

TRƯỜNG ĐẠI HỌC NHA TRANG

KHOA CÔNG NGHỆ THÔNG TIN

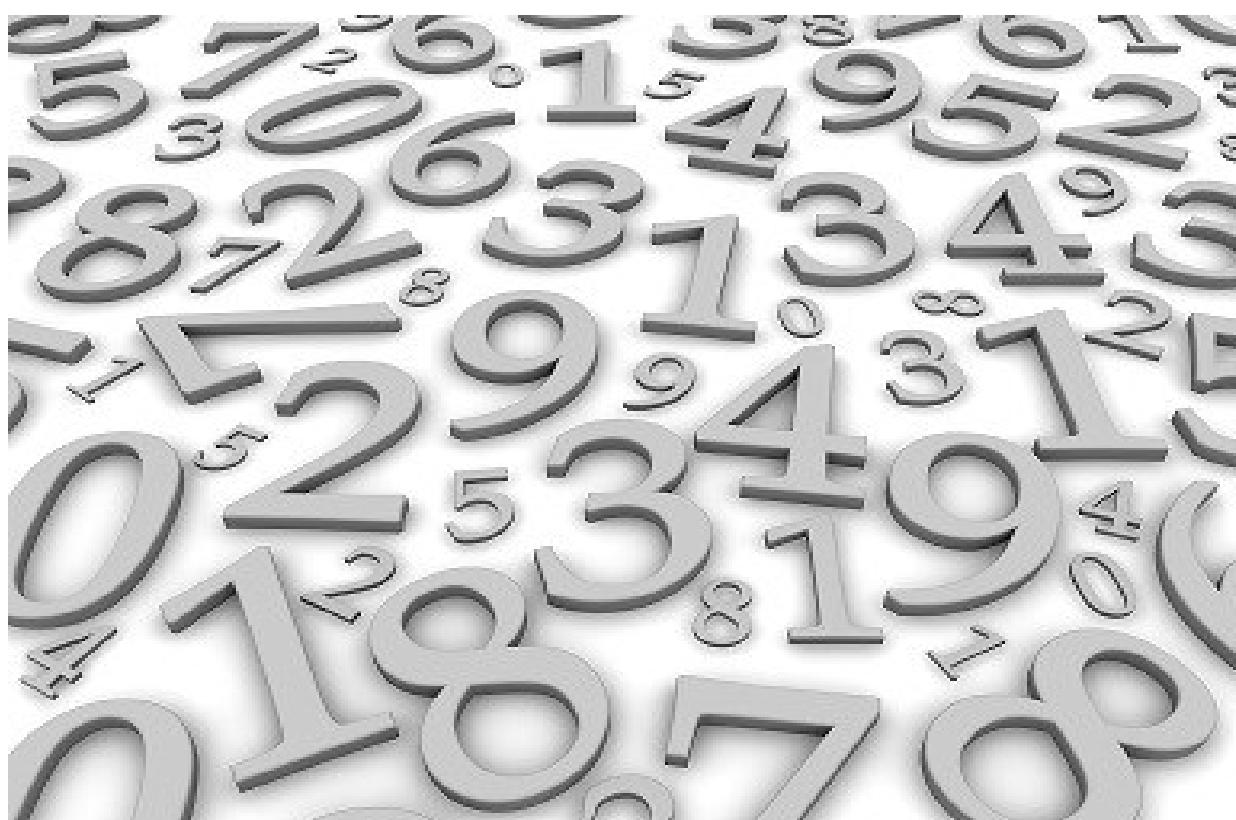


TS. NGUYỄN ĐỨC THUẦN

GIÁO TRÌNH

THỐNG KÊ MÁY TÍNH

Computational Statistics



NHA TRANG
2016

Lời nói đầu

Thống kê toán học là một khoa học biến dữ liệu thành thông tin hay tri thức. Đây là môn học không thể thiếu trong chương trình đào tạo chuyên ngành Công nghệ thông tin bậc đại học, cao học hiện nay. Không những thế, đối với các ngành học kỹ thuật khác người học cũng cần được trang bị về thống kê toán học để bố trí thí nghiệm, thu hoạch các kết quả nghiên cứu.

Ngày nay, thống kê kết hợp với máy tính góp phần làm phát triển khoa học thực nghiệm. Máy tính đã giúp khoa học thống kê thâm nhập vào thực tế, giải quyết các bài toán định lượng để xác định tính chất của các đối tượng nghiên cứu, xử lý.

Với yêu cầu là phải có một giáo trình tương đối đầy đủ, sát với chương trình Xác suất thống kê toán mà sinh viên được trang bị, chúng tôi đã biên soạn tài liệu này. Hiện nay, các tài liệu thống kê máy tính viết bằng tiếng Việt không nhiều, hơn nữa thường viết cho một lĩnh vực cụ thể nên người học sẽ gặp không ít khó khăn, nhất là sinh viên ngành công nghệ thông tin.

Với thời gian biên soạn giáo trình này tương đối hạn chế nên khó tránh khỏi khiếm khuyết. Rất mong nhận được những ý kiến đóng góp của bạn đọc để lần xuất bản sau được hoàn thiện hơn. Mọi ý kiến đóng góp xin gửi về địa chỉ ngducthanh@ntu.edu.vn

Xin chân thành cảm ơn Trường Đại học Nha Trang, Khoa Công nghệ Thông tin, cùng quý đồng nghiệp đã động viên, giúp đỡ để giáo trình kịp thời ra mắt bạn đọc.

Xin trân trọng cảm ơn!

Tháng 07 năm 2016

Nguyễn Đức Thuần

C H U O N G

1

GIỚI THIỆU NGÔN NGỮ R

1.1 NGÔN NGỮ R

R là một gói phần mềm dùng cho phân tích thống kê và đồ thị. R được tạo lập bởi Ross Ihaka và Robert Gentleman tại đại học Auckland, NewZealand vào những thập niên 90. Tiền thân của R là ngôn ngữ S được phát triển bởi John Chambers và cộng sự tại phòng thí nghiệm AT&T Bell. Tuy nhiên, giữa R và ngôn ngữ S có nhiều điểm khác biệt quan trọng (xem <http://cran.r-project.org/doc/FAQ/R-FAQ.html>).

Có thể tóm tắt về R như sau

- **R là 1 gói phần mềm thống kê**

R có nhiều công cụ hữu dụng cho mô hình hóa thống kê và đồ thị

- **R là 1 ngôn ngữ thông dịch**

R thực hiện các câu lệnh được gõ trực tiếp từ dấu nhắc

- **R là hướng đối tượng**

- Mọi thứ có thể thao tác thông qua các đối tượng đơn giản
- Các đối tượng có thể tạo lập bằng cách sử dụng lệnh gán <->
- Các đối tượng có thể là đại lượng vô hướng, vector, ma trận, danh sách, yếu tố, khung dữ liệu
- Các đối tượng có các lớp đặc trưng

- **R là gói phần mềm miễn phí, có nhiều phiên bản cho các hệ điều hành khác nhau**

- *Ưu điểm*

- Miễn phí
- Nhiều gói (packages) chuyên dụng
- Mã nguồn mở

- *Nhược điểm*

- Thuật ngữ khó hiểu
- Dùng câu lệnh
- Ký hiệu nhiều

- **R có thể download từ <http://cran.r-project.org/>**

Hiện nay có nhiều phiên bản hỗ trợ môi trường giao diện đồ họa cho R như RStudio, RCommander.. Trong giáo trình này không trình bày chi tiết đầy đủ về ngôn ngữ R, mà chỉ giới thiệu những khái niệm, đối tượng cơ bản của R để tiếp cận như một công cụ phục vụ cho các chủ đề thống kê. Có thể tham khảo chi tiết và đầy đủ về R trong <http://cran.r-project.org/>

1.2 DỮ LIỆU TRONG R

Dữ liệu trong R được lưu trữ trong các đối tượng (object). Mỗi đối tượng có một tên (gồm chữ thường, chữ hoa, số và ký hiệu “.” hay “_” (tên của đối tượng có phân biệt chữ hoa hay chữ thường). Ví dụ: kholo_age, nhatrang

Mỗi đối tượng có 2 thuộc tính nội tại (intrinsic) là **mode** và **length**.

mode là kiểu cơ sở của các phần tử/thành phần của đối tượng. Có 4 mode chính: số, ký tự, phức hợp và logic (numeric, character, complex, logical).

length là số phần tử/thành phần của đối tượng. Để xem mode, length của đối tượng dùng hàm *mode()*, *length()*.

1.2.1 Các đối tượng cơ bản

a. Các đại lượng vô hướng (Scalar)

là các đối tượng đơn, được tạo lập bởi lệnh gán

```
> x <-7
> y<-x*2+3
> x
[1] 7
> y
[1] 17
> mode(x)
[1] "numeric"
```

Trong R, các đại lượng chưa xác định hoặc chưa biết được biểu diễn bởi NA (*not available*). Một giá trị lớn có thể được biểu diễn qua lũy thừa cơ số e. R biểu diễn các giá trị số vô hạn $\pm\infty$ với Inf và -Inf, và các giá trị không phải là số bởi NaN (*not a number*)

```
> N<-3.5e23
> N
[1] 3.5e+23
> x<-5/0
> x
[1] Inf
> x-x
[1] NaN
```

Dữ liệu kiểu chuỗi được viết giữa 2 dấu nháy kép “”. Để hiển thị dấu nháy kép “” trong dữ liệu phải dùng dấu \ và hàm *cat()*. Có thể dùng dấu nháy đơn ‘ để thể hiện kiểu dữ liệu ký tự.

```
> x<- " Toi di hoc \"Thong ke may tinh\""
> x
[1] " Toi di hoc \"Thong ke may tinh\""
> cat(x)
Toi di hoc "Thong ke may tinh">
> y<-'Toi di hoc "Thong ke may tinh"'
> y
[1] "Toi di hoc \"Thong ke may tinh\""
> cat(y)
Toi di hoc "Thong ke may tinh">
```

b. Vector là một mảng các phần tử đơn có cùng kiểu. Tạo lập vector sử dụng hàm **c()** (*concatenation*), có thể đặt tên cho các phần tử

```
> x<-c(1,4,7,9,0)
> y<-c(hs1="Lan",hs2="Hue",hs3="Mai")
> y
  hs1   hs2   hs3
 "Lan" "Hue" "Mai"
> names(y)
[1] "hs1" "hs2" "hs3"
```

Nếu các phần tử được tạo lập khác kiểu, R tự động chuyên về kiểu “ràng buộc” bé nhất (*least restrictive type*).

```
> s<-c(T,12,"toi")
> s
[1] "TRUE" "12"   "toi"
> y<-c(23,"abc",T)
> y
[1] "23"   "abc"  "TRUE"
> z<-c(12,T)
> z
[1] 12   1
> t<-c("abc",12)
> t
[1] "abc" "12"
> p<-c(12,"abc")
> p
[1] "12"   "abc"
```

- Hàm hỗ trợ *seq()* tạo dãy các số:

Cú pháp: *seq(<gtri1>,<gtri2>[,<số gia>])*

Nếu số gia là 1 thì có thể viết: *<gtri1>:<gtri2>*

```
> a<-seq(2,8)
> a
[1] 2 3 4 5 6 7 8
> b<-2:8
> a
[1] 2 3 4 5 6 7 8
> c<-seq(10,20,0.5)
> c
[1] 10.0 10.5 11.0 11.5 12.0 12.5 13.0 13.5 14.0 14.5 15.0 15.5 16.0 16.5 17.0
[16] 17.5 18.0 18.5 19.0 19.5 20.0
```

- Hàm hỗ trợ *rep()* tạo dãy lặp các phần tử:

Cú pháp: *rep(<đối tượng>,<số lần lặp>)*

```
> a<-rep(5,8)
> a
[1] 5 5 5 5 5 5 5 5
> b<-rep(2:3,6)
> b
[1] 2 3 2 3 2 3 2 3 2 3
> c<-rep(c(7,9,13),3)
> c
[1] 7 9 13 7 9 13 7 9 13
> d<-rep(2:4,c(6,4,2))
> d
[1] 2 2 2 2 2 2 3 3 3 3 4 4
```

c. Ma trận (Matrix)

Một ma trận là một mảng 2 chiều. Một ma trận chính là một vector có thêm thuộc tính xác định số hàng và số cột. Ma trận được tạo lập bởi hàm matrix:

Cú pháp: matrix(<dữ liệu>, nrow=<số hàng>, ncol=<số cột>, byrow=<FALSE|TRUE>, dimnames= <NULL|tên hàng và cột>)

- Các giá trị dữ liệu sẽ điền theo thứ tự ưu tiên theo cột (mặc định) hay theo hàng (nếu byrow=TRUE), dimnames= cho tên hàng hay cột.

```
> matrix(1:12, 3, 4)
 [,1] [,2] [,3] [,4]
[1,]    1    4    7   10
[2,]    2    5    8   11
[3,]    3    6    9   12
> matrix(1:12, 3, 4, byrow=T)
 [,1] [,2] [,3] [,4]
[1,]    1    2    3    4
[2,]    5    6    7    8
[3,]    9   10   11   12
> matrix(1:12, 3, 4, byrow=T, dimnames=list(c("A", "B", "C"), c("I", "J", "H", "K")))
     I   J   H   K
A 1  2  3  4
B 5  6  7  8
C 9 10 11 12
```

Ma trận có thể được tạo lập bằng cách tiếp cận các giá trị thuộc tính dim

```
> x<-c(2,4,6,8,9,0,1,7)
> dim(x)<-c(2,4)
> x
 [,1] [,2] [,3] [,4]
[1,]    2    6    9    1
[2,]    4    8    0    7
```

Ma trận cũng có thể được tạo bằng cách gộp các vector bằng các hàm rbind(), cbind()

```
> cbind(A=1:4, B=5:8, D=9:12)
     A   B   D
[1,] 1  5  9
[2,] 2  6 10
[3,] 3  7 11
[4,] 4  8 12
> rbind(A=1:4, B=5:8, D=9:12)
     [,1] [,2] [,3] [,4]
A      1    2    3    4
B      5    6    7    8
D      9   10   11   12
```

Tính toán với ma trận:

- Tạo ma trận đơn vị

```
> A<-matrix(0, 4, 4)
> diag(A)<-1
> A
 [,1] [,2] [,3] [,4]
[1,]    1    0    0    0
[2,]    0    1    0    0
[3,]    0    0    1    0
[4,]    0    0    0    1
```

- Truy xuất phần tử của ma trận
 - Truy xuất theo hàng/cột : <tên ma trận>[<hàng>|<cột>]
 - Truy xuất phần tử <tên ma trận>[hàng,cột]

```
> A<-matrix(1:15, 5, 3)
> A
 [,1] [,2] [,3]
 [1,]    1    6   11
 [2,]    2    7   12
 [3,]    3    8   13
 [4,]    4    9   14
 [5,]    5   10   15
> A[2,2]
[1] 7
> A[,3]
[1] 11 12 13 14 15
> A[4,]
[1] 4 9 14
> A[-3,]
 [,1] [,2] [,3]
 [1,]    1    6   11
 [2,]    2    7   12
 [3,]    4    9   14
 [4,]    5   10   15
```

- Nhân 2 ma trận: phép toán nhân %*%

```
> A<-matrix(1:6, 2, 3)
> B<-matrix(c(4,2,5,1,7,9,4,3,7,8,2,3), 3, 4)
> A%*%B
 [,1] [,2] [,3] [,4]
 [1,] 35   67   48   29
 [2,] 46   84   62   42
```

- Cộng/trừ 2 ma trận

```
> A<-matrix(1:6, 2, 3)
> B<-matrix(7:12, 2, 3)
> A+B
 [,1] [,2] [,3]
 [1,]    8    12   16
 [2,]   10    14   18
> B-A
 [,1] [,2] [,3]
 [1,]    6     6     6
 [2,]    6     6     6
```

- Ma trận chuyển vị: hàm t(<ma trận>)

```
> B<-matrix(c(4,2,5,1,7,9,4,3,7,8,2,3), 3, 4)
> t(B)
 [,1] [,2] [,3]
 [1,]    4     2     5
 [2,]    1     7     9
 [3,]    4     3     7
 [4,]    8     2     3
```

- Ma trận nghịch đảo: hàm *solve(<ma trận>)*

```
> A<-matrix(c(2,3,5,1,7,8,3,9,2),3,3)
> solve(A)
      [,1]   [,2]       [,3]
[1,]  0.5272727 -0.2  0.10909091
[2,] -0.3545455  0.1  0.08181818
[3,]  0.1000000  0.1 -0.10000000
```

- Định thức: hàm *det(<ma trận>)*

```
> A<-matrix(c(2,3,5,1,7,8,3,9,2),3,3)
> det(A)
[1] -110
```

d. Yếu tố (Factor)

Yếu tố là đối tượng dữ liệu được sử dụng để phân loại dữ liệu và lưu lại như là các mức độ. Một yếu tố không chỉ bao gồm các giá trị của các biến phân loại (*categorical variable*) tương ứng, mà còn là mức độ có thể khác nhau của biến đó (ngay cả khi chúng không có mặt trong dữ liệu). Hàm tạo yếu tố:

Cú pháp: *factor(x, levels=sort(unique(x)),na.last=TRUE, labels=levels,exclude=NA, order=is.ordered(x))*

```
> factor(1:3)
[1] 1 2 3
Levels: 1 2 3
> factor(1:3,levels=1:5)
[1] 1 2 3
Levels: 1 2 3 4 5
> factor(1:3,labels=c("A", "B", "C"))
[1] A B C
Levels: A B C
> factor(1:5,exclude=4)
[1] 1     2     3     <NA> 5
Levels: 1 2 3 5
> aa<-factor(c(2,4),levels=2:5)
```

e. Danh sách (List)

Danh sách là dãy các đối tượng. Tạo danh sách bằng hàm *list()*:

Cú pháp: *list([tên 1=>đối tượng 1], [tên 2=>đối tượng 2],..,[tên n=>đối tượng n])*

```
> a<-list(1:3,c("A","B"),4:5,cuoi=c(5))
> a
[[1]]
[1] 1 2 3

[[2]]
[1] "A" "B"

[[3]]
[1] 4 5

$cuoi
[1] 5
```

f. Khung dữ liệu (Data frame)

Khung dữ liệu là một bảng dữ liệu, có vai trò như một quan hệ trong CSDL quan hệ. Mỗi khung dữ liệu có thể xem là một danh sách các vector hay các yếu tố cùng kích thước có quan hệ với nhau. Mỗi dòng ứng với các quan sát (*observation*), một cột ứng với một biến (*variable*). Mỗi cột có thể có kiểu dữ liệu khác nhau. Để tạo khung dữ liệu sử dụng hàm *data.frame()*

Cú pháp: *data.frame (<vector 1>, <vector 2>,...<vector n>)*

```
> x<-1:4;n<-10;M<-c(10,35); y<-2:4
> data.frame(x,n)
  x  n
1 1 10
2 2 10
3 3 10
4 4 10
> data.frame(x,M)
  x  M
1 1 10
2 2 35
3 3 10
4 4 35
> data.frame(x,y)
Error in data.frame(x, y) :
  arguments imply differing number of rows: 4, 3
```

- Tạo lập dataframe

Hàm *edit(data.frame())*

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

- Tách dữ liệu (*tương tự phép chọn trong CSDL QH*)

Hàm *subset(<dữ liệu>, <điều kiện>)*

```
> x<-1:4;M<-c(12,35,10,18)
> R1<-data.frame(x,M)
> subset(R1,M>12)
  x  M
2 2 35
4 4 18
```

- Kết nối dữ liệu (*tương tự phép kết nối tự nhiên trong CSDL QH*)

Hàm *merge(<dữ liệu1>, <dữ liệu2>, by="thuộc tính kết nối", all=TRUE))*

```
> x<-1:4;M<-c(12,35,10,18)
> R1<-data.frame(x,M)
> N<-c(45,23,67,89); K<-c("Nam","Nu","Nu","Nam")
> R2<-data.frame(x,N,K)
> R1
  x  M
1 1 12
2 2 35
3 3 10
4 4 18
```

```

> R2
  x  N   K
1 1 45 Nam
2 2 23 Nu
3 3 67 Nu
4 4 89 Nam
> d<-merge(R1,R2,by="x",all=T)
> d
  x  M  N   K
1 1 12 45 Nam
2 2 35 23 Nu
3 3 10 67 Nu
4 4 18 89 Nam

```

1.2.2 Các phép toán cơ bản

Các phép toán cơ bản

Phép toán					
Số học	So sánh		Logic		
+	cộng	<	nhỏ hơn	!x	$NOT(x)$
-	trừ	>	lớn hơn	x&y	$x AND y$
*	nhân	\leq	nhỏ hơn hoặc bằng	x&&y	$x AND y$
/	chia	\geq	lớn hơn hoặc bằng	x y	$x OR y$
$^{\wedge}$	lũy thừa	$=\equiv$	bằng	x y	$x OR y$
$\%\%$	modulo	!=	khác	xor(x,y)	$x XOR y$
$\%/\%$	chia lấy phần nguyên				

Các hàm số học thông thường

Căn bậc hai: \sqrt{x}	sqrt(x)	Hàm $\cos(x)$, $\sin(x)$, $\tg(x)$, $\arcsin(x)$, $\arccos(x)$, $\arctg(x)$	$\cos(x)$, $\sin(x)$, $\tan(x)$, $\text{asin}(x)$, $\text{acos}(x)$, $\text{atan}(x)$
Logarit nepe: $\ln(x)$	log(x)	Hàm tổng $\sum_{i=1}^3 x_i$	$X <- c(x_1, x_2, x_3)$ sum(X)
Logarit cơ số 10: $\log_{10}(x)$	log10(x)	Hàm e^x	exp(x)
Logarit cơ số 2: $\log_2(x)$	log2(x)	Hàm $ x $	abs(x)

Chú ý:

- So sánh các giá trị số học: So sánh bằng nhau nên dùng hàm `all.equal()`

```

> x<-0.45; y<-3*0.15
> x==y
[1] FALSE
> all.equal(x,y)
[1] TRUE

```

1.2.3 Biểu thức (Expression)

Một biểu thức là một dãy các ký tự tạo ra một cảm biến trong R. Tất cả các lệnh có hiệu lực là một biểu thức. Khi một lệnh được gõ trực tiếp từ bàn phím, nó được lượng giá bởi R và thi hành nếu hợp lý. Trong nhiều trường hợp, rất hữu dụng khi xây dựng một biểu thức chưa được lượng giá: Tạo các biểu thức bằng hàm `expression()`. Lượng giá các biểu thức sử dụng hàm `eval()`

1.2.4 Chuyển đổi đối tượng (Converting object)

Có thể chuyển đổi các kiểu dữ liệu của các đối tượng bằng cách dùng hàm có dạng: `as.<đối tượng>(<đối tượng>)`

Chuyển đổi	Hàm	Qui tắc
numeric	<code>as.numeric</code>	FALSE → 0 TRUE → 1 “1”, “2”,.. → 1,2,.. “A”,.. → NA
logical	<code>as.logical</code>	0 → FALSE Các số khác → TRUE “FALSE”, “F” → FALSE “TRUE”, “T” → TRUE Các ký tự khác → NA
character	<code>as.character</code>	1,2... → “1”, “2”,.. FALSE → “FALSE” TRUE → “TRUE”

```

> fac <- factor(c(1,10))
> fac
[1] 1 10
Levels: 1 10
> as.numeric(fac)
[1] 1 2
> fac2 <- factor(c("Male", "Female"))
> fac2
[1] Male   Female
Levels: Female Male
> as.numeric(fac2)
[1] 2 1
> as.numeric(as.character(fac))
[1] 1 10
  
```

1.2.5 Câu lệnh IF..ELSE

Cú pháp: `if(<biểu thức logic> <câu lệnh 1> [else <câu lệnh 2>])`

```

> x<-c("what","is","truth")
> if ("Truth" %in% x) {
+ print("Truth is found")
+ } else {
+ print("Truth is not found")
+ }
[1] "Truth is not found"
  
```

1.2.6 Câu lệnh SWITCH

Cú pháp: `switch(<biểu thức>, <tr.hợp 1>, <tr.hợp 2>,..)`

- Nếu giá trị của biểu thức không phải là 1 chuỗi ký tự thì nó được chuyển qua giá trị số nguyên (n), khi đó tr.hợp n sẽ được thực hiện.

- Nếu giá trị biểu thức là chuỗi ký tự, trường hợp ứng với chuỗi ký tự được thực hiện

```
> switch("a",A=1,B=5,a=6)
[1] 6
> switch(2,A=1,B=5,a=6)
[1] 5
```

1.2.7 Câu lệnh lặp repeat, while, for

a. Vòng lặp repeat

Cú pháp: repeat { <câu lệnh>
 if <biểu thức logic> {
 break
 }
 }

```
> v<- "Hello"
> x<-2
> repeat {
+   print(v)
+   x<-x+1
+   if (x>4) {
+     break
+ }
[1] "Hello"
[1] "Hello"
[1] "Hello"
```

b. Vòng lặp while

Cú pháp: while (<biểu thức logic>) {
 <câu lệnh> }

```
> v<- "Hello"
> x<-2
> while (x<5) {
+   print(v); x<-x+1
+
[1] "Hello"
[1] "Hello"
[1] "Hello"
```

c. Vòng lặp for

Cú pháp: for (<bien đếm> in <gtrị đầu:gtrị cuối>) { ds câu lệnh }

```
> v<- LETTERS[1:4]
> for (i in v) { print(i) }
[1] "A"
[1] "B"
[1] "C"
[1] "D"
> # Cộng từ 1 đến 10
> x<-0
> for (i in 1:10) {
+ x<-x+i
+
> print(x)
[1] 55
```

1.2.8 Hàm tự xây dựng:

Cú pháp: <tên hàm> <- function (<ds tham số>){ds câu lệnh}

```
> tong<-function(a,b,c) {
+ print(a+b+c)
> tong(2,5,8)
[1] 15
```

1.3 XUẤT/NHẬP DỮ LIỆU (DATA IMPORT/EXPORT)

a. Xác định thư mục làm việc

- Lấy thư mục hiện hành: Hàm `getwd()`
- Thiết lập thư mục hiện hành: Hàm `setwd(<thư mục>)`

```
> getwd()
[1] "C:/Users/admin/Documents"
> setwd("d:/Thuan/TKMT")
> getwd()
[1] "d:/Thuan/TKMT"
```

b. Lưu trữ và nhập dữ liệu từ file dữ liệu của R

Dữ liệu của ngôn ngữ R có thể lưu trữ vào file .Rdata bằng hàm `save()`. Sau đó có thể đọc lại từ R bởi hàm `load()`. Trong đoạn mã sau, hàm `rm()` xóa đối tượng a từ R

```
> a<- 1:10
> getwd()
[1] "d:/Thuan/TKMT"
> save(a,file="vidu1.Rdata")
> rm(a)
> load("vidu1.Rdata")
> print(a)
[1] 1 2 3 4 5 6 7 8 9 10
```

c. Nhập và lưu trữ dữ liệu một file dạng text

- (1) Nhập dữ liệu: sử dụng hàm `read.table(<"tên file">, header=TRUE)`

Trong lệnh này `header=TRUE`, có nghĩa là yêu cầu R đọc dòng đầu tiên của file đó như là tên của các cột

```
> qh<-read.table("ds.txt",header=T)
> qh
  Maso    Ten Namsinh Mucluong
  1     1   Binh    1975      3.12
  2     2   Lan     1982      3.10
  3     3   Hung    1980      4.20
  4     4   Tan     1980      3.12
  5     5 Chinh    1985      4.35
```

- (2) Lưu dữ liệu: sử dụng hàm `write.table(<"tên file">, row.names=F)`

```
> write.table(qh,file="dsluu1.txt",row.names=F)
```

d. Nhập và lưu trữ dữ liệu một file .CSV

- (1) Nhập dữ liệu: sử dụng hàm `read.csv(<"tên file">, header=TRUE)`

- (2) Lưu dữ liệu: sử dụng hàm `write.csv(<"tên file">, row.names=F)`

(Hàm này dùng để đọc các file excel. Các file excel trước đó phải được lưu với phần mở rộng CSV)

Ví dụ sau tạo lập một dataframe df1 và lưu trữ nó dưới dạng file .csv bằng hàm write.csv(). Sau đó, dataframe được nạp từ file df2 bằng hàm read.csv()

```
> var1<-1:5
> var2<-(1:5)/10
> var3<-c("R","and","Data Mining", "Examples", "Case Studies")
> df1<-data.frame(var1,var2,var3)
> names(df1)<-c("Biennguyen","Bienthuc","Bienkytu")
> write.csv(df1,"vidul1.csv", row.names=F)
> df2<-read.csv("vidul1.csv")
> print(df2)
  Biennguyen Bienthuc    Bienkytu
1           1      0.1          R
2           2      0.2        and
3           3      0.3  Data Mining
4           4      0.4   Examples
5           5      0.5 Case Studies
```

Chú ý: có thể kết nối và truy xuất CSDL thông qua ODBC, xem [3]. Với kết nối này có thể đọc/ghi file excel với phần mở rộng .xls.

TÀI LIỆU THAM KHẢO

- [1] Phân tích số liệu và biểu đồ bằng R, Nguyễn Văn Tuấn, cran.r-project.org/doc/contrib/Intro_to_R_Vietnamese.pdf
- [2] R for Beginners, Emmanuel Paradis, [cran.r-project.org /doc/contrib/Paradis-rdebuts_en.pdf](http://cran.r-project.org/doc/contrib/Paradis-rdebuts_en.pdf), 2005
- [3] R and Data Mining: Examples and Case Studies, Yanchang Zhao, <http://www.RDataMining.com>, 2013

C H U Ơ N G

2

THỐNG KÊ HỌC

2.1 THỐNG KÊ

Thống kê là tập hợp các phương pháp dùng để thu thập, trình bày và mô tả đặc tính của dữ liệu nhằm mục đích rút ra các quy luật chi phối các hiện tượng đang quan sát và đưa ra một quyết định.

Cơ sở của lý thuyết thống kê là lý thuyết xác suất và thống kê toán. Hiện nay thống kê được ứng dụng rộng rãi trong nhiều lĩnh vực như: thống kê dân số, thống kê xã hội, thống kê trong kinh doanh, thống kê trong y học, thống kê trong giáo dục,..

2.2 NGUỒN GỐC VÀ SỰ PHÁT TRIỂN CỦA THỐNG KÊ HỌC

Về mặt lịch sử môn thống kê ra đời và phát triển được nhờ hai hiện tượng riêng lẻ:

a. *Nhu cầu của nhà nước về các số liệu thống kê*

- Thời cổ đại: Thống kê để nhà nước nắm được nhân lực và tài lực nhằm chiêu mộ quân sĩ và thu thuế (Ai cập, Hy Lạp, La mã...)
- Thời trung cổ: các cơ sở của Giáo hội (Thiên chúa...) ghi chép và lưu trữ hồ sơ về hộ tịch...
- Ngày nay: các dữ liệu được tập hợp, phân loại, lưu trữ, truy tìm bằng các hệ thống thông tin, được khai thác có hiệu quả bằng các phương pháp thống kê và phân tích dữ liệu. Các lĩnh vực khoa học khác nhau đều sử dụng các phương pháp thống kê: vật lý, hoá học, di truyền học, khí tượng học, kinh tế học, xã hội học...

b. *Sự phát triển của lý thuyết xác suất trong Toán học:*

- Thế kỷ XVI và XVII: Các bài toán về các trò chơi may rủi đặt ra cho các nhà toán học thời đó: Pascal, Fermat, Leibnitz ..
- Thế kỷ XVIII: De Moivre, Gauss, Laplace: nghiên cứu về luật phân phối chuẩn.
- Thế kỷ XIX: Francis Galton, Karl Pearson: nghiên cứu về lý thuyết hồi quy và tương phản.
- Thế kỷ XX: Gosset: nghiên cứu phân phối t; Fischer: nghiên cứu phân phối F

2.3 CHỨC NĂNG CỦA THỐNG KÊ

Thống kê có 2 chức năng cơ bản sau:

- Thống kê mô tả (*Descriptive Statistics*) là các phương pháp sử dụng để tóm tắt hoặc mô tả một tập hợp dữ liệu. Đó là các phương pháp: thu thập dữ liệu, sắp xếp dữ liệu, trình bày tóm tắt dữ liệu, phân tích dữ liệu..

- Thống kê suy luận (*Inferential Statistics*) là các phương pháp mô hình hóa trên các dữ liệu quan sát để giải thích những biến thiên có tính ngẫu nhiên và không chắc chắn nhằm rút ra các suy diễn về quá trình, hay các đặc trưng của các đối tượng nghiên cứu. Đó là các phương pháp: ước lượng thống kê, kiểm định giả thuyết thống kê, phân tích phương sai, hồi qui tương quan, dự báo thống kê..

2.4 MỘT SỐ KHÁI NIỆM CƠ BẢN TRONG THỐNG KÊ

- **Quần thể/ Tổng thể** (*Population*) là tập hợp các đơn vị (hay các phần tử) thuộc hiện tượng nghiên cứu, cần được quan sát, thu thập và phân tích theo một số đặc trưng nào đó. Mỗi đơn vị (hay phần tử) tạo thành quần thể thống kê gọi là đơn vị quần thể.

Trong một số tài liệu thuật ngữ quần thể còn gọi là: tổng thể, dân số, đám đông.

Ví dụ: Để nghiên cứu chỉ số IQ của sinh viên Đại học Nha Trang, quần thể cần nghiên cứu là toàn bộ sinh viên Đại học Nha Trang.

Tùy theo đặc điểm của đối tượng nghiên cứu mà người ta phân loại quần thể thành các loại:

- Quần thể bộc lộ: là các quần thể mà ta có thể quan sát trực tiếp hay nhận biết được các đơn vị của quần thể (ví dụ như quần thể sinh viên 1 trường Đại học, quần thể các ngân hàng thương mại..)
- Quần thể tiềm ẩn: là quần thể mà ta không thể trực tiếp quan sát các đơn vị của quần thể (ví dụ: quần thể những người thích đi du lịch, quần thể những người nghiên thuốc lá..)
- Quần thể hữu hạn: số lượng đơn vị của quần thể là hữu hạn.
- Quần thể vô hạn: số lượng đơn vị của quần thể là vô hạn.
- Quần thể đồng chất: các đơn vị của quần thể giống nhau một số đặc điểm liên quan trực tiếp đến mục đích nghiên cứu.
- Quần thể không đồng chất: các đơn vị của quần thể không giống nhau một số đặc điểm liên quan trực tiếp đến mục đích nghiên cứu.
- **Mẫu** (*Sample*): là một số đối tượng được chọn ra từ quần thể theo một phương pháp nào đó. Các đặc trưng của mẫu được dùng để suy lý các đặc trưng của quần thể.
- **Quan sát** (*Observation*): là các đối tượng cơ sở cần thu thập dữ liệu nhằm phản ánh hiện tượng nghiên cứu. Người ta gọi 1 đơn vị của mẫu là một quan sát.
- **Đặc điểm thống kê** (*Characteristic*): là các tính chất quan trọng liên quan đến nội dung nghiên cứu được thể hiện trên các đơn vị quần thể (khái niệm này là khái niệm thuộc tính trong CSDLQH).

Có 2 loại đặc điểm thống kê:

- **Đặc điểm định tính:** là đặc điểm không thể biểu diễn trực tiếp bằng giá trị số (ví dụ: tôn giáo, dân tộc, loại hình doanh nghiệp..)
- **Đặc điểm định lượng:** là các đặc điểm có thể biểu diễn trực tiếp bằng giá trị số còn gọi là lượng biến (ví dụ: mức lương, tuổi, quy mô vốn của doanh nghiệp...). Các lượng biến số này có thể rời rạc hay liên tục
 - **Lượng biến rời rạc:** là các giá trị số đếm được (ví dụ: số nhân khẩu trong gia đình; số môn thi lại của một sinh viên..)
 - **Lượng biến liên tục:** là các giá trị không đếm được, các giá trị của lượng biến có thể chiếm 1 đoạn/khoảng trên trục số (ví dụ: trọng lượng, chiều cao của sinh viên..)
- **Chỉ tiêu thống kê:** là biểu hiện khái quát đặc điểm về mặt lượng của toàn bộ quần thể trong điều kiện không gian, thời gian nhất định. Có 2 loại chỉ tiêu thống kê:
 - **Chỉ tiêu khối lượng:** phản ánh qui mô của quần thể (ví dụ: tổng dân số của một quốc gia, tổng quy lương của một xí nghiệp, số sản phẩm của một công ty...)
 - **Chỉ tiêu chất lượng:** phản ánh về khái quát đặc điểm về tính chất, trình độ phổ biến, quan hệ so sánh trong quần thể. Chỉ tiêu này thường mang ý nghĩa phân tích và là kết quả so sánh giữa các chỉ tiêu khối lượng (ví dụ: năng suất lao động trung bình của một công nhân, tỷ lệ hộ nghèo của một địa phương, năng suất thu hoạch của một giống cây trồng..)
- **Kiểu dữ liệu:** Có nhiều kiểu dữ liệu dùng để biểu diễn các thuộc tính của các đối tượng, Có thể liệt kê sáu loại dữ liệu phổ biến:

Kiểu dữ liệu	Ý nghĩa
Định danh (Nominal data)	là tập các nhãn dùng để mô tả, phân loại (categories) các đối tượng. Ví dụ: tên màu, mã nhân viên
Nhi phân (Binary data)	là một trường hợp đặc biệt của kiểu dữ liệu định danh, các dữ liệu thuộc kiểu này chỉ mang một trong 2 giá trị. Ví dụ: kiểu boolean (true, false), giới tính(nam,nữ)
Thứ tự (Ordinary data)	là tập các phần tử chỉ định một thứ tự được sắp. Ví dụ: Xếp loại (Kém, Trung bình, Khá, Giỏi)
Số nguyên (Integer)	là tập các số nguyên. Các phần tử thuộc kiểu này có thể chịu tác động của các phép toán số học để kết xuất phần tử mới.
Khoảng (Interval data)	Dữ liệu khoảng (một đôi khi nhận giá trị là số nguyên) là một tập giá trị mà các phần tử cách nhau (thường dùng làm các thang đo). Ví dụ: Nhiệt độ được đo theo độ C;
Tỷ lệ-khoảng (Ratio-scaled data)	Tương tự kiểu dữ liệu khoảng, điểm khác biệt là các phần tử thuộc kiểu dữ liệu này có thể so sánh như là bội số với nhau. Dữ liệu kiểu tỷ lệ có thể thực hiện các phép nhân, chia. Ví dụ: Trọng lượng, 10kg là hai lần 5kg; Sự khác biệt giữa 1 và 2 tương tự như khác biệt giữa 3 và 4

Các thuộc tính được phân thành hai loại dựa vào kiểu dữ liệu của chúng:

- **Phân loại (categorical)**: các thuộc tính có dữ liệu thuộc kiểu *Định danh, Nhị phân, Thứ tự*
- **Liên tục (continuous)**: các thuộc tính có dữ liệu thuộc kiểu *Số nguyên, Khoảng, Tỷ lệ-khoảng*

2.5 QUÁ TRÌNH NGHIÊN CỨU THỐNG KÊ

Quá trình nghiên cứu thống kê thường được thực hiện theo mô hình gồm các bước:

- Xác định vấn đề nghiên cứu, mục tiêu, nội dung, đối tượng nghiên cứu
- Xây dựng hệ thống các khái niệm, Các chỉ tiêu thống kê
- Thu thập dữ liệu thống kê
- Xử lý số liệu: Kiểm tra, chỉnh lý, sắp xếp → Phân tích, thống kê sơ bộ → Phân tích thống kê thích hợp
- Phân tích & giải thích kết quả
- Trình bày các kết quả nghiên cứu.

2.6 CÁC KỸ THUẬT LẤY MẪU

Lấy mẫu nhằm rút trích các quan sát mang tính đại diện cho các quan sát tổng thể. Mục tiêu của việc chọn mẫu phải phản ánh trung thực, đại diện quần thể. Các kỹ thuật lấy mẫu đúng đắn sẽ giúp người xử lý đạt được mục tiêu này.

Có 2 nhóm kỹ thuật lấy mẫu là kỹ thuật lấy mẫu xác suất và kỹ thuật lấy mẫu phi xác suất. Phương pháp lấy mẫu xác suất dựa trên nguyên tắc lựa chọn mẫu ngẫu nhiên, nhóm kỹ thuật lấy mẫu phi xác suất dựa trên ý đồ định trước của người xử lý.

2.6.1 Kỹ thuật lấy mẫu xác suất (*Probability sampling*)

2.6.1.1 Lấy mẫu xác suất đơn giản (*Simple random sampling*)

Lấy mẫu xác suất đơn giản là phương pháp chọn mẫu trong đó mỗi đơn vị tổng thể được chọn với sự ngẫu nhiên như nhau. Để tiến hành kỹ thuật này, các bước tiến hành:

1. Tạo danh sách các đơn vị quần thể, sắp xếp theo một tiêu chí nào đó.
2. Xây dựng chỉ mục cho danh sách có được ở bước 1.
3. Tạo ra các giá trị ngẫu nhiên (dùng các phần mềm, bốc thăm, quay số..), các đơn vị tổng thể có chỉ mục tương ứng với các giá trị ngẫu nhiên tạo sẽ được chọn vào mẫu.

2.6.1.2 Lấy mẫu ngẫu nhiên hệ thống (*Systematic sampling*)

Qui trình chọn mẫu ngẫu nhiên hệ thống:

1. Tạo danh sách các đơn vị quần thể, sắp xếp theo một tiêu chí nào đó. Đánh số thứ tự cho các đơn vị trong danh sách. Tổng số đơn vị của danh sách là N.
2. Xác định cỡ mẫu muốn lấy, giả sử là n.

3. Chia N đơn vị thành k nhóm theo công thức $k=N/n$, k được gọi là khoảng cách chọn mẫu.
4. Trong k đơn vị đầu tiên ta chọn ra 1 đơn vị, các đơn vị tiếp theo được chọn có khoảng cách so với đơn vị đầu được chọn là $2k, 3k, \dots (i*k \text{ mod } N)$.

2.6.1.3 Lấy mẫu cả khối (*Clustering sampling*)

Qui trình lấy mẫu cả khối:

1. Quần thể được chia theo nhiều khối, mỗi khối được xem như một tổng thể con.
2. Khảo sát thống kê trên các mẫu tổng thể con. Phương pháp này thường được tiến hành khi không có danh sách đầy đủ quần thể cần nghiên cứu.

Ví dụ: Tổng thể là sinh viên của một trường đại học. Khi đó mỗi danh sách quần thể con là danh sách các lớp.

2.6.1.4 Lấy mẫu phân tầng (*Stratified sampling*)

Lấy mẫu phân tầng được sử dụng khi các đơn vị khảo sát quá khác nhau về tính chất liên quan đến vấn đề nghiên cứu khảo sát. Theo phương pháp này các đơn vị quần thể được chia theo từng tầng/lớp. Các đơn vị quan sát trong cùng một tầng/lớp có độ tương đồng lớn hơn giữa các đơn vị ở các tầng khác nhau. Sau đó các đơn vị mẫu được chọn từ các tầng này theo các phương pháp lấy mẫu xác suất thông thường như lấy mẫu ngẫu nhiên đơn giản hay lấy mẫu hệ thống. Chọn mẫu phân tầng có 2 vấn đề cần quan tâm là: tiêu chí phân tầng và cách thức phân bố số lượng mẫu vào các phân tầng (*xác định ngưỡng của độ tương đồng*)

2.6.2 Lấy mẫu phi xác suất (*Non-probability sampling*)

Trong thực tế, đôi khi do điều kiện về thời gian, thông tin về số lượng, cơ cấu các đơn vị quần thể, chi phí để thực hiện lấy mẫu ngẫu nhiên, người ta sử dụng các kỹ thuật lấy mẫu phi xác suất. Mẫu phi xác suất không đại diện để ước lượng cho toàn bộ quần thể, nhưng được chấp nhận trong nghiên cứu khám phá và trong kiểm định giả thuyết.

2.6.2.1 Lấy mẫu thuận tiện (*Convenient sampling*)

Lấy mẫu thuận tiện dựa trên thuận lợi hay dựa trên tính dễ tiếp cận của đối tượng. Lấy mẫu thuận tiện thường được dùng trong nghiên cứu khám phá, xác định ý nghĩa thực tiễn của vấn đề nghiên cứu hoặc khi muốn ước lượng sơ bộ của vấn đề quan tâm mà không muốn mất nhiều thời gian và chi phí.

2.6.2.2 Lấy mẫu định mức (*Quota sampling*)

Lấy mẫu định mức, người xử lý sẽ quyết định các quần thể con. Trong kỹ thuật này cần quan tâm đến tỷ lệ số lượng các đơn vị quan sát của các quần thể con với quần thể nghiên cứu.

2.6.2.3 Lấy mẫu phán đoán (*Judgement sampling*)

Lấy mẫu phán đoán dựa vào kinh nghiệm của người xử lý để chọn ra các đối tượng đối tượng phù hợp tham gia vào mẫu khảo sát. Vì vậy, tính đại diện của mẫu sẽ phụ thuộc rất nhiều về kiến thức, kinh nghiệm không những của người nghiên cứu điều tra

mà còn phụ thuộc vào kiến thức, kinh nghiệm của người đi thu thập dữ liệu.

TÀI LIỆU THAM KHẢO

- [1] Bài giảng Thống kê học, Trương Mỹ Dung (*tài liệu lưu hành nội bộ*)
- [2] Thống kê ứng dụng trong kinh tế-xã hội, Hoàng Trọng, Chu Nguyễn Mộng Ngọc, NXB Lao động-Xã hội, 2010.

C H U O N G

3

TRÌNH BÀY DỮ LIỆU BẰNG BẢNG & ĐỒ THỊ

3.1 TÓM TẮT VÀ TRÌNH BÀY DỮ LIỆU BẰNG BẢNG TẦN SỐ

Bảng tần số là bảng tổng hợp các biểu hiện có thể có của đặc điểm quan sát.

- Dạng cơ bản: bảng tần số gồm 3 cột, cột 1: mô tả các biểu hiện hoặc các khoảng giá trị được xác định cho dữ liệu, cột 2: mô tả tần số tương ứng với các biểu hiện hay giá trị, cột 3: là các tần suất (tỷ lệ %).

- Dạng mở rộng có thể có thêm cột thể hiện tính chất của dữ liệu: định tính hay định lượng, dữ liệu liên tục hay rời rạc hoặc cột chứa tần suất tích lũy.

Bảng tần số có thể có cấu trúc phức tạp hơn khi mô tả đặc điểm của mẫu nghiên cứu theo một biến (tiêu chí) dưới sự phân tích của một biến khác.

Ví dụ 1:

a. Tuổi của 30 sinh viên tại chức chuyên ngành Công nghệ phần mềm:

Độ tuổi	Tần số (SV)	Tần suất
19-24	9	30.00
24-29	10	33.33
29-34	8	26.67
34 trở lên	3	10.00
<i>Tổng</i>	30	100.00

b. Khu vực cư trú của thanh niên trong mẫu điều tra phân tách theo từng nhóm tuổi

Thanh niên trong mẫu điều tra		Nhóm tuổi					
		(14-17) tuổi		(18-21) tuổi		(22-25) tuổi	
		Tần số	Tần suất	Tần số	Tần suất	Tần số	Tần suất
Khu vực	Thành thị	1020	31.60	919	36.12	723	39.90
	Nông thôn	2208	68.40	1625	63.88	1089	60.10
<i>Tổng</i>		3228	100	2544	100	1812	100

3.2 PHÂN TỔ THÔNG KÊ

Phân tổ thông kê là căn cứ vào một số tiêu chí để sắp xếp các quan sát của 1 quần thể vào các tổ có tính chất khác nhau. Các quan sát trong cùng một tổ phải có độ tương đồng lớn hơn độ tương đồng giữa 1 cá thể cùng tổ với một cá thể khác tổ.

Khi phân tổ phải thỏa các tiêu chí:

- Các tổ không được trùng nhau, để cho các quan sát bất kỳ chỉ thuộc về 1 tổ.
- Tất cả các tổ được phân chia phải bao quát hết tất cả các giá trị hiện có của tập dữ liệu.

- Tránh không để tổ rỗng do không có quan sát nào thuộc về tổ đó.

Các bước của thủ tục phân tổ:

1. Xác định mục đích phân tổ: thông tin cần thu được sau khi phân tổ.
2. Xác định tiêu chí phân tổ
3. Xác định số tổ cần chia: Không có số qui định chính xác về số tổ cần chia. Thông thường số tổ cần chia phụ thuộc vào kinh nghiệm xử lý:
 - Nếu tiêu chí có ít lượng biến: mỗi lượng biến ứng với 1 tổ.
 - Nếu tiêu chí có nhiều lượng biến, ghép các lượng biến có độ tương đồng lớn vào cùng một tổ. Một số công thức xác định số tổ (k) mang tính chất tham khảo:
 - $k = [(2*n)^{1/3}]$
 - $k = [1+3.3*\log_{10}(n)]$
 với n là số quan sát của tập dữ liệu.
4. Xác định trị số khoảng cách tổ h :

$$h = \frac{X_{\max} - X_{\min}}{k}$$

Trong đó: X_{\max} , X_{\min} lần lượt là giá trị lớn nhất, giá trị nhỏ nhất của tập dữ liệu quan sát, và k : số tổ cần chia.

Ví dụ 2: Xét dữ liệu của 1 mẫu điều tra tuổi của 30 sinh viên tại chức ngành Công nghệ phần mềm như sau:

28	23	30	24	19	21	39	22	22	31	37
33	20	30	35	21	26	27	25	29	27	21
25	28	26	29	29	22	32	27			

- Xác định số tổ cần chia:

$$k = [(2*n)^{1/3}] = [(2*30)^{1/3}] = [3.9] = 4$$

- Xác định khoảng cách tổ h :

$$h = \frac{X_{\max} - X_{\min}}{k} = \frac{39 - 19}{4} = 5$$

- Các tổ được phân chia:

- Tổ 1: (19,24) tuổi
- Tổ 2: (24,29) tuổi
- Tổ 3: (29,34) tuổi
- Tổ 4: (34,39) tuổi

Các giá trị quan sát rơi vào cận, có thể chia như sau: giá trị rơi vào cận trên của tổ nào thì thuộc tổ đó, giá trị X_{\min} thuộc về tổ 1. (Theo kết quả thống kê bảng tần suất ở ví dụ 1.a)

3.3 TÓM TẮT VÀ TRÌNH BÀY DỮ LIỆU BẰNG BIỂU ĐỒ NHÁNH VÀ LÁ

Biểu đồ nhánh và lá là công cụ hữu hiệu trình bày dữ liệu một cách trực quan về cách thức phân tán dữ liệu.

Qui tắc lập biểu đồ: dữ liệu định lượng thể hiện bởi giá trị số sẽ được tách làm 2 phần: thân và lá. Phần thân là các chữ số “chung” của các dữ liệu cần biểu diễn, phần lá là các chữ số còn lại (*việc chọn chữ số “chung” tùy thuộc vào người xử lý số liệu*)

Hàm biểu diễn biểu đồ thân lá trong R: hàm *stem()*

Ví dụ 3: dữ liệu của 1 mẫu điều tra tuổi của 30 sinh viên tại chức ngành Công nghệ phần mềm ở ví dụ 2, biểu diễn bằng biểu đồ nhánh và lá thể hiện như sau:

Thân	Lá
1	9
2	8 3 4 1 2 2 0 1 6 7 5 9 7 1 5 8 6 9 9 2 7
3	0 9 1 7 3 0 5 2

Thực hiện bằng ngôn ngữ R:

> *stem(x, 0.3)*

Để biểu đồ dễ nhìn và hợp lý hơn người ta có thể sắp các giá trị lá theo thứ tự tăng dần, và tách thân số 2, số 3 thành 2 thân giúp biểu đồ cân đối hơn.

Thân	Lá
1	9
2	0 1 1 1 2 2 2 3 4
2	5 5 6 6 7 7 7 8 8 9 9 9
3	0 0 1 2 3
3	5 7 9

Thực hiện bằng ngôn ngữ R:

> *stem(x, 0.5)*

Chú ý:

- Nếu dữ liệu cần biểu diễn là khá lớn thì biểu diễn bằng bảng tần số hoặc đồ thị tỏ ra hợp lý hơn.
- Nếu dữ liệu có phần thập phân, thường làm tròn rồi mới biểu diễn phần thân lá

Ví dụ 4: biểu diễn các giá trị: - 2.5, 0.6, 5.7, 5.8, 12.2, 13.4, 25.3

Lấy số hàng chục làm thân, hàng đơn vị làm lá, làm tròn thập phân, biểu đồ thân lá:

> *x<-c(- 2.5, 0.6, 5.7, 5.8, 12.2, 13.4, 25.3)*
 > *stem(x)*

The decimal point is 1 digit(s) to the right of the |

```
-0 | 3
 0 | 166
 1 | 23
 2 | 5
```

3.4 TÓM TẮT VÀ TRÌNH BÀY DỮ LIỆU BẰNG ĐỒ THỊ

Trong phần trình bày này, các tham số có từ khóa giống nhau là có cùng chức năng. Vì vậy, ý nghĩa các tham số đã giải thích trong một hàm nào đó thì sẽ không được nhắc lại trong các hàm trình bày sau.

3.4.1 Biểu đồ phân phối tần số (Histogram)

Biểu đồ phân phối tần số biểu diễn tần số xuất hiện các giá trị của một biến (liên tục). Biểu đồ phân phối tần số tương tự như biểu đồ thanh (*bar chart*) nhưng điểm khác biệt là nhóm các giá trị vào các tầm (*range*) liên tiếp. Trên đồ thị này, diện tích mỗi cột tỉ lệ thuận với tần số các biểu hiện ứng với 1 đơn vị quan sát (mỗi đơn vị quan sát ứng với một nhóm giá trị tiêu chí quan sát)

Cú pháp: `hist(v,main,xlab,ylab,xlim,ylim,breaks,col,border,labels,density)`

(Trong Rstudio là hàm `histogram()`)

Các tham số thông dụng của hàm `hist()`:

v: vector dữ liệu

main: tên biểu đồ

xlab, ylab: tên trục ngang/trục đứng

xlim, ylim: giới hạn trục ngang/đứng(c(min,max))

col: màu của thanh

border: màu của đường viền

breaks

- breaks=n: biểu đồ ~ n thanh (bar)

- breaks=c(x₁,x₂,..,x_n): biểu đồ gồm n-1 thanh có các điểm phân chia theo trục hoành lần lượt là: x₁,x₂,..,x_n

density: số đường/inch (tô các thanh)

labels

- labels=TRUE: thêm vào mỗi thanh 1 số thể hiện số lượng quan sát ứng với “bins”
- labels=vectors: gán tên mỗi thanh ghi ứng với các phần tử của vector

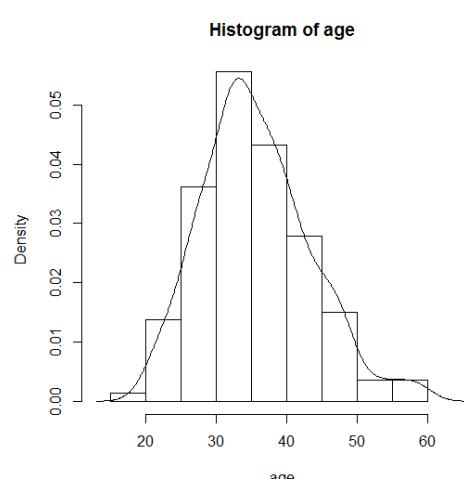
Ví dụ 5: a. (File help.Rdata download từ

www.math.smith.edu/sasr/datasets.php)

```
> setwd("d:/thuan/tkmt/")
> load("help.Rdata")
> dens<-density(age)
> xlim<-range(dens$x)
> ylim<-range(dens$y)

> hist(age,probability=TRUE,
+ xlim=xlim,ylim=ylim)

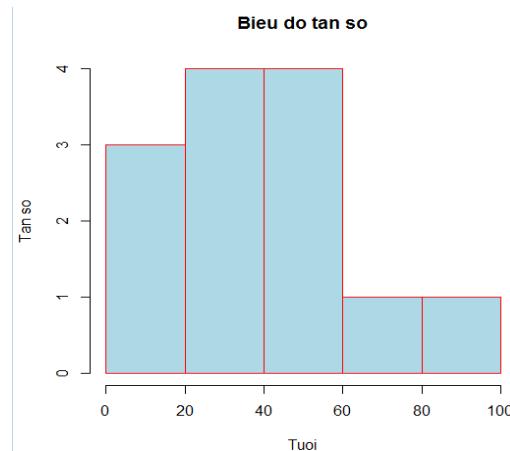
> lines(dens)
```



Để thể hiện màu, R có các bảng màu (palettes): rainbow, heat.colors, terrain.colors, topo.colors, cm.colors, grey.colors.

b.

```
> x<-c(15,24,35,18,19,53,47,68,87,45,57,39,39)
> hist(x,main="Bieu do tan so", col="lightblue",
+ border="red", xlab="Tuoi", ylab="Tan so")
```



3.4.2 Biểu đồ thanh (Bar chart)

Biểu diễn dữ liệu phân loại (*categorical data*)/ dữ liệu rời rạc (*discrete data*). Mỗi thanh tương ứng với một mức của 1 yếu tố (*factor*). Chiều cao của mỗi thanh là số lượng quan sát được ứng với mỗi mức.

Cú pháp: `barplot(h,xlab,ylab,main,names.arg,col)`

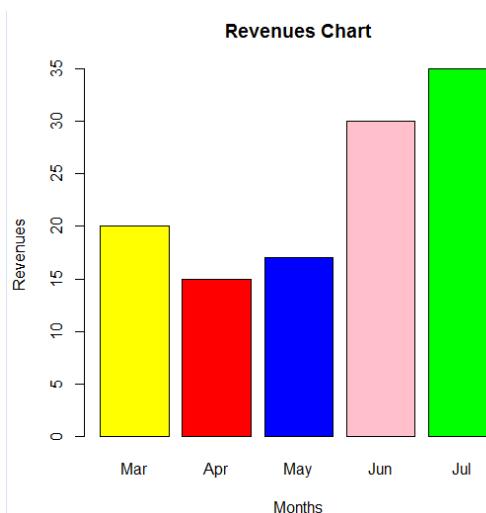
Trong đó: h: vector hay matrix

names.arg: Tên các thanh (biểu diễn bởi 1 vector)

Ví dụ 6: a. Xét dữ liệu

Tháng	3	4	5	6	7
Thu Nhập	20	15	17	30	35

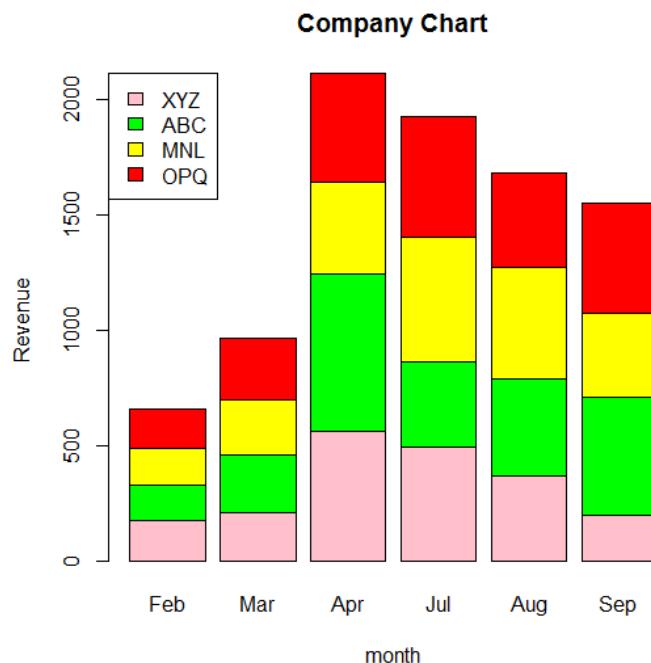
```
> h<-c(20,15,17,30,35)
> m<-c("Mar","Apr","May","Jun","Jul")
> color<-c("yellow","red","blue","pink","green")
> barplot(h,xlab="Months",ylab="Revenues", col=color,
+ names.arg=m,main="Revenues Chart")
```



b. Doanh thu 4 công ty XYZ, ABC, MNL, OPQ được thể hiện bởi bảng sau:

	2	3	4	7	8	9
XYZ	178	210	563	492	370	198
ABC	150	250	678	370	420	510
MNL	160	240	398	541	480	368
OPQ	170	265	473	520	410	472

```
> m<-matrix(c(178, 210, 563, 492, 370, 198, 150,
+ 250, 678, 370, 420, 510, 160, 240, 398, 541,
+ 480, 368, 170, 265, 473, 520, 410, 472),
+ nrow=4, ncol=6, byrow=T)
> m
[,1] [,2] [,3] [,4] [,5] [,6]
[1,] 178 210 563 492 370 198
[2,] 150 250 678 370 420 510
[3,] 160 240 398 541 480 368
[4,] 170 265 473 520 410 472
> month<-c("Feb", "Mar", "Apr","Jul", "Aug", "Sep")
> company<-c("XYZ","ABC","MNL","OPQ")
> color<-c("pink", "green","yellow","red")
> barplot(m, main="Company Chart",xlab="month", ylab="Revenue",
+ names.arg=month, col=color)
> legend("topleft",company, fill=color)
```



Một biểu đồ thanh được dùng nhiều trong quản lý dự án là biểu đồ pareto. Trong R có hàm chuyên dụng để vẽ biểu đồ này thuộc pakage qcc.

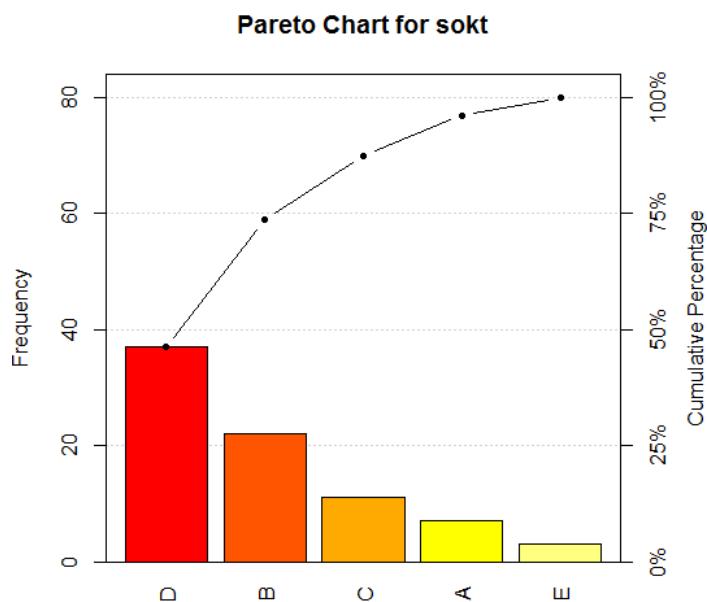
Ví dụ 7: Khi kiểm tra khuyết tật các sản phẩm ta có bảng:

Tên khuyết tật	Số lượng	Cộng dồn	% Khuyết tật	% Tích lũy
D	37	37	46.3	46.3
B	22	59	27.5	73.8
C	11	70	13.8	87.5
A	7	77	8.8	96.3
E	3	80	3.8	100
Tổng	80		100	

Biểu đồ pareto:

```
> tenkt<-c("A","B","C","D","E")
> sokt<-c(7,22,11,37,3)
> dl<-data.frame(tenkt,sokt)
> library(qcc)
Package 'qcc', version 2.5
Type 'citation("qcc")' for citing this R package in publications.
Warning message:
package 'qcc' was built under R version 3.0.3
> pareto.chart(sokt)

Pareto chart analysis for sokt
      Frequency Cum.Freq. Percentage Cum.Percent.
D          37       37     46.25      46.25
B          22       59     27.50      73.75
C          11       70     13.75      87.50
A           7       77      8.75      96.25
E           3       80      3.75      100.00
```



3.4.3 Biểu đồ dạng tròn (Piecharts)

Biểu diễn dữ liệu **dữ liệu rời rạc** (*discrete data*), có dạng hình tròn, chia thành các hình quạt, biểu diễn kết cấu của một tổng thể.

Cú pháp: `pie(x, labels, radius, main, col, clockwise)`

Trong đó: `x`:vector dữ liệu

labels: tên các mảng

radius: bán kính hình tròn (-1,1)

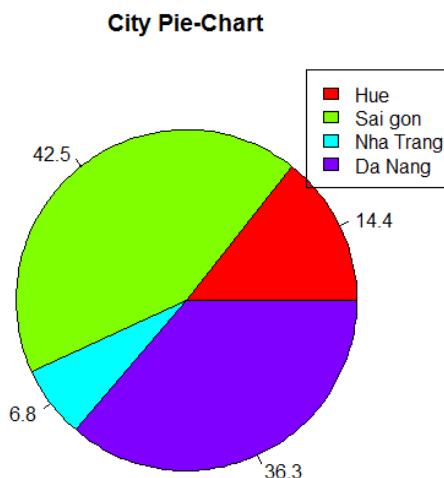
clockwise:

TRUE: các mảng cùng chiều kim đồng hồ

FALSE:các mảng ngược chiều kim đồng hồ

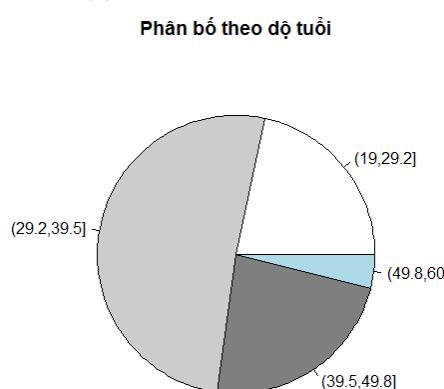
Ví dụ 8: a.

```
> x<-c(21,62,10,53)
> labels<-c("Hue","Sai gon","Nha Trang","Da Nang")
> piepercent<-round(100*x/sum(x),1)
> pie(x,labels=piepercent,main="City Pie-Chart",
+ col=rainbow(length(x)))
> legend("topright",labels,fill=rainbow(length(x)))
```



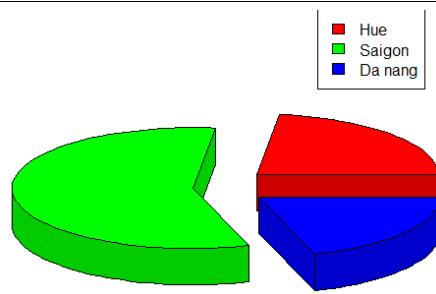
b. Biểu diễn biểu đồ thể hiện sự phân phối tần số ứng với 3 nhóm tuổi:

```
> setwd("d:/thuan/tkmt/")
> load("help.Rdata")
> attach(helpdata)
> nhomtuoi<-cut(age,4)
> slices<-c("white","grey80","grey50","lightblue")
> pie(table(nhomtuoi),main="Phân bố theo độ tuổi",col=slices)
```



Có thể vẽ biểu đồ 3D Pie Chart nhờ hàm *pie3D()* trong package: Plotrix

```
> library(plotrix)
> x<-c(23,57,20)
> pie3D(x,col=rainbow(length(x)),explode=0.2)
> legend("topright",labels,fill=rainbow(length(x)))
```



3.4.4 Biểu đồ dạng hộp (Boxplots)

Biểu đồ dạng hộp thể hiện đồng thời các thông tin: giá trị cực đại, giá trị cực tiểu, 3 tứ phân vị. Đây là biểu đồ cho thấy trực quan mối quan hệ các đại lượng thống kê kể trên.

Cú pháp: `boxplot(x, data, notch, varwidth, names, main)`

Trong đó: `x`: vector hay 1 công thức dữ liệu

`data`: dữ liệu lưu trữ dưới dạng 1 data frame

`notch`: TRUE/FALSE

`varwidth`:

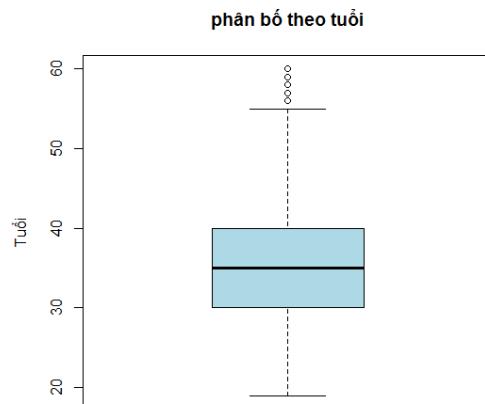
TRUE: cân chỉnh biểu đồ cân xứng với phạm vi dữ liệu

FALSE

`names`: nhãn các nhóm

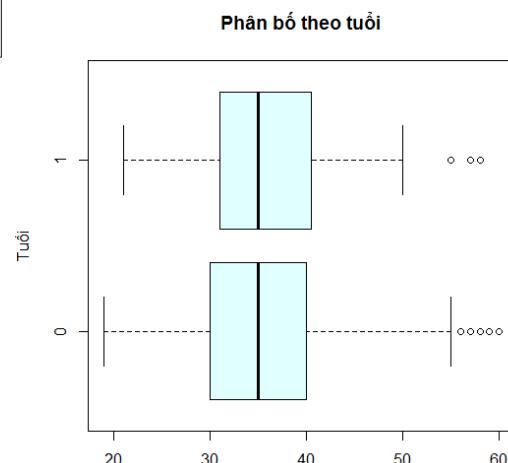
Ví dụ 9:

```
> boxplot(age, main="phân bố theo tuổi",
+ ylab="Tuổi", col="lightblue")
```



Thể hiện đồng thời hai biểu đồ dạng hộp trên cùng một đồ thị

```
> boxplot(age~female, main="Phân bố theo tuổi",
+ ylab="Tuổi", col="lightcyan", horizontal=T)
```



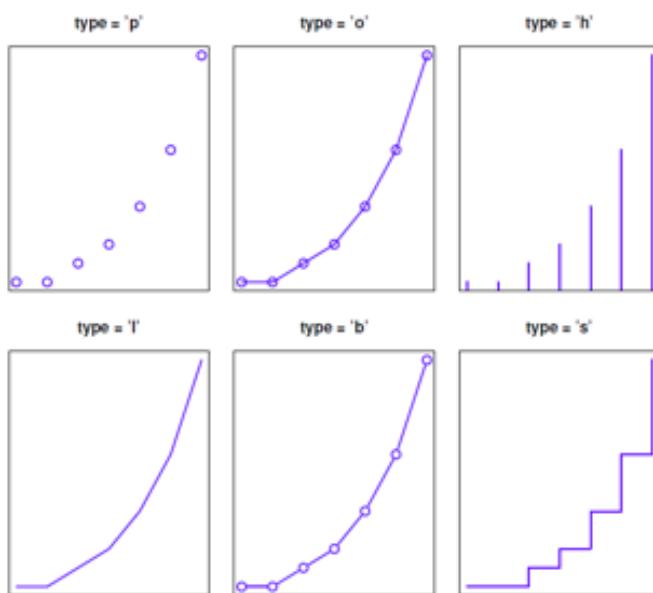
3.4.5 Biểu đồ đường (Line graphs)

Biểu đồ đường là một đồ thị nối các điểm bởi 1 đường. Các điểm này được xác định thứ tự thông qua tọa độ của chúng.

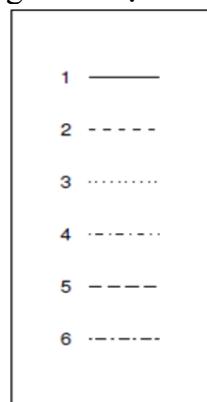
Cú pháp: `plot(v, type, col, xlab, ylab, lty)`

Trong đó: type là cách hiển thị biểu đồ điểm, cụ thể:

- type="p": vẽ điểm type="l": vẽ đường
- type="h": vẽ các đường thẳng đứng
- type="o": vẽ đường ngang qua các điểm
- type="b": vẽ đường không ngang qua các điểm
- type="s": vẽ đường theo kiểu bậc cấp

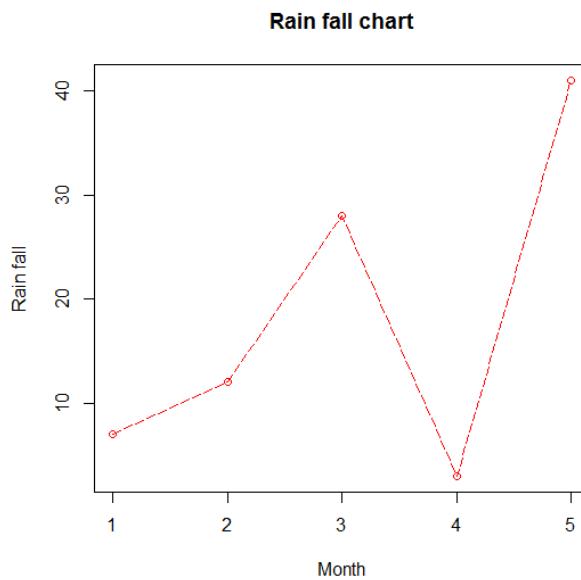


`lty=<giá trị số>`: kiểu đường hiển thị



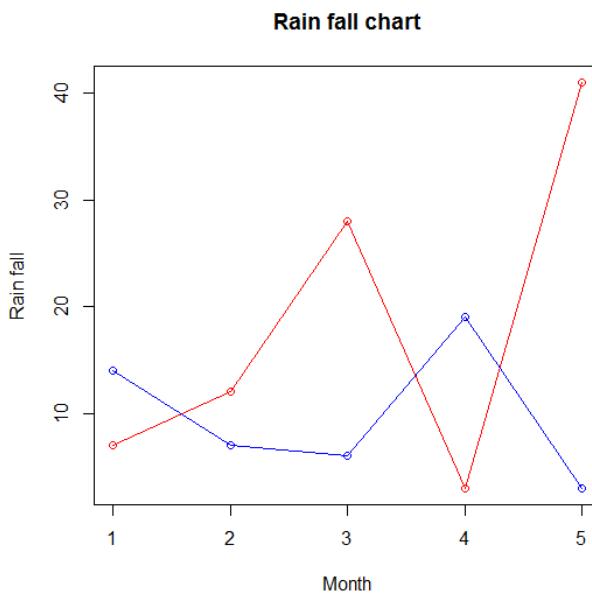
Ví dụ 10:

```
> v<-c(7,12,28,3,41)
> plot(v,type="o",col="red",xlab="Month",ylab="Rain fall",lty=5,
+ main="Rain fall chart")
```



Để vẽ nhiều đường trong cùng 1 đồ thị, kẽ từ đường thứ hai vẽ bằng hàm `lines()`, các tham số hàm `lines` như hàm `plot()`

```
> v<-c(7,12,28,3,41)
> plot(v,type="o",col="red",xlab="Month",ylab="Rain fall",
+ main="Rain fall chart")
> t<-c(14,7,6,19,3)
> lines(t,type="o",col="blue")
```



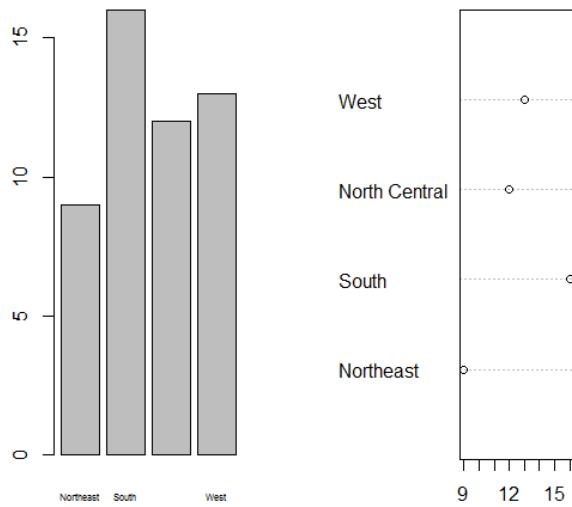
3.4.6 Biểu đồ điểm (Dot chart)

Đây là biểu đồ thể hiện thông tin tương tự như biểu đồ thanh. Tuy nhiên, ở đây thông tin của các thanh được thu gọn thành một điểm đặc trưng.

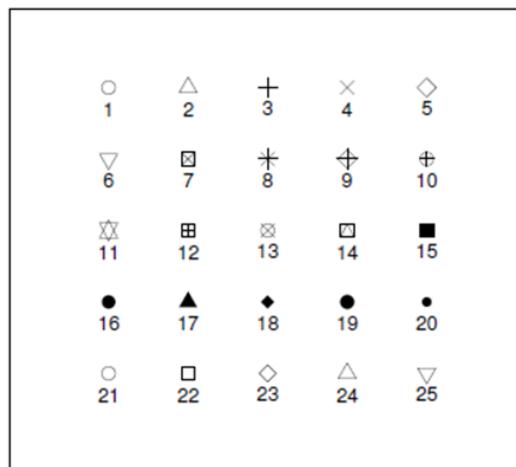
Cú pháp: `dotchart(v, labels, col, xlab, ylab, pch)`

Ví dụ 11: Xét dữ liệu **U.S state Facts and Features** trong package *gcc*

```
> require(gcc)
> par(mfrow=c(1,2))
> barplot(table(state.region),cex.names=0.5)
> x<-table(state.region)
> dotchart(as.vector(x),labels=names(x))
```



Tham số `pch=<giá trị số>`, thể hiện kiểu điểm hiển thị, cụ thể:



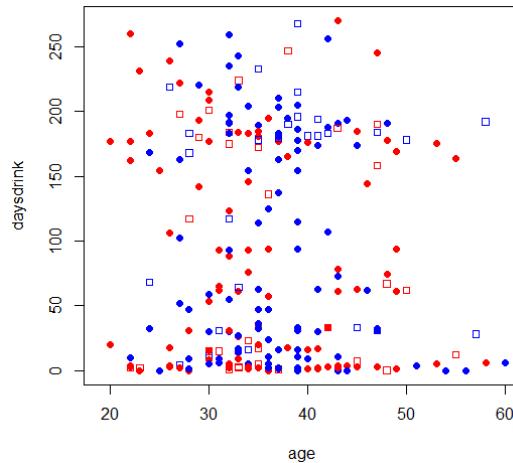
3.4.7 Biểu đồ tán xạ (Scatter plot)

Biểu đồ hiển thị các điểm trong mặt phẳng tọa độ. Mỗi điểm có tọa độ là giá trị tương ứng của 2 biến.

Cú pháp: `plot(x, y, main, xlab, ylab, xlim, ylim, pch)`

Ví dụ 12: Vẽ mối quan hệ giữa độ tuổi (*age*) và số ngày uống rượu (*daysdrink*) trong dữ liệu *help.Rdata* (nếu giới tính là nam vẽ điểm với ký tự số 16 (ô tròn), nữ ký hiệu số 22 (ô vuông))

```
> plot(age,daysdrink,pch=ifelse(female==0,16,22),col=c("red","blue"))
```



3.5 TIỀN XỬ LÝ DỮ LIỆU (DATA PREPROCESSING)

Trong thống kê, dữ liệu thu thập ban đầu thường cần phải tinh chỉnh lại mới đáp ứng được tính chính xác. Hai hiệu chỉnh phổ biến nhất là xử lý giá trị thiếu (*missing values*) và phần tử ngoại lệ (*outliers*). Quá trình hiệu chỉnh dữ liệu ban đầu trước khi tiến hành xử lý thống kê, phân tích dữ liệu gọi là tiền xử lý dữ liệu. Để đơn giản các thao tác xử lý dữ liệu trong phần này chúng tôi kết hợp sử dụng Rcmdr (R-Commander).

a. Dữ liệu thiếu (*Missing Data*)

Trong thực tế, để xử lý dữ liệu thiếu thường có 3 giải pháp:

1. Xóa các bản ghi (bộ) dữ liệu chứa giá trị thiếu.
2. Điền giá trị thiếu bởi giá trị trung bình
3. Sử dụng các công cụ toán học (như kỹ thuật phân lớp, lý thuyết tập mờ, tập thô, mạng neural.. để điền vào các giá trị thiếu)

Dễ thấy việc xóa bỏ các bản ghi mang giá trị thiếu là lãng phí và không hiệu quả khi lượng các bản ghi đó khá lớn trong mẫu thu thập được. Trong phạm vi giáo trình không đi vào các kỹ thuật của giải pháp 3 (tham khảo [3,7]), mà chỉ sử dụng giải pháp 1, 2.

Ví dụ 13: Xét tập dữ liệu Pima.tr2 (thuộc package MASS)

	npreg	glu	bp	skin	bmi	ped	age	type
198	0	106	70	37	39.4	0.605	22	No
199	1	118	58	36	33.3	0.261	23	No
200	8	155	62	26	34.0	0.543	46	Yes
201	2	134	70	NA	28.9	0.542	23	Yes
202	10	75	82	NA	33.3	0.263	38	No
203	0	146	70	NA	37.9	0.334	28	Yes
204	1	180	NA	NA	43.3	0.282	41	Yes
205	5	104	74	NA	28.8	0.153	48	No
206	9	164	78	NA	32.8	0.148	45	Yes
207	1	80	55	NA	19.1	0.258	21	No
208	4	171	72	NA	43.6	0.479	26	Yes
209	3	139	54	NA	25.6	0.402	22	Yes
210	3	122	78	NA	23.0	0.254	40	No
211	5	116	74	NA	25.6	0.201	30	No
212	6	195	70	NA	30.9	0.328	31	Yes
213	8	125	96	NA	NA	0.232	54	Yes
214	4	122	68	NA	35.0	0.394	29	No

Những dòng dữ liệu mang giá trị thiếu có giá trị NA.

1. Loại bỏ các dòng mang giá trị thiếu:

- Cú pháp: `na.omit(<tên dữ liệu>)`

Ví dụ: `na.omit(Pima.tr2)`

Trong R-Commander thực hiện như sau:

(Click) Data → Active data set → Remove cases with missing data → <điền tên thuộc tính chứa giá trị thiếu> và <cung cấp tên tập dữ liệu mới cho tập dữ liệu đã được xóa>.

Cũng có thể xóa một số bản ghi bằng cách cung cấp số hiệu các dòng:

(Click) Data → Active data set → Remove row(s) from active data → <điền số hiệu các dòng>.

2. Thay giá trị thiếu bằng giá trị trung bình

- Tính giá trị trung bình (*trừ các phần tử mang giá trị thiếu*)
 - Cú pháp: `mean(<dữ liệu>, na.rm=TRUE)`
- Thay giá trị trung bình cho các vùng thiếu dữ liệu
 - Cú pháp: `<tên dữ liệu>[is.na(Tên dữ liệu)] <- <giá trị trung bình>`

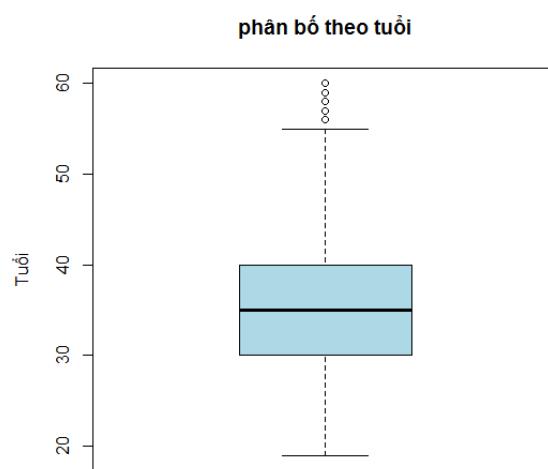
Ví dụ 14:

```
Pima.tr2$skin[is.na(Pima.tr2$skin)] <- mean(Pima.tr2$skin, na.rm=TRUE)
```

b. Phần tử ngoại lệ (*Outliers*)

Xử lý giá trị thiếu là một thách thức khi xử lý dữ liệu thô. Ngoài ra còn một thách thức khác là khi một giá trị thu thập có thể *đáng ngờ* do không theo mẫu tổng thể (*over patterns*) của phần dữ liệu còn lại. Các giá trị đó được gọi là phần tử ngoại lệ.

Ví dụ 15: Khi thống kê độ tuổi, một vài bản ghi có độ tuổi vượt hẳn (trên 58) so với các mẫu (ví dụ 8)



Biểu đồ boxplot, các độ đo tập trung (giá trị trung bình, trung vị ..) là các công cụ khá hữu hiệu trong việc phát hiện các phần tử ngoại lệ. Cụ thể như sau:

- 1) Phương pháp dựa vào phân phối chuẩn (*normal distribution*)

Nếu biến số X tuân theo luật phân phối chuẩn với giá trị trung bình μ và độ lệch chuẩn S thì 99% các giá trị của X đa phần nằm trong khoảng $[\mu-3*S, \mu+3*S]$. Các giá trị nào của X có giá trị thấp hơn $\mu-3*S$, hoặc lớn hơn $\mu+3*S$ có thể nghi ngờ là ngoại lệ.

2) Phương pháp dựa vào số trung vị

Các bước tiến hành bằng phương pháp này như sau:

- i. Tính trung vị của biến số: M
- ii. Tính độ khác biệt tuyệt đối của từng giá trị của biến là M , gọi kết quả là $d_i: d_i=|x_i-M|$
- iii. Tính trung vị của $d_i: Md$
- iv. Tính tỉ số: $t_i = \frac{d_i}{Md}$
- v. Nếu $t_i \geq 4.5$ thì có thể coi là ngoại lệ.

3) Phương pháp phi tham số (*non-parametric method*)

- i. Tìm các giá trị tứ phân vị, xác định Q1 (giá trị bách phân 25), Q3 (giá trị bách phân 75)
- ii. Tính độ trai Q1 và Q3: $IQR=Q3-Q1$
- iii. Tính giá trị thấp nhất của biến: $L=Q1-1.5*IQR$
- iv. Tính giá trị cao nhất của biến $U=Q3+1.5*IQR$
- v. Những giá trị nào của biến thấp hơn L hoặc cao hơn U, thì có thể xem là ngoại lệ.

Trong R không có hàm liệt kê các phần tử ngoại lệ. Hàm liệt kê phần tử ngoại lệ tự xây dựng

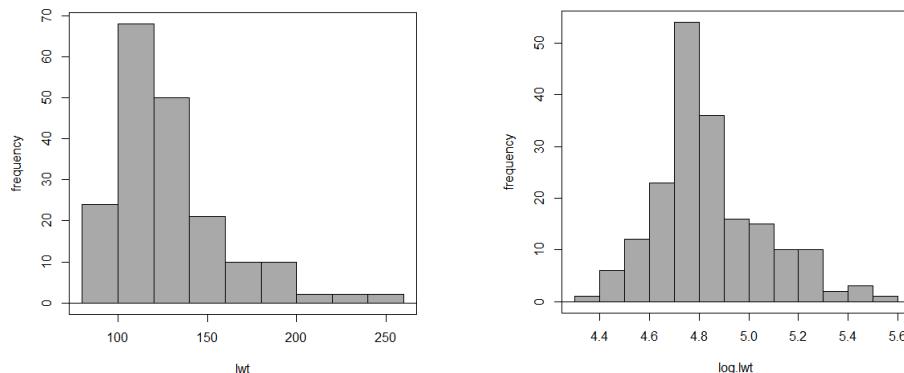
```
> gtridibiet<-function(x) {
+ kq<-summary(x)
+ Q1<-kq[2];Q2<-kq[5]
+ IQR<-Q3-Q1
+ L<-Q1-1.5*IQR;U<-Q3+1.5*IQR
+ x[x<L|x>U]}
```

c. Chuyển đổi dữ liệu (*Data Transformation*)

Nhằm giảm bớt các giá trị cực đoan (*extreme values*) trong quá trình phân tích dữ liệu, người ta sử dụng các kỹ thuật chuyển đổi dữ liệu. Hai hàm thường được dùng trong quá trình chuyển đổi dữ liệu là *hàm logarit* và *hàm lũy thừa bậc hai*. Hàm *logarit* sử dụng chuyển các biến lệch phải (*right-skewed*) dương, hàm *lũy thừa bậc hai* sử dụng chuyển các biến mang giá trị đếm được (*count variables*)

Ví dụ 16: Sử dụng tập dữ liệu birthwt trong package MASS, chuyển đổi logarit cho biến lwt như sau:

(Click) Data → Manage variables in active dataset → Compute new variable. Phía dưới thông báo New variable name, gõ: log.lwt , và dưới thông báo Expression to compute, gõ log(lwt). Sau đó dùng chức năng đồ thị Graphs để vẽ biểu đồ Histogram cho 2 biến lwt và log.lwt ta có:



Tương tự cho hàm lũy thừa bậc hai (gõ x^2 dưới thông báo Expression to compute)

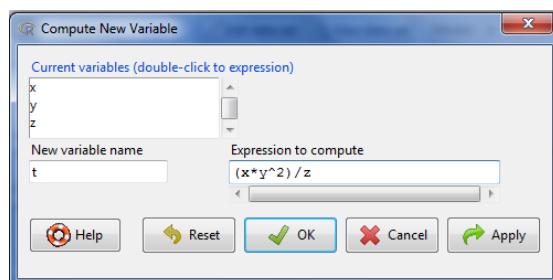
d. Tạo biến mới dựa trên hai hay nhiều biến đã có

Việc tạo ra các biến mới phục vụ thuận tiện cho việc phân tích dữ liệu cũng là một bước trong tiền xử lý dữ liệu. Trong phần này chúng tôi chỉ ra cách tạo ra biến mới từ các biến đã có

Ví dụ 17: Xét tập dữ liệu topo trong package MASS, chứa 3 biến x, y, z, chúng ta sẽ tạo biến t theo công thức $t = \frac{x * y^2}{z}$

Cách thực hiện như sau:

(Click) Data → Manage variables in active dataset → Compute new variable. Phía dưới thông báo New variable name, gõ: t, và dưới thông báo Expression to compute, gõ $(x * y^2) / z$



e. Tạo phân loại cho các biến số (*Creating categories for numerical variables*)

Một trong những kỹ thuật phổ biến được áp dụng trong tiền xử lý dữ liệu là tạo các biến phân loại (*categorical variables*) dựa trên các biến số. Điều này giúp cho việc dễ thấy mối quan hệ định danh được dễ dàng.

Ví dụ 18: Theo Center for Disease Control and Prevention (CDC) tiêu chuẩn cân nặng xác định theo BMI như sau:

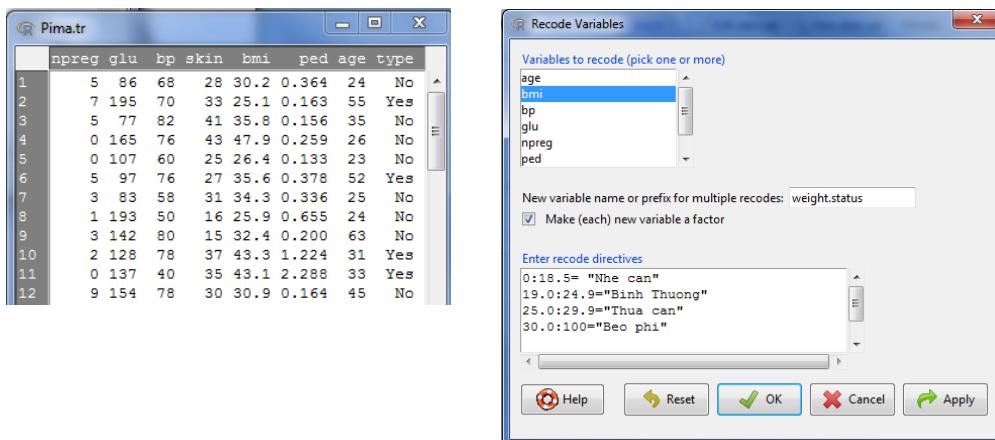
BMI	Chuẩn
Dưới 18.5	Nhẹ cân
18.5-24.9	Bình thường
25.0-29.9	Thừa cân
30.0 trở lên	Béo phì

Xét tập dữ liệu Pima.tr trong package MASS: chúng ta sẽ chia các đối tượng thành 4 nhóm Nhẹ cân, Bình thường, Thừa cân, Béo phì dựa trên chỉ số BMI bằng cách:

(Click) Data → Manage variables in active dataset → Recode variable

Chọn bmi tại cửa sổ Variables to recode (pick one or more)

Gõ weight.status vào: New variable name và điền vào cửa sổ Enter recode directives nội dung như hình sau:



Kết thúc ta có:



Lúc này vẽ đồ thị barplot với biến weight.status:

3.6 TỔNG KẾT

Phản biểu diễn dữ liệu bằng đồ thị trong ngôn ngữ R còn có khả năng thể hiện nhiều biểu đồ khác nữa, ví dụ: biểu đồ đồng hồ (*clock plot*), biểu đồ vòng (*contour plot*) biểu đồ hình hộp..

Do khuôn khổ của giáo trình, trong phần trình bày này chỉ trình bày các biểu đồ thông dụng thường gặp trong thống kê. Việc xây dựng các biểu đồ có thể thực hiện đơn giản hơn bằng cách sử dụng chức năng Graphs trong Rcmdr.

Có thể tham khảo thêm ở [1,2,4,8]

TÀI LIỆU THAM KHẢO

- [1] Biostatistics with R, Babak Shahbaba, Springer, 2012
- [2] Computational Statistics Using R and R Studio, An Introduction for Scientists, Randall Pruim, www.calvin.edu/~rpruim/talks/SC11/Seattle/IntroToR.pdf
- [3] Hung Quoc Nguyen, Duc Thuan Nguyen, An Improvement of Method Handling Missing Values in Incomplete Information System, *International Journal of Advanced Research in Computer Science and Software Engineering (IJARCSSE)*. Vol 3, Issue 9, September-2013
- [4] Introduction to Probability and Statistics Using R, G.Jay Kerns, First Edition, cran.r-project.org/web/packages/IPSUR/vignettes/IPSUR.pdf, 2010
- [5] Introductory Statistics with R, Peter Dalgaard, Second Edition, Springer, http://www.academia.dk/BiologiskAntropologi/Epidemiologi/PDF/Introductory_Statistics_with_R__2nd_ed.pdf, 2008
- [6] Learning Statistics with R: A tutorial for psychology students and other beginners, Version 0.4, Daniael Navarro, <http://learningstatisticswithr.com>, 2014
- [7] Nhập môn phát hiện tri thức & khai phá dữ liệu, Nguyễn Đức Thuần, NXB Thông tin và Truyền thông, 2013.
- [8] R for Beginners, Emmanuel Paradis, cran.r-project.org/doc/contrib/Paradis-rdebuts_en.pdf, 2005
- [9] Using R for Data Analysis and Graphics, Introduction, Code and Commentary, J.H Maindonald, cran.r-project.org/doc/contrib/usingR.pdf, 2008
- [10] Thống kê ứng dụng trong kinh tế-xã hội, Hoàng Trong, Chu Nguyễn Mộng Ngọc, NXB Lao động-Xã hội, 2010.
- [11] Phân tích số liệu và biểu đồ bằng R, Nguyễn Văn Tuân, cran.r-project.org/doc/contrib/Intro_to_R_Vietnamese.pdf

C H U O N G

4

TÓM TẮT DỮ LIỆU

4.1 CÁC ĐẠI LƯỢNG ĐO LUỒNG MỨC ĐỘ TẬP TRUNG CỦA DỮ LIỆU

4.1.1 Các đại lượng đo lường độ tập trung phổ biến

1. Trung bình cộng

a. Trung bình cộng (*Arithmetic mean*)

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$

Trong đó: x_i : giá trị quan sát thứ i của mẫu

n : kích thước mẫu.

Hàm tính trung bình cộng trong R: *mean(<đối tượng>)*

Ví dụ 19: Doanh số của 6 xí nghiệp may cho bởi bảng sau:

Xí nghiệp	A	B	C	D	E	F
Doanh số	25	17	34	26	43	35

Doanh số trung bình của 6 xí nghiệp:

$$\bar{x} = \frac{(25 + 17 + 34 + 26 + 43 + 35)}{6} = 30$$

```
> ds<-c(25,17,34,26,43,35)
> mean(ds)
[1] 30
```

Chú ý:

- Trung bình được tính trên tất cả các phần tử của tập quan sát được.
- Giá trị trung bình cộng bị chi phối bởi các giá trị cực biên, nên có thể không thể hiện đúng sự phân bố tập trung của của dữ liệu. Tuy nhiên, tập hợp dữ liệu tương đồng nhau (*homogeneous*) thì giá trị trung bình là số đo thích hợp để tóm tắt và cho biết đặc trưng của tập dữ liệu.
- Trong chương này, các độ đo sẽ tác động lên tập các quan sát. Tập quan sát có thể là: Tổng thể hay Mẫu **tùy mỗi ví dụ**, bạn đọc cần lưu ý điều này.

b. Trung bình cộng có trọng số (*Weighted mean*) hoặc dữ liệu có phân bổ

$$\bar{x}_w = \frac{\sum_{i=1}^n w_i x_i}{\sum_{i=1}^n w_i}$$

Trong đó: x_i : giá trị quan sát thứ i của mẫu

n : kích thước mẫu

w_i : trọng số/tần số xuất hiện của quan sát thứ i

Ví dụ 20: Có 5 tổ công nhân, biết sản phẩm làm ra của mỗi cá nhân trong tổ và số lượng thành viên của mỗi tổ được thể hiện bởi bảng sau:

Tổ	A	B	C	D	E
Số lượng thành viên	10	15	14	10	16
Sản phẩm của mỗi thành viên trong tổ	30	20	25	20	25

Số lượng sản phẩm trung bình mỗi công nhân làm được

$$\bar{x}_w = \frac{(10*30 + 15*20 + 14*25 + 10*20 + 16*25)}{(10+15+14+10+16)} = 23.84615$$

```
> sl<-c(10,15,14,10,16)
> sp<-c(30,20,25,20,25)
> ts<-sl*sp
> ts
[1] 300 300 350 200 400
> x<-sum(ts)/sum(sl)
> x
[1] 23.84615
```

c. Trung bình hình học (*Geometric Mean*)

Trung bình hình học của tập các quan sát x_1, x_2, \dots, x_n ($x_i > 0$) là trung bình nhân các giá trị của tập quan sát

Do đó giá trị trung bình $\bar{x} = (x_1 \times x_2 \times \dots \times x_n)^{1/n} = (\prod_{i=1}^n x_i)^{1/n}$

Độ đo trung bình hình học là độ đo hữu ích cho dữ liệu có chứa tỉ lệ tương quan (*aspect ratio*)

Mỗi quan hệ giữa trung bình hình học và trung bình logarit

$$(\prod_{i=1}^n x_i)^{1/n} = \exp\left(\frac{1}{n} \sum_{i=1}^n \ln(x_i)\right)$$

Ví dụ 21: So sánh thời gian thực hiện của 3 máy tính như sau:

	Máy 1	Máy 2	Máy 3
Chương trình 1	0.1	1	2
Chương trình 2	10	1	0.2
Trung bình cộng	5.05	1	1.1
Trung bình hình học	1	1	0.632

Trong trường hợp này máy 3 là máy chạy nhanh nhất.

Lập trình với R:

```

> A<-edit(data.frame())
> PC_fastest<-function(A) {
+ kt<-dim(A)
+ vt<-as.vector(rep(1,kt[2]))
+ for (i in 1:kt[1]) {d<-as.vector(A[i,])
+ vt<-vt*d}
+ Xt<-exp(1/kt[1]*log(vt))
+ for (i in 1:kt[2]) if (Xt[i]==min(Xt))
+ print(i)}

```

d. Trung bình cộng điều hòa (*Harmonic mean*)

Trung bình cộng điều hòa của tập các quan sát x_1, x_2, \dots, x_n ($x_i \geq 0$) là nghịch đảo trung bình cộng nghịch đảo các giá trị của tập quan sát, ký hiệu là H

$$\frac{1}{H} = \frac{\frac{1}{x_1} + \frac{1}{x_2} + \dots + \frac{1}{x_n}}{n}$$

$$\text{Do đó, } \bar{x} = H = n / \sum_{i=1}^n \frac{1}{x_i}$$

Khi một tập dữ liệu chứa các giá trị đại diện cho tốc độ thay đổi, trung bình cộng điều hòa là độ đo hữu ích thể hiện xu hướng trung tâm.

Trong trường hợp $n=2$ và $n=3$ ta có các công thức

$$H(x_1, x_2) = \frac{2x_1 x_2}{x_1 + x_2}$$

$$H(x_1, x_2, x_3) = \frac{3x_1 x_2 x_3}{x_1 x_2 + x_1 x_3 + x_2 x_3}$$

Ví dụ 22:

a. Trong Khoa học máy tính, đặc biệt là trong rút trích thông tin và học máy, trung bình điều hòa của số lượng chính xác và số lượng thu hồi dùng để lượng giá hiệu năng các thuật toán, gọi là độ đo F (*F-measure*). Xem [5], [6]

b. Nếu lái xe từ Nha Trang tới Bình Thuận với tốc độ 40 Km/h, và trở về với 60 Km/h, thì tốc độ trung bình tổng thể không phải là: 50 Km/h, mà phải là trung bình điều hòa

$$AM = (40 + 60) / 2 = 50 \quad HM = 2 / (1/40 + 1 / 60) = 48$$

để kiểm tra xem điều này là phù hợp với ví dụ đơn giản: giả sử khoảng cách Nha Trang Bình Thuận là 120km. Thời gian cần thiết cho lượt đi là 3 giờ thời gian lượt về là 2 giờ, tổng số là 5 giờ, và khoảng cách cả đi lẫn về là 240 km. Tốc độ trung bình là: $240\text{km}/5\text{h} = 48\text{km/h}$

2. Trung Vị (*Median*)

Số trung vị là giá trị ở đứng giữa của một dãy giá trị quan sát đã được sắp thứ tự tăng dần, ký hiệu M_e .

Giả sử có dãy giá trị quan sát được đã được sắp thứ tự

X	1	2	..	[n/2]	[(n+1)/2]	[(n+2)/2]	..	n
	x ₁	x ₂		x _[n/2]	x _[(n+1)/2]	x _[(n+2)/2]	..	x _n

Công thức tính M_e

$$M_e = \begin{cases} (x_{[n/2]} + x_{[(n+2)/2]})/2 & \text{nếu } n \text{ chẵn} \\ x_{[(n+1)/2]} & \text{nếu } n \text{ lẻ} \end{cases}$$

Hàm để tính trung vị trong R là *median(<đối tượng>)*

Ví dụ 23:

1. Với dãy X có n=5 phần tử

X	1	2	3	4	5
	10	12	18	20	30

Trung vị của dãy $M_e = 18 (=x[3])$

2. Với X có 6 phần tử

X	1	2	.3	4	5	6
	10	12	18	20	30	45

Trung vị của dãy $M_e = (18+20)/2=19 (=x[3]+x[4])/2$

Kết quả được cho bởi R:

```
> x<-c(10,18,12,30,20)
> median(x)
[1] 18
> x<-c(10,18,12,30,20,45)
> median(x)
[1] 19
```

Chú ý:

- Số trung vị không phụ thuộc vào các giá trị biên.
- Số trung vị chia dãy giá trị quan sát thành hai phần có tần số bằng nhau (50% giá trị lớn hơn số trung vị, 50% giá trị bé hơn số trung vị).

3. Yếu vị (*Mode*)

Là giá trị xuất hiện nhiều lần nhất trong dãy các giá trị quan sát được. Yếu vị không phụ thuộc vào các giá trị biên. Tuy vậy, yếu vị chỉ dùng với mục đích mô tả mẫu đang khảo sát. Nếu mẫu thay đổi, yếu vị có thể thay đổi. Trong một dãy các giá trị quan sát có thể có nhiều yếu vị.

Ví dụ 24:

1. Dãy các giá trị quan sát: 7 15 18 22 25 37

Không có yếu vị.

2. Dãy các giá trị quan sát: 7 15 18 25 25 37

Yếu vị là 25.

3. Dãy các giá trị quan sát: 7 7 15 18 25 25

Yếu vị là 7 và 25.

Trong R không có hàm tính yếu vị. Hàm tự xây dựng

```
> Mode<-function(x) {
+ y<-sort(x)
+ tam<-as.vector(rep(1,length(x)))
+ for (i in 1:length(x)) tam[i]<-dem(y,y[i])
+ if (min(tam)!=max(tam)){
+   for (i in 1:length(y)) if (tam[i]==min(tam)) so<-y[i]
+   for (i in 1:length(y)) if ((tam[i]==max(tam)) && (so!=y[i])){
+     print(y[i]); so<-y[i]}}}
```

Với hàm dem(x,y): có tác dụng đếm số lần y xuất hiện trong x.

4. Trung tầm (*Midrange*)

Trung tầm là trung bình cộng giá trị lớn nhất và giá trị nhỏ nhất của dãy giá trị quan sát.

Để tính giá trị trung tầm, trong R tính thông qua 2 hàm:

- *range(x)* ($\equiv c(\min(x), \max(x))$)
- *mean(x)*

Ví dụ 25: Doanh số của 6 xí nghiệp may cho bởi bảng sau:

Xí nghiệp	A	B	C	D	E	F
Doanh số	25	17	34	26	43	35

Trung tầm doanh số của 6 xí nghiệp là: $(17+43)/2=30$

Kết quả được cho bởi R:

```
> x<-c(27,17,34,26,43,35)
> range(x)
[1] 17 43
> mean(range(x))
[1] 30
```

4.1.2 Nhóm các đại lượng khác mô tả sự phân bố tập trung

1. Tứ phân vị (*Quartiles*): Chia tập dữ liệu quan sát thành 4 nhóm:

- *Tứ phân vị thứ nhất Q1*: Giá trị sao cho có 25% số quan sát nhỏ hơn nó, và 75% số quan sát lớn hơn nó.
- *Tứ phân vị thứ hai Q2*: Số trung vị, có 50% số quan sát nhỏ hơn nó và 50% số quan sát lớn hơn nó.
- *Tứ phân vị thứ ba Q3*: Giá trị sao cho có 75% số quan sát nhỏ hơn nó, và 25% số quan sát lớn hơn nó.

Để xem các giá trị tứ phân vị trong R có hàm: *summary(<đối tượng>)*

Ví dụ 26:

```
> x<-c(6, 7, 15, 36, 39, 40, 41, 42, 43, 47, 49)
> summary(x)
  Min. 1st Qu. Median Mean 3rd Qu. Max.
  6.00  25.50  40.00  33.18  42.50  49.00
> x<-c(6,7,15,36,39,40,41,42,43,47,49)
> summary(x)[c(2,5)]
  1st Qu. 3rd Qu.
  25.5    42.5
```

Có nhiều cách tính tứ phân vị (*giá trị các tứ phân vị Q1, Q2, Q3 ứng với các cách tính là khác nhau*) [<http://mathworld.wolfram.com/Quartile.html>].

method	1st quartile	1st quartile	3rd quartile	3rd quartile
	n odd	n even	n odd	n even
Minitab	$\frac{n+1}{4}$	$\frac{n+1}{4}$	$\frac{3n+3}{4}$	$\frac{3n+3}{4}$
Tukey (Hoaglin et al. 1983)	$\frac{n+3}{4}$	$\frac{n+2}{4}$	$\frac{3n+1}{4}$	$\frac{3n+2}{4}$
Moore and McCabe (2002)	$\frac{n+1}{4}$	$\frac{n+2}{4}$	$\frac{3n+3}{4}$	$\frac{3n+2}{4}$
Mendenhall and Sincich (1995)	$\left[\frac{n+1}{4} \right]$	$\left[\frac{n+1}{4} \right]$	$\left[\frac{3n+3}{4} \right]$	$\left[\frac{3n+3}{4} \right]$
Freund and Perles (1987)	$\frac{n+3}{4}$	$\frac{n+3}{4}$	$\frac{3n+1}{4}$	$\frac{3n+1}{4}$

Trong R, các giá trị tứ phân vị được tính theo [Freund and Perles].

2. Thập phân vị (*Deciles*)

Chia tập hợp đã sắp thứ tự thành những nhóm 1/10 số liệu. Phân vị thứ p ($0 < p < 10$) ở vị trí thứ i được xác định bởi công thức:

$$i = \frac{p}{10} * (n + 1)$$

Với n là số lượng số liệu.

3. Bách phân vị (*Percentiles*)

Chia tập hợp đã sắp thứ tự thành những nhóm 1/100 số liệu. Phân vị thứ p ($0 < p < 10$) ở vị trí thứ i được xác định bởi công thức:

$$i = \frac{p}{100} * (n + 1)$$

Với n là số lượng số liệu.

4.2 CÁC ĐẠI LƯỢNG ĐO LUỒNG MỨC DỘ PHÂN TÁN CỦA DỮ LIỆU

4.2.1 Tầm/Khoảng biến thiên (*Range*)

Tầm là hiệu số giữa số lớn nhất và số nhỏ nhất của tập dữ liệu quan sát.

$$R = a_{\max} - a_{\min}$$

a_{\max}, a_{\min} lần lượt là giá trị lớn nhất, giá trị nhỏ nhất của tập dữ liệu quan sát.

Tầm chỉ biết mức trải của dữ liệu, nhưng không cho biết mức độ phân bố dữ liệu.

Trong R, không có hàm tương ứng với hiệu số giữa số lớn nhất và nhỏ nhất của tập dữ liệu quan sát mà chỉ có hàm hiển thị giá trị nhỏ nhất và giá trị lớn nhất.

```
> x<-c(12,15,23,45,67,89)
> range(x)
[1] 12 89
> min(x)
[1] 12
> max(x)
[1] 89
```

4.2.2 Độ trai giữa (Interquartile Range) IQR

Là hiệu số giữa tứ phân vị thứ 3 (Q3) và tứ phân vị thứ nhất (Q1)

$$IQR = Q_3 - Q_1$$

Trong R không có hàm trực tiếp tính đại lượng này.

4.2.3 Phương sai và độ lệch chuẩn

a. Phương sai (*Variance*) là số đo đánh giá mức độ biến thiên của các giá trị quan sát quanh giá trị trung bình. Phương sai của mẫu n giá trị quan sát được:

x_1, x_2, \dots, x_n ký hiệu s^2 được xác định bởi công thức:

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{(n-1)}$$

Công thức tương đương:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n x_i^2 - \frac{n(\bar{x})^2}{n-1}$$

Trong R, hàm tính phương sai là *var()*

Ví dụ 27: Doanh số của 6 xí nghiệp may cho bởi bảng sau:

Xí nghiệp	A	B	C	D	E	F
Doanh số	25	17	34	26	43	35

Phương sai doanh số của 6 xí nghiệp may:

```
> x<-c(25, 17, 34, 26, 43, 35)
> var(x)
[1] 84
```

b. Độ lệch chuẩn (*Standard Deviation*) là căn bậc 2 của phương sai.

$$sd = \sqrt{s^2}$$

Ý nghĩa: Đa số các giá trị quan sát được nằm trong phạm vi

$$(\bar{x} - \sigma, \bar{x} + \sigma)$$

Với \bar{x} : giá trị trung bình, σ ($=sd$) độ lệch chuẩn.

Trong R, hàm tính độ lệch chuẩn là *sd()*.

Ví dụ 28: xét doanh số của 6 xí nghiệp may ở ví dụ trên, ta có

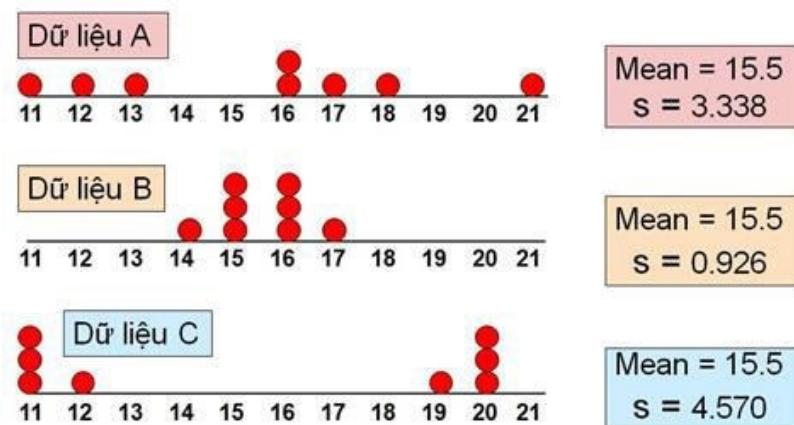
```
> x<-c(25, 17, 34, 26, 43, 35)
> var(x)
[1] 84
> sd(x)
[1] 9.165151
> mean(x)
[1] 30
```

Chú ý: Người ta xây dựng hàm tính phương sai theo công thức

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{(n-1)}$$

vì: $\sum_{i=1}^n (x_i - \bar{x}) = 0$, và công thức: $\sum_{i=1}^n |x_i - \bar{x}|$, là hàm trị tuyệt đối không thuận lợi trong các suy diễn thống kê (như lấy đạo hàm).

So sánh các độ lệch chuẩn



Nếu tập dữ liệu có số lượng quan sát n khá lớn và tương đối đối xứng, công thức tính độ lệch chuẩn có thể tính xấp xỉ theo giá trị lớn nhất (max) và giá trị nhỏ nhất (min) các quan sát như sau:

$$s = (\max - \min) / \sqrt{n}, \text{ nếu } n < 12$$

$$s = (\max - \min) / 4, \text{ nếu } 20 < n < 40$$

$$s = (\max - \min) / 5, \text{ nếu } n \approx 100$$

$$s = (\max - \min) / 6, \text{ nếu } n > 400$$

c. Sai số chuẩn (Standard Error)

Phân phối mẫu (Sampling Distribution): Nếu chúng ta lặp lại việc chọn mẫu N lần (N khá lớn) thì ta sẽ có một tập hợp N mẫu, mỗi mẫu gồm n phần tử rút từ quần thể. Giả sử ta đang khảo sát giá trị trung bình của quần thể thì với N mẫu ta có N giá trị trung bình của mẫu. Đây chính là một phân phối mẫu giá trị trung bình (*tập hợp giá trị trung bình của các mẫu*).

Sai số chuẩn (Standard error) chính là độ lệch chuẩn của tập hợp mẫu sau khi chọn mẫu N lần. Sai số chuẩn là độ lệch chuẩn của giá trị trung bình trong N lần chọn mẫu. Vì vậy sai số chuẩn phản ánh độ dao động hay biến thiên của các số trung bình mẫu

$$SE = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

Trong đó: σ là độ lệch chuẩn của quần thể đã biết.

Trong trường hợp σ của quần thể chưa biết thì ta sử dụng *độ lệch chuẩn* mẫu để ước lượng độ lệch chuẩn của quần thể

$$SE = s / \sqrt{n}$$

Sử dụng sai số chuẩn: Sai số chuẩn còn được gọi là độ lệch chuẩn của giá trị trung bình, nó chỉ ra sự khác biệt giữa giá trị trung bình mẫu và giá trị trung bình của quần thể. Trung bình của quần thể thường không xác định được do số lượng các đơn vị của quần thể rất lớn. (xem thêm [5])

4.2.4 Hệ số biến thiên CV(Coefficient of Variance)

Là độ phân tán trên một đơn vị trung bình, được xác định theo công thức

$$CV = \frac{s}{\bar{x}} * 100\%$$

Ý nghĩa: So sánh được độ phân tán của 2 tập dữ liệu trong 2 đơn vị khác nhau.

Trong R không có hàm chuẩn tương ứng. Trong ví dụ sau hàm tính hệ số biến thiên được xây dựng qua các hàm đã biết.

Ví dụ 28: Hàm tính hệ số biến thiên:

```
> CV <- function(x)
+ {   sd<-sd(x)
+   xn<-mean(x)
+   CoV<-sd/xn*100
+   c(CV=CoV)
+ }
> x<-c(12,34,45,56,67,89)
> CV(x)
      CV
52.90873
```

4.3 PHÂN TÍCH ĐỘ TẬP TRUNG VÀ PHÂN TÁN DỮ LIỆU KHI DỮ LIỆU CÓ PHÂN TỐ (HAY CÓ TRỌNG SỐ)

Khi dữ liệu có phân tố hoặc có trọng số, các công thức phân tích độ tập trung hay phân tán được mở rộng dựa trên các công thức trình bày sau:

a. Trung bình cộng (mean)

$$\bar{x}_w = \frac{\sum_{i=1}^n w_i x_i}{\sum_{i=1}^n w_i}$$

Trong đó: x_i : giá trị quan sát thứ i của mẫu

n : kích thước mẫu

w_i : trọng số/tần số xuất hiện của quan sát thứ i

b. Yếu vị (Mode)

Cách tính:

- Tìm tổ có tần số lớn nhất
- Tính yếu vị M_0 theo công thức:

$$M_0 = x_{M_0(\min)} + h_{M_0} \frac{f_{M_0} - f_{M_0-1}}{(f_{M_0} - f_{M_0-1}) + (f_{M_0} - f_{M_0+1})}$$

Trong đó

h_{M_0} : khoảng cách tổ có chứa M_0

x_{M_0} : giới hạn dưới của tổ có chứa M_0

f_{M_0} : tần số của tổ có chứa M_0

f_{M_0-1} : tần số của tổ đứng trước tổ có chứa M_0

f_{M_0+1} : tần số của tổ đứng sau tổ có chứa M_0

c. Trung vị (Median)

- Nếu không có khoảng cách tổ, trung vị M_0 sẽ là giá trị trung vị của tổ có tần số tích lũy

$$\frac{\sum f_i + 1}{2}$$

- Nếu có khoảng cách tổ, trung vị M_0 được tính như sau:

- Xác định tổ chứa trung vị: là tổ có tần số tích lũy bé nhất lớn hơn hoặc bằng

$$\frac{\sum f_i + 1}{2}$$

- Xác định trung vị theo công thức $M_e = x_{M_e \min} + h_{M_e} \frac{\sum \frac{f_i}{2} - S_{M_e-1}}{f_{M_e}}$

Trong đó:

$x_{M_e \min}$ là giá trị cận dưới của tổ chứa trung vị đã xác định ở bước trên

h_{M_e} khoảng cách tổ chứa trung vị

$\sum f_i$: tổng các tần số của các tổ ($=n$)

S_{M_e-1} : là tổng các tần số của tổ đứng trước tổ chứa trung vị

f_{M_e} : tần số của tổ chứa trung vị.

d. Phương sai và độ lệch chuẩn

$$\text{- Phương sai: } s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})f_i}{\sum_{i=1}^n f_i - 1}$$

Trong đó:

x_i : giá trị đại diện cho các tố, thường là giá trị trung bình các tố

\bar{x} : giá trị trung bình được tính của tập dữ liệu đã phân tố.

f_i : tần số các tố.

- Độ lệch chuẩn: $sd = \sqrt{s^2}$

4.4 CÁC ÚNG DỤNG CỦA THỐNG KÊ MÔ TẢ

4.4.1 Quan hệ thực nghiệm giữa trung bình, trung vị và yếu vị

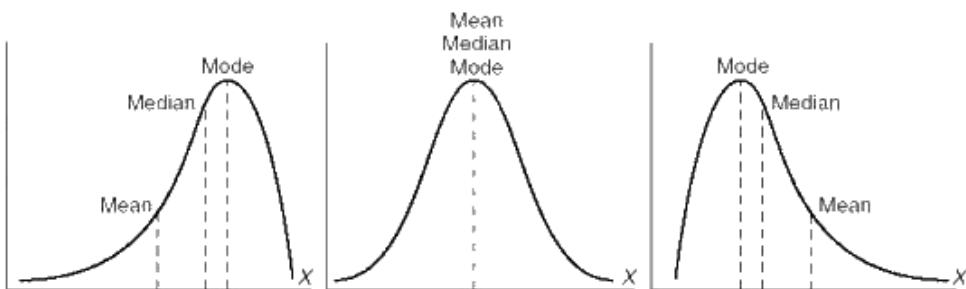
Trung bình, trung vị và yếu vị là 3 độ đo hướng tâm. Ba độ đo này, độ đo nào là thích hợp và đáng tin cậy? Câu trả lời là phụ thuộc vào sự phân bố của dữ liệu quan sát. Không độ đo nào phản ánh một cách hoàn hảo tính hướng tâm của dữ liệu. Về mặt lý thuyết thì độ đo trung bình là độ đo tốt nhất về tính hướng tâm của dữ liệu vì nó được tính toán từ dữ liệu số, sử dụng hết tất cả các quan sát và đơn nhất. Đáng lưu ý, giá trị trung bình chịu ảnh hưởng của các giá trị cực đoan, trung vị thì không chịu sự ảnh hưởng này. Tuy nhiên, trung vị là không tiêu biểu khi số lượng quan sát nhỏ, vì nó là một trung bình vị trí. Yếu vị là độ đo kém nhất thể hiện tính hướng tâm, trừ khi số lượng quan sát đủ lớn và hình ảnh phân bố dữ liệu thể hiện rõ ràng về tính hướng tâm.

Vị trí của trung vị và trung bình ảnh hưởng bởi hình dạng của phân phối dữ liệu:

a. Lệch trái
Mean < Median

b. Đối xứng
Mean = Median

c. Lệch phải
Mean > Median



Đối với tập dữ liệu đủ lớn và gần đối xứng, quan hệ giữa trung bình, trung vị và yếu vị được thể hiện bởi công thức

$$\text{Trung bình} - \text{Yếu vị} = 3(\text{Trung bình} - \text{Trung vị})$$

4.4.2 Định lý Chebychev và ước tính miền giá trị và khoảng tin cậy

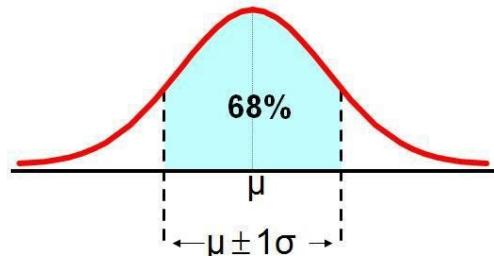
Định lý Chebychev: Với 1 quần thể bất kỳ có trung bình μ , và độ lệch chuẩn σ , k là giá trị bất kỳ lớn hơn 1. Tối thiểu $(1-1/k^2) \times 100\%$ các giá trị quan sát nằm trong khoảng $(\mu - k\sigma, \mu + k\sigma)$

k	Số phần trăm giá trị các quan sát	Thuộc khoảng
1	68%	$(\mu - \sigma, \mu + \sigma)$
2	95%	$(\mu - 2\sigma, \mu + 2\sigma)$
3	99%	$(\mu - 3\sigma, \mu + 3\sigma)$

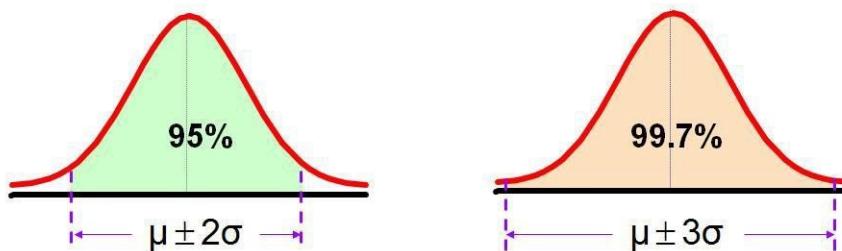
Quy tắc thực nghiệm
▪ Quy tắc

thực nghiệm (The Empirical Rule): nếu dữ liệu có phân phối có dạng hình chuông (phân phối chuẩn hoặc tiệm cận chuẩn), thì khoảng

- + $\mu \pm 1\sigma$ chứa khoảng 68% giá trị dữ liệu của mẫu hoặc quần thể.



- + $\mu \pm 2\sigma$ chứa khoảng 95% giá trị dữ liệu của mẫu hoặc quần thể.
- + $\mu \pm 3\sigma$ chứa khoảng 99.7% giá trị dữ liệu của mẫu hoặc quần thể.



4.4.3 Phép biến đổi

Xét dãy các giá trị x_1, x_2, \dots, x_n có giá trị trung bình là \bar{x} , phương sai là S_x^2 .

4.4.3.1 Phép biến đổi tuyến tính

Với mọi giá trị a, b, nếu

$$y_i = ax_i + b$$

Khi đó giá trị trung bình \bar{y} , phương sai S_y^2 ứng với dãy giá trị y_1, y_2, \dots, y_n thỏa

$$\bar{y} = a\bar{x} + b$$

$$S_y^2 = aS_x^2$$

4.4.3.2 Phép biến đổi Z:

Đối với 1 dãy các giá trị z_1, z_2, \dots, z_n , với

$$z_i = \frac{x_i - \bar{x}}{s_x^2}$$

Có giá trị trung bình $\bar{z} = 0$ và phương sai $S_z^2 = 1$

TÀI LIỆU THAM KHẢO

- [1] Introduction to Probability and Statistics Using R, G.Jay Kerns, First Edition, cran.r-project.org/web/packages/IPSUR/vignettes/IPSUR.pdf, 2010
- [2] Introductory Statistics with R, Peter Dalgaard, Second Edition, Springer, http://www.academia.dk/BiologiskAntropologi/Epidemiologi/PDF/Introductory_Statistics_with_R_2nd_ed.pdf, 2008
- [3] Thống kê ứng dụng trong kinh tế-xã hội, Hoàng Trong, Chu Nguyễn Mộng Ngọc, NXB Lao động-Xã hội, 2010.
- [4] Phân tích số liệu và biểu đồ bằng R, Nguyễn Văn Tuấn, cran.r-project.org/doc/contrib/Intro_to_R_Vietnamese.pdf
- [5] Powers, David M W (2007/2011). ["Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation"](#). *Journal of Machine Learning Technologies* 2 (1): 37–63.
- [6] POWERS, D.M.W. (February 27, 2011). ["EVALUATION: FROM PRECISION, RECALL AND F-MEASURE TO ROC, INFORMEDNESS, MARKEDNESS & CORRELATION"](#). *Journal of Machine Learning Technologies* 2 (1): 37–63.

C H U O N G

5

XÁC SUẤT

5.1 MỘT SỐ KHÁI NIỆM CƠ SỞ

- Phép thử tất định (*Deterministic experiment*): là phép thử mà kết quả có thể dự đoán chắc chắn.

Ví dụ 29: phép cộng 2+3, cho Oxy tác dụng với Hydro,..

- Phép thử ngẫu nhiên (*Random experiment*): là phép thử mà kết quả không thể dự đoán chắc chắn.

Ví dụ 30: khi tung đồng xu, kết quả không thể khẳng định là sấp hay ngửa.

- Biến cố sơ cấp (*Simple event*): là kết quả sơ đẳng một lần thực hiện phép thử, biến cố sơ cấp được thể hiện bởi một tập hợp có duy nhất 1 phần tử.

Ví dụ 31: các biến cố sơ cấp của phép thử khi tung 1 con xúc sắc là:

$$\{1\}, \{2\}, \{3\}, \{4\}, \{5\}, \{6\}$$

- Không gian mẫu (*Sample space*): Đối với một phép thử, tập hợp tất cả các biến cố sơ cấp có thể xảy của phép thử được gọi là một không gian mẫu.
- Biến cố ngẫu nhiên (*Random event*): là biến cố có thể xảy ra hoặc không xảy ra khi thực hiện phép thử. Biến cố ngẫu nhiên là tập hợp các biến cố sơ cấp có cùng 1 đặc tính.

Ví dụ 32: đối với phép thử khi tung 1 con xúc sắc, các biến cố sau là biến cố ngẫu nhiên:

- Biến cố A “xuất hiện mặt có số chấm là 5”, A={5}
- Biến cố B “xuất hiện các mặt là số chẵn”, B={2,4,6}

Chú ý: Trong ngôn ngữ R, một không gian mẫu thường được biểu diễn bởi một khung dữ liệu (*dataframe*). Mỗi một hàng của một khung dữ liệu ứng với một kết quả của phép thử. Tuy nhiên, trong một số trường hợp ứng dụng để tính xác suất của các biến cố, thì tổ chức dữ liệu bằng danh sách (*list*) cũng rất đáng quan tâm.

Để thuận tiện trong việc thực hiện các ví dụ về sau, chúng tôi sẽ sử dụng các không gian mẫu đã được tích hợp trong gói *prob* (*prob package*).

Ví dụ 33: Xem xét phép thử ngẫu nhiên tung 1 đồng xu, các biến có sơ cấp được biểu diễn là H (*head*), T (*tail*)

```
> library(prob)      > tosscoin(3)
> tosscoin(1)        toss1 toss2 toss3
toss1               1     H     H     H
1     H              2     T     H     H
2     T              3     H     T     H
                           4     T     T     H
                           5     H     H     T
                           6     T     H     T
                           7     H     T     T
                           8     T     T     T
```

5.2 BIẾN CÓ

a. Quan hệ giữa các biến có: Giả sử A, B, C, .. là các biến có ngẫu nhiên liên quan đến phép thử \varnothing nào đó, giữa các biến có có thể có các mối quan hệ sau:

- Biến có đồng nhất: biến có A, biến có B được gọi là đồng nhất, ký hiệu $A=B$, nếu với mọi kết quả có thể của phép thử chúng cùng xảy ra hay cùng không xảy ra.
- Biến có đối lập: biến có đối lập của biến có A, ký hiệu A^c hay \bar{A} là biến có “*A không xảy ra*”.
- Biến có tích: tích của 2 biến có A và B là một biến có, ký hiệu $A \cap B$ hay AB là biến có xảy ra khi cả 2 biến có A và biến có B đồng thời xảy ra. Mở rộng: tích của n biến có A_1, A_2, \dots, A_n là biến có ký hiệu $\bigcap_{i=1}^n A_i$ hay $A_1 A_2 \dots A_n$ hay $\prod_{i=1}^n A_i$ là biến có xảy ra khi các biến có A_1, A_2, \dots, A_n đồng thời xảy ra.
- Biến có xung khắc: hai biến có A, B được gọi là xung khắc khi hai biến có không đồng thời xảy ra trong phép thử. Khi đó $A \cap B = \emptyset$.
- Biến có độc lập: hai biến có A, B được gọi độc lập, nếu sự xuất hiện của biến có này không ảnh hưởng đến sự xuất hiện của biến có kia và ngược lại.
- Biến có tổng: tổng của 2 biến có A và B, ký hiệu $A+B$ hay $A \cup B$ là biến có xảy ra khi có ít nhất biến có A hay biến có B xảy ra. Mở rộng: tổng của n biến có A_1, A_2, \dots, A_n ký hiệu $\sum_{i=1}^n A_i$ hay $A_1 + A_2 + \dots + A_n$ hay $\bigcup_{i=1}^n A_i$ là biến có xảy ra khi $\exists i = 1, n$ sao cho A_i xảy ra.
- Biến có kéo theo: Nếu giả sử xảy ra biến có A kéo theo sự xảy ra biến có B thì nói A kéo theo B, ký hiệu $A \subset B$.
- Biến có chắc chắn: biến có chắc chắn xảy ra khi phép thử thực hiện, ký hiệu là Ω .
- Hệ đầy đủ các biến có: họ biến có $\{ A_1, A_2, \dots, A_n \}$ được gọi là hệ đầy đủ n nếu chúng đỏi môt xung khắc và $\sum_{i=1}^n A_i = \Omega$.

Ví dụ 34: Xét phép thử p: gieo đồng thời hai xúc xắc đều, đồng chất. Gọi A, B, C, D, E là các biến có ngẫu nhiên liên quan được xác định:

A: “Tổng các nốt xuất hiện trên hai xúc xắc là chẵn”

B: “Tổng các nốt xuất hiện trên hai xúc xắc là lẻ”

C: “Số nốt xuất hiện trên mỗi xúc xắc là lẻ”

D: “Số nốt xuất hiện trên mỗi xúc xắc là chẵn”

E: “Số nốt xuất hiện trên hai xúc xắc là cùng chẵn hoặc cùng lẻ”

Khi đó ta có các hệ thức:

$A=E, A^c=B, AB = \emptyset, A=C+D, D \subset A, \dots$

b. Các phép toán tập hợp:

1. Hàm xây dựng tập con từ một tập đã cho:

subset(<tập đã cho>, <điều kiện>)

Ví dụ 35:

```
> S<-tosscoin(3)
> S
  toss1 toss2 toss3
1     H     H     H
2     T     H     H
3     H     T     H
4     T     T     H
5     H     H     T
6     T     H     T
7     H     T     T
8     T     T     T
> subset(S,toss1=="H")
  toss1 toss2 toss3
1     H     H     H
3     H     T     H
5     H     H     T
7     H     T     T
```

2. Các hàm tìm tập con:

a. Hàm *%in%*

Cú pháp: *<vector 1> %in% <vector 2>*

Tác dụng: Tìm xem những phần tử nào của vector 1 thuộc vector 2

Ví dụ 36:

```
> x<-1:10
> y<-8:12
> y %in% x
[1] TRUE TRUE TRUE FALSE FALSE
```

b. Hàm *isin*

Cú pháp: *isin (<vector 1> ,<vector 2>)*

Tác dụng: Cho biết vector 2 có phải là vector con của vector 1 không.

Ví dụ 37:

```
> a<-2:7
> b<-1:16
> isin(a,b)
[1] FALSE
> isin(b,a)
[1] TRUE
```

Chú ý: $\text{isin}(x,y) \equiv \text{all}(y \% \in \% x)$

3. Hợp, giao và hiệu hai tập hợp:

Tên hàm	Ký hiệu	Cú pháp trong R
Hợp	$A \cup B$	<code>union(A,B)</code>
Giao	$A \cap B$	<code>intersect(A,B)</code>
Hiệu	$A \setminus B$	<code>setdiff(A,B)</code>

Ví dụ 38:

```
> A
  toss1 toss2 toss3
  1     H     H     H
  3     H     T     H
  5     H     H     T
  7     H     T     T
> B
  toss1 toss2 toss3
  3     H     T     H
  4     T     T     H
  7     H     T     T
  8     T     T     T
> C<-union(A,B)
> D<-intersect(A,B)
> E<-setdiff(A,B)
Kết quả
> C
  toss1 toss2 toss3
  1     H     H     H
  3     H     T     H
  5     H     H     T
  7     H     T     T
  8     H     T     T
> D
  toss1 toss2 toss3
  3     H     T     H
  7     H     T     T
> E
  toss1 toss2 toss3
  1     H     H     H
  5     H     H     T
```

Ví dụ 39:

```
> A
  toss1 toss2 toss3
  1     H     H     H
  3     H     T     H
  5     H     H     T
  7     H     T     T
> B
  toss1 toss2 toss3
  3     H     T     H
  4     T     T     H
  7     H     T     T
  8     T     T     T
```

Với các câu lệnh

```
> C<-union(A,B)
> D<-intersect(A,B)
> E<-setdiff(A,B)
```

Kết quả:

```
> C
  toss1 toss2 toss3
  1     H     H     H
  3     H     T     H
  4     T     T     H
  5     H     H     T
  7     H     T     T
  8     T     T     T
> D
  toss1 toss2 toss3
  3     H     T     H
  7     H     T     T
> E
  toss1 toss2 toss3
  1     H     H     H
  5     H     H     T
```

5.3 CÁC PHÉP ĐỆM

a. Hoán vị (*Permutation*): Hoán vị n phần tử là cách sắp xếp n phần tử theo một thứ tự xác định. Số lượng hoán vị n phần tử là $n!=1.2.3..n$.

Trong ngôn ngữ R, $n!$ được tính bởi hàm *factorial(n)* hay *prod(n:1)*

Ví dụ 40: tìm $3!$

```
> factorial(3)
[1] 6
```

hay

```
> prod(3:1)
[1] 6
```

Chú ý: lệnh *prod(n:m)* cho kết quả là: $n(n-1)(n-2)...m$

Ví dụ 41:

```
> prod(6:4)
[1] 120
```

b. Tổ hợp (*Combination*): Tổ hợp chập k của một tập hợp n phần tử là một tập con k phần tử của tập n phần tử đã cho. Số lượng tổ hợp chập k của n phần tử ký hiệu C_n^k hay $\binom{n}{k}$ được xác định bởi công thức:

$$C_n^k = \frac{n!}{k!(n-k)!}$$

Trong R, C_n^k được tính bởi hàm *choose(n,k)*.

Ví dụ 42: tính C_5^2

```
> choose(5,2)
[1] 10
```

5.4 XÁC SUẤT

5.4.1 Định nghĩa xác suất cỗ điện

Nếu A là biến cỗ có $n(A)$ biến cỗ sơ cấp trong một không gian biến cỗ sơ cấp gồm $n(\Omega)$ biến cỗ cùng khả năng xuất hiện, thì tỉ số: $P(A) = \frac{n(A)}{n(\Omega)}$ được gọi là xác suất của A.

Nhận xét: theo định nghĩa này xác suất của biến cỗ A có thể được xác lập mà không cần thực hiện phép thử. Điểm khó khi áp dụng công thức này là phải biết $n(\Omega)$.

5.4.2 Định nghĩa xác suất theo quan điểm thống kê

Thực hiện lặp n phép thử ngẫu nhiên và quan sát thấy biến cỗ A xuất hiện f lần. Nếu n đủ lớn thì tần suất f/n được gọi là xác suất xuất hiện của biến cỗ A:

$$P(A) = \lim_{n \rightarrow \infty} \frac{f}{n} \cong \frac{f}{n} \text{ khi } n \text{ đủ lớn}$$

- Trong ngôn ngữ R, nếu có một không gian mẫu S có cấu trúc khung dữ liệu (*dataframe*), để xác định xác suất các biến cỗ sơ cấp, bổ sung thêm tham số *makespace=TRUE* trong khi xây dựng lại không gian mẫu

Ví dụ 43:

```
> S<-tosscoin(3,makespace=TRUE)
> S
      toss1  toss2  toss3  probs
> tosscoin(3)
      toss1  toss2  toss3
 1      H      H      H
 2      T      H      H
 3      H      T      H
 4      T      T      H
 5      H      H      T
 6      T      H      T
 7      H      T      T
 8      T      T      T
```

Khi đó để tính xác suất các biến cõ ta dùng hàm $Prob(<\text{biến cõ}>)$

Ví dụ 44:

```
> S<-tosscoin(3,makespace=TRUE)
> A<-subset(S,toss1=="H"&toss2=="H")
> Prob(A)
[1] 0.25
```

5.5 NHẮC LẠI MỘT SỐ TÍNH CHẤT CỦA XÁC SUẤT

5.5.1 Hệ tiên đề Kolmogorov:

Cho không gian mẫu S , ký hiệu $P(A)$ là xác suất của biến cõ A .

Tiên đề 1: $P(A) \geq 0, \forall A \subseteq S$

Tiên đề 2: $P(S) = 1$

Tiên đề 3: Nếu các biến cõ A_1, A_2, \dots là xung khắc thì $P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i), \forall n$

Và hơn thế $P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$

5.5.2 Tính chất

Với bất kỳ 2 biến cõ A và B ,

$$B_1: P(A^c) = 1 - P(A)$$

$$B_2: P(\emptyset) = 0$$

$$B_3: \text{Nếu } A \subset B \text{ thì } P(A) \leq P(B)$$

$$B_4: 0 \leq P(A) \leq 1$$

$$B_5: P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Tổng quát, với các biến cõ A_1, A_2, \dots, A_n ,

$$P(\bigcup_{i=1}^n A_i) = \sum_{i=1}^n P(A_i) - \sum_{i=1}^{n-1} \sum_{j=i+1}^n P(A_i \cap A_j) + \dots + (-1)^{n-1} P(\bigcap_{i=1}^n A_i)$$

$$B_6: \text{hợp biến cõ } \{A_1, A_2, \dots, A_n\} \text{ hệ đầy đủ, thì } P(B) = \sum_{i=1}^n P(B \cap A_i)$$

5.6 XÁC SUẤT CÓ ĐIỀU KIỆN

5.6.1 Định nghĩa

Xác suất của biến cő A được tính với điều kiện biến cő B đã xảy ra được gọi là xác suất có điều kiện của A, ký hiệu:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

5.6.2 Qui tắc nhân

Nếu các biến cő A_i , $i = 1..n$ là độc lập thì: $P(\bigcap_{i=1}^n A_i) = \prod_{i=1}^n P(A_i)$

Nếu các biến không độc lập thì:

$$P\left(\bigcap_{i=1}^n A_i\right) = P(A_1)P(A_2|A_1)P(A_3|A_1A_2)P(A_n|A_1A_2...A_{n-1})$$

5.6.3 Qui tắc xác suất đầy đủ

Cho H_1, H_2, \dots, H_n là phân hoạch không gian mẫu M và A là biến cő bất kỳ liên quan đến phân hoạch này. Xác suất của biến cő A được tính bằng công thức xác suất đầy đủ

$$P(A) = \sum_{i=1}^n P(H_i)P(A|H_i)$$

Các xác suất $P(H_i)$ được gọi là các *xác suất tiền định (prior probability)* của A.

Các xác xuất $P(A|H_i)$ được gọi là *xác suất khả dĩ (likelihood probability)*.

Các xác xuất $P(H_i|A)$ được gọi là các *xác suất hậu định (Posterior probability)* của H_i được xác định bởi công thức

$$P(H_i|A) = \frac{P(H_i)P(A|H_i)}{P(A)}$$

Tỷ số $\frac{P(A|H_i)}{P(A)}$ được gọi là chỉ số liên quan (*irrelevance index*) dùng để đo lường sự liên quan của A và H_i . Nếu chỉ số này bằng 1, có nghĩa A và H_i không liên quan nhau.

5.6.4 Định lý Bayes

Cho H_1, H_2, \dots, H_n là phân hoạch không gian mẫu M và A là biến cő bất kỳ liên quan đến phân hoạch này. Khi đó ta có:

$$P(H_i|A) = \frac{P(H_i)P(A|H_i)}{\sum_{k=1}^n P(H_k)P(A|H_k)}$$

Tính chất: Với mọi biến cő A, B và C với $P(A)>0$

$$P(B^c|A) = 1 - P(B|A)$$

$$\text{Nếu } B \subset C \text{ thì } P(B|A) \leq P(C|A)$$

$$P([B \cup C]|A) = P(B|A) + P(C|A) - P([B \cap C]|A)$$

5.6.5 Xác suất trong ngôn ngữ R

Trong ngôn ngữ R, để tính xác suất có điều kiện $P(A|B)$ có thể sử dụng câu lệnh prob() với 2 dạng cú pháp:

- Prob(<biến có A>, given = <biến có B>)*
- Prob(<kh. gian mẫu>, <đặc tả biến có A>, given = <đặc tả biến có B>)*

Ví dụ 45:

```
> S<-tosscoin(3,makespace=TRUE)
> A<-subset(S,toss1=="H")
> B<-subset(S,toss2=="H"&toss3=="H")
> Prob(A,given=B)
[1] 0.5
> Prob(S,toss1=="H",given=(toss2=="H"&toss3=="H"))
[1] 0.5
```

TÀI LIỆU THAM KHẢO

- [1] Introduction to Probability and Statistics Using R, G.Jay Kerns, First Edition, cran.r-project.org/web/packages/IPSUR/vignettes/IPSUR.pdf, 2010
- [2] Introductory Statistics with R, Peter Dalgaard, Second Edition, Springer, http://www.academia.dk/BiologiskAntropologi/Epidemiologi/PDF/Introductory_Statistics_with_R__2nd_ed.pdf, 2008
- [3] Lý thuyết xác suất và thống kê, Đinh Văn Gắng, NXB Giáo dục, 2005.
- [4] Thống kê ứng dụng trong kinh tế-xã hội, Hoàng Trong, Chu Nguyễn Mộng Ngọc, NXB Lao động-Xã hội, 2010.
- [5] Phân tích số liệu và biểu đồ bằng R, Nguyễn Văn Tuấn, [cran.rproject.org/doc/contrib/Intro_to_R_Vietnamese.pdf](http://cran.r-project.org/doc/contrib/Intro_to_R_Vietnamese.pdf)
- [6] <http://statistics.vn/index.php/thongkecanban/>

C H U O N G

6

BIẾN NGẪU NHIÊN & PHÂN PHỐI XÁC SUẤT

6.1 BIẾN NGẪU NHIÊN (*Random variable*)

Biến ngẫu nhiên có thể được mô tả như một qui tắc biểu diễn các kết quả của phép thử ngẫu nhiên nào đó dưới dạng số.

6.1.1 Định nghĩa

Biến ngẫu nhiên X là một ánh xạ từ không gian các biến cố sơ cấp Ω vào tập số thực

$$\mathfrak{R}: \quad X: \Omega \rightarrow \mathfrak{R}$$

$$\omega \mapsto X(\omega)$$

Người ta thường ký hiệu các biến ngẫu nhiên bằng các chữ cái in hoa.

Ví dụ 46: Thực hiện phép thử gieo đồng thời 3 đồng xu cân đối, trong trường hợp này chúng ta có các biến cố sơ cấp sau

$\varpi_1=(HHH)$	$\varpi_2=(HHT)$	$\varpi_3=(HTT)$	$\varpi_4=(HTH)$
$\varpi_5=(TTT)$	$\varpi_6=(TTH)$	$\varpi_7=(THH)$	$\varpi_8=(THT)$

Nếu gọi biến ngẫu nhiên X là số đồng xu ngửa (T) xuất hiện thì X nhận các giá trị sau

$X(\varpi_1)=0$	$X(\varpi_2)=1$	$X(\varpi_3)=2$	$X(\varpi_4)=1$
$X(\varpi_5)=3$	$X(\varpi_6)=2$	$X(\varpi_7)=1$	$X(\varpi_8)=2$

6.1.2 Phân loại biến ngẫu nhiên

Biến ngẫu nhiên được phân thành 2 loại (*theo giá trị mà biến ngẫu nhiên nhận được*) là: biến ngẫu nhiên rời rạc (*discrete random variable*) và biến ngẫu nhiên liên tục (*continuous random variable*).

a. *Định nghĩa biến ngẫu nhiên rời rạc*: biến ngẫu nhiên được gọi là rời rạc nếu tập hợp các giá trị mà nó có thể nhận là một tập hữu hạn hoặc vô hạn đếm được.

Ví dụ 47: Các biến ngẫu nhiên sau là biến ngẫu nhiên rời rạc:

- Số sản phẩm kém chất lượng trong một lô hàng.
- Số con trong một gia đình.
- Số bit lỗi được truyền đi trong một kênh truyền tín hiệu số.

b. *Định nghĩa biến ngẫu nhiên liên tục*: biến ngẫu nhiên được gọi là liên tục nếu tập hợp các giá trị mà nó nhận được là một khoảng dạng (a, b) (hoặc $(a, b]$, $[a, b)$, $[a, b]$) hoặc toàn bộ \mathfrak{R} .

Ví dụ 48: Các biến ngẫu nhiên sau là biến ngẫu nhiên liên tục:

- Nhiệt độ không khí ở mỗi thời điểm nào đó.
- Thời gian hoạt động bình thường của một bóng đèn điện tử.

- Độ pH của một chất hóa học nào đó.
- Độ dài của một vật mẫu

6.2 PHÂN PHỐI XÁC SUẤT

1. Định nghĩa: Một hệ thức cho phép biểu diễn mối quan hệ giữa các giá trị có thể có của biến ngẫu nhiên với xác suất tương ứng của các giá trị gọi là quy luật phân phối xác suất của biến ngẫu nhiên.

2. Định nghĩa: Hàm phân phối xác suất (*Cumulative distribution function*) của biến ngẫu nhiên X (xác định trên không gian các biến cố sơ cấp) là hàm $F(x)$ được định nghĩa:

$$F(x) = P(X \leq x), \forall x \in (-\infty, +\infty)$$

3. Phân phối xác suất của biến ngẫu nhiên rời rạc:

- *Hàm giá trị xác suất (Probability mass function):*

Định nghĩa: Xét một biến ngẫu nhiên rời rạc X có thể nhận các giá trị x_1, x_2, \dots, x_n , một hàm giá trị xác suất (gọi tắt là hàm xác suất) là hàm thỏa:

$$\begin{aligned} f(x_i) &\geq 0, \forall i = \overline{1, n} \\ \sum_{i=1}^n f(x_i) &= 1 \\ f(x_i) &= P(X = x_i), \forall i = \overline{1, n} \end{aligned}$$

- *Bảng phân phối xác suất:*

Để mô tả biến ngẫu nhiên X nhận giá trị nào đó với xác suất tương ứng là bao nhiêu thì người ta dùng bảng phân phối xác suất. Bảng phân phối xác suất là một bảng có hai dòng:

- Dòng thứ nhất là các giá trị có thể của biến ngẫu nhiên X
- Dòng thứ hai là xác suất biến ngẫu nhiên X nhận các giá trị tương ứng

Bảng phân phối của một biến ngẫu nhiên X có dạng như sau:

X	x_1	x_2	..	x_n	..
P	$f(x_1)$	$f(x_2)$..	$f(x_n)$..

- *Hàm phân phối của biến ngẫu nhiên rời rạc*

Định nghĩa: Hàm phân phối xác suất của biến ngẫu nhiên rời rạc X, ký hiệu là $F(x)$

xác định như sau: $F(x) = P(X \leq x) = \sum_{x_i \leq x} f(x_i)$, cụ thể:

$$F(x) = P(X \leq x) = \begin{cases} 0, & x < x_1 \\ f(x_1), & x_1 \leq x < x_2 \\ f(x_1) + f(x_2), & x_2 \leq x < x_3 \\ f(x_1) + f(x_2) + f(x_3), & x_3 \leq x < x_4 \\ .. \\ f(x_1) + .. + f(x_{n-1}), & x_{n-1} \leq x < x_n \\ 1, & x \geq x_n \end{cases}$$

4. Phân phối xác suất của biến ngẫu nhiên liên tục:

- *Hàm mật độ xác suất*

Đối với biến ngẫu nhiên liên tục ngoài công cụ là hàm phân phối xác suất ta còn sử dụng hàm mật độ xác suất (*Probability density function*) của nó.

- *Định nghĩa:* Cho biến ngẫu nhiên liên tục X , hàm số $f(x)$ không âm xác định trên \mathbb{R} thỏa:

$$(1) P(X \in I) = \int_I f(x)d(x), \forall I \subset \mathbb{R}$$

$$(2) \int_{-\infty}^{+\infty} f(x)dx = 1$$

Hàm số $f(x)$ được gọi là hàm mật độ xác suất của biến ngẫu nhiên X .

6.3 CÁC ĐẶC TRƯNG SỐ CỦA BIẾN NGẪU NHIÊN

6.3.1 Kỳ vọng của biến ngẫu nhiên

1. Kỳ vọng của biến ngẫu nhiên rời rạc

Định nghĩa: Giả sử biến ngẫu nhiên rời rạc X có bảng phân phối xác suất:

X	x_1	x_2	..	x_n	..
P	$f(x_1)$	$f(x_2)$..	$f(x_n)$..

Kỳ vọng (Expectation) của X , ký hiệu $E(X)$ được định nghĩa như sau:

$$E(X) = \sum_{i=1}^{+\infty} x_i P(X=x_i) = \sum_{i=1}^{+\infty} x_i f(x_i)$$

$$E(X) = \sum_{x \in S} xf(x)$$

Với S là không gian mẫu.

2. Kỳ vọng của biến ngẫu nhiên liên tục

Định nghĩa: Giả sử biến ngẫu nhiên liên tục X có hàm mật độ xác suất $f(x)$, kỳ vọng của X là

$$E(X) = \int_{-\infty}^{+\infty} xf(x)dx$$

3. Ý nghĩa của kỳ vọng

- Là giá trị trung bình theo xác suất của tất cả giá trị có thể có của biến ngẫu nhiên.
- Kỳ vọng phản ánh giá trung bình của phân phối xác suất.

6.3.2 Phương sai của biến ngẫu nhiên

1. Định nghĩa:

Nếu biến ngẫu nhiên X có kỳ vọng $E(x)$ thì phương sai (*variance*) của X , ký hiệu $Var(X)$ được định nghĩa

$$Var(X) = E(X - E(X))^2$$

Trong thực tế, để tính phương sai của biến ngẫu nhiên X , người ta thường dùng công thức

$$Var(X) = E(X^2) - (E(X))^2$$

2. *Định nghĩa Độ lệch chuẩn* (Standard deviation): Độ lệch chuẩn của biến ngẫu nhiên X, ký hiệu $\sigma(X)$ là căn bậc hai của phương sai $Var(X)$

$$\sigma(X) = \sqrt{Var(X)}$$

3. *Ý nghĩa của phương sai:*

- Phương sai là trung bình bình phương sai lệch, nó phản ánh mức độ phân tán các giá trị của biến ngẫu nhiên xung quanh giá trị trung bình
- Trong công nghiệp phương sai biểu thị độ chính xác trong sản xuất. Trong canh tác, phương sai biểu thị mức độ ổn định của năng suất, trong đo lường phương sai thể hiện độ “ ổn định” của phép đo,..

Trong ngôn ngữ R, các đại lượng kỳ vọng, phương sai, độ lệch chuẩn có thể tính toán thủ công hoặc sử dụng gói *distrEx*.

Ví dụ 49:

- Tính toán thủ công

```
> X<-c(0,1,2,3)
> f<-c(1/8,3/8,3/8,1/8)
> mu<-sum(X*f)
> mu
[1] 1.5
> sigma2<-sum((X-mu)^2*f)
> sigma2
[1] 0.75
> sigma<-sqrt(sigma2)
> F<-cumsum(f)
> F
[1] 0.125 0.500 0.875 1.000
```

Chú ý: Hàm *cumsum()* là hàm phân phối xác suất hay còn gọi là hàm phân phối tích lũy (*Cumulative distribution function* (CDF)).

- Sử dụng gói *distrEx*

```
> library(distrEx)
> X<-DiscreteDistribution(supp=0:3,prob=c(1,3,3,1)/8)
> E(X);var(X);sd(X)
[1] 1.5
[1] 0.75
[1] 0.8660254
```

6.3.3. Một số ví dụ

1. Ứng dụng 1

Một cửa hàng buôn bán xe máy thống kê số lượng xe bán ra trong một trong khoảng thời gian 500 ngày và cho bảng thống kê:

Số xe máy bán trong 1 ngày	Tần số
0	40
1	100
2	142
3	66
4	36

5	30
6	26
7	20
8	16
9	14
10	8
11	2

- a. Gọi X là biến ngẫu nhiên chỉ số xe bán trong một ngày. Hãy lập bảng phân phối xác suất cho biến ngẫu nhiên X .
- b. Tính kỳ vọng của X tức số xe hy vọng bán được trong một ngày.
- c. Tính độ lệch chuẩn.
- d. Tính xác suất để trong 1 ngày:
- | | |
|----------------------------|-------------------------|
| 1) Có ít hơn 4 xe được bán | 4) Tối đa 4 xe được bán |
| 2) Ít nhất 4 xe được bán | 5) Đúng 4 xe được bán |
| 3) Nhiều hơn 4 xe được bán | |

Giải:

- a. Gọi X là biến ngẫu nhiên chỉ số xe bán trong 1 ngày, ta có bảng phân phối xác suất cho biến ngẫu nhiên X như sau:

X=x _i	Tần số	Xác suất
0	40	0.080
1	100	0.200
2	142	0.284
3	66	0.132
4	36	0.072
5	30	0.060
6	26	0.052
7	20	0.040
8	16	0.032
9	14	0.028
10	8	0.016
11	2	0.004
	500	1.000

- b. Kỳ vọng của X (số xe hy vọng bán được trong 1 ngày):

$$E(X) = \sum_{i=0}^{11} x_i p_i = 3.056$$

- c. Độ lệch chuẩn:

$$\text{Từ } \sigma^2 = E((X - E(X))^2) = \sum_{i=0}^{11} (x_i - E(X))^2 p_i = 6.0689$$

$$\text{Độ lệch chuẩn là: } \sigma = \sqrt{\sigma^2} = 3.056$$

- d. Tính xác suất để trong 1 ngày:

- 1) Có ít hơn 4 xe được bán:

$$P[X < 4] = P[X = 0] + P[X = 1] + P[X = 2] + P[X = 3] = 0.696$$

2) Tối đa 4 xe được bán:

$$P[X \leq 4] = P[X = 0] + P[X = 1] + P[X = 2] + P[X = 3] + P[X = 4] = 0.768$$

$$3) \text{ Ít nhất } 4 \text{ xe được bán: } P[X \geq 4] = 1 - P[X < 4] = 1 - 0.696 = 0.304$$

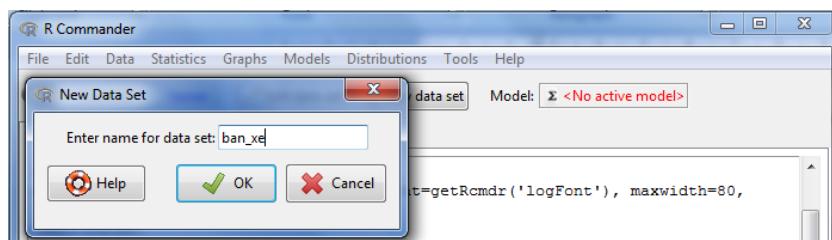
$$4) \text{ Đúng } 4 \text{ xe được bán: } P[X = 4] = 0.072$$

$$5) \text{ Nhiều hơn } 4 \text{ xe được bán: } P[X > 4] = 1 - P[\leq 4] = 1 - 0.768 = 0.232$$

Thực hiện ví dụ trên bằng ngôn ngữ R như sau:

1. Tạo dữ liệu ban đầu: Sử dụng Rcommander:

Vào Option: Data → New Data set → <Đưa vào tên data set>

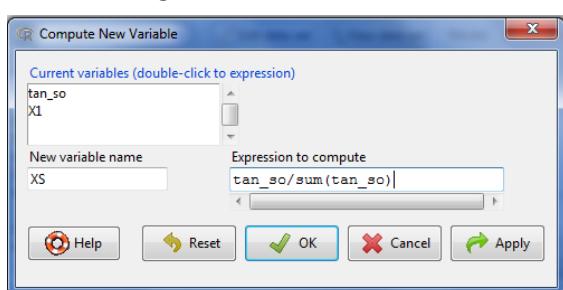


Nhập dữ liệu vào data set (có tên ban_xe)

	X1	tan_so	var3
1	0	40	
2	1	100	
3	2	142	
4	3	66	
5	4	36	
6	5	30	
7	6	26	
8	7	20	
9	8	16	
10	9	14	
11	10	8	
12	11	2	

2. Bổ sung cột xác suất để thành bảng phân phối xác suất:

Data → Manage variables in active data set → Compute new variable:



	X1	tan_so	XS
1	0	40	0.080
2	1	100	0.200
3	2	142	0.284
4	3	66	0.132
5	4	36	0.072
6	5	30	0.060
7	6	26	0.052
8	7	20	0.040
9	8	16	0.032
10	9	14	0.028
11	10	8	0.016
12	11	2	0.004

3. Sử dụng gói distrEx để tính Kỳ vọng, Phương sai, Độ lệch chuẩn:

```
> library(distrEx)
> X<-DiscreteDistribution(ban_xe$X1,ban_xe$XS)
> E(X);var(X);sd(X)
[1] 3.056
[1] 6.068864
[1] 2.463506
```

4. Tính xác suất liên quan đến số lượng bán trong 1 ngày:

- Dùng hàm `cumsum()`: để tính xác suất tích lũy
- Kết hợp với dữ liệu cột XS để tính các yêu cầu

```
> F<-cumsum(ban_xe$XS)
> F
[1] 0.080 0.280 0.564 0.696 0.768 0.828 0.880 0.920 0.952 0.980 0.996 1.000
> it_hon4<-F[4]
> toi_da4<-F[5]
> it_nhat4<-1-it_hon4
> dung_4<-ban_xe$XS[5]
> nhieuhon_4<-1-toi_da4
> it_hon4;toi_da4;it_nhat4;dung_4;nhieuhon_4
[1] 0.696
[1] 0.768
[1] 0.304
[1] 0.072
[1] 0.232
```

2. Ứng dụng 2

Ứng dụng kỳ vọng vào việc ra quyết định trong kinh doanh

(Phương pháp EMV (*Expected Monetary Value*)

Một doanh nghiệp sản xuất sản phẩm X đang cân nhắc 3 phương án xây dựng nhà máy theo qui mô nhỏ, qui mô vừa, qui mô lớn. Doanh nghiệp có đánh giá về mức lợi nhuận theo 3 phương án ứng với 3 tình huống kinh tế tăng trưởng như sau:

Tình huống nền kinh tế	Lợi nhuận		
	Qui mô lớn	Qui mô vừa	Qui mô nhỏ
Kinh tế mạnh	200	90	40
Kinh tế ổn định	50	110	30
Kinh tế suy yếu	-110	-30	20

Các chuyên gia có dự đoán khả năng (xác suất) các tình huống kinh tế như sau:

Kinh tế mạnh	0.3	Kinh tế ổn định	0.5	Kinh tế suy yếu	0.2
--------------	-----	-----------------	-----	-----------------	-----

Hãy đưa ra quyết định phương án xây dựng nhà máy tối ưu?

Giải:

- Phương pháp EMV chọn phương án nào có kỳ vọng lớn nhất.
- Thực hiện bằng Rcmdr như sau:
- Tạo bảng dữ liệu: Data→ New Data set→ <Đưa vào tên data set>
- Nhập dữ liệu: vào data set: Phuong_an

	Qmlon	Qmovua	Qmonho	XS
1	200	90	40	0.3
2	50	110	30	0.5
3	-100	-30	20	0.2
4				

Sử dụng gói distrEx để tính Kỳ vọng

```
> library(distrEx)
> PaQmlon<-DiscreteDistribution(Phuong_an$Qmlon, Phuong_an$XS)
> PaQmvua<-DiscreteDistribution(Phuong_an$Qmvua, Phuong_an$XS)
> PaQmnho<-DiscreteDistribution(Phuong_an$Qmnho, Phuong_an$XS)
> E(PaQmlon);E(PaQmvua);E(PaQmnho)
[1] 65
[1] 76
[1] 31
```

Phương án tối ưu để xây dựng nhà máy tối ưu là: Phương án xây dựng nhà máy có qui mô vừa.

6.4 PHÂN PHỐI XÁC SUẤT RỜI RẠC

6.4.1 Phân phối Bernoulli (*Bernoulli Distribution*)

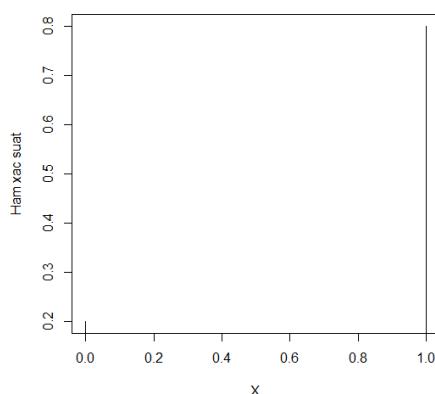
1 *Định nghĩa*: Thực hiện 1 phép thử, ta quan tâm đến biến cõ A. Nếu biến cõ A thành công thì X mang các giá trị là 1, ngược lại X mang giá trị là 0. Phép thử này gọi là phép thử Bernoulli. Giả sử biến cõ A có xác suất θ :

$$P(X=1)=\theta \text{ và } P(X=0)=1-\theta, \text{ với } 0 \leq \theta \leq 1.$$

Khi đó biến ngẫu nhiên X được gọi là biến ngẫu nhiên có phân phối Bernoulli với tham số θ , ký hiệu $X \sim \text{Bernoulli}(\theta)$.

Ví dụ 50: Biến ngẫu nhiên X biểu diễn cho tình trạng sống sót sau 5 năm của bệnh nhân ung thư vú, $X=1$ ứng với bệnh nhân sống sót, $X=0$ ứng với bệnh nhân tử vong. Giả sử xác suất $P(X=0)=0.2$, $P(X=1)=1-P(X=0)=0.8$ thì đồ thị biểu diễn $X \sim \text{Bernoulli}(0.8)$ được vẽ như sau:

```
> x<-c(0,1)
> y<-c(0.2,0.8)
> plot(x,y,xlab="X",ylab="Ham xac suat",type="h")
```



Nhận xét: Đối với phân phối Bernoulli: $X \sim \text{Bernoulli}(\theta)$, dễ dàng ta tính được:

$$E(X)=\theta; \text{Var}(X)=\theta(1-\theta)$$

6.4.2 Phân phối nhị thức (Binomial Distribution)

I Định nghĩa: Thực hiện n phép thử Bernoulli độc lập với xác suất thành công của phép thử là p. Gọi X là số lần thành công trong n phép thử thì

$$X = X_1 + X_2 + \dots + X_n$$

Với X_i ($i=1, 2, \dots, n$) là biến ngẫu nhiên có phân phối Bernoulli cùng với tham số p. Khi đó X là biến ngẫu nhiên rời rạc với miền giá trị $S = \{0, 1, \dots, n\}$ và xác suất

$$P(X = k) = C_n^k p^k q^{n-k}, \quad k \in S, \quad q = 1 - p$$

X được gọi là có phân phối nhị thức với tham số n, p ký hiệu $X \sim B(n; p)$.

Ví dụ 51: Xét 500 bệnh nhân ung thư vú, xác suất 1 bệnh nhân sống sót sau 5 năm là 0.8. Ta có $X \sim B(500; 0.8)$

Ý nghĩa: Nếu 1 phép thử được tiến hành gồm n lần, mỗi lần cho kết quả hoặc thành công (1) với xác suất p hoặc thất bại (0) với xác suất $q = 1 - p$, thì xác suất để có k lần tiến hành thành công là

$$P(X = k) = C_n^k p^k (1 - p)^{n-k}$$

Nhận xét: Điều kiện của một biến ngẫu nhiên có phân phối nhị thức:

- Số lần thử nghiệm cố định (giả sử n)
- Kết quả của mỗi thí nghiệm chỉ có 2 trạng thái thành công hay thất bại.
- Xác suất thành công là như nhau cho mỗi lần thí nghiệm (p).
- Các lần thí nghiệm là độc lập nhau (kết quả các thí nghiệm không ảnh hưởng lẫn nhau)

Các hàm trong ngôn ngữ R liên quan:

Loại hàm	Cú pháp	Tác dụng	Chú thích
PMF	dbinom(k,n,p)	$P(X = k) = C_n^k p^k (1 - p)^{n-k}$	$d \equiv distribution$
CDF	pbinom(k,n,p)	$P(X \leq k)$	$p \equiv probability$
Quantile	qbinom(p,n,prob)	$k_{min}?$ khi $F(k) = P(X \leq k) \geq prob$	$q \equiv quantile$
Simulation	rbinom(n,k,p)	Tạo ra 1 phép thử gồm n mẫu (lần), phép thử có k mẫu thành công, xác suất thành công là p	$r \equiv random$

Ví dụ 52: Trong 1 nhà máy sản xuất chip nhớ, biết rằng những cuộc thử nghiệm kiểm tra trước đó cho thấy tỉ lệ sản phẩm bị hỏng là 20%. Kiểm tra ngẫu nhiên 15 chíp. Tính xác suất:

- Có đúng 7 chíp không đạt chất lượng
- Có ít nhất 1 chíp không đạt chất lượng
- Vẽ biểu đồ của hàm xác suất (PMF) tương ứng
- Vẽ biểu đồ của hàm phân phối tích lũy (CDF) tương ứng

Giải:

Xác suất 1 chíp nhớ không đạt chất lượng là $p=0.20$.

a. Xác xuất có đúng 7 chíp không đạt chất lượng là:

$$P(X = 7) = C_{15}^7 p^7 (1-p)^{15-7} = dbinom(7,15,0.20)$$

b. Xác suất có nhất 1 chíp không đạt chất lượng:

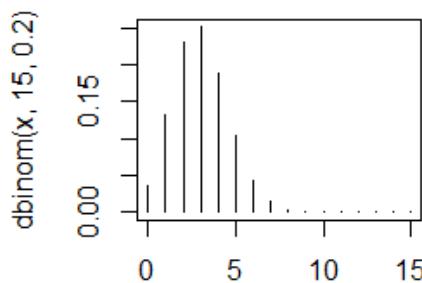
$$P(X \geq 1) = 1 - P(X < 1) = 1 - P(X = 0) = 1 - dbinom(0,15,0.20)$$

Thực hiện bằng ngôn ngữ R:

```
> dbinom(7,15,0.20)
[1] 0.01381906
> 1-dbinom(0,15,0.20)
[1] 0.9648156
```

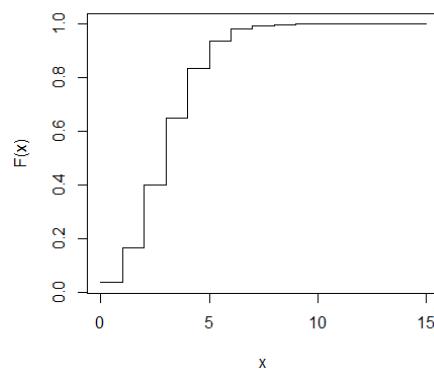
c. Biểu đồ của hàm giá trị xác suất (PMF) tương ứng:

```
> x<-0:15
> plot(x,dbinom(x,15,0.20), type="h")
```



d. Biểu đồ hàm phân phối xác suất (CDF):

```
> x<-c(0,1,2,3,4,5,6,7,8,9,10,11,12,13,14,15)
> prob<-dbinom(x,15,0.20)
> cdf<-c(0,cumsum(prob))
> cdf.plot<-stepfun(x,cdf,f=0)
> plot.stepfun(cdf.plot,xlab="x",ylab="F(x)",
+ verticals=FALSE,do.points=TRUE,main="",pch=16)
```

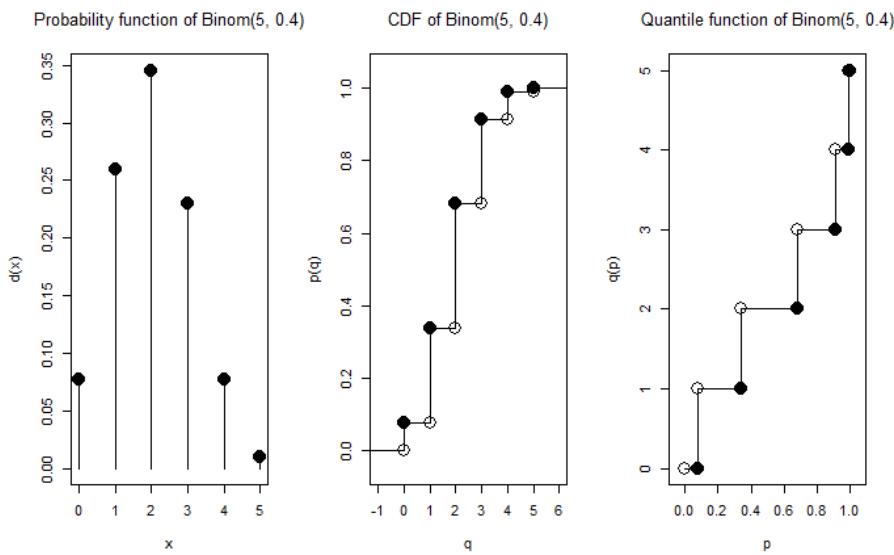


Chú ý: Biểu đồ CDF cũng có thể vẽ

```
x<-c(0,1,2,3,4,5,6,7,8,9,10,11,12,13,14,15)
plot(x,pbinom(x,15,0.20),xlab="x",ylab="F(x)",type="s")
```

- Gói distr cho phép vẽ biểu đồ khá nhanh

```
> library(distr)
> X<-Binom(5, 0.4)
> plot(X)
```



- Trong Rcmdr có chức năng vẽ các biểu đồ thông qua các mục chọn (*người đọc tự tìm hiểu*)

6.4.3 Phân phối Poisson (Poisson Distribution)

1 Định nghĩa: Phân phối Poisson là một phân phối xác suất rời rạc ứng với số sự kiện xảy ra trong một khoảng thời gian hay không gian nhất định.

- Đặt $X =$ số sự kiện xảy ra trong một khoảng (thời gian hay không gian)
- Số lượng trung bình các sự kiện xuất hiện trong mỗi khoảng (thời gian hay không gian) là λ .
- Xác suất của quan sát x sự kiện xảy ra trong một khoảng đã cho xác định bởi công thức:

$$P(X = x) = \frac{e^{-\lambda} \lambda^x}{x!}, \text{ với } x=0,1,2,3,..$$

e: hằng số Nepe, $e = 2,718282$.

Biến ngẫu nhiên rời rạc X được định nghĩa như trên gọi là biến ngẫu nhiên có phân phối Poisson với tham số λ , ký hiệu $X \sim P(\lambda)$.

- Kỳ vọng và phương sai của X lần lượt bằng:

$$E(X) = \lambda$$

$$Var(X) = \lambda$$

Chú ý:

- Một biến ngẫu nhiên Poisson có thể nhận một số nguyên không âm bất kỳ. Ngược lại, một biến ngẫu nhiên phân phối nhị thức luôn có một giới hạn trên hữu hạn.

- Khi xử lý phân phối Poisson có $\lambda \geq 20$, người ta chứng minh được có thể xấp xỉ bằng phân phối chuẩn $X \sim \mathcal{N}(\lambda, \lambda)$.

Ví dụ 53:

1. Người ta thống kê tại một tổng đài điện thoại, trung bình có 10 cuộc gọi đến trong 1 giờ. Xác suất để có 5 cuộc điện thoại gọi đến trong 1 giờ là: $P(X = 5) = \frac{e^{-10} 10^5}{5!}$.
2. Trung bình số ca sinh tại một bệnh viện phụ sản trong 1 giờ là 1.8 ca.
 - a) Xác suất để có 4 ca sinh trong 1 giờ là:

$$P(X = 4) = \frac{e^{-1.8} (1.8)^4}{4!} = 0.0723.$$

- b) Xác suất để có ít nhất 7 ca sinh trong 2 giờ là:

Gọi X là số ca sinh trong 2 giờ. X theo phân phối Poisson với giá trị trung bình $\lambda = 1.8 \times 2 = 3.6$

Xác suất để có ít nhất 7 ca sinh trong 2 giờ là:

$$P(X \geq 7) = 1 - P(X < 7) = 1 - \sum_{k=0}^6 P(X = k) = 1 - \sum_{k=0}^6 \frac{e^{-3.6} (3.6)^k}{k!}$$

Nhận xét: Một biến ngẫu nhiên X đếm số lần sự kiện xảy ra trong một đơn vị thời gian hoặc không gian có phân phối Poisson nếu:

- Mỗi một sự kiện xảy ra độc lập và ngẫu nhiên.
- Các sự kiện xảy ra với một tỉ lệ không đổi (theo nghĩa là số lượng các sự kiện xảy ra trong một khoảng thời gian nhất định là tỷ lệ thuận với độ dài của khoảng thời gian đó).
- Các sự kiện xảy ra riêng lẻ (mỗi thời điểm chỉ có 1 sự kiện xảy ra)

2. Định lý Poisson

Cho $X \sim B(n; p)$, khi $n \rightarrow \infty, p \rightarrow 0, np = \lambda$ (const) thì

$$\lim_{n \rightarrow \infty} P(X = k) = \frac{e^{-\lambda} \lambda^k}{k!}$$

Trong đó: $P(X = k) = C_n^k p^k (1-p)^{n-k}$

- Ý nghĩa định lý Poisson: Khi n khá lớn ($n \geq 1000$), và p khá nhỏ ($np \leq 10$) thì phân phối Poisson xấp xỉ với phân phối nhị phân.

Ví dụ 54: Trong một đợt tiêm chủng cho trẻ em ở một khu vực, biết xác suất một trẻ phản ứng với thuốc sau khi tiêm là 0.001. Thực hiện tiêm cho 2000 trẻ, tính xác suất có nhiều nhất 1 trẻ bị phản ứng với thuốc.

Giải: Gọi X là số trẻ em phản ứng thuốc sau khi tiêm. X là biến ngẫu nhiên có phân phối Poisson với tham số $\lambda = n.p = 2$. Vì vậy, xác suất có nhiều nhất có 1 trẻ phản ứng với thuốc là:

$$P(X \leq 1) = P(X = 0) + P(X = 1) = \sum_{k=0}^1 \frac{e^{-\lambda} \lambda^k}{k!}$$

Các hàm trong ngôn ngữ R liên quan:

Loại hàm	Cú pháp	Tác dụng	Chú thích
PMF	dpois(k,λ)	$P(X = k) = \frac{e^{-\lambda} \lambda^k}{k!}$	$d \equiv distribution$
CDF	ppois(k,λ)	$P(X \leq k)$	$p \equiv probability$
Quantile	qpois(prob,λ)	$k_{min}?$ khi $F(k) = P(X \leq k) \geq prob$	$q \equiv quantile$
Simulation	rpois(n,λ)	Tạo ra 1 phép thử gồm n mẫu (lần)	$r \equiv random$

Ví dụ 55: Kết quả của các ví dụ trên là:

5.3.a)

```
> dpois(5,10)
[1] 0.03783327
```

5.3.b)

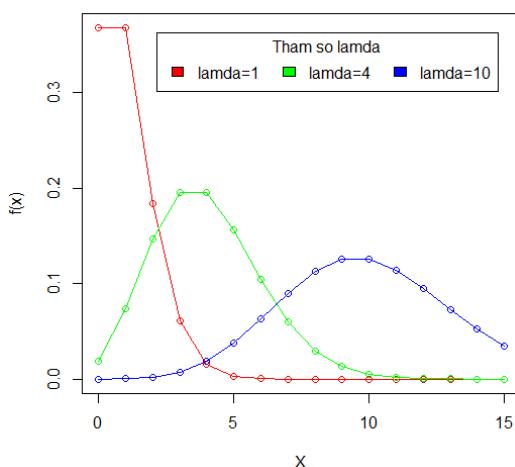
```
> 1-ppois(7,3.6)
[1] 0.03078928
```

54)

```
> ppois(1,0.001)
[1] 0.9999995
```

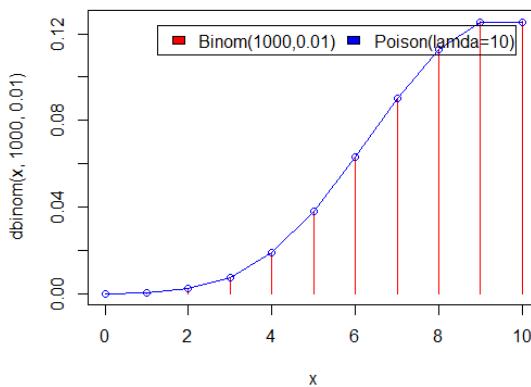
Ví dụ 56: Vẽ biểu đồ hàm giá trị xác suất của phân phối Poisson:

```
> x<-0:15
> y1<-dpois(x,1)
> y2<-dpois(x,4)
> y3<-dpois(x,10)
> plot(x,y1,xlab="X",ylab="f(x)",col="red",type="o",ylim=range(c(y1,y2,y3)))
> lines(x,y2,col="green",type="o")
> lines(x,y3,col="blue",type="o")
> legend("topright",inset=.05,title="Tham số lamda",
+ c("lamda=1","lamda=4","lamda=10"),fill=c("red","green","blue"),horiz=TRUE)
```



Ví dụ 57: Phân phối nhị thức xấp xỉ bằng phân phối Poisson

```
> x<-0:10
> plot(x,dbinom(x,1000,0.01),type="h",col="red")
> lines(x,dpois(x,10),type="o",col="blue")
> legend("topright",inset=.05,c("Binom(1000,0.01)",
+ "Poisson(lamda=10)"),fill=c("red","blue"),horiz=TRUE)
```



6.5 PHÂN PHỐI XÁC SUẤT LIÊN TỤC

6.5.1 Phân phối đều (*Uniform distribution*)

1 *Định nghĩa:* Biến ngẫu nhiên X được gọi là có phân phối đều trên đoạn $[a,b]$, ký hiệu $X \sim \text{unif}([a,b])$, nếu hàm mật độ xác suất của X có dạng

$$f(x) = \begin{cases} \frac{1}{b-a}, & x \in [a,b] \\ 0, & x \notin [a,b] \end{cases}$$

- Hàm phân phối xác xuất của $X \sim \text{unif}([a,b])$

$$F(x) = \begin{cases} 0, & x < a \\ \frac{x-a}{b-a}, & x \in [a,b] \\ 1, & x > b \end{cases}$$

Các hàm trong ngôn ngữ R liên quan:

Loại hàm	Cú pháp	Tác dụng	Chú thích
PMF	dunif(x,a,b)	$P[X=x]$	$d \equiv \text{distribution}$
CDF	punif(k,a,b)	$P(X \leq k)$	$p \equiv \text{probability}$
Quantile	qunif(prob,a,b)	$k_{\min}?$ khi $F(k) = P(X \leq k) \geq prob$	$q \equiv \text{quantile}$
Simulation	runif(n,a,b)	Tạo ra 1 phép thử gồm n mẫu (lần)	$r \equiv \text{random}$

- *Ý nghĩa:* Trong thống kê, nếu ta không biết gì về tham số cần ước lượng thì mỗi giá trị có thể có của tham số đó là đồng khả năng. Điều đó dẫn đến việc quan niệm tham số cần ước lượng như một biến ngẫu nhiên tuân theo qui luật phân phối đều.

- *Chú ý:* Khi tính hàm mật độ xác suất của phân phối đều thì:

$$P(a \leq X \leq b) = P(a < X \leq b) = P(a \leq X < b) = P(a < X < b)$$

(do đổi với đường $x=c$ (const), thì $P(X=c)=0$)

Ví dụ 58: Khi thâm nhập vào một thị trường mới, doanh nghiệp không thể khẳng định được một cách chắc chắn doanh số hàng tháng có thể đạt được sẽ là bao nhiêu mà chỉ dự kiến được doanh số tối thiểu sẽ là 20 triệu đồng/tháng và tối đa sẽ là 40 triệu đồng/tháng. Tìm xác suất để doanh nghiệp đạt được doanh số tối thiểu là 35 triệu đồng/tháng.

Giải:

Gọi X là doanh số hàng tháng mà doanh nghiệp có thể đạt được ở thị trường đó. Do chưa có thông tin gì hơn, nên có thể xem X là biến ngẫu nhiên liên tục phân phối đều trên khoảng (20,40).

Vậy X có hàm mật độ xác suất:

$$f(x) = \begin{cases} \frac{1}{40-20} = 0.05, & x \in [20,40] \\ 0, & x \notin [20,40] \end{cases}$$

Xác suất để doanh nghiệp đạt được doanh số tối thiểu là 35 triệu/tháng (tính theo hàm mật độ xác suất)

$$P(X \geq 35) = \int_{35}^{+\infty} f(x) dx = \int_{35}^{+\infty} 0.05 dx = \int_{35}^{40} 0.05 dx = 0.25$$

Thực hiện bằng ngôn ngữ R như sau:

```
> #P(X>=35)
> 1-punif(35,20,40)           > punif(35,20,40,lower.tail= FALSE)
[1] 0.25                      hoặc [1] 0.25
```

- Kỳ vọng và phương sai của phân phối đều:

Nếu X là biến ngẫu nhiên liên tục có phân phối đều: $X \sim unif([a,b])$

- Kỳ vọng của X: $E(X) = \frac{a+b}{2}$
- Phương sai của X: $Var(X) = \frac{(b-a)^2}{12}$

6.5.2 Phân phối mũ (Exponential distribution)

1. Định nghĩa: Biến ngẫu nhiên $X (X > 0)$ gọi là có phân phối mũ, ký hiệu $X \sim Exp(\lambda)$, nếu nó có hàm mật độ xác suất

$$f(t) = \lambda e^{-\lambda t}, t > 0$$

Trong đó λ : số biến cố trung bình xảy ra trong một đơn vị thời gian.

t : số đơn vị thời gian cho đến biến cố kế tiếp.

2. Các đặc trưng của phân phối mũ

Nếu $X \sim Exp(\lambda)$ thì kỳ vọng và phương sai của X lần lượt là

$$E(X) = \frac{1}{\lambda}; Var(X) = \frac{1}{\lambda^2}$$

Các hàm trong ngôn ngữ R liên quan:

Loại hàm	Cú pháp	Tác dụng	Chú thích
PMF	dexp(x,λ)	f(x)	$d \equiv distribution$
CDF	pexp(k,λ)	$P(X \leq k)$	$p \equiv probability$
Quantile	qexp(prob,λ)	$k_{min}?$ khi $F(k) = P(X \leq k) \geq prob$	$q \equiv quantile$
Simulation	rexp(n,a,b)	Tạo ra 1 phép thử gồm n mẫu (lần)	$r \equiv random$

Ví dụ 59: Trong 1 mạng máy tính ở 1 công ty, biết rằng số người dùng đăng nhập vào mạng trong 1 giờ có phân phối Poisson với trung bình bằng 25.

- Tính xác suất không có người dùng nào đăng nhập trong khoảng thời gian 6 phút.
- Tính xác suất lần đăng nhập kế tiếp cách lần đăng nhập đầu 2 đến 3 phút.

Giải:

- Gọi X là số người đăng nhập vào mạng, theo giả thiết X có phân phối Poisson với trung bình trong 6 phút là: $\lambda = 25/10 = 2.5$.

Xác suất không có người đăng nhập vào mạng trong 6 phút là:

$$P(X = 0) = \frac{e^{-2.5} (2.5)^0}{0!} = \frac{1}{e^{2.5}} = 0.82$$

```
> ppois(0, 2.5)
[1] 0.082085
```

- Gọi T là thời gian giữa 2 lần đăng nhập vào mạng, T có phân phối mũ với hàm mật độ xác suất:

$$f(t) = \lambda_1 e^{-\lambda_1 t}, \lambda_1 = \frac{25}{60} = 0.42,$$

// chú ý: λ_1 là trung bình số người đăng nhập vào mạng trong 1 phút.

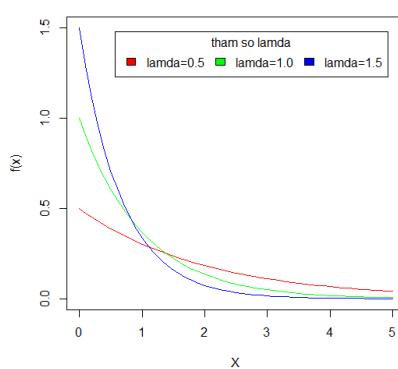
Xác suất lần đăng nhập kế tiếp cách lần đăng nhập đầu 2 đến 3 phút:

$$P(2 \leq X \leq 3) = P(X \leq 3) - P(X \leq 2) = e^{-2*0.42} - e^{-3*0.42} = 0.418$$

```
> pexp(3, 0.42)-pexp(2, 0.42)
[1] 0.1480565
```

Ví dụ 60: Biểu đồ Hàm mật độ phân phối mũ

```
> x<-seq(0, 5, 0.1)
> y1<-dexp(x, 0.5)
> y2<-dexp(x, 1)
> y3<-dexp(x, 1.5)
> plot(x, y1, col="red", ylim=range(c(y1, y2, y3)), type="l", xlab="X", ylab="f(x)")
> lines(x, y2, col="green")
> lines(x, y3, col="blue")
> legend("topright", inset=0.05, title="tham so lamda",
+ c("lamda=0.5", "lamda=1.0", "lamda=1.5"), fill=c("red", "green", "blue"),
+ horiz=TRUE)
```



6.5.3 Phân phối chuẩn (Normal distribution)

1. Định nghĩa: Biến ngẫu nhiên liên tục X nhận giá trị trong khoảng $(-\infty, +\infty)$ được gọi là có phân phối chuẩn tham số μ, σ nếu hàm mật độ xác suất có dạng

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Trong đó μ , σ là hằng số và $\sigma > 0$, $-\infty < \mu < +\infty$, ký hiệu $X \sim \mathcal{N}(\mu; \sigma^2)$.

- Nếu $X \sim \mathcal{N}(\mu; \sigma^2)$ thì kỳ vọng và phương sai của X lần lượt là

$$E(X) = \mu, \text{Var}(X) = \sigma^2$$

- Hàm phân phối xác xuất của $X \sim \mathcal{N}(\mu; \sigma^2)$:

$$F(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^x e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt$$

- Nếu $X \sim \mathcal{N}(0; 1)$, thì ta nói X có phân phối chuẩn tắc (*standardized normal distribution*). Khi đó hàm phân phối và hàm mật độ có dạng:

$$f_0(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \text{ và } F_0(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{t^2}{2}} dt$$

Với hàm Laplace

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-\frac{t^2}{2}} dt$$

- Nhận xét:

- Theo định lý về tính tuyến tính của phân phối chuẩn, nếu $X \sim \mathcal{N}(\mu; \sigma^2)$ thì $Z = \frac{X - \mu}{\sigma}$ có phân phối chuẩn hóa: $Z \sim \mathcal{N}(0; 1)$, dựa vào tính chất này ta có thể tính xác suất của biến ngẫu nhiên $X \sim \mathcal{N}(\mu; \sigma^2)$:

$$P(X \leq b) = P\left(\frac{X - \mu}{\sigma} \leq \frac{b - \mu}{\sigma}\right) = \varphi\left(\frac{b - \mu}{\sigma}\right) \Rightarrow P(a < Z < b) = \varphi(b) - \varphi(a)$$

$$P(|X - \mu| < \varepsilon) = 2\varphi\left(\frac{\varepsilon}{\sigma}\right)$$

- Người ta đưa phân phối chuẩn về phân phối chuẩn tắc nhằm chuẩn hóa các biến ngẫu nhiên để chúng độc lập với đơn vị đo lường.
- Hàm $\varphi(x)$ là một hàm số lẻ.
- Đồ thị hàm mật độ của $X \sim \mathcal{N}(\mu; \sigma^2)$ có dạng hình chuông đối xứng qua đường $x = \mu$.

Các hàm trong ngôn ngữ R liên quan:

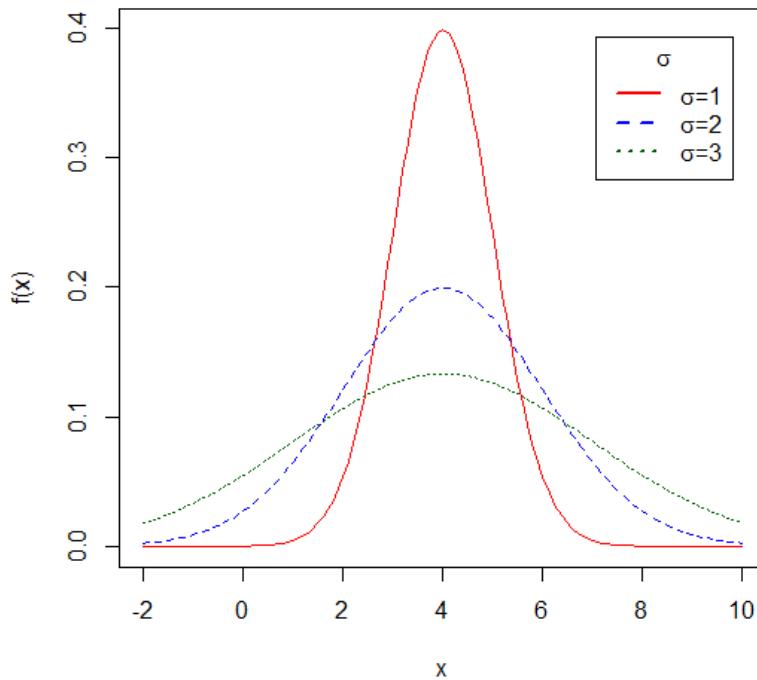
Loại hàm	Cú pháp	Tác dụng	Chú thích
PMF	dnorm(x, mu, sigma)	$f(x)$	$d \equiv \text{distribution}$
CDF	pnorm(x, mu, sigma)	$P(X \leq k)$	$p \equiv \text{probability}$
Quantile	qnorm(prob, mu, sigma)	$k_{\min}?$ khi $F(k) = P(X \leq k) \geq prob$	$q \equiv \text{quantile}$
Simulation	rnorm(n, mu, sigma)	Tạo ra 1 phép thử gồm n mẫu (lần)	$r \equiv \text{random}$

Đồ thị hàm mật độ của $X \sim \mathcal{N}(\mu; \sigma^2)$

Ví dụ 61:

```
> x<-seq(-2,10,0.1)
> y1<-dnorm(x,4,1)
> y2<-dnorm(x,4,2)
> y3<-dnorm(x,4,3)
> colors=c("red","blue","darkgreen")
> labels<-c(expression(paste(sigma,"=1")),
+ expression(paste(sigma,"=2")),
+ expression(paste(sigma,"=3")))
> plot(x,y1,xlab="x",ylab="f(x)",
+ main=expression(paste("N(",mu,",",",sigma^2," ",mu,"=4")),
+ col="red",type="l",lty=1)
> lines(x,y2,type="l",lty=2,col="blue")
> lines(x,y3,type="l",lty=3,col="darkgreen")
> legend("topright",inset=.05,title=expression(sigma),
+ labels,lwd=2,lty=c(1,2,3),col=colors)
```

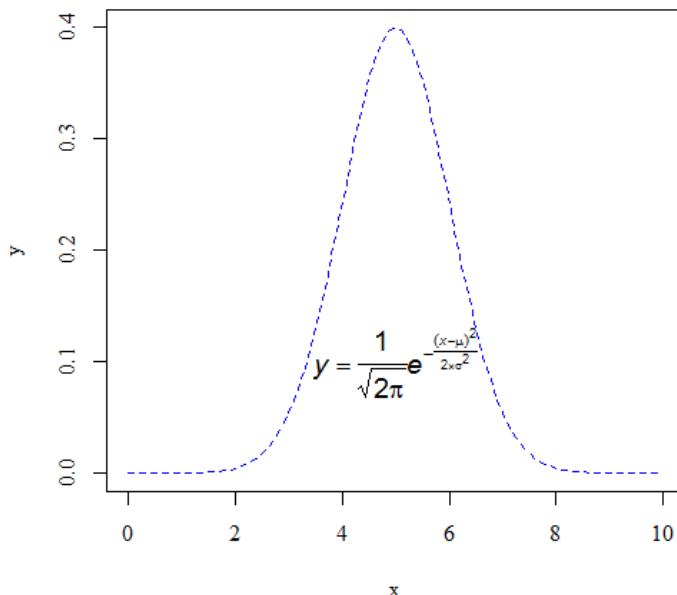
$$\mathcal{N}(\mu, \sigma^2) \mu=4$$



Ví dụ 62:

```
> x<-seq(0,10,0.1)
> y<-dnorm(x,5,1)
> plot(x,y,type="l",lty=2,col="blue",
+ main="Phân phối chuẩn",family="A")
> text(x=5, y=0.1,cex=1.2,
+ expression(italic(y == frac(1, sqrt(2 * pi)) *
+ e ^ {-frac((x-mu)^2, 2*sigma^2)} )))
```

Phân phối chuẩn



2. Qui tắc k_σ

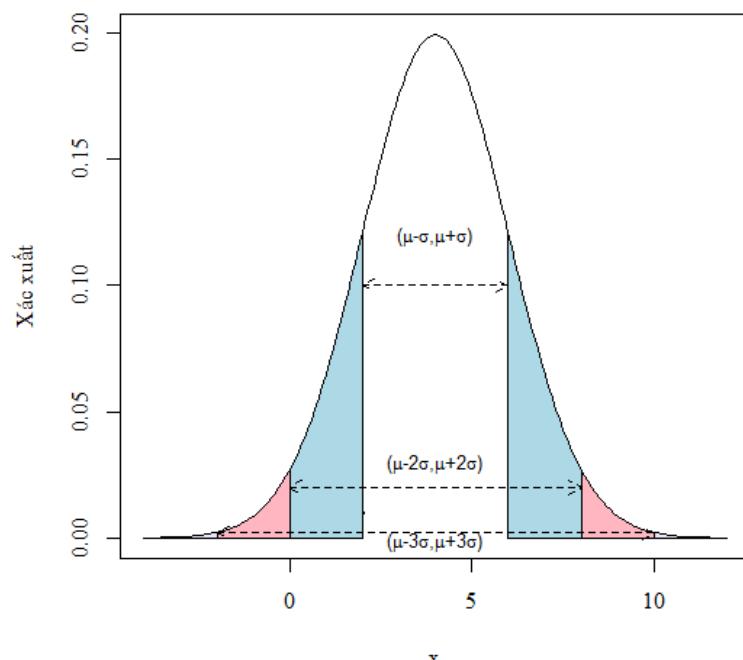
Xét $X \sim \mathcal{N}(\mu; \sigma^2)$, $\varepsilon > 0$ ta có $P(|X - \mu| < \varepsilon) = 2\varphi\left(\frac{\varepsilon}{\sigma}\right)$

- Với $\varepsilon = \sigma$, ta có $P(|X - \mu| < \sigma) = 2\varphi(1) = 2 * 0.3413 = 0.6826$. Vậy có khoảng 68.26% giá trị của X nằm trong khoảng $(\mu - \sigma, \mu + \sigma)$
- Với $\varepsilon = 2\sigma$, ta có $P(|X - \mu| < 2\sigma) = 2\varphi(2) = 2 * 0.4772 = 0.9544$. Vậy có khoảng 95.44% giá trị của X nằm trong khoảng $(\mu - 2\sigma, \mu + 2\sigma)$
- Với $\varepsilon = 3\sigma$, ta có $P(|X - \mu| < 3\sigma) = 2\varphi(3) = 2 * 0.4987 = 0.9974$. Vậy có khoảng 99.74% giá trị của X nằm trong khoảng $(\mu - 3\sigma, \mu + 3\sigma)$

Ứng dụng và ý nghĩa qui tắc k_σ :

- Nếu qui luật phân phối của một biến ngẫu nhiên được nghiên cứu chưa biết, song nó thỏa điều kiện của qui tắc $2_\sigma, 3_\sigma$ thì xem như biến ngẫu nhiên đó có phân phối chuẩn.
- Nếu biến ngẫu nhiên có phân phối chuẩn thì 95.44% các giá trị của nó nằm trong khoảng $(\mu - 2\sigma, \mu + 2\sigma)$, và hầu như các giá trị của nó nằm trong khoảng $(\mu - 3\sigma, \mu + 3\sigma)$.

Phân phối chuẩn & QT k_sigma



```

> # Thiết lập Font Việt
> windowsFonts(
+ A=windowsFont("Times New Roman"),
+ B=windowsFont("Bookman Old Style"),
+ C=windowsFont("Comic Sans MS"),
+ D=windowsFont("Symbol")
+ )
> x<-seq(-4,12,0.1)
> y<-dnorm(x,4,2)
> gh<-data.frame(z=x,gh=y)
> plot(gh, type="n",
+ ylab="Xác xuất",
+ main="Phân phối chuẩn & QT k_sigma",family="A")
> #Khoảng (mu-sigma,mu+sigma)
> z1<-6
> z2<-2
> #Khoảng (mu-2sigma,mu+2sigma)
> z3<-8
> z4<-0
> #Khoảng (mu-3sigma,mu+3sigma)
> z5<-10
> z6<--2

```

```

> # Vẽ các khoảng
> t1<-subset(gh,z>=z1)
> polygon(c(rev(t1$z),t1$z),
+ c(rep(0,nrow(t1)),t1$gh),col="lightblue")
> t2<-subset(gh,z<=z2)
> polygon(c(rev(t2$z),t2$z),
+ c(rep(0,nrow(t2)),t2$gh),col="lightblue")
> t3<-subset(gh,z>=z3)
> polygon(c(rev(t3$z),t3$z),
+ c(rep(0,nrow(t3)),t3$gh),col="lightpink")
> t4<-subset(gh,z<=z4)
> polygon(c(rev(t4$z),t4$z),
+ c(rep(0,nrow(t4)),t4$gh),col="lightpink")
> t5<-subset(gh,z>=z5)
> polygon(c(rev(t5$z),t5$z),
+ c(rep(0,nrow(t5)),t5$gh),col="lavender")
> t6<-subset(gh,z<=z6)
> polygon(c(rev(t6$z),t6$z),
+ c(rep(0,nrow(t6)),t6$gh),col="lavender")
> # Vẽ đồ thị
> lines(gh,lwd=1)

> # Vẽ các mũi tên
> arrows(2,0.10,6,0.10,
+ angle=30,length=0.1,code=3,lty=2)
> arrows(0,0.02,8,0.02,
+ angle=30,length=0.1,code=3,lty=2)
> arrows(-2,-0.002,10,0.002,
+ angle=30,length=0.1,code=3,lty=2)

> # ghi chú các khoảng
> text(4,0.03,
+ expression(paste("(",mu,"-2",sigma,
+ ",",mu,"+2",sigma,")")),cex=0.8)
> text(4,0.12,
+ expression(paste("(",mu,"-",sigma,
+ ",",mu,"+",sigma,")")),cex=0.8)
> text(4,-0.001,
+ expression(paste("(",mu,"-3",sigma,
+ ",",mu,"+3",sigma,")")),cex=0.8)

```

Ví dụ 63: Trọng lượng của một sản phẩm là biến số ngẫu nhiên có phân phối chuẩn $\mathcal{N}(10;0.25)$. Tìm tỉ lệ những sản phẩm có trọng lượng từ 9.5kg đến 11kg.

Giải: Gọi X là trọng lượng sản phẩm. Theo giả thiết X có phân phối chuẩn với $\mu=10$ và $\sigma=0.5$.

$$P(9.5 \leq X \leq 11) = \varphi\left(\frac{11-10}{0.5}\right) - \varphi\left(\frac{9.5-10}{0.5}\right) = \varphi(2) - \varphi(1) = 0.8185$$

Vậy tỉ lệ các sản phẩm có trọng lượng từ 9.5kg đến 11kg có tỉ lệ 81,85%.

Thực hiện tính toán bằng R:

```

> pnorm(11,10,0.5)-pnorm(9.5,10,0.5)
[1] 0.8185946

```

3. Quan hệ giữa phân phối nhị thức và phân phối chuẩn

Theo định lý giới hạn địa phương Moivre-Laplace và định lý tích phân Moivre-Laplace ta có các kết quả sau:

Đối $X \sim B(n, p)$ khi n lớn và p càng gần 0.5 (không quá gần 0 và 1).

1) Khi đó có thể xem X là phân phối chuẩn có hàm xác suất

$$P(X = k) = \frac{1}{\sqrt{npq}} f\left(\frac{k-\mu}{\sqrt{npq}}\right)$$

$$\text{Với } \mu = np, f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}, q = 1-p$$

(có nghĩa $X \sim \mathcal{N}(\mu; \sigma^2)$, với $\mu = np$ và $\sigma = \sqrt{npq}$)

$$2) P(a \leq X \leq b) \approx \varphi\left(\frac{b-\mu}{\sigma}\right) - \varphi\left(\frac{a-\mu}{\sigma}\right)$$

Chú ý: Các công thức xấp xỉ trên được áp dụng tốt khi $n > 100$, $npq > 20$.

Ví dụ 64: Xác suất sinh được em bé trai là 0.48. Tính xác suất sao cho trong 300 em bé sắp sinh: a) Có 170 bé trai c) Số bé trai ít nhất là 170

b) Số bé trai vào khoảng 150 đến 170

Giải: Gọi X là số bé trai trong 300 bé sắp sinh, khi đó $X \sim B(300, 0.48)$.

Với $n = 300$ khá lớn, $p = 0.48$ không quá gần 0 và 1, ta có thể xem $X \sim \mathcal{N}(\mu; \sigma^2)$ với $\mu = 300 * 0.48 = 144$, $\sigma = \sqrt{300 * 0.48 * 0.52}$

a) Xác suất có 170 bé trai: $P(X = 170) \approx \frac{1}{\sigma} f\left(\frac{170-144}{\sigma}\right)$

b) Xác suất số bé trai vào khoảng 150 đến 170:

$$P(150 \leq X \leq 170) = \varphi\left(\frac{170-144}{\sigma}\right) - \varphi\left(\frac{150-144}{\sigma}\right)$$

c) Xác suất số bé trai ít nhất là: $P(X \geq 170) = 1 - P(X < 170) = 1 - \varphi\left(\frac{b-\mu}{\sigma}\right)$

Thực hiện bằng R:

```
> dnorm(170, 144, 8.6533)
[1] 0.0005050808
> pnorm(170, 144, 8.6533) - pnorm(150, 144, 8.6533)
[1] 0.242707
> 1 - pnorm(170, 144, 8.6533)
[1] 0.001329503
```

Tóm tắt quan hệ giữa các phân phối: Nhị thức, Poisson, Phân phối chuẩn

X~B(n,p)	n rất lớn, p rất nhỏ ($p < 0.01$)
	$X \sim P(\lambda), \lambda = np$
	n rất lớn, p không gần 0 và 1
	$X \sim \mathcal{N}(\mu, \sigma^2)$, với $\mu = np$, $\sigma = \sqrt{npq}$
X~P(λ)	$\lambda \geq 20$
	$X \sim \mathcal{N}(\mu, \sigma^2) = \mathcal{N}(\lambda, \lambda)$

Chú ý: Qui luật phân phối chuẩn là qui luật phân phối xác suất được áp dụng rộng rãi trong thực tế. Điều này đã được giải thích bởi định lý giới hạn trung tâm Lyapunov ([2]). Có thể tóm tắt định lý này như sau:

Nếu biến ngẫu nhiên X là tổng của một số lớn các biến ngẫu nhiên độc lập và giá trị mỗi biến chỉ chiếm một vị trí rất nhỏ trong tổng đó thì X sẽ có phân phối xấp xỉ chuẩn.

6.5.4 Phân phối “Khi bình phương” χ^2

1. Định nghĩa:

Biến số ngẫu nhiên liên tục X với hàm mật độ của nó có dạng:

$$f(x) = \begin{cases} 0, & x \leq 0 \\ \frac{x^{k-1} e^{-\frac{x}{2}}}{2^{\frac{k}{2}} \Gamma(\frac{k}{2})}, & x > 0 \end{cases}$$

Trong đó $\Gamma(k) = \int_0^{+\infty} x^{k-1} e^{-x} dx$, được gọi là có phân phối khi bình phương với bậc k tự do hay nói gọn có phân phối $\chi^2(k)$ và được viết $X \sim \chi^2(k)$.

2. Định lý:

Giả sử biến ngẫu nhiên X_1, X_2, \dots, X_k là các biến số ngẫu nhiên độc lập và có cùng phân phối chuẩn $\mathcal{N}(0,1)$. Khi đó:

$$\sum_{i=1}^k X_i^2 = \chi^2(k)$$

Các hàm trong ngôn ngữ R liên quan:

Loại hàm	Cú pháp	Tác dụng	Chú thích
PMF	dchisq(x,df)	f(x)	$d \equiv distribution$
CDF	pchisq(k,df)	$P(X \leq k)$	$p \equiv probability$
Quantile	qchisq (prob,df)	k_{min} ? khi $F(k) = P(X \leq k) \geq prob$	$q \equiv quantile$
Simulation	rchisq (n,df)	Tạo ra 1 phép thử gồm n mẫu (lần)	$r \equiv random$

Trong đó df là bậc tự do ($df = degrees of freedom$).

- Nếu $X \sim \chi^2(k)$ thì:

$$E(X) = k$$

$$Var(X) = 2k$$

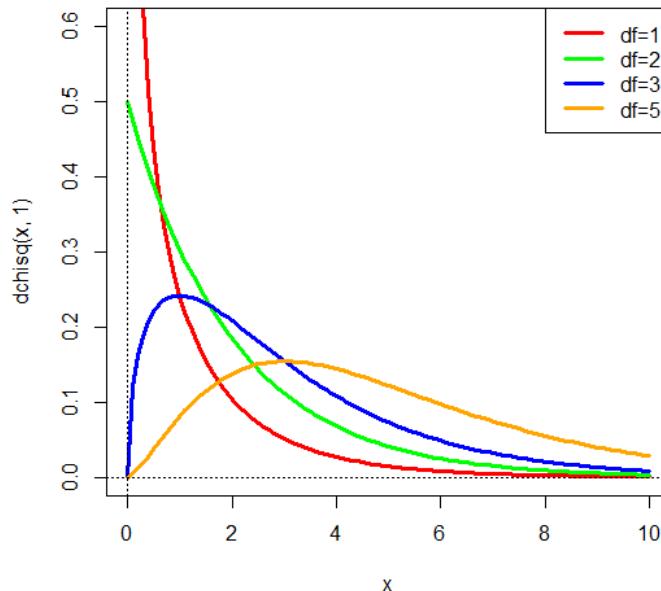
- Với bậc tự do đủ lớn, phân phối khi bình phương xấp xỉ phân phối chuẩn.

Ví dụ 65: Vẽ biểu đồ khi bình phương với bậc tự do 1,2,3,5

```

> curve(dchisq(x,1), xlim=c(0,10), ylim=c(0,0.6),
+ col="red", lwd=3)
> curve(dchisq(x,2), add=T, col="green", lwd=3)
> curve(dchisq(x,3), add=T, col="blue", lwd=3)
> curve(dchisq(x,5), add=T, col="orange", lwd=3)
> abline(h=0, lty=3)
> abline(v=0, lty=3)
> legend(par("usr")[2], par("usr")[4],
+ xjust=1, c("df=1", "df=2", "df=3", "df=5"),
+ lwd=3, lty=1,
+ col=c("red", "green", "blue", "orange"))

```



6.5.5 Phân phối Student

1. Định nghĩa:

Biến số ngẫu nhiên liên tục X với hàm mật độ có dạng

$$f(x) = \frac{\Gamma(\frac{k+1}{2})}{\sqrt{k\pi}\Gamma(\frac{k}{2})} (1 + \frac{x^2}{k})^{-\frac{(k+1)}{2}}$$

Trong đó: $\Gamma(k) = \int_0^{+\infty} x^{k-1} e^{-x} dx$

Được gọi là phân phối Student (hay phân phối t) với k bậc tự do hay nói gọn là X có phân phối t(k) và ký hiệu $X \sim t(k)$.

Các hàm trong ngôn ngữ R liên quan:

Loại hàm	Cú pháp	Tác dụng	Chú thích
PMF	dt(x,df)	f(x)	$d \equiv \text{distribution}$
CDF	pt(k,df)	$P(X \leq k)$	$p \equiv \text{probability}$
Quantile	qt (prob,df)	$k_{\min}?$ khi $F(k) = P(X \leq k) \geq prob$	$q \equiv \text{quantile}$
Simulation	rt (n,df)	Tạo ra 1 phép thử gồm n mẫu (lần)	$r \equiv \text{random}$

Trong đó df là bậc tự do ($df=degrees of freedom$).

2. Định lý:

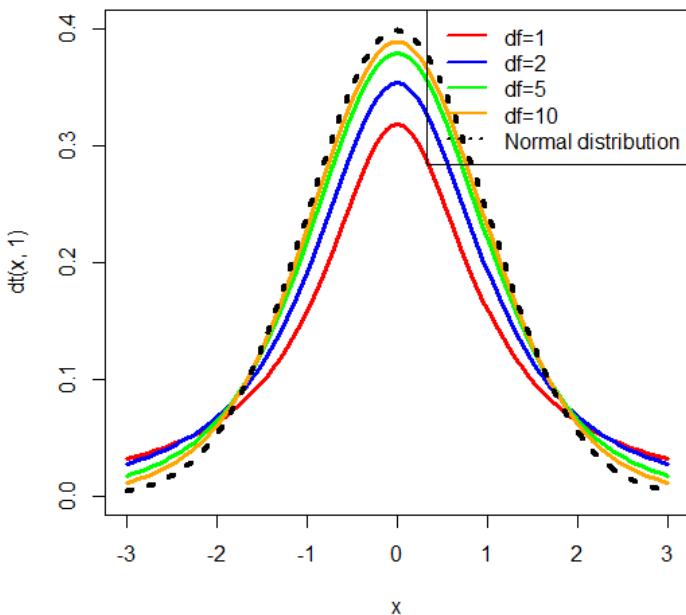
Giả sử 2 biến ngẫu nhiên độc lập, $X \sim N(0,1)$ và $Y \sim \chi^2(k)$. Khi đó $\frac{X}{\sqrt{\frac{Y}{k}}} \approx t(k)$

Ví dụ 66: Vẽ biểu đồ hàm mật độ phân phối t với bậc tự do 1, 2, 5, 10

```
> curve(dt(x,1), xlim=c(-3,3), ylim=c(0,0.4),
+ col="red", lwd=3)
> curve(dt(x,2), add=T, col="blue", lwd=3)
> curve(dt(x,5), add=T, col="green", lwd=3)
> curve(dt(x,10), add=T, col="orange", lwd=3)
> curve(dnorm(x), add=T, lwd=4, lty=3)
> title(main="Student T distribution")

> legend(par("usr")[2], par("usr")[4],
+ xjust=1, c("df=1", "df=2", "df=5", "df=10",
+ "Normal distribution"),
+ lwd=c(2,2,2,2,2), lty=c(1,1,1,1,3),
+ col=c("red", "blue", "green", "orange", par("fg"))))
```

Student T distribution



6.5.6 Phân phối Fisher-Snedecor

1. Định nghĩa:

Biến số ngẫu nhiên liên tục X với hàm mật độ có dạng

$$f(x) = \begin{cases} 0, & x \leq 0 \\ \frac{\Gamma(\frac{m+n}{2})}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})} \left(\frac{m}{n}\right)^{\frac{m}{2}} x^{\frac{m}{2}-1} \left(1 + \frac{mx}{n}\right)^{-\frac{(m+n)}{2}}, & x > 0 \end{cases}$$

Trong đó: $\Gamma(k) = \int_0^{+\infty} x^{k-1} e^{-x} dx$

Được gọi là phân phối Fisher-Snedecor (hay phân phối F) với (m,n) bậc tự do hay nói gọn là X có phân phối $F(m,n)$ và ký hiệu $X \sim F(m,n)$.

Các hàm trong ngôn ngữ R liên quan:

Loại hàm	Cú pháp	Tác dụng	Chú thích
PMF	$df(x,m,n)$	$f(x)$	$d \equiv distribution$
CDF	$pf(k,m,n)$	$P(X \leq k)$	$p \equiv probability$
Quantile	$qf(prob,m,n)$	$k_{min}?$ khi $F(k) = P(X \leq k) \geq prob$	$q \equiv quantile$
Simulation	$rf(k,m,n)$	Tạo ra 1 phép thử gồm k mẫu (lần)	$r \equiv random$

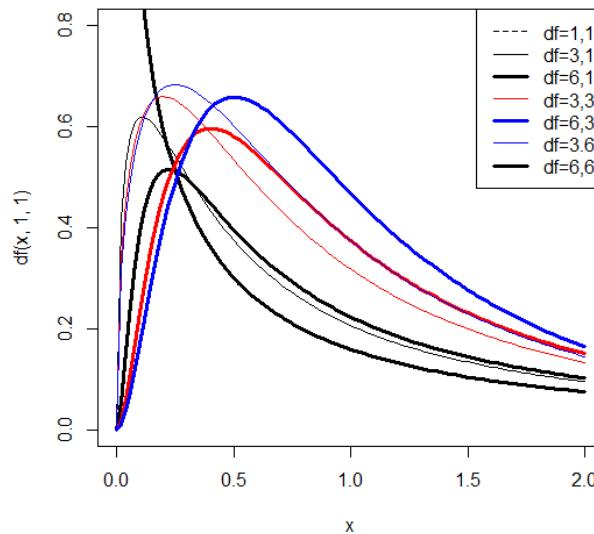
Ví dụ 67: Vẽ biểu đồ hàm mật độ phân phối t với bậc tự do (1,1), (3,1), (6,1), (3,3), (6,3), (3,6), (6,6)

```
> curve(df(x,1,1),xlim=c(0,2), ylim=c(0,0.8),lwd=3)
> curve(df(x,3,1),add=T)
> curve(df(x,6,1),add=T,lwd=3)
> curve(df(x,3,3),add=T,col="red")
> curve(df(x,6,3),add=T,col="red",lwd=3)
> curve(df(x,3,6),add=T,col="blue")
> curve(df(x,6,6),add=T,col="blue",lwd=3)

> title(main="Fisher F distribution")
> legend(par("usr")[2],par("usr")[4],
+ xjust=1,c("df=1,1","df=3,1","df=6,1","df=3,3",
+ "df=6,3","df=3,6","df=6,6"),
+ lwd=c(1,1,3,1,3,1,3),
+ lty=c(2,1,1,1,1,1,1),
+ col=c(par("fg"),par("fg"),par("fg"),"red","blue","blue"))

```

Fisher F distribution



6.6 PHÂN PHỐI XÁC SUẤT CỦA ĐẠI LƯỢNG NGẪU NHIÊN 2 CHIỀU

6.6.1. Định nghĩa

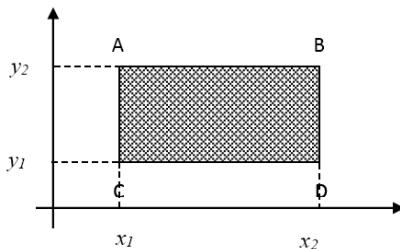
Hàm phân phối xác suất của đại lượng ngẫu nhiên 2 chiều (X, Y) hay hàm phân phối xác xuất liên hợp của X và Y (*Joint Cumulative distribution function*) là hàm 2 biến được xác định như sau:

$$F(x, y) = P(X < x; Y < y), \forall x, y \in \mathbb{R}$$

6.6.2. Tính chất

Cũng như trường hợp 1 biến, hàm phân phối xác xuất $F(x,y)$ có các tính chất:

- a. $0 \leq F(x,y) \leq 1, \forall x, y \in \mathbb{R}$
- b. $F(x,y)$ không giảm theo từng biến số.
- c. $F(-\infty, +\infty) = 0; F(+\infty, +\infty) = 1$ và $F(x, +\infty) = F_X(x); F(+\infty, y) = F_Y(y)$ ($F_X(x), F_Y(y)$: hàm phân phối xác suất biên duyên tương ứng của X và Y - Marginal Probability Mass Function)
- d. $P(x_1 < X < x_2; y_1 < Y < y_2) = F(x_1, y_1) + F(x_2, y_2) - F(x_1, y_2) - F(x_2, y_1)$. Khi đó xác suất của (X, Y) nhận giá trị là diện tích là hình chữ nhật ABCD như hình vẽ



6.6.3. Định nghĩa

Hai đại lượng ngẫu nhiên X, Y được gọi là độc lập nếu:

$$F(x,y) = F_X(x).F_Y(y)$$

6.6.4. Phân phối xác suất của đại lượng ngẫu nhiên 2 chiều rời rạc

a. *Định nghĩa*: Bảng phân phối xác suất của đại lượng ngẫu nhiên 2 chiều rời rạc (X, Y) còn gọi là bảng phân phối đồng thời của X và Y được cho như sau:

$X \backslash Y$	y_1	y_2	y_m	$p(x_i)$
x_1	p_{11}	p_{12}	p_{1m}	$p(x_1)$
x_2	p_{21}	p_{22}	p_{2m}	$p(x_2)$
....
x_n	p_{n1}	p_{n2}	p_{nm}	$p(x_n)$
$q(y_j)$	$q(y_1)$	$q(y_2)$		$q(y_m)$	

Trong đó, $p_{ij} = P(X=x_i, Y=y_j)$ (giá trị xác suất liên hợp của X và Y) và

$$\sum_{i=1}^n \sum_{j=1}^m p_{ij} = 1$$

Với bảng phân phối xác suất của (X, Y) ta có:

- $F(x,y) = \sum_{x_i < x} \sum_{y_j < y} p_{ij}$
- Các phân phối biên

- Của X :

X	x_1	x_2	x_n
P	$p(x_1)$	$p(x_2)$	$p(x_n)$

$$\text{Với } p(x_i) = \sum_{j=1}^m p_{ij}$$

- Của Y:

Y	y_1	y_2	y_n
P	$p(x_1)$	$p(x_2)$	$p(x_n)$

Với $p(y_i) = \sum_{j=1}^n p_{ij}$

- Nếu X và Y độc lập $\Leftrightarrow p_{ij} = p(x_i)p(y_j), \forall i, j$

- Phân phối có điều kiện

- Của X với điều kiện $Y=y_j$:

X	x_1	x_2	x_n
$P(X Y=y_j)$	$p_1 y_j$	$p_2 y_j$	$p_n y_j$

Trong đó

$$p_i|y_j = P(X=x_i|Y=y_j) = \frac{P(X=x_i, Y=y_j)}{P(Y=y_j)} = \frac{P_{ij}}{q(y_j)}, i = \overline{1, n}$$

- Của Y với điều kiện $X=x_i$:

Y	y_1	y_2	y_m
$P(Y X=x_i)$	$q_1 x_i$	$q_2 x_i$	$q_m x_i$

Trong đó

$$q_j|x_i = P(Y=y_j|X=x_i) = \frac{P(X=x_i, Y=y_j)}{P(X=x_i)} = \frac{P_{ij}}{p(x_i)}, j = \overline{1, m}$$

- Kỳ vọng có điều kiện

- Kỳ vọng của X với điều kiện $Y=y_j$, ký hiệu $E(X|Y=y_j)$, với

$$E(X|Y=y_j) = \sum_{i=1}^n x_i P(X=x_i|Y=y_j) = \sum_{i=1}^n x_i p_i |y_j$$

- Kỳ vọng của Y với điều kiện $X=x_i$, ký hiệu $E(Y|X=x_i)$, với

$$E(Y|X=x_i) = \sum_{j=1}^m y_j P(Y=y_j|X=x_i) = \sum_{j=1}^m y_j q_j |x_i$$

- $E(\varphi(X, Y)) = \sum_{i=1}^n \sum_{j=1}^m \varphi(x_i, y_j) p_{ij}$

Ví dụ 68: Người ta thống kê dân số một vùng theo 2 chỉ tiêu: Giới tính (X) và học vấn (Y), kết quả thu được như sau:

X \ Y	Thất học: 0	Phổ thông: 1	Đại học: 2
Nam: 0	0.12	0.24	0.15
Nữ: 1	0.13	0.22	0.14

- Lập bảng phân phối xác suất của học vấn; của giới tính
- Học vấn có độc lập với giới tính hay không?
- Tìm xác suất để lấy ngẫu nhiên một người thì người đó không bị thất học.
- Lập bảng phân phối xác suất học vấn của nữ; tính trung bình học vấn của nữ.

Giải:

a. Bảng phân phối xác suất của X và Y:

X	0	1
P	0.51	0.49

b. Do $p_{11}=P(X=0, Y=0)=0.12$ và $P(x=0).P(y=0) = 0.51 \times 0.25 = 0.1275$.

Vì $p_{11} \neq P(X=0).P(Y=0)$ nên học vẫn không độc lập với giới tính.

c. Ta có

$$P(X \geq 0, Y > 0) = \sum_{x_i > 0} \sum_{y_j > 0} p_{ij} = 0.24 + 0.22 + 0.15 + 0.14 = 0.75 = 1 - P(X \geq 0, Y \leq 0)$$

d. Lập bảng phân phối của Y với điều kiện X=1

Ta có: $P(X=1) = 0.49$, do đó

$$P(Y = 0|X = 1) = \frac{P(X = 1, Y = 0)}{P(X = 1)} = \frac{P_{21}}{P(X = 1)} = \frac{0.13}{0.49} = 0.2653$$

Và

$$P(Y = 1|X = 1) = \frac{P(X = 1, Y = 2)}{P(X = 1)} = \frac{P_{23}}{P(X = 1)} = \frac{0.16}{0.49} = 0.2857$$

Bảng phân phối điều kiện:

X	0	1	2
$P(Y X=1)$	0.2653	0.449	0.2857

Ví dụ 69: Tung 1 con xúc xắc 2 lần, định nghĩa các biến ngẫu nhiên U và V:

U = giá trị lớn nhất (của 2 mặt) 2 lần tung.

V = tổng (2 mặt) 2 lần tung

Không gian trạng thái của U là $S_U = \{1, 2, \dots, 6\}$

Không gian trạng thái của V là $S_V = \{2, 3, \dots, 12\}$

Biểu diễn các không gian trạng thái này bằng ma trận, ta có kết quả ở trang kế tiếp.

Dựa vào ma trận này, ta có thể tính hàm giá trị phân phối biên duyên của U. Để thấy, xác suất xuất hiện mỗi cặp giá trị 2 mặt ứng với 2 lần tung là $1/36$.

Do:

- 1 lần U nhận giá trị 1, nghĩa là: $f_U(1) = P(U=1) = 1/36$

- 3 lần U nhận giá trị 2, nghĩa là: $f_U(2) = P(U=2) = 3/36$

Tiếp tục nhận xét trên, ta thấy hàm phân phối biên duyên của U có thể viết:

$$f_U(u) = \frac{2u - 1}{36}, u = 1, 2, \dots, 6$$

Tương tự đối với biến ngẫu nhiên V, hàm phân phối biên duyên của V có thể viết:

$$f_V(v) = \frac{6 - |v - 7|}{36}, v = 2, 3, \dots, 12$$

U	1	2	3	4	5	6	V	1	2	3	4	5	6
1	1	2	3	4	5	6	1	2	3	4	5	6	7
2	2	2	3	4	5	6	2	3	4	5	6	7	8
3	3	3	3	4	5	6	3	4	5	6	7	8	9
4	4	4	4	4	5	6	4	5	6	7	8	9	10
5	5	5	5	5	5	6	5	6	7	8	9	10	11
6	6	6	6	6	6	6	6	7	8	9	10	11	12

(U,V)	1	2	3	4	5	6
1	(1,2)	(2,3)	(3,4)	(4,5)	(5,6)	(6,7)
2	(2,3)	(2,4)	(3,5)	(4,6)	(5,7)	(6,8)
3	(3,4)	(3,5)	(3,6)	(4,7)	(5,8)	(6,9)
4	(4,5)	(4,6)	(4,7)	(4,8)	(5,9)	(6,10)
5	(5,6)	(5,7)	(5,8)	(5,9)	(5,10)	(6,11)
6	(6,7)	(6,8)	(6,9)	(6,10)	(6,11)	(6,12)

	2	3	4	5	6	7	8	9	10	11	12	
1	1/36											1/36
2		2/36	1/36									3/36
3			2/36	2/36	1/36							5/36
4				2/36	2/36	2/36	1/36					7/36
5					2/36	2/36	2/36	2/36	1/36			9/36
6						2/36	2/36	2/36	2/36	2/36	1/36	11/36
	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36	

Hàm phân phối xác suất liên hợp của (U, V)

Thể hiện bằng ngôn ngữ R

```
> library(prob)
> S<-rolldie(2,makespace=TRUE)
> S<-addrv(S, FUN=max, invars = c("X1","X2"), name ="U")
> S<-addrv(S, FUN=sum, invars = c("X1","X2"), name ="V")
> head(S)
  X1 X2 U V      probs
1  1  1 1 2 0.02777778
2  2  1 2 3 0.02777778
3  3  1 3 4 0.02777778
4  4  1 4 5 0.02777778
5  5  1 5 6 0.02777778
6  6  1 6 7 0.02777778
> UV <- marginal(S,vars = c("U","V"))
> head(UV)
  U V      probs
1 1 2 0.02777778
2 2 3 0.05555556
3 2 4 0.02777778
4 3 4 0.05555556
5 3 5 0.05555556
6 4 5 0.05555556
```

Hiển thị dưới dạng bảng, hàm phân phối xác suất liên hợp bằng hàm xtabs

```
> xtabs(round(probs,3)~U+V, data = UV)
      V
U   2   3   4   5   6   7   8   9   10  11  12
  1 0.028 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000
  2 0.000 0.056 0.028 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000
  3 0.000 0.000 0.056 0.056 0.028 0.000 0.000 0.000 0.000 0.000 0.000
  4 0.000 0.000 0.000 0.056 0.056 0.056 0.028 0.000 0.000 0.000 0.000
  5 0.000 0.000 0.000 0.000 0.056 0.056 0.056 0.056 0.028 0.000 0.000
  6 0.000 0.000 0.000 0.000 0.000 0.056 0.056 0.056 0.056 0.056 0.028
```

Các giá trị phân phối biên duyên của U

```
> marginal(UV, vars="U")
      U      probs
  1 1 0.02777778
  2 2 0.08333333
  3 3 0.13888889
  4 4 0.19444444
  5 5 0.25000000
  6 6 0.30555556
```

Các giá trị phân phối biên duyên của V

```
> marginal(UV, vars="V")
      V      probs
  1 2 0.02777778
  2 3 0.05555556
  3 4 0.08333333
  4 5 0.11111111
  5 6 0.13888889
  6 7 0.16666667
  7 8 0.13888889
  8 9 0.11111111
  9 10 0.08333333
  10 11 0.05555556
  11 12 0.02777778
```

6.6.5. Phân phối xác suất của đại lượng ngẫu nhiên 2 chiều liên tục

Giả sử (X, Y) là đại lượng ngẫu nhiên liên tục 2 chiều có hàm phân phối xác suất $F(x, y)$ khả vi bậc hai liên tục.

1 *Định nghĩa:* Ta gọi $f(x, y) = \begin{cases} \frac{\delta^2 F(x, y)}{\delta x \delta y}, & \text{nếu tồn tại } \forall (x, y) \\ 0, & \text{nếu không tồn tại} \end{cases}$

là hàm mật độ của đại lượng ngẫu nhiên liên tục 2 chiều (X, Y) .

Khi đó,

$$P[(X, Y) \in D] = \iint_D f(x, y) dx dy, \text{ với } D \subset \mathfrak{N}$$

$$\text{và } F(x, y) = \int_{-\infty}^x \int_{-\infty}^y f(x, y) dx dy$$

2 *Tính chất:* Hàm mật độ $f(x, y)$ có các tính chất

a. $f(x, y) \geq 0; \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f(x, y) dx dy = 1$

b. Hàm mật độ của X: $f_x(x) = \int_{-\infty}^{+\infty} f(x, y) dy$; của Y: $f_y(y) = \int_{-\infty}^{+\infty} f(x, y) dx$

c. X và Y độc lập $\Leftrightarrow f(x,y) = f_X(x)f_Y(Y)$

d. Hàm phân phối có điều kiện

- Của X với điều kiện $Y=y$, ký hiệu $F(x|y)$ được định nghĩa

$$F(x|y) = P(X=x, Y=y) = \lim_{\delta \rightarrow 0} P(X < x | y \leq Y \leq y + \delta)$$

$$\text{Ta có } F(x|y) = \int_{-\infty}^x \frac{f(u,y)}{f_Y(y)} du, \text{ nếu } f_Y(y) > 0$$

- Của Y với điều kiện $X=x$, ký hiệu $F(y|x)$ được định nghĩa

$$F(y|x) = P(X=x, Y < y) = \lim_{\delta \rightarrow 0} P(x \leq X \leq x + \delta | Y < y)$$

$$\text{Ta có } F(y|x) = \int_{-\infty}^y \frac{f(u,x)}{f_X(x)} du, \text{ nếu } f_X(x) > 0$$

Ví dụ 70: Giả sử hàm phân phối xác suất liên hợp của (X,Y) được cho bởi:

$$f(x,y) = \frac{6}{5}(x+y^2), 0 < x, 1, 0 < y < 1$$

Hàm phân phối xác suất biên duyên của X là

$$f_X(x) = \int_0^1 \frac{6}{5}(x+y^2) dy = \frac{6}{5} \left(xy + \frac{y^3}{3} \right) \Big|_{y=0}^1 = \frac{6}{5} \left(x + \frac{1}{3} \right)$$

Hàm phân phối xác suất biên duyên của Y là

$$f_Y(y) = \int_0^1 \frac{6}{5}(x+y^2) dx = \frac{6}{5} \left(\frac{x^2}{2} + xy^2 \right) \Big|_{x=0}^1 = \frac{6}{5} \left(\frac{1}{2} + y^2 \right)$$

6.6.6 Hệ số tương quan

1. Định nghĩa: Hệ số tương quan của X, Y, ký hiệu $\rho(X, Y)$ được xác định.

$$\rho(X, Y) = \frac{\text{cov}(X, Y)}{\sqrt{D(X)}\sqrt{D(Y)}} = \frac{E[XY] - E[X]E[Y]}{\sqrt{D(X)}\sqrt{D(Y)}}$$

2 Tính chất: Hệ số tương quan có các tính chất

- $|\rho(aX+b, cY+d)| = |\rho(X, Y)|, \forall a, b, c, d \in \mathbb{R}$
- $|\rho(X, Y)| \leq 1$
- $|\rho(X, Y)| = 1 \Leftrightarrow X, Y \text{ có sự liên hệ tuyến tính}$
- Nếu X, Y độc lập thì $|\rho(X, Y)| = 0$

TÀI LIỆU THAM KHẢO

[1] Biostatistics with R, Babak Shahbaba, Springer, 2012

[2] Giáo trình Lý thuyết xác suất và thống kê toán, Nguyễn Cao Văn, Trần Thái Ninh, NXB Thống kê, 2005.

[3] Introduction to Probability and Statistics Using R, G.Jay Kerns, First Edition, cran.r-project.org/web/packages/IPSUR/vignettes/IPSUR.pdf, 2010

[4] Lý thuyết xác suất và thống kê (bài giảng), Hoàng Văn Hà, ĐH KHTN TP HCM.

- [5] Thống kê ứng dụng trong kinh tế-xã hội, Hoàng Trong, Chu Nguyễn Mộng Ngọc, NXB Lao động-Xã hội, 2010.
- [6] Phân tích số liệu và biểu đồ bằng R, Nguyễn Văn Tuấn, *cran.r-project.org/doc/contrib/Intro_to_R_Vietnamese.pdf*
- [7] Xác suất thống kê, Lê Bá Phi (bài giảng), ĐH Nha Trang
- [8] <http://statistics.vn/index.php/thongkecanban/>

C H U O NG

7

ƯỚC LƯỢNG

7.1 LÝ THUYẾT MẪU

1. Mẫu ngẫu nhiên (*Random Sample*)

Giả sử trên quần thể Ω có đặc tính X . Tiến hành chọn mẫu ngẫu nhiên n cá thể của Ω , ta được các kết quả:

$$\Omega^n = \{(x_1, x_2, \dots, x_n), x_i \in \Omega, i=1, n\}$$

Trên Ω^n , xác định họ biến ngẫu nhiên (X_1, X_2, \dots, X_n) bởi hệ thức:

$$X_i : \Omega^n \rightarrow \mathbb{R}$$

$$x \mapsto X_i(x) = x_i$$

trong đó $x = (x_1, x_2, \dots, x_n)$.

Họ biến ngẫu nhiên (X_1, X_2, \dots, X_n) được gọi là mẫu ngẫu nhiên kích thước n của X .

Bộ (x_1, x_2, \dots, x_n) là 1 quan sát cụ thể ứng với họ biến ngẫu nhiên (X_1, X_2, \dots, X_n) được gọi là mẫu thực nghiệm.

2. Phép chọn mẫu:

- Chọn mẫu ngẫu nhiên có hoàn lại:** trong phép chọn mẫu ngẫu nhiên n cá thể, một cá thể đã được chọn quan sát ở thứ i , vẫn được tham gia chọn ở các bước tiếp theo.
- Chọn mẫu ngẫu nhiên không hoàn lại:** trong phép chọn mẫu ngẫu nhiên n cá thể, một cá thể đã được chọn quan sát ở thứ i , không được tham gia chọn ở các bước tiếp theo.

7.2 PHÂN PHỐI MẪU

Trên quần thể Ω , xét đặc tính định lượng X , với hàm phân phối $F(x)$ chưa biết. Giả sử (X_1, X_2, \dots, X_n) là mẫu ngẫu nhiên kích thước n của X và (x_1, x_2, \dots, x_n) là mẫu thực nghiệm. Ta gọi hàm $F_n^*(x) = \frac{M(x)}{n}$, với $M(x)$: số các $x_i < x$ ($i=1, n$) là **hàm phân phối mẫu** của X qua mẫu thực hiện (x_1, x_2, \dots, x_n) .

Theo định lý Glivenko-Catelli, $F_n^*(x)$ có các tính chất sau:

- $F_n^*(x)$ xác định duy nhất với mẫu thực nghiệm (x_1, x_2, \dots, x_n)
- $\sup_{x \in \mathbb{R}} |F_n^*(x) - F(x)| \rightarrow 0 (n \rightarrow \infty)$

Ứng dụng:

- Nếu $\{X_1, X_2, \dots, X_n\}$ là một mẫu ngẫu nhiên lấy từ tổng thể của biến ngẫu nhiên X với giá trị trung bình μ và phương sai σ^2 thì:

$$E(\bar{X}) = \mu \text{ và } Var(\bar{X}) = \frac{\sigma^2}{n}$$

- Nếu $\{X_1, X_2, \dots, X_n\}$ là một mẫu ngẫu nhiên có phân phối $\mathcal{N}(\mu; \sigma^2)$ thì

$$\bar{X} \sim \mathcal{N}(\mu; \sigma^2/n)$$

7.3 UỐC LUỢNG THAM SỐ

Giả sử đặc trưng quần thể cần nghiên cứu được biểu diễn bằng một biến ngẫu nhiên X xác định trên không gian Ω . Một số đại lượng thống kê liên quan đến X cần được xác định như kỳ vọng, phương sai, giá trị trung bình..được coi là tham số của X . Do lực lượng quần thể thường quá lớn nên việc tính các tham số của X thường được ước tính dựa trên các mẫu quan sát được.

Một trong những bài toán quan trọng của thống kê là ước lượng các tham số của các đại lượng ngẫu nhiên. Giá trị ước lượng trả về của một tham số nếu là một giá trị duy nhất thì được gọi là ước lượng điểm (*point estimator*), nếu giá trị trả về của một ước lượng là một khoảng thì gọi là ước lượng khoảng (*interval estimator*).

7.4 UỐC LUỢNG ĐIỂM (POINT ESTIMATOR)

1. Định nghĩa

Giả sử cần khảo sát một đặc tính θ của một quần thể, người ta thường xem xét biến ngẫu nhiên X tương ứng có hàm phân phối $F(x; \theta)$, trong đó θ là tham số chưa biết.

Với một mẫu ngẫu nhiên kích thước n từ $X = (X_1, X_2, \dots, X_n)$

Thống kê $\hat{\theta} = h(X_1, X_2, \dots, X_n)$ phụ thuộc θ được là một **ước lượng điểm** cho θ .

Với một mẫu thực nghiệm (x_1, x_2, \dots, x_n) , thống kê $\hat{\theta} = h(x_1, x_2, \dots, x_n)$ là một **giá trị ước lượng điểm** cho θ .

Ví dụ 71: X là biến ngẫu nhiên thể hiện chiều cao dân số trong một khu vực.

$X \sim \mathcal{N}(\mu; \sigma^2)$. Thống kê trung bình mẫu và phương sai dựa trên mẫu kích thước n từ $X = (X_1, X_2, \dots, X_n)$

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

là các ước lượng điểm cho kỳ vọng μ và phương sai σ^2 .

Với một mẫu thực nghiệm $x_1=150, x_2=155, x_3=167$, giá trị ước lượng điểm của kỳ vọng μ và phương sai σ^2 là $\bar{x}=157.33$ và $s^2=76.33$. Thực hiện bằng ngôn ngữ R:

```

> X<-c(150,155,167)
> mean(X)
[1] 157.3333
> var(X)
[1] 76.33333

```

2. Phân loại ước lượng

- Uớc lượng $\hat{\theta} = h(X_1, X_2, \dots, X_n)$ cho θ được gọi là **không chêch** (*unbiased estimator*) nếu:

$$E(\hat{\theta}) = E(h(X_1, X_2, \dots, X_n)) = \theta$$

(Tùy ý nghĩa của kỳ vọng, nếu $\hat{\theta} = h(X_1, X_2, \dots, X_n)$ là ước lượng không chêch, thì ta luôn tránh được các số lệch về một phía, hoặc luôn lớn hơn, hoặc luôn nhỏ hơn tham số cần ước lượng)

Đại lượng $Bias(\hat{\theta}) = E(\hat{\theta}) - \theta$ được gọi là **độ chêch** (*Bias*) của tham số θ

- Uớc lượng $\hat{\theta} = h(X_1, X_2, \dots, X_n)$ cho θ được gọi là **vững** (*consistency estimator*) nếu:

$$\hat{\theta} \xrightarrow{P} \theta (n \rightarrow \infty)$$

hay $\lim_{n \rightarrow \infty} P(|\hat{\theta} - \theta| < \varepsilon) = 1$

Ví dụ 72: Xét một đại lượng ngẫu nhiên X có phân phối chuẩn $\mathcal{N}(\mu; \sigma^2)$, (X_1, X_2, \dots, X_n) là mẫu kích thước n của X . Uớc lượng $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ là ước lượng không chêch và vững cho μ .

$$E(\bar{X}) = \frac{1}{n} \sum_{i=1}^n E(X_i) = \frac{1}{n} n\mu = \mu$$

và $\lim_{n \rightarrow \infty} P(|\bar{X} - \mu| < \varepsilon) = 1$ (luật số lớn)

(xem thêm ở [4])

- Uớc lượng hiệu quả (*Efficiency estimator*)

Xét $\hat{\theta}, \tilde{\theta}$ là 2 ước lượng không chêch của θ , $\hat{\theta}$ được gọi là ước lượng hiệu quả hơn $\tilde{\theta}$, nếu với cỡ mẫu n cho trước: $Var(\hat{\theta}) < Var(\tilde{\theta})$

- Sai số trung bình bình phương (*MSE: Mean square Error*)

Trong trường hợp $\hat{\theta}, \tilde{\theta}$ là 2 ước lượng của θ (có thể chêch hoặc không chêch), người ta có thể sử dụng sai số trung bình bình phương (MSE) để đánh giá tính hiệu quả của tham số ước lượng [4]

$$MSE(\hat{\theta}) = E[(\hat{\theta} - \theta)^2] = Var(\hat{\theta}) + [Bias(\hat{\theta})]^2$$

Tham số nào có sai số trung bình bình phương bé hơn là ước lượng hiệu quả hơn.

Sai số chuẩn (*Standard Error*)

Sai số chuẩn (SE) của một ước lượng $\hat{\theta}$ là độ lệch tiêu chuẩn của nó, xác định bởi công thức: $SE(\hat{\theta}) = \sqrt{Var(\hat{\theta})}$

3. Các tham số và ước lượng thường gặp

Ký hiệu: X là đại lượng ngẫu nhiên trên quần thể

(X_1, X_2, \dots, X_n) được gọi là mẫu ngẫu nhiên kích thước n

Tham số	Uớc lượng T	Var(T)	SE(T)
$E(X) = \mu$	$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$	$Var(\bar{X}) = \frac{S^2}{n}$	$\frac{S}{\sqrt{n}}$
$D(X) = \sigma^2$	$S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}$	$\frac{2S^2}{n-1}$	$S^2 \sqrt{\frac{2}{n-1}}$
$P(A) = \frac{M_A}{N} = p$	$\hat{p} = \frac{m_A}{n}$	$\frac{\hat{p}(1-\hat{p})}{n}$	$\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

M_A : số phần tử có tính chất A trong quần thể; N: số lượng phần tử của quần thể

m_A : số phần tử có tính chất A trong mẫu, n kích thước của mẫu.

Ví dụ 73: So sánh ước lượng hiệu quả

Xét một đại lượng ngẫu nhiên có phân phối chuẩn, có kỳ vọng $E(X)=3$ và phương sai=2. Giả định rằng mẫu có kích thước $n=100$. So sánh 2 ước lượng giá trị trung bình của X là:

- + $X.bar$: trung bình cộng,
- + $mid.range$: trung tâm ($= \frac{\max(X) + \min(X)}{2}$) .

Thực hiện tạo mẫu mô phỏng, kết quả cho thấy ước lượng giá trị trung bình của X bằng trung bình cộng là tốt hơn do $Var(X.bar) < Var(mid.range)$

```
> mu<-3
> sig<-sqrt(2)
> X.bar<-rep(0,10^5)
> mid.range<-rep(0,10^5)
> for(i in 1:10^5)
+ {
+ X<-rnorm(100,mu,sig)
+ X.bar[i]<-mean(X)
+ mid.range[i]<- (max(X)+min(X))/2
+ }
> var(X.bar)
[1] 0.01986728
> var(mid.range)
[1] 0.1844387
```

Ví dụ 74: Xét một đại lượng ngẫu nhiên có phân phối chuẩn, có kỳ vọng $E(X)=5$ và phương sai=3. Giả định rằng mẫu có kích thước $n=20$. So sánh 2 ước lượng phương sai

$$S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1} \text{ và } S'^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n} = \frac{n-1}{n} S^2$$

Chú ý:

1. S^2 là ước lượng phương sai mẫu, trong R được tính bằng `var(X)`
2. Từ định nghĩa của S'^2 :

$$Var(S'^2) = \frac{n-1}{n} Var(S^2) \text{ và } E(S'^2) = \frac{n-1}{n} E(S^2) = \frac{n-1}{n} \sigma^2$$

3. Do ước lượng S^2 là ước lượng không chêch nên $Bias(S^2)=0$ và $E(S^2)=\sigma^2$.

$$\text{Vì vậy, } MSE(S^2) = Var(S^2) \text{ và } MSE(S'^2) = Var(S'^2) + [Bias(S'^2)]^2$$

Theo định nghĩa độ chêch:

$$Bias(S'^2) = E(S'^2) - \sigma^2 = \frac{n-1}{n} \sigma^2 - \sigma^2$$

Thể hiện bằng ngôn ngữ R:

```
> mu<-5
> std<-sqrt(3)
> X.var<-rep(0,10^5)
> for(i in 1:10^5)
+ {X<-rnorm(20,mu,std)
+ X.var[i]<-var(X)
+ }
> var(X.var)
[1] 0.9513839
> var((19/20)*X.var)
[1] 0.858624
> bias.sp2<-19/20*3-3
> MSE.sp2<-var((19/20)*X.var)+bias.sp2^2
> MSE.s2<-var(X.var)
> MSE.s2
[1] 0.9513839
> MSE.sp2
[1] 0.881124
```

Như vậy ước lượng phương sai bằng S'^2 hiệu quả hơn ước lượng bằng S^2 .

4. Phương pháp ước lượng tham số hợp lý cực đại

a. Giới thiệu:

Phương pháp hợp lý cực đại được K.Gauss và R.Fisher đề xuất.

Ý tưởng của phương pháp là biểu diễn hàm mật độ xác suất của mẫu ngẫu nhiên phụ thuộc vào tham số gọi là hàm hợp lý (*likelihood function*). Tìm các giá trị tham số để hàm hợp lý đạt cực đại, các giá trị của các tham số này là ước lượng điểm của các tham số.

b. Định nghĩa:

Giả sử (X_1, X_2, \dots, X_n) là một mẫu ngẫu nhiên của đặc tính X có hàm mật độ $f(x, \theta)$ phụ thuộc vào tham số θ . Tại một mẫu cụ thể (x_1, x_2, \dots, x_n) của vector ngẫu nhiên (X_1, X_2, \dots, X_n) . Hàm hợp lý L được xác định như sau:

$$L(\theta) = \prod_{i=1}^n f(x_i, \theta)$$

Giá trị θ làm cho hàm $L(\theta)$ đạt giá trị cực đại gọi là ước lượng cực đại (*maximum likelihood estimator - MLE*) ký hiệu $\hat{\theta}$, giá trị này gọi là ước lượng điểm của θ

c. Phương pháp ước lượng bằng hàm hợp lý cực đại

Giả sử hàm $L(\theta)$ là hàm khả vi, vì hàm $\ln()$ là hàm đơn điệu nên θ là nghiệm của phương trình:

$$\sum_{i=1}^n \frac{\partial(\ln(f(x_i, \theta)))}{\partial \theta} = 0$$

Ví dụ 75:

a. (X_1, X_2, \dots, X_n) là một mẫu ngẫu nhiên của đặc tính X có phân phối mũ

$$f(x, \theta) = \begin{cases} 0, & x \leq 0 \\ \frac{1}{\theta} e^{-\frac{x}{\theta}}, & x > 0, \theta > 0 \end{cases}$$

Tìm ước lượng hợp lý cực đại của θ

Giai:

$$\text{Ta có } \ln(f(x_i, \theta)) = -\ln(\theta) - \frac{x_i}{\theta} \Rightarrow \frac{d(\ln(f(x_i, \theta)))}{d\theta} = -\frac{1}{\theta} + \frac{x_i}{\theta^2}$$

Phương trình hợp lý có dạng:

$$\sum_{i=1}^n \left(-\frac{1}{\theta} + \frac{x_i}{\theta^2} \right) = 0 \Leftrightarrow \sum_{i=1}^n (-\theta + x_i) = 0 \Leftrightarrow \hat{\theta} = \frac{1}{n} \sum_{i=1}^n x_i = \bar{X}$$

Vậy $\hat{\theta} = \bar{X}$ là ước lượng hợp lý cực đại của θ .

b. Cho (X_1, X_2, \dots, X_n) là một mẫu ngẫu nhiên của đặc tính X có phân phối chuẩn $\mathcal{N}(\mu; \sigma^2)$.
Tìm ước lượng hợp lý cực đại cho μ, σ^2

Giai:

Ta có:

$$\ln(f(x_i, \mu, \theta)) = \frac{1}{2} \ln(2\pi\sigma^2) - \frac{(x_i - \mu)^2}{2\sigma^2}$$

$$\frac{\partial \ln(f(x, \mu, \sigma^2))}{\partial \mu} = \frac{x_i - \mu}{\sigma^2}$$

$$\frac{\partial \ln(f(x, \mu, \sigma^2))}{\partial \sigma} = -\frac{1}{2\sigma^2} + \frac{(x_i - \mu)^2}{2\sigma^2}$$

Hệ phương trình hợp lý có dạng:

$$\sum_{i=1}^n \frac{x_i - \mu}{\sigma^2} = 0 \Leftrightarrow \hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i = \bar{X}$$

$$\sum_{i=1}^n \left(-\frac{1}{2\sigma^2} + \frac{(x_i - \mu)^2}{2\sigma^2} \right) = 0 \Leftrightarrow \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{X})^2$$

Ví dụ 76: Cho (X_1, X_2, \dots, X_n) là một mẫu ngẫu nhiên của đặc tính X có phân phối Bernoulli (xác suất p) ta có:

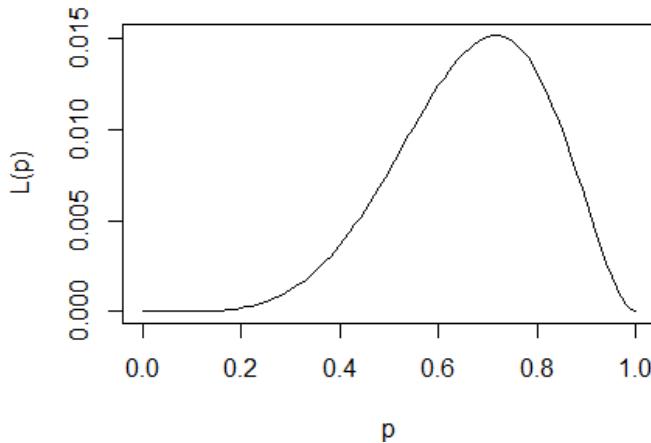
$$P(X_1 = x_1, X_2 = x_2, \dots, X_7 = x_7) = \prod_{i=1}^7 P(X_i = x_i) = p^{\sum_{i=1}^7 x_i} (1-p)^{n - \sum_{i=1}^7 x_i}$$

Hàm hợp lý tương ứng:

$$L(p) = p^{\sum_{i=1}^7 x_i} (1-p)^{n - \sum_{i=1}^7 x_i}$$

Đồ thị biểu diễn $L(p)$ với $\sum x_i = 5, n=7$

```
> x<-c(0,1,0,1,1,1,1)
> curve(x^5*(1-x)^2, from=0, to=1, xlab="p", ylab="L(p)")
```



Sử dụng hàm *optimize* để tìm giá trị p trong khoảng $(0,1)$

```
> L<-function(p,x) prod(dbinom(x, size=1, prob=p))
> optimize(L, interval=c(0,1), x=x, maximum=TRUE)
$maximum
[1] 0.7142842

$objective
[1] 0.01517832
```

Giá trị $p=0.7142842$ hoàn toàn trùng khớp với giá trị trung bình mẫu:

```
> mean(x)
[1] 0.7142857
```

Điều này có nghĩa: $\hat{p} = \frac{1}{n} \sum_{i=1}^n X_i = \bar{X}$

Ví dụ 77: Xét dữ liệu PlantGrowth, 2012 đặc tính weight có phân phối chuẩn, chúng ta sẽ tìm các ước lượng $\hat{\mu}, \hat{\sigma}^2$ bằng ước lượng hợp lý cực đại.

Trong ví dụ này ta sẽ sử dụng gói stats4 với hàm *mle* để ước lượng $\hat{\mu}, \hat{\sigma}^2$ và xét $\mu = 5, \sigma^2 = 0.5$

```

> minuslogL<-function(mu,sigma2){
+ -sum(dnorm(x,mean=mu,sd=sqrt(sigma2),log=TRUE)) }
> x<-PlantGrowth$weight
> library(stats4)
> MaxLikeEst<-mle(minuslogL,start=list(mu=5,sigma2=0.5))
Warning messages:
1: In sqrt(sigma2) : NaNs produced
2: In sqrt(sigma2) : NaNs produced
> summary(MaxLikeEst)
Maximum likelihood estimation

Call:
mle(minuslogl = minuslogL, start = list(mu = 5, sigma2 = 0.5))

Coefficients:
            Estimate Std. Error
mu      5.0729848  0.1258666
sigma2 0.4752721  0.1227108

-2 log L: 62.82084

```

Kết quả này trùng khớp với lý thuyết.

7.5 ƯỚC LUỢNG KHOẢNG (INTERVAL ESTIMATOR)

7.5.1 Định nghĩa

Một ước lượng khoảng của một tham số θ là một cặp thống kê $L(X_1, X_2, \dots, X_n)$ và $U(X_1, X_2, \dots, X_n)$ của một mẫu ngẫu nhiên X thỏa $L(X) \leq \theta \leq U(X)$

Nếu X có mẫu thực nghiệm $x=(x_1, x_2, \dots, x_n)$, khi đó $[L(x), U(x)]$ gọi là một khoảng ước lượng cho θ .

7.5.2 Khoảng tin cậy

Xét biến ngẫu nhiên $X=(X_1, X_2, \dots, X_n)$ có hàm mật độ phụ thuộc đồng thời vào θ và $L(X), U(X)$ ($L(X) \leq U(X)$). Khi đó khoảng ngẫu nhiên $[L(X), U(X)]$ gọi là có độ tin cậy $100*(1-\alpha)\%$ nếu

$$P\{L(X) \leq \theta \leq U(X)\} = 1 - \alpha$$

Ý nghĩa:

Với 100 lần lấy mẫu cỡ n thì:

- Có $100(1-\alpha)$ lần có giá trị tham số $\theta \in [L(x), U(x)]$
- Có 100α lần có giá trị tham số $\theta \notin [L(x), U(x)]$

Thông thường, khoảng tin cậy cho θ có dạng:

$$(\hat{\theta}(x) - m(x), \hat{\theta}(x) + m(x)) \approx \hat{\theta}(x) \pm m(x)$$

ở đây, hai thống kê

$\hat{\theta}(x)$: là một ước lượng điểm (*point estimate*)

$m(x)$: là sai số ước lượng (*margin of error*)

7.5.3 Ước lượng khoảng cho giá trị trung bình

1. Trường hợp biết phương sai

Giả định:

- Biến cố ngẫu nhiên X có phân phối chuẩn, tức là $X \sim \mathcal{N}(\mu; \sigma^2)$
- Phương sai σ^2 đã biết.

Xây dựng khoảng tin cậy:

- Mẫu ngẫu nhiên cỡ n của biến ngẫu nhiên X là: X_1, X_2, \dots, X_n

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i, \text{ phân phối mẫu của } \bar{X} \sim \mathcal{N}(\mu; \sigma^2/n).$$

Đặt $Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ (a), thì $Z \sim \mathcal{N}(0, 1)$. Với độ tin cậy $100(1-\alpha)\%$, ta có:

$$P\left\{-z_{1-\alpha/2} \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z_{1-\alpha/2}\right\} = 1 - \alpha$$

hay $P\left\{\bar{X} - z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}\right\} = 1 - \alpha$

với $z_{1-\alpha/2}$ là phân vị mức $1 - \alpha/2$ của Z (một số tài liệu ký hiệu $z_{1-\alpha/2}$ là $z_{\alpha/2}$)

Nói cách khác, nếu \bar{X} là trung bình mẫu của một mẫu ngẫu nhiên cỡ n được chọn từ 1 quần thể có phương sai σ^2 đã biết, với độ tin cậy $100(1-\alpha)\%$ khoảng ước lượng cho giá trị trung bình (kỳ vọng μ) được xác định như sau:

$$\bar{X} - z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}$$

$$\text{hay } \bar{X} \pm z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}$$

Trong R, sử dụng hàm định lượng (*quantile*) để xác định giá trị $z_{1-\alpha/2}$

$$z_{1-\alpha/2} = z^* = qnorm(1 - \alpha/2, mean = 0, sd = 1)$$

Với $\alpha = 0.2, 0.1, 0.05, 0.001$ các giá trị z^* lần lượt là:

```
> alpha<-c(0.2, 0.1, 0.05, 0.001)
> zstar=qnorm(1-alpha/2)
> zstar
[1] 1.281552 1.644854 1.959964 3.290527
```

Chú ý: giá trị $z^* = 1.96$ ứng với $\alpha = 0.05$ hay độ tin cậy là 95%. Đảo lại, có thể nhận các giá trị alpha từ hàm *pnorm*:

```
> 2*(1-pnorm(zstar))
[1] 0.200 0.100 0.050 0.001
```

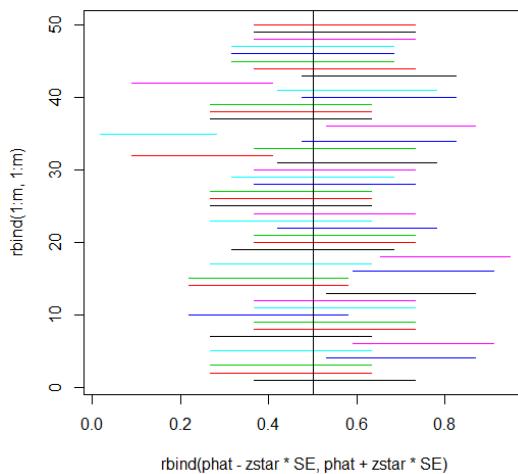
Ví dụ 78: Sử dụng dữ liệu cars.csv từ <http://pluto.huji.ac.il/~msby/StatThink/>

Ước lượng với độ tin cậy 0.95% cho giá trung bình của xe hơi:

```

> cars<-read.csv("d:/data/cars.csv")
> x.bar<-mean(cars$price,na.rm=TRUE)
> s<-sd(cars$price,na.rm=TRUE)
> n<-201
> x.bar-1.96*s/sqrt(n)
[1] 12108.47
> x.bar+1.96*s/sqrt(n)
[1] 14305.79

```



Nhận xét:

- Khi cỡ mẫu càng lớn, độ chính xác càng cao do sai số chuẩn càng nhỏ
 - Để khoảng tin cậy càng lớn, giá trị α càng nhỏ (khi đó z^* càng lớn).

Có bao nhiêu khoảng tin cậy 80% chứa \bar{X} ?

- Khoảng tin cậy không phải luôn luôn đúng: thực tế là không phải tất cả các khoảng tin cậy đều chứa giá trị đúng của tham số. Điều này có thể minh chứng bằng cách vẽ một số khoảng cách tin cậy ngẫu nhiên cùng một lúc để quan sát.

```

> m=50;n=20;p=0.5;          # gioe 20 dong xu 50 lan
> phat=rbinom(m,n,p)/n     # chia cho kích thuoc n
> SE=sqrt(phat*(1-phat)/n) # tinh sai so chuan
> alpha=0.10;zstar=qnorm(1-alpha/2)
> matplot(rbind(phat-zstar*SE,phat+zstar*SE),
+ rbind(1:m,1:m),type="l",lty=1)
> abline(v=p)

```

Ví dụ 79: Giả sử cân nặng của một chú bò trong đàn bò thí nghiệm được theo dõi trong 10 tháng là: 175 176 173 175 174 173 173 176 173 179

Giả sử độ lệch chuẩn $\sigma = 1.5$ và trọng lượng bò có phân phối chuẩn. Tính khoảng ước lượng giá trị trung bình trọng lượng với độ tin cậy 95%.

```

> x<-c(175,176,173,175,174,173,173,176,173,179)
> simple.z.test=function(x,sigma,conf.level=0.95) {
+ n=length(x);xbar=mean(x)
+ alpha=1-conf.level
+ zstar=qnorm(1-alpha/2)
+ SE=sigma/sqrt(n)
+ xbar+c(-zstar*SE,zstar*SE)
+ }
> simple.z.test(x,1.5)
[1] 173.7703 175.6297

```

Nghĩa là với độ tin cậy 95%, khoảng ước lượng giá trị trung bình trọng lượng là (173.7703, 175.6297).

2. Trường hợp không biết phương sai, mẫu lớn

Chúng ta có thể sử dụng z-khoảng (*z-interval*) để xác định khoảng tin cậy cho giá trị trung bình, trong trường hợp:

Giả định:

- Dữ liệu không có phân phối chuẩn, kích thước mẫu lớn ($n > 30$)
- Phương sai của quần thể chưa biết

Xây dựng khoảng tin cậy:

- Mẫu ngẫu nhiên cỡ n của biến ngẫu nhiên X là: X_1, X_2, \dots, X_n .

Thống kê trung bình mẫu và phương sai mẫu

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i, S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

Thay σ bởi S trong công thức (a) ở 7.5.3.1 thu được biến ngẫu nhiên:

$$T = \frac{\bar{X} - \mu}{S / \sqrt{n}}$$

T sẽ xấp xỉ với phân phối chuẩn $\mathcal{N}(0, 1)$ theo định lý giới hạn trung tâm (*Central Limit Theorem*). Do đó, khoảng tin cậy cho giá trị trung bình (kỳ vọng) μ với độ tin cậy $100(1-\alpha)\%$ cho bởi công thức

$$\bar{X} - z_{1-\alpha/2} \frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{1-\alpha/2} \frac{S}{\sqrt{n}}$$

3. Trường hợp không biết phương sai, mẫu nhỏ

Giả định:

- Dữ liệu có phân phối chuẩn $\mathcal{N}(\mu; \sigma^2)$, kích thước mẫu nhỏ ($n \leq 30$)
- Phương sai của quần thể chưa biết

Xây dựng khoảng tin cậy:

- Mẫu ngẫu nhiên cỡ n của biến ngẫu nhiên X là: X_1, X_2, \dots, X_n

Thống kê trung bình mẫu và phương sai mẫu

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i, S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

Thay σ bởi S trong công thức (a) ở 7.5.3.1 thu được biến ngẫu nhiên:

$$T = \frac{\bar{X} - \mu}{S / \sqrt{n}}$$

T là biến ngẫu nhiên có phân phối Student t với bậc $n-1$ tự do. Với độ tin cậy $100(1-\alpha)\%$, ta có:

$$P\left\{-t_{1-\alpha/2}^{n-1} \leq \frac{\bar{X} - \mu}{S / \sqrt{n}} \leq t_{1-\alpha/2}^{n-1}\right\} = 1 - \alpha, \text{ hay}$$

$$P\left\{\bar{X} - t_{1-\alpha/2}^{n-1} \frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{1-\alpha/2}^{n-1} \frac{S}{\sqrt{n}}\right\} = 1 - \alpha, \text{ với}$$

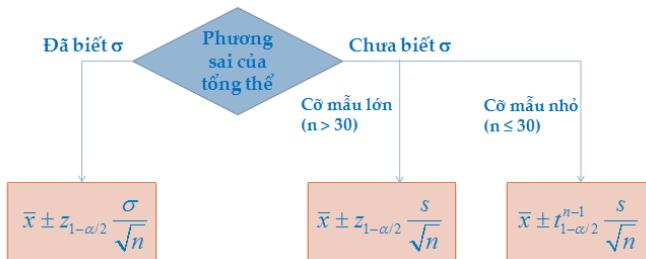
$t_{1-\alpha/2}^{n-1}$ là phân vị mức $1-\alpha/2$ của $T \sim t(n-1)$

Trong R, sử dụng hàm định lượng $qt(prob, df)$ (*quantile*) để xác định giá trị của $t_{1-\alpha}^{n-1} = qt(1-\alpha, n-1)$.

Ví dụ 80: Đối với chiều dài của 20 vi khuẩn hình que, chúng ta có giá trị trung bình $\bar{x} = 2.49$ và độ lệch chuẩn $s=0.674$. Xác định khoảng ước lượng với độ tin cậy 96%

```
> qt(0.98, 19)
[1] 2.204701
> LC<-2.490-2.204701*0.674/sqrt(20)
> UC<-2.490+2.204701*0.674/sqrt(20)
> LC;UC
[1] 2.157727
[1] 2.822273
```

Tóm tắt:



7.6 KHOẢNG TIN CẬY CHO TỶ LỆ TỔNG THỂ (POPULATION PROPORTION THEORY)

Bài toán: Tìm khoảng tin cậy cho p : tỷ lệ phần tử thỏa một tính chất \mathcal{A} của tổng thể.

- Lấy mẫu ngẫu nhiên cỡ n : $X=(X_1, X_2, \dots, X_n)$
- Đặt $Y=\text{số phần tử thỏa tính chất } \mathcal{A}$ trong n phần tử khảo sát thì $Y \sim B(n,p)$. Phân phối chuẩn sẽ được dùng như một xấp xỉ của phân phối nhị thức trong việc xây dựng khoảng tin cậy cho tỷ lệ tổng thể khi $n \geq 30$, $np \geq 5$ và $n(1-p) \geq 5$. Tuy nhiên, nhiều nhà thống kê toán đề nghị mẫu cỡ $n \geq 100$.
- Đặt $\hat{P} = \frac{Y}{n}$

- Biến ngẫu nhiên \hat{P} có kỳ vọng và phương sai lần lượt là:

$$E(\hat{P}) = p; \text{Var}(\hat{P}) = \sigma_{\hat{P}}^2 = \frac{p(1-p)}{n}$$

- Nếu cỡ mẫu n đủ lớn, theo định lý giới hạn trung tâm:

$$Z = \frac{\hat{P} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim \mathcal{N}(0, 1)$$

- Xây dựng độ tin cậy $100(1-\alpha)\%$ và $Z = \frac{\hat{P} - p}{\sqrt{\frac{p(1-p)}{n}}}$ ta có

$$P\left\{-z_{1-\alpha/2} \leq \frac{\hat{P} - p}{\sqrt{\frac{p(1-p)}{n}}} \leq z_{1-\alpha/2}\right\} = 1 - \alpha$$

Có nghĩa là

Nếu \hat{p} là tỷ lệ các phần tử thỏa một tính chất \mathcal{A} quan tâm của một mẫu ngẫu nhiên cỡ n, thì khoảng tin cậy với độ tin cậy $100(1-\alpha)\%$ cho tỷ lệ p các phần tử thỏa tính chất \mathcal{A} của tổng thể là:

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

Với $z_{1-\alpha/2}$ là phân vị mức $1-\alpha/2$ của $Z \sim \mathcal{N}(0, 1)$.

Ví dụ 81: Trong một đợt điều tra về nha khoa tại một địa phương, kiểm tra ($n=100$) em người ta thấy có 36 trẻ em bị sâu răng. Hãy tìm khoảng tin cậy 99% cho tỷ lệ trẻ bị sâu răng tại địa phương đó.

Giai: Gọi \hat{p} là tỷ lệ trẻ bị sâu răng tại địa phương đang khảo sát.

Giá trị tỷ lệ trẻ em bị sâu răng trên mẫu: $\hat{p} = 0.36$.

Do $n \hat{p} = 36 \geq 5$ và $n(1-\hat{p}) = 64 > 5$, nên khoảng tin cậy 99% cho p được viết trong R là

```
> alpha=1-0.99
> zstar<-qnorm(1-alpha/2)
> p<-0.36
> n<-100
> vr<-sqrt(p*(1-p)/n)
> LC<-p-zstar*vr
> UC<-p+zstar*vr
> LC;UC
[1] 0.2363602
[1] 0.4836398
```

Khoảng tin cậy 99% cho tỷ lệ trẻ bị sâu răng tại địa phương đó là

(0.2363602, 0.4836398).

Chú ý: khi mẫu có kích thước nhỏ không thể áp dụng ước lượng trên, người ta phải tính các khoảng tin cậy bằng phân phối nhị thức. Tuy nhiên, nếu khoảng ước lượng quá rộng ít có giá trị ứng dụng.

7.7 XÁC ĐỊNH KÍCH THƯỚC MẪU

Trong các bài toán ước lượng, chất lượng ước lượng được phản ánh qua độ tin cậy và sai số cho phép. Sai số này lại phụ thuộc vào kích thước mẫu và độ tin cậy. Bài toán đặt ra là: Để đạt được độ tin cậy $\gamma\% = (1-\alpha)$ và sai số tối đa cho phép là ε , thì kích thước mẫu là bao nhiêu?

Phương pháp giải quyết là dựa vào các công thức sai số của ước lượng và độ tin cậy.

- Trong ước lượng khoảng cho trung bình:

- Trường hợp biết phương sai:

$$\varepsilon = z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \Rightarrow n = \left\lceil \left[\frac{z_{1-\alpha/2} * \sigma}{\varepsilon} \right]^2 \right\rceil$$

- Trường hợp không biết phương sai, mẫu lớn

$$\varepsilon = z_{1-\alpha/2} \frac{S}{\sqrt{n}} \Rightarrow n = \left\lceil \left[\frac{z_{1-\alpha/2} * S}{\varepsilon} \right]^2 \right\rceil$$

- Trường hợp không biết phương sai, mẫu nhỏ

$$\varepsilon = t_{1-\alpha/2}^{n-1} \frac{S}{\sqrt{n}} \Rightarrow n = \left\lceil \left[\frac{t_{1-\alpha/2}^{n-1} * S}{\varepsilon} \right]^2 \right\rceil$$

- Trong ước lượng khoảng cho tỷ lệ tổng thể ta có:

$$\varepsilon = z_{1-\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \Rightarrow n = \left(\frac{z_{1-\alpha/2}}{\varepsilon} \right)^2 \hat{p}(1-\hat{p})$$

(Ký hiệu $\lceil x \rceil$ là số nguyên nhỏ nhất lớn hơn hoặc bằng x)

Để có các số liệu thống kê lần đầu, người ta thường chọn kích thước mẫu thăm dò là số nguyên n_1 thỏa:

$$n_1 \geq \left(\frac{z_{1-\alpha/2}}{2\varepsilon} \right)^2$$

$\hat{p}(1-\hat{p})$ được xác định bởi giá trị lớn nhất trong (0,1) là 1/4.

TÀI LIỆU THAM KHẢO

- [1] Biostatistics with R, Babak Shahbaba, Springer, 2011
- [2] Introduction to Probability and Statistics Using R, G.Jay Kerns, First Edition, cran.r-project.org/web/packages/IPSUR/vignettes/IPSUR.pdf, 2010
- [3] Introduction to Statistical Thinking (With R, Without Calculus), Benjamin, The Hebrew University, 2011.

- [4] Lý thuyết Xác suất và thống kê (bài giảng), Hoàng Văn Hà, ĐH KHTN Tp HCM.
- [5]Simple-Using R for Introductory Statistics, John Verzani, 2001,
<https://cran.r-project.org/doc/contrib/Verzani-SimpleR.pdf>
- [6] Thống kê ứng dụng trong kinh tế-xã hội, Hoàng Trong, Chu Nguyễn Mộng Ngọc, NXB Lao động-Xã hội, 2010.
- [7]Topic 16:Interval Estimation,University of Arizona,
math.arizona.edu/~jwatkins/p-interval.pdf

C H UƠNG

8

KIỂM ĐỊNH GIẢ THUYẾT THỐNG KÊ

8.1 GIỚI THIỆU

8.1.1 Các khái niệm cơ bản

a. Giả thuyết thống kê

Giả thuyết thống kê là những phát biểu về tham số, quy luật phân phối hoặc tính độc lập của các đại lượng ngẫu nhiên. Việc tìm ra kết luận để bác bỏ hoặc chấp nhận một giả thuyết gọi là kiểm định giả thuyết thống kê.

b. Giả thuyết vô hiệu (*null hypothesis*) và đối thuyết (*alternative hypothesis*)

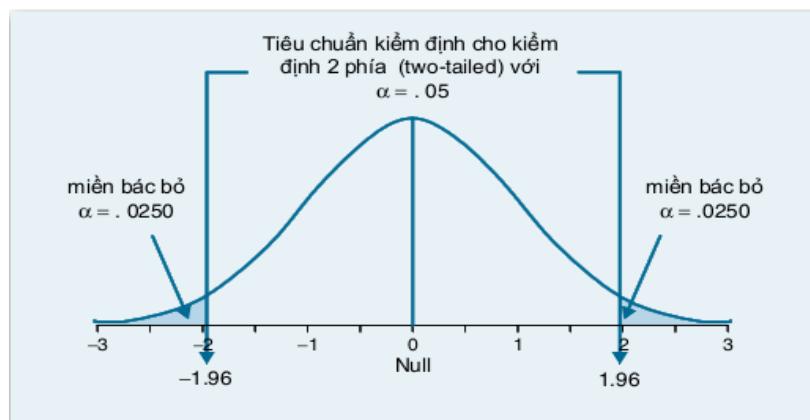
- *Giả thuyết vô hiệu* (H_0): là giả thuyết được đặt ra với ý đồ bác bỏ nó, ngược lại với điều nhà nghiên cứu muốn chứng minh, muốn thuyết phục.
- *Đối thuyết* (H_1): là mệnh đề đối lập với H_0 .

c. Cách đặt giả thuyết

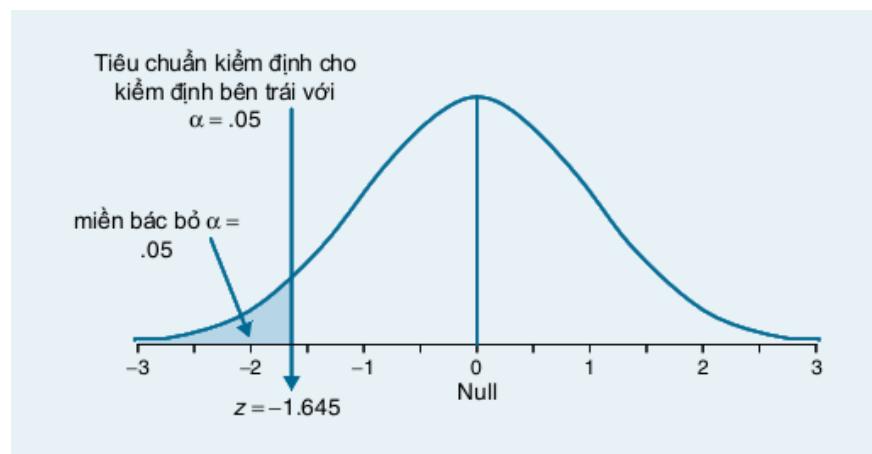
Một bài toán kiểm định giả thuyết cho tham số θ sẽ có 3 dạng:

- Hai phía (*two-tailed test*)

$$\begin{cases} H_0 : \theta = \theta_0 \\ H_1 : \theta \neq \theta_0 \end{cases}$$



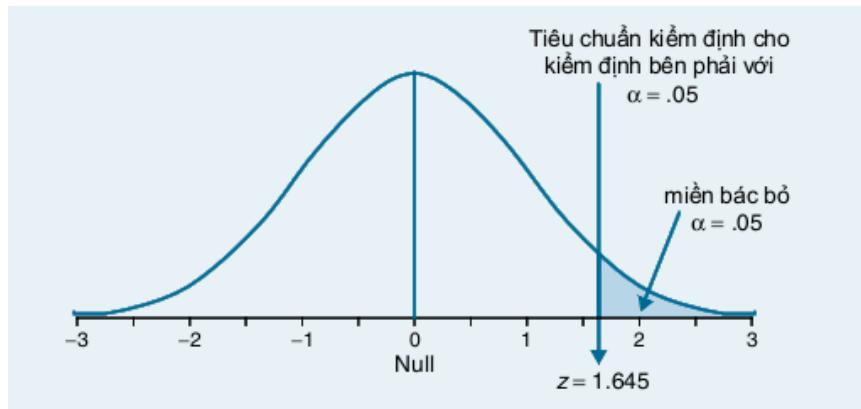
- Một phía (*one-tailed test*)
 - Một phía bên trái (*lower-tail*)



$$\begin{cases} H_0 : \theta \geq \theta_0 \\ H_1 : \theta < \theta_0 \end{cases}$$

▪ Một phía bên phải (*upper-tail*)

$$\begin{cases} H_0 : \theta \leq \theta_0 \\ H_1 : \theta > \theta_0 \end{cases}$$



d. Miền bác bỏ - Tiêu chuẩn kiểm định (*Rejection region-critical value*)

- Xét bài toán kiểm định giả thuyết có giả thuyết H_0 và đối thuyết H_1 . Giả sử H_0 đúng, từ mẫu ngẫu nhiên $X = (X_1, X_2, \dots, X_n)$ chọn hàm

$$Z = h(X_1, X_2, \dots, X_n; \theta_0)$$

sao cho với số α đủ bé ta tìm được tập W_α thỏa

$$P(Z \in W_\alpha) = \alpha$$

Trong đó:

- tập hợp W_α : là miền bác bỏ giả thuyết H_0
- đại lượng $Z = h(X_1, X_2, \dots, X_n; \theta_0)$: là tiêu chuẩn kiểm định giả thuyết H_0

- giá trị α gọi là mức ý nghĩa của bài toán kiểm định.
- Nếu từ mẫu thực nghiệm (x_1, x_2, \dots, x_n) của X ta tính được giá trị của Z là $z = h(x_1, x_2, \dots, x_n; \theta_0)$:

Nếu $z \in W_\alpha$: giả thuyết H_0 bị bác bỏ

Nếu $z \notin W_\alpha$: giả thuyết H_0 chưa đủ cơ sở để bác bỏ

8.1.2 Sai lầm loại I và loại II

Trong bài toán kiểm định giả thuyết, ta có thể mắc phải các sai lầm sau:

- *Sai lầm loại I*: bác bỏ H_0 khi thực tế H_0 đúng. Sai lầm loại I ký hiệu α (mức ý nghĩa của kiểm định)

$$\alpha = P(W_\alpha | H_0)$$

- *Sai lầm loại II*: chấp nhận H_0 khi thực tế H_0 sai. Sai lầm loại II ký hiệu β .

$$\beta = P(\bar{W}_\alpha | H_0)$$

Quyết định	Thực tế	
Bác bỏ H_0	Sai lầm loại I: α	Không có sai lầm: $(1-\beta)$
Không bác bỏ H_0	Không có sai lầm: $(1-\alpha)$	Sai lầm loại II: β

8.1.3 Số liệu để ước tính cỡ mẫu

Trong chương trước có đề cập đến ước tính cỡ mẫu phụ thuộc vào sai số của ước lượng. Việc ước tính cỡ mẫu còn được xác định phụ thuộc vào xác suất sai lầm loại I và xác suất sai lầm loại II. Ước tính này phụ thuộc các tham số:

- Xác suất sai lầm, thông thường một kết quả thống kê chấp nhận sai lầm loại I khoảng 1% hay 5% (tức $\alpha = 0.01$ hay $\alpha = 0.05$) và sai lầm loại II khoảng từ ($\beta =$) 0.1 đến ($\beta =$) 0.2 tương ứng với đại lượng power ($=1-\beta$) từ 0.8 đến 0.9.
- Độ lệch chuẩn (*standard deviation*): σ
- Độ ảnh hưởng Δ : giá trị sai số của yếu tố thống kê cần xử lý hay độ khác biệt của yếu tố thống kê giữa hai nhóm nghiên cứu.

Công thức ước tính cỡ mẫu được xác định:

a. Trường hợp xử lý thống kê trên 1 nhóm

$$n = \frac{C}{(\Delta/\sigma)^2}$$

b. Trường hợp xử lý thống kê trên 2 nhóm

$$n = 2 * \frac{C}{(\Delta/\sigma)^2}$$

Trong 2 công thức trên, hằng số C được cho bởi bảng sau:

α (sig.level)	$\beta=0.20$ (power=0.80)	$\beta=0.10$ (power=0.90)	$\beta=0.05$ (power=0.95)
0.10	6.15	8.53	10.79
0.05	7.85	10.51	13.00
0.01	13.33	16.74	19.84

Trong ngôn ngữ R có hàm *power.t.test()* để ước tính cỡ mẫu.

Cú pháp: *power.t.test(delta= ,sd=, sig.level=, power=, type=)*

Ví dụ 82: Muốn ước tính chiều cao đàn ông Việt Nam, chấp nhận sai số ước tính là 1cm ($\Delta = 1$), với khoảng tin cậy 0.95 ($\alpha = 0.05$) và sai lầm loại II là ($\beta =$) 0.20 hay power= 0.8. Các nghiên cứu trước cho biết độ lệch chuẩn chiều cao đàn ông Việt Nam là ($\sigma =$) 4.6 cm. Ước tính cỡ mẫu cần thiết cho nghiên cứu trên.

Thực hiện bằng ngôn ngữ R:

```
> power.t.test(delta=1,sd=4.6,sig.level=.05,power=.8,type='one.sample')

One-sample t test power calculation

      n = 168.0131
    delta = 1
      sd = 4.6
    sig.level = 0.05
      power = 0.8
  alternative = two.sided
```

Như vậy cỡ mẫu cần thiết là 168.

8.1.4 Phương pháp tiếp cận chấp nhận hay bác bỏ một giả thuyết thống kê

Có 2 phương pháp chính trong thực tế để tiếp cận việc chấp nhận hay bác bỏ một giả thiết thống kê [6]:

Phương pháp kiểm định ý nghĩa thống kê (*Test of significance*) do Ronald A.Fisher đề xuất. Phương pháp này được tiến hành theo 3 bước:

1. Phát biểu một giả thuyết vô hiệu (*null hypothesis*).
2. Thu thập dữ liệu liên quan đến giả thuyết. Gọi dữ liệu là D.
3. Ước tính xác suất quan sát dữ liệu D nếu giả thuyết H_0 đúng. Nói cách khác là tính $P(D|H_0)$, đây chính là trị số P (**P-value**).

Nếu giá trị $P < \alpha$ (α : mức ý nghĩa), thì giả thuyết H_0 không phù hợp với dữ liệu thu thập. (P-giá trị: là mức ý nghĩa nhỏ nhất để bác bỏ H_0)

Phương pháp kiểm định giả thuyết (*Test of hypothesis*) do J.Neyman và K.Pearson đề xuất. Phương pháp này được tiến hành theo 4 bước:

1. Phát biểu một giả thuyết vô hiệu (H_0) và một đối thuyết (H_1) (*alternative hypothesis*)
2. Quyết định mức độ a và b có thể chấp nhận được và ước tính cỡ mẫu cần của giả thuyết. a là xác suất bác bỏ giả thuyết H_1 nhưng đó là giả thuyết đúng, b là xác suất bác bỏ giả thuyết H_0 nhưng giả thuyết H_0 đúng.
3. Thu thập dữ liệu liên quan đến giả thuyết
4. Nếu dữ liệu nằm trong khoảng bác bỏ giả thuyết H_0 thì chấp nhận đối thuyết H_1 , ngược lại thì chấp nhận giả thuyết H_0 bác bỏ đối thuyết H_1 . *Chấp nhận giả thuyết không có nghĩa là tin vào giả thuyết đó, mà chỉ có nghĩa là giả thuyết đó hợp lý với dữ liệu thu thập được.*

Ngoài 2 phương pháp trên còn có **phương pháp hỗn hợp**:

1. Phát biểu một giả thuyết vô hiệu (H_0) và đối thuyết (H_1)
2. Xác định xác suất α (sai số loại I) và β (số loại II), ước tính cỡ mẫu
3. Thu thập dữ liệu liên quan đến giả thuyết. Gọi dữ liệu là D.
4. Xác định giá trị $P=P(D|H_0)$
5. Nếu $P<\alpha$ (α : mức ý nghĩa), bác bỏ giả thuyết H_0 .

8.2 KIỂM ĐỊNH GIẢ THUYẾT CHO TRƯỜNG HỢP 1 MẪU

8.2.1 Kiểm định giả thuyết cho kỳ vọng

8.2.1.1 Trường hợp biết phương sai

Các giả định:

- Mẫu ngẫu nhiên X_1, X_2, \dots, X_n được chọn từ tổng thể có phân phối chuẩn $\mathcal{N}(\mu; \sigma^2)$, với kỳ vọng μ chưa biết.
- Phương sai σ^2 đã biết
- Cho trước giá trị μ_0 , cần so sánh kỳ vọng μ với μ_0 với mức ý nghĩa α cho trước.
- Xét 1 trong 3 trường hợp kiểm định nêu ở 8.1.1 c

Các bước kiểm định

- Phát biểu giả thuyết H_0 và đối thuyết H_1
- Xác định mức ý nghĩa α
- Lấy mẫu ngẫu nhiên cỡ n: X_1, X_2, \dots, X_n và tính thống kê

$$Z_0 = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}$$

- Xác định miền bác bỏ W_α theo bảng

Giả thuyết	Miền bác bỏ
$H_0: \mu = \mu_0$ $H_1: \mu \neq \mu_0$	$W_\alpha = \{z_0 : z_0 > z_{1-\alpha/2}\}$

$H_0 : \mu \geq \mu_0$	$W_\alpha = \{z_0 : z_0 < -z_{1-\alpha}\}$
$H_0 : \mu \leq \mu_0$	$W_\alpha = \{z_0 : z_0 > z_{1-\alpha}\}$

Trong ngôn ngữ R, để tính $z_{1-\alpha/2}$ sử dụng hàm

$$qnorm(1-\alpha/2, mean=0, sd=1)$$

Hoặc sử dụng p-giá trị để bác bỏ H_0 khi p-giá trị $\leq \alpha$. Công thức p-giá trị theo bảng

Giả thuyết	p-giá trị
$H_0 : \mu = \mu_0$	
$H_1 : \mu \neq \mu_0$	$p = 2[1 - \varphi(z_0)]$
$H_0 : \mu \geq \mu_0$	
$H_1 : \mu < \mu_0$	$p = \varphi(z_0)$
$H_0 : \mu \leq \mu_0$	
$H_1 : \mu > \mu_0$	$p = 1 - \varphi(z_0)$

Với $\varphi(z)$ là hàm Laplace: $\varphi(z) = \frac{1}{\sqrt{2\pi}} \int_0^z e^{-t^2/2} dt$

Trong ngôn ngữ R, $\varphi(z) = pnorm(z, mean=0, sd=1)$

Ví dụ 83: (*kiểm định 2 phía*) Dây chuyền sản xuất kem đánh răng P/S được thiết kế để đóng hộp những tuýp kem có trọng lượng trung bình là 6 oz (1oz=28g). Một mẫu gồm 30 tuýp kem được chọn ngẫu nhiên để kiểm tra định kỳ. Nếu dây chuyền không đảm bảo 1 tuýp kem là 6 oz (nhiều hơn hoặc ít hơn) thì dây chuyền phải điều chỉnh lại. Giả sử trung bình mẫu của 30 tuýp kem là 6.1 oz và độ lệch chuẩn của tổng thể là $\sigma = 0.2$ oz. Thực hiện kiểm định giả thuyết thống kê với mức ý nghĩa 3% để xét xem dây chuyền có vận hành tốt không?

Giải:

Gọi X là trọng lượng 1 tuýp kem đánh răng, $X \sim \mathcal{N}(\mu; 0.2^2)$. Các bước kiểm định:

1. Phát biểu giả thuyết:

$$\begin{aligned} H_0 &: \mu = 6 \\ H_1 &: \mu \neq 6 \end{aligned}$$

2. Xác định mức ý nghĩa: $\alpha = 0.03$

3. Tính giá trị thống kê kiểm định

$$z_0 = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} = \frac{6.1 - 6.0}{0.2 / \sqrt{30}} = 2.74$$

4. Xác định miền bác bỏ

Từ $\alpha = 0.03$ nên $z_{1-\alpha/2} = 2.17$. Vậy H_0 bác bỏ nếu

$$z_0 < -2.17 \text{ hoặc } z_0 > 2.17$$

5. Kết luận: Do $z_0 = 2.74 > 2.17$ nên bác bỏ H_0 . Với 97% độ tin cậy trọng lượng 1 tuýp kem không bằng 6 oz.

Sử dụng p-giá trị

Tính p-giá trị bài toán kiểm định 2 phía

$$p = 2[1 - \varphi(|z_0|)] = 2[1 - \varphi(2.74)] = 2[1 - 0.9969] = 0.0062$$

Với $\alpha = 0.03$, $p=0.0062 < 0.03$ nên bác bỏ H_0 . Với 97% độ tin cậy trọng lượng 1 tuýp kem không bằng 6 oz.

Thực hiện bằng ngôn ngữ R:

```
> alpha<-0.03
> xbar<-6.1;muy0<-6.0
> sdev<-0.2;n<-30
> z0<-(xbar-muy0)/(sdev/sqrt(n))
> hs<-1-alpha/2
> zhs<-qnorm(hs,mean=0,sd=1)
> z0
[1] 2.738613
> zhs
[1] 2.17009
> phiz0<-pnorm(z0,mean=0,sd=1)
> p<-2*(1-phiz0)
> p
[1] 0.006169899
```

Ví dụ 84: (kiểm định một phía) **Metro EMS**: Một bệnh viện tại trung tâm thành phố cung cấp dịch vụ cấp cứu tại nhà. Với khoảng 20 xe cấp cứu, mục tiêu của trung tâm là cung cấp dịch vụ cấp cứu trong khoảng thời gian trung bình là 12 phút sau khi nhận được điện thoại yêu cầu. Một mẫu ngẫu nhiên gồm thời gian đáp ứng khi có yêu cầu của 40 ca cấp cứu được chọn. Trung bình mẫu là 13.25 phút. Biết rằng độ lệch chuẩn của tổng thể là $\sigma = 3.2$ phút. Giám đốc EMS muốn thực hiện kiểm định với mức ý nghĩa 5%, để xác định xem liệu thời gian một ca cấp cứu có bé hơn hoặc bằng 12 phút hay không?

Giải:

Các bước kiểm định

1. Phát biểu giả thuyết

$$H_0 : \mu \leq 12$$

$$H_1 : \mu > 12$$

2. Xác định mức ý nghĩa: $\alpha = 0.05$

3. Tính giá trị thống kê kiểm định

$$z_0 = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{13.25 - 12}{3.2/\sqrt{40}} = 2.47$$

4. Xác định miền bác bỏ

Từ $\alpha = 0.05$ nên $z_{1-\alpha/2} = 1.645$. Vậy H_0 bác bỏ vì

$$z > z_0 > 1.645$$

5. Kết luận: Với 95% độ tin cậy Metro EMS không đáp ứng mục tiêu thời gian.

Sử dụng p-giá trị

Tính p-giá trị bài toán kiểm định 1 phía

$$p = 1 - \varphi(z_0) = 1 - \varphi(2.47) = 1 - 0.9932 = 0.0068 \text{ theo là } \sigma = 3.2 \text{ phút.}$$

Do $p < 0.05$ nên với 95% độ tin cậy Metro EMS không đáp ứng mục tiêu thời gian.

Thực hiện bằng ngôn ngữ R:

```
> alpha<-0.05
> xbar<-13.25; muy0<-12
> sdev<-3.2; n<-40
> z0<-(xbar-muy0)/(sdev/sqrt(40))
> hs<-1-alpha
> zhs<-qnorm(hs,mean=0,sd=1)
> z0
[1] 2.470529
> zhs
[1] 1.644854
> phiz0<-pnorm(z0)
> p<-1-phiz0
> p
[1] 0.006745661
> alpha
[1] 0.05
```

8.2.1.2 Trường hợp không biết phương sai, mẫu nhỏ

Các giả định:

- Mẫu ngẫu nhiên X_1, X_2, \dots, X_n được chọn từ tổng thể có phân phối chuẩn $\mathcal{N}(\mu; \sigma^2)$, với kỳ vọng μ và σ^2 chưa biết.
- Sử dụng ước lượng không chêch S thay cho σ
- Cỡ mẫu nhỏ: $n \leq 30$, cần kiểm định kỳ vọng μ_0 với mức ý nghĩa α cho trước.
- Xét 1 trong 3 trường hợp kiểm định nêu ở 8.1.1 c

Các bước kiểm định

- Phát biểu giả thuyết H_0 và đối thuyết H_1
- Xác định mức ý nghĩa α
- Lấy mẫu ngẫu nhiên cỡ n: X_1, X_2, \dots, X_n và tính thống kê

$$T_0 = \frac{\bar{X} - \mu_0}{S / \sqrt{n}}$$

Biến ngẫu nhiên T_0 có phân phối Student với $n-1$ bậc tự do

- Xác định miền bác bỏ W_α theo bảng

Giả thuyết	Miền bác bỏ
$H_0 : \mu = \mu_0$ $H_1 : \mu \neq \mu_0$	$W_\alpha = \{t_0 : t_0 > t_{1-\alpha/2}^{n-1}\}$
$H_0 : \mu \geq \mu_0$ $H_1 : \mu < \mu_0$	$W_\alpha = \{t_0 : t_0 < -t_{1-\alpha}^{n-1}\}$

$H_0 : \mu \leq \mu_0$	$W_\alpha = \{t_0 : t_0 > t_{1-\alpha}^{n-1}\}$
$H_1 : \mu > \mu_0$	

Trong ngôn ngữ R để tính $t_{1-\alpha}^{n-1}$ dùng hàm: hàm định lượng $qt(prob, df)$ (*quantile*) để xác định giá trị của $t_{1-\alpha}^{n-1} = qt(1-\alpha, n-1)$.

- Sử dụng p-giá trị để bác bỏ H_0 khi p-giá trị $\leq \alpha$, với α mức ý nghĩa cho trước.
- Công thức tính p-giá trị

Giả thuyết	p-giá trị
$H_0 : \mu = \mu_0$	
$H_1 : \mu \neq \mu_0$	$p = 2P(T_{n-1} \geq t_0)$
$H_0 : \mu \geq \mu_0$	
$H_1 : \mu < \mu_0$	$p = P(T_{n-1} \leq t_0)$
$H_0 : \mu \leq \mu_0$	
$H_1 : \mu > \mu_0$	$p = P(T_{n-1} \geq t_0)$

Trong ngôn ngữ R, để tính $P(T_{n-1} \leq t_0)$ sử dụng hàm $pt(t_0, n-1)$

Ví dụ 85: Điều tra Cholesterol toàn phần trong huyết thanh của 25 bệnh nhân bị một loại bệnh B, ta có trung bình cộng lượng Cholesterol là 172mg% và độ lệch chuẩn bằng 40mg%. Theo tài liệu về hằng số sinh hóa bình thường của người Việt Nam thì lượng Cholesterol trung bình toàn phần trong huyết thanh là 156mg% và tuân theo luật phân phối chuẩn.

Hỏi lượng Cholesterol của các bệnh nhân mắc bệnh B có cao hơn bình thường không? (Kết luận ở mức $\alpha = 5\%$)

Giải:

1. Phát biểu giả thuyết

$$\begin{aligned} H_0 &: \mu = 156 \\ H_1 &: \mu > 156 \end{aligned}$$

2. Xác định mức ý nghĩa: $\alpha = 0.05$
3. Tính giá trị thống kê kiểm định

$$T = \frac{\bar{X} - 156}{S} \sqrt{25} \sim t(24)$$

4. Xác định miền bác bỏ

Từ $\alpha = 0.05$ nên $t_{1-\alpha}^{24} = 1.7109$, với mẫu cụ thể tính được

$$t_0 = \frac{172 - 156}{40} \sqrt{25} = 2 > 1.7109$$

Kết luận: Với 95% độ tin cậy, bệnh nhân mắc bệnh B lượng Cholesterol cao hơn bình thường.

Sử dụng p-giá trị

Tính p-giá trị bài toán kiểm định 1 phía

$$p = P(T_{24} \geq t_0) = 0.0285$$

Vì $p < \alpha = 0.05$. Với 95% độ tin cậy, bệnh nhân mắc bệnh B có lượng Cholesterol cao hơn bình thường.

Thực hiện bằng ngôn ngữ R:

```
> alpha<-0.05
> Xbar<-172
> muy<-156
> sde<-40; n<-25
> t0<-(Xbar-muy)/40*sqrt(n)
> t1_alpha<-qt(1-alpha,n-1)
> t0; t1_alpha
[1] 2
[1] 1.710882
> p_value<-1-pt(t0,n-1)
> p_value; alpha
[1] 0.02846992
[1] 0.05
```

8.2.1.3 Trường hợp không biết phương sai, mẫu lớn

Các giả định:

- Mẫu ngẫu nhiên X_1, X_2, \dots, X_n được chọn từ tổng thể có phân phối chuẩn $\mathcal{N}(\mu; \sigma^2)$, với kỳ vọng μ, σ^2 chưa biết.
- Sử dụng ước lượng không chêch s thay cho σ
- Cho trước giá trị μ_0 , cần so sánh kỳ vọng μ với μ_0 với mức ý nghĩa α cho trước.
- Xét 1 trong 3 trường hợp kiểm định nêu ở 8.1.1 b

Các bước kiểm định

- Phát biểu giả thuyết H_0 và đối thuyết H_1
- Xác định mức ý nghĩa α
- Lấy mẫu ngẫu nhiên cỡ n: X_1, X_2, \dots, X_n và tính thống kê

$$Z_0 = \frac{\bar{X} - \mu_0}{S / \sqrt{n}}$$

- Kết luận bác bỏ H_0 hay không tương tự như trường hợp biết phương sai (8.2.1.1)

Ví dụ 86: Trạm cảnh sát giao thông trên đường cao tốc sẽ thực hiện việc bắn tốc độ định kỳ tại các địa điểm khác nhau để kiểm tra tốc độ các phương tiện giao thông. Một mẫu tốc độ của các loại xe được chọn để kiểm định giả thuyết sau:

$$\begin{aligned} H_0 : \mu &= 65 \\ H_1 : \mu &> 65 \end{aligned}$$

Những vị trí mà bác bỏ H_0 là những vị trí tốt nhất được chọn để đặt rada kiểm soát tốc độ. Tại địa điểm F, một mẫu gồm 64 phương tiện được bắn tốc độ ngẫu nhiên có trung bình là 66.2m/ph và có độ lệch chuẩn 4.2m/ph. Sử dụng $\alpha = 5\%$ để kiểm định giả thuyết.

Giải:

Các bước kiểm định

1. Phát biểu giả thuyết $H_0 : \mu = 65$
 $H_1 : \mu > 65$

2. Xác định mức ý nghĩa: $\alpha = 0.05$
3. Tính giá trị thống kê kiểm định

$$z_0 = \frac{\bar{x} - \mu_0}{S / \sqrt{n}} = \frac{66.2 - 65}{4.2 / \sqrt{64}} = 2.286$$

4. Xác định miền bác bỏ

Từ $\alpha = 0.05$ nên $z_{1-\alpha/2} = 1.645$. Vậy H_0 bị bác bỏ vì

$$z > z_0 > 1.645$$

5. Kết luận: Với 95% độ tin cậy giả thuyết H_0 bị bác bỏ.

Sử dụng p-giá trị

Tính p-giá trị bài toán kiểm định 1 phía

$$p = 1 - \varphi(z_0) = 1 - \varphi(2.286) = 0.0111$$

Do $p < 0.05$ nên với 95% độ tin cậy giả thuyết H_0 bị bác bỏ.

Hãy thực hiện ví dụ trên bằng ngôn ngữ R.

8.2.2 Kiểm định cho phương sai

Các giả định:

- Mẫu ngẫu nhiên X_1, X_2, \dots, X_n được chọn từ tổng thể có phân phối chuẩn $\mathcal{N}(\mu; \sigma^2)$, với phương sai σ^2 chưa biết.
- Cho trước giá trị σ_0^2 , cần so sánh phương sai σ^2 với σ_0^2 với mức ý nghĩa α cho trước.
- Xét 3 trường hợp kiểm định
 - + Hai phía (*two-tailed test*)

$$\begin{cases} H_0 : \sigma^2 = \sigma_0^2 \\ H_1 : \sigma^2 \neq \sigma_0^2 \end{cases}$$

- + Một phía (*one-tailed test*)
 - Một phía bên trái (*lower-tail*)

$$\begin{cases} H_0 : \sigma^2 \geq \sigma_0^2 \\ H_1 : \sigma^2 < \sigma_0^2 \end{cases}$$

- Một phía bên phải (*upper-tail*)

$$\begin{cases} H_0 : \sigma^2 \leq \sigma_0^2 \\ H_1 : \sigma^2 > \sigma_0^2 \end{cases}$$

Các bước kiểm định

- Phát biểu giả thuyết H_0 và đối thuyết H_1
- Xác định mức ý nghĩa α
- Lấy mẫu ngẫu nhiên cỡ n: X_1, X_2, \dots, X_n và tính thống kê

$$K_0 = \frac{(n-1)S^2}{\sigma_0^2} \sim \chi^2(n-1)$$

- Miền bác bỏ:

Giả thuyết	Miền bác bỏ
$\begin{cases} H_0 : \sigma^2 = \sigma_0^2 \\ H_1 : \sigma^2 \neq \sigma_0^2 \end{cases}$	$W_\alpha = \{k_0 : 0 < k_0 < k_{\alpha/2} \vee k_{1-\alpha/2} < k_0\}$
$\begin{cases} H_0 : \sigma^2 \geq \sigma_0^2 \\ H_1 : \sigma^2 < \sigma_0^2 \end{cases}$	$W_\alpha = \{k_0 : k_0 < k_\alpha\}$
$\begin{cases} H_0 : \sigma^2 \leq \sigma_0^2 \\ H_1 : \sigma^2 > \sigma_0^2 \end{cases}$	$W_\alpha = \{k_0 : k_0 > k_{1-\alpha}\}$

Trong ngôn ngữ R, để tính $k_{1-\alpha/2}$ sử dụng hàm

$$qchisq(1 - \alpha/2, n-1)$$

Ví dụ 87: Khi máy hoạt động bình thường, trọng lượng sản phẩm X phân phối chuẩn với $\sigma_0^2 = 0.12$. Nghi ngờ máy hoạt động không ổn định, người ta cân thử 25 sản phẩm và tính được $s^2 = 0.13$. Với mức ý nghĩa $\alpha = 0.05$, hãy kết luận về điều nghi ngờ trên.

Giải:

Các bước kiểm định

1. Phát biểu giả thuyết

$$\begin{aligned} H_0 &: \sigma^2 = 0.12 \\ H_1 &: \sigma^2 \neq 0.12 \end{aligned}$$

2. Xác định mức ý nghĩa: $\alpha = 0.05$

3. Tính giá trị thống kê kiểm định

$$k_0 = \frac{(n-1)s^2}{\sigma_0^2} = \frac{24 * 0.13}{0.12} = 26$$

4. Xác định miền bác bỏ

Từ $\alpha = 0.05$, tìm $K_{\alpha/2} = \chi^2_{\alpha/2} = 12.401$, $K_{1-\alpha/2} = \chi^2_{1-\alpha/2} = 39.3641$.

Vậy H_0 bác bỏ vì $k_0 = 26 \notin W_\alpha$

5. Kết luận: Với 95% độ tin cậy giả thuyết H_0 không bị bác bỏ: xem như máy móc hoạt động bình thường.

Thực hiện bằng ngôn ngữ R:

```

> alpha<-0.05
> sd02<-0.12; n<-25
> sd2<-0.13
> k0<-(n-1)*sd2/sd02
> hs1<-alpha/2
> hs2<-1-alpha/2
> khs1<-qchisq(hs1,n-1)
> khs2<-qchisq(hs2,n-1)
> k0;khs1;khs2
[1] 26
[1] 12.40115
[1] 39.36408

```

8.2.3 Kiểm định giả thuyết cho tỷ lệ

Bài toán: Cho tổng thể X, trong đó tỷ lệ phần tử thỏa một tính chất \mathcal{A} của tổng thể là p (p chưa biết). Từ mẫu ngẫu nhiên cỡ n: $X=(X_1, X_2, \dots, X_n)$. Hãy kiểm định:

+ Hai phía (*two-tailed test*)

$$\begin{cases} H_0 : p = p_0 \\ H_1 : p \neq p_0 \end{cases}$$

+ Một phía (*one-tailed test*)

- Một phía bên trái (*lower-tail*)

$$\begin{cases} H_0 : p \geq p_0 \\ H_1 : p < p_0 \end{cases}$$

- Một phía bên phải (*upper-tail*)

$$\begin{cases} H_0 : p \leq p_0 \\ H_1 : p > p_0 \end{cases}$$

Với mức ý nghĩa α .

- **Giả định**

- Cỡ mẫu n lớn: $n \geq 30$, $np_0 \geq 5$ và $n(1-p_0) \geq 5$
- Đặt Y=số phần tử thỏa tính chất \mathcal{A} trong n phần tử khảo sát thì $Y \sim B(n,p)$. Đặt

$$\hat{P} = \frac{Y}{n}$$

Là một ước lượng không chênh cho p .

- Nếu cỡ mẫu n đủ lớn, theo định lý giới hạn trung tâm:

$$Z_0 = \frac{\hat{P} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \sim \mathcal{N}(0,1)$$

Các bước kiểm định

1. Phát biểu giả thuyết H_0 và đối thuyết H_1
2. Xác định mức ý nghĩa α
3. Lấy mẫu ngẫu nhiên cỡ n: X_1, X_2, \dots, X_n và tính thống kê

$$Z_0 = \frac{\hat{P} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

4. Xác định miền bắc bỏ W_α theo bảng

Giả thuyết	Miền bắc bỏ
$H_0 : p = p_0$ $H_1 : p \neq p_0$	$W_\alpha = \{z_0 : z_0 > z_{1-\alpha/2}\}$
$H_0 : p \geq p_0$ $H_1 : p < p_0$	$W_\alpha = \{z_0 : z_0 < -z_{1-\alpha}\}$
$H_0 : p \leq p_0$ $H_1 : p > p_0$	$W_\alpha = \{z_0 : z_0 > z_{1-\alpha}\}$

Trong ngôn ngữ R, để tính $z_{1-\alpha/2}$ sử dụng hàm

$$qnorm(1-\alpha/2, mean=0, sd=1)$$

Hoặc sử dụng p-giá trị để bắc bỏ H_0 khi p-giá trị $\leq \alpha$. Công thức p-giá trị theo bảng

Giả thuyết	p-giá trị
$H_0 : p = p_0$ $H_1 : p \neq p_0$	$p = 2[1 - \varphi(z_0)]$
$H_0 : p \geq p_0$ $H_1 : p < p_0$	$p = \varphi(z_0)$
$H_0 : p \leq p_0$ $H_1 : p > p_0$	$p = 1 - \varphi(z_0)$

Với $\varphi(z)$ là hàm Laplace: $\varphi(z) = \frac{1}{\sqrt{2\pi}} \int_0^z e^{-t^2/2} dt$

Trong ngôn ngữ R, $\varphi(z) = pnorm(z, mean=0, sd=1)$

Ví dụ 88: Trong kỳ nghỉ giáng sinh và đầu năm mới. Cục an toàn giao thông đã thống kê được rằng có 500 người chết và 25000 người bị thương do các vụ tai nạn giao thông trên toàn quốc. Theo thông cáo của Cục An toàn giao thông (ATGT) thì khoảng 50% số vụ tai nạn có liên quan đến bia rượu. Khảo sát ngẫu nhiên 120 vụ tai nạn thấy có 67 vụ do ảnh hưởng của bia rượu. Sử dụng số liệu trên để kiểm định thông báo của Cục An toàn giao thông với mức ý nghĩa $\alpha = 5\%$

Giải:

Các bước kiểm định

1. Phát biểu giả thuyết

$$\begin{aligned} H_0 &: p = 0.5 \\ H_1 &: p \neq 0.5 \end{aligned}$$

2. Xác định mức ý nghĩa: $\alpha = 0.05$

3. Tính giá trị thống kê kiểm định

$$z_0 = \frac{\hat{P} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{(67/120) - 0.5}{\sqrt{\frac{0.5(1-0.5)}{120}}} = 1.28$$

4. Xác định miền bác bỏ

Từ $\alpha = 0.05$ nên $z_{1-\alpha/2} = z_{0.975} = 1.96$. Vậy $|z_0| < z_{0.975}$

5. Kết luận: Với 95% độ tin cậy giả thuyết H_0 không bị bác bỏ.

Sử dụng p-giá trị

Tính p-giá trị bài toán kiểm định 2 phía

$$p = 2(1 - \varphi(z_0)) = 2(1 - \varphi(1.28)) = 2(1 - 0.8994) = 0.2012$$

Do $p > 0.05$ nên với 95% độ tin cậy giả thuyết H_0 không bác bỏ, nghĩa là chưa đủ cơ sở để bác bỏ giả thuyết H_0 .

Thực hiện bằng ngôn ngữ R:

```
> alpha<-0.05
> p0<-0.5; n=120
> pmu<-67/120
> sdmu<-sqrt(p0*(1-p0)/120)
> z0<-(pmu-p0)/sdmu
> hs1<-1-alpha/2
> zhs1<-qnorm(hs1,mean=0,sd=1)
> z0;zhs1
[1] 1.278019
[1] 1.959964
> p<-2*(1-pnorm(z0,mean=0,sd=1))
> p; alpha
[1] 0.2012426
[1] 0.05
```

8.3 KIỂM ĐỊNH GIẢ THUYẾT CHO TRƯỜNG HỢP HAI MẪU ĐỘC LẬP

8.3.1 So sánh hai kỳ vọng, trường hợp biết phương sai

Các giả định

- X_1, X_2, \dots, X_n là các mẫu ngẫu nhiên được chọn từ tổng thể 1 có phân phối chuẩn với kỳ vọng μ_1 và phương sai σ_1^2
- Y_1, Y_2, \dots, Y_m là các mẫu ngẫu nhiên được chọn từ tổng thể 2 có phân phối chuẩn với kỳ vọng μ_2 và phương sai σ_2^2
- Tổng thể 1 và tổng thể 2 (đại diện bởi X và Y) độc lập nhau
- Các phương sai σ_1^2 và σ_2^2 đã biết

Bài toán kiểm định giả thuyết trên hai mẫu độc lập gồm các dạng sau:

+ Hai phía (*two-tailed test*)

$$\begin{cases} H_0 : \mu_1 - \mu_0 = D \\ H_1 : \mu_1 - \mu_0 \neq D \end{cases}$$

+ Một phía (*one-tailed test*)

- Một phía bên trái (*lower-tail*)

$$\begin{cases} H_0 : \mu_1 - \mu_0 = D \\ H_1 : \mu_1 - \mu_0 < D \end{cases}$$

- Một phía bên phải (*upper-tail*)

$$\begin{cases} H_0 : \mu_1 - \mu_0 = D \\ H_1 : \mu_1 - \mu_0 > D \end{cases}$$

Với mức ý nghĩa α .

Các bước kiểm định

1. Phát biểu giả thuyết H_0 và đối thuyết H_1
2. Xác định mức ý nghĩa α
3. Tính thống kê

$$Z_0 = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}}$$

4. Xác định miền bác bỏ W_α giá trị p-giá trị theo bảng

Đối thuyết	Miền bác bỏ	p-giá trị
$H_1 : \mu_1 - \mu_2 \neq D_0$	$W_\alpha = \{z_0 : z_0 > z_{1-\alpha/2}\}$	$p = 2[1 - \varphi(z_0)]$
$H_1 : \mu_1 - \mu_2 < D_0$	$W_\alpha = \{z_0 : z_0 < -z_{1-\alpha}\}$	$p = \varphi(z_0)$
$H_1 : \mu_1 - \mu_2 > D_0$	$W_\alpha = \{z_0 : z_0 > z_{1-\alpha}\}$	$p = 1 - \varphi(z_0)$

5. Kết luận: Nếu bác bỏ H_0 , ta kết luận H_1 chấp nhận với $(1-\alpha)100\%$ độ tin cậy, ngược lại chưa đủ cơ sở để bác bỏ H_0 với α cho trước.

Ví dụ 89: Tại một xí nghiệp người ta xây dựng 2 phương án gia công cùng một loại chi tiết. Để đánh giá xem chi phí trung bình về nguyên liệu theo hai phương án ấy có khác nhau hay không, người ta tiến hành sản xuất thử và thu được các kết quả sau:

P. án 1	2.5	3.2	3.5	3.8	3.5	
P. án 2	2.0	2.7	2.5	2.9	2.3	2.6

Với mức ý nghĩa $\alpha = 0.05$, hãy kết luận về vấn đề trên biết rằng chi phí nguyên liệu theo cả hai phương án gia công đều là các biến ngẫu nhiên phân phối chuẩn với $\sigma_1^2 = \sigma_2^2 = 0.16$.

Giải:

Các bước kiểm định

1. Phát biểu giả thuyết

$$\begin{cases} H_0 : \mu_1 - \mu_0 = 0 \\ H_1 : \mu_1 - \mu_0 \neq 0 \end{cases}$$

2. Xác định mức ý nghĩa: $\alpha = 0.05$
3. Tính giá trị thống kê kiểm định

$$\bar{X} = \frac{2.5 + 3.2 + 3.5 + 3.8 + 3.5}{5} = 3.3$$

$$\bar{Y} = \frac{2.0 + 2.7 + 2.5 + 2.9 + 2.3 + 2.6}{6} = 2.5$$

$$z_0 = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}} = \frac{3.3 - 2.5}{\sqrt{\frac{0.16}{5} + \frac{0.16}{6}}} = 3.30$$

4. Xác định miền bác bỏ

Từ $\alpha = 0.05$ nên $z_{1-\alpha/2} = z_{0.975} = 1.96$. Vậy $|z_0| > z_{0.975}$

5. Kết luận: Với 95% độ tin cậy giả thuyết H_0 bị bác bỏ.

Sử dụng p-giá trị

Tính p-giá trị bài toán kiểm định 2 phía

$$p = 2(1 - \varphi(z_0)) = 2(1 - \varphi(3.30)) = 2(1 - 0.9995) = 0.000957$$

Do $p < 0.05$ nên với 95% độ tin cậy giả thuyết H_0 bị bác bỏ, nghĩa là chi phí nguyên liệu theo 2 phương án gia công trên thực sự khác nhau.

Thực hiện bằng ngôn ngữ R:

```
> pa1<-c(2.5,3.2,3.5,3.8,3.5)
> pa2<-c(2.0,2.7,2.5,2.9,2.3,2.6)
> Xbar<-mean(pa1)
> Ybar<-mean(pa2)
> z0<-(Xbar-Ybar)/sqrt(0.16/5+0.16/6)
> alpha<-0.05
> hs1<-1-alpha/2
> zhs1<-qnorm(hs1,mean=0,sd=1)
> z0;zhs1
[1] 3.302891
[1] 1.959964
> p<-2*(1-pnorm(z0,mean=0,sd=1))
> p;alpha
[1] 0.0009569348
[1] 0.05
```

8.3.2 So sánh hai kỳ vọng, trong trường hợp chưa biết phương sai

Với giả định tương tự như 8.3.1 tuy nhiên phương sai của các mẫu chưa biết

1. So sánh hai kỳ vọng, trong trường hợp chưa biết phương sai, mẫu lớn

- Đối với mẫu lớn, khi phương sai tổng thể σ_1^2, σ_2^2 chưa biết, ta thay bằng các phương sai mẫu S_1^2, S_2^2 .
- Khi $n > 30, m > 30$, tính đại lượng thống kê

$$Z_0 = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n} + \frac{S_2^2}{m}}}$$

- Các bước kiểm định tương tự như trường hợp 8.3.1

2. So sánh hai kỳ vọng, trong trường hợp chưa biết phương sai, mẫu nhỏ

a. Giả định phương sai 2 mẫu bằng nhau: $\sigma_1^2 = \sigma_2^2 = \sigma^2$

Dùng ước lượng chung cho σ_1^2, σ_2^2 là S_p^2 gọi là phương sai mẫu chung (*pooled sample variance*)

$$S_p^2 = \frac{(n-1)S_1^2 + (m-1)S_2^2}{n+m-2}$$

Tính đại lượng thống kê

$$T_0 = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{S_p \sqrt{\frac{1}{n} + \frac{1}{m}}} \sim t(n+m-2)$$

(phân phối Student bậc $df=n+m-2$)

Miền bác bỏ và p-giá trị

Đối thuyết	Miền bác bỏ	p-giá trị
$H_1: \mu_1 - \mu_2 \neq D_0$	$W_\alpha = \{t_0 : t_0 > t_{1-\alpha/2}^{df}\}$	$p = 2P(T_{df} \geq t_0)$
$H_1: \mu_1 - \mu_2 < D_0$	$W_\alpha = \{t_0 : t_0 < -t_{1-\alpha}^{df}\}$	$p = P(T_{df} \leq t_0)$
$H_1: \mu_1 - \mu_2 > D_0$	$W_\alpha = \{t_0 : t_0 > t_{1-\alpha}^{df}\}$	$p = P(T_{df} \geq t_0)$

b. Giả định phương sai 2 mẫu khác nhau: $\sigma_1^2 \neq \sigma_2^2$

Tính đại lượng thống kê

$$T_0 = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n} + \frac{S_2^2}{m}}}$$

Khi đó T_0 có phân phối Student với bậc tự do:

$$df = \frac{\left[(S_1^2/n) + (S_2^2/m) \right]^2}{\frac{(S_1^2/n)^2}{n-1} + \frac{(S_2^2/m)^2}{m-1}}$$

Miền bác bỏ và p-giá trị tương tự như phần a.

8.3.3 So sánh 2 phương sai

Các giả định

- X_1, X_2, \dots, X_n là các mẫu ngẫu nhiên được chọn từ tổng thể 1 có phân phối chuẩn với kỳ vọng μ_1 và phương sai σ_1^2
- Y_1, Y_2, \dots, Y_m là các mẫu ngẫu nhiên được chọn từ tổng thể 2 có phân phối chuẩn với kỳ vọng μ_2 và phương sai σ_2^2

Người ta chứng minh được:

$$\frac{(n-1)S_1^2}{\sigma_1^2} \sim \chi^2(n-1) \text{ và } \frac{(m-1)S_2^2}{\sigma_2^2} \sim \chi^2(m-1)$$

Khi đó đại lượng

$F = \frac{S_1^2 / \sigma_1^2}{S_2^2 / \sigma_2^2}$ có phân phối Fisher_Snedecor bậc tự do (n-1,m-1): F(n-1,m-1)

Các bước kiểm định

1. Phát biểu giả thuyết H_0 và đối thuyết H_1

- a. $H_0 : \sigma_1^2 = \sigma_2^2$; $H_1 : \sigma_1^2 \neq \sigma_2^2$
- b. $H_0 : \sigma_1^2 = \sigma_2^2$; $H_1 : \sigma_1^2 > \sigma_2^2$
- c. $H_0 : \sigma_1^2 = \sigma_2^2$; $H_1 : \sigma_1^2 < \sigma_2^2$

2. Xác định mức ý nghĩa α

3. Tính thống kê $F_0 = \frac{S_1^2}{S_2^2}$

4. Xác định miền bác bỏ W_α giá trị p-giá trị theo bảng

Đối thuyết	Miền bác bỏ
$H_1 : \sigma_1^2 \neq \sigma_2^2$	$W_\alpha = \{F_0 : F_0 < f_{1-\alpha/2}^{(n-1,m-1)} \vee F_0 > f_{\alpha/2}^{(n-1,m-1)}\}$
$H_1 : \sigma_1^2 < \sigma_2^2$	$W_\alpha = \{F_0 : F_0 > f_{1-\alpha}^{(n-1,m-1)}\}$
$H_1 : \sigma_1^2 > \sigma_2^2$	$W_\alpha = \{F_0 : F_0 > f_\alpha^{(n-1,m-1)}\}$

Trong đó: $f_\alpha^{(n,m)}$ là giá trị của biến ngẫu nhiên F có phân phối Fisher_Snedecor bậc tự do (n,m) thỏa:

$$P(F > f_\alpha^{(n,m)}) = \alpha \text{ và có tính chất } f_\alpha^{(n,m)} = \frac{1}{f_{1-\alpha}^{(n,m)}}$$

Trong ngôn ngữ R, $f_\alpha^{(n-1,m-1)} = qf(1-\alpha, n, m)$

Ví dụ 90: Người ta lấy 2 mẫu ứng với 2 giống lúa và thu được kết quả

Giống lúa	Kích thước mẫu	Phương sai mẫu
A	n=41	$S_1^2 = 11.41$
B	m=40	$S_2^2 = 6.52$

Biết hai giống lúa có năng suất trung bình xấp xỉ nhau, nhưng mức độ phân tán về năng suất (phương sai) có thể khác nhau.

Gọi X, Y lần lượt là các năng suất của 2 giống lúa A và B, X, Y có phân phối chuẩn với các phương sai chưa biết lần lượt là σ_1^2, σ_2^2 . Với mức ý nghĩa 5%, độ thiêu ổn định (độ phân tán) năng suất giống lúa thứ 1 có cao hơn của giống lúa thứ 2 không?

Giải:

Các bước kiểm định

1. Phát biểu giả thuyết

$$\begin{cases} H_0 : \sigma_1^2 = \sigma_2^2 \\ H_1 : \sigma_1^2 > \sigma_2^2 \end{cases}$$

2. Xác định mức ý nghĩa: $\alpha = 0.05$
3. Tính giá trị thống kê kiểm định

$$f_0 = \frac{S_1^2}{S_2^2} = \frac{11,41}{6,52} = 1.75$$

4. Xác định miền bác bỏ

Từ $\alpha = 0.05$ nên $f_{0.05}^{40,39} = 1.7004 \Rightarrow f_0 > f_{0.05}^{40,39}$

5. Kết luận: Với 95% độ tin cậy giả thuyết H_0 bị bác bỏ, có nghĩa độ thiếu ổn định năng suất giống lúa thứ 1 cao hơn giống lúa thứ 2.

Thực hiện bằng ngôn ngữ R:

```
> alpha=0.05
> sd1<-11.41
> sd2<-6.52
> f0<-sd1/sd2
> falpha4039<-qf(1-alpha, 40, 39)
> f0;falpha4039
[1] 1.75
[1] 1.700385
```

8.3.4 So sánh hai tỷ lệ

Khảo sát những phần tử thỏa một tính chất \mathcal{A} nào đó trên hai tổng thể độc lập với tỷ lệ tương ứng là p_1, p_2 . Từ 2 tổng thể chọn ra 2 mẫu với cỡ lần lượt là n và m . Gọi X và Y là số phần tử thỏa tính chất \mathcal{A} trong mẫu 1 và mẫu 2, khi đó $X \sim B(n, p_1)$ và $Y \sim B(m, p_2)$

Bài toán: so sánh tỷ lệ p_1 và p_2

Bài toán kiểm định gồm các trường hợp sau

- $H_0 : p_1 - p_2 = D_0$; $H_1 : p_1 - p_2 \neq D_0$
- $H_0 : p_1 - p_2 = D_0$; $H_1 : p_1 - p_2 < D_0$
- $H_0 : p_1 - p_2 = D_0$; $H_1 : p_1 - p_2 > D_0$

Các giả định

- Hai mẫu độc lập
- Cỡ mẫu lớn: $np_1 > 5; n(1-p_1) > 5; mp_2 > 5; m(1-p_2) > 5$

Các bước kiểm định

1. Phát biểu giả thuyết H_0 và đối thuyết H_1
2. Xác định mức ý nghĩa α
3. Tính thống kê

$$Z_0 = \frac{\hat{P}_1 - \hat{P}_2 - D_0}{\sqrt{\hat{P}(1-\hat{P})\left(\frac{1}{n} + \frac{1}{m}\right)}}$$

Với $\hat{P}_1 = \frac{X}{n}; \hat{P}_2 = \frac{Y}{m}; \hat{P} = \frac{X+Y}{n+m}$; và $Z \sim \mathcal{N}(0, 1)$

- Xác định miền bác bỏ W_α giá trị p-giá trị theo bảng

Đối thuyết	Miền bác bỏ	p-giá trị
$H_1: p_1 - p_2 \neq D_0$	$W_\alpha = \{z_0 : z_0 > z_{1-\alpha/2}\}$	$p = 2[1 - \varphi(z_0)]$
$H_1: p_1 - p_2 < D_0$	$W_\alpha = \{z_0 : z_0 < -z_{1-\alpha}\}$	$p = \varphi(z_0)$
$H_1: p_1 - p_2 > D_0$	$W_\alpha = \{z_0 : z_0 > z_{1-\alpha}\}$	$p = 1 - \varphi(z_0)$

- Kết luận: Nếu bác bỏ H_0 , ta kết luận H_1 chấp nhận với $(1-\alpha)100\%$ độ tin cậy, ngược lại chưa đủ cơ sở để bác bỏ H_0 với α cho trước.

Ví dụ 91: Một công ty sản xuất thuốc cần kiểm tra một loại thuốc có tác dụng làm giảm xuất hiện cơn đau ngực ở các bệnh nhân. Công ty thực hiện thí nghiệm trên 400 người, chia làm 2 nhóm:

- Nhóm 1: gồm 200 người được uống thuốc
- Nhóm 2: gồm 200 người được uống giả dược.

Theo dõi thấy ở nhóm 1 có 8 người lâm cơn đau ngực và nhóm 2 có 25 người lâm cơn đau ngực. Với $\alpha = 0.05$, hãy cho kết luận về hiệu quả của thuốc mới sản xuất.

Thực hiện bằng ngôn ngữ R:

```
> alpha<-0.05
> n<-200;m<-200
> p1<-192/n;p2<-175/200
> n*p1
[1] 192
> m*p2
[1] 175
> n*(1-p1)
[1] 8
> m*(1-p2)
[1] 25
> X<-192;Y<-175
> p1hat<-X/n;p2hat<-Y/m;phat<- (X+Y) / (n+m)
> z0<-(p1hat-p2hat-0)/sqrt(phat*(1-phat)*(1/n+1/m))
> hs<-1-alpha
> zhs<-qnorm(hs,mean=0,sd=1)
> z0;zhs
[1] 3.089505
[1] 1.644854
> p<-1-pnorm(z0,mean=0,sd=1)
> p;alpha
[1] 0.001002451
[1] 0.05
```

Kết luận:

Với giả thuyết

$$H_0: p_1 - p_2 = D_0$$

$$H_1: p_1 - p_2 > D_0$$

Do $z_0 > z_{1-\alpha}$ ($p < \alpha$) nên giả thuyết H_0 bị bác bỏ, nghĩa là tỷ lệ bệnh nhân giảm cơn đau ngực được uống thuốc cao hơn tỷ lệ bệnh nhân uống giả dược với độ tin cậy $(1-0.05)100\% = 95\%$.

8.4 KIỂM ĐỊNH GIẢ THUYẾT CHO TRƯỜNG HỢP HAI MẪU KHÔNG ĐỘC LẬP (PAIRED t-TEST)

Khi quan sát hai mẫu không độc lập thì mỗi giá trị quan trắc trong một mẫu có mối liên hệ tương ứng với một giá trị quan trắc ở mẫu thứ hai. Giữa hai mẫu có thể ghép cặp từng đôi giá trị trong hai mẫu với nhau.

Kiểm định cho sự khác biệt của hai trung bình tổng thể

Bài toán:

Xét (X_{1i}, X_{2i}) , với $i=1,2,\dots,n$ là tập gồm n cặp giá trị quan trắc với giả sử rằng kỳ vọng và phương sai của tổng thể của X_1 là μ_1 và σ_1^2 ; kỳ vọng và phương sai của tổng thể của X_2 là μ_2 và σ_2^2 , X_{1i}, X_{2j} ($i \neq j$) độc lập.

Độ sai khác giữa mỗi cặp trong tập hợp các giá trị quan trắc là

$$D_i = X_{1i} - X_{2i}$$

Các D_i ($i=1,2,\dots,n$) là các biến ngẫu nhiên độc lập và có cùng phân phối chuẩn. Gọi $\mu_D = E(D_i)$, d_i là giá trị của D_i ta định nghĩa

$$\bar{d} = \frac{1}{n} \sum_{i=1}^n d_i$$

$$s_d^2 = \frac{1}{n-1} \sum_{i=1}^n (d_i - \bar{d})^2 = \frac{1}{n-1} \sum_{i=1}^n d_i^2 - \frac{n}{n-1} (\bar{d})^2$$

- Bài toán kiểm định gồm các trường hợp sau:

- $H_0 : \mu_D = D_0$; $H_1 : \mu_D \neq D_0$
- $H_0 : \mu_D = D_0$; $H_1 : \mu_D < D_0$
- $H_0 : \mu_D = D_0$; $H_1 : \mu_D > D_0$

Các bước kiểm định

1. Phát biểu giả thuyết H_0 và đối thuyết H_1
2. Xác định mức ý nghĩa α
3. Tính thống kê $T_0 = \frac{\bar{D} - D_0}{S_D / \sqrt{n}} \sim t(n-1)$
4. Xác định miền bác bỏ W_α giá trị p-giá trị theo bảng

Giả thuyết	Miền bác bỏ	p-giá trị
$H_1 : \mu_D \neq D_0$	$W_\alpha = \{t_0 : t_0 > t_{1-\alpha/2}^{n-1}\}$	$p = 2P(T_{n-1} \geq t_0)$
$H_1 : \mu_D < D_0$	$W_\alpha = \{t_0 : t_0 < -t_{1-\alpha}^{n-1}\}$	$p = P(T_{n-1} \leq t_0)$
$H_1 : \mu_D > D_0$	$W_\alpha = \{t_0 : t_0 > t_{1-\alpha}^{n-1}\}$	$p = P(T_{n-1} \geq t_0)$

Lưu ý: Trường hợp cỡ mẫu $n > 30$, bài toán kiểm định hai mẫu phụ thuộc thực hiện tương tự như trường hợp 1 mẫu dựa trên mẫu ngẫu nhiên

$$(D_1, D_2, \dots, D_n)$$

Ví dụ 92: Để so sánh tốc độ xử lý của 2 phần mềm thông kê SW1 và SW2, người ta lấy 10 bộ dữ liệu thực hiện bởi hai phần mềm này trên cùng một máy tính. Kết quả thu được thời gian xử lý của 2 phần mềm ứng với lần lượt 10 bộ dữ liệu là

SW1	9.98	9.88	9.84	9.99	9.94	9.84	9.86	10.12	9.90	9.91
SW2	9.88	9.86	9.75	9.8	9.87	9.84	9.87	9.86	9.83	9.86

Hãy kiểm định phần mềm SW2 có chạy nhanh hơn SW1 không? Với độ tin cậy 95%.

Giải:

1. Phát biểu giả thuyết

$$\begin{cases} H_0 : \mu_D = 0 \\ H_1 : \mu_D > 0 \end{cases}$$

2. Xác định mức ý nghĩa: $\alpha = 0.05$

3. Tính giá trị thống kê kiểm định

$$\bar{d} = \frac{1}{n} \sum_{i=1}^n d_i = \frac{0.84}{10} = 0.084$$

$$s_d^2 = \frac{1}{n-1} \sum_{i=1}^n (d_i - \bar{d})^2 = \frac{1}{n-1} \sum_{i=1}^n d_i^2 - \frac{n}{n-1} (\bar{d})^2 = \frac{0.0641}{9}$$

$$T_0 = \frac{\bar{D} - D_0}{S_D / \sqrt{n}} = \frac{0.084 - 0}{0.0844 / \sqrt{10}} = 3.15$$

4. Xác định miền bác bỏ

Từ $\alpha = 0.05$ nên $t_{0.05}^9 = 1.8331 \Rightarrow t_0 > t_{0.05}^9$

5. Kết luận: Với 95% độ tin cậy giả thuyết H_0 bị bác bỏ, có nghĩa có thể kết luận SW2 chạy nhanh hơn SW1 với độ tin cậy 95%.

Thực hiện bằng ngôn ngữ R:

Có thể thực hiện thủ công như 8.3.2.2, trong ví dụ này sử dụng hàm `t.test()` có sẵn của R.

```
> SW1<-c(9.98,9.88,9.84,9.99,9.94,9.84,9.86,10.12,9.90,9.91)
> SW2<-c(9.88,9.86,9.75,9.8,9.87,9.84,9.87,9.86,9.83,9.86)
> t.test(SW1,SW2,paired=TRUE,alt="greater")
```

```
Paired t-test

data: SW1 and SW2
t = 3.149, df = 9, p-value = 0.005878
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 0.03510171      Inf
sample estimates:
mean of the differences
 0.084
```

8.5 KIỂM ĐỊNH CHI BÌNH PHƯƠNG (χ^2 -TESTING)

8.5.1 Kiểm định giả thuyết về phân phối

Bài toán Khảo sát một đại lượng ngẫu nhiên X với qui luật phân phối xác suất chưa biết. Cần kiểm định xem $X \sim \mathcal{D}(x; \theta)$?

Các bước kiểm định

Chọn mẫu ngẫu nhiên cỡ n: (X_1, X_2, \dots, X_n)

- Chia miền giá trị của các biến ngẫu nhiên X_i thành K khoảng không trùng nhau l_1, l_2, \dots, l_K
 - Trường hợp X là biến ngẫu nhiên rời rạc, ta chia thành K điểm x_1, x_2, \dots, x_K
1. Gọi O_j là số các giá trị mẫu nằm trong khoảng l_j , nếu X là biến ngẫu nhiên rời rạc thì O_j là tần số lặp lại của giá trị x_j . O_j gọi là giá trị tần số thực nghiệm.
 2. Phát biểu giả thuyết

$$H_0 : X \text{ tuân theo qui luật } \mathcal{D}(x; \theta)$$

$$H_1 : X \text{ không tuân theo qui luật } \mathcal{D}(x; \theta)$$

Tính các thống kê:

$$p_j = P(X \in l_j) \text{ hoặc } p_j = P(X = x_j) \text{ (nếu } X \text{ rời rạc)}$$

(P0): là **hàm xác suất ứng với phân phối** $\mathcal{D}(x; \theta)$). Đặt $E_j = np_j$: tần số lý thuyết. **Điều kiện** $E_j \geq 5$

$$3. \text{ Thống kê kiểm định } Q^2 = \sum_{j=1}^K \frac{(O_j - E_j)^2}{E_j} \sim \chi^2(K-1)$$

$$4. \text{ Bác bỏ } H_0 \text{ nếu } Q^2 \geq \chi^2_{\alpha, K-r-1}, \text{ với } r \text{ là số tham số ước lượng.}$$

Ví dụ 93: Đo chiều cao của một loại cây trồng cùng độ tuổi có số liệu:

Chiều cao (cm)	0-3	3-6	6-9	9-12	12-15	15-18	18-21	21-24	24-27	27-30
Số cây có cùng chiều cao (n_j)	10	30	40	60	110	100	70	50	20	10

Với mức ý nghĩa $\alpha = 0.05$, có thể coi chiều cao của loại cây này là phân phối chuẩn không?

Giải:

1. Phát biểu giả thuyết

Gọi X là chiều cao của loại cây này. Cần kiểm định

$$H_0 : X \sim \mathcal{N}(\mu; \sigma^2)$$

$$H_1: X \neq \mathcal{N}(\mu; \sigma^2)$$

ở đây $\mu \approx \bar{x}; \sigma^2 \approx s^2$

2. Xác định mức ý nghĩa: $\alpha = 0.05$
3. Tính giá trị thống kê kiểm định

$$\bar{x} = \frac{\sum_{i=1}^{10} n_i x_i}{n} = \frac{7500}{500} = 15; s^2 = \frac{\sum_{i=1}^{10} n_i x_i^2 - n\bar{x}^2}{n} = 34.65$$

Số tham số cần xấp xỉ là 2

$$P_i = P(X \in (x_i, x_{i+1})) = \varphi\left(\frac{x_{i+1} - \bar{x}}{s}\right) - \varphi\left(\frac{x_i - \bar{x}}{s}\right)$$

Biên cỡ	Xác suất P_i	T/số lý thuyết $n_i P_i$	n_i
$X \in (-\infty, 3)$	0.0210	10.50	10
$X \in (3, 6)$	0.0426	21.30	30
$X \in (6, 9)$	0.0910	45.50	40
$X \in (9, 12)$	0.1510	75.50	60
$X \in (12, 15)$	0.1944	97.20	110
$X \in (15, 18)$	0.1944	97.20	100
$X \in (18, 21)$	0.1510	75.50	70
$X \in (21, 24)$	0.0910	45.50	50
$X \in (24, 27)$	0.0426	21.30	20
$X \in (27, +\infty)$	0.0210	10.50	10
Σ			n=500

$$Q^2 = 10.1394$$

4. Xác định miền bác bỏ

Từ $\alpha = 0.05$ nên $\chi^2_{0.05, 10-2-1} = \chi^2_{0.05, 7} = 14.0617 \Rightarrow Q^2 < \chi^2_{0.05, 7}$

5. Kết luận: $Q^2 \notin W_\alpha$: Giả thuyết H_0 được chấp thuận:

X có phân phối chuẩn

Thực hiện theo ví dụ trên bằng ngôn ngữ R. Trong R, có hàm *shapiro.test()* có thể dùng để kiểm định một biến ngẫu nhiên có phân phối chuẩn hay không. Với ví dụ trên ta có lời giải:

```
> Oj<-c(10,30,40,60,110,100,70,50,20,10)
> Xj<-c(1.5,4.5,7.5,10.5,13.5,16.5,19.5,22.5,22.5,28.5)
> n<-500
> xbar<-sum(Oj*Xj)/n;sd<-(sum(Oj*Xj^2)-n*xbar^2)/n
> Xjnormal<-(Xj-xbar)/sqrt(sd)
> shapiro.test(Xjnormal)
```

Shapiro-Wilk normality test

```
data: Xjnormal
W = 0.972, p-value = 0.9087
```

Với $p > \alpha$, nên giả thuyết H_0 được chấp thuận, nghĩa là có thể coi chiều cao của loại cây này là đại lượng ngẫu nhiên theo qui luật phân phối chuẩn.

Chú ý: Khi kiểm tra một biến ngẫu nhiên X có phân phối chuẩn $\mathcal{N}(0,1)$ hay không, ta chuẩn hóa:

$$X_{normal} = \frac{X - \mu}{\sigma}$$

8.5.2 Kiểm định giả thuyết về tính độc lập

Bài toán Giả sử mỗi phần tử trong một tổng thể có thể được phân loại theo hai đặc tính khác nhau, gọi là các đặc tính X và đặc tính Y . X có r giá trị và Y có s giá trị. Đặt

$$P_{ij} = P(X = x_i, Y = y_j) \text{ với } i=1,2,\dots,r; j=1,2,\dots,s$$

$$p_i = P(X = x_i) = \sum_{j=1}^s P_{ij}, i=1,2,\dots,n \text{ và } q_j = P(Y = y_j) = \sum_{i=1}^r P_{ij}, j=1,2,\dots,m$$

- **Kiểm định X có độc lập với Y :** là kiểm định với giả thuyết

$$H_0 : P_{ij} = p_i q_j, \forall i=1,2,\dots,r; j=1,2,\dots,s$$

$$H_1 : \exists (i,j), P_{ij} \neq p_i q_j$$

Khảo sát N phần tử, ta có bảng kết quả gọi là bảng liên hợp (*contingency table*)

$X \backslash Y$	y_1	y_2	...	y_s	Tổng hàng
x_1	n_{11}	n_{12}		n_{1s}	n_1
x_2	n_{21}	n_{22}		n_{2s}	n_2
....					
x_r	n_{r1}	n_{r2}		n_{rs}	n_r
Tổng cột	m_1	m_2		m_s	N

Trong đó n_{ij} gọi là tần số thực nghiệm.

- Các đại lượng thống kê

Ước lượng của p_i và q_j lần lượt là

$$\hat{p}_i = \frac{n_i}{N}, i=1,\dots,r$$

$$\hat{q}_j = \frac{m_j}{N}, j=1,\dots,s$$

- Gọi N_{ij} là số phần tử có đặc tính (x_i, y_j) trong N phần tử khảo sát, thì

$$N_{ij} \sim B(N, P_{ij}). \text{ Khi đó khi } H_0 \text{ đúng}$$

Đặt $e_{ij} = N\hat{p}_i \hat{q}_j = \frac{n_i m_j}{N}, e_{ij}$ gọi là tần số lý thuyết.

Theo Pearson, với và $E_{ij} = NP_{ij}$ biến ngẫu nhiên

$$\sum_{i=1}^r \sum_{j=1}^s \frac{(N_{ij} - E_{ij})^2}{E_{ij}} \rightarrow \chi^2_{(r-1)(s-1)}$$

Các bước kiểm định

1. Phát biểu giả thuyết H_0 và đối thuyết H_1
2. Xác định mức ý nghĩa α
3. Xác định tần số thực nghiệm n_{ij} và tần số lý thuyết

$$e_{ij} = \frac{n_i m_j}{N} \quad (\text{điều kiện } e_{ij} \geq 5)$$

4. Tính thống kê kiểm định

$$Q^2 = \sum_{i=1}^r \sum_{j=1}^s \frac{(n_{ij} - e_{ij})^2}{e_{ij}} = \sum_{i=1}^r \sum_{j=1}^s \frac{n_{ij}^2}{e_{ij}} - N \sim \chi^2_{(r-1)(s-1)}$$

5. Xác định miền bác bỏ :

$$H_0 \text{ bác bỏ khi } Q^2 > \chi^2_{(r-1)(s-1)}(\alpha)$$

6. Sử dụng p-giá trị $p = P(\chi^2_{(r-1)(s-1)} \geq Q^2)$

$$H_0 \text{ bác bỏ khi } p < \alpha.$$

Ví dụ 94: Một báo cáo khoa học tuyên bố rằng việc sở hữu một thú cưng trong nhà sẽ làm tăng khả năng sống sót của những người chủ thường bị lâm cơn đau tim. Một mẫu ngẫu nhiên gồm 95 người đã lâm cơn đau tim được chọn để khảo sát. Dữ liệu của mỗi người được khảo sát được chia làm 2 loại:

- Người sống sót/tử vong 1 năm sau khi lâm cơn đau tim
- Người sống sót/tử vong có nuôi thú cưng trong nhà hay không

	Có nuôi thú cưng	Không nuôi thú cưng
Sống sót	28	44
Tử vong	8	15

Giải:

1. Phát biểu giả thuyết H_0 : Bệnh lâm cơn đau tim độc lập với việc nuôi thú cưng
2. Tính tần số thực nghiệm: $n_1 = 72; n_2 = 23; m_1 = 36; m_2 = 59$

$$e_{11} = \frac{n_1 m_1}{N} = \frac{72 * 36}{95} = 27.284; \quad e_{12} = \frac{n_1 m_2}{N} = \frac{72 * 59}{95} = 44.716$$

$$e_{21} = \frac{n_2 m_1}{N} = \frac{23 * 36}{95} = 8.716; \quad e_{22} = \frac{n_2 m_2}{N} = \frac{23 * 59}{95} = 14.284$$

3. Tính giá trị thống kê

$$Q^2 = \sum_{i=1}^r \sum_{j=1}^s \frac{n_{ij}^2}{e_{ij}} - n = \left(\frac{28^2}{27.284} + \frac{44^2}{44.716} + \frac{8^2}{8.716} + \frac{15^2}{14.284} \right) - 95 = 0.125$$

$$\chi^2_{(r-1)(s-1)}(\alpha) = \chi^2_1(0.05) = 3.81$$

4. Kết luận: Do $Q^2 < \chi^2_1$ nên chưa đủ cơ sở để bác bỏ H_0 : bệnh nhân lên cơn đau tim độc lập với việc nuôi thú cưng.

Hãy thực hiện ví dụ trên bằng ngôn ngữ R.

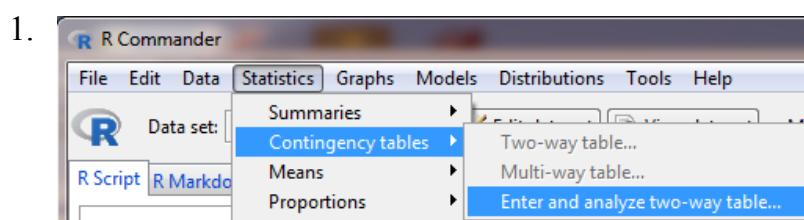
Ở đây trình bày cách kiểm định dùng hàm *chisq.test()*

```
> table<-matrix(c(28,44,8,15),2,2,byrow=TRUE)
> chisq.test(table,correct=FALSE)

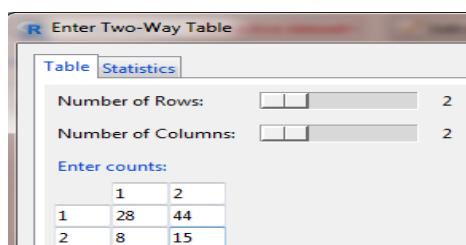
Pearson's Chi-squared test

data: table
X-squared = 0.12489, df = 1, p-value = 0.7238
```

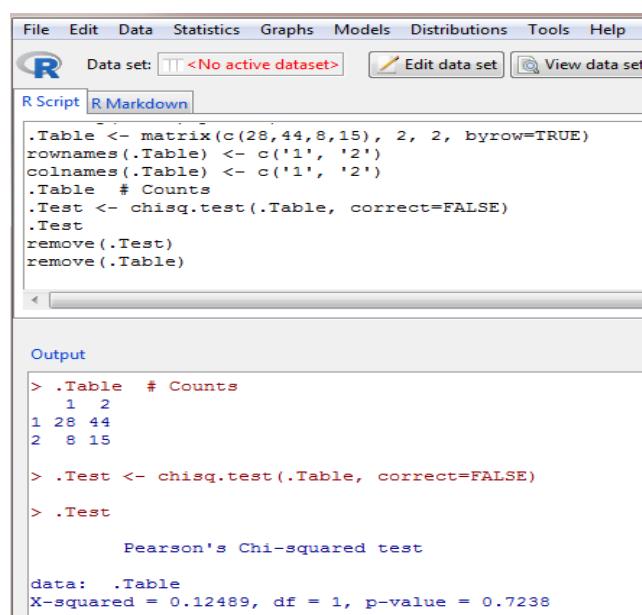
Hoặc trong R-commander



2.



3.



TÀI LIỆU THAM KHẢO

- [1] Biostatistics with R, Babak Shahbaba, Springer, 2011
- [2] <https://home.ubalt.edu/.../chapter8/chapter8.ppt>
- [3] Giáo trình Lý thuyết Xác suất và thống kê toán, PGS.TS Nguyễn Cao Văn, TS Trần Thái Ninh, NXB Thống kê, 2005
- [4] Introduction to Probability and Statistics Using R, G.Jay Kerns, First Edition, *cran.r-project.org/web/packages/IPSUR/vignettes/IPSUR.pdf*, 2010
- [5] Introduction to Statistical Thinking (With R, Without Calculus), Benjamin, The Hebrew University, 2011.
- [6] Lý thuyết Xác suất và thống kê (bài giảng), Hoàng Văn Hà, ĐH KHTN Tp HCM, 2012
- [7] Phân tích dữ liệu với R, Nguyễn Văn Tuấn, NXB Tổng hợp Tp HCM, 2014
- [8] Simple-Using R for Introductory Statistics, John Verzani, 2001,
<https://cran.r-project.org/doc/contrib/Verzani-SimpleR.pdf>
- [9] www.sagepub.com/sites/default/.../40007_Chapter8.pdf

C H U O N G

9

HỒI QUI & TƯƠNG QUAN

9.1 TƯƠNG QUAN (CORRELATION)

Phân tích tương quan (*Correlation Analysis*) dùng để đo độ mạnh của mối liên hệ tuyến tính giữa hai biến ngẫu nhiên. Độ mạnh của mối liên hệ này được thể hiện qua hệ số tương quan (*coefficient of correlation*).

Trong phần này chúng tôi trình bày 3 hệ số tương quan thông dụng là: hệ số tương quan Pearson r , Spearman ρ , Kendall τ .

9.1.1 Hệ số tương quan Pearson r

Cho hai biến ngẫu nhiên X và Y tuân theo qui luật phân phối chuẩn, được quan trắc thực nghiệm bởi hai mẫu cỡ n:

$$X: X_1, X_2, \dots, X_n; Y: Y_1, Y_2, \dots, Y_n$$

Hệ số tương quan Pearson r được xác định:

$$r_{XY} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2 \sum_{i=1}^n (Y_i - \bar{Y})^2}}$$

Hệ số tương quan $r_{XY} \in [-1, +1]$, thực tế người ta qui ước

$|r_{XY}| > 0.8$: tương quan tuyến tính rất mạnh

$|r_{XY}| \in (0.6, 0.8)$: tương quan tuyến tính mạnh

$|r_{XY}| \in (0.4, 0.6)$: có tương quan tuyến tính

$|r_{XY}| \in (0.2, 0.4)$: tương quan tuyến tính yếu

$|r_{XY}| < 0.2$: tương quan tuyến tính rất yếu hoặc không có tương quan tuyến tính.

Trong ngôn ngữ R để tính hệ số tương quan r của hai biến x, y người ta sử dụng hàm $cor(x, y)$

Ví dụ 95: Xét dữ liệu nghiên cứu độ Cholesterol trong máu của 18 đối tượng nam như sau (BMI: tỷ số trọng lượng (kg) với chiều cao bình phương (cm^2)). Ước tính hệ số tương quan giữa độ tuổi và Cholesterol

Mã số ID (id)	Độ tuổi (age)	BMI (bmi)	Cholesterol (chol)
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	4.6
15	40	26.4	3.2
16	48	21.2	4.2
17	28	21.2	2.3
18	49	22.8	4.0

Thực hiện bằng ngôn ngữ R như sau:

```
> 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)
> chol<-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,chol)
> cor(age,chol)
[1] 0.9367261
```

Trong R còn có phép kiểm định giả thiết hệ số tương quan bằng 0, dựa vào hàm *cor.test(x,y)*

```
> cor.test(age,chol)

Pearson's product-moment correlation

data: age and chol
t = 10.704, df = 16, p-value = 1.058e-08
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
0.8350463 0.9765306
sample estimates:
cor
0.9367261
```

Kết quả phân tích cho thấy giữa độ tuổi và cholesterol có ý nghĩa thống kê.

9.1.2 Hệ số tương quan Spearman ρ

Cho hai biến ngẫu nhiên X và Y , được quan trắc thực nghiệm bởi hai mẫu cỡ n :

$$X: X_1, X_2, \dots, X_n; Y: Y_1, Y_2, \dots, Y_n$$

Ký hiệu: $D_i = X_i - Y_i$

Hệ số tương quan Spearman được tính

$$r_{\rho} = 1 - \frac{6 \sum_{i=1}^n D_i^2}{n(n-1)}$$

Trong công thức tính toán này, các biến ngẫu nhiên X, Y được biến đổi thành thứ bậc (rank) và xem độ tương quan giữa hai dãy số bậc. (X, Y có thể không tuân theo phân phối chuẩn).

Trong ngôn ngữ R, tính hệ số tương quan r của hai biến x, y người ta sử dụng hàm `cor.test(x,y,method="spearman")`

Ví dụ 96:

```
> cor.test(age, chol, method="spearman")

  Spearman's rank correlation rho

data: age and chol
S = 51.158, p-value = 2.57e-09
alternative hypothesis: true rho is not equal to 0
sample estimates:
rho
0.947205

Warning message:
In cor.test.default(age, chol, method = "spearman") :
  Cannot compute exact p-value with ties
```

Kết quả này cũng cho thấy mối liên hệ giữa độ tuổi và cholesterol có ý nghĩa thống kê rất cao.

9.1.3 Hệ số tương quan Kendall τ

Hệ số tương quan Kendall τ được tính dựa vào các cặp số “song hành” (*concordant* và *discordant*). Cách thức tính toán có thể tham khảo [8]. Chú ý, khi tính toán hệ số tương quan Kendall τ , một biến ngẫu nhiên X (hoặc Y) phải được sắp thứ tự. Trong ngôn ngữ R sử dụng hàm `cor.test(x,y,method="kendall")` để tính hệ số tương quan

```
> cor.test(age, chol, method="kendall")

  Kendall's rank correlation tau

data: age and chol
z = 4.755, p-value = 1.984e-06
alternative hypothesis: true tau is not equal to 0
sample estimates:
tau
0.8333333

Warning message:
In cor.test.default(age, chol, method = "kendall") :
  Cannot compute exact p-value with ties
```

Kết quả cũng cho thấy mối liên hệ giữa độ tuổi (*age*) và cholesterol (*chol*) có sự tương quan tuyến tính.

9.2 HỒI QUI (REGRESSION)

Phân tích hồi qui là nghiên cứu mối liên hệ phụ thuộc của 1 biến (gọi là *biến phụ thuộc*) vào một hay nhiều biến khác (gọi là các biến *độc lập*), với ý tưởng ước lượng

hoặc/và dự đoán giá trị trung bình (*tổng thể*) của các biến phụ thuộc trên cơ sở các biến độc lập (*dựa trên mẫu*).

9.2.1 Hồi qui tuyến tính đơn giản (*Simple linear regression*)

Cho hai biến ngẫu nhiên X và Y , được quan trắc thực nghiệm bởi hai mẫu cỡ n:

$$X : X_1, X_2, \dots, X_n; Y : Y_1, Y_2, \dots, Y_n$$

Y có quan hệ tuyến tính với X , nếu

$$Y_i = \alpha + \beta X_i + \varepsilon_i, i=1,2,\dots,n$$

Với: ε_i là biến ngẫu nhiên theo luật phân phối chuẩn $\mathcal{N}(0; \sigma^2)$

α : gọi là chặn (*intercept*)

β : gọi là độ dốc (*slop hay gradient*)

Các hệ số này được ước tính từ dữ liệu. Phương pháp ước tính là phương pháp bình phương bé nhất (*least squares method*). Phương pháp này tìm α, β để

$$\sum_{i=1}^n [y_i - (\alpha + \beta x_i)]^2 \text{ đạt giá trị bé nhất}$$

Khi đó

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

$$\hat{\alpha} = \bar{y} - \hat{\beta} \bar{x}$$

Chú ý: $\hat{\alpha}, \hat{\beta}$ là các ước lượng xấp xỉ của α, β . Từ $\hat{\alpha}, \hat{\beta}$ ta có $\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$, khi đó đại lượng $(y_i - \hat{y}_i)$ gọi là phần dư (*residual*). Phương sai của phần dư có thể ước lượng bằng

$$s^2 = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n - 2}.$$

Trong phân tích hồi qui tuyến tính, thông thường chúng ta muốn biết hệ số $\beta = 0$ hay $\beta \neq 0$. Nếu $\beta = 0$, thì cũng có nghĩa là không có mối liên hệ gì giữa X và Y ; nếu $\beta \neq 0$, chúng ta có bằng chứng để phát biểu rằng X và Y có liên quan nhau. Để kiểm định giả thiết $\beta = 0$ ta xét nghiệm t sau đây:

$$t = \frac{\hat{\beta}}{SE(\hat{\beta})}$$

$SE(\hat{\beta})$ là sai số chuẩn (standard error) của $\hat{\beta}$. $t \sim t(n-2)$

9.2.2 Phân tích hồi qui tuyến tính đơn giản bằng R

Trong ngôn ngữ R sử dụng hàm lm (*linear model*) để tính các giá trị $\hat{\alpha}, \hat{\beta}$ và s^2 .

Ví dụ 97: Phân tích hồi qui tuyến tính đơn giản cho hai đại lượng age , $chol$ (nêu trong ví dụ 95)

```
> lm(chol~age)

Call:
lm(formula = chol ~ age)

Coefficients:
(Intercept)      age
1.08922        0.05779
```

Trong kết quả này mô tả $chol$ là một hàm số của age , với $\hat{\alpha} = 1.0892$; $\hat{\beta} = 0.05779$, nghĩa là ta có phương trình tuyến tính

$$\hat{y}_i = 1.08922 + 0.05779 * \hat{x}_i$$

Có thể lấy nhiều thông tin hơn bằng cách

```
> reg<-lm(chol~age)
> summary(reg)

Call:
lm(formula = chol ~ age)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.40729 -0.24133 -0.04522  0.17939  0.63040 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 1.089218   0.221466   4.918 0.000154 ***
age         0.057788   0.005399  10.704 1.06e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3027 on 16 degrees of freedom
Multiple R-squared:  0.8775,    Adjusted R-squared:  0.8698 
F-statistic: 114.6 on 1 and 16 DF,  p-value: 1.058e-08
```

Thông tin hiển thị mô tả phần dư (*residuals*), giá trị chẵn và hệ số góc $\bar{\alpha}, \bar{\beta}$, kiểm định $\beta = 0$ ($p=1.6e-08$) cho thấy giả thuyết $\beta = 0$ bị bác bỏ, nói cách khác là $chol$ và age có quan hệ tuyến tính.

Có thể tính \hat{y}_i cho từng cá thể bởi lệnh: *fitted()*

```
> reg<-lm(chol~age)
> fitted(reg)
     1      2      3      4      5      6      7      8 
3.747483 2.244985 4.094214 2.822869 4.383156 2.533927 2.707292 3.169600 
     9     10     11     12     13     14     15     16 
2.360562 3.574118 4.383156 2.996234 2.360562 4.729886 3.400753 3.863060 
     17     18 
2.707292 3.920849
```

(trong tính toán trên $\hat{y}_1 = 1.08922 + 0.05779 * 46 = 3.747483$)

Để tính phần dư cho từng cá thể sử dụng lệnh: `resid()`

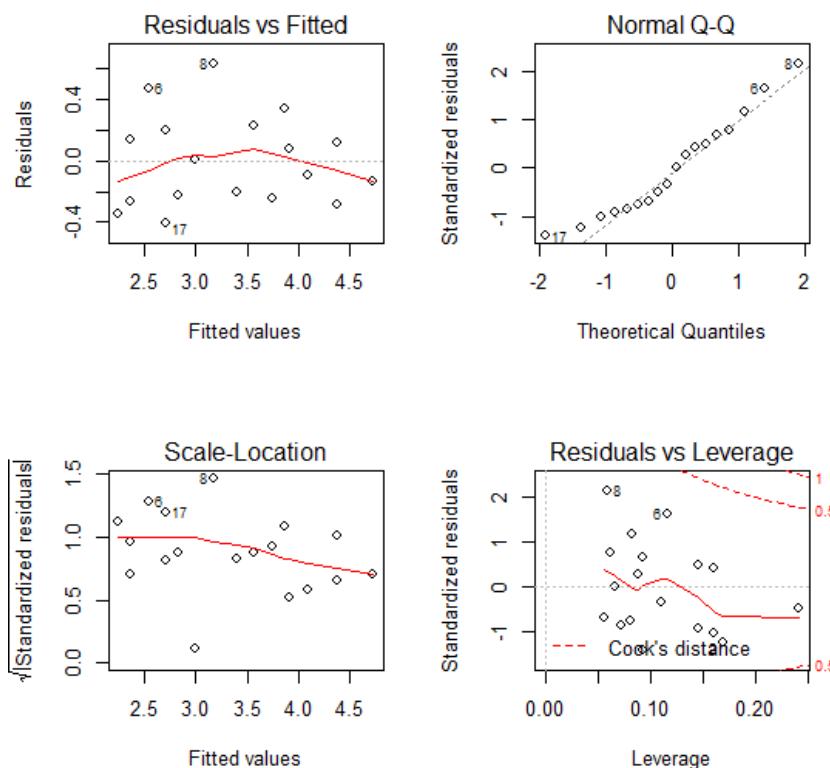
```
> resid(reg)
   1          2          3          4          5          6
-0.247483426 -0.344985415 -0.094213736 -0.222869265  0.116844338  0.466072660
   7          8          9         10         11         12
  0.192707505  0.630400424 -0.260562185  0.225881729 -0.283155662  0.003765579
  13         14         15         16         17         18
  0.139437815 -0.129885972 -0.200753116  0.336939804 -0.407292495  0.079151419
```

(ở đây $e_1 = 3.5 - 3.74748 = -0.24748$)

Hiển thị đồ thị phân tích phần dư kiểm tra các giả định trong phân tích hồi qui tuyến tính

```
> op<-par(mfrow=c(2,2))
> plot(reg)
```

Kết quả



Ý nghĩa đồ thị:

- Đồ thị bên trái dòng 1: vẽ phần dư e_i và giá trị tiên đoán \hat{y}_i . Trên đồ thị này các giá trị phần dư e_i tập trung xung quanh đường $y=0$ có nghĩa ε_i có giá trị trung bình bằng 0.
- Đồ thị bên phải dòng 1: vẽ giá trị phần dư và giá trị kỳ vọng dựa vào phân phối chuẩn. Trên đồ thị này, các giá trị phần dư tập trung gần các giá trị trên đường phân phối chuẩn, tức ε_i phân phối theo qui luật phân phối chuẩn.
- Đồ thị bên trái dòng 2 vẽ căn số phần dư chuẩn (*standardized residual*) và giá trị của \hat{y}_i . Đồ thị này cho thấy không có sự khác nhau giữa các số phần dư

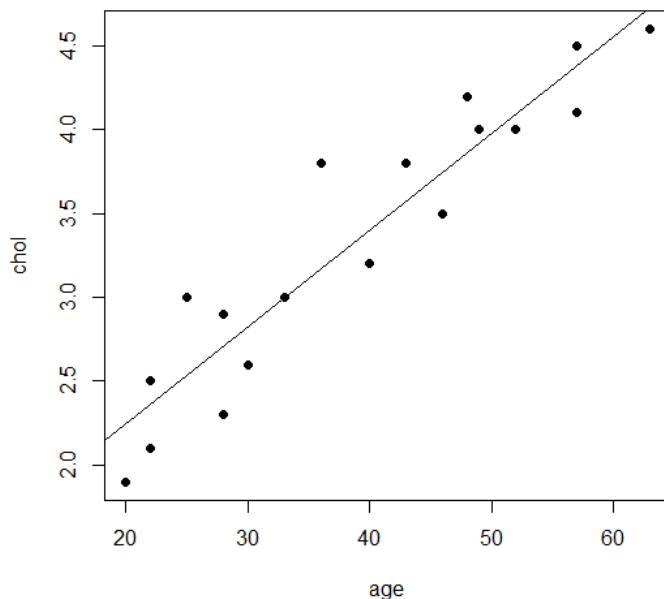
chuẩn cho các giá trị \hat{y}_i , có nghĩa các giá trị ε_i có phương sai σ^2 cố định cho tất cả các x_i .

9.2.3 Mô hình tiên đoán

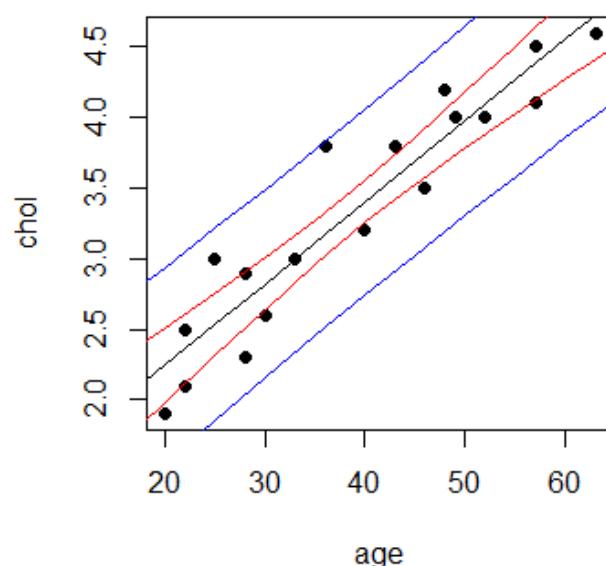
Sau khi kiểm định giữa *chol* và *age*, có thể vẽ đường biểu diễn mối liên hệ giữa *age* và *chol* bằng lệnh *abline()*

```
> plot(chol~age, pch=16)
> abline(reg)
```

Kết quả:



Do \hat{y}_i được tính từ $\hat{\alpha}, \hat{\beta}$, mà $\hat{\alpha}, \hat{\beta}$ đều có sai số chuẩn nên \hat{y}_i cũng có sai số. Khoảng tin cậy 95% ước lượng được thể hiện qua R bởi đồ thị xây dựng bằng các câu lệnh:



Biểu đồ trên hiển thị giá trị tiên đoán trung bình \hat{y}_i : đường màu đen (đường chính giữa), và khoảng tin cậy 95% của giá trị này là đường màu đỏ (hai đường đối xứng trong cùng). Đường màu xanh (hai đường đối xứng ngoài cùng) là khoảng tin cậy của giá trị tiên đoán cholesterol cho một độ tuổi mới. (*Cách thức ước lượng các khoảng tin cậy có thể tham khảo ở [7] trang 103-105*)

9.2.4 Hồi qui tuyến tính đa biến (Multiple linear regression)

Trong trường hợp tổng quát, một biến ngẫu nhiên Y có thể phụ thuộc tuyến tính với nhiều biến ngẫu nhiên X_1, X_2, \dots, X_k . Khi đó có thể biểu diễn dưới dạng

$$Y = \alpha + X\beta + \varepsilon$$

Với

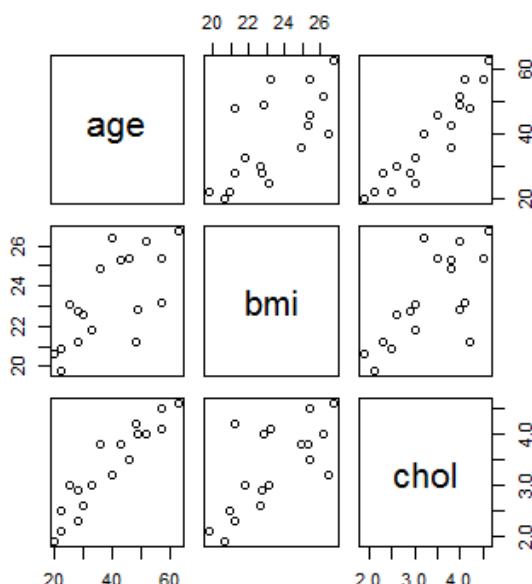
$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{bmatrix}, X = \begin{bmatrix} x_{11} & x_{21} & \dots & x_{k1} \\ x_{12} & x_{22} & \dots & x_{k2} \\ \dots & \dots & \dots & \dots \\ x_{1n} & x_{2n} & \dots & x_{kn} \end{bmatrix}, \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \dots \\ \beta_k \end{bmatrix}, \alpha = \begin{bmatrix} \alpha \\ \alpha \\ \dots \\ \alpha \end{bmatrix}, \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \dots \\ \varepsilon_n \end{bmatrix}$$

Để tìm các giá trị xấp xỉ $\hat{\alpha}, \hat{\beta}$ cũng dựa vào phương pháp bình phương bé nhất tương tự như trường hợp hồi qui tuyến tính đơn giản (*một biến*).

Để thấy biểu đồ biểu diễn mối quan hệ giữa các biến, dùng lệnh: *pairs(data)*

Ví dụ 98: Biểu đồ biểu diễn mối quan hệ giữa các biến *age*, *bmi*, *chol* với dữ liệu nêu ở ví dụ 95.

> *pairs(data)*



Tính toán các giá trị $\hat{\alpha}, \hat{\beta}$ cũng sử dụng câu lệnh trong ngôn ngữ R tương tự hồi qui tuyến tính đơn giản :*lm()*, *summary()*

Ví dụ 99: Tính toán $\hat{\alpha}, \hat{\beta}$ ứng với trường hợp tìm quan hệ tuyến tính giữa *chol* với *age* và *bmi*

```
> mreg<-lm(chol~age+bmi)
> summary(mreg)

Call:
lm(formula = chol ~ age + bmi)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.3762 -0.2259 -0.0534  0.1698  0.5679 

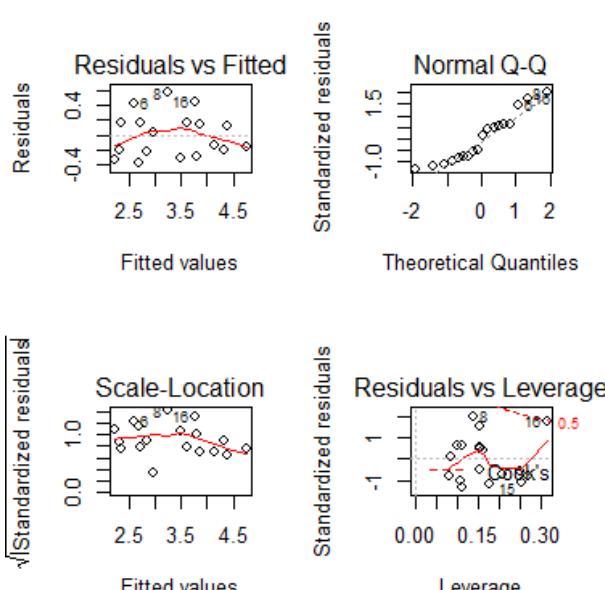
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 0.455458   0.918230   0.496   0.627    
age         0.054052   0.007591   7.120  3.5e-06 ***  
bmi         0.033364   0.046866   0.712   0.487    
---
Signif. codes:  0 '****' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3074 on 15 degrees of freedom
Multiple R-squared:  0.8815,   Adjusted R-squared:  0.8657 
F-statistic: 55.77 on 2 and 15 DF,  p-value: 1.132e-07
```

Với kết quả tính toán ta có:

$$chol = 0.455458 + 0.054052 * age + 0.033364 * bmi$$

Ví dụ 100: Hiển thị đồ thị phân tích phần dư kiểm tra các giả định trong phân tích hồi qui tuyến tính



```
> op<-par(mfrow=c(2,2))
> plot(mreg)
```

9.3 HỒI QUI ĐA THỨC (POLYNOMIAL REGRESSION ANALYSIS)

Trong thực tế, phần lớn các hàm thể hiện sự phụ thuộc giữa hai biến ngẫu nhiên X, Y thường là *hàm phi tuyến* (*non-linear function*).

Để thể hiện mối quan hệ phi tuyến, người ta thường sử dụng các hàm đa thức:

$$y_i = \alpha + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \dots + \beta_p x_i^p + \varepsilon_i$$

Với giả định $\varepsilon_i \sim \mathcal{N}(0; \sigma^2)$. Để ước tính các hệ số β_i người ta cũng sử dụng phương pháp bình phương bé nhất (xem [9]).

Trong ngôn ngữ R, sử dụng hàm `lm(<dữ liệu 1> ~ poly(<dữ liệu 2>, <bậc p>))` để tính các giá trị β_i .

Ví dụ 101: Sử dụng dữ liệu dưới đây tìm mối liên hệ giữa hàm lượng gỗ cứng (*hardwood concentration*) và độ căng (*tensile strength*) của vật liệu

Id	Hàm lượng gỗ cứng (x): <i>conc</i>	Độ căng mạnh (y): <i>strength</i>
1	1.0	6.3
2	1.5	11.1
3	2.0	20.0
4	3.0	24.0
5	4.0	26.1
6	4.5	30.0
7	5.0	33.8
8	5.5	34.0
9	6.0	38.1
10	6.5	39.9
11	7.0	42.0
12	8.0	46.1
13	9.0	53.1
14	10.0	52.0
15	11.0	52.5
16	12.0	48.0
17	13.0	42.8
18	14.0	27.8
19	15.0	21.9

- a. Thủ nghiệm với mối quan hệ tuyến tính đơn giản giữa *strength* và *conc*
- b. Thủ nghiệm với mối quan hệ đa thức bậc hai giữa *strength* và *conc*
- c. Thủ nghiệm với mối quan hệ đa thức bậc ba giữa *strength* và *conc*

- Xây dựng dữ liệu

```
> id<-1:19
> conc<-c(1.0,1.5,2.0,3.0,4.0,4.5,5.0,5.5,6.0,6.5,7.0,8.0,9.0,
+ 10.0,11.0,12.0,13.0,14.0,15.0)
> strength<-c(6.3,11.1,20.0,24.0,26.1,30.0,33.8,34.0,38.1,39.9,
+ 42.0,46.1,53.1,52.0,52.5,48.0,42.8,27.8,21.9)
> data<-data.frame(id,conc,strength)
```

- a. Thủ nghiệm với mối quan hệ tuyến tính đơn giản giữa *strength* và *conc*

```

> hq.dongian<-lm(strength~conc)
> summary(hq.dongian)

Call:
lm(formula = strength ~ conc)

Residuals:
    Min      1Q  Median      3Q     Max 
-25.986 -3.749  2.938  7.675 15.840 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 21.3213    5.4302   3.926  0.00109 **  
conc         1.7710    0.6478   2.734  0.01414 *   
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.82 on 17 degrees of freedom
Multiple R-squared:  0.3054,   Adjusted R-squared:  0.2645 
F-statistic: 7.474 on 1 and 17 DF,  p-value: 0.01414

```

- Kết quả giá trị ước lượng phương sai của mô hình là $s^2 = (11.82)^2 = 139.7$, vì vậy mối quan hệ tuyến tính là không phù hợp. Điều này cũng có thể thấy trực quan qua biểu đồ

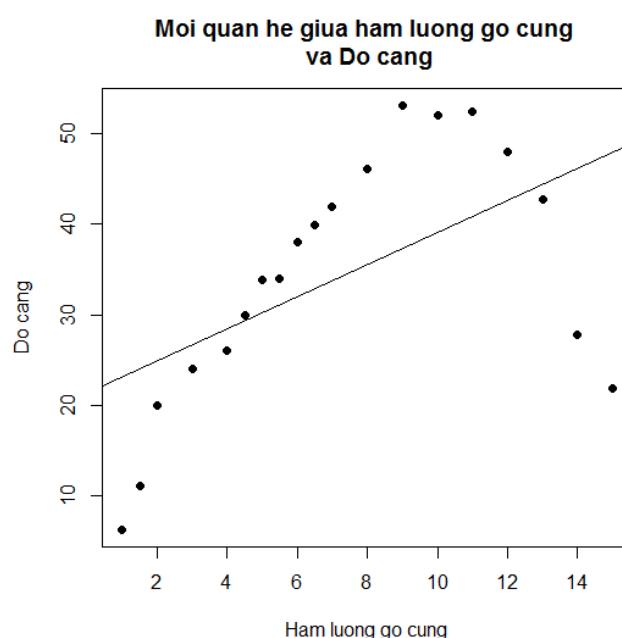
```

> plot(strength~conc, xlab="Ham luong go cung",
+       ylab="Do cang", main="Moi quan he giua ham luong go cung \n va Do cang",
+       pch=16)
> abline(hq.dongian)

```

- Kết quả cho các giá trị của mô hình có thể biểu diễn

$$y = 21.32 + 1.77 \cdot x$$



b. Thủ nghiệm với mối quan hệ đa thức bậc hai giữa *strength* và *conc*

```
> hq.bac2<-lm(strength~poly(conc,2))
> summary(hq.bac2)

Call:
lm(formula = strength ~ poly(conc, 2))

Residuals:
    Min      1Q  Median      3Q     Max 
-5.8503 -3.2482 -0.7267  4.1350  6.5506 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   34.184     1.014   33.709 2.73e-16 ***
poly(conc, 2)1 32.302     4.420    7.308 1.76e-06 ***
poly(conc, 2)2 -45.396     4.420   -10.270 1.89e-08 ***
---
Signif. codes:  0 '****' 0.001 '***' 0.01 '**' 0.05 '*' 0.1 '.' 1

Residual standard error: 4.42 on 16 degrees of freedom
Multiple R-squared:  0.9085,    Adjusted R-squared:  0.8971 
F-statistic: 79.43 on 2 and 16 DF,  p-value: 4.912e-09
```

- Kết quả cho các giá trị của mô hình có thể biểu diễn

$$y = 34.18 + 32.30 * x - 45.4 * x^2$$

giá trị ước lượng phương sai của mô hình là $s^2 = (4.42)^2 = 19.5$, có thể chấp nhận hơn mô hình quan hệ tuyến tính

c. Thủ nghiệm với mối quan hệ đa thức bậc ba giữa *strength* và *conc*

- Kết quả cho các giá trị của mô hình có thể biểu diễn

$$y = 34.1842 + 32.3021 * x - 45.3963 * x^2 - 14.5740 * x^3$$

giá trị ước lượng phương sai của mô hình là $s^2 = (2.59)^2 = 6.71$, có thể chấp nhận hơn hai mô hình trước

```

> hq.bac3<-lm(strength~poly(conc,3))
> summary(hq.bac3)

Call:
lm(formula = strength ~ poly(conc, 3))

Residuals:
    Min      1Q  Median      3Q     Max 
-4.6250 -1.6109  0.0413  1.5892  5.0216 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 34.1842   0.5931  57.641 < 2e-16 ***
poly(conc, 3)1 32.3021   2.5850 12.496 2.48e-09 ***
poly(conc, 3)2 -45.3963   2.5850 -17.561 2.06e-11 ***
poly(conc, 3)3 -14.5740   2.5850  -5.638 4.72e-05 ***
---
Signif. codes:  0 '****' 0.001 '***' 0.01 '**' 0.05 '*' 0.1 '.' 1

Residual standard error: 2.585 on 15 degrees of freedom
Multiple R-squared:  0.9707,    Adjusted R-squared:  0.9648 
F-statistic: 165.4 on 3 and 15 DF,  p-value: 1.025e-11

```

Vẽ biểu đồ ứng với ba mô hình:

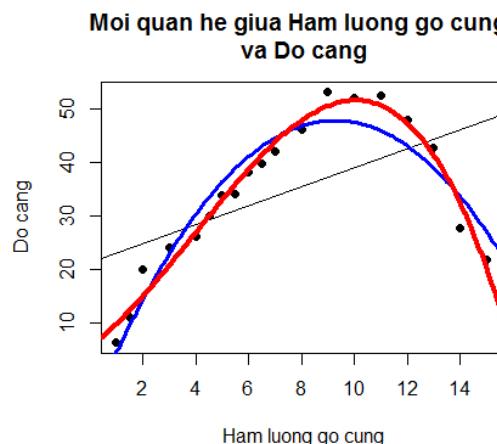
```

> hq.dg<-lm(strength~conc)
> hq.bac2<-lm(strength~poly(conc,2))
> hq.bac3<-lm(strength~poly(conc,3))
> # Tao cac bien lam tron do thi
> xnew<-(0:160)/10
> # Tinh cac gia tri tien doan cua y
> y2<-predict(hq.bac2,data.frame(conc=xnew))
> y3<-predict(hq.bac3,data.frame(conc=xnew))
> #ve bieu do
> plot(strength~conc, xlab="Ham luong go cung",
+ ylab="Do cang", main="Moi quan he giua Ham luong go cung \n va Do cang",
+ pch=16)
> abline(hq.dg,col="black")
> lines(xnew,y2,col="blue",lwd=3)
> lines(xnew,y3,col="red",lwd=4)

```

Kết quả hiển thị:

- Đường thẳng: Ứng với hồi qui tuyến tính đơn giản
- Đường cong đậm(màu xanh nhạt): ứng với hàm hồi qui bậc hai.
- Đường cong màu nhạt (màu đỏ): ứng với hàm hồi qui bậc ba.



TÀI LIỆU THAM KHẢO

- [1] Biostatistics with R, Babak Shahbaba, Springer, 2011
- [2] Giáo trình Lý thuyết Xác suất và thống kê toán, PGS.TS Nguyễn Cao Văn, TS Trần Thái Ninh, NXB Thống kê, 2005
- [3] Introduction to Probability and Statistics Using R, G.Jay Kerns, First Edition, cran.r-project.org/web/packages/IPSUR/vignettes/IPSUR.pdf, 2010
- [4] Introduction to Statistical Thinking (With R, Without Calculus), Benjamin, The Hebrew University, 2011.
- [5] Lý thuyết Xác suất và thống kê (bài giảng), Hoàng Văn Hà, ĐH KHTN Tp HCM, 2012
- [6] Phân tích dữ liệu với R, Nguyễn Văn Tuấn, NXB Tổng hợp Tp HCM, 2014
- [7] Simple-Using R for Introductory Statistics, John Verzani, 2001,
<https://cran.r-project.org/doc/contrib/Verzani-SimpleR.pdf>
- [8] Thống kê ứng dụng trong kinh tế-xã hội, Hoàng Trong, Chu Nguyễn Mộng Ngọc, NXB Lao động-Xã hội, 2010.
- [9] <http://home.iitk.ac.in/~shalab/regression/Chapter12-Regression-PolynomialRegression.pdf>

C H UƠNG

10

PHÂN TÍCH PHƯƠNG SAI

10.1 GIỚI THIỆU

Phân tích phương sai (*Analysis of Variance-ANOVA*) được phát triển bởi Ronald Fisher vào năm 1918 và là phần mở rộng của t-kiểm định (*t-test*) và z-kiểm định (*z-test*). Trước khi sử dụng ANOVA, t-kiểm định và z-kiểm định đã được sử dụng phổ biến, nhưng t-kiểm định không thể áp dụng cho nhiều hơn hai nhóm tổng thể.

Phân tích phương sai là phương pháp thống kê dùng để kiểm tra sự khác biệt giữa hai hay nhiều giá trị trung bình. Phân tích phương sai thông qua kiểm định giả thuyết thống kê để kết luận về sự bằng nhau của các số trung bình này.

Trong nghiên cứu, phân tích phương sai được dùng như là một công cụ để xem xét ảnh hưởng của một hay một số yếu tố nguyên nhân (*định tính*) đến một yếu tố kết quả (*định lượng*).

Ví dụ 102:

- Nghiên cứu ảnh hưởng của thời gian tự học đến kết quả học tập của sinh viên.
- Nghiên cứu ảnh hưởng của bậc thợ tới năng suất lao động.
- Nghiên cứu ảnh hưởng của phương pháp bán hàng, trình độ (kinh nghiệm) của nhân viên bán hàng đến doanh số

10.2 PHÂN TÍCH PHƯƠNG SAI MỘT YẾU TỐ (ONE-FACTOR/WAY ANOVA)

Phân tích phương sai một yếu tố là phân tích ảnh hưởng của một yếu tố nguyên nhân (dạng biến định tính định tính) đến một yếu tố kết quả (dạng biến định lượng) đang nghiên cứu.

Xét yếu tố \mathcal{A} được thể hiện qua k mức ứng với k tổng thể, minh họa qua bảng sau (*):

X_1	X_2	..	X_k	
x_{11}	x_{21}	..	x_{k1}	
x_{12}	x_{k2}	
...	x_{2n_2}	..	x_{k3}	
x_{1n_1}		
			x_{kn_k}	$n = \sum_{i=1}^k n_i$
$\bar{x}_1 = \sum_{j=1}^{n_1} x_{1j} / n_1$	$\bar{x}_2 = \sum_{j=1}^{n_2} x_{2j} / n_2$		$\bar{x}_k = \sum_{j=1}^{n_k} x_{kj} / n_k$	$\bar{x} = \sum_{i=1}^k n_i \bar{x}_i / n$

$S_1^2 = \frac{\left[\sum_{j=1}^{n_1} (x_{1j} - \bar{x}_1) \right]^2}{n_1 - 1}$	$S_2^2 = \frac{\left[\sum_{j=1}^{n_2} (x_{2j} - \bar{x}_2) \right]^2}{n_2 - 1}$	$S_k^2 = \frac{\left[\sum_{j=1}^{n_k} (x_{kj} - \bar{x}_k) \right]^2}{n_k - 1}$	
$SS_1 = \sum_{j=1}^{n_1} (x_{1j} - \bar{x}_1)$	$SS_2 = \sum_{j=1}^{n_2} (x_{2j} - \bar{x}_2)$	$SS_k = \sum_{j=1}^{n_k} (x_{kj} - \bar{x}_k)$	
<i>Tổng các chênh lệch trong nội bộ các nhóm</i>		$SSW = \sum_{i=1}^k SS_i = \sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2$	
<i>Tổng bình phương độ lệch riêng của các nhóm so với \bar{x}</i>		$SSG = \sum_{i=1}^k n_i (\bar{x}_i - \bar{x})^2$	
<i>Tổng các chênh lệch bình phương toàn bộ các tổng thể</i>		$SST = \sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \bar{x})^2$	

Có thể chứng minh: $SST = SSW + SSG$.

(Các đại lượng trên phát xuất từ biểu diễn: $x_{ij} = \bar{x} + (\bar{x}_i - \bar{x}) + (x_{ij} - \bar{x}_i)$ thành dạng tổng quát $X_{ij} = \mu + \alpha_i + \varepsilon_{ij}$)

Bài toán phân tích phương sai một yếu tố:

- Giả định:

- k tổng thể có phân phối chuẩn $\mathcal{N}(\mu, \sigma^2)$
- Lấy k mẫu độc lập từ k tổng thể, mỗi mẫu tuân theo phân phối chuẩn, mẫu j được quan sát n_j lần.
- Các phương sai tổng thể bằng nhau.

- Bài toán:

Ký hiệu μ_i là giá trị trung bình của tổng thể j . Kiểm định giả thuyết

$$\begin{cases} H_0 : \mu_1 = \mu_2 = \dots = \mu_k \\ H_1 : \exists i \neq j \mid \mu_i \neq \mu_j \end{cases}$$

- Các bước kiểm định:

1. Tính các đại lượng được nêu ở (*) trình bày trên.

2. Tính

▪ Phương sai trong nội bộ nhóm $MSW = \frac{SSW}{n - k}$

▪ Phương sai giữa các nhóm $MSG = \frac{SSG}{k - 1}$

3. Xác định mức ý nghĩa α

4. Kiểm định giả thuyết

Giả thuyết về sự bằng nhau của k trung bình tổng thể được quyết định dựa trên tỉ số của hai phương sai: phương sai giữa các nhóm (MSG) và phương sai trong nội bộ nhóm

(MSW). Tỉ số này thỏa qui luật Fisher-Snedecor với bậc tự do là $k-1$ ở tử số, $n-k$ ở mẫu số.

- Giả thuyết H_0 bị bác bỏ khi $F = \frac{MSG}{MSW} > F_{(k-1, n-k)}^{1-\alpha}$

Chú ý: Khi tính toán thủ công ta có thể áp dụng các công thức:

$$T_i = \sum_{j=1}^{n_k} x_{ij}; \quad T = \sum_{i=1}^k \sum_{j=1}^{n_i} x_{ij}; \quad SST = \sum_{i=1}^k \sum_{j=1}^{n_i} x_{ij}^2 - \frac{T^2}{n}; \quad SSG = \sum_{i=1}^k \frac{T_i^2}{n_i} - \frac{T^2}{n}$$

$$SSW = SST - SSG;$$

Ví dụ 103: Hàm lượng Alcaloid (mg) trong một loại dược liệu được thu hái từ 3 vùng khác nhau được số liệu sau:

Vùng 1 :	7,5	6,8	7,1	7,5	6,8	6,6	7,8
Vùng 2 :	5,8	5,6	6,1	6,0	5,7		
Vùng 3 :	6,1	6,3	6,5	6,4	6,5	6,3	

Hỏi hàm lượng Alcaloid có khác nhau theo vùng hay không?

Giải:

Ta có số liệu:

	Vùng 1	Vùng 2	Vùng 3	
	7.5	5.8	6.1	
	6.8	5.6	6.3	
	7.1	6.1	6.5	
	7.5	6.0	6.4	
	6.8	5.7	6.5	
	6.6		6.3	
	7.8			
n_i	7	5	6	$\sum_{i=1}^3 n_i = 18$
$\sum_{j=1}^{n_i} x_{ij}^2$	359,79	170,7	242,05	$\sum_{i=1}^3 \sum_{j=1}^{n_i} x_{ij}^2 = 772.54$
T_i	50.1	29.2	38.1	

Tính các đại lượng thống kê:

$$SST = 6.831111; SSG = 5.326968; SSW = 1.5041428; F = 26.561504$$

Kiểm định giả thiết:

$$\begin{cases} H_0 : \mu_1 = \mu_2 = \mu_3 \\ H_1 : \exists i \neq j \mid \mu_i \neq \mu_j \end{cases}; \quad \alpha = 0.05$$

Do $F_{(1-\alpha)}^{(k, n-k)} = 3.68$ nên H_0 bị bác bỏ nghĩa là hàm lượng Alcaloid có sai khác theo vùng.

Kết quả phân tích bằng ngôn ngữ R

- Tạo chức dữ liệu

- Khai báo 3 nhóm tổng thể lần lượt có số lượng 7,5,6
- Đại lượng thể hiện nhóm của các cá thể phải là kiểu *factor*

```
> s1<-rep(1,7);s2<-rep(2,5);s3<-rep(3,6)
> group<-c(s1,s2,s3)
> group<-as.factor(group)
> alcal<-c(7.5,6.8,7.1,7.5,6.8,6.6,7.8,
+ 5.8,5.6,6.1,6.0,5.7,
+ 6.1,6.3,6.5,6.4,6.5,6.3)
> data<-data.frame(group,alcal)
> attach(data)
```

Xem alcal là một hàm tuyến tính ứng với biến group

Và gọi hàm *anova()* để có kết quả phân tích anova

```
> analysis<- lm(alcal~group)
> anova(analysis)
Analysis of Variance Table

Response: alcal
          Df Sum Sq Mean Sq F value    Pr(>F)
group      2 5.3270 2.66348  26.561 1.178e-05 ***
Residuals 15 1.5041 0.10028
---
Signif. codes:  0 '****' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Giá trị $F_{(1-\alpha)}^{(2,15)}$ có thể tính bởi hàm: $qf(1-\alpha, k, n-k) = qf(1-0.05, 2, 15)$

```
> qf(1-0.05, 2, 15)
[1] 3.68232
```

Vì $F > F_{(1-0.05)}^{(2,15)}$ nên giả thuyết H_0 bị bác bỏ.

10.3 PHÂN TÍCH PHƯƠNG SAI HAI YẾU TỐ (TWO-FACTOR/WAY ANOVA)

Phân tích phương sai 2 yếu tố nhằm xem xét cùng lúc hai yếu tố nguyên nhân (dưới dạng dữ liệu định tính) ảnh hưởng đến yếu tố kết quả (dưới dạng dữ liệu định lượng) đang nghiên cứu. Ví dụ: Nghiên cứu ảnh hưởng của *thời gian tự học* và *mức độ yêu thích ngành học* đến kết quả học tập của sinh viên.

Xét hai yếu tố \mathcal{A} , \mathcal{B} , yếu tố thứ nhất (\mathcal{A}) được thể hiện qua k nhóm, yếu tố thứ hai (\mathcal{B}) được thể hiện qua h khối được minh họa qua bảng sau (**):

yếu tố thứ hai (\mathcal{B}) được thể hiện qua h khối	yếu tố thứ nhất (\mathcal{A}) được thể hiện qua k nhóm			
	1	2	..	K
1	$x_{111}x_{112}\dots x_{11L}$	$x_{211}x_{212}\dots x_{21L}$..	$x_{K11}x_{K12}\dots x_{K1L}$
2	$x_{121}x_{122}\dots x_{12L}$	$x_{221}x_{222}\dots x_{22L}$..	$x_{K21}x_{K22}\dots x_{K2L}$
..
H	$x_{1H1}x_{1H2}\dots x_{1HL}$	$x_{2H1}x_{2H2}\dots x_{2HL}$..	$x_{KH1}x_{KH2}\dots x_{KHL}$

Chú ý: Trong trường hợp

- a. L=1: **phân tích phương sai hai yếu tố không lặp**
- b. L>1: **phân tích phương sai hai yếu tố có lặp**

Bài toán phân tích phương sai hai yếu tố:

- Giả định:

- Các mẫu có phân phối chuẩn $\mathcal{N}(\mu, \sigma^2)$
- Lấy K mẫu độc lập từ K tổng thể, H mẫu độc lập từ H tổng thể

- Bài toán:

Ký hiệu μ_j là giá trị trung bình của nhóm (tổng thể) j ($j = \overline{1, K}$)

Ký hiệu μ_j^* là giá trị trung bình của khối (tổng thể) j ($j = \overline{1, H}$)

Kiểm định ba giả thuyết

- | | | | |
|---|---|--|---|
| a. | $\begin{cases} H_0 : \mu_1 = \mu_2 = \dots = \mu_K \\ H_1 : \exists i \neq j \mid \mu_i \neq \mu_j \end{cases}$ | b. | $\begin{cases} H_0 : \mu_1^* = \mu_2^* = \dots = \mu_H^* \\ H_1 : \exists i \neq j \mid \mu_i^* \neq \mu_j^* \end{cases}$ |
| Hai yếu tố \mathcal{A}, \mathcal{B} có ảnh hưởng như nhau đến yếu tố
kết quả | | | |
| c. | $\begin{cases} H_0 : \\ H_1 : \end{cases}$ | Hai yếu tố \mathcal{A}, \mathcal{B} có ảnh hưởng khác nhau đến
yếu tố kết quả | |

Với mức ý nghĩa α

- Các bước kiểm định:

1. Tính các trung bình

- Trung bình mẫu của từng nhóm (*cột*)

$$\bar{x}_i = \frac{\sum_{j=1}^H \sum_{s=1}^L x_{ijs}}{H * L}; \quad i = \overline{1, K}$$

- Trung bình mẫu của từng khối (*dòng*)

$$\bar{x}_j^* = \frac{\sum_{i=1}^K \sum_{s=1}^L x_{ijs}}{K * L}; \quad j = \overline{1, H}$$

- Trung bình mẫu của từng ô

$$\bar{x}_{ij} = \frac{\sum_{s=1}^L x_{ijs}}{L}$$

- Trung bình chung của toàn bộ quan sát

$$\bar{x} = \frac{\sum_{i=1}^K \sum_{j=1}^H \sum_{s=1}^L x_{ijs}}{K * H * L}$$

2. Tính tổng các chênh lệch bình phương

1. Tổng các chênh lệch bình phương toàn bộ

$$SST = SSG + SSB + SSI + SSE = \sum_{i=1}^K \sum_{j=1}^H \sum_{s=1}^L (x_{ijs} - \bar{x})^2$$

2. Tổng các chênh lệch bình phương giữa các nhóm (between –groups)

$$SSG = H * L * \sum_{i=1}^K (\bar{x}_i - \bar{x})^2$$

SSG: phản ánh phần biến thiên của yếu tố định lượng kết quả đang nghiên cứu do ảnh hưởng của yếu tố nguyên nhân thứ nhất (dùng để phân nhóm ở cột)

3. Tổng các chênh lệch bình phương giữa các khối (between –blocks)

$$SSB = K * L * \sum_{j=1}^H (\bar{x}_j^* - \bar{x})^2$$

SSB: phản ánh phần biến thiên của yếu tố định lượng kết quả đang nghiên cứu do ảnh hưởng của yếu tố nguyên nhân thứ hai (dùng để phân khối ở dòng)

Nếu trong trường hợp phân tích phương sai hai yếu tố không lặp, không xét 4.

4. Tổng các chênh lệch bình phương giữa các ô (giao giữa nhóm và khối)

$$SSI = L * \sum_{i=1}^K \sum_{j=1}^H (\bar{x}_{ij} - \bar{x}_i - \bar{x}_j^* + \bar{x})^2$$

SSI: phản ánh phần biến thiên do tác động giữa hai yếu tố đang nghiên cứu.

5. Tổng các chênh lệch bình phương phần dư

Nếu phân tích phương sai hai yếu tố có lặp:

$$SSE = \sum_{i=1}^K \sum_{j=1}^H \sum_{s=1}^L (x_{ijs} - \bar{x}_{ij})^2 = SST - SSG - SSB - SSI$$

Nếu phân tích phương sai hai yếu tố không lặp:

$$SSE = \sum_{i=1}^K \sum_{j=1}^H (\bar{x}_{ij} - \bar{x}_i - \bar{x}_j^* + \bar{x})^2 = SST - SSG - SSB$$

3. Tính các phương sai

1. Phương sai giữa các nhóm: $MSG = \frac{SSG}{K-1}$

2. Phương sai giữa các khối: $MSB = \frac{SSB}{H-1}$

Nếu phân tích phương sai hai yếu tố không lặp không tính 3:

$$3. Phuong sai giua cac ô: MSI = \frac{SSI}{(K-1)*(H-1)}$$

4. Phuong sai du:

Nếu phân tích phương sai hai yếu tố có lắp:

$$MSE = \frac{SSE}{K * H * (L-1)}$$

Nếu phân tích phương sai hai yếu tố không lắp:

$$MSE = \frac{SSE}{(K-1)*(H-1)}$$

4. Tính các yếu tố kiểm định

$$F_1 = \frac{MSG}{MSE}, tỉ số này thỏa qui luật Fisher-Snedecor với bậc tự do$$

- K-1, KH(L-1), *Nếu phân tích phương sai hai yếu tố có lắp*
- K-1,(K-1)(H-1), *Nếu phân tích phương sai hai yếu tố không lắp*

$$F_2 = \frac{MSB}{MSE}, tỉ số này thỏa qui luật Fisher-Snedecor với bậc tự do$$

- H-1, KH(L-1), *Nếu phân tích phương sai hai yếu tố có lắp*
- H-1,(K-1)(H-1), *Nếu phân tích phương sai hai yếu tố không lắp*

Nếu phân tích phương sai hai yếu tố không lắp, không tính đại lượng F₃.

$$F_3 = \frac{MSI}{MSE}, tỉ số này thỏa qui luật Fisher-Snedecor với bậc tự do$$

$$(K-1)(H-1), KH(L-1)$$

5. Kết luận

1. Giả thuyết H_0 của kiểm định a. là bác bỏ khi

$$F_1 > F_{1-\alpha}^{K-1, KH(L-1)}, *Nếu phân tích phương sai hai yếu tố có lắp*$$

$$F_1 > F_{1-\alpha}^{K-1, (K-1)(H-1)}, *Nếu phân tích phương sai hai yếu tố không lắp*$$

2. Giả thuyết H_0 của kiểm định b. là bác bỏ khi

$$F_2 > F_{1-\alpha}^{H-1, KH(L-1)}, *Nếu phân tích phương sai hai yếu tố có lắp*$$

$$F_2 > F_{1-\alpha}^{H-1, (K-1)(H-1)}, *Nếu phân tích phương sai hai yếu tố không lắp*$$

3. Giả thuyết H_0 của kiểm định c. là bác bỏ khi

$$F_3 > F_{1-\alpha}^{(K-1)(H-1), KH(L-1)}, *Nếu phân tích phương sai hai yếu tố có lắp*$$

Nếu phân tích phương sai hai yếu tố không lắp, không kiểm định c.

Ví dụ 104: Giả sử ta có dữ liệu thể hiện điểm trung bình của sinh viên phân theo thời gian tự học và mức độ yêu thích ngành học.

1. Xét xem điểm trung bình học tập của sinh viên có thời gian tự học khác nhau có khác nhau không?

2. Điểm trung bình học tập của sinh viên có mức độ yêu thích ngành đang học có khác nhau không?
3. Giữa thời gian tự học và mức độ yêu thích ngành học có ảnh hưởng tương tác với nhau không?

Mức độ yêu thích Ngành học	Thời gian tự học		
	Ít giờ	Trung bình	Nhiều giờ
Không thích	5.8	6.0	6.2
	6.2	6.6	5.8
	5.4	6.1	6.5
	6.0	5.8	6.2
	5.2	5.9	6.4
	5.3	6.0	5.7
	5.4	5.9	6.1
Thích	5.6	6.0	6.8
	6.2	6.7	7.1
	5.7	6.5	6.5
	5.5	6.3	7.1
	6.1	6.1	7.2
	6.0	6.8	6.7
	5.2	6.4	7.0
Rất thích	6.4	6.8	7.6
	5.5	6.6	7.7
	5.0	6.4	7.8
	5.6	6.2	6.8
	6.2	7.1	7.3
	6.1	7.0	7.1
	5.3	7.2	7.2

Giải:

0. Xác định giả thuyết H_0 ứng với ba yêu cầu của đề bài

- a. Điểm trung bình học tập của sinh viên có thời gian tự học khác nhau đều bằng nhau
- b. Điểm trung bình học tập của sinh viên có mức độ yêu thích ngành học khác nhau đều bằng nhau
- c. Không có ảnh hưởng tương tác giữa thời gian tự học và mức độ yêu thích ngành đang học của sinh viên.

Mức ý nghĩa $\alpha = 0.05$

1. Tính các giá trị trung bình

Trung bình mẫu của từng nhóm (Group means)

- Điểm trung bình của nhóm thời gian tự học ít

$$\bar{x}_1 = \frac{5.8 + 6.2 + 5.4 + \dots + 6.2 + 6.1 + 5.3}{3 * 7} = 5.7$$

- Điểm trung bình của nhóm thời gian tự học trung bình

$$\bar{x}_2 = 6.4$$

- Điểm trung bình của nhóm thời gian tự học nhiều

$$\bar{x}_3 = 6.8$$

Trung bình mẫu của từng khối (Block means)

- Điểm trung bình của khối không yêu ngành học

$$\bar{x}_1^* = \frac{5.8 + 6.0 + 6.2 + \dots + 5.4 + 5.9 + 6.1}{3 * 7} = 5.93$$

- Điểm trung bình của khối yêu ngành học

$$\bar{x}_2^* = 6.36$$

- Điểm trung bình của khối rất yêu ngành học

$$\bar{x}_3^* = 6.6$$

Trung bình mẫu của từng ô (Cell means)

		Thời gian tự học			x_j
		Ít	TBình	Nhiều	
Yêu thích	Không	5.61	6.04	6.13	5.93
	Thích	5.76	6.40	6.91	6.36
	Rất	5.73	6.76	7.36	6.60
	x_i	5.70	6.40	6.80	6.30

2. Tính tổng các chênh lệch bình phương

$$SST = 27.02; SSG = 13.02; SSB = 4.84; SSI = 2.23; SSE = 6.93$$

3. Tính các phương sai

$$MSG = \frac{13.02}{3-1} = 6.52; MSB = \frac{4.84}{3-1} = 2.42; MSI = \frac{2.23}{(3-1)(3-1)} = 0.558$$

$$MSE = \frac{6.93}{3 * 3 * (7-1)} = 0.128$$

4. Tính các tỷ số F

$$F_1 = 50.86 > F_{1-0.05}^{2,54} = 3.17$$

$$F_2 = 18.91 > F_{1-0.05}^{2,54} = 3.17$$

$$F_3 = 4.36 > F_{1-0.05}^{4,54} = 2.54$$

Kết luận:

1. Bác bỏ giả thuyết H_0 trường hợp a: Điểm trung bình của sinh viên có thời gian tự học khác nhau là khác nhau. Nói khác hơn, thời gian tự học có ảnh hưởng đến điểm trung bình.
2. Bác bỏ giả thuyết H_0 trường hợp b: Điểm trung bình của sinh viên có thích mức độ yêu thích ngành học khác nhau là khác nhau. Nói khác hơn, có ảnh hưởng đến điểm trung bình.

3. Bác bỏ giả thuyết H_0 trường hợp c: Thời gian tự học và mức độ yêu thích ngành học có ảnh hưởng tương tác với nhau.

Thực hiện bằng ngôn ngữ R:

Trong ngôn ngữ R, hàm *anova()* chỉ trả lời được 2 kiểm định a, b. Còn kiểm định c hàm *anova()* không trả lời. Với ví dụ trên thực hiện bằng ngôn ngữ R như sau:

- Xây dựng dữ liệu

```
> tuhoc<-gl(3,21,63)
> yeuthich<-gl(3,7,63)
> id<-1:63
> dtb<-c(5.8,6.2,5.4,6.0,5.2,5.3,5.4,
+ 5.6,6.2,5.7,5.5,6.1,6.0,5.2,
+ 6.4,5.5,5.0,5.6,6.2,6.1,5.3,
+ 6.0,6.6,6.1,5.8,5.9,6.0,5.9,
+ 6.0,6.7,6.5,6.3,6.1,6.8,6.4,
+ 6.8,6.6,6.4,6.2,7.1,7.0,7.2,
+ 6.2,5.8,6.5,6.2,6.4,5.7,6.1,
+ 6.8,7.1,6.5,7.1,7.2,6.7,7.0,
+ 7.6,7.7,7.8,6.8,7.3,7.1,7.2)
> data<-data.frame(tuhoc,yeuthich,id,dtb)
> attach(data)
```

- Phân tích dữ liệu bằng hàm *anova()*

```
> twoway<-lm(dtb ~ tuhoc+yeuthich)
> anova(twoway)
Analysis of Variance Table

Response: dtb
          Df Sum Sq Mean Sq F value    Pr(>F)
tuhoc      2   13.02   6.5100  42.141 4.990e-12 ***
yeuthich   2     5.04   2.5200  16.312 2.395e-06 ***
Residuals 58    8.96   0.1545
---
Signif. codes:  0 '****' 0.001 '***' 0.01 '**' 0.05 '*' 0.1 ' ' 1
```

Giá trị $F_{1-0.05}^{2,54} = qf(1-0.05, 2, 54) = 3.17$

```
> qf(1-0.05, 2, 54)
[1] 3.168246
```

Chú ý: các giá trị tính toán so với tính toán thủ công trên có thể không trùng khớp do quá trình làm tròn và sai số.

Kết luận:

- yếu tố tự học có $F_1=42.141 > F_{1-0.05}^{2,54} = qf(1-0.05, 2, 54) = 3.17$
- yếu tố yêu thích $F_2=16.312 > F_{1-0.05}^{2,54} = qf(1-0.05, 2, 54) = 3.17$

Nên giả thuyết H_0 cho trường hợp a, b đều bị bác bỏ.

Có thể sử dụng p-giá trị hiển thị ở kết quả để có được kết quả kiểm định.

Trong ngôn ngữ R có nhiều hàm khác để hỗ trợ phân tích phương sai.[7,9]

(Giáo trình này chỉ giới thiệu các hàm thông dụng).

TÀI LIỆU THAM KHẢO

- [1] Biostatistics with R, Babak Shahbaba, Springer, 2011
- [2] <http://files.tranminhtam74.webnode.vn/200000029-1286c13824/CHUONG%204-KIEM%20DINH%20GIA%20THUYET%20THONG%20KE.ppt>
- [3] Introduction to Probability and Statistics Using R, G.Jay Kerns, First Edition, *cran.r-project.org/web/packages/IPSUR/vignettes/IPSUR.pdf*, 2010
- [5] Introduction to Statistical Thinking (With R, Without Calculus), Benjamin, The Hebrew University, 2011.
- [6] Lý thuyết Xác suất và thống kê (bài giảng), Hoàng Văn Hà, ĐH KHTN Tp HCM, 2012
- [7] Phân tích dữ liệu với R, Nguyễn Văn Tuấn, NXB Tổng hợp Tp HCM, 2014
- [8] Simple-Using R for Introductory Statistics, John Verzani, 2001,
<https://cran.r-project.org/doc/contrib/Verzani-SimpleR.pdf>
- [9] Thống kê ứng dụng trong kinh tế-xã hội, Hoàng Trong, Chu Nguyễn Mộng Ngọc, NXB Lao động-Xã hội, 2010.
- [10] www2.hcmuaf.edu.vn/data/dtdanh/Anova.pdf

MỤC LỤC

LỜI MỞ ĐẦU

CHƯƠNG 1: GIỚI THIỆU NGÔN NGỮ R	1
1.1 NGÔN NGỮ R	1
1.2 DỮ LIỆU TRONG R	2
1.2.1 Các đối tượng cơ bản.....	2
1.2.2 Các phép toán cơ bản.....	8
1.2.3 Biểu thức (<i>Expression</i>)	8
1.2.4 Chuyển đổi đối tượng (<i>Converting object</i>)	9
1.2.5 Câu lệnh IF..ELSE.....	9
1.2.6 Câu lệnh SWITCH	9
1.2.7 Câu lệnh lặp repeat, while, for	10
1.2.8 Hàm tự xây dựng:	11
1.3 XUẤT/NHẬP DỮ LIỆU (DATA IMPORT/EXPORT).....	11
CHƯƠNG II: THỐNG KÊ HỌC.....	13
2.1 THỐNG KÊ	13
2.2 NGUỒN GỐC VÀ SỰ PHÁT TRIỂN CỦA THỐNG KÊ HỌC	13
2.3 CHỨC NĂNG CỦA THỐNG KÊ	13
2.4 MỘT SỐ KHÁI NIỆM CƠ BẢN TRONG THỐNG KÊ	14
2.5 QUÁ TRÌNH NGHIÊN CỨU THỐNG KÊ	16
2.6 CÁC KỸ THUẬT LẤY MẪU	16
2.6.1 Kỹ thuật lấy mẫu xác suất (<i>Probability sampling</i>)	16
2.6.2 Lấy mẫu phi xác suất (<i>Non-probability sampling</i>).....	17
CHƯƠNG III: TRÌNH BÀY DỮ LIỆU BẰNG BẢNG & ĐỒ THỊ.....	19
3.1 TÓM TẮT VÀ TRÌNH BÀY DỮ LIỆU BẰNG BẢNG TẦN SỐ	19
3.2 PHÂN TỔ THỐNG KÊ	19
3.3 TÓM TẮT VÀ TRÌNH BÀY DỮ LIỆU BẰNG BIỂU ĐỒ NHÁNH VÀ LÁ	21

3.4 TÓM TẮT VÀ TRÌNH BÀY DỮ LIỆU BẰNG ĐỒ THỊ	22
3.4.1 Biểu đồ phân phối tần số (<i>Histogram</i>)	22
3.4.2 Biểu đồ thanh (<i>Bar chart</i>).....	23
3.4.3 Biểu đồ dạng tròn (<i>Piecharts</i>)	25
3.4.4 Biểu đồ dạng hộp (<i>Boxplots</i>)	27
3.4.5 Biểu đồ đường (<i>Line graphs</i>)	28
3.4.6 Biểu đồ điểm (<i>Dot chart</i>)	29
3.4.7 Biểu đồ tán xạ (<i>Scatter plot</i>).....	30
3.5 TIỀN XỬ LÝ DỮ LIỆU (DATA PREPROCESSING)	31
3.6 TỔNG KẾT.....	35
CHƯƠNG IV: TÓM TẮT DỮ LIỆU	37
4.1 CÁC ĐẠI LƯỢNG ĐO LUỒNG MỨC ĐỘ TẬP TRUNG CỦA DỮ LIỆU.....	37
4.1.1 Các đại lượng đo lường độ tập trung phổ biến.....	37
4.1.2 Nhóm các đại lượng khác mô tả sự phân bố tập trung	41
4.2 CÁC ĐẠI LƯỢNG ĐO LUỒNG MỨC DỘ PHÂN TÁN CỦA DỮ LIỆU	42
4.2.1 Tầm/Khoảng biến thiên (<i>Range</i>)	42
4.2.2 Độ trai giữa (<i>Interquartile Range</i>) IQR.....	43
4.2.3 Phương sai và độ lệch chuẩn	44
4.2.4 Hệ số biến thiên CV(<i>Coefficient of Variance</i>)	45
4.3 PHÂN TÍCH ĐỘ TẬP TRUNG VÀ PHÂN TÁN DỮ LIỆU KHI DỮ LIỆU CÓ PHÂN TỐ (HAY CÓ TRỌNG SỐ).....	45
4.4 CÁC ỨNG DỤNG CỦA THÔNG KÊ MÔ TẢ	47
4.4.1 Quan hệ thực nghiệm giữa trung bình, trung vị và yếu vị	48
4.4.2 Định lý Chebychev và ước tính miền giá trị và khoảng tin cậy	48
4.4.3 Phép biến đổi	48
CHƯƠNG V: XÁC SUẤT	50
5.1 MỘT SỐ KHÁI NIỆM CƠ SỞ.....	50
5.2 BIẾN CÓ.....	51
5.3 CÁC PHÉP ĐÉM.....	54
5.4 XÁC SUẤT.....	54

5.4.1 Định nghĩa xác suất cỗ điển.....	54
5.4.2 Định nghĩa xác suất theo quan điểm thống kê.....	54
5.5 NHẮC LẠI MỘT SỐ TÍNH CHẤT CỦA XÁC SUẤT.....	55
5.6 XÁC SUẤT CÓ ĐIỀU KIỆN.....	56
5.6.1 Định nghĩa	56
5.6.2 Qui tắc nhân.....	56
5.6.3 Qui tắc xác suất đầy đủ.....	56
5.6.4 Định lý Bayes	56
5.6.5 Xác suất trong ngôn ngữ R	57
CHƯƠNG VI: BIẾN NGẪU NHIÊN & PHÂN PHỐI XÁC SUẤT	58
6.1 BIẾN NGẪU NGHIÊN (<i>Random variable</i>).....	58
6.1.1 Định nghĩa	58
6.1.2 Phân loại biến ngẫu nhiên.....	58
6.2 PHÂN PHỐI XÁC SUẤT	59
6.3 CÁC ĐẶC TRƯNG SÓ CỦA BIẾN NGẪU NHIÊN	60
6.3.1 Kỳ vọng của biến ngẫu nhiên	60
6.3.2 Phương sai của biến ngẫu nhiên	60
6.3.3. Một số ví dụ.....	61
6.4 PHÂN PHỐI XÁC SUẤT RỒI RẠC	65
6.4.1 Phân phối Bernoulli (<i>Bernoulli Distribution</i>)	65
6.4.2 Phân phối nhị thức (<i>Binomial Distribution</i>).....	66
6.4.3 Phân phối Poisson (<i>Poisson Distribution</i>).....	68
6.5 PHÂN PHỐI XÁC SUẤT LIÊN TỤC	71
6.5.1 Phân phối đều (<i>Uniform distribution</i>)	71
6.5.2 Phân phối mũ (<i>Exponential distribution</i>)	72
6.5.3 Phân phối chuẩn (<i>Normal distribution</i>)	73
6.5.4 Phân phối “Khi bình phương” χ^2	80
6.5.5 Phân phối Student.....	81
6.5.6 Phân phối Fisher-Snedecor.....	82
6.6 PHÂN PHỐI XÁC SUẤT CỦA ĐẠI LUỢNG NGẪU NHIÊN 2 CHIỀU	83

6.6.1. Định nghĩa	83
6.6.2. Tính chất.....	84
6.6.3. Định nghĩa	84
6.6.4. Phân phối xác suất của đại lượng ngẫu nhiên 2 chiều rời rạc	84
6.6.5. Phân phối xác suất của đại lượng ngẫu nhiên 2 chiều liên tục.....	88
6.6.6 Hệ số tương quan	89
CHƯƠNG VII: ƯỚC LUỢNG.....	91
7.1 LÝ THUYẾT MẪU.....	91
7.2 PHÂN PHỐI MẪU.....	91
7.3 ƯỚC LUỢNG THAM SỐ.....	92
7.4 ƯỚC LUỢNG ĐIỂM (POINT ESTIMATOR).....	92
7.5 ƯỚC LUỢNG KHOẢNG (INTERVAL ESTIMATOR).....	98
7.5.1 Định nghĩa	98
7.5.2 Khoảng tin cậy.....	98
7.5.3 Ước lượng khoảng cho giá trị trung bình	99
7.6 KHOẢNG TIN CẬY CHO TỶ LỆ TỔNG THÊ	102
7.7 XÁC ĐỊNH KÍCH THƯỚC MẪU	104
CHƯƠNG VIII: KIỂM ĐỊNH GIẢ THUYẾT THỐNG KÊ	106
8.1 GIỚI THIỆU	106
8.1.1 Các khái niệm cơ bản	106
8.1.2 Sai lầm loại I và loại II	108
8.1.3 Số liệu để ước tính cỡ mẫu	108
8.1.4 Phương pháp tiếp cận chấp nhận hay bác bỏ một giả thuyết thống kê	109
8.2 KIỂM ĐỊNH GIẢ THUYẾT CHO TRƯỜNG HỢP 1 MẪU	110
8.2.1 Kiểm định giả thuyết cho kỳ vọng	110
8.2.2 Kiểm định cho phương sai.....	116
8.2.3 Kiểm định giả thuyết cho tỷ lệ	118
8.3 KIỂM ĐỊNH GIẢ THUYẾT CHO TRƯỜNG HỢP HAI MẪU ĐỘC LẬP	120
8.3.1 So sánh hai kỳ vọng, trường hợp biết phương sai	120
8.3.2 So sánh hai kỳ vọng, trong trường hợp chưa biết phương sai.....	122

8.3.3 So sánh 2 phương sai	123
8.3.4 So sánh hai tỷ lệ.....	125
8.4 KIỂM ĐỊNH GIẢ THUYẾT CHO TRƯỜNG HỢP HAI MẪU KHÔNG ĐỘC LẬP	127
8.5 KIỂM ĐỊNH CHI BÌNH PHƯƠNG.....	129
8.5.1 Kiểm định giả thuyết về phân phối	130
8.5.2 Kiểm định giả thuyết về tính độc lập	131
CHƯƠNG X: HỒI QUI & TƯƠNG QUAN	135
9.1 TƯƠNG QUAN	135
9.1.1 Hệ số tương quan Pearson r	135
9.1.2 Hệ số tương quan Spearman ρ	136
9.1.3 Hệ số tương quan Kendall τ	137
9.2 HỒI QUI (REGRESSION)	137
9.2.1 Hồi qui tuyến tính đơn giản.....	138
9.2.2 Phân tích hồi qui tuyến tính đơn giản bằng R	138
9.2.3 Mô hình tiên đoán.....	141
9.2.4 Hồi qui tuyến tính đa biến	142
9.3 HỒI QUI ĐA THÚC (POLYNOMIAL REGRESSION ANALYSIS).....	143
CHƯƠNG XI: PHÂN TÍCH PHƯƠNG SAI.....	149
10.1 GIỚI THIỆU	149
10.2 PHÂN TÍCH PHƯƠNG SAI MỘT YẾU TỐ (ONE-FACTOR/WAY ANOVA) ..	149
10.3 PHÂN TÍCH PHƯƠNG SAI HAI YẾU TỐ	152
MỤC LỤC	161