**BÀI TẬP KIỂM ĐỊNH PHI THAM SỐ, CHI BÌNH PHƯƠNG+ HỒI QUY TUYẾN TÍNH**

**I.Kiểm định wilcoxon**

**Bài toán 1: Kiểm định so sánh trung vị với 1 số khi cỡ mẫu nhỏ, tổng thể không có phân phối chuẩn, sử dụng hàm wilcox.test**

**Usage**

wilcox.test(x, ...)

## Default S3 method:

wilcox.test(x, y = NULL,

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

mu = 0, paired = FALSE, exact = NULL, correct = TRUE,

conf.int = FALSE, conf.level = 0.95, ...)

*trong đó:*

alt="t" (two-side): kiểm định 2 phía

alt="g" (greater): kiểm định lớn hơn

alt="l" (less): kiểm định nhỏ hơn

**Ví dụ:** Kiểm định giả thiết 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 không là chuẩn.

#Gọi M là thể tích của các hộp đựng loại dầu nhờn

#Bài toán kiểm định giả thiết về 1 tổng thể không có phân phối chuẩn, cỡ mẫu nhỏ

#H0: M=10; H1:M khác 10

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

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

Wilcoxon signed rank test with continuity correction

data: x

V = 35, p-value = 0.4721

alternative hypothesis: true location is not equal to 10

Warning message:

In wilcox.test.default(x, alt = "t", mu = 10, conf.level = 0.99) :

cannot compute exact p-value with ties

Do p-value = 0.4721> 0.01 nên chấp nhận gt H0

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

**Chú ý:** Nếu muốn tính chính xác p-value, ta dùng thêm hàm jitter, R sẽ không xuất hiện cảnh báo tính toán không chính xác p-value:

“Warning message:

In wilcox.test.default(x, alt = "t", mu = 10, conf.level = 0.99) :

cannot compute exact p-value with ties”.

Tuy nhiên mỗi lần tính sẽ cho 1 kết quả xấp xỉ nhau nên nếu ta muốn chỉ tính ra 1 giá trị p-value thì bỏ hàm jitter.

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

> wilcox.test(jitter(x),alt="t",mu=10,conf.level = 0.99)

Wilcoxon signed rank test

data: jitter(x)

V = 37, p-value = 0.375

alternative hypothesis: true location is not equal to 10

Do p-value = 0.375>0.01 nên chấp nhận gt H0

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

**Bài toán 2: Kiểm định giả thiết cho mẫu theo đôi(quan sát cặp đôi, 2 mẫu không độc lập) khi cỡ mẫu nhỏ, tổng thể không có phân phối chuẩn, sử dụng hàm wilcox.test**

**Usage**

wilcox.test(x, ...)

## Default S3 method:

wilcox.test(x, y = NULL,

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

mu = 0, paired = TRUE, exact = NULL, correct = TRUE,

conf.int = FALSE, conf.level = 0.95, ...)

*trong đó:*

alt="t" (two-side): kiểm định 2 phía

alt="g" (greater): kiểm định lớn hơn

alt="l" (less): kiểm định nhỏ hơn

**Ví dụ:** Một nhóm các sinh viên muốn du học ở Anh đã đăng ký thi IELTS chuẩn bị cho khóa học. Lấy một mẫu kiểm tra vào ngày đầu tiên đi học và sau kiểm tra lại vào cuối khóa học. Kết quả thu được như sau:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Trước | 5.5 | 5 | 4.5 | 6.5 | 6 | 5 |
| Sau | 6.5 | 6 | 4 | 7 | 6.5 | 6.5 |

Sử dụng mức ý nghĩa 0,05 và giả sử phân phối không là chuẩn, kiểm định xem liệu khoá học có giúp sinh viên học IELTS tốt hơn không?

#Bài toán kiểm định giả thiết cho mẫu theo đôi khi tổng thể không có phân phối chuẩn, cỡ mẫu nhỏ

#H0: M1 - M2 = 0; H1:M1 - M2 <0

> T=scan()

1: 5.5 5 4.5 6.5 6 5

7:

Read 6 items

> S=scan()

1: 6.5 6 4 7 6.5 6.5

7:

Read 6 items

> wilcox.test(x,y, alt="less",mu=0,paired = TRUE,conf.level = 0.95)

Wilcoxon signed rank test

V = 2, p-value = 0.04688

alternative hypothesis: true location shift is less than 0

Do p-value = 0.04688<0.05 nên bác bỏ gt H0

Vậy có thể xem khoá học giúp sinh viên học IELTS tốt hơn.

**Bài toán 3: Kiểm định giả thiết cho 2 mẫu độc lập khi cỡ mẫu nhỏ, tổng thể không có phân phối chuẩn, sử dụng hàm wilcox.test**

**Usage**

wilcox.test(x, ...)

## Default S3 method:

wilcox.test(x, y = NULL,

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

mu = 0, paired = FALSE, exact = NULL, correct = TRUE,

conf.int = FALSE, conf.level = 0.95, ...)

*trong đó:*

alt="t" (two-side): kiểm định 2 phía

alt="g" (greater): kiểm định lớn hơn

alt="l" (less): kiểm định nhỏ hơn

***Ví dụ:*** Một nghiên cứu được thực hiện bởi Trung tâm Thủy lợi và được phân tích bởi một Trung tâm Thống kê, 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% và giả sử phân phối không là chuẩn, có thể kết luận rằng có sự khác nhau giữa lượng nước thải được xử lý ở vùng có thu nhập thấp và vùng có thu nhập cao không.

#Bài toán kiểm định giả thiết cho 2 mẫu độc lập khi tổng thể không có phân phối chuẩn, cỡ mẫu nhỏ

#H0: M1 - M2 = 0; H1:M1 - M2 khác 0

> x=scan()

1: 21 19 20 23 22 28 32 19 13 18

11:

Read 10 items

> y=scan()

1: 20 39 24 33 30 28 30 22 33 24

11:

Read 10 items

> wilcox.test(x,y, alt="t",mu=0,paired = FALSE,conf.level = 0.95)

Wilcoxon rank sum test

W = 17, p-value = 0.0115

alternative hypothesis: true location shift is not equal to 0

Do p-value = 0.0115<0.05 nên bác bỏ gt H0

Vậy 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.

***Bài tập luyện tập***

**1.** Một máy sản xuất các mảnh kim loại có hình trụ. Một mẫu các mảnh được lấy ra với các đường kính là 1.01, 0.97, 1.03, 1.04, 0.99, 0.98, 0.99, 1.01 , 1.03cm. Sử dụng mức ý nghĩa 0,01 và giả sử phân phối của thể tích không là chuẩn, kiểm định giả thiết đường kính các mảnh kim loại là 1cm.

**2.** Năm mẫu quặng sắt, mỗi mẫu được chia thành hai phần, rồi lần lượt được xác định hàm lượng sắt bằng hai cách là dùng tia X và dùng phân tích hóa học, kết quả thu được là

**Số thứ tự mẫu**

**Cách phân tích 1 2 3 4 5**

Tia X 2,0 2,0 2,3 2,1 2,4

Phân tích hóa học 2,2 1,9 2,5 2,3 2,4

Giả sử các số liệu ở mỗi cách phân tích không theo phân phối chuẩn. Hãy kiểm định rằng hai phương pháp cho kết quả giống nhau, với mức ý nghĩa 0,05

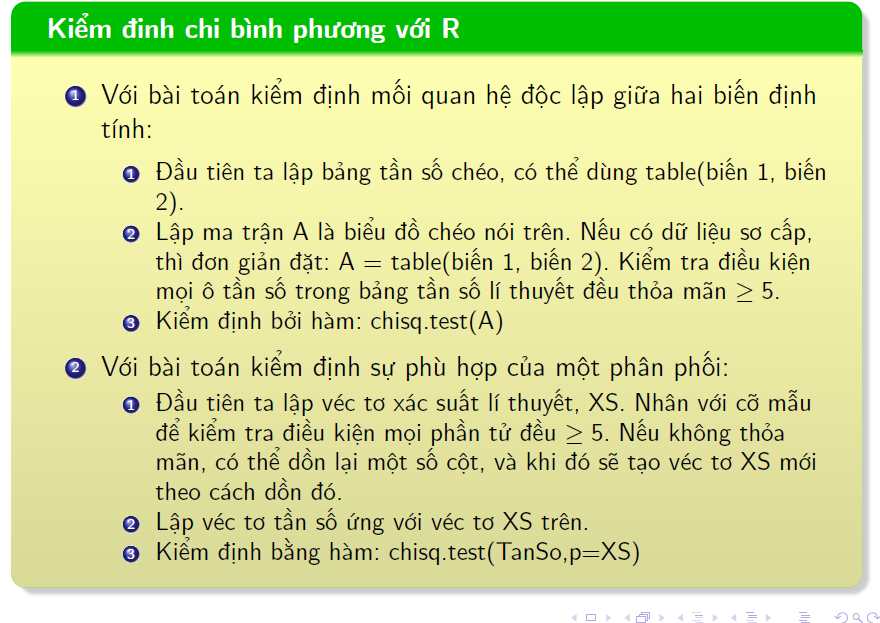
**3.** Một nghiên cứu của Khoa Giáo dục thể chất, nhằm xác định xem sau 8 tuần luyện tập, lượng cholesterol của mỗi người tham gia luyện tập có thực sự giảm không. Một nhóm 15 người tham gia luyện tập 2 lần một tuần, lượng cholesterol trước và sau luyện tập được ghi lại như sau:

**Trước luyện tập**: 129 131 154 172 115 126 175 191 122 238 159 156 176 175 126

**Sau luyện tập**: 151 132 196 195 188 198 187 168 115 165 137 208 133 217 191

Ta có thể kết luận, với mức ý nghĩa 4% rằng, lượng cholesterol thực sự sẽ giảm sau khi thực hiện chương trình luyện tập không? giả sử phân phối không là chuẩn.

**II.Kiểm định Chi bình phương**

****

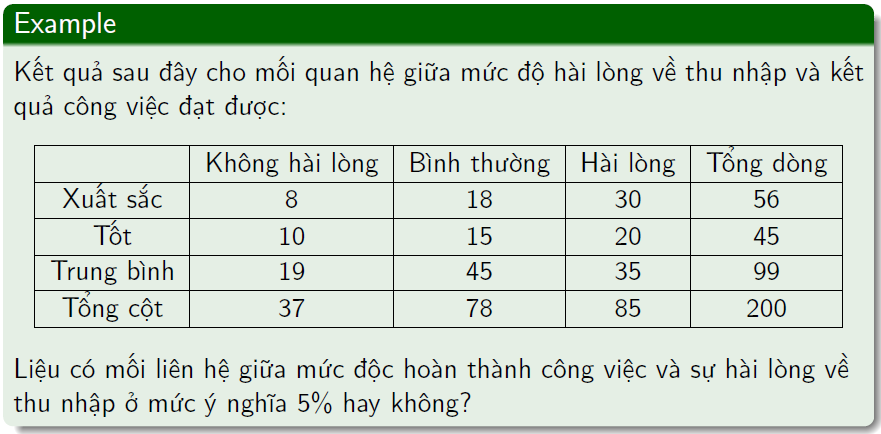
**Bài toán 1: Kiểm định giả thiết về tính độc lập**

Xét bài toán:

H0: Hai biến định tính là độc lập(không có mối liên hệ với nhau)

H1: Hai biến định tính là phụ thuộc

Dùng lệnh chisq.test(A), với A là ma trận dữ liệu.



H0: Hai yếu tố mức độ hài lòng về thu nhập và kết quả công việc là độc lập với nhau

H1: Hai yếu tố trên có mối liên hệ với nhau

> x=c(8,18,30,10,15,20,19,45,35)

> A=matrix(x, nrow = 3, ncol = 3, byrow = TRUE)

> A

[,1] [,2] [,3]

[1,] 8 18 30

[2,] 10 15 20

[3,] 19 45 35

> chisq.test(A)

Pearson's Chi-squared test

data: A

X-squared = 5.8483, df = 4, p-value = 0.2108

Do p-value = 0.2108 > 0.05 nên chấp nhận gt H0

Hai yếu tố mức độ hài lòng về thu nhập và kết quả công việc là độc lập hay không có mối liên hệ với nhau.

**Bài toán 2: So sánh nhiều tỷ lệ**

Xét bài toán:

H0: p1=p2=…=pk

H1: Tồn tại pi khác pj (i khác j)

Dùng lệnh chisq.test(A), với A là ma trận dữ liệu.

**Ví dụ**: Kiểm tra sản phẩm của 3 nhà máy cùng sản xuất một loại sản phẩm thu được số liệu sau

|  |  |  |  |
| --- | --- | --- | --- |
| Nhà máy  Chất lượng | B1 | B2 | B3 |
| Phế phẩm | 11 | 17 | 18 |
| Chính phẩm | 89 | 103 | 112 |

Có thể khẳng định tỷ lệ phế phẩm của 3 nhà máy là như nhau hay không với mức ý nghĩa 0.05?

Gọi p1, p2, p3 lần lượt là tỷ lệ phế phẩm của 3 nhà máy

H0: p1=p2=p3

H1: Tồn tại pi khác pj (i khác j) với i,j thuộc {1,2,3}

> x=scan()

1: 11 17 18

4: 89 103 112

7:

Read 6 items

> A=matrix(x,nrow = 2,ncol = 3,byrow = TRUE)

> A

[,1] [,2] [,3]

[1,] 11 17 18

[2,] 89 103 112

> chisq.test(A)

Pearson's Chi-squared test

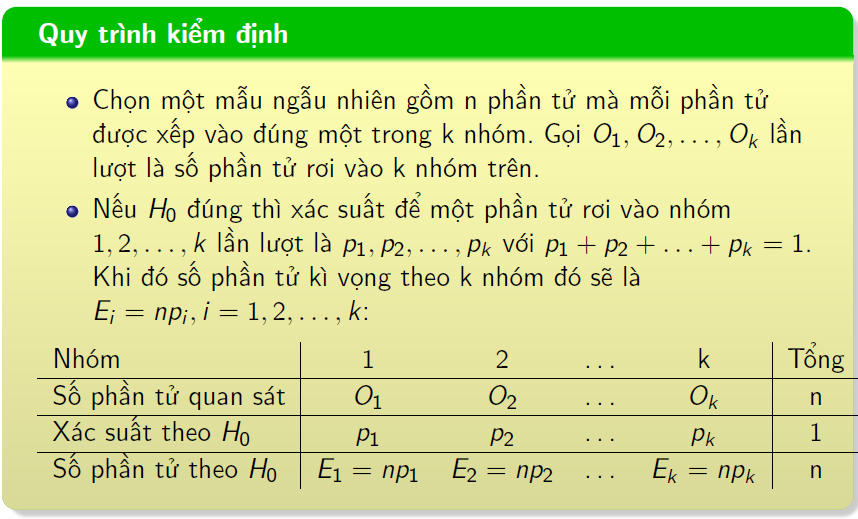
data: A

X-squared = 0.56876, df = 2, p-value = 0.7525

Do p-value = 0.7525> 0.05 nên chấp nhận gt H0

Có thể khẳng định tỷ lệ phế phẩm của 3 nhà máy là như nhau.

**Bài toán 3: Kiểm định sự phù hợp của một phân phối**

****

Xét bài toán:

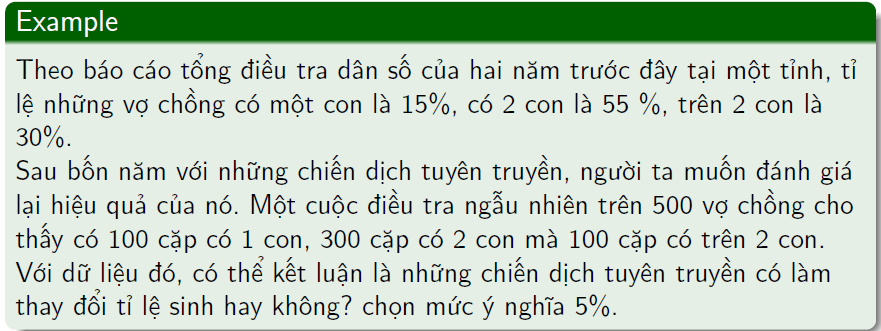
H0: Tổng thể tuân theo phân phối A

H1: Tổng thể không tuân theo phân phối A

x=c(O1, O2,…, Ok)

p0=(p1,p2,…,pk)

Dùng hàm chisq.test(x,p=p0)

****

H0: Tỷ lệ số cặp vợ chồng có 1 con là 0.15, có 2 con là 0.55, có trên 2 con là 0.3

H1: Tỷ lệ nói trên nay đã khác

> x=c(100,300,100)

> p0=c(0.15,0.55,0.3)

> chisq.test(x,p=p0)

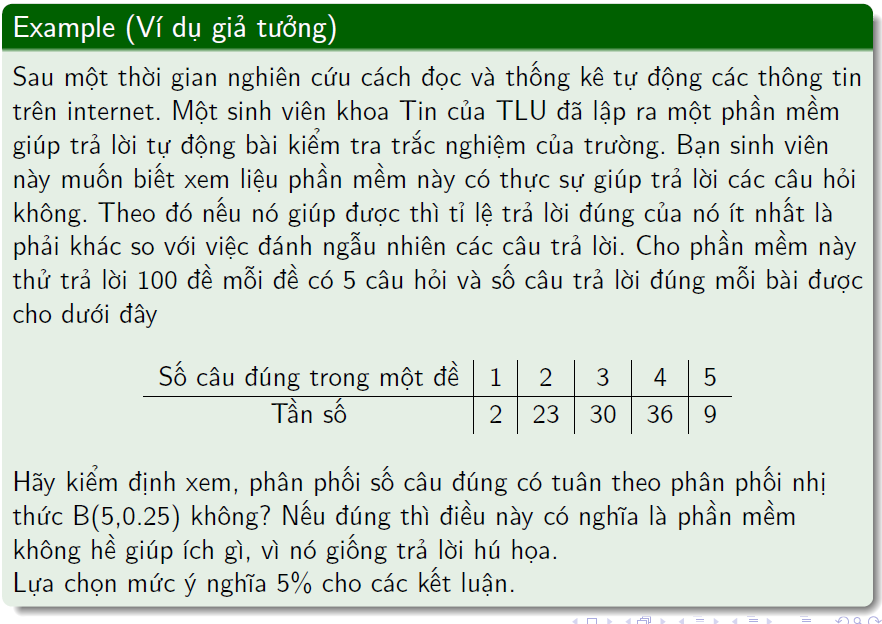
Chi-squared test for given probabilities

data: x

X-squared = 27.273, df = 2, p-value = 1.196e-06

Do p-value = 0.211.196e-0608 < 0.05 nên bác bỏ gt H0

Tỷ lệ nói trên nay đã khác



H0: Số câu trả lời đúng tuân theo phân phối nhị thức B(5,0.25)

H1: Số câu trả lời đúng không tuân theo phân phối nhị thức B(5,0.25)

Nếu số câu trả lời đúng tuân theo phân phối nhị thức B(5,0.25), ta có

> dbinom(0:5,5,0.25)

[1] 0.2373046875 0.3955078125 0.2636718750 0.0878906250 0.0146484375 0.0009765625

> p0=c(dbinom(0:5,5,0.25))

> p0

[1] 0.2373046875 0.3955078125 0.2636718750 0.0878906250 0.0146484375 0.0009765625

> x=scan()

1: 0 2 23 30 36 9

7:

Read 6 items

> x

[1] 0 2 23 30 36 9

> chisq.test(x,p=p0)

Chi-squared test for given probabilities

data: x

X-squared = 1736.7, df = 5, p-value < 2.2e-16

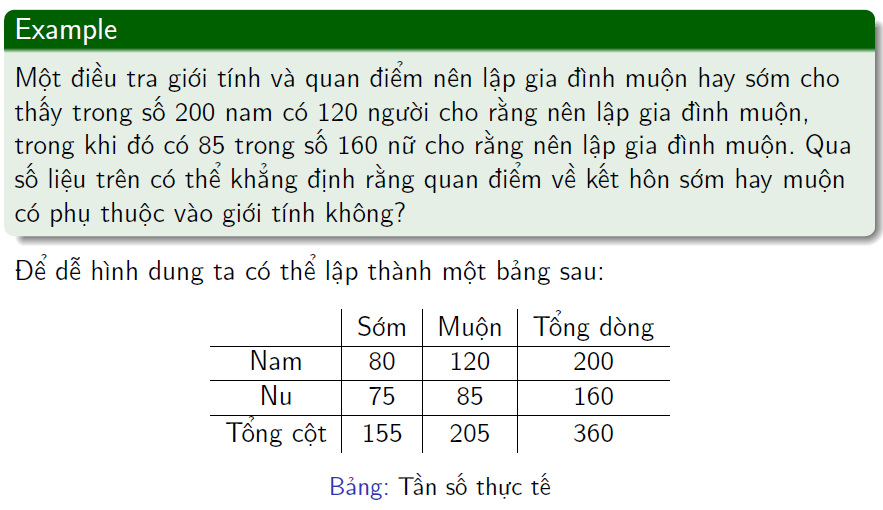
Warning message:

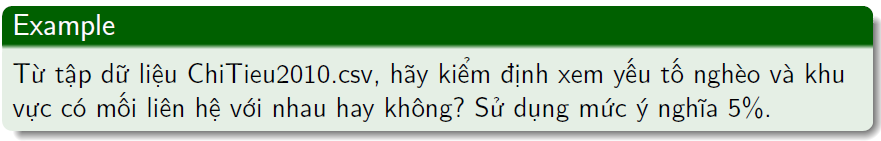
In chisq.test(x, p = p0) : Chi-squared approximation may be incorrect

Do p-value < 2.2e-16 < 0.05 nên bác bỏ gt H0

Số câu trả lời đúng không tuân theo phân phối nhị thức B(5,0.25).

***Bài tập luyện tập***





HD:

H0: Hai yếu tố hộ nghèo và khu vực là độc lập với nhau

H1: Hai yếu tố trên phụ thuộc (có mối liên hệ với nhau)

> DL=read.csv("ChiTieu2010.csv")

> attach(DL)

> A=table(DL$KhuVuc,DL$HoNgheo)

> A

0 1

1 2459 188

2 4830 1921

> chisq.test(A)

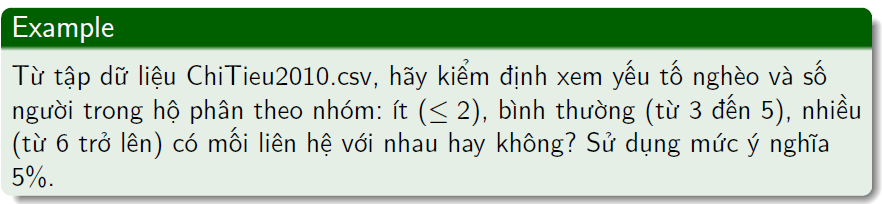
Pearson's Chi-squared test with Yates' continuity correction

data: A

X-squared = 496.87, df = 1, p-value < 2.2e-16

Do p-value < 2.2e-16<0.05 nên bác bỏ gt H0.

Hai yếu tố trên có mối liên hệ với nhau



HD:

H0: Hai yếu tố hộ nghèo và số người trong hộ phân theo 3 nhóm độc lập với nhau

H1: Hai yếu tố trên có mối liên hệ với nhau

**Cách 1:**

> B=table(DL$HoNgheo,DL$SoNguoiTrongHo)

> B

1 2 3 4 5 6 7 8 9 10 11 12 13 15

0 245 800 1531 2522 1235 601 215 93 31 8 4 3 1 0

1 213 378 326 557 305 169 85 39 17 11 5 2 0 2

#Lap ma tran theo 3 nhom

> a11=sum(245, 800 )

> a12=sum(1531, 2522, 1235)

> a13=sum(601, 215, 93, 31, 8, 4, 3, 1, 0)

> a21=sum(213, 378)

> a22=sum(326, 557, 305)

> a23=sum(169, 85, 39, 17, 11, 5, 2, 0, 2)

> x=c(a11,a12,a13,a21,a22,a23)

> A=matrix(x,nrow = 2,ncol = 3,byrow = TRUE)

> A

[,1] [,2] [,3]

[1,] 1045 5288 956

[2,] 591 1188 330

> chisq.test(A)

Pearson's Chi-squared test

data: A

X-squared = 246.1, df = 2, p-value < 2.2e-16

p-value < 2.2e-16<0.05 nên bác bỏ gt H0

Hai yếu tố trên có mối liên hệ với nhau

**Cách 2:**

> B=table(DL$HoNgheo,DL$SoNguoiTrongHo)

> B

1 2 3 4 5 6 7 8 9 10 11 12 13 15

0 245 800 1531 2522 1235 601 215 93 31 8 4 3 1 0

1 213 378 326 557 305 169 85 39 17 11 5 2 0 2

#Lap ma tran theo 3 nhom

> a11=scan()

1: 245 800

3:

Read 2 items

> sum(a11)

[1] 1045

> a12=scan()

1: 1531 2522 1235

4:

Read 3 items

> sum(a12)

[1] 5288

> a13=scan()

1: 601 215 93 31 8 4 3 1 0

10:

Read 9 items

> sum(a13)

[1] 956

> a21=scan()

1: 213 378

3:

Read 2 items

> a22=scan()

1: 326 557 305

4:

Read 3 items

> a23=scan()

1: 169 85 39 17 11 5 2 0 2

10:

Read 9 items

> sum(a21)

[1] 591

> sum(a22)

[1] 1188

> sum(a23)

[1] 330

> x=c(sum(a11),sum(a12),sum(a13),sum(a21),sum(a22),sum(a23))

> A=matrix(x,nrow = 2,ncol = 3,byrow = TRUE)

> A

[,1] [,2] [,3]

[1,] 1045 5288 956

[2,] 591 1188 330

> chisq.test(A)

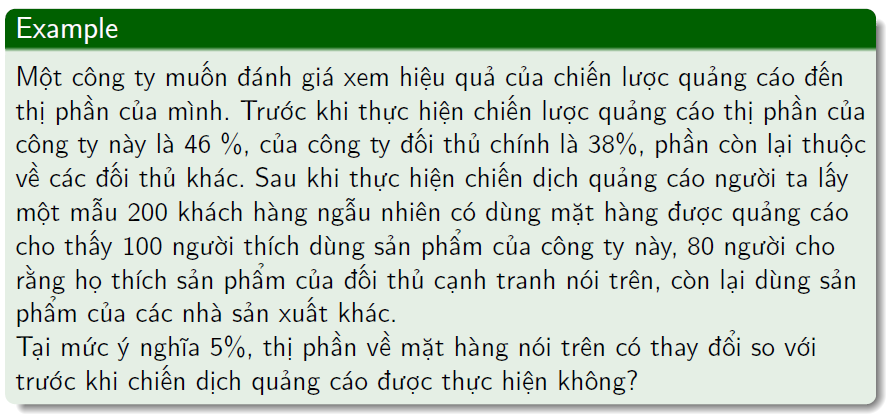
Pearson's Chi-squared test

data: A

X-squared = 246.1, df = 2, p-value < 2.2e-16

p-value < 2.2e-16<0.05 nên bác bỏ gt H0

Hai yếu tố trên có mối liên hệ với nhau



**III.Hồi quy tuyến tính**

**Bài toán 1: Hồi quy tuyến tính đơn**

Cho mẫu {(*xi, yi*) = 1,2, ...,*n*}, mô hình HQTT đơn biến của biến phụ thuộc Y theo biến độc lập X là phương trình có dạng 

Giá trị trung bình của Y khi X nhận giá trị 

PT đường hồi quy tuyến tính mẫu: 

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

**+ Lệnh >plot(x,y):** Vẽ các điểm.

+Covariance



> cov(x,y)

+Hệ số tương quan(Pearson)



> cor(x,y)

+Sai số chuẩn của ước lượng



+Hệ số xác định đường hồi qui mẫu



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

> summary(lm(y ~ x))

**+KTC cho **

> confint(lm(y ~ x),level = 1-alpha)

**+Kiểm định giả thuyết hệ số độ dốc **

*H*0: = 0

*H*1:  ≠ 0

> summary(lm(y ~ x))$coefficients[2,]

**+Kiểm định giả thuyết hệ số **

*H*0: = 0

*H*1:  ≠ 0

> summary(lm(y ~ x))$coefficients[1,]

+Kiểm định giả thuyết hệ số tương quan 

*H*0: = 0

*H*1: 

cor.test(x, y,

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

method = c("pearson", "kendall", "spearman"),

exact = NULL, conf.level = 0.95, continuity = FALSE, ...)

**+Dự báo**

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

**+Khoảng dự đoán**  **cho y khi **

**>predict(lm(y~x),data.frame(****),interval=c("prediction"),level =** **)**

**+Khoảng tin cậy**  **cho giá trị trung bình của y khi** 

**>predict(lm(y~x),data.frame(****),interval=c("confidence"),level =** **)**

(chú ý thay c("prediction") cho **c("confidence")**

**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 |

Tìm phương trình đường hồi quy tuyến tính mẫu của y với x; Tính

+Covariance

+Hệ số tương quan(Pearson)

+ Sai số chuẩn của ước lượng

+Hệ số xác định đường hồi quy mẫu

+Khoảng tin cậy 99% cho 

+Kiểm định giả thuyết 95% hệ số 

*H*0: = 0

*H*1:  ≠ 0

+Kiểm định giả thuyết 95% hệ số tương quan 

*H*0: = 0

*H*1: 

+Dự báo y khi 

+ Khoảng dự đoán 99% cho y khi 

+ Khoảng tin cậy 99% cho 

HD:

PT đường hồi quy tuyến tính mẫu

> 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)

> 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

Covariance

> cov(x,y)

[1] 9.698

Hệ số tương quan

> cor(x,y)

[1] 0.997908

+Sai số chuẩn của ước lượng, Hệ số xác định đường hồi qui mẫu



> 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

Tính được sai số chuẩn của ước lượng: 

Hệ số xác định đường hồi qui mẫu 

Khoảng tin cậy 99% cho 

> confint(lm(y ~ x),level = 0.99)

0.5 % 99.5 %

(Intercept) -0.8171482 1.515557

x 1.6411391 2.216458

Kiểm định giả thuyết 95% hệ số 

*H*0: = 0

*H*1:  ≠ 0

Với mức ý nghĩa 0,05.

Lấy p-value trong lệnh > summary(lm(y ~ x)) hoặc cụ thể hơn

> summary(lm(y ~ x))$coefficients[2,]

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

1.928799e+00 6.247909e-02 3.087111e+01 6.560101e-06

Do đó p-value=nên bác bỏ gt H0.

Kiểm định giả thuyết 95% hệ số tương quan 

*H*0: = 0

*H*1: 

> cor.test(x,y,alt="g",method = "pearson",conf.level = 0.95)

Pearson's product-moment correlation

data: x and y

t = 30.871, df = 4, p-value = 3.28e-06

alternative hypothesis: true correlation is greater than 0

95 percent confidence interval:

0.9861053 1.0000000

sample estimates:

cor

0.997908

**Hoặc là**

> cor.test(x,y,alt="g",conf.level = 0.95)

Pearson's product-moment correlation

data: x and y

t = 30.871, df = 4, p-value = 3.28e-06

alternative hypothesis: true correlation is greater than 0

95 percent confidence interval:

0.9861053 1.0000000

sample estimates:

cor

0.997908

Do đó p-value=nên bác bỏ gt H0.

Dự báo y khi 

> predict(lm(y ~ x),data.frame(x=8.5))

1

16.74399

Tìm khoản dự đoán 99% cho y khi 

> predict(lm(y ~ x),data.frame(x=8.5),interval = c("prediction"),level = 0.99)

fit lwr upr

1 16.74399 14.62369 18.8643

KDĐ là (14.62369, 18.8643)

Tìm khoảng tin cậy 99% cho 

> predict(lm(y ~ x),data.frame(x=8.5),interval = c("confidence"),level = 0.99)

fit lwr upr

1 16.74399 15.18983 18.29815

KTC là (15.18983, 18.29815)

***Bài tập luyện tập: Câu hỏi tương tự như ví dụ trên***

**1.**Điểm của một lớp học gồm 9 sinh viên trong bài báo cáo giữa kỳ (*x*) và bài thi (*y*) như sau:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| x | 77 | 50 | 71 | 72 | 81 | 94 | 96 | 77 | 50 |
| y | 82 | 66 | 78 | 34 | 47 | 85 | 99 | 82 | 66 |



**2.** Một cuộc nghiên cứu về lượng mưa và lượng ô nhiễm không khí thải ra đã cho các số liệu sau:

|  |  |
| --- | --- |
| **Lượng mưa hàng ngày, *x* (0,01 cm)** | **Lượng hạt ô nhiễm thải ra, *y***  **(mcg/cum)** |
| 4,3 | 126 |
| 4,5 | 121 |
| 5,9 | 116 |
| 5,6 | 118 |
| 6,1 | 114 |
| 5,2 | 118 |
| 3,8 | 132 |
| 2,1 | 141 |
| 7,5 | 108 |



**3.**  Số lượng hợp chất hóa học *y* hòa tan trong 100g nước tại các nhiệt độ biến thiên *x*, được ghi lại như sau:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ***x* (oC)** |  | ***y* (gram)** | | |
| 0 |  | 8 | 6 | 8 |
| 15 |  | 12 | 10 | 14 |
| 30 |  | 25 | 31 | 24 |
| 45 |  | 31 | 33 | 28 |
| 60 |  | 44 | 39 | 42 |
| 75 |  | 48 | 51 | 44 |



**4.** Từ tập dữ liệu ChiTieu2010.csv

a) ChiTieuGiaoDucTrongNam(x); SoNguoiTrongHo(y)



b) ChiTieuYTe(x); SoNguoiTrongHo(y)



c) DieuTriNgoaiTru(x); Thuoc(y)



***HD:***

a)

> DL=read.csv("ChiTieu2010.csv")

> attach(DL)

> x=ChiTieuGiaoDucTrongNam

> y=SoNguoiTrongHo

> lm(y ~ x)

Call:

lm(formula = y ~ x)

Coefficients:

(Intercept) x

3.8757574 0.0002555

**Bài toán 2: Hồi quy tuyến tính đa biến**

Cho mẫu , mô hình HQTT đa biến của biến phụ thuộc y theo các biến độc lập  là phương trình có dạng 

Giá trị trung bình của Y khi X nhận giá trị 

PT đường hồi quy tuyến tính mẫu: 

**+ Hàm** (lm là viết tắt của linear model) **tính toán các giá trị của** **.**

**Trong R, các lệnh tương tự như bài toán HQTT đơn biến.**

**Ví dụ:** Cho mẫu

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| x1 | 0,5 | 1,6 | 3,5 | 4,2 | 5,3 | 6,9 |
| x2 | 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 |

Tìm phương trình đường hồi quy tuyến tính mẫu của y với x1, x2; Tính

+ Sai số chuẩn của ước lượng

+Hệ số bội xác định đường hồi quy mẫu

+Khoảng tin cậy 98% cho 

+Kiểm định giả thuyết 95% hệ số 

*H*0: = 0

*H*1:  ≠ 0

+Kiểm định giả thuyết 95% hệ số 

*H*0: = 0

*H*1:  ≠ 0

+Kiểm định giả thuyết 95% hệ số

*H*0: = 0

*H*1:  ≠ 0 (i=1, 2)

+Dự báo y khi 

+ Khoảng dự đoán 96% cho y khi 

+ Khoảng tin cậy 96% cho giá trị trung bình của y khi 

HD:

+Phương trình đường hồi quy tuyến tính mẫu

> x1=c(0.5, 1.6, 3.5, 4.2, 5.3, 6.9)

> x2=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)

> lm(y ~ x1+x2)

Call:

lm(formula = y ~ x1 + x2)

Coefficients:

(Intercept) x1 x2

0.3591 2.1540 -0.3306

Ta có



+ Sai số chuẩn của ước lượng; Hệ số bội xác định đường hồi quy mẫu

> summary(lm(y ~ x1+x2))

Call:

lm(formula = y ~ x1 + x2)

Residuals:

1 2 3 4 5 6

0.02924 0.09042 -0.14018 -0.01738 -0.08926 0.12716

Coefficients:

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

(Intercept) 0.3591 0.1076 3.336 0.0445 \*

x1 2.1540 0.4920 4.378 0.0221 \*

x2 -0.3306 0.5167 -0.640 0.5678

---

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

Residual standard error: 0.1331 on 3 degrees of freedom

Multiple R-squared: 0.9994, Adjusted R-squared: 0.9991

F-statistic: 2651 on 2 and 3 DF, p-value: 1.345e-05

Ta có



+Khoảng tin cậy 98% cho 

> confint(lm(y ~ x1+x2),level = 0.98)

1 % 99 %

(Intercept) -0.12966025 0.847776

x1 -0.07991614 4.388002

x2 -2.67698695 2.015721

+Kiểm định giả thuyết 95% hệ số 

*H*0: = 0

*H*1:  ≠ 0

Lấy p-value trong lệnh >summary(lm(y ~ x1+x2))

Do p-value=0.5678 > 0.05 nên chấp nhận gt H0 .

Tương tự Kiểm định giả thuyết 95% hệ số 

+Kiểm định giả thuyết 95% hệ số

*H*0: = 0

*H*1:  ≠ 0 (i=1, 2)

Do p-value: 1.345e-05< 0.05 nên bác bỏ gt H0.

+Dự báo y khi 

> predict(lm(y ~ x1+x2),data.frame(x1=8.5,x2=9))

1

15.69273

+ Khoảng dự đoán 96% cho y khi 

> predict(lm(y ~ x1+x2),data.frame(x1=8.5,x2=9),interval = c("prediction"),level = 0.96)

fit lwr upr

1 15.69273 13.9382 17.44726

Ta được (13.9382; 17.44726)

+ Khoảng tin cậy 96% cho giá trị trung bình của y khi 

> predict(lm(y ~ x1+x2),data.frame(x1=8.5,x2=9),interval = c("confidence"),level = 0.96)

fit lwr upr

1 15.69273 14.00048 17.38497

Ta được KTC (14.00048; 17.38497).

***Bài tập luyện tập: Câu hỏi tương tự như ví dụ trên***

**1.**Điểm của một lớp học gồm 9 sinh viên trong bài kiểm tra giữa kỳ (*x1*), điểm chuyên cần (x2), điểm tích cực (x3) và bài thi (*y*) như sau:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| x1 | 72 | 60 | 81 | 77 | 91 | 93 | 86 | 72 | 60 |
| x2 | 75 | 54 | 76 | 82 | 86 | 84 | 56 | 75 | 54 |
| x3 | 77 | 50 | 71 | 72 | 81 | 94 | 96 | 77 | 50 |
| y | 82 | 66 | 78 | 34 | 47 | 85 | 99 | 82 | 66 |



(HD:<lm(y ~ x1+x2+x3))

> x1=scan()

1: 72 60 81 77 91 93 86 72 60

10:

Read 9 items

> x2=scan()

1: 75 54 76 82 86 84 56 75 54

10:

Read 9 items

> x3=scan()

1: 77 50 71 72 81 94 96 77 50

10:

Read 9 items

> y=scan()

1: 82 66 78 34 47 85 99 82 66

10:

Read 9 items

> summary(lm(y ~ x1+x2+x3))

Call:

lm(formula = y ~ x1 + x2 + x3)

Residuals:

1 2 3 4 5 6 7 8 9

5.55178 0.05461 19.36773 -24.72780 -8.54927 10.76062 -8.06405 5.55178 0.05461

Coefficients:

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

(Intercept) 93.9117 39.1354 2.400 0.0616 .

x1 -0.9196 1.1961 -0.769 0.4767

x2 -0.8388 0.6011 -1.396 0.2217

x3 1.4501 0.7679 1.888 0.1176

---

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

Residual standard error: 16.14 on 5 degrees of freedom

Multiple R-squared: 0.6013, Adjusted R-squared: 0.3621

F-statistic: 2.514 on 3 and 5 DF, p-value: 0.1726

**2.** Một cuộc nghiên cứu về lượng mưa và lượng ô nhiễm không khí thải ra: lượng mưa hàng ngày tại khu vực A , *x1* (0,01 cm); lượng mưa hàng ngày tại khu vực B , *x2* (0,01 cm); lượng hạt ô nhiễm thải ra, *y*(mcg/cum), đã cho các số liệu sau:

|  |  |  |
| --- | --- | --- |
| x1 | x2 | y |
| 4,3 | 4,4 | 126 |
| 4,5 | 4,6 | 121 |
| 5,9 | 4,9 | 116 |
| 5,6 | 5,6 | 118 |
| 6,1 | 7,1 | 114 |
| 5,2 | 6,2 | 118 |
| 3,8 | 3,9 | 132 |
| 2,1 | 2,2 | 141 |
| 7,5 | 6,5 | 108 |



**3.** Từ tập dữ liệu ChiTieu2010.csv

a) ChiTieuGiaoDucTrongNam(x1); CTAnUongDipLeTrongNam(x2); SoNguoiTrongHo(y)



b) ChiTieuGiaoDucTrongNam(x1); CTAnUongDipLeTrongNam(x2); ChiTieuYTe(x3); SoNguoiTrongHo(y)



c) DieuTriNgoaiTru(x1); DieuTriNoiTru(x2); Thuoc(y)



***HD:***

***a)*** <lm(y ~ x1+x2)