**7**

**Văn phạm biểu đồ trong R**

Yếu tố thị giác rất quan trọng. Quả thật, biểu đồ tốt có khả năng gây ấn tượng cho người đọc báo khoa học rất lớn, và thường có giá trị đại diện cho cả công trình nghiên cứu. Vì thế biểu đồ là một phương tiện hữu hiệu nhất để nhấn mạnh thông điệp của bài báo. Biểu đồ thường được sử dụng để thể hiện xu hướng và kết quả cho từng nhóm, nhưng cũng có thể dùng để trình bày dữ kiện một cách gọn gàng. Các biểu đồ dễ hiểu, nội dung phong phú là những phương tiện vô giá. Do đó, nhà nghiên cứu cần phải suy nghĩ một cách sáng tạo cách thể hiện số liệu quan trọng bằng biểu đồ. Vì thế, phân tích biểu đồ đóng một vai trò rất quan trọng trong phân tích thống kê. Có thể nói, không có đồ thị là phân tích dữ liệu không có nghĩa.

Trong ngôn ngữ R có rất nhiều cách để thiết kế một biểu đồ gọn và đẹp. Phần lớn những hàm để thiết kế biểu đồ có sẵn trong R, nhưng một số loại biểu đồ tinh vi và phức tạp khác có thể thiết kế bằng các package chuyên dụng như lattice và ggplopt2 có thể tải từ website của R. Chương này sẽ giới thiệu văn phạm biểu đồ và minh hoạ bằng một số biểu đồ cơ bản.

**7.1 Môi trường và thiết kế biểu đồ**

Chúng ta sẽ chỉ quan tâm đến các biểu đồ 2D, và không bàn đến các biểu đồ 3D. Các biểu đồ phổ biến bao gồm:

* biểu đồ phân bố (histogram);
* biểu đồ tương quan (scatter plot);
* biểu đồ hộp (box plot); và
* biểu đồ thanh (bar plot).

Một biểu đồ tiêu biểu thường có những yếu tố như sau:

* trục tung (y-axis) và trục hoành (x-axis);
* giới hạn của trục tung và trục hoành (limits);
* nhãn của trục tung và trục hoành (label);
* chú giải (legend);
* màu sắc; và
* đối tượng hình học.

Chúng ta bắt đầu bằng một dữ liệu thực tế được trích từ một nghiên cứu về trọng lượng trẻ sơ sinh của 189 bà mẹ. Mỗi trẻ có những thông tin (biến số) như sau:

* low: biến chỉ trọng lượng trẻ dưới 2.5 kg (mã số 1) hoặc trên 2.5 kg (mã số 0)
* age: tuổi của bà mẹ;
* lwt: trọng lượng của mẹ (đơn vị pound);
* race: sắc tộc, 1=trắng, 2=đen, 3=khác;
* smoke: bà mẹ hút thuốc lá (1) hay không hút thuốc lá (0) lúc mang thai;
* ptl: số lần sanh non trước đây;
* ht: tiền sử cao huyết áp
* ui: sản giật;
* ftv: số lần ghé phòng mạch bác sĩ để khám thai;
* bwt: trọng lượng trẻ khi sanh (gram).

Dữ liệu này có tên là "birthwt" trong package "MASS". Do đó, để truy cập dữ liệu, bạn đọc cần phải cài đặt MASS bằng hàm install.packages như hướng dẫn trong Chương 1. Các lệnh dưới đây truy cập dữ liệu và liệt kê 6 dòng đầu của dữ liệu:

library(MASS)

data(birthwt)

Chúng ta sẽ tạo ra một biến mới, gọi là cân nặng của trẻ và của mẹ bằng cách hoán chuyển từ đơn vị pound sang kg:

birthwt$**lwt** = birthwt$**lwt**\*0.453592

birthwt$**bwt** = birthwt$**bwt**\*0.453592

> head(birthwt)

low age lwt race smoke ptl ht ui ftv bwt

85 0 19 82.55374 2 0 0 0 1 0 1144.413

86 0 33 70.30676 3 0 0 0 0 3 1157.113

87 0 20 47.62716 1 1 0 0 0 1 1159.835

88 0 21 48.98794 1 1 0 0 1 2 1176.618

89 0 18 48.53434 1 1 0 0 1 0 1179.339

91 0 21 56.24541 3 0 0 0 0 0 1189.318

Để minh hoạ cho văn phạm biểu đồ trong R, chúng ta sẽ thiết kế một biểu đồ mô tả mối tương quan giữa cân nặng của mẹ và cân nặng của con. Trong R, có hàm plot có thể giúp chúng ta thiết kế biểu đồ này.

**Trục tung và trục hoành**

Việc đầu tiên là chúng ta phải xác định trục tung và trục hoành. Trong ví dụ trên, cân nặng của trẻ sẽ là trục tung (*y*) và cân nặng của mẹ là trục hoành (*x*). Hàm plot có cú pháp chung với hai cách viết như sau:

plot(y ~ x)

plot(x, y)

Trong hàm trên, với cách viết mô hình biến *y* xuất hiện trước, theo sau là biến *x* và cách nhau bằng dấu tilde. Trong hàm thứ hai, nếu dùng dấu phẩy thì *x* xuất hiện trước và *y* sau. Chúng ta sẽ chọn cách viết đầu tiên cho ví dụ:

plot(birthwt$**bwt** ~ birthwt$**lwt**)



**Nhãn cho trục tung và trục hoành (xlab, ylab)**

Như chúng ta thấy, nhãn của trục tung và trục hoành lấy từ tên của biến số. Do đó, chúng ta cần phải đặt nhãn (label) cho biểu đồ để dễ hiểu hơn. Đối số xlab và ylab giúp chúng ta thực hiện việc đó. Ngoài ra, có thể đặt tiêu đề cho biểu đồ qua đối số main:

plot(birthwt$bwt ~ birthwt$lwt,

**xlab**="Cân nặng của mẹ",

**ylab**="Cân nặng của con",

**main**="Mối tương quan giữa cân nặng của mẹ và con")  


**Giới hạn cho trục tung và trục hoành (xlim, ylim)**

Thỉnh thoảng, chúng ta có thể dùng đối số xlim và ylim để định nghĩa giới hạn của biến số:

plot(birthwt$bwt ~ birthwt$lwt,

**xlab**="Cân nặng của mẹ",

**ylab**="Cân nặng của con",

**main**="Mối tương quan giữa cân nặng của mẹ và con",

**xlim**=c(20, 120), **ylim**=c(200, 2500))



**Đối tượng hình học (pch)**

Trong biều đồ trên, đối tượng hình học là ô tròn. Trong thực tế, chúng ta có thể chọn những kí hiệu hình học khác theo ý muốn bằng đối số pch.



plot(birthwt$bwt ~ birthwt$lwt, **pch=16**,

**xlab**="Cân nặng của mẹ",

**ylab**="Cân nặng của con",

**main**="Mối tương quan giữa cân nặng của mẹ và con",

**xlim**=c(20, 120), **ylim**=c(200, 2500))



**Màu sắc (col)**

Biểu đồ trên có màu đen là màu mặc định. Tuy nhiên, chúng ta có thể thay đổi màu của biểu đồ bằng đối số col. R có hơn 3000 màu khác nhau, nhưng các cơ bản bao gồm white, blue, green, cyan, red, brown, yellow.

plot(birthwt$bwt ~ birthwt$lwt, pch=16, **col="blue",**

xlab="Cân nặng của mẹ",

ylab="Cân nặng của con",

main="Mối tương quan giữa cân nặng của mẹ và con",

xlim=c(20, 120), ylim=c(200, 2500))

****

**Đường biểu diễn (abline)**

Đối với biểu đồ tương quan giữa 2 biến *x* và *y,* một cách để thể hiện mối liên quan là dùng mô hình hồi qui tuyến tính. Mô hình này có 2 tham số là intercept (kí hiệu *a*) và slope (kí hiệu *b*). Do đó, trong R, có hàm abline để vẽ đường biểu diễn của một mối tương quan.

plot(birthwt$bwt ~ birthwt$lwt, pch=16, **col="blue",**

xlab="Cân nặng của mẹ",

ylab="Cân nặng của con",

main="Mối tương quan giữa cân nặng của mẹ và con",

xlim=c(20, 120), ylim=c(200, 2500))

**abline**(**lm**(birthwt$bwt ~ birthwt$lwt), col="red")

Trong hàm trên lm(birthwt$bwt ~ birthwt$lwt) có nghĩa là phân tích mô hình hồi qui tuyến tính giữa cân nặng của con như là hàm số cân nặng của mẹ, và dùng tham số của mô hình để vẽ đường biểu diễn:



**Viết chữ trong biểu đồ**

Phầnlớn các biểu đồ không cung cấp phương tiện để viết chữ hay ghi chú trong biểu đồ, hay có cung cấp nhưng rất hạn chế. Trong R có hàn mtext() cho phép chúng ta đặt chữ viết hay giải thích bên cạnh hay trong biểu đồ.

Bắt đầu từ phía dưới của biểu đồ (side=1), chúng ta chuyển theo hướng kim đồng hồ đến cạnh số 4. Lệnh plot trong ví dụ sau đây không in tên của trục và tên của biểu đồ, nhưng chỉ cung cấp một cái khung. Trong ví dụ này, chúng ta sử dụng cex (character expansion) để kiểm soát kích thước của chữ viết. Theo mặc định thì cex=1, nhưng với cex=2, chữ viết sẽ có kích thước gấp hai lần kích thước mặc định. Lệnh text() cho phép chúng ta đặt chữ viết vào một vị trí cụ thể. Lệnh thứ nhất đặt chữ viết trong ngoặc kép và trung tâm tại x=15, y=4.3. Qua sử dụng adj, chúng ta còn có thể sắp xếp về phía trái (adj=0) sao cho tọa độ là điểm xuất phát của chữ viết.

plot(y, xlab=" ", ylab=" ", type="n")   
 mtext("Text on side 1, cex=1", side=1,cex=1)   
 mtext("Text on side 2, cex=1.2", side=2,cex=1.2)   
 mtext("Text on side 3, cex=1.5", side=3,cex=1.5)   
 mtext("Text on side 4, cex=2", side=4,cex=2)   
 text(15, 4.3, "text(15, 4.3)")   
 text(35, 3.5, adj=0, "text(35, 3.5), left aligned")   
 text(40, 5, adj=1, "text(40, 5), right aligned")



**7.2 Biểu đồ phân bố**

Biểu đồ phân bố (histogram) là một phương pháp rất hữu hiệu để mô tả sự phân bố của một biến số liên tục. Giả dụ như trong dữ liệu birthwt, chúng ta muốn mô tả phân của biến cân nặng trẻ sơ sinh, chúng ta có thể dùng hàm hist trong R như sau:



Kết quả là một biểu đồ phân bố với trục tung là số trẻ em, trục hoành là cân nặng (đơn vị gram), và những nhãn mang tính mặc định. Chúng ta có thể làm cho biểu đồ sống động hơn và dễ hiểu hơn bằng cách thêm màu và nhãn, và thêm số thanh qua đối số breaks:

hist(birthwt$bwt, col="blue",

border="yellow",

breaks=15,

xlab="Cân nặng (gram)",

ylab="Số trẻ em",

main="Phân bố cân nặng trẻ em")



Thỉnh thoảng, chúng ta có thể thêm đường biểu diễn phân bố qua hàm lines(density). Tuy nhiên, để vẽ đường biểu diễn, trục tung cần phải hoán chuyển sang đơn vị xác suất với đối số prob=T (thay vì số đếm như trong biểu đồ trên).

hist(birthwt$bwt, col="blue",

**prob=T**, breaks=15,

border="yellow",

xlab="Cân nặng (gram)",

ylab="Số trẻ em", ylim=c(0, 0.0011),

main="Phân bố cân nặng trẻ em")

lines(density(birthwt$bwt), col="red", lwd=3)



**7.3 Biểu đồ hộp**

Một cách khác để thể hiện phân bố của một biến liên tục là biểu đồ hộp. Đây là một sáng kiến tuyệt vời nhằm cung cấp những chỉ số thống kê trong một biểu đồ đơn giản và dễ hiểu. Một biểu đồ hộp tiêu biểu như sau:

A picture containing clock

Description automatically generated

Trục tung thể hiện giá trị của biến số. Nội dung bao gồm giá trị trung vị (bác phân vị 50%), bách phân vị 25% (Q1), bách phân vị 75% (Q3). Ngoài ra, còn có 2 hàng rào (fences) thể hiện bằng đường ngang phía trên và phía dưới của biểu đồ. Khoảng cách từ Q1 đến fence bằng 1.5\*IQR.

Chẳng hạn như chúng ta có thể mô tả biến số bwt bằng hàm summary hay biểu đồ hộp qua hàm boxplot như sau:

summary(birthwt$bwt)

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

321.6 1095.0 1350.3 1335.6 1581.7 2263.4

boxplot(birthwt$bwt)



Như có thể thấy qua biểu đồ trên, 25% có trọng lượng thấp hơn hoặc bằng 1095 g, 50% trẻ em có trọng lượng thấp hơn hoặc bằng 1350 g, và 75% trẻ em có cân nặng thấp hơn hoặc bằng 1582 g. Có một trẻ với cân nặng khá thấp, và có thể là một *outlier* – giá trị ngoại vi.

Chúng ta cũng có thể trang trí biểu đồ hộp bằng màu và nhãn:

boxplot(birthwt$bwt,

col="red",

border="purple",

ylab="gram")

****

Biểu đồ hộp cũng có thể sử dụng để so sánh phân bố của một biến số. Chẳng hạn như chúng ta muốn so sánh cân nặng của trẻ em ở những bà mẹ hút thuốc là và không hút thuốc lá. Biến hút thuốc lá trong dữ liệu là smoke với giá trị 0 và 1. Do đó, chúng ta cần tạo ra một biến mới để dễ hiểu hơn, tạm gọi là biến smoker:

birthwt$**smoker** [birthwt$smoke==1] = "Yes"

birthwt$**smoker** [birthwt$smoke==0] = "No"

Sau đó, chúng ta vẽ biểu đồ hộp so sánh cân nặng của trẻ giữa 2 nhóm bà mẹ:

boxplot(birthwt$**bwt** ~ birthwt$**smoker**)

Nhưng với một số 'trang trí' như màu và nhãn:

boxplot(birthwt$bwt ~ birthwt$smoker,

col=c("blue", "red"),

xlab="Hút thuốc lá",

ylab="Cân nặng (g)")

****

Biểu đồ trên cho thấy những bà mẹ hút thuốc lá sanh con với cân nặng thấp hơn những bà mẹ không thuốc lá. Tuy nhiên, chúng ta chưa biết sự khác biệt có ý nghĩa thống kê, vì chưa làm kiểm định t (sẽ bàn trong một chương sau).

**7.4 Biểu đồ thanh**

Biểu đồ thanh (barplot) là một cách thể hiện số đếm rất tốt. Một cách thể hiện số đếm là dùng hàm table. Chẳng hạn chúng ta muốn biết số bà mẹ theo từng nhóm sắc tộc:

> table(birthwt$race)

1 2 3

96 26 67

hàm trên cho thấy 96 người Mĩ da trắng, 26 da đen, và 67 người thuộc sắc tộc khác (chủ yếu là người Mễ Tây Cơ). Hoặc có thể tính phần trăm cho mỗi nhóm:

> prop.table(table(birthwt$race))

1 2 3

0.5079365 0.1375661 0.3544974

Với R, chúng ta có thể lưu trữ những kết quả trên để thể hiện bằng biểu đồ thanh. Hàm dưới đây lưu trữ tần số theo sắc tộc trong đối tượng temp, và dùng hàm barplot vẽ dữ liệu của temp:

temp = prop.table(table(birthwt$race))

barplot(temp)



Biểu đồ trên chỉ là một cách thể hiện kết quả của hàm table. Trục tung là tỉ lệ và trục hoành là 3 nhóm sắc tộc được mã hoá theo 1, 2, và 3. Chúng ta có thể trang trí cho biểu đồ dễ hiểu hơn bằng cách thêm đối số names.arg và col:

barplot(temp, names.arg = c("White", "Black", "Others"), col=c("blue", "black", "brown"))



Biểu đồ thanh cũng có thể sử dụng để thể hiện hai biến số (hay hơn) cùng một lúc. Ví dụ như chúng ta muốn biết số bà mẹ hút thuốc lá (biến smoker) cho từng sắc tộc (race), đầu tiên chúng ta sẽ 'đếm' với hàm table:

birthwt$**smoker** [birthwt$smoke==1] = "Yes"

birthwt$**smoker** [birthwt$smoke==0] = "No"

> temp = prop.table(table(birthwt$smoker, birthwt$race), 2)

> temp

1 2 3

No 0.3826087 0.1391304 0.4782609

Yes 0.7027027 0.1351351 0.1621622

Hàm trên tính tỉ lệ hút thuốc lá cho mỗi sắc tộc. Đối số "2" trong hàm prop.table có nghĩa là tính phần trăm theo cột; nếu là "1" sẽ tính phần trăm theo dòng. Bây giờ, chúng ta có thể thể hiện bằng biểu đồ thanh:

barplot(temp, beside=T,

names.arg = c("White", "Black", "Others"),

col=c("green", "red"))

legend("top",

c("Non-smokers","Smokers"),

fill = c("green","red"))



Trong biểu đồ trên, chúng ta dùng đối số beside=T. Nhưng nếu không có đối số đó, biểu đồ sẽ là một 'stack bar' như sau:

barplot(temp,

names.arg = c("White", "Black", "Others"),

col=c("green", "red"))

legend("top",

c("Non-smokers","Smokers"),

fill = c("green","red"))



**7.5. Nhiều biểu đồ cho một cửa sổ**

Thông thường, R vẽ một biểu đồ cho một cửa sổ. Nhưng chúng ta có thể vẽ nhiều biểu đồ trong một cửa sổ bằng cách sử dụng hàm par. Chẳng hạn như par(mfrow=c(1,2))có hiệu năng chia cửa sổ ra thành 1 dòng và hai cột, tức là chúng ta có thể trình bày hai biểu đồ kề cạnh bên nhau. Còn par(mfrow=c(2,3)) chia cửa sổ ra thành 2 dòng và 3 cột, tức chúng ta có thể trình bày 6 biểu đồ trong mộ cửa sổ. Sau khi đã vẽ xong, chúng ta có thể quay về với “chế độ” 1 cửa sổ bằng lệnh par(mfrow=c(1,1).

Ví dụ sau đây tạo ra một dữ liệu gồm hai biến *x* và *y* bằng phương pháp mô phỏng (tức số liệu hoàn toàn được tạo ra bằng R). Sau đó, chúng ta chia cửa sổ thành 2 dòng và 2 cột, và trình bày bốn loại biểu đồ từ dữ liệu được mô phỏng:

par(mfrow=c(2,2))

N = 200

x = runif(N, -4, 4)

y = sin(x) + 0.5\*rnorm(N)

plot(x,y, main="Scatter plot of y and x")

hist(x, main="Histogram of x")

boxplot(y, main="Box plot of y")

barplot(x, main="Bar chart of x")

par(mfrow=c(1,1))

|  |
| --- |
|  |
| Cách chia cửa sổ thành 2 dòng và 2 cột và trình bày 4 biểu đồ trong cùng một cửa số: par(mfrow=c(2,2)) |

**\*\*\***

Chương này giới thiệu những cú pháp cơ bản cho việc thiết kế các biểu đồ phổ biến. Như có thể thấy, chúng ta không dùng các package chuyên dụng để soạn các biểu đồ tương quan, phân bố, hay biểu đồ hộp, mà chỉ dùng những hàm có sẵn trong R cơ bản. Tuy nhiên, trong thực tế nhiều dữ liệu phức tạp đòi hỏi package chuyên dụng như ggplot2 để thiết kế các biểu đồ có chất lượng cao, và chúng ta sẽ học trong chương kế tiếp.