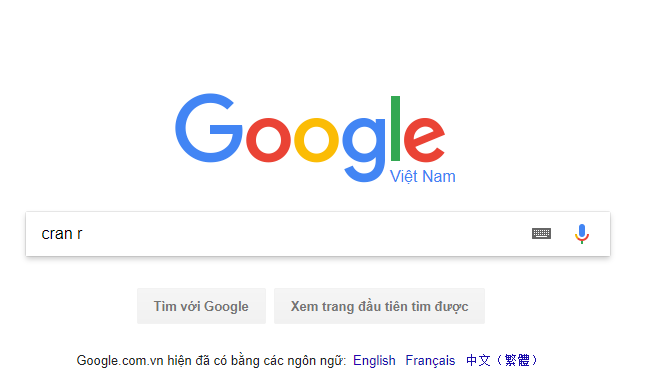
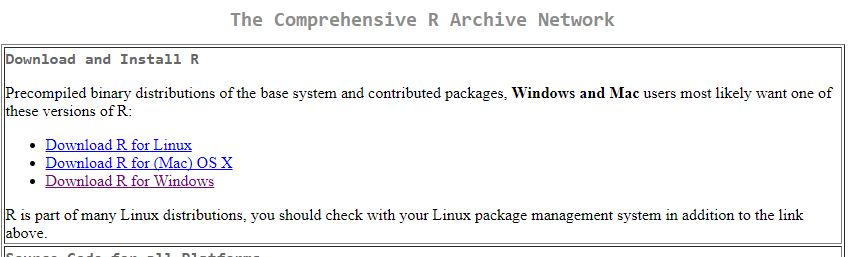
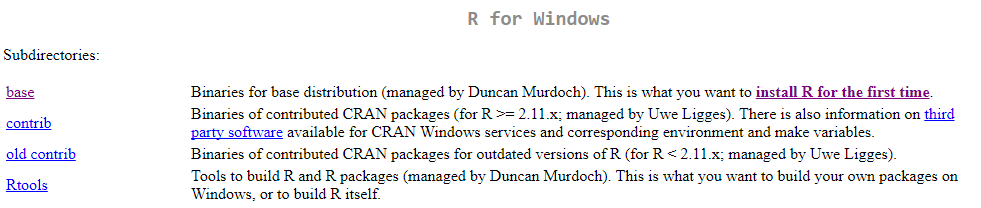
|  |  |
| --- | --- |
| Mục lục | Trang |
| Cài đặt phần mềm R cho máy tính nối mạng | 2 |
| I.Các PPXS thường gặp | 5 |
| II.Nhập dữ liệu, thống kê mô tả, vẽ biểu đồ | 6 |
| 1.Nhập dữ liệu vào R | 6 |
| 2.Thống kê mô tả (descriptive statistics, summary) | 9 |
| 3.Thống kê mô tả theo từng nhóm | 11 |
| 4. Biểu đồ cột: <barplot( ) | 12 |
| 5. Biểu đồ tròn: <pie( ) | 14 |
| III. Ước lượng và kiểm định thống kê | 15 |
| 1.Tính TB mẫu mean(x), độ lệch mẫu sd(x) | 15 |
| 2. Sử dụng hàm zsum.test kiểm định, tìm KTC | 15 |
| 3. Sử dụng hàm tsum.test kiểm định, tìm KTC | 21 |
| 4. Sử dụng hàm prop.test kiểm định, tìm KTC | 26 |
| IV. Hồi quy tuyến tính | 29 |
| 1.Ước lượng điểm cho các hệ số hồi quy | 29 |
| 2.Tương quan tuyến tính | 31 |
| 3.Ước lượng không chệch của phương sai | 32 |

**Cài đặt phần mềm R cho máy tính nối mạng.**

1. **Gõ từ khóa “cran r” trong google**

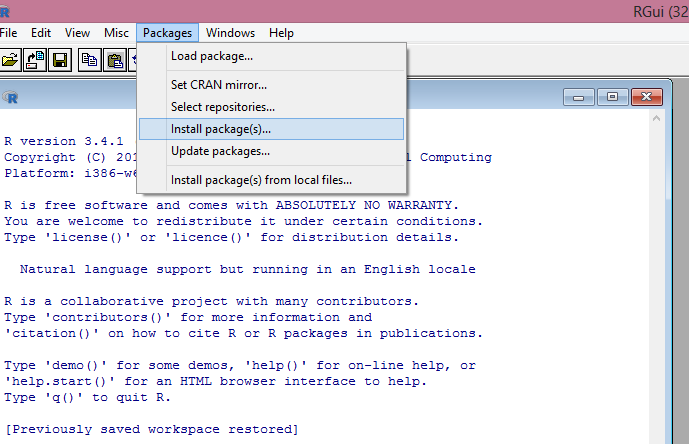
****

**2) Click vào “Download R for Windows”**

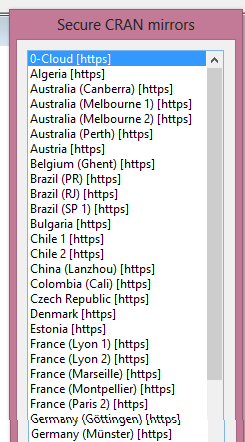
**3) Click vào “Install R for the first time”**

**4) Click vào “Download R 3.4.1 for Windows”**

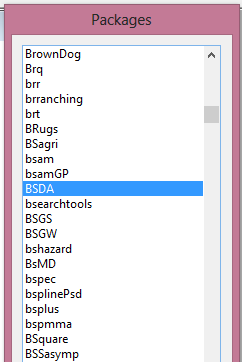
**5) Cài đặt R**

**6) Cài đặt gói BSDA**

**7)**

****

**8)**

****

**Kiểm tra cài đặt thành công:**

**> library(BSDA)**

**I. Các PPXS thường gặp**

**1.PP Nhị thức**

Cho BNN 



***Ví dụ:*** Xác suất để một bệnh nhân sống sót sau khi mắc một loại bệnh hiếm thấy về máu là 0,4. Nếu biết rằng đã có 15 người mắc loại bệnh này, tìm xác suất để

* 1. có đúng 6 người sống sót
  2. có ít nhất 10 người sống sót

Giải: a)



> dbinom(6,15,0.4)

[1] 0.2065976

b) Xác suất để có ít nhất 10 người sống sót là:



> 1-pbinom(9,15,0.4)

[1] 0.0338333

**2.PP Chuẩn**

Cho BNN 





 Cho **,** nếu k thỏa mãn thì 

****

***Ví dụ:*** Cho một phân phối chuẩn có 

a)Tìm xác suất để *X* nhận giá trị trong khoảng 45 và 62.

b) Tìm k để 

Giải: a)



> pnorm(62,50,10)-pnorm(45,50,10)

[1] 0.5763928

b)****

> qnorm(0.7,50,10)

[1] 55.24401

**3.PP Poisson**

BNN  là số lần 1 biến cố xảy ra trong khoảng thời gian nhất định





****

***Ví dụ 1:*** Trong suốt phép thử ở phòng thí nghiệm, trung bình số hạt phóng xạ đi qua một máy đếm trong một phần nghìn giây là 4. Xác suất để 6 hạt đi vào máy đếm trong vòng một phần nghìn giây cho trước là bao nhiêu?

Giải:

****

> dpois(6,4)

[1] 0.1041956

***Ví dụ 2:*** Số thùng dầu trung bình được đưa đến một cảng của thành phố nào đó trong mỗi ngày là 10. Năng lực bốc dỡ của cảng nhiều nhất là 15 thùng một ngày. Tính xác suất để một ngày cho trước có thùng dầu phải chở quay lại.

Giải:

Gọi *X* là số thùng dầu được đưa đến cảng trong một ngày.

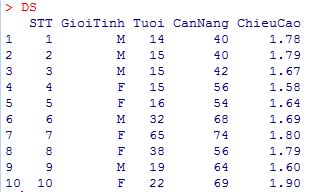


> 1-ppois(15,10)

[1] 0.0487404

**II. Nhập dữ liệu, thống kê mô tả, vẽ biểu đồ**

Trong phần này ta sẽ nhập dữ liệu và phân tích ví dụ cho bảng thống kê DS sau đây

****

**1. Nhập dữ liệu vào R**

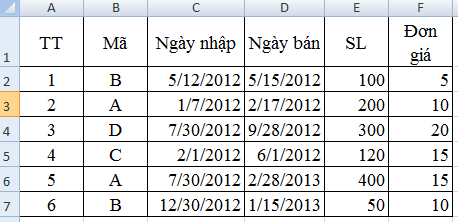
**a) Nhập số liệu từ Excel:** Đưa dữ liệu bảng excel có sẵn vào R

Để nhập số liệu từ phần mềm Excel, chúng ta cần tiến hành 2 bước

• Bước 1: Dùng lệnh “Save as” trong Excel và lưu file số liệu dưới dạng “csv”;

• Bước 2: Dùng lệnh “read.csv(“tenfile.csv”)” trong R để đưa bảng excel đó vào R.

Ví dụ: Ta muốn đưa file excel DonHang.xls dưới đây vào R



Ta làm như sau:

+ Dùng lệnh > setwd() trong R để kiểm tra thư mục R đang làm việc

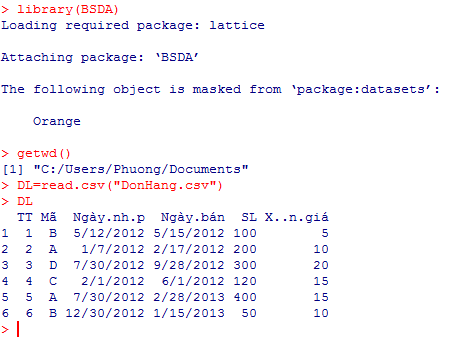
> setwd()

[1] "C:/Users/Phuong/Documents"

+Như vậy R đang làm việc với thư mục "C:/Users/Phuong/Documents", nên ta copy file excel DonHang.xls vào thư mục này.

+Mở file excel DonHang.xls và lưu dưới dạng đuôi .CSV(comma delimited).

+Đọc file DonHang.csv trong R

****

**b) Nhập số liệu trực tiếp dạng vectơ: c()**

Ví dụ: Cho mẫu số liệu

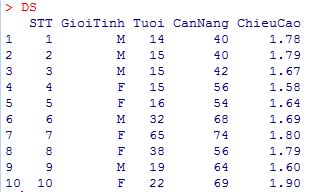
7.8, 6.6, 6.5, 7.4, 7.3, 7.0, 6.4, 7.1, 6.7, 7.6, 6.8

+Nhập dữ liệu:

> x=c(7.8, 6.6, 6.5, 7.4, 7.3, 7.0, 6.4, 7.1, 6.7, 7.6, 6.8)

**c) Nhập số liệu trực tiếp dạng bảng: edit(data.frame())**

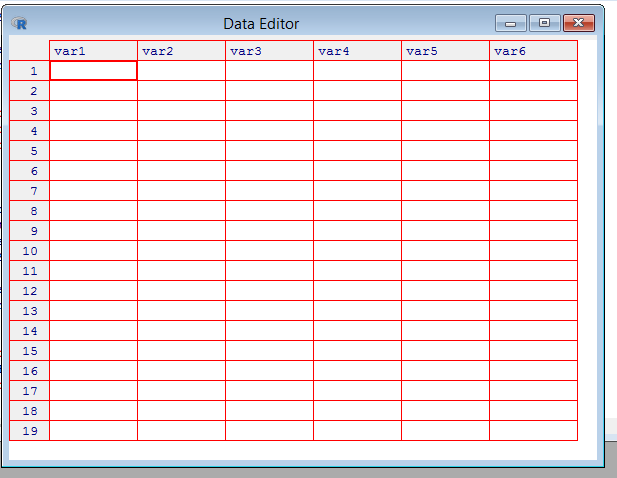
Ví dụ: Nhập bảng dữ liệu có tên là DS



-Ta làm như sau

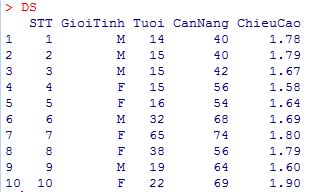
> DS=edit(data.frame())

Cửa sổ hiện ra



Nhấp chuột vào var1, var2,… nhập các trường STT, GioiTinh, Tuoi, CanNang, ChieuCao và nhập số liệu cho bảng, đóng cửa sổ nhập liệu.

-Xem lại bảng DS

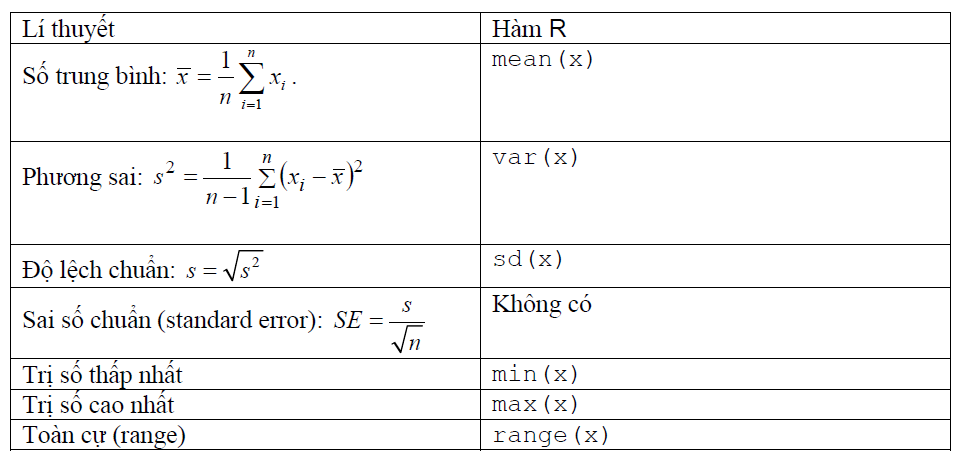


**Dưới đây ta sẽ xử lý số liệu thống kê cho bảng DS**

**2.Thống kê mô tả (descriptive statistics, summary)**

a) Cho một biến số chúng ta có thể tính toán một số chỉ số thống kê mô tả

như sau:

****

**Ví dụ**: **Để tìm giá trị phương sai mẫu của độ tuổi trong bảng trên**

+Gọi thư viện BSDA

+Attach bảng DS

+Tính phương sai

**> library(BSDA)**

**> attach(DS)**

**The following objects are masked from DS (pos = 3):**

**CanNang, ChieuCao, GioiTinh, STT, Tuoi**

**The following objects are masked from DS (pos = 4):**

**CanNang, ChieuCao, GioiTinh, STT, Tuoi**

**> var(Tuoi)**

**[1] 262.7667**

**b) Lệnh >summary**

Cho chúng ta tất cả thông tin thống kê về một biến số

**Ví dụ: Phân tích độ tuổi**

> summary(Tuoi)

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

14.0 15.0 17.5 25.1 29.5 65.0

Trong kết quả trên, có hai chỉ số “1st Qu” và “3rd Qu” có nghĩa là first quartile (tương

đương với vị trí 25%) và third quartile (tương đương với vị trí 75%) của một biến số.

Các chỉ số còn lại là Min, Trung vị, Trung bình mẫu, Max.

**Ví dụ: Phân tích cả bảng DS**

> summary(DS)

STT GioiTinh Tuoi CanNang

Min. : 1.00 Length:10 Min. :14.0 Min. :40.0

1st Qu.: 3.25 Class :character 1st Qu.:15.0 1st Qu.:45.0

Median : 5.50 Mode :character Median :17.5 Median :56.0

Mean : 5.50 Mean :25.1 Mean :56.3

3rd Qu.: 7.75 3rd Qu.:29.5 3rd Qu.:67.0

Max. :10.00 Max. :65.0 Max. :74.0

ChieuCao

Min. :1.580

1st Qu.:1.647

Median :1.735

Mean :1.724

3rd Qu.:1.790

Max. :1.900

**c) Lệnh >by(Tên bảng, trường , summary)**

Cho kết quả từng nhóm riêng biệt

**Ví dụ: Nếu chúng ta muốn kết quả cho từng nhóm nam và nữ riêng biệt**

> by(DS,GioiTinh,summary)

GioiTinh: F

STT GioiTinh Tuoi CanNang

Min. : 4.0 Length:5 Min. :15.0 Min. :54.0

1st Qu.: 5.0 Class :character 1st Qu.:16.0 1st Qu.:56.0

Median : 7.0 Mode :character Median :22.0 Median :56.0

Mean : 6.8 Mean :31.2 Mean :61.8

3rd Qu.: 8.0 3rd Qu.:38.0 3rd Qu.:69.0

Max. :10.0 Max. :65.0 Max. :74.0

ChieuCao

Min. :1.580

1st Qu.:1.640

Median :1.790

Mean :1.742

3rd Qu.:1.800

Max. :1.900

------------------------------------------------------------

GioiTinh: M

STT GioiTinh Tuoi CanNang ChieuCao

Min. :1.0 Length:5 Min. :14 Min. :40.0 Min. :1.600

1st Qu.:2.0 Class :character 1st Qu.:15 1st Qu.:40.0 1st Qu.:1.670

Median :3.0 Mode :character Median :15 Median :42.0 Median :1.690

Mean :4.2 Mean :19 Mean :50.8 Mean :1.706

3rd Qu.:6.0 3rd Qu.:19 3rd Qu.:64.0 3rd Qu.:1.780

Max. :9.0 Max. :32 Max. :68.0 Max. :1.790

**3.Thống kê mô tả theo từng nhóm**

**a)Lệnh > tapply(Trường, list(Trường), Thống kê mẫu)**

**Ví dụ: Tính trung bình chiều cao của mỗi nhóm nam và nữ giới**

**> tapply(ChieuCao, list(GioiTinh), mean)**

**F M**

**1.742 1.706**

Nếu muốn tính cho nhiều nhóm, ta chỉ cần thêm một biến số trong hàm list.

**Ví dụ: Tính trung bình chiều cao theo tuổi của mỗi nhóm nam và nữ giới**

**> tapply(ChieuCao, list(Tuoi,GioiTinh), mean)**

**F M**

**14 NA 1.78**

**15 1.58 1.73**

**16 1.64 NA**

**19 NA 1.60**

**22 1.90 NA**

**32 NA 1.69**

**38 1.79 NA**

**65 1.80 NA**

Nhận xét: Trong bảng DS không có người nào 14 tuổi giới tính F nên chiều cao tính ra NA(không xác định).

**4. Biểu đồ cột: <barplot( )**

**Usage**

barplot(height, ...)

## Default S3 method:

barplot(height, width = 1, space = NULL,

names.arg = NULL, legend.text = NULL, beside = FALSE,

horiz = FALSE, density = NULL, angle = 45,

col = NULL, border = par("fg"),

main = NULL, sub = NULL, xlab = NULL, ylab = NULL,

xlim = NULL, ylim = NULL, xpd = TRUE, log = "",

axes = TRUE, axisnames = TRUE,

cex.axis = par("cex.axis"), cex.names = par("cex.axis"),

inside = TRUE, plot = TRUE, axis.lty = 0, offset = 0,

add = FALSE, args.legend = NULL, ...)

**Ta sẽ thực hành với bảng DS ở trên**

**Ví dụ 1: Từ bảng DS, lọc và vẽ biểu đồ cột tần suất của giới tính**

**+Gọi bảng DS**

> attach(DS)

The following objects are masked from DS (pos = 3):

CanNang, ChieuCao, GioiTinh, STT, Tuoi

The following objects are masked from DS (pos = 4):

CanNang, ChieuCao, GioiTinh, STT, Tuoi

The following objects are masked from DS (pos = 5):

CanNang, ChieuCao, GioiTinh, STT, Tuoi

**+Lọc dữ liệu giới tính**

> table(GioiTinh)

GioiTinh

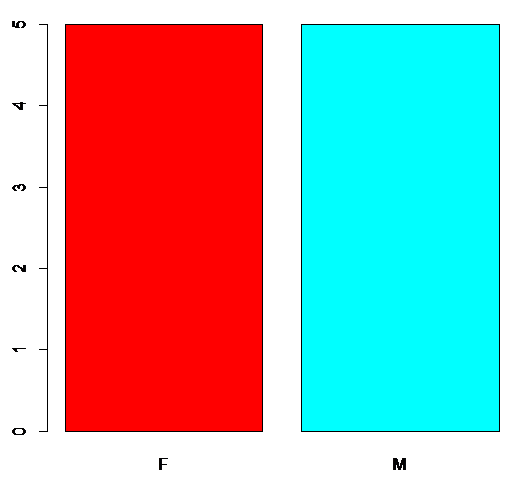
F M

5 5

**+Vẽ**

> barplot(table(GioiTinh))

> barplot(table(GioiTinh),col=rainbow(2))



Nhận xét: Biểu đồ biểu thị 5 Nam và 5 Nữ.

**Ví dụ 2: Từ bảng DS, lọc và vẽ biểu đồ tần suất của giới tính theo cân nặng**

> table(CanNang,GioiTinh)

GioiTinh

CanNang F M

40 0 2

42 0 1

54 1 0

56 2 0

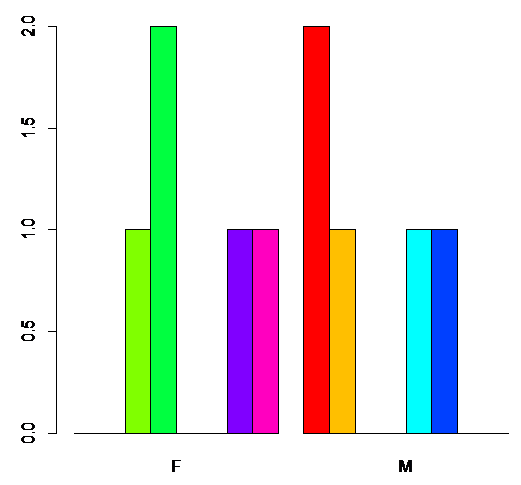
64 0 1

68 0 1

69 1 0

74 1 0

> barplot(table(CanNang,GioiTinh),beside=TRUE,col=rainbow(8))



Nhận xét: Biểu đồ biểu thị Nữ có 4 loại cân nặng; gồm có 54, 69, 74 mỗi hạng cân 1 người; hạng cân 56 có 2 người.

Tương tự cho Nam.

**5. Biểu đồ tròn: <pie( )**

**Usage**

pie(x, labels = names(x), edges = 200, radius = 0.8,

clockwise = FALSE, init.angle = if(clockwise) 90 else 0,

density = NULL, angle = 45, col = NULL, border = NULL,

lty = NULL, main = NULL, ...)

**Ví dụ: Từ bảng DS, lọc và vẽ biểu đồ tròn tần suất của chiều cao**

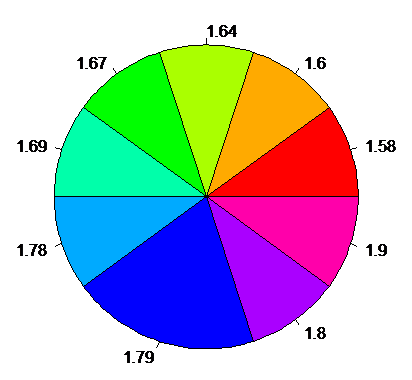
> table(ChieuCao)

ChieuCao

1.58 1.6 1.64 1.67 1.69 1.78 1.79 1.8 1.9

1 1 1 1 1 1 2 1 1

> pie(table(ChieuCao),col=rainbow(9))



**III. Ước lượng và kiểm định thống kê**

1. **Tính TB mẫu mean(x), độ lệch mẫu sd(x)**

**Ví dụ**: Cho mẫu số liệu

7.8, 6.6, 6.5, 7.4, 7.3, 7.0, 6.4, 7.1, 6.7, 7.6, 6.8

Tính TB mẫu mean(x), độ lệch mẫu sd(x)

+Gọi thư viện BSDA:

> library(BSDA)

Loading required package: lattice

Attaching package: ‘BSDA’

The following object is masked from ‘package:datasets’:

Orange

+Nhập dữ liệu:

> x=c(7.8, 6.6, 6.5, 7.4, 7.3, 7.0, 6.4, 7.1, 6.7, 7.6, 6.8)

> mean(x)

[1] 7.018182

> sd(x)

[1] 0.4643666

**2. Sử dụng hàm zsum.test kiểm định, tìm KTC cho các trường hợp sau**

**+Kiểm định, tìm KTC về một trung bình, khi  đã biết**

**+Kiểm định, tìm KTC về một trung bình, khi  chưa biết, cỡ mẫu lớn (n  30)**

Trường hợp này, vì cỡ mẫu lớn (*n * 30) nên ta có thể coi  và đưa về như trường hợp  đã biết.

**+Kiểm định, tìm KTC về hiệu hai trung bình, khi phương sai  đã biết**

**+Kiểm định, tìm KTC về hiệu hai trung bình, khi phương sai chưa biết : ta xấp xỉ phân phối chuẩn.**

**Usage**

zsum.test(mean.x, sigma.x = NULL, n.x = NULL, mean.y = NULL,

sigma.y = NULL, n.y = NULL, alt="t", mu = 0,

conf.level = 0.95)

Trong đó:

alt="t" (two-side): kiểm định 2 phía và cho ước lượng KTC

alt="g" (greater): kiểm định lớn hơn

alt="l" (less): kiểm định nhỏ hơn

**Chú ý: Nếu dữ liệu cho ở dạng số liệu quan sát (dữ liệu sơ cấp), ta có thể sử dụng hàm z.test (xem VD3).**

***Ví dụ 1(KĐ 2 phía) :***Một nhà sản xuất khẳng định rằng khối lượng trung bình của dây câu có thể chịu là 8 kg, biết khối lượng dây có thể chịu có PP chuẩn với độ lệch chuẩn của tổng thể là 0,5 kg. Để kiểm định giả thiết  = 8 kg với đối thiết  ≠ 8 kg, 50 dây ngẫu nhiên được kiểm tra và khối lượng trung bình dây có thể chịu là 7,8 kg. Hãy kiểm định khẳng định của nhà sản xuất với mức ý nghĩa 0,01.

**GIẢI:**

+Từ giả thiết ta có: 

Đây là bài toán kiểm định giả thiết cho một giá trị trung bình, khi  đã biết

Đặt bài toán: .

+Gọi thư viện BSDA:

> library(BSDA)

Loading required package: lattice

Attaching package: ‘BSDA’

The following object is masked from ‘package:datasets’:

Orange

+Sử dụng hàm zsum.test trong R

> zsum.test(mean.x=7.8,sigma.x=0.5, n.x=50, alt="t", mu=8,conf.level = 0.99)

One-sample z-Test

data: Summarized x

z = -2.8284, p-value = 0.004678

alternative hypothesis: true mean is not equal to 8

99 percent confidence interval:

7.617861 7.982139

sample estimates:

mean of x

7.8

**+Có thể kết luận theo 3 cách**

1) Chỉ tiêu kiểm định z = -2.8284 thuộc D là miền bác bỏ giả thiết  , nên bác bỏ giả thiết  .

2) p-value = 0.004678< mức ý nghĩa alpha=0.01 nên bác bỏ giả thiết 

3) KTC cho mu99 percent confidence interval:

7.617861 7.982139 không chứa 8 nên bác bỏ giả thiết 

Kết luận: Với mức ý nghĩa 0,01 ta bác bỏ giả thiết  tức là trọng lượng trung bình dây có thể chịu là khác 8kg.

***Ví dụ 2(KĐ 1 phía) :***Một nhà sản xuất khẳng định rằng khối lượng trung bình của dây câu có thể chịu là nhỏ hơn 8 kg, biết khối lượng dây có thể chịu có PP chuẩn với độ lệch chuẩn của tổng thể là 0,5 kg. Để kiểm định giả thuyết  = 8 kg với đối thuyết  < 8 kg, 50 dây ngẫu nhiên được kiểm tra và khối lượng trung bình dây có thể chịu là 7,8 kg. Hãy kiểm định khẳng định của nhà sản xuất với mức ý nghĩa 0,01.

**GIẢI:**

+Từ giả thiết ta có: 

Đây là bài toán kiểm định giả thiết cho một giá trị trung bình, khi  đã biết

Đặt bài toán: .

+Sử dụng hàm zsum.test trong R

> zsum.test(mean.x=7.8,sigma.x=0.5, n.x=50, alt="l", mu=8,conf.level = 0.99)

One-sample z-Test

data: Summarized x

z = -2.8284, p-value = 0.002339

alternative hypothesis: true mean is less than 8

99 percent confidence interval:

NA 7.964498

sample estimates:

mean of x

7.8

+Có thể kết luận theo 3 cách

1) Chỉ tiêu kiểm định z = -2.8284 thuộc D nên bác bỏ giả thiết 

2) p-value = 0.002339 < mức ý nghĩa alpha=0.01 nên bác bỏ giả thiết

3) Các giá trị trong KTC 99 percent confidence interval:

NA 7.964498 nhỏ hơn 8 nên bác bỏ giả thiết 

Kết luận: Với mức ý nghĩa 0,01 ta bác bỏ giả thuyết  tức là có thể xem trọng lượng trung bình dây có thể chịu là nhỏ hơn 8kg.

***Ví dụ 3(*dữ liệu sơ cấp)*:***Một nhà sản xuất khẳng định rằng khối lượng trung bình của dây câu có thể chịu là 8 kg, biết khối lượng dây có thể chịu có PP chuẩn với độ lệch chuẩn của tổng thể là 0,5 kg. Để kiểm định giả thiết  = 8 kg với đối thiết  ≠ 8 kg, 11 dây ngẫu nhiên được kiểm tra và cho số liệu như sau:7.8, 6.6, 6.5, 7.4, 7.3, 7.0, 6.4, 7.1, 6.7, 7.6, 6.8 .

Hãy kiểm định khẳng định của nhà sản xuất với mức ý nghĩa 0,01.

**Giải:**

+Gọi thư viện BSDA:

> library(BSDA)

Loading required package: lattice

Attaching package: ‘BSDA’

The following object is masked from ‘package:datasets’:

Orange

+Nhập dữ liệu:

> x=c(7.8, 6.6, 6.5, 7.4, 7.3, 7.0, 6.4, 7.1, 6.7, 7.6, 6.8)

+Sử dụng hàm zsum.test trong R

> zsum.test(mean(x),sigma.x=0.5, n.x=11, alt="t", mu=8,conf.level = 0.99)

One-sample z-Test

data: Summarized x

z = -6.5126, p-value = 7.384e-11

alternative hypothesis: true mean is not equal to 8

99 percent confidence interval:

6.629861 7.406503

sample estimates:

mean of x

7.018182

Cho ta cho kết luận tương tự.

Kết luận: Với mức ý nghĩa 0,01 ta bác bỏ giả thiết tức là có thể xem trọng lượng trung bình dây có thể chịu là khác 8kg.

**Nhận xét : VD này cho dữ liệu sơ cấp, ta có thể sử dụng hàm z.test để giải quyết bài toán.**

Khi đó không cần nhập mean(x), n.x.

+Sử dụng hàm z.test trong R

> z.test(x,sigma.x=0.5, alt="t", mu=8,conf.level = 0.99)

One-sample z-Test

data: x

z = -6.5126, p-value = 7.384e-11

alternative hypothesis: true mean is not equal to 8

99 percent confidence interval:

6.629861 7.406503

sample estimates:

mean of x

7.018182

Kết luận: Với mức ý nghĩa 0,01 ta bác bỏ giả thiết  tức là có thể xem trọng lượng trung bình dây có thể chịu là khác 8kg.

***Ví dụ 4(KĐ hiệu 2 TB):***Một mẫu ngẫu nhiên  lấy từ tổng thể có phân phối chuẩn với độ lệch chuẩn là , có giá trị trung bình . Một mẫu ngẫu nhiên thứ hai  lấy từ tổng thể có phân phối chuẩn với độ lệch chuẩn là , có giá trị trung bình . Kiểm định giả thiết rằng không có sự sai khác về chất lượng giữa hai tổng thể, với mức ý nghĩa 0,05.

**GIẢI:**

- Gọi  là giá trị trung bình của hai tổng thể 1, 2 tương ứng

- Từ giả thiết:

 ;  ; 

- Đây là bài toán kiểm định hiệu hai giá trị trung bình khi đã biết.

Đặt bài toán: 

+Sử dụng hàm zsum.test trong R

> zsum.test(mean.x=80, sigma.x=5, n.x=25, mean.y=75, sigma.y=3, n.y=6,mu=0,conf.level=0.95)

Two-sample z-Test

data: Summarized x and y

z = 3.1623, p-value = 0.001565

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

1.901025 8.098975

sample estimates:

mean of x mean of y

80 75

Vậy có thể xem có sự sai khác về chất lượng giữa hai tổng thể, với mức ý nghĩa 0,05.

**3. Sử dụng hàm tsum.test kiểm định, tìm KTC cho các trường hợp sau**

**+Kiểm định, tìm KTC về một trung bình, khi  chưa biết, cỡ mẫu nhỏ (n < 30)**

**+Kiểm định, tìm KTC về hiệu hai giá trị trung bình, khi  chưa biết.**

**+Kiểm định, tìm KTC về hiệu hai giá trị trung bình, khi  chưa biết (cỡ mẫu nhỏ).**

### Usage

tsum.test(mean.x, s.x = NULL, n.x = NULL, mean.y = NULL, s.y = NULL,

n.y = NULL, alternative = "two.sided", mu = 0, var.equal = FALSE,

conf.level = 0.95)

Trong đó:

alt="t" : kiểm định 2 phía và cho ước lượng KTC

alt="g" : kiểm định lớn hơn

alt="l" : kiểm định nhỏ hơn

**Chú ý: Nếu dữ liệu cho ở dạng số liệu quan sát (dữ liệu sơ cấp), ta có thể sử dụng hàm t.test.**

***Ví dụ 1:***Một báo cáo khẳng định mỗi máy hút bụi tiêu thụ khoảng 46 kWh/1 năm. Từ một mẫu gồm 12 gia đình được nghiên cứu, cho thấy máy hút bụi tiêu thụ trung bình 42 kWh mỗi năm với độ lệch chuẩn mẫu 11,9 kWh. Liệu có thể nói, với mức ý nghĩa 0,05, trung bình máy hút bụi tiêu thụ không bằng 46 kWh mỗi năm hay không? Giả sử tổng thể đang xét có phân phối chuẩn.

Giải:

- Từ giả thiết:



- Đây là bài toán kiểm định một giá trị trung bình, khi  chưa biết, .

Đặt bài toán  .

> tsum.test(mean.x=42, s.x=11.9, n.x=12, mu=46, alternative="t",conf.level = 0.95)

One-sample t-Test

data: Summarized x

t = -1.1644, df = 11, p-value = 0.2689

alternative hypothesis: true mean is not equal to 46

95 percent confidence interval:

34.4391 49.5609

sample estimates:

mean of x

42

+Kết luận: Do p-value = 0.2689 > alpha=0.05 nên chấp nhận giả thiết .

Vậy có thể khẳng định mỗi máy hút bụi tiêu thụ TB khoảng 46 kWh/1 năm.

***Ví dụ 2(dữ liệu sơ cấp):*** Kiểm định giả thuyết rằng thể tích của các hộp đựng loại dầu nhờn nào đó là 10 lít, nếu từ mẫu ngẫu nhiên gồm 10 hộp ta có các thể tích là: 10,2 9,7 10,1 10,3 10,1 9,8 9,9 10,4 10,3 9,8. Sử dụng mức ý nghĩa 0,01 và giả sử phân phối của thể tích là chuẩn.

Giải:

- Đây là bài toán kiểm định một giá trị trung bình, khi  chưa biết, .

Đặt bài toán 

+Nhập dữ liệu:

> x=c(10.2, 9.7, 10.1, 10.3, 10.1, 9.8, 9.9, 10.4, 10.3, 9.8)

+Sử dụng hàm tsum.test trong R

> tsum.test(mean(x), sd(x), n.x=10, mu=10, alternative="t",conf.level = 0.99)

One-sample t-Test

data: Summarized x

t = 0.77174, df = 9, p-value = 0.46

alternative hypothesis: true mean is not equal to 10

99 percent confidence interval:

9.807338 10.312662

sample estimates:

mean of x

10.06

Vậy có thể xem thể tích của các hộp đựng loại dầu nhờn nào đó là 10 lít.

**Nhận xét : VD này cho dữ liệu sơ cấp, ta có thể sử dụng hàm t.test để giải quyết bài toán:**

> x=c(10.2, 9.7, 10.1, 10.3, 10.1, 9.8, 9.9, 10.4, 10.3, 9.8)

> t.test(x, mu=10, alternative="t",conf.level = 0.99)

One Sample t-test

data: x

t = 0.77174, df = 9, p-value = 0.46

alternative hypothesis: true mean is not equal to 10

99 percent confidence interval:

9.807338 10.312662

sample estimates:

mean of x

10.06

Vậy có thể xem thể tích của các hộp đựng loại dầu nhờn nào đó là 10 lít.

***Ví dụ 3 (dữ liệu có tần số):*** Điều tra ngẫu nhiên lương tháng (triệu đồng) của giám đốc công ty địa ốc, ta có bảng số liệu sau:

Lương tháng 90 95 105 110 115 130

Số lượng 2 5 2 3 1 3

Tìm khoảng tin cậy 99% cho mức lương tháng trung bình giám đốc công ty địa ốc. Giả sử PP là xấp xỉ chuẩn.

**Nhận xét: VD này dùng lệnh rep( ) ghép Luongthang và Tanso.**

Giải:

+Nhập dữ liệu:

> Luongthang=c(90, 95, 105, 110, 115, 130)

> Tanso=c(2, 5, 2, 3, 1, 3)

> Dulieu=rep(Luongthang,Tanso)

> Dulieu

[1] 90 90 95 95 95 95 95 105 105 110 110 110 115 130 130 130

> t.test(Dulieu, alternative="t",conf.level = 0.99)

One Sample t-test

data: Dulieu

t = 30.178, df = 15, p-value = 7.619e-15

alternative hypothesis: true mean is not equal to 0

99 percent confidence interval:

95.8753 116.6247

sample estimates:

mean of x

106.25

KL: khoảng tin cậy 99% cho mức lương tháng trung bình giám đốc công ty địa ốc

99 percent confidence interval:

95.8753 116.6247

***Ví dụ 4(Phương sai bằng nhau chưa biết*)*:*** Để so sánh mức độ mài mòn của hai loại kim loại khác nhau, người ta lấy 12 miếng loại 1 và 10 miếng loại 2. Mẫu ứng với kim loại 1 có trung bình mài mòn là 85 đơn vị, với độ lệch chuẩn mẫu bằng 4; trong khi mẫu ứng với kim loại 2 có trung bình là 81 và độ lệch chuẩn mẫu là 5. Có thể kết luận, với mức ý nghĩa 0,05, rằng hiệu số trung bình mức độ mài mòn của kim loại 1 và kim loại 2 là hơn 2 đơn vị được không? Giả sử các mật độ đều xấp xỉ chuẩn với **phương sai bằng nhau.**

**Nhận xét: Do phương sai bằng nhau ta chỉ thay đổi một thông số trong hàm t với var.equal=TRUE**

Trong VD này bậc tự do (degree freedom) df = 20 được tính theo công thức



**Giải:**

- Đặt  là kỳ vọng cho độ mài mòn của hai kim loại 1 và 2

- Từ giả thiết:



- Đây là bài toán kiểm định hiệu hai giá trị trung bình, với phương sai bằng nhau chưa biết

Đặt bài toán: .

> tsum.test(mean.x=85, s.x=4, n.x=12, mean.y=81, s.y=5, n.y=10, mu=2, alternative="greater",**var.equal = TRUE**,conf.level=0.95)

Standard Two-Sample t-Test

data: Summarized x and y

t = 1.0432, df = 20, p-value = 0.1547

alternative hypothesis: true difference in means is greater than 2

95 percent confidence interval:

0.6932903 NA

sample estimates:

mean of x mean of y

85 81

Kết luận: Do p-value = 0.1547 > alpha = 0.05 nên chấp nhận . Ta không thể kết luận rằng mức độ mài mòn của kim loại 1 hơn kim loại 2 là 2 đơn vị.

***Ví dụ 5(Phương sai khác nhau chưa biết*)*:*** Để so sánh mức độ mài mòn của hai loại kim loại khác nhau, người ta lấy 12 miếng loại 1 và 10 miếng loại 2. Mẫu ứng với kim loại 1 có trung bình mài mòn là 85 đơn vị, với độ lệch chuẩn mẫu bằng 4; trong khi mẫu ứng với kim loại 2 có trung bình là 81 và độ lệch chuẩn mẫu là 5. Có thể kết luận, với mức ý nghĩa 0,05, rằng hiệu số trung bình mức độ mài mòn của kim loại 1 và kim loại 2 là hơn 2 đơn vị được không? Giả sử các mật độ đều xấp xỉ chuẩn với **phương sai khác nhau.**

**Nhận xét: Do phương sai khác nhau ta để mặc định (var.equal=FALSE)**

Trong VD này bậc tự do (degree freedom) df = 17.165 được tính theo công thức



**Giải:**

- Đặt  là kỳ vọng cho độ mài mòn của hai kim loại 1 và 2

- Từ giả thiết:



- Đây là bài toán kiểm định hiệu hai giá trị trung bình, với phương sai chưa biết

Đặt bài toán: .

> tsum.test(mean.x=85, s.x=4, n.x=12, mean.y=81, s.y=5, n.y=10, mu=2, alternative="greater", conf.level=0.95)

Welch Modified Two-Sample t-Test

data: Summarized x and y

t = 1.0215, df = 17.165, p-value = 0.1606

alternative hypothesis: true difference in means is greater than 2

95 percent confidence interval:

0.5959261 NA

sample estimates:

mean of x mean of y

85 81

Kết luận: Do p-value = 0.1606 > alpha = 0.05 nên chấp nhận . Ta không thể kết luận rằng mức độ mài mòn của kim loại 1 hơn kim loại 2 là 2 đơn vị.

***Ví dụ 6(KĐ hiệu 2 TB, dữ liệu sơ cấp):*** Một nghiên cứu được thực hiện bởi Trung tâm Thủy lợi và được phân tích bởi Trung tâm Thống kê, thuộc Đại học Virginia, nhằm so sánh hai thiết bị xử lý nước thải. Thiết bị A được đặt ở vùng dân cư có thu nhập trung bình thấp. Thiết bị B được đặt ở vùng dân cư có thu nhập trung bình cao. Lượng nước thải được xử lý bởi mỗi thiết bị (tính theo nghìn ga-lông/ ngày) được đo trong 10 ngày như sau:

Thiết bị A: 21 19 20 23 22 28 32 19 13 18

Thiết bị B: 20 39 24 33 30 28 30 22 33 24

Với mức ý nghĩa 5%, có thể kết luận rằng có sự khác nhau giữa lượng nước thải trung bình được xử lý ở vùng có thu nhập thấp và vùng có thu nhập cao không. Giả sử các mật độ đều xấp xỉ chuẩn với **phương sai bằng nhau.**

Giải:

+Nhập dữ liệu:

> x=c(21, 19, 20, 23, 22, 28, 32, 19, 13, 18)

> y=c(20, 39, 24, 33, 30, 28, 30, 22, 33, 24)

> t.test(x, y, mu=0, alternative="t",var.equal = TRUE, conf.level=0.95)

Two Sample t-test

data: x and y

t = -2.7149, df = 18, p-value = 0.01419

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-12.0621 -1.5379

sample estimates:

mean of x mean of y

21.5 28.3

KL: Do p-value = 0.01419 < 0.05 nên bác bỏ . Với mức ý nghĩa 5%, có thể kết luận rằng có sự khác nhau giữa lượng nước thải trung bình được xử lý ở vùng có thu nhập thấp và vùng có thu nhập cao.

**Nhận xét: Trong VD này 2 mẫu là độc lập.**

**Nếu 2 mẫu là phụ thuộc (quan sát cặp đôi), ví dụ một nhóm đối tượng được theo dõi theo thời gian thì chúng ta cần sử dụng một kiểm định t có tên là paired t-test: t.test(x, y, paired=TRUE).**

**4. Sử dụng hàm prop.test** **kiểm định, tìm KTC cho tỷ lệ**

**+** **Kiểm định, tìm KTC cho 1 tỷ lệ, n lớn**

**+ Kiểm định, tìm KTC cho hiệu 2 tỷ lệ, n lớn**

**Usage**

+prop.test(x, n, p = NULL,

alternative = c("two.sided", "less", "greater"),

conf.level = 0.95, correct = TRUE)

+prop.test(c(x1,x2), c(n1,n2), p = NULL,

alternative = c("two.sided", "less", "greater"),

conf.level = 0.95, correct = TRUE)

Trong đó:

alt="t" : kiểm định 2 phía và cho ước lượng KTC

alt="g" : kiểm định lớn hơn

alt="l" : kiểm định nhỏ hơn

correct: tham số dạng logic chỉ xem có hay không sự điều chỉnh liên tục Yate, mặc định là correct = TRUE.

Nếu và , chọn correct = FALSE.

***Ví dụ 1(KĐ 1 phía):*** Loại thuốc an thần cũ chỉ có tác động tới 60% người sử dụng. Kết quả thử nghiệm loại thuốc mới với 100 người thì thấy thuốc có tác dụng với 70 người. Có thể tin được hay không rằng loại thuốc mới tốt hơn loại thường dùng? Sử dụng mức ý nghĩa 0,05.

**GIẢI**

Gọi *p* là tỷ lệ người nhận được sự tác động của loại thuốc mới.

Từ giả thiết ta có: .

Đây là bài toán kiểm định một phía cho một tỷ lệ với cỡ mẫu lớn.

Đặt bài toán: 

Do và , chọn correct = FALSE

> prop.test(70, 100, p = 0.6, alternative = "greater",conf.level = 0.95, correct = F)

1-sample proportions test without continuity correction

data: 70 out of 100, null probability 0.6

X-squared = 4.1667, df = 1, p-value = 0.02061

alternative hypothesis: true p is greater than 0.6

95 percent confidence interval:

0.6201679 1.0000000

sample estimates:

p

0.7

Kết luận: Do p-value = 0.02061 < 0.05, nên bác bỏ .Với mức ý nghĩa 0,05 ta có thể nói rằng loại thuốc mới tốt hơn loại thuốc thường dùng.

***Ví dụ 2(ƯL, KĐ 2 phía):*** Điều tra ngẫu nhiên 500 gia đình có tivi ở thành phố Hamilton, Canada, thấy rằng có 340 gia đình thuê bao chương trình HBO. Hãy tìm khoảng tin cậy 95% cho tỷ lệ gia đình thuê bao chương trình HBO trong số những gia đình có tivi trong thành phố này.

**GIẢI**

- Đặt *p* là tỷ lệ cần ước lượng.

- Từ giả thiết: ta có: .

- Đây là bài toán ước lượng cho 1 tỷ lệ, cỡ mẫu lớn

Do và , chọn correct = FALSE

**-Chọn alt="t" cho ước lượng KTC.**

> prop.test(340, 500, alternative = "t",conf.level = 0.95, correct = F)

1-sample proportions test without continuity correction

data: 340 out of 500, null probability 0.5

X-squared = 64.8, df = 1, p-value = 8.29e-16

alternative hypothesis: true p is not equal to 0.5

95 percent confidence interval:

0.6378730 0.7193822

sample estimates:

p

0.68

KL: 95 percent confidence interval:

0.6378730 0.7193822

***Ví dụ 3(KĐ 2 tỷ lệ):*** Một cuộc bỏ phiếu được đưa ta để xác định vị trí xây dựng một nhà máy hóa chất ở trong thành phố hay ở ngoại ô. Có 120 trên 200 cử tri trong thành phố đồng ý cho xây dựng trong thành phố và 240 trên 500 cử tri ở ngoại ô đồng ý với đề xuất này. Liệu có thể cho rằng tỷ lệ cử tri trong thành phố và ngoại ô đồng ý với đề xuất này là như nhau không? Sử dụng mức ý nghĩa 0,025?

**GIẢI:**

* Gọi *p* là tỷ lệ cử tri trong thành phố và ngoại ô đồng ý.
* Từ giả thiết: 
* Đây là bài toán kiểm định về sự bằng nhau giữa hai tỷ lệ với cỡ mẫu lớn

Do và  và và , chọn correct = FALSE

Đặt bài toán: 

> prop.test(c(120,240), c(200,500), alternative = "t",conf.level = 0.975, correct = F)

2-sample test for equality of proportions without continuity correction

data: c(120, 240) out of c(200, 500)

X-squared = 8.2353, df = 1, p-value = 0.004108

alternative hypothesis: two.sided

97.5 percent confidence interval:

0.02760635 0.21239365

sample estimates:

prop 1 prop 2

0.60 0.48

Kết luận: Do p-value = 0.004108 < 0.025 nên bác bỏ . Với mức ý nghĩa 0,025, ta bác bỏ giả thiết tức là có thể cho rằng tỷ lệ cử tri trong và ngoài thị trấn đồng ý là không bằng nhau.

**IV. Hồi quy tuyến tính**

**1. Ước lượng điểm cho các hệ số hồi quy theo phương pháp bình phương tối thiểu**

+ Cho mẫu {(*xi, yi*) = 1,2, ...,*n*}, các ước lượng bình phương tối thiểu *a* và *b* của hệ số hồi quy  được tính từ công thức sau:



Và



Phương trình đường hồi quy tuyến tính thực nghiệm



**+ Hàm > lm(y ~ x)** (lm là viết tắt của linear model) **tính toán các giá trị của a và b.**

**+ Lệnh >plot(x,y):** Vẽ các điểm.

**+ Lệnh > predict(y ~ x, data.frame(****):** Dự báo y khi 

**Ví dụ:** Cho mẫu

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| x | 0,5 | 1,5 | 3,2 | 4,2 | 5,1 | 6,5 |
| y | 1,3 | 3,4 | 6,7 | 8,0 | 10,0 | 13,2 |

1. Vẽ các điểm trên.
2. Tìm phương trình đường hồi quy tuyến tính thực nghiệm.
3. Dự báo y khi 

Giải:

a)

> library(BSDA)

Loading required package: lattice

Attaching package: ‘BSDA’

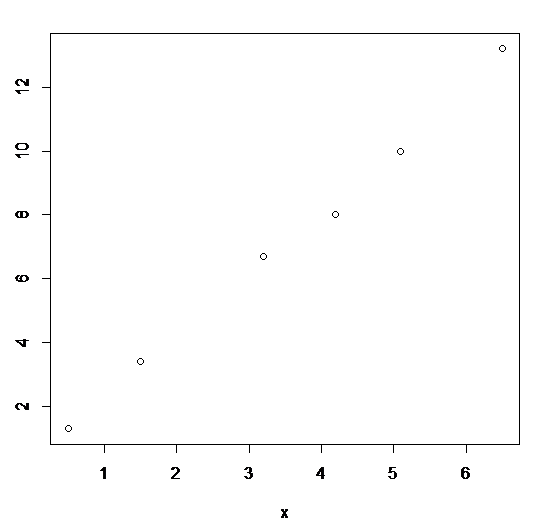
The following object is masked from ‘package:datasets’:

Orange

**> x=c(0.5, 1.5, 3.2, 4.2, 5.1, 6.5)**

**> y=c(1.3, 3.4, 6.7, 8, 10, 13.2)**

**> plot(x,y)**

****

b)

> lm(y ~ x)

Call:

lm(formula = y ~ x)

Coefficients:

(Intercept) x

0.3492 1.9288

Do đó, PT đường hồi quy tuyến tính thực nghiệm

*ŷ* = 0.3492+ 1.9288x

c)

> res=lm(y ~ x)

> new=data.frame(x = 8.5)

> pred=predict(res, new)

> pred

1

16.74399.

**2. Tương quan tuyến tính**

+ Khái niệm hệ số tương quan dùng để đo mức độ phụ thuộc tuyến tính giữa hai biến ngẫu nhiên



* 
* Khi , thì không có tương quan tuyến tính giữa *X* và *Y* ;
* Khi || càng gần 1, thì sự phụ thuộc tuyến tính giữa *X* và *Y* càng mạnh;
* Khi || = 1, sự phụ thuộc tuyến tính là mạnh nhất. Khi đó, *Y* là một hàm tuyến tính của *X* tức là tồn tại hai số *a* và *b* sao cho

*Y = aX + b*.

Muốn biết được , ta phải có phân phối xác suất của (*X*, *Y*). Khi ta không biết phân phối xác suất của (*X, Y*), thì đặt ra bài toán là ước lượng  và kiểm định giả thuyết về giá trị của  dựa vào một mẫu cụ thể (*x*1, *y*1), (*x*2, *y*2), (*x*3, *y*3),..., (*x*n, *y*n).

+ Ước lượng cho  bởi giá trị từ một mẫu:



*r* được gọi là hệ số tương quan tuyến tính mẫu.

**+ Hàm > cor(x,y) tính toán hệ số tương quan tuyến tính mẫu r.**

**Ví dụ:** Cho mẫu

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| x | 0,5 | 1,5 | 3,2 | 4,2 | 5,1 | 6,5 |
| y | 1,3 | 3,4 | 6,7 | 8,0 | 10,0 | 13,2 |

Tính hệ số tương quan tuyến tính mẫu r.

Giải:

> x=c(0.5, 1.5, 3.2, 4.2, 5.1, 6.5)

> y=c(1.3, 3.4, 6.7, 8, 10, 13.2)

> cor(x,y)

[1] 0.997908

Cho r=0.997908.

**3. Ước lượng không chệch của phương sai**

+ Phương sai của phần dư:  cho ta biết mức độ phân tán của biến ngẫu nhiên Y xung quanh đường hồi quy. Ước lượng không chệch của  là



**+ Lệnh <summary(lm(y ~ x)) liệt kê các thông tin tính toán trong lm(y ~ x), trong đó có giá trị** 

**Ví dụ:** Cho mẫu

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| x | 0,5 | 1,5 | 3,2 | 4,2 | 5,1 | 6,5 |
| y | 1,3 | 3,4 | 6,7 | 8,0 | 10,0 | 13,2 |

Tính được phương sai mẫu của phần dư: 

ở phần ba của kết quả:

> x=c(0.5, 1.5, 3.2, 4.2, 5.1, 6.5)

> y=c(1.3, 3.4, 6.7, 8, 10, 13.2)

> summary(lm(y ~ x))

Call:

lm(formula = y ~ x)

Residuals:

1 2 3 4 5 6

-0.0136 0.1576 0.1786 -0.4502 -0.1861 0.3136

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 0.34920 0.25333 1.378 0.24

x 1.92880 0.06248 30.871 6.56e-06 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

**Residual standard error: 0.3133 on 4 degrees of freedom**

Multiple R-squared: 0.9958, Adjusted R-squared: 0.9948

F-statistic: 953 on 1 and 4 DF, p-value: 6.56e-06

**Nhận xét:** Trong lệnh summary <summary(lm(y ~ x)), phần một mô tả phần dư (Residuals) của mô hình hồi quy, phần hai là ước lượng hệ số a; b cùng với sai số và giá trị của kiểm định t.