
> cholesterol<-
c(3.5,1.9,4.0,2.6,4.5,3.0,2.9,3.8,2.1,3.8,4.1,
```

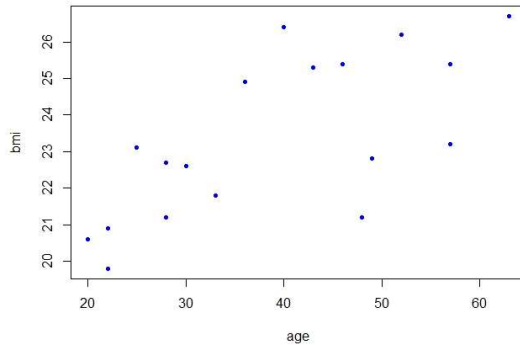
```
3.0, 2.5, 4.6, 3.2, 4.2, 2.3, 4.0)
```

```
> data <- data.frame(age, bmi, cholesterol)
```

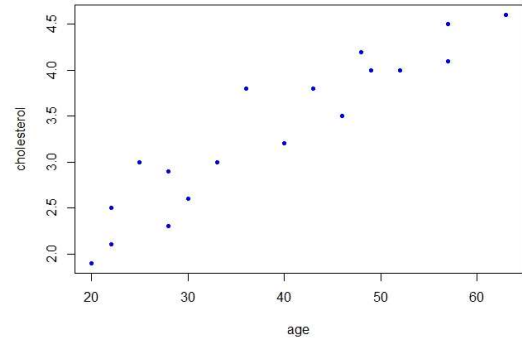
Bước 2: Vẽ biểu đồ phân tán

```
> plot(bmi ~ age, pch=20, col='blue')
```

```
> plot(cholesterol ~ age, pch=20, col='blue')
```



H7a. Biểu đồ phân tán giữa độ tuổi và chỉ số bmi



H7b. Biểu đồ phân tán giữa độ tuổi và nồng độ cholesterol

Hình 7. Biểu đồ phân tán

Bước 3: Tìm hệ số tương quan

```
> cor(age, cholesterol)
```

```
[1] 0.9367261
```

```
> cor(age, bmi)
```

```
[1] 0.6914202
```

5.2. Hàm hồi quy tuyến tính

- Hàm hồi quy tuyến tính thực nghiệm có dạng

$$y = \beta_0 + \beta_1 x$$

- Cho n điểm quan sát $(x_1, y_1); (x_2, y_2); \dots; (x_n, y_n)$. Các hệ số hồi quy β_0, β_1 nhận được bằng cách cực tiểu hóa hàm E

$$E = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$$

- Ước lượng của các hệ số của mô hình hồi quy tuyến tính là

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

- Để tìm hàm hồi quy tuyến tính, ta sử dụng hàm

lm(y~x)

- Để vẽ đường hồi quy tuyến tính ta dùng hàm

abline(object)

Ví dụ 5.2: Xét Ví dụ 5.1, tìm hàm hồi quy tuyến tính của hàm lượng cholesterol theo độ tuổi và vẽ đường hồi quy.

Bước 1: Tìm hàm hồi quy tuyến tính

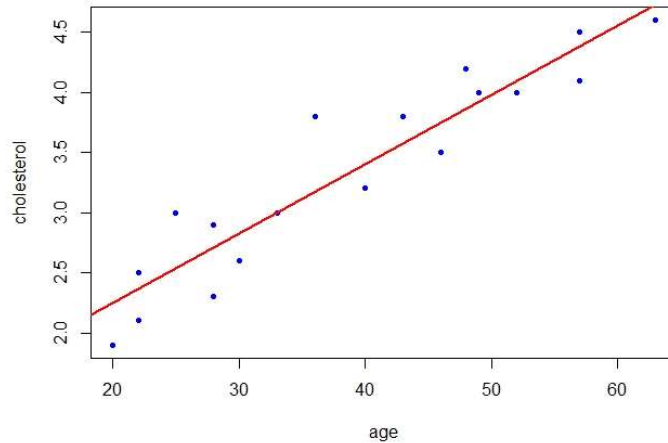
```
> reg <- lm(cholesterol~age)
> reg
Call:
lm(formula = cholesterol ~ age)
Coefficients:
(Intercept)          age
    1.08922      0.05779
```

Trong lệnh trên, “cholesterol ~ age” có nghĩa là mô tả *cholesterol* là một hàm số của *age*. Kết quả tính toán của lm cho thấy $\beta_0 = 1,08922$ và $\beta_1 = 0.05779$. Nói cách khác, với hai thông số này, chúng ta có thể ước tính độ *cholesterol* cho bất cứ độ tuổi nào trong khoảng tuổi của mẫu bằng phương trình tuyến tính:

$$y = 1,08922 + 0,05779 \cdot x$$

Bước 2: Vẽ đường hồi quy

```
> plot(cholesterol ~ age, pch=20, col='blue')
> abline(reg,col='red',lwd=2)
```



Hình 8. Đường biểu diễn mối liên hệ giữa cholesterol và độ tuổi (age)

5.3. Khoảng tin cậy cho giá trị dự báo

- Giá trị dự báo tại điểm x_0 là: $y_0 = \beta_0 + \beta_1 x_0$.
- Công thức tính khoảng tin cậy cho giá trị dự báo tại x_0 với độ tin cậy γ là

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

trong đó, $s = \sqrt{\frac{1}{n-2} \left(S_{yy} - \frac{S_{xy}^2}{S_{xx}} \right)}$ là sai số trung bình của mô hình

với $S_{xy} = \sum_{i=1}^n x_i y_i - \frac{1}{n} \sum_{i=1}^n x_i \cdot \sum_{i=1}^n y_i$; $S_{xx} = \sum_{i=1}^n x_i^2 - \frac{1}{n} \left(\sum_{i=1}^n x_i \right)^2$;

$$S_{yy} = \sum_{i=1}^n y_i^2 - \frac{1}{n} \left(\sum_{i=1}^n y_i \right)^2$$

- Để thực hành trong R, ta dùng hàm

predict(object, newdata, interval, level)

Ví dụ 5.3: Tìm khoảng tin cậy 95% cho nồng độ cholesterol cho độ tuổi 47.

```
>predict(reg, newdata=data.frame(age=47), interval='prediction', level=0.95)
```

```
      fit      lwr      upr
1 3.805272 3.13949 4.471053
```

C. Bài tập

Bài 1. Kiểm tra về độ xốp ($y, \%$) của một loại đá được nung trong các điều kiện nhiệt độ (x , nhân 100 độ C) khác nhau ta có bảng số liệu sau:

Nhiệt độ	11	12	13	14	15	16	17
Độ xốp	30,8	28,5	25,8	27,4	20,5	19,8	18,6

- Tính hệ số tương quan giữa nhiệt độ và độ xốp.
- Tìm phương trình hồi quy tuyến tính của y theo x .
- Tìm khoảng tin cậy 98% cho độ xốp ở nhiệt độ 1800 độ C.
- Vẽ biểu đồ phân tán và đường hồi quy.

Bài 2. Chỉ số lạm phát có phản ánh tăng trưởng kinh tế? Dữ liệu dưới đây thể hiện chỉ số lạm phát và tốc độ tăng trưởng kinh tế trong giai đoạn 2010 - 2017 của Việt Nam:

Năm	2010	2011	2012	2013	2014	2015	2016	2017
Lạm phát (x)	7.78	13.62	8.19	4.77	3.31	2.05	1.83	1.41
Tăng trưởng (y)	6.4	6.2	5.3	5.4	6.0	6.7	6.2	6.8

- Tính hệ số tương quan giữa lạm phát và tốc độ tăng trưởng.
- Tìm phương trình hồi quy tuyến tính của y theo x .
- Tìm khoảng tin cậy 95% cho tăng trưởng ứng với lạm phát 1.3.
- Vẽ biểu đồ phân tán và đường hồi quy.

Bài 3. Dữ liệu dưới đây là mực nước (đơn vị: cm) thấp nhất của sông Hồng ghi nhận được tại trạm Hà Nội từ năm 2002 đến năm 2011:

Năm	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
Mực nước	257	234	186	158	136	112	80	66	10	10

- Tìm phương trình hồi quy tuyến tính của mực nước theo thời gian.
- Tìm khoảng tin cậy 95% cho mực nước năm 2015.
- Vẽ biểu đồ phân tán và đường hồi quy.

Bài 4. Tốc 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 (tấn/ha/năm) 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	38	78	55	84	52	111	94

- Tính hệ số tương quan giữa tốc độ xói mòn và độ dốc.

- b) Tìm phương trình hồi quy tuyến tính của tốc độ xói mòn theo độ dốc.
c) Vẽ biểu đồ phân tán và đường hồi quy.

Bài 5. Người ta tiến hành thử nghiệm các mẫu than đá khác nhau trong một khu vực địa chất và ghi lại dung trọng ($x, t/m^3$) của than đá và độ tro ($y, \%$) của các mẫu đó dưới đây

Dung trọng	9	20	20	17	24	24	25	25	30	30	36
Độ tro	1	1.4	1.5	1.3	1.6	1.4	1.6	1.7	1.6	2	1.9

- a) Tính hệ số tương quan thực nghiệm.
b) Tìm phương trình hồi quy tuyến tính của y theo x .
c) Vẽ biểu đồ phân tán và đường hồi quy.

Bài 6. Người ta lấy một mẫu thực nghiệm của đại lượng ngẫu nhiên hai chiều (X, Y) trong đó X là số tiền đầu tư và Y là doanh thu tương ứng của 7 dự án trong lĩnh vực cầu đường (tính theo nghìn tỷ đồng) và thu được kết quả

(3, 2; 4, 5); (3, 8; 4, 8); (3, 7; 4, 63); (6, 6; 9, 8); (7; 10, 2); (8, 5; 11, 6); (12; 14, 3)

- a) Tính hệ số tương quan thực nghiệm.
b) Tìm phương trình hồi quy tuyến tính của y theo x .
c) Vẽ biểu đồ phân tán và đường hồi quy.

Bài 7. Người ta lấy một mẫu thực nghiệm của đại lượng ngẫu nhiên hai chiều (X, Y) và thu được kết quả

X	4.15	4.46	4.65	4.98	5.12	5.25
Y	18.2	19.6	19.7	20.1	22.3	22.9

- a) Tính hệ số tương quan thực nghiệm.
b) Tìm phương trình hồi quy tuyến tính của y theo x .
c) Vẽ biểu đồ phân tán và đường hồi quy.

Bài 8. Để nghiên cứu về quan hệ giữa khối lượng đào đắp X (nghìn m^3) và thời gian thi công Y (giờ) người ta lấy một mẫu thực nghiệm và thu được kết quả

(10; 25); (12; 28); (11; 27); (9; 23); (9,5; 24); (8; 20); (12; 30); (8,5; 22)

- a) Tính hệ số tương quan thực nghiệm.
b) Tìm phương trình hồi quy tuyến tính của y theo x .
c) Vẽ biểu đồ phân tán và đường hồi quy.

Bài 9. Số liệu về số lượt nghe một bài hát của ca sĩ A sau khi bài hát được đưa lên youtube như sau

Ngày thứ (x)	1	2	3	4	5	6	7
Số lượt nghe (y)	2112	2523	2265	2032	1983	1928	1765

- Tính hệ số tương quan thực nghiệm.
- Tìm phương trình hồi quy tuyến tính của y theo x .
- Vẽ biểu đồ phân tán và đường hồi quy.

Bài 10. Số liệu về lượng vận chuyển của một công ty vận tải trong các năm qua (tính theo triệu tấn) là như sau

Năm	2007	2008	2009	2010	2011	2012	2013
Khối lượng	42	44	45.5	46	38.5	50	51

- Tìm phương trình hồi quy tuyến tính của khối lượng theo thời gian.
- Tìm khoảng tin cậy 95% cho khối lượng vận chuyển năm 2015.
- Vẽ biểu đồ phân tán và đường hồi quy.

Bài 11. Người ta lấy một mẫu thực nghiệm của đại lượng ngẫu nhiên (X, Y) trong đó X là số tháng được sử dụng của máy in và Y là số trang đã in (tính theo nghìn trang) của 8 máy in văn phòng và thu được kết quả

(8; 3,2), (10; 4,1), (11; 4,6), (14; 5,2), (18; 7,3), (24; 8,5), (21; 8,7), (15; 6,3)

- Tính hệ số tương quan thực nghiệm.
- Tìm phương trình hồi quy tuyến tính của y theo x .
- Vẽ biểu đồ phân tán và đường hồi quy.

Bài 12. Người ta lấy một mẫu thực nghiệm của đại lượng ngẫu nhiên hai chiều (X, Y) trong đó X là số tiền đầu tư và Y là doanh thu tương ứng của 7 dự án trong lĩnh vực cầu đường (tính theo nghìn tỷ đồng) và thu được kết quả

(2,3; 3,08), (4,5; 5,12), (3,7; 4,63), (7,1; 9,04), (12; 13,2), (8,5; 9,6), (10; 11,3)

- Tính hệ số tương quan thực nghiệm.
- Tìm phương trình hồi quy tuyến tính của y theo x .
- Vẽ biểu đồ phân tán và đường hồi quy.

Bài 13. Người ta lấy một mẫu thực nghiệm của đại lượng ngẫu nhiên (X, Y) trong đó X là số giờ vắng mặt trên lớp và Y là điểm thi của 7 sinh viên và thu được kết quả

(8; 6,1), (10; 6,0), (15; 5,5), (20; 4,2), (25; 1,3), (24; 3,5), (21; 2,7)

- a) Tính hệ số tương quan thực nghiệm.
- b) Tìm phương trình hồi quy tuyến tính của y theo x .
- c) Vẽ biểu đồ phân tán và đường hồi quy.

Bài 14. 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 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) Tính hệ số tương quan thực nghiệm.
- b) Tìm phương trình hồi quy tuyến tính của y theo x .
- c) Tìm khoảng tin cậy 95% cho thời gian đi làm với người có khoảng cách là 13 km.
- d) Vẽ biểu đồ phân tán và đường hồi quy.