**Tóm tắt:**

## C1 Introduction and Framework

## 1.1 Population, Sample, and Observations

Nghiên cứu quan sát: một nghiên cứu được thực hiện trong đó nghiên cứu không có sự kiểm soát đối với các yếu tố được nghiên cứu.

Dân số: Vũ trụ của các giá trị tiềm năng mà từ đó một mẫu được rút ra.

Mẫu xác suất: mẫu trong đó mọi thành viên trong quần thể đều có xác suất được đưa vào mẫu.

### 1.2 Variables

Là đại lượng có giá trị luôn thay đổi trong quá trình thực hiện chương trình

#### Example 1.2.1

Biến định tính: **là biến miêu tả các giá trị đo lường bằng các chữ cái, chữ số hay ký hiệu được xếp vào các nhóm khác nhau**. + Biến thứ hạng: khi các nhóm biến được sắp xếp theo thứ tự nhất định (mức độ lâm sàng 1,2,3...) - Biến độc lập: là các yếu tố nguy cơ (yếu tố phơi nhiễm..) ảnh hưởng đến vấn đề nghiên cứu.

### 1.2.1 Qualitative and Quantitative Variables

Biến định lượng là gì? Biến định lượng (quantiative variable) **những biến mà các giá trị của chúng là các con số giá trị thực**. Ví dụ: Cũng là biến độ tuổi nhưng dữ liệu chúng ta thu thập là con số tuổi chính xác của đáp viên thì biến này sẽ là biến định lượng.

### 1.2.2 Discrete and Continuous Variables

Biến rời rạc (discrete variable) **là biến số chỉ nhận các giá trị nguyên**. Ví dụ: Số con, số người trong gia đình, số lần xét nghiệm, số công nhân trong một doanh nghiệp, số sản phẩm sản xuất ra trong một ngày của 1 phân xưởng may.

Biến liên tục (continuous variable) **là biến số có thể nhận bất kỳ giá trị nào trong một khoảng nhất định, tức biến thiên mà không bị gián đoạn**, Ngược với biến liên tục là biến rời rạc.

### 1.2.3 Scales

Nhóm dữ liệu đóng một vai trò quan trọng khi chúng ta phải xử lý dữ liệu lớn. Thông tin này cũng có thể được hiển thị bằng hình ảnh hoặc biểu đồ thanh. Dữ liệu được hình thành bằng cách sắp xếp các quan sát riêng lẻ về một biến thành các nhóm, để bảng phân phối tần suất của các nhóm này cung cấp một cách thuận tiện để tóm tắt hoặc phân tích dữ liệu được gọi là dữ liệu nhóm.

Phần mềm thống kê

setwd('C:/directory')

read.table('pizza\_delivery.dat')

read.table('pizza\_delivery.txt')

read.csv('pizza\_delivery.csv')

fix() # option 1

View() # option 2

### 1.2.4 Grouped Data

Nếu dữ liệu có sẵn ở dạng nhóm, chúng tôi gọi là bắt biến tương ứng thông tin này một biến được nhóm. Đôi khi, những biến này còn được gọi là các biến phân loại. Tuy nhiên, đây không phải là một định nghĩa hoàn chỉnh vì phân loại biến đề cập đến bất kỳ loại biến nào có số lượng hữu hạn, có thể nhỏ, các giá trị. Do đó, mọi biến số rời rạc và / hoặc danh nghĩa và / hoặc thứ tự và / hoặc định tính có thể được coi là một biến phân loại. Bất kỳ biến được nhóm hoặc phân loại nào chỉ có thể nhận hai giá trị được gọi là một biến nhị phân.

1.6 Bài Tập:

Exercise 1.1 Describe both the population and the observations for the

following

research questions:

Evaluation of the satisfaction of employees from an airline.

Population: quản trị viên-nhân viên quản lý, phi công, quản lý, nhân viên vệ sinh, và những người khác. Mỗi đơn

Observations: nhân viên liên quan đến một quan sát trong cuộc khảo sát.

Description of the marks of students from an assignment.

Population : Dân số bao gồm tất cả học sinh tham gia kỳ kiểm tra

Observations: Mỗi học sinh đại diện cho một quan sát.

Comparison of two drugs which deal with high blood pressure.

Population: Tất cả những người bị cao huyết áp trong khu vực nghiên cứu (thành phố, tỉnh, quốc gia, ...), là dân số quan tâm

Observations :Mỗi người trong số những người này là một quan sát.

1.2 A national park conducts a study on the behaviour of their leopards. A few of the park’s leopards are registered and receive a GPS device which allows measuring the position of the leopard. Use this example to describe the following concepts: population, sample, observation, value, and variable.

Dân số : tất cả con vật trong sở thú

Mẫu : các con báo

Quan sát : các con báo dc gắn định vị GPS

giá trị và biến : là những thông tin của GPS

Exercise 1.3 Which of the following variables are qualitative, and which are

quan-titative? Specify which of the quantitative variables are discrete and

which are continuous:

Định tính: Đảng phái chính trị ưu tiên, màu mắt, giới tính, nhóm máu, dòng tiêu đề của email.

Định lượng và rời rạc: Cỡ giày, mức độ hài lòng của khách hàng trên thang điểm từ 1 đến 10, số bàn thắng trong một trận đấu khúc côn cầu.

Định lượng và liên tục: Thời gian đi làm, giá bữa ăn canteen, bước sóng ánh sáng, thời gian chuyển phát bưu kiện, chiều cao của một đứa trẻ.

Exercise 1.4 Identify the scale of the following variables:

(a) Political party voted for in an election: yes or no

(b) The difficulty of different levels in a computer game: hard or easy

(c) Production time of a car : about 5 year

(d) Age of turtles : about 500 year

(e) Calender year : 2022 year

(f) Price of a chocolate bar : 10 dollars

(g) Identification number of a student

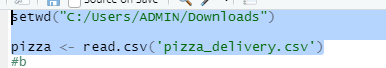
(h) Final ranking at a beauty contest

(i) Intelligence quotient. : tall or short Câu 1.5:

Input:

(a) Đầu tiên, hãy duyệt qua phần giới thiệu về R trong Phụ lục A. Sau đó, đọc trong phần

dữ liệu.

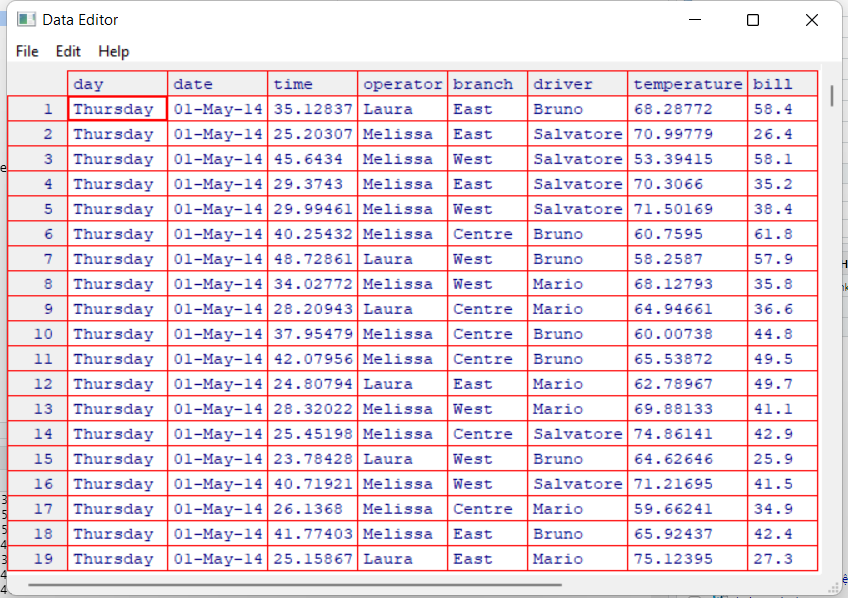


(b) Xem dữ liệu cả trong trình chỉnh sửa dữ liệu R và trong bảng điều khiển R

Input:

fix(pizza)

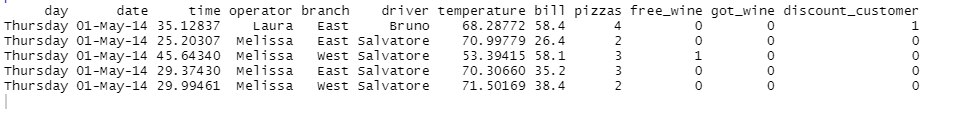
pizza



(c) Tạo một ma trận dữ liệu mới bao gồm 5 hàng đầu tiên và 5 biến đầu tiên của dữ liệu. In tập dữ liệu này trên bảng điều khiển R. Bây giờ, hãy lưu tập dữ liệu này vào định dạng ưa thích.

pizza2=head(pizza,5)

pizza2



Input

write.csv(pizza2,file='pizza2.csv')

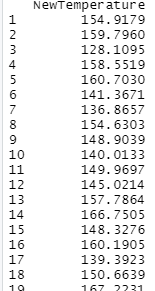
write.table(pizza2,file='pizza2.dat')

save(pizza2,file='pizza2.Rdata')

(d) Add a new variable “NewTemperature” to the data set which converts the tem- perature from ◦C to ◦F.

pizza$NewTemperature <- 32+1.8\*pizza$temperature

print(pizza)

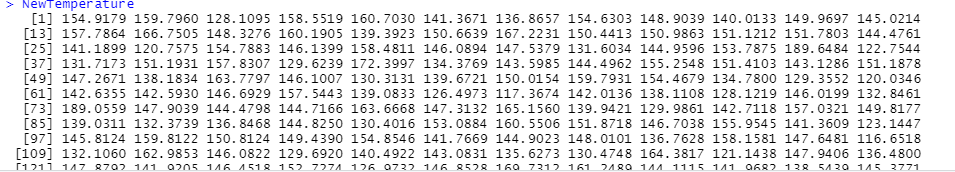


(e) Attach the data and list the values from the variable “NewTemperature”.

Input

attach(pizza)

NewTemperature



C2 Frequency Measures and Graphical 2

2.1 Absolute and Relative Frequencies

Tần suất tuyệt đối **là số lần một sự kiện được lặp lại trong một thời gian nhất định hoặc trong một đơn vị thời gian**. Tần suất tương đối được biết đến với định nghĩa đó chính lag là tần số của một sự kiện đối với sự kiện khác. Tần số tương đối là một khái niệm được thảo luận theo thống kê.

Tần số tương đối là một khái niệm được thảo luận theo thống kê. Theo thống kê, tần suất tương đối **là số lần xuất hiện của sự kiện 1 khi tổng số sự kiện được chuẩn hóa**. Khi một tập hợp tần số tuyệt đối được chuẩn hóa, mỗi giá trị tương ứng với giá trị tần số tuyệt đối ban đầu là tần số tương đối của hệ thống.

Exiced 2.1.1

ECDF". Như chính cái tên cho thấy, nó cho chúng ta một ý tưởng về quan hệ tích lũy-

tần số tive lên đến một thời điểm nhất định. Ví dụ: giả sử chúng tôi muốn biết có bao nhiêu

### 2.2 Empirical Cumulative Distribution Function

ECDF cho dữ liệu này có thể thu được bằng cách tóm tắt dữ liệu trong một vectơ và

bằng cách sử dụng hàm plot.ecdf () trong R,

code:

sv <- c(rep(1,4),rep(2,16),rep(3,90),rep(4,70),rep(5,20))

plot.ecdf(sv)

ECDF có thể được sử dụng để thu được các tần số tương đối cho các giá trị có trong

khoảng thời gian nhất định như

H (c ≤ x ≤ d) = tần số tương đối của các giá trị x với c ≤ x ≤ d.

### 2.2.1 ECDF for Ordinal Variables

Hãy xem xét một cuộc khảo sát mức độ hài lòng của khách hàng từ một công ty dịch vụ xe hơi.

200 khách hàng đã thực hiện dịch vụ ô tô trong vòng 30 ngày qua đã được yêu cầu

phản hồi về mức độ hài lòng chung của họ đối với chất lượng dịch vụ xe hơi

trên thang điểm từ 1 đến 5 dựa trên các tùy chọn sau: 1 = không hài lòng chút nào, 2 =

không hài lòng, 3 = hài lòng, 4 = rất hài lòng, và 5 = hoàn toàn hài lòng. Dựa trên

tần số của mỗi tùy chọn, chúng tôi có thể tính toán các tần số tương đối và sau đó

XEM hàm phân phối tích lũy theo kinh nghiệm, hoặc theo cách thủ công (mất nhiều thời gian hơn)

hoặc bằng cách sử dụng R (nhanh):

2.3 Biểu diễn đồ thị bằng một biến: Các bảng tần suất và các hàm phân phối tích lũy theo kinh nghiệm rất hữu ích trong việc cung cấp một bản tóm tắt số của một biến. Đồ thị là một cách thay thế để tóm tắt thể loại. Chiều cao của mỗi thanh được xác định bởi tần số tuyệt đối hoặc thông tin của một biến. Trong nhiều tình huống, chúng có lợi thế trong việc truyền tải tần suất tương đối của danh mục tương ứng và được hiển thị trên trục y. Nếu thông tin ẩn trong dữ liệu nhỏ gọn hơn. Tương tự, tâm trạng của ai đó có thể biến được đo ở mức thứ tự, sau đó nên sắp xếp các thanh dễ hiểu hơn khi nhìn vào một mặt cười hơn là đọc một bài luận về trên trục x theo thứ hạng hoặc giá trị của chúng. Nếu số lượng danh mục lớn, tâm trạng của một người trong một đoạn văn dài.

ECDF cho các biến liên tục 2.3 Biểu diễn bằng đồ thị của một biến

Các bảng tần số và các hàm phân phối tích lũy theo kinh nghiệm rất hữu ích trong việc cung cấp-

ing một bản tóm tắt số của một biến. Đồ thị là một cách thay thế để tóm tắt

thông tin của một biến. Trong nhiều tình huống, chúng có lợi thế trong việc truyền tải

thông tin ẩn trong dữ liệu nhỏ gọn hơn. Tương tự, tâm trạng của ai đó có thể

dễ hiểu hơn khi nhìn vào một mặt cười hơn là đọc một bài luận về

tâm trạng của một người trong một đoạn văn dài.

2.4 Đồ thị mật độ nhân: Một nhược điểm của biểu đồ là dữ liệu liên tục được phân loại giả tạo. Các

việc lựa chọn các khoảng thời gian của lớp là rất quan trọng đối với giao diện cuối cùng của biểu đồ. Một thanh lịch hơn cách để đối phó với vấn đề này là làm mịn biểu đồ theo nghĩa là mỗi quan sát có thể đóng góp vào các lớp khác nhau với các trọng số khác nhau và phân phối âm mưu có thể được tạo ra bằng cách sử dụng chức năng sau: việc lựa chọn các khoảng thời gian của lớp là rất quan trọng đối với giao diện cuối cùng của biểu đồ. Một thanh lịch hơn được biểu diễn bằng một hàm liên tục chứ không phải một hàm bước. Mật độ nhân.

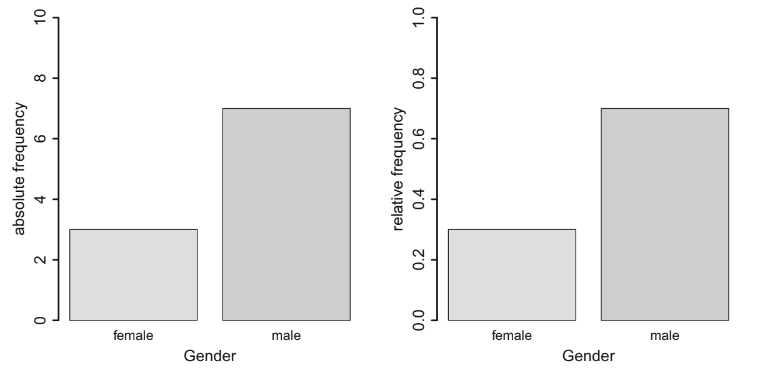
2.4Biểu đồ cột

barplot(table(branch))

barplot(table(branch)/length(branch))

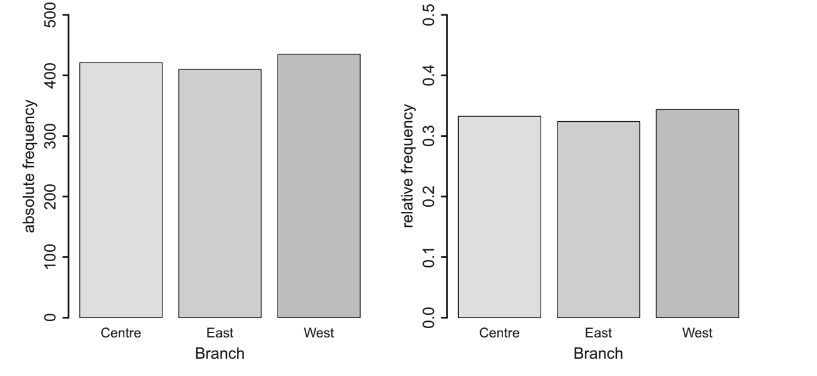
2.4.1Pie Chart

pie(table(sv))



2.4.2Histogram

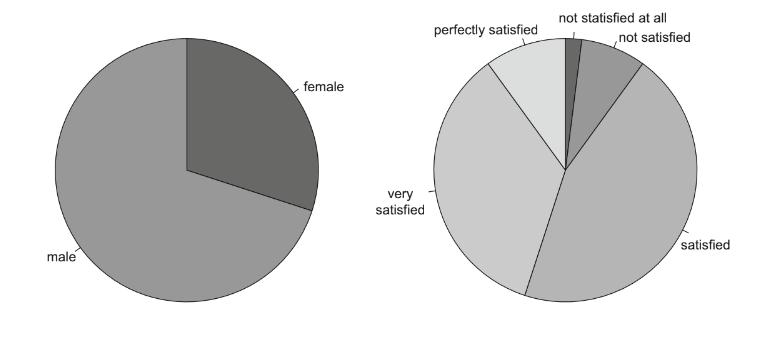
hist(time) # show abs. frequencies

hist(time, freq=F) # show rel. frequencies 

2.5 Kernel Density Plots

example <- c(3,6,7,8,10)

plot(density(example, kernel='gaussian'))



2.6 Bài tập

Exercise 2.1 Consider the results of the national elections in South Africa in 2014 and 2009:

1. Summarize the results of the 2014 elections in a bar chart. Do it manually and by using R.

Code :

#input

Party= c("ANC","DA","EFF","IFP","COPE","Others")

Results2014 = c(62.15 ,22.23 ,6.35 , 2.40, 0.67, 6.20)

Results2009= c(65.9 ,16.66, 4.55, "N/A", 7.42, 5.47)

acid= data.frame(Party,Results2014,Results2009)

acid

#output

Party Results2014 Results2009

1 ANC 62.15 65.9

2 DA 22.23 16.66

3 EFF 6.35 4.55

4 IFP 2.40 N/A

5 COPE 0.67 7.42

6 Others 6.20 5.47

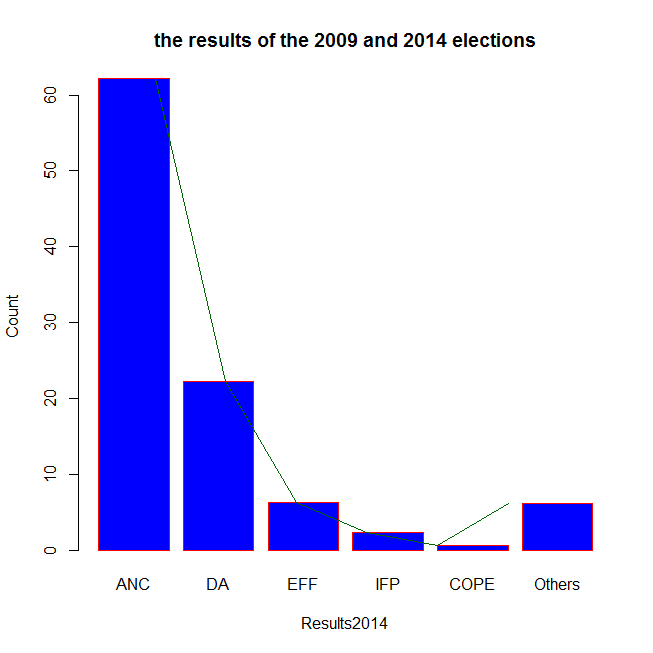
1. How would you compare the results of the 2009 and 2014 elections? Offer asimple solution that can be represented in a single plot. Construct this plot in R.

Code:

barplot(Results2014,xlab="Results2014",ylab="Count"

,main="the results of the 2009 and 2014 elections",col="blue",names.arg=Party,border="red")

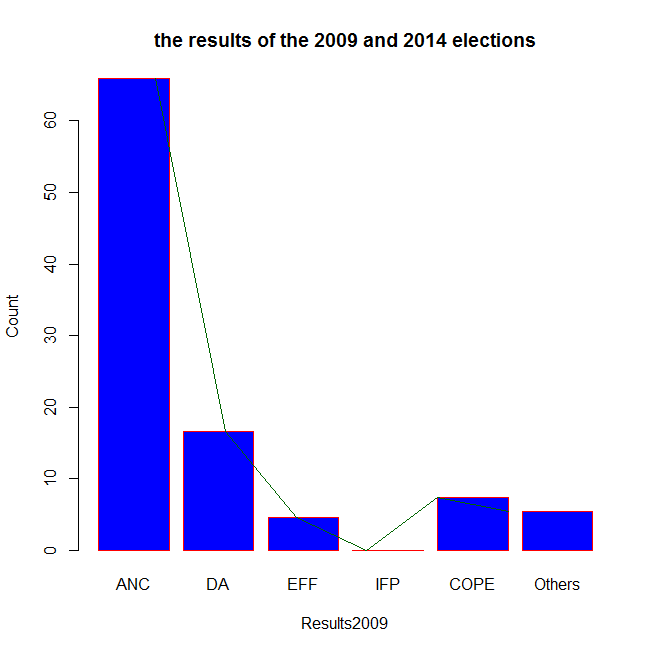
lines(Results2014,col="darkgreen")



barplot(Results2009,xlab="Results2019",ylab="Count"

,main="the results of the 2009 and 2014 elections",col="blue",names.arg=Party,border="red")

lines(Results2009,col="darkgreen")



Exercise 2.2 Consider a variable X describing the time until the first goal was scored in the matches of the 2006 football World Cup competition. Only matches with at least one goal are considered, and goals during the xth minute of extra time are denoted as 90 + x:

#a

What is the scale of X?

Code:

#input

acid2=c(6,24,90+1,84, 25, 3 ,83, 89, 34, 25, 24, 18, 6,

23 ,10 ,28 ,4 ,63 ,6 ,60 ,5 ,40, 2 ,22 ,26 ,23 ,26,

44, 49, 34, 2 ,33 ,9 ,16, 55, 23, 13, 23, 4, 8, 26,

70, 4 ,6 ,60 ,23 ,90+5 ,28 ,49, 6 ,57 ,33 ,56, 7)

mean(acid2)

#output

[1] 31.16667

2.2 Write down the frequency table of X based on the following categories: [0, 15),[15, 30), [30, 45), [45, 60), [60, 75), [75, 90), [90, 96).

Code:

factor(acid2)

acid3=table(acid2)

acid3

#install.packages('plyr')

#library('plyr')

count(acid2)

a0=sum(count(acid3[0:10])[2])

a1=sum(count(acid3[11:18])[2])

a2=sum(count(acid3[19:22])[2])

a3=sum(count(acid3[23:27])[2])

a4=sum(count(acid3[28:29])[2])

a5=sum(count(acid3[30:32])[2])

a6=sum(count(acid3[33:34])[2])

count1= c(a0,a1,a2,a3,a4,a5,a6)

minute=c("[0,15)" , "[15, 30)","[30, 45)" , "[45, 60)", "[60, 75)", "[75, 90)", "[90, 96)")

continue <- data.frame(minute,count1)

continue

output:

minute count1

1 [0,15) 17

2 [15, 30) 17

3 [30, 45) 6

4 [45, 60) 7

5 [60, 75) 2

6 [75, 90) 3

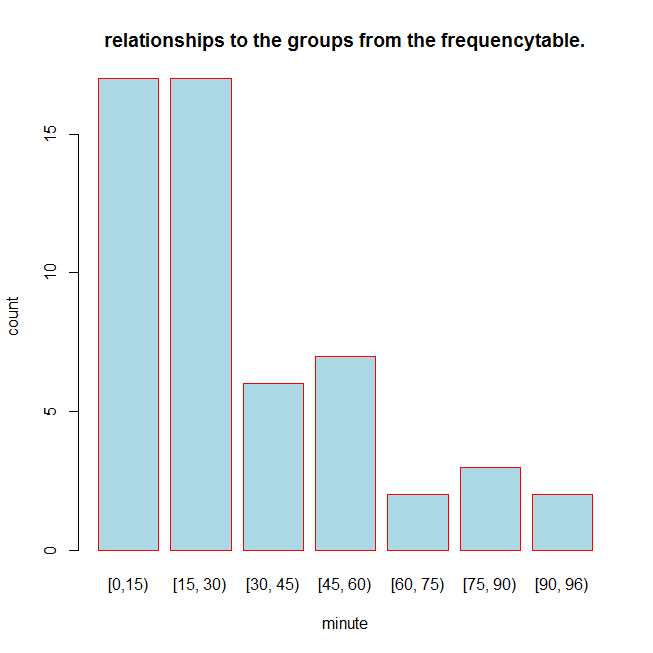
7 [90, 96) 2

2.3Draw the histogram for X with intervals relating to the groups from the frequency table.

Code:

barplot(count1,names.arg = minute,xlab="minute",ylab="count",col="lightblue"

,main="relationships to the groups from the frequencytable.",border = "red")



#d

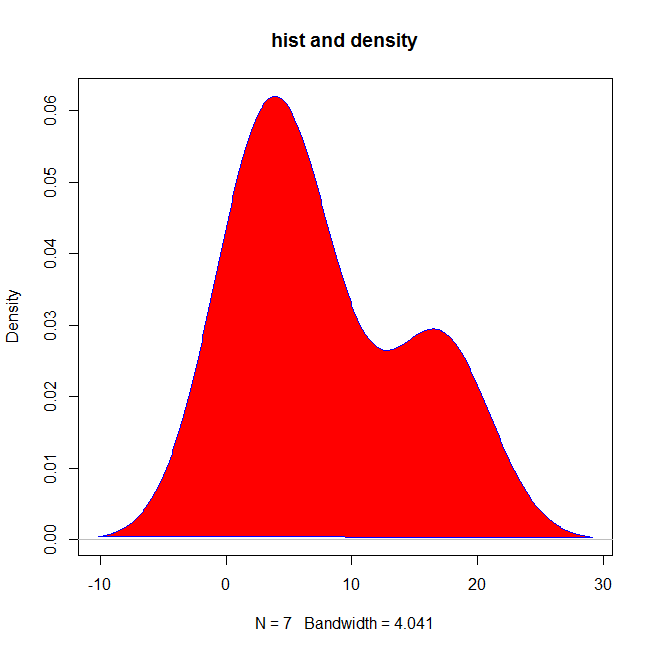
code

#input

d <- density(count1)

plot(d, main="hist and density")

polygon(d, col="red", border="blue")



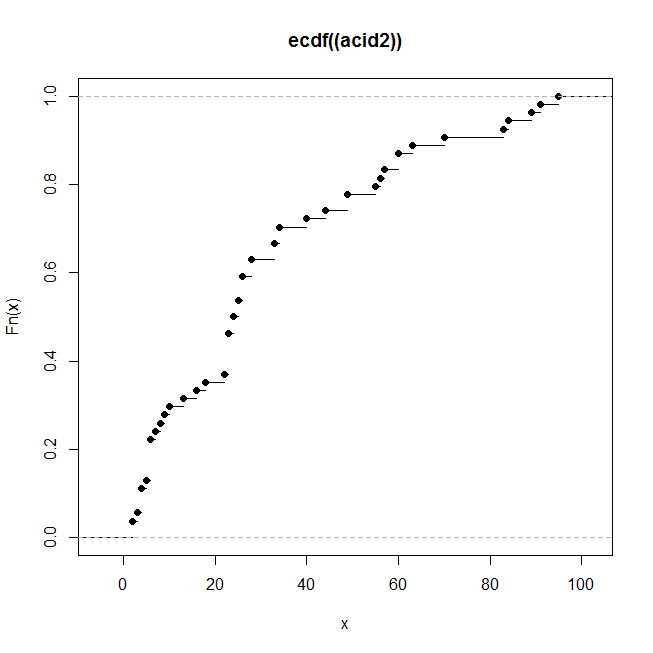
Use R to plot the ECDF (via a step function) for

(i) the original data and

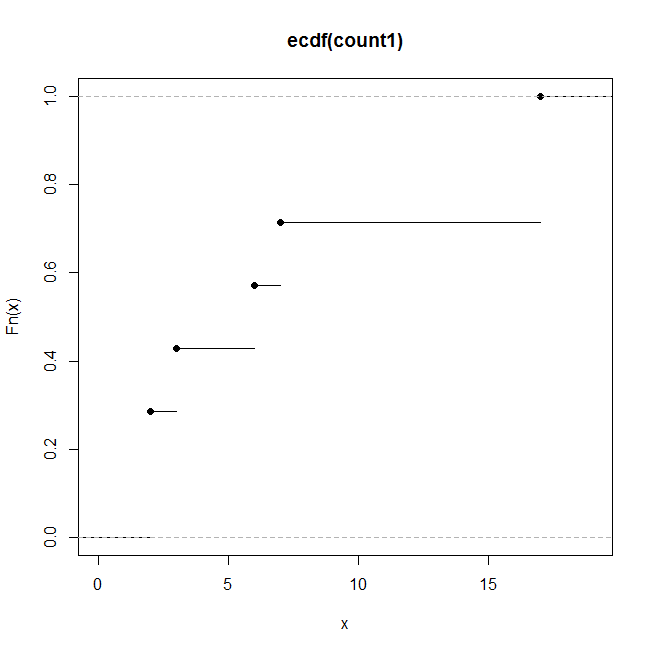
(ii) the grouped data.

code

plot(ecdf((acid2)))



plot(ecdf(count1))



(g) Consider the grouped data. Now assume that the values within each interval are

distributed uniformly. Determine the proportion of first goals which occurred

1. in the first half, i.e. during the first 45 min,

(ii) in the last 10 min or during the extra time,

1. between the 20th and 65th min, i.e. what is H(20 ≤ X ≤ 65)?

a7=sum(count(acid3[0:22])[2])

a8=sum(count(acid3[0:9])[2])

a6=sum(count(acid3[33:34])[2])

a9=sum(count(acid3[12:28])[2])

mount1=c("first 45 min","first 10 min","the extra time","the 20th and 65th min")

count2= c(a7,a8,a6,a9)

continue2= c(mount1,count2)

continue2

output

[1] "first 45 min" "first 10 min" "the extra time" "the 20th and 65th min"

[5] "40" "16" "2" "30"

(h) Determine the time point at which in 80 % of the matches the first goal was scored at or before this time point.

Code:

max(count(acid3)[2])

23 and 6 min

Exercise 2.4 A university survey was conducted on 500 first-year students to obtainknowledge about the size of their accommodation (in square metres).

code

Fx=c(0.25,0.40 ,0.75 , 0.97, 1)

bazo=c("8–14","14–22","22–34","34–50","50–82")

continue3=data.frame(Size\_of\_accommodation=bazo,Fx)

a=continue3[1,2]\*500

b=continue3[2,2]\*500

c=continue3[3,2]\*500

d=continue3[4,2]\*500

e=continue3[5,2]\*500

tansotuyetdoi=c(a,b,c,d,e)

continue3$tansotuyetdoi= tansotuyetdoi

continue3

output

Size\_of\_accommodation Fx tansotuyetdoi

1 8–14 0.25 125

2 14–22 0.40 200

3 22–34 0.75 375

4 34–50 0.97 485

5 50–82 1.00 500

What proportion of people live in a flat of at least 34 m2?

h=continue3[4,2]

h

output

[1] 0.97

Exercise 2.5 Consider a survey in which 100 people were asked to rate on a scale from 1 to 10 how much they agree with the statement that “there is too much footballon television”. The results are summarized below:

Calculate and draw the ECDF of the scores.

code :

#input

Score =c(0, 1, 2, 3 ,4 ,5 ,6, 7, 8, 9, 10)

Responses= c(0,1, 3, 8, 8, 27, 30, 11, 6, 4, 2)

continue4=data.frame(Score,Responses)

continue4

plot(ecdf(Score))

plot(ecdf(Responses))

#output

Score Responses

1 0 0

2 1 1

3 2 3

4 3 8

5 4 8

6 5 27

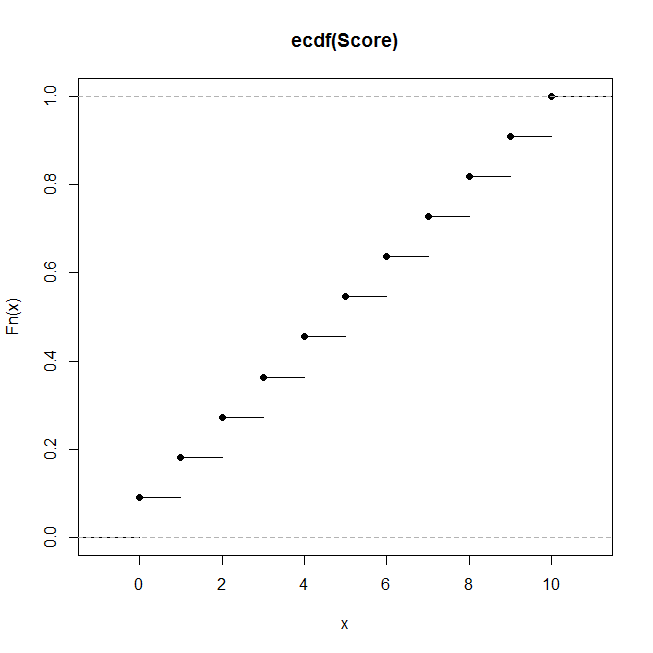
7 6 30

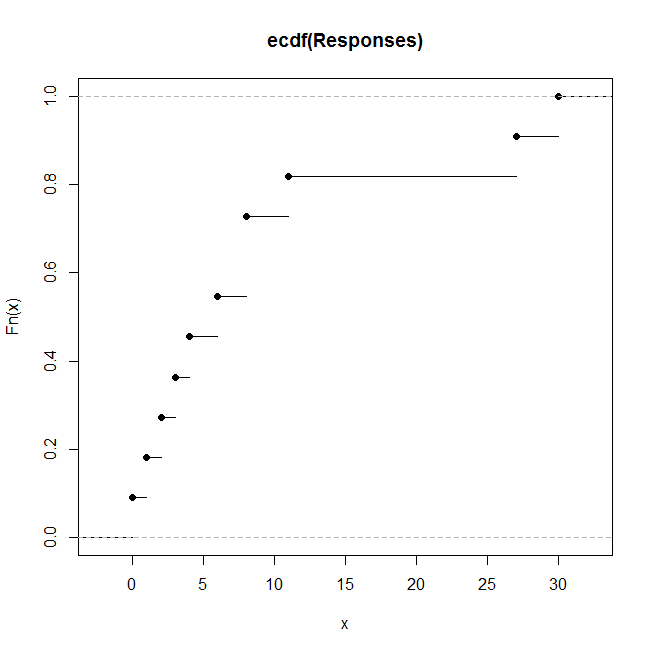
8 7 11

9 8 6

10 9 4

11 10 2



(b) Determine F(3) and F(9).  
aa= Responses[4]/100

ab= Responses[10]/100

aa

ab

output :

[1] 0.08

[1] 0.04

(c) Consider the situation, where the data is summarized in the two categories “dis- agree” (score ≤ 5) and “agree” (score > 5). What would the ECDF look like under the approach outlined in (2.11)? Determine F(3) and F(9) for the sum-marized data.

n1=if( continue4[1,2] >5){

print("dong y")

} else{print("khong dong y")

}

n2=if( continue4[2,2] >5){

print("dong y")

} else{print("khong dong y")

}

n3=if( continue4[3,2] >5){

print("dong y")

} else{print("khong dong y")

}

n4=m=if( continue4[4,2] >5){

print("dong y")

} else{print("khong dong y")

}

n5=if( continue4[5,2] >5){

print("dong y")

} else{print("khong dong y")

}

n6=if( continue4[6,2] >5){

print("dong y")

} else{print("khong dong y")

}

n7=if( continue4[7,2] >5){

print("dong y")

} else{print("khong dong y")

}

n8=if( continue4[8,2] >5){

print("dong y")

} else{print("khong dong y")

}

n9=if( continue4[9,2] >5){

print("dong y")

} else{print("khong dong y")

}

n10=if( continue4[10,2] >5){

print("dong y")

} else{print("khong dong y")

}

n11=if( continue4[11,2] >5){

print("dong y")

} else{print("khong dong y")

}

continue4$agree\_or\_disagree=c(n1,n2,n3,n4,n5,n6,n7,n8,n9,n10,n11)

continue4

output:

Score Responses agree\_or\_disagree

1 0 0 khong dong y

2 1 1 khong dong y

3 2 3 khong dong y

4 3 8 dong y

5 4 8 dong y

6 5 27 dong y

7 6 30 dong y

8 7 11 dong y

9 8 6 dong y

10 9 4 khong dong y

11 10 2 khong dong y

What would the ECDF look like under the approach outlined in (2.11)?

code

oneor2=c(m1,m2,m3,m4,m5,m6,m7,m8,m9,m10,m11)

continue4$oneor2=c(m1,m2,m3,m4,m5,m6,m7,m8,m9,m10,m11)

continue4

plot(ecdf(oneor2))

Score Responses agree\_or\_disagree oneor2

1 0 0 khong dong y 1

2 1 1 khong dong y 1

3 2 3 khong dong y 1

4 3 8 dong y 2

5 4 8 dong y 2

6 5 27 dong y 2

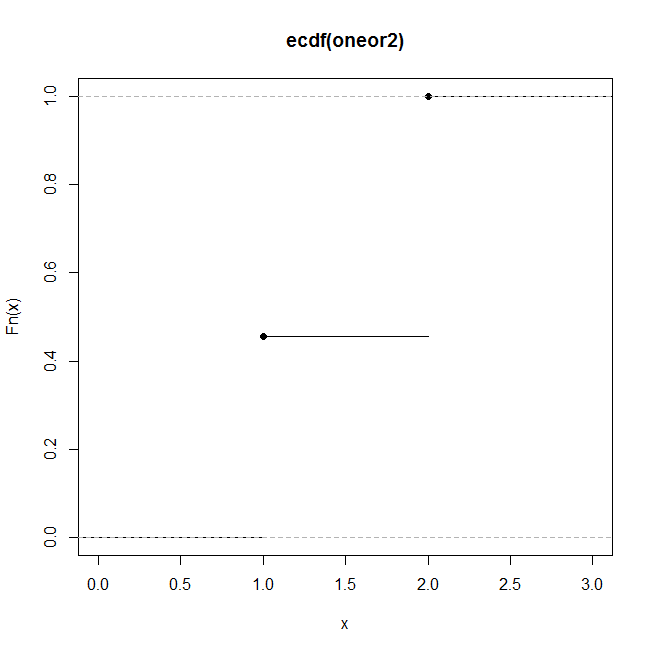
7 6 30 dong y 2

8 7 11 dong y 2

9 8 6 dong y 2

10 9 4 khong dong y 1

11 10 2 khong dong y 1



Determine F(3) and F(9) for the sum-marized data.

agree\_or\_disagree[4]

agree\_or\_disagree[10]

[1] "dong y

[1] "khong dong y"

(b) nhìn có thể suy ra (C) và ngược lại

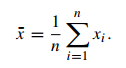
3. Measures of Central Tendency and dispersion

3.1 Các biện pháp của xu hướng trung tâm

Xu hướng của con người là so sánh với “mức trung bình”. Ví dụ, một học sinh đạt 40% trong một bài kiểm tra sẽ hài lòng với kết quả nếu điểm trung bình điểm của lớp là 25%. Nếu điểm trung bình của lớp là 90%, thì học sinh có thể không cảm thấy hạnh phúc ngay cả khi anh ấy đúng 70%.

3.1.1 Trung bình số học

Trung bình cộng là một trong những thước đo trực quan nhất của khuynh hướng trọng tâm. Giả sử một biến có kích thước n gồm các giá trị x1, x2, …, xn. Số học trong bình dữ liệu này được định nghĩa là:



Trong R, chúng ta sử dụng hàm weighted.mean để thu được kết quả. Chức năng yêu cầu để chỉ định phương tiện (giả thuyết) cho mỗi nhóm, ví dụ: các giá trị giữa của các khoảng thời gian trong lớp, cũng như trọng số.



3.1.2 Trung vị và lượng tử

Giá trị trung bình là giá trị chia các quan sát thành hai phần bằng nhau như rằng ít nhất 50% giá trị lớn hơn hoặc bằng giá trị trung bình và ít nhất 50% giá trị nhỏ hơn hoặc bằng giá trị trung vị. Trung vị được ký hiệu là x˜0,5; sau đó, về mặt của hàm phân phối tích lũy theo kinh nghiệm, điều kiện F (x˜0,5) = 0,5 thỏa mãn. Xét n quan sát x1, x2, ..., xn có thể được sắp xếp là x (1) ≤ x (2) ≤ ··· ≤ x (n). Việc tính toán số trung vị phụ thuộc vào số quan sát n là lẻ hay chẵn. Khi n lẻ, thì x˜0,5 là giá trị đặt hàng giữa. Khi n chẵn thì x˜0,5 là trung bình cộng của hai các giá trị có thứ tự giữa:

So sánh Trung bình với Trung vị. Trong các ví dụ trên, giá trị trung bình và trung tuyến hóa ra khá giống nhau. Điều này là do chúng tôi đã xem xét dữ liệ được phân bố đối xứng xung quanh tâm của nó, tức là ở mức trung bình, chúng ta có thể mong đợi 26 C với độ lệch tương tự trên và dưới nhiệt độ trung bình Một ví dụ tương tự được đưa ra trong Hình 3.1a: chúng ta thấy rằng dữ liệu thô được tóm tắt bởi bằng cách sử dụng dấu tích ở cuối biểu đồ và bằng cách sử dụng công cụ ước tính mật độ hạt nhân. Các giá trị trung bình và giá trị trung bình tương tự nhau ở đây vì sự phân bố của các quan sát là đối xứng xung quanh tâm. Nếu chúng ta có dữ liệu sai lệch (Hình 3.1b), thì giá trị trung bình và trung vị có thể khác nhau. Nếu dữ liệu có nhiều hơn một trung tâm, chẳng hạn như trong Hình 3.1c cả giá trị trung bình và giá trị trung bình đều không có cách diễn giải có ý nghĩa. Nếu chúng ta có ngoại lệ (Hình 3.1d), thì nên sử dụng giá trị trung bình vì giá trị trung bình nhạy cảm với các giá trị ngoại lệ. Những ví dụ này cho thấy rằng tùy thuộc vào tình huống quan tâm hoặc giá trị trung bình, trung bình, cả hai hoặc cả hai hoặc cả hai đều có thể hữu ích. Lượng tử. Lượng tử là sự tổng quát hóa ý tưởng về trung vị. Giá trị trung bình là giá trị chia dữ liệu thành hai phần bằng nhau. Tương tự, một phân vùng lượng tử dữ liệu thành các tỷ lệ khác. Ví dụ: 25% -quantile chia dữ liệu thành hai phần sao cho ít nhất 25% giá trị nhỏ hơn hoặc bằng lượng tử và ít nhất 75% giá trị lớn hơn hoặc bằng lượng tử. Nói chung, cho α là một số từ 0 đến 1. Lượng tử (α × 100)%, được ký hiệu là x˜α, được định nghĩa là giá trị chia dữ liệu theo tỷ lệ (α × 100)% và (1 - α) × 100% sao cho ít nhất là α × 100 % giá trị nhỏ hơn hoặc bằng lượng tử và ít nhất (1 - α) × 100% giá trị lớn hơn hoặc bằng lượng tử. Về mặt của hàm phân phối tích lũy thực nghiệm, chúng ta có thể viết F (x˜α) = α. Ngay sau đó là đối với n quan sát, ít nhất giá trị nα nhỏ hơn hoặc bằng x˜α và ít nhất n (1 - α) quan sát lớn hơn hoặc bằng x˜α. Giá trị trung bình là 50% -quantile x˜0,5. Nếu α nhận các giá trị 0,1, 0,2, ..., 0,9, các lượng tử được gọi là deciles. Nếu α · 100 là một số nguyên (ví dụ: α × 100 = 95), các lượng tử được gọi là phân vị, tức là dữ liệu được chia thành 100 phần bằng nhau. Nếu α nhận các giá trị 0,2, 0,4, 0,6 và 0,8, các lượng tử được gọi là ngũ phân vị và chúng chia dữ liệu thành năm phần bằng nhau. Nếu α nhận các giá trị 0,25, 0,5 và 0,75, các lượng tử được gọi là tứ phân vị.

3.1.5 Trung bình hình học

Ý nghĩa hình học đóng một vai trò quan trọng trong các lĩnh vực mà chúng ta quan tâm đến sản phẩm của các quan sát, chẳng hạn như khi chúng ta xem xét phần trăm thay đổi về số lượng. Chúng tôi minh họa cách giải thích và sử dụng của nó bằng cách xem xét mức tăng trưởng trung bình của một lượng theo nghĩa chúng tôi cho phép một giá trị ban đầu, chẳng hạn như một số tiền nhất định hoặc một dân số cụ thể, thay đổi theo thời gian. Giả sử chúng ta có một giá trị bắt đầu ở một số mốc thời gian cơ sở 0 (không), có thể được ký hiệu là B0. Tại thời điểm t, giá trị này có thể đã thay đổi và do đó chúng tôi ký hiệu là Bt, t = 1, 2, ..., T. Tỷ lệ của Bt và

3.2 Các biện pháp phân tán

Các phép đo về xu hướng trung tâm, như đã giới thiệu trước đó, cho chúng ta một ý tưởng về loca tion nơi tập trung phần lớn dữ liệu. Tuy nhiên, hai tập dữ liệu khác nhau có thể có cùng giá trị đối với thước đo của xu hướng trung tâm, giả sử cùng một số học có nghĩa là, nhưng chúng có thể có nồng độ khác nhau xung quanh giá trị trung bình. Trong trường hợp này, các biện pháp vị trí có thể không đủ thích hợp để mô tả sự phân bố của dữ liệu. Nồng độ hoặc độ phân tán của các quan sát xung quanh bất kỳ giá trị cụ thể nào là một thuộc tính khác đặc trưng cho dữ liệu và sự phân phối của nó.

3.3 Biểu đồ boxplot

Cho đến nay chúng tôi đã mô tả các thước đo khác nhau về xu hướng trung tâm và sự phân tán. Có thể rất tẻ nhạt khi liệt kê các biện pháp đó trong các bảng tóm tắt. Một biểu đồ đơn giản và mạnh mẽ là biểu đồ hộp tóm tắt sự phân bố của một biến liên tục (hoặc đôi khi là thứ tự) bằng cách sử dụng giá trị trung vị, tứ phân vị, giá trị nhỏ nhất, lớn nhất và cực trị của nó. Nhìn vào toàn bộ biểu đồ hộp cho chúng ta biết về sự phân bố dữ liệu, phạm vi và sự biến thiên của các quan sát. Đôi khi, có thể nên hiểu những giá trị nào là cực đoan theo nghĩa là chúng “ở rất xa” so với trung tâm của điểm phân phối. Trong nhiều gói phần mềm, bao gồm cả R, các giá trị được xác định là cực trị nếu chúng lớn hơn 1,5 chiều dài hộp tính từ phần tư thứ nhất hoặc thứ ba. Đôi khi, chúng được gọi là ngoại lệ. Giá trị ngoại lai và giá trị cực trị được xác định khác nhau trong một số gói phần mềm và sách.

3.4 Các biện pháp tập trung

Một khái niệm hoàn toàn khác được sử dụng để mô tả một biến định lượng là ý tưởng về tứ phân vị (25, 29C). Giá trị nhỏ nhất và lớn nhất là 21C và 31C. Cái hộp một số thời gian giao hàng đặc biệt ngắn và dài. nồng độ. Đối với một biến X, nó tóm tắt tỷ lệ của mỗi quan sát biểu đồ cho dữ liệu này được thể hiện trong Hình 3.5a. Có thể thấy rằng sự phân bố nhiệt độ đối với tổng tất cả các quan sát Chúng ta hãy xem một ví dụ đơn giản xi . để chứng minh tính hữu ích của nó.

3.6 Bài tập

3.1 Bài tập 3.1 Một người đam mê đi bộ đường dài có một ứng dụng mới cho điện thoại thông minh của anh ấy, ứng dụng này có thể tóm tắt các chuyến đi bộ của anh ấy bằng cách sử dụng thiết bị GPS. Hãy để chúng tôi xem xét khoảng cách đi bộ đường dài (tính bằng km) và độ cao tối đa (tính bằng m) trong 10 lần đi bộ đường dài gần nhất:



3.1

Distance =c(12.5 ,29.9 ,14.8 ,18.7 ,7.6, 16.2 ,16.5 ,27.4 ,12.1 ,17.5)

Altitude =c(342, 1245, 502, 555, 398, 670, 796, 912, 238, 466)

df= data.frame(Distance,Altitude)

df

mean(Distance)

mean(Altitude)

median(Distance)

median(Altitude)

quantile(Distance)[2]

quantile(Altitude)[2]

quantile(Altitude)[4]

quantile(Altitude)[4]

#b

qqplot(Distance,Altitude)

qqplot(quantile(Altitude), quantile(Altitude))

#c

IQR(Distance)

IQR(Altitude)

mad(Distance)

mad(Altitude)

#d

var(Distance)\*9/10

var(Altitude)\*9/10

Altitude\_feet =Altitude \*3.28

mean(Altitude\_feet)

#e

boxplot(Altitude, main="box plot for altitude",col="lightblue")

#trung binh thuong tap trung o muc 500

boxplot(Distance, main="box plot for distance ",col="lightblue")

#trung binh thuong tap trung o muc 17

#f

a0= if(Distance[1] >= 5 & Distance[1] <15){print("ngắn")

}else if(Distance[1] >= 15 & Distance[1] <20){print("trung bình")

}else{print("dài")}

a1= if(Distance[2] >= 5 & Distance[2] <15){print("ngắn")

}else if(Distance[2] >= 15 & Distance[2] <20){print("trung bình")

}else{print("dài")}

a2= if(Distance[3] >= 5 & Distance[3] <15){print("ngắn")

}else if(Distance[3] >= 15 & Distance[3] <20){print("trung bình")

}else{print("dài")}

a3= if(Distance[4] >= 5 & Distance[4] <15){print("ngắn")

}else if(Distance[4] >= 15 & Distance[4] <20){print("trung bình")

}else{print("dài")}

a4= if(Distance[5] >= 5 & Distance[5] <15){print("ngắn")

}else if(Distance[5] >= 15 & Distance[5] <20){print("trung bình")

}else{print("dài")}

a5= if(Distance[6] >= 5 & Distance[6] <15){print("ngắn")

}else if(Distance[6] >= 15 & Distance[6] <20){print("trung bình")

}else{print("dài")}

a6= if(Distance[7] >= 5 & Distance[7] <15){print("ngắn")

}else if(Distance[7] >= 15 & Distance[7] <20){print("trung bình")

}else{print("dài")}

a7= if(Distance[8] >= 5 & Distance[8] <15){print("ngắn")

}else if(Distance[8] >= 15 & Distance[8] <20){print("trung bình")

}else{print("dài")}

a8= if(Distance[9] >= 5 & Distance[9] <15){print("ngắn")

}else if(Distance[9] >= 15 & Distance[9] <20){print("trung bình")

}else{print("dài")}

a9= if(Distance[10] >= 5 & Distance[10] <15){print("ngắn")

}else if(Distance[10] >= 15 & Distance[10] <20){print("trung bình")

}else{print("dài")}

long\_or\_short=c(a0,a1,a2,a3,a4,a5,a6,a7,a8,a9)

df$long\_or\_short=c(a0,a1,a2,a3,a4,a5,a6,a7,a8,a9)

df

weighted.mean(c(10,17.5,25),c(4/10,4/10,2/10))

#g

aa=sum(table(Distance)[1:4])

ab=sum(table(Distance)[5:8])

ac=sum(table(Distance)[9:10])

count1= c(aa,ab,ac)

about=c("(5; 15]", "(15; 20]"," (20; 30]")

compelete= data.frame(about,count1)

bbb=c(count(Distance)[1])

ccc=c(7.6, 12.1, 12.5, 14.8, 16.2 ,16.5, 17.5, 18.7, 27.4, 29.9)

ad= var(ccc[1:4])

ae=var(ccc[5:8])

af=var(ccc[9:10])

var1=c(ad,ae,af)

compelete$var1=c(ad,ae,af)

compelete

3.2

Exercise 3.2 A gambler notes down his wins and losses (in e) from playing 10

games of roulette in a casino.

Round 1 2 3 4 5 6 7 8 9 10

Won/Lost 200 600 −200 −200 −200 −100 −100 −400 0

Assume x ̄ = − e90 and s = e294.7881. What is the result of round 10?

−90 = 1/10 (200 + 600 − 200 − 200 − 200 − 100 − 100 − 400 + 0 + R)

−90 = 1/10 (−400 + R)

⇒ R = −500.

b) (b) The mode is x ̄M = −200. Using nα = 2.5 and nα = 7.5, respectively, we can

determine the quartiles as follows:

x ̃0.25 = x(3) = −200,

x ̃0.75 = x(8) = 0.

The interquartile range is dQ = 0 − (−200) = 200.

c)

nm = n − nw − nc = 100 − 45 − 20 = 35.

Using the formula for the arithmetic mean for grouped data,

35(100 · 15 − 45 · 16 − 20 · 7.5) = 18.

Similarly, we can use the Theorem of Variance Decomposition, Eq. (3.27), to calcu-

late the variance for the grouped data:

This yields

= 1/35(100 · 19.55 − 45 · 6 − 20 · 3 − 45 · 12 − 35 · 32 − 20 · 7.52) = 4.

Exercise 3.3 A fashion boutique has summarized its daily sales of designer socks in

different groups: men’s socks, women’s socks, and children’s socks. Unfortunately,

the data for men’s socks was lost. Determine the missing values.

n Arithmetic mean in e Standard deviation in e

Women’s wear 45 16 √6

Men’s wear ? ? ?

Children’s wear 20 7.5 √3

Total 100 15 √19.55

Men’s wear (n)=100 − 45 − 20 = 35.

Using the formula for the arithmetic mean for grouped data,

35(100 · 15 − 45 · 16 − 20 · 7.5) = 18.

Standard deviation in e = 1/35(100 · 19.55 − 45 · 6 − 20 · 3 − 45 · 12 − 35 · 32 − 20 · 7.52) = 4.

#3.4

Bt /Bt−1 = (−, 1.04, 1.125, 0.925, 1.2, 0.933)

x ̄G = (1.04 · 1.125 · 0.925 · 1.2 · 0.933)

1/5 = 1.04.

b)

r = 1.04 − 1 = 4 %.

b) B2018 = ̄xG B2017, B2017 = ̄xG B2016, ⇒ B2018 = ̄x2G B2016

B2018 = 1.042 · 28 = 30.28 ≈ 31.

(i) 1 2 3 4

Number of members 10 8 8 4

Rel. number of members f j 10/30 8/30 8/30 4/30

u ̃i =

j f j 1 0/30 18/30 26/30 1

Money invested xi 40 60 70 80

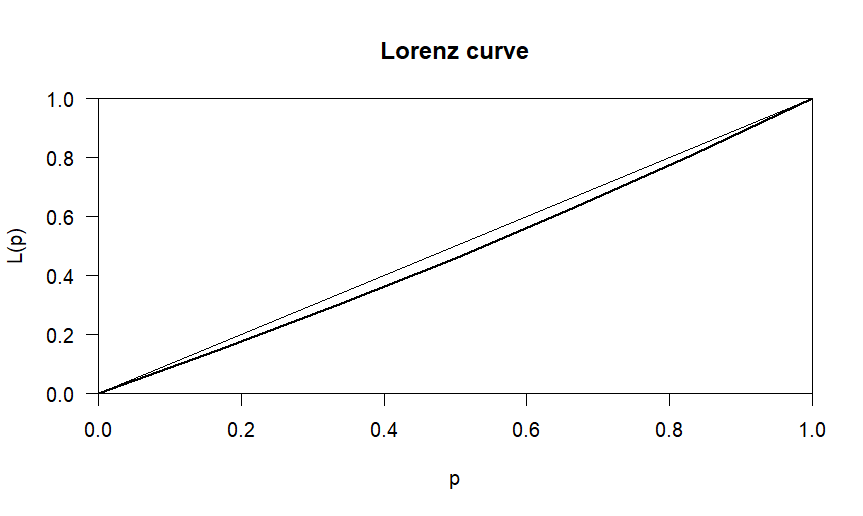
Rel. amount per group 40/250 60/250 70/250 80/250

vi 40/250 100/250 170/250 1

#e

G+ (3.39) = 30/29 · 107/375 = 214/725 = 0.2952.

#e



Code:

library(ineq)

plot(Lc(Members,n=Year))

#3.5

The quantiles coincide approx-

imately with the bisection line. This means that the distribution of “length of service”

is similar for men and women. They have worked for about the same amount of time

for the company. For example, the median service time should be approximately

the same for men and women, the first quartile should be very similar too, the third

quartile should be similar too, and so on. However, Fig. 3.8a shows a somewhat dif-

ferent pattern: the quantiles for men are consistently higher than those for women.

For example, the median salary will be higher for men. In conclusion, we see that

men and women have a similar length of service in the company, but earn less.

#3.6

There are many ways in which a “mode” function can be

programmed. We present a simple and understandable solution, not the most efficient

one. Recall that table constructs a frequency table. This table shows immediately

which value(s) occur(s) most often in the data. How can we extract it? Applying the

names function on the table returns the names of the values represented in the table.

The only thing we need to do is to choose the name which corresponds to the value

which has been counted most. This yields the following function:

code:

mymode <- function(vec){

mt <- table(vec)

names(mt)[mt == max(mt)]

}

#3.7

Consider a country in which 90% of the wealth is owned by 20% of

the population, the so-called upper class. For simplicity, let us assume that the wealth

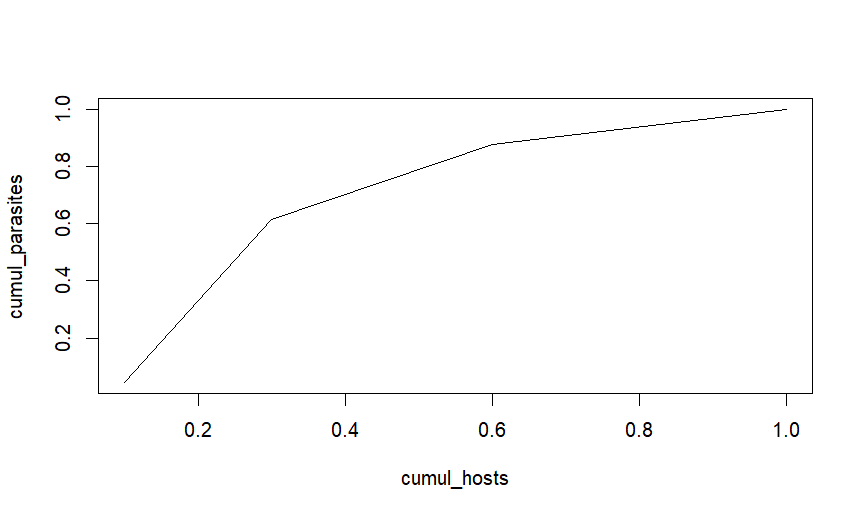
is distributed equally within this class.

#3.9

cumul\_parasites <- cumsum(Contribution)/max(cumsum(Contribution))

cumul\_hosts <- cumsum(Person)/max(cumsum(Person))

plot(cumul\_hosts, cumul\_parasites, type= "l”)



(b) Determine and interpret the standardized Gini coefficient.

(b) We can calculate the Gini coefficient as follows:

G (3.37) = 1 – 1/4

[(0 + 0.044) + (0.044 + 0.168) + (0.168 + 0.428)

+ (0.428 + 1)] = 1 – 1/4 · 2.28 = 0.43.

G+ (3.39) = n/(n – 1(G = 4/3 · 0.43 = 0.57.))

© Hệ số Gini không đổi khi khoản đầu tư tương đối không đổi.

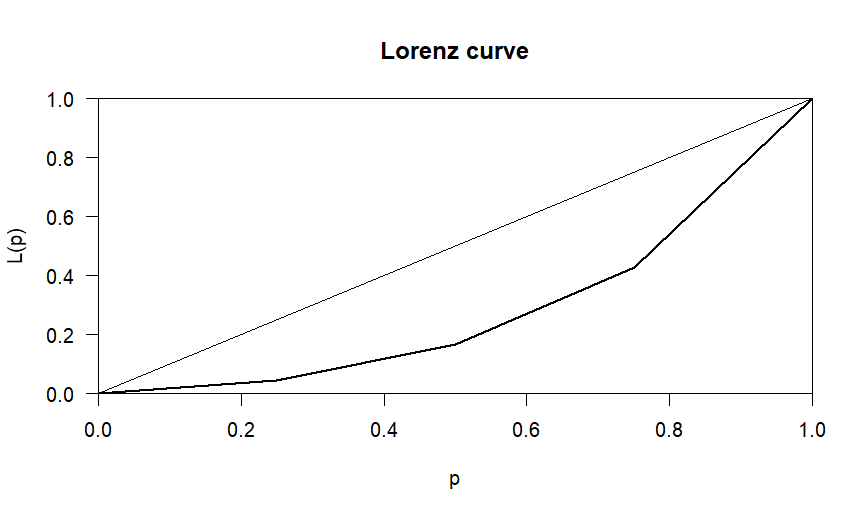
d)

library(ineq)

investment <- c(800,10300,4700,2200)

plot(Lc(investment))

ineq(investment)



#3.10

setwd('C:/yourpath')

pizza <- read.csv('pizza\_delivery.csv')

attach(pizza)

summary(pizza[,c('time','temperature','bill','pizzas')])

b)

quantile(time,probs=0.99)

quantile(temperature,probs=0.99)

C)

amdev <- function(mv){1/length(mv)\*sum(abs(mv-mean(mv)))}

amdev(temperature)

d)

sc.time <- scale(time)

mean(sc.time)

var(sc.time)

e)

boxplot(temperature,range=0)

boxplot(time,range=0)

F)

tc <- cut(time,breaks=seq(10,60,10))

weighted.mean(c(15,25,35,45,55),table(tc)/sum(table(tc)))

[1] 34.18641

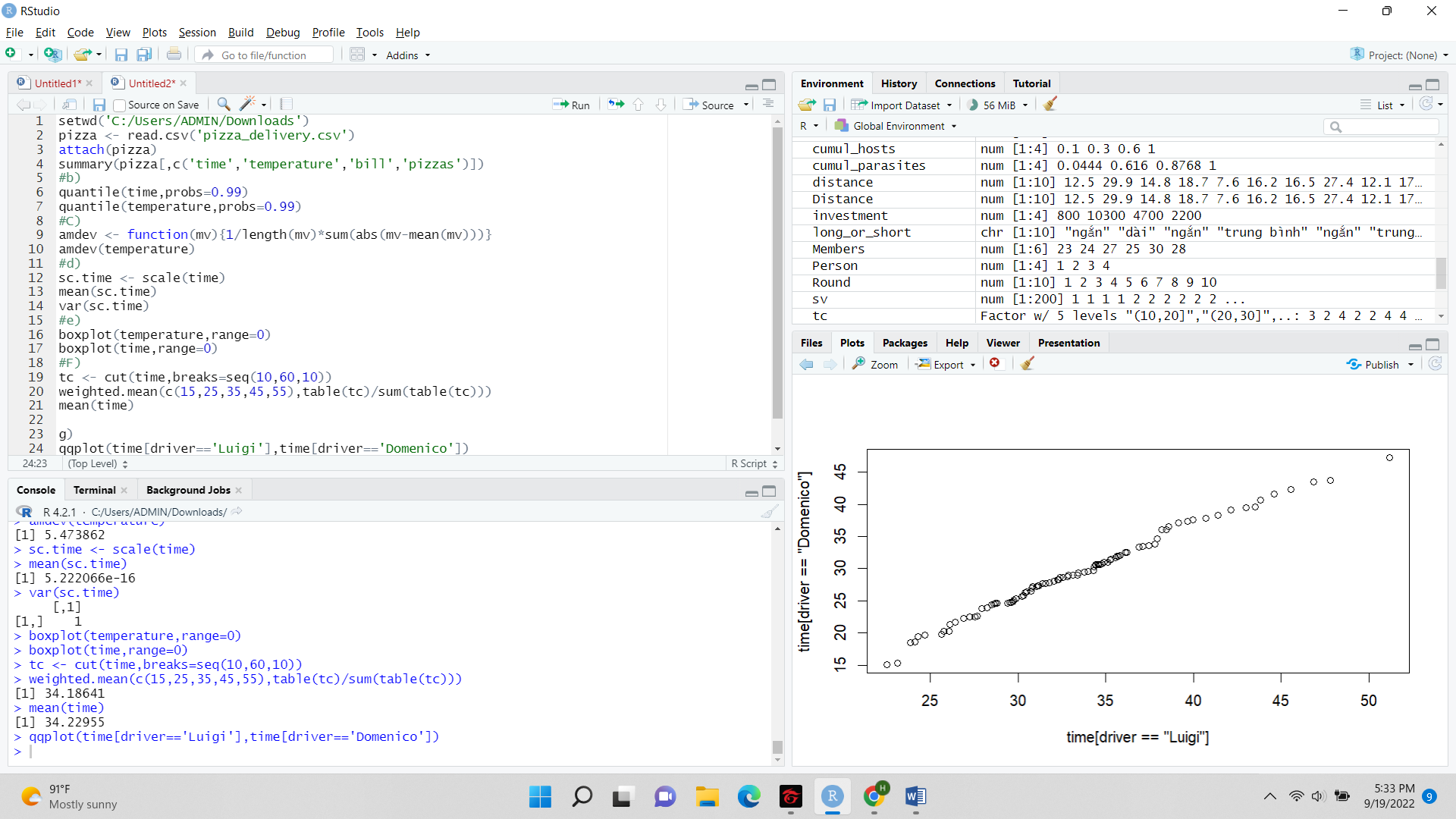
mean(time)

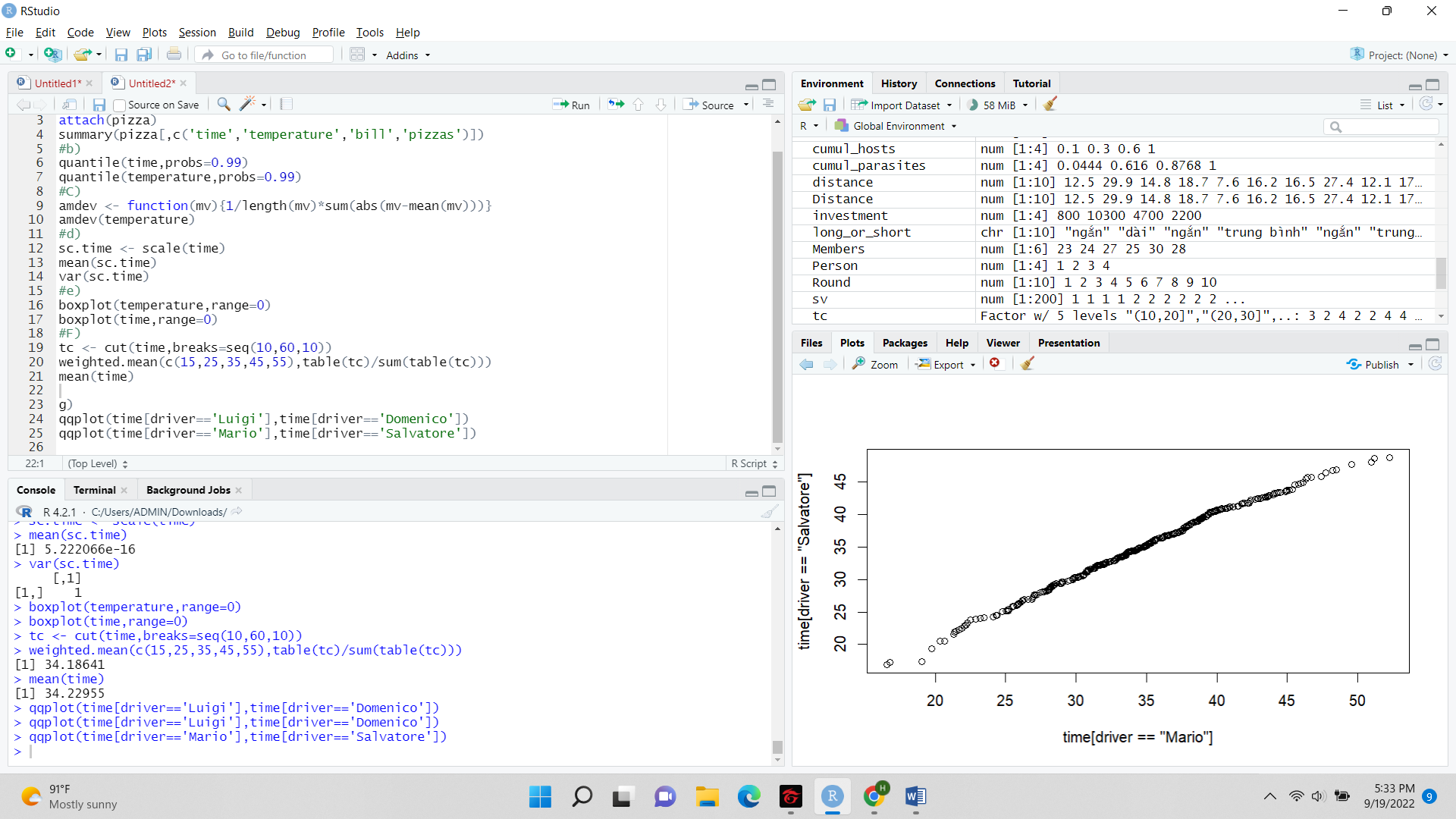
[1] 34.22955

g)

qqplot(time[driver=='Luigi'],time[driver=='Domenico'])

qqplot(time[driver=='Mario'],time[driver=='Salvatore'])





4. Association of two variables

4.1 Tóm tất sự phân bố của hai biến số rời rạc

Khi cả hai biến đều rời rạc, thì có thể liệt kê tất cả các kết hợp giá trị của hai biến và đếm tần suất các kết hợp này xảy ra trong dữ liệu Hãy xem xét ví dụ về tiền lương trong phần giới thiệu chương này, trong đó cả hai biến đều là nhị phân. Có bốn sự kết hợp có thể có của các loại biến (nhóm nữ và nhóm lương thấp, nhóm nữ và nhóm lương cao, nhóm nam và nhóm lương thấp và nhóm nam và nhóm lương cao). Có thể đưa ra mô tả đầy đủ về sự kết hợp xảy ra khi hai biến này có thể được đưa ra bằng cách đếm, đối với mỗi kết hợp, số lượng đơn vị mà sự kết hợp này được đo lường. Sau đây, chúng ta tổng quát hóa khái niệm này thành hai biến mà mỗi biến có thể có một số giá trị hoặc danh mục tùy ý (nhưng cố định).

4.1.1 Bảng dự phòng cho dữ liệu rời rạc

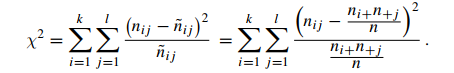
4.1.2 Phân phối tần số chung, cận biên và có điều kiện

Khi dữ liệu về hai biến được tóm tắt trong một bảng dự phòng, có một số khái niệm có thể giúp chúng tôi nghiên cứu các đặc điểm của dữ liệu. Vì ví dụ, cách các giá trị của cả hai biến hoạt động cùng nhau, cách các giá trị của một biến hoạt động khi một biến khác được giữ cố định, v.v. Những tính năng này có thể được nghiên cứu bằng cách sử dụng các khái niệm về phân phối tần số chung, phân bố tần số biên và phân phối tần số có điều kiện. Nếu tần số tương đối được sử dụng thay thế của tần số tuyệt đối, sau đó chúng ta nói về phân bố tần số tương đối chung, phân bố tần số tương đối mar ginal và phân bố tần số tương đối có điều kiện.

4.2 Các biện pháp liên kết cho hai biến rời rạc

Khi hai biến không độc lập, thì chúng được liên kết với nhau. Hiệp hội của họ có thể yếu hoặc mạnh. Bây giờ chúng tôi mô tả một số biện pháp phổ biến của liên kết Các phép đo liên kết mô tả mức độ liên kết giữa hai biến và cũng có thể có một hướng đi. Lưu ý rằng nếu các biến được xác định trên thang đo danh nghĩa, thì chưa thể nói gì về hướng liên kết, chỉ nói về sức mạnh.

4.2.1 Thống kê x2 của pearson

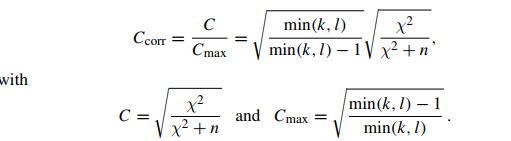


4.2.2 Thống kê carmer’s V



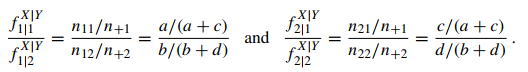
4.2.3 Hệ số dự phòng C

Một tùy chọn khác để chuẩn hóa χ2 được đưa ra bởi một phiên bản hiệu chỉnh của hệ  
số liên tục Pearson:



4.2.4 Tỷ lệ rủi ro và tỷ lệ cược tương đối

Bây giờ chúng tôi giới thiệu các khái niệm về tỷ lệ chênh lệch và rủi ro tương đối. Hãy xem xét một 2 × 2 bảng dự phòng như đã giới thiệu trong Bảng 4.5. Bây giờ, giả sử chúng ta có hai biến X và Y với các phân phối có điều kiện f và f



4.3 Sự kết hợp giữa các biến thứ tự và biến liên tục

Một cách đơn giản để tóm tắt bằng đồ thị sự liên kết giữa hai liên tục biến là để vẽ biểu đồ các quan sát được ghép nối của hai biến trong một hai chiều hệ tọa độ. Nếu n kết hợp quan sát cho hai biến liên tục X và Y có sẵn dưới dạng (xi, yi), i = 1, 2, ..., n, sau đó tất cả các quan sát như vậy có thể được vẽ biểu đồ



4.3.3 Hệ số tương quan cấp bậc của Spearman

Hãy xem xét một tình huống trong đó n đối tượng được xếp hạng theo hai biến X và Y trong một cuộc thi tài năng, người xếp hạng những người tham gia theo hiệu suất của họ. Điều này có nghĩa là đối với mỗi giám khảo, người tham gia kém nhất (với điểm thấp nhất là xi) là xếp hạng 1, người tham gia kém thứ hai (với điểm thấp thứ hai là xi) sẽ nhận được hạng 2, v.v. Vì vậy, mỗi người tham gia đã được xếp hai cấp bậ các thẩm phán khác nhau. Giả sử chúng ta muốn đo lường mức độ liên kết giữa hai bản án khác nhau; nghĩa là, hai nhóm cấp bậc khác nhau. Chúng tôi mong đợi rằng dưới thỏa thuận hoàn hảo.



4.3.4 Các biện pháp sử dụng các cập bất hòa với hòa hợp

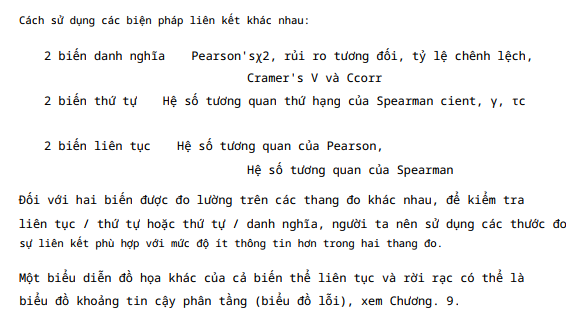
Một khái niệm khác sử dụng các cấp bậc để đo lường mối liên hệ giữa các ables biến thiên theo thứ tự dựa trên các cặp quan sát phù hợp và bất hòa . Nó được minh họa tốt nhất bằng một ví dụ.



4.4 Hình dung các biến từ các thang đo khác nhau

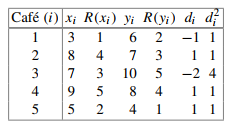
Nếu chúng ta muốn cùng nhau trực quan hóa sự liên kết giữa một biến X, thì đó là danh nghĩa hoặc thứ tự và một biến Y khác liên tục, thì chúng ta có thể sử dụng bất kỳ đồ thị phù hợp với biến liên tục (xem Chaps. 2 và 3) và sản xuất nó cho mỗi loại của biến danh nghĩa / thứ tự. Chúng tôi khuyên bạn nên sử dụng phân tầng các ô hộp hoặc các ECDF phân tầng, vì chúng dễ đọc khi được tóm tắt trong một nhân vât; tuy nhiên, cũng có thể đặt các biểu đồ bên cạnh nhau hoặc trên nhau hoặc chồng lên các biểu đồ mật độ hạt nhân, nhưng chúng tôi không minh họa điều này ở đây trong chi tiết.

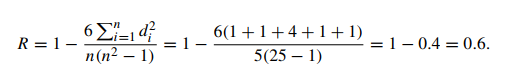
4.5 Các điểm chính và các vấn đề khác



4.6 Bài tập

4.1 a. chúng ta cần bảng sau để tính hệ số tương quan R:



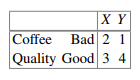


Có mối tương quan thuận từ trung bình đến mạnh giữa xếp hạng của cả hai những người đam mê cà phê. Nói chung, đánh giá cao từ một nhân viên có nghĩa là đánh giá khá cao từ các nhân viên khác.

b) Ở trên, chúng tôi đã xếp các cấp bậc theo thứ tự tăng dần; tức là xi / yi thấp nhất nhận được thứ hạng thấp nhất (1) và xi / yi cao nhất được xếp hạng cao nhất (5). Nếu chúng ta sử dụng thứ tự giảm dần và gán thứ hạng thấp nhất cho các giá trị cao nhất, chúng tôi nhận được các tần số dự kiến (không độc lập) là:

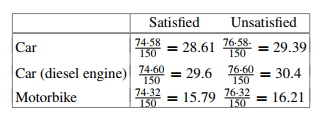
Như trong (a), chúng ta có = 8 và do đó, kết quả là giống hệt nhau: R = 0,6. Tùy thuộc vào việc xếp hạng được chỉ định theo thứ tự tăng hay giảm thứ tự, dấu của di khác nhau, nhưng tính toán của R không bị ảnh hưởng vì các giá trị bình phương của di được sử dụng để tính toán nó và dấu của di do đó không quan trọng.

c. Chúng ta có thể tóm tắt các xếp hạng trong bảng 2 × 2:



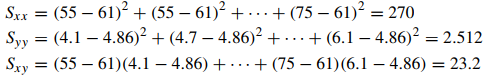
Tỷ lệ chênh lệch là OR = (2 × 4) / (3 × 1) = 2. Cơ hội xếp hạng cà phê là tốt có khả năng xảy ra với người X gấp đôi so với người Y.

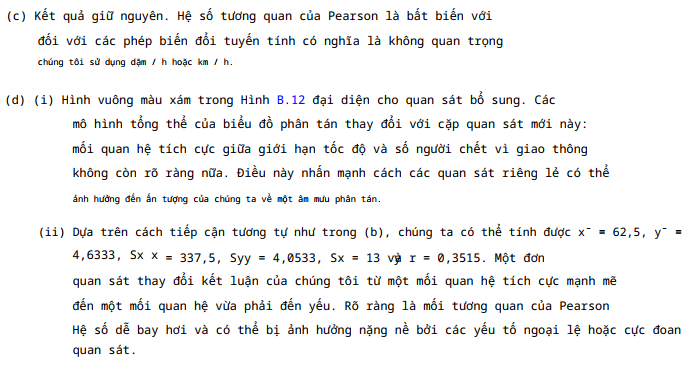
4.2 a) Các tần số dự kiến (không độc lập) là:



4.3 a) Biểu đồ phân tán được cho trong Hình B.12. Các vòng tròn màu đen hiển thị năm quan sát. Một mối quan hệ tích cực có thể được phát hiện: giới hạn tốc độ càng cao, số người chết. Tuy nhiên, “Ý” (quan sát ở trên cùng bên phải) là quan sát mang lại cho biểu đồ một mô hình dễ thấy và thúc đẩy hiển thị của chúng tôi về mối quan hệ tiềm năng.

(b) Sử dụng x¯ = 61 và y¯ = 4,86, chúng ta thu được





5. Tổ hợp

5.2 Hoán vị

Định nghĩa 5.2.1 Xét một tập hợp gồm n phần tử. Mỗi thành phần có thứ tự của n phần tử này được gọi là một hoán vị

Chúng ta phân biệt giữa hai trường hợp: Nếu tất cả các phần tử có thể phân biệt được, thì chúng ta nói về hoán vị mà không cần thay thế. Tuy nhiên, nếu một số hoặc tất cả các phần tử không thể phân biệt được, thì chúng ta nói về hoán vị với sự thay thế. Xin lưu ý rằng ý nghĩa của "thay thế" ở đây chỉ là một quy ước và không đề cập trực tiếp đến các hình vẽ, ví dụ như từ mô hình bình

5.2.1 Hoán vị không cần thay thế

Nếu tất cả n phần tử đều có thể phân biệt được thì N! (5.2) các thành phần khác nhau của các yếu tố này

5.3 Kết hợp

Giả sử rằng không phải tất cả n phần tử đều có thể phân biệt được. Các yếu tố được chia thành các nhóm và các nhóm này có thể phân biệt được. Giả sử, có s nhóm kích thước n1, n2, ..., ns. Tổng số cách sắp xếp n phần tử trong nhóm s là:



5.3.1 Sự kết hợp mà không có sự thay thế và không có sự xem xét của đơn đặt hang

Định nghĩa 5.3.1 Hệ số nhị thức cho bất kỳ số nguyên m và n nào với n ≥ m ≥ 0 được biểu thị và định nghĩa là



5.3.2 Kết hợp mà không cần thay thế và có xem xét đơn đặt hang

Nó được đọc là “n chọn m” và có thể được tính bằng R bằng lệnh sau:



Có một số quy tắc tính toán cho hệ số nhị thức:



Bây giờ chúng ta trả lời câu hỏi có bao nhiêu khả năng khác nhau để rút m trong số n phần tử, tức là m trong số n quả bóng từ một cái bình. Cần phân biệt bốn trường hợp sau:

1. Sự kết hợp không thay thế và không xem xét thứ tự của các yếu tố
2. Sự kết hợp không thay thế và có xem xét thứ tự của của các yếu tố
3. Kết hợp với thay thế và không xem xét thứ tự của các yếu tố
4. Kết hợp với thay thế và xem xét thứ tự của các ments

5.3.1 Sự kết hợp mà không có sự thay thế và không có sự xem xét của đơn đặt hàng

Khi không có sự thay thế và thứ tự của các phần tử cũng không có liên quan thì tổng số các tổ hợp có thể phân biệt được khi vẽ m trong số n phần tử là



5.3.2 Kết hợp mà không cần thay thế và có xem xét đơn đặt hàng

Tổng số các kết hợp khác nhau cho cài đặt mà không cần thay thế và có xem xét thứ tự là

Chart, scatter chart

Description automatically generated

5.3.3 Kết hợp với thay thế và không xem xét đơn đặt hàng

Tổng số các kết hợp khác nhau có thay thế và không có ý niệm về thứ tự là

A picture containing graphical user interface

Description automatically generated

Lưu ý rằng đây là hai biểu diễn tuân theo định nghĩa của hệ số nhị thức nhưng thường chỉ biểu diễn đầu tiên được sử dụng trong sách giáo khoa

5.3.4 Kết hợp với thay thế và xem xét đơn đặt hàng

Tổng số các kết hợp khác nhau của các số nguyên m và n có thay thế và khi thứ tự có liên quan là



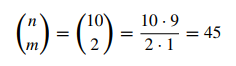
5.4 Các điểm chính và các vấn đề khác

Text, table

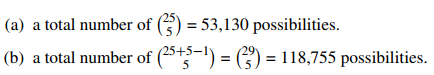
Description automatically generated with medium confidence

5.5 Bài tập

5.1 Có n = 10 khách và m = 2 khách bắt tay nhau. Thứ tự không quan trọng: hai vị khách bắt tay nhau bất kể ai được “rút thăm trước”. Cũng không thể bắt tay với chính mình, có nghĩa là về mô hình bình, chúng tôi có trường hợp “không thể thay thế”. Do đó, chúng tôi nhận được giải pháp là



Chúng tôi giả định rằng việc biết học sinh nào được kiểm tra vào thời điểm nào là không phù hợp. Do đó, chúng tôi đang xử lý các kết hợp mà không cần xem xét thứ tự. Ta có n = 25 và m = 5 nên thu được:



5.3 Tấm ván gồm n = 381 hải lý. Mỗi nút đều có hoặc không. Do đó, chúng ta có thể giả định là một "bản vẽ" không có thay thế theo nghĩa là mỗi nút chỉ có thể được vẽ (chiếm) một lần. Nếu chúng ta đặt m = 64 bộ đếm trên bàn cờ, thì chúng ta có thể đồng thời nghĩ đến việc “vẽ” 64 hải lý bị chiếm đóng trong tổng số 381 hải lý. Thứ tự mà chúng tôi vẽ không liên quan — một nút thắt là 381 ≈ 4,35 · đã chiếm hoặc không. Tổng số kết hợp là 64 lấy số cuối cùngsửtrong dụng 1073. Rlệnh bằng Chúng select cách tôi (381,64).

5.4 (a) Khách hàng lấy bia “thay thế” vì khách hàng có thể chọn trong số bất kỳ loại bia nào cho từng vị trí trong khay. Người ta cũng có thể nghĩ đến một mô hình bình đựng 6 quả bóng liên quan đến sáu loại bia khác nhau, nơi các loại bia được rút ra thay thế và đặt trên khay. Thứ tự đặt các loại bia trên khay không liên quan. Do đó chúng tôi có.



(b) Nếu khách hàng khăng khăng muốn có ít nhất một cốc bia cho mỗi nhà máy bia trên khay của mình, thì 6 trong số 20 vị trí của khay đã được sử dụng. Dựa trên những suy nghĩ tương tự như trong (a), chúng tôi tính tổng số kết hợp là.



5.6 Có n = 12 chữ cái khác nhau cho m = 4 vị trí của mã thành viên. Mỗi chữ cái có thể được sử dụng nhiều lần nếu muốn và do đó chúng tôi thu được nm = 124 = 20,736 kết hợp có thể có. Do đó, chúng tôi kết luận rằng còn lại đủ mã thành viên. Tuy nhiên, điều này có thể không kéo dài và cửa hàng sách có thể vẫn muốn tạo ra một giải pháp khác cho các mã thành viên của mình.

# 6.1 Basic Concepts and Set Theory

Một định nghĩa đơn giản (không nghiêm ngặt) về một thử nghiệm ngẫu nhiên yêu cầu rằng chuyên gia- iment có thể được lặp lại bất kỳ số lần nào trong cùng một tập hợp các điều kiện, và kết quả của nó chỉ được biết sau khi hoàn thành thử nghiệm. Một đơn giản và ví dụ cổ điển về một thí nghiệm ngẫu nhiên là tung đồng xu hoặc tung chết. Khi tung một đồng xu, không biết kết quả sẽ như thế nào, đầu hay đuôi, cho đến khi đồng xu được tung. Thử nghiệm có thể được lặp lại và các kết quả khác nhau có thể được quan sát trong mỗi lần lặp lại. Tương tự, khi lăn một con súc sắc, không biết bao nhiêu các chấm sẽ xuất hiện trên bề mặt trên cho đến khi cán khuôn. Một lần nữa, cái chết có thể là được cuộn nhiều lần và số lượng chấm khác nhau thu được trong mỗi lần thử. Một khả thi

A ∪ B Hợp các sự kiện A ∪ B là tập hợp tất cả các sự kiện đơn giản của A và B mà xảy ra nếu ít nhất một trong các sự kiện đơn giản của A hoặc B xảy ra (Hình 6.1, bên trái bên, khu vực bóng mờ màu xám). Xin lưu ý rằng chúng tôi sử dụng từ “hoặc” từ quan điểm thống kê: “A hoặc B” có nghĩa là một sự kiện đơn giản từ A xảy ra, hoặc một sự kiện đơn giản từ B xảy ra hoặc một sự kiện đơn giản là một phần của cả A và B xảy ra.

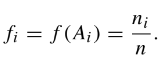
A ∩ B Giao của các biến cố A ∩ B là tập hợp tất cả các biến cố đơn giản A và B xảy ra khi một sự kiện đơn giản xảy ra thuộc về A và B

A\B Sự kiện A \ B chứa tất cả các sự kiện đơn giản của A, không chứa trong B. Sự kiện “A nhưng không phải B” hoặc “A trừ B” xảy ra, nếu A xảy ra nhưng B không xảy ra. Ngoài ra A \ B = A ∩ B ̄ (Hình 6.2, bên trái, vùng bóng mờ màu xám). A ̄ Biến cố A ̄ chứa tất cả các biến cố đơn giản của Ω, biến cố không chứa trong A. Sự kiện bổ sung của A (là “Không phải A” hoặc “A ̄” xảy ra bất cứ khi nào A không xảy ra

A ⊆ B A là một tập con của B. Điều này có nghĩa là tất cả các sự kiện đơn giản của A cũng là một phần của không gian mẫu của B.

# 6.2 Relative Frequency and Laplace Probability

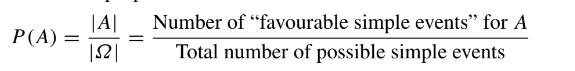
Có một mối liên hệ chặt chẽ giữa tần số tương đối và xác suất của một sự kiện. Ví dụ, một thử nghiệm ngẫu nhiên được mô tả bằng các kết quả có thể có của nó nhận được một số từ 1 đến 6 khi lăn một con súc sắc.



Nếu chúng ta giả định rằng thử nghiệm được lặp lại nhiều lần (toán học- về mặt matic, điều này có nghĩa là n có xu hướng đến vô cùng) và các điều kiện thực nghiệm giữ nguyên (ít nhất là gần đúng) qua tất cả các lần lặp lại, sau đó tương đối tần số f (A) hội tụ đến một giá trị giới hạn cho A. Giá trị giới hạn này được giải thích là xác suất của A và được ký hiệu là P (A), tức là

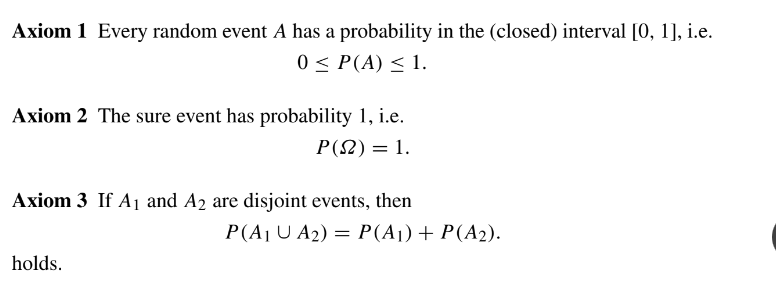


## Definition 6.2.1 The proportion

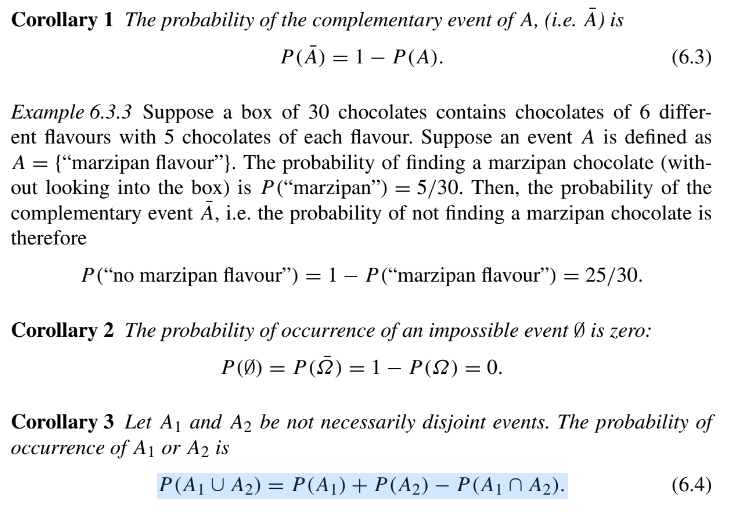


được gọi là xác suất Laplace, trong đó | A | là số chính của A, tức là số sự kiện đơn giản có trong tập A, và | Ω | là số chính của Ω, tức là số sự kiện đơn giản có trong tập Ω.

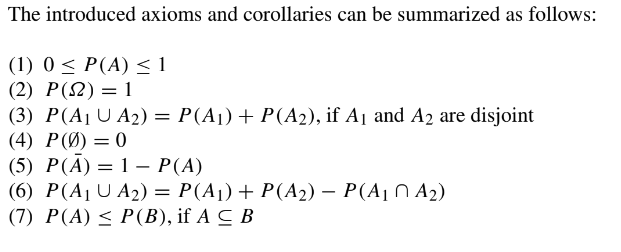
# 6.3 The Axiomatic Definition of Probability



## 6.3.1 Corollaries Following from Kolomogorov’s Axioms



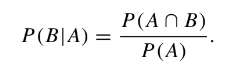
## 6.3.2 Calculation Rules for Probabilities

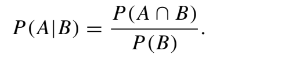


# 6.4 Conditional Probability

Hãy xem xét ví dụ sau để hiểu khái niệm xác suất có điều kiện: Giả sử một xét nghiệm y tế mới được phát triển để chẩn đoán một bệnh nhiễm trùng cụ thể máu. Thử nghiệm được tiến hành trên các mẫu máu của 100 bệnh nhân được chọn ngẫu nhiên

## Definition 6.4.1





## Theorem 6.4.1



# 6.5 Independence

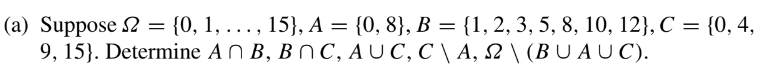
Về mặt trực quan, hai sự kiện là độc lập nếu sự xuất hiện hoặc không xảy ra của một sự kiện sự kiện không ảnh hưởng đến sự xuất hiện hoặc không xảy ra của sự kiện kia. Trong khác từ, hai sự kiện A và B là độc lập nếu xác suất xảy ra của B có không ảnh hưởng đến xác suất xảy ra của A. Trong tình huống như vậy, người ta hy vọng rằng

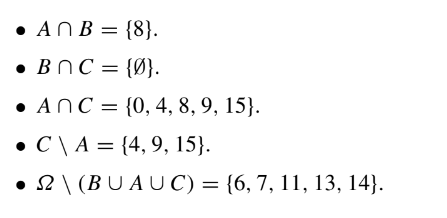


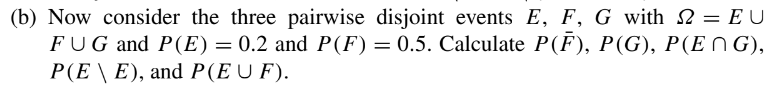


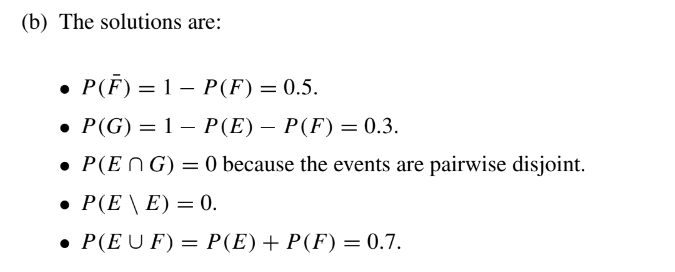
### Bai tap

### 6.1









Bài tập 6.2 Kỳ thi sát hạch cấp giấy phép lái xe bao gồm hai phần dựa trên kiểm tra lý thuyết và kiểm tra thực hành. Giả sử 25% số người thất bại trong việc thực hành kiểm tra, 15% số người không đạt trong kỳ thi lý thuyết và 10% số người không đạt cả các kỳ thi. Nếu một người được chọn ngẫu nhiên, thì xác suất là bao nhiêu rằng người này

(a) không đạt ít nhất một trong các kỳ thi?

(b) Chỉ thi trượt phần thi thực hành chứ không thi phần thi lý thuyết?

(c) vượt qua thành công cả hai bài kiểm tra?

(d) không đạt bất kỳ trong hai kỳ thi?

Bài tập 6.3 Một trò chơi board mới sử dụng một con xúc sắc mười hai mặt. Giả sử cái chết được cuộn

một lần, xác suất nhận được là bao nhiêu

(a) một số chẵn?

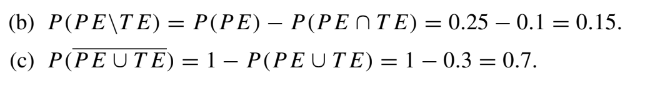
(b) một số lớn hơn 9?

(c) một số chẵn lớn hơn 9?

(d) một số chẵn hoặc một số lớn hơn 9?

Chúng tôi biết rằng xác suất trượt kỳ thi thực hành-ination là P (P E) = 0,25, không đạt yêu cầu phần thi lý thuyết là P (T E) = 0,15, và không đạt cả hai là P (P E ∩ T E) = 0,1.

1. Nếu chúng ta yêu cầu xác suất trượt trong ít nhất một bài kiểm tra, chúng ta ngụ ý kiểm tra lý thuyết hoặc kiểm tra thực hành, hoặc cả hai không được thông qua. Do đó, chúng ta có thể tính toán sự kết hợp của các sự kiện P (P E ∪ T E) =P(P E) + P(T E) − P(T E ∩ P E) = 0.25 + 0.15 − 0.1 = 0.3.



(d) Chúng tôi quan tâm đến xác suất của người đó trượt chính xác trong một kỳ thi. Đây tương ứng với P (M \ C ∪ C \ M) = P (M ∪ C) - P (C ∩ M) = 0,3 - 0,1 = 0,2.

(b) P(P E\T E) = P(P E) − P(P E ∩ T E) = 0.25 − 0.1 = 0.15.

(c) P(P E ∪ T E) = 1 − P(P E ∪ T E) = 1 − 0.3 = 0.7.

(d) Chúng tôi quan tâm đến xác suất của người đó trượt chính xác trong một kỳ thi. Đây tương ứng với P (M \ C ∪ C \ M) = P (M ∪ C) - P (C ∩ M) = 0,3 - 0,1 = 0,2.

Bài tập 6.3 Một trò chơi board mới sử dụng một con xúc sắc mười hai mặt. Giả sử cái chết được cuộn

một lần, xác suất nhận được là bao nhiêu

(a) một số chẵn?

(b) một số lớn hơn 9?

(c) một số chẵn lớn hơn 9?

(d) một số chẵn hoặc một số lớn hơn 9?

Tổng số các biến cố đơn giản có thể xảy ra là || = 12. Cái số sự kiện đơn giản thuận lợi là

(a) | A | = 6 (tức là các số 2, 4, 6, 8, 10, 12). Do đó, P (A) = 6/12 = 1/2.

(b) | B | = 3 (tức là các số 10, 11, 12). Do đó, P (B) = 3/12 = 1/4 .

(c) | C | = 2 (tức là các số 10, 12). Do đó, P (A ∩ B) = 2/12 = 1/6.

(d) | D | = 7 (tức là các số 2, 4, 6, 8, 10, 11, 12). Do đó, P (A ∪ B) = 7/12.

Bài tập 6.4 Nhà Smith là một gia đình có sáu người. Họ đang tổ chức lễ Giáng sinh và

có 12 phần quà, mỗi phần quà cho mỗi thành viên trong gia đình. Các thẻ tên cho mỗi gia đình

thành viên đã được đính kèm với quà tặng. Thật không may, thẻ tên trên quà tặng là

hư hỏng do nước. Giả sử mỗi thành viên trong gia đình bốc thăm ngẫu nhiên hai phần quà. Gì

là xác suất mà ai đó

(a) nhận hai món quà của anh ấy / cô ấy, thay vì nhận những món quà cho một thành viên khác trong gia đình?

(b) không nhận quà của anh ấy / cô ấy, mà là nhận quà cho các thành viên khác trong gia đình?

Lời giải bài tập 6.4 Tổng số sự kiện đơn giản là 12/2

1. Số sự kiện đơn giản thuận lợi là một và do đó P (đúng hai quà tặng

Bài tập 6.5 Một đầu bếp của một chương trình nấu ăn nổi tiếng trên truyền hình đôi khi cho quá nhiều muối

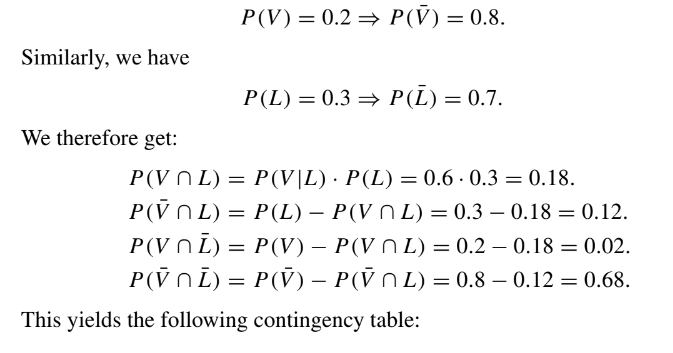
trong món súp bí ngô của mình và xác suất điều này xảy ra là 0,2. Nếu anh ấy đang yêu (mà anh ta là với xác suất 0,3), thì xác suất sử dụng quá nhiều muối là 0,6.

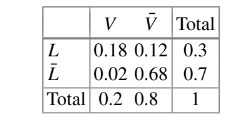
(a) Tạo một bảng dự phòng cho các xác suất của hai biến số "tình yêu" và "quá nhiều muối".

(b) Xác định xem hai biến có độc lập ngẫu nhiên hay không.

### Solution to Exercise 6.5

Let V denote the event that there is too much salt in the soup and let L denote the event that the chef is in love. We know that





(b) Các biến không độc lập ngẫu nhiên vì, ví dụ, P (V)P (L) = 0,3 · 0,2 = 0,06 = 0,18 = P (V ∩ L).

Bài tập 6.6 Tiến sĩ Obermeier yêu cầu người hàng xóm chăm sóc cây húng quế của mình trong khi

anh ấy đi nghỉ phép. Anh ta cho rằng người hàng xóm của anh ta không chăm sóc cây húng quế

với xác suất 1/3. Húng quế chết với xác suất 1/2 khi ai đó chăm sóc của nó và với xác suất 3/4 nếu không có ai chăm sóc nó.

(a) Tính xác suất cây húng quế sống sót sau khi chủ nhân của nó bỏ đi.

(b) Cuối cùng thì cây húng quế cũng chết. Xác suất mà

Hàng xóm của Tiến sĩ Obermeier đã không chăm sóc cây trồng?

### Solution to Exercise 6.6

Bài tập 6.6 Tiến sĩ Obermeier yêu cầu người hàng xóm chăm sóc cây húng quế của mình trong khi

anh ấy đi nghỉ phép. Anh ta cho rằng người hàng xóm của anh ta không chăm sóc cây húng quế

với xác suất 1/3. Húng quế chết với xác suất 1/2 khi ai đó chăm sóc của nó và với xác suất 3/4 nếu không có ai chăm sóc nó.

(a) Tính xác suất cây húng quế sống sót sau khi chủ nhân của nó bỏ đi.

(b) Cuối cùng thì cây húng quế cũng chết. Xác suất mà

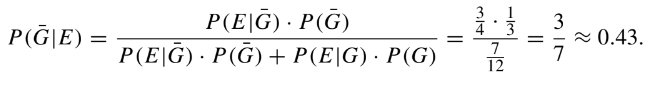
Hàng xóm của Tiến sĩ Obermeier đã không chăm sóc cây trồng?

### Solution to Exercise 6.6

1. Chúng tôi xác định các sự kiện sau: G = Basil được đối xử tốt, G ̄ = Basil thì không được đối xử tốt; E = Basil sống sót, E ̄ = Basil chết. Chúng ta biết rằng;



Chúng ta có thể sử dụng Định lý Bayes để trả lời câu hỏi:



Bài tập 6.7 Một ngân hàng xem xét việc thay đổi chính sách thẻ tín dụng của mình. Hiện tại là 5% tín dụng chủ sở hữu thẻ không thể thanh toán hóa đơn của họ trong bất kỳ tháng nào, tức là họ không bao giờ thanh toán hóa đơn của mình.Trong số những người nói chung có thể thanh toán hóa đơn của họ, vẫn có 20% xác suất rằng hóa đơn được thanh toán quá muộn trong một tháng cụ thể.

(a) Xác suất mà một người nào đó không thanh toán hóa đơn của mình trong một tháng cụ thể là bao nhiêu?

(b) Một chủ thẻ tín dụng đã không thanh toán hóa đơn của mình trong một tháng cụ thể. Cái gì là xác suất rằng anh ta không bao giờ trả lại tiền?

(c) Ngân hàng có nên xem xét việc khóa thẻ tín dụng nếu khách hàng không thanh toán hóa đơn đúng hạn?

Chúng ta xác định các sự kiện và xác suất sau

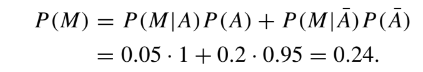
• A: Hóa đơn chưa bao giờ được thanh toán, P (A) = 0,05 ⇒ P (A ̄) = 0,95.

• M: Hóa đơn thanh toán quá muộn, P (M) =?

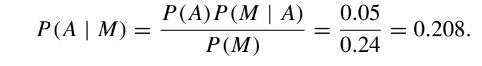
• P (M | A ̄) = 0,2.

• P (M | A) = 1 bởi vì người không bao giờ trả tiền sẽ luôn trả quá muộn.

(a) Chúng tôi quan tâm đến P (M), xác suất mà một người nào đó không thanh toán hóa đơn của mình trong một tháng cụ thể, hoặc vì anh ta không có khả năng hoặc anh ta trả quá muộn. Chúng ta có thể sử dụng định luật xác suất toàn phần để thu được kết quả:

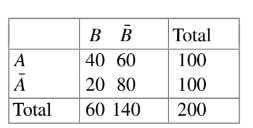


1. Chúng ta có thể sử dụng Định lý Bayes để thu được kết quả:



(c) Nếu hóa đơn không được thanh toán trong một tháng cụ thể, xác suất là 20,8% rằng hóa đơn đó sẽ không bao giờ được thanh toán, và 78,2% rằng nó sẽ được thanh toán. Người ta có thể tranh luận rằng một sự ngăn chặn biện pháp ảnh hưởng đến gần 79% khách hàng đáng tin cậy là không lý tưởng và do đó ngân hàng không nên chặn thẻ tín dụng nếu hóa đơn không được thanh toán đúng hạn.

Bài tập 6.8 Có dịch bệnh ảnh hưởng đến động vật như bò, lợn và động vậters. Giả sử 200 con bò được kiểm tra để xem chúng có bị nhiễm vi rút hay không. Hãy để sự kiện A mô tả xem gần đây một con bò có được chở bằng xe tải hay không và gọi B là trường hợp một con bò được xét nghiệm dương tính với vi rút. Dữ liệu được tóm tắt trong bảng sau:

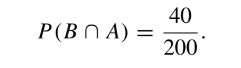


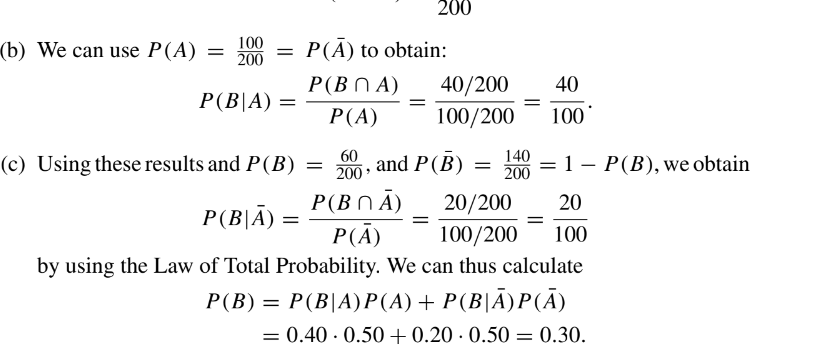
(a) Xác suất để một con bò bị nhiễm bệnh và được chở bằng xe tải là bao nhiêu?gần đây?

(b) Xác suất để có một con bò bị nhiễm bệnh cho rằng nó đã được chuyển gen là bao nhiêu

được chuyển bằng xe tải?

1. Xác định và giải thích P (B).
2. Toán tử “và” đề cập đến phân phối chung của hai biến. Trong kỳ thi của chúng tôi-xin vui lòng, chúng tôi muốn biết xác suất bị nhiễm bệnh và đã được chuyển đổi được chuyển bằng xe tải. Xác suất này có thể trực tiếp thu được từ





Nghĩa là xác suất bò mắc bệnh là 30%. Ngoài ra, chúng ta có thể chỉ cần xem xét phân phối biên của trường hợp dự phòng bảng lấy P (B) = 60/200 = 0,3.

Bài tập 6.9 Mục tiêu tập luyện bóng đá là một bức tường di động có hai lỗ (đó là

mục tiêu) trong đó để chụp ảnh huấn luyện. Giả sử có hai người chơi A và B.

xác suất bắn trúng mục tiêu của A và B lần lượt là 0,4 và 0,5.

(a) Xác suất để ít nhất một trong số các đấu thủ thành công với cú đánh của anh ta là bao nhiêu?

(b) Xác suất để chính xác một trong các người chơi bắn trúng mục tiêu là bao nhiêu?

(c) Xác suất chỉ B đạt điểm là bao nhiêu?

Solution to Exercise 6.9

(a) The two shots are independent of each other and thus

P(A ∩ B) = 0.4 · 0.5 = 0.2.

P(A ∪ B) = 0.4 + 0.5 − 0.2 = 0.7.

(b) We need to calculate P(A\B ∪ B\A) = 0.4 − 0.2 + 0.5 − 0.2 = 0.5.

©P(B\A) = 0.5 − 0.2 = 0.3.

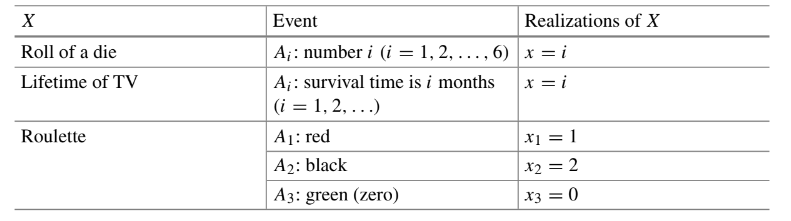
# Chương 7

# 7.1 Random Variables

Các biến ngẫu nhiên giúp chúng ta xem dữ liệu thu thập được là kết quả của một biến ngẫu nhiên

cuộc thí nghiệm.

Table 7.1 Examples of random variables

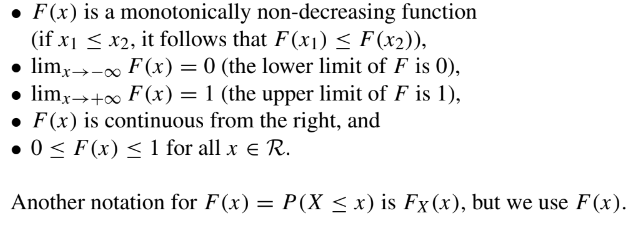


7.2 Cumulative Distribution Function (CDF)

Định nghĩa 7.2.1 Hàm phân phối tích lũy (CDF) của một biến ngẫu nhiên

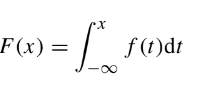
X được định nghĩa là





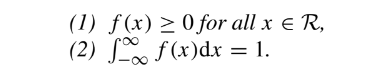
7.2.1 CDF of Continuous Random Variables

Một biến ngẫu nhiên X được cho là liên tục nếu có một hàm f(x) sao cho với mọi x ∈ R



nắm giữ. F (x) là hàm phân phối tích lũy (CDF) của X và f (x) là hàm mật độ xác suất (PDF) của x và d dx F (x) = f (x) với mọi x là điểm liên tục của f

Đối với một hàm f (x) là một hàm mật độ xác suất (PDF) của X cần thỏa mãn các điều kiện sau:



Gọi X là biến ngẫu nhiên với CDF F (x). Nếu x1 <x2, trong đó x1 và x2 là các hằng số đã biết,



Xác suất của một biến ngẫu nhiên liên tục lấy một biến cụ thể giá trị x0 bằng 0



iệc xác định một chức năng, ví dụ như CDF, rất đơn giản trong R: Người ta có thể sử dụng

lệnh hàm theo sau bằng cách chỉ định các biến mà hàm đánh giá trong dấu ngoặc tròn (ví dụ: x) và chính hàm trong dấu ngoặc nhọn (ví dụ: x / 20). Các chức năng có thể được vẽ bằng lệnh curve:

cdf <- function(x){1/20 ∗ x}

curve(cdf,from=0,to=20)

# 7.2.2 CDF of Discrete Random Variables

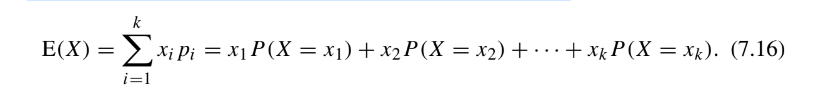
Một biến ngẫu nhiên X được xác định là rời rạc nếu xác suất của nó không gian là hữu hạn hoặc có thể đếm được, tức là nếu nó chỉ chiếm một số hữu hạn hoặc có thể đếm được các giá trị. Lưu ý rằng một tập hợp V được cho là có thể đếm được, nếu các phần tử của nó có thể được liệt kê, tức là có sự tương ứng 1-1 giữa V và các số nguyên dương.

# 7.3 Expectation and Variance of a Random Variable

## 7.3.1 Expectation

The expectation of a continuous random variable X, having the probability density function f (x) with 

Đối với một biến ngẫu nhiên rời rạc X, nhận các giá trị x1, x2, ... tương ứng với xác suất p2, p2, ..., kỳ vọng của X được xác định là



## 7.3.2 Variance

Phương sai mô tả sự thay đổi của một biến ngẫu nhiên. Nó đưa ra một ý tưởng về nồng độ hoặc độ phân tán của các giá trị xung quanh trung bình cộng của phân phối.



## 7.3.3 Quantiles of a Distribution

## 7.4 Tschebyschev’s Inequality

Nếu chúng ta không biết phân phối của một biến ngẫu nhiên X, chúng ta vẫn có thể đưa ra trạng thái- ments về xác suất X nhận các giá trị trong một khoảng thời gian nhất định (phải đối xứng xung quanh kỳ vọng μ) nếu giá trị trung bình μ và phương sai σ2 của X là đã biết.

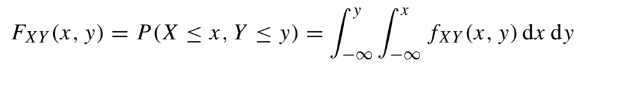


## 7.5 Bivariate Random Variables

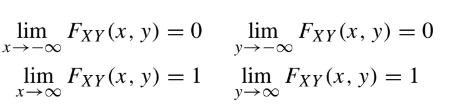
Có nhiều tình huống mà chúng tôi muốn phân tích nhiều hơn một biến, nói hai biến. Khi chúng ta có nhiều hơn một biến, thì không chỉ sự phân phối riêng lẻ của họ cũng như sự phân phối chung của họ có thể được quan tâm

## Definition 7.5.1

Một biến ngẫu nhiên hai biến (X, Y) là liên tục nếu có một hàm tion fXY (x, y) sao cho

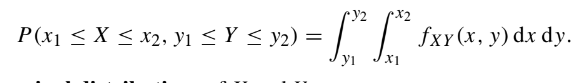


Hàm FXY (x, y) là hàm phân phối tích lũy chung của X và Y; hàm phân phối chung được ký hiệu là fXY (x, y) và fXY (x, y) phải đáp ứng các điều kiện thông thường của một hàm mật độ. Điều kiện cần và đủ mà một hàm FXY (x, y) là một hàm phân phối tích lũy hai biến như sau:

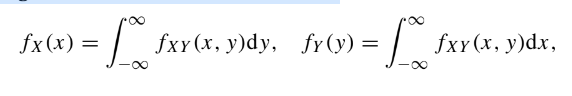


Điều kiện cuối cùng đôi khi được gọi là bất đẳng thức hình chữ nhật. Như trong trường hợp đơn biến, chúng ta có thể sử dụng hàm phân phối tích lũy để tính xác suất trễ khoảng thời gian; tương tự, chúng tôi xem xét khu vực hình chữ nhật được xác định bởi (x1, y1), (x1, y2), (x2, y1) và (x2, y2) trong trường hợp hai biến (thay vì một khoảng[a, b]),

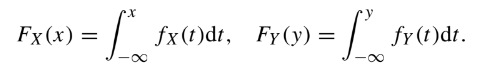
Chúng ta có thể tính toán các xác suất mong muốn như sau



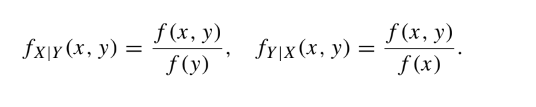
Phân phối cận biên của X và Y là



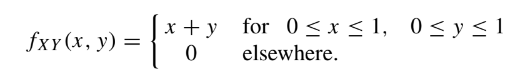
tương ứng. Tương tự như trường hợp rời rạc, fX (x) và fY (y) cũng mô tả phân phối sự phân bố của X vô điều kiện trên Y và sự phân bố của Y vô điều kiện trên X.phân phối cận biên tích lũy laf



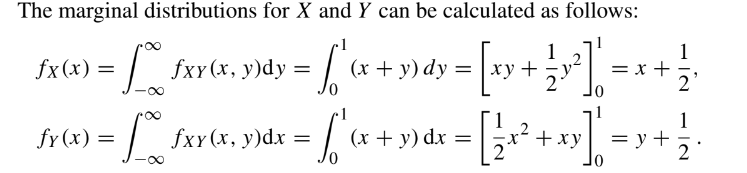
Các phân phối có điều kiện có thể nhận được bằng tỷ lệ của khớp và biên phân phối:



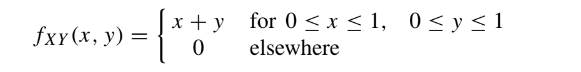
Example 7.5.2 Xét hàm



Giả sử X và Y đại diện cho nồng độ của hai loại thuốc trong cơ thể người. Sau đó, fXY (x, y) có thể đại diện cho tổng của hai nồng độ thuốc trong cơ thể người thân hình. Vì có vô hạn khả năng nhận ra của cả X và Y, chúng tôi đại diện cho sự phân bố chung của chúng trong một hình chứ không phải một bảng, xem



### Definition 7.5.2



với phân phối cận biên của X và Y là 

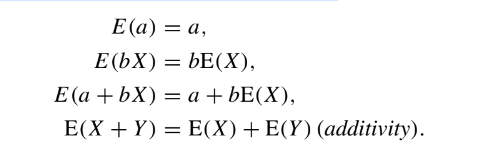
 theo đó X và Y không

sống độc lập. Giải thích là nồng độ của hai loại thuốc không

sống độc lập

## 7.6 Calculation Rules for Expectation and Variance

Quy tắc tính toán cho kỳ vọng. Với mọi giá trị không đổi a và b, và bất kỳ các biến ngẫu nhiên X và Y, các quy tắc sau giữ nguyên



Example 7.6.1: Ví dụ 7.2.3 trong đó chúng tôi đã minh họa kết quả như thế nào của một thử nghiệm cuộn khuôn có thể được ghi lại bởi một biến ngẫu nhiên. Có 6 sự kiện, và X có thể nhận các giá trị x1 = 1, x2 = 2, ..., x6 = 6. Xác suất của sự xuất hiện của bất kỳ số nào là P (X = xi) = 1/6 và kỳ vọng đã được tính như 3.5. Hãy xem xét hai tình huống khác nhau:

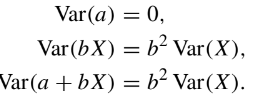
1. Giả sử con súc sắc nhận giá trị 10, 20, 30, 40, 50 và 60 thay vì các giá trị 1, 2, 3, 4, 5 và 6. Biến ngẫu nhiên Y = 10X mô tả điều này một cách phù hợp, và kỳ vọng của nó là



Nếu chúng ta tung hai con xúc xắc X1 và X2, thì kỳ vọng cho tổng của hai kết quả là



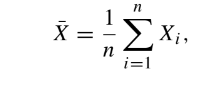
Quy tắc tính toán cho phương sai. Gọi a và b là bất kỳ hằng số nào đã biết và X là một biến ngẫu nhiên (rời rạc hoặc liên tục). Sau đó, chúng tôi có các quy tắc sau:



## 7.6.1 Expectation and Variance of the Arithmetic Mean

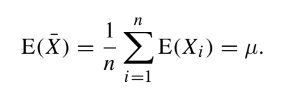
Definition 7.6.1 Chúng tôi xác định các biến ngẫu nhiên X1, X2, ..., Xn là i.i.d. (vô hạn- được phân phối giống nhau), nếu tất cả Xi tuân theo cùng một phân phối và gẫu nhiên độc lập với nhau.

Cho X1, X2, ..., Xn là n i.i.d. biến ngẫu nhiên với E (Xi) = μ và Var (Xi) = σ2, i = 1, 2, ..., n. Giá trị trung bình cộng của các biến này được cho bởi

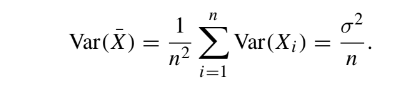


lại là một biến ngẫu nhiên tuân theo một phân phối với kỳ vọng nhất định

và phương sai. Một hàm của các biến ngẫu nhiên được gọi là thống kê. Bằng cách sử dụng (7.29) và (7.31), chúng tôi thu được



Nếu chúng ta áp dụng (7.34) và nhớ rằng các biến độc lập với nhau, chúng ta có thể cũng tính toán phương sai như



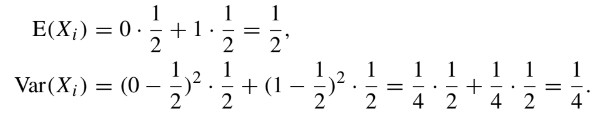
### Example 7.6.3

Nếu chúng ta tung một đồng xu, chúng ta thu được đầu hoặc đuôi, và do đó,

P (“đầu”) = P (“đuôi”) = 1/2. Nếu chúng ta tung đồng xu n lần, chúng ta có mỗi lần tung

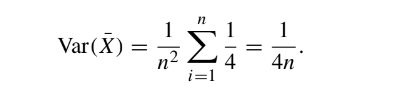


Thật đơn giản để tính toán kỳ vọng và phương sai cho mỗi lần tung đồng xu:



 mô tả tần số tương đối của các đầu khi

đồng xu được tung n lần. Bây giờ chúng ta có thể áp dụng (7,35) và (7,36) để tính toán



## 7.7.1 Covariance

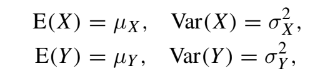
### Definition 7.7.1

Hiệp phương sai giữa X và Y được định nghĩa là

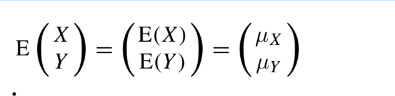


Hiệp phương sai là dương nếu, trung bình, các giá trị lớn hơn của X tương ứng với các giá trị của Y; nó là âm nếu, trung bình, các giá trị lớn hơn của X tương ứng với các giá trị của Y.

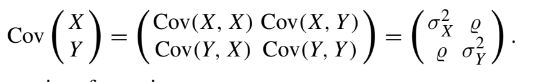
Hàm mật độ xác suất của bất kỳ biến ngẫu nhiên lưỡng biến nào (X, Y) là char-ddược thực hiện bởi kỳ vọng và phương sai của cả X và Y



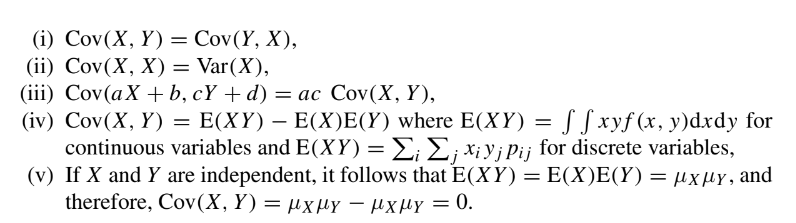
cũng như hiệp phương sai của chúng. Chúng tôi có thể tóm tắt các đặc điểm này bằng cách sử dụng kỳ vọng vectơ



và ma trận hiệp phương sai



Các tính chất quan trọng của hiệp phương sai là



### Theorem 7.7.1

(Định lý cộng) Phương sai của tổng (phép trừ) của X và Y được đưa ra bởi



Nếu X và Y độc lập, thì Cov (X, Y) = 0 và do đó Var (X ±Y) = Var (X) + Var (Y). Chúng ta bỏ qua phần chứng minh của định lý này.

## 7.7.2 Correlation Coefficient

Nếu X và Y độc lập, chúng cũng không tương quan. Tuy nhiên, nếu chúng không liên quan thì chúng không nhất thiết phải độc lập.

### Example 7.7.2

Trong ví dụ 7.6.2, chúng tôi ước tính hiệp phương sai giữa thời gian chờ

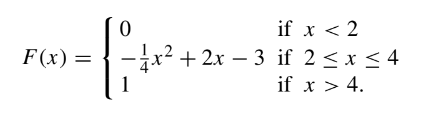
Thời gian xe buýt và thời gian chờ tàu: Cov (X, Y) = 0. Mối tương quan

do đó hệ số cũng bằng 0 cho thấy không có mối quan hệ tuyến tính giữa sự chờ đợi

thời gian cho xe buýt và xe lửa.

7. Bài tập

Hãy xem xét hàm phân phối tích lũy sau của một ngẫu nhiên biến X:

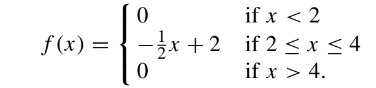


(a) PDF của X là gì?

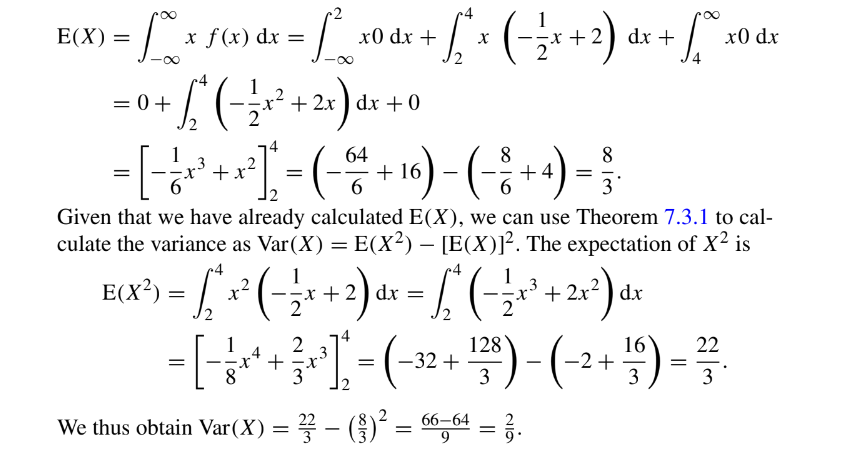
(b) Tính P (X <3) và P (X = 4).

(c) Xác định E (X) và Var (X).

1. The first derivative of the CDF yields the PDF, F (x) = f (x):



1. Từ Định lý 7.2.3, chúng ta biết rằng với mọi biến liên tục P (X = x0) = 0 và do đó P (X = 4) = 0. Ta tính được P (X <3) = P (X ≤ 3) - P (X = 3) = F (3) - 0 = −9/4 + 6 - 3 = 0,75.

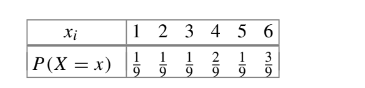


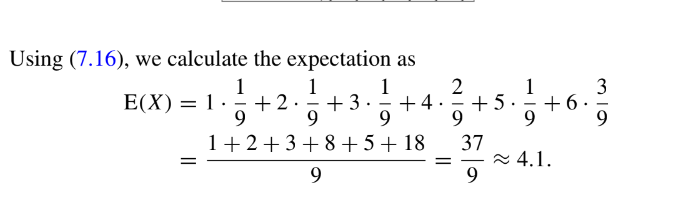
Bài tập 7.2 Joey điều khiển một con súc sắc để tăng cơ hội thắng một ván cờ chống lại bạn bè của mình. Trong mỗi vòng, một con súc sắc được lăn và những con số lớn hơn thường là thuận lợi. Hãy xem xét biến ngẫu nhiên X biểu thị kết quả của viên súc sắc và các xác suất tương ứng P (X = 1 = 2 = 3 = 5) = 1/9, P (X = 4) = 2/9,và P (X = 6) = 3/9.

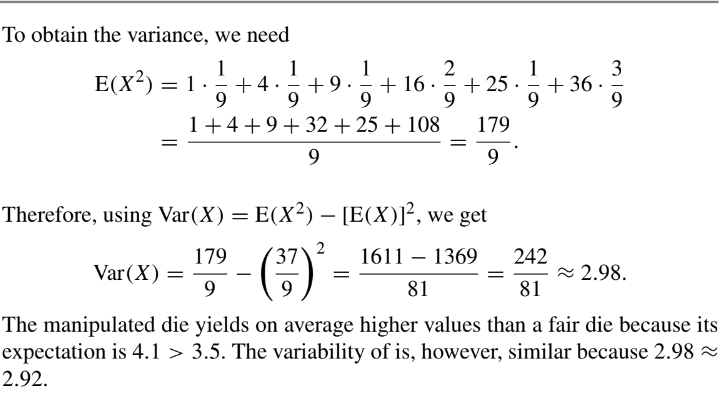
(a) Tính và giải thích kỳ vọng và phương sai của X.

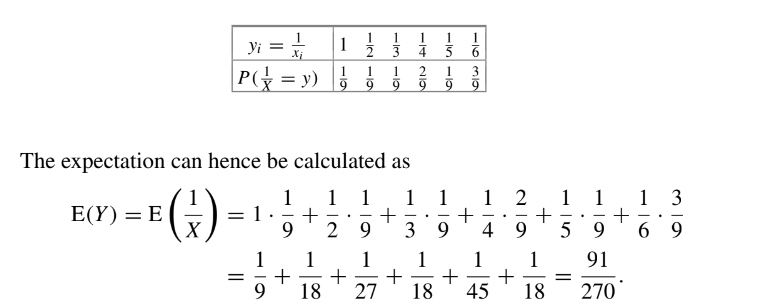
(b) Hãy tưởng tượng rằng trò chơi trên bàn cờ chứa một hành động khiến người chơi sử dụng 1 / X chứ không phải X. Kỳ vọng của Y = 1 / X là bao nhiêu? Có phải E (Y) = E (1 / X) =1 / E (X)?

1. Hàm khối lượng xác suất của X là



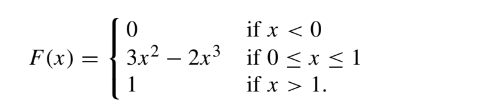






So sánh kết quả từ (a) và (b) cho thấy rõ ràng rằng E (1/X) = 1 VÍ DỤ). Hồi tưởng rằng E (bX) = bE (X). Thật thú vị khi thấy rằng đối với một số phép biến đổi T (X) nó cho rằng E (T (X)) = T (E (X)), nhưng đối với một số người thì không. Điều này nhắc nhở chúng tôi hãy cẩn thận khi nghĩ đến ý nghĩa của các phép biến hình.

Bài tập 7.3 Một nhà sản xuất rượu sáng tạo thử nghiệm với nho mới và thêm một rượu mới về kho của anh ấy. Phần trăm bán được vào cuối mùa phụ thuộc vào thời tiết và nhiều yếu tố khác. Nó có thể được mô hình hóa bằng cách sử dụng biến ngẫu nhiên X với CDF là



(a) Vẽ đồ thị của hàm phân phối tích lũy với R.

(b) Xác định f (x).

(c) Xác suất để bán được ít nhất một phần ba số rượu của anh ta, nhưng không nhiều hơn hơn hai phần ba?

(d) Xác định CDF trong R và tính xác suất của c) một lần nữa.

(e) Phương sai của X là bao nhiêu?

### Solution to Exercise 7.3

(a) Có một số cách để vẽ CDF. Một khả năng là xác định hàm

và vẽ nó bằng lệnh curve. Vì hàm có các định nghĩa khác nhau- tions cho các khoảng [∞, 0), [0, 1], (1, ∞], chúng ta cần tính đến điều này Hãy nhớ rằng một câu lệnh logic trong R tương ứng với một số, tức là TRUE = 1 và FALSE = 0; do đó, chúng tôi có thể chỉ cần thêm các phần khác nhau của hàm và

nhân chúng với một điều kiện xác định xem X có nằm trong khoảng không hoặc không .



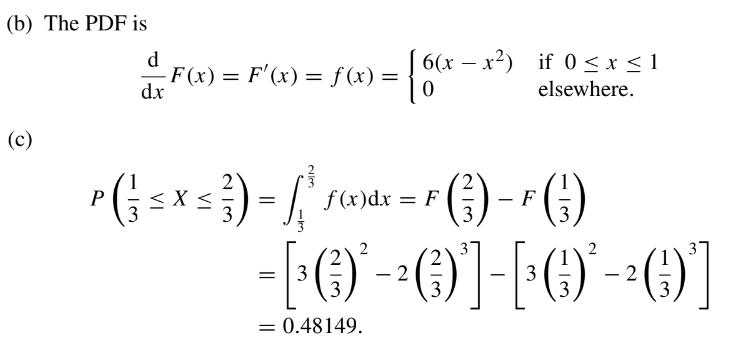
Code:

cdf < −function(x){

(3 ∗ x ˆ2 − 2 ∗ xˆ3) ∗ (x <= 1 & x >= 0) + 1 ∗ (x > 1) + 0 ∗ (x < 0)

}

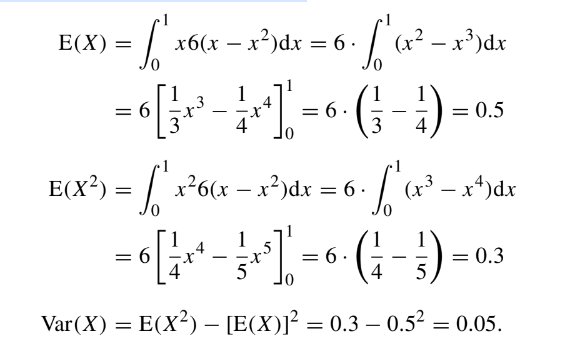
curve(cdf,from=-0.5,to=1.5)



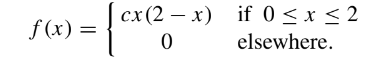
(d) Chúng tôi đã xác định CDF trong (a). Bây giờ chúng ta có thể chỉ cần cắm các giá trị x vào lãi:

cdf(2/3)-cdf(1/3)

1. Phương sai có thể được tính như sau:



Bài tập 7.4 Một chỉ số chất lượng tóm tắt các tính năng khác nhau của một sản phẩm bằng của một điểm số. Các chuyên gia khác nhau có thể ấn định các điểm chất lượng khác nhau tùy thuộc vào trải nghiệm với sản phẩm. Gọi X là chỉ số chất lượng của máy tính bảng. Giả sử hàm mật độ xác suất tương ứng được đưa ra như sau:



(a) Xác định c sao cho f (x) là một PDF thích hợp.

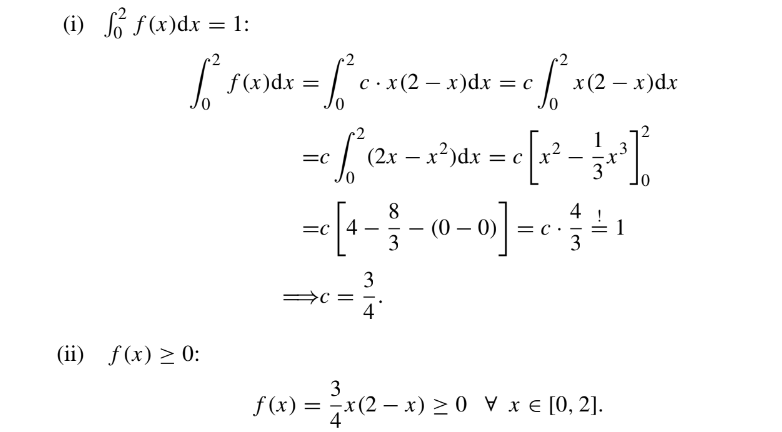
(b) Xác định hàm phân phối tích lũy.

(c) Tính kỳ vọng và phương sai của X.

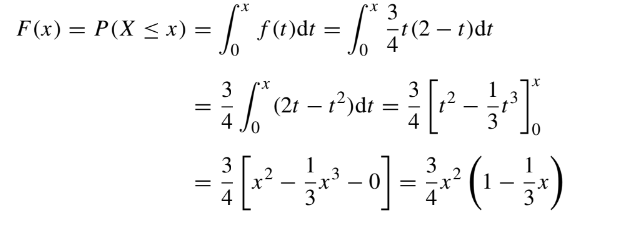
(d) Sử dụng bất đẳng thức Tschebyschev để xác định xác suất X không lệch hơn 0,5 so với kỳ vọng của nó.

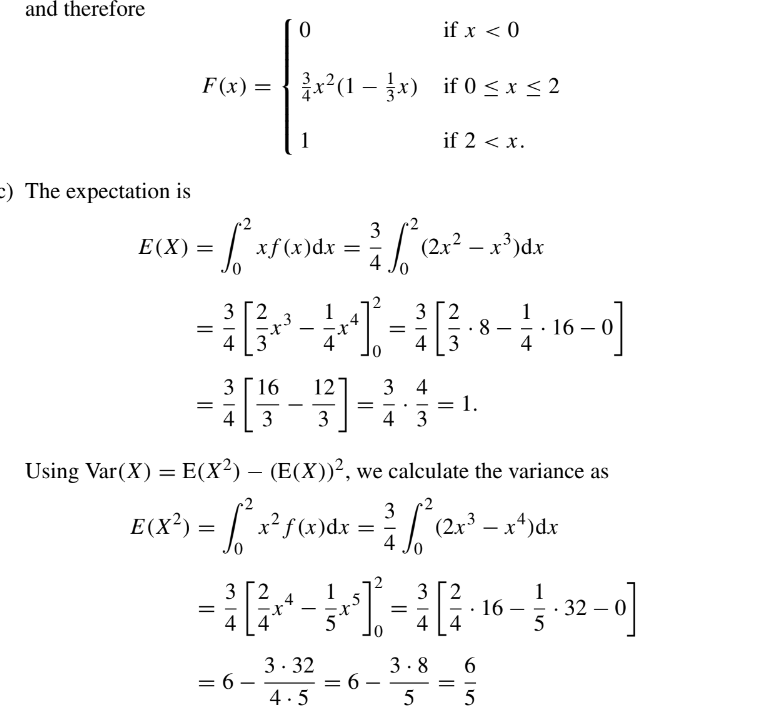
Solution to Exercise 7.4

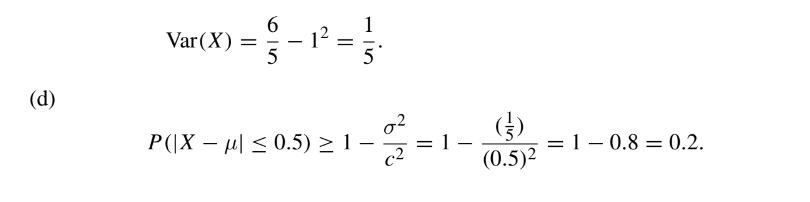
1. Cần thỏa mãn hai điều kiện để f (x) là một tệp PDF thích hợp:



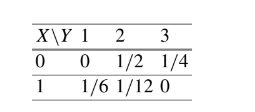
1. We calculate







Bài tập 7.5 Hãy xem xét bản PDF chung cho loại dịch vụ khách hàng X (0 = điện thoại đường dây nóng, 1 = Email) và điểm hài lòng Y (1 = không hài lòng, 2 = sat- hợp nhất, 3 = rất hài lòng):



(a) Xác định và giải thích phân phối biên của cả X và Y.

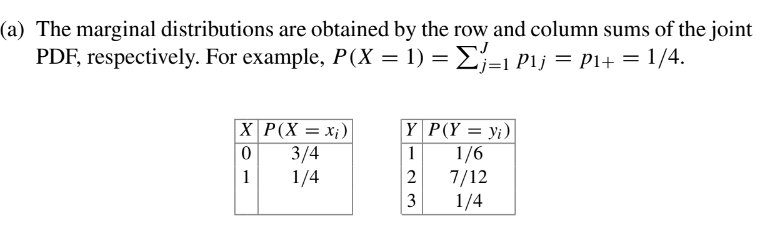
(b) Tính lượng tử 75% cho phân phối biên của Y.

(c) Xác định và giải thích phân phối có điều kiện của mức độ thỏa mãn cho X =1.

(d) Hai biến có độc lập không?

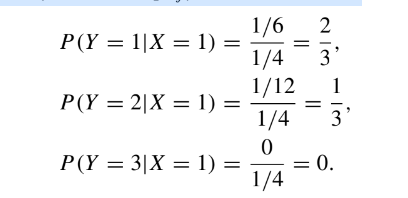
(e) Tính và giải thích hiệp phương sai của X và Y

### Solution to Exercise 7.5



Phân phối cận biên của X cho chúng ta biết có bao nhiêu khách hàng đã tìm kiếm sự trợ giúp thông qua qua điện thoại đường dây nóng (75%) và qua email (25%). Phân phối cận biên của Y đại diện cho sự phân bổ của mức độ hài lòng, làm nổi bật rằng nhiều hơn một nửa số khách hàng (7/12) “hài lòng”.

1. Để xác định lượng tử 75% đối với Y, chúng ta cần tìm giá trị y0,75 với F (y0,75) ≥ 0,75 và F (y) <0,75 với y <y0,75. Y lấy các giá trị 1, 2, 3. Lượng tử không thể là y0,75 = 1 vì F (1) = 1/6 <0,75. Lượng tử 75% là y0,75 = 2 vì F (2) = 1/6 + 7/12 = 3/4 ≥ 0,75 và với tất cả các giá trị nhỏ hơn 2, ta nhận được F (x) <0,75.
2. We can calculate the conditional distribution using P(Y = y j|X = 1) = p1 j / p1+ = p1 j /(1/6 + 1/12 + 0) = p1 j /(0.25). Therefore,



Trong số những người đã sử dụng dịch vụ email khách hàng, hai phần ba không hài lòng,

một phần ba hài lòng và không ai hài lòng lắm.

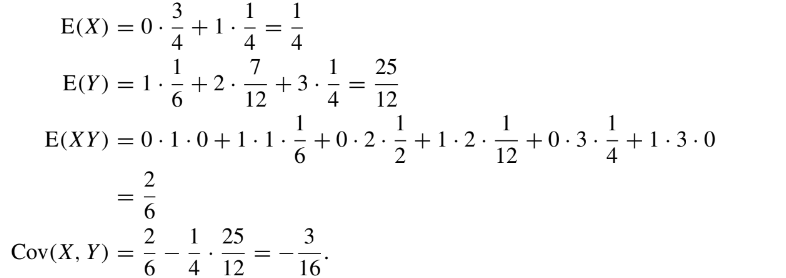
(d) Như chúng ta đã biết từ (7.27), hai biến ngẫu nhiên rời rạc được cho là không thể tách rời-

vết lõm nếu P (X = xi, Y = y j) = P (X = xi) P (Y = y j). Tuy nhiên, trong ví dụ của chúng tôi,

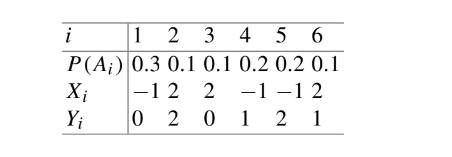
P (X = 0, Y = 1) = P (X = 0) P (X = 1) = 3/4 · 1/6 = 0. Điều này có nghĩa là X v Y không độc lập.

(e) Hiệp phương sai của X và Y được xác định là Cov (X, Y) = E (XY) - E (X) E (Y). chúng tôi

tính toán



Bài tập 7.6 Xét một biến ngẫu nhiên liên tục X với kỳ vọng 15 và phương sai 4. Xác định khoảng [15 - c, 15 + c] nhỏ nhất chứa ít nhất 90% các giá trị của X. Bài tập 7.7 Cho X và Y là hai biến ngẫu nhiên chỉ có 6 biến cố có thể xảy ra— A1, A2, A3, A4, A5, A6 — được định nghĩa:



(a) PDF chung của X và Y là gì?

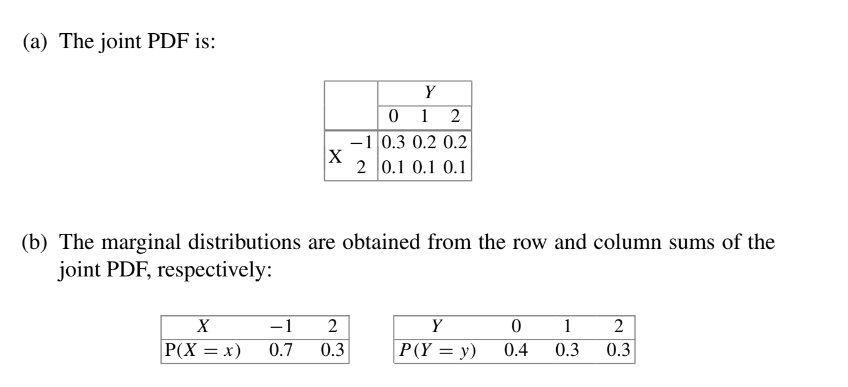
(b) Tính phân phối biên của X và Y.

(c) Cả hai biến có độc lập không?

(d) Xác định PDF liên kết cho U = X + Y.

(e) Tính E (U) và Var (U) và so sánh nó với E (X) + E (Y) và Var (X) +Var (Y), tương ứng.

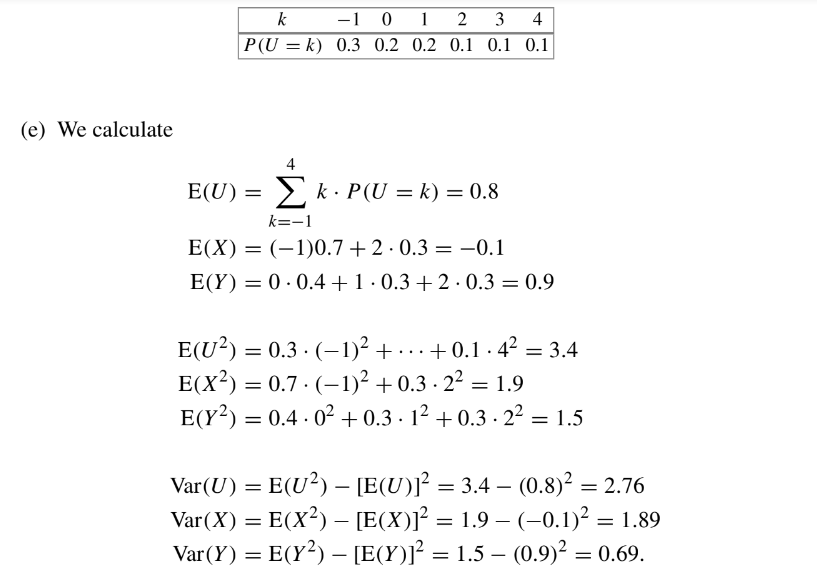
### Solution to Exercise 7.7



(c) Các biến ngẫu nhiên X và Y là độc lập nếu P (X = x, Y = y) = P (X = x) P (Y = y) ∀x, y.

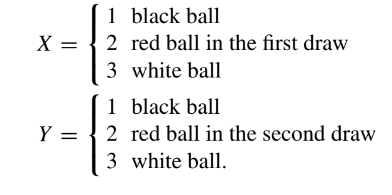
Tuy nhiên, trong ví dụ của chúng tôi, ví dụ,P (X = −1, Y = 0) = 0,3 = P (X = −1) · P (Y = 0) = 0,7 · 0,4= 0,28. Do đó, hai biến không độc lập.

1. Phân phối chung của X và Y có thể được sử dụng để có được phân phối mong muốn của U. Ví dụ, Nếu X = −1 và Y = 0, thì U = X + Y = −1. Hồi đáp- xác suất tive là P (U = −1) = 0,3 vì P (U = −1) = P (X = −1, Y =0) = 0,3 và không có sự kết hợp nào khác của các giá trị X- và Y tạo ra X + Y = −1. Do đó, sự phân bố của U như sau:



Có thể thấy rằng E (X) + E (Y) = −0,1 + 0,9 = 0,8 = E (U). Điều này làm cho có nghĩa vì từ (7.31) chúng ta biết rằng E (X + Y) = E (X) + E (Y). Tuy nhiên, Var (U) = 2,76 = Var (X) + Var (Y) = 1,89 + 0,69. Điều này tuân theo (7.7.1) điều này nói rằng Var (X ± Y) = Var (X) + Var (Y) ± 2Cov (X, Y) và do đó, Var (X ± Y) = Var (X) + Var (Y) chỉ khi hiệp phương sai là 0. Chúng ta biết từ (c) rằng X và Y không độc lập và do đó Cov (X, Y) = 0.

Bài tập 7.8 Hãy nhớ lại mô hình cái bình mà chúng ta đã giới thiệu trong Chap. 5. Xem xét một chiếc bình đựng 8 quả bóng: 4 quả bóng màu trắng, 3 quả bóng màu đen và một quả bóng màu đỏ. Bây giờ, hai quả bóng được rút ra từ bình. Các biến ngẫu nhiên X và Y được xác định như sau:

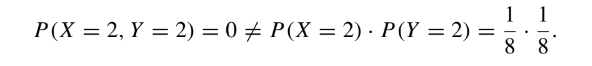


(a) Khi nào X và Y độc lập — khi nào hai quả bóng được rút ra thay thế hoặc không có thay thế?

(b) Giả sử các quả bóng được rút ra sao cho X và Y phụ thuộc. Sử dụng điều kiện phân phối P (Y | X) để xác định PDF chung của X và Y.

(c) Tính E (X), E (Y) và ρ (X, Y). Lời giải cho bài tập 7.8

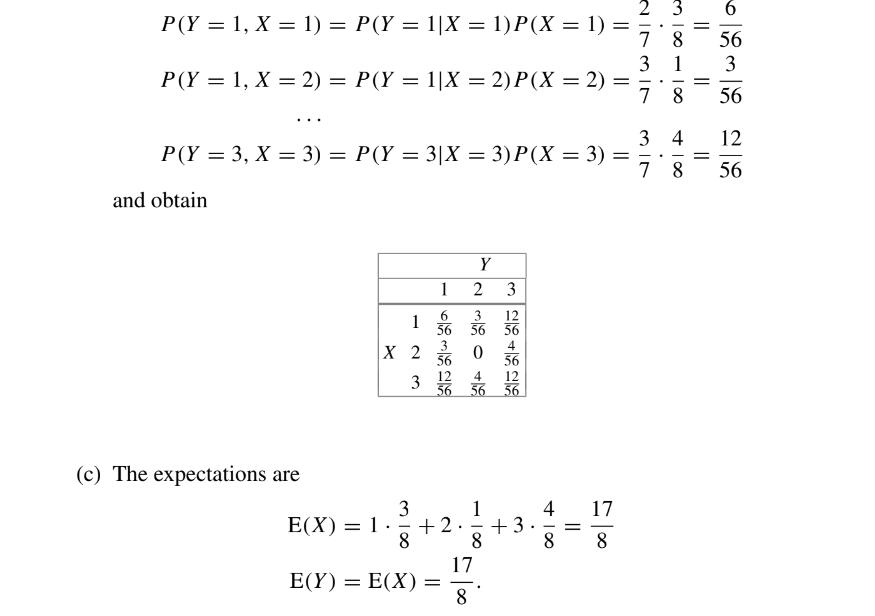
(a) Các biến ngẫu nhiên X và Y là độc lập nếu các quả bóng được rút ra với sự thay thế. Điều này trở nên rõ ràng bằng cách hiểu rằng bản vẽ có thay thế- đề cập đến ngụ ý rằng đối với cả hai lần rút thăm, các quả bóng giống nhau ở trong bình và điều kiện trong mỗi lần rút thăm được giữ nguyên. Lần rút thăm đầu tiên không liên quan đến lần rút thăm thứ hai. Nếu chúng tôi vẽ các quả bóng mà không có vật thay thế, thì lần vẽ đầu tiên có thểcó ý nghĩa đối với lần rút thứ hai: ví dụ: nếu quả bóng đầu tiên được rút ra có màu đỏ, thì cái thứ hai không thể đỏ được vì trong bình chỉ có một viên bi màu đỏ.Điều này có nghĩa là bản vẽ không có thay thế ngụ ý sự phụ thuộc của X và Y. Đâycũng có thể được thấy bằng cách đánh giá giả định về tính độc lập (7.27):

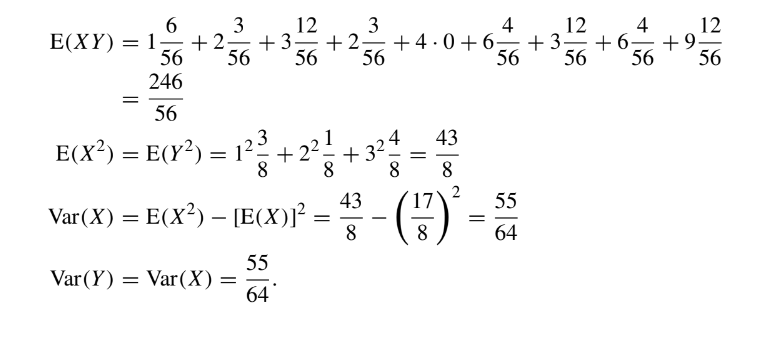


(b) Xác suất cận biên P (X = xi) có thể thu được từ thông tin đã cho

sự. Ví dụ, 3 trong số 8 quả bóng là màu đen và do đó P (X = 1) = 3/8. Các

phân phối có điều kiện P (Y | X = xi) có thể được tính toán dễ dàng bằng cách nhận ra rằng dưới sự phụ thuộc giả định của X và Y, lần rút thứ hai luôn dựa trên 7 quả bóng (8 quả bóng trừ đi một quả được rút ra với điều kiện X = xi) - ví dụ. nếu quả bóng đầu tiên được rút ra là màu đen, sau đó 7 quả bóng, trong đó có 2 quả màu đen, vẫn ở trong bình và P (Y = 1 | X = 1) = 2/7. Do đó chúng tôi tính toán





### Chương 8

# Chap 8

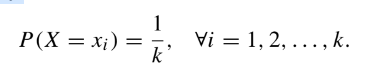
## 8.1 Standard Discrete Distributions

Đầu tiên, chúng ta thảo luận về một số phân phối chuẩn cho các biến ngẫu nhiên rời rạc.

Definition 8.1.1 Một biến ngẫu nhiên rời rạc X với k kết quả có thể xảy ra x1, x2, ...,

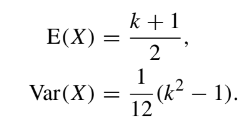
xk được cho là tuân theo một phân phối đồng đều rời rạc nếu hàm khối lượng xác suất

(PMF) của X được cho bởi



Nếu kết quả là các số tự nhiên xi = i (i = 1, 2, ..., k), giá trị trung bình và phương sai

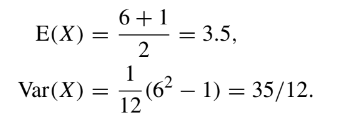
ance của X thu được là



Example 8.1.1 Nếu chúng ta tung một con xúc xắc công bằng, các kết quả “1”, “2”, ..., “6” có xác suất như nhau- khả năng xảy ra, và do đó, biến ngẫu nhiên X "số chấm được quan sát trên bề mặt trên của khuôn ”có sự phân bố rời rạc đồng đều với PMF

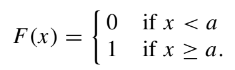


Giá trị trung bình và phương sai của X là



### 8.1.2 Degenerate Distribution

Definition 8.1.2 Một biến ngẫu nhiên X có phân phối suy biến tại a, nếu a laf kết quả duy nhất có thể xảy ra với P (X = a) = 1. CDF trong trường hợp này được đưa ra bởi

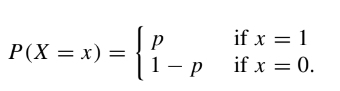


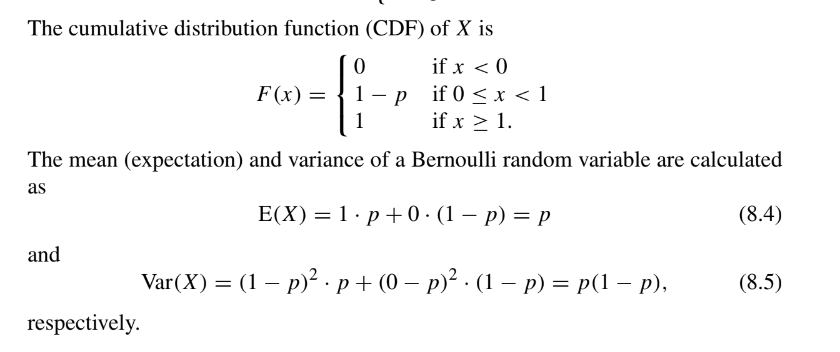
Hơn nữa, E (X) = a và Var (X) = 0.

Phân phối suy biến chỉ ra rằng chỉ có một kết quả cố định có thể xảy ra, và do đó, không có sự ngẫu nhiên nào có liên quan. Theo đó, chúng ta cần ít nhất hai các kết quả có thể xảy ra để có tính ngẫu nhiên trong các quan sát của một biến ngẫu nhiên hoặc thí nghiệm ngẫu nhiên. Phân phối Bernoulli là một phân phối như vậy khi có chỉ có hai kết quả, ví dụ: thành công và thất bại hoặc nam và nữ. Những kết quả này là thường được biểu thị bằng các giá trị “0” và “1”.

## 8.1.3 Bernoulli Distribution

Definition 8.1.3 Một biến ngẫu nhiên X có phân phối Bernoulli nếu PMF của X được đưa ra như



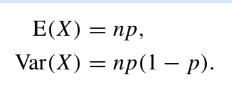


Phân phối Bernoulli hữu ích khi chỉ có hai kết quả có thể xảy ra vàsự quan tâm của chúng tôi nằm ở bất kỳ kết quả nào trong hai kết quả, ví dụ: liệu một khách hàng có mua một thứ nhất định sản phẩm hay không, hoặc liệu một cơn bão có tấn công một hòn đảo hay không. Kết quả của một sự kiện A thường được mã hóa là 1 xảy ra với xác suất p. Nếu sự kiện quan tâm không không xảy ra, tức là sự kiện bổ sung A ̄ xảy ra, kết quả được mã hóa là 0 xảy ra với xác suất 1 - p. Vậy p là xác suất để biến cố A quan tâm xảy ra.

### 8.1.4 Binomial Distribution

### Definition 8.1.4

Một biến ngẫu nhiên rời rạc X được cho là tuân theo phân phối nhị thức tion với các tham số n và p nếu PMF của nó được cho bởi (8.6). Ta cũng viết X ∼ B (n; p).Giá trị trung bình và phương sai của một biến ngẫu nhiên nhị thức X được cho bởi



Nhận xét 8.1.1 Một biến ngẫu nhiên Bernoulli do đó có phân phối B (1; p)

### Theorem 8.1.1

Cho X ∼ B (n; p) và Y ∼ B (m; p) và giả sử rằng X và Y là (ngẫu nhiên) độc lập. sau đó



Điều này rõ ràng bằng trực giác vì chúng ta có thể giải thích định lý này như mô tả người nghiện-

kết hợp tive của hai thử nghiệm nhị thức độc lập với n và m thử nghiệm, với

xác suất p tương ứng bằng nhau. Vì mọi thử nghiệm nhị thức là một chuỗi bất định

thí nghiệm Bernoulli không hoạt động, điều này tương đương với một chuỗi n + m độc lập

Các thử nghiệm Bernoulli với xác suất thành công không đổi p, lần lượt tương đương với

phân phối nhị thức với n + m nghiệm.

## 8.1.5 Poisson Distribution

Hãy xem xét một tình huống trong đó số lượng các sự kiện là rất lớn và xác suất thành công là rất nhỏ: ví dụ: số lượng hạt alpha được phát ra bởi một chất phóng xạ đi vào một vùng cụ thể trong một khoảng thời gian ngắn nhất định. Ghi chú rằng số lượng các hạt alpha được phát ra là rất cao nhưng chỉ có một số hạt là truyền qua vùng trong một khoảng thời gian ngắn nhất định. Một số ví dụ khác trong đó các phân phối Poisson hữu ích là số ca bệnh cúm ở một quốc gia trong một năm, số lượng cơn bão nhiệt đới trong một khu vực nhất định trong một năm, hoặc số vi khuẩn được tìm thấy trong một cuộc điều tra sinh học.

### Definition 8.1.5

A discrete random variable X is said to follow a Poisson distribution with parameter λ > 0 if its PMF is given by



Chúng ta cũng viết X ∼ Po (λ). Giá trị trung bình và phương sai của biến ngẫu nhiên Poisson là

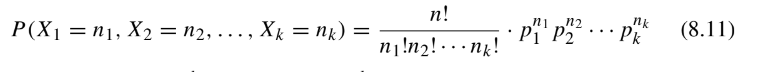
giống hệt nhau:

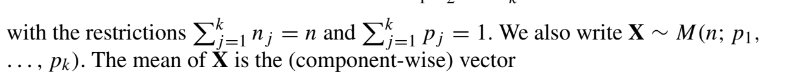


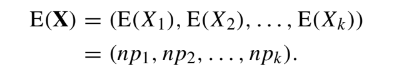
## 8.1.6 Multinomial Distribution

### Definition 8.1.6

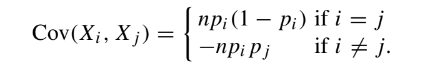
The random vector X = (X1, X2,..., Xk ) is said to follow a multinomial distribution if its PMF is given as







Phần tử thứ (i, j) của ma trận hiệp phương sai V (X) là



### Remark 8.1.2

ta sử dụng phân phối đa thức để mô tả tính ngẫu nhiên của phân loại biến. Giả sử chúng ta quan tâm đến biến "đảng chính trị"; có thể là tám đảng phái chính trị và do đó chúng tôi có thể tóm tắt biến này bằng tám các biến, mỗi biến mô tả sự kiện bên Aj, j = 1, 2, ..., 8, được bỏ phiếu vì. Theo nghĩa này, X = (X1, X2, ..., X8) tuân theo một phân phối đa thức.

### 8.1.7 Geometric Distribution

Hãy xem xét một tình huống mà chúng ta quan tâm đến việc xác định có bao nhiêu Cần thử nghiệm Bernoulli cho đến khi sự kiện quan tâm xảy ra lần đầu tiên. Vì ví dụ, chúng tôi có thể quan tâm đến việc mua bao nhiêu vé trong một cuộc xổ số cho đến khi chúng tôi giành chiến thắng lần đầu tiên hoặc bao nhiêu loại thuốc khác nhau để cố gắng giải quyết thành công tình trạng nghiêm trọng đau nửa đầu, v.v. Phân bố hình học có thể được sử dụng để xác định xác suất rằng sự kiện quan tâm xảy ra ở lần thử thứ k lần đầu tiên.

### Definition 8.1.7

Một biến ngẫu nhiên rời rạc X được cho là tuân theo một phân biệt hình học bution với tham số p nếu PMF của nó được cho bởi

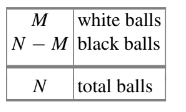


Giá trị trung bình (kỳ vọng) và phương sai được cho bởi E (X) = 1 / p và Var (X) =1 / p (1 / p - 1), tương ứng.

### 8.1.8 Hypergeometric Distribution

Một lần nữa, chúng ta có thể sử dụng mô hình urn để thúc đẩy một phân phối khác, siêu hình học

phân bổ. Hãy xem xét một chiếc bình có Chúng tôi rút ngẫu nhiên n quả bóng mà không cần thay thế,

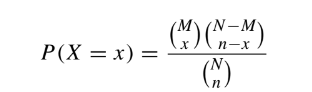


tức là chúng tôi không đặt một quả bóng trở lại bình sau khi nó được rút ra. Thứ tự mà

bóng được rút ra được coi là không quan tâm; chỉ số lượng quả bóng trắng được rút ra

có liên quan. Chúng tôi xác định biến ngẫu nhiên sau

X: “số quả bóng trắng (x) trong số n quả bóng được rút ra”.



for x ∈ {max(0, n − (N − M)), . . . , min(n, M)}.

### Definition 8.1.8

A random variable X is said to follow a hypergeometric distribu- tion with parameters n, M, N, i.e. X ∼ H(n, M, N), if its PMF is given by (8.14). A random variable X is said to follow a hypergeometric distribu- tion with parameters n, M, N, i.e. X ∼ H(n, M, N), if its PMF is given by (8.14).

## 8.2 Standard Continuous Distributions

Bây giờ, chúng ta thảo luận về một số phân phối xác suất tiêu chuẩn của (tuyệt đối) liên tục

biến ngẫu nhiên. Đặc điểm của biến ngẫu nhiên liên tục là số các kết quả có thể xảy ra là vô hạn và chúng có phân phối liên tục- hàm tion F (x). Theo đó xác suất điểm bằng 0, tức là P (X = x) = 0.

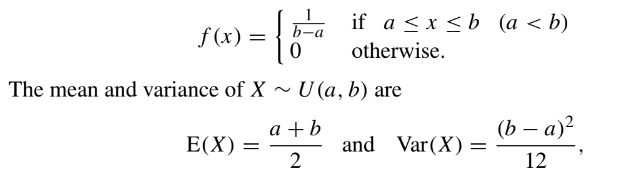
Hơn nữa, chúng tôi giả sử tồn tại một hàm mật độ f duy nhất, sao cho 

### 8.2.1 Continuous Uniform Distribution

Một chất tương tự liên tục với phân bố đồng đều rời rạc là đồng nhất liên tục phân phối trên một khoảng đóng trong R.

### Definition 8.2.1

Một biến ngẫu nhiên liên tục X được cho là tuân theo a (liên tục) phân phối đều trong khoảng [a, b], tức là X ∼ U (a, b), nếu mật độ xác suất của nó hàm (PDF) được cung cấp bởi

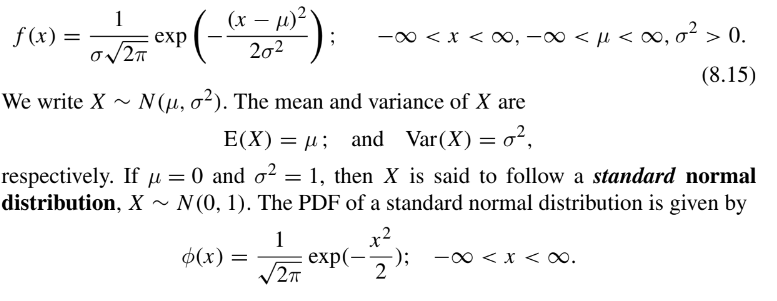


## 8.2.2 Normal Distribution

Phân phối chuẩn là một trong những phân phối quan trọng nhất được sử dụng trong thống kê. Tên được đặt bởi Carl Friedrich Gauss (1777–1855), một nhà toán học người Đức, nhà thiên văn học, nhà trắc địa và nhà vật lý, người đã quan sát thấy các phép đo trong ngành trắc địa và thiên văn học lệch một cách ngẫu nhiên đối xứng với các giá trị thực của chúng. Thường, đơn giản phân phối do đó cũng thường được gọi là phân phối Gaussian.

## Definition 8.2.2

A random variable X is said to follow a normal distribution with parameters μ and σ2 if its PDF is given by



Remark 8.2.1 Không có công thức rõ ràng nào để giải tích phân trong phương trình. (8.16). Nó có

được giải bằng phương pháp số (hoặc tính toán). Đây là lý do tại sao CDF

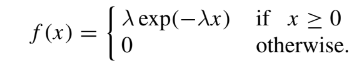
bảng được trình bày trong hầu hết các sách giáo khoa thống kê, xem Bảng C.1 trong Phụ lục C.

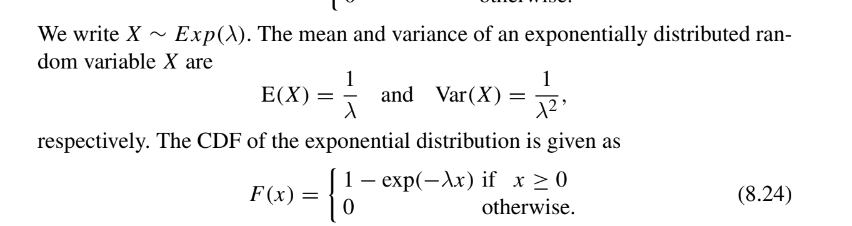
### 8.2.3 Exponential Distribution

Phân phối theo cấp số nhân hữu ích trong nhiều trường hợp, chẳng hạn như khi một quan tâm đến thời gian chờ đợi, hoặc thời gian tồn tại, cho đến khi một sự kiện quan tâm xảy ra. Nếu chúng ta giả định rằng thời gian tồn tại trong tương lai độc lập với thời gian tồn tại đã địa điểm (tức là không có quá trình “lão hóa” nào đang hoạt động), thời gian chờ đợi có thể được coi là được phân phối theo cấp số nhân.

### Definition 8.2.3

Một biến ngẫu nhiên X được cho là tuân theo phân phối hàm mũ với tham số λ> 0 nếu PDF của nó được cung cấp bởi

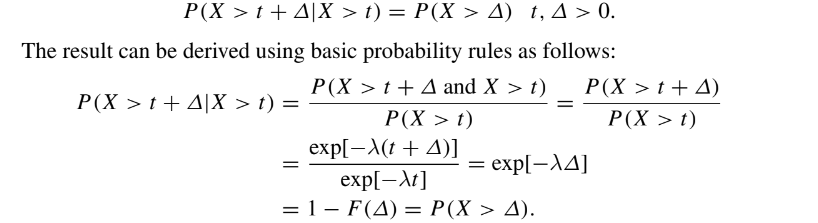




Lưu ý rằng P (X> x) = 1 - F (x) = exp (−λx) (x ≥ 0). Một tài sản thú vị của

phân phối hàm mũ là sự không nhớ của nó: nếu đã đạt đến thời gian t,

xác suất đến thời điểm lớn hơn t + Δ không phụ thuộc vào t. Cái này có thể được viết là



## Theorem 8.2.1

Số sự kiện Y xảy ra trong một khoảng thời gian liên tục là Poisson được phân phối với tham số λ nếu và chỉ khi thời gian giữa hai sự kiện là phân phối theo hàm mũ với tham số λ. Thời gian liên tục phụ thuộc vào vấn đề hiện tại. Nó có thể là một giây, một phút, 3 tháng, một năm hoặc bất kỳ khoảng thời gian nào khác

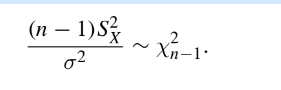
### 8.3 Sampling Distributions

Tất cả các phân phối được giới thiệu trong chương này cho đến nay đều được thúc đẩy bởi các ứng dụng. Tuy nhiên, có những phân bố lý thuyết đóng một vai trò quan trọng vai trò trong việc xây dựng và phát triển các công cụ thống kê khác nhau như được giới thiệu trong Chaps. 9–11.Chúng tôi gọi các phân phối này là “phân phối lấy mẫu”. Hiện nay, chúng ta thảo luận về các phân phối χ2-, t- và F-.

### 8.3.1 χ2-Distribution

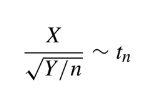
### Theorem 8.3.1

Xét hai biến ngẫu nhiên độc lập là χ2 m- và χ2 N- được phân phối tương ứng. Tổng của hai biến ngẫu nhiên này là χ2n + m-phân phối. Một ví dụ quan trọng về biến ngẫu nhiên có phân phối χ2 là phương sai mẫu (S2 X) của một i.i.d. mẫu có kích thước n từ một quần thể phân bố chuẩn, tức là



## 8.3.2 t-Distribution

### Definition 8.3.2

Gọi X và Y là hai biến ngẫu nhiên độc lập trong đó X ∼ N (0, 1) và Y ∼ χ2 N. Tỉ lệ 

## 8.3.3 F-Distribution

Định nghĩa 8.3.3 Cho X và Y độc lập χ2

m và χ2biến ngẫu nhiên phân phối n- ables, sau đó là phân phối của tỷ lệ



tuân theo phân phối Fisher F với (m, n) bậc tự do. PDF của Phân phối F được cho trong Công thức. (C.9) trong Phụ lục C.3.

8.Bài tập

Bài tập 8.1 Một công ty sản xuất ngũ cốc cung cấp một món đồ chơi trong mỗi gói ngũ cốc thứ sáu nhân kỷ niệm 50 năm thành lập của họ. Một ông bố mua ngay 20 gói.

(a) Xác suất tìm được 4 đồ chơi trong 20 gói là bao nhiêu?

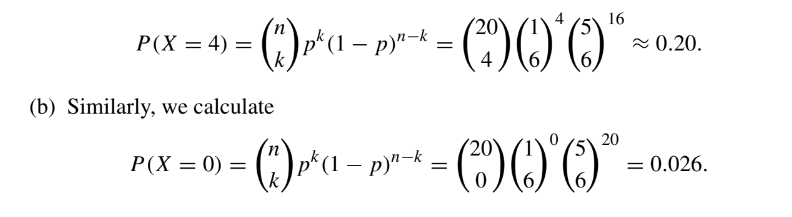
(b) Xác suất không tìm thấy đồ chơi nào là bao nhiêu?

(c) Các gói chứa ba đồ chơi. Xác suất để trong số 5 gói- độ tuổi cho con gái út của gia đình, cô ấy tìm thấy hai món đồ chơi?

bài tập 8.1 Biến ngẫu nhiên X: "số gói đồ chơi" là

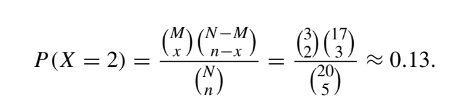
phân phối nhị thức. Trong mỗi n = 20 “lần thử nghiệm”, một món đồ chơi có thể được tìm thấy với xác suất p = 1/6.

1. Do đó, chúng tôi nhận được



(c) Câu hỏi này liên quan đến phân phối siêu đại: có N = 20 gói- lứa tuổi có M = 3 gói có đồ chơi và N - M = 17 gói không có đồ chơi.

Con gái nhận được n = 5 gói và chúng tôi quan tâm đến P (X = 2). Kể từ đây, chúng tôi nhận được



Bài tập 8.2 Một nghiên cứu về chim sinh sản thu thập thông tin chẳng hạn như chiều dài của

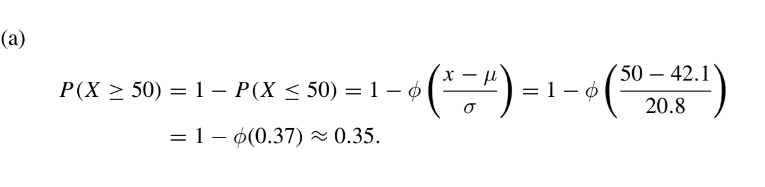
trứng của chúng (tính bằng mm). Giả sử rằng chiều dài được phân phối chuẩn với μ = 42,1 mm

và σ2 = 20,82. Xác suất của

(a) Tìm một quả trứng có chiều dài lớn hơn 50 mm?

(b) Tìm một quả trứng có chiều dài từ 30 đến 40 mm?

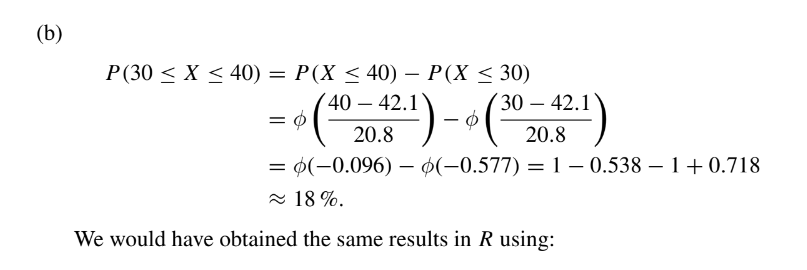
Tính toán kết quả cả bằng tay và bằng R.



Chúng tôi thu được kết quả tương tự trong R như sau:

code

1-pnorm(50,42.1,20.8)

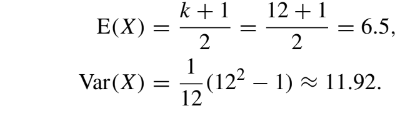


Code

pnorm(40,42.1,20.8)-pnorm(30,42.1,20.8)

Bài tập 8.3 Một khối tứ diện là một khối có 12 cạnh. Giả sử các số trên súc sắc là 1–12. Hãy xem xét biến ngẫu nhiên X mô tả số nào được hiển thị sau khi lăn cái chết một lần. Phân phối của X là gì? Xác định E (X) và Var (X).

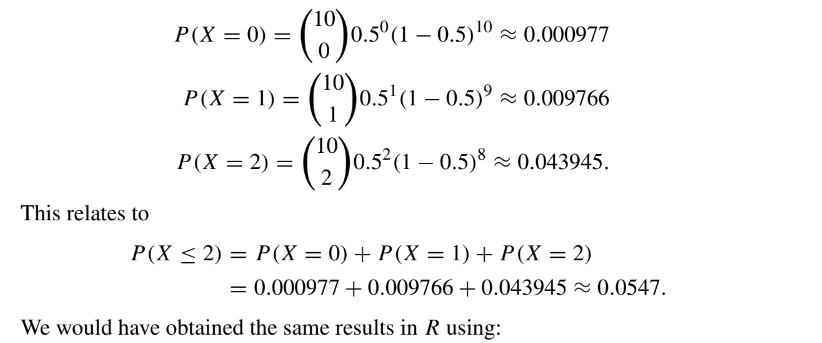
Lời giải bài tập 8.3 Biến ngẫu nhiên X tuân theo một phân phối đồng đều rời rạc- thận trọng vì pi = 1/12 cho mỗi xi. Do đó, kỳ vọng và phương sai là:



Bài tập 8.4 Felix nói rằng anh ta có thể phân biệt được hỗn hợp cà phê mới xa từ một siêu thị cà phê bình thường. Một người bạn của anh ấy yêu cầu anh ấy nếm thử 10 cốc về cà phê và cho anh ta biết anh ta đã thưởng thức loại cà phê nào. Giả sử rằng Felix thực sự không có manh mối về cà phê và chỉ đơn giản là đoán nhãn hiệu. Xác suất của ít nhất 8 là bao nhiêu đoán chính xác?

Lời giải cho bài tập 8.4 Mỗi lần đoán là một thử nghiệm Bernoulli trong đó quyền câu trả lời được đưa ra với xác suất là 50%. Do đó, số lần đoán đúng tuân theo phân phối nhị thức, tức là X ∼ B (10; 0,5). Xác suất đưa ra trả lời đúng ít nhất 8 lần trùng với xác suất không sai nhiều hơn

hơn 2 lần. Do đó, chúng ta có thể tính P (X ≥ 8) là P (X ≤ 2):



Code:

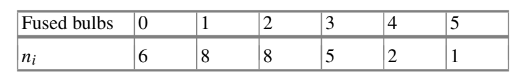
pbinom(2,10,0.5)

1-pbinom(7,10,0.5)

Bài tập 8.5 Một bảng quảng cáo được chiếu sáng bởi vài trăm bóng đèn. Một số của các bóng đèn được nung chảy hoặc đập vỡ thường xuyên. Nếu có nhiều hơn 5 bóng đèn hợp nhất trên

một ngày, chủ sở hữu của hội đồng quản trị thay thế họ, nếu không thì không. Hãy xem những điều sau dữ liệu được thu thập trong một tháng ghi lại số ngày (ni) mà tôi bóng đèn

đã bị phá vỡ



(a) Đề xuất một phân phối thích hợp cho X: “số bóng đèn bị hỏng mỗi ngày”.Số bóng đèn bị hỏng trung bình trong một ngày là bao nhiêu? Phương sai là gì?

(c) Xác định xác suất P (X = x) bằng cách sử dụng phân phối bạn đã chọn trong và sử dụng số bóng đèn bị hỏng trung bình mà bạn tính được trong . So sánh xác suất với tỷ lệ thu được từ dữ liệu.

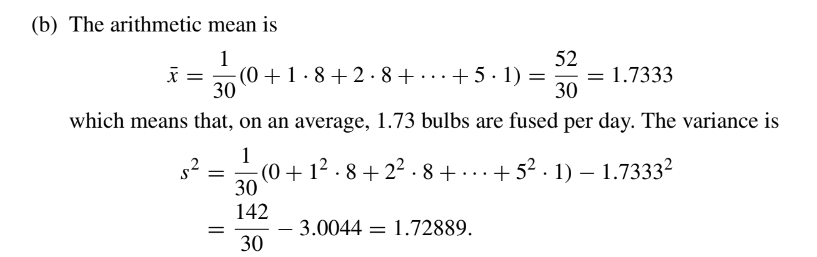
(d) Tính xác suất để có ít nhất 6 bóng đèn được hợp nhất, nghĩa là chúng cần được thay thế.

(e) Xem xét biến ngẫu nhiên Y: “thời gian cho đến khi bóng đèn tiếp theo bị hỏng”. Cái gì là phân phối của Y?

(f) Tính và giải thích E (Y).

### Solution to Exercise 8.5

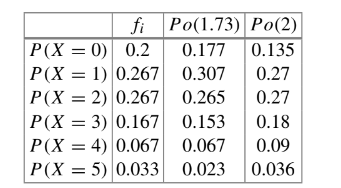
1. Có vẻ thích hợp để mô hình số lượng bóng đèn hợp nhất với một đĩa Poisson cống hiến. Tuy nhiên, chúng tôi giả định rằng xác suất của bóng đèn hợp nhất trên hai các ngày liên tiếp độc lập với nhau; tức là chúng chỉ phụ thuộc vào λ nhưng không vào thời gian t.



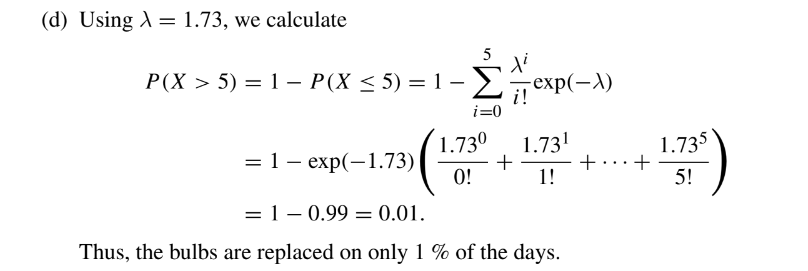
Chúng tôi thấy rằng giá trị trung bình và phương sai là tương tự nhau, đó là một dấu hiệu cho thấy

lựa chọn phân phối Poisson là phù hợp vì chúng ta giả sử E (X) = λ và Var (X) = λ.

1. Bảng sau liệt kê các tỷ lệ (tức là tần số tương đối f j) với nhau với xác suất P (X = x) từ phân phối Po (1,73). Như một tài liệu tham khảo, chúng tôi cũng liệt kê các xác suất từ ​​phân phối Po (2) vì nó không thực tế có thể 1,73 bóng đèn ngừng hoạt động và do đó nó có thể là một lựa chọn để làm tròn nghĩa.

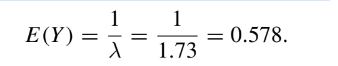


Người ta có thể thấy rằng tỷ lệ quan sát được và xác suất dự kiến ​​là gần nhau cùng với nhau, điều này chỉ ra một lần nữa rằng lựa chọn phân phối Poisson là phù hợp. Chương 9 cung cấp thêm chi tiết về cách ước tính các tham số, chẳng hạn như dưới dạng λ, từ dữ liệu nếu nó chưa biết.



(e) Nếu X tuân theo phân phối Poisson thì, cho Định lý 8.2.1, Y tuân theo một phân phối hàm mũ với λ = 1,73.

(f) Kỳ vọng của một biến phân phối theo cấp số nhân là



Điều này có nghĩa là, trung bình, phải mất hơn nửa ngày cho đến khi một trong các bóng đèn

được hợp nhất

Bài tập 8.6 Công ty của Marco tổ chức sổ xố vào một chức năng cuối năm. Ở đó 4000 vé xổ số sẽ được bán, trong đó có 500 vé trúng giải. Giá của mỗi vé là e1,50. Giá trị của các giải thưởng, phần lớn là các thiết bị điện được sản xuất của công ty, thay đổi giữa e80 và e250, với giá trị trung bình là e142.

(a) Marco muốn đảm bảo 99% sẽ nhận được ba giải thưởng. Bao nhiêu tiền mà anh ta cần phải tiêu? Sử dụng R để giải quyết câu hỏi.

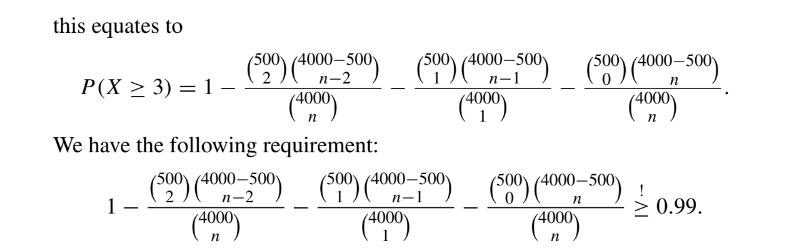
(b) Sử dụng R để vẽ biểu đồ của hàm mô tả mối quan hệ giữa số số vé đã mua và xác suất trúng ít nhất ba giải.

(c) Với giá trị của giải thưởng và chi phí của vé, nó có đáng để tham gia không trong xổ số?

### Solution to Exercise 8.6

1. Gọi X là biến ngẫu nhiên mô tả “số x vé trúng thưởng trong số n vé đã mua ”; thì X tuân theo phân phối siêu bội X ∼ H (n, 500, 4000). Chúng ta cần xác định n cho các điều kiện được chỉ định. Chúng tôi rất hứng thú với P (X ≥ 3) = 1 - P (X = 2) - P (X = 1) - P (X = 0). Sử dụng PMF của phân phối siêu đại áp





Để giải phương trình này, chúng ta có thể lập trình hàm này cho P (X> 3; n) trong R và

đánh giá nó cho các số lượng vé đã bán khác nhau, ví dụ: từ 50 đến 100 vé:

code :

raffle <- function(n){

p <- 1-((choose(500,2)\*choose(3500,n-2))/(choose(4000,n)))

-((choose(500,1)\*choose(3500,n-1))/(choose(4000,n)))

-((choose(500,0)\*choose(3500,n))/(choose(4000,n)))

return(p)

}

raffle(50:100)

raffle(63:64)

Kết quả cho thấy cần mua ít nhất 64 vé để có 99%

đảm bảo rằng ít nhất ba vé giành chiến thắng. Điều này tương đương với chi tiêu e 96.

Bài tập 8.7 Một quốc gia có tỉ số giữa số nam và số sinh nữ là 1,05, có nghĩa là 51,22% trẻ sinh ra là nam.

(a) Xác suất để người mẹ sinh con gái đầu lòng trong ba lần đầu là bao nhiêu? sinh?

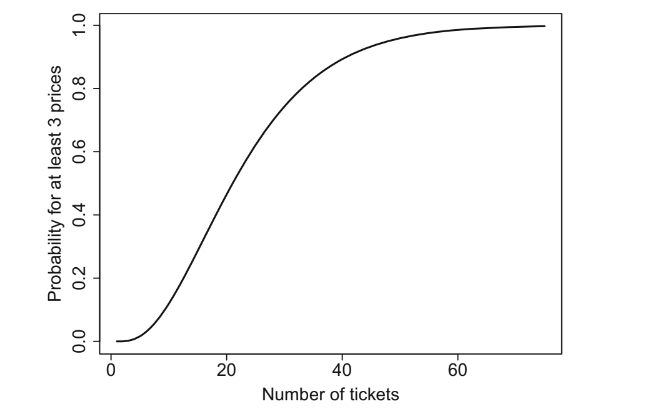
(b) Xác suất sinh được 2 bé gái trong số 4 em bé là bao nhiêu?

1. Chúng ta có thể vẽ biểu đồ của hàm như sau:

Code :

nb <- seq(1:75)

plot(nb,tombola(nb),type='l')



### Solution to Exercise 8.7

Xác suất sinh được một bé gái là p = 1 - 0,5122 =

0,4878.

(a) Chúng tôi đang giải quyết một phân bố hình học ở đây. Vì chúng tôi quan tâm đến

P (X ≤ 3), chúng ta có thể tính được:

P (X = 1) = 0,4878

P (X = 2) = 0,4878 (1 - 0,4878) = 0,2498512

P (X = 3) = 0,4878 (1 - 0,4878) 2 = 0,1279738

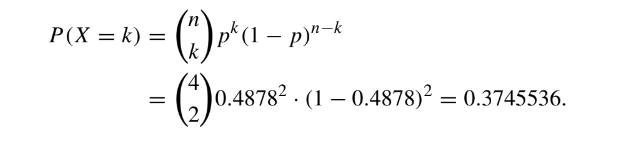
P (X ≤ 3) = P (X = 1) + P (X = 2) + P (X = 3) = 0,865625.

Chúng tôi sẽ thu được cùng một kết quả trong R bằng cách sử dụng:

Code :

pgeom(2,0.4878)

Ở đây chúng ta giải quyết một phân phối nhị thức với k = 2 và n = 4. Chúng ta có thể tính P (X = 2) như sau:



R sẽ cho chúng ta cùng một kết quả bằng cách sử dụng dbinom (2,4,0.4878) com-

sự ủy thác

Bài tập 8.8 Trung bình một giờ ngư dân đánh bắt được 3 con cá. Hãy để Y là một biến ngẫu nhiên biểu thị số lượng cá đánh bắt được trong một giờ và gọi X là khoảng thời gian giữa hai lần câu được con cá. Chúng tôi giả định rằng X theo cấp số nhân phân bổ.

(a) Phân phối của Y là gì?

(b) Xác định E (Y) và E (X).

(c) Tính P (Y = 5) và P (Y <1).

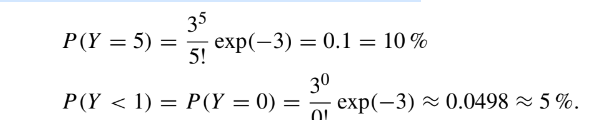
(a) Biến ngẫu nhiên Y tuân theo phân phối Poisson, xem Định lý 8.2.1 để biết

chi tiết hơn.

(b) Trung bình một giờ ngư dân đánh bắt được 3 con cá. Do đó, chúng tôi có thể giả định rằng

tỷ lệ λ là 3 và do đó E (Y) = λ = 3. Tương tự, E (X) = 1/ λ = 1 /3 có nghĩa là trung bình phải mất 20 phút để bắt một con cá khác.

1. Sử dụng bản PDF của phân phối Poisson, chúng tôi nhận được:

Chúng tôi s

thu được kết quả tương tự trong R bằng cách sử dụng dpois (5,3) và

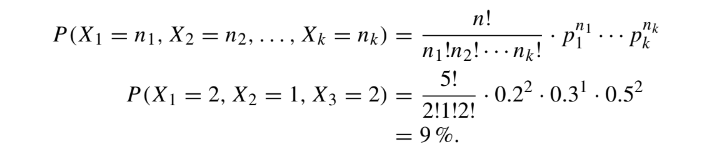
lệnh dpois (0,3).

Bài tập 8.9 Một nhà hàng bán ba loại tráng miệng khác nhau: sô cô la, bánh hạnh nhân, sữa chua với trái cây theo mùa và bánh tart chanh. Nhiều năm kinh nghiệm đã chỉ ra rằng xác suất mà món tráng miệng được chọn lần lượt là 0,2, 0,3 và 0,5.

(a) Xác suất để trong 5 khách, 2 khách chọn bánh hạnh nhân, 1 khách chọn sữa chua, và 2 khách còn lại chọn bánh tart chanh?

(b) Giả sử hai trong số năm vị khách được biết là luôn chọn bánh tart chanh. Gì xác suất của những người khác cũng chọn chanh chua?

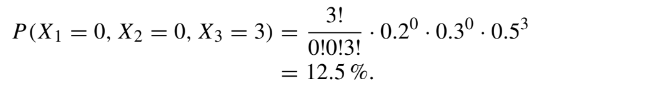
(c) Xác định kỳ vọng và phương sai giả sử một nhóm có 20 khách.



Code :

dmultinom(c(2,1,2),prob=c(0.2,0.3,0.5))

1. Xác suất chọn được bánh chanh cho hai vị khách đầu tiên là 1. Do đó chúng ta cần để xác định xác suất để 3 trong số 3 khách còn lại gọi chanh Chua cay:

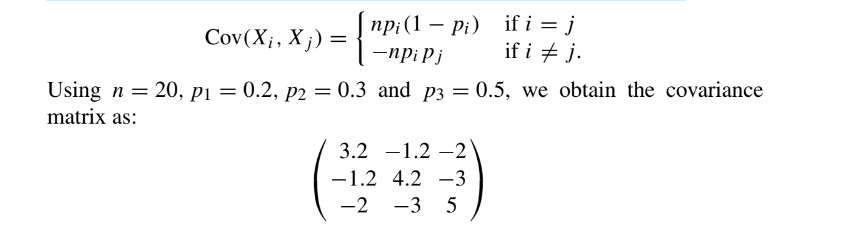


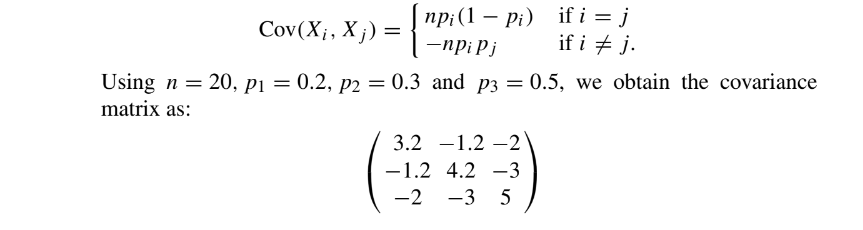
Sử dụng dmultinom (c (0,0,3), prob = c (0,2,0.3,0,5)) trong R, chúng tôi nhận được tương tự kết quả.

(c) The expectation vector is

E (X) = (np1, ..., npk) = (20 · 0,2, 20 · 3, 20 · 0,5) = (4, 6, 10).

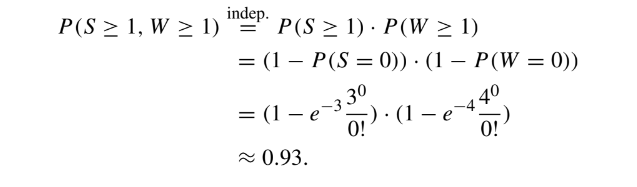
Điều này có nghĩa là chúng tôi dự kiến ​​sẽ có 4 khách gọi bánh hạnh nhân, 6 khách gọi sữa chua và 10 khách gọi món đặt bánh tart chanh. Ma trận hiệp phương sai có thể được xác định như sau:





Bài tập 8.10 Một công ty tái bảo hiểm hoạt động theo chính sách trả phí cho các khoản chênh lệch tự nhiên ters. Dựa trên kinh nghiệm, người ta biết rằng W = “số lượng thiên tai từ Tháng 10 đến tháng 3 ”(mùa đông) được phân bố Poisson với λW = 4. Tương tự, ran- biến dom S = “số lượng thiên tai từ tháng 4 đến tháng 9” (mùa hè) có phân phối Poisson với λS = 3. Xác định xác suất để có ít nhất 1

thiên tai trong cả mùa hè và mùa đông dựa trên giả định rằng hai các biến là độc lập.



Exercise 8.11 Read Appendix C.3 to learn about the Theorem of Large Numbers

and the Central Limit Theorem.

1. Vẽ 1000 nhận thức từ phân phối chuẩn chuẩn bằng cách sử dụng R và tính trung bình cộng. Lặp lại quá trình này 1000 lần. Đánh giá sự phân phối của trung bình số học bằng cách vẽ biểu đồ mật độ hạt nhân và bằng cách tính nghĩa và phương sai của nó.

### Solution to Exercise 8.11

Các số ngẫu nhiên của phân phối chuẩn có thể được tạo ra bằng cách sử dụng rnorm yêu cầu. Theo mặc định μ = 0 và σ = 1 (xem? Rnorm), vì vậy chúng ta không cần chỉ định các tham số này. Chúng ta chỉ cần đặt n = 1000. Giá trị trung bình của Do đó, có thể thu được 1000 hiện thực bằng cách sử dụng giá trị trung bình (rnorm (1000)). Chúng tôi có thể, cho ví dụ, viết một vòng lặp for để lặp lại quá trình này 1000 lần. Một trống (= NA) vectơ có kích thước 1000 có thể được sử dụng để lưu trữ và đánh giá kết quả:

(b) Lặp lại quy trình trong (a) với phân phối hàm mũ với λ = 1. Giải thích

phát hiện của bạn theo Định lý Giới hạn Trung tâm.

(c) Lặp lại quy trình trong (b) sử dụng 10.000 thay vì 1000 hiện thực. Làm thế nào để

kết quả thay đổi và tại sao?

Code:

set.seed(24121980)

R <- 1000

means <- c(rep(NA,R))

for(i in 1:R){means[i] <- mean(rnorm(1000))}

mean(means)

[1] -0.0007616465

var(means)

[1] 0.0009671311

plot(density(means))

1. Chúng ta có thể sử dụng mã tương tự như trên, ngoại trừ chúng ta sử dụng hàm mũ thay vì phân phối chuẩn:

Code:

means2 <- c(rep(NA,R))

for(i in 1:R){means2[i] <- mean(rexp(1000))}

mean(means2)

var(means2)

plot(density(means))

1. Việc tăng số lần lặp lại làm cho phân phối gần giống với phân phối chuẩn, xem Hình B.18b. Điều này hình dung rằng khi n có xu hướng đến vô cùng X ̄ n tiến gần hơn đến N (μ, σ2 n) -phân phối.

### Cheap 9

## Cheap 9 Properties of Point Estimators

### 9.2.1 Unbiasedness and Efficiency

### Definition 9.2.1

Công cụ ước lượng T (X) được gọi là công cụ ước lượng không chệch của θ nếu

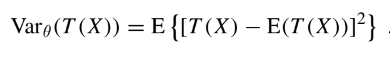


Chỉ số θ biểu thị rằng kỳ vọng được tính toán liên quan đến phân phối có tham số là θ.

Độ chệch của công cụ ước lượng T (X) được định nghĩa là



### Definition 9.2.2

Phương sai của T (X) được định nghĩa là 

9. Bài tập

Bài tập 9.1 Hãy xem xét một i.i.d. mẫu có kích thước n từ một ngẫu nhiên có phân phối Po (λ) biến X.

(a) Xác định ước lượng khả năng xảy ra tối đa cho λ.

(b) Hàm log-khả năng trông như thế nào đối với các nhận thức sau: x1 = 4, x2 = 3, x3 = 8, x4 = 6, x5 = 6? Vẽ đồ thị hàm bằng R.

Gợi ý:Lệnh curve có thể được sử dụng để vẽ các hàm.

(c) Sử dụng Định lý nhân tố hóa Neyman-Fisher để lập luận rằng giá trị tối đa ước lượng lihood thu được trong (a) là một thống kê đủ cho λ.

Solution to Exercise 9.1

(b) Sử dụng kết quả từ (a), chúng ta có thể viết hàm log-khả năng xảy ra cho x1 =

4, x2 = 3, x3 = 8, x4 = 6, x5 = 6 là:

ln L = 27 ln λ - ln (4! 3! 8! 6! 6!) - 5λ. tại vì xi = 27. Chúng ta có thể viết hàm này trong R như sau:

code :

MLP <- function(lambda){

27∗log(lambda) - log(factorial(4)∗...∗factorial(6)) -

5∗lambda

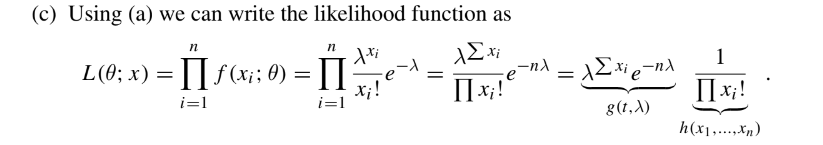
}

Hàm có thể được vẽ bằng cách sử dụng lệnh curve:

curve(MLP, from=0, to=10)

Hình B.19 cho thấy hàm khả năng ghi nhật ký. Có thể thấy rằng hàm đạt cực đại tại x ̄ = 5,4.

phân biệt chức năng, ước lượng khả năng xảy ra tối đa của θ là bao nhiêu.

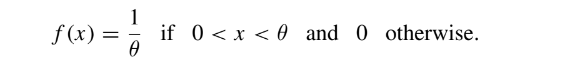


Điều này có nghĩa là T = Ni = 1 xi là đủ cho λ. Trung bình cộng, là

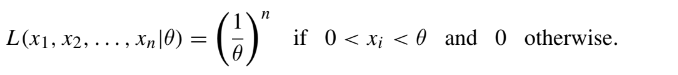
ước tính khả năng xảy ra tối đa, là một hàm một đối một của T và do đó quá đủ

Bài tập 9.3 Cho X1, X2, ..., Xn là n i.i.d. các biến ngẫu nhiên tuân theo một phân phối dạng, U (0, θ). Viết ra hàm khả năng và tranh luận, không có

The probability density function of U(0, θ) is



Lưu ý rằng điều này tương đương với PDF từ Định nghĩa 8.2.1 cho a = 0 và b = θ. Các do đó, hàm khả năng là:



Người ta có thể thấy rằng L (x1, x2, ..., xn | θ) tăng khi θ giảm. Tối đa của do đó, hàm khả năng đạt được với giá trị θ nhỏ nhất. Đặc biệt, θ là cực tiểu khi θ ≥ max (x1, x2, ..., xn) = x (n). Điều này theo sau từ định nghĩa của PDF yêu cầu rằng 0 <xi <θ và do đó θ> xi. Như vậy, tối đa ước lượng khả năng xảy ra của θ là x (n), giá trị quan sát được lớn nhất trong mẫu.

Bài tập 9.4 Cho X1, X2, ..., Xn là n i.i.d. các biến ngẫu nhiên tuân theo một triển lãm- phân phối theo cấp số nhân. Một nhà thống kê thông minh đề xuất sử dụng hai công cụ ước lượng để ước tính μ = 1 / λ:

(i) Tn (X) = nXmin với Xmin = min (X1, ..., Xn) và Xmin ∼ Exp (nλ),

(ii) Vn (X) = n − 1 n i = 1 Xi.

(a) Cả Tn (X) và Vn (X) (tiệm cận) đều không thiên vị đối với μ?

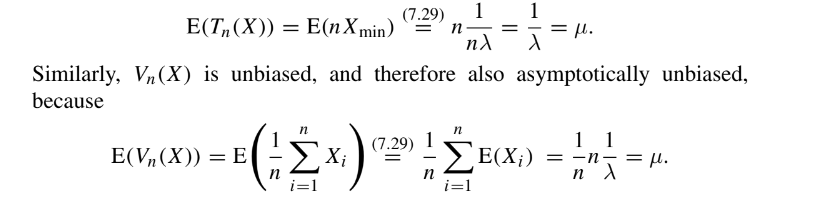
(b) Tính sai số bình phương trung bình của cả hai bộ ước lượng. Công cụ ước tính nào nhiều hơn

Có hiệu quả?

Vn (X) MSE có nhất quán, nhất quán yếu, cả hai hay hoàn toàn không nhất quán?

### Solution to Exercise 9.4

1. Tn (X) là không thiên vị, và do đó cũng không thiên vị về mặt tiệm cận, bởi vì



(b) Để tính toán MSE, chúng ta cần xác định độ chệch và phương sai của ước tính như chúng ta biết từ Eq. (9,5). Theo sau từ (a) cả hai công cụ ước tính không thiên vị và do đó độ chệch là 0. Đối với các phương sai, chúng ta nhận được:

Chương 10 Kiểm định giả thuyết

# 10.1 Giới thiệu

Chúng tôi đã giới thiệu ước lượng điểm và khoảng của các tham số trong chương trước.

Đôi khi, câu hỏi nghiên cứu ít tham vọng hơn theo nghĩa là chúng tôi không quan tâm đến các ước tính chính xác của một tham số, nhưng chúng tôi chỉ muốn kiểm tra xem liệu

một tuyên bố về một tham số quan tâm hoặc giả thuyết nghiên cứu có đúng hay không

(mặc dù chúng ta sẽ thấy ở phần sau của chương này rằng có một mối liên hệ giữa khoảng tin cậy và kiểm định thống kê, được gọi là đối ngẫu). Một vấn đề liên quan khác là một khi

một nhà phân tích ước tính các thông số trên cơ sở một mẫu ngẫu nhiên, (các) anh ta sẽ

muốn suy luận điều gì đó về giá trị của tham số trong tổng thể. Kiểm định giả thuyết thống kê tạo điều kiện thuận lợi cho việc so sánh các giá trị ước tính với giả thuyết

các giá tr

Ví dụ 10.1.1 Như một ví dụ đơn giản, hãy xem xét trường hợp chúng ta muốn tìm hiểu

liệu tỷ lệ phiếu bầu cho một bên P trong một cuộc bầu cử sẽ vượt quá 30% hay

không phải. Thông thường, trước cuộc bầu cử, chúng tôi sẽ cố gắng lấy dữ liệu đại diện về

tỷ lệ bầu cử cho các đảng khác nhau (ví dụ: phỏng vấn qua điện thoại) và sau đó thực hiện

một tuyên bố như "có", chúng tôi hy vọng rằng P sẽ nhận được hơn 30% số phiếu bầu hoặc "không",

chúng tôi không có đủ bằng chứng rằng P sẽ nhận được hơn 30% số phiếu bầu. Trong

trong trường hợp như vậy, chúng tôi sẽ chỉ biết sau cuộc bầu cử liệu tuyên bố của chúng tôi có đúng hay không

Sai lầm. Lưu ý rằng thuật ngữ đại diện dữ liệu chỉ có nghĩa là mẫu tương tự

cho dân số liên quan đến phân phối của một số biến chính, ví dụ: tuổi tác,

giới tính và giáo dục. Vì chúng tôi sử dụng một mẫu để so sánh nó với một giá trị cố định

(30%), chúng tôi gọi đó là bài toán một mẫu.

# 10.2 Định nghĩa cơ bản

## 10.2.1 Vấn đề một và hai mẫu

Trong các bài toán một mẫu, dữ liệu thường được giả định là phát sinh dưới dạng một mẫu từ

dân số xác định. Trong các bài toán hai mẫu, dữ liệu bắt nguồn ở dạng hai

mẫu có thể từ hai quần thể khác nhau. Tính không đồng nhất thường được mô hình hóa

bằng cách giả định rằng hai quần thể chỉ khác nhau về một số thông số hoặc số lượng chính

## 10.2.2 Giả thuyết

Một nhà nghiên cứu có thể có một câu hỏi nghiên cứu mà sự thật về dân số quan tâm chưa được biết. Giả sử dữ liệu có thể được thu thập bằng cách sử dụng một cuộc khảo sát. Nếu, với mức độ không đảm bảo xác định trước, một thử nghiệm thống kê dựa trên

dữ liệu hỗ trợ giả thuyết về dân số, chúng tôi nói rằng giả thuyết này được chứng minh về mặt thống kê. Lưu ý rằng câu hỏi nghiên cứu phải được vận hành trước khi nó có thể được kiểm tra bằng một bài kiểm tra thống kê.

Một là kiểm tra xem mức giảm trung bình (từ mức cơ bản đến

3 tháng) huyết áp của thuốc B cao hơn (và dương tính) với thuốc A. Chúng tôi

sau đó nêu giả thuyết của chúng tôi dưới dạng giá trị kỳ vọng (tức là μ). Tại sao chúng ta phải sử dụng

các giá trị mong đợi μ và không chỉ đơn giản là so sánh các giá trị số học x¯? Lý do là

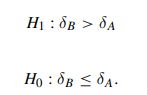
rằng tính ưu việt của B thể hiện trong mẫu sẽ chỉ có giá trị đối với mẫu này và

không nhất thiết cho một mẫu khác. Chúng ta cần thể hiện sự vượt trội của B trong toàn bộ

dân số, và do đó, giả thuyết của chúng tôi cần phản ánh điều này.

Điều ngược lại với giả thuyết nghiên cứu phải được xây dựng thành giả thuyết vô hiệu thống kê, ký hiệu là H0. Trong ví dụ về thuốc, giả thuyết thay thế và giả thuyết vô hiệu

tương ứng là:

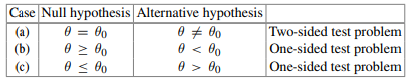


## 10.2.3 Kiểm tra một mặt và hai mặt

Chúng tôi phân biệt giữa các giả thuyết và thử nghiệm một phía và hai phía. Trước đây, chúng tôi đã đưa ra một ví dụ về kiểm tra một phía.

Đối với thông số tổng thể không xác định θ (ví dụ: μ) và giá trị cố định θ0 (ví dụ 5),

Ba trường hợp sau phải được phân biệt:



*Ví dụ 10.2.1* Các bài toán một mẫu thường kiểm tra xem có đạt được giá trị mục tiêu hay không

hay không. Ví dụ: hãy coi giả thuyết rỗng là

• H0: khối lượng nhân trung bình của gói bột = 1 kg

• H0: chiều cao cơ thể trung bình (nam) = 178 cm.

Giả thuyết thay thế H1 được xây dựng dưới dạng độ lệch so với giá trị mục tiêu. Nếu

độ lệch theo cả hai hướng là điều thú vị, khi đó H1 được xây dựng dưới dạng hai mặt

giả thuyết,

• H1: chiều cao cơ thể trung bình (nam) = 178 cm.

Nếu sự sai lệch theo một hướng cụ thể là chủ đề được quan tâm, thì H1 được xây dựng

như một giả thuyết một phía, chẳng hạn,

• H1: khối lượng nhân trung bình của gói bột nhỏ hơn 1 kg.

• H1: khối lượng nhân trung bình của gói bột lớn hơn 1 kg.

Bài toán hai mẫu thường kiểm tra sự khác biệt của hai mẫu. Giả sử

Giả thuyết vô hiệu H0 liên quan đến khối lượng trung bình của các gói bột được lấp đầy bởi hai

máy móc, giả sử 1 và 2. Sau đó, giả thuyết vô hiệu là

• H0: khối lượng trung bình của gói bột do máy 1 làm đầy = khối lượng trung bình của

gói bột bằng máy 2.

Sau đó, H1 có thể được xây dựng dưới dạng giả thuyết một phía hoặc hai phía. Nếu chúng ta muốn

chứng minh rằng máy 1 và máy 2 có khối lượng điền đầy khác nhau, khi đó H1 sẽ

được xây dựng như một giả thuyết hai mặt

• H1: khối lượng chiết rót trung bình của máy 1 = khối lượng chiết rót trung bình của máy 2.

Nếu chúng ta muốn chứng minh rằng máy 1 có khối lượng chiết rót trung bình thấp hơn máy 2,

H1 sẽ được xây dựng như một giả thuyết một chiều

• H1: khối lượng chiết rót trung bình của máy 1 <khối lượng chiết rót trung bình của máy 2.

Nếu chúng ta muốn chứng minh rằng máy 2 có khối lượng chiết rót thấp hơn máy 1, H1

sẽ được xây dựng như một giả thuyết một chiều

• H1: khối lượng chiết rót trung bình của máy 1> khối lượng chiết rót trung bình của máy 2.

Nhận xét 10.2.1 Lưu ý rằng chúng ta chưa xem xét tình huống sau: H0: θ =

θ0, H1: θ = θ0. Nói chung, với các thử nghiệm được mô tả trong chương này, chúng tôi không thể chứng minh

sự bình đẳng của một tham số với một giá trị được xác định trước và chúng tôi cũng không thể chứng minh

bằng nhau của hai tham số, như trong H0: θ1 = θ2, H1: θ1 = θ2. Ví dụ, chúng ta có thể

10.2 Các định nghĩa cơ bản 213

không chứng minh (về mặt thống kê) rằng máy 1 và 2 trong ví dụ trước cung cấp

trọng lượng lấp đầy. Điều này sẽ dẫn đến lớp kiểm tra tương đương phức tạp hơn,

là một chủ đề nằm ngoài phạm vi của cuốn sách này.

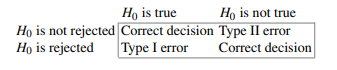
## 10.2.4 Lỗi loại I và loại II

Nếu chúng tôi thực hiện kiểm tra thống kê, hai loại lỗi có thể xảy ra.

• Giả thuyết H0 đúng nhưng bị bác bỏ; lỗi này được gọi là lỗi loại I.

• Giả thuyết H0 không bị bác bỏ mặc dù nó sai; đây được gọi là lỗi loại II.

Khi một giả thuyết được kiểm tra, thì bốn trường hợp sau đây có thể xảy ra:



Mức ý nghĩa là xác suất của lỗi loại I, P (H1 | H0) = α, là

xác suất bác bỏ H0 (chấp nhận H1) nếu H0 là đúng. Nếu chúng tôi xây dựng một bài kiểm tra,

mức ý nghĩa α được xác định trước, ví dụ: α = 0,05. Một bài kiểm tra ý nghĩa được xây dựng

sao cho xác suất của lỗi loại I không vượt quá α trong khi xác suất của một

lỗi loại II phụ thuộc vào các giá trị tham số đúng nhưng không xác định trong (các) tập hợp

và kích thước mẫu. Do đó, hai lỗi không được xử lý đối xứng trong một

kiểm tra ý nghĩa. Trên thực tế, lỗi loại II β, P (H0 | H1) = β không được kiểm soát bởi

việc xây dựng bài kiểm tra và có thể trở nên rất cao, đôi khi lên đến 1 - α.

Đây là lý do tại sao một thử nghiệm không bác bỏ H0 không phải là một bằng chứng (thống kê) của H0.

Trong thống kê toán học, một người tìm kiếm thử nghiệm tốt nhất duy trì α và

giảm thiểu β. Không thể giảm thiểu đồng thời cả α và β. Các

lý do là khi α tăng thì β giảm và ngược lại. Vì vậy, một trong những lỗi

cần phải được sửa và các lỗi khác được giảm thiểu. Do đó, lỗi mà

được coi là nghiêm trọng hơn được khắc phục và sau đó các lỗi khác được giảm thiểu. Các bài kiểm tra

được thảo luận trong các phần dưới đây được thu thập dựa trên giả định rằng loại I

lỗi nghiêm trọng hơn lỗi loại II. Vì vậy, thống kê thử nghiệm thu được bằng cách sửa

α và sau đó cực tiểu hóa β. Trên thực tế, giả thuyết vô hiệu được đóng khung theo cách

nó ngụ ý rằng lỗi loại I nghiêm trọng hơn lỗi loại II. Xác suất

1 - β = P (H1 | H1) được gọi là lũy thừa của phép thử. Nó là xác suất để tạo ra một

quyết định có lợi cho giả thuyết nghiên cứu H1, nếu nó đúng, tức là xác suất của

phát hiện một giả thuyết nghiên cứu đúng.

## 10.2.5 Cách tiến hành kiểm tra thống kê

Nói chung, chúng ta có thể làm theo các bước được mô tả bên dưới để kiểm tra giả thuyết về một

tham số dân số dựa trên một mẫu dữ liệu.

(1) Xác định các giả định phân phối cho các biến ngẫu nhiên quan tâm, và

chỉ định chúng dưới dạng các tham số tổng thể (ví dụ: θ hoặc μ và σ). Điều này là cần thiết cho các bài kiểm tra tham số. Có những loại kiểm tra khác, được gọi là không tham số

các bài kiểm tra, trong đó các giả định có thể được nới lỏng theo nghĩa là chúng ta không cần phải

chỉ định một phân phối cụ thể, xem Sect. 10,6ff. Hơn nữa, đối với một số thử nghiệm,

Các giả định về phân phối có thể được nới lỏng nếu cỡ mẫu lớn.

(2) Hình thành giả thuyết vô hiệu và giả thuyết thay thế như được mô tả trong

Các môn phái. 10.2.2 và 10.2.3.

(3) Sửa một giá trị ý nghĩa (thường được gọi là lỗi loại I) α, ví dụ α = 0,05, xem

cũng Sect. 10.2.4.

(4) Xây dựng thống kê kiểm định T (X) = T (X1, X2, ..., Xn). Phân phối của T có

được biết theo giả thuyết vô hiệu H0. Chúng tôi lưu ý một lần nữa rằng (X1, X2, ...,

Xn) đề cập đến các biến ngẫu nhiên trước khi vẽ mẫu thực và

x1, x2, ..., xn là các giá trị thực (quan sát được) trong mẫu.

(5) Xây dựng vùng tới hạn K cho thống kê T, tức là vùng mà — nếu T giảm

trong khu vực này — H0 bị từ chối, như vậy



Kí hiệu PH0 (·) có nghĩa là bất đẳng thức này phải giữ cho tất cả các giá trị tham số

θ thuộc giả thuyết rỗng H0. Vì chúng tôi giả định rằng chúng tôi biết

phân phối của T (X) dưới H0, vùng tới hạn được xác định bởi các giá trị của

T (X) khó có thể quan sát được (tức là với xác suất nhỏ hơn α)

theo giả thuyết vô hiệu. Lưu ý rằng mặc dù T (X) là một biến ngẫu nhiên, K là một

vùng xác định rõ, xem Hình 10.1 để làm ví dụ.

(6) Tính t (x) = T (x1, x2, ..., xn) dựa trên các giá trị mẫu nhận được X1 =

x1, X2 = x2, ..., Xn = xn.

(7) Quy tắc quyết định: nếu t (x) rơi vào vùng tới hạn K thì giả thuyết vô hiệu H0

bị từ chối. Giả thuyết thay thế sau đó được chứng minh về mặt thống kê. Nếu t (x) giảm

ngoài vùng tới hạn, H0 không bị từ chối.

t (x) ∈ K: H0 bị từ chối ⇒ H1 có ý nghĩa thống kê,

t (x) / ∈ K: H0 không bị bác bỏ và do đó được chấp nhận.

Hai đoạn tiếp theo chỉ ra cách đi đến các quyết định kiểm tra từ bước 7 trong

một cách khác. Độc giả quan tâm đến một ví dụ về kiểm tra thống kê có thể nhảy

đến Sect. 10.3.1 và có thể cả Ví dụ 10.3.1.

## 10.2.6 Quyết định kiểm tra bằng cách sử dụng giá trị p

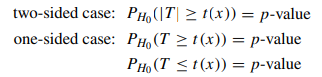
Phần mềm thống kê thường không hiển thị cho chúng ta tất cả các bước kiểm tra giả thuyết như

nêu trong Sect. 10.2.5. Thông thường, thay vì tính toán và báo cáo

giá trị tới hạn, thống kê thử nghiệm được in cùng với cái gọi là giá trị p. Nó là

có thể sử dụng giá trị p thay vì các vùng quan trọng để đưa ra quyết định kiểm tra. Các

Giá trị p của thống kê thử nghiệm T (X) được xác định như sau:



Nó có thể được hiểu là xác suất của các kết quả quan sát bằng hoặc cao hơn

so với những điều thực sự được quan sát nếu giả thuyết vô hiệu là đúng. Sau đó, quy tắc quyết định

Là

H0 bị bác bỏ nếu giá trị p nhỏ hơn mức ý nghĩa xác định trước α.

Nếu không, H0 không thể bị từ chối.

## 10.2.7 Kiểm tra quyết định sử dụng khoảng tin cậy

Có một mối quan hệ thú vị và hữu ích giữa khoảng tin cậy và

kiểm định giả thuyết. Nếu giả thuyết vô hiệu H0 bị bác bỏ ở mức ý nghĩa α, thì

tồn tại khoảng tin cậy 100 (1 - α)% đưa ra kết luận tương tự như

kiểm tra: nếu khoảng tin cậy thích hợp không chứa giá trị θ0 được nhắm mục tiêu

trong giả thuyết, thì H0 bị bác bỏ. Chúng tôi gọi đây là tính hai mặt. Ví dụ, nhớ lại

Ví dụ 10.1.2 trong đó chúng tôi quan tâm đến việc liệu sự thay đổi trung bình trong máu

áp lực đối với thuốc B cao hơn đối với thuốc A, tức là H1: δB> δA. Giả thuyết này là

tương đương với H1: δB - δA> δ0 = 0. Trong phần sau, chúng tôi phát triển các bài kiểm tra để

quyết định xem H1 có ý nghĩa thống kê hay không. Ngoài ra, chúng tôi có thể xây dựng

khoảng tin cậy 100 (1 - α)% cho sự khác biệt δB - δA và đánh giá xem liệu

khoảng có chứa δ0 = 0 hoặc không; nếu có, chúng tôi chấp nhận H0; nếu không, chúng tôi từ chối nó. Vì

một số bài kiểm tra được giới thiệu trong phần sau, chúng tôi đề cập đến khoảng tin cậy

dẫn đến kết quả tương tự như bài kiểm tra tương ứng.

# 10.3 Kiểm tra tham số cho các tham số vị trí

## 10.3.1 Kiểm định giá trị trung bình khi phương sai được biết (một mẫu Kiểm tra Gauss)

Có một mối quan hệ thú vị và hữu ích giữa khoảng tin cậy và

kiểm định giả thuyết. Nếu giả thuyết vô hiệu H0 bị bác bỏ ở mức ý nghĩa α, thì

tồn tại khoảng tin cậy 100 (1 - α)% đưa ra kết luận tương tự như

kiểm tra: nếu khoảng tin cậy thích hợp không chứa giá trị θ0 được nhắm mục tiêu

trong giả thuyết, thì H0 bị bác bỏ. Chúng tôi gọi đây là tính hai mặt. Ví dụ, nhớ lại

Ví dụ 10.1.2 trong đó chúng tôi quan tâm đến việc liệu sự thay đổi trung bình trong máu

áp lực đối với thuốc B cao hơn đối với thuốc A, tức là H1: δB> δA. Giả thuyết này là

tương đương với H1: δB - δA> δ0 = 0. Trong phần sau, chúng tôi phát triển các bài kiểm tra để

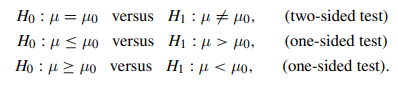
quyết định xem H1 có ý nghĩa thống kê hay không. Ngoài ra, chúng tôi có thể xây dựng

khoảng tin cậy 100 (1 - α)% cho sự khác biệt δB - δA và đánh giá xem liệu

khoảng có chứa δ0 = 0 hoặc không; nếu có, chúng tôi chấp nhận H0; nếu không, chúng tôi từ chối nó. Vì

một số bài kiểm tra được giới thiệu trong phần sau, chúng tôi đề cập đến khoảng tin cậy

dẫn đến kết quả tương tự như bài kiểm tra tương ứng.



3. Nêu xác suất của lỗi loại I là α: Thường chọn α = 0,05 = 5%.

4. Xây dựng thống kê thử nghiệm: Giá trị trung bình chưa biết, tức là kỳ vọng μ, thường là

ước tính bởi trung bình mẫu x¯. Chúng tôi đã biết rằng nếu X là i.i.d., thì

giá trị trung bình của mẫu được phân phối chuẩn. Theo giả định rằng H0 là đúng,



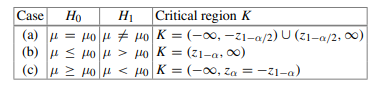
∼ có nghĩa là "phân phối dưới H0". Nếu chúng ta chuẩn hóa giá trị trung bình theo H0,

chúng tôi nhận được thống kê thử nghiệm được phân phối N (0, 1)



5. Vùng tới hạn: Vì thống kê thử nghiệm T (X) là N (0, 1) được phân phối, chúng tôi nhận được

sau các vùng tới hạn, tùy thuộc vào giả thuyết:



Đối với trường hợp (a) với H0: μ = μ0 và H1: μ = μ0, chúng ta quan tâm đến cực trị

giá trị của thống kê thử nghiệm trên cả hai phần: giá trị rất nhỏ và giá trị rất lớn của

thống kê kiểm định cung cấp cho chúng tôi bằng chứng rằng H0 là sai (vì thống kê chủ yếu là

bằng sự khác biệt của giá trị trung bình mẫu và giá trị thử nghiệm μ0 đối với phương sai cố định),

xem Hình 10.1. Trong thử nghiệm hai mặt như vậy, khi phân phối của thống kê thử nghiệm là

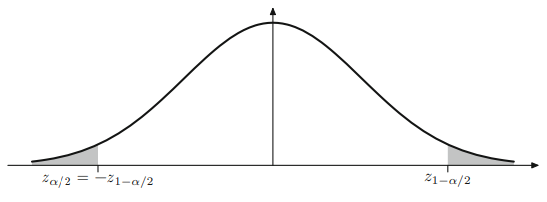
đối xứng, chúng tôi chia vùng tới hạn thành hai phần bằng nhau và gán cho mỗi vùng

có kích thước α / 2 ở phía bên trái và bên phải của phân phối. Đối với α = 0,05, 2,5% của

giá trị cực đoan nhất về phía cuối bên phải của phân phối và 2,5% giá trị cao nhất

các giá trị cực đoan về phía cuối bên trái của phân phối cung cấp cho chúng tôi đủ bằng chứng rằng

H0 sai và có thể bị bác bỏ và H1 được chấp nhận. Cũng rõ ràng tại sao α là



## 10.3.2 Kiểm tra giá trị trung bình khi phương sai không xác định (Thử nghiệm t một mẫu)

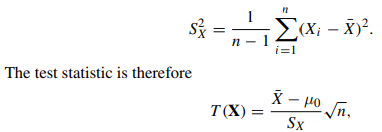
Nếu phương sai σ2 không xác định, giả thuyết về giá trị trung bình μ của một ngẫu nhiên bình thường

biến X ∼ N (μ, σ2) có thể được kiểm tra theo cách tương tự với Gauss một mẫu

kiểm tra. Sự khác biệt là phương sai chưa biết được ước tính từ mẫu. Một

220 10 Kiểm tra giả thuyết

ước lượng không chệch của σ2 là phương sai mẫu



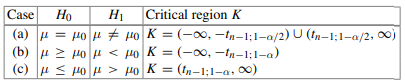
tuân theo phân phối t với n - 1 bậc tự do nếu H0 là đúng, vì chúng ta

biết từ Định lý 8.3.2.

Các khu vực quan trọng và các quyết định kiểm tra

Vì T (X) tuân theo phân phối t dưới H0, các vùng tới hạn đề cập đến các vùng

của phân phối t mà khó có thể quan sát được theo H0:



Giả thuyết H0 bị bác bỏ nếu thống kê kiểm định được thực hiện, tức là:



## 

## 10.3.3 So sánh phương tiện của hai mẫu độc lập

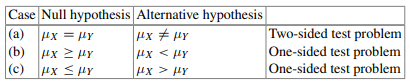
Trong bài toán hai mẫu, chúng ta có thể quan tâm đến việc so sánh giá trị trung bình của hai mẫu độc lập. Giả sử rằng chúng ta có hai mẫu của hai biến phân phối chuẩn X ∼ N (μX, σ2

X) và Y ∼ N (μY, σ2

Y) có kích thước n1 và n2, tức là X1, X2, ..., Xn1

là i.i.d. có cùng phân phối với X và Y1, Y2, ..., Yn2 là i.i.d. với cùng một

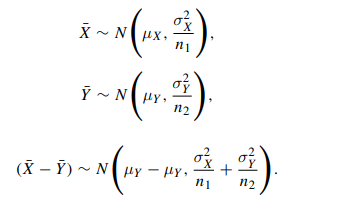
phân phối như Y. Chúng tôi có thể xác định các giả thuyết sau:



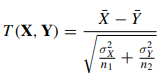
Trường hợp 1: Các phương sai đã biết (kiểm định Gauss hai mẫu).

Nếu giả thuyết rỗng H0: μX = μY là đúng, thì sử dụng các quy tắc thông thường cho thông thường

sự phân phối và tính độc lập của các mẫu,

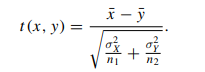


Theo sau đó là thống kê thử nghiệm



tuân theo phân phối chuẩn chuẩn, T (X, Y) ∼ N (0, 1). Thống kê thử nghiệm hiện thực

Là

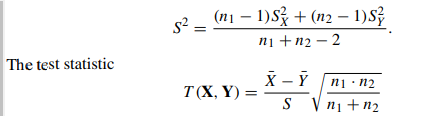


Trường hợp 2: Các phương sai không xác định, nhưng bằng nhau (kiểm định t hai mẫu).

Chúng tôi biểu thị phương sai chưa biết của cả hai phân phối là σ2 (tức là cả hai quần thể

được giả định là có phương saiσ2). Chúng tôi ước tínhσ2 bằng cách sử dụng phương sai mẫu tổng hợp

trong đó mỗi mẫu được ấn định trọng lượng tương ứng với cỡ mẫu:



với S như trong (10.3) tuân theo phân phối t với n1 + n2 - 2 bậc tự do nếu

H0 là đúng. Thống kê thử nghiệm được thực hiện là



Trường hợp 3: Các phương sai không xác định và không bằng nhau (Kiểm định Welch).

Chúng tôi kiểm tra H0: μX = μY so với H1: μX = μY cho trước σ2

X = σ2

Y và cả hai σ2

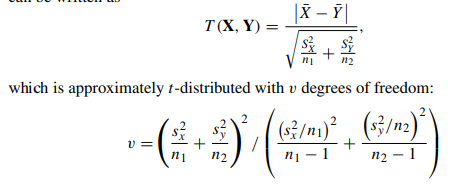
X và σ2

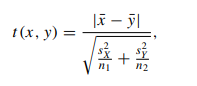
Y

không rõ. Vấn đề này còn được gọi là vấn đề Behrens-Fisher và là

kiểm tra thường được sử dụng nhất khi so sánh hai phương tiện trong thực tế. Thống kê thử nghiệm

có thể được viết như





## 10.3.4 Thử nghiệm so sánh phương tiện của hai mẫu phụ thuộc (t-Test)

Giả sử có hai biến ngẫu nhiên liên tục phụ thuộc X và Y với E (X) =

μX và E (Y) = μY. Chúng có thể phụ thuộc vì chúng tôi đo lường cùng một biến

hai lần trên cùng một đối tượng vào những thời điểm khác nhau. Thông thường, đây là trường hợp trước khi đăng

thí nghiệm, ví dụ: khi chúng tôi đo trọng lượng của một người trước khi bắt đầu

chế độ ăn kiêng đặc biệt và sau khi kết thúc chế độ ăn kiêng; hoặc khi đánh giá các khoản chi tiêu của hộ gia đình

trên các thiết bị điện tử trong hai năm liên tiếp. Sau đó, chúng tôi nói rằng các mẫu

được ghép nối hoặc phụ thuộc. Vì cùng một biến được đo hai lần trên cùng một

chủ đề, nó có ý nghĩa khi tính toán sự khác biệt giữa hai giá trị tương ứng.

Gọi D = X - Y biểu thị biến ngẫu nhiên "sự khác biệt của X và Y". Nếu H0: μX = μY

là true, khi đó sự khác biệt kỳ vọng bằng 0 và chúng ta nhận được E (D) = μD = 0. Điều này có nghĩa là

thử nghiệm H0: μX = μY giống với thử nghiệm μX - μY = μD = 0. Chúng tôi giả thiết thêm

rằng D được phân phối chuẩn nếu H0: μX = μY là đúng (hoặc tương đương nếu H0: μD = 0 là

true), tức là D ∼ N (0, σ2

D). Đối với một mẫu ngẫu nhiên (D1, D2, ..., Dn) về sự khác biệt,

thống kê thử nghiệm



ist-phân phối với n - 1 bậc tự do. Giá trị trung bình mẫu là D¯ = n

i = 1 / Din

và phương sai mẫu là



là một ước tính của σ2

D. Thống kê thử nghiệm được thực hiện là như vậy



# 10.4 Kiểm tra tham số cho xác suất

## 10.4.1 Phép thử nhị thức một mẫu cho xác suất p

Thử nghiệm xây dựng và các giả thuyết.

Gọi X là biến ngẫu nhiên Bernoulli B (1; p) với hai kết quả có thể xảy ra là 1

và 0, biểu thị sự xuất hiện và không xảy ra của một sự kiện quan tâm A.

xác suất để A trong quần thể là p. Từ mẫu X = (X1, X2, ..., Xn) của

các biến ngẫu nhiên phân phối B (1; p) độc lập, chúng tôi tính giá trị trung bình (tương đối

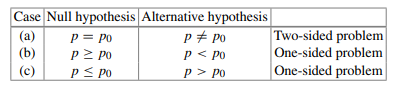
tần số) dưới dạng pˆ = 1

N

n

i = 1 Xi là ước lượng không chệch của p. Sau

Do đó, các giả thuyết có thể được quan tâm:



Sau đây, chúng tôi mô tả hai giải pháp khả thi, một cách tiếp cận chính xác và một

lời giải gần đúng. Giải pháp gần đúng dựa trên giá trị gần đúng của

phân phối nhị thức theo phân phối chuẩn, thích hợp nếu n là

đủ lớn và điều kiện np (1 - p) ≥ 9 giữ nguyên (tức là p không quá nhỏ

cũng không quá lớn). Đầu tiên, chúng tôi trình bày giải pháp gần đúng và sau đó là giải pháp chính xác.

Kiểm tra thống kê và kiểm tra quyết định.

(a) Kiểm tra gần đúng nhị thức. Chúng tôi xác định thống kê thử nghiệm chuẩn hóa là



Nó xấp xỉ rằng T (X) ∼ N (0, 1), với điều kiện (i)

n đủ lớn và (ii) np (1 - p) ≥ 9 thỏa mãn. Sau đó kiểm tra có thể

được tiến hành dọc theo đường của bài kiểm tra Gauss ở Sect. 10.3.1; đó là, bài kiểm tra

quyết định dựa trên Bảng 10.1.

Ví dụ 10.4.1 Chúng ta quay lại Ví dụ 10.1.1. Hãy để chúng tôi giả định rằng một đại diện

mẫu cỡ n = 2000 đã được lấy từ số lượng cử tri đủ điều kiện, từ

trong đó 700 (35%) đã bỏ phiếu cho bên quan tâm P. Giả thuyết nghiên cứu

(phải được nêu là H1) là hơn 30% (tức là p0 = 0,3) đủ điều kiện

cử tri bỏ phiếu cho đảng P. Mẫu có lợi cho H1 vì pˆ = 35%,

nhưng để đưa ra kết luận về tỷ lệ cử tri của đảng P trong dân số,

chúng ta phải tiến hành một phép thử nhị thức. Vì n lớn và np (1 - p) = 2000 · 0,35 ·

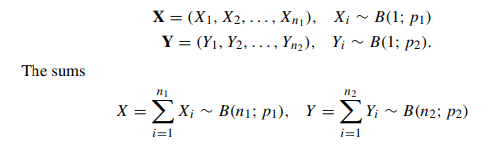
0,65 = 455 ≥ 9, các giả thiết cho việc sử dụng thống kê thử nghiệm (10,11) được thỏa mãn.

Chúng tôi có thể viết ra thống kê thử nghiệm đã nhận ra dưới dạng

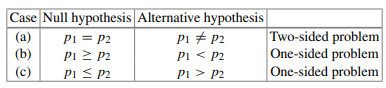


## 10.4.2 Kiểm tra nhị thức hai mẫu

Bây giờ chúng tôi xem xét trường hợp của hai i.i.d độc lập. mẫu từ các phân phối Bernoulli với các tham số p1 và p2.



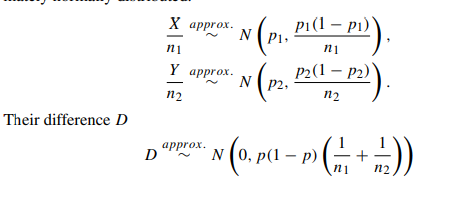
tuân theo các phân phối nhị thức. Một trong những giả thuyết sau đây có thể được quan tâm:



Tương tự như trường hợp một mẫu, cả hai thử nghiệm chính xác và gần đúng đều tồn tại. Ở đây, chúng tôi chỉ

trình bày bài kiểm tra gần đúng. Thử nghiệm chính xác của Fisher được trình bày trong Phụ lục C.5,

P. 428. Gọi n1 và n2 là kích thước mẫu. Khi đó, X / n1 và Y / n2 được phân phối xấp xỉ bình thường:



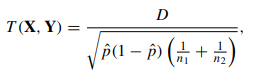
được phân phối bình thường quá dưới H0 (cho trước p = p1 = p2 giữ). Vì xác suất p1 và p2 giống hệt nhau theo H0, chúng ta có thể gộp hai mẫu lại và ước lượng

p bởi



Kiểm tra thống kê và kiểm tra quyết định.

Thống kê thử nghiệm



tuân theo phân phối N (0, 1) nếu n1 và n2 đủ lớn và p không gần

ranh giới 0 và 1 (ví dụ, người ta có thể sử dụng lại điều kiện np (1 - p)> 9

với n = n1 + n2). Thống kê thử nghiệm thực tế có thể được tính toán bằng cách sử dụng

hiệu dˆ = ˆp1 - ˆp2. Thử nghiệm có thể được tiến hành đối với trường hợp một mặt và hai mặt như thử nghiệm Gauss được giới thiệu trong Sect. 10.3.1; nghĩa là, các quy tắc quyết định từ

Bảng 10.1 có thể được áp dụng.

# 10.5 Kiểm tra các thông số tỷ lệ

Có nhiều bài kiểm tra khác nhau có sẵn để kiểm tra các giả thuyết về các thông số thang đo. Như là

các thử nghiệm hữu ích khi một người quan tâm đến sự phân tán của một biến, ví dụ

trong kiểm soát chất lượng, nơi mà sự thay đổi của một quá trình có thể được quan tâm. Một mâu

kiểm định các giả thuyết về phương sai của phân phối chuẩn, ví dụ: giả thuyết như vậy

như H0: σ2 = σ2

0, có thể được kiểm tra bằng phép thử χ2 cho phương sai, xem Phụ lục C.5,

P. 430. Các vấn đề về hai mẫu có thể được giải quyết bằng F-test (được giải thích trong

Phụ lục C.5, tr. 431); hoặc bằng các thử nghiệm khác, chẳng hạn như thử nghiệm Levene hoặc thử nghiệm Bartlett,

cũng có sẵn trong R (leveneTest trong xe trọn gói, bartlett trong

phân phối cơ sở của R).

# 10.6 Wilcoxon–Mann–Whitney (WMW) U-Test

Thử nghiệm xây dựng và các giả thuyết.

Kiểm tra WMW U thường được đề xuất thay thế cho kiểm tra t vì nó cũng

tập trung vào vị trí nhưng không tập trung vào giá trị mong đợi μ. Nó là một bài kiểm tra phi tham số

và hữu ích trong các tình huống mà các phân phối lệch được so sánh với nhau.

Chúng tôi coi hai mẫu ngẫu nhiên độc lập X = (X1, X2, ..., Xn1) và Y =

(Y1, Y2 ,, ..., Yn2) từ hai quần thể có giá trị quan sát (x1, x2, ..., xn1) và

(y1, y2, ..., yn2), tương ứng. Trong trường hợp này, giả thuyết vô hiệu H0 xem xét

vị trí có thể được xây dựng dưới dạng

H0: P (X> Y) = P (Y> X) = 1

2.

Giả thuyết rỗng có thể được giải thích theo cách sau: xác suất để a

quan sát được rút ra ngẫu nhiên từ quần thể đầu tiên có giá trị x lớn hơn

(hoặc thấp hơn) so với giá trị y của một đối tượng được lấy ngẫu nhiên từ tập hợp thứ hai

là 1

2. Giả thuyết thay thế H1 sau đó là

H1: P (X> Y) = P (Y> X).

Điều này có nghĩa là chúng tôi đang so sánh toàn bộ phân phối của hai biến. Nếu có một

dịch chuyển vị trí theo nghĩa là một phân phối được dịch chuyển sang trái (hoặc phải) so với

với phân phối khác, giả thuyết rỗng sẽ bị bác bỏ vì sự thay đổi này

có thể được xem như một phần của giả thuyết thay thế P (X> Y) = P (Y> X). Trong thực tế,

theo một số giả định, giả thuyết thậm chí có thể được hiểu là so sánh hai

trung gian, và đây là điều thường được thực hiện trong thực tế.

Thống kê thử nghiệm quan sát.

Để xây dựng thống kê thử nghiệm, cần phải hợp nhất (x1, x2, ..., xn1) và (y1, y2,

..., yn2) vào một mẫu được sắp xếp, thường theo thứ tự tăng dần, trong khi vẫn giữ thông tin giá trị thuộc về mẫu nào. Hiện tại, chúng tôi giả định rằng tất cả các giá trị của

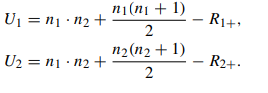
hai mẫu là khác biệt; nghĩa là, không có ràng buộc nào hiện tại. Sau đó, mỗi quan sát có

10.6 Thử nghiệm U Wilcoxon – Mann – Whitney (WMW) 233

hạng từ 1 đến (n1 + n2). Gọi R1 + là tổng các bậc của mẫu x và

gọi R2 + là tổng các bậc của mẫu y. Thống kê thử nghiệm được định nghĩa là U, trong đó

U là giá trị nhỏ nhất trong hai giá trị U1, U2, U = min (U1, U2) với



Quyết định kiểm tra.

H0 bị loại nếu U <un1, n2; α. Ở đây, un1, n2; α là giá trị tới hạn xuất phát từ

phân phối của U theo giả thuyết không. Phân phối chính xác (phức tạp) có thể,

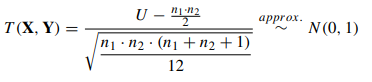
ví dụ, được dẫn xuất tính toán (trong R). Chúng tôi đang trình bày một giá trị gần đúng

giải pháp cùng với việc triển khai nó trong R.

Vì U1 + U2 = n1 · n2, nên chỉ tính Ri + và U = min {Ui,

n1n2 - Ui} (i = 1 hoặc i = 2 được chọn sao cho Ri + được tính cho mẫu

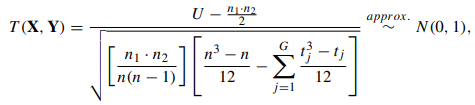
với kích thước mẫu thấp hơn). Đối với n1, n2 ≥ 8, người ta có thể sử dụng tính gần đúng



như thống kê thử nghiệm. Đối với giả thuyết hai vế, H0 bị bác bỏ nếu | t (x, y) | > z1 − α / 2;

đối với các giả thuyết một phía H0 bị bác bỏ nếu | t (x, y) | > z1 − α. Trong trường hợp quan hệ,

mẫu số của thống kê thử nghiệm trong (10.16) có thể được sửa đổi thành



# 10.7 χ2-Kiểm tra độ phù hợp

Thi công thử nghiệm.

Thử nghiệm χ2-độ phù hợp là một trong những thử nghiệm phổ biến nhất để kiểm tra độ tốt

sự phù hợp của dữ liệu quan sát được với một phân phối. Nguyên tắc xây dựng rất chung chung

và có thể được sử dụng cho các biến của bất kỳ thang đo nào. Thống kê thử nghiệm được suy ra sao cho

tần số tuyệt đối quan sát được so sánh với tần số tuyệt đối dự kiến

theo giả thuyết vô hiệu H0.

Ví dụ 10.7.1 Xét một thí nghiệm trong đó một con súc sắc được lăn n = 60 lần. Dưới

giả thuyết vô hiệu H0, chúng tôi giả định rằng con súc sắc là công bằng, tức là pi = 1

6, i = 1, 2, ..., 6,

trong đó pi = P (X = i). Chúng tôi cũng có thể nói rằng H0 là giả thuyết rằng

các cuộn đang tuân theo một phân bố đồng đều rời rạc. Do đó, các tần số tuyệt đối dự kiến ​​dưới H0 là npi = 60 · 1

6 = 10, trong khi các tần số quan sát được trong

mẫu là Ni, i = 1, 2, ..., 6. Ni thường lệch với npi. Thống kê χ2

dựa trên sự khác biệt bình phương, 6

i = 1 (Ni - npi) 2, và trở nên lớn khi

sự khác biệt giữa tần số quan sát và tần số mong đợi trở nên lớn hơn. Các

Thống kê χ2-test là sự sửa đổi của tổng này bằng cách chia tỷ lệ mỗi chênh lệch bình phương bằng

tần số mong đợi, npi, và được giải thích bên dưới.

Với một biến danh nghĩa, chúng ta có thể tiến hành như trong Ví dụ 10.7.1. Nếu quy mô của

biến là thứ tự hoặc liên tục, số lượng các giá trị khác nhau có thể lớn. Ghi chú

rằng trong trường hợp khắc nghiệt nhất, chúng ta có thể có nhiều giá trị khác nhau như các quan sát

(n), dẫn đến Ni = 1 với mọi i = 1, 2, ..., n. Sau đó, cần phải nhóm dữ liệu

thành k khoảng thời gian trước khi áp dụng phép thử χ2. Lý do là lý thuyết chung về

phép thử χ2 giả định rằng số k (là 6 trong Ví dụ 10.7.1 ở trên) là

cố định và không phát triển với số lần quan sát n; đó là, lý thuyết nói

rằng phép thử χ2 chỉ hoạt động đúng nếu k cố định và n lớn. Vì lý do này, chúng tôi

nhóm mẫu X = (X1, X2, ..., Xn) thành k lớp như trong Sect. 2.1.

Lớp 1 2 ·· k Tổng số

Số lần quan sát n1 n2 ··· nk n

Việc lựa chọn các khoảng thời gian trong lớp hơi tùy ý. Theo quy tắc ngón tay cái npi> 5

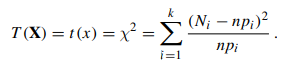
nên giữ trong hầu hết các khoảng thời gian của lớp học. Các giả thuyết chung có thể được xây dựng trong

dạng của các hàm phân phối:



Thử nghiệm thống kê.

Thống kê thử nghiệm được định nghĩa là



• Ni (i = 1, 2, ..., k) là tần số tuyệt đối của các quan sát của mẫu X

trong lớp i, Ni là một biến ngẫu nhiên với nhận thức ni trong mẫu quan sát;

• pi (i = 1, 2, ..., k) được tính từ phân phối dưới H0, F0 (x), và là

xác suất (giả thiết) mà một quan sát X thuộc loại i;

• npi là tần số tuyệt đối dự kiến ​​trong lớp i dưới H0.

Quyết định kiểm tra.

Đối với mức ý nghĩa α, H0 bị loại bỏ nếu t (x) lớn hơn (1 - α) -quantile

của phân phối χ2 với k - 1 - r bậc tự do, tức là nếu

t (x) = χ2> ck − 1 − r, 1 − α.

Lưu ý rằng r là số tham số của F0 (x), nếu các tham số này được ước lượng

từ mẫu. Thống kê χ2 thử nghiệm chỉ được phân phối tiệm cận χ2 trong

H0.

Ví dụ 10.7.2 Gọi F0 (x) là hàm phân phối của phân phối thử nghiệm. Nếu

người ta chỉ định phân phối chuẩn, chẳng hạn như F0 (x) = N (3, 10) hoặc đồng nhất rời rạc

phân phối với pi = 0,25 (i = 1, 2, 3, 4), thì r = 0, vì không có tham số nào có

được ước tính từ dữ liệu. Ngược lại, nếu chúng ta chỉ muốn kiểm tra xem

dữ liệu được tạo ra từ phân phối chuẩn N (μ, σ2) hoặc dữ liệu tuân theo chuẩn

phân phối N (μ, σ2), thì μ và σ2 có thể được ước lượng từ mẫu theo x¯ và s2.

Khi đó, r = 2 và số bậc tự do bị giảm đi.

# 10.8 Thử nghiệm χ2-Độc lập và Thử nghiệm χ2-Thử nghiệm khác

Trong Chap. 4, chúng tôi đã giới thiệu các phương pháp khác nhau để mô tả mối liên hệ giữa

hai biến. Một số biện pháp kết hợp có thể phù hợp nếu các biến

phân loại, ví dụ như Cramer’s V, Goodman’s và Kruskal’s γ, Spearman’s rank

hệ số tương quan và tỷ lệ chênh lệch. Nếu chúng ta không quan tâm đến sức mạnh của

nhưng thay vì tìm hiểu xem liệu có một liên kết nào không, người ta có thể

sử dụng thử nghiệm χ2-độc lập.

Thi công thử nghiệm.

Trong phần sau, chúng tôi giả định rằng chúng tôi quan sát một mẫu từ một biến phân biệt rời rạc

phân phối của hai biến X và Y có thể được tóm tắt trong trường hợp dự phòng

bảng với tần số tuyệt đối ni j, (i = 1, 2, ..., I; j = 1, 2 ..., J)

Nhớ lấy

ni + là tổng hàng thứ i,

n + j là tổng cột thứ j, và

n là tổng số lần quan sát.

Các giả thuyết là H0: X và Y là độc lập so với H1: X và Y thì không

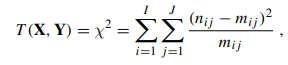
sống độc lập. Nếu X và Y độc lập, thì tần số mong đợi mi j là

mˆ i j = nπˆi j = ni + n + j

N . (10.18)

Thử nghiệm thống kê.

Thống kê thử nghiệm χ2 của Pearson đã được giới thiệu trong Chap. 4, Phương trình (4,6). Nó là



trong đó mi j = nπi j = nπi + π + j (tần số ô tuyệt đối dự kiến dưới H0).

Nói một cách chính xác, mi j là tần số dự kiến thực, chưa biết dưới H0 và

được ước tính bởi mˆ i j = nπi + π + j, sao cho thống kê thử nghiệm thực tế tương đương với



Quyết định kiểm tra.

Số bậc tự do theo H0 là (I - 1) (J - 1), trong đó I - 1 là

các tham số phải được ước tính cho phân phối biên của X và J - 1

là số lượng các tham số cho phân phối biên của Y. Quyết định kiểm tra là:

Bác bỏ H0, nếu t (x, y) = χ2> c (I − 1) (J − 1); 1 − α.

Lưu ý rằng giả thuyết thay thế H1 rất chung chung. Nếu H0 bị từ chối, không gì có thể

được cho biết về cấu trúc của sự phụ thuộc của X và Y từ giá trị χ2 của chính nó.

Ví dụ 10.8.1 Xem xét bảng dự phòng sau đây. Ở đây, X mô tả

trình độ học vấn (1: tiểu học, 2: trung học, 3: đại học) và Y ưu tiên cho một

chính đảng cụ thể (1: Bên A, 2: Bên B, 3: Bên C). Giả thuyết vô hiệu của chúng tôi là

hai biến độc lập và chúng tôi muốn đưa ra giả thuyết thay thế

mà nói rằng có một mối quan hệ giữa chúng.

# 10.9 Các điểm chính và các vấn đề khác

Tóm tắt đồ họa về thời điểm sử dụng các bài kiểm tra được giới thiệu trong chương này

được nêu trong Phụ lục D.2 và D.3.

Để đi đến quyết định kiểm tra, tức là chấp nhận H0 hoặc từ chối nó, điều đó không quan trọng

cho dù người ta so sánh thống kê thử nghiệm với vùng quan trọng, người ta sử dụng

p-value thu được từ phần mềm thống kê hoặc đánh giá khoảng tin cậy thích hợp. Tuy nhiên, điều quan trọng là không hiểu sai

p-value (xem Sect. 10.2.6) và để chọn khoảng tin cậy đúng.

Có sự khác biệt giữa mức độ liên quan và ý nghĩa. Một bài kiểm tra có thể là

đáng kể, nhưng ước tính điểm về số lượng quan tâm có thể không

có liên quan từ một quan điểm thực chất. Tương tự, một bài kiểm tra có thể không

quan trọng, nhưng các ước tính điểm và khoảng thời gian vẫn có thể mang lại kết quả phù hợp

kết luận.

Thống kê kiểm định của phép thử t (một mẫu, hai mẫu, được ghép nối) được phân phối theo phương pháp tiệm cận bình thường. Điều này có nghĩa là đối với n tương đối lớn (như

quy tắc ngón tay cái> 30 cho mỗi nhóm) mẫu không cần phải đến từ

phân phối bình thường. Tuy nhiên, việc áp dụng phép thử t có ý nghĩa

chỉ khi kỳ vọng μ có thể được giải thích một cách có nghĩa; điều này có thể

không phải là trường hợp phân phối lệch hoặc phân phối có giá trị ngoại lai.

dù đó

là ngày cuối tuần, v.v. Một câu hỏi hợp lý cần đặt ra là: Các tài xế khác nhau có

(trung bình) thời gian giao hàng khác nhau hay không — do họ giao dịch với cùng một

nhà điều hành, số lượng thực phẩm được đặt như nhau, ngày giống nhau, v.v. Những câu

hỏi kiểu này có thể được trả lời với nhiều hồi quy tuyến tính. Mô hình chứa nhiều hơn

một, chẳng hạn p, hiệp biến X1, X2, ..., X Mô hình tuyến rộngtính như(11.2) có thể được mở

Y = β0 + β1X1 + ··· + βpX p

P.

newX <- sqrt (X) lm

(YnewX) #option 1 lm (YI (sqrt

(X))) #option 2

(11.15

10.10 Bài tập

(a) Các giả thuyết là



(b) Đây là bài toán một mẫu đối với μ: do đó, đối với phương sai đã biết, có thể sử dụng  
kiểm định Gauss; ngược lại, kiểm định t, xem thêm Phụ lục D. Vì, trong bài tập này, σ  
được giả định là đã biết, chúng ta có thể sử dụng kiểm định Gauss; tức là chúng ta có  
thể so sánh thống kê kiểm định với phân phối chuẩn (trái ngược với phân phối t khi sử  
dụng kiểm định t). Kích thước mẫu nhỏ: do đó chúng ta phải giả định một phân phối  
chuẩn cho dữ liệu.

(c) Để tính toán thống kê thử nghiệm đã thực hiện, chúng ta cần trung bình cộng x¯ = 98,08.



Ta bác bỏ H0 nếu | t (x) | > z1 α = 1,96. Kể từ khi | t (x) | = 3,72> 1,96, giả thuyết vô hiệu có thể bị bác bỏ. Thử nghiệm là đáng kể. Có đủ bằng chứng cho thấy rằng các trọng lượng quan sát được không đến từ phân phối chuẩn với giá trị trung bình 100 g.

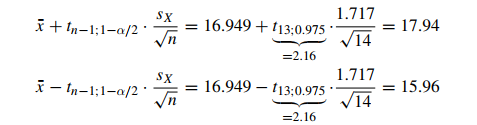
(d) Để chỉ ra rằng μ <100, chúng ta cần tiến hành thử nghiệm một phía. Quan trọng nhất, giả thuyết nghiên cứu quan tâm cần được phát biểu là giả thuyết thay thế :



(e) Thống kê thử nghiệm giống như trong (b): t (x) = 3,72. Tuy nhiên, vùng quan trọng những thay đổi. H0 bị loại nếu t (x) < z1 α = 1,64. Một lần nữa, giả thuyết vô hiệu bị từ chối. Nhà sản xuất đã đúng khi đưa ra giả thuyết rằng trọng lượng trung bình của thanh sô cô la của anh ta thấp hơn 100 g.

Lưu ý rằng chúng tôi sử dụng các ước tính không chệch cho phương sai và tiêu chuẩn devia tiêu chuẩn như được giải thích trong Chap. Số 9; tức là chúng ta sử dụng 1 / (n - 1) thay vì 1 / n. Nó là điều hiển nhiên rằng giá trung bình của các cuộc đấu giá (μa) thấp hơn giá trung bình không đấu giá.

(c) Sử dụng (9.6) chúng ta có thể tính toán giới hạn trên và giới hạn dưới của khoảng tin cậy như:



11. Hồi quy tuyến tính.

# 11.7 Quan điểm quy nạp của hồi quy

Cho đến nay, chúng tôi đã giới thiệu hồi quy tuyến tính như một phương pháp mô tả mối

quan hệ tương đối giữa các biến phụ thuộc và độc lập thông qua sự phù hợp nhất với dữ

liệu. Tuy nhiên, như chúng tôi đã nhấn mạnh trong các chương trước, chúng tôi thường

không chỉ quan tâm đến việc mô tả dữ liệu mà còn quan tâm đến việc đưa ra kết luận từ

một mẫu về dân số quan tâm. Ví dụ: trong một ví dụ trước đó, chúng tôi đã ước tính sự

liên kết giữa chi nhánh và thời gian giao hàng cho dữ liệu giao bánh pizza. Mô hình tuyến

tính được ước tính là thời gian giao hàng = 36,3 - 5,2 Đông - 1,1 Tây; điều này cho thấy

thời gian giao hàng trung bình ngắn hơn 5,2 phút đối với chi nhánh ở phía Đông của thị

trấn so với chi nhánh trung tâm và ngắn hơn 1,1 phút đối với chi nhánh ở phía Tây so với

chi nhánh trung tâm. Khi xem xét tất cả các lần giao bánh pizza (dân số), không chỉ những

lần giao hàng do chúng tôi thu thập trong khoảng thời gian một tháng (mẫu), chúng tôi có

thể tự hỏi liệu 1,1 phút là chênh lệch thực sự hợp lệ cho toàn bộ dân số hay chỉ do lỗi

ngẫu nhiên liên quan trong mẫu chúng tôi đã chọn. Đặc biệt, chúng tôi cũng muốn biết

khoảng tin cậy 95% cho ước tính tham số của chúng tôi là bao nhiêu.

Các biến đổi của Biến kết quả. Cũng có thể áp dụng các biến đổi cho biến kết quả. Trong

nhiều tình huống, điều này làm cho việc diễn giải khá khó khăn, một phép biến đổi nhật

ký là khá phổ biến và dễ diễn giải. Xem xét mô hình log-tuyến tính

đường hồi quy cho thấy thời gian hồi phục tăng lên nhiều đối với thời gian tập thể dục

lớn hơn 65 phút. Mặc dù có thể tập thể dục quá nhiều gây ra tổn thương cho đầu gối và

làm chậm quá trình phục hồi, nhưng dữ liệu dường như không đủ rõ ràng để hỗ trợ hình dạng

mà mô hình đề xuất. Ví dụ này minh họa sự cân bằng giữa sự phù hợp (nghĩa là đường hồi

quy phù hợp với dữ liệu tốt như thế nào) và parsimony (tức là mô hình có thể được hiểu

tốt như thế nào nếu nó trở nên phức tạp hơn). Phần 11.8 khám phá vấn đề không nhỏ này

một cách chi tiết hơn.

Có thể dễ dàng nhận thấy rằng xj tăng một đơn vị sẽ nhân y với eβj . Do đó, nếu y được

biến đổi thành log, người ta chỉ cần giải thích exp (β) thay vì β. Ví dụ: nếu y là thu

nhập hàng năm, x biểu thị giới tính (1 = nam, 0 = nữ) và β1 = 0,2 thì người ta có thể

nói rằng thu nhập của một người đàn ông là exp (0,2) = 1,22 lần (tức là 22%) so với của

một người phụ nữ.

Liên quan đến Chaps. 9 và 10, bây giờ chúng ta áp dụng các khái niệm về suy luận và kiểm

định thống kê trong bối cảnh phân tích hồi quy

# 11.8 So sánh các mô hình khác nhau

Sai số được ước tính từ dữ liệu là eˆ = y - Xβˆ và được gọi là phần dư.

khác 0 hoặc liệu độ lệch của các ước tính từ 0 có thể được giải thích hay không

(bao gồm một cột của 1 cho phần đánh chặn). Ma trận nhận dạng I bao gồm 1 trên

Chúng tôi muốn ước lượng β để đưa ra kết luận về một mối quan hệ trong pop ulation. Tương

tự, e phản ánh các sai số ngẫu nhiên trong tổng thể. Quan trọng nhất,

βˆ = (X X) 1X

Có thể chứng minh rằng βˆ tuân theo phân phối chuẩn với trung bình E (βˆ) = β và ma trận

covari ance Var (βˆ) = σ2I là

tính hợp lệ trên cơ sở một mẫu dữ liệu nhất định được giải thích trong Sect. 11,9.

y.

mô hình tuyến tính hiện chứa các giả định về các lỗi. Họ được cho là

Ước lượng điểm và khoảng trong mô hình tuyến tính. Bây giờ chúng tôi viết lại mô hình cho

βˆ N (β, σ2 (X X) 1). (11,26)

n - (p + 1)

=

kết luận về dân số quan tâm từ mô hình tuyến tính. Như chúng ta đã thấy,

và khi đó, mô hình sẽ không chứa thuật ngữ β jxj. Điều này có nghĩa là hiệp biến xj

là 0, E (ei) = 0, phương sai là Var (ei) = σ2 (và do đó giống nhau đối với mọi ei), và

trong Phụ lục C.6. Một ước tính không chệch của σ2 là

y = β01 + β1x1 + ··· + βpxp + e

2 σˆ

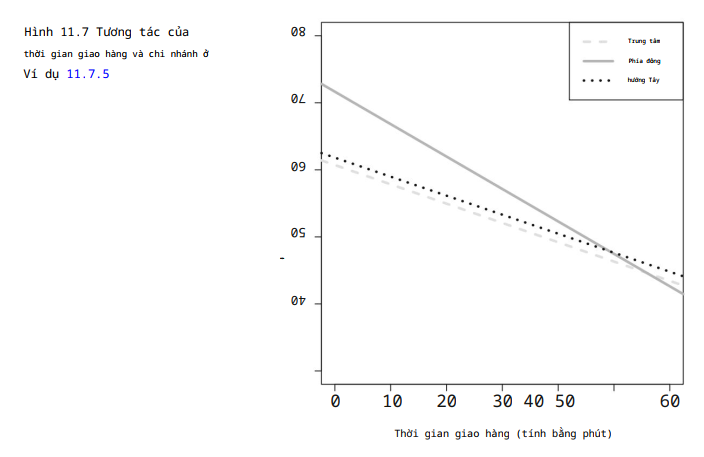
n - (p + 1)

do đó tương đương với việc kiểm tra xem xj có liên kết với y hay không theo nghĩa

Trước khi đưa ra một ví dụ về dữ liệu chi tiết, chúng tôi muốn phác thảo cách vẽ

theo biến thể ngẫu nhiên, chúng tôi có các tùy chọn sau:

# 11.9 Kiểm tra mô hình giả định



Có thể thấy pizza do chi nhánh phía Đông giao hàng về tổng thể hot hơn các hộp từ nhánh phía đông không được đóng đúng cách và do đó — mặc dù

nó có vẻ ngoài lúc đầu. Nếu lãi suất là x1, thì ước tính đề cập đến loại

Ví dụ 11.7.6 Hãy xem xét lại dữ liệu bánh pizza. Nếu chúng ta có giả thuyết rằng

Việc giải thích các tham số hồi quy cần quan tâm ít phức tạp hơn

nhưng thời gian giao hàng lâu hơn mức có lợi. Người ta có thể suy đoán rằng việc giao hàng

khác nhau đối với các nhánh khác nhau, sau đó là một mô hình hồi quy với nhánh (3 loại, 2

biến giả), toán tử (2 danh mục, một biến giả) và tương tác của chúng

lm (bộ điều hành thời gian \* nhánh)

danh mục tham chiếu của x2. Ảnh hưởng của x1i có thể thay đổi đối với x2, và

Categorical – Tương tác phân loại. Đối với hai biến phân loại x1 và x2, với

tổng các hiệu ứng chính và tương tác tương ứng cho mỗi loại sản lượng x2

hiệu suất tổng thể tốt — nhiệt độ giảm nhanh hơn theo thời gian vì điều này

chi nhánh.

lãi suất (tức là x1i) được hiểu như bình thường — với sự khác biệt mà nó chỉ liên quan đến

(2 biến mới) có thể được sử dụng.

ví dụ sau đây.

• Phía Đông: Nhiệt độ = 70,7 + 10,9 - (0,29 + 0,195) × thời gian,

cần được tạo như sau:

loại k và l , tương ứng, (k - 1) × (l - 1) biến giả mới x1i × x2 j

• Trung tâm: Nhiệt độ = 70,7 - 0,29 × thời gian,

Nếu chúng tôi quan tâm đến nhà điều hành, chúng tôi thấy rằng thời gian giao hàng ở độ

tuổi khác ngắn hơn 0,21 phút đối với nhà điều hành “Melissa”. Khi toán tử này xử lý một

nhánh khác với tham chiếu (Trung tâm), ước tính thay đổi thành 0,21 + 0,86 = 0,64 phút

phân phối dài hơn trong trường hợp nhánh “Đông” và 0,21 + 0,48 = 0,26 phút cho nhánh

“Tây

Nếu chúng tôi quan tâm đến các chi nhánh, chúng tôi nhận thấy rằng thời gian giao hàng là

ngắn nhất đối với chi nhánh miền Đông, nơi có thời gian giao hàng trung bình ngắn hơn chi

nhánh trung tâm là 5,66 phút. Tuy nhiên, đây là ước tính chỉ dành cho toán tử tham chiếu;

nếu nhà điều hành “Melissa” phụ trách, thì chênh lệch thời gian giao hàng của hai chi

nhánh chỉ là 5,66 + 0,86 = 4,8 phút. Điều tương tự cũng áp dụng khi so sánh nhánh phía

tây với nhánh trung tâm: tùy thuộc vào nhà điều hành, sự khác biệt giữa hai nhánh này

được ước tính là 1,36 hoặc 1,36 + 0,48 = 0,88 phút, tương ứng.

Các điều khoản tương tác không khác 0 đáng kể. Do đó, chúng tôi kết luận rằng giả thuyết

về thời gian giao hàng khác nhau của hai nhà khai thác, có thể thay đổi theo chi nhánh,

không thể được xác nhận.

Liên tục – Tương tác liên tục. Cũng có thể thêm tương tác của hai biến liên tục. Điều này

được thực hiện bằng cách thêm tích của hai biến này vào mô hình. Nếu x1 và x2 là hai biến

liên tục, thì x1x2 là biến mới được thêm vào mô hình dưới dạng hiệu ứng tương tác như

Đầu ra R ở trên cho thấy rằng có một hiệu ứng tương tác đáng kể. Việc thử nghiệm giữa

các khó khăn hơn ở đây. Rõ ràng là thời gian giao hàng lâu hơn và hóa đơn lớn hơn sẽ làm

giảm nhiệt độ của bánh pizza. Tuy nhiên, đối với một tích số lượng lớn của hóa đơn và thời

gian (tức là khi cả hai đều lớn), các liên kết này ít rõ rệt hơn vì các hệ số âm ngày càng

lớn hơn thuật ngữ tương tác dương. Ngược lại, đối với một hóa đơn nhỏ và thời gian giao

hàng ngắn, nhiệt độ có thể được dự đoán là khá cao

# 11.10 Sự liên kết và Kết quả

Có nhiều tình huống trong đó nhiều mô hình tuyến tính khác nhau có thể được lắp vào một

tập dữ liệu nhất định, nhưng không rõ đâu là mô hình tốt nhất. Điều này liên quan đến

câu hỏi những biến nào nên được đưa vào một mô hình và những biến nào có thể bị loại bỏ.

mymodel <- lm (Temperaturetime \* branch) vcov

(mymodel)

khoảng tin cậy 95% của họ. Như chúng ta đã biết từ Định lý 7.7.1, Phương sai

của tổ hợp hai biến ngẫu nhiên có thể nhận được bằng Var (X ± Y ) = σ2 Var (X)

+ Var (Y ) ± 2 Cov (X, Y ). Do đó, chúng ta có thể ước tính khoảng tin cậy cho

ước lượng kết hợp βˆ i + βˆj là

Ví dụ 11.7.8 Nhớ lại Ví dụ 11.7.5 trong đó chúng tôi ước tính mối liên hệ giữa nhiệt độ

bánh pizza, thời gian giao hàng và chi nhánh. Có một hiệu ứng tương tác đáng kể đối với

thời gian và chi nhánh. Sử dụng R, chúng ta thu được ma trận hiệp phương sai như sau:

= Var (βi) + Var (βj) + 2 Cov (βi, βj).

(11,29)

Machine Translated by GoogleR2 R2

adj

Tuyến tính 0,6584 Bậc hai0,6243 0,6787

0,6074 Khối 0,7151 0,6083

11.8 So sánh các mô hình khác nhau 281

Căn bậc hai 0,6694 0,6363

R2

điều chỉnh

Ví dụ 11.8.1 Trong hình 11.6, mối liên hệ giữa thời gian tập thể dục và thời gian phục hồi

sau khi phẫu thuật đầu gối được mô hình hóa một cách tuyến tính, bậc hai và hình khối;

trong Hình 11.5, sự kết hợp này được mô hình hóa bằng một phép biến đổi căn bậc hai. Tóm

tắt mô hình trong R trả về cả R2 (trong “Nhiều R bình phương”) và R2 đã điều chỉnh (trong

“R bình phương được điều chỉnh”). Kết quả như sau:

SQError / (n - p - 1)

= 1 -

Có thể thấy rằng R2 lớn hơn đối với các mô hình có nhiều biến hơn; tức là mô hình khối

(bao gồm ba biến) có R2 lớn nhất. R2 điều chỉnh , có tính đến các kích thước mô hình khác

nhau, ủng hộ mô hình có phép biến đổi căn bậc hai. Do đó, mô hình này cung cấp sự phù hợp

nhất với dữ liệu trong số bốn mô hình được xem xét sử dụng R2.

Có nhiều tiêu chí lựa chọn mô hình khác nhau để so sánh chất lượng của các mô hình được

trang bị khác nhau. Chúng có thể hữu ích trong nhiều trường hợp, nhưng cũng có một số hạn

chế khiến việc lựa chọn mô hình trở thành một nhiệm vụ khó khăn và đầy thử thách.

Các biến được thêm vào có thể không phù hợp, nhưng R2 vẫn sẽ tăng, sai chỉ ra sự cải thiện

trong mô hình được điều chỉnh. Tiêu chí này do đó không phải là thước đo tốt để lựa chọn

mô hình. Phiên bản điều chỉnh của R2 được định nghĩa là

. (11.30)

Ví dụ, khi chúng tôi lập mô hình liên kết giữa thời gian phục hồi và thời gian thực hiện

bằng cách sử dụng các phép biến đổi khác nhau, vẫn chưa rõ biến đổi nào được đề xuất là tốt

nhất. Trong ví dụ về giao bánh pizza, chúng ta có khoảng 10 hiệp biến: Liệu có hợp lý khi

đưa tất cả chúng vào mô hình hay chúng ta hiểu các mối liên kết trong dữ liệu tốt hơn bằng

cách giới hạn bản thân ở một vài biến quan trọng nhất? Có cần thiết phải bao gồm các tương

tác hay không?

R2 điều chỉnh . Mặc dù R2 là một thước đo hợp lý để đánh giá mức độ phù hợp của một mô

hình, nhưng nó cũng có một số hạn chế. Một hạn chế là nếu số hiệp biến tăng, R2 cũng tăng

(chúng ta bỏ qua định lý và chứng minh cho phát biểu này).

Hệ số xác định (R2). Một tiêu chí có thể có để kiểm tra mức độ phù hợp của mô hình là hệ số

xác định (R2). Chúng tôi đã thảo luận về sự phát triển, giải thích và triết lý của nó trong

Sect. 11.3 (xem trang 256 để biết thêm chi tiết).

# 11.11 Điểm nhấn và Các vấn đề tương lai

AIC càng nhỏ, mô hình càng tốt. AIC không chỉ tính đến sự phù hợp với dữ liệu qua SQError mà

còn tính đến phân tích cú pháp của mô hình thông qua thuật ngữ 2 (p + 1).

AIC, R2 hoặc giá trị p có thể đạt được.

Chúng tôi không đưa ra động lực chi tiết của tiêu chí này ở đây. Nó có thể được viết cho mô

hình tuyến tính là

Ví dụ 11.8.2 Xem xét lại Ví dụ 11.8.1. R2 thích mô hình bao gồm thời gian thực hiện thông

qua phép biến đổi căn bậc hai hơn các tùy chọn khác. Giá trị AIC cho mô hình trong đó thời

gian thực hành được mô hình hóa tuyến tính là 60,59, khi được mô hình hóa theo phương pháp

bậc hai là 61,84, khi được mô hình hóa theo phương thức lập phương là 62,4 và 60,19 đối với

phép biến đổi căn bậc hai. Do đó, phù hợp với R2 , AIC tínhcũng từ .thích phép biến đổi căn bậc hai ,

tương tự.

Tiêu chí Thông tin của Akaike (AIC) là một tiêu chí khác để lựa chọn mô hình. AIC dựa trên

phương pháp luận khả năng xảy ra và có nguồn gốc từ lý thuyết thông tin.

AIC = n log

Lựa chọn ngược. Hai mô hình, chỉ khác nhau một biến xj , có thể được so sánh bằng cách đơn

giản nhìn vào kết quả kiểm định cho βj = 0: nếu giả thuyết rỗng bị bác bỏ, biến được giữ

nguyên trong mô hình; nếu không, mô hình khác được chọn. Nếu có nhiều hơn hai mô hình, thì

tốt hơn là nên xem xét một cách tiếp cận có hệ thống để so sánh chúng. Ví dụ: giả sử chúng

ta có 10 biến tiềm năng có liên quan và chúng ta không chắc nên đưa chúng vào mô hình cuối

cùng. Có thể có 210 = 1024 sự kết hợp khác nhau của các biến số và lần lượt có rất nhiều sự

lựa chọn về mô hình!

3. Lặp lại bước 2. cho đến khi tiêu chí dừng được đáp ứng, nghĩa là cho đến khi không có cải tiến nào liên quan đến

1. Bắt đầu với mô hình đầy đủ chứa tất cả các biến quan tâm, Υ = {x1, x2

Do đó, chúng tôi kết luận rằng mô hình "tốt nhất" cuối cùng bao gồm tất cả các biến được xem xét,

mô hình cho cùng một tập dữ liệu. Tùy thuộc vào chiến lược được sử dụng để thêm và bớt

α, thì việc lựa chọn α là quan trọng), người ta có thể thu được các kết quả khác nhau.

các mô hình phù hợp, nhưng không có cách tốt nhất duy nhất để làm như vậy. Các tiêu chí khác nhau được dựa trên

diễn giải đơn giản như được nêu trong Sect. 11.7.3.

• Một sự cân nhắc tương tự đề cập đến các tương tác. Nếu một tương tác được thêm vào

thiết lập (ví dụ: nếu chiến lược là thêm / bớt một biến nếu giá trị p nhỏ hơn

Những hạn chế và những cân nhắc thực tế hơn nữa. Các cuộc thảo luận ở trên về các phương pháp tiếp

cận lựa chọn có thể thay đổi làm rõ rằng có nhiều lựa chọn khác nhau để so sánh

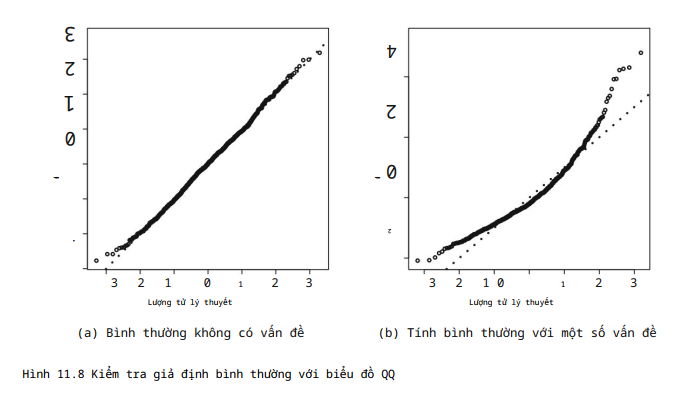
mô hình, sau đó các hiệu ứng chính cũng nên được giữ trong mô hình. Điều này cho phép

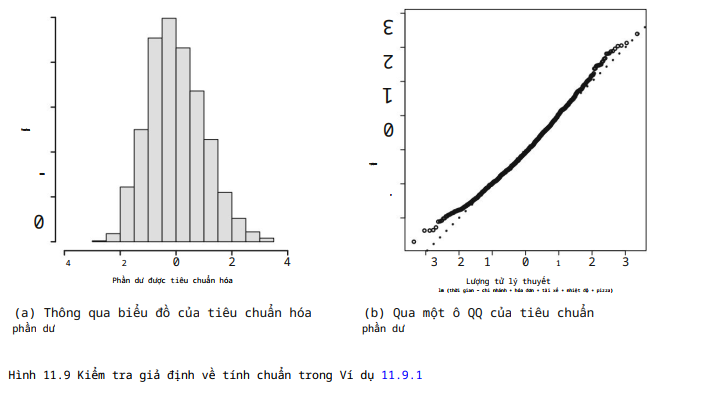
• Tất cả các cân nhắc ở trên cho thấy rằng việc lựa chọn mô hình là một nhiệm vụ không hề tầm thường mà

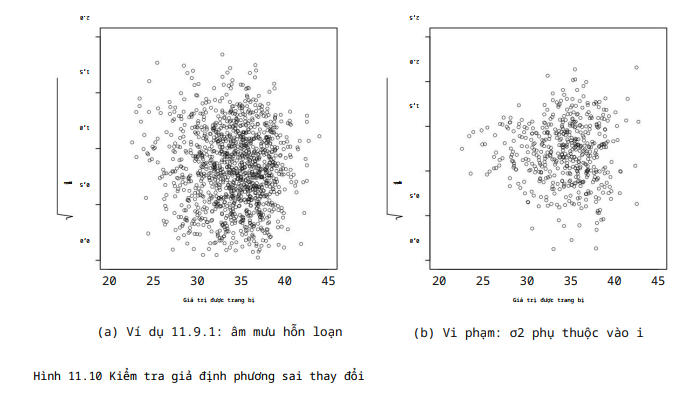
• Nếu các biến phân loại được bao gồm trong mô hình, thì tất cả các ables biến giả tương ứng cần được

xem xét như một tổng thể: tất cả các biến giả của biến này phải

biến đổi cũng như hiệu ứng tuyến tính nên được giữ trong mô hình.







11.1 Mô hình tuyến tính.

Xem xét các biểu đồ phân tán từ Hình 4.2 trên P. 80. Lập đồ thị giá trị X so với giá trị Y cho phép chúng ta hình dung mối quan hệ giữa hai biến. Hình 11.1a xem xét mối quan hệ thuận (tuyến tính) giữa X và Y trông như thế nào: giá trị X càng cao thì giá trị Y càng cao (và ngược lại). Cốt truyện chỉ ra rằng có thể có một mối quan hệ tuyến tính giữa X và Y. Mặt khác, Hình 11.1b hiển thị thước đo R2 , rõ biểu ràng đồ giữa phân X tán và Y không được cho thể thấy hiện mối trong quan hai hệ thích chi tiết hơn sau đó, tương đương với hệ số tương quan bình phương hình của Bravais- và được Pearson.

Tóm tắt mối quan hệ được quan sát giữa X và Y chúng ta có thể mặc định



Công thức này. (11.1) biểu diễn một đường thẳng trong đó α là điểm giao nhau và β biểu diễn hệ số góc của đường thẳng. Độ dốc cho biết sự thay đổi trong giá trị Y khi giá trị X thay đổi bởi một đơn vị. Nếu dấu của β là dương, nó chỉ ra rằng giá trị của Y tăng khi giá trị của X tăng. Nếu dấu của β là âm, nó chỉ ra rằng giá trị của Y giảm khi giá trị của X tăng. Khi X = 0 thì Y = α. Nếu α = 0, thì Y = βX đại diện cho một đường thẳng đi qua điểm gốc

11.2 Phương pháp bình phương nhỏ nhất

Giả sử n bộ quan sát Pi = (xi, yi), i = 1, 2, ..., n, thu được trên hai biến P = (X, Y ) và được vẽ trong biểu đồ phân tán. Ví dụ về bốn tions quan sát, (x1, y1), (x2, y2), (x3, y3) và (x4, y4), được đưa ra trong Hình 11.2.

Phương pháp ít tốn công nhất để một dòng có thể được khớp với tập dữ liệu đã cho để các lỗi được giảm thiểu. Điều này ngụ ý rằng người ta có thể xác định α và β sao cho tổng bình phương khoảng cách giữa các điểm dữ liệu và đường thẳng Y = α + βX là nhỏ nhất. Ví dụ, trong Hình 11.2, điểm dữ liệu đầu tiên (x1, y1) không nằm trên đường được vẽ và độ lệch là e1 = y1 - (α + βx1). Tương tự, chúng tôi có được

Chart

Description automatically generated

11.2.1 Thuộc tính của đường hồi quy tuyến tính

Có một số kết quả thú vị liên quan đến đường hồi quy và ước lượng bình phương nhỏ nhất. (i) Theo quy tắc chung, người ta nên giải thích đường hồi quy yˆi = ˆα + βˆxi chỉ trong khoảng [x (1), x (n)]. Ví dụ: nếu X biểu thị "Năm", với các giá trị quan sát từ năm 1999 đến năm 2015, và Y biểu thị “khối lượng bán hàng hàng năm của một công ty ”, thì một dự đoán cho năm 2030 có thể không có ý nghĩa hoặc không hợp lệ bởi vì mối quan hệ tuyến tính được phát hiện trong quá khứ có thể không tiếp tục duy trì tương lai

11.3 Độ tốt của sự vừa vặn

Mặc dù người ta có thể dễ dàng phù hợp với mô hình hồi quy tuyến tính, nhưng điều này không có nghĩa là mô hìnhnhất thiết phải tốt. Xem xét lại Hình 11.1: Trong Hình 11.1a, đường hồi quy là nắm bắt rõ ràng xu hướng tuyến tính được nhìn thấy trong biểu đồ phân tán. Sự phù hợp của mô hình với dữ liệu có vẻ tốt. Hình 11.1b cho thấy tuy nhiên, các điểm dữ liệu thay đổi đáng kể xung quanh dòng. Chất lượng của mô hình có vẻ không được tốt cho lắm. Nếu chúng tôi muốn sử dụng đường hồi quy để dự đoán dữ liệu, chúng tôi có thể sẽ nhận được kết quả kém. Nó xung quanh dòng. Chất lượng của mô hình có vẻ không được tốt cho lắm. Nếu chúng tôi muốnsử dụng đường hồi quy để dự đoán dữ liệu, chúng tôi có thể sẽ nhận được kết quả kém. Nó Rõ ràng là từ Hình 11.1 rằng mô hình cung cấp sự phù hợp tốt với dữ liệu khi quan sát gần đường thẳng và nắm bắt xu hướng tuyến tính.

Chart, scatter chart

Description automatically generated

11.4 Hồi quy tuyến tính với biến số nhị phân

Cho đến nay, chúng ta đã giả định rằng hiệp biến X là liên tục. Tuy nhiên, cũng đơn giản để phù hợp với mô hình hồi quy tuyến tính khi X là nhị phân, tức là nếu X có hai loại. Trong trường hợp này, các giá trị của X trong danh mục đầu tiên thường được mã hóa là 0 và các giá trị của X trong danh mục thứ hai được mã hóa là 1. Ví dụ: nếu biến nhị phân là "giới tính", chúng tôi thay thế từ "nam" với số 0 và từ “giống cái” với 1. Sau đó, chúng ta có thể phù hợp với mô hình hồi quy tuyến tính bằng cách sử dụng những con số này, nhưng cách diễn giải khác với cách diễn giải trong trường hợp biến liên tục. Nhắc lại định nghĩa của mô hình tuyến tính, Y = α + βX + e với E (e) = 0; nếu X = 0 (nam) thì E (Y | X = 0) = α và nếu X = 1 (nữ) thì E (Y | X = 1) = α + β · 1 = α + β. Do đó, α là giá trị trung bình của Y đối với nam, tức là E (Y | X = 0), và β = E (Y | X = 1) - E (Y | X = 0). Kết quả là những đối tượng có X = 1 (ví dụ: nữ) có giá trị Y trung bình cao hơn β đơn vị so với các đối tượng có X = 0 (ví dụ: nam).

11.5 Hồi quy tuyến tính với biến biến đổi của một mô hình tuyến tính hoàn toàn không liên quan đến tuyến tính trong hiệp biến

Chart, scatter chart

Description automatically generated

Nhớ lại rằng một mô hình được cho là tuyến tính khi nó tuyến tính trong các tham số của nó. Định nghĩa hàm trong β. không phải là một mô hình tuyến tính vì vế phải của phương trình không phải là một tuyến tính Tuy nhiên

là một mô hình tuyến tính. Mô hình này có thể được trang bị như cho bất kỳ mô hình tuyến tính nào khác: chúng tôi có được αˆ và βˆ như bình thường và chỉ cần sử dụng các giá trị bình phương của X thay vì X. Điều này có thể được chứng minh bằng cách xem xét X tăng đơn vị trong giá trị bình phương của X, tức là X thông thường thậm chí không cần thiết. Người ta có thể đơn giản vẽ đường hồi quy của Y trên X.

11.6 Hồi quy tuyến tính với nhiều hiệp biến

Cho đến nay, chúng tôi chỉ xem xét hai biến - Y và X. Do đó, chúng tôi giả định rằng Y chỉ bị ảnh hưởng bởi một hiệp biến X. Tuy nhiên, trong nhiều trường hợp, có thể có nhiều hơn một hiệp biến ảnh hưởng đến Y và do đó tất cả chúng đều liên quan đến phân tích (xem thêm Phần 11.10 về sự khác biệt giữa liên kết và nhân quả). Ví dụ, năng suất của một loại cây trồng phụ thuộc vào một số yếu tố như lượng phân bón, nước tưới và nhiệt độ, trong số những yếu tố khác. Trong một ví dụ khác, trong tập dữ liệu bánh pizza (được mô tả trong Phụ lục A.4), nhiều biến có thể liên quan đến thời gian giao hàng — người lái xe, lượng thức ăn đã đặt, người điều hành xử lý đơn hàng, cho dù đó là ngày cuối tuần, v.v. Một câu hỏi hợp lý cần đặt ra là: Các tài xế khác nhau có (trung bình) thời gian giao hàng khác nhau hay không — do họ giao dịch với cùng một nhà điều hành, số lượng thực phẩm được đặt như nhau, ngày giống nhau, v.v. Những câu hỏi kiểu này có thể được trả lời với nhiều hồi quy tuyến tính. Mô hình chứa nhiều hơn một, chẳng hạn p, hiệp biến X1, X2, ..., X Mô hình tuyến tính (11.2) có thể được mở rộng như



Lưu ý rằng thuật ngữ chặn ở đây được ký hiệu là β0. So với (11.2), α = β0 và β = β1.

11.6.1 Ký hiệu ma trận

Nếu chúng ta có tập dữ liệu gồm n quan sát, thì mọi tập hợp quan sát đều thỏa mãn mô hình (11.15) và chúng ta có thể viết

Text

Description automatically generated

Có thể viết n phương trình trong (11.16) dưới dạng ký hiệu ma trận là

Graphical user interface, text

Description automatically generated

11.6.2 Các biến số phân loại

Bây giờ chúng ta xem xét trường hợp khi các hiệp biến là phân loại thay vì liên tục.

11.6.3 Biến đổi

Như chúng tôi đã trình bày trong Sect. 11.5, nếu x được biến đổi bởi một hàm T (x) = (T (x1), T (x2), ... T (xn)) sau đó



vẫn là một mô hình tuyến tính vì mô hình là tuyến tính trong các tham số β của nó. Các dạng trans phổ biến T (x) là log (x), √x, exp (x), xp, trong số những dạng khác. Việc lựa chọn một chức năng như vậy thường được thúc đẩy bởi ứng dụng và dữ liệu trong tay. Ngoài ra, mối quan hệ giữa y và x có thể được mô hình hóa thông qua một đa thức như sau:

Chart, scatter chart

Description automatically generated

11.7 Cách nhìn quy nạp của hồi quy tuyến tính

Cho đến nay, chúng tôi đã giới thiệu hồi quy tuyến tính như một phương pháp mô tả mối quan hệ tương đối giữa các biến phụ thuộc và độc lập thông qua sự phù hợp nhất với dữ liệu. Tuy nhiên, như chúng tôi đã nhấn mạnh trong các chương trước, chúng tôi thường không chỉ quan tâm đến việc mô tả dữ liệu mà còn quan tâm đến việc đưa ra kết luận từ một mẫu về dân số quan tâm. Ví dụ: trong một ví dụ trước đó, chúng tôi đã ước tính sự liên kết giữa chi nhánh và thời gian giao hàng cho dữ liệu giao bánh pizza. Mô hình tuyến tính được ước tính là thời gian giao hàng = 36,3 - 5,2 Đông - 1,1 Tây; điều này cho thấy thời gian giao hàng trung bình ngắn hơn 5,2 phút đối với chi nhánh ở phía Đông thị trấn so với chi nhánh trung tâm và ngắn hơn 1,1 phút đối với chi nhánh ở phía Tây so với chi nhánh trung tâm. Khi xem xét tất cả các lần giao bánh pizza (dân số), không chỉ những lần giao hàng do chúng tôi thu thập trong khoảng thời gian một tháng (mẫu), chúng tôi có thể tự hỏi liệu 1,1 phút là chênh lệch thực sự hợp lệ cho toàn bộ dân số hay chỉ do lỗi ngẫu nhiên có liên quan trong mẫu chúng tôi đã chọn. Đặc biệt, chúng tôi cũng muốn biết khoảng tin cậy 95% cho ước tính tham số của chúng tôi là bao nhiêu

Khoảng thời gian có bao gồm "không" không? Nếu có, chúng tôi có thể kết luận rằng chúng tôi không thể chắc chắn rằng có một liên kết và không bác bỏ giả thuyết rỗng rằng βWest = 0. Liên quan đến Chaps. 9 và 10, bây giờ chúng ta áp dụng các khái niệm về suy luận và kiểm định thống kê trong bối cảnh phân tích hồi quy.

Ước lượng điểm và khoảng trong mô hình tuyến tính. Bây giờ chúng tôi viết lại mô hình cho

mục đích của việc giới thiệu suy luận thống kê đối với hồi quy tuyến tính như

A picture containing text

Description automatically generated.

11.7.1 Thuộc tính của Bình phương nhỏ nhất và Công cụ ước tính khả năng xảy ra tối đa

Công cụ ước lượng bình phương nhỏ nhất của β có một số thuộc tính:

1. Công cụ ước lượng bình phương nhỏ nhất của β là không chệch, E (βˆ) = β (xem Phụ lục C.6 cho chứng minh )
2. Ước lượng σˆ 2 như được giới thiệu trong phương trình (11.27) cũng không chệch, tức là E (σˆ 2) = σ2.
3. Ước lượng bình phương nhỏ nhất của β là nhất quán, tức là βˆ hội tụ với β khi n tiến tới vô cực
4. Công thức ước lượng bình phương nhỏ nhất của β là tiệm cận phân phối chuẩn.
5. Bộ ước lượng bình phương nhỏ nhất của β có phương sai nhỏ nhất trong số tất cả các bộ ước lượng tuyến tính và không chệch (công cụ ước lượng không chệch tuyến tính tốt nhất) của β

Chúng tôi không thảo luận chi tiết về chi tiết kỹ thuật của các thuộc tính này. Điều quan trọng hơn là phải biết rằng công cụ ước lượng bình phương nhỏ nhất có các tính năng tốt và đó là lý do tại sao chúng tôi chọn nó để phù hợp với mô hình. Vì chúng tôi sử dụng công cụ ước tính "tốt", chúng tôi hy vọng rằng mô hình cũng sẽ "tốt". Người ta có thể hỏi liệu có thể sử dụng một công cụ ước tính khác hay không. Chúng tôi đã đưa ra các giả định về phân phối trong mô hình: chúng tôi yêu cầu các lỗi phải được phân phối bình thường, với điều kiện là thực sự có thể áp dụng nguyên tắc khả năng xảy ra tối đa để có được các ước tính cho β và σ2 trong thiết lập của chúng tôi.

11.7.2 Bảng ANOVA

Một bảng thường được các gói phần mềm hiển thị khi thực hiện phân tích hồi quy là bảng phân tích phương sai (ANOVA). Bảng này có thể có một số cách hiểu và diễn giải trung bình và có thể trông hơi khác một chút tùy thuộc vào ngữ cảnh mà nó được sử dụng. Ở đây chúng tôi tập trung vào cách giải thích của nó i) như một cách để tóm tắt ảnh hưởng của các biến phân loại và ii) như một bài kiểm tra giả thuyết để so sánh k nghĩa. Điều này được minh họa trong ví dụ sau đây.

11.7.3 Tương tác

Có thể tác dụng chung của một số đồng biến ảnh hưởng đến phản ứng. Ví dụ 11.7.5 Xem xét lại dữ liệu bánh pizza được mô tả trong Phụ lục A.4. Chúng tôi có thể quan tâm đến việc liệu mối liên hệ giữa thời gian và nhiệt độ giao hàng có thay đổi tùy theo chi nhánh hay không. Ngoài thời gian và nhánh, do đó chúng tôi cần các biến tương tác bổ sung. Vì có 3 nhánh nên chúng ta cần 3 - 1 = 2 biến tương tác về cơ bản là tích của (i) thời gian và nhánh “Đông” và (ii) thời gian và nhánh “Tây”. Điều này có thể đạt được trong R bằng cách sử dụng toán tử “” (sẽ tạo ra cả hiệu ứng chính và tương hoặc tác) toán tử “:” (chỉ tạo ra thuật ngữ tương tác). Theo đó, đối với loại tham chiếu của x1, ảnh hưởng của x2 chỉ là βk (vì mỗi x2x1i bằng 0 vì mỗi x1 bằng 0). Tuy nhiên, hiệu βj ứng đề cho cập tất đến cả x1jx2. các hạng Do đó, mục mối khác liên là hệ β2 giữa + βj x2 trong và kết đó quả y khác nhau bởi βj giữa loại j và loại tham chiếu. Kiểm tra H0 : βj = 0 do đó giúp xác định liệu có tác động tương tác đối với loại j hay không. Ví dụ, nồng độ thuốc có thể có ảnh hưởng khác nhau đối với nam giới, phụ nữ và trẻ em; một loại phân bón có thể hoạt động khác nhau ở các khu vực địa lý khác nhau; hoặc một chương trình giáo dục mới có thể chỉ cho thấy lợi ích với một số giáo viên nhất định. Khá đơn giản để nhắm mục tiêu các câu hỏi như vậy trong phân tích hồi quy tuyến tính bằng cách sử dụng các tương tác. Các tions tương tác được đo lường dưới dạng tích của hai (hoặc nhiều) biến. Nếu một hoặc cả hai biến là phân loại, thì người ta chỉ cần xây dựng các sản phẩm cho mỗi biến giả, do đó tạo (k - 1) × (l - 1) biến mới khi xử lý hai biến phân loại (với k và l loại tương ứng). Các thuật ngữ sản phẩm này được gọi là các hành động giữa các hành động và ước tính sự khác nhau giữa một liên kết của một biến đối với các giá trị của biến kia. Bây giờ, chúng tôi đưa ra các ví dụ cho các tương tác liên tục – phân loại, phân loại – phân loại và liên tục – liên tục cho hai biến x1 và x2.

11.8 So sánh các mô hình khác nhau

Có nhiều tình huống trong đó nhiều mô hình tuyến tính khác nhau có thể được lắp vào một tập dữ liệu nhất định, nhưng không rõ đâu là mô hình tốt nhất. Điều này liên quan đến câu hỏi những biến nào nên được đưa vào một mô hình và những biến nào có thể bị loại bỏ.

11.9 Kiểm tra các giả định về mô hình

Các giả định được thực hiện cho mô hình tuyến tính chủ yếu liên quan đến các thuật ngữ sai số: e N (0, σ2I). Điều này ngụ ý một số điều: i) sự khác biệt giữa y và Xβ (đó là e) cần tuân theo phân phối chuẩn, ii) phương sai cho mỗi lỗi ei là hằng số σ2 và do đó không phụ thuộc vào i, và iii) không có mối tương quan giữa các điều khoản lỗi, Cov (ei, ei ) = 0 cho tất cả i, i . Chúng tôi cũng giả định rằng Xβ là đầy đủ theo nghĩa là chúng tôi đã bao gồm tất cả các biến và tương tác có liên quan trong mô hình và công cụ ước lượng βˆ ổn định và đầy đủ do các quan sát có ảnh hưởng eˆ hoặc các biến có tương quan cao. Bây giờ, chúng tôi trình bày cách kiểm tra các giả định i) Giả định chuẩn mực. Các lỗi e được giả định tuân theo một phân phối chuẩn.

Bản thân chúng ta không bao giờ biết lỗi. Khả năng duy nhất là ước tính các lỗi bằng các phần dư eˆ = y - Xβˆ. Để đảm bảo rằng phần còn lại ước tính là có thể so sánh với quy mô của phân phối chuẩn chuẩn , chúng có thể được tiêu chuẩn hóa

11.10 Nguyên nhân so với hiệp hội

Có sự khác biệt giữa liên kết và nhân quả: liên kết nói rằng giá trị cao hơn của X liên quan đến giá trị cao hơn của Y (hoặc ngược lại), trong khi quan hệ nhân quả nói rằng vì giá trị của X cao hơn nên giá trị của Y cũng cao hơn. Phân tích hồi quy, như được giới thiệu trong chương này, thiết lập mối liên hệ giữa Xi và Y trừ khi đưa ra các giả định mạnh mẽ. Ví dụ, nhớ lại ví dụ giao bánh pizza từ Phụ lục A.4. Chúng tôi đã thấy trong một số ví dụ và bài tập rằng thời gian giao hàng khác nhau tùy theo tài xế. Trên thực tế, trong một mô hình hồi quy tuyến tính đa bội trong đó Y là thời gian giao hàng và các hiệp biến là chi nhánh, hóa đơn, nhà điều hành, tài xế và nhiệt độ, chúng tôi nhận được sự khác biệt đáng kể về thời gian giao hàng trung bình của một số tài xế (ví dụ: Domenico và Bruno). Điều này có nghĩa là vì Domenico là người lái xe nên bánh pizza được giao nhanh hơn? Không cần thiết. Nếu Domenico lái chiếc xe tay ga nhanh nhất, thì đây có thể là nguyên nhân khiến anh ta tăng tốc độ. Tuy nhiên, chúng tôi chưa đo lường biến số "loại xe tay ga" và do đó không thể tính đến giá trị này. Kết quả mà chúng tôi thu được từ mô hình hồi quy vẫn hữu ích vì nó cung cấp cho người quản lý những giả thuyết và dự đoán hữu ích về thời gian giao hàng của anh ta, nhưng việc giải thích rằng khả năng của người lái xe gây ra thời gian giao hàng ngắn hơn có thể không nhất thiết là phù hợp. plot (fm, which = 3) ,không phải nhân quả, Ví dụ này nhấn mạnh rằng một trong những giả định quan trọng nhất để giải thích các tham số hồi quy theo cách nhân quả là đã đo lường (và sử dụng) tất cả các biến Xi có ảnh hưởng đến cả kết quả Y và biến quan tâm A (ví dụ: biến “trình điều khiển” ở trên) . Một giả định khác là mối quan hệ giữa tất cả Xi và Y được mô hình hóa một cách chính xác, ví dụ phi tuyến tính nếu thích hợp. Hơn nữa, chúng ta cần một số giả định tiên nghiệm về cách các biến liên quan với nhau và một số giả định kỹ thuật cũng cần được đáp ứng. Bạn đọc quan tâm có thể tham khảo Hernan and Robins (2017) để biết thêm chi tiết.

11.11

(a) Tính x¯ = Do đó, 66% độ biến thiên của dữ liệu có thể được giải thích bằng mô hình. Sự tốt lành αˆ = ¯y - βˆ x¯ = 160 - 4,57 · 25,5 = 43,465. (26 + 23 + 27 + 28 + 24 + 25) = 25,5 và y¯ = (170 + 150 + 160 + 175 + 155 + 150) = 160, cần có bảng sau  
=  
để ước lượng αˆ và βˆ:

