

Phân tích số liệu và tạo biểu đồ bằng



hướng dẫn thực hành

Mục lục

1	Lời nói đầu
2	Giới thiệu ngôn ngữ R
2.1	R là gì ?
2.2	Tải và cài đặt R vào máy tính
2.3	Package cho các phân tích đặc biệt
2.4	Khởi động và ngưng chạy R
2.5	“Văn phạm” ngôn ngữ R
2.6	Cách đặt tên trong R
2.7	Hỗ trợ trong R
2.8	Môi trường vận hành
3	Nhập dữ liệu
3.1	Nhập số liệu trực tiếp: <code>c()</code>
3.2	Nhập số liệu trực tiếp: <code>edit(data.frame())</code>
3.3	Nhập số liệu từ một textfile: <code>read.table()</code>
3.4	Nhập số liệu từ Excel: <code>read.csv</code>
3.5	Nhập số liệu từ SPSS: <code>read.spss</code>
3.6	Tìm thông tin cơ bản về dữ liệu
4	Biên tập dữ liệu
4.1	Kiểm tra số liệu trống không: <code>na.omit()</code>
4.2	Tách rời dữ liệu: <code>subset</code>
4.3	Chiết số liệu từ một <code>data.frame</code>
4.4	Nhập hai <code>data.frame</code> thành một: <code>merge</code>
4.5	Mã hóa số liệu (data coding)
4.5.1	Mã hóa bằng hàm <code>replace</code>
4.5.2	Đổi một biến liên tục thành biến rời rạc
4.6	Chia một biến liên tục thành nhóm: <code>cut</code>
4.7	Tập hợp số liệu bằng <code>cut2</code> (Hmisc)

5	Sử R cho các phép tính đơn giản và ma trận
5.1	Tính toán đơn giản
5.2	Số liệu về ngày tháng
5.3	Tạo dãy số bằng seq, rep và gl
5.4	Sử dụng R cho các phép tính ma trận
5.4.1	Chiết phân tử từ ma trận
5.4.2	Tính toán với ma trận
6	Tính toán xác suất và mô phỏng (simulation)
6.1	Tính toán đơn giản
6.1.1	Phép hoán vị (permutation)
6.1.2	Tổ hợp (combination)
6.2	Biến số ngẫu nhiên và hàm phân phối
6.3	Các hàm phân phối xác suất (probability distribution function)
6.3.1	Hàm phân phối nhị phân (Binomial distribution)
6.3.2	Hàm phân phối Poisson (Poisson distribution)
6.3.3	Hàm phân phối chuẩn (Normal distribution)
6.3.4	Hàm phân phối chuẩn hóa (Standardized Normal distribution)
6.3.5	Hàm phân phối t, F và χ^2
6.4.	Mô phỏng (simulation)
6.4.1	Mô phỏng phân phối nhị phân
6.4.2	Mô phỏng phân phối Poisson
6.4.3	Mô phỏng phân phối χ^2 , t, F, gamma, beta, Weibull, Cauchy
6.5	Chọn mẫu ngẫu nhiên (random sampling)
7	Kiểm định giả thiết thống kê và ý nghĩa trị số P
7.1	Trị số P
7.2	Giả thiết khoa học và phản nghiệm
7.3	Ý nghĩa của trị số P qua mô phỏng
7.4	Vấn đề logic của trị số P
7.5	Vấn đề kiểm định nhiều giả thiết (multiple tests of hypothesis)
8	Phân tích số liệu bằng biểu đồ
8.1	Môi trường và thiết kế biểu đồ
8.1.1	Nhiều biểu đồ cho một cửa sổ (windows)
8.1.2	Đặt tên cho trục tung và trục hoành
8.1.3	Cho giới hạn của trục tung và trục hoành
8.1.4	Thể loại và đường biểu diễn
8.1.5	Màu sắc, khung, và kí hiệu
8.1.6	Ghi chú (legend)
8.17	Viết chữ trong biểu đồ

8.2	Số liệu cho phân tích biểu đồ
8.3	Biểu đồ cho một biến số rời rạc (discrete variable): barplot
8.4.	Biểu đồ cho hai biến số rời rạc (discrete variable): barplot
8.5	Biểu đồ hình tròn
8.6	Biểu đồ cho một biến số liên tục: stripchart và hist
8.6.1	Stripchart
8.6.2	Histogram
8.6.3	Biểu đồ hộp (boxplot)
8.6.4	Biểu đồ thanh (barchart)
8.6.5	Biểu đồ điểm (dotchart)
8.7	Phân tích biểu đồ cho hai biến liên tục
8.7.1	Biểu đồ tán xạ (scatter plot)
8.8	Phân tích Biểu đồ cho nhiều biến: pairs
8.9	Một số biểu đồ “đa năng”
8.9.1	Biểu đồ tán xạ và hình hộp
8.9.2	Biểu đồ tán xạ với kích thước biến thứ ba
8.9.3	Biểu đồ thanh và xác suất tích lũy
8.9.4	Biểu đồ hình đồng hồ (clock plot)
8.9.5	Biểu đồ với sai số chuẩn (standard error)
8.9.6	Biểu đồ vòng (contour plot)
8.9.10	Biểu đồ với kí hiệu toán

9

9.0	Phân tích thống kê mô tả
9.1	Khái niệm về tổng thể (population) và mẫu (sample)
9.2	Thống kê mô tả: summary
9.3	Kiểm định xem một biến có phải phân phối chuẩn
9.4	Thống kê mô tả theo từng nhóm
9.4.1	Kiểm định t (t.test)
9.4.2	Kiểm định t một mẫu
9.5	Kiểm định t hai mẫu
9.6	So sánh phương sai (var.test)
9.7	Kiểm định Wilcoxon cho hai mẫu (wilcox.test)
9.8	Kiểm định t cho các biến số theo cặp (paired t-test, t.test)
9.9	Kiểm định Wilcoxon cho các biến số theo cặp (wilcox.test)
9.10	Tần số (frequency)
9.11	Kiểm định tỉ lệ (proportion test, prop.test, binom.test)
9.12	So sánh hai tỉ lệ (prop.test, binom.test)
9.12.1	So sánh nhiều tỉ lệ (prop.test, chisq.test)
9.12.2	Kiểm định Chi bình phương
	Kiểm định Fisher

10	Phân tích hồi qui tuyến tính (regression analysis)
10.1	Hệ số tương quan
10.1.1	Hệ số tương quan Pearson
10.1.2	Hệ số tương quan Spearman
10.1.3	Hệ số tương quan Kendall
10.2	Mô hình của hồi qui tuyến tính đơn giản
10.2.1	Vài dòng lí thuyết
10.2.2	Phân tích hồi qui tuyến tính đơn giản bằng R
10.2.3	Giả định của phân tích hồi qui tuyến tính
10.2.4	Mô hình tiên đoán
10.3	Mô hình hồi qui tuyến tính đa biến (multiple linear regression)
10.4	Phân tích hồi qui đa thức (Polynomial regression analysis)
10.5	Xây dựng mô hình tuyến tính từ nhiều biến
10.6	Xây dựng mô hình tuyến tính bằng Bayesian Model Average (BMA)
11	Phân tích phương sai (analysis of variance)
11.1	Phân tích phương sai đơn giản (one-way analysis of variance - ANOVA)
11.1.1	Mô hình phân tích phương sai
11.1.2	Phân tích phương sai đơn giản với R
11.2	So sánh nhiều nhóm (multiple comparisons) và điều chỉnh trị số p
11.2.1	So sánh nhiều nhóm bằng phương pháp Tukey
11.2.2	Phân tích bằng biểu đồ
11.3	Phân tích bằng phương pháp phi tham số
11.4	Phân tích phương sai hai chiều (two-way analysis of variance - ANOVA)
11.4.1	Phân tích phương sai hai chiều với R
11.5	Phân tích hiệp biến (analysis of covariance - ANCOVA)
11.5.1	Mô hình phân tích hiệp biến
11.5.2	Phân tích bằng R
11.6	Phân tích phương sai cho thí nghiệm gai thửa (factorial experiment)
11.7	Phân tích phương sai cho thí nghiệm hình vuông Latin (Latin square experiment)
11.8	Phân tích phương sai cho thí nghiệm giao chéo (cross-over experiment)
11.9	Phân tích phương sai cho thí nghiệm tái đo lường (repeated measure experiment)
12	Phân tích hồi qui logistic (logistic regression analysis)
12.1	Mô hình hồi qui logistic

12.2	Phân tích hồi qui logistic bằng R
12.3	Uớc tính xác suất bằng R
12.4	Phân tích hồi qui logistic từ số liệu giản lược bằng R
12.5	Phân tích hồi qui logistic đa biến và chọn mô hình
12.6	Chọn mô hình hồi qui logistic bằng Bayesian Model Average
12.7	Số liệu dùng cho phân tích
13	Phân tích biến cố (survival analysis)
13.1	Mô hình phân tích số liệu mang tính thời gian
13.2	Uớc tính Kaplan-Meier bằng R
13.3	So sánh hai hàm xác suất tích lũy: kiểm định log-rank (log-rank test)
13.4	Kiểm định log-rank bằng R
13.5	Mô hình Cox (hay Cox's proportional hazards model)
13.6	Xây dựng mô hình Cox bằng Bayesian Model Average (BMA)
14	Phân tích tổng hợp (meta-analysis)
14.1	Nhu cầu cho phân tích tổng hợp
14.2	Ảnh hưởng ngẫu nhiên và ảnh hưởng bất biến (Fixed-effects và Random-effects)
14.3	Qui trình của một phân tích tổng hợp
14.4	Phân tích tổng hợp ảnh hưởng bất biến cho một tiêu chí liên tục (Fixed-effects meta-analysis for a continuous outcome)
14.4.1	Phân tích tổng hợp bằng tính toán “thủ công”
14.4.2	Phân tích tổng hợp bằng R
14.5	Phân tích tổng hợp ảnh hưởng bất biến cho một tiêu chí nhị phân (Fixed-effects meta-analysis for a dichotomous outcome)
14.5.1	Mô hình phân tích
14.5.2	Phân tích bằng R
15	Uớc tính cỡ mẫu (estimation of sample size)
15.1	Khái niệm về “power”
15.2	Thử nghiệm giả thiết thống kê và chẩn đoán bệnh
15.3	Số liệu để ước tính cỡ mẫu
15.4	Uớc tính cỡ mẫu
15.4.1	Uớc tính cỡ mẫu cho một chỉ số trung bình
15.4.2	Uớc tính cỡ mẫu cho so sánh hai số trung bình
15.4.3	Uớc tính cỡ mẫu cho phân tích phương sai
15.4.4	Uớc tính cỡ mẫu cho ước tính một tỉ lệ
15.4.5	Uớc tính cỡ mẫu cho so sánh hai tỉ lệ
16	Phụ lục 1: Lập trình và viết hàm bằng ngôn ngữ R

- 17 Phụ lục 2: Một số lệnh thông dụng trong R**
- 18 Phụ lục 3: Thuật ngữ dùng trong sách**
- 19 Lời bạt (tài liệu tham khảo và đọc thêm)**

1

Lời nói đầu

Trái với quan điểm của nhiều người, thống kê là một bộ môn khoa học: *Khoa học thống kê* (Statistical Science). Các phương pháp phân tích dù dựa vào nền tảng của toán học và xác suất, nhưng đó chỉ là phần “kỹ thuật”, phần quan trọng hơn là thiết kế nghiên cứu và diễn dịch ý nghĩa dữ liệu. Người làm thống kê, do đó, không chỉ là người đơn thuần làm phân tích dữ liệu, mà phải là một nhà khoa học, một nhà suy nghĩ (“thinker”) về nghiên cứu khoa học. Chính vì thế, mà khoa học thống kê đóng một vai trò cực kì quan trọng, một vai trò không thể thiếu được trong các công trình nghiên cứu khoa học, nhất là khoa học thực nghiệm. Có thể nói rằng ngày nay, nếu không có thống kê thì các thử nghiệm gen với triệu triệu số liệu chỉ là những con số vô hồn, vô nghĩa.

Một công trình nghiên cứu khoa học, cho dù có tôn kém và quan trọng cỡ nào, nếu không được phân tích đúng phương pháp sẽ không có ý nghĩa khoa học gì cả. Chính vì thế mà ngày nay, chỉ cần nhìn qua tất cả các tạp san nghiên cứu khoa học trên thế giới, hầu như bất cứ bài báo y học nào cũng có phần “Statistical Analysis” (Phân tích thống kê), nơi mà tác giả phải mô tả cẩn thận phương pháp phân tích, tính toán như thế nào, và giải thích ngắn gọn tại sao sử dụng những phương pháp đó để hàm ý “bảo kê” hay tăng trọng lượng khoa học cho những phát biểu trong bài báo. Các tạp san y học có uy tín càng cao yêu cầu về phân tích thống kê càng nặng. Xin nhắc lại để nhấn mạnh: không có phần phân tích thống kê, bài báo không có ý nghĩa khoa học.

Một trong những phát triển quan trọng nhất trong khoa học thống kê là ứng dụng máy tính cho phân tích và tính toán thống kê. Có thể nói không ngoa rằng không có máy tính, khoa học thống kê vẫn chỉ là một khoa học buồn tẻ khô khan, với những công thức rắc rối mà thiếu tính ứng dụng vào thực tế. Máy tính đã giúp khoa học thống kê làm một cuộc cách mạng lớn nhất trong lịch sử của bộ môn: đó là đưa khoa học thống kê vào thực tế, giải quyết các vấn đề gai góc nhất và góp phần làm phát triển khoa học thực nghiệm.

Người viết còn nhớ hơn 20 năm về trước khi còn là một sinh viên theo học chương trình thạc sĩ thống kê ở Úc, một vị giáo sư khả kính kể một câu chuyện về nhà thống kê danh tiếng người Mĩ, Fred Mosteller, nhận được một hợp đồng nghiên cứu từ Bộ Quốc phòng Mĩ để cải tiến độ chính xác của vũ khí Mĩ vào thời Thế chiến thứ II, mà trong đó ông phải giải một bài toán thống kê gồm khoảng 30 thông số. Ông phải mướn 20 sinh viên sau đại học làm việc này: 10 sinh viên chỉ việc suốt ngày tính toán bằng tay; còn 10 sinh viên khác kiểm tra lại tính toán của 10 sinh viên kia. Công việc kéo dài gần một tháng trời. Ngày nay, với một máy tính cá nhân (personal computer) khiêm tốn, phân tích thống kê đó có thể giải trong vòng trên dưới 1 giây.

Nhưng nếu máy tính mà không có phần mềm thì máy tính cũng chỉ là một đồng sắt hay silicon “vô hồn” và vô dụng. Một phần mềm đã, đang và sẽ làm cách mạng thống kê là R. Phần mềm này được một số nhà nghiên cứu thống kê và khoa học trên thế giới phát triển và hoàn thiện trong khoảng 10 năm qua để sử dụng cho việc học tập, giảng dạy và nghiên cứu. Cuốn sách này sẽ giới thiệu bạn đọc cách sử dụng R cho phân tích thống kê và đồ thị.

Tại sao R? Trước đây, các phần mềm dùng cho phân tích thống kê đã được phát triển và khá thông dụng. Những phần mềm nổi tiếng từ thời “xa xưa” như MINITAB, BMD-P đến những phần mềm tương đối mới như STATISTICA, SPSS, SAS, STAT, v.v... thường rất đắt tiền (giá cho một đại học có khi lên đến hàng trăm ngàn đô-la hàng năm), một cá nhân hay thậm chí cho một đại học không khả năng mua. Nhưng R đã thay đổi tình trạng này, vì R hoàn toàn miễn phí. Trái với cảm nhận thông thường, miễn phí không có nghĩa là chất lượng kém. Thật vậy, chẳng những hoàn toàn miễn phí, R còn có khả năng làm tất cả (xin nói lại: tất cả), thậm chí còn hơn cả, những phân tích mà các phần mềm thương mại làm. R có thể tải xuống máy tính cá nhân của bất cứ cá nhân nào, bất cứ lúc nào, và bất cứ ở đâu trên thế giới. Chỉ vài phút cài đặt là R có thể đưa vào sử dụng. Chính vì thế mà đại đa số các đại học Tây phương và thế giới càng ngày càng chuyển sang sử dụng R cho học tập, nghiên cứu và giảng dạy. Trong xu hướng đó, cuốn sách này có một mục tiêu khiêm tốn là giới thiệu đến bạn đọc trong nước để kịp thời cập nhật hóa những phát triển về tính toán và phân tích thống kê trên thế giới.

Cuốn sách này được soạn chủ yếu cho sinh viên đại học và các nhà nghiên cứu khoa học, những người cần một phần mềm để học thống kê, để phân tích số liệu, hay vẽ đồ thị từ số liệu khoa học. Cuốn sách này không phải là sách giáo khoa về lý thuyết thống kê, hay nhằm chỉ bạn đọc cách làm phân tích thống kê, nhưng sẽ giúp bạn đọc làm phân tích thống kê hữu hiệu hơn và hào hứng hơn. Mục đích chính của tôi là cung cấp cho bạn đọc những kiến thức cơ bản về thống kê, và cách ứng dụng R cho giải quyết vấn đề, và qua đó làm nền tảng để bạn đọc tìm hiểu hay phát triển thêm R.

Tôi cho rằng, cũng như bất cứ ngành nghề nào, cách học phân tích thống kê hay nhất là tự mình làm phân tích. Vì thế, sách này được viết với rất nhiều ví dụ và dữ liệu thực. Bạn đọc có thể vừa đọc sách, vừa làm theo những chỉ dẫn trong sách (bằng cách gõ các lệnh vào máy tính) và sẽ thấy hào hứng hơn. Nếu bạn đọc đã có sẵn một dữ liệu nghiên cứu của chính mình thì việc học tập sẽ hữu hiệu hơn bằng cách ứng dụng ngay những phép tính trong sách. Đối với sinh viên, nếu chưa có số liệu sẵn, các bạn có thể dùng các phương pháp mô phỏng (simulation) để hiểu thống kê hơn.

Khoa học thống kê ở nước ta tương đối còn mới, cho nên một số thuật ngữ chưa được diễn dịch một cách thống nhất và hoàn chỉnh. Vì thế, bạn đọc sẽ thấy đây đó trong sách một vài thuật ngữ “lạ”, và trong trường hợp này, tôi cố gắng kèm theo thuật ngữ gốc

tiếng Anh để bạn đọc tham khảo. Ngoài ra, trong phần cuối của sách, tôi có liệt kê các thuật ngữ Anh – Việt đã được đề cập đến trong sách.

Tất cả các dữ liệu sử dụng trong sách này đều có thể tải từ internet xuống máy tính cá nhân, hay có thể truy nhập trực tiếp qua trang web: <http://www.ykhoa.net/R>.

Tôi hi vọng bạn đọc sẽ tìm thấy trong sách một vài thông tin bổ ích, một vài kĩ thuật hay phép tính có ích cho việc học tập, giảng dạy và nghiên cứu của mình. Nhưng có lẽ chẳng có cuốn sách nào hoàn thiện hay không có thiếu sót; thành ra, nếu bạn đọc phát hiện một sai sót trong sách, xin báo cho tôi biết qua điện thư t.nguyen@garvan.org.au hay rkguyen@gmail.com. Thành thật cảm ơn các bạn đọc trước.

Tôi muốn nhân dịp này cảm ơn Tiến sĩ Nguyễn Hoàng Dzũng thuộc khoa Hóa, Đại học Bách khoa Thành phố Hồ Chí Minh, người đã gợi ý và giúp đỡ tôi in cuốn sách này ở trong nước. Tôi cảm ơn Bác sĩ Nguyễn Đình Nguyên, người đã đọc một phần lớn bản thảo của cuốn sách, góp nhiều ý kiến thiết thực, và đã thiết kế bìa sách. Tôi cũng cảm ơn Nhà xuất bản Đại học Bách khoa Thành phố Hồ Chí Minh đã giúp tôi in cuốn sách này.

Bây giờ, tôi mời bạn đọc cùng đi với tôi một “hành trình thống kê” ngắn bằng R.

*Sydney, 31 Tháng Ba Năm 2006
Nguyễn Văn Tuấn*

2 Giới thiệu ngôn ngữ R

2.1 R là gì ?

Nói một cách ngắn gọn, R là một phần mềm sử dụng cho phân tích thống kê và đồ thị. Thật ra, về bản chất, R là ngôn ngữ máy tính đa năng, có thể sử dụng cho nhiều mục tiêu khác nhau, từ tính toán đơn giản, toán học giải trí (recreational mathematics), tính toán ma trận (matrix), đến các phân tích thống kê phức tạp. Vì là một ngôn ngữ, cho nên người ta có thể sử dụng R để phát triển thành các phần mềm chuyên môn cho một vấn đề tính toán cá biệt.

Hai người sáng tạo ra R là hai nhà thống kê học tên là Ross Ihaka và Robert Gentleman. Kể từ khi R ra đời, rất nhiều nhà nghiên cứu thống kê và toán học trên thế giới ủng hộ và tham gia vào việc phát triển R. Chủ trương của những người sáng tạo ra R là theo định hướng mở rộng (Open Access). Cũng một phần vì chủ trương này mà R hoàn toàn miễn phí. Bất cứ ai ở bất cứ nơi nào trên thế giới đều có thể truy nhập và tải toàn bộ mã nguồn của R về máy tính của mình để sử dụng. Cho đến nay, chỉ qua chưa đầy 5 năm phát triển, càng ngày càng có nhiều các nhà thống kê học, toán học, nghiên cứu trong mọi lĩnh vực đã chuyển sang sử dụng R để phân tích dữ liệu khoa học. Trên toàn cầu, đã có một mạng lưới gần một triệu người sử dụng R, và con số này đang tăng theo cấp số nhân. Có thể nói trong vòng 10 năm nữa, chúng ta sẽ không cần đến các phần mềm thống kê đắt tiền như SAS, SPSS hay Stata (các phần mềm này rất đắt tiền, có thể lên đến 100.000 USD một năm) để phân tích thống kê nữa, vì tất cả các phân tích đó có thể tiến hành bằng R.

Vì thế, những ai làm nghiên cứu khoa học, nhất là ở các nước còn nghèo khó như nước ta, cần phải học cách sử dụng R cho phân tích thống kê và đồ thị. Bài viết ngắn này sẽ hướng dẫn bạn đọc cách sử dụng R. Tôi giả định rằng bạn đọc không biết gì về R, nhưng tôi kì vọng bạn đọc biết qua về cách sử dụng máy tính.

2.2 Tải R xuống và cài đặt vào máy tính

Để sử dụng R, việc đầu tiên là chúng ta phải cài đặt R trong máy tính của mình. Để làm việc này, ta phải truy nhập vào mạng và vào website có tên là “Comprehensive R Archive Network” (CRAN) sau đây:

<http://cran.R-project.org>.

Tài liệu cần tải về, tùy theo phiên bản, nhưng thường có tên bắt đầu bằng mẫu tự R và số phiên bản (version). Chẳng hạn như phiên bản tôi sử dụng vào cuối năm 2005 là 2.2.1, nên tên của tài liệu cần tải là:

R-2.2.1-win32.zip

Tài liệu này khoảng 26 MB, và địa chỉ cụ thể để tải là:

<http://cran.r-project.org/bin/windows/base/R-2.2.1-win32.exe>

Tại website này, chúng ta có thể tìm thấy rất nhiều tài liệu chỉ dẫn cách sử dụng R, đủ trình độ, từ sơ đẳng đến cao cấp. Nếu chưa quen với tiếng Anh, tài liệu này của tôi có thể cung cấp những thông tin cần thiết để sử dụng mà không cần phải đọc các tài liệu khác.

Khi đã tải R xuống máy tính, bước kế tiếp là cài đặt (set-up) vào máy tính. Để làm việc này, chúng ta chỉ đơn giản nhấp chuột vào tài liệu trên và làm theo hướng dẫn cách cài đặt trên màn hình. Đây là một bước rất đơn giản, chỉ cần 1 phút là việc cài đặt R có thể hoàn tất.

2.3 Package cho các phân tích đặc biệt

R cung cấp cho chúng ta một “ngôn ngữ” máy tính và một số *function* để làm các phân tích căn bản và đơn giản. Nếu muốn làm những phân tích phức tạp hơn, chúng ta cần phải tải về máy tính một số *package* khác. *Package* là một phần mềm nhỏ được các nhà thống kê phát triển để giải quyết một vấn đề cụ thể, và có thể chạy trong hệ thống R. Chẳng hạn như để phân tích hồi qui tuyến tính, R có *function lm* để sử dụng cho mục đích này, nhưng để làm các phân tích sâu hơn và phức tạp hơn, chúng ta cần đến các *package* như **lme4**. Các *package* này cần phải được tải về máy tính và cài đặt.

Địa chỉ để tải các *package* vẫn là: <http://cran.r-project.org>, rồi bấm vào phần “Packages” xuất hiện bên trái của mục lục trang web. Một số *package* cần tải về máy tính để sử dụng cho các ví dụ trong sách này là:

Tên package	Chức năng
Trellis	Dùng để vẽ đồ thị và làm cho đồ thị đẹp hơn
lattice	Dùng để vẽ đồ thị và làm cho đồ thị đẹp hơn
Hmisc	Một số phương pháp mô hình dữ liệu của F. Harrell
Design	Một số mô hình thiết kế nghiên cứu của F. Harrell
Epi	Dùng cho các phân tích dịch tễ học
epitools	Một <i>package</i> khác chuyên cho các phân tích dịch tễ học
foreign	Dùng để nhập dữ liệu từ các phần mềm khác như SPSS, Stata, SAS, v.v...
Rmeta	Dùng cho phân tích tổng hợp (meta-analysis)
meta	Một <i>package</i> khác cho phân tích tổng hợp
survival	Chuyên dùng cho phân tích theo mô hình Cox (Cox's proportional hazard model)

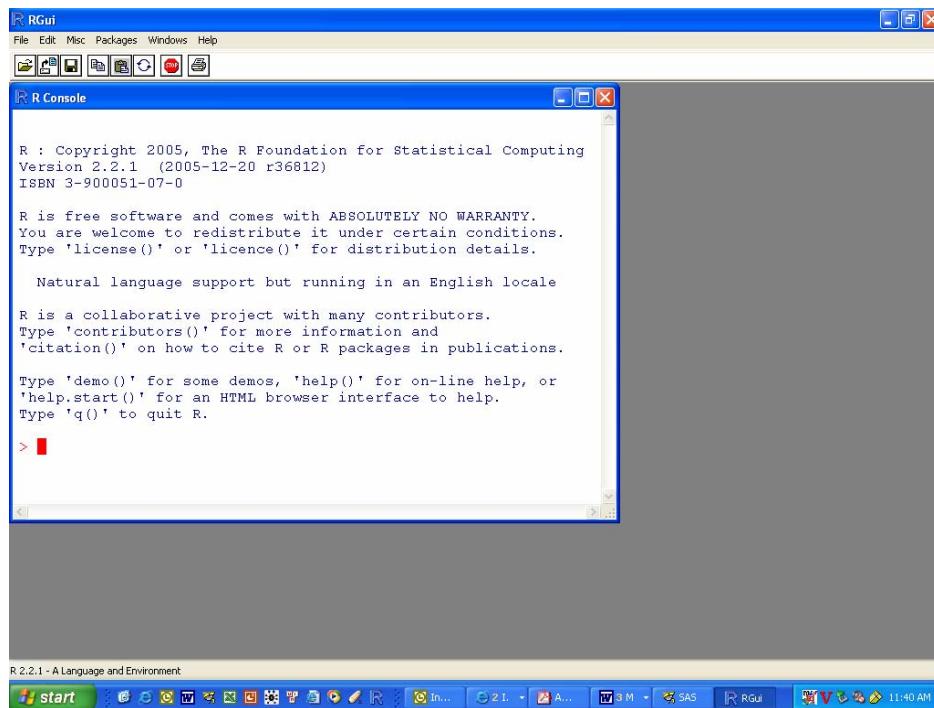
splines	Package cho survival vận hành
Zelig	Package dùng cho các phân tích thống kê trong lĩnh vực xã hội học
genetics	Package dùng cho phân tích số liệu di truyền học
BMA	Bayesian Model Average
leaps	Package dùng cho BMA

2.4 Khởi động và ngưng chạy R

Sau khi hoàn tất việc cài đặt, một *icon*



sẽ xuất hiện trên *desktop* của máy tính. Đến đây thì chúng ta đã sẵn sàng sử dụng R. Có thể nhấp chuột vào icon này và chúng ta sẽ có một *window* như sau:



R thường được sử dụng dưới dạng "*command line*", có nghĩa là chúng ta phải trực tiếp gõ lệnh vào cái *prompt* màu đỏ trên. Các lệnh phải tuân thủ nghiêm ngặt theo “văn phạm” và ngôn ngữ của R. Có thể nói toàn bộ bài viết này là nhằm hướng dẫn bạn đọc hiểu và viết theo ngôn ngữ của R. Một trong những văn phạm này là R phân biệt giữa Library và library. Nói cách khác, R phân biệt lệnh viết bằng chữ hoa hay chữ thường. Một văn phạm khác nữa là khi có hai chữ rời nhau, R thường dùng dấu chấm để

thay vào khoảng trống, chẳng hạn như `data.frame`, `t.test`, `read.table`, v.v... Điều này rất quan trọng, nếu không để ý sẽ làm mất thì giờ của người sử dụng.

Nếu lệnh gõ ra đúng “văn phạm” thì R sẽ cho chúng ta một cái prompt khác hay cho ra kết quả nào đó (tùy theo lệnh); nếu lệnh không đúng văn phạm thì R sẽ cho ra một thông báo ngắn là không đúng hay không hiểu. Ví dụ, nếu chúng ta gõ:

```
> x <- rnorm(20)
>
```

thì R sẽ hiểu và làm theo lệnh đó, rồi cho chúng ta một prompt khác: `>.`. Nhưng nếu chúng ta gõ:

```
> R is great
```

R sẽ không “đồng ý” với lệnh này, vì ngôn ngữ này không có trong thư viện của R, một thông báo sau đây sẽ xuất hiện:

```
Error: syntax error
>
```

Khi muốn rời khỏi R, chúng ta có thể đơn giản nhấn nút chéo (x) bên góc trái của window, hay gõ lệnh `q()`.

2.5 “Văn phạm” ngôn ngữ R

“Văn phạm” chung của R là một lệnh (command) hay function (tôi sẽ think thoáng đề cập đến là “hàm”). Mà đã là hàm thì phải có thông số; cho nên sau hàm là những thông số mà chúng ta phải cung cấp. Chẳng hạn như:

```
> reg <- lm(y ~ x)
```

thì `reg` là một object, còn `lm` là một hàm, và `y ~ x` là thông số của hàm. Hay:

```
> setwd("c:/works/stats")
```

thì `setwd` là một hàm, còn `"c:/works/stats"` là thông số của hàm.

Để biết một hàm cần có những thông số nào, chúng ta dùng lệnh `args(x)`, (`args` viết tắt chữ `arguments`) mà trong đó `x` là một hàm chúng ta cần biết:

```
> args(lm)
function (formula, data, subset, weights, na.action, method = "qr",
model = TRUE, x = FALSE, y = FALSE, qr = TRUE, singular.ok = TRUE,
contrasts = NULL, offset, ...)
```

NULL

R là một ngôn ngữ “đối tượng” (object oriented language). Điều này có nghĩa là các dữ liệu trong R được chứa trong object. Định hướng này cũng có vài ảnh hưởng đến cách viết của R. Chẳng hạn như thay vì viết `x = 5` như thông thường chúng ta vẫn viết, thì R yêu cầu viết là `x == 5`.

Đối với R, `x = 5` tương đương với `x <- 5`. Cách viết sau (dùng kí hiệu `<-`) được khuyến khích hơn là cách viết trước (`=`). Chẳng hạn như:

```
> x <- rnorm(10)
```

có nghĩa là mô phỏng 10 số liệu và chứa trong object x. Chúng ta cũng có thể viết `x = rnorm(10)`.

Một số kí hiệu hay dùng trong R là:

<code>x == 5</code>	x bằng 5
<code>x != 5</code>	x không bằng 5
<code>y < x</code>	y nhỏ hơn x
<code>x > y</code>	x lớn hơn y
<code>z <= 7</code>	z nhỏ hơn hoặc bằng 7
<code>p >= 1</code>	p lớn hơn hoặc bằng 1
<code>is.na(x)</code>	Có phải x là biến số missing
<code>A & B</code>	A và B (AND)
<code>A B</code>	A hoặc B (OR)
<code>!</code>	Không là (NOT)

Với R, tất cả các câu chữ hay lệnh sau kí hiệu # đều không có hiệu ứng, vì # là kí hiệu dành cho người sử dụng thêm vào các ghi chú, ví dụ:

```
> # lệnh sau đây sẽ mô phỏng 10 giá trị normal  
> x <- rnorm(10)
```

2.6 Cách đặt tên trong R

Đặt tên một đối tượng (object) hay một biến số (variable) trong R khá linh hoạt, vì R không có nhiều giới hạn như các phần mềm khác. Tên một object phải được viết liền nhau (tức không được cách rời bằng một khoảng trống). Chẳng hạn như R chấp nhận `myobject` nhưng không chấp nhận `my object`.

```
> myobject <- rnorm(10)  
> my object <- rnorm(10)  
Error: syntax error in "my object"
```

Nhưng đôi khi tên myobject khó đọc, cho nên chúng ta nên tác rời bằng “.” Như my.object.

```
> my.object <- rnorm(10)
```

Một điều quan trọng cần lưu ý là R phân biệt mẫu tự viết hoa và viết thường. Cho nên My.object khác với my.object. Ví dụ:

```
> My.object.u <- 15  
> my.object.L <- 5  
> My.object.u + my.object.L  
[1] 20
```

Một vài điều cần lưu ý khi đặt tên trong R là:

- Không nên đặt tên một biến số hay variable bằng kí hiệu “_” (underscore) như my_object hay my-object.
- Không nên đặt tên một object giống như một biến số trong một dữ liệu. Ví dụ, nếu chúng ta có một data.frame (dữ liệu hay dataset) với biến số age trong đó, thì không nên có một object trùng tên age, tức là không nên viết: age <- age. Tuy nhiên, nếu data.frame tên là data thì chúng ta có thể đề cập đến biến số age với một kí tự \$ như sau: data\$age. (Tức là biến số age trong data.frame data), và trong trường hợp đó, age <- data\$age có thể chấp nhận được.

2.7 Hỗ trợ trong R

Ngoài lệnh args() R còn cung cấp lệnh help() để người sử dụng có thể hiểu “văn phạm” của từng hàm. Chẳng hạn như muốn biết hàm lm có những thông số (arguments) nào, chúng ta chỉ đơn giản lệnh:

```
> help(lm)
```

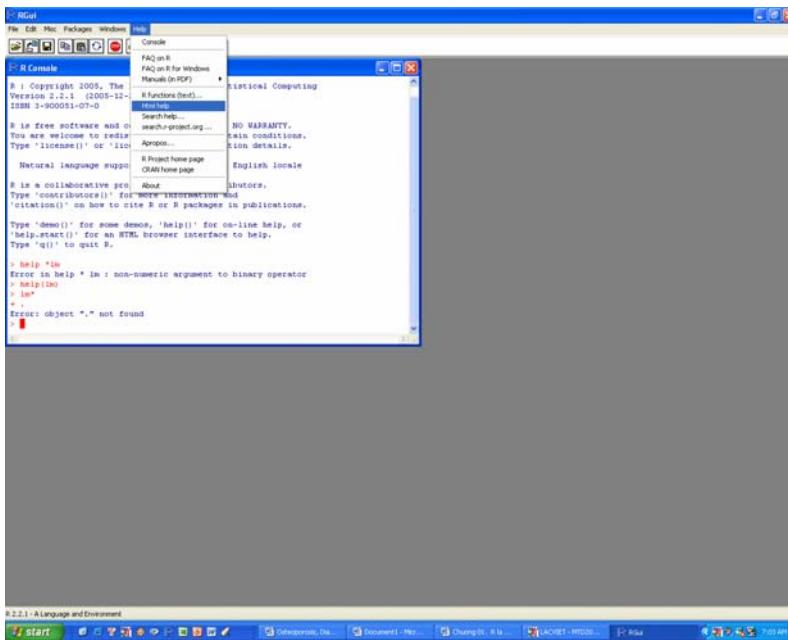
hay

```
> ?lm
```

Một cửa sổ sẽ hiện ra bên phải của màn hình chỉ rõ cách sử dụng ra sao và thậm chí có cả ví dụ. Bạn đọc có thể đơn giản copy và dán ví dụ vào R để xem cách vận hành.

Trước khi sử dụng R, ngoài sách này nếu cần bạn đọc có thể đọc qua phần chỉ dẫn có sẵn trong R bằng cách chọn mục help và sau đó chọn Html help như hình dưới

đây để biết thêm chi tiết. Bạn đọc cũng có thể copy và dán các lệnh trong mục này vào R để xem cho biết cách vận hành của R.



Thay vì chọn mục trên, bạn đọc cũng có thể đơn giản lệnh:

```
> help.start()
```

và một cửa sổ sẽ xuất hiện chỉ dẫn toàn bộ hệ thống R.

Hàm apropos cũng rất có ích vì nó cung cấp cho chúng ta tất cả các hàm trong R bắt đầu bằng kí tự mà chúng ta muốn tìm. Chẳng hạn như chúng ta muốn biết hàm nào trong R có kí tự "lm" thì chỉ đơn giản lệnh:

```
> apropos(lm)
```

Và R sẽ báo cáo các hàm với kí tự lm như sau có sẵn trong R:

```
[1] ".__C__anova.glm"      ".__C__anova.glm.null" ".__C__glm"
[4] ".__C__glm.null"       ".__C__lm"                 ".__C__mlm"
[7] "anova.glm"           "anova.glm.list"        "anova.lm"
[10] "anova.lm.list"        "anova.mlm"             "anovalist.lm"
[13] "contr.helmert"        "glm"                  "glm.control"
[16] "glm.fit"               "glm.fit.null"          "hatvalues.lm"
[19] "KalmanForecast"        "KalmanLike"            "KalmanRun"
[22] "KalmanSmooth"          "lm"                   "lm.fit"
[25] "lm.fit.null"           "lm.influence"         "lm.wfit"
[28]     "lm.wfit.null"        "model.frame.glm"
```

"model.frame.lm"

```
[31] "model.matrix.lm"          "nlm"                  "nlminb"
[34] "plot.lm"                  "plot.mlm"              "predict.glm"
[37] "predict.lm"               "predict.mlm"           "print.glm"
[40] "print.lm"                  "residuals.glm"         "residuals.lm"
[43] "rstandard.glm"            "rstandard.lm"           "rstudent.glm"
[46] "rstudent.lm"               "summary.glm"            "summary.lm"
[49] "summary.mlm"               "kappa.lm"
```

2.8 Môi trường vận hành

Dữ liệu phải được chứa trong một khu vực (directory) của máy tính. Trước khi sử dụng R, có lẽ cách hay nhất là tạo ra một directory để chứa dữ liệu, chẳng hạn như c:\works\stats. Để R biết dữ liệu nằm ở đâu, chúng ta sử dụng lệnh setwd (set working directory) như sau:

```
> setwd("c:/works/stats")
```

Lệnh trên báo cho R biết là dữ liệu sẽ chứa trong directory có tên là c:\works\stats. Chú ý rằng, R dùng forward slash “/” chứ không phải backward slash “\” như trong hệ thống Windows.

Để biết hiện nay, R đang “làm việc” ở directory nào, chúng ta chỉ cần lệnh:

```
> getwd()
[1] "C:/Program Files/R/R-2.2.1"
```

Cái prompt mặc định của R là “>”. Nhưng nếu chúng ta muốn có một prompt khác theo cá tính cá nhân, chúng ta có thể thay thế dễ dàng:

```
> options(prompt="R> ")
R>
```

Hay:

```
> options(prompt="Tuan> ")
Tuan>
```

Màn ảnh R mặc định là 80 characters, nhưng nếu chúng ta muốn màn ảnh rộng hơn, thì chỉ cần ra lệnh:

```
> options(width=100)
```

Hay muốn R trình bày các số liệu ở dạng 3 số thập phân:

```
> options(scipen=3)
```

Các lựa chọn và thay đổi này có thể dùng lệnh `options()`. Để biết các thông số hiện tại của R là gì, chúng ta chỉ cần lệnh:

```
> options()
```

Tìm hiểu ngày tháng:

```
> Sys.Date()  
[1] "2006-03-31"
```

Nếu bạn đọc cần thêm thông tin, một số tài liệu trên mạng (viết bằng tiếng Anh) cũng rất có ích. Các tài liệu này có thể tải xuống máy miễn phí:

R for beginners (của Emmanuel Paradis):

http://cran.r-project.org/doc/contrib/rdebuts_en.pdf

Using R for data analysis and graphics (của John Maindonald):

<http://cran.r-project.org/doc/contrib/usingR.pdf>

3

Nhập dữ liệu

Muốn làm phân tích dữ liệu bằng R, chúng ta phải có sẵn dữ liệu ở dạng mà R có thể hiểu được để xử lí. Dữ liệu mà R hiểu được phải là dữ liệu trong một `data.frame`. Có nhiều cách để nhập số liệu vào một `data.frame` trong R, từ nhập trực tiếp đến nhập từ các nguồn khác nhau. Sau đây là những cách thông dụng nhất:

3.1 Nhập số liệu trực tiếp: `c()`

Ví dụ 1: chúng ta có số liệu về độ tuổi và insulin cho 10 bệnh nhân như sau, và muốn nhập vào R.

50	16.5
62	10.8
60	32.3
40	19.3
48	14.2
47	11.3
57	15.5
70	15.8
48	16.2
67	11.2

Chúng ta có thể sử dụng function có tên `c` như sau:

```
> age <- c(50, 62, 60, 40, 48, 47, 57, 70, 48, 67)
> insulin <- c(16.5, 10.8, 32.3, 19.3, 14.2, 11.3, 15.5, 15.8, 16.2, 11.2)
```

Lệnh thứ nhất cho R biết rằng chúng ta muốn tạo ra một cột dữ liệu (từ nay tôi sẽ gọi là *biến số*, tức *variable*) có tên là `age`, và lệnh thứ hai là tạo ra một cột khác có tên là `insulin`. Tất nhiên, chúng ta có thể lấy một tên khác mà mình thích.

Chúng ta dùng function `c` (viết tắt của chữ *concatenation* – có nghĩa là “móc nối vào nhau”) để nhập dữ liệu. Chú ý rằng mỗi số liệu cho mỗi bệnh nhân được cách nhau bằng một dấu phẩy.

Kí hiệu `insulin <-` (cũng có thể viết là `insulin =`) có nghĩa là các số liệu sau sẽ có nằm trong biến số `insulin`. Chúng ta sẽ gặp kí hiệu này rất nhiều lần trong khi sử dụng R.

R là một ngôn ngữ cấu trúc theo dạng đối tượng (thuật ngữ chuyên môn là “*object-oriented language*”), vì mỗi cột số liệu hay mỗi một `data.frame` là một đối tượng (*object*) đối với R. Vì thế, `age` và `insulin` là hai đối tượng riêng lẻ. Bây giờ

chúng ta cần phải nhập hai đối tượng này thành một `data.frame` để R có thể xử lý sau này. Để làm việc này chúng ta cần đến function `data.frame`:

```
> tuan <- data.frame(age, insulin)
```

Trong lệnh này, chúng ta muốn cho R biết rằng nhập hai cột (hay hai đối tượng) `age` và `insulin` vào một đối tượng có tên là `tuan`.

Đến đây thì chúng ta đã có một đối tượng hoàn chỉnh để tiến hành phân tích thống kê. Để kiểm tra xem trong `tuan` có gì, chúng ta chỉ cần đơn giản gõ:

```
> tuan
```

Và R sẽ báo cáo:

	age	insulin
1	50	16.5
2	62	10.8
3	60	32.3
4	40	19.3
5	48	14.2
6	47	11.3
7	57	15.5
8	70	15.8
9	48	16.2
10	67	11.2

Nếu chúng ta muốn lưu lại các số liệu này trong một file theo dạng R, chúng ta cần dùng lệnh `save`. Giả dụ như chúng ta muốn lưu số liệu trong directory có tên là “`c:\works\stats`”, chúng ta cần gõ như sau:

```
> setwd("c:/works/stats")
> save(tuan, file="tuan.rda")
```

Lệnh đầu tiên (`setwd` – chữ `wd` có nghĩa là *working directory*) cho R biết rằng chúng ta muốn lưu các số liệu trong directory có tên là “`c:\works\stats`”. Lưu ý rằng thông thường Windows dùng dấu backward slash “`/`”, nhưng trong R chúng ta dùng dấu forward slash “`/`”.

Lệnh thứ hai (`save`) cho R biết rằng các số liệu trong đối tượng `tuan` sẽ lưu trong file có tên là “`tuan.rda`”). Sau khi gõ xong hai lệnh trên, một file có tên `tuan.rda` sẽ có mặt trong directory đó.

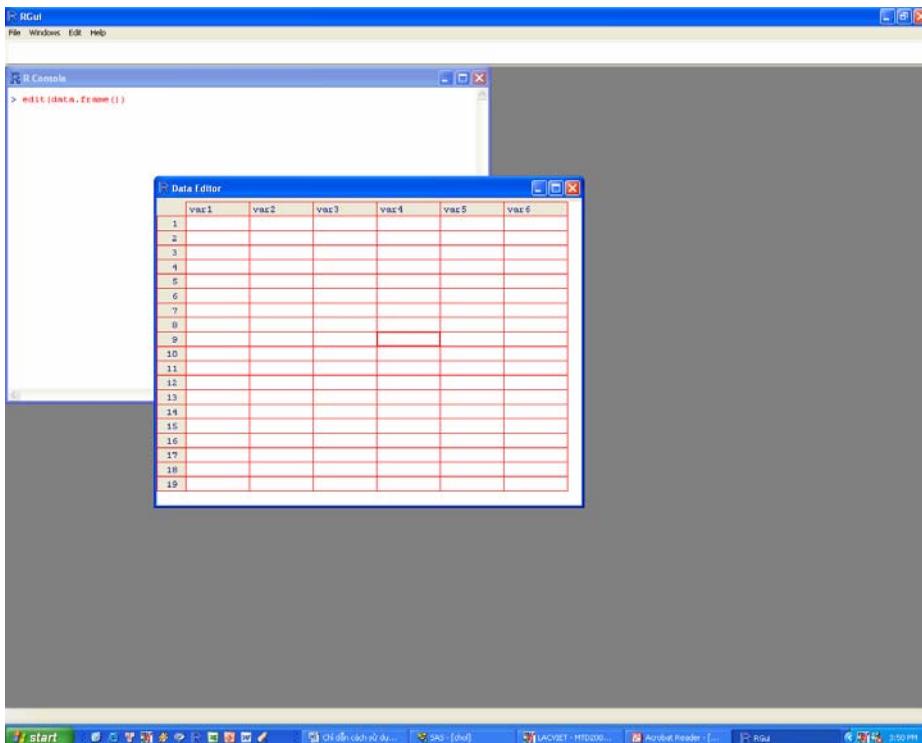
3.2 Nhập số liệu trực tiếp: `edit(data.frame())`

Ví dụ 1 (tiếp tục): chúng ta có thể nhập số liệu về độ tuổi và insulin cho 10 bệnh nhân bằng một function rất có ích, đó là: `edit(data.frame())`. Với function này,

R sẽ cung cấp cho chúng ta một window mới với một dãy cột và dòng giống như Excel, và chúng ta có thể nhập số liệu trong bảng đó. Ví dụ:

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

Chúng ta sẽ có một window như sau:



Ở đây, R không biết chúng ta có biến số nào, cho nên R liệt kê các biến số var1, var2, v.v... Nhập chuột vào cột var1 và thay đổi bằng cách gõ vào đó age. Nhập chuột vào cột var2 và thay đổi bằng cách gõ vào đó insulin. Sau đó gõ số liệu cho từng cột. Sau khi xong, bấm nút chéo X ở góc phải của spreadsheet, chúng ta sẽ có một data.frame tên ins với hai biến số age và insulin.

3.3 Nhập số liệu từ một **text file**: `read.table`

Ví dụ 2: Chúng ta thu thập số liệu về độ tuổi và cholesterol từ một nghiên cứu ở 50 bệnh nhân mắc bệnh cao huyết áp. Các số liệu này được lưu trong một text file có tên là chol.txt tại directory c :\works\stats. Số liệu này như sau: cột 1 là mã số của bệnh nhân, cột 2 là giới tính, cột 3 là body mass index (bmi), cột 4 là HDL cholesterol (viết tắt là hdl), kế đến là LDL cholesterol, total cholesterol (tc) và triglycerides (tg).

id	sex	age	bmi	hdl	ldl	tc	tg
1	Nam	57	17	5.000	2.0	4.0	1.1
2	Nu	64	18	4.380	3.0	3.5	2.1

3	Nu	60	18	3.360	3.0	4.7	0.8
4	Nam	65	18	5.920	4.0	7.7	1.1
5	Nam	47	18	6.250	2.1	5.0	2.1
6	Nu	65	18	4.150	3.0	4.2	1.5
7	Nam	76	19	0.737	3.0	5.9	2.6
8	Nam	61	19	7.170	3.0	6.1	1.5
9	Nam	59	19	6.942	3.0	5.9	5.4
10	Nu	57	19	5.000	2.0	4.0	1.9
11	Nu	63	20	4.217	5.0	6.2	1.7
12	Nam	51	20	4.823	1.3	4.1	1.0
13	Nu	60	20	3.750	1.2	3.0	1.6
14	Nam	42	20	1.904	0.7	4.0	1.1
15	Nam	64	20	6.900	4.0	6.9	1.5
16	Nu	49	20	0.633	4.1	5.7	1.0
17	Nu	44	21	5.530	4.3	5.7	2.7
18	Nu	45	21	6.625	4.0	5.3	3.9
19	Nu	80	21	5.960	4.3	7.1	3.0
20	Nu	48	21	3.800	4.0	3.8	3.1
21	Nu	61	21	5.375	3.1	4.3	2.2
22	Nu	45	21	3.360	3.0	4.8	2.7
23	Nu	70	21	5.000	1.7	4.0	1.1
24	Nu	51	21	2.608	2.0	3.0	0.7
25	Nam	63	22	4.130	2.1	3.1	1.0
26	Nam	54	22	5.000	4.0	5.3	1.7
27	Nu	57	22	6.235	4.1	5.3	2.9
28	Nam	70	22	3.600	4.0	5.4	2.5
29	Nu	47	22	5.625	4.2	4.5	6.2
30	Nu	60	22	5.360	4.2	5.9	1.3
31	Nu	60	22	6.580	4.4	5.6	3.3
32	Nam	50	22	7.545	4.3	8.3	3.0
33	Nam	60	22	6.440	2.3	5.8	1.0
34	Nu	55	22	6.170	6.0	7.6	1.4
35	Nu	74	23	5.270	3.0	5.8	2.5
36	Nam	48	23	3.220	3.0	3.1	0.7
37	Nu	46	23	5.400	2.6	5.4	2.4
38	Nam	49	23	6.300	4.4	6.3	2.4
39	Nu	69	23	9.110	4.3	8.2	1.4
40	Nu	72	23	7.750	4.0	6.2	2.7
41	Nam	51	23	6.200	3.0	6.2	2.4
42	Nu	58	23	7.050	4.1	6.7	3.3
43	nam	60	24	6.300	4.4	6.3	2.0
44	Nam	45	24	5.450	2.8	6.0	2.6
45	Nam	63	24	5.000	3.0	4.0	1.8
46	Nu	52	24	3.360	2.0	3.7	1.2
47	Nam	64	24	7.170	1.0	6.1	1.9
48	Nam	45	24	7.880	4.0	6.7	3.3
49	Nu	64	25	7.360	4.6	8.1	4.0
50	Nu	62	25	7.750	4.0	6.2	2.5

Chúng ta muốn nhập các dữ liệu này vào R để tiện việc phân tích sau này. Chúng ta sẽ sử dụng lệnh `read.table` như sau:

```
> setwd("c:/works/stats")
```

```
> chol <- read.table("chol.txt", header=TRUE)
```

Lệnh thứ nhất chúng ta muốn đảm bảo R truy nhập đúng directory mà số liệu đang được lưu giữ. Lệnh thứ hai yêu cầu R nhập số liệu từ file có tên là “chol.txt” (trong directory c:\works\stats) và cho vào đối tượng chol. Trong lệnh này, header=TRUE có nghĩa là yêu cầu R đọc dòng đầu tiên trong file đó như là tên của từng cột dữ kiện.

Chúng ta có thể kiểm tra xem R đã đọc hết các dữ liệu hay chưa bằng cách ra lệnh:

```
> chol
```

Hay

```
> names(chol)
```

R sẽ cho biết có các cột như sau trong dữ liệu (name là lệnh hỏi trong dữ liệu có những cột nào và tên gì):

```
[1] "id"   "sex"  "age"  "bmi"  "hdl"  "ldl"  "tc"   "tg"
```

Bây giờ chúng ta có thể lưu dữ liệu dưới dạng R để xử lí sau này bằng cách ra lệnh:

```
> save(chol, file="chol.rda")
```

3.4 Nhập số liệu từ Excel: `read.csv`

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

- Bước 1: Dùng lệnh “Save as” trong Excel và lưu số liệu dưới dạng “csv”;
- Bước 2: Dùng R (lệnh `read.csv`) để nhập dữ liệu dạng csv.

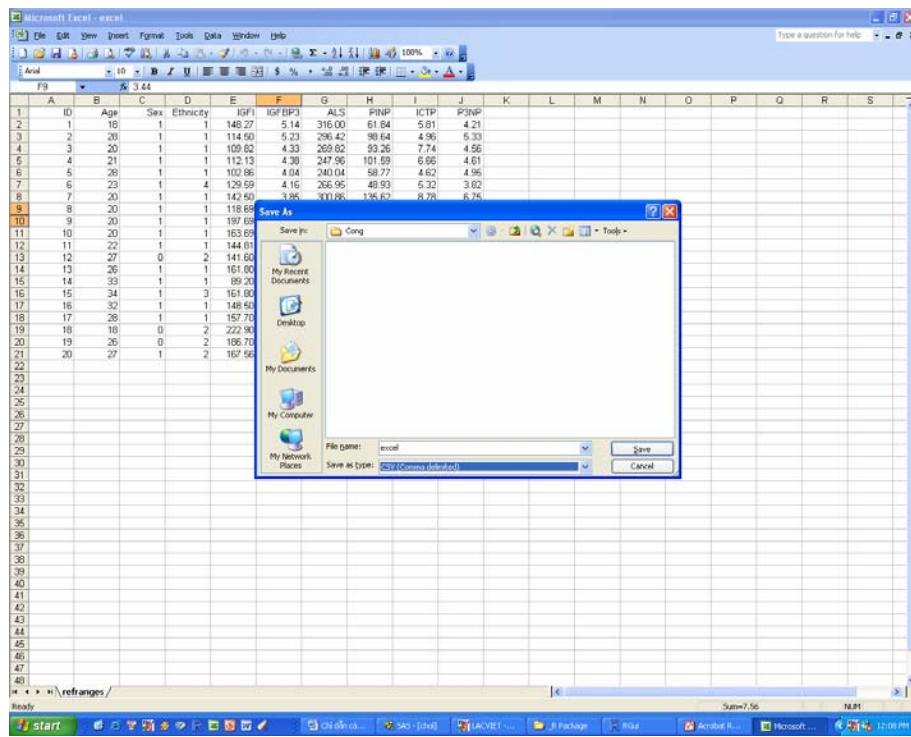
Ví dụ 3: Một dữ liệu gồm các cột sau đây đang được lưu trong Excel, và chúng ta muốn chuyển vào R để phân tích. Dữ liệu này có tên là `excel.xls`.

ID	Age	Sex	Ethnicity	IGFI	IGFBP3	ALS	PINP	ICTP	P3NP
1	18	1	1	148.27	5.14	316.00	61.84	5.81	4.21
2	28	1	1	114.50	5.23	296.42	98.64	4.96	5.33
3	20	1	1	109.82	4.33	269.82	93.26	7.74	4.56
4	21	1	1	112.13	4.38	247.96	101.59	6.66	4.61
5	28	1	1	102.86	4.04	240.04	58.77	4.62	4.95
6	23	1	4	129.59	4.16	266.95	48.93	5.32	3.82
7	20	1	1	142.50	3.85	300.86	135.62	8.78	6.75
8	20	1	1	118.69	3.44	277.46	79.51	7.19	5.11
9	20	1	1	197.69	4.12	335.23	57.25	6.21	4.44
10	20	1	1	163.69	3.96	306.83	74.03	4.95	4.84

11	22	1		1	144.81	3.63	295.46	68.26	4.54	3.70
12	27	0		2	141.60	3.48	231.20	56.78	4.47	4.07
13	26	1		1	161.80	4.10	244.80	75.75	6.27	5.26
14	33	1		1	89.20	2.82	177.20	48.57	3.58	3.68
15	34	1		3	161.80	3.80	243.60	50.68	3.52	3.35
16	32	1		1	148.50	3.72	234.80	83.98	4.85	3.80
17	28	1		1	157.70	3.98	224.80	60.42	4.89	4.09
18	18	0		2	222.90	3.98	281.40	74.17	6.43	5.84
19	26	0		2	186.70	4.64	340.80	38.05	5.12	5.77
20	27	1		2	167.56	3.56	321.12	30.18	4.78	6.12

Việc đầu tiên là chúng ta cần làm, như nói trên, là vào Excel để lưu dưới dạng csv:

- Vào Excel, chọn File → Save as
- Chọn Save as type “CSV (Comma delimited)”



Sau khi xong, chúng ta sẽ có một file với tên “excel.csv” trong directory “c:\works\stats”.

Việc thứ hai là vào R và ra những lệnh sau đây:

```
> setwd("c:/works/stats")
> gh <- read.csv ("excel.txt", header=TRUE)
```

Lệnh thứ hai `read.csv` yêu cầu R đọc số liệu từ “excel.csv”, dùng dòng thứ nhất là tên cột, và lưu các số liệu này trong một object có tên là `gh`.

Bây giờ chúng ta có thể lưu gh dưới dạng R để xử lí sau này bằng lệnh sau đây:

```
> save(gh, file="gh.rda")
```

3.5 Nhập số liệu từ một SPSS: `read.spss`

Phần mềm thống kê SPSS lưu dữ liệu dưới dạng “sav”. Chẳng hạn như nếu chúng ta đã có một dữ liệu có tên là `testo.sav` trong directory `C:\works\stats`, và muốn chuyển dữ liệu này sang dạng R có thể hiểu được, chúng ta cần sử dụng lệnh `read.spss` trong package có tên là `foreign`. Các lệnh sau đây sẽ hoàn tất dễ dàng việc này:

Việc đầu tiên chúng ta cho truy nhập `foreign` bằng lệnh `library`:

```
> library(foreign)
```

Việc thứ hai là lệnh `read.spss`:

```
> setwd("C:/works/stats")
> testo <- read.spss("testo.sav", to.data.frame=TRUE)
```

Lệnh thứ hai `read.spss` yêu cầu R đọc số liệu từ “`testo.sav`”, và cho vào một `data.frame` có tên là `testo`.

Bây giờ chúng ta có thể lưu `testo` dưới dạng R để xử lí sau này bằng lệnh sau đây:

```
> save(testo, file="testo.rda")
```

3.6 Thông tin cơ bản về dữ liệu

Giả dụ như chúng ta đã nhập số liệu vào một `data.frame` có tên là `chol` như trong ví dụ 1. Để tìm hiểu xem trong dữ liệu này có gì, chúng ta có thể nhập vào R như sau:

- Dẫn cho R biết chúng ta muốn xử lí `chol` bằng cách dùng lệnh `attach(arg)` với `arg` là tên của dữ liệu..

```
> attach(chol)
```

- Chúng ta có thể kiểm tra xem `chol` có phải là một `data.frame` không bằng lệnh `is.data.frame(arg)` với `arg` là tên của dữ liệu. Ví dụ:

```
> is.data.frame(chol)
[1] TRUE
```

R cho biết chol quả là một data.frame.

- Có bao nhiêu cột (hay *variable = biến số*) và dòng số liệu (*observations*) trong dữ liệu này? Chúng ta dùng lệnh `dim(arg)` với `arg` là tên của dữ liệu. (`dim` viết tắt chữ `dimension`). Ví dụ (kết quả của R trình bày ngay sau khi chúng ta gõ lệnh):

```
> dim(chol)
[1] 50   8
```

- Như vậy, chúng ta có 50 dòng và 8 cột (hay biến số). Vậy những biến số này tên gì? Chúng ta dùng lệnh `names(arg)` với `arg` là tên của dữ liệu. Ví dụ:

```
> names(chol)
[1] "id"  "sex" "age" "bmi" "hdl" "ldl" "tc"  "tg"
```

- Trong biến số `sex`, chúng ta có bao nhiêu nam và nữ? Để trả lời câu hỏi này, chúng ta có thể dùng lệnh `table(arg)` với `arg` là tên của biến số. Ví dụ:

```
> table(sex)
sex
nam Nam Nu
    1   21 28
```

Kết quả cho thấy dữ liệu này có 21 nam và 28 nữ.

4 Biên tập dữ liệu

Biên tập số liệu ở đây không có nghĩa là thay đổi số liệu gốc (vì đó là một tội lớn, một sự gian dối trong khoa học không thể chấp nhận được), mà chỉ có nghĩa tổ chức số liệu sao cho R có thể phân tích một cách hữu hiệu. Nhiều khi trong phân tích thống kê, chúng ta cần phải tập trung số liệu thành một nhóm, hay tách rời thành từng nhóm, hay thay thế từ kí tự (characters) sang số (numeric) cho tiện việc tính toán. Trong chương này, tôi sẽ bàn qua một số lệnh căn bản cho việc biên tập số liệu.

Chúng ta sẽ quay lại với dữ liệu chol trong ví dụ 1. Để tiện việc theo dõi và hiểu “câu chuyện”, tôi xin nhắc lại rằng chúng ta đã nhập số liệu vào trong một dữ liệu R có tên là chol từ một text file có tên là chol.txt:

```
> setwd("c:/works/stats")
> chol <- read.table("chol.txt", header=TRUE)
> attach(chol)
```

4.1 Kiểm tra số liệu trống không (missing value)

Trong nghiên cứu, vì nhiều lý do số liệu không thể thu thập được cho tất cả đối tượng, hay không thể đo lường tất cả biến số cho một đối tượng. Trong trường hợp đó, số liệu trống được xem là “missing value” (mà tôi tạm dịch là số liệu trống không). R xem các số liệu trống không là NA. Có một số kiểm định thống kê đòi hỏi các số liệu trống không phải được loại ra (vì không thể tính toán được) trước khi phân tích. R có một lệnh rất có ích cho việc này: na.omit, và cách sử dụng như sau:

```
> chol.new <- na.omit(chol)
```

Trong lệnh trên, chúng ta yêu cầu R loại bỏ các số liệu trống không trong data.frame chol và đưa các số liệu không trống vào data.frame mới tên là chol.new. Chú ý lệnh trên chỉ là ví dụ, vì trong dữ liệu chol không có số liệu trống không.

4.2 Tách rời dữ liệu: subset

Nếu chúng ta, vì một lý do nào đó, chỉ muốn phân tích riêng cho nam giới, chúng ta có thể tách chol ra thành hai data.frame, tạm gọi là nam và nu. Để làm chuyện này, chúng ta dùng lệnh subset (data, cond), trong đó data là data.frame mà chúng ta muốn tách rời, và cond là điều kiện. Ví dụ:

```
> nam <- subset(chol, sex=="Nam")
> nu <- subset(chol, sex=="Nu")
```

Sau khi ra hai lệnh này, chúng ta đã có 2 dữ liệu (hai data.frame) mới tên là nam và nu. Chú ý điều kiện sex == "Nam" và sex == "Nu" chúng ta dùng == thay vì = để chỉ điều kiện chính xác.

Tất nhiên, chúng ta cũng có thể tách dữ liệu thành nhiều data.frame khác nhau với những điều kiện dựa vào các biến số khác. Chẳng hạn như lệnh sau đây tạo ra một data.frame mới tên là old với những bệnh nhân trên 60 tuổi:

```
> old <- subset(chol, age>=60)
> dim(old)
[1] 25 8
```

Hay một data.frame mới với những bệnh nhân trên 60 tuổi và nam giới:

```
> n60 <- subset(chol, age>=60 & sex=="Nam")
> dim(n60)
[1] 9 8
```

4.3 Chiết số liệu từ một data .frame

Trong chol có 8 biến số. Chúng ta có thể chiết dữ liệu chol và chỉ giữ lại những biến số cần thiết như mã số (id), độ tuổi (age) và total cholesterol (tc). Để ý từ lệnh names(chol) rằng biến số id là cột số 1, age là cột số 3, và biến số tc là cột số 7. Chúng ta có thể dùng lệnh sau đây:

```
> data2 <- chol[, c(1,3,7)]
```

Ở đây, chúng ta lệnh cho R biết rằng chúng ta muốn chọn cột số 1, 3 và 7, và đưa tất cả số liệu của hai cột này vào data.frame mới có tên là data2. Chú ý chúng ta sử dụng ngoặc kép vuông [] chứ không phải ngoặc kép vòng (), vì chol không phải là một function. Dấu phẩy phía trước c, có nghĩa là chúng ta chọn tất cả các dòng số liệu trong data.frame chol.

Nhưng nếu chúng ta chỉ muốn chọn 10 dòng số liệu đầu tiên, thì lệnh sẽ là:

```
> data3 <- chol[1:10, c(1,3,7)]
> print(data3)
  id sex  tc
1   1  Nam 4.0
2   2   Nu 3.5
3   3   Nu 4.7
4   4  Nam 7.7
5   5  Nam 5.0
6   6   Nu 4.2
7   7  Nam 5.9
8   8  Nam 6.1
```

```
9   9 Nam 5.9  
10 10 Nu 4.0
```

Chú ý lệnh `print(arg)` đơn giản liệt kê tất cả số liệu trong `data.frame arg`. Thật ra, chúng ta chỉ cần đơn giản gõ `data3`, kết quả cũng giống y như `print(data3)`.

4.4 Nhập hai `data.frame` thành một: `merge`

Giả dụ như chúng ta có dữ liệu chứa trong hai `data.frame`. Dữ liệu thứ nhất tên là `d1` gồm 3 cột: `id`, `sex`, `tc` như sau:

```
id sex tc  
1 Nam 4.0  
2 Nu 3.5  
3 Nu 4.7  
4 Nam 7.7  
5 Nam 5.0  
6 Nu 4.2  
7 Nam 5.9  
8 Nam 6.1  
9 Nam 5.9  
10 Nu 4.0
```

Dữ liệu thứ hai tên là `d2` gồm 3 cột: `id`, `sex`, `tg` như sau:

```
id sex tg  
1 Nam 1.1  
2 Nu 2.1  
3 Nu 0.8  
4 Nam 1.1  
5 Nam 2.1  
6 Nu 1.5  
7 Nam 2.6  
8 Nam 1.5  
9 Nam 5.4  
10 Nu 1.9  
11 Nu 1.7
```

Hai dữ liệu này có chung hai biến số `id` và `sex`. Nhưng dữ liệu `d1` có 10 dòng, còn dữ liệu `d2` có 11 dòng. Chúng ta có thể nhập hai dữ liệu thành một `data.frame` bằng cách dùng lệnh `merge` như sau:

```
> d <- merge(d1, d2, by="id", all=TRUE)  
> d  
  id sex.x  tc sex.y  tg
```

1	1	Nam	4.0	Nam	1.1
2	2	Nu	3.5	Nu	2.1
3	3	Nu	4.7	Nu	0.8
4	4	Nam	7.7	Nam	1.1
5	5	Nam	5.0	Nam	2.1
6	6	Nu	4.2	Nu	1.5
7	7	Nam	5.9	Nam	2.6
8	8	Nam	6.1	Nam	1.5
9	9	Nam	5.9	Nam	5.4
10	10	Nu	4.0	Nu	1.9
11	11	<NA>	NA	Nu	1.7

Trong lệnh `merge`, chúng ta yêu cầu R nhập 2 dữ liệu `d1` và `d2` thành một và đưa vào `data.frame` mới tên là `d`, và dùng biến số `i_d` làm chuẩn. Chúng ta để ý thấy bệnh nhân số 11 không có số liệu cho `tc`, cho nên R cho là NA (một dạng “not available”).

4.5 Mã hóa số liệu (data coding)

Trong việc xử lí số liệu dịch tễ học, nhiều khi chúng ta cần phải biến đổi số liệu từ biến liên tục sang biến mang tính cách phân loại. Chẳng hạn như trong chẩn đoán loãng xương, những phụ nữ có chỉ số T của mật độ chất khoáng trong xương (bone mineral density hay BMD) bằng hay thấp hơn -2.5 được xem là “loãng xương”, những ai có BMD giữa -2.5 và -1.0 là “xốp xương” (osteopenia), và trên -1.0 là “bình thường”. Ví dụ, chúng ta có số liệu BMD từ 10 bệnh nhân như sau:

```
-0.92, 0.21, 0.17, -3.21, -1.80, -2.60, -2.00, 1.71, 2.12, -2.11
```

Để nhập các số liệu này vào R chúng ta có thể sử dụng *function* `c` như sau:

```
bmd <- c(-0.92, 0.21, 0.17, -3.21, -1.80, -2.60, -2.00, 1.71, 2.12, -2.11)
```

Để phân loại 3 nhóm loãng xương, xốp xương, và bình thường, chúng ta có thể dùng mã số 1, 2 và 3. Nói cách khác, chúng ta muốn tạo nên một biến số khác (hãy gọi là `diagnosis`) gồm 3 giá trị trên dựa vào giá trị của `bmd`. Để làm việc này, chúng ta sử dụng lệnh:

```
# tạm thời cho biến số diagnosis bằng bmd
> diagnosis <- bmd

# biến đổi bmd thành diagnosis
> diagnosis[bmd <= -2.5] <- 1
> diagnosis[bmd > -2.5 & bmd <= 1.0] <- 2
> diagnosis[bmd > -1.0] <- 3

# tạo thành một data frame
> data <- data.frame(bmd, diagnosis)

# liệt kê để kiểm tra xem lệnh có hiệu quả không
```

```

> data
      bmd diagnosis
1   -0.92      3
2    0.21      3
3    0.17      3
4   -3.21      1
5   -1.80      2
6   -2.60      1
7   -2.00      2
8    1.71      3
9    2.12      3
10  -2.11      2

```

4.5.1 Biến đổi số liệu bằng cách dùng *replace*

Một cách biến đổi số liệu khác là dùng *replace*, dù cách này có vẻ rườm rà chút ít. Tiếp tục ví dụ trên, chúng ta biến đổi từ *bmd* sang *diagnosis* như sau:

```

> diagnosis <- bmd
> diagnosis <- replace(diagnosis, bmd <= -2.5, 1)
> diagnosis <- replace(diagnosis, bmd > -2.5 & bmd <= 1.0, 2)
> diagnosis <- replace(diagnosis, bmd > -1.0, 3)

```

4.5.2 Biến đổi thành yếu tố (*factor*)

Trong phân tích thống kê, chúng ta phân biệt một biến số mang tính *yếu tố* (*factor*) và biến số liên tục bình thường. Biến số yếu tố không thể dùng để tính toán như cộng trừ nhân chia, nhưng biến số số học có thể sử dụng để tính toán. Chẳng hạn như trong ví dụ *bmd* và *diagnosis* trên, *diagnosis* là yếu tố vì giá trị trung bình giữa 1 và 2 chẳng có ý nghĩa thực tế gì cả; còn *bmd* là biến số số học.

Nhưng hiện nay, *diagnosis* được xem là một biến số số học. Để biến thành biến số yếu tố, chúng ta cần sử dụng *function factor* như sau:

```

> diag <- factor(diagnosis)
> diag
[1] 3 3 3 1 2 1 2 3 3 2
Levels: 1 2 3

```

Chú ý R bây giờ thông báo cho chúng ta biết *diag* có 3 bậc: 1, 2 và 3. Nếu chúng ta yêu cầu R tính số trung bình của *diag*, R sẽ không làm theo yêu cầu này, vì đó không phải là một biến số số học:

```

> mean(diag)
[1] NA
Warning message:
argument is not numeric or logical: returning NA in: mean.default(diag)

```

Dĩ nhiên, chúng ta có thể tính giá trị trung bình của *diagnosis*:

```
> mean(diagnosis)
[1] 2.3
```

nhưng kết quả 2.3 này không có ý nghĩa gì trong thực tế cả.

4.6 Chia nhóm bằng cut

Với một biến liên tục, chúng ta có thể chia thành nhiều nhóm bằng hàm `cut`. Ví dụ, chúng ta có biến `age` như sau:

```
> age <- c(17,19,22,43,14,8,12,19,20,51,8,12,27,31,44)
```

Độ tuổi thấp nhất là 8 và cao nhất là 51. Nếu chúng ta muốn chia thành 2 nhóm tuổi:

```
> cut(age, 2)
[1] (7.96,29.5] (7.96,29.5] (7.96,29.5] (29.5,51]      (7.96,29.5] (7.96,29.5]
(7.96,29.5] (7.96,29.5]
[9] (7.96,29.5] (29.5,51]      (7.96,29.5] (7.96,29.5] (7.96,29.5] (29.5,51]
(29.5,51]
Levels: (7.96,29.5] (29.5,51]
```

`cut` chia biến `age` thành 2 nhóm: nhóm 1 tuổi từ 7.96 đến 29.5; nhóm 2 từ 29.5 đến 51. Chúng ta có thể đếm số đối tượng trong từng nhóm tuổi bằng hàm `table` như sau:

```
> table(cut(age, 2))
(7.96,29.5] (29.5,51]
11             4

> ageg <- cut(age, 3, labels=c("low", "medium", "high"))
[1] low   low   low   high  low   low   low   low   high
low   low   medium medium
[15] high
Levels: low medium high

> ageg <- cut(age, 3, labels=c("low", "medium", "high"))
> table(ageg)
ageg
  low medium   high
    10      2       3
```

Tất nhiên, chúng ta cũng có thể chia `age` thành 4 nhóm (quartiles) bằng cách cho những thông số 0, 0.25, 0.50 và 0.75 như sau:

```
cut(age,
  breaks=quantiles(age, c(0, 0.25, 0.50, 0.75, 1)),
  labels=c("q1", "q2", "q3", "q4"),
```

```

include.lowest=TRUE)

cut(age,
  breaks=quantiles(c(0, 0.25, 0.50, 0.75, 1)),
  labels=c("q1", "q2", "q3", "q4"),
  include.lowest=TRUE)

```

4.7. Tập hợp số liệu bằng cut2 (Hmisc)

Hàm `cut` trên chia biến số theo giá trị của biến, chứ không dựa vào số mẫu, cho nên số lượng mẫu trong từng nhóm không bằng nhau. Tuy nhiên, trong phân tích thống kê, có khi chúng ta cần phải phân chia một biến số liên tục thành nhiều nhóm dựa vào phân phối của biến số nhưng số mẫu bằng hay tương đương nhau. Chẳng hạn như đối với biến số `bmd` chúng ta có thể “cắt” dãy số thành 3 nhóm với số mẫu tương đương nhau bằng cách dùng function `cut2` (trong thư viện `Hmisc`) như sau:

```

> # nhập thư viện Hmisc để có thể dùng function cut2
> library(Hmisc)
> bmd <- c(-0.92,0.21,0.17,-3.21,-1.80,-2.60,-2.00,1.71,2.12,-2.11)
> # chia biến số bmd thành 2 nhóm và để trong đối tượng group
> group <- cut2(bmd, g=2)
> table(group)
group
[-3.21,-0.92)  [-0.92,  2.12]
                5                  5

```

Như thấy qua ví dụ trên, `g = 2` có nghĩa là chia thành 2 nhóm (`g = group`). R tự động chia thành nhóm 1 gồm giá trị `bmd` từ -3.21 đến -0.92, và nhóm 2 từ -0.92 đến 2.12. Mỗi nhóm gồm có 5 số.

Tất nhiên, chúng ta cũng có thể chia thành 3 nhóm bằng lệnh:

```
> group <- cut2(bmd, g=3)
```

Và với lệnh `table` chúng ta sẽ biết có 3 nhóm, nhóm 1 gồm 4 số, nhóm 2 và 3 mỗi nhóm có 3 số:

```

> table(group)
group
[-3.21,-1.80)  [-1.80,  0.21)  [ 0.21,  2.12]
                4                  3                  3

```

5

Dùng R cho các phép tính đơn giản và ma trận

Một trong những lợi thế của R là có thể sử dụng như một ... máy tính cầm tay. Thật ra, hơn thế nữa, R có thể sử dụng cho các phép tính ma trận và lập chương. Trong chương này tôi chỉ trình bày một số phép tính đơn giản mà học sinh hay sinh viên có thể sử dụng lập tức trong khi đọc những dòng chữ này.

5.1 Tính toán đơn giản

Cộng hai số hay nhiều số với nhau: > 15+2997 [1] 3012	Cộng và trừ: > 15+2997-9768 [1] -6756
Nhân và chia > -27*12/21 [1] -15.42857	Số lũy thừa: $(25 - 5)^3$ > (25 - 5)^3 [1] 8000
Căn số bậc hai: $\sqrt{10}$ > sqrt(10) [1] 3.162278	Số pi (π) > pi [1] 3.141593 > 2+3*pi [1] 11.42478
Logarit: \log_e > log(10) [1] 2.302585	Logarit: \log_{10} > log10(100) [1] 2
Số mũ: $e^{2.7689}$ > exp(2.7689) [1] 15.94109 > log10(2+3*pi) [1] 1.057848	Hàm số lượng giác > cos(pi) [1] -1
Vector > x <- c(2,3,1,5,4,6,7,6,8) > x [1] 2 3 1 5 4 6 7 6 8 > sum(x) [1] 42 > x*2	> exp(x/10) [1] 1.221403 1.349859 1.105171 1.648 1.491825 1.822119 2.013753 1.822119 [9] 2.225541 > exp(cos(x/10)) [1] 2.664634 2.599545 2.704736 2.405 2.511954 2.282647 2.148655 2.282647 [9] 2.007132

[1] 4 6 2 10 8 12 14 12 16	
Tính tổng bình phương (sum of squares): $1^2 + 2^2 + 3^2 + 4^2 + 5^2 = ?$ > x <- c(1, 2, 3, 4, 5) > sum(x^2) [1] 55	Tính tổng bình phương điều chỉnh (adjusted sum of squares): $\sum_{i=1}^n (x_i - \bar{x})^2 = ?$ > x <- c(1, 2, 3, 4, 5) > sum((x-mean(x))^2) [1] 10 Trong công thức trên mean(x) là số trung bình của vector x.
Tính sai số bình phương (mean square): $\sum_{i=1}^n (x_i - \bar{x})^2 / n = ?$ > x <- c(1, 2, 3, 4, 5) > sum((x-mean(x))^2) / length(x) [1] 2 Trong công thức trên, length(x) có nghĩa là tổng số phần tử (elements) trong vector x.	Tính phương sai (variance) và độ lệch chuẩn (standard deviation): Phương sai: $s^2 = \sum_{i=1}^n (x_i - \bar{x})^2 / (n-1) = ?$ > x <- c(1, 2, 3, 4, 5) > var(x) [1] 2.5 Độ lệch chuẩn: $\sqrt{s^2}$: > sd(x) [1] 1.581139

5.2 Số liệu về ngày tháng

Trong phân tích thống kê, các số liệu ngày tháng có khi là một vấn đề nan giải, vì có rất nhiều cách để mô tả các dữ liệu này. Chẳng hạn như 01/02/2003, có khi người ta viết 1/2/2003, 01/02/03, 01FEB2003, 2003-02-01, v.v... Thật ra, có một qui luật chuẩn để viết số liệu ngày tháng là tiêu chuẩn ISO 8601 (nhưng rất ít ai tuân theo!) Theo qui luật này, chúng ta viết:

2003-02-01

Lí do đằng sau cách viết này là chúng ta viết số với đơn vị lớn nhất trước, rồi dần dần đến đơn vị nhỏ nhất. Chẳng hạn như với số “123” thì chúng ta biết ngay rằng “một trăm hai mươi ba”: bắt đầu là hàng trăm, rồi đến hàng chục, v.v... Và đó cũng là cách viết ngày tháng chuẩn của R.

```
> date1 <- as.Date("01/02/06", format="%d/%m/%y")
> date2 <- as.Date("06/03/01", format="%y/%m/%d")
```

Chú ý chúng ta nhập hai số liệu khác nhau về thứ tự ngày tháng năm, nhưng chúng ta cũng cho biết cụ thể cách đọc bằng %d (ngày), %m (tháng), và %y (năm). Chúng ta có thể tính số ngày giữa hai thời điểm:

```
> days <- date2-date1
> days
Time difference of 28 days
```

Chúng ta cũng có thể tạo một dãy số liệu ngày tháng như sau:

```
> seq(as.Date("2005-01-01"), as.Date("2005-12-31"), by="month")
[1] "2005-01-01" "2005-02-01" "2005-03-01" "2005-04-01" "2005-05-01"
[6] "2005-06-01" "2005-07-01" "2005-08-01" "2005-09-01" "2005-10-01"
[11] "2005-11-01" "2005-12-01"

> seq(as.Date("2005-01-01"), as.Date("2005-12-31"), by="2 weeks")
[1] "2005-01-01" "2005-01-15" "2005-01-29" "2005-02-12" "2005-02-26"
[6] "2005-03-12" "2005-03-26" "2005-04-09" "2005-04-23" "2005-05-07"
[11] "2005-05-21" "2005-06-04" "2005-06-18" "2005-07-02" "2005-07-16"
[16] "2005-07-30" "2005-08-13" "2005-08-27" "2005-09-10" "2005-09-24"
[21] "2005-10-08" "2005-10-22" "2005-11-05" "2005-11-19" "2005-12-03"
[26] "2005-12-17" "2005-12-31"
```

5.3 Tạo dãy số bằng hàm seq, rep và gl

R còn có công dụng tạo ra những dãy số rất tiện cho việc mô phỏng và thiết kế thí nghiệm. Những hàm thông thường cho dãy số là seq (sequence), rep (repetition) và gl (generating levels):

Áp dụng seq

- Tạo ra một vector số từ 1 đến 12:

```
> x <- (1:12)
> x
[1] 1 2 3 4 5 6 7 8 9 10 11 12

> seq(12)
[1] 1 2 3 4 5 6 7 8 9 10 11 12
```

- Tạo ra một vector số từ 12 đến 5:

```
> x <- (12:5)
> x
[1] 12 11 10 9 8 7 6 5

> seq(12,7)
[1] 12 11 10 9 8 7
```

Công thức chung của hàm seq là seq(from, to, by=) hay seq(from, to, length.out=). Cách sử dụng sẽ được minh họa bằng vài ví dụ sau đây:

- Tạo ra một vector số từ 4 đến 6 với khoảng cách bằng 0.25:

```
> seq(4, 6, 0.25)
[1] 4.00 4.25 4.50 4.75 5.00 5.25 5.50 5.75 6.00
```

- Tạo ra một vector 10 số, với số nhỏ nhất là 2 và số lớn nhất là 15

```
> seq(length=10, from=2, to=15)
[1] 2.000000 3.444444 4.888889 6.333333 7.777778 9.222222
10.666667 12.111111 13.555556 15.000000
```

Áp dụng rep

Công thức của hàm rep là `rep(x, times, ...)`, trong đó, `x` là một biến số và `times` là số lần lặp lại. Ví dụ:

- Tạo ra số 10, 3 lần:

```
> rep(10, 3)
[1] 10 10 10
```

- Tạo ra số 1 đến 4, 3 lần:

```
> rep(c(1:4), 3)
[1] 1 2 3 4 1 2 3 4 1 2 3 4
```

- Tạo ra số 1.2, 2.7, 4.8, 5 lần:

```
> rep(c(1.2, 2.7, 4.8), 5)
[1] 1.2 2.7 4.8 1.2 2.7 4.8 1.2 2.7 4.8 1.2 2.7 4.8 1.2 2.7 4.8
```

- Tạo ra số 1.2, 2.7, 4.8, 5 lần:

```
> rep(c(1.2, 2.7, 4.8), 5)
[1] 1.2 2.7 4.8 1.2 2.7 4.8 1.2 2.7 4.8 1.2 2.7 4.8 1.2 2.7 4.8
```

Áp dụng gl

`gl` được áp dụng để tạo ra một biến thứ bậc (categorical variable), tức biến không để tính toán, mà là đếm. Công thức chung của hàm `gl` là `gl(n, k, length = n*k, labels = 1:n, ordered = FALSE)` và cách sử dụng sẽ được minh họa bằng vài ví dụ sau đây:

- Tạo ra biến gồm bậc 1 và 2; mỗi bậc được lặp lại 8 lần:

```
> gl(2, 8)
[1] 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2
Levels: 1 2
```

Hay một biến gồm bậc 1, 2 và 3; mỗi bậc được lặp lại 5 lần:

```
> gl(3, 5)
[1] 1 1 1 1 1 2 2 2 2 2 3 3 3 3 3
Levels: 1 2 3
```

- Tạo ra biến gồm bậc 1 và 2; mỗi bậc được lặp lại 10 lần (do đó `length=20`):

```
> gl(2, 10, length=20)
```

```
[1] 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2  
Levels: 1 2
```

Hay:

```
> gl(2, 2, length=20)  
[1] 1 1 2 2 1 1 2 2 1 1 2 2 1 1 2 2 1 1 2 2  
Levels: 1 2
```

- Cho thêm kí hiệu:

```
> gl(2, 5, label=c("C", "T"))  
[1] C C C C C T T T T T  
Levels: C T
```

- Tạo một biến gồm 4 bậc 1, 2, 3, 4. Mỗi bậc lặp lại 2 lần.

```
> rep(1:4, c(2,2,2,2))  
[1] 1 1 2 2 3 3 4 4
```

Cũng tương đương với:

```
> rep(1:4, each = 2)  
[1] 1 1 2 2 3 3 4 4
```

- Với ngày giờ tháng:

```
> x <- .leap.seconds[1:3]  
> rep(x, 2)  
[1] "1972-06-30 17:00:00 Pacific Standard Time" "1972-12-31 16:00:00  
Pacific Standard Time"  
[3] "1973-12-31 16:00:00 Pacific Standard Time" "1972-06-30 17:00:00  
Pacific Standard Time"  
[5] "1972-12-31 16:00:00 Pacific Standard Time" "1973-12-31 16:00:00  
Pacific Standard Time"  
  
> rep(as.POSIXlt(x), rep(2, 3))  
[1] "1972-06-30 17:00:00 Pacific Standard Time" "1972-06-30 17:00:00  
Pacific Standard Time"  
[3] "1972-12-31 16:00:00 Pacific Standard Time" "1972-12-31 16:00:00  
Pacific Standard Time"  
[5] "1973-12-31 16:00:00 Pacific Standard Time" "1973-12-31 16:00:00  
Pacific Standard Time"
```

5.4 Sử dụng R cho các phép tính ma trận

Như chúng ta biết ma trận (matrix), nói đơn giản, gồm có dòng (row) và cột (column). Khi viết $A[m, n]$, chúng ta hiểu rằng ma trận A có m dòng và n cột. Trong R, chúng ta cũng có thể thể hiện như thế. Ví dụ: chúng ta muốn tạo một ma trận vuông A gồm 3 dòng và 3 cột, với các phần tử (element) 1, 2, 3, 4, 5, 6, 7, 8, 9, chúng ta viết:

$$A = \begin{pmatrix} 1 & 4 & 7 \\ 2 & 5 & 8 \\ 3 & 6 & 9 \end{pmatrix}$$

Và với R:

```
> y <- c(1,2,3,4,5,6,7,8,9)
> A <- matrix(y, nrow=3)
> A
[,1] [,2] [,3]
[1,]    1    4    7
[2,]    2    5    8
[3,]    3    6    9
```

Nhưng nếu chúng ta lệnh:

```
> A <- matrix(y, nrow=3, byrow=TRUE)
> A
```

thì kết quả sẽ là:

```
[,1] [,2] [,3]
[1,]    1    2    3
[2,]    4    5    6
[3,]    7    8    9
```

Tức là một **ma trận chuyển vị (transposed matrix)**. Một cách khác để tạo một ma trận hoán vị là dùng t(). Ví dụ:

```
> y <- c(1,2,3,4,5,6,7,8,9)
> A <- matrix(y, nrow=3)
> A
[,1] [,2] [,3]
[1,]    1    4    7
[2,]    2    5    8
[3,]    3    6    9
```

và $B = A'$ có thể diễn tả bằng R như sau:

```
> B <- t(A)
> B
[,1] [,2] [,3]
[1,]    1    2    3
[2,]    4    5    6
[3,]    7    8    9
```

Ma trận vô hướng (scalar matrix) là một ma trận vuông (tức số dòng bằng số cột), và tất cả các phần tử ngoài đường chéo (off-diagonal elements) là 0, và phần tử đường chéo là 1. Chúng ta có thể tạo một ma trận như thế bằng R như sau:

```
> # tạo ra một ma trận 3 x 3 với tất cả phần tử là 0.
> A <- matrix(0, 3, 3)

> # cho các phần tử đường chéo bằng 1
```

```

> diag(A) <- 1
> diag(A)
[1] 1 1 1

> # bây giờ ma trận A sẽ là:
> A
     [,1] [,2] [,3]
[1,]    1    0    0
[2,]    0    1    0
[3,]    0    0    1

```

5.4.1 Chiết phần tử từ ma trận

```

> y <- c(1,2,3,4,5,6,7,8,9)
> A <- matrix(y, nrow=3)
> A
     [,1] [,2] [,3]
[1,]    1    4    7
[2,]    2    5    8
[3,]    3    6    9

> # cột 1 của ma trận A
> A[,1]
[1] 1 4 7

> # cột 3 của ma trận A
> A[3,]
[1] 7 8 9

> # dòng 1 của ma trận A
> A[1,]
[1] 1 2 3

> # dòng 2, cột 3 của ma trận A
> A[2,3]
[1] 6

> # tất cả các dòng của ma trận A, ngoại trừ dòng 2
> A[-2,]
     [,1] [,2] [,3]
[1,]    1    4    7
[2,]    3    6    9

> # tất cả các cột của ma trận A, ngoại trừ cột 1
> A[,-1]
     [,1] [,2]
[1,]    4    7
[2,]    5    8
[3,]    6    9

```

```

> # xem phần tử nào cao hơn 3.
> A>3
 [,1] [,2] [,3]
[1,] FALSE TRUE TRUE
[2,] FALSE TRUE TRUE
[3,] FALSE TRUE TRUE

```

5.4.2 Tính toán với ma trận

Cộng và trừ hai ma trận. Cho hai ma trận A và B như sau:

```

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

> B <- matrix(-1:-12, 3, 4)
> B
 [,1] [,2] [,3] [,4]
[1,]   -1   -4   -7  -10
[2,]   -2   -5   -8  -11
[3,]   -3   -6   -9  -12

```

Chúng ta có thể cộng A+B:

```

> C <- A+B
> C
 [,1] [,2] [,3] [,4]
[1,]    0    0    0    0
[2,]    0    0    0    0
[3,]    0    0    0    0

```

Hay A-B:

```

> D <- A-B
> D
 [,1] [,2] [,3] [,4]
[1,]    2    8   14   20
[2,]    4   10   16   22
[3,]    6   12   18   24

```

Nhân hai ma trận. Cho hai ma trận:

$$A = \begin{pmatrix} 1 & 4 & 7 \\ 2 & 5 & 8 \\ 3 & 6 & 9 \end{pmatrix} \quad \text{và} \quad B = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix}$$

Chúng ta muốn tính AB , và có thể triển khai bằng R bằng cách sử dụng `%*%` như sau:

```
> y <- c(1,2,3,4,5,6,7,8,9)
> A <- matrix(y, nrow=3)
> B <- t(A)
> AB <- A %*% B
> AB
      [,1] [,2] [,3]
[1,]    66   78   90
[2,]    78   93  108
[3,]    90  108  126
```

Hay tính BA , và có thể triển khai bằng R bằng cách sử dụng `%*%` như sau:

```
> BA <- B %*% A
> BA
      [,1] [,2] [,3]
[1,]    14   32   50
[2,]    32   77  122
[3,]    50  122  194
```

Nghịch đảo ma trận và giải hệ phương trình. Ví dụ chúng ta có hệ phương trình sau đây:

$$\begin{aligned} 3x_1 + 4x_2 &= 4 \\ x_1 + 6x_2 &= 2 \end{aligned}$$

Hệ phương trình này có thể viết bằng kí hiệu ma trận: $AX = Y$, trong đó:

$$A = \begin{pmatrix} 3 & 4 \\ 1 & 6 \end{pmatrix}, \quad X = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \quad \text{và} \quad Y = \begin{pmatrix} 4 \\ 2 \end{pmatrix}$$

Nghiệm của hệ phương trình này là: $X = A^{-1}Y$, hay trong R:

```
> A <- matrix(c(3,1,4,6), nrow=2)
> Y <- matrix(c(4,2), nrow=2)
> X <- solve(A) %*% Y
> X
      [,1]
[1,] 1.1428571
[2,] 0.1428571
```

Chúng ta có thể kiểm tra:

```
> 3*X[1,1]+4*X[2,1]
[1] 4
```

Trị số eigen cũng có thể tính toán bằng function `eigen` như sau:

```
> eigen(A)
$values
[1] 7 2

$vectors
[,1]      [,2]
[1,] -0.7071068 -0.9701425
[2,] -0.7071068  0.2425356
```

Định thức (determinant). Làm sao chúng ta xác định một ma trận có thể đảo nghịch hay không? Ma trận mà định thức bằng 0 là **ma trận suy biến (singular matrix)** và không thể đảo nghịch. Để kiểm tra định thức, R dùng lệnh `det()`:

```
> E <- matrix((1:9), 3, 3)
> E
[,1] [,2] [,3]
[1,] 1 4 7
[2,] 2 5 8
[3,] 3 6 9
> det(E)
[1] 0
```

Nhưng ma trận F sau đây thì có thể đảo nghịch:

```
> F <- matrix((1:9)^2, 3, 3)
> F
[,1] [,2] [,3]
[1,] 1 16 49
[2,] 4 25 64
[3,] 9 36 81
> det(F)
[1] -216
```

Và nghịch đảo của ma trận F (F^{-1}) có thể tính bằng function `solve()` như sau:

```
> solve(F)
[,1]      [,2]      [,3]
[1,] 1.291667 -2.166667  0.9305556
[2,] -1.166667  1.666667 -0.6111111
[3,]  0.375000 -0.500000  0.1805556
```

Ngoài những phép tính đơn giản này, R còn có thể sử dụng cho các phép tính phức tạp khác. Một lợi thế đáng kể của R là phần mềm cung cấp cho người sử dụng tự do tạo ra những phép tính phù hợp cho từng vấn đề cụ thể. Trong vài chương sau, tôi sẽ quay lại vấn đề này chi tiết hơn.

R có một package `Matrix` chuyên thiết kế cho tính toán ma trận. Bạn đọc có thể tải package xuống, cài vào máy, và sử dụng, nếu cần. Địa chỉ để tải là:

http://cran.au.r-project.org/bin/windows/contrib/r-release/Matrix_0.995-8.zip

cùng với tài liệu chỉ dẫn cách sử dụng (dài khoảng 80 trang):

<http://cran.au.r-project.org/doc/packages/Matrix.pdf>

6

Tính toán xác suất và mô phỏng (simulation)

Xác suất là nền tảng của phân tích thống kê. Tất cả các phương pháp phân tích số liệu và suy luận thống kê đều dựa vào lí thuyết xác suất. Lí thuyết xác suất quan tâm đến việc mô tả và thể hiện qui luật phân phối của một biến số ngẫu nhiên. “Mô tả” ở đây trong thực tế cũng có nghĩa đơn giản là đếm những trường hợp hay khả năng xảy ra của một hay nhiều biến. Chẳng hạn như khi chúng ta chọn ngẫu nhiên 2 đối tượng, và nếu 2 đối tượng này có thể được phân loại bằng hai đặc tính như giới tính và sở thích, thì vấn đề đặt ra là có bao nhiêu tất cả “phối hợp” giữa hai đặc tính này. Hay đối với một biến số liên tục như huyết áp, mô tả có nghĩa là tính toán các chỉ số thống kê của biến như trị số trung bình, trung vị, phương sai, độ lệch chuẩn, v.v... Từ những chỉ số mô tả, lí thuyết xác suất cung cấp cho chúng ta những mô hình để thiết lập các hàm phân phối cho các biến số đó. Trong chương này, tôi sẽ bàn qua hai lĩnh vực chính là phép đếm và các hàm phân phối.

6.1 Các phép đếm

6.1.1 Phép hoán vị (permutation).

Theo định nghĩa, hoán vị n phần tử là cách sắp xếp n phần tử theo một thứ tự định sẵn. Định nghĩa này thật là ... khó hiểu, chẳng khác gì ... đó! Có lẽ một ví dụ cụ thể sẽ làm rõ định nghĩa hơn. Hãy tưởng tượng một trung tâm cấp cứu có 3 bác sĩ (x , y và z), và có 3 bệnh nhân (a , b và c) đang ngồi chờ được khám bệnh. Cả ba bác sĩ đều có thể khám bất cứ bệnh nhân a , b hay c . Câu hỏi đặt ra là có bao nhiêu cách sắp xếp bác sĩ – bệnh nhân? Để trả lời câu hỏi này, chúng ta xem xét vài trường hợp sau đây:

- Bác sĩ x có 3 lựa chọn: khám bệnh nhân a , b hoặc c ;
- Khi bác sĩ x đã chọn một bệnh nhân rồi, thì bác sĩ y có hai lựa chọn còn lại;
- Và sau cùng, khi 2 bác sĩ kia đã chọn, bác sĩ z chỉ còn 1 lựa chọn.
- Tổng cộng, chúng ta có 6 lựa chọn.

Một ví dụ khác, trong một buổi tiệc gồm 6 bạn, hỏi có bao nhiêu cách sắp xếp cách ngồi trong một bàn với 6 ghế? Qua cách lí giải của ví dụ trên, đáp số là: $6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 720$ cách. (Chú ý dấu “.” có nghĩa là dấu nhân hay tích số). Và đây chính là phép đếm *hoán vị*.

Chúng ta biết rằng $3! = 3 \cdot 2 \cdot 1 = 6$, và $0! = 1$. Nói chung, công thức tính hoán vị cho một số n là: $n! = n(n-1)(n-2)(n-3) \times \dots \times 1$. Trong R cách tính này rất đơn giản với lệnh `prod()` như sau:

- Tìm 3!

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

- Tìm 10!

```
> prod(10:1)
[1] 3628800
```

- Tìm 10.9.8.7.6.5.4

```
> prod(10:4)
[1] 604800
```

- Tìm $(10.9.8.7.6.5.4) / (40.39.38.37.36)$

```
> prod(10:4) / prod(40:36)
[1] 0.007659481
```

6.1.2 Tổ hợp (combination).

Tổ hợp n phần tử chập k là mọi tập hợp con gồm k phần tử của tập hợp n phần tử. Định nghĩa này phải nói là rất khó hiểu và ... rườm rà! Cách dễ hiểu nhất là qua một ví dụ như sau: Cho 3 người (hãy cho là A , B , và C) ứng viên vào 2 chức chủ tịch và phó chủ tịch, hỏi: có bao nhiêu cách để chọn 2 chức này trong số 3 người đó. Chúng ta có thể tưởng tượng có 2 ghế mà phải chọn 3 người:

Cách chọn	Chủ tịch	Phó chủ tịch
1	A	B
2	B	A
3	A	C
4	C	A
5	B	C
6	C	B

Như vậy có 6 cách chọn. Nhưng chú ý rằng cách chọn 1 và 2 trong thực tế chỉ là 1 cặp, và chúng ta chỉ có thể đếm là 1 (chứ không 2 được). Tương tự, 3 và 4, 5 và 6 cũng chỉ có thể đếm là 1 cặp. Tổng cộng, chúng ta có 3 cách chọn 3 người cho 2 chức vụ. Đáp số này được gọi là tổ hợp.

Thật ra tổng số lần chọn có thể tính bằng công thức sau đây:

$$\binom{3}{2} = \frac{3!}{2!(3-2)!} = \frac{6}{2} = 3 \text{ lần.}$$

Nói chung, số lần chọn k người từ n người là:

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

Công thức này cũng có khi viết là C_k^n thay vì $\binom{n}{k}$. Với R, phép tính này rất đơn giản bằng hàm `choose(n, k)`. Sau đây là vài ví dụ minh họa:

- Tìm $\binom{5}{2}$

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

- Tìm xác suất cặp A và B trong số 5 người được đắc cử vào hai chức vụ:

```
> 1/choose(5, 2)
[1] 0.1
```

6.2 Biến số ngẫu nhiên và hàm phân phối

Phần lớn phân tích thống kê dựa vào các luật phân phối xác suất để suy luận. Hai chữ “phân phối” (distribution) có lẽ cũng cần vài dòng giải thích ở đây. Nếu chúng ta chọn ngẫu nhiên 10 bạn trong một lớp học và ghi nhận chiều cao và giới tính của 10 bạn đó, chúng ta có thể có một dãy số liệu như sau:

	1	2	3	4	5	6	7	8	9	10
Giới tính	Nữ	Nữ	Nam	Nữ	Nữ	Nữ	Nam	Nam	Nữ	Nam
Chiều cao (cm)	156	160	175	145	165	158	170	167	178	155

Nếu tính gộp chung lại, chúng ta có 6 bạn gái và 4 bạn trai. Nói theo phần trăm, chúng ta có 60% nữ và 40% nam. Nói theo ngôn ngữ xác suất, xác suất nữ là 0.6 và nam là 0.4.

Về chiều cao, chúng ta có giá trị trung bình là 162.9 cm, với chiều cao thấp nhất là 155 cm và cao nhất là 178 cm.

Nói theo ngôn ngữ thống kê xác suất, biến số giới tính và chiều cao là hai *biến số ngẫu nhiên* (random variable). Ngẫu nhiên là vì chúng ta không đoán trước một cách chính xác các giá trị này, nhưng chỉ có thể đoán giá trị tập trung, giá trị trung bình, và độ dao động của chúng. Biến giới tính chỉ có hai “giá trị” (nam hay nữ), và được gọi là biến *không liên tục*, hay *biến rời rạc* (discrete variable), hay *biến thứ bậc* (categorical variable). Còn biến chiều cao có thể có bất cứ giá trị nào từ thấp đến cao, và do đó có tên là *biến liên tục* (continuous variable).

Khi nói đến “phân phối” (hay distribution) là đề cập đến các giá trị mà biến số có thể có. Các *hàm phân phối* (distribution function) là hàm nhằm mô tả các biến số đó một cách có hệ thống. “Có hệ thống” ở đây có nghĩa là theo mô hình toán học cụ thể với những thông số cho trước. Trong xác suất thống kê có khá nhiều hàm phân phối, và ở đây chúng ta sẽ xem xét qua một số hàm quan trọng nhất và thông dụng nhất: đó là phân

phối nhị phân, phân phối Poisson, và phân phối chuẩn. Trong mỗi luật phân phối, có 4 loại hàm quan trọng mà chúng ta cần biết:

- hàm mật độ xác suất (probability density distribution);
- hàm phân phối tích lũy (cumulative probability distribution);
- hàm định bậc (quantile); và
- hàm mô phỏng (simulation).

R có những hàm sẵn trên có thể ứng dụng cho tính toán xác suất. Tên mỗi hàm được gọi bằng một tiếp đầu ngữ để chỉ loại hàm phân phối, và viết tắt tên của hàm đó. Các tiếp đầu ngữ là `d` (chỉ distribution hay xác suất), `p` (chỉ cumulative probability, xác suất tích lũy), `q` (chỉ định bậc hay quantile), và `r` (chỉ random hay số ngẫu nhiên). Các tên viết tắt là `norm` (normal, phân phối chuẩn), `binom` (binomial , phân phối nhị phân), `pois` (Poisson, phân phối Poisson), v.v... Bảng sau đây tóm tắt các hàm và thông số cho từng hàm:

Hàm phân phối	Mật độ	Tích lũy	Định bậc	Mô phỏng
Chuẩn	<code>dnorm(x, mean, sd)</code>	<code>pnorm(q, mean, sd)</code>	<code>qnorm(p, mean, sd)</code>	<code>rnorm(n, mean, sd)</code>
Nhị phân	<code>dbinom(k, n, p)</code>	<code>pbinom(q, n, p)</code>	<code>qbinom(p, n, p)</code>	<code>rbinom(k, n, prob)</code>
Poisson	<code>dpois(k, lambda)</code>	<code>ppois(q, lambda)</code>	<code>qpois(p, lambda)</code>	<code>rpois(n, lambda)</code>
Uniform	<code>dunif(x, min, max)</code>	<code>runif(q, min, max)</code>	<code>qunif(p, min, max)</code>	<code>runif(n, min, max)</code>
Negative binomial	<code>dnbinom(x, k, p)</code>	<code>pnbinom(q, k, p)</code>	<code>qnbinom(p, k, prob)</code>	<code>rbinom(n, n, prob)</code>
Beta	<code>dbeta(x, shape1, shape2)</code>	<code>pbeta(q, shape1, shape2)</code>	<code>qbeta(p, shape1, shape2)</code>	<code>rbeta(n, shape1, shape2)</code>
Gamma	<code>dgamma(x, shape, rate, scale)</code>	<code>gamma(q, shape, rate, scale)</code>	<code>qgamma(p, shape, rate, scale)</code>	<code>rgamma(n, shape, rate, scale)</code>
Geometric	<code>dgeom(x, p)</code>	<code>pgeom(q, p)</code>	<code>qgeom(p, prob)</code>	<code>rgeom(n, prob)</code>
Exponential	<code>dexp(x, rate)</code>	<code>pexp(q, rate)</code>	<code>qexp(p, rate)</code>	<code>rexp(n, rate)</code>
Weibull	<code>dnorm(x, mean, sd)</code>	<code>pnorm(q, mean, sd)</code>	<code>qnorm(p, mean, sd)</code>	<code>rnorm(n, mean, sd)</code>
Cauchy	<code>dcauchy(x, location, scale)</code>	<code>pcauchy(q, location, scale)</code>	<code>qcauchy(p, location, scale)</code>	<code>rcauchy(n, location, scale)</code>
F	<code>df(x, df1, df2)</code>	<code>pf(q, df1, df2)</code>	<code>qf(p, df1, df2)</code>	<code>rf(n, df1, df2)</code>
T	<code>dt(x, df)</code>	<code>pt(q, df)</code>	<code>qt(p, df)</code>	<code>rt(n, df)</code>
Chi-squared	<code>dchisq(x, df)</code>	<code>pchi(q, df)</code>	<code>qchisq(p, df)</code>	<code>rchisq(n, df)</code>

Chú thích: Trong bảng trên, `df` = degrees of freedom (bậc tự do); `prob` = probability (xác suất); `n` = sample size (số lượng mẫu). Các thông số khác có thể tham khảo thêm cho từng luật phân phối. Riêng các luật phân phối F, t, Chi-squared còn có một thông số khác nữa là non-centrality parameter (`ncp`) được cho số 0. Tuy nhiên người sử dụng có thể cho một thông số khác thích hợp, nếu cần.

6.3 Các hàm phân phối xác suất (probability distribution function)

6.3.1 Hàm phân phối nhị phân (Binomial distribution)

Như tên gọi, hàm phân phối nhị phân chỉ có hai giá trị: nam / nữ, sống / chết, có / không, v.v... Hàm nhị phân được phát biểu bằng định lí như sau: Nếu một thử nghiệm

được tiến hành n lần, mỗi lần cho ra kết quả hoặc là thành công hoặc là thất bại, và gồm xác suất thành công được biết trước là p , thì xác suất có k lần thử nghiệm thành công là: $P(k|n,p) = C_k^n p^k (1-p)^{n-k}$, trong đó $k = 0, 1, 2, \dots, n$. Để hiểu định lí đó rõ ràng hơn, chúng ta sẽ xem qua vài ví dụ sau đây.

Ví dụ 1: Hàm mật độ nhị phân (Binomial density probability function). Trong ví dụ trên, lớp học có 10 người, trong đó có 6 nữ. Nếu 3 bạn được chọn một cách ngẫu nhiên, xác suất mà chúng ta có 2 bạn nữ là bao nhiêu? Chúng ta có thể trả lời câu hỏi này một cách tương đối thủ công bằng cách xem xét tất cả các trường hợp có thể xảy ra. Mỗi lần chọn có 2 khả năng (nam hay nữ), và 3 lần chọn, chúng ta có $2^3 = 8$ trường hợp như sau.

Bạn 1	Bạn 2	Bạn 3	Xác suất
Nam	Nam	Nam	(0.4)(0.4)(0.4) = 0.064
Nam	Nam	Nữ	(0.4)(0.4)(0.6) = 0.096
Nam	Nữ	Nam	(0.4)(0.6)(0.4) = 0.096
Nam	Nữ	Nữ	(0.4)(0.6)(0.6) = 0.144
Nữ	Nam	Nam	(0.6)(0.4)(0.4) = 0.096
Nữ	Nam	Nữ	(0.6)(0.4)(0.6) = 0.144
Nữ	Nữ	Nam	(0.6)(0.6)(0.4) = 0.144
Nữ	Nữ	Nữ	(0.6)(0.6)(0.6) = 0.216
Tất cả các trường hợp			1.000

Chúng ta biết trước rằng trong nhóm 10 học sinh có 6 nữ, và do đó, xác suất nữ là 0.60. (Nói cách khác, xác suất chọn một bạn nam là 0.4). Do đó, xác suất mà tất cả 3 bạn được chọn đều là nam giới là: $0.4 \times 0.4 \times 0.4 = 0.064$. Trong bảng trên, chúng ta thấy có 3 trường hợp mà trong đó có 2 bạn gái: đó là trường hợp Nam-Nữ-Nữ, Nữ-Nữ-Nam, và Nữ-Nam-Nữ, cả 3 đều có xác suất 0.144. Thành ra, xác suất chọn đúng 2 bạn nữ trong số 3 bạn được chọn là $3 \times 0.144 = 0.432$.

Trong R, có hàm `dbinom(k, n, p)` có thể giúp chúng ta tính công thức $P(k|n,p) = C_k^n p^k (1-p)^{n-k}$ một cách nhanh chóng. Trong trường hợp trên, chúng ta chỉ cần đơn giản lệnh:

```
> dbinom(2, 3, 0.60)
[1] 0.432
```

Ví dụ 2: Hàm nhị phân tích lũy (Cumulative Binomial probability distribution). Xác suất thuộc chủng loãng xương có hiệu nghiệm là khoảng 70% (tức là $p = 0.70$). Nếu chúng ta điều trị 10 bệnh nhân, xác suất có tối thiểu 8 bệnh nhân với kết quả tích cực là bao nhiêu? Nói cách khác, nếu gọi X là số bệnh nhân được điều trị thành công, chúng ta cần tìm $P(X \geq 8) = ?$. Để trả lời câu hỏi này, chúng ta sử dụng hàm `pbinom(k, n, p)`. Xin nhắc lại rằng hàm `pbinom(k, n, p)` cho chúng ta $P(X \leq k)$. Do đó, $P(X \geq 8) = 1 - P(X \leq 7)$. Thành ra, đáp số bằng R cho câu hỏi là:

```
> 1-pbinom(7, 10, 0.70)
[1] 0.3827828
```

Ví dụ 3: Mô phỏng hàm nhị phân: Biết rằng trong một quần thể dân số có khoảng 20% người mắc bệnh cao huyết áp; nếu chúng ta tiến hành chọn mẫu 1000 lần, mỗi lần chọn 20 người trong quần thể đó một cách ngẫu nhiên, sự phân phối số bệnh nhân cao huyết áp sẽ như thế nào? Để trả lời câu hỏi này, chúng ta có thể ứng dụng hàm `rbinom(n, k, p)` trong R với những thông số như sau:

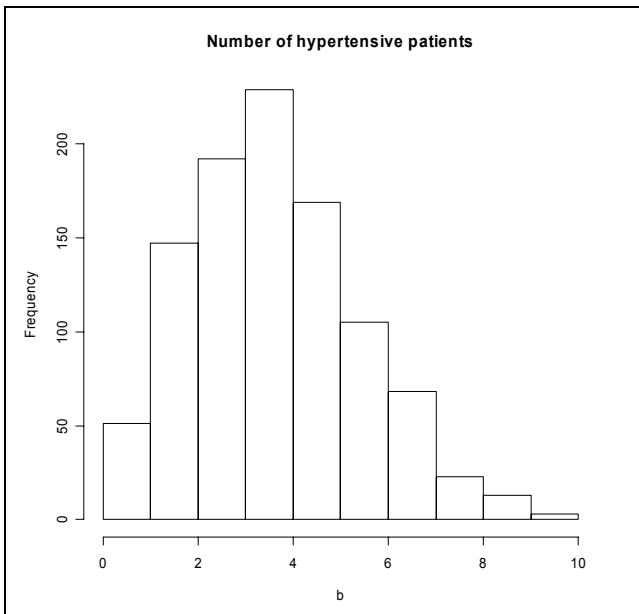
```
> b <- rbinom(1000, 20, 0.20)
```

Trong lệnh trên, kết quả mô phỏng được tạm thời chứa trong đối tượng tên là `b`. Để biết `b` có gì, chúng ta đếm bằng lệnh `table`:

```
> table(b)
b
 0   1   2   3   4   5   6   7   8   9   10
 6  45 147 192 229 169 105  68  23  13    3
```

Dòng số liệu thứ nhất (`0, 1, 2, ..., 10`) là số bệnh nhân mắc bệnh cao huyết áp trong số 20 người mà chúng ta chọn. Dòng số liệu thứ hai cho chúng ta biết số lần chọn mẫu trong 1000 lần xảy ra. Do đó, có 6 mẫu không có bệnh nhân cao huyết áp nào, 45 mẫu với chỉ 1 bệnh nhân cao huyết áp, v.v... Có lẽ cách để hiểu là vẽ đồ thị các tần số trên bằng lệnh `hist` như sau:

```
> hist(b, main="Number of hypertensive patients")
```



Biểu đồ 1. Phân phối số bệnh nhân cao huyết áp trong số 20 người được chọn ngẫu nhiên trong một quần thể gồm 20% bệnh nhân cao

huyết áp, và chọn mẫu được lặp lại 1000 lần.

Qua biểu đồ trên, chúng ta thấy xác suất có 4 bệnh nhân cao huyết áp (trong mỗi lần chọn mẫu 20 người) là cao nhất (22.9%). Điều này cũng có thể hiểu được, bởi vì tỉ lệ cao huyết áp là 20%, cho nên chúng ta kì vọng rằng trung bình 4 người trong số 20 người được chọn phải là cao huyết áp. Tuy nhiên, điều quan trọng mà biểu đồ trên thể hiện là có khi chúng ta quan sát đến 10 bệnh nhân cao huyết áp dù xác suất cho mẫu này rất thấp (chỉ 3/1000).

Ví dụ 4: Úng dụng hàm phân phối nhị phân: Hai mươi khách hàng được mời uống hai loại bia A và B, và được hỏi họ thích bia nào. Kết quả cho thấy 16 người thích bia A. Vấn đề đặt ra là kết quả này có đủ để kết luận rằng bia A được nhiều người thích hơn bia B, hay là kết quả chỉ là do các yếu tố ngẫu nhiên gây nên?

Chúng ta bắt đầu giải quyết vấn đề bằng cách giả thiết rằng nếu không có khác nhau, thì xác suất $p=0.50$ thích bia A và $q=0.5$ thích bia B. Nếu giả thiết này đúng, thì xác suất mà chúng ta quan sát 16 người trong số 20 người thích bia A là bao nhiêu. Chúng ta có thể tính xác suất này bằng R rất đơn giản:

```
> 1- pbinom(15, 20, 0.5)
[1] 0.005908966
```

Đáp số là xác suất 0.005 hay 0.5%. Nói cách khác, nếu quả thật hai bia giống nhau thì xác suất mà 16/20 người thích bia A chỉ 0.5%. Tức là, chúng ta có bằng chứng cho thấy khả năng bia A quả thật được nhiều người thích hơn bia B, chứ không phải do yếu tố ngẫu nhiên. Chú ý, chúng ta dùng 15 (thay vì 16), là bởi vì $P(X \geq 16) = 1 - P(X \leq 15)$. Mà trong trường hợp ta đang bàn, $P(X \leq 15) = pbinom(15, 20, 0.5)$.

6.3.2 Hàm phân phối Poisson (Poisson distribution)

Hàm phân phối Poisson, nói chung, rất giống với hàm nhị phân, ngoại trừ thông số p thường rất nhỏ và n thường rất lớn. Vì thế, hàm Poisson thường được sử dụng để mô tả các biến số rất hiếm xảy ra (như số người mắc ung thư trong một dân số chẵng hạn). Hàm Poisson còn được ứng dụng khá nhiều và thành công trong các nghiên cứu kĩ thuật và thị trường như số lượng khách hàng đến một nhà hàng mỗi giờ.

Ví dụ 5: Hàm mật độ Poisson (Poisson density probability function). Qua theo dõi nhiều tháng, người ta biết được tỉ lệ đánh sai chính tả của một thư ký đánh máy. Tính trung bình cứ khoảng 2.000 chữ thì thư ký đánh sai 1 chữ. Hỏi xác suất mà thư ký đánh sai chính tả 2 chữ, hơn 2 chữ là bao nhiêu?

Vì tần số khá thấp, chúng ta có thể giả định rằng biến số “sai chính tả” (tạm đặt tên là biến số X) là một hàm ngẫu nhiên theo luật phân phối Poisson. Ở đây, chúng ta có

tỉ lệ sai chính tả trung bình là 1($\lambda = 1$). Luật phân phối Poisson phát biểu rằng xác suất mà $X = k$, với điều kiện tỉ lệ trung bình λ :

$$P(X = k | \lambda) = \frac{e^{-\lambda} \lambda^k}{k!}$$

Do đó, đáp số cho câu hỏi trên là: $P(X = 2 | \lambda = 1) = \frac{e^{-2} 1^2}{2!} = 0.1839$. Đáp số này có thể tính bằng R một cách nhanh chóng hơn bằng hàm dpois như sau:

```
> dpois(2, 1)
[1] 0.1839397
```

Chúng ta cũng có thể tính xác suất sai 1 chữ, và xác suất không sai chữ nào:

```
> dpois(1, 1)
[1] 0.3678794
```

```
> dpois(0, 1)
[1] 0.3678794
```

Chú ý trong hàm trên, chúng ta chỉ đơn giản cung cấp thông số $k = 2$ và ($\lambda = 1$). Trên đây là xác suất mà thư ký đánh sai chính tả đúng 2 chữ. Nhưng xác suất mà thư ký đánh sai chính tả hơn 2 chữ (tức 3, 4, 5, ... chữ) có thể ước tính bằng:

$$\begin{aligned} P(X > 2) &= P(X = 3) + P(X = 4) + P(X = 5) + \dots \\ &= 1 - P(X \leq 2) \\ &= 1 - 0.3678 - 0.3678 - 0.1839 \\ &= 0.08 \end{aligned}$$

Bằng R, chúng ta có thể tính như sau:

```
# P(X ≤ 2)
> ppois(2, 1)
[1] 0.9196986
```

```
# 1 - P(X ≤ 2)
> 1 - ppois(2, 1)
[1] 0.0803014
```

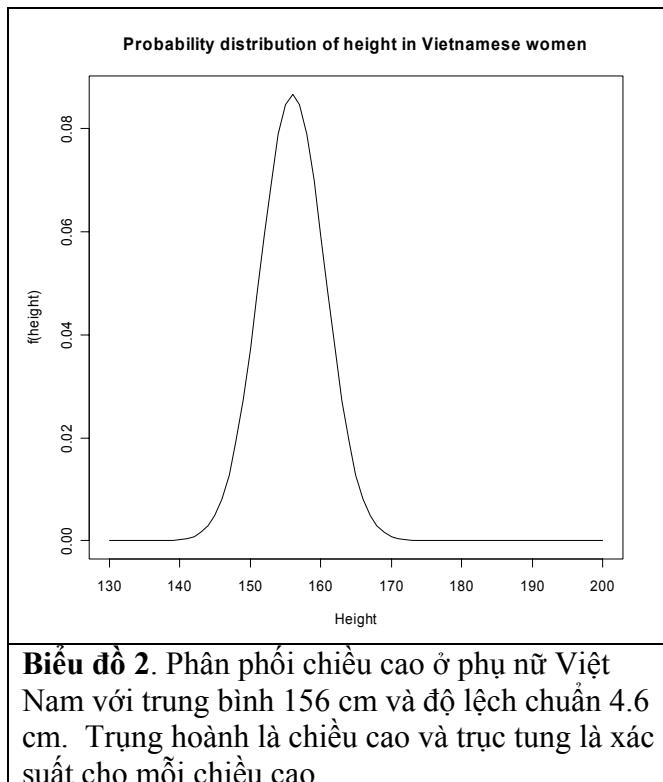
6.3.3 Hàm phân phối chuẩn (Normal distribution)

Hai luật phân phối mà chúng ta vừa xem xét trên đây thuộc vào nhóm phân phối áp dụng cho các biến số phi liên tục (discrete distributions), mà trong đó biến số có những giá trị theo bậc thứ hay thể loại. Đối với các biến số liên tục, có vài luật phân phối

thích hợp khác, mà quan trọng nhất là phân phối chuẩn. Phân phối chuẩn là nền tảng quan trọng nhất của phân tích thống kê. Có thể nói không ngoa rằng hầu hết lí thuyết thống kê được xây dựng trên nền tảng của phân phối chuẩn. Hàm mật độ phân phối chuẩn có hai thông số: trung bình μ và phương sai σ^2 (hay độ lệch chuẩn σ). Gọi X là một biến số (như chiều cao chặng hạn), hàm mật độ phân phối chuẩn phát biểu rằng xác suất mà $X=x$ là:

$$P(X=x|\mu,\sigma^2) = f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$

Ví dụ 6: Hàm mật độ phân phối chuẩn (Normal density probability function). Chiều cao trung bình hiện nay ở phụ nữ Việt Nam là 156 cm, với độ lệch chuẩn là 4.6 cm. Cũng biết rằng chiều cao này tuân theo luật phân phối chuẩn. Với hai thông số $\mu=156$, $\sigma=4.6$, chúng ta có thể xây dựng một hàm phân phối chiều cao cho toàn bộ quần thể phụ nữ Việt Nam, và hàm này có hình dạng như sau:



Biểu đồ trên được vẽ bằng hai lệnh sau đây. Lệnh đầu tiên nhằm tạo ra một biến số `height` có giá trị 130, 131, 132, ..., 200 cm. Lệnh thứ hai là vẽ biểu đồ với điều kiện trung bình là 156 cm và độ lệch chuẩn là 4.6 cm.

```
> height <- seq(130, 200, 1)
> plot(height, dnorm(height, 156, 4.6),
       type="l",
       ylab="f(height)",
       xlab="Height",
```

```
main="Probability distribution of height in Vietnamese women")
```

Với hai thông số trên (và biểu đồ), chúng ta có thể ước tính xác suất cho bất cứ chiều cao nào. Chẳng hạn như xác suất một phụ nữ Việt Nam có chiều cao 160 cm là:

$$P(X = 160 | \mu=156, \sigma=4.6) = \frac{1}{4.6\sqrt{2\times 3.1416}} \exp\left[-\frac{(160-156)^2}{2(4.6)^2}\right] \\ = 0.0594$$

Hàm `dnorm(x, mean, sd)` trong R có thể tính toán xác suất này cho chúng ta một cách gọn nhẹ:

```
> dnorm(160, mean=156, sd=4.6)
[1] 0.05942343
```

Hàm xác suất chuẩn tích lũy (cumulative normal probability function). Vì chiều cao là một biến số liên tục, trong thực tế chúng ta ít khi nào muốn tìm xác suất cho một giá trị cụ thể x , mà thường tìm xác suất cho một khoảng giá trị a đến b . Chẳng hạn như chúng ta muốn biết xác suất chiều cao từ 150 đến 160 cm (tức là $P(160 \leq X \leq 150)$), hay xác suất chiều cao thấp hơn 145 cm, tức $P(X < 145)$. Để tìm đáp số các câu hỏi như thế, chúng ta cần đến hàm xác suất chuẩn tích lũy, được định nghĩa như sau:

$$P(a \leq X \leq b) = \int_a^b f(x) dx$$

Thành ra, $P(160 \leq X \leq 150)$ chính là diện tích tính từ trục hoành = 150 đến 160 của **biểu đồ 2**. Trong R có hàm `pnorm(x, mean, sd)` dùng để tính xác suất tích lũy cho một phân phối chuẩn rất có ích.

$$\text{pnorm}(a, mean, sd) = \int_{-\infty}^a f(x) dx = P(X \leq a | \text{mean}, \text{sd})$$

Chẳng hạn như xác suất chiều cao phụ nữ Việt Nam bằng hoặc thấp hơn 150 cm là 9.6%:

```
> pnorm(150, 156, 4.6)
[1] 0.0960575
```

Hay xác suất chiều cao phụ nữ Việt Nam bằng hoặc cao hơn 165 cm là:

```
> 1-pnorm(164, 156, 4.6)
[1] 0.04100591
```

Nói cách khác, chỉ có khoảng 4.1% phụ nữ Việt Nam có chiều cao bằng hay cao hơn 165 cm.

Ví dụ 7: Ứng dụng luật phân phối chuẩn: Trong một quần thể, chúng ta biết rằng áp suất máu trung bình là 100 mmHg và độ lệch chuẩn là 13 mmHg, hỏi: có bao nhiêu người trong quần thể này có áp suất máu bằng hoặc cao hơn 120 mmHg? Câu trả lời bằng R là:

```
> 1-pnorm(120, mean=100, sd=13)
[1] 0.0619679
```

Tức khoảng 6.2% người trong quần thể này có áp suất máu bằng hoặc cao hơn 120 mmHg.

6.3.4 Hàm phân phối chuẩn chuẩn hóa (Standardized Normal distribution)

Một biến X tuân theo luật phân phối chuẩn với trung bình μ và phương sai σ^2 thường được viết tắt là:

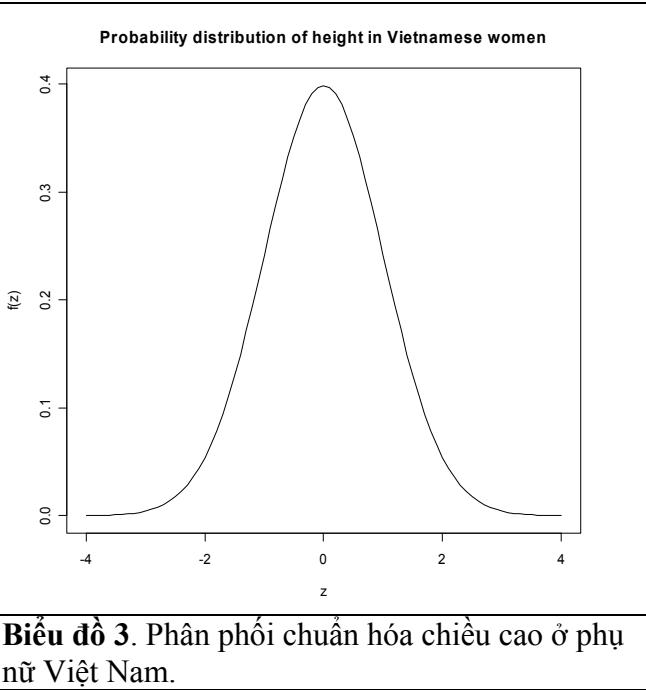
$$X \sim N(\mu, \sigma^2)$$

Ở đây μ và σ^2 tùy thuộc vào đơn vị đo lường của biến số. Chẳng hạn như chiều cao được tính bằng cm (hay m), huyết áp được đo bằng mmHg, tuổi được đo bằng năm, v.v... cho nên đôi khi mô tả một biến số bằng đơn vị gốc rất khó so sánh. Một cách đơn giản hơn là chuẩn hóa (standardized) X sao cho số trung bình là 0 và phương sai là 1. Sau vài thao tác số học, có thể chứng minh dễ dàng rằng, cách biến đổi X để đáp ứng điều kiện trên là:

$$Z = \frac{X - \mu}{\sigma}$$

Nói theo ngôn ngữ toán: nếu $X \sim N(\mu, \sigma^2)$, thì $(X - \mu)/\sigma \sim N(0, 1)$. Như vậy qua công thức trên, Z thực chất là độ khác biệt giữa một số và trung bình tính bằng số độ lệch chuẩn. Nếu $Z = 0$, chúng ta biết rằng X bằng số trung bình μ . Nếu $Z = -1$, chúng ta biết rằng X thấp hơn μ đúng 1 độ lệch chuẩn. Tương tự, $Z = 2.5$, chúng ta biết rằng X cao hơn μ đúng 2.5 độ lệch chuẩn. v.v...

Biểu đồ phân phối chiều cao của phụ nữ Việt Nam có thể mô tả bằng một đơn vị mới, đó là chỉ số z như sau:



Biểu đồ 3. Phân phối chuẩn hóa chiều cao ở phụ nữ Việt Nam.

Biểu đồ trên được vẽ bằng hai lệnh sau đây:

```
> height <- seq(-4, 4, 0.1)
> plot(height, dnorm(height, 0, 1),
       type="l",
       ylab="f(z)",
       xlab="z",
       main="Probability distribution of height in Vietnamese women")
```

Với phân phối chuẩn hóa, chúng ta có một tiện lợi là có thể dùng nó để mô tả và so sánh mật độ phân phối của bất cứ biến nào, vì tất cả đều được chuyển sang chỉ số z.

Trong biểu đồ trên, trục tung là xác suất z và trục hoành là biến số z. Chúng ta có thể tính toán xác suất z nhỏ hơn một hằng số (constant) nào đó dễ dàng bằng R. Ví dụ, chúng ta muốn tìm $P(z \leq -1.96) = ?$ cho một phân phối mà trung bình là 0 và độ lệch chuẩn là 1.

```
> pnorm(-1.96, mean=0, sd=1)
[1] 0.02499790
```

Hay $P(z \leq 1.96) = ?$

```
> pnorm(1.96, mean=0, sd=1)
[1] 0.9750021
```

Do đó, $P(-1.96 < z < 1.96)$ chính là:

```
> pnorm(1.96) - pnorm(-1.96)
[1] 0.9500042
```

Nói cách khác, xác suất 95% là z nằm giữa -1.96 và 1.96 . (Chú ý trong lệnh trên tôi không cung cấp `mean=0, sd=1`, bởi vì trong thực tế, `pnorm` giá trị mặc định (default value) của thông số `mean` là `0` và `sd` là `1`).

Ví dụ 6 (tiếp tục). Xin nhắc lại để tiện việc theo dõi, chiều cao trung bình ở phụ nữ Việt Nam là 156 cm và độ lệch chuẩn là 4.6 cm. Do đó, một phụ nữ có chiều cao 170 cm cũng có nghĩa là $z = (170 - 156) / 4.6 = 3.04$ độ lệch chuẩn, và tỉ lệ các phụ nữ Việt Nam có chiều cao cao hơn 170 cm là rất thấp, chỉ khoảng 0.1%.

```
> 1-pnorm(3.04)
[1] 0.001182891
```

Tìm định lượng (quantile) của một phân phối chuẩn. Đôi khi chúng ta cần làm một tính toán đảo ngược. Chẳng hạn như chúng ta muốn biết: nếu xác suất Z nhỏ hơn một hằng số z nào đó cho trước bằng p , thì z là bao nhiêu? Diễn tả theo kí hiệu xác suất, chúng ta muốn tìm z trong nếu:

$$P(Z < z) = p$$

Để trả lời câu hỏi này, chúng ta sử dụng hàm `qnorm(p, mean=, sd=)`.

Ví dụ 8: Biết rằng $Z \sim N(0, 1)$ và nếu $P(Z < z) = 0.95$, chúng ta muốn tìm z .

```
> qnorm(0.95, mean=0, sd=1)
[1] 1.644854
```

Hay $P(Z < z) = 0.975$ cho phân phối chuẩn với trung bình 0 và độ lệch chuẩn 1:

```
> qnorm(0.975, mean=0, sd=1)
[1] 1.959964
```

6.3.5 Hàm phân phối t, F và χ^2

Các hàm phân phối t, F và χ^2 trong thực tế là hàm của hàm phân phối chuẩn. Mọi liên hệ và cách tính các hàm này có thể được mô tả bằng vài ghi chú sau đây:

- **Phân phối Chi bình phương (χ^2).** Phân phối χ^2 xuất phát từ tổng bình phương của một biến phân phối chuẩn. Nếu nếu $x_i \sim N(0, 1)$, và gọi $u = \sum_{i=1}^n x_i^2$, thì u tuân theo luật phân phối Chi bình phương với bậc tự do n (thường viết tắt là df). Nói theo ngôn ngữ toán, $u \sim \chi_n^2$.

Ví dụ 9: Tìm xác suất của một biến Chi bình phương, do đó, chỉ cần hai thông số u và n . Chẳng hạn như nếu chúng ta muốn tìm xác suất $P(u=21, df=13)$, chỉ đơn giản dùng hàm `pchisq` như sau:

```
> dchisq(21, 13)
[1] 0.01977879
```

Tìm xác suất mà một biến số u nhỏ hơn 21 với bậc tự do 13 df. Tức là tìm $P(u \leq 21 | df=13) = ?$

```
> pchisq(21, 13)
[1] 0.9270714
```

Cũng có thể nói kết quả trên cho biết $P(\chi^2_{13} < 21) = 0.927$.

Tìm quantile của một trị số u tương đương với 90% của một phân phối χ^2 với 15 bậc tự do:

```
> qchisq(0.95, 15)
[1] 24.99579
```

Nói cách khác, $P(\chi^2_{15} < 24.99) = 0.95$.

Phi trung tâm (Non-centrality). Chú ý trong định nghĩa trên, phân phối χ^2 xuất phát từ tổng bình phương của một biến phân phối chuẩn có trung bình 0 và phương sai 1. Nhưng nếu một biến phân phối chuẩn có trung bình không phải là 0 và phương sai không phải là 1, thì chúng ta sẽ có một **phân phối Chi bình**

phi trung tâm. Nếu $x_i \sim N(\mu_i, 1)$ và đặt $u = \sum_{i=1}^n x_i^2$, thì u tuân theo luật phân phối Chi bình phương phi trung tâm với bậc tự do n và thông số phi trung tâm (non-centrality parameter) λ như sau:

$$\lambda = \sum_{i=1}^n \mu_i^2$$

Và kí hiệu là $u \sim \chi^2_{n,\lambda}$. Có thể nói thêm rằng, trung bình của u là $n+\lambda$, và phương sai của u là $2(n+2\lambda)$.

Tìm xác suất mà u nhỏ hơn hoặc bằng 21, với điều kiện bậc tự do là 13 và thông số non-centrality bằng 5.4:

```
> pchisq(21, 13, 5.4)
[1] 0.6837649
```

Tức là, $P(\chi^2_{13,5.4} < 21) = 0.684$.

Tìm quantile của một trị số tương đương với 50% của một phân phối χ^2 với 7 bậc tự do và thông số non-centrality bằng 3.

```
> qchisq(0.5, 7, 3)
[1] 9.180148
```

Do đó, $P(\chi^2_{7,3} < 9.180148) = 0.50$

- **Phân phối t (t distribution).** Chúng ta vừa biết rằng nếu $X \sim N(\mu, s^2)$ thì $(X - \mu)/\sigma \sim N(0, 1)$. Nhưng phát biểu đó đúng (hay chính xác) khi chúng ta biết phương sai σ^2 . Trong thực tế, ít khi nào chúng ta biết chính xác phương sai, mà chỉ ước tính từ số liệu thực nghiệm. Trong trường hợp phương sai được ước tính từ số liệu nghiên cứu, và hãy gọi ước tính này là s^2 , thì chúng ta có thể phát biểu rằng: $(X - \mu)/s \sim t(0, v)$, trong đó v là bậc tự do.

Ví dụ 10. Tìm xác suất mà x lớn hơn 1, trong biến theo luật phân phối t với 6 bậc tự do:

```
> 1-pt(1.1, 6)
[1] 0.1567481
```

Tức là, $P(t_6 > 1.1) = 1 - P(t_6 < 1.1) = 0.157$.

Tìm định lượng của một trị số tương đương với 95% của một phân phối t với 15 bậc tự do:

```
> qt(0.95, 15)
[1] 1.753050
```

Nói cách khác, $P(t_{19} < 1.75035) = 0.95$.

- **Phân phối F.** Tỉ số giữa hai biến số theo luật phân phối χ^2 có thể chứng minh là tuân theo luật phân phối F . Nói cách khác, nếu $u \sim \chi^2_n$ và $v \sim \chi^2_m$, thì $u/v \sim F_{n,m}$, trong đó n là bậc tự do tử số (numerator degrees of freedom) và m là bậc tự do mẫu số (denominator degrees of freedom).

Ví dụ 11: Tìm xác suất mà một trị số F lớn hơn 3.24, biết rằng biến số đó tuân theo luật phân phối F với bậc tự do 3 và 15 df và thông số non-centrality 5:

```
> 1-pf(3.24, 3, 15, 5)
[1] 0.3558721
```

Do đó, $P(F_{3, 15} > 3.24) = 1 - P(F_{3, 15} \geq 3.24) = 0.355338$.

Với bậc tự do 3 và 15, tìm C sao cho $P(F_{3, 15} > C) = 0.05$. Lời giải của R là:

```
> qf(1-0.05, 3, 15)
[1] 3.287382
```

Nói cách khác, $P(F_{3,15} > 3.287382) = 1 - P(F_{3,15} \geq 3.287382) = 1 - 0.95 = 0.05$

6.4 Mô phỏng (simulation)

Trong phân tích thống kê, đôi khi vì hạn chế số mẫu chúng ta khó có thể ước tính một cách chính xác các thông số, và trong trường hợp bất định đó, chúng ta cần đến mô phỏng để biết được độ dao động của một hay nhiều thông số. Mô phỏng thường dựa vào các luật phân phối. Đây là một lĩnh vực khá phức tạp mà tôi không có ý định trình bày đầy đủ trong chương này. Ở đây, tôi chỉ trình bày một số mô hình mô phỏng mang tính minh họa để bạn đọc có thể dựa vào đó mà phát triển thêm.

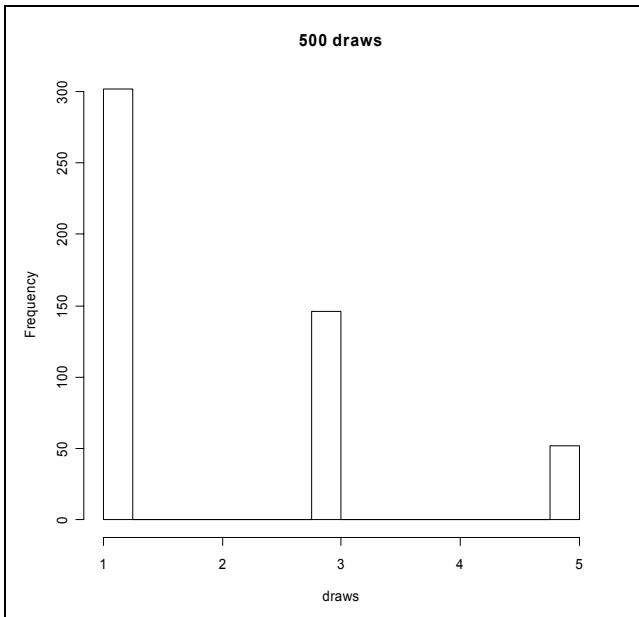
Ví dụ 11: Mô phỏng để chứng minh phương sai của số trung bình bằng phương sai chia cho n ($\text{var}(\bar{X}) = \sigma^2 / n$). Chúng ta sẽ xem một biến số không liên tục với giá trị 1, 3 và 5 với xác suất như sau:

x	$P(x)$
1	0.60
3	0.30
5	0.10

Qua số liệu này, chúng ta biết rằng giá trị trung bình là $(1 \times 0.60) + (3 \times 0.30) + (5 \times 0.10) = 2.0$ và phương sai (bạn đọc có thể tự tính) là 1.8.

Bây giờ chúng ta sử dụng hai thông số này để thử mô phỏng 500 lần. Lệnh thứ nhất tạo ra 3 giá trị của x . Lệnh thứ hai nhập số xác suất cho từng giá trị của x . Lệnh sample yêu cầu R tạo nên 500 số ngẫu nhiên và cho vào đối tượng draws.

```
x <- c(1, 3, 5)
px <- c(0.6, 0.3, 0.1)
draws <- sample(x, size=500, replace=T, prob=px)
hist(draws, breaks=seq(1,5, by=0.25), main="1000 draws")
```



Từ luật phân phối xác suất chúng ta biết rằng tính trung bình sẽ có 60% lần có giá trị “1”, 30% có giá trị “2”, và 10% có giá trị “5”. Do đó, chúng ta kì vọng sẽ quan sát 300, 150 và 50 lần cho mỗi giá trị. Biểu đồ trên cho thấy phân phối các giá trị này gần với giá trị mà chúng ta kì vọng. Ngoài ra, chúng ta cũng biết rằng phương sai của biến số này là khoảng 1.8. Bây giờ chúng ta kiểm tra xem có đúng như kì vọng hay không:

```
> var(draws)
[1] 1.835671
```

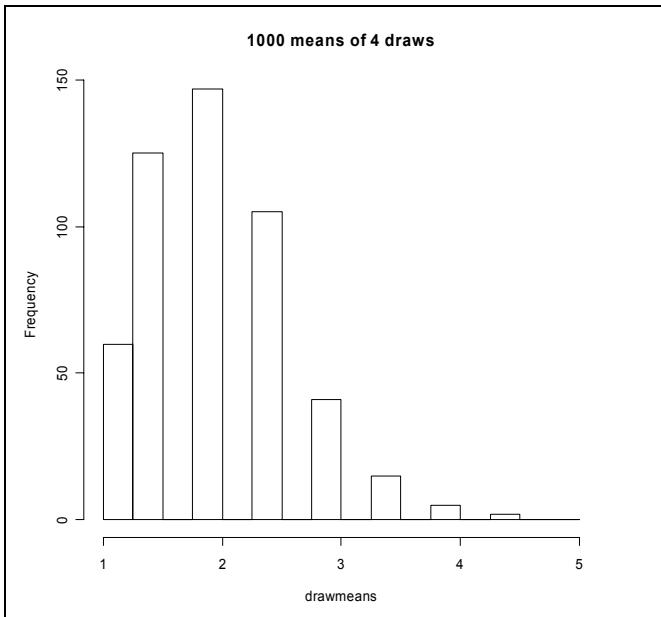
Kết quả trên cho thấy phương sai của 500 mẫu là 1.836, tức không xa mấy so với giá trị kì vọng.

Bây giờ chúng ta thử mô phỏng 500 giá trị trung bình \bar{x} (\bar{x} là số trung bình của 4 số liệu mô phỏng) từ quần thể trên:

```
> draws <- sample(x, size=4*500, replace=T, prob=px)
> draws = matrix(draws, 4)
> drawmeans = apply(draws, 2, mean)
```

Lệnh thứ nhất và thứ hai tạo nên đối tượng tên là `draws` với 4 dòng, mỗi dòng có 500 giá trị từ luật phân phối trên. Nói cách khác, chúng ta có $4 \times 500 = 2000$ số. 500 số cũng có nghĩa là 500 cột: 1 đến 500. Tức mỗi cột có 4 số. Lệnh thứ ba tìm trị số trung bình cho mỗi cột. Lệnh này sẽ cho ra 500 số trung bình và chứa trong đối tượng `drawmeans`. Biểu đồ sau đây cho thấy phân phối của 500 số trung bình:

```
> hist(drawmeans, breaks=seq(1,5,by=0.25), main="1000 means of 4 draws")
```



Chúng ta thấy rằng phương sai của phân phối này nhỏ hơn. Thật ra, phương sai của 500 số trung bình này là 0.45.

```
> var(drawmeans)
[1] 0.4501112
```

Đây là giá trị tương đương với giá trị 0.45 mà chúng ta kì vọng từ công thức $\text{var}(\bar{X}) = \sigma^2 / 4 = 1.8 / 4 = 0.45$.

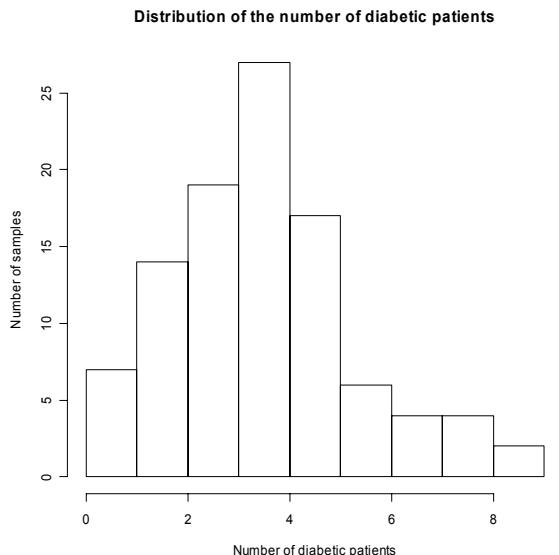
6.4.1 Mô phỏng phân phối nhị phân

Ví dụ 12: Mô phỏng mẫu từ một quần thể với luật phân phối nhị phân. Giả dụ chúng ta biết một quần thể có 20% người bị bệnh đái đường (xác suất $p=0.2$). Chúng ta muốn lấy mẫu từ quần thể này, mỗi mẫu có 20 đối tượng, và phương án chọn mẫu được lặp lại 100 lần:

```
> bin <- rbinom(100, 20, 0.2)
> bin
[1] 4 4 5 3 2 2 3 2 5 4 3 6 7 3 4 4 1 5 3 5 3 4 4 5 1 4 4 4 3 2 4 2 2 5 4 5
[38] 7 3 5 3 3 4 3 2 4 5 2 4 5 5 4 2 2 2 8 5 5 5 3 4 5 7 4 3 6 4 6 6 8 8 3 3 1
[75] 1 4 4 2 3 9 7 4 4 0 0 8 6 9 3 1 4 5 6 4 5 3 2 4 3 2
```

Kết quả trên là số lần đầu, chúng ta sẽ có 4 người mắc bệnh; lần 2 cũng 4 người; lần 3 có 5 người mắc bệnh; v.v... kết quả này có thể tóm lược trong một biểu đồ như sau:

```
> hist(bin,
      xlab="Number of diabetic patients",
      ylab="Number of samples",
      main="Distribution of the number of diabetic patients")
```



```
> mean(bin)
[1] 3.97
```

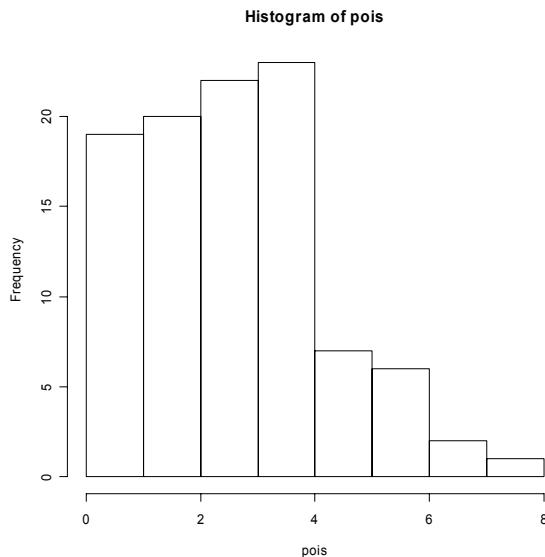
Đúng như chúng ta kì vọng, vì chọn mỗi lần 20 đối tượng và xác suất 20%, nên chúng ta tiên đoán trung bình sẽ có 4 bệnh nhân đái đường.

6.4.2 Mô phỏng phân phối Poisson

Ví dụ 13: Mô phỏng mẫu từ một quần thể với luật phân phối Poisson. Trong ví dụ sau đây, chúng ta mô phỏng 100 mẫu từ một quần thể tuân theo luật phân phối Poisson với trung bình $\lambda=3$:

```
> pois <- rpois(100, lambda=3)
> pois
> pois
[1] 4 3 2 4 2 3 4 4 0 7 5 0 3 3 3 4 2 2 6 1 4 2 3 3 5 4 2 1 4 0 2 1 5 1 2 2 2 6
[38] 1 3 6 3 3 5 4 3 2 2 5 3 3 3 1 4 7 3 4 3 2 6 1 4 1 0 5 2 2 3 6 8 4 4 1 4
[75] 1 0 0 4 3 3 2 3 3 4 1 5 4 4 1 3 1 6 4 4 4 2 2 2 4
```

Và mật độ phân phối:



Phân phối Poisson và phân phối mũ. Trong ví dụ sau đây, chúng ta mô phỏng thời gian bệnh nhân đến một bệnh viện. Biết rằng bệnh nhân đến bệnh viện một cách ngẫu nhiên theo luật phân phối Poisson, với trung bình 15 bệnh nhân cho mỗi 150 phút. Có thể chứng minh dễ dàng rằng thời gian giữa hai bệnh nhân đến bệnh viện tuân theo luật phân phối mũ. Chúng ta muốn biết thời gian mà bệnh nhân ghé bệnh viện; do đó, chúng ta mô phỏng 15 thời gian giữa hai bệnh nhân từ luật phân phối mũ với tỉ lệ $15/150 = 0.1$ mỗi phút. Các lệnh sau đây đáp ứng yêu cầu đó:

```
# Tạo thời gian đến bệnh viện
> appoint <- rexp(15, 0.1)
> times <- round(appoint, 0)
> times
[1] 37   5   8  10  24   5   1   7   8   6  12   6   3  25  15
```

6.4.3 Mô phỏng phân phối χ^2 , t, F

Cách mô phỏng trên đây còn có thể áp dụng cho các luật phân phối khác như nhị phân âm (negative binomial distribution với `rnbnom`), gamma (`rgamma`), beta (`rbeta`), Chi bình phương (`rchisq`), hàm mũ (`rexp`), t (`rt`), F (`rf`), v.v... Các thông số cho các hàm mô phỏng này có thể tìm trong phần đầu của chương.

Các lệnh sau đây sẽ minh họa các luật phân phối thông thường đó:

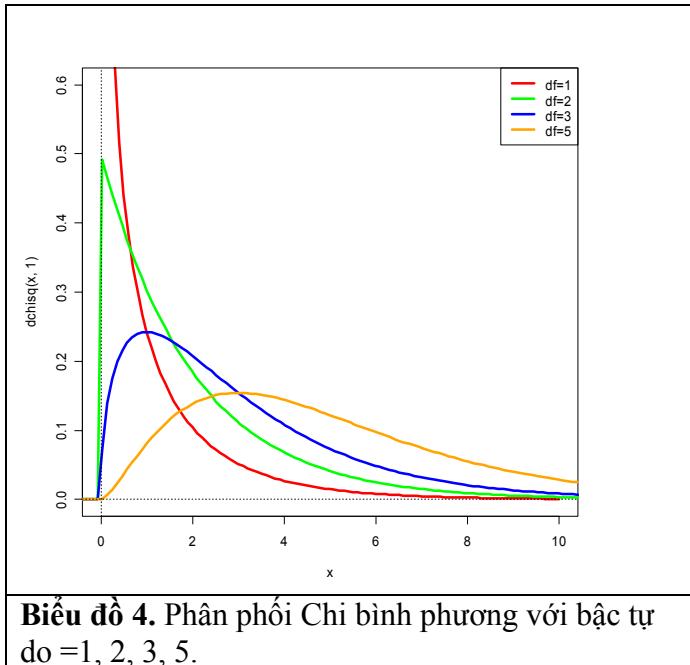
- Phân phối Chi bình phương với một số bậc tự do:

```
> curve(dchisq(x, 1), xlim=c(0,10), ylim=c(0,0.6), col="red", lwd=3)
> curve(dchisq(x, 2), add=T, col="green", lwd=3)
> curve(dchisq(x, 3), add=T, col="blue", lwd=3)
> curve(dchisq(x, 5), add=T, col="orange", lwd=3)
> abline(h=0, lty=3)
```

```

> abline(v=0, lty=3)
> legend(par("usr") [2], par("usr") [4],
  xjust=1,
  c("df=1", "df=2", "df=3", "df=5"), lwd=3, lty=1,
  col=c("red", "green", "blue", "orange"))

```

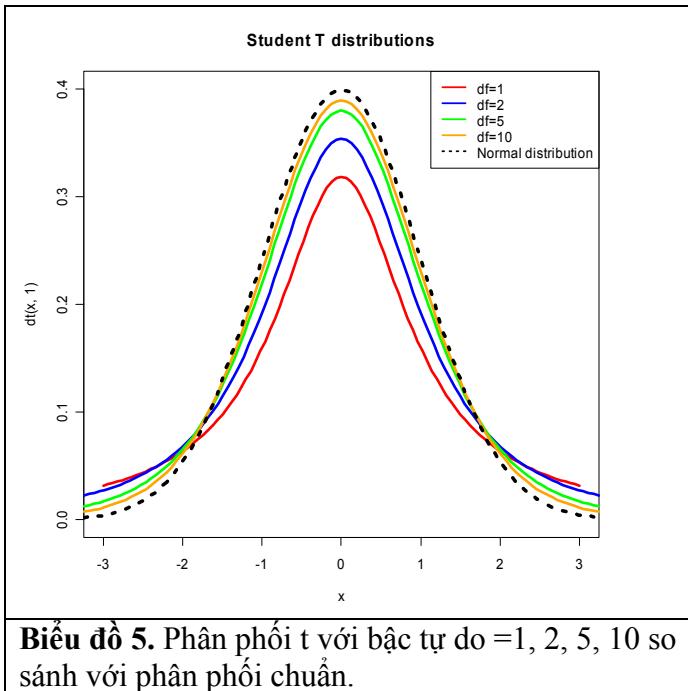


- Phân phối t :

```

> curve(dt(x, 1), xlim=c(-3,3), ylim=c(0,0.4), col="red", lwd=3)
> curve(dt(x, 2), add=T, col="blue", lwd=3)
> curve(dt(x, 5), add=T, col="green", lwd=3)
> curve(dt(x, 10), add=T, col="orange", lwd=3)
> curve(dnorm(x), add=T, lwd=4, lty=3)
> title(main="Student T distributions")
> legend(par("usr") [2], par("usr") [4],
  xjust=1,
  c("df=1", "df=2", "df=5", "df=10", "Normal distribution"),
  lwd=c(2,2,2,2,2),
  lty=c(1,1,1,1,3),
  col=c("red", "blue", "green", "orange", par("fg")))

```

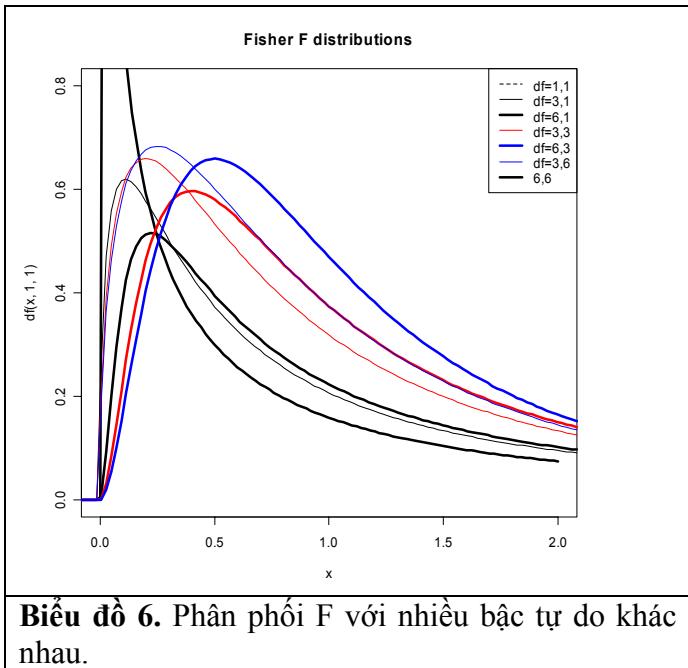


- Phân phối F:

```

> curve(df(x,1,1), xlim=c(0,2), ylim=c(0,0.8), lwd=3)
> curve(df(x,3,1), add=T)
> curve(df(x,6,1), add=T, lwd=3)
> curve(df(x,3,3), add=T, col="red")
> curve(df(x,6,3), add=T, col="red", lwd=3)
> curve(df(x,3,6), add=T, col="blue")
> curve(df(x,6,6), add=T, col="blue", lwd=3)
> title(main="Fisher F distributions")
> legend(par("usr")[2], par("usr")[4],
  xjust=1,
  c("df=1,1", "df=3,1", "df=6,1", "df=3,3", "df=6,3",
    "df=3,6", df="6,6"),
  lwd=c(1,1,3,1,3,1,3),
  lty=c(2,1,1,1,1,1,1),
  col=c(par("fg"), par("fg"), "red", "blue", "blue"))

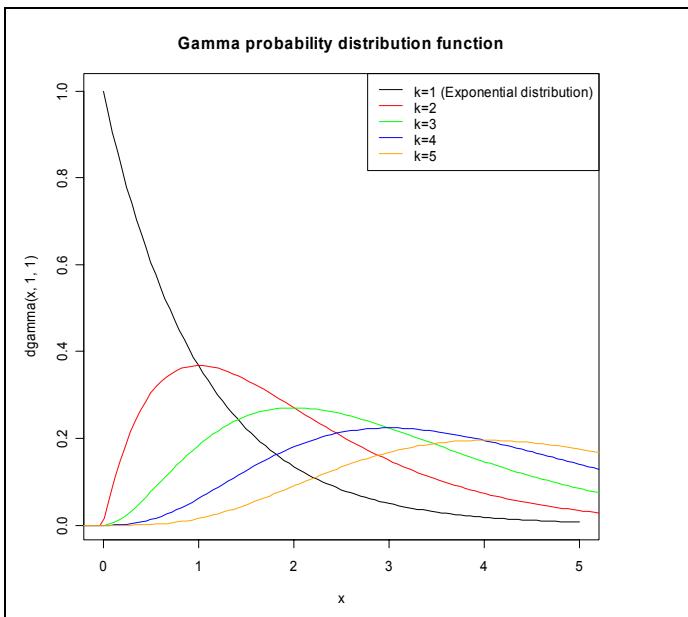
```



Biểu đồ 6. Phân phối F với nhiều bậc tự do khác nhau.

- Phân phối *gamma*:

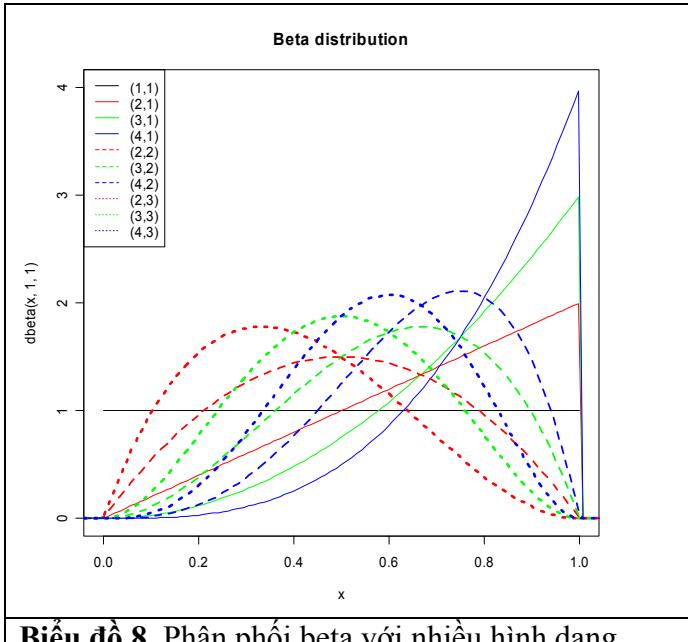
```
> curve( dgamma(x,1,1), xlim=c(0,5) )
> curve( dgamma(x,2,1), add=T, col='red' )
> curve( dgamma(x,3,1), add=T, col='green' )
> curve( dgamma(x,4,1), add=T, col='blue' )
> curve( dgamma(x,5,1), add=T, col='orange' )
> title(main="Gamma probability distribution function")
> legend(par('usr')[2], par('usr')[4], xjust=1,
         c('k=1 (Exponential distribution)', 'k=2', 'k=3', 'k=4', 'k=5'),
         lwd=1, lty=1,
         col=c(par('fg'), 'red', 'green', 'blue', 'orange') )
```



Biểu đồ 7. Phân phối Gamma với nhiều hình dạng.

- Phân phối *beta*:

```
> curve( dbeta(x,1,1), xlim=c(0,1), ylim=c(0,4) )
> curve( dbeta(x,2,1), add=T, col='red' )
> curve( dbeta(x,3,1), add=T, col='green' )
> curve( dbeta(x,4,1), add=T, col='blue' )
> curve( dbeta(x,2,2), add=T, lty=2, lwd=2, col='red' )
> curve( dbeta(x,3,2), add=T, lty=2, lwd=2, col='green' )
> curve( dbeta(x,4,2), add=T, lty=2, lwd=2, col='blue' )
> curve( dbeta(x,2,3), add=T, lty=3, lwd=3, col='red' )
> curve( dbeta(x,3,3), add=T, lty=3, lwd=3, col='green' )
> curve( dbeta(x,4,3), add=T, lty=3, lwd=3, col='blue' )
> title(main="Beta distribution")
> legend(par('usr')[1], par('usr')[4], xjust=0,
  c('(1,1)', '(2,1)', '(3,1)', '(4,1)',
    '(2,2)', '(3,2)', '(4,2)',
    '(2,3)', '(3,3)', '(4,3)' ),
  lwd=1, #c(1,1,1,1, 2,2,2, 3,3,3),
  lty=c(1,1,1,1, 2,2,2, 3,3,3),
  col=c(par('fg'), 'red', 'green', 'blue',
    'red', 'green', 'blue',
    'red', 'green', 'blue' ))
```



Biểu đồ 8. Phân phối beta với nhiều hình dạng.

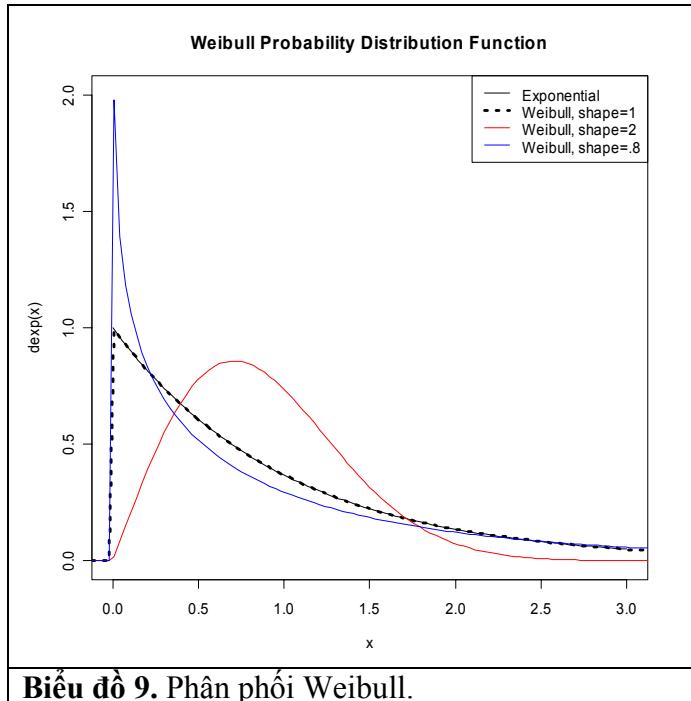
- Phân phối *Weibull*:

```
> curve(dexp(x), xlim=c(0,3), ylim=c(0,2))
> curve(dweibull(x,1), lty=3, lwd=3, add=T)
> curve(dweibull(x,2), col='red', add=T)
```

```

> curve(dweibull(x,.8), col='blue', add=T)
> title(main="Weibull Probability Distribution Function")
> legend(par('usr')[2], par('usr')[4], xjust=1,
  c('Exponential', 'Weibull, shape=1',
    'Weibull, shape=2', 'Weibull, shape=.8'),
  lwd=c(1,3,1,1),
  lty=c(1,3,1,1),
  col=c(par("fg"), par("fg"), 'red', 'blue'))

```

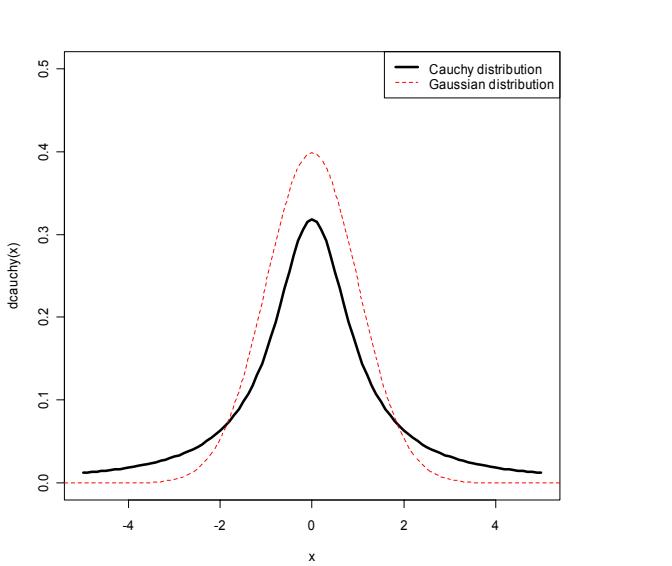


- Phân phối *Cauchy*:

```

> curve(dcauchy(x), xlim=c(-5,5), ylim=c(0,.5), lwd=3)
> curve(dnorm(x), add=T, col='red', lty=2)
> legend(par('usr')[2], par('usr')[4], xjust=1,
  c('Cauchy distribution', 'Gaussian distribution'),
  lwd=c(3,1),
  lty=c(1,2),
  col=c(par("fg"), 'red'))

```



Biểu đồ 9. Phân phối Cauchy so sánh với phân phối chuẩn.

6.5 Chọn mẫu ngẫu nhiên (random sampling)

Trong xác suất và thống kê, lấy mẫu ngẫu nhiên rất quan trọng, vì nó đảm bảo tính hợp lí của các phương pháp phân tích và suy luận thống kê. Với R, chúng ta có thể lấy mẫu một mẫu ngẫu nhiên bằng cách sử dụng hàm `sample`.

Ví dụ: Chúng ta có một quần thể gồm 40 người (mã số 1, 2, 3, ..., 40). Nếu chúng ta muốn chọn 5 đối tượng trong quần thể đó, ai sẽ là người được chọn? Chúng ta có thể dùng lệnh `sample()` để trả lời câu hỏi đó như sau:

```
> sample(1:40, 5)
[1] 32 26 6 18 9
```

Kết quả trên cho biết 5 đối tượng 32, 26, 8, 18 và 9 được chọn. Mỗi lần ra lệnh này, R sẽ chọn một mẫu khác, chứ không hoàn toàn giống như mẫu trên. Ví dụ:

```
> sample(1:40, 5)
[1] 5 22 35 19 4

> sample(1:40, 5)
[1] 24 26 12 6 22

> sample(1:40, 5)
[1] 22 38 11 6 18
```

v.v...

Trên đây là lệnh để chúng ta chọn mẫu ngẫu nhiên mà không thay thế (random sampling without replacement), tức là mỗi lần chọn mẫu, chúng ta không bỏ lại các mẫu đã chọn vào quần thể.

Nhưng nếu chúng ta muốn chọn mẫu thay thế (tức mỗi lần chọn ra một số đối tượng, chúng ta bỏ vào lại trong quần thể để chọn tiếp lần sau). Ví dụ, chúng ta muốn chọn 10 người từ một quần thể 50 người, bằng cách lấy mẫu với thay thế (random sampling with replacement), chúng ta chỉ cần thêm tham số `replace = TRUE`:

```
> sample(1:50, 10, replace=T)
[1] 31 44 6 8 47 50 10 16 29 23
```

Hay ném một đồng xu 10 lần; mỗi lần, dĩ nhiên đồng xu có 2 kết quả H và T; và kết quả 10 lần có thể là:

```
> sample(c("H", "T"), 10, replace=T)
[1] "H" "T" "H" "H" "H" "T" "H" "H" "T" "T"
```

Cũng có thể tưởng tượng chúng ta có 5 quả banh màu xanh (X) và 5 quả banh màu đỏ (D) trong một bao. Nếu chúng ta chọn 1 quả banh, ghi nhận màu, rồi để lại vào bao; rồi lại chọn 1 quả banh khác, ghi nhận màu, và bỏ vào bao lại. Cứ như thế, chúng ta chọn 20 lần, kết quả có thể là:

```
> sample(c("X", "D"), 20, replace=T)
[1] "X" "D" "D" "D" "D" "X" "X" "X" "X" "D" "X" "X" "D" "X" "X" "X" "X"
[20] "D"
```

Ngoài ra, chúng ta còn có thể lấy mẫu với một xác suất cho trước. Trong hàm sau đây, chúng ta chọn 10 đối tượng từ dãy số 1 đến 5, nhưng xác suất không bằng nhau:

```
> sample(5, 10, prob=c(0.3, 0.4, 0.1, 0.1, 0.1), replace=T)
[1] 3 1 3 2 2 2 2 5 1
```

Đối tượng 1 được chọn 2 lần, đối tượng 2 được chọn 5 lần, đối tượng 3 được chọn 2 lần, v.v... Tuy không hoàn toàn phù hợp với xác suất 0.3, 0.4, 0.1 như cung cấp vì số mẫu còn nhỏ, nhưng cũng không quá xa với kì vọng.

Kiểm định giả thiết thống kê và ý nghĩa của trị số P (P-value)

7.1 Trị số P

Trong nghiên cứu khoa học, ngoài những dữ kiện bằng số, biểu đồ và hình ảnh, con số mà chúng ta thường hay gặp nhất là trị số P (mà tiếng Anh gọi là P-value). Trong các chương sau đây, bạn đọc sẽ gặp trị số P rất nhiều lần, và đại đa số các suy luận phân tích thống kê, suy luận khoa học đều dựa vào trị số P. Do đó, trước khi bàn đến các phương pháp phân tích thống kê bằng R, tôi thấy cần phải có đôi lời về ý nghĩa của trị số này.

Trị số P là một con số xác suất, tức là viết tắt chữ “probability value”. Chúng ta thường gặp những phát biểu được kèm theo con số, chẳng hạn như “Kết quả phân tích cho thấy tỉ lệ gãy xương trong nhóm bệnh nhân được điều trị bằng thuốc Alendronate là 2%, thấp hơn tỉ lệ trong nhóm bệnh nhân không được chữa trị (5%), và mức độ khác biệt này có ý nghĩa thống kê ($p = 0.01$)”, hay một phát biểu như “Sau 3 tháng điều trị, mức độ giảm áp suất máu trong nhóm bệnh nhân là 10% ($p < 0.05$)”. Trong văn cảnh trên đây, đại đa số nhà khoa học hiểu rằng trị số P phản ánh xác suất sự hiệu nghiệm của thuốc Alendronate hay một thuật điều trị, họ hiểu rằng câu văn trên có nghĩa là “xác suất mà thuốc Alendronate tốt hơn giả dược là 0.99” (lấy 1 trừ cho 0.01). Nhưng cách hiểu đó hoàn toàn sai!

Trong “Từ điển toán kinh tế thống kê, kinh tế lượng Anh – Việt” (Nhà xuất bản Khoa học và Kỹ thuật, 2004), tác giả định nghĩa trị số P như sau: “*P – giá trị (hoặc giá trị xác suất). P giá trị là mức ý nghĩa thống kê thấp nhất mà ở đó giá trị quan sát được của thống kê kiểm định có ý nghĩa*” (trang 690). Định nghĩa này thật là khó hiểu! Thật ra đó cũng là định nghĩa chung mà các sách khoa Tây phương thường hay viết. Lật bất cứ sách giáo khoa nào bằng tiếng Anh, chúng ta sẽ thấy một định nghĩa về trị số P na ná giống nhau như “Trị số P là xác suất mà mức độ khác biệt quan sát do các yếu tố ngẫu nhiên gây ra (*P value is the probability that the observed difference arose by chance*)”. Thực ra định nghĩa này chưa đầy đủ, nếu không muốn nói là ... sai. Chính vì sự mù mờ của định nghĩa cho nên rất nhiều nhà khoa học hiểu sai ý nghĩa của trị số P.

Thật vậy, rất nhiều người, không chỉ người đọc mà ngay cả chính các tác giả của những bài báo khoa học, không hiểu ý nghĩa của trị số P. Theo một nghiên cứu được công bố trên tạp san danh tiếng *Statistics in Medicine* [1], tác giả cho biết 85% các tác giả khoa học và bác sĩ nghiên cứu không hiểu hay hiểu sai ý nghĩa của trị số P. Đọc đến đây có lẽ bạn đọc rất ngạc nhiên, bởi vì điều này có nghĩa là nhiều nhà nghiên cứu khoa học có khi không hiểu hay hiểu sai những gì chính họ viết ra có nghĩa gì! Thê thì, câu hỏi cần đặt ra một cách nghiêm chỉnh: Ý nghĩa của trị số P là gì? Để trả lời cho câu hỏi này,

chúng ta cần phải xem xét qua khái niệm phản nghiệm và tiến trình của một nghiên cứu khoa học.

7.2 Giả thiết khoa học và phản nghiệm

Một giả thiết được xem là mang tính “khoa học” nếu giả thiết đó có khả năng “phản nghiệm”. Theo Karl Popper, nhà triết học khoa học, đặc điểm duy nhất để có thể phân biệt giữa một lí thuyết khoa học thực thụ với ngụy khoa học (pseudoscience) là thuyết khoa học luôn có đặc tính có thể “bị bác bỏ” (hay bị phản bác – falsified) bằng những thực nghiệm đơn giản. Ông gọi đó là “khả năng phản nghiệm” (falsifiability, có tài liệu ghi là falsibility). Phép phản nghiệm là phương cách tiến hành những thực nghiệm không phải để xác minh mà để phê phán các lí thuyết khoa học, và có thể coi đây như là một nền tảng cho khoa học thực thụ. Chẳng hạn như giả thiết “Tất cả các quả đều màu đen” có thể bị bác bỏ nếu ta tìm ra có một con quả màu đỏ.

Có thể xem qui trình phản nghiệm là một cách học hỏi từ sai lầm! Thật vậy, trong khoa học chúng ta học hỏi từ sai lầm. Khoa học phát triển cũng một phần lớn là do học hỏi từ sai lầm mà giới khoa học không ai chối cãi. Sai lầm là điểm mạnh của khoa học. Có thể xác định nghiên cứu khoa học như là một qui trình thử nghiệm giả thuyết, theo các bước sau đây:

Bước 1, nhà nghiên cứu cần phải định nghĩa một giả thuyết đảo (null hypothesis), tức là một giả thuyết ngược lại với những gì mà nhà nghiên cứu tin là sự thật. Thí dụ trong một nghiên cứu lâm sàng, gồm hai nhóm bệnh nhân: một nhóm được điều trị bằng thuốc A, và một nhóm được điều trị bằng placebo, nhà nghiên cứu có thể phát biểu một giả thuyết đảo rằng sự hiệu nghiệm thuốc A tương đương với sự hiệu nghiệm của placebo (có nghĩa là thuốc A không có tác dụng như mong muốn).

Bước 2, nhà nghiên cứu cần phải định nghĩa một giả thuyết phụ (alternative hypothesis), tức là một giả thuyết mà nhà nghiên cứu nghĩ là sự thật, và điều cần được “chứng minh” bằng dữ kiện. Chẳng hạn như trong ví dụ trên đây, nhà nghiên cứu có thể phát biểu giả thuyết phụ rằng thuốc A có hiệu nghiệm cao hơn placebo.

Bước 3, sau khi đã thu thập đầy đủ những dữ kiện liên quan, nhà nghiên cứu dùng một hay nhiều phương pháp thống kê để kiểm tra xem trong hai giả thuyết trên, giả thuyết nào được xem là khả dĩ. Cách kiểm tra này được tiến hành để trả lời câu hỏi: nếu giả thuyết đảo đúng, thì xác suất mà những dữ kiện thu thập được phù hợp với giả thuyết đảo là bao nhiêu. Giá trị của xác suất này thường được đề cập đến trong các báo cáo khoa học bằng kí hiệu “P value”. Điều cần chú ý ở đây là nhà nghiên cứu không thử nghiệm giả thuyết khác, mà chỉ thử nghiệm giả thuyết đảo mà thôi.

Bước 4, quyết định chấp nhận hay loại bỏ giả thuyết đảo, bằng cách dựa vào giá trị xác suất trong bước thứ ba. Chẳng hạn như theo truyền thống lựa chọn trong một nghiên cứu y học, nếu giá trị xác suất nhỏ hơn 5% thì nhà nghiên cứu sẵn sàng bác bỏ giả thuyết đảo: sự hiệu nghiệm của thuốc A khác với sự hiệu nghiệm của placebo. Tuy nhiên, nếu giá trị xác suất cao hơn 5%, thì nhà nghiên cứu chỉ có thể phát biểu rằng chưa

có bằng chứng đầy đủ để bác bỏ giả thuyết đảo, và điều này không có nghĩa rằng giả thuyết đảo là đúng, là sự thật. Nói một cách khác, thiếu bằng chứng không có nghĩa là không có bằng chứng.

Bước 5, nếu giả thuyết đảo bị bác bỏ, thì nhà nghiên cứu mặc nhiên thừa nhận giả thuyết phụ. Nhưng vẫn đề khởi đi từ đây, bởi vì có nhiều giả thuyết phụ khác nhau. Chẳng hạn như so sánh với giả thuyết phụ ban đầu (A khác với Placebo), nhà nghiên cứu có thể đặt ra nhiều giả thuyết phụ khác nhau như thuốc sự hiệu nghiệm của thuốc A cao hơn Placebo 5%, 10% hay nói chung X%. Nói tóm lại, một khi nhà nghiên cứu bác bỏ giả thuyết đảo, thì giả thuyết phụ được mặc nhiên công nhận, nhưng nhà nghiên cứu không thể xác định giả thuyết phụ nào là đúng với sự thật.

7.3 Ý nghĩa của trị số P qua mô phỏng

Để hiểu ý nghĩa thực tế của trị số P, tôi sẽ nêu một ví dụ đơn giản như sau:

Ví dụ 1. Một thí nghiệm được tiến hành để tìm hiểu sở thích của người tiêu thụ đối với hai loại cà phê (hãy tạm gọi là cà phê A và B). Các nhà nghiên cứu cho 50 khách hàng uống thử hai loại cà phê trong cùng một điều kiện, và hỏi họ thích loại cà phê nào. Kết quả cho thấy 35 người thích cà phê A, và 15 người thích cà phê B. Vấn đề đặt ra là qua kết quả này, các nhà nghiên cứu có thể kết luận rằng cà phê loại A được ưa chuộng hơn cà phê B, hay kết quả trên chỉ là do ngẫu nhiên mà ra?

“Do ngẫu nhiên mà ra” có nghĩa là theo luật nhị phân, khả năng mà kết quả trên xảy ra là bao nhiêu? Do đó, lí thuyết xác suất nhị phân có phần ứng dụng trong trường hợp này, bởi vì kết quả của nghiên cứu chỉ có hai “giá trị” (hoặc là thích A, hoặc thích B).

Nói theo ngôn ngữ của phản nghiệm, giả thiết đảo là nếu không có sự khác biệt về sở thích, xác suất mà một khách hàng ưa chuộng một loại cà phê là 0.5. Nếu giả thiết này là đúng (tức $p = 0.5$, p ở đây là xác suất thích cà phê A), và nếu nghiên cứu trên được lặp đi lặp lại (chẳng hạn như) 1000 lần, và mỗi lần vẫn 50 khách hàng, thì có bao nhiêu lần với 35 khách hàng ưa chuộng cà phê A? Gọi số lần nghiên cứu mà 35 (hay nhiều hơn) trong số 50 thích cà phê A là “biến cố” X, nói theo ngôn ngữ xác suất, chúng ta muốn tìm $P(X | p=0.50) = ?$

Để trả lời câu hỏi này, chúng ta có thể ứng dụng hàm `rbinom` để mô phỏng vì như nói trên thực chất của vấn đề là một phân phối nhị phân:

```
> bin <- rbinom(1000, 50, 0.5)
```

Trong lệnh trên, chúng ta yêu cầu R mô phỏng 1000 lần nghiên cứu, mỗi lần có 50 khách hàng, và theo giả thiết đảo, xác suất thích A là 0.50. Để biết kết quả của mô phỏng đó, chúng ta sử dụng hàm `table` như sau:

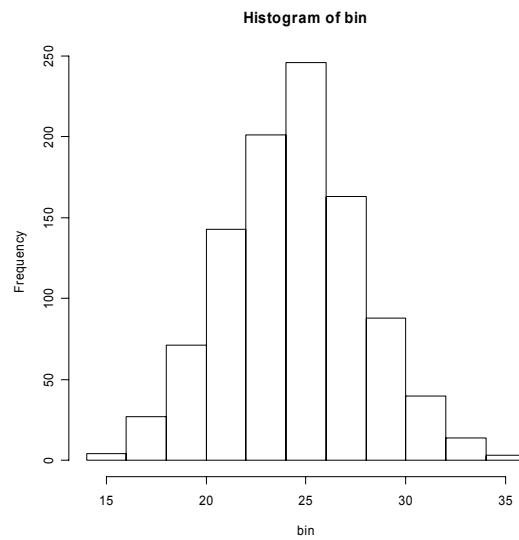
```
> table(bin)
```

bin	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33
1	1	1	2	11	16	24	47	60	83	94	107	132	114	98	65	44	44	26	14	12
2	34	35																		

Qua kết quả trên, chúng ta thấy trong số 1000 “nghiên cứu” đó, chỉ có 3 nghiên cứu mà số khách hàng thích cà phê A là 35 người (với điều kiện không có khác biệt giữa hai loại cà phê, hay nói đúng hơn là nếu $p=0.5$). Nói cách khác:

$$P(X \geq 35 | p=0.50) = 3/1000 = 0.003$$

Chúng ta cũng có thể thể hiện tần số trên bằng một biểu đồ tần số như sau:



Tất nhiên chúng ta có thể làm một mô phỏng khác với số lần tái thí nghiệm là 100.000 lần (thay vì 1000 lần) và tính xác suất $P(X \geq 35 | p=0.50)$.

```
bin <- rbinom(100000, 50, 0.5)
> bin <- rbinom(100000, 50, 0.5)
> table(bin)
bin
  11     12     13     14     15     16     17     18     19     20     21     22     23 
    4      17     40     83    197    462    946   1592   2719   4098   5892   7937  9733 
  24     25     26     27     28     29     30     31     32     33     34     35     36 
10822 11191 10799  9497  7925  5904  4185  2682  1562  893   455   223   98 
  37     38     39     40 
    31      5      7      1
```

Lần này, chúng ta có nhiều khả năng hơn (vì số lần mô phỏng tăng lên). Chẳng hạn như có thể có nghiên cứu cho ra 11 khách hàng (tối thiểu) hay 40 khách hàng (tối đa) thích cà

phê A. Nhưng chúng ta muốn biết số lần nghiên cứu mà 35 khách hàng trở lên thích cà phê A, và kết quả trên cho chúng ta biết, xác suất đó là:

```
> (223+98+21+5+7+1) / 100000  
[1] 0.00355
```

Nói cách khác, xác suất $P(X \geq 35 | p=0.50)$ quá thấp (chỉ 0.3%), chúng ta có bằng chứng để cho rằng kết quả trên có thể không do các yếu tố ngẫu nhiên gây nên; tức có một sự khác biệt về sở thích của khách hàng đối với hai loại cà phê.

Con số $P = 0.0035$ chính là trị số P. Theo một qui ước khoa học, tất cả các trị số P thấp hơn 0.05 (tức thấp hơn 5%) được xem là “significant”, tức là “có ý nghĩa thống kê”.

Cần phải nhấn mạnh một lần nữa để hiểu ý nghĩa của trị số P như sau: Mục đích của phân tích trên là nhằm trả lời câu hỏi: *nếu hai loại cà phê có xác suất ưa chuộng bằng nhau ($p = 0.5$, giả thuyết đảo), thì xác suất mà kết quả trên (35 trong số 50 khách hàng thích A) xảy ra là bao nhiêu?* Nói cách khác, đó chính là phương pháp đi tìm trị số P. Do đó, diễn dịch trị số P phải có điều kiện, và điều kiện ở đây là $p = 0.50$. Bạn đọc có thể làm thí nghiệm thêm với $p = 0.6$ hay $p = 0.7$ để thấy kết quả khác nhau ra sao.

Trong thực tế, trị số P có một ảnh hưởng rất lớn đến số phận của một bài báo khoa học. Nhiều tập san và nhà khoa học xem một nghiên cứu khoa học với trị số P cao hơn 0.05 là một “kết quả tiêu cực” (“negative result”) và bài báo có thể bị từ chối cho công bố. Chính vì thế mà đối với đại đa số nhà khoa học, con số “ $P < 0.05$ ” đã trở thành một cái “giấy thông hành” để công bố kết quả nghiên cứu. Nếu kết quả với $P < 0.05$, bài báo có cơ may xuất hiện trên một tập san nào đó và tác giả có thể sẽ nổi tiếng; nếu kết quả $P > 0.05$, số phận bài báo và công trình nghiên cứu có cơ may đi vào lăng quên!

7.4 Vấn đề logic của trị số P

Nhưng đứng trên phương diện lí trí và khoa học nghiêm chỉnh, chúng ta có nên đặt tầm quan trọng vào trị số P như thế hay không? Theo tôi, câu trả lời là không. Trị số P có nhiều vấn đề, và việc phụ thuộc vào nó trong quá khứ (cũng như hiện nay) đã bị rất nhiều người phê phán gay gắt. Cái khiếm khuyết số 1 của trị số P là nó thiếu tính logic. Thật vậy, nếu chúng ta chịu khó xem xét lại ví dụ trên, chúng ta có thể khái quát tiến trình của một nghiên cứu y học (dựa vào trị số P) như sau:

- Đề ra một giả thuyết chính (H_+)
- Từ giả thuyết chính, đề ra một giả thuyết đảo (H_-)
- Tiến hành thu thập dữ kiện (D)
- Phân tích dữ kiện: tính toán xác suất D xảy ra nếu H_- là sự thật. Nói theo ngôn ngữ toán xác suất, bước này xác định $P(D | H_-)$.

Vì thế, con số P có nghĩa là xác suất của dữ kiện D xảy ra *nếu* (nhấn mạnh: “nếu”) giả thuyết đảo H- là sự thật. Như vậy, con số P không trực tiếp cho chúng ta một ý niệm gì về sự thật của giả thuyết chính H; nó chỉ gián tiếp cung cấp bằng chứng để chúng ta chấp nhận giả thuyết chính và bác bỏ giả thuyết đảo.

Cái logic đằng sau của trị số P có thể được hiểu như là một tiến trình *chứng minh đảo ngược* (proof by contradiction):

- Mệnh đề 1: Nếu giả thuyết đảo là sự thật, thì dữ kiện này không thể xảy ra;
- Mệnh đề 2: Dữ kiện xảy ra;
- Mệnh đề 3 (kết luận): Giả thuyết đảo không thể là sự thật.

Nếu bạn đọc cảm thấy khó hiểu cách lập luận trên, tôi xin lấy thêm một ví dụ trong y khoa để minh họa cho tiến trình này:

- Nếu ông Tuấn bị cao huyết áp, thì ông không thể có triệu chứng rụng tóc (hai hiện tượng sinh học này không liên quan với nhau, ít ra là theo kiến thức y khoa hiện nay);
- Ông Tuấn bị rụng tóc;
- Do đó, ông Tuấn không thể bị cao huyết áp.

Trị số P, do đó, gián tiếp phản ánh xác suất của mệnh đề 3. Và đó cũng chính là một khiếm khuyết quan trọng của trị số P, bởi vì con số P nó ước tính mức độ khả dĩ của dữ kiện, chứ không nói cho chúng ta biết mức độ khả dĩ của một giả thuyết. Điều này làm cho việc suy luận dựa vào trị số P rất xa rời với thực tế, xa rời với khoa học thực nghiệm. Trong khoa học thực nghiệm, điều mà nhà nghiên cứu muốn biết là với dữ kiện mà họ có được, xác suất của giả thuyết chính là bao nhiêu, chứ họ không muốn biết nếu giả thuyết đảo là sự thật thì xác suất của dữ kiện là bao nhiêu. Nói cách khác và dùng kí hiệu mô tả trên, nhà nghiên cứu muốn biết $P(H+ | D)$, chứ không muốn biết $P(D | H+)$ hay $P(D | H-)$.

7.5. Vấn đề kiểm định nhiều giả thuyết (multiple tests of hypothesis)

Như đã nói trên, nghiên cứu y học là một quá trình thử nghiệm giả thuyết. Trong một nghiên cứu, ít khi nào chúng ta thử nghiệm chỉ một giả thuyết duy nhất, mà rất nhiều giả thuyết một lược. Chẳng hạn như trong một nghiên cứu về mối liên hệ giữa vitamin D và nguy cơ gãy xương đùi, các nhà nghiên cứu có thể phân tích mối liên hệ tương quan giữa vitamin D và mật độ xương (bone mineral density), giữa vitamin D và nguy cơ gãy xương theo từng giới tính, từng nhóm tuổi, hay phân tích theo các đặc tính lâm sàng của bệnh nhân, v.v... (Xem ví dụ dưới đây). Mỗi một phân tích như thế có thể xem là một thử nghiệm giả thuyết. Ở đây, chúng ta phải đổi diện với vấn đề nhiều giả thuyết (multiple tests of hypothesis hay còn gọi là **multiple comparisons**).

Bảng 2. Phân tích hiệu quả của vitamin D và calcium theo đặc tính của bệnh nhân

Đặc tính bệnh nhân	Nhóm được điều trị bằng calcium và vitamin D ¹	Nhóm giả dược (placebo) ¹	Tỉ số nguy cơ (relative risk) và khoảng tin cậy 95% ²
Độ tuổi			
50-59	29 (0.06)	13 (0.03)	2,17 (1.13-4.18)
60-69	53 (0.09)	71 (0.13)	0.74 (0.52-1.06)
70-79	93 (0.44)	115 (0.54)	0.82 (0.62-1.08)
Body mass index			
<25	69 (0.20)	66 (0.19)	1.05 (0.75-1.47)
25-30	63 (0.14)	74 (0.16)	0.87 (0.62-1.22)
≥30	43 (0.09)	59 (0.13)	0.73 (0.49-1.09)
Hút thuốc lá			
Không hút thuốc	159 (0.14)	178 (0.15)	0.90 (0.71-1.11)
Hiện hút thuốc	14 (0.14)	16 (0.17)	0.85 (0.41-1.74)

Chú thích: ¹ số ngoài ngoặc là số bệnh nhân bị gãy xương đùi trong thời gian theo dõi (7 năm) và số trong ngoặc là tỉ lệ gãy xương tính bằng phần trăm mỗi năm. ² Tỉ số nguy cơ tương đối (hay relative risk – RR – sẽ giải thích trong một chương sau) được ước tính bằng cách lấy tỉ lệ gãy xương trong nhóm can thiệp chia cho tỉ lệ trong nhóm giả dược; nếu khoảng tin cậy 95% bao gồm 1 thì mức độ khác biệt giữa 2 nhóm không có ý nghĩa thống kê; nếu khoảng tin cậy 95% không bao gồm 1 thì mức độ khác biệt giữa 2 nhóm được xem là có ý nghĩa thống kê (hay $p < 0.05$).

Xin nhắc lại rằng trong mỗi lần thử nghiệm một giả thuyết, chúng ta chấp nhận một sai sót 5% (giả dụ chúng ta chấp nhận tiêu chuẩn $p = 0.05$ để tuyên bố có ý nghĩa hay không có ý nghĩa thống kê). Vấn đề đặt ra là trong bối cảnh thử nghiệm nhiều giả thuyết là như sau: **nếu trong số n thử nghiệm, chúng ta tuyên bố k thử nghiệm “có ý nghĩa thống kê” (tức là $p < 0.05$), thì xác suất có ít nhất một giả thuyết sai là bao nhiêu?**

Để trả lời câu hỏi này tôi sẽ bắt đầu bằng một ví dụ đơn giản. Mỗi thử nghiệm chúng ta chấp nhận một xác suất sai lầm là 0.05. Nói cách khác, chúng ta có xác suất đúng là 0.95. Nếu chúng ta thử nghiệm 3 giả thuyết, xác suất mà chúng ta đúng cả ba là [dĩ nhiên]: $0.95 \times 0.95 \times 0.95 = 0.8574$. Như vậy, xác xuất có ít nhất một sai lầm trong ba tuyên bố “có ý nghĩa thống kê” là: $1 - 0.8574 = 0.1426$ (tức khoảng 14%).

Nói chung, nếu chúng ta thử nghiệm n giả thuyết, và mỗi lần thử nghiệm chúng ta chấp nhận một xác suất sai lầm là p , thì xác suất có ít nhất 1 sai lầm trong n lần thử nghiệm đó là $1 - (1 - p)^n$. Khi $n = 10$ và $p = 0.05$ thì xác suất có ít nhất một sai lầm lên đến: 40%.

“Bài học” rút ra từ cách lí giải trên là như sau: nếu chúng ta đọc một bài báo khoa học mà trong đó nhà nghiên cứu tiến hành nhiều thử nghiệm khác nhau với các kết quả trị số $p < 0.05$, chúng ta có lí do để cho rằng xác suất mà một trong những cái-gọi-là

“significant” (hay “có ý nghĩa thống kê”) đó rất cao. Chúng ta cần phải dè dặt với những kết quả phân tích như thế.

Đối với một người làm nghiên cứu, ý nghĩa của vấn đề thử nghiệm nhiều giả thuyết là: không nên “câu cá”. Tôi xin nói thêm về khái niệm “câu cá” trong khoa học. Hãy tưởng tượng, một nhà nghiên cứu muốn tìm hiểu hiệu quả của một thuật điều trị mới cho các bệnh nhân đau khớp. Sau khi xem xét các nghiên cứu đã công bố trong y văn, nhà nghiên cứu quyết định tiến hành một nghiên cứu trên 300 bệnh nhân: phân nửa được điều trị bằng thuật mới, phân nửa chỉ sử dụng giả dược. Sau thời gian theo dõi, thu thập dữ liệu, nhà nghiên cứu phân tích và phát hiện sự khác biệt giữa hai nhóm không có ý nghĩa thống kê. Nói cách khác, thuật điều trị không có hiệu quả. Nhà nghiên cứu không chịu “đầu hàng”, nên tìm cách tìm cho được một kết quả có ý nghĩa thống kê. Ông chia bệnh nhân thành nhiều nhóm theo độ tuổi (trên 50 hay dưới 50), theo giới tính (nam hay nữ), thành phần kinh tế (có thu nhập cao hay thấp), và thói quen (choi thể thao hay không). Tính chung, ông có 16 nhóm khác nhau, và có thể thử nghiệm 16 lần. Ông “khám phá” thuật điều trị có ý nghĩa thống kê trong nhóm phụ nữ tuổi trên 50 và có thu nhập cao. Và, ông công bố kết quả. Đó là một qui trình làm việc mà giới nghiên cứu khoa học gọi là “fishing expedition” (một chuyến đi câu cá). Tất nhiên, một kết quả như thế không có giá trị khoa học và không thể tin được. (Với 16 thử nghiệm khác nhau và với $p = 0.05$, xác suất mà một thử nghiệm có kết quả “significant” lên đến 55%, do đó chúng ta chẳng ngạc nhiên khi thấy có một “con cá” được bắt!)

Để cho kết quả trị số P có ý nghĩa nguyên thủy của nó trong bối cảnh thử nghiệm nhiều giả thuyết, các nhà nghiên cứu đề nghị sử dụng thuật điều chỉnh Bonferroni (tên của một nhà thống kê học người Ý từng đề nghị cách làm này). Theo đề nghị này, **trước khi** tiến hành nghiên cứu, nhà nghiên cứu phải xác định rõ giả thuyết nào là chính, và giả thuyết nào là phụ. Ngoài ra, nhà nghiên cứu còn phải đề ra kế hoạch sẽ thử nghiệm bao nhiêu giả thuyết **trước khi bắt tay vào phân tích dữ liệu**. Chẳng hạn như nếu nhà nghiên cứu có kế hoạch thử nghiệm 20 so sánh và muốn giữ cho trị số p ở 0.05, thì thay vì dựa vào 0.05 là tiêu chuẩn để tuyên bố “significant”, nhà nghiên cứu phải dựa vào tiêu chuẩn 0.0025 (tức lấy 0.05 chia cho 20) để tuyên bố “significant”. Nói cách khác, chỉ khi nào một kết quả có trị số p thấp hơn 0.0025 (hay nói chung là p/n) thì nhà nghiên cứu mới có “quyền” tuyên bố kết quả đó có ý nghĩa thống kê.

Trị số P, dù cực kì thông dụng trong nghiên cứu khoa học, không phải là một phán xét cuối cùng của một công trình nghiên cứu hay một giả thuyết. Thế nhưng trong thực tế, các nhà khoa học đã quá lệ thuộc vào trị số P để suy luận trong nghiên cứu và tuyên bố những khám phá mà sau này được chứng minh là sai lầm. Có thể nói không ngoa rằng chính vì sự lạm dụng và phụ thuộc một cách mù quáng vào trị số P mà khoa học, nhất là y sinh học, đã trở nên nghèo nàn. Hàng ngày chúng ta đọc hay nghe những phát hiện khoa học trái ngược nhau (như lúc thì có nghiên cứu cho thấy cà phê có tác dụng tốt cho sức khỏe, lúc khác có nghiên cứu cho biết cà phê có hại cho sức khỏe; hay lúc thì thuốc giảm đau aspirin có hiệu năng làm giảm nguy cơ ung thư, nhưng mới đây có nghiên cứu cho thấy aspirin có thể làm tăng nguy cơ bị ung thư vú, v.v...). Có khi công chúng không biết phát hiện nào là thực và phát hiện nào là “dương tính giả”. Theo phân

tích của Berger và Sellke, khoảng 25% các phát hiện với “ $p < 0.05$ ” là các phát hiện dương tính giả [2].

Do đó, chúng ta không nên quá phụ thuộc vào trị số P . Không phải cứ nghiên cứu nào với $p < 0.05$ là thành công và $p > 0.05$ là thất bại. Có khi một phát hiện với $p > 0.05$ nhưng lại là một phát hiện có ý nghĩa. Vấn đề quan trọng là làm sao để ước tính mức độ khả dĩ của một giả thuyết một khi có dữ kiện thật trong tay, tức là ước tính $P(H+ | D)$. Để ước tính $P(H+ | D)$, chúng ta phải áp dụng Định lí Bayes, và cách tiếp cận định lí này không nằm trong phạm trù của cuốn sách này. Bạn đọc muốn tham khảo thêm có thể đọc một vài bài báo của tôi hay các bài báo của James Berger mà tài liệu tham khảo dưới đây có thể cung cấp thêm.

Tài liệu tham khảo:

- [1] Wulff et al., *Statistics in Medicine* 1987; 6:3-10.
- [2] Berger JO, Sellke T. Testing a point null hypothesis: the irreconcilability of P-values and evidence. *Journal of the American Statistical Association* 1987; 82:112-20.

8

Phân tích số liệu bằng biểu đồ

Yếu tố thị giác rất quan trọng. Người Trung Quốc có câu “một biểu đồ có giá trị bằng cả vạn chữ viết”. 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 cực kì quan trọng trong phân tích thống kê. Có thể nói, không có đồ thị là phân tích thống kê 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* hay *trellis* có thể tải từ website của R. Trong chương này tôi sẽ chỉ cách vẽ các biểu đồ thông dụng bằng cách sử dụng các hàm phổ biến trong R.

8.1 Môi trường và thiết kế biểu đồ

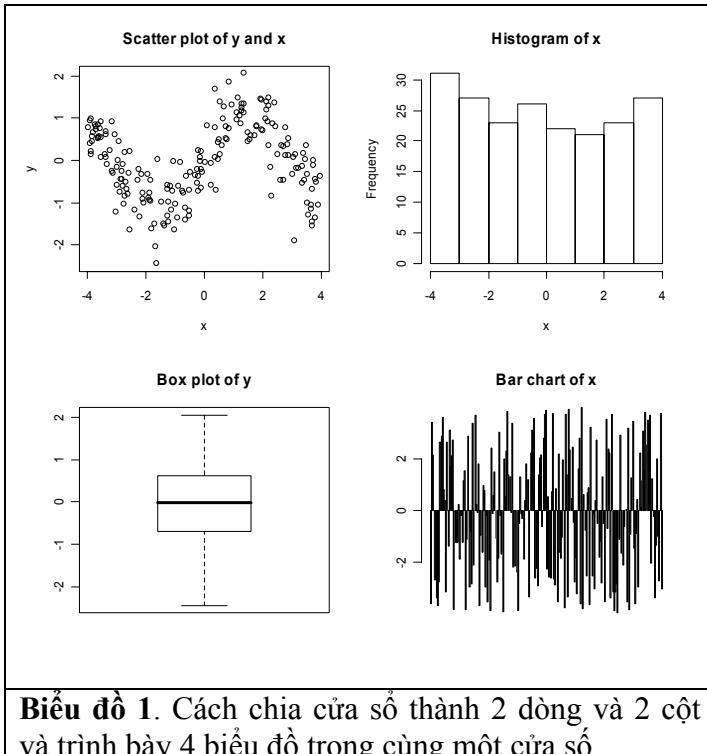
8.1.1 Nhiều biểu đồ cho một cửa sổ (windows)

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 một 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))
```



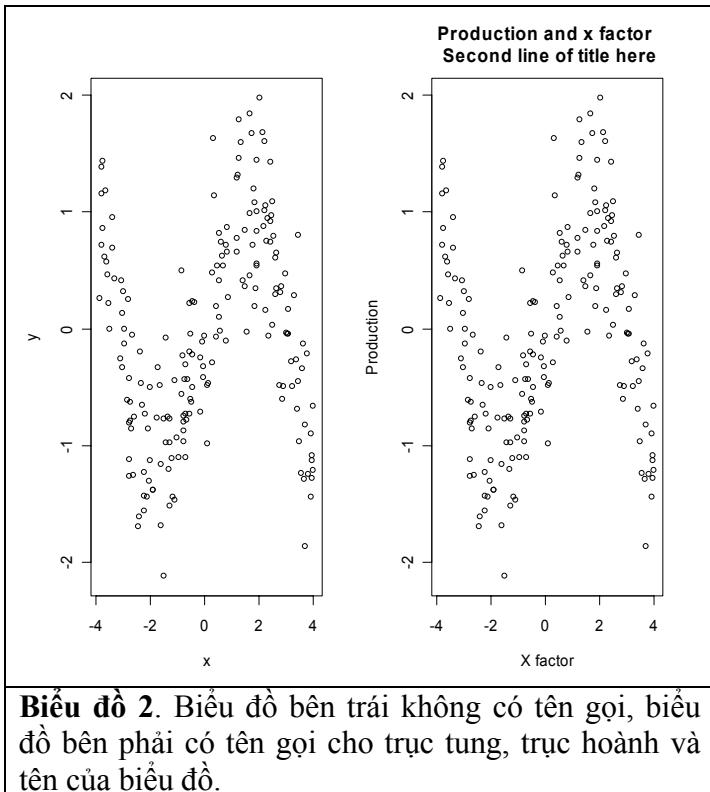
Biểu đồ 1. Cách chia cửa sổ thành 2 dòng và 2 cột và trình bày 4 biểu đồ trong cùng một cửa sổ.

8.1.2 Đặt tên cho trục tung và trục hoành

Biểu đồ thường có trục tung (y-axis) và trục hoành. Vì dữ liệu thường được gọi bằng các chữ viết tắt, cho nên biểu đồ cần phải có tên cho từng biến để dễ theo dõi. Trong ví dụ sau đây, biểu đồ bên trái không có tên mà chỉ dùng tên của biến gốc (tức x và y), còn bên phải có tên dễ hiểu hơn.

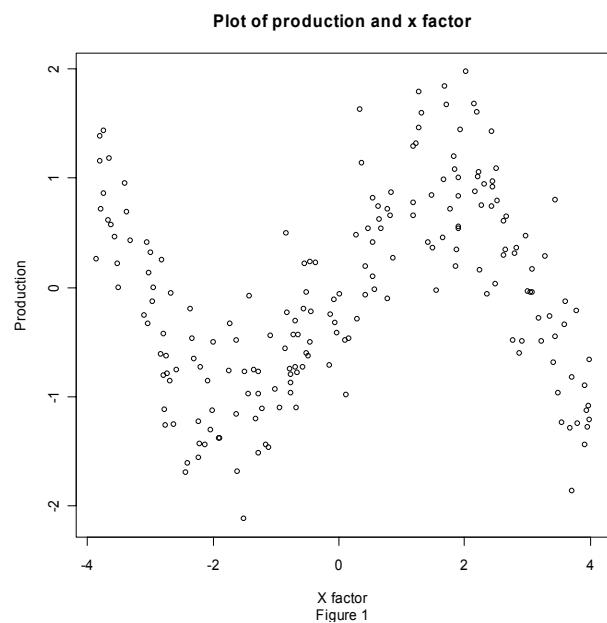
```
> par(mfrow=c(1,2))
> N <- 200
> x <- runif(N, -4, 4)
> y <- sin(x) + 0.5*rnorm(N)
> plot(x,y)
> plot(x, y, xlab="X factor",
       ylab="Production",
       main="Production and x factor \n Second line of title here")
> par(mfrow=c(1,1))
```

Trong các lệnh trên, `xlab` (viết tắt từ x label) và `ylab` (viết tắt từ y label) dùng để đặt tên cho trục hoành và trục tung. Còn `main` được dùng để đặt tên cho biểu đồ. Chú ý rằng trong `main` có kí hiệu `\n` dùng để viết dòng thứ hai (nếu tên gọi biểu đồ quá dài).



Ngoài ra, chúng ta còn có thể sử dụng hàm title và sub để đặt tên:

```
> plot(x, y, xlab="Time",
      ylab="Production")
> title(main="Plot of production and x factor",
      sub="Figure 1")
```



8.1.3 Cho giới hạn của trục tung và trục hoành

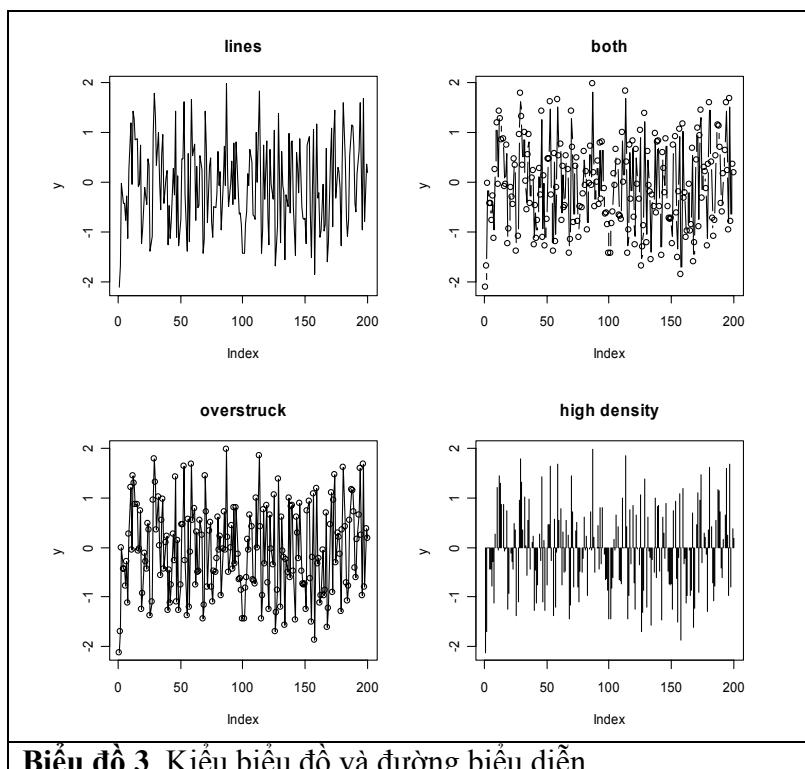
Nếu không cung cấp giới hạn của trục tung và trục hoành, R sẽ tự động tìm điều chỉnh và cho các số liệu này. Tuy nhiên, chúng ta cũng có thể kiểm soát biểu đồ bằng cách sử dụng `xlim` và `ylim` để cho R biết cụ thể giới hạn của hai trục này:

```
> plot(x, y, xlab="X factor",
       ylab="Production",
       main="Plot of production and x factor",
       xlim=c(-5, 5),
       ylim=c(-3, 3))
```

8.1.4 Thể loại và đường biểu diễn

Trong một dãy biểu đồ, chúng ta có thể yêu cầu R vẽ nhiều kiểu và đường biểu diễn khác nhau.

```
> par(mfrow=c(2,2))
> plot(y, type="l"); title("lines")
> plot(y, type="b"); title("both")
> plot(y, type="o"); title("overstruck")
> plot(y, type="h"); title("high density")
```



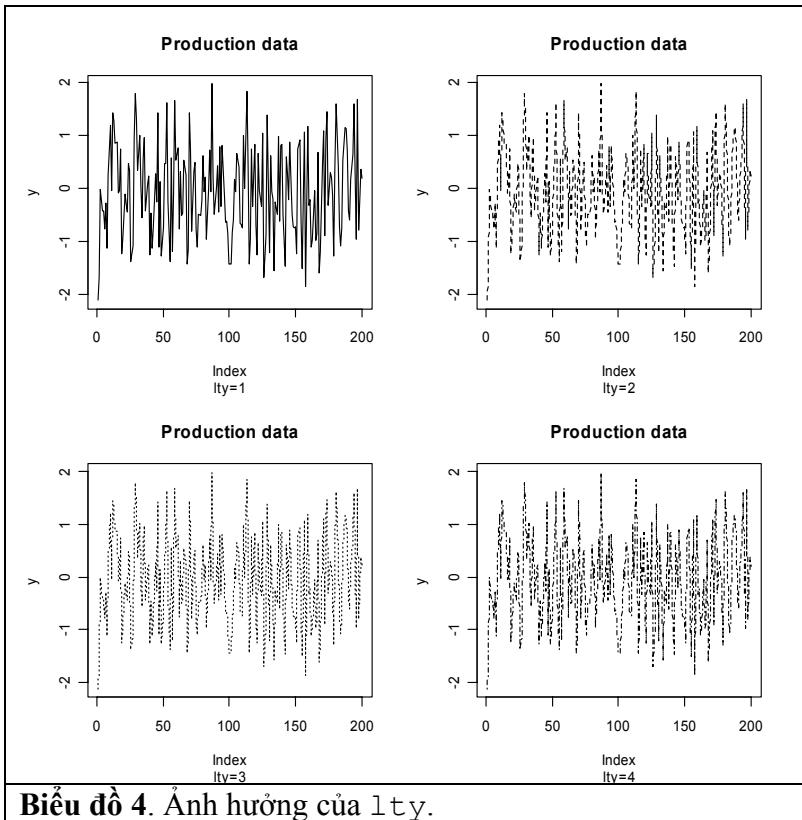
Biểu đồ 3. Kiểu biểu đồ và đường biểu diễn.

Ngoài ra, chúng ta cũng có thể nhiều đường biểu diễn bằng `lty` như sau:

```

> par(mfrow=c(2,2))
> plot(y, type="l", lty=1); title(main="Production data", sub="lty=1")
> plot(y, type="l", lty=2); title(main="Production data", sub="lty=2")
> plot(y, type="l", lty=3); title(main="Production data", sub="lty=3")
> plot(y, type="l", lty=4); title(main="Production data", sub="lty=4")

```



8.1.5 Màu sắc, khung, và kí hiệu

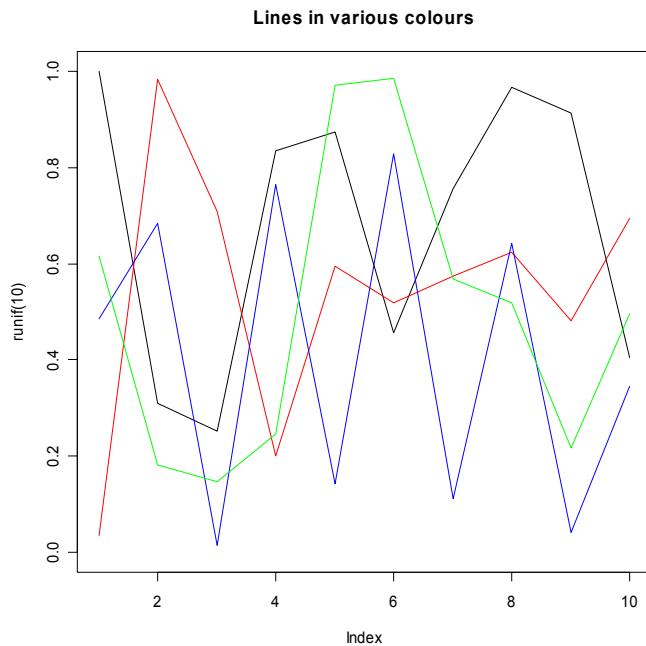
Chúng ta có thể kiểm soát màu sắc của một biểu đồ bằng lệnh `col`. Giá trị mặc định của `col` là 1. Tuy nhiên, chúng ta có thể thay đổi các màu theo ý muốn hặc bằng cách cho số hoặc bằng cách viết ra tên màu như "red", "blue", "green", "orange", "yellow", "cyan", v.v...

Ví dụ sau đây dùng một hàm để vẽ ba đường biểu diễn với ba màu đỏ, xanh nước biển, và xanh lá cây:

```

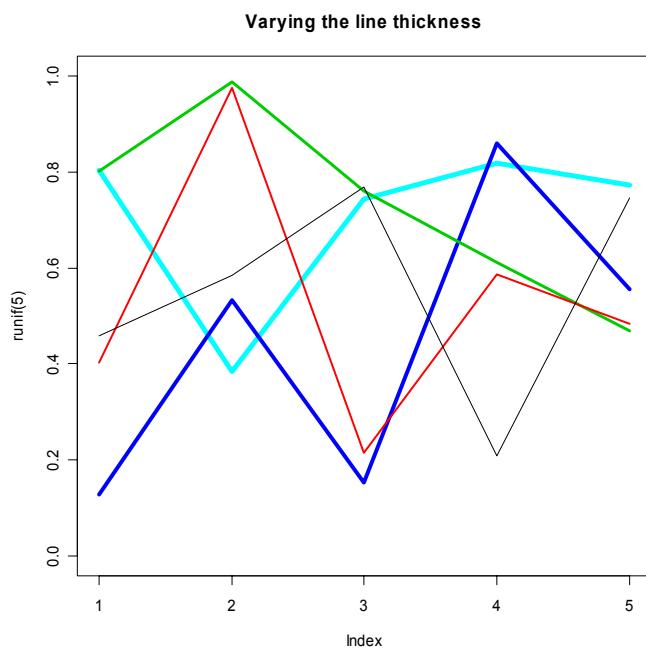
> plot(runif (10), ylim=c(0,1), type='l')
> for (i in c('red', 'blue', 'green'))
> {
>   lines(runif (10), col=i )
> }
> title(main="Lines in various colours")

```



Ngoài ra, chúng ta còn có thể vẽ đường biểu diễn bằng cách tăng bè dày của mỗi đường:

```
> plot(runif(5), ylim=c(0,1), type='n')
> for (i in 5:1)
{ 
  lines( runif(5), col=i, lwd=i )
}
> title(main="Varying the line thickness")
```



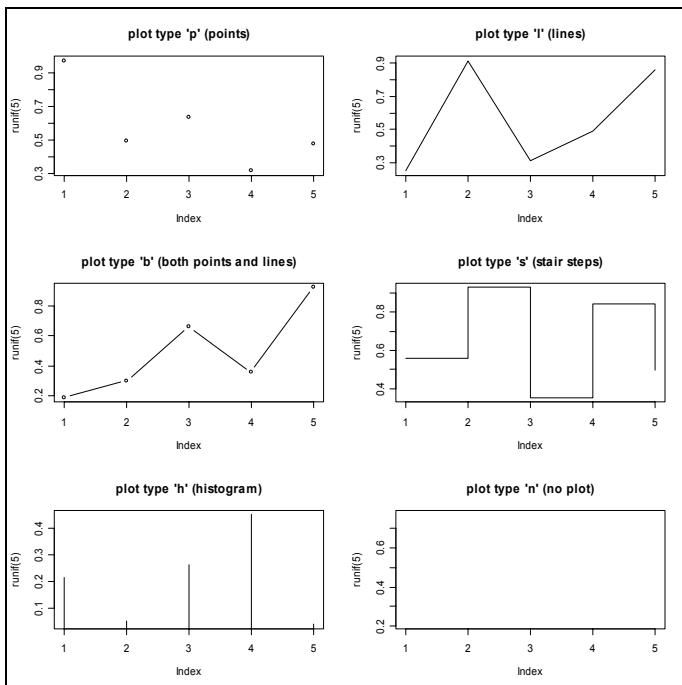
Hình dạng của biểu đồ cũng có thể thay đổi bằng `type` như sau:

```
> op <- par(mfrow=c(3, 2))
```

```

> plot(runif(5), type = 'p',
       main = "plot type 'p' (points)")
> plot(runif(5), type = 'l',
       main = "plot type 'l' (lines)")
> plot(runif(5), type = 'b',
       main = "plot type 'b' (both points and lines)")
> plot(runif(5), type = 's',
       main = "plot type 's' (stair steps)")
> plot(runif(5), type = 'h',
       main = "plot type 'h' (histogram)")
> plot(runif(5), type = 'n',
       main = "plot type 'n' (no plot)")
> par(op)

```

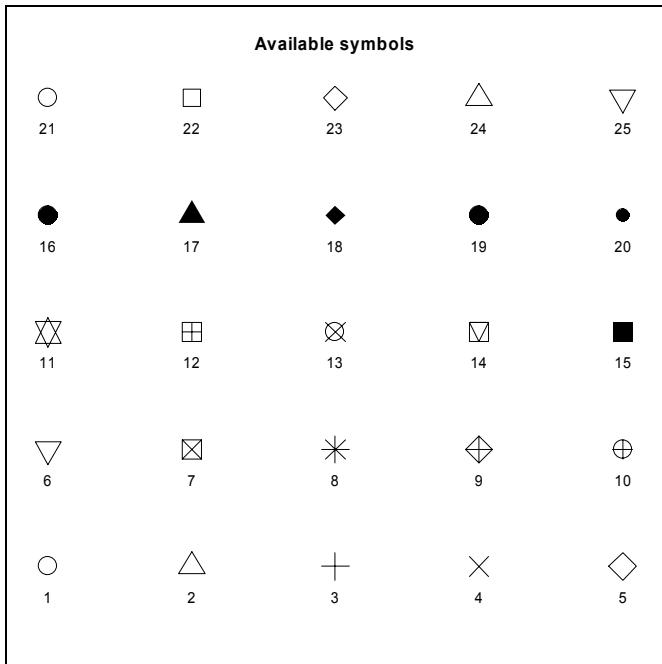


Khung biểu đồ có thể kiểm soát bằng lệnh `bty` với các thông số như sau:

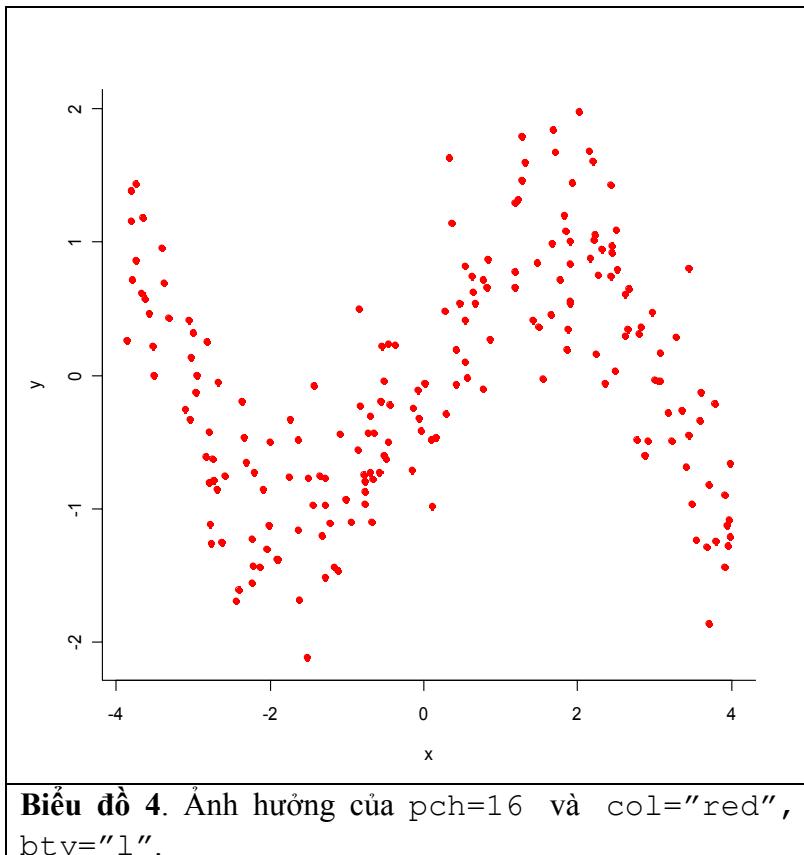
<code>bty="n"</code>	Không có vòng khung chung quanh biểu đồ
<code>bty="o"</code>	Có 4 khung chung quanh biểu đồ
<code>bty="c"</code>	Vẽ một hộp gồm 3 cạnh chung quanh biểu đồ theo hình chữ C
<code>bty="l"</code>	Vẽ hộp 2 cạnh chung quanh biểu đồ theo hình chữ L
<code>bty="7"</code>	Vẽ hộp 2 cạnh chung quanh biểu đồ theo hình số 7

Cách hay nhất để bạn đọc làm quen với các cách vẽ biểu đồ này là bằng cách thử trên R để biết rõ hơn.

Kí hiệu của một biểu đồ cũng có thể thay thế bằng cách cung cấp số cho `pch` (plotting character) trong R. Các kí hiệu thông dụng là:



```
> plot(x, y, col="red", pch=16, bty="l")
```

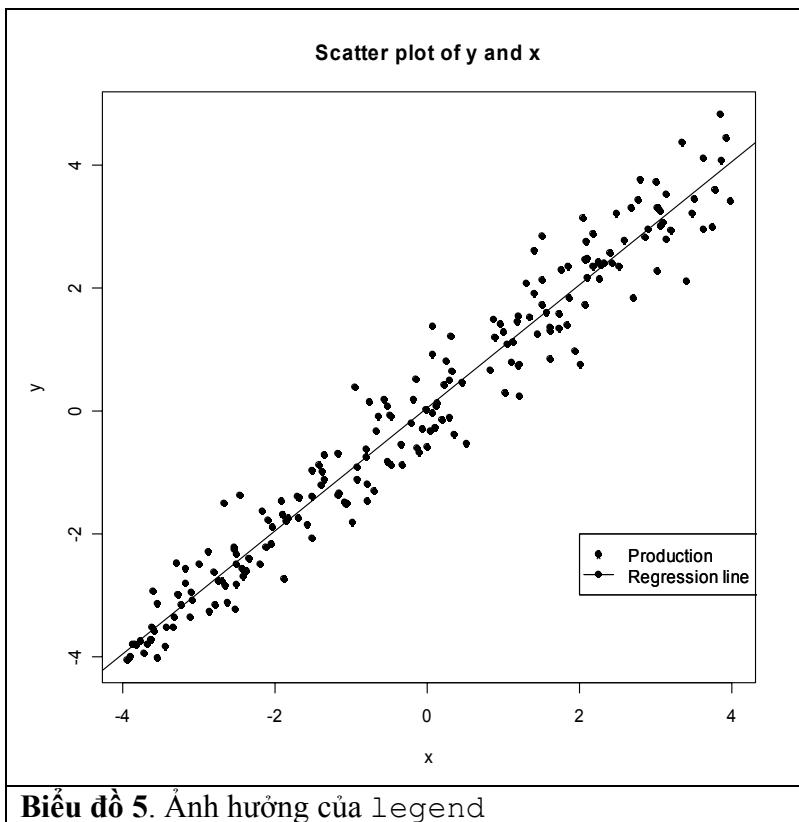


8.1.6 Ghi chú (legend)

Hàm legend rất có ích cho việc ghi chú một biểu đồ và giúp người đọc hiểu được ý nghĩa của biểu đồ tốt hơn. Cách sử dụng legend có thể minh họa bằng ví dụ sau đây:

```
> N <- 200  
> x <- runif(N, -4, 4)  
> y <- x + 0.5*rnorm(N)  
> plot(x,y, pch=16, main="Scatter plot of y and x")  
> reg <- lm(y~x)  
> abline(reg)  
> legend(2,-2, c("Production","Regression line"), pch=16, lty=c(0,1))
```

Thông số legend(2, -2) có nghĩa là đặt phần ghi chú vào trực hoành (x-axis) bằng 2 và trực tung (y-axis) bằng -2.

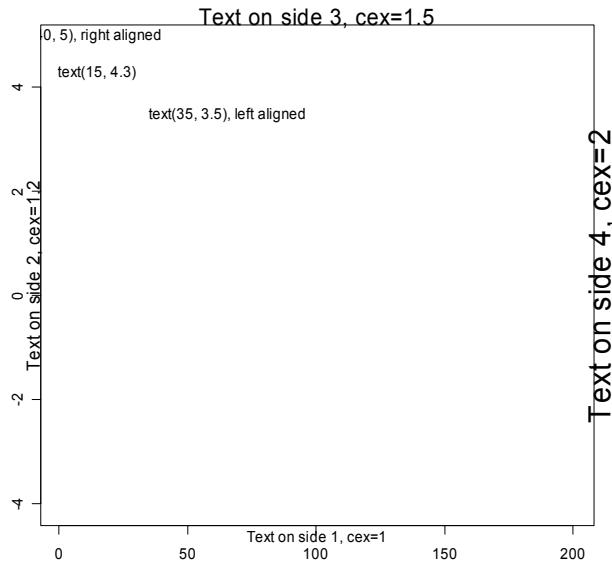


8.1.7 Viết chữ trong biểu đồ

Phần lớn 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 hàm `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")
```



8.1.8 Đặt kí hiệu vào biểu đồ. `abline()` có thể sử dụng để vẽ một đường thẳng, với những thông số như sau:

`abline(a, b)` : đường hồi qui tuyến tính $a=\text{intercept}$ và $b=\text{slope}$.

`abline(h=30)` vẽ một đường ngang tại $y=30$.

`abline(v=12)` vẽ một đường thẳng đứng tại điểm $x=12$.

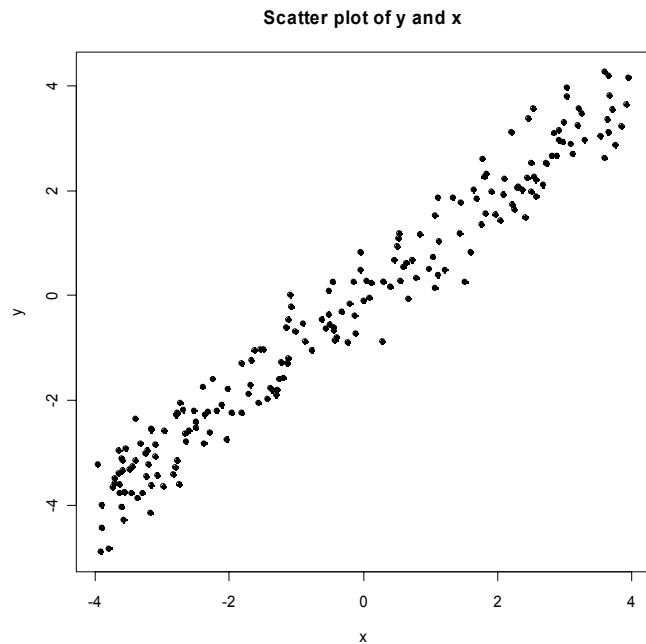
Ngoài ra, chúng ta còn có thể cho vào biểu đồ một mũi tên để ghi chú một điểm số liệu nào đó.

```
> N <- 200
```

```

> x <- runif(N, -4, 4)
> y <- x + 0.5*rnorm(N)
> plot(x,y, pch=16, main="Scatter plot of y and x")

```

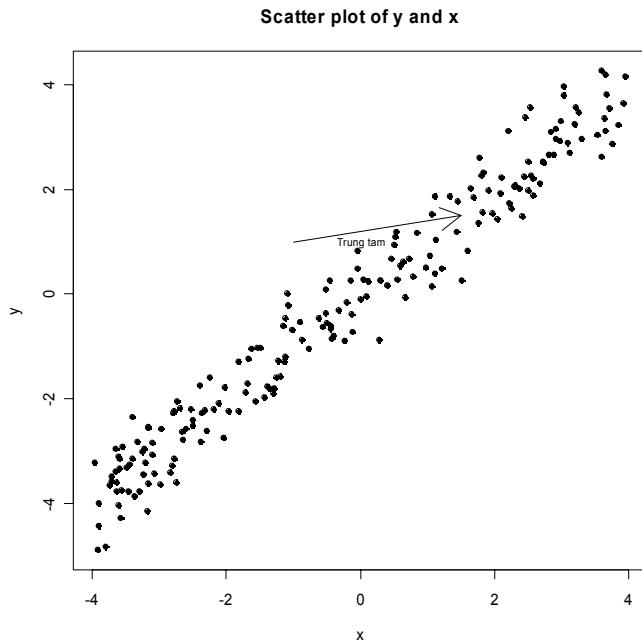


Giả sử chúng ta muốn ghi chú ngay tại $x=0$ và $y=0$ là điểm trung tâm, chúng ta trước hết dùng arrows để vẽ mũi tên. Trong lệnh sau đây, `arrows(-1, 1, 1.5, 1.5)` có nghĩa như sau tọa độ $x=-1$, $y=1$ bắt đầu vẽ mũi tên và chấm dứt tại tọa độ $x=1.5$, $y=1.5$. Phần `text(0, 1)` yêu cầu R viết chữ tại tọa độ $x=0$, $y=1$.

```

> arrows(-1, 1.0, 1.5, 1.5)
> text(0, 1, "Trung tam", cex=0.7)

```



8.2 Số liệu cho phân tích biểu đồ

Sau khi đã biết qua môi trường và những lựa chọn để thiết kế một biểu đồ, bây giờ chúng ta có thể sử dụng một số hàm thông dụng để vẽ các biểu đồ cho số liệu. Theo tôi, biểu đồ có thể chia thành 2 loại chính: biểu đồ dùng để mô tả một biến số và biểu đồ về mối liên hệ giữa hai hay nhiều biến số. Tất nhiên, biến số có thể là liên tục hay không liên tục, cho nên, trong thực tế, chúng ta có 4 loại biểu đồ. Trong phần sau đây, tôi sẽ điểm qua các loại biểu đồ, từ đơn giản đến phức tạp.

Có lẽ cách tốt nhất để tìm hiểu cách vẽ đồ thị bằng R là bằng một dữ liệu thực tế. Tôi sẽ quay lại **ví dụ 2** trong chương trước. Trong ví dụ đó, chúng ta có dữ liệu gồm 8 cột (hay biến số): id, sex, age, bmi, hdl, ldl, tc, và tg. (Chú ý, id là mã số của 50 đối tượng nghiên cứu; sex là giới tính (nam hay nữ); age là độ tuổi; bmi là tỉ số trọng lượng; hdl là high density cholesterol; ldl là low density cholesterol; tc là tổng số - total – cholesterol; và tg triglycerides). Dữ liệu được chứa trong directory directory c:\\works\\insulin dưới tên chol.txt. Trước khi vẽ đồ thị, chúng ta bắt đầu bằng cách nhập dữ liệu này vào R.

```
> setwd("c:/works/stats")
> cong <- read.table("chol.txt", header=TRUE, na.strings=".")
> attach(cong)
```

Hay để tiện việc theo dõi tôi sẽ nhập các dữ liệu đó bằng các lệnh sau đây:

```
sex <- c("Nam", "Nu", "Nu", "Nam", "Nam", "Nu", "Nam", "Nam", "Nu",
       "Nu", "Nam", "Nu", "Nam", "Nam", "Nu", "Nu", "Nu", "Nu",
       "Nu", "Nu", "Nu", "Nu", "Nu", "Nam", "Nu", "Nu", "Nu", "Nu",
       "Nu", "Nu", "Nu", "Nu", "Nu", "Nu", "Nu", "Nu", "Nu", "Nu",
       "Nu", "Nu", "Nu", "Nu", "Nu", "Nu", "Nu", "Nu", "Nu", "Nu",
       "Nu", "Nu", "Nu", "Nu", "Nu", "Nu", "Nu", "Nu", "Nu", "Nu")
```

```

"Nam", "Nu", "Nam", "Nam", "Nu", "Nam", "Nam", "Nu", "Nu")

age <- c(57, 64, 60, 65, 47, 65, 76, 61, 59, 57,
       63, 51, 60, 42, 64, 49, 44, 45, 80, 48,
       61, 45, 70, 51, 63, 54, 57, 70, 47, 60,
       60, 50, 60, 55, 74, 48, 46, 49, 69, 72,
       51, 58, 60, 45, 63, 52, 64, 45, 64, 62)

bmi <- c( 17, 18, 18, 18, 18, 19, 19, 19, 19, 20, 20, 20, 20, 20,
         20, 21, 21, 21, 21, 21, 21, 21, 22, 22, 22, 22, 22, 22,
         22, 22, 22, 22, 23, 23, 23, 23, 23, 23, 23, 23, 24, 24, 24,
         24, 24, 24, 25, 25)

hdl <- c(5.000,4.380,3.360,5.920,6.250,4.150,0.737,7.170,6.942,5.000,
        4.217,4.823,3.750,1.904,6.900,0.633,5.530,6.625,5.960,3.800,
        5.375,3.360,5.000,2.608,4.130,5.000,6.235,3.600,5.625,5.360,
        6.580,7.545,6.440,6.170,5.270,3.220,5.400,6.300,9.110,7.750,
        6.200,7.050,6.300,5.450,5.000,3.360,7.170,7.880,7.360,7.750)

ldl <- c(2.0, 3.0, 3.0, 4.0, 2.1, 3.0, 3.0, 3.0, 3.0, 2.0,
        5.0, 1.3, 1.2, 0.7, 4.0, 4.1, 4.3, 4.0, 4.3, 4.0,
        3.1, 3.0, 1.7, 2.0, 2.1, 4.0, 4.1, 4.0, 4.2, 4.2,
        4.4, 4.3, 2.3, 6.0, 3.0, 3.0, 2.6, 4.4, 4.3, 4.0,
        3.0, 4.1, 4.4, 2.8, 3.0, 2.0, 1.0, 4.0, 4.6, 4.0)

tc <-c (4.0, 3.5, 4.7, 7.7, 5.0, 4.2, 5.9, 6.1, 5.9, 4.0,
        6.2, 4.1, 3.0, 4.0, 6.9, 5.7, 5.7, 5.3, 7.1, 3.8,
        4.3, 4.8, 4.0, 3.0, 3.1, 5.3, 5.3, 5.4, 4.5, 5.9,
        5.6, 8.3, 5.8, 7.6, 5.8, 3.1, 5.4, 6.3, 8.2, 6.2,
        6.2, 6.7, 6.3, 6.0, 4.0, 3.7, 6.1, 6.7, 8.1, 6.2)

tg <- c(1.1, 2.1, 0.8, 1.1, 2.1, 1.5, 2.6, 1.5, 5.4, 1.9,
       1.7, 1.0, 1.6, 1.1, 1.5, 1.0, 2.7, 3.9, 3.0, 3.1,
       2.2, 2.7, 1.1, 0.7, 1.0, 1.7, 2.9, 2.5, 6.2, 1.3,
       3.3, 3.0, 1.0, 1.4, 2.5, 0.7, 2.4, 2.4, 1.4, 2.7,
       2.4, 3.3, 2.0, 2.6, 1.8, 1.2, 1.9, 3.3, 4.0, 2.5)

cong <- data.frame(sex, age, bmi, hdl, ldl, tc, tg)

```

Sau khi đã có số liệu, chúng ta sẵn sàng tiến hành phân tích số liệu bằng biểu đồ như sau:

8.3 Biểu đồ cho một biến số rời rạc (discrete variable): barplot

Biến sex trong dữ liệu trên có hai giá trị (nam và nu), tức là một biến không liên tục. Chúng ta muốn biết tần số của giới tính (bao nhiêu nam và bao nhiêu nữ) và vẽ một biểu đồ đơn giản. Để thực hiện ý định này, trước hết, chúng ta cần dùng hàm table để biết tần số:

```

> sex.freq <- table(sex)
> sex.freq
sex
Nam   Nu
22    28

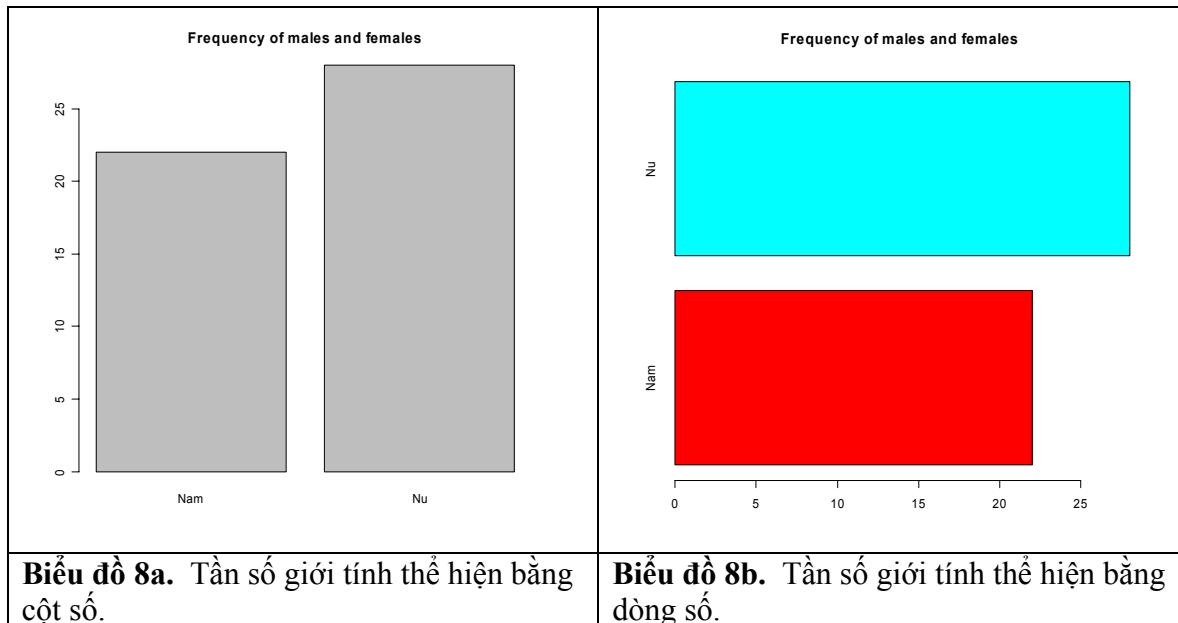
```

Có 22 nam và 28 nữ trong nghiên cứu. Sau đó dùng hàm `barplot` để thể hiện tần số này như sau:

```
> barplot(sex.freq, main="Frequency of males and females")
```

Biểu trên cũng có thể có được bằng một lệnh đơn giản hơn (**Biểu đồ 8a**):

```
> barplot(table(sex), main="Frequency of males and females")
```



Thay vì thể hiện tần số nam và nữ bằng 2 cột, chúng ta có thể thể hiện bằng hai dòng bằng thông số `horiz = TRUE`, như sau (xem kết quả trong **Biểu đồ 6b**):

```
> barplot(sex.freq,
          horiz = TRUE,
          col = rainbow(length(sex.freq)),
          main="Frequency of males and females")
```

8.4 Biểu đồ cho hai biến số rời rạc (discrete variable): `barplot`

Age là một biến số liên tục. Chúng ta có thể chia bệnh nhân thành nhiều nhóm dựa vào độ tuổi. Hàm `cut` có chức năng “cắt” một biến liên tục thành nhiều nhóm rời rạc. Chẳng hạn như:

```
> ageg <- cut(age, 3)
> table(ageg)
ageg
(42,54.7] (54.7,67.3] (67.3,80]
    19         24         7
```

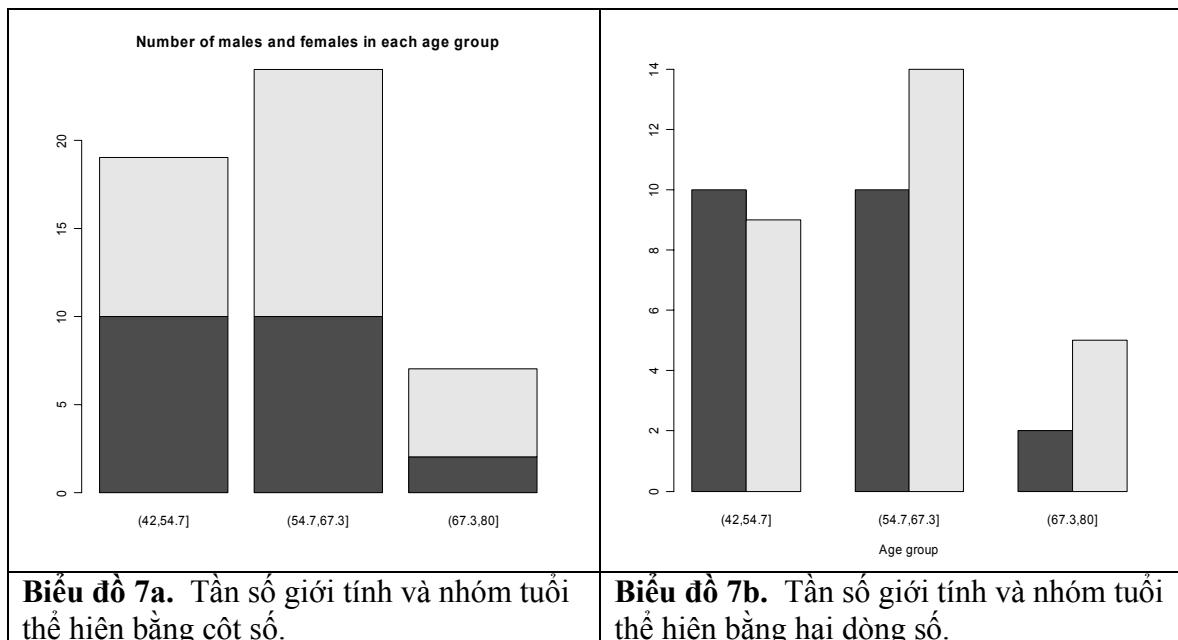
Có hiệu quả chia biến age thành 3 nhóm. Tần số của ba nhóm này là: 42 tuổi đến 54.7 tuổi thành nhóm 1, 54.7 đến 67.3 thành nhóm 2, và 67.3 đến 80 tuổi thành nhóm 3. Nhóm 1 có 19 bệnh nhân, nhóm 2 và 3 có 24 và 7 bệnh nhân.

Bây giờ chúng ta muốn biết có bao nhiêu bệnh nhân trong từng độ tuổi và từng giới tính bằng lệnh table:

```
> age.sex <- table(sex, ageg)
> age.sex
  ageg
sex   (42,54.7] (54.7,67.3] (67.3,80]
  Nam      10      10      2
  Nu       9       14      5
```

Kết quả trên cho thấy chúng ta có 10 bệnh nam và 9 nữ trong nhóm tuổi thứ nhất, 10 nam và 14 nữ trong nhóm tuổi thứ hai, v.v... Để thể hiện tần số của hai biến này, chúng ta vẫn dùng barplot:

```
> barplot(age.sex, main="Number of males and females in each age group")
```



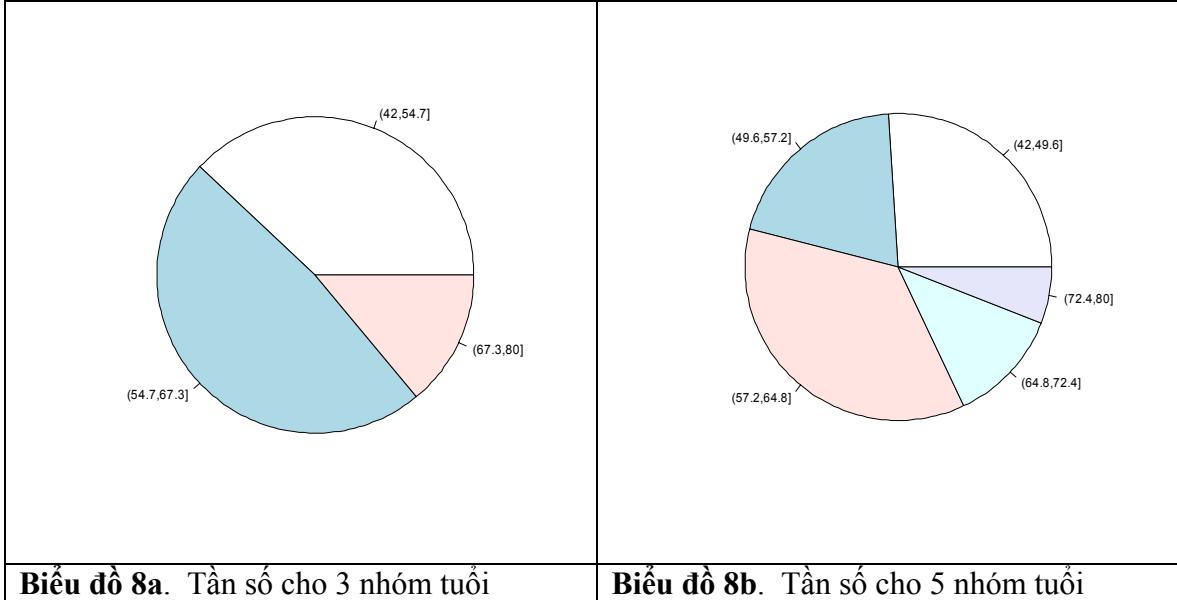
Trong **Biểu đồ 7a**, mỗi cột là cho một độ tuổi, và phần đậm của cột là nữ, và phần màu nhạt là tần số của nam giới. Thay vì thể hiện tần số nam nữ trong một cột, chúng ta cũng có thể thể hiện bằng 2 cột với beside=T như sau (**Biểu đồ 7b**):

```
barplot(age.sex, beside=TRUE, xlab="Age group")
```

8.5 Biểu đồ hình tròn

Tần số một biến rời rạc cũng có thể thể hiện bằng biểu đồ hình tròn. Ví dụ sau đây vẽ biểu đồ tần số của độ tuổi. **Biểu đồ 8a** là 3 nhóm độ tuổi, và **Biểu đồ 8b** là biểu đồ tần số cho 5 nhóm tuổi:

```
> pie(table(ageg))
pie(table(cut(age, 5)))
```

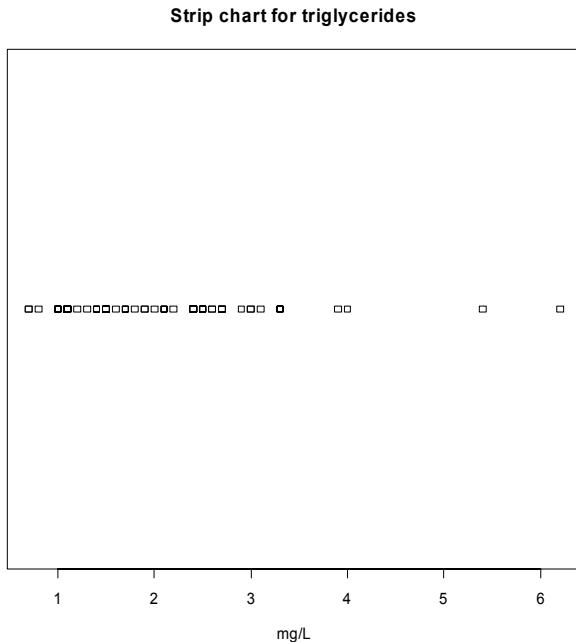


8.6 Biểu đồ cho một biến số liên tục: **stripchart** và **hist**

8.6.1 Stripchart

Biểu đồ strip cho chúng ta thấy tính liên tục của một biến số. Chẳng hạn như chúng ta muốn tìm hiểu tính liên tục của triglyceride (tg), hàm **stripchart()** sẽ giúp trong mục tiêu này:

```
> stripchart(tg,
             main="Strip chart for triglycerides", xlab="mg/L")
```

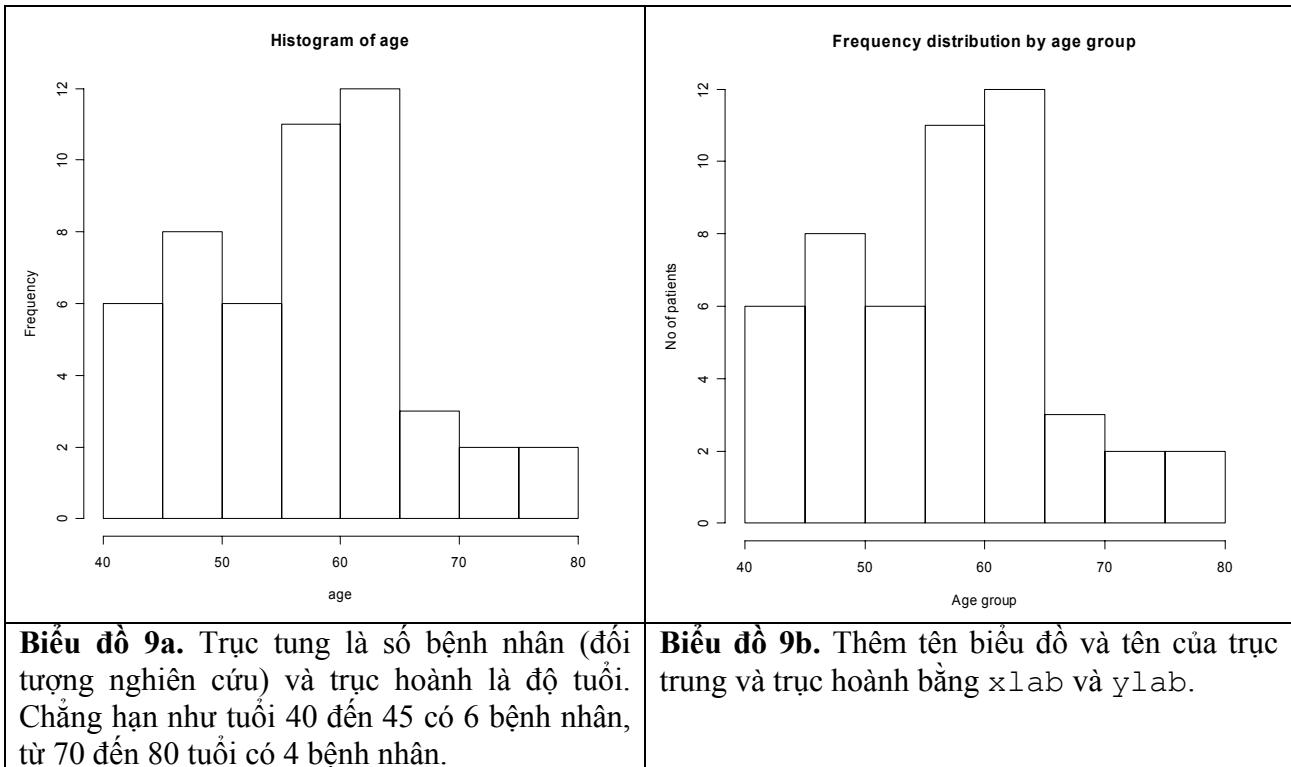


Chúng ta thấy biến số tg có sự bất liên tục, nhất là các đối tượng có tg cao. Trong khi phần lớn đối tượng có độ tg thấp hơn 5, thì có 2 đối tượng với tg rất cao (>5).

8.6.2 Histogram

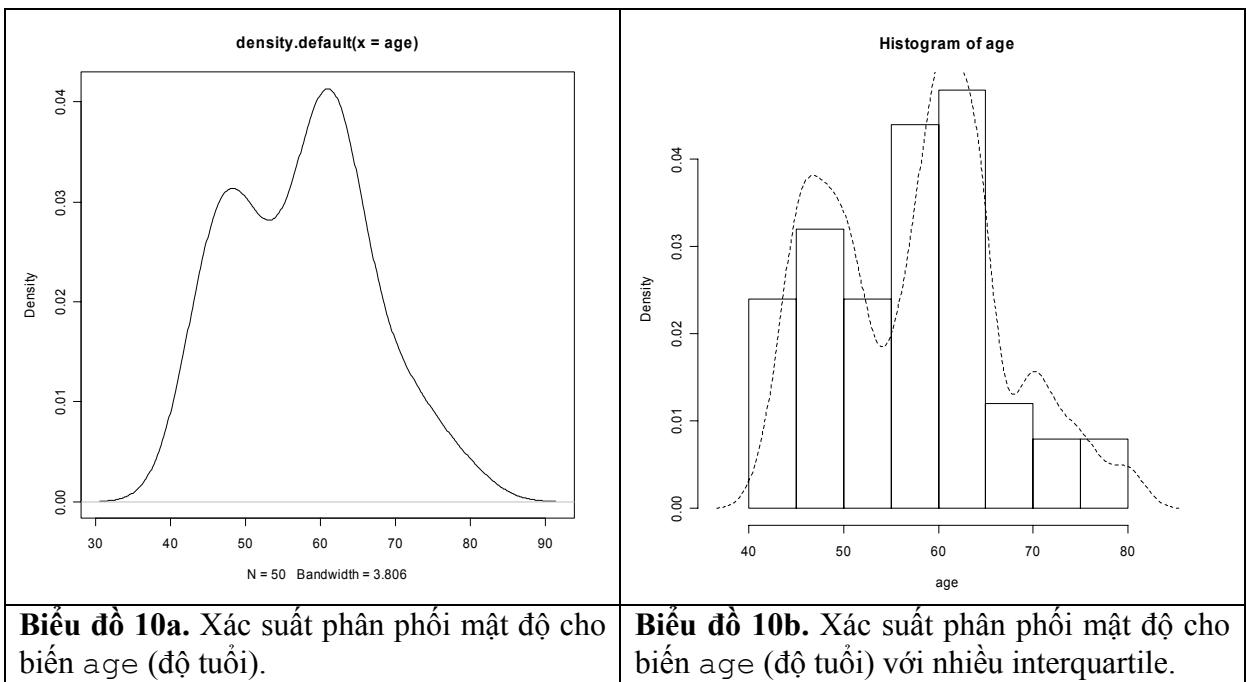
Age là một biến số liên tục. Để vẽ biểu đồ tần số của biến số age, chúng ta chỉ đơn giản lệnh `hist(age)`. Như đã đề cập trên, chúng ta có thể cải tiến đồ thị này bằng cách cho thêm tựa đề chính (`main`) và tựa đề của trục hoành (`xlab`) và trục tung (`ylab`):

```
> hist(age)
> hist(age, main="Frequency distribution by age group", xlab="Age
group", ylab="No of patients")
```



Chúng ta cũng có thể biến đổi biểu đồ thành một đồ thị phân phối xác suất bằng hàm plot (density) như sau (kết quả trong **Biểu đồ 10a**):

```
> plot(density(age), add=TRUE)
```

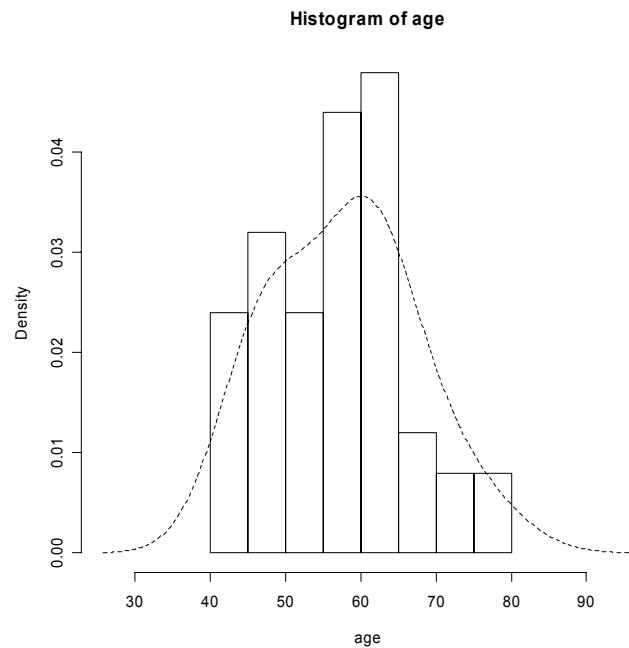


Chúng ta có thể vẽ hai đồ thị chồng lên bằng cách dùng hàm interquartile như sau (kết quả xem **Biểu đồ 10b**):

```
> iqr <- diff(summary(age) [c(2,5)])
> des <- density(age, width=0.5*iqr)
> hist(age, xlim=range(des$x), probability=TRUE)
> lines(des, lty=2)
```

Trong đồ thị trên, chúng ta dùng khoảng cách $0.5 \times \text{iqr}$ (tương đối “gần” nhau). Nhưng chúng ta có thể biến đổi thông số này thành $1.5 \times \text{iqr}$ để làm cho phân phối thực tế hơn:

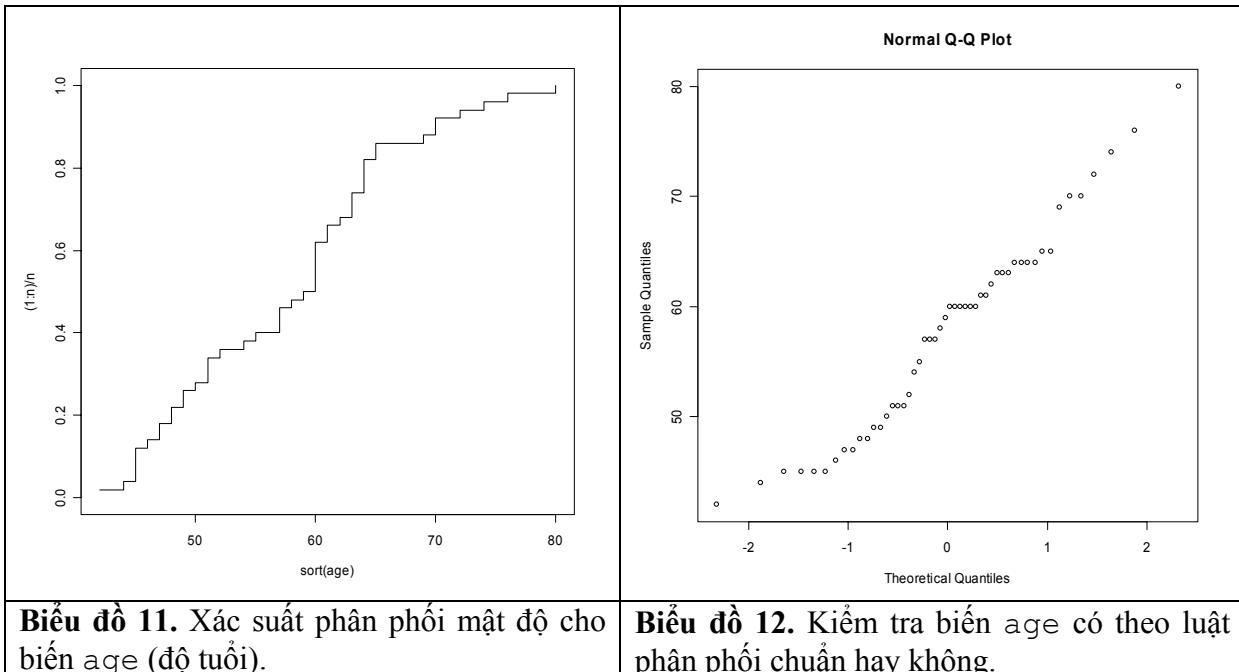
```
> iqr <- diff(summary(age) [c(2,5)])
> des <- density(age, width=1.5*iqr)
> hist(age, xlim=range(des$x), probability=TRUE)
> lines(des, lty=2)
```



Chúng ta có thể biến đổi biểu đồ thành một đồ thị phân phối xác suất tích lũy (cumulative distribution) bằng hàm `plot` và `sort` như sau:

```
> n <- length(age)
> plot(sort(age), (1:n)/n, type="s", ylim=c(0,1))
```

Kết quả được trình bày trong phần trái của biểu đồ sau đây (**Biểu đồ 11**).



Trong đồ thị trên, trục tung là xác suất tích lũy và trục hoành là độ tuổi từ thấp đến cao. Chẳng hạn như nhìn qua biểu đồ, chúng ta có thể thấy khoảng 50% đối tượng có tuổi thấp hơn 60.

Để biết xem phân phối của `age` có theo luật phân phối chuẩn (normal distribution) hay không chúng ta có thể sử dụng hàm `qqnorm`.

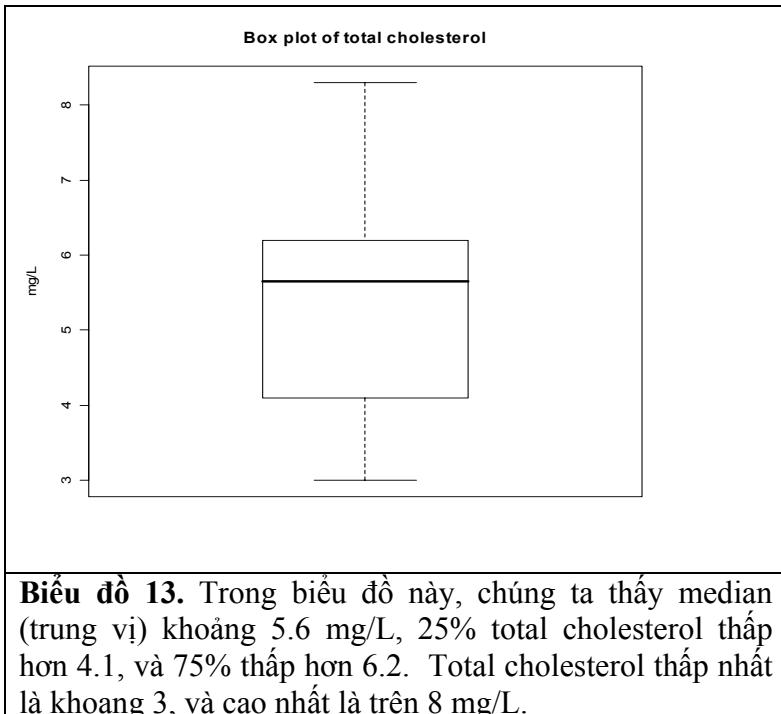
```
> qqnorm(age)
```

Trục hoành của biểu đồ trên là định lượng theo luật phân phối chuẩn (theoretical quantile) và trục hoành định lượng của số liệu (sample quantiles). Nếu phân phối của `age` theo luật phân phối chuẩn, thì đường biểu diễn phải theo một đường thẳng chéo 45 độ (tức là định lượng phân phối và định lượng số liệu bằng nhau). Nhưng qua **Biểu đồ 12**, chúng ta thấy phân phối của `age` không hẳn theo luật phân phối chuẩn.

8.6.3 Biểu đồ hộp (boxplot)

Để vẽ biểu đồ hộp của biến số `tc`, chúng ta chỉ đơn giản lệnh:

```
> boxplot(tc, main="Box plot of total cholesterol", ylab="mg/L")
```



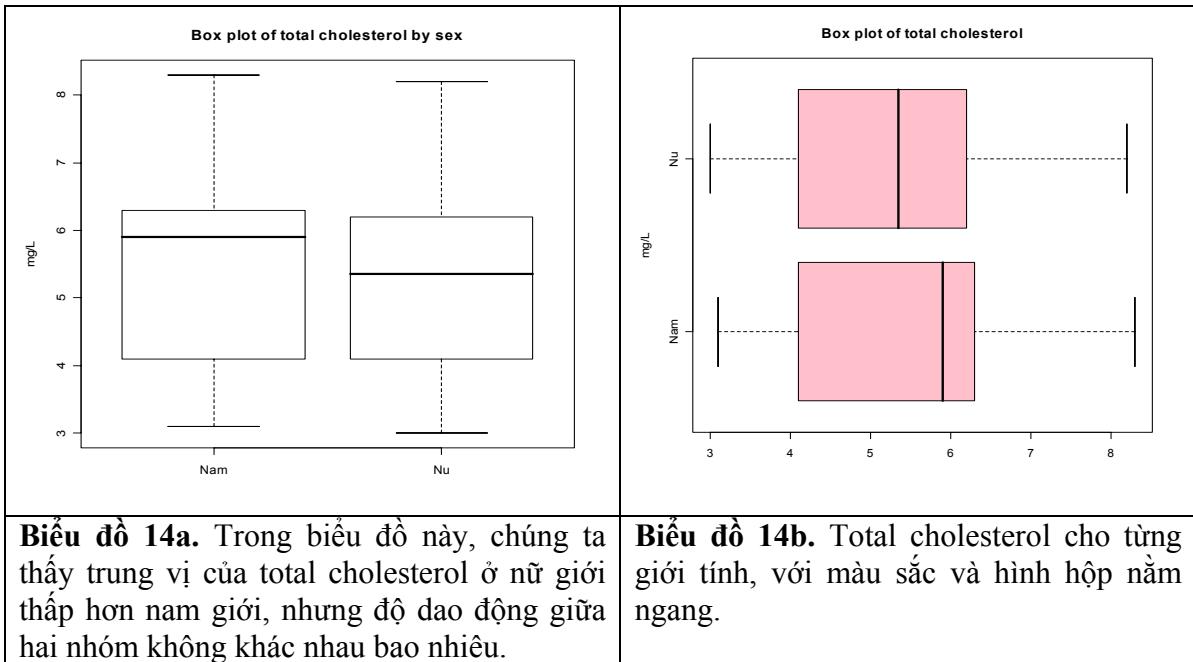
Biểu đồ 13. Trong biểu đồ này, chúng ta thấy median (trung vị) khoảng 5.6 mg/L, 25% total cholesterol thấp hơn 4.1, và 75% thấp hơn 6.2. Total cholesterol thấp nhất là khoang 3, và cao nhất là trên 8 mg/L.

Trong biểu đồ sau đây, chúng ta so sánh tc giữa hai nhóm nam và nữ:

```
> boxplot(tc ~ sex, main="Box plot of total cholesterol by sex",
  ylab="mg/L")
```

Kết quả trình bày trong **Biểu đồ 14a**. Chúng ta có thể biến đổi giao diện của đồ thị bằng cách dùng thông số horizontal=TRUE và thay đổi màu bằng thông số col như sau (**Biểu đồ 14b**):

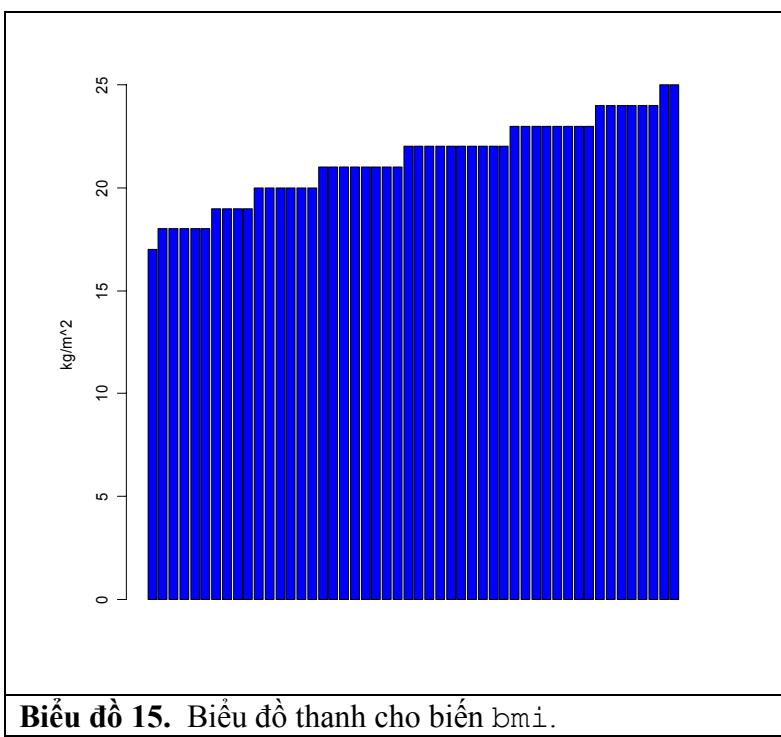
```
> boxplot(tc~sex, horizontal=TRUE, main="Box plot of total
cholesterol", ylab="mg/L", col = "pink")
```



8.6.4 Biểu đồ thanh (bar chart)

Để vẽ biểu đồ thanh của biến số `bmi`, chúng ta chỉ đơn giản lệnh:

```
> barplot(bmi, col="blue")
```

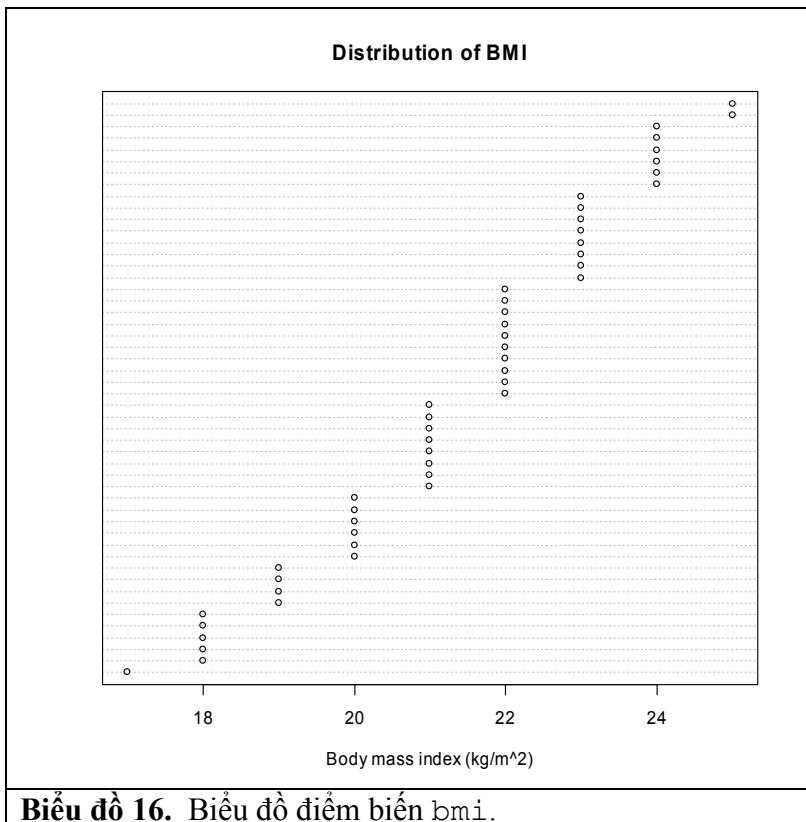


Biểu đồ 15. Biểu đồ thanh cho biến `bmi`.

8.6.5 Biểu đồ điểm (dotchart)

Một đồ thị khác cung cấp thông tin giống như barplot là dotchart:

```
> dotchart(bmi, xlab="Body mass index (kg/m^2)", main="Distribution of BMI")
```

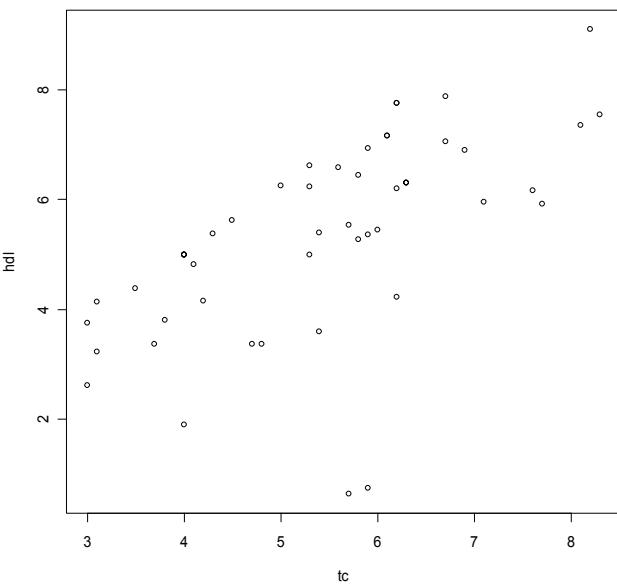


8.7 Phân tích biểu đồ cho hai biến liên tục

8.7.1 Biểu đồ tán xạ (scatter plot)

Để tìm hiểu mối liên hệ giữa hai biến, chúng ta dùng biểu đồ tán xạ. Để vẽ biểu đồ tán xạ về mối liên hệ giữa biến số `tc` và `hdl`, chúng ta sử dụng hàm `plot`. Thông số thứ nhất của hàm `plot` là trục hoành (x-axis) và thông số thứ 2 là trục tung. Để tìm hiểu mối liên hệ giữa `tc` và `hdl` chúng ta đơn giản lệnh:

```
> plot(tc, hdl)
```



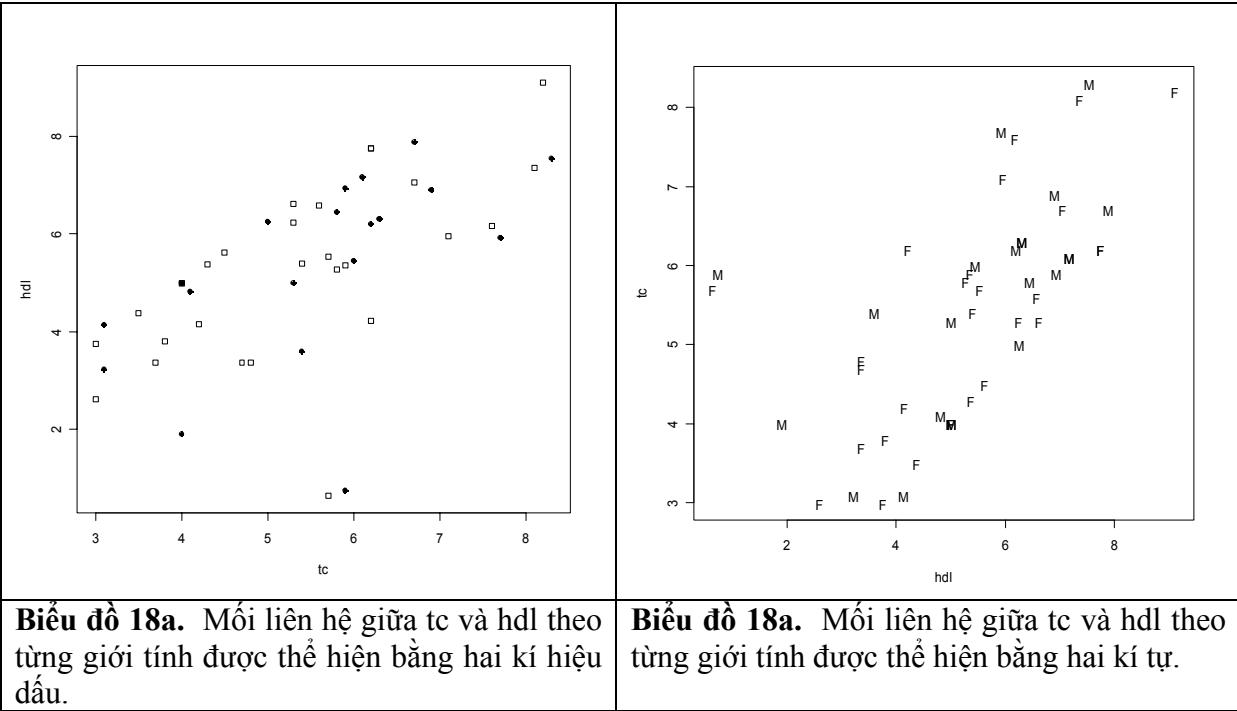
Biểu đồ 17. Mối liên hệ giữa tc và hdl. Trong biểu đồ này, chúng ta vẽ biến số hdl trên trục tung và tc trên trục hoành.

Chúng ta muốn phân biệt giới tính (nam và nữ) trong biểu đồ trên. Để vẽ biểu đồ đó, chúng ta phải dùng đến hàm `ifelse`. Trong lệnh sau đây, nếu `sex=="Nam"` thì vẽ kí tự số 16 (ô tròn), nếu không nam thì vẽ kí tự số 22 (tức ô vuông):

```
> plot(hdl, tc, pch=ifelse(sex=="Nam", 16, 22))
```

Kết quả là **Biểu đồ 18a**. Chúng ta cũng có thể thay kí tự thành “M” (nam) và “F” (nữ) (xem **Biểu đồ 18b**):

```
> plot(hdl, tc, pch=ifelse(sex=="Nam", "M", "F"))
```



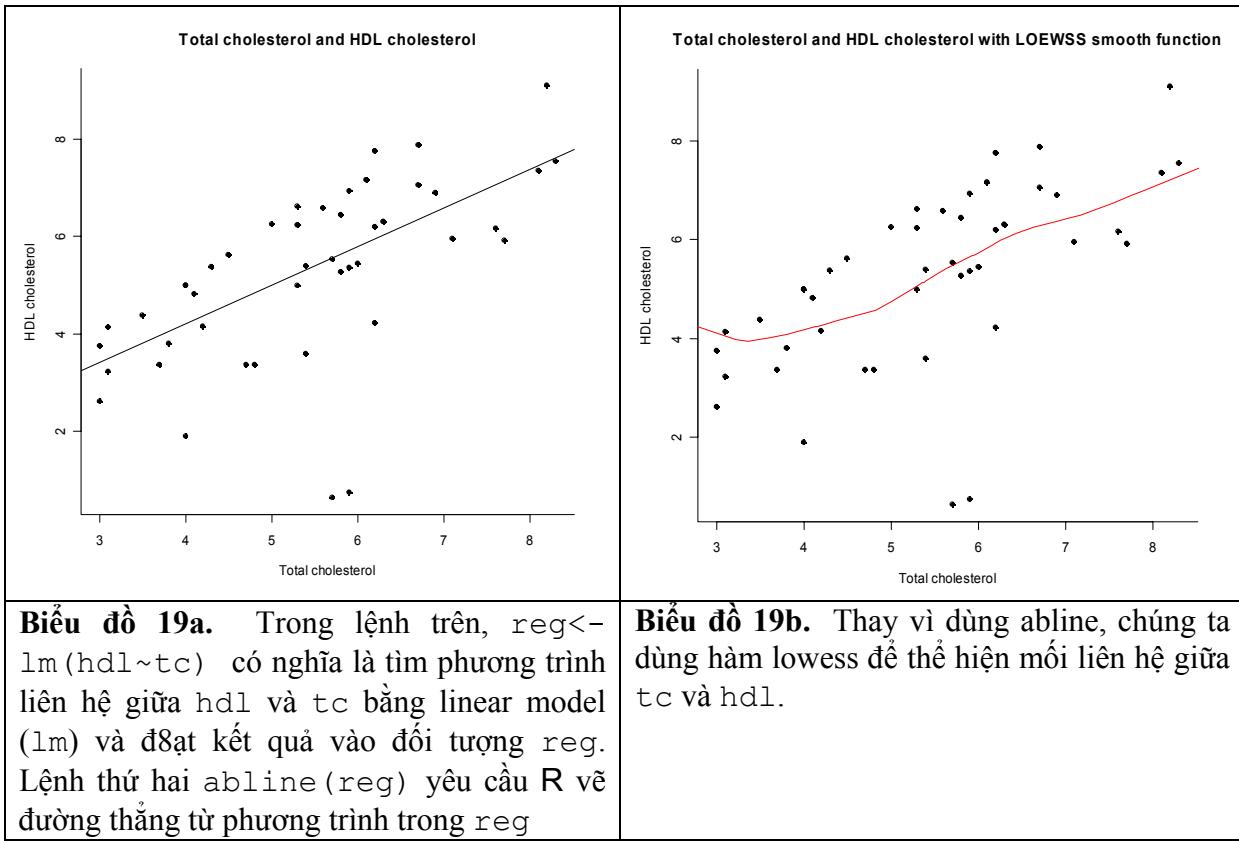
Chúng ta cũng có thể vẽ một đường biểu diễn hồi qui tuyến tính (regression line) qua các điểm trên bằng cách tiếp ra các lệnh sau đây:

```
> plot(hdl ~ tc, pch=16, main="Total cholesterol and HDL cholesterol",
       xlab="Total cholesterol", ylab="HDL cholesterol", bty="l")
> reg <- lm(hdl ~ tc)
> abline(reg)
```

Kết quả là **Biểu đồ 19a** dưới đây. Chúng ta cũng có thể dùng hàm trơn (smooth function) để biểu diễn mối liên hệ giữa hai biến số. Đồ thị sau đây sử dụng lowess (một hàm thông thường nhất) trong việc “làm trơn” số liệu tc và hdl (**Biểu đồ 19b**).

```
> plot(hdl ~ tc, pch=16,
       main="Total cholesterol and HDL cholesterol with LOESS smooth
function",
       xlab="Total cholesterol", ylab="HDL cholesterol", bty="l")

> lines(lowess(hdl, tc, f=2/3, iter=3), col="red")
```



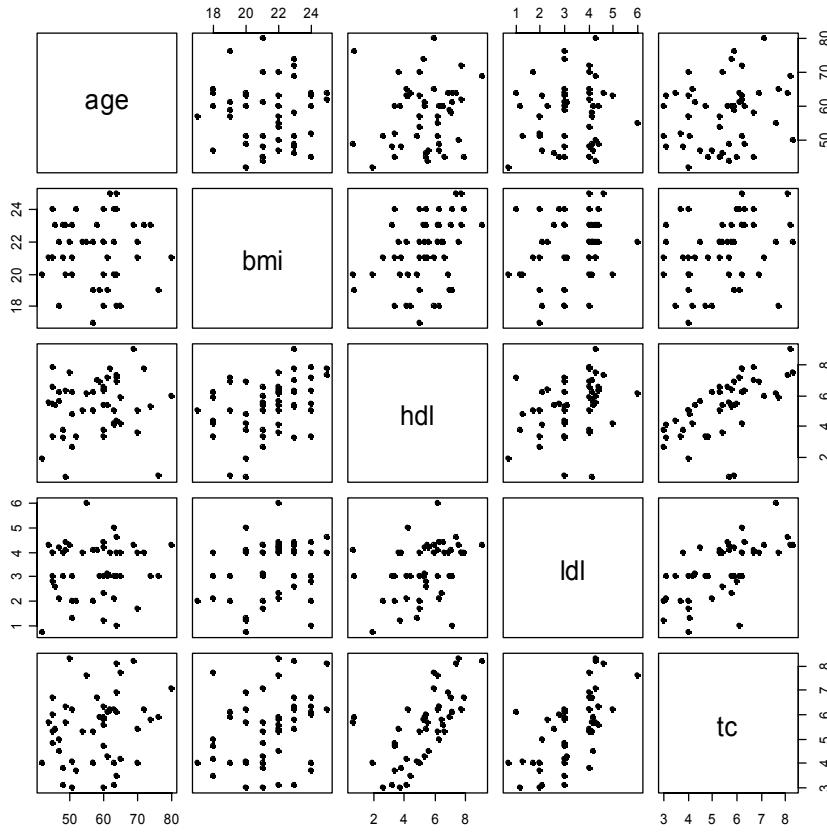
Bạn đọc có thể thí nghiệm với nhiều thông số $f=1/2$, $f=2/5$, hay thậm chí $f=1/10$ sẽ thấy đồ thị biến đổi một cách “thú vị”.

8.8 Phân tích Biểu đồ cho nhiều biến: pairs

Chúng ta có thể tìm hiểu mối liên hệ giữa các biến số như `age`, `bmi`, `hdl`, `ldl` và `tc` bằng cách dùng lệnh `pairs`. Nhưng trước hết, chúng ta phải đưa các biến số này vào một `data.frame` chỉ gồm những biến số có thể vẽ được, và sau đó sử dụng hàm `pairs` trong `R`.

```
> lipid <- data.frame(age,bmi,hdl,ldl,tc)
> pairs(lipid, pch=16)
```

Kết quả sẽ là:



Biểu đồ trên đây có thể cải tiến bằng hàm `matrix.cor` (do một tác giả trên mạng soạn) sau đây để cho ra nhiều thông tin thú vị.

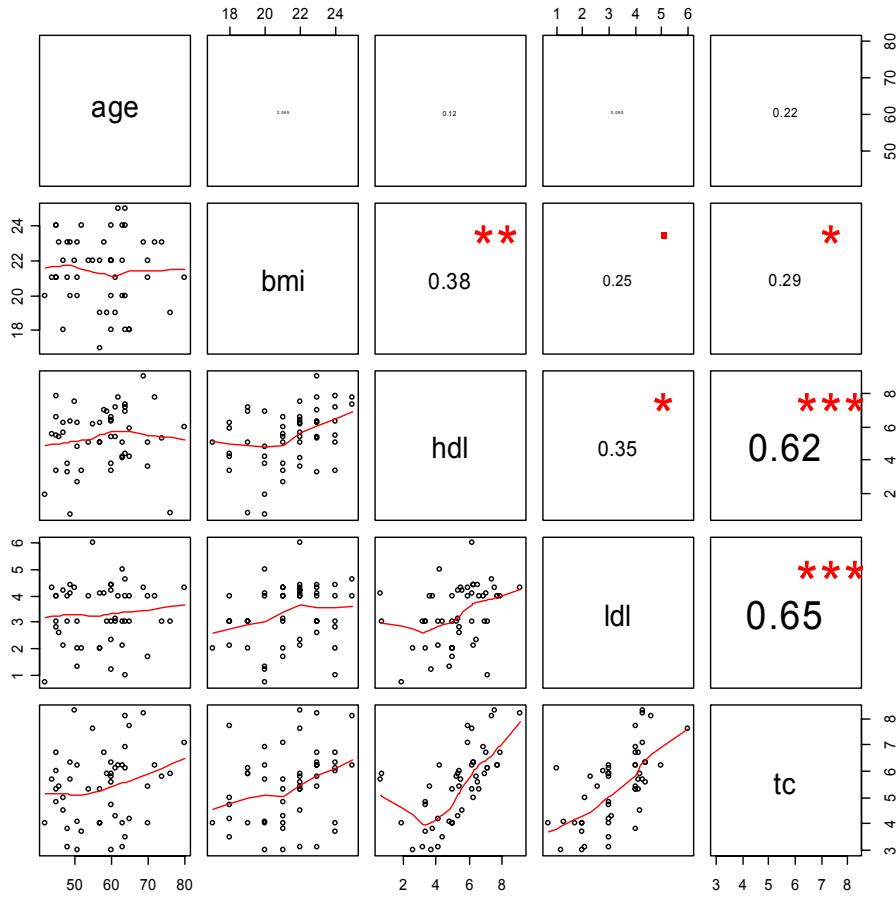
```
matrix.cor <- function(x, y, digits=2, prefix="", cex.cor)
{
  usr <- par("usr"); on.exit(par(usr))
  par(usr = c(0, 1, 0, 1))
  r <- abs(cor(x, y))
  txt <- format(c(r, 0.123456789), digits=digits)[1]
  txt <- paste(prefix, txt, sep="")
  if(missing(cex.cor)) cex <- 0.8/strwidth(txt)

  test <- cor.test(x,y)
  # borrowed from printCoefmat
  Signif <- symnum(test$p.value, corr = FALSE, na = FALSE,
    cutpoints = c(0, 0.001, 0.01, 0.05, 0.1, 1),
    symbols = c("***", "**", "*", ".", ""))

  text(0.5, 0.5, txt, cex = cex * r)
  text(.8, .8, Signif, cex=cex, col=2)
}
```

Chúng ta quay lại với dữ liệu lipid bằng cách gọi hàm `matrix.cor` như sau:

```
pairs(lipid, lower.panel=panel.smooth, upper.panel=matrix.cor)
```



Đồ thị này cung cấp cho chúng ta tất cả hệ số tương quan giữa tất cả các biến số. Chẳng hạn như, hệ số tương quan giữa `age` và `bmi` quá thấp và không có ý nghĩa thống kê; giữa `age` và `hdl` hay giữa `age` và `ldl` cũng không có ý nghĩa thống kê; nhưng giữa `age` và `tc` thì bằng 0.22. Hệ số tương quan cao nhất là giữa `ldl` và `tc` (0.65) và `hdl` và `tc` (0.62). Giữa `hdl` và `ldl`, hệ số tương quan chỉ 0.35, nhưng có ý nghĩa thống kê (có sao!)

Chú ý biểu đồ trên chẳng những cung cấp hai thông tin chính (hệ số tương quan hay correlation coefficient, và vẽ biểu đồ tán xạ cho từng cặp biến số), mà còn cho biết hệ số tương quan nào có ý nghĩa thống kê (những kí hiệu sao). Hệ số tương quan càng cao, kích thước của font chữ càng lớn. Một biểu đồ rất ấn tượng!

8.9 Một số biểu đồ “đa năng”

8.9.1 Biểu đồ tán xạ và hình hộp

Như trên đã trình bày, biểu đồ tán xạ giúp cho chúng ta hình dung ra mối liên hệ giữa hai biến số liên tục như độ tuổi age và hdl chẳng hạn. Và để làm việc này, chúng ta dùng hàm plot. Để tìm hiểu phân phối cho từng biến age hay hdl chúng ta có thể dùng hàm boxplot. Nhưng nếu chúng ta muốn xem phân phối của hai biến và đồng thời mối liên hệ giữa hai biến, thì chúng ta cần phải viết một vài lệnh để thực hiện việc này. Các lệnh sau đây vẽ biểu đồ tán xạ về mối liên quan giữa age và hdl, đồng thời vẽ biểu đồ hình hộp cho từng biến.

```
op <- par()
layout( matrix( c(2,1,0,3), 2, 2, byrow=T ),
        c(1,6), c(4,1),
        )

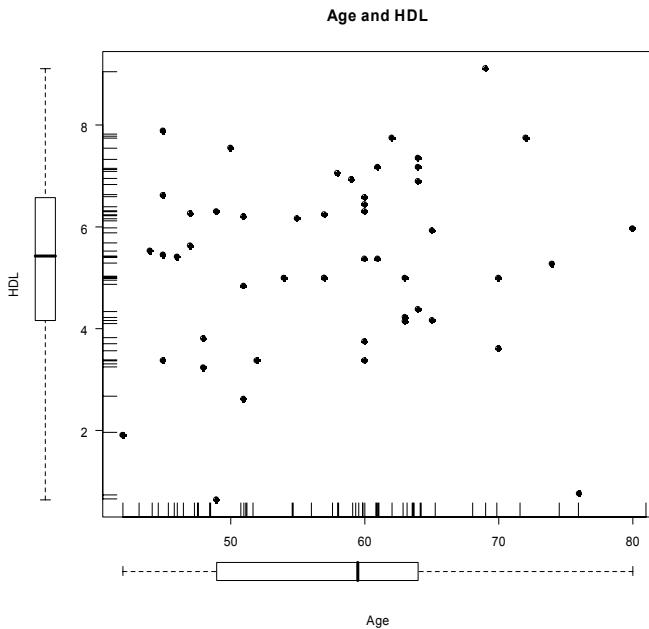
par(mar=c(1,1,5,2))
plot(hdl ~ age,
     xlab='', ylab='',
     las = 1,
     pch=16)
rug(side=1, jitter(age, 5) )
rug(side=2, jitter(hdl, 20) )
title(main = "Age and HDL")

par(mar=c(1,2,5,1))
boxplot(hdl, axes=F)
title(ylab='HDL', line=0)

par(mar=c(5,1,1,2))
boxplot(age, horizontal=T, axes=F)
title(xlab='Age', line=1)

par(op)
```

Và kết quả là:

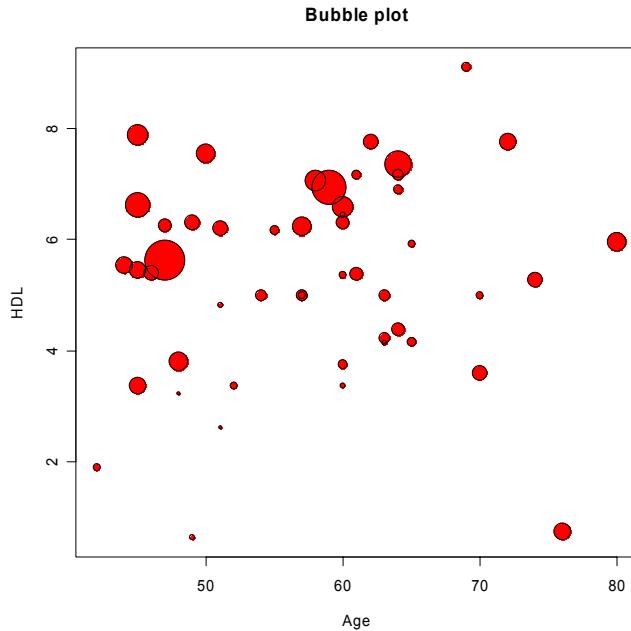


8.9.2 Biểu đồ tán xạ với kích thước biến ba

Biểu đồ trên thể hiện mối liên hệ giữa `age` và `hdl`, với mỗi điểm chấm có kích thước nhau nhau. Nhưng chúng ta biết rằng `hdl` cũng có liên hệ với triglyceride (`tg`). Để thể hiện một phần nào mối liên hệ 3 chiều này, một cách làm là vẽ kích thước của điểm tùy theo giá trị của `tg`. Chúng ta sẽ sử dụng thông số `cex` đã bàn trong phần đầu để vẽ mối liên hệ ba chiều này như sau:

```
> plot(age, hdl, cex=tg,
      pch=16,
      col="red",
      xlab="Age", ylab="HDL",
      main="Bubble plot")

> points(age, hdl, cex=tg)
```



8.9.3 Biểu đồ thanh và xác suất tích lũy

Để vẽ biểu đồ tần số của một biến liên tục chúng ta chủ yếu sử dụng hàm hist. Hàm này cho ra kết quả tần số cho từng nhóm (như nhóm độ tuổi chẵng hạn). Nhưng đôi khi chúng ta cần biết cả xác suất tích lũy cho từng nhóm, và muốn vẽ cả hai kết quả trong một biểu đồ. Để làm việc này chúng ta cần phải viết một hàm bằng ngôn ngữ R. Hàm sau đây được gọi là pareto (tất nhiên bạn đọc có thể cho một tên khác) được soạn ra để thực hiện mục tiêu trên. Mã cho hàm pareto như sau:

```
pareto <- function (x, main = "", ylab = "Value")
{
  op <- par(mar = c(5, 4, 4, 5) + 0.1,
            las = 2)
  if( ! inherits(x, "table") ) {
    x <- table(x)
  }
  x <- rev(sort(x))
  plot( x, type = 'h', axes = F, lwd = 16,
        xlab = "", ylab = ylab, main = main )
  axis(2)
  points( x, type = 'h', lwd = 12,
          col = heat.colors(length(x)) )
  y <- cumsum(x)/sum(x)
  par(new = T)
  plot(y, type = "b", lwd = 3, pch = 7,
        axes = FALSE,
        xlab='', ylab='', main='')
  points(y, type = 'h')
  axis(4)
  par(las=0)
  mtext("Cumulated frequency", side=4, line=3)
```

```

print(names(x))
axis(1, at=1:length(x), labels=names(x))
par(op)
}

```

Bây giờ chúng ta sẽ áp dụng hàm pareto vào việc vẽ tần số cho biến tg (triglyceride) như sau. Trước hết, chúng ta chia tg thành 10 nhóm bằng cách dùng hàm `cut` và cho kết quả vào đối tượng `tg.group`.

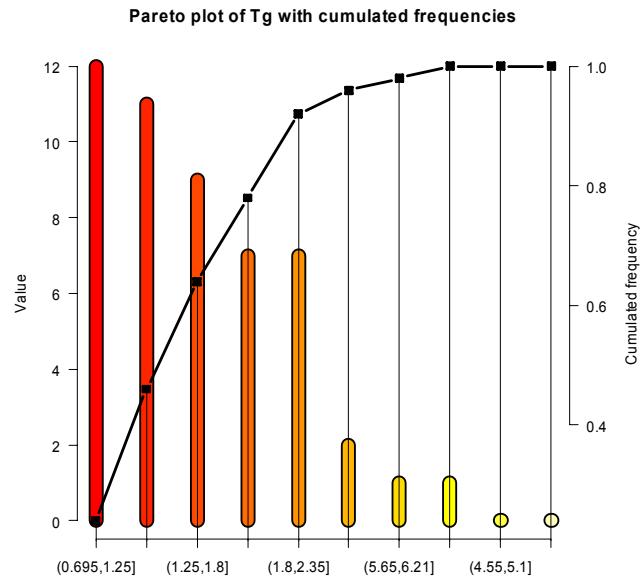
```
> tg.group <- cut(tg, 10)
```

Kế đến, chúng ta ứng dụng hàm pareto:

```

> pareto(tg.group)
[1] "(0.695,1.25]" "(2.35,2.9]" "(1.25,1.8]" "(2.9,3.45]" "(1.8,2.35]"
[6] "(3.45,4]" "(5.65,6.21]" "(5.1,5.65]" "(4.55,5.1]" "(4,4.55]"
> title(main="Pareto plot of Tg with cumulated frequencies")

```



Trong biểu đồ này, chúng ta có hai trục tung. Trục tung phía trái là tần số (số bệnh nhân) cho từng nhóm tg, và trục tung bên phải là tần số tích lũy tích bằng xác suất (do đó, số cao nhất là 1).

8.9.4 Biểu đồ hình đồng hồ (clock plot)

Biểu đồ hình đồng hồ, như tên gọi là biểu đồ dùng để vẽ một biến số liên tục bằng kim đồng hồ. Tức là thay vì thể hiện bằng cột hay bằng dòng, biểu đồ này thể hiện bằng đồng hồ. Hàm sau đây (`clock`) được soạn để thực hiện biểu đồ hình đồng hồ:

```
clock.plot <- function (x, col = rainbow(n), ...) {
```

```

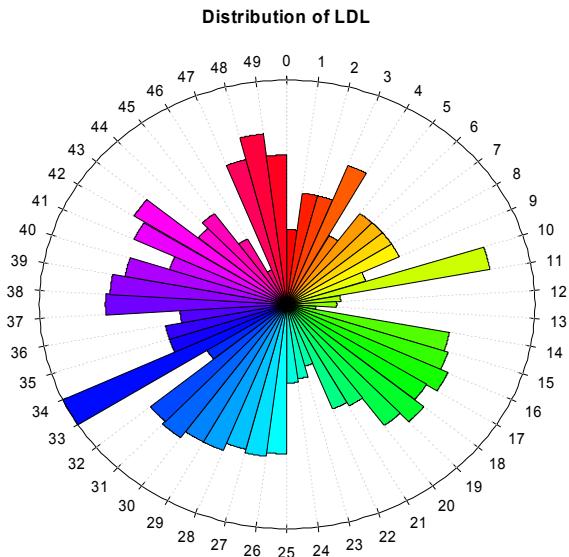
if( min(x)<0 ) x <- x - min(x)
if( max(x)>1 ) x <- x/max(x)
n <- length(x)
if(is.null(names(x))) names(x) <- 0:(n-1)
m <- 1.05
plot(0,
      type = 'n', # do not plot anything
      xlim = c(-m,m), ylim = c(-m,m),
      axes = F, xlab = '', ylab = '', ...)
a <- pi/2 - 2*pi/200*0:200
polygon( cos(a), sin(a) )
v <- .02
a <- pi/2 - 2*pi/n*0:n
segments( (1+v)*cos(a), (1+v)*sin(a),
          (1-v)*cos(a), (1-v)*sin(a) )
segments( cos(a), sin(a),
          0, 0,
          col = 'light grey', lty = 3)
ca <- -2*pi/n*(0:50)/50
for (i in 1:n) {
  a <- pi/2 - 2*pi/n*(i-1)
  b <- pi/2 - 2*pi/n*i
  polygon( c(0, x[i]*cos(a+ca), 0),
            c(0, x[i]*sin(a+ca), 0),
            col=col[i] )
  v <- .1
  text((1+v)*cos(a), (1+v)*sin(a), names(x)[i])
}
}

```

Chúng ta sẽ ứng dụng hàm clock để vẽ biểu đồ cho biến ldl như sau:

```
> clock.plot(ldl,
             main = "Distribution of LDL")
```

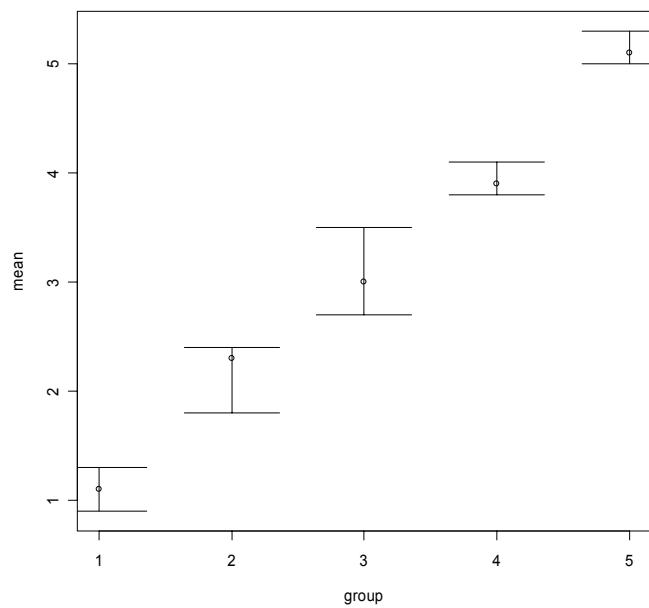
Và kết quả là:



8.9.5 Biểu đồ với sai số chuẩn (standard error)

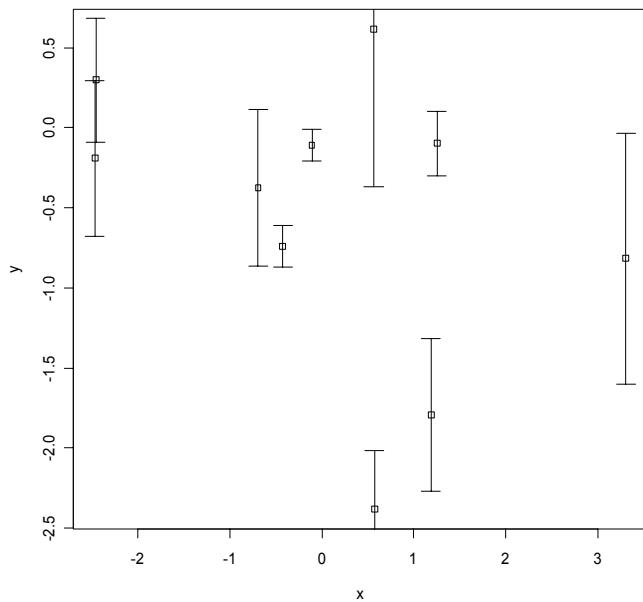
Trong biểu đồ sau đây, chúng ta có 5 nhóm (biến số x được mô phỏng chứ không phải số liệu thật), và mỗi nhóm có giá trị trung bình mean, và độ tin cậy 95% (lcl và ucl). Thông thường $lcl = \text{mean} - 1.96 * SE$ và $ucl = \text{mean} + 1.96 * SE$ (SE là sai số chuẩn). Chúng ta muốn vẽ biểu đồ cho 5 nhóm với sai số chuẩn đó. Các lệnh và hàm sau đây sẽ cần thiết:

```
> group <- c(1,2,3,4,5)
> mean <- c(1.1, 2.3, 3.0, 3.9, 5.1)
> lcl <- c(0.9, 1.8, 2.7, 3.8, 5.0)
> ucl <- c(1.3, 2.4, 3.5, 4.1, 5.3)
> plot(group, mean, ylim=range(c(lcl, ucl)))
> arrows(group, ucl, group, lcl, length=0.5, angle=90, code=3)
```



Sau đây là một mô phỏng khác. Chúng ta tạo ra 10 giá trị x và y theo luật phân phối chuẩn, và 10 giá trị sai số theo luật phân phối đều (`se.x` và `se.y` uniform distribution).

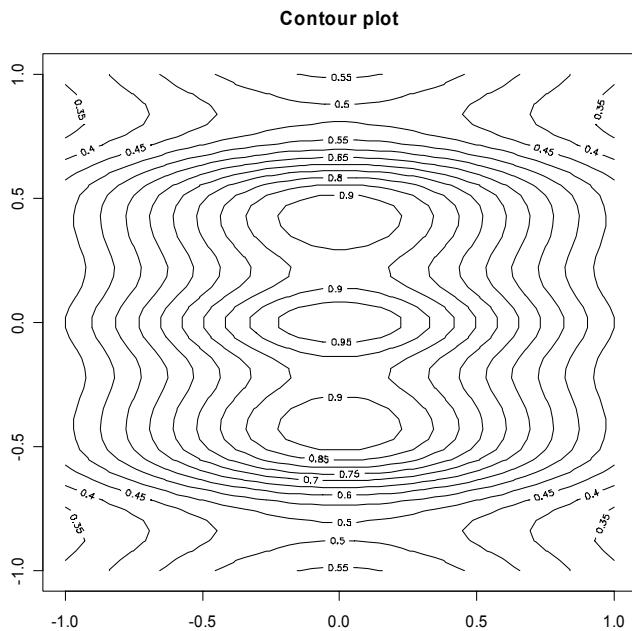
```
> x <- rnorm(10)
> y <- rnorm(10)
> se.x <- runif(10)
> se.y <- runif(10)
> plot(x, ypch=22)
> arrows(x, y-se.y, x, y+se.y, code=3, angle=90, length=0.1)
```



8.9.6 Biểu đồ vòng (contour plot)

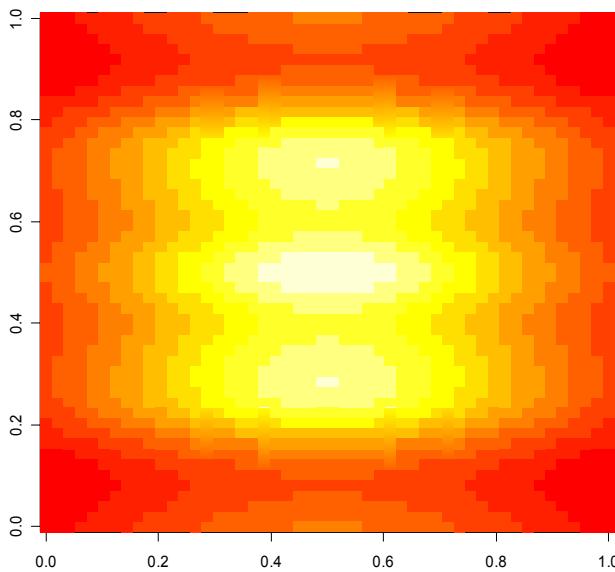
R có thể vẽ các đồ thị vòng với nhiều hình dạng khác nhau, tùy theo ý thích và dữ liệu. Trong các lệnh sau đây, chúng ta sử dụng kỹ thuật mô phỏng để vẽ đồ thị vòng cho ba biến số x, y và z.

```
> N <- 50
> x <- seq(-1, 1, length=N)
> y <- seq(-1, 1, length=N)
> xx <- matrix(x, nr=N, nc=N)
> yy <- matrix(y, nr=N, nc=N, byrow=TRUE)
> z <- 1 / (1 + xx^2 + (yy + .2 * sin(10*yy))^2)
> contour(x, y, z, main = "Contour plot")
```



Đồ thị này có thể chuyển thành một hình (image) bằng hàm `image`.

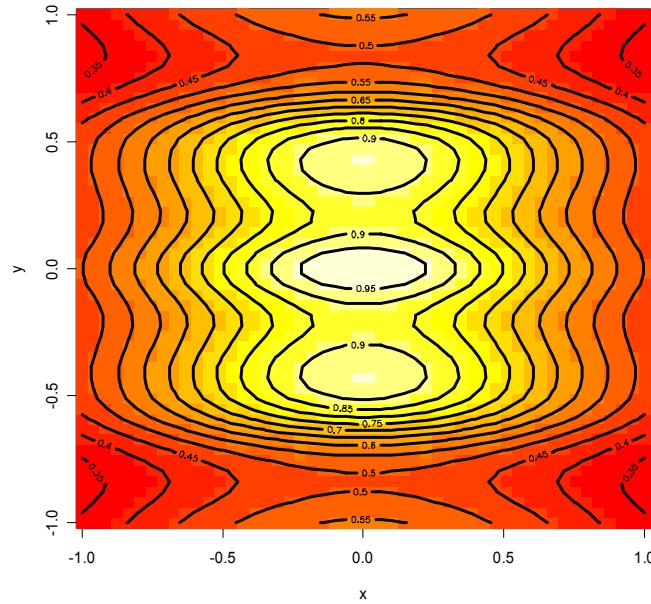
```
> image(z)
```



Một vài thay đổi nhỏ nhưng quan trọng:

```
> image(x, y, z,
       xlab="x",
       ylab="y")
```

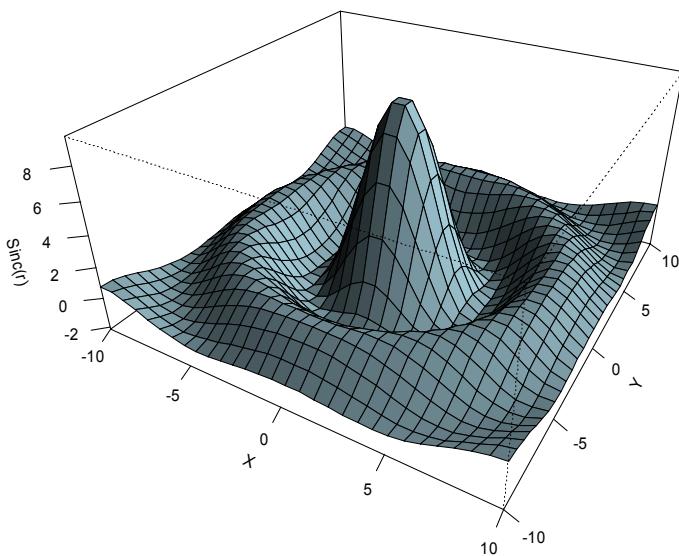
```
> contour(x, y, z, lwd=3, add=TRUE)
```



Sau đây là một vài thay đổi để vẽ biểu đồ theo hàm số sin và 3 chiều. Đồ thị này tuy xem “hấp dẫn”, nhưng trong thực tế có lẽ ít sử dụng. Tuy nhiên, tôi trình bày ở đây để cho thấy một ví dụ về tính đa dụng của R.

```
> x <- seq(-10, 10, length= 30)
> y <- x
> f <- function(x,y) { r <- sqrt(x^2+y^2); 10 * sin(r)/r }
> z <- outer(x, y, f)
> z[is.na(z)] <- 1
> op <- par(bg = "white", mar=c(0,2,3,0)+.1)
> persp(x, y, z,
+         theta = 30, phi = 30,
+         expand = 0.5,
+         col = "lightblue",
+         ltheta = 120,
+         shade = 0.75,
+         ticktype = "detailed",
+         xlab = "X", ylab = "Y", zlab = "Sinc(r)",
+         main = "The sinc function"
+     )
> par(op)
```

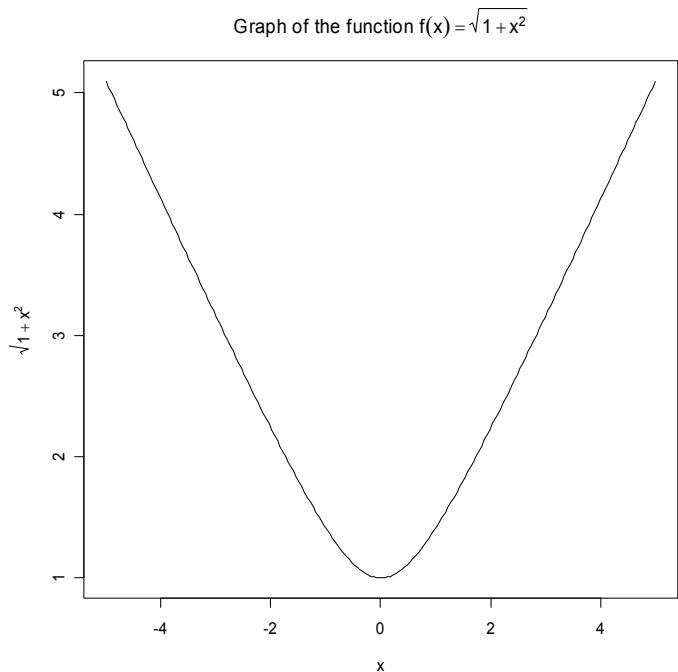
The sinc function



8.9.10 Biểu đồ với kí hiệu toán

Đôi khi chúng ta cần vẽ biểu đồ với tựa đề có kí hiệu toán học. Trong đồ thị sau đây, chúng ta tạo ra một biến số x với 200 giá trị từ -5 đến 5, và $y = \sqrt{1+x^2}$. Để viết công thức trên, chúng ta cần sử dụng hàm expression như sau:

```
> x <- seq(-5,5,length=200)
> y <- sqrt(1+x^2)
> plot(y~x, type='l', ylab=expression(sqrt(1+x^2)))
> title(main=expression("Graph of the function
f"(x)==sqrt(1+x^2)))
```



Ngay cả tiếng Nhật cũng có thể thể hiện bằng R:

```
> plot(1:9, type="n", axes=FALSE, frame=TRUE, ylab="",
      main= "example(Japanese)", xlab= "using Hershey fonts")
> par(cex=3)
> Vf <- c("serif", "plain")
> text(4, 2, "\\#J2438\\#J2421\\#J2451\\#J2473", vfont = Vf)
> text(4, 4, "\\#J2538\\#J2521\\#J2551\\#J2573", vfont = Vf)
> text(4, 6, "\\#J467c\\#J4b5c", vfont = Vf)
> text(4, 8, "Japan", vfont = Vf)
> par(cex=1)
> text(8, 2, "Hiragana")
> text(8, 4, "Katakana")
> text(8, 6, "Kanji")
> text(8, 8, "English")
```

example(Japanese)



using Hershey fonts

Chuong này chỉ giới thiệu một số biểu đồ thông thường trong nghiên cứu khoa học. Ngoài các biểu đồ thông dụng này, R còn có khả năng vẽ những đồ thị phức tạp và tinh vi hơn nữa. Hiện nay, R có một package tên là `lattice` có thể vẽ những biểu đồ chất lượng cao hơn. `lattice`, cũng như bất cứ package nào của R, đều miễn phí, có thể tải về máy tính và cài đặt để sử dụng khi cần thiết.

9 Phân tích thống kê mô tả

Trong chương này, chúng ta sẽ sử dụng R cho mục đích phân tích thống kê mô tả. Nói đến thống kê mô tả là nói đến việc mô tả dữ liệu bằng các phép tính và chỉ số thống kê thông thường mà chúng ta đã làm quen qua từ thuở trung học như số trung bình (mean), số trung vị (median), phương sai (variance) độ lệch chuẩn (standard deviation) ... cho các biến số liên tục, và tỉ số (proportion) cho các biến số không liên tục. Nhưng trước khi hướng dẫn phân tích thống kê mô tả, tôi muốn bạn đọc phải phân biệt cho được hai khái niệm *tổng thể* (population) và *mẫu* (sample).

9.0 Khái niệm tổng thể (population) và mẫu (sample)

Sách giáo khoa thống kê thường giải thích hai khái niệm này một cách mù mờ và có khi vô nghĩa. Chẳng hạn như cuốn “Modern Mathematical Statistics” (E. J. Dudewicz và S. N. Mishra, Nhà xuất bản Wiley, 1988) giải thích tổng thể rằng “population is a set of n distinct elements (points) $a_1, a_2, a_3, \dots, a_n$.” (trang 24, tạm dịch: “tổng thể là tập hợp gồm n phần tử hay điểm $a_1, a_2, a_3, \dots, a_n$ ”), còn L. Fisher và G. van Belle trong “Biostatistics – A Methodology for the Health Science” (Nhà xuất bản Wiley, 1993), giải thích rằng “The sample space or population is the set of all possible values of a variable” (trang 38, tạm dịch “Không gian mẫu hay tổng thể là tập hợp tất cả các giá trị khả dĩ của một biến”). Đối với một nhà nghiên cứu thực nghiệm phải nói những định nghĩa loại này rất trừu tượng và khó hiểu, và dường như chẳng có liên quan gì với thực tế! Trong phần này tôi sẽ giải thích hai khái niệm này bằng mô phỏng và hi vọng là bạn đọc sẽ hiểu rõ hơn.

Có thể nói mục tiêu của nghiên cứu khoa học thực nghiệm là nhằm tìm hiểu và khám phá những cái *chưa được biết* (unknown), trong đó bao gồm những qui luật hoạt động của tự nhiên. Để khám phá, chúng ta sử dụng đến các phương pháp *phân loại*, *so sánh*, và *phỏng đoán*. Tất cả các phương pháp khoa học, kể cả thống kê học, được phát triển nhằm vào ba mục tiêu trên. Để phân loại, chúng ta phải đo lường một yếu tố hay tiêu chí có liên quan đến vấn đề cần nghiên cứu. Để so sánh và phỏng đoán, chúng ta cần đến các phương pháp kiểm định giả thiết và mô hình thống kê học.

Cũng như bất cứ mô hình nào, mô hình thống kê phải có thông số. Và muốn có thông số, chúng ta trước hết phải tiến hành đo lường, và sau đó là ước tính thông số từ đo lường. Chẳng hạn như để biết sinh viên nữ có chỉ số thông minh (IQ) bằng sinh viên nam hay không, chúng ta có thể làm nghiên cứu theo hai phương án:

- (a) Một là lập danh sách tất cả sinh viên nam và nữ trên toàn quốc, rồi đo lường chỉ số IQ ở từng người, và sau đó so sánh giữa hai nhóm;
- (b) Hai là chọn ngẫu nhiên một mẫu gồm n nam và m nữ sinh viên, rồi đo lường chỉ số IQ ở từng người, và sau đó so sánh giữa hai nhóm.

Phương án (a) rất tốn kém và có thể nói là không thực tế, vì chúng ta phải tập hợp tất cả sinh viên của cả nước, một việc làm rất khó thực hiện được. Nhưng giả dụ như chúng ta có thể làm được, thì phương án này không cần đến thống kê học. Giá trị IQ trung bình của nữ và nam sinh viên tính từ phương án (a) là giá trị cuối cùng, và nó trả lời câu hỏi của chúng ta một cách trực tiếp, chúng ta không cần phải suy luận, không cần đến kiểm định thống kê gì cả!

Phương án (b) đòi hỏi chúng ta phải chọn n nam và m nữ sinh viên sao cho *đại diện* (representative) cho toàn quần thể sinh viên của cả nước. Tính “đại diện” ở đây có nghĩa là các số n nam và m nữ sinh viên này phải có cùng đặc tính như độ tuổi, trình độ học vấn, thành phần kinh tế, xã hội, nơi sinh sống. v.v... so với tổng thể sinh viên của cả nước. Bởi vì chúng ta không biết các đặc tính này trong toàn bộ tổng thể sinh viên, chúng ta không thể so sánh trực tiếp được, cho nên một phương pháp rất hữu hiệu là lấy mẫu một cách ngẫu nhiên. Có nhiều phương pháp lấy mẫu ngẫu nhiên đã được phát triển và tôi sẽ không bàn qua chi tiết của các phương pháp này, ngoại trừ muốn nhấn mạnh rằng, nếu cách lấy mẫu không ngẫu nhiên thì các ước số từ mẫu sẽ không có ý nghĩa khoa học cao, bởi vì các phương pháp phân tích thống kê dựa vào giả định rằng mẫu phải được chọn một cách ngẫu nhiên.

Tôi sẽ lấy một ví dụ cụ thể về tổng thể và mẫu qua ứng dụng R như sau. Giả dụ chúng ta có một tổng thể gồm 20 người và biết rằng chiều cao của họ như sau (tính bằng cm): 162, 160, 157, 155, 167, 160, 161, 153, 149, 157, 159, 164, 150, 162, 168, 165, 156, 157, 154 và 157. Như vậy, chúng ta biết rằng chiều cao trung bình của tổng thể là 158.65 cm. Xin nhấn mạnh đó là tổng thể.

Vì thiếu thốn phương tiện chúng ta không thể nghiên cứu trên toàn tổng thể mà chỉ có thể lấy mẫu từ tổng thể để ước tính chiều cao. Hàm `sample()` cho phép chúng ta lấy mẫu. Và ước tính chiều cao trung bình từ mẫu tất nhiên sẽ khác với chiều cao trung bình của tổng thể.

- Chọn 5 người từ tổng thể:
> `sample5 <- sample(height, 5)`
> `sample5`
[1] 153 157 164 156 149

Ước tính chiều cao trung bình từ mẫu này:

```
> mean(sample5)  
[1] 155.8
```

- Chọn 5 người khác từ tổng thể và tính chiều cao trung bình:
> `sample5 <- sample(height, 5)`
> `sample5`
[1] 157 162 167 161 150
> `mean(sample5)`
[1] 159.4

Chú ý ước tính chiều cao của mẫu thứ hai là 159.4 cm (thay vì 155.8 cm), bởi vì chọn ngẫu nhiên, cho nên đối tượng được chọn lần hai không nhất thiết phải là đối tượng lần thứ nhất, cho nên ước tính trung bình khác nhau.

- Bây giờ chúng ta thử lấy mẫu 10 người từ tổng thể và tính chiều cao trung bình:

```
> sample10 <- sample(height, 10)  
> sample10  
[1] 153 160 150 165 159 160 164 156 162 157  
> mean(sample10)  
[1] 158.6
```

Chúng ta có thể lấy nhiều mẫu, mỗi mẫu gồm 10 người và ước tính số trung bình từ mẫu, bằng một lệnh đơn giản hơn như sau:

```
> mean(sample(height, 10))  
[1] 156.7  
> mean(sample(height, 10))  
[1] 157.1  
> mean(sample(height, 10))  
[1] 159.3  
> mean(sample(height, 10))  
[1] 159.3  
> mean(sample(height, 10))  
[1] 158.3  
> mean(sample(height, 10))
```

Chú ý độ dao động của số trung bình từ 156.7 đến 159.3 cm.

- Chúng ta thử lấy mẫu 15 người từ tổng thể và tính chiều cao trung bình:

```
> mean(sample(height, 15))  
[1] 158.6667  
> mean(sample(height, 15))  
[1] 159.4  
> mean(sample(height, 15))  
[1] 158.0667  
> mean(sample(height, 15))  
[1] 158.1333  
> mean(sample(height, 15))  
[1] 156.4667
```

Chú ý độ dao động của số trung bình bây giờ từ 158.0 đến 158.7 cm, tức thấp hơn mẫu với 10 đối tượng.

- Tăng cỡ mẫu lên 18 người (tức gần số đối tượng trong tổng thể)

```
> mean(sample(height, 18))  
[1] 158.2222  
> mean(sample(height, 18))  
[1] 158.7222  
> mean(sample(height, 18))  
[1] 158.0556  
> mean(sample(height, 18))  
[1] 158.4444  
> mean(sample(height, 18))
```

```
[1] 158.6667
> mean(sample(height, 18))
[1] 159.0556
> mean(sample(height, 18))
[1] 159
```

Bây giờ thì ước tính chiều cao khá ổn định, nhưng không khác gì so với cỡ mẫu với 15 người, do độ dao động từ 158.2 đến 159 cm.

Từ các ví dụ trên đây, chúng ta có thể rút ra một nhận xét quan trọng: Ước số từ các mẫu được chọn một cách ngẫu nhiên sẽ khác với thông số của tổng thể, nhưng khi số cỡ mẫu tăng lên thì độ khác biệt sẽ nhỏ lại dần. Do đó, một trong những vấn đề then chốt của thiết kế nghiên cứu là nhà nghiên cứu phải ước tính cỡ mẫu sao cho ước số mà chúng ta tính từ mẫu gần (hay chính xác) so với thông số của tổng thể. Tôi sẽ quay lại vấn đề này trong Chương 15.

Trong ví dụ trên số trung bình của tổng thể là 158.65 cm. Trong thống kê học, chúng ta gọi đó là *thông số* (parameter). Và các số trung bình ước tính từ các mẫu chọn từ tổng thể đó được gọi là *ước số mẫu* (sample estimate). Do đó, xin nhắc lại để nhấn mạnh: những chỉ số liên quan đến tổng thể là thông số, còn những số ước tính từ các mẫu là ước số. Như thấy trên, ước số có độ dao động chung quanh thông số, và vì trong thực tế chúng ta không biết thông số, cho nên chúng mục tiêu chính của phân tích thống kê là sử dụng ước số để suy luận về thông số.

Mục tiêu chính của phân tích thống kê mô tả là tìm những ước số của mẫu. Có hai loại đo lường: liên tục (continuous measurement) và không liên tục hay rời rạc (discrete measurement). Các biến liên tục như độ tuổi, chiều cao, trọng lượng cơ thể, v.v... là biến số liên tục, còn các biến mang tính phân loại như có hay không có bệnh, thích hay không thích, trắng hay đen, v.v... là những biến số không liên tục. Cách tính hai loại biến số này cũng khác nhau.

Ước số thông thường nhất dùng để mô tả một biến số liên tục là số trung bình (mean). Chẳng hạn như chiều cao của nhóm 1 gồm 5 đối tượng là 160, 160, 167, 156, và 161, do đó số trung bình là 160.8 cm. Nhưng chiều cao của nhóm 2 cũng gồm 5 đối tượng khác như 142, 150, 187, 180 và 145, thì số trung bình vẫn là 160.8. Do đó, số trung bình không thể phản ánh đầy đủ sự phân phối của một biến liên tục, vì ở đây tuy hai nhóm có cùng trung bình nhưng độ khác biệt của nhóm 2 cao hơn nhóm 1 rất nhiều. Và chúng ta cần một ước số khác gọi là phương sai (variance). Phương sai của nhóm 1 là 15.7 cm^2 và nhóm 2 là 443.7 cm^2 .

Với một biến số không liên tục như 0 và 1 (0 kí hiệu còn sống, và 1 kí hiệu tử vong) thì ước số trung bình không còn ý nghĩa “trung bình” nữa, cho nên chúng ta có ước số tỉ lệ (proportion). Chẳng hạn như trong số 10 người có 2 người tử vong, thì tỉ lệ tử vong là 0.2 (hay 20%). Trong số 200 người có 40 người qua đời thì tỉ lệ tử vong vẫn 0.2. Do đó, cũng như trường hợp trung bình, tỉ lệ không thể mô tả một biến không liên tục đầy đủ được. Chúng ta cần đến phương sai để, cùng với tỉ lệ, mô tả một biến không liên tục. Trong trường hợp 2/10 phương sai là 0.016, còn trong trường hợp 40/200, phương sai là

0.0008. Trong chương này, chúng ta sẽ làm quen với một số lệnh trong R để tiến hành những tính toán đơn giản trên.

9.1 Thống kê mô tả (descriptive statistics, summary)

Để minh họa cho việc áp dụng R vào thống kê mô tả, tôi sẽ sử dụng một dữ liệu nghiên cứu có tên là igfdata. Trong nghiên cứu này, ngoài các chỉ số liên quan đến giới tính, độ tuổi, trọng lượng và chiều cao, chúng tôi đo lường các hormone liên quan đến tình trạng tăng trưởng như igfi, igfbp3, als, và các markers liên quan đến sự chuyển hóa của xương pinp, ictp và pinp. Có 100 đối tượng nghiên cứu. Dữ liệu này được chứa trong directory c:\works\stats. Trước hết, chúng ta cần phải nhập dữ liệu vào R với những lệnh sau đây (các câu chữ theo sau dấu # là những chú thích để bạn đọc theo dõi):

```
> options(width=100)
# chuyển directory
> setwd("c:/works/stats")

# đọc dữ liệu vào R
> igfdata <- read.table("igf.txt", header=TRUE, na.strings=".")
> attach(igfdata)

# xem xét các cột số trong dữ liệu
> names(igfdata)
[1] "id"          "sex"         "age"        "weight"      "height"      "ethnicity"
[7] "igfi"        "igfbp3"      "als"        "pinp"       "ictp"       "p3np"

> igfdata
   id sex age weight height ethnicity igfi igfbp3 als pinp ictp p3np
1  1 Female 15    42    162 Asian 189.000 4.00000 323.667 353.970 11.2867 8.3367
2  2 Male 16    44    160 Caucasian 160.000 3.75000 333.750 375.885 10.4300 6.7450
3  3 Female 15    43    157 Asian 146.833 3.43333 248.333 199.507 8.3633 12.5000
4  4 Female 15    42    155 Asian 185.500 3.40000 251.000 483.607 13.3300 14.2767
5  5 Female 16    47    167 Asian 192.333 4.23333 322.000 105.430 7.9233 4.5033
6  6 Female 25    45    160 Asian 110.000 3.50000 284.667 76.487 4.9833 4.9367
7  7 Female 19    45    161 Asian 157.000 3.20000 274.000 75.880 6.3500 5.3200
8  8 Female 18    43    153 Asian 146.000 3.40000 303.000 86.360 7.3700 4.6700
9  9 Female 15    41    149 Asian 197.667 3.56667 308.500 254.803 11.8700 6.8200
10 10 Female 24    45    157 African 148.000 3.40000 273.000 44.720 3.7400 6.1600
...
...
97 97 Female 17    54    168 Caucasian 204.667 4.96667 441.333 64.130 5.1600 4.4367
98 98 Male 18    55    169 Asian 178.667 3.86667 273.000 185.913 7.5267 8.8333
99 99 Female 18    48    151 Asian 237.000 3.46667 324.333 105.127 5.9867 5.6600
100 100 Male 15    54    168 Asian 130.000 2.70000 259.333 325.840 10.2767 6.5933
```

Trên đây chỉ là một phần số liệu trong số 100 đối tượng.

Cho một biến số $x_1, x_2, x_3, \dots, x_n$ chúng ta có thể tính toán một số chỉ số thống kê mô tả như sau:

Lí thuyết	Hàm R
Số trung bình: $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$.	mean(x)
Phương sai: $s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$	var(x)
Độ lệch chuẩn: $s = \sqrt{s^2}$	sd(x)
Sai số chuẩn (standard error): $SE = \frac{s}{\sqrt{n}}$	Không có
Trị số thấp nhất	min(x)
Trị số cao nhất	max(x)
Toàn cự (range)	range(x)

Ví dụ 1: Để tìm giá trị trung bình của độ tuổi, chúng ta chỉ đơn giản lệnh:

```
> mean(age)
[1] 19.17
```

Hay phương sai và độ lệch chuẩn của tuổi:

```
> var(age)
[1] 15.33444
```

```
> sd(age)
[1] 3.915922
```

Tuy nhiên, R có lệnh `summary` có thể cho chúng ta tất cả thông tin thống kê về một biến số:

```
> summary(age)
   Min. 1st Qu. Median     Mean 3rd Qu.    Max.
 13.00    16.00   19.00    19.17   21.25   34.00
```

Nói chung, kết quả này đơn giản và các viết tắt cũng có thể dễ hiểu. Chú ý, trong kết quả trên, có hai chỉ số “1st Qu.” và “3rd Qu.” có nghĩa là first quartile (tương đương với vị trí 25%) và third quartile (tương đương với vị trí 75%) của một biến số. First quartile = 16 có nghĩa là 25% đối tượng nghiên cứu có độ tuổi bằng hoặc nhỏ hơn 16 tuổi. Tương tự, Third quartile = 34 có nghĩa là 75% đối tượng có độ tuổi bằng hoặc thấp hơn 34 tuổi. Tất nhiên số trung vị (median) 19 cũng có nghĩa là 50% đối tượng có độ tuổi 19 trở xuống (hay 19 tuổi trở lên).

R không có hàm tính sai số chuẩn, và trong hàm summary, R cũng không cung cấp độ lệch chuẩn. Để có các số này, chúng ta có thể tự viết một hàm đơn giản (hãy gọi là desc) như sau:

```
desc <- function(x)
{
  av <- mean(x)
  sd <- sd(x)
  se <- sd/sqrt(length(x))
  c(MEAN=av, SD=sd, SE=se)
}
```

Và có thể gọi hàm này để tính bất cứ biến nào chúng ta muốn, như tính biến als sau đây:

```
> desc(als)
      MEAN           SD           SE
  301.841120   58.987189   5.898719
```

Để có một “quang cảnh” chung về dữ liệu igfdata chúng ta chỉ đơn giản lệnh summary như sau:

```
> summary(igfdata)
    id          sex         age        weight       height     ethnicity
Min. : 1.00  Female:69  Min. :13.00  Min. :41.00  Min. :149.0  African : 8
1st Qu.: 25.75  Male :31  1st Qu.:16.00  1st Qu.:47.00  1st Qu.:157.0  Asian :60
Median : 50.50                    Median :19.00  Median :50.00  Median :162.0  Caucasian:30
Mean   : 50.50                    Mean   :19.17  Mean   :49.91  Mean   :163.1  Others  : 2
3rd Qu.: 75.25                    3rd Qu.:21.25  3rd Qu.:53.00  3rd Qu.:168.0
Max.   :100.00                    Max.   :34.00   Max.   :60.00   Max.   :196.0

      igfi        igfbp3        als        pinp       ictp
Min.   : 85.71  Min.   :2.000  Min.   :192.7  Min.   : 26.74  Min.   : 2.697
1st Qu.:137.17  1st Qu.:3.292  1st Qu.:256.8  1st Qu.: 68.10  1st Qu.: 4.878
Median :161.50  Median :3.550  Median :292.5  Median :103.26  Median : 6.338
Mean   :165.59  Mean   :3.617  Mean   :301.8  Mean   :167.17  Mean   : 7.420
3rd Qu.:186.46  3rd Qu.:3.875  3rd Qu.:331.2  3rd Qu.:196.45  3rd Qu.: 8.423
Max.   :427.00  Max.   :5.233  Max.   :471.7  Max.   :742.68  Max.   :21.237

      p3np
Min.   : 2.343
1st Qu.: 4.433
Median : 5.445
Mean   : 6.341
3rd Qu.: 7.150
Max.   :16.303
```

R tính toán tất cả các biến số nào có thể tính toán được! Thành ra, ngay cả cột id (tức mã số của đối tượng nghiên cứu) R cũng tính luôn! (và chúng ta biết kết quả của cột id chẳng có ý nghĩa thống kê gì). Đối với các biến số mang tính phân loại như sex và ethnicity (sắc tộc) thì R chỉ báo cáo tần số cho mỗi nhóm.

Kết quả trên cho tất cả đối tượng nghiên cứu. Nếu chúng ta muốn kết quả cho từng nhóm nam và nữ riêng biệt, hàm `by` trong R rất hữu dụng. Trong lệnh sau đây, chúng ta yêu cầu R tóm lược dữ liệu `igfdata` theo `sex`.

```
> by(igfdata, sex, summary)
```

sex: Female

	id	sex	age	weight	height
Min.	: 1.0	Female: 69	Min. :13.00	Min. :41.00	Min. :149.0
1st Qu.	:21.0	Male : 0	1st Qu.:17.00	1st Qu.:47.00	1st Qu.:156.0
Median	:47.0		Median :19.00	Median :50.00	Median :162.0
Mean	:48.2		Mean :19.59	Mean :49.35	Mean :161.9
3rd Qu.	:75.0		3rd Qu.:22.00	3rd Qu.:52.00	3rd Qu.:166.0
Max.	:99.0		Max. :34.00	Max. :60.00	Max. :196.0
ethnicity					
African	: 4	igfi	igfbp3	als	
Asian	:43	Min. : 85.71	Min. :2.767	Min. :204.3	
Caucasian	:22	1st Qu.:136.67	1st Qu.:3.333	1st Qu.:263.8	
Others	: 0	Median :163.33	Median :3.567	Median :302.7	
		Mean :167.97	Mean :3.695	Mean :311.5	
		3rd Qu.:186.17	3rd Qu.:3.933	3rd Qu.:361.7	
		Max. :427.00	Max. :5.233	Max. :471.7	
pinp					
Min.	: 26.74	ictp	p3np		
1st Qu.	: 62.75	Min. : 2.697	Min. : 2.343		
Median	: 78.50	1st Qu.: 4.717	1st Qu.: 4.337		
Mean	:108.74	Median : 5.537	Median : 5.143		
3rd Qu.	:115.26	Mean : 6.183	Mean : 5.643		
Max.	:502.05	3rd Qu.: 7.320	3rd Qu.: 6.143		
		Max. :13.633	Max. :14.420		

sex: Male

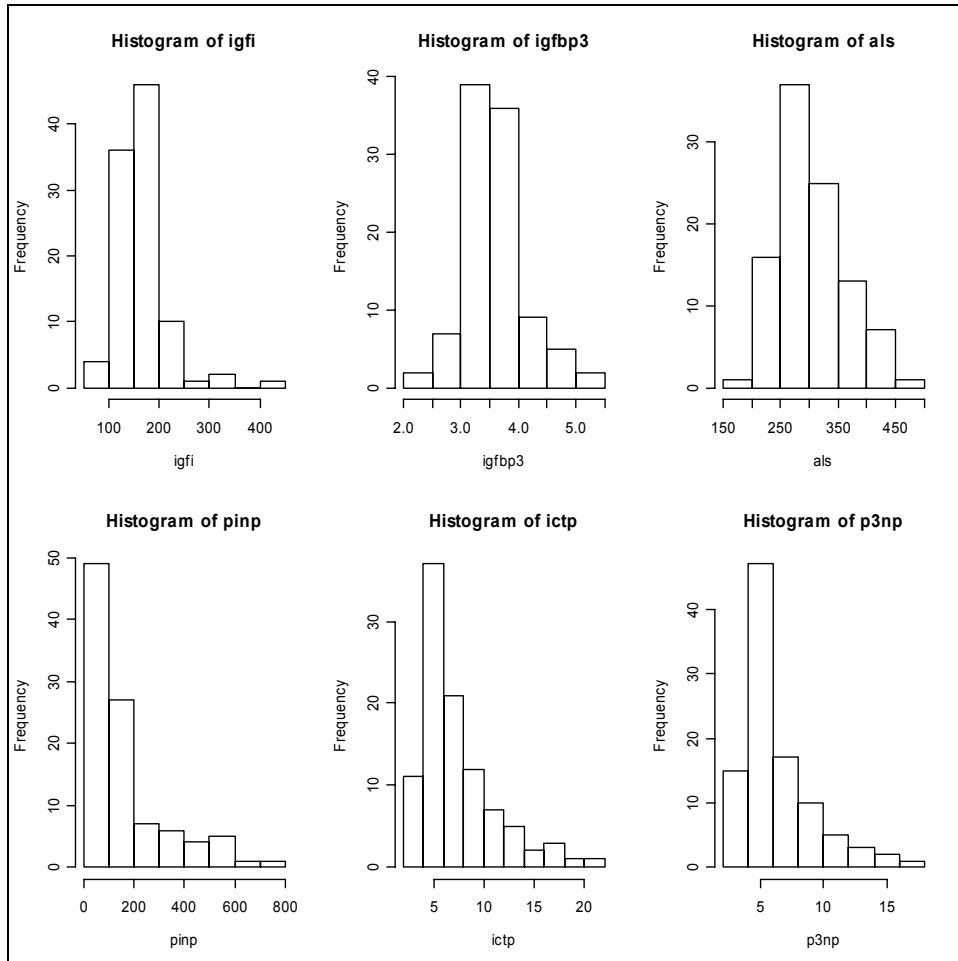
	id	sex	age	weight	height
Min.	: 2.00	Female: 0	Min. :14.00	Min. :44.00	Min. :155.0
1st Qu.	: 34.50	Male :31	1st Qu.:15.00	1st Qu.:48.50	1st Qu.:161.5
Median	: 56.00		Median :17.00	Median :51.00	Median :164.0
Mean	: 55.61		Mean :18.23	Mean :51.16	Mean :165.6
3rd Qu.	: 75.00		3rd Qu.:20.00	3rd Qu.:53.50	3rd Qu.:169.0
Max.	:100.00		Max. :27.00	Max. :59.00	Max. :191.0
ethnicity					
African	: 4	igfi	igfbp3	als	
Asian	:17	Min. : 94.67	Min. :2.000	Min. :192.7	
Caucasian	: 8	1st Qu.:138.67	1st Qu.:3.183	1st Qu.:249.8	
Others	: 2	Median :160.00	Median :3.500	Median :276.0	
		Mean :160.29	Mean :3.443	Mean :280.2	
		3rd Qu.:183.00	3rd Qu.:3.775	3rd Qu.:311.3	
		Max. :274.00	Max. :4.500	Max. :388.7	
pinp					
Min.	: 56.28	ictp	p3np		
1st Qu.	:135.07	Min. : 3.650	Min. : 3.390		
Median	:245.92	1st Qu.: 6.900	1st Qu.: 5.375		
Mean	:297.21	Median : 9.513	Median : 7.140		
3rd Qu.	:450.38	Mean :10.173	Mean : 7.895		
Max.	:742.68	3rd Qu.:13.517	3rd Qu.:10.010		
		Max. :21.237	Max. :16.303		

Để xem qua phân phối của các hormones và chỉ số sinh hóa cùng một lúc, chúng ta có thể vẽ đồ thị cho tất cả 6 biến số. Trước hết, chia màn ảnh thành 6 cửa sổ (với 2 dòng và 3 cột); sau đó lần lượt vẽ:

```

> op <- par(mfrow=c(2,3))
> hist(igfi)
> hist(igfbp3)
> hist(als)
> hist(pinp)
> hist(ictp)
> hist(p3np)

```



9.2 Kiểm định xem một biến có phải phân phối chuẩn

Trong phân tích thống kê, phần lớn các phép tính dựa vào giả định biến số phải là một biến số phân phối chuẩn (normal distribution). Do đó, một trong những việc quan trọng khi xem xét dữ kiện là phải kiểm định giả thiết phân phối chuẩn của một biến số. Trong đồ thị trên, chúng ta thấy các biến số như igfi, pinp, ictp và p3np có vẻ tập trung vào các giá trị thấp và không cân đối, tức dấu hiệu của một sự phân phối không chuẩn.

Để kiểm định nghiêm chỉnh, chúng ta cần phải sử dụng kiểm định thống kê có tên là “Shapiro test” và trong R gọi là hàm shapiro.test. Chẳng hạn như kiểm định giả thiết phân phối chuẩn của biến số pinp,

```
> shapiro.test(pinp)

Shapiro-Wilk normality test

data: pinp
W = 0.748, p-value = 8.314e-12
```

Vì trị số p (p-value) thấp hơn 0.05, chúng ta có thể kết luận rằng biến số pinp không đáp ứng luật phân phối chuẩn.

Nhưng với biến số weight (trọng lượng cơ thể) thì kiểm định này cho biết đây là một biến số tuân theo luật phân phối chuẩn vì trị số p > 0.05.

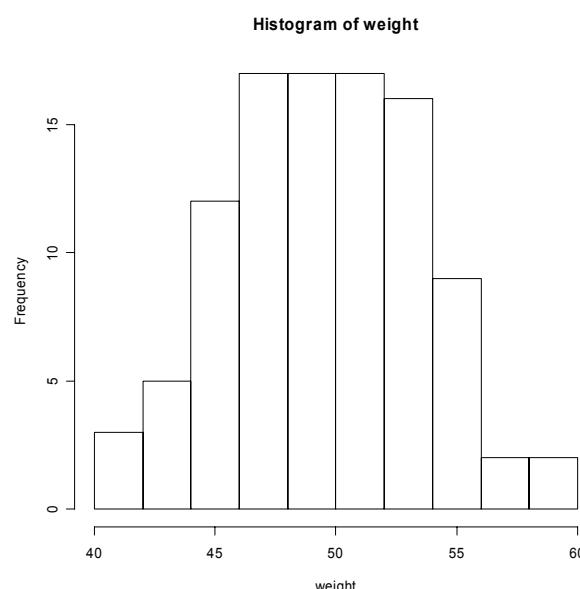
```
> shapiro.test(weight)

Shapiro-Wilk normality test

data: weight
W = 0.9887, p-value = 0.5587
```

Thật ra, kết quả trên cũng phù hợp với đồ thị của weight:

```
> hist(weight)
```



9.3 Thống kê mô tả theo từng nhóm

Nếu chúng ta muốn tính trung bình của một biến số như `igfi` cho mỗi nhóm nam và nữ giới, hàm `tapply` trong R có thể dùng cho việc này:

```
> tapply(igfi, list(sex), mean)
   Female      Male
 167.9741 160.2903
```

Trong lệnh trên, `igfi` là biến số chúng ta cần tính, biến số phân nhóm là `sex`, và chỉ số thống kê chúng ta muốn là trung bình (`mean`). Qua kết quả trên, chúng ta thấy số trung bình của `igfi` cho nữ giới (167.97) cao hơn nam giới (160.29).

Nhưng nếu chúng ta muốn tính cho từng giới tính và sắc tộc, chúng ta chỉ cần thêm một biến số trong hàm `list`:

```
> tapply(igfi, list(ethnicity, sex), mean)
   Female      Male
African    145.1252 120.9168
Asian      165.6589 160.4999
Caucasian 176.6536 169.4790
Others        NA 200.5000
```

Trong kết quả trên, NA có nghĩa là “not available”, tức không có số liệu cho phụ nữ trong các sắc tộc “others”.

9.4 Kiểm định t (`t.test`)

Kiểm định t dựa vào giả thiết phân phối chuẩn. Có hai loại kiểm định t: kiểm định t cho một mẫu (one-sample t-test), và kiểm định t cho hai mẫu (two-sample t-test). Kiểm định t một mẫu nằm trả lời câu hỏi dữ liệu từ một mẫu có phải thật sự bằng một thông số nào đó hay không. **Còn kiểm định t hai mẫu thì nhằm trả lời câu hỏi hai mẫu có cùng một luật phân phối, hay cụ thể hơn là hai mẫu có thật sự có cùng trị số trung bình hay không.** Tôi sẽ lần lượt minh họa hai kiểm định này qua số liệu `igfdata` trên.

9.1.1 Kiểm định t một mẫu

Ví dụ 2. Qua phân tích trên, chúng ta thấy tuổi trung bình của 100 đối tượng trong nghiên cứu này là 19.17 tuổi. Chẳng hạn như trong quần thể này, trước đây chúng ta biết rằng tuổi trung bình là 30 tuổi. Vấn đề đặt ra là có phải mẫu mà chúng ta có được có đại diện cho quần thể hay không. Nói cách khác, chúng ta muốn biết giá trị trung bình 19.17 có thật sự khác với giá trị trung bình 30 hay không.

Để trả lời câu hỏi này, chúng ta sử dụng kiểm định t. Theo lí thuyết thống kê, kiểm định t được định nghĩa bằng công thức sau đây:

$$t = \frac{\bar{x} - \mu}{s / \sqrt{n}}$$

Trong đó, \bar{x} là giá trị trung bình của mẫu, μ là trung bình theo giả thiết (trong trường hợp này, 30), s là độ lệch chuẩn, và n là số lượng mẫu (100). Nếu giá trị t cao hơn giá trị lí thuyết theo phân phối t ở một tiêu chuẩn có ý nghĩa như 5% chẳng hạn thì chúng ta có lí do để phát biểu khác biệt có ý nghĩa thống kê. Giá trị này cho mẫu 100 có thể tính toán bằng hàm `qt` của R như sau:

```
> qt(0.95, 100)
[1] 1.660234
```

Nhưng có một cách tính toán nhanh gọn hơn để trả lời câu hỏi trên, bằng cách dùng hàm `t.test` như sau:

```
> t.test(age, mu=30)

One Sample t-test

data: age
t = -27.6563, df = 99, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 30
95 percent confidence interval:
 18.39300 19.94700
sample estimates:
mean of x
 19.17
```

Trong lệnh trên `age` là biến số chúng ta cần kiểm định, và `mu=30` là giá trị giả thiết. R trình bày trị số $t = -27.66$, với 99 bậc tự do, và trị số $p < 2.2e-16$ (tức rất thấp). R cũng cho biết độ tin cậy 95% của `age` là từ 18.4 tuổi đến 19.9 tuổi (30 tuổi nằm quá ngoài khoảng tin cậy này). Nói cách khác, chúng ta có lí do để phát biểu rằng độ tuổi trung bình trong mẫu này thật sự thấp hơn độ tuổi trung bình của quần thể.

9.4.2 Kiểm định t hai mẫu

Ví dụ 3. Qua phân tích mô tả trên (phàm `summary`) chúng ta thấy phụ nữ có độ hormone `igf1` cao hơn nam giới (167.97 và 160.29). Câu hỏi đặt ra là có phải thật sự đó là một khác biệt có hệ thống hay do các yếu tố ngẫu nhiên gây nên. Trả lời câu hỏi này, chúng ta cần xem xét mức độ khác biệt trung bình giữa hai nhóm và độ lệch chuẩn của độ khác biệt.

$$t = \frac{\bar{x}_2 - \bar{x}_1}{SED}$$

Trong đó \bar{x}_1 và \bar{x}_2 là số trung bình của hai nhóm nam và nữ, và `SED` là độ lệch chuẩn của $(\bar{x}_1 - \bar{x}_2)$. Thực ra, `SED` có thể ước tính bằng công thức:

$$SED = \sqrt{SE_1^2 + SE_2^2}$$

Trong đó SE_1 và SE_2 là sai số chuẩn (standard error) của hai nhóm nam và nữ. Theo lí thuyết xác suất, t tuân theo luật phân phối t với bậc tự do $n_1 + n_2 - 2$, trong đó n_1 và n_2 là số mẫu của hai nhóm. Chúng ta có thể dùng R để trả lời câu hỏi trên bằng hàm `t.test` như sau:

```
> t.test(igfi ~ sex)
```

```
Welch Two Sample t-test
```

```
data: igfi by sex
t = 0.8412, df = 88.329, p-value = 0.4025
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-10.46855 25.83627
sample estimates:
mean in group Female mean in group Male
167.9741      160.2903
```

R trình bày các giá trị quan trọng trước hết:

```
t = 0.8412, df = 88.329, p-value = 0.4025
```

df là bậc tự do. Trị số $p = 0.4025$ cho thấy mức độ khác biệt giữa hai nhóm nam và nữ không có ý nghĩa thống kê (vì cao hơn 0.05 hay 5%).

```
95 percent confidence interval:
-10.46855 25.83627
```

là khoảng tin cậy 95% về độ khác biệt giữa hai nhóm. Kết quả tính toán trên cho biết độ igf ở nữ giới có thể thấp hơn nam giới 10.5 ng/L hoặc cao hơn nam giới khoảng 25.8 ng/L. Vì độ khác biệt quá lớn và đó là thêm bằng chứng cho thấy không có khác biệt có ý nghĩa thống kê giữa hai nhóm.

Kiểm định trên dựa vào giả thiết hai nhóm nam và nữ có khác phương sai. Nếu chúng ta có lí do để cho rằng hai nhóm có cùng phương sai, chúng ta chỉ thay đổi một thông số trong hàm `t` với `var.equal=TRUE` như sau:

```
> t.test(igfi ~ sex, var.equal=TRUE)
```

```
Two Sample t-test
```

```
data: igfi by sex
t = 0.7071, df = 98, p-value = 0.4812
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-13.88137 29.24909
```

```

sample estimates:
mean in group Female    mean in group Male
           167.9741          160.2903

```

Về mặc số, kết quả phân tích trên có khác chút ít so với kết quả phân tích dựa vào giả định hai phương sai khác nhau, nhưng trị số p cũng đi đến một kết luận rằng độ khác biệt giữa hai nhóm không có ý nghĩa thống kê.

9.5 So sánh phương sai (`var.test`)

Bây giờ chúng ta thử kiểm định xem phương sai giữa hai nhóm có khác nhau không. Để tiến hành phân tích, chúng ta chỉ cần lệnh:

```

> var.test(igfi ~ sex)

F test to compare two variances

data: igfi by sex
F = 2.6274, num df = 68, denom df = 30, p-value = 0.004529
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
1.366187 4.691336
sample estimates:
ratio of variances
2.627396

```

Kết quả trên cho thấy độ khác biệt về phương sai giữa hai nhóm cao 2.62 lần. Trị số p = 0.0045 cho thấy phương sai giữa hai nhóm khác nhau có ý nghĩa thống kê. Như vậy, chúng ta chấp nhận kết quả phân tích của hàm `t.test(igfi ~ sex)`.

9.6 Kiểm định Wilcoxon cho hai mẫu (`wilcox.test`)

Kiểm định t dựa vào giả thiết là phân phối của một biến phải tuân theo luật phân phối chuẩn. Nếu giả định này không đúng, kết quả của kiểm định t có thể không hợp lí (valid). Để kiểm định phân phối của `igfi`, chúng ta có thể dùng hàm `shapiro.test` như sau:

```

> shapiro.test(igfi)

Shapiro-Wilk normality test

data: igfi
W = 0.8528, p-value = 1.504e-08

```

Trị số p nhỏ hơn 0.05 rất nhiều, cho nên chúng ta có thể nói rằng phân phối của `igfi` không tuân theo luật phân phối chuẩn. Trong trường hợp này, việc so sánh giữa hai nhóm có thể dựa vào phương pháp phi tham số (non-parametric) có tên là kiểm định

Wilcoxon, vì kiểm định này (không như kiểm định t) không tùy thuộc vào giả định phân phối chuẩn.

```
> wilcox.test(igfi ~ sex)

Wilcoxon rank sum test with continuity correction

data: igfi by sex
W = 1125, p-value = 0.6819
alternative hypothesis: true mu is not equal to 0
```

Trị số $p = 0.682$ cho thấy quả thật độ khác biệt về igfi giữa hai nhóm nam và nữ không có ý nghĩa thống kê. Kết luận này cũng không khác với kết quả phân tích bằng kiểm định t.

9.7 Kiểm định t cho các biến số theo cặp (paired t-test, t.test)

Kiểm định t vừa trình bày trên là cho các nghiên cứu gồm hai nhóm độc lập nhau (như giữa hai nhóm nam và nữ), nhưng không thể ứng dụng cho các nghiên cứu mà một nhóm đối tượng được theo dõi theo thời gian. Tôi tạm gọi các nghiên cứu này là nghiên cứu theo cặp. Trong các nghiên cứu này, chúng ta cần sử dụng một kiểm định t có tên là paired t-test.

Ví dụ 4. Một nhóm bệnh nhân gồm 10 người được điều trị bằng một thuốc nhằm giảm huyết áp. Huyết áp của bệnh nhân được đo lúc khởi đầu nghiên cứu (lúc chưa điều trị), và sau khi điều trị. Số liệu huyết áp của 10 bệnh nhân như sau:

Trước khi điều trị (x_0)	180, 140, 160, 160, 220, 185, 145, 160, 160, 170
Sau khi điều trị (x_1)	170, 145, 145, 125, 205, 185, 150, 150, 145, 155

Câu hỏi đặt ra là độ biến chuyển huyết áp trên có đủ để kết luận rằng thuốc điều trị có hiệu quả giảm áp huyết. Để trả lời câu hỏi này, chúng ta dùng kiểm định t cho từng cặp như sau:

```
> # nhập dữ kiện
> before <- c(180, 140, 160, 160, 220, 185, 145, 160, 160, 170)
> after <- c(170, 145, 145, 125, 205, 185, 150, 150, 145, 155)
> bp <- data.frame(before, after)

> # kiểm định t
> t.test(before, after, paired=TRUE)

Paired t-test

data: before and after
t = 2.7924, df = 9, p-value = 0.02097
```

```

alternative hypothesis: true difference in means is not equal to
0
95 percent confidence interval:
 1.993901 19.006099
sample estimates:
mean of the differences
 10.5

```

Kết quả trên cho thấy sau khi điều trị áp suất máu giảm 10.5 mmHg, và khoảng tin cậy 95% là từ 2.0 mmHg đến 19 mmHg, với trị số $p = 0.0209$. Như vậy, chúng ta có bằng chứng để phát biểu rằng mức độ giảm huyết áp có ý nghĩa thống kê.

Chú ý nếu chúng ta phân tích sai bằng kiểm định thống kê cho hai nhóm độc lập dưới đây thì trị số $p = 0.32$ cho biết mức độ giảm áp suất không có ý nghĩa thống kê!

```

> t.test(before, after)

Welch Two Sample t-test

data: before and after
t = 1.0208, df = 17.998, p-value = 0.3209
alternative hypothesis: true difference in means is not equal to
0
95 percent confidence interval:
 -11.11065 32.11065
sample estimates:
mean of x mean of y
 168.0      157.5

```

9.8 Kiểm định Wilcoxon cho các biến số theo cặp (**wilcox.test**)

Thay vì dùng kiểm định t cho từng cặp, chúng ta cũng có thể sử dụng hàm **wilcox.test** cho cùng mục đích:

```

> wilcox.test(before, after, paired=TRUE)

Wilcoxon signed rank test with continuity correction

data: before and after
V = 42, p-value = 0.02291
alternative hypothesis: true mu is not equal to 0

```

Kết quả trên một lần nữa khẳng định rằng độ giảm áp suất máu có ý nghĩa thống kê với trị số ($p=0.023$) chẳng khác mấy so với kiểm định t cho từng cặp.

9.9 Tân số (frequency)

Hàm `table` trong R có chức năng cho chúng ta biết về tần số của một biến số mang tính phân loại như `sex` và `ethnicity`.

```
> table(sex)
sex
Female    Male
  69      31

> table(ethnicity)
ethnicity
  African     Asian Caucasian   Others
    8          60        30         2
```

Một bảng thống kê 2 chiều:

```
> table(sex, ethnicity)
      ethnicity
      sex      African Asian Caucasian Others
Female      4     43       22       0
Male       4     17        8       2
```

Chú ý trong các bảng thống kê trên, hàm `table` không cung cấp cho chúng ta số phần trăm. Để tính số phần trăm, chúng ta cần đến hàm `prop.table` và cách sử dụng có thể minh họa như sau:

```
# tạo ra một object tên là freq để chứa kết quả tần số
> freq <- table(sex, ethnicity)

# kiểm tra kết quả
> freq
      ethnicity
      sex      African Asian Caucasian Others
Female      4     43       22       0
Male       4     17        8       2

# dùng hàm margin.table để xem kết quả
> margin.table(freq, 1)
sex
Female    Male
  69      31

> margin.table(freq, 2)
ethnicity
  African     Asian Caucasian   Others
    8          60        30         2
```

```
# tính phần trăm bằng hàm prop.table
> prop.table(freq, 1)
  ethnicity
sex      African     Asian Caucasian   Others
Female  0.05797101 0.62318841 0.31884058 0.00000000
Male    0.12903226 0.54838710 0.25806452 0.06451613
```

Trong bảng thống kê trên, prop.table tính tỉ lệ sắc tộc cho từng giới tính. Chẳng hạn như ở nữ giới (female), 5.8% là người Phi châu, 62.3% là người Á châu, 31.8% là người Tây phương da trắng . Tổng cộng là 100%. Tương tự, ở nam giới tỉ lệ người Phi châu là 12.9%, Á châu là 54.8%, v.v...

```
# tính phần trăm bằng hàm prop.table
> prop.table(freq, 2)
  ethnicity
sex      African     Asian Caucasian   Others
Female  0.5000000 0.7166667 0.7333333 0.0000000
Male    0.5000000 0.2833333 0.2666667 1.0000000
```

Trong bảng thống kê trên, prop.table tính tỉ lệ giới tính cho từng sắc tộc. Chẳng hạn như trong nhóm người Á châu, 71.7% là nữ và 28.3% là nam.

```
# tính phần trăm cho toàn bộ bảng
> freq/sum(freq)
  ethnicity
sex      African Asian Caucasian   Others
Female  0.04  0.43      0.22  0.00
Male    0.04  0.17      0.08  0.02
```

9.10 Kiểm định tỉ lệ (proportion test, prop.test, binom.test)

Kiểm định một tỉ lệ thường dựa vào giả định phân phối nhị phân (binomial distribution). Với một số mẫu n và tỉ lệ p , và nếu n lớn (tức hơn 50 chẳng hạn), thì phân phối nhị phân có thể tương đương với phân phối chuẩn với số trung bình np và phương sai $np(1 - p)$. Gọi x là số biến cỏ mà chúng ta quan tâm, kiểm định giả thiết $p = \pi$ có thể sử dụng thống kê sau đây:

$$z = \frac{x - n\pi}{\sqrt{n\pi(1-\pi)}}$$

Ở đây, z tuân theo luật phân phối chuẩn với trung bình 0 và phương sai 1. Cũng có thể nói z^2 tuân theo luật phân phối Chi bình phương với bậc tự do bằng 1.

Ví dụ 5. Trong nghiên cứu trên, chúng ta thấy có 69 nữ và 31 nam. Như vậy tỉ lệ nữ là 0.69 (hay 69%). Để kiểm định xem tỉ lệ này có thật sự khác với tỉ lệ 0.5 hay không, chúng ta có thể sử dụng hàm `prop.test(x, n, pi)` như sau:

```
> prop.test(69, 100, 0.50)

1-sample proportions test with continuity correction

data: 69 out of 100, null probability 0.5
X-squared = 13.69, df = 1, p-value = 0.0002156
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.5885509 0.7766330
sample estimates:
    p
 0.69
```

Trong kết quả trên, `prop.test` ước tính tỉ lệ nữ giới là 0.69, và khoảng tin cậy 95% là 0.588 đến 0.776. Giá trị Chi bình phương là 13.69, với trị số $p = 0.00216$. Như vậy, nghiên cứu này có tỉ lệ nữ cao hơn 50%.

Một cách tính chính xác hơn kiểm định tỉ lệ là kiểm định nhị phân `binom.test(x, n, pi)` như sau:

```
> binom.test(69, 100, 0.50)

Exact binomial test

data: 69 and 100
number of successes = 69, number of trials = 100, p-value = 0.0001831
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.5896854 0.7787112
sample estimates:
probability of success
 0.69
```

Nói chung, kết quả của kiểm định nhị phân không khác gì so với kiểm định Chi bình phương, với trị số $p = 0.00018$, chúng ta càng có bằng chứng để kết luận rằng tỉ lệ nữ giới trong nghiên cứu này thật sự cao hơn 50%.

9.11 So sánh hai tỉ lệ (`prop.test`, `binom.test`)

Phương pháp so sánh hai tỉ lệ có thể khai triển trực tiếp từ lí thuyết kiểm định một tỉ lệ vừa trình bày trên. Cho hai mẫu với số đối tượng n_1 và n_2 , và số biến cố là x_1 và x_2 . Do đó, chúng ta có thể ước tính hai tỉ lệ p_1 và p_2 . Lí thuyết xác suất cho phép chúng ta phát biểu rằng độ khác biệt giữa hai mẫu $d = p_1 - p_2$ tuân theo luật phân phối chuẩn với số trung bình 0 và phương sai bằng:

$$V_d = \left(\frac{1}{n_1} + \frac{1}{n_2} \right) p(1-p)$$

Trong đó:

$$p = \frac{x_1 + x_2}{n_1 + n_2}$$

Thành ra, $z = d/V_d$ tuân theo luật phân phối chuẩn với trung bình 0 và phương sai 1. Nói cách khác, z^2 tuân theo luật phân phối Chi bình phương với bậc tự do bằng 1. Do đó, chúng ta cũng có thể sử dụng `prop.test` để kiểm định hai tỉ lệ.

Ví dụ 6. Một nghiên cứu được tiến hành so sánh hiệu quả của thuốc chống gãy xương. Bệnh nhân được chia thành hai nhóm: nhóm A được điều trị gồm có 100 bệnh nhân, và nhóm B không được điều trị gồm 110 bệnh nhân. Sau thời gian 12 tháng theo dõi, nhóm A có 7 người bị gãy xương, và nhóm B có 20 người gãy xương. Vấn đề đặt ra là tỉ lệ gãy xương trong hai nhóm này bằng nhau (tức thuốc không có hiệu quả)? Để kiểm định xem hai tỉ lệ này có thật sự khác nhau, chúng ta có thể sử dụng hàm `prop.test(x, n, pi)` như sau:

```
> fracture <- c(7, 20)
> total <- c(100, 110)
> prop.test(fracture, total)

 2-sample test for equality of proportions with continuity
correction

data: fracture out of total
X-squared = 4.8901, df = 1, p-value = 0.02701
alternative hypothesis: two.sided
95 percent confidence interval:
-0.20908963 -0.01454673
sample estimates:
prop 1    prop 2
0.0700000 0.1818182
```

Kết quả phân tích trên cho thấy tỉ lệ gãy xương trong nhóm 1 là 0.07 và nhóm 2 là 0.18. Phân tích trên còn cho thấy xác suất 95% rằng độ khác biệt giữa hai nhóm có thể 0.01 đến 0.20 (tức 1 đến 20%). Với trị số $p = 0.027$, chúng ta có thể nói rằng tỉ lệ gãy xương trong nhóm A quả thật thấp hơn nhóm B.

9.12 So sánh nhiều tỉ lệ (`prop.test`, `chisq.test`)

Kiểm định `prop.test` còn có thể sử dụng để kiểm định nhiều tỉ lệ cùng một lúc. Trong nghiên cứu trên, chúng ta có 4 nhóm sắc tộc và tần số cho từng giới tính như sau:

```
> table(sex, ethnicity)
    ethnicity
    sex      African Asian Caucasian Others
Female        4     43       22       0
Male         4     17        8       2
```

Chúng ta muốn biết tỉ lệ nữ giới giữa 4 nhóm sắc tộc có khác nhau hay không, và để trả lời câu hỏi này, chúng ta lại dùng `prop.test` như sau:

```
> female <- c( 4, 43, 22, 0)
> total <- c(8, 60, 30, 2)
> prop.test(female, total)

 4-sample test for equality of proportions without continuity
 correction

data: female out of total
X-squared = 6.2646, df = 3, p-value = 0.09942
alternative hypothesis: two.sided
sample estimates:
prop 1   prop 2   prop 3   prop 4
0.5000000 0.7166667 0.7333333 0.0000000

Warning message:
Chi-squared approximation may be incorrect in: prop.test(female, total)
```

Tuy tỉ lệ nữ giới giữa các nhóm có vẻ khác nhau lớn (73% trong nhóm 3 (người da trắng) so với 50% trong nhóm 1 (Phi châu) và 71.7% trong nhóm Á châu, nhưng kiểm định Chi bình phương cho biết trên phương diện thống kê, các tỉ lệ này không khác nhau, vì trị số $p = 0.099$.

9.12.1 Kiểm định Chi bình phương (Chi squared test, `chisq.test`)

Thật ra, kiểm định Chi bình phương còn có thể tính toán bằng hàm `chisq.test` như sau:

```
> chisq.test(sex, ethnicity)

Pearson's Chi-squared test

data: sex and ethnicity
X-squared = 6.2646, df = 3, p-value = 0.09942

Warning message:
Chi-squared approximation may be incorrect in: chisq.test(sex,
ethnicity)
```

Kết quả này hoàn toàn giống với kết quả từ hàm `prop.test`.

9.12.2 Kiểm định Fisher (Fisher's exact test, `fisher.test`)

Trong kiểm định Chi bình phương trên, chúng ta chú ý cảnh báo:

"Warning message:
Chi-squared approximation may be incorrect in: prop.test(female, total)"

Vì trong nhóm 4, không có nữ giới cho nên tỉ lệ là 0%. Hơn nữa, trong nhóm này chỉ có 2 đối tượng. Vì số lượng đối tượng quá nhỏ, cho nên các ước tính thống kê có thể không đáng tin cậy. Một phương pháp khác có thể áp dụng cho các nghiên cứu với tần số thấp như trên là kiểm định fisher (còn gọi là Fisher's exact test). Bạn đọc có thể tham khảo lí thuyết đằng sau kiểm định fisher để hiểu rõ hơn về logic của phương pháp này, nhưng ở đây, chúng ta chỉ quan tâm đến cách dùng R để tính toán kiểm định này. Chúng ta chỉ đơn giản lệnh:

```
> fisher.test(sex, ethnicity)

Fisher's Exact Test for Count Data

data: sex and ethnicity
p-value = 0.1048
alternative hypothesis: two.sided
```

Chú ý trị số p từ kiểm định Fisher là 0.1048, tức rất gần với trị số p của kiểm định Chi bình phương. Cho nên, chúng ta có thêm bằng chứng để khẳng định rằng tỉ lệ nữ giới giữa các sắc tộc không khác nhau một cách đáng kể.

10

Phân tích hồi qui tuyến tính

Phân tích hồi qui tuyến tính (linear regression analysis) có lẽ là một trong những phương pháp phân tích số liệu thông dụng nhất trong thống kê học. Anon từng viết “Cho con người 3 vũ khí – hệ số tương quan, hồi qui tuyến tính và một cây bút, con người sẽ sử dụng cả ba”! Trong chương này, tôi sẽ giới thiệu cách sử dụng R để phân tích hồi qui tuyến tính và các phương pháp liên quan như hệ số tương quan và kiểm định giả thiết thống kê.

Ví dụ 1. Để minh họa cho vấn đề, chúng ta thử xem xét nghiên cứu sau đây, mà trong đó nhà nghiên cứu đo lường độ cholesterol trong máu của 18 đối tượng nam. Tỉ trọng cơ thể (body mass index) cũng được ước tính cho mỗi đối tượng bằng công thức tính BMI là lấy trọng lượng (tính bằng kg) chia cho chiều cao bình phương (m^2). Kết quả đo lường như sau:

Bảng 1. Độ tuổi, tỉ trọng cơ thể và cholesterol

Mã số ID (id)	Độ tuổi (age)	BMI (bmi)	Cholesterol (cho1)
1	46	25.4	3.5
2	20	20.6	1.9
3	52	26.2	4.0
4	30	22.6	2.6
5	57	25.4	4.5
6	25	23.1	3.0
7	28	22.7	2.9
8	36	24.9	3.8
9	22	19.8	2.1
10	43	25.3	3.8
11	57	23.2	4.1
12	33	21.8	3.0
13	22	20.9	2.5
14	63	26.7	4.6
15	40	26.4	3.2
16	48	21.2	4.2
17	28	21.2	2.3
18	49	22.8	4.0

Nhìn sơ qua số liệu chúng ta thấy người có độ tuổi càng cao độ cholesterol cũng càng cao. Chúng ta thử nhập số liệu này vào R và vẽ một biểu đồ tán xạ như sau:

```
> age <- c(46,20,52,30,57,25,28,36,22,43,57,33,22,63,40,48,28,49)
```

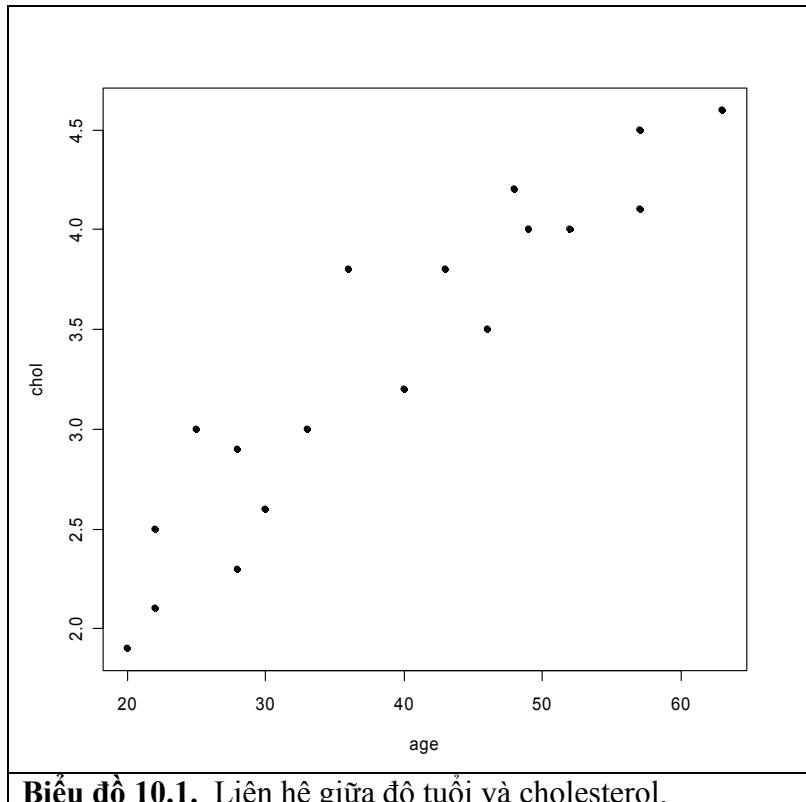
```

> bmi <- c(25.4,20.6,26.2,22.6,25.4,23.1,22.7,24.9,19.8,25.3,23.2,
  21.8,20.9,26.7,26.4,21.2,21.2,22.8)

> chol <- c(3.5,1.9,4.0,2.6,4.5,3.0,2.9,3.8,2.1,3.8,4.1,3.0,
  2.5,4.6,3.2, 4.2,2.3,4.0)

> data <- data.frame(age, bmi, chol)
> plot(chol ~ age, pch=16)

```



Biểu đồ 10.1. Liên hệ giữa độ tuổi và cholesterol.

Biểu đồ 10.1 trên đây gợi ý cho thấy mối liên hệ giữa độ tuổi (age) và cholesterol là một đường thẳng (tuyến tính). Để “đo lường” mối liên hệ này, chúng ta có thể sử dụng hệ số tương quan (coefficient of correlation).

10.1 Hệ số tương quan

Hệ số tương quan (r) là một chỉ số thống kê đo lường mối liên hệ tương quan giữa hai biến số, như giữa độ tuổi (x) và cholesterol (y). Hệ số tương quan có giá trị từ -1 đến 1. Hệ số tương quan bằng 0 (hay gần 0) có nghĩa là hai biến số không có liên hệ gì với nhau; ngược lại nếu hệ số bằng -1 hay 1 có nghĩa là hai biến số có một mối liên hệ tuyệt đối. Nếu giá trị của hệ số tương quan là âm ($r < 0$) có nghĩa là khi x tăng cao thì y giảm (và ngược lại, khi x giảm thì y tăng); nếu giá trị hệ số tương quan là dương ($r > 0$) có nghĩa là khi x tăng cao thì y cũng tăng, và khi x tăng cao thì y cũng giảm theo.

Thực ra có nhiều hệ số tương quan trong thống kê, nhưng ở đây tôi sẽ trình bày 3 hệ số tương quan thông dụng nhất: hệ số tương quan Pearson r , Spearman ρ , và Kendall τ .

10.1.1 Hệ số tương quan Pearson

Cho hai biến số x và y từ n mẫu, hệ số tương quan Pearson được ước tính bằng công thức sau đây:

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}$$

Trong đó, như định nghĩa phần trên, \bar{x} và \bar{y} là giá trị trung bình của biến số x và y . Để ước tính hệ số tương quan giữa độ tuổi age và cholesterol, chúng ta có thể sử dụng hàm `cor(x, y)` như sau:

```
> cor(age, chol)
[1] 0.936726
```

Chúng ta có thể kiểm định giả thiết hệ số tương quan bằng 0 (tức hai biến x và y không có liên hệ). Phương pháp kiểm định này thường dựa vào phép biến đổi Fisher mà R đã có sẵn một hàm `cor.test` để tiến hành việc tính toán.

```
> cor.test(age, chol)

Pearson's product-moment correlation

data: age and chol
t = 10.7035, df = 16, p-value = 1.058e-08
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
0.8350463 0.9765306
sample estimates:
cor
0.936726
```

Kết quả phân tích cho thấy kiểm định $t = 10.70$ với trị số $p = 1.058e-08$; do đó, chúng ta có bằng chứng để kết luận rằng mối liên hệ giữa độ tuổi và cholesterol có ý nghĩa thống kê. Kết luận này cũng chính là kết luận chúng ta đã đi đến trong phần phân tích hồi qui tuyến tính trên.

10.1.2 Hệ số tương quan Spearman ρ

Hệ số tương quan Pearson chỉ hợp lý nếu biến số x và y tuân theo luật phân phối chuẩn. Nếu x và y không tuân theo luật phân phối chuẩn, chúng ta phải sử dụng một hệ số tương quan khác tên là Spearman, một phương pháp phân tích phi tham số. Hệ số này

được ước tính bằng cách biến đổi hai biến số x và y thành thứ bậc (rank), và xem độ tương quan giữa hai dãy số bậc. Do đó, hệ số còn có tên tiếng Anh là Spearman's Rank correlation. R ước tính hệ số tương quan Spearman bằng hàm `cor.test` với thông số `method="spearman"` như sau:

```
> cor.test(age, chol, method="spearman")

  Spearman's rank correlation rho

data: age and chol
S = 51.1584, p-value = 2.57e-09
alternative hypothesis: true rho is not equal to 0
sample estimates:
rho
0.947205

Warning message:
Cannot compute exact p-values with ties in: cor.test.default(age,
chol, method = "spearman")
```

Kết quả phân tích cho thấy giá trị rho = 0.947, và trị số p = 2.57e-09. Kết quả từ phân tích này cũng không khác với phân tích hồi qui tuyến tính: mối liên hệ giữa độ tuổi và cholesterol rất cao và có ý nghĩa thống kê.

10.1.3 Hệ số tương quan Kendall τ

Hệ số tương quan Kendall (cũng là một phương pháp phân tích phi tham số) được ước tính bằng cách tìm các cặp số (x, y) “song hành” với nhau. Một cặp (x, y) song hành ở đây được định nghĩa là hiệu (độ khác biệt) trên trực hoành có cùng dấu hiệu (dương hay âm) với hiệu trên trực tung. Nếu hai biến số x và y không có liên hệ với nhau, thì số cặp song hành bằng hay tương đương với số cặp không song hành.

Bởi vì có nhiều cặp phải kiểm định, phương pháp tính toán hệ số tương quan Kendall đòi hỏi thời gian của máy tính khá cao. Tuy nhiên, nếu một dữ liệu dưới 5000 đối tượng thì một máy vi tính có thể tính toán khá dễ dàng. R dùng hàm `cor.test` với thông số `method="kendall"` để ước tính hệ số tương quan Kendall:

```
> cor.test(age, chol, method="kendall")

  Kendall's rank correlation tau

data: age and chol
z = 4.755, p-value = 1.984e-06
alternative hypothesis: true tau is not equal to 0
sample estimates:
tau
0.8333333

Warning message:
```

```
Cannot compute exact p-value with ties in: cor.test.default(age,
chol, method = "kendall")
```

Kết quả phân tích hệ số tương quan Kendall một lần nữa khẳng định mối liên hệ giữa độ tuổi và cholesterol có ý nghĩa thống kê, vì hệ số tau = 0.833 và trị số p = 1.98e-06.

Các hệ số tương quan trên đây đo mức độ tương quan giữa hai biến số, nhưng không cho chúng ta một phương trình để nối hai biến số đó với nhau. Thành ra, vẫn đề đặt ra là chúng ta tìm một phương trình tuyến tính để mô tả mối liên hệ này. Chúng ta sẽ ứng dụng mô hình hồi qui tuyến tính.

10.2 Mô hình của hồi qui tuyến tính đơn giản

10.2.1 vài dòng lí thuyết

Để tiện việc theo dõi và mô tả mô hình, gọi độ tuổi cho cá nhân i là x_i và cholesterol là y_i . Ở đây $i = 1, 2, 3, \dots, 18$. Mô hình hồi qui tuyến tính phát biểu rằng:

$$y_i = \alpha + \beta x_i + \varepsilon_i \quad [1]$$

Nói cách khác, phương trình trên giả định rằng độ cholesterol của một cá nhân bằng một hằng số α cộng với một hệ số β liên quan đến độ tuổi, và một sai số ε_i . Trong phương trình trên, α là *chặn* (intercept, tức giá trị lúc $x_i = 0$), và β là độ dốc (slope hay gradient). Trong thực tế, α và β là hai thông số (paramater, còn gọi là *regression coefficient* hay hệ số hồi qui), và ε_i là một biến số theo luật phân phối chuẩn với trung bình 0 và phương sai σ^2 .

Các thông số α , β và σ^2 phải được ước tính từ dữ liệu. Phương pháp để ước tính các thông số này là phương pháp *bình phương nhỏ nhất* (least squares method). Như tên gọi, phương pháp bình phương nhỏ nhất tìm giá trị α , β sao cho $\sum_{i=1}^n [y_i - (\alpha + \beta x_i)]^2$ nhỏ nhất. Sau vài thao tác toán, có thể chứng minh dễ dàng rằng, ước số cho α và β đáp ứng điều kiện đó là:

$$\hat{\beta} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad [2]$$

và

$$\hat{\alpha} = \bar{y} - \hat{\beta} \bar{x} \quad [3]$$

Ở đây, \bar{x} và \bar{y} là giá trị trung bình của biến số x và y . Chú ý, tôi viết $\hat{\alpha}$ và $\hat{\beta}$ (với dấu mũ phía trên) là để nhắc nhở rằng đây là hai ước số (estimates) của α và β , chứ không phải α và β (chúng ta không biết chính xác α và β , nhưng chỉ có thể ước tính mà thôi).

Sau khi đã có ước số $\hat{\alpha}$ và $\hat{\beta}$, chúng ta có thể ước tính độ cholesterol trung bình cho từng độ tuổi như sau:

$$\hat{y}_i = \hat{\alpha} + \hat{\beta}x_i$$

Tất nhiên, \hat{y}_i ở đây chỉ là số trung bình cho độ tuổi x_i , và phần còn lại (tức $y_i - \hat{y}_i$) gọi là *phần dư (residual)*. Và phương sai của phần dư có thể ước tính như sau:

$$s^2 = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n-2} \quad [4]$$

s^2 chính là ước số của σ^2 .

Trong phân tích hồi qui tuyến tính, thông thường chúng ta muốn biết hệ số $\beta = 0$ hay khác 0. Nếu β bằng 0, thì cũng có nghĩa là không có mối liên hệ gì giữa x và y ; nếu β khác với 0, chúng ta có bằng chứng để phát biểu rằng x và y có liên quan nhau. Để kiểm định giả thiết $\beta = 0$ chúng ta dùng xét nghiệm t sau đây:

$$t = \frac{\hat{\beta}}{SE(\hat{\beta})} \quad [5]$$

$SE(\hat{\beta})$ có nghĩa là sai số chuẩn (standard error) của ước số $\hat{\beta}$. Trong phương trình trên, t tuân theo luật phân phối t với bậc tự do $n-2$ (nếu thật sự $\beta = 0$).

10.2.2 Phân tích hồi qui tuyến tính đơn giản bằng R

Hàm `lm` (viết tắt từ linear model) trong R có thể tính toán các giá trị của $\hat{\alpha}$ và $\hat{\beta}$, cũng như s^2 một cách nhanh gọn. Chúng ta tiếp tục với ví dụ bằng R như sau:

```
> lm(chol ~ age)
```

Call:

```
lm(formula = chol ~ age)
```

Coefficients:

(Intercept)	age
1.08922	0.05779

Trong lệnh trên, “`chol ~ age`” có nghĩa là mô tả `chol` là một hàm số của `age`. Kết quả tính toán của `lm` cho thấy $\hat{\alpha} = 1.0892$ và $\hat{\beta} = 0.05779$. Nói cách khác, với hai thông số này, chúng ta có thể ước tính độ cholesterol cho bất cứ độ tuổi nào trong khoảng tuổi của mẫu bằng phương trình tuyến tính:

$$\hat{y}_i = 1.08922 + 0.05779 \times \text{age}$$

Phương trình này có nghĩa là khi độ tuổi tăng 1 năm thì độ cholesterol tăng khoảng 0.058 mmol/L.

Thật ra, hàm `lm` còn cung cấp cho chúng ta nhiều thông tin khác, nhưng chúng ta phải đưa các thông tin này vào một object. Gọi object đó là `reg`, thì lệnh sẽ là:

```
> reg <- lm(chol ~ age)
> summary(reg)

Call:
lm(formula = chol ~ age)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.40729 -0.24133 -0.04522  0.17939  0.63040 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 1.089218   0.221466  4.918 0.000154 ***
age         0.057788   0.005399 10.704 1.06e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3027 on 16 degrees of freedom
Multiple R-Squared:  0.8775,    Adjusted R-squared:  0.8698 
F-statistic: 114.6 on 1 and 16 DF,  p-value: 1.058e-08
```

Lệnh thứ hai, `summary(reg)`, yêu cầu R liệt kê các thông tin tính toán trong `reg`. Phần kết quả chia làm 3 phần:

(a) Phần 1 mô tả phần dư (residuals) của mô hình hồi qui:

```
Residuals:
    Min      1Q  Median      3Q     Max 
-0.40729 -0.24133 -0.04522  0.17939  0.63040 
```

Chúng ta biết rằng trung bình phần dư phải là 0, và ở đây, số trung vị là -0.04, cũng không xa 0 bao nhiêu. Các số quantiles 25% (1Q) và 75% (3Q) cũng khá cân đối chung quan số trung vị, cho thấy phần dư của phương trình này tương đối cân đối.

(b) Phần hai trình bày ước số của $\hat{\alpha}$ và $\hat{\beta}$ cùng với sai số chuẩn và giá trị của kiểm định t. Giá trị kiểm định t cho $\hat{\beta}$ là 10.74 với trị số p = 1.06e-08, cho thấy β không phải bằng 0. Nói cách khác, chúng ta có bằng chứng để cho rằng có một mối liên hệ giữa cholesterol và độ tuổi, và mối liên hệ này có ý nghĩa thống kê.

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	1.089218	0.221466	4.918	0.000154 ***
age	0.057788	0.005399	10.704	1.06e-08 ***

Signif. codes:	0 '***'	0.001 '**'	0.01 '*'	0.05 '.'
	0.1	'	'	1

(c) Phần ba của kết quả cho chúng ta thông tin về phương sai của phần dư (residual mean square). Ở đây, $s^2 = 0.3027$. Trong kết quả này còn có kiểm định F, cũng chỉ là một kiểm định xem có quả thật β bằng 0, tức có ý nghĩa tương tự như kiểm định t trong phần trên. Nói chung, trong trường hợp phân tích hồi qui tuyến tính đơn giản (với một yếu tố) chúng ta không cần phải quan tâm đến kiểm định F.

Residual standard error: 0.3027 on 16 degrees of freedom
Multiple R-Squared: 0.8775, Adjusted R-squared: 0.8698
F-statistic: 114.6 on 1 and 16 DF, p-value: 1.058e-08

Ngoài ra, phần 3 còn cho chúng ta một thông tin quan trọng, đó là trị số R^2 hay *hệ số xác định bội* (coefficient of determination). Hệ số này được ước tính bằng công thức:

$$R^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad [6]$$

Tức là bằng tổng bình phương giữa số ước tính và trung bình chia cho tổng bình phương số quan sát và trung bình. Trị số R^2 trong ví dụ này là 0.8775, có nghĩa là phương trình tuyến tính (với độ tuổi là một yếu tố) giải thích khoảng 88% các khía cạnh riêng về độ cholesterol giữa các cá nhân. Tất nhiên trị số R^2 có giá trị từ 0 đến 100% (hay 1). Giá trị R^2 càng cao là một dấu hiệu cho thấy mối liên hệ giữa hai biến số độ tuổi và cholesterol càng chặt chẽ.

Một hệ số cũng cần đề cập ở đây là *hệ số điều chỉnh xác định bội* (mà trong kết quả trên R gọi là “Adjusted R-squared”). Đây là hệ số cho chúng ta biết mức độ cải tiến của phương sai phần dư (residual variance) do yếu tố độ tuổi có mặt trong mô hình tuyến tính. Nói chung, hệ số này không khác mấy so với hệ số xác định bội, và chúng ta cũng không cần chú tâm quá mức.

10.2.3 Giả định của phân tích hồi qui tuyến tính

Tất cả các phân tích trên dựa vào một số giả định quan trọng như sau:

- (a) x là một biến số cố định hay fixed, (“cố định” ở đây có nghĩa là không có sai sót ngẫu nhiên trong đo lường);
- (b) ε_i phân phối theo luật phân phối chuẩn;
- (c) ε_i có giá trị trung bình (mean) là 0;
- (d) ε_i có phương sai σ^2 cố định cho tất cả x_i ; và
- (e) các giá trị liên tục của ε_i không có liên hệ tương quan với nhau (nói cách khác, ε_1 và ε_2 không có liên hệ với nhau).

Nếu các giả định này không được đáp ứng thì phương trình mà chúng ta ước tính có vấn đề hợp lý (validity). Do đó, trước khi trình bày và diễn dịch mô hình trên, chúng ta cần phải kiểm tra xem các giả định trên có đáp ứng được hay không. Trong trường hợp này, giả định (a) không phải là vấn đề, vì độ tuổi không phải là một biến số ngẫu nhiên, và không có sai số khi tính độ tuổi của một cá nhân.

Đối với các giả định (b) đến (e), cách kiểm tra đơn giản nhưng hữu hiệu nhất là bằng cách xem xét mối liên hệ giữa \hat{y}_i , x_i , và phần dư e_i ($e_i = y_i - \hat{y}_i$) bằng những đồ thị tán xạ.

Với lệnh `fitted()` chúng ta có thể tính toán \hat{y}_i cho từng cá nhân như sau (ví dụ đối với cá nhân 1, 46 tuổi, độ cholesterol có thể tiên đoán như sau: $1.08922 + 0.05779 \times 46 = 3.747$).

```
> fitted(reg)
     1          2          3          4          5          6          7          8
3.747483 2.244985 4.094214 2.822869 4.383156 2.533927 2.707292 3.169600
     9         10         11         12         13         14         15         16
2.360562 3.574118 4.383156 2.996234 2.360562 4.729886 3.400753 3.863060
     17        18
2.707292 3.920849
```

Với lệnh `resid()` chúng ta có thể tính toán phần dư e_i cho từng cá nhân như sau (với đối tượng 1, $e_1 = 3.5 - 3.74748 = -0.24748$):

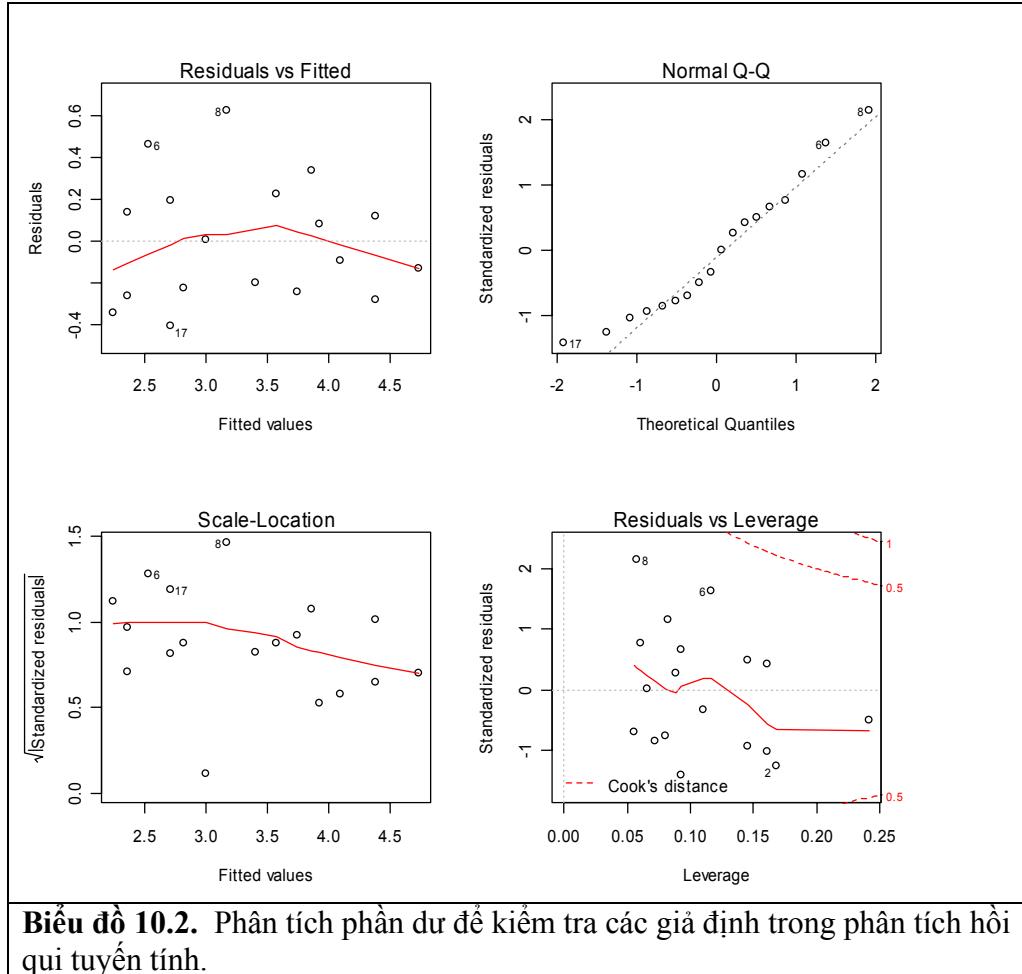
```
> resid(reg)
     1          2          3          4          5          6
-0.247483426 -0.344985415 -0.094213736 -0.222869265 0.116844338 0.466072660
     7          8          9          10         11         12
0.192707505 0.630400424 -0.260562185 0.225881729 -0.283155662 0.003765579
     13         14         15         16         17         18
0.139437815 -0.129885972 -0.200753116 0.336939804 -0.407292495 0.079151419
```

Để kiểm tra các giả định trên, chúng ta có thể vẽ một loạt 4 đồ thị mà tôi sẽ giải thích sau đây:

```

> op <- par(mfrow=c(2, 2))
# yêu cầu R dành ra 4 cửa sổ
> plot(reg) # vẽ các đồ thị trong reg

```



Biểu đồ 10.2. Phân tích phần dư để kiểm tra các giả định trong phân tích hồi qui tuyến tính.

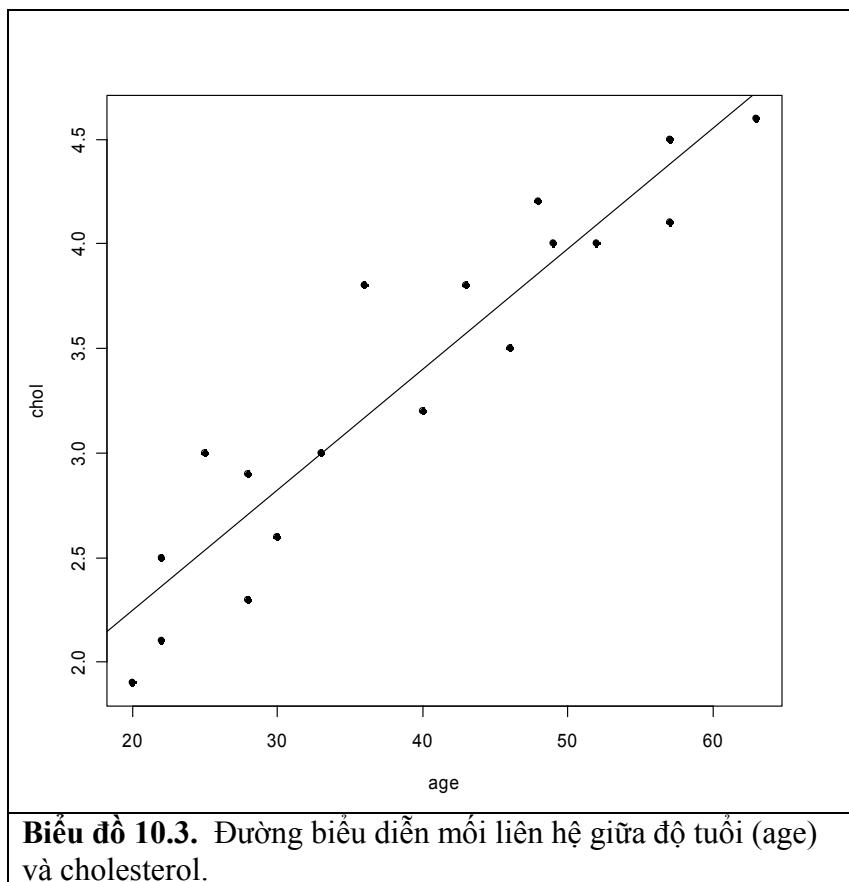
- Đồ thị bên trái dòng 1 vẽ phần dư e_i và giá trị tiên đoán cholesterol \hat{y}_i . Đồ thị này cho thấy các giá trị phần dư tập chung quanh đường $y = 0$, cho nên giả định (c), hay ε_i có giá trị trung bình 0, là có thể chấp nhận được.
- Đồ thị bên phải dòng 1 vẽ giá trị phần dư và giá trị kì vọng dựa vào phân phối chuẩn. Chúng ta thấy các số phần dư tập trung rất gần các giá trị trên đường chuẩn, và do đó, giả định (b), tức ε_i phân phối theo luật phân phối chuẩn, cũng có thể đáp ứng.
- Đồ thị bên trái dòng 2 vẽ căn số phần dư chuẩn (standardized residual) và giá trị của \hat{y}_i . Đồ thị này cho thấy không có gì khác nhau giữa các số phần dư chuẩn cho các giá trị của \hat{y}_i , và do đó, giả định (d), tức ε_i có phương sai σ^2 cố định cho tất cả x_i , cũng có thể đáp ứng.

Nói chung qua phân tích phần dự, chúng ta có thể kết luận rằng mô hình hồi qui tuyến tính mô tả mối liên hệ giữa độ tuổi và cholesterol một cách khá đầy đủ và hợp lí.

10.2.4 Mô hình tiên đoán

Sau khi mô hình tiên đoán cholesterol đã được kiểm tra và tính hợp lí đã được thiết lập, chúng ta có thể vẽ đường biểu diễn của mối liên hệ giữa độ tuổi và cholesterol bằng lệnh `abline` như sau (xin nhắc lại object của phân tích là `reg`):

```
> plot(chol ~ age, pch=16)
> abline(reg)
```



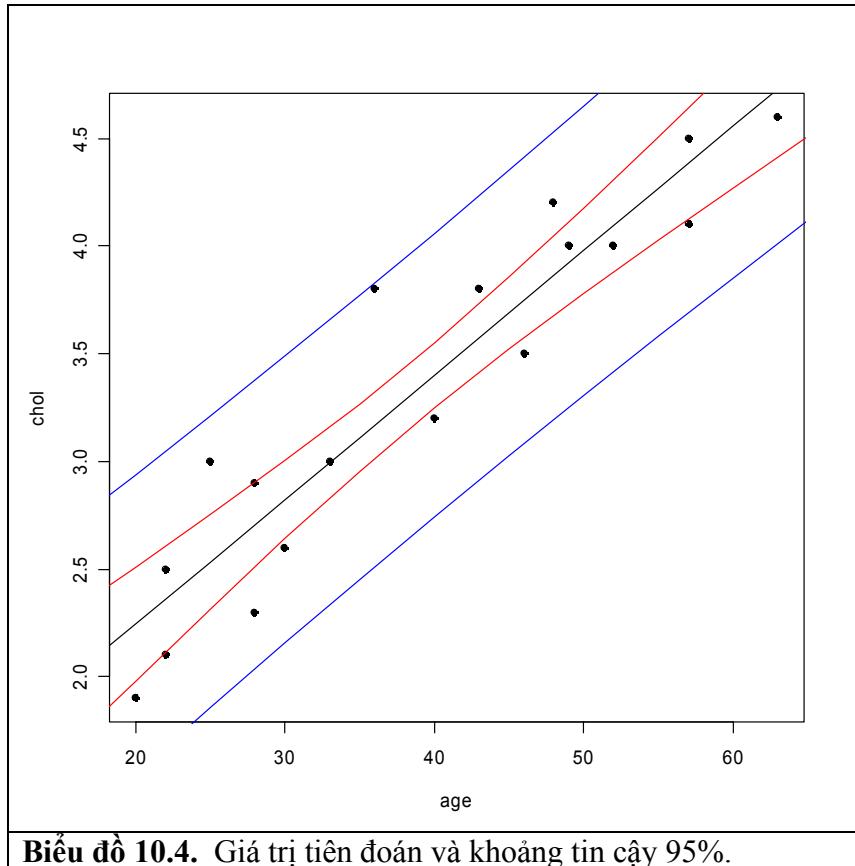
Nhưng mỗi giá trị \hat{y}_i được tính từ ước số $\hat{\alpha}$ và $\hat{\beta}$, mà các ước số này đều có sai số chuẩn, cho nên giá trị tiên đoán \hat{y}_i cũng có sai số. Nói cách khác, \hat{y}_i chỉ là trung bình, nhưng trong thực tế có thể cao hơn hay thấp hơn tùy theo chọn mẫu. Khoảng tin cậy 95% này có thể ước tính qua R bằng các lệnh sau đây:

```
> reg <- lm(chol ~ age)
> new <- data.frame(age = seq(15, 70, 5))
```

```

> pred.w.plim <- predict.lm(reg, new, interval="prediction")
> pred.w.clim <- predict.lm(reg, new, interval="confidence")
> resc <- cbind(pred.w.clim, new)
> resp <- cbind(pred.w.plim, new)
> plot(chol ~ age, pch=16)
> lines(resc$fit ~ resc$age)
> lines(resc$lwr ~ resc$age, col=2)
> lines(resc$upr ~ resc$age, col=2)
> lines(resp$lwr ~ resp$age, col=4)
> lines(resp$upr ~ resp$age, col=4)

```



Biểu đồ trên vẽ giá trị tiên đoán trung bình \hat{y}_i (đường thẳng màu đen), và khoảng tin cậy 95% của giá trị này là đường màu đỏ. Ngoài ra, đường màu xanh là khoảng tin cậy của giá trị tiên đoán cholesterol cho một độ tuổi mới trong quần thể.

10.3 Mô hình hồi qui tuyến tính đa biến (multiple linear regression)

Mô hình được diễn đạt qua phương trình [1] $y_i = \alpha + \beta x_i + \varepsilon_i$ có một yếu tố duy nhất (đó là x), và vì thế thường được gọi là mô hình hồi qui tuyến tính đơn giản (simple

linear regression model). Trong thực tế, chúng ta có thể phát triển mô hình này thành nhiều biến, chứ không chỉ giới hạn một biến như trên, chẳng hạn như:

$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ki} + \varepsilon_i \quad [7]$$

nói cụ thể hơn:

$$\begin{aligned} y_1 &= \alpha + \beta_1 x_{11} + \beta_2 x_{21} + \dots + \beta_k x_{k1} + \varepsilon_1 \\ y_2 &= \alpha + \beta_1 x_{12} + \beta_2 x_{22} + \dots + \beta_k x_{k2} + \varepsilon_2 \\ y_3 &= \alpha + \beta_1 x_{13} + \beta_2 x_{23} + \dots + \beta_k x_{k3} + \varepsilon_3 \\ &\dots \\ y_n &= \alpha + \beta_1 x_{1n} + \beta_2 x_{2n} + \dots + \beta_k x_{kn} + \varepsilon_n \end{aligned}$$

Chú ý trong phương trình trên, chúng ta có nhiều biến x (x_1, x_2, \dots đến x_k), và mỗi biến có một thông số β_j ($j = 1, 2, \dots, k$) cần phải ước tính. Vì thế mô hình này còn được gọi là mô hình hồi qui tuyến tính đa biến.

Phương pháp ước tính β_j cũng chủ yếu dựa vào phương pháp bình phương nhỏ nhất. Gọi $\hat{y}_i = \hat{\alpha} + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \dots + \hat{\beta}_k x_{ki}$ là ước tính của y_i , phương pháp bình phương nhỏ nhất tìm giá trị $\hat{\alpha}, \hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_k$ sao cho $\sum_{i=1}^n (y_i - \hat{y}_i)^2$ nhỏ nhất. Đối với mô hình hồi qui tuyến tính đa biến, cách viết và mô tả mô hình gọn nhất là dùng kí hiệu ma trận. Mô hình [7] có thể thể hiện bằng kí hiệu ma trận như sau:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

Trong đó: \mathbf{Y} là một vector $n \times 1$, \mathbf{X} là một matrix $n \times k$ phần tử, $\boldsymbol{\beta}$ và một vector $k \times 1$, và $\boldsymbol{\varepsilon}$ là vector gồm $n \times 1$ phần tử:

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{bmatrix}, \quad X = \begin{bmatrix} 1 & x_{11} & x_{21} & \dots & x_{k1} \\ 1 & x_{12} & x_{22} & \dots & x_{k2} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & x_{1n} & x_{2n} & \dots & x_{kn} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \dots \\ \beta_k \end{bmatrix}, \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \dots \\ \varepsilon_n \end{bmatrix}$$

Phương pháp bình phương nhỏ nhất giải vector $\boldsymbol{\beta}$ bằng phương trình sau đây:

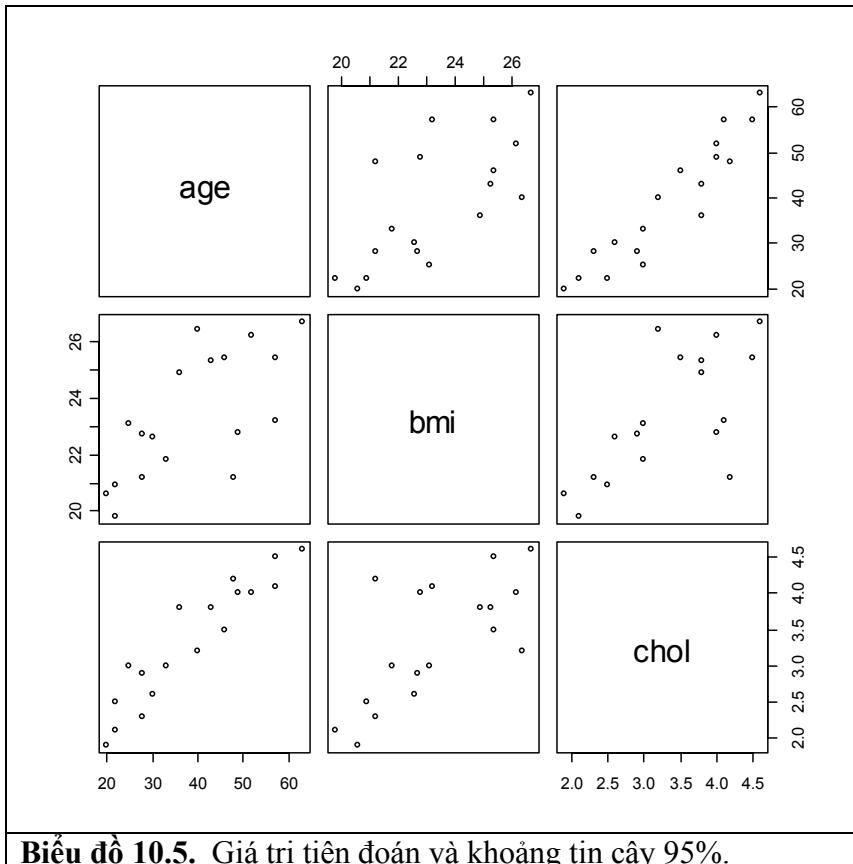
$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$$

và tổng bình phương phần dư:

$$\boldsymbol{\varepsilon}^T \boldsymbol{\varepsilon} = \|Y - \hat{Y}\|^2$$

Ví dụ 2. Chúng ta quay lại nghiên cứu về mối liên hệ giữa độ tuổi, bmi và cholesterol. Trong ví dụ, chúng ta chỉ mới xét mối liên hệ giữa độ tuổi và cholesterol, mà chưa xem đến mối liên hệ giữa cả hai yếu tố độ tuổi và bmi và cholesterol. Biểu đồ sau đây cho chúng ta thấy mối liên hệ giữa ba biến số này:

```
> pairs(data)
```



Biểu đồ 10.5. Giá trị tiên đoán và khoảng tin cậy 95%.

Cũng như giữa độ tuổi và cholesterol, mối liên hệ giữa bmi và cholesterol cũng gần tuân theo một đường thẳng. Biểu đồ trên còn cho chúng ta thấy độ tuổi và bmi có liên hệ với nhau. Thật vậy, phân tích hồi qui tuyến tính đơn giản giữa bmi và cholesterol cho thấy như mối liên hệ này có ý nghĩa thống kê:

```
> summary(lm(chol ~ bmi))
```

Call:

```
lm(formula = chol ~ bmi)
```

Residuals:

Min	1Q	Median	3Q	Max
-0.9403	-0.3565	-0.1376	0.3040	1.4330

Coefficients:

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

```

(Intercept) -2.83187      1.60841   -1.761   0.09739 .
bmi          0.26410      0.06861    3.849   0.00142 **

---
Signif. codes:  0 '****' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.623 on 16 degrees of freedom
Multiple R-Squared: 0.4808,     Adjusted R-squared: 0.4483
F-statistic: 14.82 on 1 and 16 DF,  p-value: 0.001418

```

BMI giải thích khoảng 48% độ dao động về cholesterol giữa các cá nhân. Nhưng vì BMI cũng có liên hệ với độ tuổi, chúng ta muốn biết nếu hai yếu tố này được phân tích cùng một lúc thì yếu tố nào quan trọng hơn. Để biết ảnh hưởng của cả hai yếu tố age (x_1) và bmi (tạm gọi là x_2) đến cholesterol (y) qua một mô hình hồi qui tuyến tính đa biến, và mô hình đó là:

$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$$

hay phương trình cũng có thể mô tả bằng kí hiệu ma trận: $\mathbf{Y} = \mathbf{X}\beta + \boldsymbol{\varepsilon}$ mà tôi vừa trình bày trên. Ở đây, \mathbf{Y} là một vector vector 18×1 , \mathbf{X} là một matrix 18×2 phần tử, β và một vector 2×1 , và $\boldsymbol{\varepsilon}$ là vector gồm 18×1 phần tử. Để ước tính hai hệ số hồi qui, β_1 và β_2 chúng ta cũng ứng dụng hàm `lm()` trong R như sau:

```

> mreg <- lm(chol ~ age + bmi)
> summary(mreg)

Call:
lm(formula = chol ~ age + bmi)

Residuals:
    Min      1Q  Median      3Q      Max 
-0.3762 -0.2259 -0.0534  0.1698  0.5679 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 0.455458  0.918230  0.496   0.627    
age         0.054052  0.007591  7.120 3.50e-06 ***  
bmi         0.033364  0.046866  0.712   0.487    
---
Signif. codes:  0 '****' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3074 on 15 degrees of freedom
Multiple R-Squared: 0.8815,     Adjusted R-squared: 0.8657 
F-statistic: 55.77 on 2 and 15 DF,  p-value: 1.132e-07

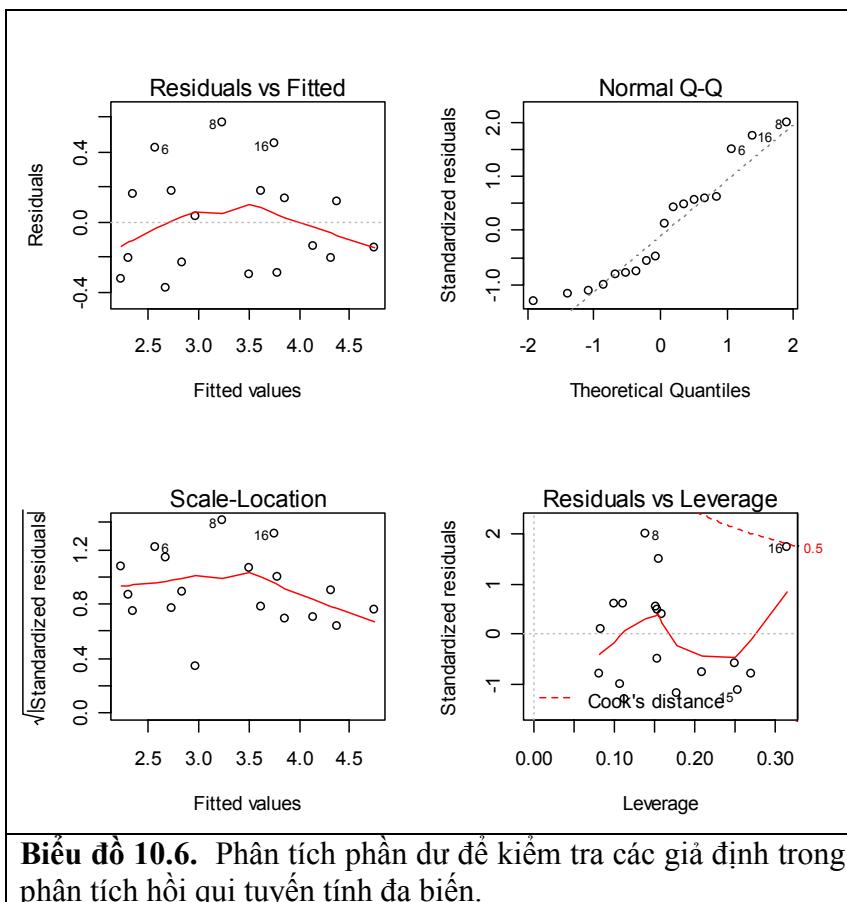
```

Kết quả phân tích trên cho thấy ước số $\hat{\alpha} = 0.455$, $\hat{\beta}_1 = 0.054$ và $\hat{\beta}_2 = 0.0333$. Nói cách khác, chúng ta có phương trình ước đoán độ cholesterol dựa vào hai biến số độ tuổi và bmi như sau:

$$\text{Cholesterol} = 0.455 + 0.054(\text{age}) + 0.0333(\text{bmi})$$

Phương trình cho biết khi độ tuổi tăng 1 năm thì cholesterol tăng 0.054 mg/L (ước số này không khác mấy so với 0.0578 trong phương trình chỉ có độ tuổi), và mỗi $1 \text{ kg}/\text{m}^2$ tăng BMI thì cholesterol tăng 0.0333 mg/L. Hai yếu tố này “giải thích” khoảng 88.2% ($R^2 = 0.8815$) độ dao động của cholesterol giữa các cá nhân.

Chúng ta chú ý phương trình với độ tuổi (trong phân tích phần trước) giải thích khoảng 87.7% độ dao động cholesterol giữa các cá nhân. Khi chúng ta thêm yếu tố BMI, hệ số này tăng lên 88.2%, tức chỉ 0.5%. Câu hỏi đặt ra là 0.5% tăng trưởng này có ý nghĩa thống kê hay không. Câu trả lời có thể xem qua kết quả kiểm định yếu tố bmi với trị số $p = 0.487$. Như vậy, bmi không cung cấp cho chúng thêm thông tin hay tiên đoán cholesterol hơn những gì chúng ta đã có từ độ tuổi. Nói cách khác, khi độ tuổi đã được xem xét, thì ảnh hưởng của bmi không còn ý nghĩa thống kê. Điều này có thể hiểu được, bởi vì qua Biểu đồ 10.5 chúng ta thấy độ tuổi và bmi có một mối liên hệ khá cao. Vì hai biến này có tương quan với nhau, chúng ta không cần cả hai trong phương trình. (Tuy nhiên, ví dụ này chỉ có tính cách minh họa cho việc tiến hành phân tích hồi qui tuyến tính đa biến bằng R, chứ không có ý định mô phỏng dữ liệu theo định hướng sinh học).



Tuy BMI không có ý nghĩa thống kê trong trường hợp này, **Biểu đồ 10.6** cho thấy các giả định về mô hình hồi qui tuyến tính có thể đáp ứng.

10.4 Phân tích hồi qui đa thức (Polynomial regression analysis)

Một khai triển tất nhiên từ phân tích hồi qui đa biến độc lập là phân tích hồi qui đa thức. Mô hình hồi qui đa biến mô tả một biến phụ thuộc như là một *hàm số tuyến tính* (linear function) của nhiều biến độc lập, trong khi đó mô hình hồi qui đa thức mô tả một biến phụ thuộc là *hàm số phi tuyến tính* (non-linear function) của một biến độc lập.

Nói theo ngôn ngữ toán học, mô hình hồi qui đa thức tìm mối liên hệ giữa biến phụ thuộc y và biến độc lập x theo những hàm số sau đây:

$$y_i = \alpha + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \dots + \beta_p x^p + \varepsilon_i$$

Trong đó các thông số β_j ($j = 1, 2, 3, \dots, p$) là hệ số đo lường mối liên hệ giữa y và x ; và ε_i là phần dư của mô hình, với giả định ε_i tuân theo luật phân phối chuẩn với trung bình 0 và phương sai σ^2 . Cho một dãy cặp số $(y_1, x_1), (y_2, x_2), (y_3, x_3), \dots, (y_n, x_n)$, chúng ta có thể áp dụng phương pháp bình phương nhỏ nhất để ước tính β_j và σ^2 .

Trong mô hình trên, chúng ta có thể dễ dàng thấy rằng mô hình hồi qui đa thức còn là một phát triển trực tiếp từ mô hình hồi qui tuyến tính đơn giản. Tức là nếu $\beta_2 = 0, \beta_3 = 0, \dots, \beta_p = 0$, thì mô hình trên đơn giản thành mô hình hồi qui tuyến tính một biến mà chúng ta gặp trong phần đầu của chương này. Nếu $y_i = \alpha + \beta_1 x + \beta_2 x^2 + \varepsilon_i$ thì mô hình đơn giản là một phương trình bậc hai, v.v.

Ví dụ 3. Thí nghiệm sau đây tìm mối liên hệ giữa hàm lượng gỗ cứng (hardwood concentration) và độ căng (tensile strength) của vật liệu. Mười chín vật liệu khác nhau với nhiều hàm lượng gỗ cứng được thử nghiệm để đo độ căng mạnh của vật liệu, và kết quả được tóm lược trong bảng số liệu sau đây:

Id	Hàm lượng gỗ cứng (x)	Độ căng mạnh (y)
1	1.0	6.3
2	1.5	11.1
3	2.0	20.0
4	3.0	24.0
5	4.0	26.1
6	4.5	30.0
7	5.0	33.8
8	5.5	34.0
9	6.0	38.1
10	6.5	39.9
11	7.0	42.0
12	8.0	46.1

13	9.0	53.1
14	10.0	52.0
15	11.0	52.5
16	12.0	48.0
17	13.0	42.8
18	14.0	27.8
19	15.0	21.9

Trước khi phân tích các số liệu này, chúng ta cần nhập số liệu vào R với những lệnh thông thường như sau:

```
> id <- 1:19
> conc <- c(1.0, 1.5, 2.0, 3.0, 4.0,
4.5, 5.0, 5.5, 6.0,
6.5, 7.0, 8.0, 9.0, 10.0, 11.0, 12.0, 13.0, 14.0, 15.0)
> strength <- c(6.3, 11.1, 20.0, 24.0, 26.1, 30.0, 33.8, 34.0, 38.1,
39.9, 42.0, 46.1, 53.1, 52.0, 52.5, 48.0, 42.8, 27.8, 21.9)
> data <- data.frame(id, conc, strength)
```

Chúng ta thử xem mô hình hồi qui tuyến tính đơn giản bằng lệnh:

```
> simple.model <- lm(strength ~ conc)
> summary(simple.model)

Call:
lm(formula = strength ~ conc)

Residuals:
    Min      1Q  Median      3Q     Max 
-25.986 -3.749  2.938  7.675 15.840 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 21.3213   5.4302   3.926  0.00109 **  
conc        1.7710   0.6478   2.734  0.01414 *   
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

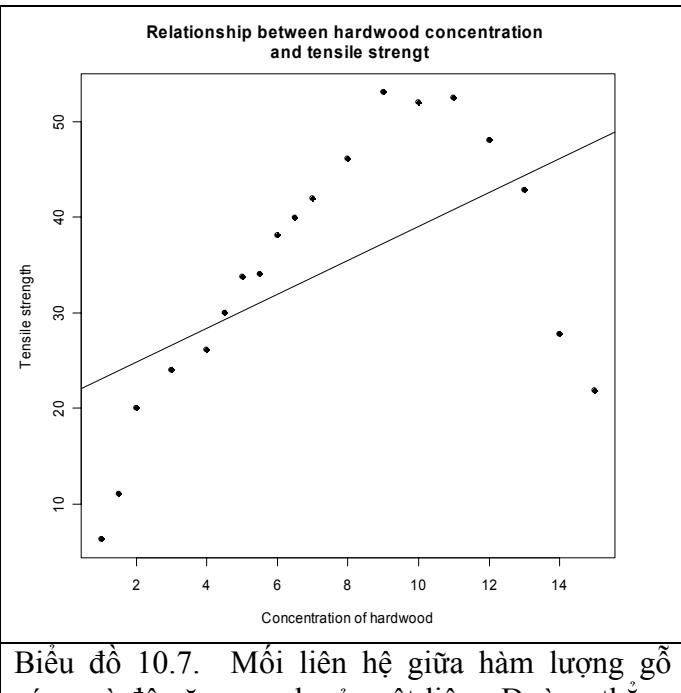
Residual standard error: 11.82 on 17 degrees of freedom
Multiple R-Squared:  0.3054,    Adjusted R-squared:  0.2645 
F-statistic: 7.474 on 1 and 17 DF,  p-value: 0.01414
```

Kết quả trên cho thấy mô hình hồi qui tuyến tính đơn giản này ($\text{strength} = 21.32 + 1.77 \times \text{conc}$) giải thích khoảng 31% phương sai của `strength`. Ước số phương sai của mô hình này là: $s^2 = (11.82)^2 = 139.7$.

Bây giờ chúng ta xem qua biểu đồ và đường biểu diễn của mô hình trên:

```
> plot(strength ~ conc,
       xlab="Concentration of hardwood",
       ylab="Tensile strength",
       main="Relationship between hardwood concentration \n and tensile
strength", pch=16)

> abline(simple.model)
```



Biểu đồ 10.7. Mối liên hệ giữa hàm lượng gỗ cứng và độ căng mạnh của vật liệu. Đường thẳng là đường biểu diễn của mô hình hồi qui tuyến tính đơn giản.

```
lm(formula = strength ~ poly(conc, 2))
```

Residuals:

Min	1Q	Median	3Q	Max
-5.8503	-3.2482	-0.7267	4.1350	6.5506

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	34.184	1.014	33.709	2.73e-16 ***
poly(conc, 2)1	32.302	4.420	7.308	1.76e-06 ***
poly(conc, 2)2	-45.396	4.420	-10.270	1.89e-08 ***

Signif. codes:	0 '***'	0.001 '**'	0.01 '*'	0.05 '.'
	0.1	'	'	1

Residual standard error: 4.42 on 16 degrees of freedom
Multiple R-Squared: 0.9085, Adjusted R-squared: 0.8971
F-statistic: 79.43 on 2 and 16 DF, p-value: 4.912e-09

Như vậy, mô hình mới này $y = 34.18 + 32.30*x - 45.4*x^2$ giải thích khoảng 91% phương sai của y. Phương sai của y bây giờ là $s^2 = (4.42)^2 = 19.5$. So với mô hình tuyến tính, mô hình này rõ ràng là tốt hơn rất nhiều.

Chúng ta thử xét một mô hình cubic (bậc ba) $y_i = \alpha + \beta_1x + \beta_2x^2 + \beta_3x^3$ xem có mô tả y tốt hơn mô hình phương trình bậc hai hay không.

```
> cubic <- lm(strength ~ poly(conc, 3))
> summary(cubic)
```

Qua biểu đồ này, chúng ta thấy rõ ràng mô hình hồi qui tuyến tính không thích hợp cho số liệu, bởi vì mối liên hệ giữa hai biến này không tuân theo một phương trình đường thẳng, mà là một đường cong. Nói cách khác, một mô hình phương trình bậc hai có lẽ thích hợp hơn. Gọi y là strength và x là conc, chúng ta có thể viết mô hình đó như sau:

$$y_i = \alpha + \beta_1x + \beta_2x^2$$

Bây giờ chúng ta sẽ sử dụng R để ước tính ba thông số trên.

```
> quadratic <- lm(strength ~ poly(conc, 2))
> summary(quadratic)
```

Call:

```

Call:
lm(formula = strength ~ poly(conc, 3))

Residuals:
    Min      1Q  Median      3Q     Max 
-4.62503 -1.61085  0.04125  1.58922  5.02159 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 34.1842    0.5931  57.641 < 2e-16 ***
poly(conc, 3)1 32.3021    2.5850 12.496 2.48e-09 ***
poly(conc, 3)2 -45.3963    2.5850 -17.561 2.06e-11 ***
poly(conc, 3)3 -14.5740    2.5850 -5.638 4.72e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.585 on 15 degrees of freedom
Multiple R-Squared:  0.9707,    Adjusted R-squared:  0.9648 
F-statistic: 165.4 on 3 and 15 DF,  p-value: 1.025e-11

```

Mô hình cubic này thậm chí có khả năng mô tả y tốt hơn hai mô hình trước, với hệ số xác định bội (R^2) bằng 0.97, và tất cả các thông số trong mô hình đều có ý nghĩa thống kê. Biểu đồ sau đây so sánh 3 mô hình trên:

```

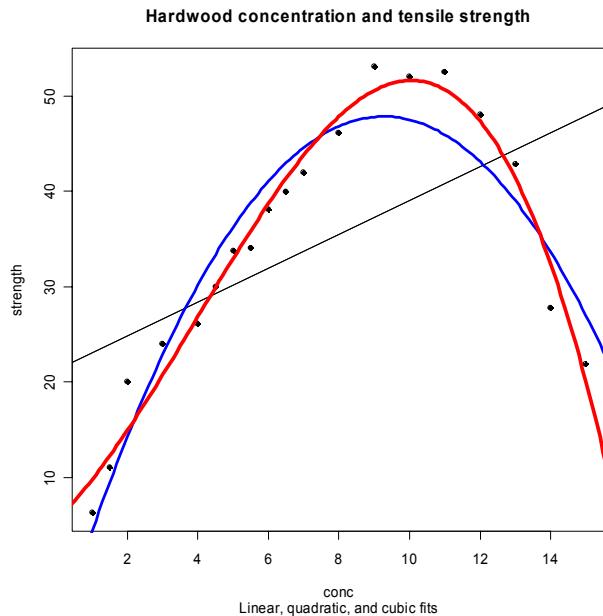
# Lặp lại các mô hình trên:
> linear <- lm(strength ~ conc)
> quadratic <- lm(strength ~ poly(conc, 2))
> cubic <- lm(strength ~ poly(conc, 3))

# Tạo nên một biến x với nhiều số gần nhau
> xnew <- (0:160)/10

# Tính giá trị tiên đoán (predictive values) của y
> y2 = predict(quadratic, data.frame(conc=xnew))
> y3 = predict(cubic, data.frame(conc=xnew))

# Vẽ 3 đường thẳng, bậc hai và bậc 3
> plot(strength ~ conc, pch=16,
       main="Hardwood concentration and tensile strength",
       sub="Linear, quadratic, and cubic fits")
> abline(linear, col="black")
> lines(xnew, y2, col="blue", lwd=3)
> lines(xnew, y3, col="red", lwd=4)

```



10.5 Xây dựng mô hình tuyến tính từ nhiều biến

Trong một nghiên cứu thông thường với một biến số phụ thuộc, nhiều biến số độc lập $x_1, x_2, x_3, \dots, x_k$, mà k có thể lên đến hàng chục, thậm chí hàng trăm. Các biến độc lập đó thường liên hệ với nhau. Có rất nhiều tổ hợp biến độc lập có khả năng tiên đoán biến phụ thuộc y . Ví dụ nếu chúng ta có 3 biến độc lập x_1, x_2 , và x_3 , để xây dựng mô hình tiên đoán y , chúng ta có thể phải xem xét các mô hình sau đây: $y = f_1(x_1)$, $y = f_2(x_2)$, $y = f_3(x_3)$, $y = f_4(x_1, x_2)$, $y = f_5(x_1, x_3)$, $y = f_6(x_2, x_3)$, $y = f_7(x_1, x_2, x_3)$, v.v... trong đó f_k là những hàm số được định nghĩa bởi hệ số liên quan đến các biến cụ thể. Khi k cao, số lượng mô hình cũng lên rất cao.

Vấn đề đặt ra là trong các mô hình đó, mô hình nào có thể tiên đoán y một cách đầy đủ, đơn giản và hợp lý. Tôi sẽ quay lại ba tiêu chuẩn này trong chương phân tích hồi qui logistic. Ở đây, tôi chỉ muốn bàn đến một tiêu chuẩn thống kê để xây dựng mô hình hồi qui tuyến tính. Trong trường hợp có nhiều mô hình như thế, tiêu chuẩn thống kê để chọn một mô hình tối ưu thường dựa vào tiêu chuẩn thông tin Akaike (còn gọi là AIC hay Akaike Information Criterion).

Cho một mô hình hồi qui tuyến tính $\hat{y}_i = \hat{\alpha} + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \dots + \hat{\beta}_k x_k$, chúng ta có $k+1$ thông số $\hat{\alpha}, \hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_k$, và có thể tính tổng bình phương phần dư (residual sum of squares, RSS):

$$RSS = \sum_{i=1}^n (\hat{y}_i - y_i)^2$$

Trong đó, n là số lượng mẫu. Công thức trên cho thấy nếu mô hình mô tả y đầy đủ thì RSS sẽ thấp, vì độ khác biệt giữa giá trị tiên đoán \hat{y} và giá trị quan sát y gần nhau. Một qui luật chung của phân tích hồi qui tuyến tính là một mô hình với k biến độc lập sẽ có RSS thấp hơn mô hình với $k-1$ biến; và tương tự mô hình với $k-1$ biến sẽ có RSS thấp hơn mô hình với $k-2$ biến, v.v... Nói cách khác, mô hình càng có nhiều biến độc lập sẽ “giải thích” y càng tốt hơn. Nhưng vì một số biến độc lập x liên hệ với nhau, cho nên có thêm nhiều biến không có nghĩa là RSS sẽ giảm một cách có ý nghĩa. Một phép tính để dung hòa RSS và số biến độc lập trong một mô hình là AIC, được định nghĩa như sau:

$$AIC = \log\left(\frac{RSS}{n}\right) + \frac{2k}{n}$$

Mô hình nào có giá trị AIC thấp nhất được xem là mô hình “tối ưu”. Trong ví dụ sau đây, chúng ta sẽ dùng hàm `step` để tìm một mô hình tối ưu dựa vào giá trị AIC.

Ví dụ 4. Để nghiên cứu ảnh hưởng của các yếu tố như nhiệt độ, thời gian, và thành phần hóa học đến sản lượng CO₂. Số liệu của nghiên cứu này có thể tóm lược trong bảng số 2. Mục tiêu chính của nghiên cứu là tìm một mô hình hồi qui tuyến tính để tiên đoán sản lượng CO₂, cũng như đánh giá độ ảnh hưởng của các yếu tố này.

Bảng 2. Sản lượng CO₂ và một số yếu tố có thể ảnh hưởng đến CO₂

Id	y	X1	X2	X3	X4	X5	X6	X7
1	36.98	5.1	400	51.37	4.24	1484.83	2227.25	2.06
2	13.74	26.4	400	72.33	30.87	289.94	434.90	1.33
3	10.08	23.8	400	71.44	33.01	320.79	481.19	0.97
4	8.53	46.4	400	79.15	44.61	164.76	247.14	0.62
5	36.42	7.0	450	80.47	33.84	1097.26	1645.89	0.22
6	26.59	12.6	450	89.90	41.26	605.06	907.59	0.76
7	19.07	18.9	450	91.48	41.88	405.37	608.05	1.71
8	5.96	30.2	450	98.60	70.79	253.70	380.55	3.93
9	15.52	53.8	450	98.05	66.82	142.27	213.40	1.97
10	56.61	5.6	400	55.69	8.92	1362.24	2043.36	5.08
11	26.72	15.1	400	66.29	17.98	507.65	761.48	0.60
12	20.80	20.3	400	58.94	17.79	377.60	566.40	0.90
13	6.99	48.4	400	74.74	33.94	158.05	237.08	0.63
14	45.93	5.8	425	63.71	11.95	130.66	1961.49	2.04
15	43.09	11.2	425	67.14	14.73	682.59	1023.89	1.57
16	15.79	27.9	425	77.65	34.49	274.20	411.30	2.38
17	21.60	5.1	450	67.22	14.48	1496.51	2244.77	0.32
18	35.19	11.7	450	81.48	29.69	652.43	978.64	0.44
19	26.14	16.7	450	83.88	26.33	458.42	687.62	8.82
20	8.60	24.8	450	89.38	37.98	312.25	468.38	0.02
21	11.63	24.9	450	79.77	25.66	307.08	460.62	1.72
22	9.59	39.5	450	87.93	22.36	193.61	290.42	1.88
23	4.42	29.0	450	79.50	31.52	155.96	233.95	1.43
24	38.89	5.5	460	72.73	17.86	1392.08	2088.12	1.35
25	11.19	11.5	450	77.88	25.20	663.09	994.63	1.61
26	75.62	5.2	470	75.50	8.66	1464.11	2196.17	4.78

27	36.03	10.6	470	83.15	22.39	720.07	1080.11	5.88
----	-------	------	-----	-------	-------	--------	---------	------

Chú thích: y = sản lượng CO_2 ; $X1$ = thời gian (phút); $X2$ = nhiệt độ (C); $X3$ = phần trăm hòa tan; $X4$ = lượng dầu (g/100g); $X5$ = lượng than đá; $X6$ = tổng số lượng hòa tan; $X7$ = số hydrogen tiêu thụ.

Trước khi phân tích số liệu, chúng ta cần nhập số liệu vào R bằng các lệnh thông thường. Số liệu sẽ chứa trong đối tượng REGdata.

```
> y <- c(36.98,13.74,10.08, 8.53,36.42,26.59,19.07, 5.96,15.52,56.61,
       26.72,20.80, 6.99,45.93,43.09,15.79,21.60,35.19,26.14, 8.60,
       11.63, 9.59, 4.42,38.89,11.19,75.62,36.03)
> x1 <- c(5.1,26.4,23.8,46.4, 7.0,12.6,18.9,30.2,53.8,5.6,15.1,20.3,48.4,
       5.8,11.2,27.9,5.1,11.7,16.7,24.8,24.9,39.5,29.0, 5.5, 11.5,
       5.2,10.6)
> x2 <- c(400,400, 400, 400, 450, 450, 450, 450, 450, 400, 400, 400,
       400, 425, 425, 425, 450, 450, 450, 450, 450, 450, 450, 460,
       450, 470, 470)
> x3 <- c(51.37,72.33,71.44,79.15,80.47,89.90,91.48,98.60,98.05,55.69,
       66.29,58.94,74.74,63.71,67.14,77.65,67.22,81.48,83.88,89.38,
       79.77,87.93,79.50,72.73,77.88,75.50,83.15)
> x4 <- c(4.24,30.87,33.01,44.61,33.84,41.26,41.88,70.79,66.82,
       8.92,17.98,17.79,33.94,11.95,14.73,34.49,14.48,29.69,26.33,
       37.98,25.66,22.36,31.52,17.86,25.20, 8.66,22.39)
> x5 <- c(1484.83, 289.94, 320.79, 164.76, 1097.26, 605.06, 405.37,
       253.70, 142.27,1362.24, 507.65, 377.60, 158.05, 130.66,
       682.59, 274.20, 1496.51, 652.43, 458.42, 312.25, 307.08,
       193.61, 155.96,1392.08, 663.09,1464.11, 720.07)
> x6 <- c(2227.25, 434.90, 481.19, 247.14,1645.89, 907.59, 608.05,
       380.55, 213.40,2043.36, 761.48, 566.40, 237.08,1961.49,1023.89,
       411.30,2244.77, 978.64, 687.62, 468.38, 460.62, 290.42,
       233.95,2088.12, 994.63,2196.17,1080.11)
> x7 <- c(2.06,1.33,0.97,0.62,0.22,0.76,1.71,3.93,1.97,5.08,0.60,0.90,
       0.63,2.04,1.57,2.38,0.32,0.44,8.82,0.02,1.72,1.88,1.43,
       1.35,1.61,4.78,5.88)

> REGdata <- data.frame(y, x1,x2,x3,x4,x5,x6,x7)
```

Trước khi phân tích số liệu, chúng ta cần nhập số liệu vào R bằng các lệnh thông thường. Số liệu sẽ chứa trong đối tượng REGdata.

Bây giờ chúng ta bắt đầu phân tích. Mô hình đầu tiên là mô hình gồm tất cả 7 biến độc lập như sau:

```
> reg <- lm(y ~ x1+x2+x3+x4+x5+x6+x7, data=REGdata)
> summary(reg)
```

Call:

```
lm(formula = y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7, data = REGdata)
```

Residuals:

Min	1Q	Median	3Q	Max
-20.035	-4.681	-1.144	4.072	21.214

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	53.937016	57.428952	0.939	0.3594

```

x1      -0.127653   0.281498  -0.453   0.6553
x2      -0.229179   0.232643  -0.985   0.3370
x3       0.824853   0.765271   1.078   0.2946
x4      -0.438222   0.358551  -1.222   0.2366
x5      -0.001937   0.009654  -0.201   0.8431
x6       0.019886   0.008088   2.459   0.0237 *
x7      1.993486   1.089701   1.829   0.0831 .
---
Signif. codes:  0 '****' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Residual standard error: 10.61 on 19 degrees of freedom
 Multiple R-Squared: 0.728, Adjusted R-squared: 0.6278
 F-statistic: 7.264 on 7 and 19 DF, p-value: 0.0002674

Kết quả trên cho thấy tất cả 7 biến số “giải thích” khoảng 73% phương sai của y . Nhưng trong 7 biến đó, chỉ có x_6 là có ý nghĩa thống kê ($p = 0.024$). Chúng ta thử giảm mô hình thành một mô hình hồi qui tuyến tính đơn giản với chỉ biến x_6 .

```

> summary(lm(y ~ x6, data=REGdata))

Call:
lm(formula = y ~ x6, data = REGdata)

Residuals:
    Min      1Q      Median      3Q      Max 
-28.081  -5.829  -0.839   5.522  26.882 

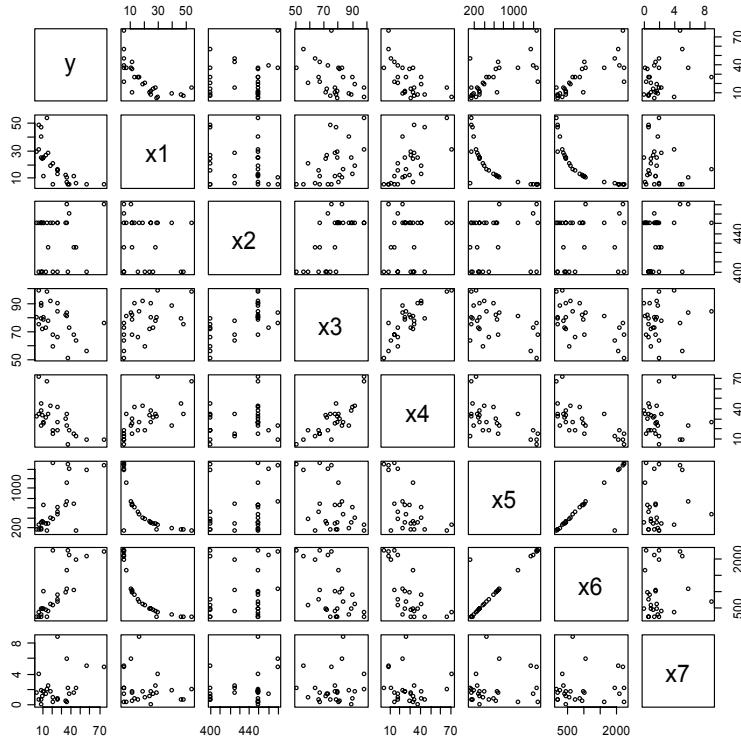
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 6.144181  3.483064  1.764   0.09 .  
x6          0.019395  0.002932  6.616 6.24e-07 *** 
---
Signif. codes:  0 '****' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Residual standard error: 10.7 on 25 degrees of freedom
 Multiple R-Squared: 0.6365, Adjusted R-squared: 0.6219
 F-statistic: 43.77 on 1 and 25 DF, p-value: 6.238e-07

Chỉ với một biến x_6 mà mô hình có thể giải thích khoảng 64% phương sai của y . Chúng ta chấp nhận mô hình này? Trước khi chấp nhận mô hình này, chúng ta phải xem xét độ tương quan giữa các biến độc lập:

```
> pairs(REGdata)
```



Kết quả trên cho thấy y có liên hệ với các biến như x_1 , x_5 và x_6 . Ngoài ra, biến x_5 và x_6 có một mối liên hệ rất mật thiết (gần như là một đường thẳng) với hệ số tương quan là 0.88. Ngoài ra, x_5 và x_1 hay x_6 và x_5 cũng có liên hệ với nhau nhưng theo một hàm số nghịch đảo. Điều này có nghĩa là biến x_5 và x_6 cung cấp một lượng thông tin như nhau để tiên đoán y , tức là chúng ta không cần cả hai trong mô hình.

Để tìm một mô hình tối ưu trong bối cảnh có nhiều mối tương quan như thế, chúng ta ứng dụng step như sau. Chú ý cách cung cấp thông số lm($y \sim .$), dấu “.” có nghĩa là yêu cầu R xem xét tất cả biến trong đối tượng REGdata.

```
> reg <- lm(y ~ ., data=REGdata)
> step(reg, direction="both")
```

<p>Start: AIC= 134.07</p> <table border="1"> <thead> <tr> <th></th> <th>Df</th> <th>Sum of Sq</th> <th>RSS</th> <th>AIC</th> </tr> </thead> <tbody> <tr> <td>$y \sim x_1 + x_2 + x_3 + x_4 + x_5 + x_6 + x_7$</td> <td></td> <td></td> <td></td> <td></td> </tr> <tr> <td>- x_5</td> <td>1</td> <td>4.54</td> <td>2145.37</td> <td>132.13</td> </tr> <tr> <td>- x_1</td> <td>1</td> <td>23.17</td> <td>2164.00</td> <td>132.36</td> </tr> <tr> <td>- x_2</td> <td>1</td> <td>109.34</td> <td>2250.18</td> <td>133.42</td> </tr> <tr> <td>- x_3</td> <td>1</td> <td>130.90</td> <td>2271.74</td> <td>133.68</td> </tr> <tr> <td><none></td> <td></td> <td>2140.83</td> <td>134.07</td> <td></td> </tr> <tr> <td>- x_4</td> <td>1</td> <td>168.31</td> <td>2309.14</td> <td>134.12</td> </tr> <tr> <td>- x_7</td> <td>1</td> <td>377.09</td> <td>2517.92</td> <td>136.45</td> </tr> <tr> <td>- x_6</td> <td>1</td> <td>681.09</td> <td>2821.92</td> <td>139.53</td> </tr> </tbody> </table>		Df	Sum of Sq	RSS	AIC	$y \sim x_1 + x_2 + x_3 + x_4 + x_5 + x_6 + x_7$					- x_5	1	4.54	2145.37	132.13	- x_1	1	23.17	2164.00	132.36	- x_2	1	109.34	2250.18	133.42	- x_3	1	130.90	2271.74	133.68	<none>		2140.83	134.07		- x_4	1	168.31	2309.14	134.12	- x_7	1	377.09	2517.92	136.45	- x_6	1	681.09	2821.92	139.53	<p>Step 1: AIC= 132.13</p> <table border="1"> <thead> <tr> <th></th> <th>Df</th> <th>Sum of Sq</th> <th>RSS</th> <th>AIC</th> </tr> </thead> <tbody> <tr> <td>$y \sim x_1 + x_2 + x_3 + x_4 + x_6 + x_7$</td> <td></td> <td></td> <td></td> <td></td> </tr> <tr> <td>- x_1</td> <td>1</td> <td>22.7</td> <td>2168.1</td> <td>130.4</td> </tr> <tr> <td>- x_2</td> <td>1</td> <td>113.8</td> <td>2259.1</td> <td>131.5</td> </tr> <tr> <td>- x_3</td> <td>1</td> <td>133.5</td> <td>2278.9</td> <td>131.8</td> </tr> <tr> <td><none></td> <td></td> <td>2145.4</td> <td>132.1</td> <td></td> </tr> <tr> <td>- x_4</td> <td>1</td> <td>170.8</td> <td>2316.2</td> <td>132.2</td> </tr> <tr> <td>+ x_5</td> <td>1</td> <td>4.5</td> <td>2140.8</td> <td>134.1</td> </tr> <tr> <td>- x_7</td> <td>1</td> <td>375.7</td> <td>2521.1</td> <td>134.5</td> </tr> <tr> <td>- x_6</td> <td>1</td> <td>1058.5</td> <td>3203.8</td> <td>141.0</td> </tr> </tbody> </table>		Df	Sum of Sq	RSS	AIC	$y \sim x_1 + x_2 + x_3 + x_4 + x_6 + x_7$					- x_1	1	22.7	2168.1	130.4	- x_2	1	113.8	2259.1	131.5	- x_3	1	133.5	2278.9	131.8	<none>		2145.4	132.1		- x_4	1	170.8	2316.2	132.2	+ x_5	1	4.5	2140.8	134.1	- x_7	1	375.7	2521.1	134.5	- x_6	1	1058.5	3203.8	141.0
	Df	Sum of Sq	RSS	AIC																																																																																																	
$y \sim x_1 + x_2 + x_3 + x_4 + x_5 + x_6 + x_7$																																																																																																					
- x_5	1	4.54	2145.37	132.13																																																																																																	
- x_1	1	23.17	2164.00	132.36																																																																																																	
- x_2	1	109.34	2250.18	133.42																																																																																																	
- x_3	1	130.90	2271.74	133.68																																																																																																	
<none>		2140.83	134.07																																																																																																		
- x_4	1	168.31	2309.14	134.12																																																																																																	
- x_7	1	377.09	2517.92	136.45																																																																																																	
- x_6	1	681.09	2821.92	139.53																																																																																																	
	Df	Sum of Sq	RSS	AIC																																																																																																	
$y \sim x_1 + x_2 + x_3 + x_4 + x_6 + x_7$																																																																																																					
- x_1	1	22.7	2168.1	130.4																																																																																																	
- x_2	1	113.8	2259.1	131.5																																																																																																	
- x_3	1	133.5	2278.9	131.8																																																																																																	
<none>		2145.4	132.1																																																																																																		
- x_4	1	170.8	2316.2	132.2																																																																																																	
+ x_5	1	4.5	2140.8	134.1																																																																																																	
- x_7	1	375.7	2521.1	134.5																																																																																																	
- x_6	1	1058.5	3203.8	141.0																																																																																																	
<p>Step 2: AIC= 130.42</p> <table border="1"> <thead> <tr> <th></th> <th>Df</th> <th>Sum of Sq</th> <th>RSS</th> <th>AIC</th> </tr> </thead> <tbody> <tr> <td>$y \sim x_2 + x_3 + x_4 + x_6 + x_7$</td> <td></td> <td></td> <td></td> <td></td> </tr> </tbody> </table>		Df	Sum of Sq	RSS	AIC	$y \sim x_2 + x_3 + x_4 + x_6 + x_7$					<p>Step 3: AIC= 129.59</p> <table border="1"> <thead> <tr> <th></th> <th>Df</th> <th>Sum of Sq</th> <th>RSS</th> <th>AIC</th> </tr> </thead> <tbody> <tr> <td>$y \sim x_3 + x_4 + x_6 + x_7$</td> <td></td> <td></td> <td></td> <td></td> </tr> </tbody> </table>		Df	Sum of Sq	RSS	AIC	$y \sim x_3 + x_4 + x_6 + x_7$																																																																																				
	Df	Sum of Sq	RSS	AIC																																																																																																	
$y \sim x_2 + x_3 + x_4 + x_6 + x_7$																																																																																																					
	Df	Sum of Sq	RSS	AIC																																																																																																	
$y \sim x_3 + x_4 + x_6 + x_7$																																																																																																					

<table border="1"> <thead> <tr> <th></th><th>Df</th><th>Sum of Sq</th><th>RSS</th><th>AIC</th></tr> </thead> <tbody> <tr><td>- x2</td><td>1</td><td>96.8</td><td>2264.9</td><td>129.6</td></tr> <tr><td>- x3</td><td>1</td><td>122.0</td><td>2290.0</td><td>129.9</td></tr> <tr><td><none></td><td></td><td>2168.1</td><td>130.4</td><td></td></tr> <tr><td>- x4</td><td>1</td><td>187.4</td><td>2355.5</td><td>130.7</td></tr> <tr><td>+ x1</td><td>1</td><td>22.7</td><td>2145.4</td><td>132.1</td></tr> <tr><td>+ x5</td><td>1</td><td>4.1</td><td>2164.0</td><td>132.4</td></tr> <tr><td>- x7</td><td>1</td><td>385.0</td><td>2553.1</td><td>132.8</td></tr> <tr><td>- x6</td><td>1</td><td>1526.2</td><td>3694.3</td><td>142.8</td></tr> </tbody> </table> <p>Step 4: AIC= 127.9 $y \sim x_4 + x_6 + x_7$</p> <table border="1"> <thead> <tr> <th></th><th>Df</th><th>Sum of Sq</th><th>RSS</th><th>AIC</th></tr> </thead> <tbody> <tr><td>- x4</td><td>1</td><td>73.5</td><td>2363.8</td><td>126.7</td></tr> <tr><td><none></td><td></td><td>2290.3</td><td>127.9</td><td></td></tr> <tr><td>+ x3</td><td>1</td><td>25.4</td><td>2264.9</td><td>129.6</td></tr> <tr><td>+ x1</td><td>1</td><td>11.3</td><td>2279.0</td><td>129.8</td></tr> <tr><td>+ x5</td><td>1</td><td>6.3</td><td>2284.0</td><td>129.8</td></tr> <tr><td>+ x2</td><td>1</td><td>0.3</td><td>2290.0</td><td>129.9</td></tr> <tr><td>- x7</td><td>1</td><td>486.6</td><td>2776.9</td><td>131.1</td></tr> <tr><td>- x6</td><td>1</td><td>1993.8</td><td>4284.1</td><td>142.8</td></tr> </tbody> </table>		Df	Sum of Sq	RSS	AIC	- x2	1	96.8	2264.9	129.6	- x3	1	122.0	2290.0	129.9	<none>		2168.1	130.4		- x4	1	187.4	2355.5	130.7	+ x1	1	22.7	2145.4	132.1	+ x5	1	4.1	2164.0	132.4	- x7	1	385.0	2553.1	132.8	- x6	1	1526.2	3694.3	142.8		Df	Sum of Sq	RSS	AIC	- x4	1	73.5	2363.8	126.7	<none>		2290.3	127.9		+ x3	1	25.4	2264.9	129.6	+ x1	1	11.3	2279.0	129.8	+ x5	1	6.3	2284.0	129.8	+ x2	1	0.3	2290.0	129.9	- x7	1	486.6	2776.9	131.1	- x6	1	1993.8	4284.1	142.8	<table border="1"> <thead> <tr> <th></th><th>Df</th><th>Sum of Sq</th><th>RSS</th><th>AIC</th></tr> </thead> <tbody> <tr><td>- x3</td><td>1</td><td>25.4</td><td>2290.3</td><td>127.9</td></tr> <tr><td>- x4</td><td>1</td><td>90.9</td><td>2355.8</td><td>128.7</td></tr> <tr><td><none></td><td></td><td>2264.9</td><td>129.6</td><td></td></tr> <tr><td>+ x2</td><td>1</td><td>96.8</td><td>2168.1</td><td>130.4</td></tr> <tr><td>+ x5</td><td>1</td><td>8.3</td><td>2256.5</td><td>131.5</td></tr> <tr><td>+ x1</td><td>1</td><td>5.7</td><td>2259.1</td><td>131.5</td></tr> <tr><td>- x7</td><td>1</td><td>384.9</td><td>2649.7</td><td>131.8</td></tr> <tr><td>- x6</td><td>1</td><td>2015.6</td><td>4280.5</td><td>144.8</td></tr> </tbody> </table> <p>Step 5: AIC= 126.75 $y \sim x_6 + x_7$</p> <table border="1"> <thead> <tr> <th></th><th>Df</th><th>Sum of Sq</th><th>RSS</th><th>AIC</th></tr> </thead> <tbody> <tr><td><none></td><td></td><td>2363.8</td><td>126.7</td><td></td></tr> <tr><td>+ x4</td><td>1</td><td>73.5</td><td>2290.3</td><td>127.9</td></tr> <tr><td>+ x1</td><td>1</td><td>33.4</td><td>2330.4</td><td>128.4</td></tr> <tr><td>+ x3</td><td>1</td><td>8.1</td><td>2355.8</td><td>128.7</td></tr> <tr><td>+ x5</td><td>1</td><td>7.7</td><td>2356.1</td><td>128.7</td></tr> <tr><td>+ x2</td><td>1</td><td>7.3</td><td>2356.6</td><td>128.7</td></tr> <tr><td>- x7</td><td>1</td><td>497.3</td><td>2861.2</td><td>129.9</td></tr> <tr><td>- x6</td><td>1</td><td>4477.0</td><td>6840.8</td><td>153.4</td></tr> </tbody> </table>		Df	Sum of Sq	RSS	AIC	- x3	1	25.4	2290.3	127.9	- x4	1	90.9	2355.8	128.7	<none>		2264.9	129.6		+ x2	1	96.8	2168.1	130.4	+ x5	1	8.3	2256.5	131.5	+ x1	1	5.7	2259.1	131.5	- x7	1	384.9	2649.7	131.8	- x6	1	2015.6	4280.5	144.8		Df	Sum of Sq	RSS	AIC	<none>		2363.8	126.7		+ x4	1	73.5	2290.3	127.9	+ x1	1	33.4	2330.4	128.4	+ x3	1	8.1	2355.8	128.7	+ x5	1	7.7	2356.1	128.7	+ x2	1	7.3	2356.6	128.7	- x7	1	497.3	2861.2	129.9	- x6	1	4477.0	6840.8	153.4
	Df	Sum of Sq	RSS	AIC																																																																																																																																																																																	
- x2	1	96.8	2264.9	129.6																																																																																																																																																																																	
- x3	1	122.0	2290.0	129.9																																																																																																																																																																																	
<none>		2168.1	130.4																																																																																																																																																																																		
- x4	1	187.4	2355.5	130.7																																																																																																																																																																																	
+ x1	1	22.7	2145.4	132.1																																																																																																																																																																																	
+ x5	1	4.1	2164.0	132.4																																																																																																																																																																																	
- x7	1	385.0	2553.1	132.8																																																																																																																																																																																	
- x6	1	1526.2	3694.3	142.8																																																																																																																																																																																	
	Df	Sum of Sq	RSS	AIC																																																																																																																																																																																	
- x4	1	73.5	2363.8	126.7																																																																																																																																																																																	
<none>		2290.3	127.9																																																																																																																																																																																		
+ x3	1	25.4	2264.9	129.6																																																																																																																																																																																	
+ x1	1	11.3	2279.0	129.8																																																																																																																																																																																	
+ x5	1	6.3	2284.0	129.8																																																																																																																																																																																	
+ x2	1	0.3	2290.0	129.9																																																																																																																																																																																	
- x7	1	486.6	2776.9	131.1																																																																																																																																																																																	
- x6	1	1993.8	4284.1	142.8																																																																																																																																																																																	
	Df	Sum of Sq	RSS	AIC																																																																																																																																																																																	
- x3	1	25.4	2290.3	127.9																																																																																																																																																																																	
- x4	1	90.9	2355.8	128.7																																																																																																																																																																																	
<none>		2264.9	129.6																																																																																																																																																																																		
+ x2	1	96.8	2168.1	130.4																																																																																																																																																																																	
+ x5	1	8.3	2256.5	131.5																																																																																																																																																																																	
+ x1	1	5.7	2259.1	131.5																																																																																																																																																																																	
- x7	1	384.9	2649.7	131.8																																																																																																																																																																																	
- x6	1	2015.6	4280.5	144.8																																																																																																																																																																																	
	Df	Sum of Sq	RSS	AIC																																																																																																																																																																																	
<none>		2363.8	126.7																																																																																																																																																																																		
+ x4	1	73.5	2290.3	127.9																																																																																																																																																																																	
+ x1	1	33.4	2330.4	128.4																																																																																																																																																																																	
+ x3	1	8.1	2355.8	128.7																																																																																																																																																																																	
+ x5	1	7.7	2356.1	128.7																																																																																																																																																																																	
+ x2	1	7.3	2356.6	128.7																																																																																																																																																																																	
- x7	1	497.3	2861.2	129.9																																																																																																																																																																																	
- x6	1	4477.0	6840.8	153.4																																																																																																																																																																																	
<p>Call: $lm(formula = y \sim x_6 + x_7, data = REGdata)$</p> <p>Coefficients:</p> <table> <thead> <tr> <th>(Intercept)</th> <th>x6</th> <th>x7</th> </tr> </thead> <tbody> <tr> <td>2.52646</td> <td>0.01852</td> <td>2.18575</td> </tr> </tbody> </table>	(Intercept)	x6	x7	2.52646	0.01852	2.18575																																																																																																																																																																															
(Intercept)	x6	x7																																																																																																																																																																																			
2.52646	0.01852	2.18575																																																																																																																																																																																			

Quá trình tìm mô hình tối ưu dừng ở mô hình với hai biến x_6 và x_7 , vì mô hình này có giá trị AIC thấp nhất. Phương trình tuyến tính tiên đoán y là: $y = 2.526 + 0.0185(x_6) + 2.186(x_7)$.

```

> summary(lm(y ~ x6+x7, data=REGdata))

Call:
lm(formula = y ~ x6 + x7, data = REGdata)

Residuals:
    Min      1Q   Median      3Q     Max 
-23.2035 -4.3713  0.2513  4.9339 21.9682 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 2.526460   3.610055   0.700   0.4908    
x6          0.018522   0.002747   6.742 5.66e-07 ***
x7          2.185753   0.972696   2.247   0.0341 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.924 on 24 degrees of freedom
Multiple R-Squared:  0.6996,    Adjusted R-squared:  0.6746 
F-statistic: 27.95 on 2 and 24 DF,  p-value: 5.391e-07

```

Phân tích chi tiết (kết quả trên) cho thấy hai biến này giải thích khoảng 70% phương sai của y .

10.6 Xây dựng mô hình tuyến tính bằng Bayesian Model Average (BMA)

Một vấn đề trong cách xây dựng mô hình trên là mô hình với x_6 và x_7 được xem là mô hình sau cùng, trong khi đó chúng ta biết rằng một mô hình x_5 và x_7 cũng có thể là một mô hình khả dĩ, bởi vì x_5 và x_6 có mối tương quan rất gần nhau. Nếu nghiên cứu được tiến hành tiếp và với thêm số liệu mới, có lẽ một mô hình khác sẽ “ra đời”.

Để đánh giá sự bất định trong việc xây dựng mô hình thống kê, một phép tính khác có triển vọng tốt hơn cách phép tính trên là BMA (Bayesian Model Average). Bạn đọc muốn tìm hiểu thêm về phép tính này có thể tham khảo vài bài báo khoa học dưới đây. Nói một cách ngắn gọn, phép tính BMA tìm tất cả các mô hình khả dĩ (với 7 biến độc lập, số mô hình khả dĩ là $2^7 = 128$, chưa tính đến các mô hình tương tác!) và trình bày kết quả của các mô hình được xem là “tối ưu” nhất về lâu về dài. Tiêu chuẩn tối ưu cũng dựa vào giá trị AIC.

Để tiến hành phép tính BMA, chúng ta phải dùng đến package BMA (có thể tải về từ trang web của R <http://cran.R-project.org>). Sau khi đã cài đặt package BMA trong máy tính, chúng ta ra phải nhập BMA vào môi trường vận hành của R bằng lệnh:

```
> library(BMA)
```

Sau đó, tạo ra một ma trận chỉ gồm các biến độc lập. Trong data frame chúng ta biết REGdata có 8 biến, với biến số 1 là y . Do đó, lệnh `REGdata[, -1]` có nghĩa là tạo ra một data frame mới ngoại trừ cột thứ nhất (tức y).

```
> xvars <- REGdata[, -1]
```

Kế tiếp, chúng ta định nghĩa biến phụ thuộc tên `co2` từ REGdata:

```
> co2 <- REGdata[, 1]
```

Bây giờ chúng ta đã sẵn sàng phân tích bằng phép tính BMA. Hàm `bicreg` được viết đặc biệt cho phân tích hồi qui tuyến tính. Cách áp dụng hàm `bicreg` như sau:

```
> bma <- bicreg(xvars, co2, strict=FALSE, OR=20)
```

Chúng ta sử dụng hàm `summary` để biết kết quả:

```
> summary(bma)
Call:
bicreg(x = xvars, y = co2, strict = FALSE, OR = 20)
```

16 models were selected

Best 5 models (cumulative posterior probability = 0.6599):

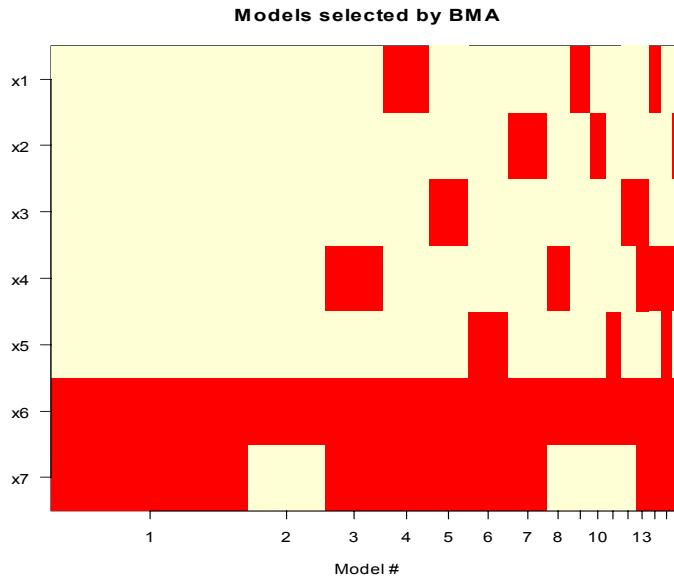
	p!=0	EV	SD	model 1	model 2	model 3	model 4	model 5
Intercept	100.0	5.75672	14.6244	2.5264	6.1441	8.6120	7.5936	7.3537
x1	12.4	-0.01807	0.1008	.	.	.	-0.1393	.
x2	10.4	-0.00075	0.0282
x3	10.7	0.00011	0.0791	-0.0572
x4	20.2	-0.03059	0.1020	.	.	-0.1419	.	.
x5	10.5	-0.00023	0.0030
x6	100.0	0.01815	0.0040	0.0185	0.0193	0.0164	0.0162	0.0179
x7	73.7	1.60766	1.2821	2.1857	.	2.1628	2.1233	2.2382
nVar				2	1	3	3	3
r2				0.700	0.636	0.709	0.704	0.701
BIC				-25.8832	-24.0238	-23.4412	-22.9721	-22.6801
post prob				0.311	0.123	0.092	0.072	0.063

BMA trình bày kết quả của 5 mô hình được đánh giá là tối ưu nhất cho tiên đoán y (model 1, model 2, ... model 5).

- Cột thứ nhất liệt kê danh sách các biến số độc lập;
- Cột 2 trình bày xác suất giả thiết một biến độc lập có ảnh hưởng đến y . Chẳng hạn như xác suất là x_6 có ảnh hưởng đến y là 100%; trong khi đó xác suất mà x_7 có ảnh hưởng đến y là 73.7%. Tuy nhiên xác suất các biến khác thấp hơn chỉ bằng 20%. Do đó, chúng ta có thể nói rằng mô hình với x_6 và x_7 có lẽ là mô hình tối ưu nhất.
- Cột 3 (EV) và 4 (SD) trình bày trị số trung bình và độ lệch chuẩn của hệ số cho mỗi biến số độc lập.
- Cột 5 là ước tính hệ số ảnh hưởng (regression coefficient) của mô hình 1. Như thấy trong cột này, mô hình 1 gồm intercept (tức α), và hai biến x_6 và x_7 . Mô hình này giải thích (như chúng ta đã biết qua phân tích phần trên) 70% phương sai của y . Trị số BIC (Bayesian Information Criterion) thấp nhất. Trong số tất cả mô hình mà BMA tìm, mô hình này có xác suất xuất hiện là 31.1%.
- Cột 6 là ước tính hệ số ảnh hưởng của mô hình 2. Như thấy trong cột này, mô hình 2 gồm intercept (tức α), và biến x_6 . Mô hình này giải thích 64% phương sai của y . Trong số tất cả mô hình mà BMA tìm, mô hình này có xác suất xuất hiện chỉ là 12.3%.
- Các mô hình khác cũng có thể diễn dịch một cách tương tự.

Một cách thể hiện kết quả trên là qua một biểu đồ như sau:

```
> imageplot.bma(bma)
```



Biểu đồ này trình bày 13 mô hình. Trong 13 mô hình đó, biến x_6 xuất hiện một cách nhất quán. Kế đến là biến x_7 cũng có xuất hiện trong một số mô hình, nhưng như chúng ta biết xác suất là 74%.

Trong ví dụ này, cả hai phép tính đều cho ra một kết quả nhất quán, nhưng trong nhiều trường hợp, hai phép tính có thể cho ra kết quả khác nhau. Nhiều nghiên cứu lý thuyết gần đây cho thấy kết quả từ phép tính BMA rất đáng tin cậy, và trong tương lai, có lẽ là phương pháp chuẩn để xây dựng mô hình.

Tài liệu tham khảo cho BMA

Raftery, Adrian E. (1995). Bayesian model selection in social research (with Discussion). *Sociological Methodology 1995* (Peter V. Marsden, ed.), pp. 111-196, Cambridge, Mass.: Blackwells.

Một số bài báo liên quan đến BMA có thể tải từ trang web sau đây:
www.stat.colostate.edu/~jah/papers.

11

Phân tích phương sai (Analysis of variance)

Phân tích phương sai, như tên gọi, là một số phương pháp phân tích thống kê mà trọng điểm là phương sai (thay vì số trung bình). Phương pháp phân tích phương sai nằm trong “đại gia đình” các phương pháp có tên là mô hình tuyến tính (hay general linear models), bao gồm cả hồi qui tuyến tính mà chúng ta đã gặp trong chương trước. Trong chương này, chúng ta sẽ làm quen với cách sử dụng R trong phân tích phương sai. Chúng ta sẽ bắt đầu bằng một phân tích đơn giản, sau đó sẽ xem đến phân tích phương sai hai chiều, và các phương pháp phi tham số thông dụng.

11.1 Phân tích phương sai đơn giản (one-way analysis of variance - ANOVA)

Ví dụ 1. Bảng thống kê 11.1 dưới đây so sánh độ galactose trong 3 nhóm bệnh nhân: nhóm 1 gồm 9 bệnh nhân với bệnh Crohn; nhóm 2 gồm 11 bệnh nhân với bệnh viêm ruột kết (colitis); và nhóm 3 gồm 20 đối tượng không có bệnh (gọi là nhóm đối chứng). Câu hỏi đặt ra là độ galactose giữa 3 nhóm bệnh nhân có khác nhau hay không? Gọi giá trị trung bình của ba nhóm là μ_1 , μ_2 , và μ_3 , và nói theo ngôn ngữ của kiểm định giả thiết thì giả thiết đảo là:

$$H_0: \mu_1 = \mu_2 = \mu_3$$

Và giả thiết chính là:

$$H_A: \text{có một khác biệt giữa } 3 \mu_j \ (j=1,2,3)$$

Bảng 11.2. Độ galactose cho 3 nhóm bệnh nhân Crohn, viêm ruột kết và đối chứng

Nhóm 1: bệnh Crohn	Nhóm 2: bệnh viêm ruột kết	Nhóm 3: đối chứng (control)
1343	1264	1809 2850
1393	1314	1926 2964
1420	1399	2283 2973
1641	1605	2384 3171
1897	2385	2447 3257
2160	2511	2479 3271
2169	2514	2495 3288
2279	2767	2525 3358
2890	2827	2541 3643
	2895	2769 3657

3011		
$n=9$	$n=11$	$n=20$
Trung bình: 1910	Trung bình: 2226	Trung bình: 2804
SD: 516	SD: 727	SD: 527

Chú thích: SD là độ lệch chuẩn (standard deviation).

Thoạt đầu có lẽ bạn đọc, sau khi đã học qua phương pháp so sánh hai nhóm bằng kiểm định t, sẽ nghĩ rằng chúng ta cần làm 3 so sánh bằng kiểm định t: giữa nhóm 1 và 2, nhóm 2 và 3, và nhóm 1 và 3. Nhưng phương pháp này không hợp lí, vì có ba phương sai khác nhau. Phương pháp thích hợp cho so sánh là phân tích phương sai. Phân tích phương sai có thể ứng dụng để so sánh nhiều nhóm cùng một lúc (simultaneous comparisons).

11.1.1 Mô hình phân tích phương sai

Để minh họa cho phương pháp phân tích phương sai, chúng ta phải dùng kí hiệu. Gọi độ galactose của bệnh nhân i thuộc nhóm j ($j = 1, 2, 3$) là x_{ij} . Mô hình phân tích phương sai phát biểu rằng:

$$x_{ij} = \mu + \alpha_i + \varepsilon_{ij} \quad [1]$$

Hay cụ thể hơn:

$$x_{i1} = \mu + \alpha_1 + \varepsilon_{i1}$$

$$x_{i2} = \mu + \alpha_2 + \varepsilon_{i2}$$

$$x_{i3} = \mu + \alpha_3 + \varepsilon_{i3}$$

Tức là, giá trị galactose của bất cứ bệnh nhân nào bằng giá trị trung bình của toàn quần thể (μ) cộng/trừ cho ảnh hưởng của nhóm j được đo bằng hệ số ảnh hưởng α_i , và sai số ε_{ij} . Một giả định khác là ε_{ij} phải tuân theo luật phân phối chuẩn với trung bình 0 và phương sai σ^2 . Hai thông số cần ước tính là μ và α_i . Cũng như phân tích hồi qui tuyến tính, hai thông số này được ước tính bằng phương pháp bình phương nhỏ nhất; tức là tìm ước số $\hat{\mu}$ và $\hat{\alpha}_j$ sao cho $\sum(x_{ij} - \hat{\mu} - \hat{\alpha}_j)^2$ nhỏ nhất.

Quay lại với số liệu nghiên cứu trên, chúng ta có những tóm tắt thống kê như sau:

Nhóm	Số đối tượng (n_i)	Trung bình	Phương sai
1 – Crohn	$n_1 = 9$	$\bar{x}_1 = 1910$	$s_1^2 = 265944$
2 – Viêm ruột kết	$n_2 = 11$	$\bar{x}_2 = 2226$	$s_2^2 = 473387$
3 – Đối chứng	$n_3 = 20$	$\bar{x}_3 = 2804$	$s_3^2 = 277500$
Toàn bộ mẫu	$n = 40$	$\bar{x} = 2444$	

$$\text{Chú ý rằng: } x_{ij} = \bar{x} + (\bar{x}_j - \bar{x}) + (x_{ij} - \bar{x}_j) \quad [2]$$

Trong đó, \bar{x} là số trung bình của toàn mẫu, và \bar{x}_j là số trung bình của nhóm j . Nói cách khác, phần $(\bar{x}_j - \bar{x})$ phản ánh độ khác biệt (hay cũng có thể gọi là hiệu số) giữa trung bình trung bình nhóm và trung bình toàn mẫu, và phần $(x_{ij} - \bar{x}_j)$ phản ánh hiệu số giữa một galactose của một đôi tượng và số trung bình của từng nhóm. Theo đó,

- tổng bình phương cho toàn bộ mẫu là:

$$\begin{aligned} SST &= \sum_i \sum_j (x_{ij} - \bar{x})^2 \\ &= (1343 - 2444)^2 + (1393 - 2444)^2 + (1343 - 2444)^2 + \dots + (3657 - 2444)^2 \\ &= 12133923 \end{aligned}$$

- tổng bình phương vì khác nhau *giữa* các nhóm:

$$\begin{aligned} SSB &= \sum_i \sum_j (\bar{x}_i - \bar{x})^2 = \sum_j n_j (\bar{x}_j - \bar{x})^2 \\ &= 9(1910 - 2444)^2 + 11(2226 - 2444)^2 + 20(2804 - 2444)^2 \\ &= 5681168 \end{aligned}$$

- tổng bình phương vì dao động *trong* mỗi nhóm:

$$\begin{aligned} SSW &= \sum_i \sum_j (x_{ij} - \bar{x}_j)^2 = \sum_j (n_j - 1) s_j^2 \\ &= (9-1)(265944) + (11-1)(473387) + (20-1)(277500) \\ &= 12133922 \end{aligned}$$

Có thể chứng minh dễ dàng rằng: $SST = SSB + SSW$.

SSW được tính từ mỗi bệnh nhân cho 3 nhóm, cho nên trung bình bình phương cho từng nhóm (mean square – MSW) là:

$$MSW = SSW / (N - k) = 12133922 / (40 - 3) = 327944$$

và trung bình bình phương *giữa* các nhóm là:

$$MSB = SSB / (k - 1) = 5681168 / (3 - 1) = 2841810$$

Trong đó N là tổng số bệnh nhân ($N = 40$) của ba nhóm, và $k = 3$ là số nhóm bệnh nhân. Nếu có sự khác biệt giữa các nhóm, thì chúng ta kì vọng rằng MSB sẽ lớn hơn MSW . Thành ra, để kiểm tra giả thiết, chúng ta có thể dựa vào kiểm định F :

$$F = MSB / MSW = 8.67 \quad [3]$$

Với bậc tự do $k-1$ và $N-k$. Các số liệu tính toán trên đây có thể trình bày trong một bảng phân tích phương sai (ANOVA table) như sau:

Nguồn biến thiên (source of variation)	Bậc tự do (degrees of freedom)	Tổng bình phương (sum of squares)	Trung bình bình phương (mean square)	Kiểm định F
Khác biệt giữa các nhóm (between-group)	2	5681168	2841810	8.6655
Khác biệt trong từng nhóm (with-group)	37	12133923	327944	
Tổng số	39	12133923		

11.1.2 Phân tích phương sai đơn giản với R

Tất cả các tính toán trên tương đối rườm rà, và tốn khá nhiều thời gian. Tuy nhiên với R, các tính toán đó có thể làm trong vòng 1 giây, sau khi dữ liệu đã được chuẩn bị đúng cách.

(a) Nhập dữ liệu. Trước hết, chúng ta cần phải nhập dữ liệu vào R. Bước thứ nhất là báo cho R biết rằng chúng ta có ba nhóm bệnh nhân (1, 2 và), nhóm 1 gồm 9 người, nhóm 2 có 11 người, và nhóm 3 có 20 người:

```
> group <- c(1,1,1,1,1,1,1,1,1, 2,2,2,2,2,2,2,2,2,2,2,2,  
3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3)
```

Để phân tích phương sai, chúng ta phải định nghĩa biến group là một yếu tố - factor.

```
> group <- as.factor(group)
```

Bước kế tiếp, chúng ta nạp số liệu galactose cho từng nhóm như định nghĩa trên (gọi object là galactose):

```
> galactose <- c(1343,1393,1420,1641,1897,2160,2169,2279,2890,  
1264,1314,1399,1605,2385,2511,2514,2767,2827,2895,3011,  
1809,2850,1926,2964,2283,2973,2384,3171,2447,3257,2479,3271,2495,3288,  
2525,3358,2541,3643,2769,3657)
```

Đưa hai biến group và galactose vào một dataframe và gọi là data:

```
> data <- data.frame(group, galactose)  
> attach(data)
```

Sau khi đã có dữ liệu sẵn sàng, chúng ta dùng hàm lm() để phân tích phương sai như sau:

```
> analysis <- lm(galactose ~ group)
```

Trong hàm trên chúng ta cho R biết biến galactose là một hàm số của group. Gọi kết quả phân tích là analysis.

(b) Kết quả phân tích phương sai. Bây giờ chúng ta dùng lệnh anova để biết kết quả phân tích:

```
> anova(analysis)
Analysis of Variance Table

Response: galactose
          Df  Sum Sq Mean Sq F value    Pr(>F)
group      2  5683620 2841810 8.6655 0.0008191 ***
Residuals 37 12133923   327944
---
Signif. codes: 0 '****' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Trong kết quả trên, có ba cột: Df (degrees of freedom) là bậc tự do; Sum Sq là tổng bình phương (sum of squares), Mean Sq là trung bình bình phương (mean square); F value là giá trị F như định nghĩa [3] vừa đề cập phần trên; và Pr (>F) là trị số P liên quan đến kiểm định F.

Dòng group trong kết quả trên có nghĩa là bình phương giữa các nhóm (between-groups) và residual là bình phương trong mỗi nhóm (within-group). Ở đây, chúng ta có:

$$SSB = 5683620 \text{ và } MSB = 2841810$$

và:

$$MSB = 2841810 \text{ và } MSW = 327944$$

Thành ra, $F = 2841810 / 327944 = 8.6655$.

Trị số p = 0.00082 có nghĩa là tín hiệu cho thấy có sự khác biệt về độ galactose giữa ba nhóm.

(c) Ước số. Để biết thêm chi tiết kết quả phân tích, chúng ta dùng lệnh summary như sau:

```
> summary(analysis)

Call:
lm(formula = galactose ~ group)

Residuals:
    Min     1Q Median     3Q    Max
-995.5 -437.9  102.0  456.0  979.8

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1910.2     190.9 10.007 4.5e-12 ***
group2       316.3     257.4  1.229 0.226850
group3       894.3     229.9  3.891 0.000402 ***
```

```

---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 572.7 on 37 degrees of freedom
Multiple R-Squared: 0.319, Adjusted R-squared: 0.2822
F-statistic: 8.666 on 2 and 37 DF, p-value: 0.0008191

```

Theo kết quả trên đây, intercept chính là $\hat{\mu}$ trong mô hình [1]. Nói cách khác, $\hat{\mu} = 1910$ và sai số chuẩn là 190.9.

Để ước tính thông số $\hat{\alpha}_j$, R đặt $\hat{\alpha}_1 = 0$, và $\hat{\alpha}_2 = \hat{\alpha}_2 - \hat{\alpha}_1 = 316.3$, với sai số chuẩn là 257, và kiểm định $t = 316.3 / 257 = 1.229$ với trị số $p = 0.2268$. Nói cách khác, so với nhóm 1 (bệnh nhân Crohn), bệnh nhân viêm ruột kết có độ galactose trung bình cao hơn 257, nhưng độ khác biệt này không có ý nghĩa thống kê.

Tương tự, $\hat{\alpha}_3 = \hat{\alpha}_3 - \hat{\alpha}_1 = 894.3$, với sai số chuẩn là 229.9, kiểm định $t = 894.3/229.9 = 3.89$, và trị số $p = 0.00040$. So với bệnh nhân Crohn, nhóm đối chứng có độ galactose cao hơn 894, và mức độ khác biệt này có ý nghĩa thống kê.

11.2 So sánh nhiều nhóm (multiple comparisons) và điều chỉnh trị số p

Cho k nhóm, chúng ta có ít nhất là $k(k-1)/2$ so sánh. Ví dụ trên có 3 nhóm, cho nên tổng số so sánh khả dĩ là 3 (giữa nhóm 1 và 2, nhóm 1 và 3, và nhóm 2 và 3). Khi $k=10$, số lần so sánh có thể lên rất cao. Như đã đề cập trong chương 7, khi có nhiều so sánh, trị số p tính toán từ các kiểm định thống kê không còn ý nghĩa ban đầu nữa, bởi vì các kiểm định này có thể cho ra kết quả dương tính giả (tức kết quả với $p<0.05$ nhưng trong thực tế không có khác nhau hay ảnh hưởng). Do đó, trong trường hợp có nhiều so sánh, chúng ta cần phải điều chỉnh trị số p sao cho hợp lý.

Có khá nhiều phương pháp điều chỉnh trị số p , và 4 phương pháp thông dụng nhất là: Bonferroni, Scheffé, Holm và Tukey (tên của 4 nhà thống kê học danh tiếng). Phương pháp nào thích hợp nhất? Không có câu trả lời dứt khoát cho câu hỏi này, nhưng hai điểm sau đây có thể giúp bạn đọc quyết định tốt hơn:

- (a) Nếu $k < 10$, chúng ta có thể áp dụng bất cứ phương pháp nào để điều chỉnh trị số p . Riêng cá nhân tôi thì thấy phương pháp Tukey thường rất hữu ích trong so sánh.
- (b) Nếu $k > 10$, phương pháp Bonferroni có thể trở nên rất “bảo thủ”. Bảo thủ ở đây có nghĩa là phương pháp này rất ít khi nào tuyên bố một so sánh có ý nghĩa thống kê, dù trong thực tế là có thật! Trong trường hợp này, hai phương pháp Tukey, Holm và Scheffé có thể áp dụng.

Ở đây, tôi sẽ không giải thích lí thuyết đằng sau các phương pháp này (vì bạn đọc có thể tham khảo trong các sách giáo khoa về thống kê), nhưng sẽ chỉ cách sử dụng R để tiến hành các so sánh theo phương pháp của Tukey.

Quay lại ví dụ trên, các trị số p trên đây là những trị số chưa được điều chỉnh cho so sánh nhiều lần. Trong chương về trị số p, tôi đã nói các trị số này phỏng đại ý nghĩa thống kê, không phản ánh trị số p lúc ban đầu (tức 0.05). Để điều chỉnh cho nhiều so sánh, chúng ta phải sử dụng đến phương pháp điều chỉnh Bonferroni.

Chúng ta có thể dùng lệnh pairwise.t.test để có được tất cả các trị số p so sánh giữa ba nhóm như sau:

```
> pairwise.t.test(galactose, group, p.adj="bonferroni")  
  
Pairwise comparisons using t tests with pooled SD  
  
data: galactose and group  
  
 1      2  
2 0.6805 -  
3 0.0012 0.0321  
  
P value adjustment method: bonferroni
```

Kết quả trên cho thấy trị số p giữa nhóm 1 (Crohn) và viêm ruột kết là 0.6805 (tức không có ý nghĩa thống kê); giữa nhóm Crohn và đối chứng là 0.0012 (có ý nghĩa thống kê), và giữa nhóm viêm ruột kết và đối chứng là 0.0321 (tức cũng có ý nghĩa thống kê).

Một phương pháp điều chỉnh trị số p khác có tên là phương pháp Holm:

```
> pairwise.t.test(galactose, group)  
  
Pairwise comparisons using t tests with pooled SD  
  
data: galactose and group  
  
 1      2  
2 0.2268 -  
3 0.0012 0.0214  
  
P value adjustment method: holm
```

Kết quả này cũng không khác so với phương pháp Bonferroni.

Tất cả các phương pháp so sánh trên sử dụng một sai số chuẩn chung cho cả ba nhóm. Nếu chúng ta muốn sử dụng cho từng nhóm thì lệnh sau đây (pool.sd=F) sẽ đáp ứng yêu cầu đó:

```
> pairwise.t.test(galactose, group, pool.sd=FALSE)  
  
Pairwise comparisons using t tests with non-pooled SD
```

```

data: galactose and group

 1      2
2 0.2557 -
3 0.0017 0.0544

P value adjustment method: holm

```

Một lần nữa, kết quả này cũng không làm thay đổi kết luận.

11.2.1 So sánh nhiều nhóm bằng phương pháp Tukey

Trong các phương pháp trên, chúng ta chỉ biết trị số p so sánh giữa các nhóm, nhưng không biết mức độ khác biệt cũng như khoảng tin cậy 95% giữa các nhóm. Để có những ước số này, chúng ta cần đến một hàm khác có tên là `aov` (viết tắt từ analysis of variance) và hàm `TukeyHSD` (`HSD` là viết tắt từ Honest Significant Difference, tạm dịch nôm na là “Khác biệt có ý nghĩa thành thật”) như sau:

```

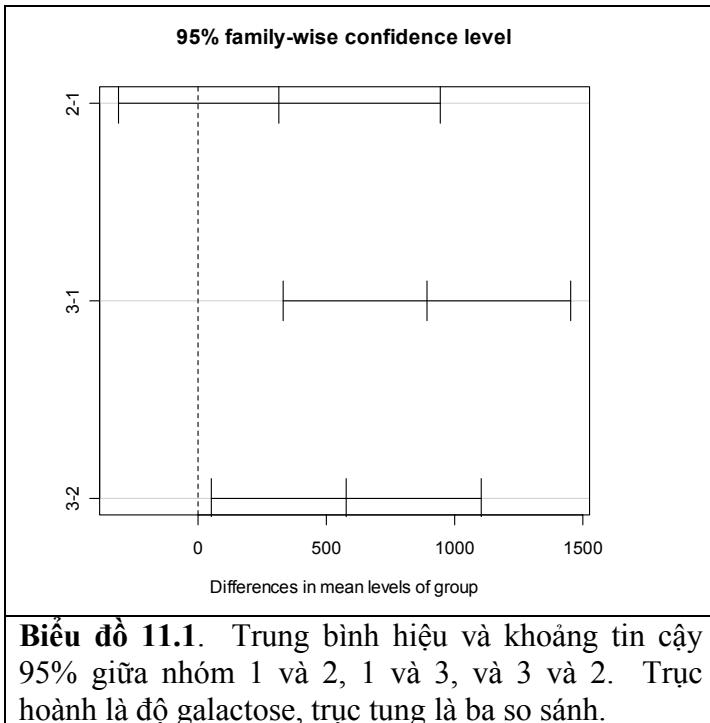
> res <- aov(galactose ~ group)
> TukeyHSD(res)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = galactose ~ group)

$group
   diff      lwr      upr      p adj
2-1 316.3232 -312.09857 944.745 0.4439821
3-1 894.2778  333.07916 1455.476 0.0011445
3-2 577.9545   53.11886 1102.790 0.0281768

```

Kết quả trên cho chúng ta thấy nhóm 3 và 1 khác nhau khoảng 894 đơn vị, và khoảng tin cậy 95% từ 333 đến 1455 đơn vị. Tương tự, galactose trong nhóm bệnh nhân viêm ruột kết thấp hơn nhóm đối chứng (nhóm 3) khoảng 578 đơn vị, và khoảng tin cậy 95% từ 53 đến 1103.



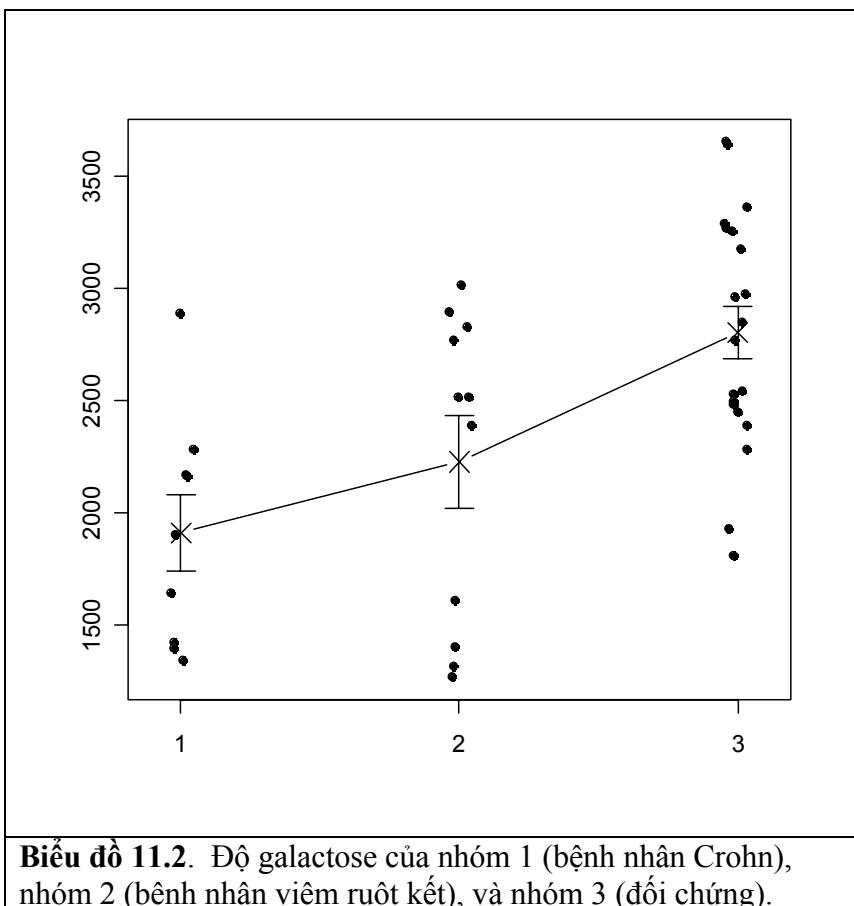
11.2.2 Phân tích bằng biểu đồ

Một phân tích thống kê không thể nào hoàn tất nếu không có một đồ thị minh họa cho kết quả. Các lệnh sau đây vẽ đồ thị thể hiện độ galactose trung bình và sai số chuẩn cho từng nhóm bệnh nhân. Biểu đồ này cho thấy, nhóm bệnh nhân Crohn có độ galactose thấp nhất (nhưng không thấp hơn nhóm viêm ruột kết), và cả hai nhóm thấp hơn nhóm đối chứng và sứ khác biệt này có ý nghĩa thống kê.

```

> xbar <- tapply(galactose, group, mean)
> s <- tapply(galactose, group, sd)
> n <- tapply(galactose, group, length)
> sem <- s/sqrt(n)
> stripchart(galactose ~ group, "jitter", jit=0.05, pch=16, vert=TRUE)
> arrows(1:3, xbar+sem, 1:3, xbar-sem, angle=90, code=3, length=0.1)
> lines(1:3, xbar, pch=4, type="b", cex=2)

```



Biểu đồ 11.2. Độ galactose của nhóm 1 (bệnh nhân Crohn), nhóm 2 (bệnh nhân viêm ruột kết), và nhóm 3 (đối chứng).

11.3 Phân tích bằng phương pháp phi tham số

Phương pháp so sánh nhiều nhóm phi tham số (non-parametric statistics) tương đương với phương pháp phân tích phương sai là Kruskal-Wallis. Cũng như phương pháp Wilcoxon so sánh hai nhóm theo phương pháp phi tham số, phương pháp Kruskal-Wallis cũng biến đổi số liệu thành thứ bậc (ranks) và phân tích độ khác biệt thứ bậc này giữa các nhóm. Hàm `kruskal.test` trong R có thể giúp chúng ta trong kiểm định này:

```
> kruskal.test(galactose ~ group)

Kruskal-Wallis rank sum test

data: galactose by group
Kruskal-Wallis chi-squared = 12.1381, df = 2, p-value = 0.002313
```

Trị số p từ kiểm định này khá thấp ($p = 0.002313$) cho thấy có sự khác biệt giữa ba nhóm như phân tích phương sai qua hàm `lm` trên đây. Tuy nhiên, một bất tiện của kiểm định phi tham số Kruskal-Wallis là phương pháp này không cho chúng ta biết hai nhóm nào khác nhau, mà chỉ cho một trị số p chung. Trong nhiều trường hợp, phân tích

phi tham số như kiểm định Kruskal-Wallis thường không có hiệu quả như các phương pháp thống kê tham số (parametric statistics).

11.4 Phân tích phương sai hai chiều (two-way analysis of variance - ANOVA)

Phân tích phương sai đơn giản hay một chiều chỉ có một yếu tố (factor). Nhưng phân tích phương sai hai chiều (two-way ANOVA), như tên gọi, có hai yếu tố. Phương pháp phân tích phương sai hai chiều chỉ đơn giản khai triển từ phương pháp phân tích phương sai đơn giản. Thay vì ước tính phương sai của một yếu tố, phương pháp phân sai hai chiều ước tính phương sai của hai yếu tố.

Ví dụ 2. Trong ví dụ sau đây, để đánh giá hiệu quả của một kỹ thuật sơn mới, các nhà nghiên cứu áp dụng sơn trên 3 loại vật liệu (1, 2 và 3) trong hai điều kiện (1, 2). Mỗi điều kiện và loại vật liệu, nghiên cứu được lặp lại 3 lần. Độ bền được đo là chỉ số bền bỉ (tạm gọi là score). Tổng cộng, có 18 số liệu như sau:

Bảng 11.2. Độ bền bỉ của sơn cho 2 điều kiện và 3 vật liệu

Điều kiện (i)	Vật liệu (j)		
	1	2	3
1	4.1, 3.9, 4.3	3.1, 2.8, 3.3	3.5, 3.2, 3.6
2	2.7, 3.1, 2.6	1.9, 2.2, 2.3	2.7, 2.3, 2.5

Số liệu này có thể tóm lược bằng số trung bình cho từng điều kiện và vật liệu trong bảng thống kê sau đây:

Bảng 11.3. Tóm lược số liệu từ thí nghiệm độ bền bỉ của nước sơn

Điều kiện (i)	Vật liệu (j)			Trung bình cho 3 vật liệu
	1	2	3	
Trung bình				
1	4.10	3.07	3.43	3.533
2	2.80	2.13	2.50	2.478
Trung bình 2 nhóm	3.450	2.600	2.967	3.00
Phương sai				
1	0.040	0.063	0.043	
2	0.070	0.043	0.040	

Những tính toán sơ khởi trên đây cho thấy có thể có sự khác nhau (hay ảnh hưởng) của điều kiện và vật liệu thí nghiệm.

Gọi x_{ij} là score của điều kiện i ($i = 1, 2$) cho vật liệu j ($j = 1, 2, 3$). (Để đơn giản hóa vấn đề, chúng ta tạm thời bỏ qua k đối tượng). Mô hình phân tích phương sai hai chiều phát biểu rằng:

$$x_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij} \quad [4]$$

Hay cụ thể hơn:

$$x_{11} = \mu + \alpha_1 + \beta_1 + \varepsilon_{11}$$

$$x_{12} = \mu + \alpha_1 + \beta_2 + \varepsilon_{12}$$

$$x_{13} = \mu + \alpha_1 + \beta_3 + \varepsilon_{13}$$

$$x_{21} = \mu + \alpha_2 + \beta_1 + \varepsilon_{21}$$

$$x_{22} = \mu + \alpha_2 + \beta_2 + \varepsilon_{22}$$

$$x_{23} = \mu + \alpha_2 + \beta_3 + \varepsilon_{23}$$

μ là số trung bình cho toàn quần thể, các hệ số α_i (ảnh hưởng của điều kiện i) và β_j (ảnh hưởng của vật liệu j) cần phải ước tính từ số liệu thực tế. ε_{ij} được giả định tuân theo luật phân phối chuẩn với trung bình 0 và phương sai σ^2 .

Trong phân tích phương sai hai chiều, chúng ta cần chia tổng bình phương ra thành 3 nguồn:

- nguồn thứ nhất là tổng bình phương do biến đổi giữa 2 điều kiện:

$$\begin{aligned} SSc &= \sum_i n_i (\bar{x}_i - \bar{x})^2 \\ &= 9(3.533 - 3.00)^2 + 9(2.478 - 3.00)^2 \\ &= 5.01 \end{aligned}$$

- nguồn thứ hai là tổng bình phương do biến đổi giữa 3 vật liệu:

$$\begin{aligned} SSm &= \sum_j n_j (\bar{x}_j - \bar{x})^2 \\ &= 6(3.45 - 3.00)^2 + 6(2.60 - 3.00)^2 + 6(2.967 - 3.00)^2 \\ &= 2.18 \end{aligned}$$

- nguồn thứ ba là tổng bình phương phần dư (residual sum of squares):

$$\begin{aligned}
SSe &= \sum_i \sum_j (x_{ij} - \bar{x}_i - \bar{x}_j + \bar{x})^2 = \sum (n_{ij} - 1) s_{ij}^2 \\
&= 2(0.040) + 2(0.063) + 2(0.043) + 2(0.070) + 2(0.043) + 2(0.040) \\
&= 0.73
\end{aligned}$$

Trong các phương trình trên, $n = 3$ (lặp lại 3 lần cho mỗi điều kiện và vật liệu), $m = 3$ vật liệu, \bar{x} là số trung bình cho toàn mẫu, \bar{x}_i là số trung bình cho từng điều kiện, \bar{x}_j là số trung bình cho từng vật liệu. Vì SSc có $m-1$ bậc tự do, SSm có $(n-1)$ bậc tự do, và SSe có $N-nm+2$ bậc tự do, trong đó N là tổng số mẫu (tức 18). Do đó, các trung bình bình phương

- giữa hai điều kiện: $MSc = SSc / (m-1) = 5.01 / 1 = 5.01$
- giữa ba vật liệu: $MSm = SSm / (n-1) = 2.18 / 2 = 1.09$
- phần dư: $MSe = SSE / (N-nm+2) = 0.73 / 14 = 0.052$

Do đó, so sánh độ khác biệt giữa hai điều kiện dựa vào kiểm định $F = MSc/Mse$ với bậc tự do 1 và 14. Tương tự, so sánh độ khác biệt giữa ba vật liệu có thể dựa vào kiểm định $F = MSm/Mse$ với bậc tự do 2 và 14. Các phân tích trên có thể trình bày trong một bảng phân tích phương sai như sau:

Nguồn biến thiên (source of variation)	Bậc tự do (degrees of freedom)	Tổng bình phương (sum of squares)	Trung bình bình phương (mean square)	Kiểm định F
Khác biệt giữa 2 điều kiện	1	5.01	5.01	95.6
Khác biệt giữa 3 vật liệu	2	2.18	1.09	20.8
Phần dư (residual)	14	0.73	0.052	
Tổng số	17	7.92		

11.4.1 Phân tích phương sai hai chiều với R

(a) Bước đầu tiên là nhập số liệu từ bảng 11.2 vào R. Chúng ta cần phải tổ chức dữ liệu sao cho có 4 biến như sau:

Condition (điều kiện)	Material (vật liệu)	Đối tượng	Score
1	1	1	4.1
1	1	2	3.9
1	1	3	4.3
1	2	4	3.1
1	2	5	2.8
1	2	6	3.3
1	3	7	3.5
1	3	8	3.2

1	3	9	3.6
2	1	10	2.7
2	1	11	3.1
2	1	12	2.6
2	2	13	1.9
2	2	14	2.2
2	2	15	2.3
2	3	16	2.7
2	3	17	2.3
2	3	18	2.5

Chúng ta có thể tạo ra một dãy số bằng cách sử dụng hàm `gl` (generating levels). Cách sử dụng hàm này có thể minh họa như sau:

```
> gl(9, 1, 18)
[1] 1 2 3 4 5 6 7 8 9 1 2 3 4 5 6 7 8 9
Levels: 1 2 3 4 5 6 7 8 9
```

Trong lệnh trên, chúng ta tạo ra một dãy số 1,2,3, ... 9 hai lần (với tổng số 18 số). Mỗi một lần là một nhóm. Trong khi lệnh:

```
> gl(4, 9, 36)
[1] 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 4 4 4 4 4 4 4 4 4
Levels: 1 2 3 4
```

Trong lệnh trên, chúng ta tạo ra một dãy số với 4 bậc (1,2,3,4) 9 lần (với tổng số 36 số).

Do đó, để tạo ra các bậc cho điều kiện và vật liệu, chúng ta lệnh như sau:

```
> condition <- gl(2, 9, 18)
> material <- gl(3, 3, 18)
```

Và tạo nên 18 mã số (từ 1 đến 18):

```
> id <- 1:18
```

Sau cùng là số liệu cho score:

```
> score <- c(4.1, 3.9, 4.3, 3.1, 2.8, 3.3, 3.5, 3.2, 3.6,
   2.7, 3.1, 2.6, 1.9, 2.2, 2.3, 2.7, 2.3, 2.5)
```

Tất cả cho vào một dataframe tên là data:

```
> data <- data.frame(condition, material, id, score)
> attach(data)
```

(b) Phân tích và kết quả sơ khởi. Bây giờ số liệu đã sẵn sàng cho phân tích. Để phân tích phương sai hai chiều, chúng ta vẫn sử dụng lệnh `lm` với các thông số như sau:

```
> twoway <- lm(score ~ condition + material)
> anova(twoway)
Analysis of Variance Table
```

```

Response: score
      Df Sum Sq Mean Sq F value    Pr(>F)
condition  1 5.0139  5.0139  95.575 1.235e-07 ***
material   2 2.1811  1.0906  20.788 6.437e-05 ***
Residuals 14 0.7344  0.0525
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Ba nguồn dao động (variation) của score được phân tích trong bảng trên. Qua trung bình bình phương (mean square), chúng ta thấy ảnh hưởng của điều kiện có vẻ quan trọng hơn là ảnh hưởng của vật liệu thí nghiệm. Tuy nhiên, cả hai ảnh hưởng đều có ý nghĩa thống kê, vì trị số p rất thấp cho hai yếu tố.

(c) Ước số. Chúng ta yêu cầu R tóm lược các ước số phân tích bằng lệnh **summary**:

```

> summary(twoway)

Call:
lm(formula = score ~ condition + material)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.32778 -0.16389  0.03333  0.16111  0.32222 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  3.9778    0.1080  36.841 2.43e-15 ***
condition2   -1.0556    0.1080  -9.776 1.24e-07 ***
material2   -0.8500    0.1322  -6.428 1.58e-05 ***
material3   -0.4833    0.1322  -3.655  0.0026 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.229 on 14 degrees of freedom
Multiple R-Squared:  0.9074,    Adjusted R-squared:  0.8875 
F-statistic: 45.72 on 3 and 14 DF,  p-value: 1.761e-07

```

Kết quả trên cho thấy so với điều kiện 1, điều kiện 2 có score thấp hơn khoảng 1.056 và sai số chuẩn là 0.108, với trị số p = 1.24e-07, tức có ý nghĩa thống kê. Ngoài ra, so với vật liệu 1, score cho vật liệu 2 và 3 cũng thấp hơn đáng kể với độ thấp nhất ghi nhận ở vật liệu 2, và ảnh hưởng của vật liệu thí nghiệm cũng có ý nghĩa thống kê.

Giá trị có tên là “Residual standard error” được ước tính từ trung bình bình phương phần dư trong phần (a), tức là $\sqrt{0.0525} = 0.229$, tức là ước số của $\hat{\sigma}$.

Hệ số xác định bội (R^2) cho biết hai yếu tố điều kiện và vật liệu giải thích khoảng 91% độ dao động của toàn bộ mẫu. Hệ số này được tính từ tổng bình phương trong kết quả phần (a) như sau:

$$R^2 = \frac{5.0139 + 2.1811}{5.0139 + 2.1811 + 0.7344} = 0.9074$$

Và sau cùng, hệ số R^2 điều chỉnh phản ánh độ “cải tiến” của mô hình. Để hiểu hệ số này tốt hơn, chúng ta thấy phương sai của toàn bộ mẫu là $s^2 = (5.0139 + 2.1811 + 0.7344) / 17 = 0.4644$. Sau khi điều chỉnh cho ảnh hưởng của điều kiện và vật liệu, phương sai này còn 0.0525 (tức là residual mean square). Như vậy hai yếu tố này làm giảm phương sai khoảng $0.4644 - 0.0525 = 0.4119$. Và hệ số R^2 điều chỉnh là:

$$\text{Adj } R^2 = 0.4119 / 0.4644 = 0.88$$

Tức là sau khi điều chỉnh cho hai yếu tố điều kiện và vật liệu phương sai của score giảm khoảng 88%.

(d) Hiệu ứng tương tác (interaction effects)

Để cho phân tích hoàn tất, chúng ta còn phải xem xét đến khả năng ảnh hưởng của hai yếu tố này có thể tương tác nhau (interactive effects). Tức là mô hình score trở thành:

$$x_{ij} = \mu + \alpha_i + \beta_j + (\alpha_i \beta_j)_{ij} + \varepsilon_{ij}$$

Chú ý phương trình trên có phần $(\alpha_i \beta_j)_{ij}$ phản ánh sự tương tác giữa hai yếu tố. Và chúng ta chỉ đơn giản lệnh R như sau:

```
> anova(twoway <- lm(score ~ condition+ material+condition*material))
Analysis of Variance Table

Response: score
            Df Sum Sq Mean Sq F value    Pr(>F)
condition      1 5.0139  5.0139 100.2778 3.528e-07 ***
material       2 2.1811  1.0906  21.8111 0.0001008 ***
condition:material 2 0.1344  0.0672   1.3444  0.2972719
Residuals     12 0.6000  0.0500
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Kết quả phân tích trên ($p = 0.297$ cho ảnh hưởng tương tác). Chúng ta có bằng chứng để kết luận rằng ảnh hưởng tương tác giữa vật liệu và điều kiện không có ý nghĩa thống kê, và chúng ta chấp nhận mô hình [4], tức không có tương tác.

(e) So sánh giữa các nhóm. Chúng ta sẽ ước tính độ khác biệt giữa hai điều kiện và ba vật liệu bằng hàm TukeyHSD với aov:

```
> res <- aov(score ~ condition+ material+condition)
> TukeyHSD(res)
Tukey multiple comparisons of means
 95% family-wise confidence level
```

```

Fit: aov(formula = score ~ condition + material + condition)

$condition
      diff      lwr      upr p adj
2-1 -1.055556 -1.287131 -0.8239797 1e-07

$material
      diff      lwr      upr p adj
2-1 -0.8500000 -1.19610279 -0.5038972 0.0000442
3-1 -0.4833333 -0.82943612 -0.1372305 0.0068648
3-2  0.3666667  0.02056388  0.7127695 0.0374069

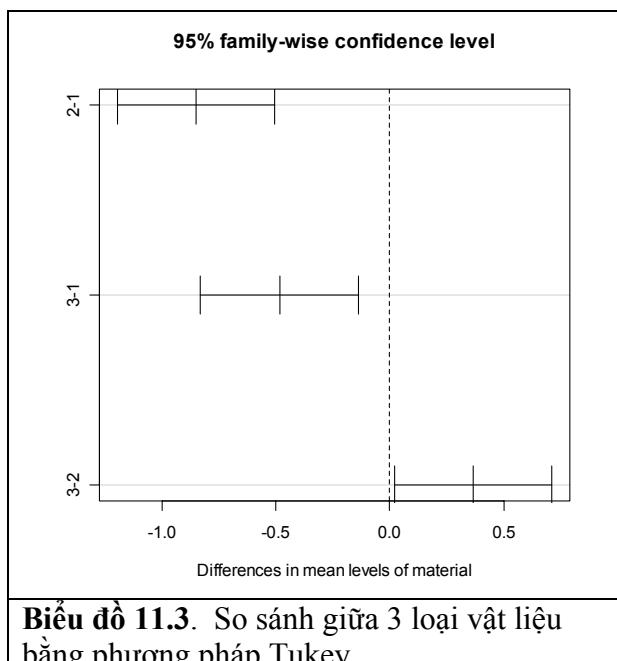
```

Biểu đồ sau đây sẽ minh họa cho các kết quả trên:

```

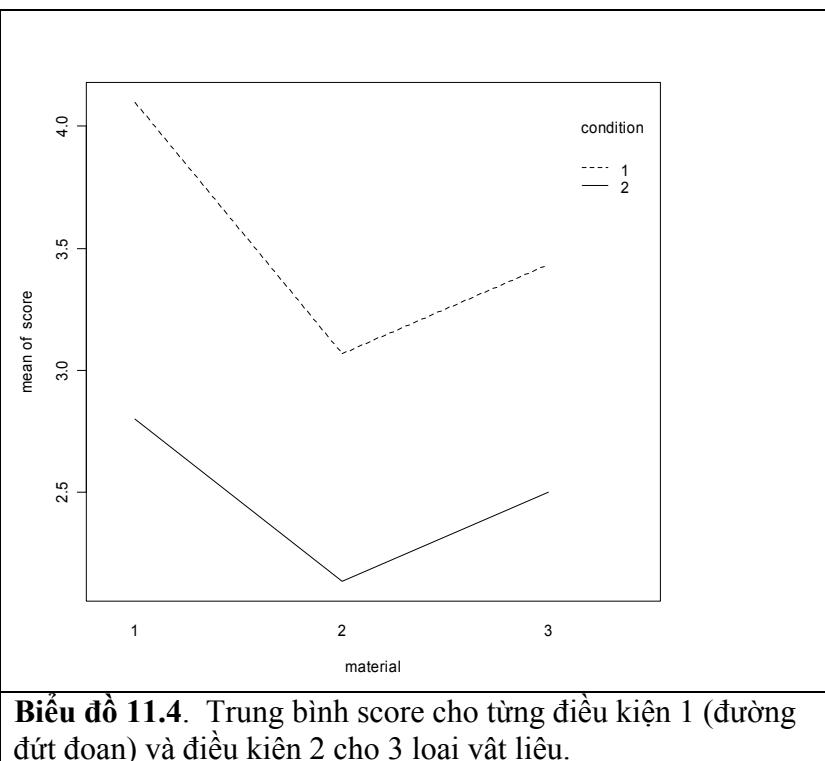
> plot(TukeyHSD(res), ordered=TRUE)
There were 16 warnings (use warnings() to see them)

```



(f) Biểu đồ. Để xem qua độ ảnh hưởng của hai yếu tố điều kiện và vật liệu, chúng ta cần phải có một đồ thị, mà trong phân tích phương sai gọi là đồ thị tương tác. Hàm interaction.plot cung cấp phương tiện để vẽ biểu đồ này:

```
> interaction.plot(score, condition, material)
```



Biểu đồ 11.4. Trung bình score cho từng điều kiện 1 (đường đứt đoạn) và điều kiện 2 cho 3 loại vật liệu.

11.5 Phân tích hiệp biến (analysis of covariance - ANCOVA)

Phân tích hiệp biến (sẽ viết tắt là ANCOVA) là phương pháp phân tích sử dụng cả hai mô hình hồi qui tuyến tính và phân tích phương sai. Trong phân tích hồi qui tuyến tính, cả hai biến phụ thuộc (dependent variable, cũng có thể gọi là “biến ứng” – response variable) và biến độc lập (independent variable hay predictor variable) phần lớn là ở dạng liên tục (continuous variable), như độ cholesterol và độ tuổi chẳng hạn. Trong phân tích phương sai, biến phụ thuộc là biến liên tục, còn biến độc lập thì ở dạng thứ bậc và thể loại (categorical variable), như độ galactose và nhóm bệnh nhân trong ví dụ 1 chẳng hạn. Trong phân tích hiệp biến, biến phụ thuộc là liên tục, nhưng biến độc lập có thể là liên tục và thể loại.

Ví dụ 3. Trong nghiên cứu mà kết quả được trình bày dưới đây, các nhà nghiên cứu đo chiều cao và độ tuổi của 18 học sinh thuộc vùng thành thị (urban) và 14 học trò thuộc vùng nông thôn (rural).

Bảng 11.4. Chiều cao của học trò vùng thành thị và nông thôn			
Area	ID	Age (months)	Height (cm)
urban	1	109	137.6
urban	2	113	147.8
urban	3	115	136.8
urban	4	116	140.7

Câu hỏi đặt ra là có sự khác biệt nào về chiều cao giữa trẻ em ở thành thị và nông thôn hay không. Nói cách khác, môi trường cư trú có ảnh hưởng đến chiều cao hay không, và nếu có thì mức độ ảnh hưởng là bao nhiêu?

Một yếu tố có ảnh hưởng lớn đến chiều cao là độ tuổi. Trong độ tuổi trưởng thành, chiều cao tăng theo độ tuổi. Do đó, so sánh chiều cao giữa hai nhóm chỉ có thể khách quan nếu độ tuổi giữa hai nhóm phải tương đương nhau. Để đảm bảo tính khách quan của so sánh, chúng ta cần phải phân tích số liệu bằng mô hình hiệp biến.

Việc đầu tiên là chúng ta phải nhập số liệu vào R với những lệnh sau đây:

urban	5	119	132.7
urban	6	120	145.4
urban	7	121	135.0
urban	8	124	133.0
urban	9	126	148.5
urban	10	129	148.3
urban	11	130	147.5
urban	12	133	148.8
urban	13	134	133.2
urban	14	135	148.7
urban	15	137	152.0
urban	16	139	150.6
urban	17	141	165.3
urban	18	142	149.9
rural	1	121	139.0
urban	2	121	140.9
urban	3	128	134.9
urban	4	129	149.5
urban	5	131	148.7
urban	6	132	131.0
urban	7	133	142.3
urban	8	134	139.9
urban	9	138	142.9
urban	10	138	147.7
urban	11	138	147.7
urban	12	140	134.6
urban	13	140	135.8
urban	14	140	148.5

```

> # tạo ra dãy số id
> id <- c(1:18, 1:14)
> # group 1=urban 2=rural và cần phải xác định group là một "factor"
> group <- c(1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,
             2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2)
> group <- as.factor(group)

> # nhập dữ liệu
> age <- c(109,113,115,116,119,120,121,124,126,129,130,133,134,135,
           137,139,141,142,
           121,121,128,129,131,132,133,134,138,138,138,140,140,140)

> height <- c(137.6,147.8,136.8,140.7,132.7,145.4,135.0,133.0,148.5,
              148.3,147.5,148.8,133.2,148.7,152.0,150.6,165.3,149.9,
              139.0,140.9,134.9,149.5,148.7,131.0,142.3,139.9,142.9,
              147.7,147.7,134.6,135.8,148.5)

> # tạo một data frame
> data <- data.frame(id, group, age, height)
> attach(data)

```

Chúng ta thử xem qua vài chỉ số thống kê mô tả bằng cách ước tính độ tuổi và chiều cao trung bình cho từng nhóm học sinh:

```

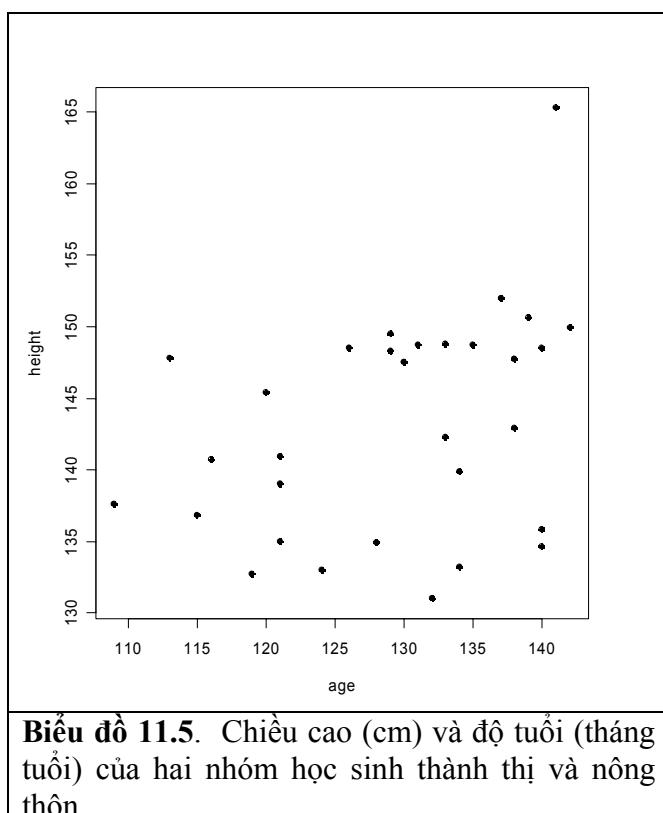
> tapply(age, group, mean)
  1      2
126.8333 133.0714

> tapply(height, group, mean)
  1      2
144.5444 141.6714

```

Kết quả trên cho thấy nhóm học sinh thành thị có độ tuổi thấp hơn học sinh nông thôn khoảng 6.3 tháng (126.8 – 133.1). Tuy nhiên, chiều cao của học sinh thành thị cao hơn học sinh nông thôn khoảng 2.8 cm (144.5 – 141.7). Bạn đọc có thể dùng kiểm định t để thấy rằng sự khác biệt về độ tuổi giữa hai nhóm có ý nghĩa thống kê ($p = 0.045$).

Ngoài ra, biểu đồ sau đây còn cho thấy có một mối liên hệ tương quan giữa tuổi và chiều cao:



Vì hai nhóm khác nhau về độ tuổi, và tuổi có liên hệ với chiều cao, cho nên chúng ta không thể phát biểu hay so sánh chiều cao giữa 2 nhóm học sinh mà không điều chỉnh cho độ tuổi. Để điều chỉnh độ tuổi, chúng ta sử dụng phương pháp phân tích hiệp biến.

11.5.1 Mô hình phân tích hiệp biến

Gọi y là chiều cao, x là độ tuổi, và g là nhóm. Mô hình căn bản của ANCOVA giả định rằng mối liên hệ giữa y và x là một đường thẳng, và độ dốc (gradient hay slope)

của hai nhóm trong mối liên hệ này không khác nhau. Nói cách khác, viết theo kí hiệu của hồi qui tuyến tính, chúng ta có:

$$\begin{aligned} y_1 &= \alpha_1 + \beta x + e_1 && \text{in group 1} \\ y_2 &= \alpha_2 + \beta x + e_2 && \text{in group 2.} \end{aligned} \quad [5]$$

Trong đó:

- α_1 : là giá trị trung bình của y khi $x=0$ của nhóm 1;
- α_2 : là giá trị trung bình của y khi $x=0$ của nhóm 2;
- β : độ dốc của mối liên hệ giữa y và x ;
- e_1 và e_2 : biến số ngẫu nhiên với trung bình 0 và phương sai σ^2 .

Gọi \bar{x} là số trung bình của độ tuổi cho cả 2 nhóm, \bar{x}_1 và \bar{x}_2 là tuổi trung bình của nhóm 1 và nhóm 2. Như nói trên, nếu $\bar{x}_1 \neq \bar{x}_2$, thì so sánh chiều cao trung bình của nhóm 1 và 2 (\bar{y}_1 và \bar{y}_2) sẽ thiếu khách quan, vì

$$\begin{aligned} \bar{y}_1 &= \alpha_1 + \beta \bar{x}_1 + e_1 \\ \bar{y}_2 &= \alpha_2 + \beta \bar{x}_2 + e_2 \end{aligned}$$

và mức độ khác biệt giữa hai nhóm bây giờ tùy thuộc vào hệ số β :

$$\bar{y}_1 - \bar{y}_2 = \alpha_1 - \alpha_2 + \beta(\bar{x}_1 - \bar{x}_2)$$

Chú ý rằng trong mô hình [5], chúng ta có thể diễn dịch $\alpha_1 - \alpha_2$ là độ khác biệt chiều cao trung bình giữa hai nhóm *nếu* cả hai nhóm có cùng tuổi trung bình. Mức khác biệt này *thì hiện ảnh hưởng* của hai nhóm *nếu* không có một yếu tố nào liên hệ đến y . Thành ra, để ước tính $\alpha_1 - \alpha_2$, chúng ta không thể đơn giản trừ hai số trung bình \bar{y}_1 - \bar{y}_2 , nhưng phải điều chỉnh cho x . Gọi x^* là một giá trị chung cho cả hai nhóm, chúng ta có thể ước tính giá trị điều chỉnh y cho nhóm 1 (kí hiệu \bar{y}_{1a}) như sau:

$$\bar{y}_{1a} = \bar{y}_1 - \beta(\bar{x}_1 - x^*)$$

\bar{y}_{1a} có thể xem là một ước số cho chiều cao trung bình của nhóm 1 (thành thị) cho giá trị x là x^* . Tương tự,

$$\bar{y}_{2a} = \bar{y}_2 - \beta(\bar{x}_2 - x^*)$$

là số cho chiều cao trung bình của nhóm 1 (nông thôn) với cùng giá trị x^* . Từ đây, chúng ta có thể ước tính ảnh hưởng của thành thị và nông thôn bằng công thức sau đây:

$$\bar{y}_{1a} - \bar{y}_{2a} = \bar{y}_2 - \bar{y}_1 - \beta(\bar{x}_1 - \bar{x}_2)$$

Do đó, vấn đề là chúng ta phải ước tính β . Có thể chứng minh rằng ước số β từ phương pháp bình phương nhỏ nhất cũng là ước tính khách quan cho $\alpha_1 - \alpha_2$. Khi viết bằng mô hình tuyến tính, mô hình biến có thể mô tả như sau:

$$y = \alpha + \beta x + \gamma g + \delta(xg) + e \quad [6]$$

Nói cách khác, mô hình trên phát biểu rằng chiều cao của một học sinh bị ảnh hưởng bởi 3 yếu tố: độ tuổi (β), thành thị hay nông thôn (γ), và tương tác giữa hai yếu tố đó (δ). Nếu $\delta = 0$ (tức ảnh hưởng tương tác không có ý nghĩa thống kê), mô hình trên giảm xuống thành:

$$y = \alpha + \beta x + \gamma g + e \quad [7]$$

Nếu $\gamma = 0$ (tức ảnh hưởng của thành thị không có ý nghĩa thống kê), mô hình trên giảm xuống thành:

$$y = \alpha + \beta x + e \quad [8]$$

11.5.2 Phân tích bằng R

Các thảo luận vừa trình bày trên xem ra khá phức tạp, nhưng trong thực tế, với R, cách ước tính rất đơn giản bằng hàm lm. Chúng ta sẽ phân tích ba mô hình [6], [7] và [8]:

```
> # model 6
> model6 <- lm(height ~ group + age + group:age)

> # model 7
> model7 <- lm(height ~ group + age)

> # model 8
> model8 <- lm(height ~ age)
```

Chúng ta cũng có thể so sánh cả ba mô hình cùng một lúc bằng lệnh anova như sau:

```
> anova(model6, model7, model8)
Analysis of Variance Table

Model 1: height ~ group + age + group:age
Model 2: height ~ group + age
Model 3: height ~ age
  Res.Df   RSS Df Sum of Sq    F   Pr(>F)
  1      28 1270.44
  2      29 1338.02 -1     -67.57 1.4893 0.23251
  3      30 1545.95 -1    -207.93 4.5827 0.04114 *
```

```
---
Signif. codes: 0 '****' 0.001 '***' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Chú ý “model 1” chính là mô hình [6], “model 2” là mô hình [7], và “model 3” là mô hình [8]. RSS là residual sum of squares, tức tổng bình phương phần dư cho mỗi mô hình. Kết quả phân tích trên cho thấy:

- Toàn bộ mẫu có $18+14=32$ học sinh, mô hình [6] có 4 thông số (α, β, γ và δ), cho nên mô hình này có $32-4 = 28$ bậc tự do. Tổng bình phương của mô hình là 1270.44.
- mô hình [7] có 3 thông số (tức còn 29 bậc tự do), cho nên tổng bình phương phần dư cao hơn mô hình [7]. Tuy nhiên, đứng trên phương diện xác suất thì trung bình bình phương phần dư của mô hình này $1338.02 / 29 = 46.13$, không khác mấy so với mô hình [6] (trung bình bình phương là: $1270.44 / 28 = 45.36$), vì trị số $p = 0.2325$, tức không có ý nghĩa thống kê. Nói cách khác, bỏ hệ số tương tác δ không làm thay đổi khả năng tiên đoán của mô hình một cách đáng kể.
- Mô hình [8] chỉ có 2 thông số (và do đó có 30 bậc tự do), với tổng bình phương là 1545.95. Trung bình bình phương phần dư của mô hình này là 51.53 ($1545.95 / 30$), tức cao hơn hai mô hình [6] một cách đáng kể, vì trị số $p = 0.0411$.

Qua phân tích trên, chúng ta thấy mô hình [7] là tối ưu hơn cả, vì chỉ cần 3 thông số mà có thể “giải thích” được dữ liệu một cách đầy đủ. Bây giờ chúng ta sẽ chú tâm vào phân tích kết quả của mô hình này.

```
> summary(model7)
```

Call:

```
lm(formula = height ~ group + age)
```

Residuals:

Min	1Q	Median	3Q	Max
-14.324	-3.285	0.879	3.956	14.866

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	91.8171	17.9294	5.121	1.81e-05 ***
group2	-5.4663	2.5749	-2.123	0.04242 *
age	0.4157	0.1408	2.953	0.00619 **

```
Signif. codes: 0 '****' 0.001 '***' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Residual standard error: 6.793 on 29 degrees of freedom

Multiple R-Squared: 0.2588, Adjusted R-squared: 0.2077

F-statistic: 5.063 on 2 and 29 DF, p-value: 0.01300

Qua phần ước tính thông số trình bày trên đây, chúng ta thấy tính trung bình chiều cao học sinh tăng khoảng 0.41 cm cho mỗi tháng tuổi. Chú ý trong kết quả trên, phần “group2” có nghĩa là hệ số hồi qui (regression coefficient) cho nhóm 2 (tức là nông thôn), vì R phải đặt hệ số cho nhóm 1 bằng 0 để tiện việc tính toán. Vì thế, chúng ta có hai phương trình (hay hai đường biểu diễn) cho hai nhóm học sinh như sau:

Đối với học sinh thành thị:

$$\text{Height} = 91.817 + 0.4157(\text{age})$$

Và đối với học sinh nông thôn:

$$\text{Height} = 91.817 - 5.4663(\text{rural}) + 0.4157(\text{age})$$

Nói cách khác, sau khi điều chỉnh cho độ tuổi, nhóm học sinh nông thôn (rural) có chiều cao thấp hơn nhóm thành thị khoảng 5.5 cm và mức độ khác biệt này có ý nghĩa thống kê vì trị số $p = 0.0424$. (Chú ý là trước khi điều chỉnh cho độ tuổi, mức độ khác biệt là 2.8 cm).

Các biểu đồ sau đây sẽ minh họa cho các mô hình trên:

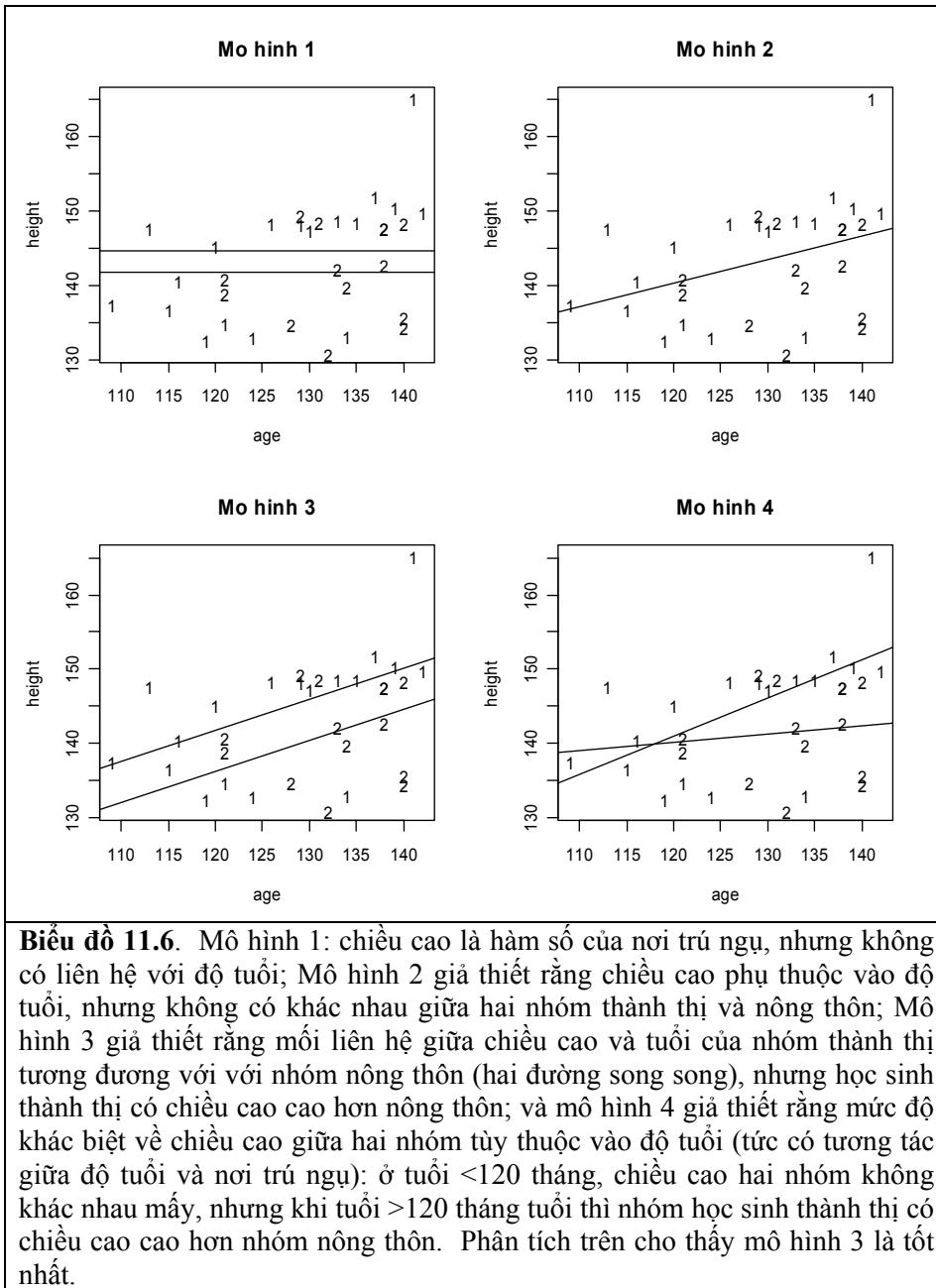
```
> par(mfrow=c(2,2))

> plot(age, height, pch=as.character(group),
      main="Mô hình 1")
> abline(144.54, 0) #mean value for urban
> abline(141.67, 0) #mean value for rural

> plot(age, height, pch=as.character(group),
      main="Mô hình 2")
> abline(102.63, 0.3138) #single line for dependence on age

> plot(age, height, pch=as.character(group),
      main="Mô hình 3")
> abline(91.8, 0.416) #line for males
> abline(91.8-5.46, 0.416) #line for females parallel

> plot(age, height, pch=as.character(group),
      main="Mô hình 4")
> abline(79.7, 0.511) #line for males
> abline(79.7+47.08, 0.511-0.399) #line for females parallel
> par(mfrow=c(1,1))
```



11.6 Phân tích phương sai cho thí nghiệm giai thừa (factorial experiment)

Ví dụ 4. Để khảo sát ảnh hưởng của 4 loại thuốc trừ sâu (1, 2, 3 và 4) và ba loại giống (B1, B2 và B3) đến sản lượng của cam, các nhà nghiên cứu tiến hành một thí nghiệm loại giai thừa. Trong thí nghiệm này, mỗi giống cam có 4 cây cam được chọn một cách ngẫu nhiên, và 4 loại thuốc trừ sâu áp dụng (cũng ngẫu nhiên) cho mỗi cây cam. Kết quả nghiên cứu (sản lượng cam) cho từng giống và thuốc trừ sâu như sau:

Bảng 11.5. Sản lượng cam cho 3 loại giống và 4 loại thuốc trừ sâu

Mô hình phân tích thí nghiệm giai thừa cũng không khác	Giống cam (variety)	Thuốc trừ sâu (pesticide)				Tổng số
		1	2	3	4	
B1		29	50	43	53	175
B2		41	58	42	73	214
B3		66	85	63	85	305
Tổng số		136	193	154	211	694

giờ so với phân tích phương sai hai chiều như trình bày trong phần trên. Cụ thể hơn, mô hình mà chúng ta xem xét là:

$$\text{product} = \alpha + \beta(\text{variety}) + \gamma(\text{pesticide}) + \varepsilon$$

Trong đó, α là hằng số biểu hiện trung bình toàn mẫu, β là hệ số ảnh hưởng của ba giống cam, và γ là hệ số ảnh hưởng của 4 loại thuốc trừ sâu, và ε là phần dư (residual) của mô hình.

Chúng ta có thể sử dụng hàm `aov` của R để ước tính các thông số trên như sau:

```
# trước hết chúng ta nhập số liệu
> variety <- c(1, 1, 1, 1, 2, 2, 2, 2, 3, 3, 3, 3)
> pesticide <- c(1, 2, 3, 4, 1, 2, 3, 4, 1, 2, 3, 4)
> product <- c(29,50,43,53,41,58,42,73,66,85,69,85)

# định nghĩa variety và pesticide là hai yếu tố (factors)
> variety <- as.factor(variety)
> pesticide <- as.factor(pesticide)

# cho vào một data frame tên là data
> data <- data.frame(variety, pesticide, product)

# phân tích phương sai bằng aov và cho vào object analysis
> analysis <- aov(product ~ variety + pesticide)
> anova(analysis)
```

Analysis of Variance Table

```
Response: product
          Df  Sum Sq Mean Sq F value    Pr(>F)
variety     2 2225.17 1112.58  44.063 0.000259 ***
pesticide   3 1191.00  397.00  15.723 0.003008 **
Residuals   6   151.50    25.25
---
Signif. codes:  0 '****' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Kết quả trên cho thấy cả hai yếu tố giống cây (variety) và thuốc trừ sâu (pesticide) đều có ảnh hưởng đến sản lượng cam, vì trị số $p < 0.05$. Để so sánh cụ thể cho từng hai nhóm, chúng ta sử dụng hàm TukeyHSD như sau:

```
> TukeyHSD(analysis)
```

```

Tukey multiple comparisons of means
 95% family-wise confidence level

Fit: aov(formula = product ~ variety + pesticide)

$variety
    diff      lwr      upr     p adj
2-1  9.75 -1.152093 20.65209 0.0749103
3-1 32.50 21.597907 43.40209 0.0002363
3-2 22.75 11.847907 33.65209 0.0016627

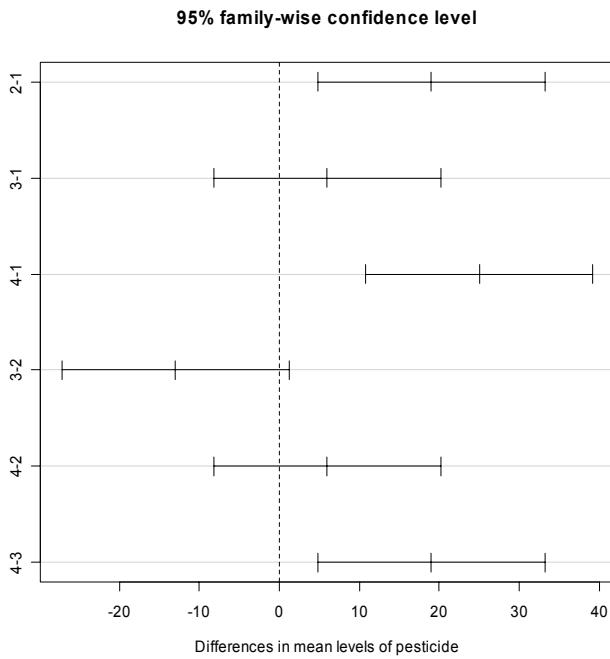
$pesticide
    diff      lwr      upr     p adj
2-1   19  4.797136 33.202864 0.0140509
3-1    6 -8.202864 20.202864 0.5106152
4-1   25 10.797136 39.202864 0.0036109
3-2  -13 -27.202864  1.202864 0.0704233
4-2    6 -8.202864 20.202864 0.5106152
4-3   19  4.797136 33.202864 0.0140509

```

Kết quả phân tích giữa các loại giống cho thấy giống B3 có sản lượng cao hơn giống B1 khoảng 32 đơn vị với khoảng tin cậy 95% từ 21 đến 43 ($p = 0.0002$). Giống cam B3 cũng tốt hơn giống B2, với độ khác biệt trung bình khoảng 22 đơn vị ($p = 0.0017$). Nhưng không có khác biệt đáng kể giữa giống B2 và B1.

So sánh giữa các loại thuốc trừ sâu, kết quả trên cho chúng ta biết các thuốc trừ sâu 4 có hiệu quả cao hơn thuốc 1 và 3. Ngoài ra, thuốc 2 cũng có hiệu quả cao hơn thuốc 1. Còn các so sánh khác không có ý nghĩa thống kê. Biểu đồ Tukey sau đây minh họa cho kết luận trên.

```
> plot(TukeyHSD(analysis), ordered=TRUE)
```



11.7 Phân tích phương sai cho thí nghiệm hình vuông Latin (Latin square experiment)

Ví dụ 5. Để so sánh hiệu quả của 2 loại phân bón (A và B) cùng 2 phương pháp canh tác (a và b), các nhà nghiên cứu tiến hành một thí nghiệm hình vuông Latin. Theo đó, có 4 nhóm can thiệp tổng hợp từ hai loại phân bón và phương pháp canh tác: Aa, Ab, Ba, và Bb (sẽ cho mã số, lần lược, là 1=Aa, 2=Ab, 3=Ba, 4=Bb). Bón phương (treatment) đó được áp dụng trong 4 mẫu ruộng (sample = 1, 2, 3, 4) và 4 loại cây trồng (variety = 1, 2, 3, 4). Tổng cộng, thí nghiệm có $4 \times 4 = 16$ mẫu. Tiêu chí để đánh giá là sản lượng, và kết quả sản lượng được tóm tắt trong bảng sau đây:

Bảng 11.6. Sản lượng cho 2 loại phân bón và 2 phương pháp canh tác

Mẫu ruộng (sample)	Giống (variety)			
	1	2	3	4
1	175	143	128	166
	Aa	Ba	Bb	Ab
2	170	178	140	131
	Ab	Aa	Ba	Bb
3	135	173	169	141
	Bb	Ab	Aa	Ba
4	145	136	165	173
	Ba	Bb	Ab	Aa

Câu hỏi đặt ra là các phương pháp canh tác và phân bón có ảnh hưởng đến sản lượng hay không. Để trả lời câu hỏi đó, chúng ta phải xem xét đến các nguồn làm cho sản lượng thay đổi hay biến thiên. Nhìn qua thí nghiệm và bảng số liệu trên, rất dễ dàng hình dung ra 3 nguồn biến thiên chính:

- Nguồn thứ nhất là khác biệt giữa các phương pháp canh tác và phân bón;
- Nguồn thứ hai là khác biệt giữa các loại giống cây;
- Nguồn thứ ba là khác biệt giữa các mẫu ruộng;

Và phần còn lại là khác biệt trong mỗi mẫu ruộng và loại giống. Để có một cái nhìn chung về số liệu, chúng ta hãy tính trung bình cho từng nhóm qua bảng số sau đây:

Trung bình cho từng loại giống	Trung bình cho từng mẫu	Trung bình cho từng phương pháp
1: 156.25	1: 153.00	1: 173.75
2: 157.50	2: 154.75	2: 168.50
3: 150.50	3: 154.50	3: 142.25
4: 152.75	4: 154.75	4: 132.50
Tổng trung bình: 154.25	Tổng trung bình: 154.25	Tổng trung bình: 154.25

Bảng tóm lược trên cho phép chúng ta tính tổng bình phương cho từng nguồn biến thiên. Khởi đầu là tổng bình phương cho toàn bộ thí nghiệm (tôi sẽ tạm gọi là SStotal):

- Tổng bình phương chung cho toàn thí nghiệm:

$$\begin{aligned} SStotal &= (175 - 154.25)^2 + (143 - 154.25)^2 + \dots + (165 - 154.25)^2 + (173 - 154.25)^2 \\ &= 4941 \end{aligned}$$

- Tổng bình phương do khác biệt giữa các loại giống (SSvariety). Chú ý là vì trung bình mỗi giống được tính từ 4 số, cho nên chúng ta phải nhân cho 4 khi tính tổng bình phương:

$$\begin{aligned} SSvariety &= 4(156.25 - 154.25)^2 + 4(157.50 - 154.25)^2 + \\ &\quad 4(150.50 - 154.25)^2 + 4(152.75 - 154.25)^2 \\ &= 123.5 \end{aligned}$$

Vì có 4 loại giống và một thông số, cho nên bậc tự do là $4-1=3$. Theo đó, trung bình bình phương (mean square) là: $123.5 / 3 = 41.2$.

- Tổng bình phương do khác biệt giữa giống (SSsample). Chú ý là vì trung bình mỗi mẫu được tính từ 4 số, cho nên khi tính tổng bình phương, cần phải nhân cho 4:

$$SSsample = 4(153.00 - 154.25)^2 + 4(154.75 - 154.25)^2 +$$

$$= 4(154.50 - 154.25)^2 + 4(154.75 - 154.25)^2 \\ = 8.5$$

Vì có 4 mẫu và một thông số, cho nên bậc tự do là $4-1=3$, và theo đó trung bình bình phương là: $8.5 / 3 = 2.8$.

- Tổng bình phương do khác biệt giữa các phương pháp (SSmethod). Chú ý là vì trung bình mỗi phương pháp được tính từ 4 số, cho nên khi tính tổng bình phương, cần phải nhân cho 4:

$$\text{SSsample} = 4(173.75 - 154.25)^2 + 4(168.50 - 154.25)^2 + \\ 4(142.25 - 154.25)^2 + 4(132.50 - 154.25)^2 \\ = 4801.50$$

Vì có 4 phương pháp và một thông số, cho nên bậc tự do là $4-1=3$, và theo đó trung bình bình phương là: $4801.5 / 3 = 1600.5$.

- Tổng bình phương phần dư (residual sum of squares):

$$\text{SSresidual} = \text{SStotal} - \text{SSmethod} - \text{SSsample} - \text{SSvariety} \\ = 4941.0 - 4801.5 - 8.5 - 123.5 \\ = 7.5$$

Những ước tính trên đây có thể trình bày trong một bảng phân tích phương sai như sau:

Nguồn biến thiên	Bậc tự do (degrees of freedom)	Tổng bình phương (Sum of squares)	Trung bình bình phương (Mean square)	Kiểm định F
Giữa 4 mẫu ruộng	3	8.5	2.8	2.3
Giữa 4 loại giống	3	123.5	41.2	32.9
Giữa 4 phương pháp	3	4801.5	1600.5	1280.4
Phần dư (residual)	6	7.5		
Tổng số	16	4941.0		

Qua phân tích thủ công và đơn giản trên, chúng ta dễ dàng thấy phương pháp canh tác và loại giống có ảnh hưởng lớn đến sản lượng. Để tính toán chính xác trị số p, chúng ta có thể sử dụng R để tiến hành phân tích phương sai cho thí nghiệm hình vuông Latin.

Vấn đề tổ chức số liệu sao cho thích hợp để R có thể tính toán rất quan trọng. Nói một cách ngắn gọn, mỗi số liệu phải là một số đặc thù (unique), hiểu theo nghĩa nó có một “căn cước” độc nhất vô nhị. Trong thí nghiệm trên, chúng ta có 4 loại giống, 4 mẫu, cho nên tổng số là 16 số liệu. Và, 16 số liệu này phải được định nghĩa cho từng loại giống, từng mẫu, và quan trọng hơn là cho từng phương pháp canh tác. Chẳng hạn như,

trong ví dụ bảng số liệu 10.6 trên, 175 là sản lượng của phương pháp canh tác 1 (tức Aa), loại giống 1, và mẫu 1; nhưng 173 (số ở góc mặc cuối bảng) là sản lượng của phương pháp canh tác 1, nhưng từ loại giống 4, và mẫu 4; v.v...

- Trước hết, chúng ta nhập số liệu sản lượng, và gọi đó là y:

```
> y <- c(175, 143, 128, 166,
       170, 178, 140, 131,
       135, 173, 169, 141,
       145, 136, 165, 173)
```

- Ké đến, gọi variety là giống gồm 4 bậc (1,2,3,4) cho từng số liệu trong y (và cũng định nghĩa rằng variety là một factor, tức biến thứ bậc):

```
> variety <- c(1,2,3,4,
      1,2,3,4,
      1,2,3,4,
      1,2,3,4,
      1,2,3,4)
> variety <- as.factor(variety)
```

- Gọi sample là mẫu gồm 4 bậc (1,2,3,4) cho từng số liệu trong y (và cũng định nghĩa rằng sample là một factor, tức biến thứ bậc):

```
> sample <- c(1,1,1,1,
      2,2,2,2,
      3,3,3,3,
      4,4,4,4)
> sample <- as.factor(sample)
```

- Nhập số liệu cho phương pháp, method, cũng gồm 4 bậc (1,2,3,4) cho từng số liệu trong y (và cũng định nghĩa rằng method là một factor, tức biến thứ bậc):

```
> method <- c(1, 3, 4, 2,
      2, 1, 3, 4,
      4, 2, 1, 3,
      3, 4, 2, 1)
> method <- as.factor(method)
```

- Tổng hợp tất cả các số liệu trên vào một data frame và gọi là data:

```
> data <- data.frame(sample, variety, method, y)
```

- In ra data để kiểm tra xem số liệu có đúng và thích hợp hay chưa:

```
> data
   sample variety method  y
1       1        1      1 175
2       1        2      3 143
3       1        3      4 128
4       1        4      2 166
5       2        1      2 170
```

6	2	2	1	178
7	2	3	3	140
8	2	4	4	131
9	3	1	4	135
10	3	2	2	173
11	3	3	1	169
12	3	4	3	141
13	4	1	3	145
14	4	2	4	136
15	4	3	2	165
16	4	4	1	173

Bây giờ chúng ta đã sẵn sàng dùng hàm lm hay aov để phân tích số liệu. Ở đây tôi sẽ sử dụng hàm aov để tính các nguồn biến thiên trên (kết quả tính toán sẽ chứa trong đối tượng latin):

```
> latin <- aov(y ~ sample + variety + method)
> summary(latin)
      Df Sum Sq Mean Sq   F value    Pr(>F)
sample       3     8.5     2.8    2.2667 0.1810039
variety      3 123.5    41.2   32.9333 0.0004016 ***
method       3 4801.5   1600.5 1280.4000 8.293e-09 ***
Residuals    6     7.5     1.3
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Tất cả các kết quả này (dĩ nhiên) là những kết quả mà chúng ta đã tóm tắt trong bảng phân tích phương sai một cách “thủ công” trên đây. Tuy nhiên, ở đây R cung cấp cho chúng ta trị số p (trong Pr > F) để có thể suy luận thống kê. Và, qua trị số p, chúng ta có thể phát biểu rằng mẫu ruộng không có ảnh hưởng đến sản lượng, nhưng loại giống và phương pháp canh tác thì có ảnh hưởng đến sản lượng.

Để biết mức độ khác biệt giữa các phương pháp canh tác và giữa các loại giống, chúng ta dùng hàm TukeyHSD như sau:

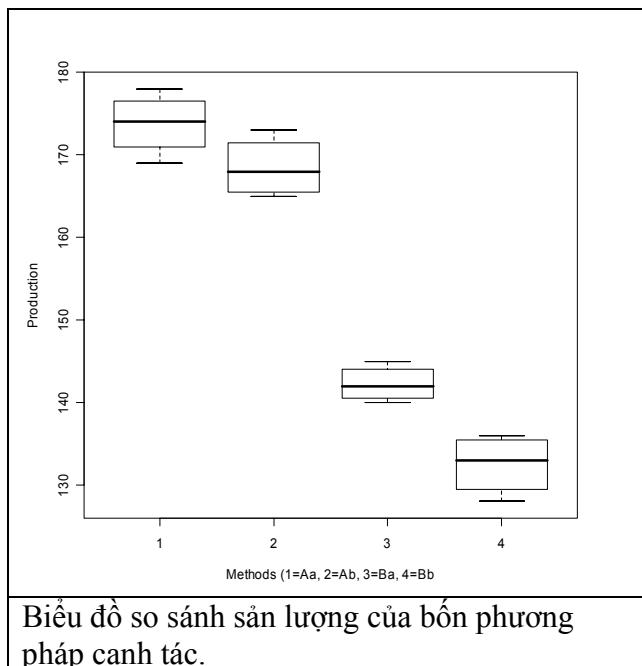
```
> TukeyHSD(latin)
$variety
  diff      lwr      upr      p adj
2-1  1.25 -1.4867231  3.9867231 0.4528549
3-1 -5.75 -8.4867231 -3.0132769 0.0014152
4-1 -3.50 -6.2367231 -0.7632769 0.0173206
3-2 -7.00 -9.7367231 -4.2632769 0.0004803
4-2 -4.75 -7.4867231 -2.0132769 0.0038827
4-3  2.25 -0.4867231  4.9867231 0.1034761

$method
  diff      lwr      upr      p adj
2-1 -5.25 -7.986723  -2.513277 0.0023016
3-1 -31.50 -34.236723 -28.763277 0.0000001
4-1 -41.25 -43.986723 -38.513277 0.0000000
3-2 -26.25 -28.986723 -23.513277 0.0000004
4-2 -36.00 -38.736723 -33.263277 0.0000000
4-3 -9.75 -12.486723 -7.013277 0.0000730
```

So sánh giữa các loại giống cho thấy có sự khác biệt giữa giống 3 và 1, 4 và 1, 3 và 2, 4 và 2.

Tất cả các so sánh giữa các phương pháp canh tác đều có ý nghĩa thống kê. Nhưng loại nào có sản lượng cao nhất? Để trả lời câu hỏi này, chúng ta sẽ sử dụng biểu đồ hộp:

```
> boxplot(y ~ method, xlab="Methods (1=Aa, 2=Ab, 3=Ba, 4=Bb",  
ylab="Production")
```



11.8 Phân tích phương sai cho thí nghiệm giao chéo (cross-over experiment)

Ví dụ 6. Để thử nghiệm hiệu ứng của một thuốc mới đối với chứng ra mồ hôi (thuốc này được bào chế để chữa trị bệnh tim, nhưng ra mồ hôi là một ảnh hưởng phụ), các nhà nghiên cứu tiến hành một nghiên cứu trên 16 bệnh nhân. Số bệnh nhân này được chia thành 2 nhóm (tạm gọi là nhóm AB và BA) một cách ngẫu nhiên. Mỗi nhóm gồm 8 bệnh nhân. Bệnh nhân được theo dõi hai lần: tháng thứ nhất và tháng thứ 2. Đối với bệnh nhân nhóm AB, tháng thứ nhất họ được điều trị bằng thuốc, tháng thứ hai họ được cho sử dụng giả dược (placebo). Ngược lại, với bệnh nhân nhóm BA, tháng thứ nhất sử dụng giả dược, và tháng thứ hai được điều trị bằng thuốc. Tiêu chí để đánh giá là thời gian ra mồ hôi trên trán (tính từ lúc uống thuốc đến khi ra mồ hôi) sau khi sử dụng thuốc hay giả dược. Kết quả nghiên cứu được trình bày trong bảng số liệu sau đây:

Bảng 11.7. Kết quả nghiên cứu hiệu ứng ra mồ hôi của thuốc điều trị bệnh tim

Nhóm	Thời gian (phút) ra mồ hôi trên trán
------	--------------------------------------

	Mã số bệnh nhân số (id)	Tháng 1	Tháng 2
AB		A	Placebo
	1	6	4
3	8	7	
5	12	6	
6	7	8	
9	9	10	
10	6	4	
13	11	6	
15	8	8	
BA		Placebo	A
	2	5	7
4	9	6	
7	7	11	
8	4	7	
11	9	8	
12	5	4	
14	8	9	
16	9	13	

Câu hỏi chính là có sự khác biệt về thời gian ra mồ hôi giữa hai nhóm điều trị bằng thuốc và giả dược hay không.

Để trả lời câu hỏi trên, chúng ta cần tiến hành phân tích phương sai. Nhưng vì cách thiết kế nghiên cứu khá đặc biệt (hai nhóm bệnh nhân với cách sắp xếp can thiệp theo hai thứ tự khác nhau), nên các phương pháp phân tích trên không thể áp dụng được. Có một phương pháp thông dụng là phân tích phương sai trong từng nhóm, rồi sau đó so sánh giữa hai nhóm. Một trong những vấn đề chúng ta cần phải lưu ý là khả năng hiệu ứng kéo dài (còn gọi là carry-over effect), tức là trong nhóm AB, hiệu quả của tháng thứ 2 có thể chịu ảnh hưởng kéo dài từ tháng thứ nhất khi bệnh được điều trị bằng thuốc thật. Trước hết, chúng ta thử tóm lược dữ liệu bằng bảng sau đây:

Bảng 11.8. Tóm lược kết quả thí nghiệm hiệu ứng ra mồ hôi của thuốc điều trị bệnh tim

Nhóm	Mã số bệnh nhân số (id)	Thời gian (phút) ra mồ hôi trên trán		Trung bình cho từng bệnh nhân
		Tháng 1	Tháng 2	
AB		A	Placebo	
	1	6	4	5.0
	3	8	7	7.5
	5	12	6	9.0
	6	7	8	7.5
	9	9	10	9.5
	10	6	4	5.0
	13	11	6	8.5
	15	8	8	8.0

	Trung bình	8.375	6.625	7.50
BA		Placebo	A	
2	5	7	6.0	
4	9	6	7.5	
7	7	11	9.0	
8	4	7	5.5	
11	9	8	8.5	
12	5	4	4.5	
14	8	9	8.5	
16	9	13	11.0	
Trung bình	7.000	8.125	7.5625	
Trung bình cho 2 nhóm	7.6875	7.3750	7.5312	

Trung bình cho nhóm A = $(8.375 + 8.125) / 2 = 8.25$

Trung bình cho nhóm P (giả dược) = $(6.625 + 7.000) / 2 = 6.8125$

Qua bảng tóm lược trên, chúng ta có thể tính toán một số tổng bình phương:

- Tổng bình phương do khác biệt giữa hai nhóm điều trị bằng thuốc và giả dược:

$$SSTreat = 16(8.25 - 7.5312)^2 + 16(8.8125 - 7.5312)^2 = 16.53$$

- Tổng bình phương do khác biệt giữa tháng 1 và tháng 2:

$$SSPeriod = 16(7.6875 - 7.5312)^2 + 16(7.3750 - 7.5312)^2 = 0.781$$

- Tổng bình phương do khác biệt giữa hai nhóm AB và BA (thứ tự):

$$SSseq = 16(7.50 - 7.5312)^2 + 16(7.5625 - 7.5312)^2 = 0.031$$

- Tổng bình phương do khác biệt giữa các bệnh nhân trong cùng nhóm AB hay BA:

$$\begin{aligned} SSw &= (5.0 - 7.50)^2 + (7.5 - 7.50)^2 + (9.0 - 7.50)^2 + \dots + (8.0 - 7.50)^2 + \\ &\quad (6.0 - 7.5625)^2 + (7.5 - 7.5625)^2 + (9.0 - 7.5625)^2 + \dots + (11.0 - 7.5625)^2 \\ &= 103.44 \end{aligned}$$

- Tổng bình phương cho toàn bộ mẫu:

$$\begin{aligned} SStotal &= (6 - 7.5312)^2 + (9 - 7.5312)^2 + \dots + (13 - 7.5312)^2 + (9 - 7.5312)^2 \\ &= 167.97 \end{aligned}$$

- Tổng bình phương còn lại (tức phần dư):

$$SSres = 167.97 - 16.53 - 0.781 - 0.031 - 103.44 = 47.19$$

Đến đây, chúng ta có thể lập bảng phân tích phương sai như sau:

Bảng 11.9. Kết quả phân tích phương sai số liệu trong bảng 11.7

Nguồn biến thiên	Bậc tự do (degrees of freedom)	Tổng bình phương (Sum of squares)	Trung bình bình phương (Mean square)	Kiểm định F
Giữa hai nhóm điều trị	1	16.53	16.53	4.90
Giữa hai tháng	1	0.781	0.781	0.23
Giữa AB và BA	1	0.031	0.031	0.004
Trong mỗi nhóm	14	103.44	7.39	
Phản dư (residual)	14	47.19	3.37	
Tổng số	31	167.97		

Qua phân tích trên, chúng ta thấy độ khác biệt giữa thuốc và giả dược lớn hơn là độ khác biệt giữa hai tháng hay hai nhóm AB và BA. Kiểm định F để thử nghiệm giả thiết thuốc và giả dược có hiệu quả như nhau là kiểm định $F = 16.53 / 3.37 = 4.90$ với bậc tự do 1 và 14. Dựa trên lí thuyết xác suất, trị số F với bậc tự do 1 và 14 là 4.60. Do đó, chúng ta có thể kết luận rằng thuốc này có hiệu ứng làm ra mồ hôi lâu hơn nhóm giả dược.

Tất cả các tính toán “thủ công” trên chỉ là minh họa cho cách phân tích phương sai cho thí nghiệm giao chéo. Trong thực tế, chúng ta có thể sử dụng R để tiến hành các tính toán đó như cách tính phương sai cho các thí nghiệm đơn giản. Vấn đề chính là tổ chức số liệu cho phân tích. R (cũng như nhiều phần mềm khác) yêu cầu người sử dụng phải nhập từng số liệu một, và **mỗi số liệu phải gắn liền với một bệnh nhân, một nhóm điều trị, một tháng (hay giai đoạn), và một nhóm thứ tự**. Đó là một yêu cầu rất quan trọng, vì nếu tổ chức số liệu không đúng, kết quả phân tích có thể sai.

Trong phần sau đây, tôi sẽ mô tả từng bước một:

bước 1: nhập dữ liệu và đặt tên object là y

```
> y <- c(6,8,12,7,9,6,11,8,
      4,7,6,8,10,4,6,8,
      5,9,7,4,9,5,8,9
      7,6,11,7,8,4,9,13)
```

bước 2: cứ mỗi số liệu trong bước 1, chỉ ra nhóm AB hay BA (mã số 1 và 2)

```
> seq <- c(1,1,1,1,1,1,1,1,
      1,1,1,1,1,1,1,1,
      2,2,2,2,2,2,2,2,
      2,2,2,2,2,2,2,2)
> seq <- as.factor(seq)
```

```

# bước 3: cùm mỗi số liệu trong bước 1, chỉ ra tháng 1 hay tháng 2

> period <- c(1,1,1,1,1,1,1,1,
   2,2,2,2,2,2,2,2,
   2,2,2,2,2,2,2,2,
   1,1,1,1,1,1,1,1)
> period <- as.factor(period)

# bước 4: cùm mỗi số liệu trong bước 1, chỉ ra nhóm A hay placebo
bằng mã số 1 và 2:

> treat <- c(1,1,1,1,1,1,1,1,
   2,2,2,2,2,2,2,2,
   1,1,1,1,1,1,1,1,
   2,2,2,2,2,2,2,2)
> treat <- as.factor(treat)

# bước 5: cùm mỗi số liệu trong bước 1, chỉ ra mã số cho từng bệnh
nhân

> id <- c(1,3,5,6,9,10,13,15,
   1,3,5,6,9,10,13,15,
   2,4,7,8,11,12,14,16,
   2,4,7,8,11,12,14,16)
> id <- as.factor(id)

# bước 6: lập thành một data frame tên là data và in ra để kiểm
tra một lần nữa.

> data <- data.frame(seq, period, treat, id, y)
> data
  seq period treat id  y
1    1      1     1  1  6
2    1      1     1  3  8
3    1      1     1  5 12
4    1      1     1  6  7
5    1      1     1  9  9
6    1      1     1 10  6
7    1      1     1 13 11
8    1      1     1 15  8
9    1      2     2  1  4
10   1      2     2  3  7
11   1      2     2  5  6
12   1      2     2  6  8
13   1      2     2  9 10
14   1      2     2 10  4
15   1      2     2 13  6
16   1      2     2 15  8
17   2      2     1  2  7
18   2      2     1  4  6
19   2      2     1  7 11
20   2      2     1  8  7
21   2      2     1 11  8

```

```

22   2      2      1 12   4
23   2      2      1 14   9
24   2      2      1 16  13
25   2      1      2   2   5
26   2      1      2   4   9
27   2      1      2   7   7
28   2      1      2   8   4
29   2      1      2 11   9
30   2      1      2 12   5
31   2      1      2 14   8
32   2      1      2 16   9

```

Bây giờ chúng ta đã sẵn sàng dùng hàm `lm` của R để phân tích số liệu. Chú ý rằng cách dùng hàm `lm` cho phân tích phương sai áp dụng cho thí nghiệm giao chéo hoàn toàn không khác gì với cách dùng cho các thí nghiệm khác. Khía cạnh khác biệt duy nhất là cách tổ chức dữ liệu cho phân tích như trình bày trên.

```

> xover <- lm(y ~ treat+seq+period)
> anova(xover)

```

Analysis of Variance Table

```

Response: y
          Df  Sum Sq Mean Sq F value    Pr(>F)
treat      1 16.531 16.531  4.9046 0.04388 *
seq        1  0.031  0.031  0.0093 0.92466
period     1  0.781  0.781  0.2318 0.63764
id         14 103.438  7.388  2.1921 0.07711 .
Residuals 14  47.187  3.371
---
Signif. codes:  0 '****' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Phân tích trên đây một lần nữa khẳng định cách tính thủ công mà tôi đã trình bày phần trên. Nói tóm lại, mức độ khác biệt giữa thuốc và giả được có ý nghĩa thống kê, với trị số F là 0.044.

Chúng ta cũng có thể yêu cầu khoảng tin cậy 95% cho độ khác biệt giữa hai nhóm (bằng cách lệnh `TukeyHSD`) như sau (chú ý là với `TukeyHSD` chúng ta chỉ sử dụng hàm `aov` chứ không phải `lm`):

```

> TukeyHSD(aov(y ~ treat+seq+period+id))
Tukey multiple comparisons of means
 95% family-wise confidence level

Fit: aov(formula = y ~ treat + seq + period + id)

$treat
       diff      lwr      upr      p adj
2-1 -1.4375 -2.829658 -0.04534186 0.0438783

$seq

```

```

diff      lwr      upr      p adj
2-1 0.0625 -1.329658 1.454658 0.924656

$period
diff      lwr      upr      p adj
2-1 -0.3125 -1.704658 1.079658 0.6376395

```

Chú ý kết quả:

```

$treat
diff      lwr      upr      p adj
2-1 -1.4375 -2.829658 -0.04534186 0.0438783

```

cho biết tính trung bình thời gian ra mồ hôi của nhóm được điều trị cao hơn nhóm giả dược khoảng 1.44 phút, và khoảng tin cậy 95% là từ 0.05 phút đến 2.8 phút. Còn các kết quả so sánh giữa hai nhóm AB và BA (seq) hay giữa tháng 1 và tháng 2 (period) không có ý nghĩa thống kê.

11.9 Phân tích phương sai cho thí nghiệm tái đo lường (repeated measure experiment)

Ví dụ 7. Một nghiên cứu sơ khởi (pilot study) được tiến hành để đánh giá hiệu nghiệm của một vắc-xin mới chống bệnh thấp khớp. Nghiên cứu gồm 8 bệnh nhân, được chia thành 2 nhóm một cách ngẫu nhiên. Nhóm 1 gồm 4 bệnh nhân được điều trị bằng vắc-xin; nhóm 2 cũng gồm 4 bệnh nhân nhưng được nhận giả dược (placebo, hay đối chứng). Bệnh nhân được theo dõi trong 3 tháng, và cứ mỗi tháng, bệnh nhân được hỏi về tình trạng của bệnh ra sao. Tình trạng bệnh được “đo lường” bằng một chỉ số có giá trị từ 0 (không có hiệu nghiệm, bệnh vẫn như trước) đến 10 (có hiệu nghiệm tuyệt đối, hết bệnh). Kết quả nghiên cứu có thể tóm tắt trong bảng số liệu sau đây:

Bảng 11.10. Kết quả nghiên cứu vắc-xin chống đau thấp khớp

Nhóm	Mã số bệnh nhân số (id)	Chỉ số bệnh qua từng tháng		
		Tháng 1	Tháng 2	Tháng 3
Vắc-xin				
	1	6	3	0
	2	7	3	1
	3	4	1	2
	4	8	4	3
Placebo				
	5	6	5	5
	6	9	4	6
	7	5	3	4
	8	6	2	3

Câu hỏi chính là có sự khác biệt nào giữa hai nhóm vắc-xin và giả dược hay không.

Để đơn giản hóa cách phân tích phương sai cho thí nghiệm tái đo lường, tôi sẽ tránh dùng kí hiệu toán, mà chỉ minh họa bằng vài phép tính “thủ công” để bạn đọc có thể theo dõi. Trước hết, chúng ta cần phải tóm lược số liệu bằng cách tính trung bình cho mỗi bệnh nhân, mỗi nhóm điều trị, và mỗi tháng như sau:

Bảng 11.11. Tóm lược số liệu nghiên cứu vắc-xin chống đau thấp khớp

Nhóm điều trị	id	Chỉ số bệnh qua từng tháng			Trung bình
		1	2	3	
Vắc-xin	1	6	3	0	3.000
	2	7	3	1	3.667
	3	4	1	2	2.333
	4	8	4	3	5.000
	Trung bình	6.25	2.75	1.50	3.500
	SD	1.71	1.26	1.29	
Placebo	5	6	5	5	5.333
	6	9	4	6	6.333
	7	5	3	4	4.000
	8	6	2	3	3.667
	Trung bình	6.50	3.50	4.50	4.833
	SD	1.73	1.29	1.29	
Trung bình cho hai nhóm		6.375	3.125	3.000	4.167

Qua bảng trên, chúng ta có thể thấy ngay rằng có 5 nguồn làm cho kết quả thí nghiệm khác nhau:

- (a) giữa vắc-xin và giả dược (có lẽ là nguồn mà chúng ta cần biết!);
 - (b) giữa 3 tháng theo dõi;
 - (c) giữa mỗi ba tháng trong mỗi nhóm điều trị, mà giới thống kê thường đề cập đến là “interaction” (tương tác), và trong trường hợp này, tương tác giữa nhóm điều trị và thời gian;
 - (d) giữa các bệnh nhân trong cùng một nhóm điều trị;
 - (e) và sau cùng là phần dư, tức phần mà chúng ta không thể “giải thích” sau khi xem xét các nguồn (a) đến (d) trên.
- Trước hết là tổng bình phương giữa hai nhóm điều trị (vắc-xin và giả dược), tôi sẽ gọi là *SStreat*:

$$\begin{aligned} SStreat &= 12(3.500 - 4.167)^2 + \\ &\quad 12(4.833 - 4.167)^2 = 10.667 \end{aligned}$$

- Kế đến là tổng bình phương giữa 3 tháng điều trị, tôi sẽ gọi là *SStime*:

$$\begin{aligned} SStime &= 8(6.375 - 4.167)^2 + \\ &\quad 8(3.125 - 4.167)^2 + \\ &\quad 8(3.000 - 4.167)^2 = 58.583 \end{aligned}$$

- Nguồn thứ ba là tổng bình phương do tương tác giữa điều trị và thời gian, tôi sẽ gọi là SS_{int}

$$\begin{aligned}
 SS_{int} &= 4(6.25 - 4.167)^2 + \\
 &\quad 4(2.75 - 4.167)^2 + \\
 &\quad 4(1.50 - 4.167)^2 + \\
 &\quad 4(6.50 - 4.167)^2 + \\
 &\quad 4(3.50 - 4.167)^2 + \\
 &\quad 4(4.50 - 4.167)^2 - \\
 &\quad SS_{\text{vácxin}} - SS_{\text{time}} \\
 &= 77.833 - 10.667 - 58.583 \\
 &= 8.583
 \end{aligned}$$

- Nguồn thứ tư là tổng bình phương do tương tác giữa bệnh nhân trong mỗi nhóm điều trị, tôi sẽ gọi là $SS_{patient(treat)}$:

$$\begin{aligned}
 SS_{patient(treat)} &= 3(3.000 - 3.350)^2 + 3(3.667 - 3.350)^2 + 3(2.333 - 3.350)^2 + 3(5.000 - 3.350)^2 + \\
 &\quad 3(5.333 - 4.833)^2 + 3(6.333 - 4.833)^2 + 3(4.000 - 4.833)^2 + 3(3.667 - 4.833)^2 \\
 &= 25.333
 \end{aligned}$$

- Ngoài ra, tổng bình phương cho toàn mẫu là:

$$SS_{total} = (6-4.167)^2 + (3-4.167)^2 + (0-4.167)^2 + \dots + (3-4.167)^2 = 115.333$$

- Từ đó, chúng ta có thể ước tính tổng bình phương cho phần dư:

$$\begin{aligned}
 SSE &= SS_{total} - SS_{\text{vácxin}} - SS_{\text{time}} - SS_{patient(\text{vácxin})} - SS_{\text{vácxin-time}} \\
 &= 115.333 - 10.667 - 58.583 - 25.333 - 8.583 \\
 &= 12.167
 \end{aligned}$$

Đến đây, chúng ta có thể lập bảng phân tích phương sai như sau:

Nguồn biến thiên	Bậc tự do (degrees of freedom)	Tổng bình phương (Sum of squares)	Trung bình bình phương (Mean square)	Kiểm định F
Giữa vácxin và placebo	1	10.667	10.667	2.53
Bệnh nhân (nhóm điều trị)	6	25.333	4.222	-
Giữa 3 tháng	2	58.583	29.292	28.89
Thời gian và nhóm điều trị	2	8.583	4.292	4.23
Phần dư (residual)	12	12.167	1.014	-
Tổng số	23	115.333		

Tất cả các tính toán thủ công trên, như bạn đọc có thể thấy, khá rườm rà, và rất dễ sai sót. Nhưng trong R, chúng ta có thể có kết quả trong vòng 1 giây, sau khi số liệu đã được sắp xếp một cách thích hợp. Sau đây, tôi sẽ trình bày cách phân tích phương sai tái do lường bằng R:

- Trước hết, chúng ta nhập dữ liệu cho từng bệnh nhân. Cũng như bất cứ phần mềm thống kê nào, mỗi giá trị phải được kèm theo những biến số đặc trưng như cho mỗi bệnh nhân, mỗi nhóm, và mỗi thời gian:

```
y <- c(6, 7, 4, 8,
      3, 3, 1, 4,
      0, 1, 2, 3,
      6, 9, 5, 6,
      5, 4, 3, 2,
      5, 6, 4, 3)
```

- Trong mỗi số liệu trên, cho R biết thuộc nhóm điều trị (mã số 1) hay giả dược (mã số 2). Cũng nên cho R biết `treat` là một biến thứ bậc (categorical variable) chứ không phải biến số (numerical variable):

```
treat <- c(1, 1, 1, 1,
          1, 1, 1, 1,
          1, 1, 1, 1,
          2, 2, 2, 2,
          2, 2, 2, 2,
          2, 2, 2, 2)
treat <- as.factor(treat)
```

- Trong mỗi số liệu trên, cho R biết thuộc tháng nào (mã số 1, 2, 3), và định nghĩa `time` là một biến thứ bậc.

```
time <- c(1, 1, 1, 1,
         2, 2, 2, 2,
         3, 3, 3, 3,
         1, 1, 1, 1,
         2, 2, 2, 2,
         3, 3, 3, 3)
time <- as.factor(time)
```

- Trong mỗi số liệu trên, cho R biết thuộc bệnh nhân nào (mã số 1, 2, 3, ..., 8), và định nghĩa `id` là một biến thứ bậc.

```
id <- c(1, 2, 3, 4, 1, 2, 3, 4, 1, 2, 3, 4,
       5, 6, 7, 8, 5, 6, 7, 8, 5, 6, 7, 8)
id <- as.factor(id)
```

- Nhập tất cả biến vào một data frame và đặt tên là, chẳng hạn như, `data`. Kiểm tra một lần nữa xem số liệu đã đúng với ý định sắp xếp hay chưa. Xin nhắc lại, trước khi phân tích số liệu, việc quan trọng là phải kiểm tra lại cho thật kĩ số liệu để đảm bảo số liệu đã được tổ chức đúng và thích hợp.

```

data <- data.frame(id, time, treat, y)
data
  id time treat y
1   1     1    1 6
2   2     1    1 7
3   3     1    1 4
4   4     1    1 8
5   1     2    1 3
6   2     2    1 3
7   3     2    1 1
8   4     2    1 4
9   1     3    1 0
10  2     3    1 1
11  3     3    1 2
12  4     3    1 3
13  5     1    2 6
14  6     1    2 9
15  7     1    2 5
16  8     1    2 6
17  5     2    2 5
18  6     2    2 4
19  7     2    2 3
20  8     2    2 2
21  5     3    2 5
22  6     3    2 6
23  7     3    2 4
24  8     3    2 3

```

Bây giờ, chúng ta đã sẵn sàng sử dụng R để phân tích. Hàm chính để phân tích phương sai là `aov` (analysis of variance). Trong hàm này, chú ý cách cung cấp thông số bằng cách dùng một hàm khác có tên là `Error`. Trong hàm `Error`, chúng ta cho R biết rằng mỗi bệnh nhân (`id`) “thuộc” vào một nhóm điều trị và do đó thuộc vào biến `time`. Cách để cho R biết là: `Error(id/time)`. Cụ thể hơn:

```
> repeated <- aov(y ~ treat*time + Error(id/time))
```

Lệnh trên đây yêu cầu R phân tích theo mô hình: $y = \text{treat} + \text{time} + \text{treat} * \text{time}$ (chú ý `treat*time` tương đương với `treat+time+treat*time`), và trung bình bình phương phần dư phải được tách thành hai phần: một phần trong các bệnh nhân, và một phần giữa các tháng điều trị (viết tắt bằng kí hiệu `id/time`). Tất cả kết quả cho vào đó itượng có tên là `repeated`. Chúng ta yêu cầu một bảng tóm lược kết quả từ đối tượng `repeated`:

```
> summary(repeated)
```

```
Error: id
  Df Sum Sq Mean Sq F value Pr(>F)
treat      1 10.6667 10.6667  2.5263 0.1631
Residuals  6 25.3333  4.2222
```

```
Error: id:time
  Df Sum Sq Mean Sq F value     Pr(>F)
```

```

time      2 58.583 29.292 28.8904 2.586e-05 ***
treat:time 2   8.583    4.292   4.2329    0.04064 *
Residuals 12 12.167    1.014
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Kết quả phân tích trong phần đầu của bảng trên cho thấy sự khác biệt giữa nhóm điều trị bằng thuốc và giả dược không có ý nghĩa thống kê ($p = 0.16$). Như vậy chúng ta có thể kết luận thuốc không có hiệu nghiệm giảm đau thấp khớp?

Câu trả lời là “không”, bởi vì phần thứ hai của bảng phân tích phương sai cho thấy mối tương tác giữa treat và time (trị số $p = 0.041$). Điều này có nghĩa là độ khác biệt giữa thuốc và giả dược tùy thuộc vào tháng điều trị. Thật vậy, nếu chúng ta xem lại bảng 10.11 sẽ thấy trong tháng 1, trung bình của nhóm vắc-xin và giả dược không mấy khác nhau (6.25 và 6.50), nhưng đến tháng thứ 2 và nhất là tháng thứ 3 thì độ khác biệt giữa hai nhóm rất cao (như tháng thứ ba: 1.50 cho vắc-xin và 4.50 cho nhóm giả dược). Như vậy, độ hiệu nghiệm trong nhóm được điều trị tăng dần theo thời gian, còn trong nhóm giả dược thì hầu như không có khác biệt giữa 3 tháng. Nói cách khác và tóm lại, qua thí nghiệm sơ khởi này chúng ta có thể nói vắc-xin có vẻ có hiệu quả giảm đau trong các bệnh nhân thấp khớp.

Trên đây là vài cách sử dụng cho việc phân tích phương sai với các thí nghiệm thông dụng. Thiết kế và phân tích thí nghiệm (experimental design) là một lĩnh vực nghiên cứu tương đối chuyên sâu, những chỉ dẫn trên đây không thể và cũng không có tham vọng mô tả tất cả các phép tính cũng như phương pháp cho tất cả thí nghiệm. Tuy nhiên, trong thực tế, các phương pháp và thí nghiệm rất thường được áp dụng trong khoa học thực nghiệm. R có một package tên là `n1me` (non-linear mixed-effects) cũng có thể sử dụng cho các phân tích trên và các mô hình phức tạp hơn với đa biến và đa thứ bậc. Package này cũng có thể tải về máy miễn phí tại website của R: <http://cran.R-project.org>.

12

Phân tích hồi qui logistic

Trong các chương trước về phân tích hồi qui tuyến tính và phân tích phương sai, chúng ta tìm mô hình và mối liên hệ giữa một biến phụ thuộc liên tục (continuous dependent variable) và một hay nhiều biến độc lập (independent variable) hoặc là liên tục hoặc là không liên tục. Nhưng trong nhiều trường hợp, biến phụ thuộc không phải là biến liên tục mà là biến mang tính đo lường nhị phân: có/không, mắc bệnh/không mắc bệnh, chết/sống, xảy ra/không xảy ra, v.v..., còn các biến độc lập có thể là liên tục hay không liên tục. Chúng ta cũng muốn tìm hiểu mối liên hệ giữa các biến độc lập và biến phụ thuộc.

Ví dụ 1. Trong một nghiên cứu do tôi tiến hành để tìm hiểu mối liên hệ giữa nguy cơ gãy xương (fracture, viết tắt là fx) và mật độ xương cùng một số chỉ số sinh hóa khác, 139 bệnh nhân nam (hay nói đúng hơn là đối tượng nghiên cứu) tuổi từ 60 trở lên. Năm 1990, các số liệu sau đây được thu thập cho mỗi đối tượng: độ tuổi (age), tỉ trọng cơ thể (body mass index hay BMI), mật độ chất khoáng trong xương (bone mineral density hay BMD), chỉ số hủy xương ICTP, chỉ số tạo xương PINP. Các đối tượng nghiên cứu được theo dõi trong vòng 15 năm. Trong thời gian theo dõi, các bệnh nhân bị gãy xương hay không gãy xương được ghi nhận. Câu hỏi đặt ra ban đầu là có một mối liên hệ gì giữa BMD và nguy cơ gãy xương hay không. Số liệu của nghiên cứu này được trình bày trong phần cuối của chương này, và sẽ trình bày một phần dưới đây để bạn đọc nắm được vấn đề.

Bảng 12.1. Một phần số liệu nghiên cứu về các yếu tố nguy cơ cho gãy xương

id	fx	age	bmi	bmd	ictp	pinp
1	1	79	24.7252	0.818	9.170	37.383
2	1	89	25.9909	0.871	7.561	24.685
3	1	70	25.3934	1.358	5.347	40.620
4	1	88	23.2254	0.714	7.354	56.782
5	1	85	24.6097	0.748	6.760	58.358
6	0	68	25.0762	0.935	4.939	67.123
7	0	70	19.8839	1.040	4.321	26.399
8	0	69	25.0593	1.002	4.212	47.515
9	0	74	25.6544	0.987	5.605	26.132
10	0	79	19.9594	0.863	5.204	60.267
...						
137	0	64	38.0762	1.086	5.043	32.835
138	1	80	23.3887	0.875	4.086	23.837
139	0	67	25.9455	0.983	4.328	71.334

Ở đây, vì biến phụ thuộc (gãy xương) không được đo lường theo tính liên tục (mà chỉ là *có* hay *không*), cho nên phương pháp phân tích hồi qui tuyến tính để phân tích mối liên hệ giữa biến phụ thuộc và biến độc lập. Một phương pháp phân tích được phát triển tương đối gần đây (vào thập niên 1970s) có tên là logistic regression analysis (hay phân tích hồi qui logistic) có thể áp dụng cho trường hợp trên.

Trong nghiên cứu này, sau 15 năm theo dõi, có 38 bệnh nhân bị gãy xương. Tính theo phần trăm, tỉ lệ gãy xương là $38 / 139 = 0.273$ (hay 27.3%).

12.1 Mô hình hồi qui logistic

Cho một tần số biến cõi x ghi nhận từ n đối tượng, chúng ta có thể tính xác suất của biến cõi đó là:

$$p = \frac{x}{n}$$

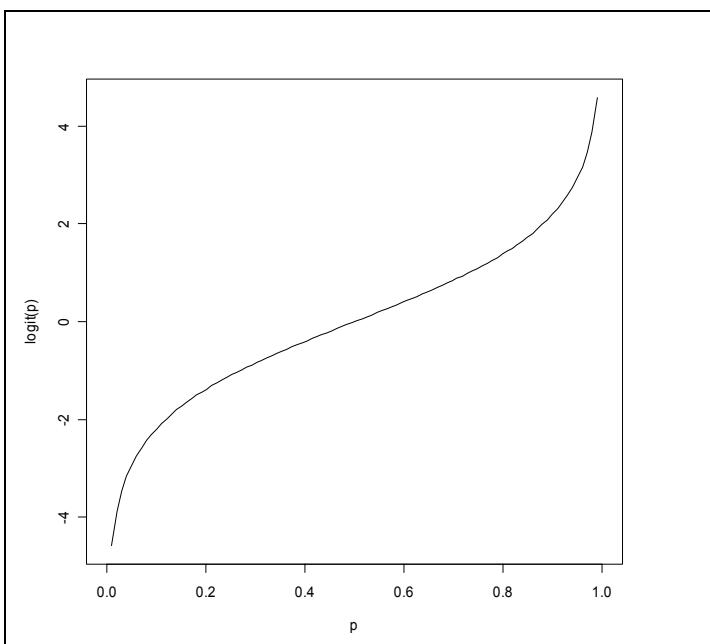
p có thể xem là một chỉ số đo lường nguy cơ của một biến cõi. Một cách thể hiện nguy cơ khác là *odds* (một danh từ, nếu tôi không làm, chỉ có trong tiếng Anh – ngay cả tiếng Pháp, Đức, Tây Ban Nha ... cũng không có danh từ tương đương với *odds*). Tôi tạm dịch *odds* là *khả năng*. Khả năng của một biến cõi được định nghĩa đơn giản bằng tỉ số xác suất biến cõi xảy ra trên xác suất biến cõi không xảy ra:

$$\text{odds} = \frac{p}{1-p} \quad [1]$$

Hàm *logit* của *odds* được định nghĩa như sau:

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right) \quad [2]$$

Mối liên hệ giữa p và $\text{logit}(p)$ là một mối liên hệ liên tục (dĩ nhiên!) và theo dạng như sau:



Biểu đồ 12.1. Mối liên hệ giữa logit(p) và p , cho $0 < p < 1$.

Chú ý: biểu đồ trên được vẽ bằng các lệnh sau đây:

```

p <- seq(0, 1, length=100)
p <- p[2:(length(p)-1)]
logit <- function(t)
{
  log(t / (1-t))
}
plot(logit(p) ~ p, type="l")

```

Cho một biến độc lập x (x có thể là liên tục hay không liên tục), mô hình hồi qui logistic phát biểu rằng:

$$\text{logit}(p) = \alpha + \beta x \quad [3]$$

Tương tự như mô hình hồi qui tuyến tính, α và β là hai thông số tuyến tính cần phải ước tính từ dữ liệu nghiên cứu. Nhưng ý nghĩa của thông số này, đặc biệt là thông số β , rất khác với ý nghĩa mà ta đã quen với mô hình hồi qui tuyến tính. Để hiểu ý nghĩa của hai thông số này, tôi sẽ quay lại với ví dụ 1.

Ví dụ 1 (tiếp theo). Vấn đề mà chúng ta muốn biết là mối liên hệ giữa mật độ xương bmd và nguy cơ gãy xương (fx). Để tiện cho việc minh họa, gọi bmd là x , vấn đề mà chúng ta cần biết có thể viết bằng ngôn ngữ mô hình như sau

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right)\alpha + \beta x \quad [4]$$

Nói cách khác:

$$\text{odds}(p) = \frac{p}{1-p} = e^{\alpha+\beta x}$$

Nói cách khác, mô hình hồi qui logistic vừa trình bày trên phát biểu rằng mối liên hệ giữa xác suất gãy xương (p) và mật độ xương bmd là một mối liên hệ theo hình chữ S. Mô hình trên còn cho thấy xác suất gãy xương p tùy thuộc vào giá trị của x . Thành ra, mô hình trên có thể viết một cách chính xác hơn rằng *khả năng gãy xương* với điều kiện x là:

$$\text{odds}(p | x) = e^{\alpha+\beta x}$$

Khi $x = x_0$, khả năng gãy xương là: $\text{odds}(p | x = x_0) = e^{\alpha+\beta x_0}$

Khi $x = x_0 + 1$ (tức tăng 1 đơn vị từ x_0), khả năng gãy xương là:

$$\text{odds}(p | x = x_0 + 1) = e^{\alpha+\beta(x_0+1)}$$

Và, tỉ số của hai xác suất gãy xương:

$$\frac{\text{odds}(p | x = x_0 + 1)}{\text{odds}(p | x = x_0)} = \frac{e^{\alpha + \beta(x_0 + 1)}}{e^{\alpha + \beta x_0}} = e^\beta \quad [5]$$

Trong dịch tễ học, e^β được gọi là *odds ratio*. *Odds ratio*, như tên gọi là, *tỉ số khả năng* hay *tỉ số khả dĩ*. Nói cách khác, hệ số β trong mô hình hồi qui logistic chính là tỉ số khả dĩ.

Phương pháp để ước tính thông số trong mô hình [3] khá phức tạp (dùng phương pháp maximum likelihood – tức phương pháp *Hợp lí cực đại*) và không nằm trong phạm vi của cuốn sách này, nên tôi sẽ không trình bày ở đây (bạn đọc có thể tham khảo sách giáo khoa để biết thêm, nếu cần thiết). Tuy nhiên, tôi muốn đề cập ngắn gọn là phương pháp hợp lí cực đại cung cấp cho chúng ta một hệ phương trình như sau:

$$\begin{cases} \sum_{i=1}^n y_i = \sum_{i=1}^n \left(1 + e^{-(\hat{\alpha} + \hat{\beta}x_i)}\right)^{-1} \\ \sum_{i=1}^n x_i y_i = \sum_{i=1}^n x_i \left(1 + e^{-(\hat{\alpha} + \hat{\beta}x_i)}\right) \end{cases}$$

Trong đó, Trong đó, y_i là biến phụ thuộc (gãy xương với giá trị 0 hay 1), và x_i là biến độc lập (mật độ xương), và n là số mẫu. Để tìm ước số $\hat{\alpha}$ và $\hat{\beta}$, một trong những phép tính hay sử dụng là iterative weighted least square hay Newton-Raphson. R sử dụng phép tính Newton-Raphson để tìm hai ước số đó.

Sau khi đã có ước số $\hat{\alpha}$ và $\hat{\beta}$ chúng ta có thể ước tính xác suất p cho bất cứ giá trị nào của x như sau (sau vài thao tác đơn giản):

$$\hat{p} = \frac{e^{\hat{\alpha} + \hat{\beta}x}}{1 + e^{\hat{\alpha} + \hat{\beta}x}} = \frac{1}{1 + e^{-(\hat{\alpha} + \hat{\beta}x)}}$$

Chú ý tôi dùng dấu mũ \hat{p} để chỉ số ước tính (predicted value), chứ không phải p là xác suất quan sát. Nếu mô hình mô tả dữ liệu tốt và đầy đủ, độ khác biệt giữa p và \hat{p} nhỏ; nếu mô hình không thích hợp hay không tốt, độ khác biệt đó có thể sẽ cao. Độ khác biệt giữa p và \hat{p} được gọi là *deviance*. Phương pháp tính deviance khá phức tạp, nhưng đó không phải là chủ đề ở đây, cho nên tôi chỉ nói qua khái niệm mà thôi. Khi chúng ta có nhiều mô hình để mô phỏng một hay nhiều mối liên hệ, deviance có thể được sử dụng để đánh giá sự thích hợp của một mô hình, hay để chọn một mô hình “tối ưu”.

12.2 Phân tích hồi qui logistic bằng R

Ví dụ 1 (tiếp theo). Bây giờ, chúng ta quay lại với ví dụ 1, dùng số liệu trong Bảng 12.1 để ước tính hai thông số α và β bằng R. Trước hết chúng ta phải nhập toàn bộ số liệu vào một data frame, và cho một cái tên, chẳng hạn như fracture. Trong trường hợp của tôi, dữ liệu được chứa trong directory c:\works\stats dưới tên fracture.txt, do đó, các lệnh sau đây cần thiết để nhập số liệu:

```
# báo cho R biết nơi chứa số liệu
> setwd("c:/works/stats")

# nhập số liệu và cho vào một data frame tên fracture
> fracture <- read.table("fracture.txt", header=TRUE, na.string=".")

# kiểm tra xem có bao nhiêu biến trong dữ liệu fracture
> names(fracture)
[1] "id"    "fx"    "age"   "bmi"   "bmd"   "ictp"  "pinp"

# Chọn những bệnh nhân có đầy đủ số liệu cho phân tích
> fulldata <- na.omit(fracture)
> attach(fulldata)
```

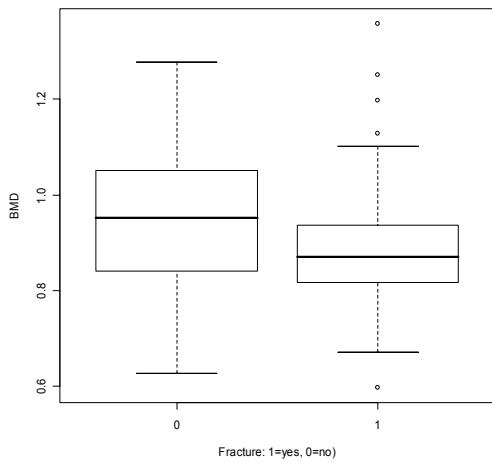
Hai biến mà chúng ta quan tâm trong ví dụ này là: fx (gãy xương) và bmd (mật độ xương). Chúng ta kiểm tra xem có bao nhiêu bệnh nhân gãy xương:

```
> table(fx)
fx
 0   1
101 38
```

Kết quả, xem mật độ xương trong nhóm gãy xương và không gãy xương ra sao:

```
> tapply(bmd, fx, mean)
      0       1
0.9444851 0.9016667

> boxplot(bmd ~ fx,
           xlab="Fracture: 1=yes, 0=no",
           ylab="BMD")
```



Kết quả trên cho thấy, bmd trong nhóm bệnh nhân bị gãy xương thấp hơn so với nhóm không bị gãy xương (0.90 và 0.94). Và, kiểm định t sau đây cho thấy mức độ khác biệt này không có ý nghĩa thống kê ($p = 0.15$).

```
> t.test(bmd~fx)

Welch Two Sample t-test

data: bmd by fx
t = 1.4572, df = 53.952, p-value = 0.1508
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-0.01609226 0.10172922
sample estimates:
mean in group 0 mean in group 1
0.9444851     0.9016667
```

Để ước tính thông số trong mô hình [4], hàm số `glm` (viết tắt từ *generalized linear model*) trong R có thể áp dụng, với “cú pháp” như sau:

```
> logistic <- glm(fx ~ bmd, family="binomial")
> summary(logistic)

Call:
glm(formula = fx ~ bmd, family = "binomial")

Deviance Residuals:
    Min      1Q   Median      3Q      Max 
-1.0287 -0.8242 -0.7020  1.3780  2.0709 

Coefficients:
            Estimate Std. Error z value Pr(>|z|)    
(Intercept)  1.063     1.342    0.792    0.428    
bmd        -2.270     1.455   -1.560    0.119    
                                                        
(Dispersion parameter for binomial family taken to be 1)
```

```

Null deviance: 157.81 on 136 degrees of freedom
Residual deviance: 155.27 on 135 degrees of freedom
AIC: 159.27

```

Number of Fisher Scoring iterations: 4

Tôi sẽ lần lượt giải thích các kết quả trên:

- (a) Trong lệnh `logistic <- glm(fx ~ bmd, family="binomial")` chúng ta yêu cầu R phân tích theo mô hình `fx` là một hàm số với `bmd` như mô hình [4]. Trong `glm` có nhiều luật phân phối, mà trong đó phân phối nhị phân (`binomial`) là một luật phân phối chuẩn cho hồi qui logistic. Do đó, `family="binomial"` cần thiết cho R.
- (b) Deviance: phần thứ nhất của kết quả cho biết qua về deviance.

Deviance Residuals:

	Min	1Q	Median	3Q	Max
	-1.0287	-0.8242	-0.7020	1.3780	2.0709

Deviance như giải thích trên phản ánh độ khác biệt giữa mô hình và dữ liệu (cũng tương tự như mean square residual trong phân tích hồi qui tuyến tính vậy). Đối với một mô hình đơn lẻ như ví dụ này thì giá trị của deviance không có ý nghĩa gì nhiều.

- (c) Phần kế tiếp cung cấp ước số của $\hat{\alpha}$ (mà R đặt tên là `intercept`) và $\hat{\beta}$ (`bmd`) và sai số chuẩn (standard error).

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	1.063	1.342	0.792	0.428
bmd	-2.270	1.455	-1.560	0.119

Qua kết quả này, chúng ta có $\hat{\alpha} = 1.063$ và $\hat{\beta} = -2.27$. Ước số $\hat{\beta}$ là số âm cho thấy mối liên hệ giữa nguy cơ gãy xương và `bmd` là mối liên hệ nghịch đảo: xác suất gãy xương tăng khi giá trị của `bmd` giảm. Tuy nhiên, kiểm định z (tính bằng cách lấy ước số chia cho sai số chuẩn) cho chúng ta thấy ảnh hưởng của `bmd` không có ý nghĩa thống kê, vì trị số p = 0.119.

Nhớ rằng tỉ số khả dĩ (odds ratio hay viết tắt là OR) chính là $e^{-2.27} = 0.1033$. Nói cách khác, khi `bmd` tăng 1 g/cm^2 (đơn vị đo lường của `bmd` là g/cm^2) thì tỉ số OR giảm 0.9067 hay 90.67%. Nhưng tăng 1 g/cm^2 là mật độ rất cao trong xương và không thực tế. Cho nên một cách tính khác là tính trên độ lệch chuẩn (standard deviation) của `bmd`. Chúng ta sẽ tìm hiểu độ lệch chuẩn của `bmd`:

```

> sd(bmd)
[1] 0.1406543

```

Do đó, OR sẽ tính trên mỗi 0.14 g/cm^2 . Và OR cho mỗi độ lệch chuẩn, do đó, là:

$$e^{-2.27*0.1406} = 0.7267$$

Tức là, khi bmd tăng một độ lệch chuẩn thì tỉ số khả dĩ gãy xương giảm khoảng 28%. Cũng có thể nói cách khác, là khi bmd giảm một độ lệch chuẩn thì tỉ số khả dĩ tăng $e^{2.27*0.1406} = 1.376$ hay khoảng 38%.

Một cách khác để biết ảnh hưởng của bmd là ước tính xác suất gãy xương qua phương trình:

$$\hat{p} = \frac{e^{1.063 - 2.27(bmd)}}{1 + e^{1.063 - 2.27(bmd)}}$$

Theo đó, khi $bmd = 1.00$, $p = 0.23$. Khi $bmd = 0.86$ (tức giảm 1 độ lệch chuẩn), $p = 0.291$. Tức là, nếu BMD giảm 1 độ lệch chuẩn thì xác suất gãy xương tăng $0.291/0.23 = 1.265$ hay 26%5.

(d) Phần cuối của kết quả cung cấp deviance cho hai mô hình: mô hình không có biến độc lập (null deviance), và mô hình với biến độc lập, tức là bmd trong ví dụ (residual deviance).

```
Null deviance: 157.81 on 136 degrees of freedom
Residual deviance: 155.27 on 135 degrees of freedom
AIC: 159.27
```

Qua hai số này, chúng ta thấy bmd ảnh hưởng rất thấp đến việc tiên đoán gãy xương, chỉ làm giảm deviance từ 157.8 xuống còn 155.27, và mức độ giảm này không có ý nghĩa thống kê.

Ngoài ra, R còn cung cấp giá trị của AIC (Akaike Information Criterion) được tính từ deviance và bậc tự do. Tôi sẽ quay lại ý nghĩa của AIC trong phần sắp đến khi so sánh các mô hình.

12.3 Ước tính xác suất bằng R

Xin nhắc lại trong phân tích trên, chúng ta cho các kết quả vào đối tượng logistic. Trong đối tượng này có nhiều thông tin có ích, nhưng nếu muốn xem các thông tin này chúng ta phải dùng đến các lệnh như `summary` chẳng hạn. Trong phần này, tôi sẽ trình bày một vài hàm để xem xét các thông tin liên quan đến việc tiên đoán xác suất.

- `predict` dùng để liệt kê các giá trị ước tính (predicted values) của mô hình
 $\log\left(\frac{p}{1-p}\right) = \alpha + \beta x$ cho từng bệnh nhân.

```
> predict(logistic)
```

```

1          2          3          4          5          6
2.377576584 1.085694014 -2.141117756 1.492824115 0.965379946 -0.941253280
7          8          9          10         11         12
-1.733686514 -1.675645430 -0.665282957 -0.507046129 -0.941854868 -0.648740461
...

```

Các số trên là $\log(p / (1 - p))$, tức *log odds*, không có ý nghĩa hực tế bao nhiêu. Chúng ta muốn biết giá trị tiên đoán xác suất p tính từ phương trình $\hat{p} = \frac{e^{1.063-2.27(bmd)}}{1+e^{1.063-2.27(bmd)}}$. Để có giá trị này cho từng bệnh nhân, chúng ta cho thông số `type="response"` vào hàm `predict` như sau:

```

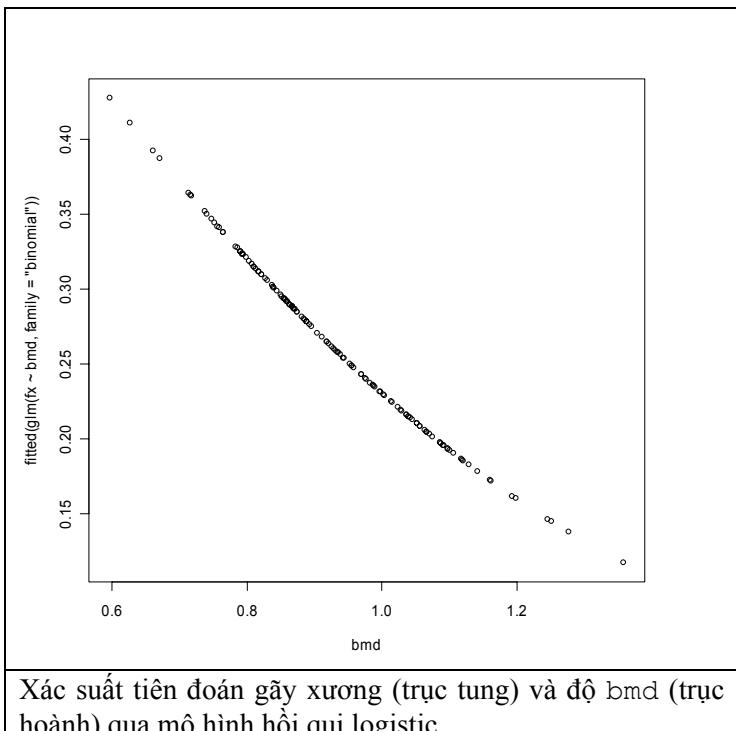
> predict(logistic, type="response")
1          2          3          4          5          6          7
0.91510135 0.74757001 0.10516416 0.81650178 0.72419767 0.28064726 0.15011664
8          9          10         11         12         13         14
0.15767295 0.33955387 0.37588624 0.28052582 0.34327343 0.44305196 0.23830776
...

```

Trong kết quả trên (chỉ in một phần) ước tính xác suất gãy xương cho bệnh nhân 1 là 0.915, cho bệnh nhân 2 là 0.747, v.v...

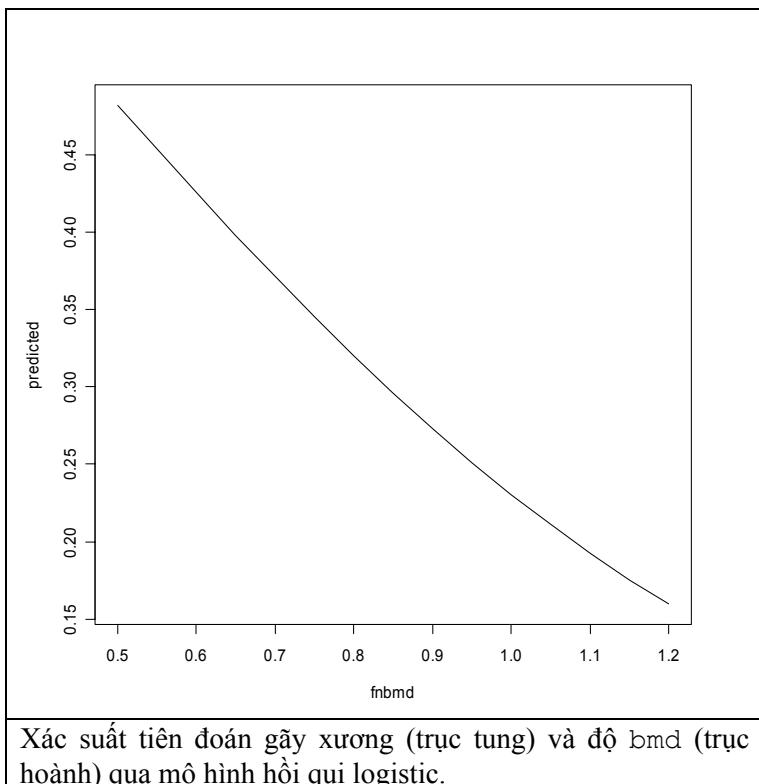
- Chúng ta có thể xem xét các giá trị tiên đoán này với độ b_{MD} bằng cách dùng hàm `plot` thông thường:

```
> plot(bmd, fitted(glm(fx ~ bmd, family="binomial")))
```



Biểu đồ trên có thể cải tiến bằng cách cho các khoảng cách giá trị bmd gần nhau hơn (như 0.50, 0.55, 0.60, ..., 1.20 chẳng hạn), và dùng đường biểu diễn thay vì dùng dấu chấm. Các lệnh sau đây sẽ cải tiến biểu đồ.

```
> logistic <- glm(fx ~ bmd, family="binomial")
> fnbmd <- seq(0.5, 1.2, 0.05) #cho fnbmd từ > 0.50, 0.55, 0.6,...,1.2
> new.data <- data.frame(bmd = fnbmd) #cho vào một datafram mới
> predicted <- predict(logistic, new.data, type="response")
> plot(predicted ~ fnbmd, type="l")
```



12.4 Phân tích hồi qui logistic từ số liệu giản lược bằng R

Trong quá trình phân tích số liệu vừa trình bày trên đây, chúng ta có số liệu cho từng bệnh nhân và các biến độc lập đều là biến liên tục. Nhưng trong nhiều trường hợp biến độc lập là bậc thứ (và bởi vì biến phụ thuộc chỉ có hai giá trị 0 và 1) cho nên trên lý thuyết chúng ta có thể tóm lược dữ liệu bằng các bảng tần số (frequency table).

Ví dụ 2. Trong một nghiên cứu về ảnh hưởng của thói quen hút thuốc lá, tình trạng béo phì, thở ngáy (trong khi ngủ) đến nguy cơ bệnh cao huyết áp, các nhà nghiên cứu tóm lược số liệu như sau (số liệu trích từ Altman, trang 353):

smoking	obesity	snoring	ntotal	nhyper
0	0	0	60	5
1	0	0	17	2

0	1	0	8	1
1	1	0	2	0
0	0	1	187	35
1	0	1	85	13
0	1	1	51	15
1	1	1	23	8
Tổng số			433	79

Bảng 12.2. Tóm lược số liệu liên quan đến hút thuốc lá (smoking), béo phì (obesity), ngáy (snoring), và cao huyết áp. ntotal là tổng số bệnh nhân cho từng nhóm, và nhyper là số bệnh nhân trong tổng số bị bệnh cao huyết áp. Các biến số smoking, obesity, và snoring có giá trị 0=no và 1=yes.

Trong nghiên cứu có 433 bệnh nhân, và trong số này 79 người (hay 18%) bị bệnh cao huyết áp. Tuy nhiên, tỉ lệ này dao động khá cao theo từng nhóm bệnh nhân. Chẳng hạn như trong nhóm không hút thuốc lá (smoking=0), không béo phì (obesity=0) và không ngáy (snoring=0), tỉ lệ cao huyết áp là 8.3% (5/60). Trong khi đó nhóm với 3 yếu tố nguy cơ trên (smoking=1, obesity=1, snoring=0) thì có hơn 1 phần 3 hay 35% (8/23) bị bệnh cao huyết áp.

Để phân tích mối liên hệ giữa 3 yếu tố nguy cơ đó và bệnh cao huyết áp, trước hết cần phải cho số liệu vào R theo đúng như bảng số liệu trên.

```
> noyes <- c("no", "yes") #định nghĩa biến noyes có 2 giá trị
> smoking <- gl(2,1,8, noyes) #biến smoking
> obesity <- gl(2,2,8, noyes) #biến obesity
> snoring <- gl(2,4,8, noyes) #biến snoring
> ntotal <- c(60, 17, 8, 2, 187, 85, 51, 23)
> nhyper <- c(5, 2, 1, 0, 35, 13, 15, 8)
> data <- data.frame(smoking, obesity, snoring, ntotal, nhyper)
> data
  smoking obesity snoring ntotal nhyper
1      no       no       no     60      5
2     yes      no       no     17      2
3      no      yes      no      8      1
4     yes      yes      no      2      0
5      no      no      yes    187     35
6     yes      no      yes     85     13
7      no      yes      yes     51     15
8     yes      yes      yes     23      8
```

Bây giờ chúng ta có thể sử dụng hàm glm để phân tích số liệu. Trước hết, chúng ta phải tạo thêm một biến số proportion như sau:

```
> proportion <- nhyper/ntotal
> logistic <- glm(proportion ~ smoking+obesity+snoring,
  family="binomial",
  weight=ntotal)
```

Chú ý trong hàm glm trên, chúng ta mô phỏng proportion như là một hàm số của smoking, obesity và snoring, vẫn với phân phối nhị phân (binomial), nhưng

có thêm một thông số `weight=ntotal`. Thông số `weight` yêu cầu R sử dụng `ntotal` là một số tóm lược (thay vì một bệnh nhân). Nay giờ, chúng ta có thể xem qua kết quả phân tích:

```
> summary(logistic)

Call:
glm(formula = proportion ~ smoking + obesity + snoring, family = "binomial",
     weights = ntotal)

Deviance Residuals:
    1      2      3      4      5      6      7      8 
-0.04344  0.54145 -0.25476 -0.80051  0.19759 -0.46602 -0.21262  0.56231 

Coefficients:
            Estimate Std. Error z value Pr(>|z|)    
(Intercept) -2.37766   0.38018 -6.254   4e-10 ***  
smokingyes  -0.06777   0.27812 -0.244   0.8075    
obesityyes   0.69531   0.28509  2.439   0.0147 *   
snoringyes   0.87194   0.39757  2.193   0.0283 *   
---
Signif. codes:  0 '****' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 14.1259  on 7  degrees of freedom
Residual deviance: 1.6184  on 4  degrees of freedom
AIC: 34.537

Number of Fisher Scoring iterations: 4
```

Kết quả trên cho thấy biến `smoking` không có ý nghĩa thống kê; cho nên có lẽ chúng ta nên bỏ biến này ra ngoài mô hình và có một mô hình đơn giản hơn:

```
> logistic <- glm(proportion ~ obesity+snoring,
                     family="binomial",
                     weight=ntotal)

> summary(logistic)

Call:
glm(formula = proportion ~ obesity + snoring, family = "binomial",
     weights = ntotal)

Deviance Residuals:
    1      2      3      4      5      6      7      8 
-0.01247  0.47756 -0.24050 -0.82050  0.30794 -0.62742 -0.14449  0.45770 

Coefficients:
            Estimate Std. Error z value Pr(>|z|)    
(Intercept) -2.3921    0.3757  -6.366 1.94e-10 ***  
obesityyes   0.6954    0.2851   2.440   0.0147 *   
snoringyes   0.8655    0.3967   2.182   0.0291 *   
---
Signif. codes:  0 '****' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 14.1259  on 7  degrees of freedom
Residual deviance: 1.6781  on 5  degrees of freedom
```

AIC: 32.597

Number of Fisher Scoring iterations: 4

Phân tích phương sai trên deviance sau đây cũng khẳng định obesity và snoring là hai biến có ảnh hưởng đến cao huyết áp:

```
> anova(logistic, test="Chisq")
Analysis of Deviance Table

Model: binomial, link: logit

Response: proportion

Terms added sequentially (first to last)

          Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL           7    14.1259
obesity      1     6.8260      6    7.2999  0.0090
snoring       1     5.6218      5    1.6781  0.0177
```

12.5 Phân tích hồi qui logistic đa biến và chọn mô hình

Một trong những vấn đề khó khăn và có khi khá nan giải trong việc phân tích hồi qui logistic đa biến là chọn một mô hình để có thể mô tả đầy đủ dữ liệu. Một nghiên cứu với một biến phụ thuộc y và 3 biến độc lập x_1, x_2 và x_3 , chúng ta có thể có những mô hình sau đây để tiên đoán y : $y = f(x_1)$, $y = f(x_2)$, $y = f(x_3)$, $y = f(x_1, x_2)$, $y = f(x_1, x_3)$, $y = f(x_2, x_3)$, và $y = f(x_1, x_2, x_3)$, trong đó f là hàm số. Nói chung với k biến độc lập $x_1, x_2, x_3, \dots, x_k$, chúng ta có rất nhiều mô hình (2^k) để tiên đoán y . Trong điều kiện có nhiều mô hình khả dĩ như thế, vấn đề đặt ra là mô hình nào được xem là tối ưu?

Câu hỏi trên đặt ra một câu hỏi cơ bản khác: thế nào là “tối ưu”? Nói một cách ngắn gọn một mô hình tối ưu phải đáp ứng ba tiêu chuẩn sau đây:

- Đơn giản
- Đầy đủ
- Có ý nghĩa thực tế

Tiêu chuẩn đơn giản đòi hỏi mô hình có ít biến số độc lập, vì nếu quá nhiều biến số thì vấn đề diễn dịch sẽ trở nên khó khăn, và có khi thiếu thực tế. Nói một cách ví von, nếu chúng ta bỏ ra 50.000 đồng để mua 500 trang sách tốt hơn là bỏ ra 60.000 ngàn để mua cùng số trang sách. Tương tự, một mô hình với 3 biến độc lập mà có khả năng mô tả dữ liệu tương đương với mô hình với 5 biến độc lập, thì mô hình đầu được chọn. Một mô hình đơn giản là một mô hình ... tiết kiệm! (Tiếng Anh gọi là *parsimonious model*).

Tiêu chuẩn đầy đủ ở đây có nghĩa là mô hình đó phải mô tả dữ liệu một cách thỏa đáng, tức phải tiên đoán gần (hay càng gần càng tốt) với giá trị thực tế quan sát của biến

phù thuộc y . Nếu giá trị quan sát của y là 10, và nếu có một mô hình tiên đoán là 9 và một mô hình tiên đoán là 6 thì mô hình đầu phải được xem là đầy đủ hơn.

Tiêu chuẩn “có ý nghĩa thực tế”, như cách gọi, có nghĩa là mô hình đó phải được yểm trợ bằng lí thuyết hay có ý nghĩa sinh học (nếu là nghiên cứu sinh học), ý nghĩa lâm sàng (nếu là nghiên cứu lâm sàng), v.v... Có thể số điện thoại một cách nào đó có liên quan đến tỉ lệ gãy xương, nhưng tất nhiên một mô hình như thế hoàn toàn vô nghĩa. Đây là một tiêu chuẩn quan trọng, bởi vì nếu một phân tích thống kê dẫn đến một mô hình dù rất có ý nghĩa toán học mà không có ý nghĩa thực tế thì mô hình đó cũng chỉ là một trò chơi con số, trò chơi toán học không hơn không kém, chứ không có giá trị khoa học thật sự.

Tiêu chuẩn thứ ba (có ý nghĩa thực tế) thuộc về lĩnh vực lí thuyết, và tôi sẽ không bàn ở đây. Tôi sẽ bàn qua tiêu chuẩn đơn giản và đầy đủ. Một thước đo quan trọng và có ích để chúng ta quyết định một mô hình đơn giản và đầy đủ là Akaike Information Criterion (AIC) mà chúng ta đã gặp trong phần đầu của chương này. Để hiểu AIC, chúng ta quay lại với ví dụ 1.

Xin nhắc lại trong ví dụ 1, chúng ta muốn tiên đoán gãy xương (biến fx) từ các biến độc lập sau đây: độ tuổi (age), tỉ số cơ thể (bmi), mật độ chất khoáng trong xương (bmd), và hai chỉ số hủy xương ($ictp$) và tạo xương ($pinp$).

(a) Chúng ta thử mô hình fx là hàm số của độ tuổi:

```
> attach(fulldata)
> summary(glm(fx ~ age, family="binomial", data=fulldata))

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -8.06447   2.72559 -2.959  0.00309 ***
age         0.09806   0.03766  2.604  0.00922 ***
---
Signif. codes:  0 '****' 0.001 '***' 0.01 '**' 0.05 '*' 0.1 '.' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 157.81 on 136 degrees of freedom
Residual deviance: 150.74 on 135 degrees of freedom
AIC: 154.74
```

Chúng ta để ý thấy residual deviance = 150.74, và AIC = 154.74. Thật ra, AIC được ước tính từ công thức:

$$AIC = \text{Residual Deviance} + 2(\text{số thông số})$$

Trong mô hình trên, chúng ta có 2 thông số (`intercept` và `age`), cho nên $AIC = 150.74 + 4 = 154.74$.

(b) Mô hình thứ hai mà chúng ta muốn so sánh là fx là hàm số của `ictp`:

```
> summary(glm(fx ~ ictp, family="binomial", data=fulldata))
```

```

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.9206     0.7726  -5.074 3.89e-07 ***
ictp         0.6066     0.1527   3.973 7.11e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 157.81  on 136  degrees of freedom
Residual deviance: 139.15  on 135  degrees of freedom
AIC: 143.15

```

Cũng với hai thông số, nhưng mô hình này có giá trị residual deviance (139.15) nhỏ hơn mô hình với độ tuổi (150.74), và do đó AIC cũng thấp hơn (143.15 so với 154.74). Kết quả này cho thấy mô hình với ictp mô tả fx đầy đủ hơn là mô hình với độ tuổi. So sánh này cho thấy trong hai mô hình này, chúng ta sẽ chọn mô hình với ictp.

(c) Nay giờ chúng ta thử xem mô hình với ictp và age.

```

> summary(glm(fx ~ ictp + age, family="binomial", data=fulldata))

(Intercept) -8.25707    2.91403  -2.834 0.004603 **
ictp          0.55461    0.15665   3.540 0.000399 ***
age           0.06398    0.04067   1.573 0.115701
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 157.81  on 136  degrees of freedom
Residual deviance: 136.61  on 134  degrees of freedom
AIC: 142.61

```

Mô hình này với 3 thông số (intercept, age và ictp), nhưng trị số AIC chỉ giảm xuống 142.61 (so với mô hình với ictp là 143.15), một độ giảm rất khiêm tốn, trong khi chúng ta phải “tiêu” thêm một thông số! Chúng ta có thể kết luận rằng age không cần thiết trong mô hình này. Thật vậy, trị số p cho age là 0.115, tức không có ý nghĩa thống kê.

Qua ba trường hợp trên, chúng ta có thể rút ra một nhận xét chung: một mô hình đơn giản và đầy đủ phải là mô hình có trị số AIC càng thấp càng tốt và các biến độc lập phải có ý nghĩa thống kê. Thành ra, vấn đề đi tìm một mô hình đơn giản và đầy đủ là thật sự đi tìm một (hay nhiều) mô hình với trị số AIC thấp nhất hay gần thấp nhất.

Tất nhiên, chúng ta có thể xem xét nhiều mô hình khác bằng cách thay thế hay tổng hợp các biến số độc lập với nhau. Nhưng một việc làm như thế rất phức tạp, đòi hỏi nhiều thời gian và có khi rườm rà. R có một hàm gọi là step có thể giúp chúng ta đi tìm một mô hình đơn giản và đầy đủ. Trong ví dụ trên, cách sử dụng hàm step sẽ được viết như sau:

```
> temp <- glm(fx ~ ., family="binomial", data=fulldata)
```

Trong lệnh trên, thông số “`fx ~ .`” có nghĩa là tìm tất cả các biến độc lập (kí hiệu “.”) để tiên đoán `fx` trong dataframe `fulldata`. Kết quả cho vào đối tượng `temp`. Để xem kết quả trong `temp`, chúng ta lệnh `search` như sau:

```
> search <- step(temp)

> search <- step(temp)
Start: AIC= 146.09
  fx ~ id + age + bmi + bmd + ictp + pinp

      Df Deviance    AIC
- pinp  1    132.45 144.45
- id    1    132.47 144.47
- age   1    132.63 144.63
- bmi   1    133.41 145.41
- bmd   1    133.87 145.87
<none>        132.09 146.09
- ictp  1    148.90 160.90

Step: AIC= 144.45
  fx ~ id + age + bmi + bmd + ictp

      Df Deviance    AIC
- id    1    132.81 142.81
- age   1    133.14 143.14
- bmi   1    133.66 143.66
- bmd   1    134.00 144.00
<none>        132.45 144.45
- ictp  1    149.05 159.05

Step: AIC= 142.81
  fx ~ age + bmi + bmd + ictp

      Df Deviance    AIC
- age   1    133.32 141.32
- bmi   1    133.67 141.67
- bmd   1    134.33 142.33
<none>        132.81 142.81
- ictp  1    149.88 157.88

Step: AIC= 141.33
  fx ~ bmi + bmd + ictp

      Df Deviance    AIC
- bmi   1    134.34 140.34
<none>        133.32 141.32
- bmd   1    135.65 141.65
- ictp  1    155.18 161.18

Step: AIC= 140.34
  fx ~ bmd + ictp

      Df Deviance    AIC
<none>        134.34 140.34
- bmd   1    139.15 143.15
- ictp  1    155.27 159.27
```

Trong kết quả trên, R báo cáo cho chúng ta biết từng bước trong quá trình đi tìm mô hình tối ưu. Khởi đầu là mô hình với tất cả 6 biến, và trị số AIC = 146.09. Bước thứ hai

chỉ gồm 5 biến (loại bỏ pinp) và AIC = 144.45. Và vân vân. Kết quả có thể tóm lược trong bảng sau đây:

Mô hình	AIC
$fx \sim id + age + bmi + bmd + ictp + pinp$	146.09
$fx \sim id + age + bmi + bmd + ictp$	144.45
$fx \sim age + bmi + bmd + ictp$	142.81
$fx \sim bmi + bmd + ictp$	141.33
$fx \sim bmd + ictp$	140.34

Kết quả 5 bước tìm mô hình, R dừng lại với mô hình gồm 2 biến bmd và ictp vì có giá trị AIC thấp nhất. Thật ra, nếu không muốn in tất cả các bước đi tìm mô hình, chúng ta chỉ cần lệnh `summary` như sau:

```
> summary(search)

Call:
glm(formula = fx ~ bmd + ictp, family = "binomial", data = fulldata)

Deviance Residuals:
    Min      1Q  Median      3Q     Max 
-1.9126 -0.7317 -0.5559  0.4212  2.1242 

Coefficients:
            Estimate Std. Error z value Pr(>|z|)    
(Intercept) -1.0651    1.5029 -0.709   0.4785    
bmd         -3.4998    1.6638 -2.103   0.0354 *  
ictp          0.6876    0.1704  4.036 5.43e-05 *** 
---
Signif. codes:  0 '****' 0.001 '***' 0.01 '**' 0.05 '*' 0.1 '.' 1 

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 157.81  on 136  degrees of freedom
Residual deviance: 134.34  on 134  degrees of freedom
AIC: 140.34

Number of Fisher Scoring iterations: 4
```

Kết quả này đơn giản hơn kết quả của hàm `search`, vì `summary` chỉ trình bày mô hình sau cùng. Nói tóm lại, trong phân tích này, chúng ta kết luận rằng bmd (mật độ chất khoáng trong xương) và ictp (marker về chu trình hủy xương) là hai yếu tố có liên hệ hay ảnh hưởng đến nguy cơ gãy xương.

12.6 Chọn mô hình hồi qui logistic bằng Bayesian Model Average (BMA)

Trong chương 10, tôi đã nói qua cách chọn và xây dựng một mô hình hồi qui tuyến tính bằng ứng dụng phép tính BMA. Chúng ta cũng có thể ứng dụng BMA vào việc xây dựng một mô hình hồi qui logistic.

Tiếp tục ví dụ 1, chúng ta sẽ chuẩn bị dữ liệu cho phân tích BMA bằng cách chọn ra biến phụ thuộc (trong trường hợp này là `fx`) và một ma trận gồm các biến độc lập. Tiếp theo đó, chúng ta sử dụng hàm `bic.glm` để tìm các biến có ảnh hưởng đến `fx`.

```
> attach(fulldata)
> names(fulldata)
[1] "id"    "fx"    "age"   "bmi"   "bmd"   "ictp"  "pinp"

# Chọn cột 3 đến 7 (từ age đến pinp) làm ma trận biến độc lập
> xvars <- fulldata[,3:7]

# Chọn fx làm biến phụ thuộc
> y <- fx

# Gọi hàm bic.glm với các thông số như sau:
> bma.search <- bic.glm(xvars, y, strict=F, OR=20, glm.family="binomial")

# Tóm lược kết quả phân tích:
> summary(bma.search)

Call:
bic.glm.data.frame(x = xvars, y = y, glm.family = "binomial",      strict = F, OR = 20)

 9 models were selected
Best 5 models (cumulative posterior probability = 0.8836 ):

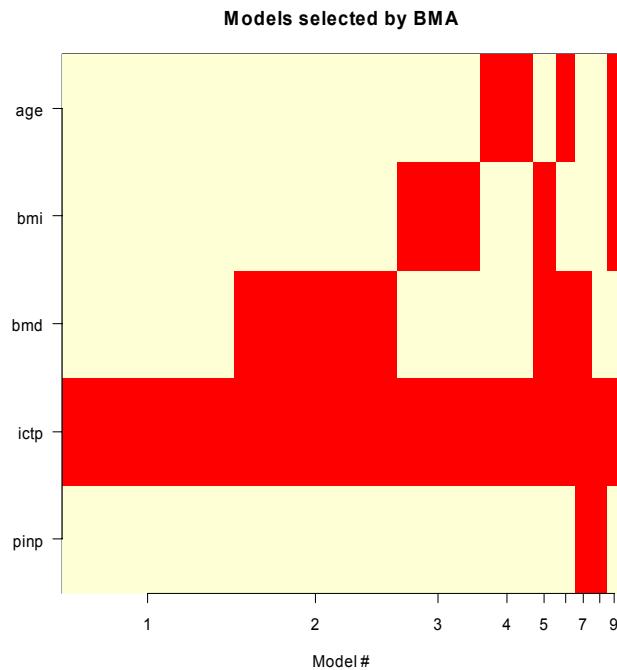
          p!=0      EV       SD     model 1     model 2     model 3     model 4     model 5
Intercept 100 -2.85012 2.8651 -3.920     -1.065    -1.201    -8.257   -0.072
age        15.3  0.00845 0.0261      .         .         .         0.063     .
bmi        21.7 -0.02302 0.0541      .         .         -0.116     .         -0.070
bmd        39.7 -1.34136 1.9762      .         -3.499     .         .         -2.696
ictp       100.0  0.64575 0.1699  0.606     0.687    0.680     0.554    0.714
pinp       5.7  -0.00037 0.0041      .         .         .         .         .

nVar           1       2       2       2       3
BIC      -525.044 -524.939 -523.625 -522.672 -521.032
post prob  0.307   0.291   0.151   0.094   0.041
```

Kết quả phân tích trên đây cho thấy xác suất mà `ictp` là liên quan đến gãy xương là 100%, trong khi đó, xác suất cho `bmd` chỉ khoảng 40%. Nhưng quan trọng hơn, mô hình “tối ưu” nhất là mô hình với `ictp`, và xác suất cho mô hình này là 0.307. Mô hình tối ưu thứ hai gồm có `ictp` và `bmd` (cũng là mô hình dựa vào tiêu chuẩn AIC như mô tả phần trên), nhưng xác suất cho mô hình này thường đối thấp hơn (0.291). Ba mô hình khác cũng có thể là “ứng viên” để mô tả xác suất gãy xương đầy đủ. Rõ ràng, qua phân tích BMA, chúng ta có nhiều lựa chọn mô hình hơn, và ý thức được sự bất định của một mô hình thống kê.

Biểu đồ sau đây thể hiện kết quả trên. Qua biểu đồ này chúng ta thấy `ictp` là yếu tố có ảnh hưởng đến nguy cơ gãy xương nhất quán nhất. Yếu tố quan trọng thứ hai có lẽ là `bmd` hay `bmi`. Các yếu tố như `age` và `pinp` tuy có khả năng ảnh hưởng đến nguy cơ gãy xương, nhưng các yếu tố này không có độ nhất quán cao như các yếu tố vừa kể trên.

```
> imageplot.bma(bma.search)
```



Xây dựng mô hình thống kê là một nghệ thuật toán học. Vì tính nghệ thuật của việc làm, nhà nghiên cứu phải cân nhắc rất nhiều yếu tố để đi đến một mô hình đẹp. Bởi vì mô hình là nhằm mục đích mô tả thực tế, một mô hình đẹp là mô hình mô tả sát với thực tế. Tuy nhiên nếu một mô hình phản ánh 100% thực tế thì đó không còn là “mô hình” nữa, hay quá phức tạp không thể ứng dụng được. Ngược lại một mô hình chỉ mô tả thực tế khoảng 1% thì cũng không thể sử dụng được. Xây dựng mô hình phải làm sao tìm điểm cân bằng cho hai thái cực đó. Đó là một yêu cầu rất cao, cho nên xây dựng mô hình không chỉ tùy thuộc vào các phép tính thống kê, toán học, mà còn phải xem xét đến các yếu tố thực tế để bảo đảm cho sự hữu ích của mô hình. Nói như nhà thống kê học nổi tiếng George Box: “Mô hình nào cũng sai so với thực tế, nhưng trong số các mô hình sai đó, có một vài mô hình có ích”.

12.7 Số liệu nghiên cứu về nguy cơ gãy xương trong nam giới trên 60 tuổi

- id: mã số bệnh nhân
- fx: gãy xương hay không (0=không gãy xương, 1=gãy xương)
- age: độ tuổi
- bmi: body mass index, tính bằng trọng lượng chia cho chiều cao bình phương
- bmd: (bone mineral density) mật độ chất khoáng trong xương đùi.
- ictp: chỉ số sinh hóa đo lường hoạt tính hủy xương
- pinp: chỉ số sinh hóa đo lường hoạt tính tạo xương

id	fx	age	bmi	bmd	ictp	pinp
1	1	79	24.7252	0.818	9.170	37.383
2	1	89	25.9909	0.871	7.561	24.685
3	1	70	25.3934	1.358	5.347	40.620
4	1	88	23.2254	0.714	7.354	56.782
5	1	85	24.6097	0.748	6.760	58.358
6	0	68	25.0762	0.935	4.939	67.123
7	0	70	19.8839	1.040	4.321	26.399
8	0	69	25.0593	1.002	4.212	47.515
9	0	74	25.6544	0.987	5.605	26.132
10	0	79	19.9594	0.863	5.204	60.267
11	1	76	22.5981	0.889	4.704	27.026
12	0	76	26.4236	0.886	5.115	43.256
13	1	62	20.3223	0.889	5.741	51.097
14	0	69	19.3698	0.790	3.880	49.678
15	0	72	24.2215	0.988	5.844	41.672
16	0	67	32.1120	1.119	4.160	60.356
17	0	74	25.3934	1.037	6.728	40.225
18	0	69	23.8895	0.893	4.203	27.334
19	1	78	24.6755	0.850	7.347	38.893
20	0	71	27.1314	0.790	4.476	38.173
21	1	74	23.0518	0.597	4.835	35.141
22	1	76	23.4568	0.889	5.354	27.568
23	1	75	23.5457	0.803	3.773	36.762
24	0	70	23.3234	0.919	3.672	40.093
25	1	69	22.8625	0.870	4.552	29.627
26	0	71	22.0384	0.811	4.286	30.380
27	1	80	24.6914	0.859	5.706	37.529
28	1	79	26.8519	0.867	3.563	43.924
29	0	72	27.1809	0.717	3.760	39.714
30	0	78	23.9512	0.822	3.453	27.294
31	1	80	28.3874	1.004	5.948	33.376
32	0	79	23.5102	0.738	4.193	65.640
33	1	67	19.7232	0.865	4.443	36.252
34	1	84	27.4406	0.808	5.482	33.539
35	0	78	28.6661	0.955	8.815	42.398
36	0	65	23.7812	0.912	4.704	39.254
37	0	70	23.4493	0.857	4.138	75.947
38	0	67	25.5354	0.855	3.727	41.851
39	0	74	24.7409	0.959	3.967	42.293
40	0	73	22.2291	1.036	4.438	40.222
41	0	74	34.4753	1.092	7.271	45.434
42	1	68	32.1929	.	4.269	50.841
43	0	80	23.3355	0.759	4.856	31.114
44	0	78	22.7903	0.757	4.831	73.343
45	1	79	24.6097	0.671	4.870	69.924
46	0	72	27.5802	0.814	3.012	27.088
47	1	67	30.1205	1.101	7.538	35.487
48	0	70	25.8166	0.818	3.564	36.001
49	0	69	30.4218	1.088	3.826	33.833
50	0	67	28.7132	0.934	3.996	56.167

51	0	74	34.5429	0.969	6.762	43.099
52	0	71	24.6097	0.794	4.350	39.023
53	0	67	23.5294	0.830	3.176	36.595
54	0	67	25.6173	1.057	3.738	32.550
55	0	65	25.3086	1.160	3.060	44.757
56	0	66	24.8358	0.811	3.263	26.941
57	0	69	22.3094	0.977	3.106	27.951
58	0	72	26.5285	1.063	6.970	41.188
59	0	75	25.8546	1.091	4.798	36.045
60	0	70	20.6790	0.741	3.908	30.198
61	0	74	28.3675	1.045	4.784	31.339
62	0	71	29.0688	1.066	4.527	24.252
63	0	65	23.9995	0.841	3.089	79.910
64	0	77	22.9819	1.015	4.041	57.147
65	1	67	33.3598	1.129	7.239	67.103
66	0	66	27.1314	1.030	4.096	29.435
67	0	70	24.7676	0.896	4.352	44.291
68	0	70	24.4193	1.106	2.823	37.348
69	0	69	28.2570	0.869	2.974	46.229
70	1	65	23.6614	0.837	2.689	28.738
71	1	75	26.0262	0.921	3.917	29.667
72	0	67	26.5731	1.118	3.832	50.292
73	0	67	24.8591	0.765	7.112	45.778
74	0	73	22.5710	0.752	4.249	39.950
75	1	63	31.8342	1.251	7.303	48.697
76	1	72	24.8016	0.839	3.860	41.055
77	0	73	25.0574	0.662	3.138	36.312
78	0	69	23.9512	0.844	4.069	39.926
79	0	75	23.4586	0.852	4.176	51.394
80	0	65	28.7347	0.795	3.328	27.679
81	0	71	25.3350	0.867	2.349	36.506
82	0	66	28.0899	0.997	4.171	53.094
83	0	66	25.5650	0.827	4.569	25.157
84	0	71	28.7274	1.023	4.111	19.557
85	0	73	32.4074	1.066	5.680	36.995
86	0	64	27.9155	0.874	4.298	43.872
87	0	68	25.5937	0.882	4.056	30.523
88	1	67	28.0428	0.718	9.739	66.974
89	0	66	30.7174	0.856	4.180	34.597
90	0	77	28.3737	1.052	3.737	28.102
91	0	75	28.6990	0.929	3.527	23.008
92	0	67	29.1687	0.953	3.593	16.132
93	0	73	27.4145	0.784	4.332	47.410
94	0	68	29.0688	1.120	6.510	45.674
95	0	70	26.1738	1.040	3.161	36.302
96	0	66	30.1038	1.028	3.930	38.301
97	0	77	24.6559	0.884	3.880	36.560
98	1	71	25.3934	0.943	4.692	69.500
99	0	74	26.4721	1.075	4.561	25.948
100	0	70	29.0253	1.057	3.709	41.322
101	0	78	29.0253	1.098	5.247	23.896
102	0	76	26.2346	1.014	3.958	24.344
103	1	64	26.4915	0.998	4.218	29.390
104	0	67	27.0416	0.905	3.553	23.020
105	0	66	22.7732	0.627	2.333	53.621
106	0	70	30.5241	1.052	5.425	44.352
107	0	66	25.3069	1.086	4.945	64.788
108	1	65	22.3863	0.818	3.786	96.360
109	1	64	34.0136	1.066	5.792	37.473
110	1	70	26.5668	1.198	7.257	28.406
111	1	70	27.6361	0.926	5.746	17.228
112	0	70	25.4017	1.193	2.437	35.432
113	0	68	30.3673	0.938	2.658	32.293
114	0	67	28.0428	0.863	4.246	48.702
115	1	73	27.7778	0.799	3.934	26.709
116	0	71	29.0006	0.969	4.054	22.769
117	1	71	35.2941	0.931	3.631	18.629
118	0	75	29.3658	1.071	4.222	36.555
119	0	76	26.2649	1.161	2.548	24.217
120	0	71	25.6055	0.786	3.832	32.023
121	0	73	29.9136	0.839	4.215	26.507

122	0	64	34.5271	1.042	6.436	53.080
123	0	70	33.4554	0.976	4.541	26.619
124	0	80	29.0688	0.765	3.998	67.388
125	0	67	25.7276	1.277	3.877	22.159
126	0	68	25.6801	1.097	3.782	42.286
127	0	66	25.9701	0.793	2.991	38.673
128	0	64	26.4490	0.989	3.196	31.456
129	1	69	28.6990	0.822	3.565	45.044
130	0	69	25.6173	0.944	6.512	49.557
131	0	67	30.3871	1.245	3.603	46.769
132	0	67	33.6901	1.142	3.666	38.839
133	0	68	28.4005	0.860	2.890	32.140
134	1	59	25.4017	1.172	.	104.579
135	0	66	22.5710	0.956	3.354	36.253
136	1	71	24.4473	0.918	4.633	53.881
137	0	64	38.0762	1.086	5.043	32.835
138	1	80	23.3887	0.875	4.086	23.837
139	0	67	25.9455	0.983	4.328	71.334

13

Phân tích sự kiện (event history hay survival analysis)

Qua ba chương trước, chúng ta đã làm quen với các mô hình thống kê cho các biến phụ thuộc liên tục (như áp suất máu) và biến bậc thứ (như có/không, bệnh hay không bệnh). Trong nghiên cứu khoa học, và đặc biệt là y học và kĩ thuật, có khi nhà nghiên cứu muốn tìm hiểu ảnh hưởng đến các biến phụ thuộc mang tính thời gian. Nhà kinh tế học John Maynard Keynes từng nói một câu có liên quan đến chủ đề mà tôi sẽ mô tả trong chương này như sau: “Về lâu về dài tất cả chúng ta đều chết, cái khác nhau là chết sớm hay chết muộn mà thôi.” Thành ra, ở đây việc theo dõi hay mô tả một biến bậc thứ như sống hay chết tuy quan trọng, nhưng … không chính xác. Cái biến số quan trọng hơn và chính xác hơn là thời gian dẫn đến việc sự kiện xảy ra.

Trong các nghiên cứu y học, kể cả nghiên cứu lâm sàng, các nhà nghiên cứu thường theo dõi bệnh nhân trong một thời gian, có khi lên đến vài mươi năm. Biến cố xảy ra trong thời gian đó như có bệnh hay không có bệnh, sống hay chết, v.v... là những biến cố có ý nghĩa lâm sàng nhất định, nhưng thời gian dẫn đến bệnh nhân mắc bệnh hay chết còn quan trọng hơn cho việc đánh giá ảnh hưởng của một thuật điều trị hay một yếu tố nguy cơ. Nhưng thời gian này khác nhau giữa các bệnh nhân. Chẳng hạn như thời điểm từ lúc điều trị ung thư đến thời điểm bệnh nhân chết rất khác nhau giữa các bệnh nhân, và độ khác biệt đó có thể tùy thuộc vào các yếu tố như độ tuổi, giới tính, tình trạng bệnh, và các yếu tố mà có khi chúng ta không/chưa đo lường được như tương tác giữa các gen.

Mô hình chính để thể hiện mối liên hệ giữa thời gian dẫn đến bệnh (hay không bệnh) và các yếu tố nguy cơ (risk factors) là mô hình có tên là “survival analysis” (có thể tạm dịch là *phân tích sống sót*). Cụm từ “survival analysis” xuất phát từ nghiên cứu trong bảo hiểm, và giới nghiên cứu y khoa từ đó dùng cụm từ cho bộ môn của mình. Nhưng như nói trên, sống/chết không phải là biến duy nhất, vì trong thực tế chúng ta cũng có những biến như có bệnh hay không bệnh, xảy ra hay không xảy ra, và do đó, trong giới tâm lý học, người ta dùng cụm từ “event history analysis” (*phân tích biến cố*) mà tôi thấy có vẻ thích hợp hơn là *phân tích sống sót*. Ngoài ra, trong các bộ môn kĩ thuật, người ta dùng một cụm từ khác, *reliability analysis* (*phân tích độ tin cậy*), để chỉ cho khái niệm *survival analysis*. Tuy nhiên, trong chương này tôi sẽ dùng cụm từ *phân tích biến cố*.

13.1 Mô hình phân tích số liệu mang tính thời gian

Ví dụ 1. Thời gian dẫn đến ngưng sử dụng IUD. Một nghiên cứu về hiệu quả của một y cụ ngừa thai trên 18 phụ nữ, tuổi từ 18 đến 35. Một số phụ nữ ngưng sử dụng y cụ vì bị chảy máu. Còn số khác thì tiếp tục sử dụng. Bảng số liệu sau đây là thời gian

(tính bằng tuần) kể từ lúc bắt đầu sử dụng y cụ đến khi chảy máu (tức ngưng sử dụng) hay đến khi kết thúc nghiên cứu (tức vẫn còn sử dụng đến khi chấm dứt nghiên cứu).

Bảng 13.1 Thời gian dẫn đến ngưng sử dụng hay tiếp tục sử dụng y cụ IUD

Mã số bệnh nhân	Thời gian (tuần)	Tình trạng (ngưng=1 hay tiếp tục=0)
1	18	0
2	10	1
3	13	0
4	30	1
5	19	1
6	23	0
7	38	0
8	54	0
9	36	1
10	107	1
11	104	0
12	97	1
13	107	0
14	56	0
15	59	1
16	107	0
17	75	1
18	93	1

Câu hỏi đặt ra là mô tả thời gian ngưng sử dụng y cụ. Thuật ngữ “mô tả” ở đây có nghĩa là ước tính số trung vị thời gian dẫn đến ngưng sử dụng, hay xác suất mà phụ nữ ngưng sử dụng vào một thời điểm nào đó. Tình trạng tiếp tục sử dụng có khi gọi là “survival” (tức “sống sót”).

Để giải quyết vấn đề trên, đối với những phụ nữ đã ngưng sử dụng vấn đề ước tính thời gian không phải là khó. Nhưng vấn đề quan trọng trong dữ liệu mang tính thời gian này là một số phụ nữ vẫn còn tiếp tục sử dụng, bởi vì chúng ta không biết họ còn sử dụng bao lâu nữa, trong khi nghiên cứu phải “đóng sổ” theo một thời điểm định trước. Những trường hợp đó được gọi bằng một thuật ngữ khó hiểu là “censored” hay “survival” (tức còn sống, hay còn tiếp tục sử dụng, hay biến cố chưa xảy ra).

Gọi T là thời gian còn tiếp tục sử dụng (có khi gọi là *survival time*). T là một biến ngẫu nhiên, với hàm mật độ (probability density distribution) $f(t)$, và hàm phân phối tích lũy (cumulative distribution) là:

$$F(t) = \int'_{-\infty} f(s) ds$$

Đây là xác suất mà một cá nhân ngưng sử dụng (hay kinh qua biến cố) tại thời điểm t . Hàm bổ sung $S(t) = 1 - F(t)$ thường được gọi là hàm “sống sót” (survival function).

Số liệu thời gian T thường được mô phỏng bằng hai hàm xác suất: hàm sống sót và hàm nguy cơ (*hazard function*). Hàm sống sót như định nghĩa trên là xác suất một cá nhân còn “sống sót” (hay trong ví dụ trên, còn sử dụng y cụ) đến một thời điểm t . Hàm nguy cơ, thường được viết bằng ký hiệu $h(t)$ hay $\lambda(t)$ là xác suất mà cá nhân đó ngưng sử dụng (hay xảy ra biến cố) ngay tại thời điểm t .

$$h(t) = \lim_{\Delta t \rightarrow 0} \frac{\Pr[(t \leq T < t + \Delta t) | T \geq t]}{\Delta t} = \frac{f(t)}{S(t)}$$

sao cho $h(t)$ là xác suất một cá nhân ngưng sử dụng trong khoảng thời gian ngắn δt với điều kiện cá nhân đó sống đến thời điểm t . Từ mối liên hệ:

$$\Pr(\text{sống sót đến } t+\delta t) = \Pr(\text{sống sót đến } t) \cdot \Pr(\text{sống sót đến } \delta t | \text{sống đến } t)$$

chúng ta có:

$$1 - F(t + \delta t) = (1 - F(t)) \times (1 - h(t) \delta t)$$

Từ đó, chúng ta có:

$$\delta t F'(t) = (1 - F(t)) h(t) \delta t$$

Thành ra, hàm nguy cơ là:

$$h(t) = \frac{f(t)}{1 - F(t)}$$

Và hàm nguy cơ tích lũy:

$$\Lambda(t) = \int_{-\infty}^t \lambda(u) du$$

Từ định nghĩa hàm nguy cơ $-h(t) = \frac{-f(t)}{1 - F(t)}$, chúng ta có thể viết:

$$\Lambda(t) = -\log(1 - F(t))$$

Một số hàm nguy cơ có thể ứng dụng để mô tả thời gian này. Hàm đơn giản nhất là một hằng số, dẫn đến một mô hình Poisson (thuộc nhóm các luật phân phối mõi):

$$f(t) = \lambda e^{-\lambda t} \quad (t \geq 0)$$

Do đó:

$$F(t) = 1 - e^{-\lambda t}$$

Thành ra:

$$h(t) = \lambda$$

Những lí thuyết trên đây thoạt đầu mới xem qua có vẻ tương đối rắc rối, nhưng với số liệu thực tế thì sẽ dễ theo dõi hơn. Bây giờ chúng ta quay lại với số liệu từ **Ví dụ 1**. Để tiện việc theo dõi và tính toán, chúng ta cần phải sắp xếp lại dữ liệu trên theo thứ tự thời gian, bắt kể đó là thời gian ngưng sử dụng hay còn tiếp tục sử dụng:

10	13*	18*	19	23*	30	36	38*	54*
56*	59	75	93	97	104*	107	107*	

Trong dãy số liệu trên dấu “*” là để đánh dấu thời gian censored (tức còn tiếp tục sử dụng IUD). Cách đơn giản nhất là chia thời gian từ 10 tuần (ngắn nhất) đến 107 tuần (lâu nhất) thành nhiều khoảng thời gian như trong bảng phân tích sau đây:

Bảng 13.2. Ước tính xác suất tích lũy cho mỗi khoảng thời gian

Mốc thời gian (t)	Khoảng thời gian (tuần)	Số phụ nữ lúc bắt đầu thời điểm (n_t)	Số phụ nữ ngưng sử dụng (d_t)	Xác suất ngưng sử dụng $h(t)$	Xác suất còn sử dụng p_t	Xác suất tích lũy $S(t)$
1	0 – 9	18	0	0.0000	1.0000	1.0000
2	10 – 18	18	1	0.0555	0.9445	0.9445
3	19 – 29	15	1	0.0667	0.9333	0.8815
4	30 – 35	13	1	0.0769	0.9231	0.8137
5	36 – 58	12	1	0.0833	0.9167	0.7459
6	59 – 74	8	1	0.1250	0.8750	0.6526
7	75 – 92	7	1	0.1428	0.8572	0.5594
8	93 – 96	6	1	0.1667	0.8333	0.4662
9	97 – 106	5	1	0.2000	0.8000	0.3729
10	107 –	3	1	0.3333	0.6667	0.2486

Trong bảng tính toán trên, chúng ta có:

- Cột thứ nhất là mốc thời gian (tạm kí hiệu là t). Cột này không có ý nghĩa gì, ngoại trừ sử dụng để làm chỉ số;
- Cột thứ 2 là khoảng thời gian (duration) tính bằng tuần. Như đề cập trên, chúng ta chia thời gian thành nhiều khoảng để tính toán, chẳng hạn như từ 0 đến 9 tuần, 10 đến 18 tuần, v.v... Chú ý rằng trong thực tế, chúng ta không có số liệu cho thời gian từ 0 đến 9 tuần, nhưng khoảng thời gian này đặt ra để làm cái mốc khởi đầu để tiện cho việc ước tính sau này. Đây chỉ là những phân chia tương đối tùy tiện và chỉ có tính cách minh họa; trong thực tế máy tính có thể làm việc đó cho chúng ta;
- Cột thứ 3 là số đối tượng nghiên cứu n_t (hay cụ thể hơn là số phụ nữ trong nghiên cứu này) bắt đầu một khoảng thời gian. Chẳng hạn như khoảng thời gian 0-9, tại thời điểm bắt đầu 0, có 18 phụ nữ (hay cũng có thể hiểu rằng số phụ nữ được theo dõi/quan sát ít nhất 0 tuần là 18 người).

Trong khoảng thời gian 10–18, ngay tại thời điểm bắt đầu 10, chúng ta có 18 phụ nữ; nhưng trong khoảng thời gian 19–29, ngay tại thời điểm bắt đầu 19, chúng ta có 15 phụ nữ (cụ thể là: 19 23* 30 36 38* 54* 56* 59 75 93 97 104* 107 107* 107*); vân vân.

Nói cách khác, cột này thể hiện số đối tượng với thời gian quan sát tối thiểu là t . Do đó, trong khoảng thời gian 97 – 106, chúng ta có 5 phụ nữ với thời gian theo dõi từ 97 tuần trở lên ($97 \ 104^* \ 107 \ 107^* \ 107^*$).

- Cột thứ 4 trình bày số phụ nữ ngưng sử dụng y cụ d_t (hay biến cố xảy ra) trong một *khoảng thời gian*. Chẳng hạn như trong khoảng thời gian 10–18 tuần, có một phụ nữ ngưng sử dụng(tại 10 tuần); trong khoảng thời gian 19 – 29 tuần cũng có một trường hợp ngưng sử dụng (tại 19 tuần), v.v...
- Cột thứ 5 là xác suất nguy cơ $h(t)$ trong một khoảng thời gian. Một cách đơn giản, $h(t)$ được ước tính bằng cách lấy d_t chia cho n_t . Ví dụ trong khoảng thời gian 10-18 có 1 phụ nữ ngưng sử dụng (trong số 18 phụ nữ), và xác suất nguy cơ là $1/18 = 0.0555$. Xác suất này được ước tính cho từng khoảng thời gian.
- Cột thứ 6 là xác suất còn sử dụng cho một khoảng thời gian, tức lấy 1 trừ cho $h(t)$ trong cột thứ 5. Xác suất này không cung cấp nhiều thông tin, nhưng chỉ được trình bày để dễ theo dõi tính toán trong cột kế tiếp.
- Cột thứ 7 là xác suất tích lũy còn sử dụng y cụ $S(t)$ (hay cumulative survival probability). Đây là cột số liệu quan trọng nhất cho phân tích. Vì tính chất “tích lũy”, cho nên cách ước tính được nhân từ hai hay nhiều xác suất.

Trong khoảng thời gian 0-9, xác suất tích lũy chính là xác suất còn sử dụng trong cột 6, (vì không có ai ngưng sử dụng).

Trong khoảng thời gian 10-18, xác suất tích lũy được ước tính bằng cách lấy xác suất còn sử dụng trong thời gian 0-9 nhân cho xác suất còn sử dụng trong thời gian 10-18, tức là: $1.000 \times 0.9445 = 0.9445$. Ý nghĩa của ước tính này là: xác suất còn sử dụng cho đến thời gian 9 tuần là 94.45%.

Tương tự, trong khoảng thời gian 19-29 tuần, xác suất tích lũy còn sử dụng được tính bằng cách lấy xác suất tích lũy còn sử dụng đến tuần 10-18 nhân cho xác suất còn sử dụng trong khoảng thời gian 19-29: $0.9445 \times 0.9333 = 0.8815$. Tức là, xác suất còn sử dụng đến tuần 29 là 88.15%.

Nói chung, công thức ước tính $S(t)$ là $\hat{S}(t) = \prod_{t=1}^k \left(\frac{n_t - d_t}{n_t} \right)$. Chú ý dấu mũ “ k ”

trên $S(t)$ là để nhắc nhở rằng đó là ước số. Nếu gọi xác suất còn sử dụng trong khoảng thời gian t là p_t (tức cột 6), thì $S(t)$ cũng có thể tính bằng công thức:

$$\hat{S}(t) = \prod_{t=1}^k p_t .$$

Phép ước tính được mô tả trên thường được gọi là *ước tính Kaplan-Meier* (Kaplan-Meier estimates), hay thỉnh thoảng cũng được gọi là *product-limit estimate*.

13.2 Ước tính Kaplan-Meier bằng R

Tất cả các tính toán trên, tất nhiên, có thể được tiến hành bằng R. Trong R có một package tên là `survival` (do Terry Therneau và Thomas Lumley phát triển) có thể ứng dụng để phân tích biến cố. Trong phần sau đây tôi sẽ hướng dẫn cách sử dụng package này.

Quay lại với Ví dụ 1, việc đầu tiên mà chúng ta cần làm là nhập dữ liệu vào R. Nhưng trước hết, chúng ta phải nhập package `survival` vào môi trường làm việc:

```
> library(survival)
```

Kế đến, chúng ta tạo ra hai biến số: biến thứ nhất gồm thời gian (hãy gọi là `weeks` cho trường hợp này), và biến thứ hai là chỉ số cho biết đối tượng ngưng sử dụng y cụ (cho giá trị 1) hay còn tiếp tục sử dụng (cho giá trị 0) và đặt tên biến này là `status`. Sau đó nhập hai biến vào một dataframe (và gọi là `data`) để tiện việc phân tích.

```
> weeks <- c(10, 13, 18, 19, 23, 30, 36, 38, 54,
   56, 59, 75, 93, 97, 104, 107, 107, 107)
> status <- c(1, 0, 0, 1, 0, 1, 1, 0, 0, 1, 1, 1, 1, 0, 1, 0, 0)
> data <- data.frame(duration, status)
```

Bây giờ, chúng ta đã sẵn sàng phân tích. Để ước tính Kaplan-Meier, chúng ta sẽ sử dụng hai hàm `Surv` và `survfit` trong package `survival`. Hàm `Surv` dùng để tạo ra một biến số hợp (combined variable) với thời gian và tình trạng. Ví dụ, trong lệnh sau đây:

```
> survtime <- Surv(weeks, status==1)
> survtime
[1] 10 13+ 18+ 19 23+ 30 36 38+ 54+ 56+ 59 75 93 97
[15] 104+ 107 107+ 107+
```

chúng ta sẽ có `survtime` là một biến với thời gian và dấu “+” (chỉ còn sống sót, hay censored observation, hay trong trường hợp này là còn sử dụng y cụ). Biến số này chỉ có giá trị và ý nghĩa cho phân tích của R, chứ trong thực tế, có lẽ chúng ta không cần nó.

Còn hàm `survfit` cũng khá đơn giản, chúng ta chỉ cần cung cấp hai thông số: thời gian và chỉ số như ví dụ sau đây:

```
> survfit(Surv(weeks, status==1))
```

Hay nếu đã có object `survtime` thì chúng ta chỉ đơn giản “gọi”:

```
> survfit(survtime)
Call: survfit(formula = survtime)
```

```
n events median 0.95LCL 0.95UCL
```

18 9 93 59 Inf

Kết quả trên đây chẳng có gì hấp dẫn, vì nó cung cấp những thông tin mà chúng ta đã biết: có 9 biến có (ngưng sử dụng y cụ) trong số 18 đối tượng. Thời gian (median - trung vị) ngưng sử dụng là 93 tuần, với khoảng tin cậy 95% từ 59 tuần đến vô cực (`Inf` = infinity). Để có thêm kết quả chúng ta cần phải đưa kết quả phân tích vào một object chặng hạn như `kp` và dùng hàm `summary` để biết thêm chi tiết:

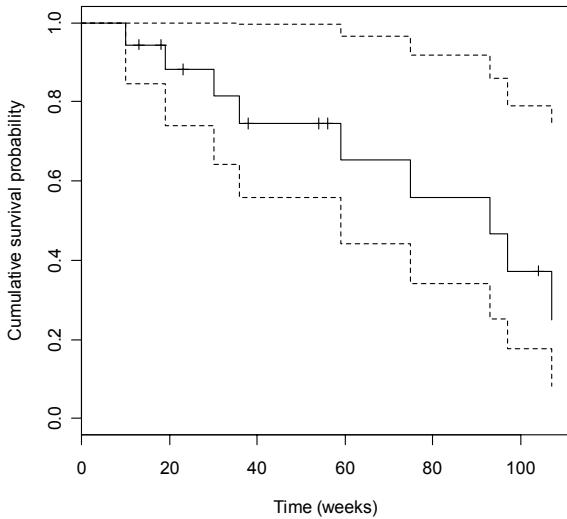
```
> kp <- survfit(Surv(weeks, status==1))
> summary(kp)
Call: survfit(formula = Surv(weeks, status == 1))
```

time	n.risk	n.event	survival	std.err	lower	95% CI	upper	95% CI
10	18	1	0.944	0.0540	0.844	1.000		
19	15	1	0.881	0.0790	0.739	1.000		
30	13	1	0.814	0.0978	0.643	1.000		
36	12	1	0.746	0.1107	0.558	0.998		
59	8	1	0.653	0.1303	0.441	0.965		
75	7	1	0.559	0.1412	0.341	0.917		
93	6	1	0.466	0.1452	0.253	0.858		
97	5	1	0.373	0.1430	0.176	0.791		
107	3	1	0.249	0.1392	0.083	0.745		

Một phần của kết quả này (cột `time`, `n.risk`, `n.event`, `survival`) chúng ta đã tính toán “thủ công” trong bảng trên. Tuy nhiên R còn cung cấp cho chúng ta sai số chuẩn (standard error) của $\hat{S}(t)$ và khoảng tin cậy 95%.

Khoảng tin cậy 95% được ước tính từ công thức $\hat{S}(t) \pm 1.96 \times se[\hat{S}(t)]$, mà trong đó, $se[\hat{S}(t)] = \hat{S}(t) \times \left\{ \sum_{t=1}^k \frac{d_t}{n_t(n_t - d_t)} \right\}$. Công thức sai số chuẩn này còn được gọi là *công thức Greenwood* (hay *Greenwood's formula*). Chúng ta có thể thể hiện kết quả trên bằng một biểu đồ bằng hàm `plot` như sau:

```
> plot(kp,
       xlab="Time (weeks)",
       ylab="Cumulative survival probability")
```



Trong biểu đồ trên, trục hoành là thời gian (tính bằng tuần) và trục tung là xác suất tích lũy còn sử dụng y cụ. Đường chính giữa chính là xác suất tích lũy $\hat{S}(t)$, hai đường chấm là khoảng tin cậy 95% của $\hat{S}(t)$. Qua kết quả phân tích này, chúng ta có thể phát biểu rằng xác suất sử dụng y cụ đến tuần 107 là khoảng 25% và khoảng tin cậy từ 8% đến 74.5%. Khoảng tin cậy khá rộng cho biết ước số có độ dao động cao, đơn giản vì số lượng đối tượng nghiên cứu còn tương đối thấp.

13.3 So sánh hai hàm xác suất tích lũy: kiểm định log-rank (log-rank test)

Phân tích trên chỉ áp dụng cho một nhóm đối tượng, và mục đích chính là ước tính $S(t)$ cho từng khoảng thời gian. Trong thực tế, nhiều nghiên cứu có mục đích so sánh $S(t)$ giữa hai hay nhiều nhóm khác nhau. Chẳng hạn như trong các nghiên cứu lâm sàng, nhất là nghiên cứu chữa trị ung thư, các nhà nghiên cứu thường so sánh thời gian sống sót giữa hai nhóm bệnh nhân để đánh giá mức độ hiệu nghiệm của một thuật điều trị.

Ví dụ 2. Một nghiên cứu trên 48 bệnh nhân với bệnh mụn giật (herpes) ở bộ phận sinh dục nhằm xét nghiệm hiệu quả của một loại vắc-xin mới (tạm gọi bằng mã danh gd2). Bệnh nhân được chia thành 2 nhóm một cách ngẫu nhiên: nhóm 1 được điều trị bằng gd2 (gồm 25 người), và 23 người còn lại trong nhóm hai nhận giả dược (placebo). Tình trạng bệnh được theo dõi trong vòng 12 tháng. Bảng số liệu sau đây trình bày thời gian (tính bằng tuần và gọi tắt là *time*) đến khi bệnh tái phát. Ngoài ra, mỗi bệnh nhân còn cung cấp số liệu về số lần bị nhiễm trong vòng 12 tháng trước khi tham gia công trình nghiên cứu (*episodes*). Theo kinh nghiệm lâm sàng, episodes có liên hệ mật thiết đến xác suất bị nhiễm (và chúng ta sẽ quay lại với cách phân tích biến số này một phần sau). Câu hỏi đặt ra là gd2 có hiệu nghiệm làm giảm nguy cơ bệnh tái phát hay không.

Bảng 13.1. Thời gian đến nhiễm trùng ở bệnh nhân với bệnh mụn giộp cho nhóm gd2 và giả dược

id	episodes	time	infected	id	episodes	time	infected
1	12	8	1	2	9	15	1
3	10	12	0	4	10	44	0
6	7	52	0	5	12	2	0
7	10	28	1	9	7	8	1
8	6	44	1	11	7	12	1
10	8	14	1	13	7	52	0
12	8	3	1	16	7	21	1
14	9	52	1	17	11	19	1
15	11	35	1	19	16	6	1
18	13	6	1	21	16	10	1
20	7	12	1	22	6	15	0
23	13	7	0	25	15	4	1
24	9	52	0	27	9	9	0
26	12	52	0	29	10	27	1
28	13	36	1	30	17	1	1
31	8	52	0	32	8	12	1
33	10	9	1	35	8	20	1
34	16	11	0	37	8	32	0
36	6	52	0	38	8	15	1
39	14	15	1	41	14	5	1
40	13	13	1	43	13	35	1
42	13	21	1	45	9	28	1
44	16	24	0	47	15	6	1
46	13	52	0				
48	9	28	1				

Chú thích: trong biến infected (nhiễm), 1 có nghĩa là bị nhiễm, và 0 là không bị nhiễm.

Trong trường hợp trên chúng ta có hai nhóm để so sánh. Một cách phân tích đơn giản là ước tính $S(t)$ cho từng nhóm và từng khoảng thời gian, rồi so sánh hai nhóm bằng một kiểm định thống kê thích hợp. Song, phương pháp phân tích này có nhược điểm là nó không cung cấp cho chúng ta một “bức tranh” chung của tất cả các khoảng thời gian. Ngoài ra, vấn đề so sánh giữa hai nhóm trong nhiều khoảng thời gian khác nhau làm cho kết quả rất khó diễn dịch.

Để khắc phục hai nhược điểm so sánh trên, một phương pháp phân tích được phát triển có tên là log-rank test (kiểm định log-rank). Đây là một phương pháp phân tích phi thông số để kiểm định giả thiết rằng hai nhóm có cùng $S(t)$. Phương pháp này cũng chia thời gian ra thành k khoảng thời gian, $t_1, t_2, t_3, \dots, t_k$, mà khoảng thời gian t_j ($j = 1, 2, 3, \dots, k$) phản ánh thời điểm j khi một hay nhiều đối tượng của hai nhóm cộng lại. Gọi d_{ij} là số bệnh nhân trong nhóm i ($i=1, 2$) bị bệnh trong khoảng thời gian t_j . Gọi $d_j = d_{1j} + d_{2j}$ là tổng số bệnh nhân mắc bệnh và đặt $n_j = n_{1j} + n_{2j}$ là tổng số bệnh nhân của hai nhóm trong khoảng thời gian t_j . Với $j = 1, 2, 3, \dots, k$, chúng ta có thể ước tính:

$$e_{1j} = \frac{n_{1j}d_j}{n_j} \quad \text{và} \quad e_{2j} = \frac{n_{2j}d_j}{n_j}$$

$$v_j = \frac{n_{1j}n_{2j}d_j(n_j - d_j)}{n_j^2(n_j - 1)}$$

(ở đây, e_{1j} , e_{2j} là số bệnh nhân trong nhóm 1 và 2 mà chúng ta tiên đoán là sẽ mắc bệnh nếu có cùng xác suất mắc bệnh trong cả hai nhóm (tức xác suất trung bình), v_j là phuơng sai). Ngoài ra, chúng ta có thể ước tính tổng số bệnh nhân mắc bệnh cho nhóm 1 và 2:

$$O_1 = \sum_{j=1}^k d_{1j} \quad \text{và} \quad O_2 = \sum_{j=1}^k d_{2j}$$

Và tổng số bệnh nhân mắc bệnh nếu có cùng chung xác suất mắc bệnh cho cả hai nhóm:

$$E_1 = \sum_{j=1}^k v_j \quad \text{và} \quad V = \sum_{j=1}^k v_j$$

Gọi T_i là một biến ngẫu nhiên phản ánh thời gian từ khi được điều trị đến khi mắc bệnh cho nhóm i , và gọi $S_i(t) = \Pr(T_i \geq t)$, kiểm định log-rank được định nghĩa như sau:

$$\chi^2 = \frac{(O_1 - E_1)^2}{V}$$

Nếu $\chi^2 > \chi^2_{1,\alpha}$ (trong đó, $\chi^2_{1,\alpha}$ là trị số Chi bình phuơng với độ ý nghĩa thống kê $\alpha=0.95$), chúng ta có bằng chứng để kết luận rằng độ khác biệt về $S(t)$ giữa hai nhóm có ý nghĩa thống kê.

13.4 Kiểm định log-rank bằng R

Ví dụ 2 (tiếp tục). Chúng ta quay lại với ví dụ 2 và sẽ sử dụng R để tính toán kiểm định log-rank. Trước hết, chúng ta phải nhập các dữ liệu cần thiết bằng các lệnh thông thường như sau:

```
> group <- c(1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
   1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
   2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2,
   2, 2, 2, 2, 2, 2, 2)

> episode <- c(12, 10, 7, 10, 6, 8, 8, 9, 11, 13, 7, 13, 9,
   12, 13, 8, 10, 16, 6, 14, 13, 13, 16, 13, 9,
   9, 10, 12, 7, 7, 7, 7, 11, 16, 16, 6, 15,
   9, 10, 17, 8, 8, 8, 8, 14, 13, 9, 15)

> time <- c(8, 12, 52, 28, 44, 14, 3, 52, 35, 6, 12, 7, 52,
   52, 36, 52, 9, 11, 52, 15, 13, 21, 24, 52, 28,
   15, 44, 2, 8, 12, 52, 21, 19, 6, 10, 15, 4, 9, 27, 1,
```

```

12,20,32,15, 5,35,28, 6)

> infected <- c(1, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 1,
   0, 1, 0, 0, 1, 1, 1, 0, 0, 1,
   1, 0, 0, 1, 1, 0, 1, 1, 1, 1, 0, 1, 0, 1, 1, 1,
   1, 1, 0, 1, 1, 1, 1)

```

```

> data <- data.frame(group, episode, time, infected)

```

(a) Chúng ta ứng dụng hàm `survfit` để ước tính xác suất tích lũy $S(t)$ cho từng nhóm bệnh nhân và cho kết quả vào đối tượng `kp.by.group` như sau (chú ý cách cung cấp thông số ~ `group`):

```

> library(survival)
> kp.by.group <- survfit(Surv(time, infected==1) ~ group)
> summary(kp.by.group)
Call: survfit(formula = Surv(time, infected == 1) ~ group)

```

group=1								
time	n.risk	n.event	survival	std.err	lower	95% CI	upper	95% CI
3	25	1	0.960	0.0392		0.886		1.000
6	24	1	0.920	0.0543		0.820		1.000
8	22	1	0.878	0.0660		0.758		1.000
9	21	1	0.836	0.0749		0.702		0.997
12	19	1	0.792	0.0829		0.645		0.973
13	17	1	0.746	0.0902		0.588		0.945
14	16	1	0.699	0.0958		0.534		0.915
15	15	1	0.653	0.1001		0.483		0.882
21	14	1	0.606	0.1033		0.434		0.846
28	12	2	0.505	0.1080		0.332		0.768
35	10	1	0.454	0.1083		0.285		0.725
36	9	1	0.404	0.1074		0.240		0.680
44	8	1	0.353	0.1052		0.197		0.633
52	7	1	0.303	0.1016		0.157		0.584

group=2								
time	n.risk	n.event	survival	std.err	lower	95% CI	upper	95% CI
1	23	1	0.957	0.0425		0.8767		1.000
4	21	1	0.911	0.0601		0.8004		1.000
5	20	1	0.865	0.0723		0.7346		1.000
6	19	2	0.774	0.0889		0.6183		0.970
8	17	1	0.729	0.0946		0.5650		0.940
10	15	1	0.680	0.1000		0.5099		0.907
12	14	2	0.583	0.1067		0.4072		0.835
15	12	2	0.486	0.1088		0.3132		0.754
19	9	1	0.432	0.1093		0.2630		0.709
20	8	1	0.378	0.1082		0.2156		0.662
21	7	1	0.324	0.1053		0.1712		0.613
27	6	1	0.270	0.1007		0.1300		0.561
28	5	1	0.216	0.0939		0.0921		0.506
35	3	1	0.144	0.0859		0.0447		0.463

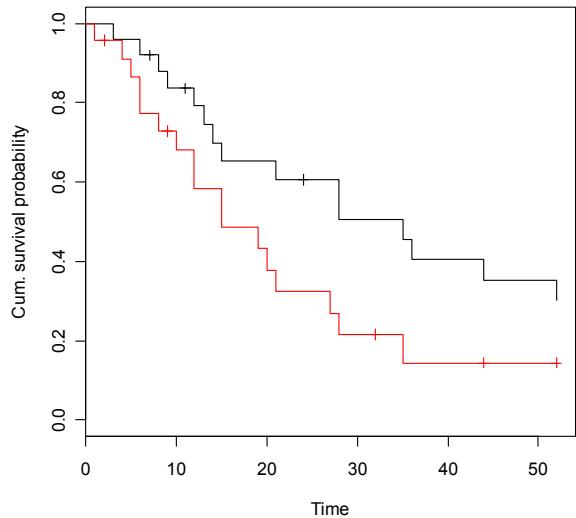
Và vẽ biểu đồ Kaplan-Meier cho từng nhóm như sau:

```
> plot(kp.by.group,
```

```

xlab="Time",
ylab="Cum. survival probability",
col=c("black", "red"))

```



Qua biểu đồ trên, chúng ta có thể thấy khá rõ là nhóm được điều trị bằng gd2 (đường màu đen phía trên) có xác suất nhiễm (hay bệnh tái phát) thấp hơn nhóm giả dược (đường màu đỏ, phía dưới). Nhưng phân tích trên không cung cấp trị số p để chúng ta phát biểu kết luận.

(b) Để có trị số p, chúng ta cần phải sử dụng hàm `survdiff` như sau:

```

> survdiff(Surv(time, infected==1) ~ group)
Call:
survdiff(formula = Surv(time, infected == 1) ~ group)

      N Observed Expected (O-E)^2/E (O-E)^2/V
group=1 25      15    20.0     1.26     3.65
group=2 23      17    12.0     2.11     3.65

  Chisq= 3.7 on 1 degrees of freedom, p= 0.056

```

Kết quả phân tích log-rank cho trị số p=0.056. Vì p > 0.05, chúng ta vẫn chưa có bằng chứng thuyết phục để kết luận rằng gd2 quả thật có hiệu nghiệm giảm nguy cơ tái phát bệnh.

13.5 Mô hình Cox (hay Cox's proportional hazards model)

Kiểm định log-rank là phương pháp cho phép chúng ta so sánh $S(t)$ giữa hai hay nhiều nhóm. Nhưng trong thực tế, $S(t)$ hay hàm nguy cơ $h(t)$ có thể không chỉ khác nhau giữa các nhóm, mà còn chịu sự chi phối của các yếu tố khác. Vấn đề đặt ra là làm sao ước tính mức độ ảnh hưởng của các yếu tố nguy cơ (risk factors) đến $h(t)$. Chẳng hạn

như trong nghiên cứu trên, số lần bệnh nhân từng bị nhiễm (biến `episode`) được xem là có ảnh hưởng đến nguy cơ bệnh tái phát. Do đó, vấn đề đặt ra là nếu chúng ta xem xét và điều chỉnh cho ảnh hưởng của `episode` thì mức độ khác biệt về $S(t)$ giữa hai nhóm có thật sự tồn tại hay không?

Vào khoảng giữa thập niên 1970s, David R. Cox, giáo sư thống kê học thuộc Đại học Imperial College (London, Anh) phát triển một phương pháp phân tích dựa vào mô hình hồi qui (regression) để trả lời câu hỏi trên (D.R. Cox, Regression models and life tables (with discussion), Journal of the Royal Statistical Society series B, 1972; 74:187-220). Phương pháp phân tích đó, sau này được gọi là *Mô hình Cox*. Mô hình Cox được đánh giá là một trong những phát triển quan trọng nhất của khoa học nói chung (không chỉ khoa học thống kê) trong thế kỉ 20! Không thể kể hết bao nhiêu số lần trích dẫn bài báo của David Cox, vì bài báo gây ảnh hưởng cho toàn bộ hoạt động nghiên cứu khoa học.

Vì mô tả chi tiết mô hình Cox nằm ngoài phạm vi của chương sách này, nên tôi chỉ phát họa vài nét chính để bạn đọc có thể nắm vấn đề. Gọi $x_1, x_2, x_3, \dots, x_p$ là p yếu tố nguy cơ. x có thể là các biến liên tục hay không liên tục. Mô hình Cox phát biểu rằng:

$$h(t) = \lambda(t) e^{\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \dots + \beta_p x_p}$$

$h(t)$ được định nghĩa như phần trên (tức hàm nguy cơ), β_j ($j = 1, 2, 3, \dots, p$) là hệ số ảnh hưởng liên quan đến x_j , và $\lambda(t)$ là hàm số nguy cơ nếu các yếu tố nguy cơ x không tồn tại (còn gọi là baseline hazard function). Vì mức độ ảnh hưởng của một yếu tố nguy cơ x_j thường được thể hiện bằng tỉ số nguy cơ (*hazard ratio*, HR, cũng tương tự như odds ratio trong phân tích hồi qui logistic), hệ số $\exp(\beta_j)$ chính là HR cho khi x_j tăng một đơn vị.

Hàm `coxph` trong package `R` có thể được ứng dụng để ước tính hệ số β_j . Trong lệnh sau đây:

```
> analysis <- coxph(Surv(time, infected==1) ~ group)
```

Trong lệnh trên, chúng ta muốn kiểm định ảnh hưởng của hai nhóm điều trị đến hàm nguy cơ $h(t)$ và kết quả được chứa trong đối tượng `analysis`. Để tóm lược `analysis`, chúng ta sử dụng hàm `summary`:

```
> summary(analysis)
Call:
coxph(formula = Surv(time, infected == 1) ~ group)

n= 48
      coef exp(coef)  se(coef)     z      p
group 0.684      1.98      0.363  1.88 0.06

      exp(coef) exp(-coef) lower .95 upper .95
group      1.98      0.505    0.973     4.04

Rsquare= 0.071   (max possible= 0.986 )
```

```

Likelihood ratio test= 3.55  on 1 df,    p=0.0597
Wald test            = 3.55  on 1 df,    p=0.0596
Score (logrank) test = 3.67  on 1 df,    p=0.0553

```

Nên nhớ nhóm điều trị được cho mã số 1, và nhóm giả dược có mã số 2. Do đó, kết quả phân tích trên cho biết khi group tăng 1 đơn vị thì $h(t)$ tăng 1.98 lần (với khoảng tin cậy 95% dao động từ 0.97 đến 4.04). Nói cách khác, nguy cơ bệnh tái phát trong nhóm giả dược cao hơn nhóm điều trị gd2 gần 2 lần. Tuy nhiên vì khoảng tin cậy 95% bao gồm cả 1 và trị số $p = 0.06$, cho nên chúng ta vẫn không thể kết luận rằng mức độ ảnh hưởng này có ý nghĩa thống kê.

Nhưng chúng ta cần phải xem xét (và điều chỉnh) cho ảnh hưởng của quá trình bệnh trong quá khứ được đo lường bằng biến số episode. Để tiến hành phân tích này, chúng ta cho thêm episode vào hàm coxph như sau:

```

> analysis <- coxph(Surv(time, infected==1) ~ group + episode)
> summary(analysis)
Call:
coxph(formula = Surv(time, infected == 1) ~ group + episode)

n= 48
      coef exp(coef)  se(coef)     z      p
group  0.874      2.40   0.3712  2.35 0.0190
episode 0.172      1.19   0.0648  2.66 0.0079

      exp(coef) exp(-coef) lower .95 upper .95
group      2.40      0.417     1.16      4.96
episode    1.19      0.842     1.05      1.35

Rsquare= 0.196  (max possible= 0.986 )
Likelihood ratio test= 10.5  on 2 df,    p=0.00537
Wald test            = 10.4  on 2 df,    p=0.00555
Score (logrank) test = 10.6  on 2 df,    p=0.00489

```

Kết quả phân tích trên cho chúng ta một diễn dịch khác và có lẽ chính xác hơn. Mô hình $h(t)$ bây giờ là:

$$h(t | group, episode) = \lambda(t) e^{0.874(group) + 0.172(episode)}$$

Nếu episode tạm thời giữ cố định, tỉ số $h(t)$ giữa hai nhóm là:

$$\frac{h(t | group = 2)}{h(t | group = 1)} = e^{0.874(2-1)} = 2.40$$

Tương tự, nếu group tạm thời giữ cố định, khi episode tăng một đơn vị, tỉ số nguy cơ sẽ tăng 1.14 lần.

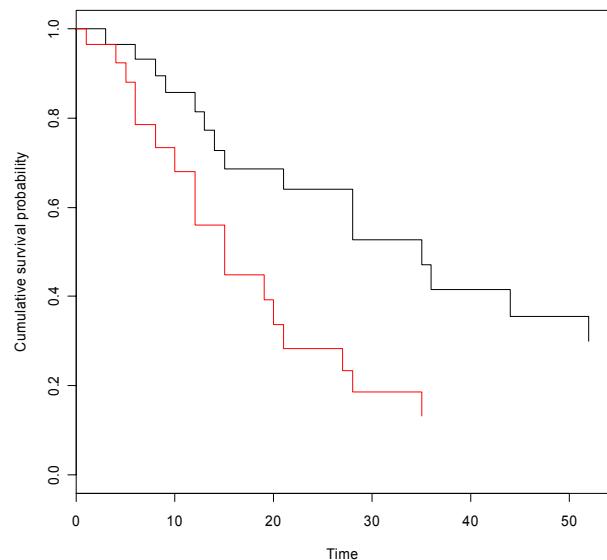
Nói cách khác, mỗi lần mắc bệnh trong quá khứ (tức episode tăng 1 đơn vị) làm tăng nguy cơ tái phát bệnh 19% (với khoảng tin cậy 95% dao động từ 5% đến 35%). Nhóm giả dược có nguy cơ bệnh tái phát tăng gấp 2.4 lần so với nhóm điều trị bằng gd2 (và khoảng tin cậy 95% có thể từ 1.2 đến gần 5 lần). Cả hai yếu tố (nhóm điều trị) và episode đều có ý nghĩa thống kê, vì trị số $p < 0.05$.

Nhưng episode là một biến liên tục. Vấn đề đặt ra là sau khi điều chỉnh cho episode thì hàm $S(t)$ cho từng nhóm sẽ ra sao? Cách khác quan nhất là giả định cả hai nhóm gd2 và giả dược có cùng số lần episode (như số trung bình chặng hạn), và hàm $S(t)$ cho từng nhóm có thể ước tính bằng:

```
> Cox.model <- survfit(coxph(Surv(time, infected==1)~episode+strata(group)))
> plot(Cox.model,
      xlab="Time",
      ylab="Cumulative survival probability",
      col=c("black", "red"))
```

hay đơn giản hơn:

```
> plot(survfit(coxph(Surv(time, infected==1)~episode+strata(group))),
      xlab="Time",
      ylab="Cumulative survival probability",
      col=c("black", "red"))
```



13.6 Xây dựng mô hình Cox bằng Bayesian Model Average (BMA)

Cũng như trường hợp của phân tích hồi qui tuyến tính đa biến và phân tích hồi qui logistic đa biến, vấn đề tìm một mô hình “tối ưu” để tiên đoán biến cố trong điều kiện có nhiều biến độc lập là một vấn đề nan giải. Phần lớn sách giáo khoa thống kê học

trình bày ba phương án chính để tìm một mô hình tối ưu: forward algorithm, backward algorithm, và tiêu chuẩn AIC.

Với phương án forward algorithm, chúng ta khởi đầu tìm biến độc lập x có ảnh hưởng lớn đến biến phụ thuộc y , rồi từng bước thêm các biến độc lập khác x cho đến khi mô hình không còn cải tiến thêm nữa.

Với phương án backward algorithm, chúng ta khởi đầu bằng cách xem xét tất cả biến độc lập x trong dữ liệu có thể có ảnh hưởng lớn đến biến phụ thuộc y , rồi từng bước loại bỏ từng biến độc lập x cho đến khi mô hình chỉ còn lại những biến có ý nghĩa thống kê.

Hai phương án trên (forward và backward algorithm) dựa vào phần dư (residual) và trị số P để xét một mô hình tối ưu. Một phương án thứ ba là dựa vào tiêu chuẩn Aikaike Information Criterion (AIC) mà tôi đã trình bày trong chương trước. Để hiểu phương pháp xây dựng mô hình dựa vào AIC tôi sẽ lấy một ví dụ thực tế như sau. Giả sử chúng ta muốn đi từ tỉnh A đến tỉnh B qua huyện C, và mỗi tuyến đường chúng ta có 3 lựa chọn: bằng xe hơi, bằng đường thủy, và bằng xe gắn máy. Tất nhiên, đi xe hơi đắt tiền hơn đi xe gắn máy. Mặt khác, đi đường thủy tuy ít tốn kém nhưng chậm hơn đi bằng xe hơi hay xe gắn máy. Nếu có tất cả 6 phương án đi, vấn đề đặt ra là chúng ta muốn tìm một phương án đi sao cho ít tốn kém nhất, nhưng tiêu ra một thời gian ngắn nhất! Tương tự, phương pháp xây dựng mô hình dựa vào tiêu chuẩn AIC là đi tìm một mô hình sao cho ít thông số nhất nhưng có khả năng tiên đoán biến phụ thuộc đầy đủ nhất.

Nhưng cả ba phương án trên có vấn đề là mô hình “tối ưu” nhất được xem là mô hình sau cùng, và tất cả suy luận khoa học đều dựa vào ước số của mô hình đó. Trong thực tế, bất cứ mô hình nào (kể cả mô hình “tối ưu”) cũng có độ bất định của nó, và khi chúng ta có thêm số liệu, mô hình tối ưu chưa chắc là mô hình sau cùng, và do đó suy luận có thể sai lầm. Một cách tốt hơn và có triển vọng hơn để xem xét đến yếu tố bất định này là Bayesian Model Average (BMA).

Với phân tích BMA, thay vì chúng ta hỏi yếu tố độc lập x ảnh hưởng đến biến phụ thuộc có ý nghĩa thống kê hay không, chúng ta hỏi: xác suất mà biến độc lập x có ảnh hưởng đến y là bao nhiêu. Để trả lời câu hỏi đó BMA xem xét tất cả các mô hình có khả năng giải thích y , và xem trong các mô hình đó, biến x xuất hiện bao nhiêu lần.

Ví dụ 3. Trong ví dụ sau đây, chúng ta sẽ mô phỏng một nghiên cứu với 5 biến độc lập x_1, x_2, x_3, x_4 , và x_5 . Ngoại trừ x_1 , 4 biến kia được mô phỏng theo luật phân phối chuẩn. Biến y là thời gian và kèm theo biến tử vong (death). Trong 5 biến x này, chỉ có biến x_1 có liên hệ với xác suất tử vong bằng mối liên hệ $\exp(3*x_1 + 1)$, còn các biến x_2, x_3, x_4 , và x_5 được mô phỏng toàn độc lập với nguy cơ tử vong. Chúng ta sẽ sử dụng phương pháp xây dựng mô hình theo tiêu chuẩn AIC và BMA để so sánh.

```
# Nhập package survival và BMA để phân tích
> library(survival)
> library(BMA)
```

```

# Tạo ra 5 biến số độc lập
> x1 <- (1:50)/2 - 3
> x2 <- rnorm(50)
> x3 <- rnorm(50)
> x4 <- rnorm(50)
> x5 <- rnorm(50)

# Mô phỏng mối liên hệ risk=exp(beta*x1 + 1)
> model <- exp(3*x1 + 1)

# Tạo ra biến số phụ thuộc y
> y <- rexp(50, rate = model)

# Tạo ra biến sự kiện theo luật phân phối mũ, tỉ lệ 0.3
> censored <- rexp(50, rate=0.3)
> ycencored <- pmin(y, censored)
> death <- as.numeric(y <= censored)

# Cho tất cả biến số vào data frame tên simdata
> simdata <- data.frame(y, death, x1,x2,x3,x4,x5)

# Phân tích bằng mô hình Cox
> cox <- coxph(Surv(y, death) ~ ., data=simdata)
> summary(cox)
Call:
coxph(formula = Surv(y, death) ~ ., data = simdata)

n= 50
      coef  exp(coef)  se(coef)      z      p
x1  3.2325   25.344    0.568  5.6908 1.3e-08
x2 -0.0319    0.969    0.331 -0.0963 9.2e-01
x3  0.3112   1.365    0.327  0.9518 3.4e-01
x4  0.1364   1.146    0.297  0.4600 6.5e-01
x5  0.4898   1.632    0.313  1.5643 1.2e-01

      exp(coef)  exp(-coef) lower .95 upper .95
x1     25.344     0.0395    8.325    77.16
x2      0.969     1.0324    0.506     1.85
x3     1.365     0.7326    0.719     2.59
x4     1.146     0.8725    0.641     2.05
x5     1.632     0.6127    0.883     3.01

Rsquare= 0.992 (max possible= 0.997 )
Likelihood ratio test= 241 on 5 df,  p=0
Wald test            = 33.3 on 5 df,  p=3.36e-06
Score (logrank) test = 107 on 5 df,  p=0

```

Kết quả trên cho thấy biến x1,x3 và x5 có ảnh hưởng có ý nghĩa thống kê đến biến y. Tất nhiên, đây làm một kết quả sai vì chúng ta biết rằng chỉ có x1 là có ý nghĩa thống kê mà thôi. Nay giờ chúng ta thử áp dụng cách xây dựng mô hình dựa vào tiêu chuẩn AIC:

```

# Tìm mô hình dựa vào tiêu chuẩn AIC
> searchAIC <- step(cox, direction="both")
> summary(searchAIC)
Call:
coxph(formula = Surv(y, death) ~ x1 + x5, data = simdata)

```

```

n= 50
      coef exp(coef)   se(coef)      z      p
x1 3.126    22.79     0.529  5.91 3.4e-09
x5 0.429     1.54     0.297  1.45 1.5e-01

      exp(coef) exp(-coef) lower .95 upper .95
x1     22.79     0.0439    8.080    64.27
x5     1.54     0.6510    0.858    2.75

Rsquare= 0.992 (max possible= 0.997 )
Likelihood ratio test= 240 on 2 df,  p=0
Wald test            = 35.3 on 2 df,  p=2.18e-08
Score (logrank) test = 104 on 2 df,  p=0

```

Kết quả này cho thấy x_1 và x_5 là hai yếu tố độc lập có ảnh hưởng có ý nghĩa thống kê đến biến y . Một lần nữa, kết quả này sai! Nay giờ chúng ta sẽ áp dụng phép tính BMA:

```

#tìm mô hình bằng phép tính BMA
> time <- simdata$y
> death <- simdata$death
> xvars <- simdata[,c(3,4,5,6,7)]
> bma <- bic.surv(xvars, time, death)
> summary(bma)
> imageplot.bma(bma)

Call:
bic.surv.data.frame(x = xvars, surv.t = time, cens = death)

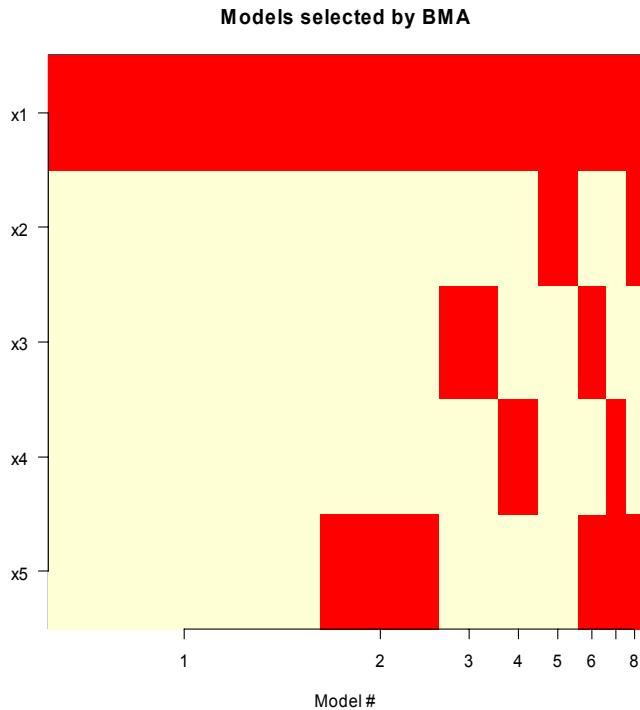
8 models were selected
Best 5 models (cumulative posterior probability = 0.8911 ):

      p!=0    EV    SD model 1 model 2 model 3 model 4 model 5
x1 100.0 3.0360 0.509 2.98048 3.12625 3.03900 2.98288 2.98098
x2 9.6 0.0008 0.096 .
x3 14.6 0.0410 0.155 .
x4 10.0 0.0063 0.092 .
x5 31.0 0.1349 0.261 .        0.42920 .

      nVar      1      2      2      2      2
      BIC -233.774 -232.126 -230.713 -229.933 -229.930
      post prob 0.458 0.201 0.099 0.067 0.067

```

Kết quả phân tích BMA cho thấy mô hình tối ưu là mô hình 1 chỉ có một biến có ý nghĩa thống kê: đó là biến x_1 . Xác suất mà yếu tố này có ảnh hưởng đến nguy cơ tử vong là 100%. Đây chính là kết quả mà chúng ta kì vọng, bởi vì chúng ta đã mô phỏng chỉ có x_1 có ảnh hưởng đến y mà thôi. Mô hình 2 có hai biến x_1 và x_5 (tức cũng chính là mô hình mà tiêu chuẩn AIC xác định), nhưng mô hình này chỉ có xác suất 0.201 mà thôi. Các mô hình 3(x_1 và x_3), mô hình 4 (x_1 và x_4) và mô hình 5 (x_1 và x_2) cũng có khả năng nhưng xác suất quá thấp (dưới 0.1) cho nên chúng ta không thể chấp nhận được. Biểu đồ sau đây thể hiện các kết quả trên:



Biểu đồ trên trình bày 8 mô hình, và trong tất cả 8 mô hình, biến x_1 xuất hiện một cách nhất quán (xác suất 100%). Còn các biến khác có ảnh hưởng nhưng không nhất quán. Qua so sánh giữa hai phương pháp xây dựng mô hình rõ ràng cho thấy cách phân tích BMA cung cấp cho chúng ta mô hình phù hợp đáng tin cậy nhất và có vẻ phù hợp với thực tế nhất.

Trên đây là những phương pháp phân tích biến có thông dụng nhất trong khoa học thực nghiệm với mô hình Cox và kiểm định log-rank. Mô hình Cox có thể khai triển thành những mô hình phức tạp và tinh vi hơn cho các nghiên cứu phức tạp với nhiều biến và tương tác giữa các yếu tố nguy cơ. Tài liệu hướng dẫn cách sử dụng package `survival` có thể giúp bạn đọc tìm hiểu sâu hơn. Tài liệu này có tại trang web www.cran.R-project.org.

14

Phân tích tổng hợp

Ông bà ta vẫn thường nói “*Một cây làm chằng nên non, ba cây chụm lại lên hòn núi cao*” để đề cao tinh thần hợp lực, đoàn kết nhằm hoàn tất một công việc quan trọng cần đến nhiều người. Trong nghiên cứu khoa học nói chung và y học nói riêng, nhiều khi chúng ta cần phải xem xét nhiều kết quả nghiên cứu từ nhiều nguồn khác nhau để giải quyết một vấn đề cụ thể.

14.1 Nhu cầu cho phân tích tổng hợp

Trong mấy năm gần đây, trong nghiên cứu khoa học xuất hiện khá nhiều nghiên cứu dưới danh mục “meta-analysis”, mà tôi tạm dịch là phân tích tổng hợp. Vậy phân tích tổng hợp là gì, mục đích là gì, và cách tiến hành ra sao ... là những câu hỏi mà rất nhiều bạn đọc muốn biết. Trong bài này tôi sẽ mô tả sơ qua vài khái niệm và cách tiến hành một phân tích tổng hợp, với hi vọng bạn đọc có thể tự mình làm một phân tích mà không cần đến các phần mềm đắt tiền.

Nguồn gốc và ý tưởng tổng hợp dữ liệu khởi đầu từ thế kỉ 17, chứ chẳng phải là một ý tưởng mới. Thời đó, các nhà thiên văn học nghĩ rằng cần phải hệ thống hóa dữ liệu từ nhiều nguồn để có thể đi đến một quyết định chính xác và hợp lí hơn các nghiên cứu riêng lẻ. Nhưng phương pháp phân tích tổng hợp hiện đại phải nói là bắt đầu từ hơn nửa thế kỉ trước trong ngành tâm lí học. Năm 1952, nhà tâm lí học trứ danh Hans J. Eysenck tuyên bố rằng tâm lí trị liệu (psychotherapy) chẳng có hiệu quả gì cả. Hơn hai mươi năm sau, năm 1976, Gene V. Glass, một nhà tâm lí học người Mĩ, muốn chứng minh rằng Eysenck sai, nên ông tìm cách thu thập dữ liệu của hơn 375 nghiên cứu về tâm lí trị liệu trong quá khứ, và tiến hành tổng hợp chúng bằng một phương pháp mà ông đặt tên là “meta-analysis” [1]. Qua phương pháp phân tích này, Glass tuyên bố rằng tâm lí trị liệu có hiệu quả và giúp ích cho bệnh nhân.

Phân tích tổng hợp – hay meta-analysis – từ đó được các bộ môn khoa học khác, nhất là y học, ứng dụng để giải quyết các vấn đề như hiệu quả của thuốc trong việc điều trị bệnh nhân. Cho đến nay, các phương pháp phân tích tổng hợp đã phát triển một bước dài, và trở thành một phương pháp chuẩn để thẩm định các vấn đề gai góc, các vấn đề mà sự nhất trí giữa các nhà khoa học vẫn chưa đạt được. Có người xem phân tích tổng hợp có thể cung cấp một câu trả lời sau cùng cho một câu hỏi y học. Người viết bài này không lạc quan và tự tin như thế, nhưng vẫn cho rằng phân tích tổng hợp là một phương pháp rất có ích cho chúng ta giải quyết những vấn đề còn trong vòng tranh cãi. Phân tích tổng hợp cũng có thể giúp cho chúng ta nhận ra những lĩnh vực nào cần phải nghiên cứu thêm hay cần thêm bằng chứng.

Kết quả của mỗi nghiên cứu đơn lẻ thường được đánh giá hoặc là “tích cực” (tức là, chẳng hạn như, thuật điều trị có hiệu quả), hoặc là “tiêu cực” (tức là thuật điều trị không có hiệu quả), và sự đánh giá này dựa vào trị số P. Thuật ngữ tiếng Anh gọi qui

trình đó là “significance testing” – thử nghiệm ý nghĩa thống kê. Nhưng ý nghĩa thống kê tùy thuộc vào số mẫu được chọn trong nghiên cứu, và một kết quả “tiêu cực” không có nghĩa là giả thiết của nghiên cứu sai, mà có thể đó là tín hiệu cho thấy số lượng mẫu chưa đầy đủ để đi đến một kết luận đáng tin cậy. Cái logic của phân tích tổng hợp, do đó, là chuyển hướng từ **significance testing** sang ước tính **effect size** - mức độ ảnh hưởng. Câu trả lời mà phân tích tổng hợp muốn đưa ra không chỉ đơn giản là có hay không có ý nghĩa thống kê (significant hay insignificant) mà là mức độ ảnh hưởng bao nhiêu, có đáng để chúng ta quan tâm, có thích hợp để chúng ta ứng dụng vào thực tế lâm sàng trong việc chăm sóc bệnh nhân.

14.2 Fixed-effects và Random-effects

Hai thuật ngữ mà bạn đọc thường gặp trong các phân tích tổng hợp là fixed-effects (tạm dịch là **ảnh hưởng bất biến**) và random-effects (**ảnh hưởng biến thiên**). Để hiểu hai thuật ngữ này tôi sẽ đưa ra một ví dụ tương đối đơn giản. Hãy tưởng tượng chúng ta muốn ước tính chiều cao của người Việt Nam trong độ tuổi trưởng thành (18 tuổi trở lên). Chúng ta có thể tiến hành 100 nghiên cứu tại nhiều địa điểm khác nhau trên toàn quốc; mỗi nghiên cứu chọn mẫu (samples) một cách ngẫu nhiên từ 10 người đến vài chục ngàn người; và cứ mỗi nghiên cứu chúng ta tính toán chiều cao trung bình. Như vậy, chúng ta có 100 số trung bình, và chắc chắn những con số này không giống nhau: một số nghiên cứu có chiều cao trung bình thấp, cao hay ... trung bình. Phân tích tổng hợp là nhằm mục đích sử dụng 100 số trung bình đó để ước tính chiều cao cho toàn thể người Việt. Có hai cách để ước tính: fixed-effects meta-analysis (phân tích tổng hợp ảnh hưởng bất biến) và random-effects meta-analysis (phân tích tổng hợp ảnh hưởng biến thiên) [2].

Phân tích tổng hợp ảnh hưởng bất biến xem sự khác biệt giữa 100 con số trung bình đó là do các yếu tố ngẫu nhiên liên quan đến mỗi nghiên cứu (còn gọi là within-study variance) gây nên. Cái giả định đằng sau cách nhận thức này là: nếu 100 nghiên cứu đó đều được tiến hành y chang nhau (như có cùng số lượng đối tượng, cùng độ tuổi, cùng tỉ lệ giới tính, cùng chế độ dinh dưỡng, v.v...) thì sẽ không có sự khác biệt giữa các số trung bình.

Nếu chúng ta gọi số trung bình của 100 nghiên cứu đó là x_1, x_2, \dots, x_{100} , quan điểm của phân tích tổng hợp ảnh hưởng bất biến cho rằng mỗi x_i là một biến số gồm hai phần: một phần phản ánh số trung của toàn bộ quần thể dân số (tạm gọi là M), và phần còn lại (khác biệt giữa x_i và M là một biến số e_i). Nói cách khác:

$$\begin{aligned}x_1 &= M + e_1 \\x_2 &= M + e_2 \\&\dots \\x_{100} &= M + e_{100}\end{aligned}$$

Hay nói chung là:

$$x_i = M + e_i$$

Tất nhiên e_i có thể <0 hay >0 . Nếu M và e_i độc lập với nhau (tức không có tương quan gì với nhau) thì phương sai của x_i (gọi là $\text{var}[x_i]$) có thể viết như sau:

$$\text{var}[x_i] = \text{var}[M] + \text{var}[e_i] = 0 + s_e^2$$

Chú ý $\text{var}[M] = 0$ vì M là một hằng số bất biến, s_e^2 là phương sai của e_i . Mục đích của phân tích tổng hợp là ước tính M và s_e^2 .

Phân tích tổng hợp ảnh hưởng biến thiên xem mức độ khác biệt (còn gọi là variance hay phương sai) giữa các số trung bình là do hai nhóm yếu tố gây nên: các yếu tố liên quan đến mỗi nghiên cứu (within-study variance) và các yếu tố giữa các nghiên cứu (between-study variance). Các yếu tố khác biệt giữa các nghiên cứu như địa điểm, độ tuổi, giới tính, dinh dưỡng, v.v... cần phải được xem xét và phân tích. Nói cách khác, phân tích tổng hợp ảnh hưởng biến thiên đi xa hơn phân tích tổng hợp ảnh hưởng bất biến một bước bằng cách xem xét đến những khác biệt giữa các nghiên cứu. Do đó, kết quả từ phân tích tổng hợp ảnh hưởng biến thiên thường “bảo thủ” hơn các phân tích tổng hợp ảnh hưởng bất biến.

Quan điểm của phân tích tổng hợp ảnh hưởng biến thiên cho rằng mỗi nghiên cứu có một giá trị trung bình cá biệt phải ước tính, gọi là m_i . Do đó, x_i là một biến số gồm hai phần: một phần phản ánh số trung của quần thể mà mẫu được chọn (m_i , chú ý ở đây có chỉ từ i để chỉ một nghiên cứu riêng lẻ i), và phần còn lại (khác biệt giữa x_i và m_i là một biến số e_i). Ngoài ra, phân tích tổng hợp ảnh hưởng biến thiên còn phát biểu rằng m_i dao động chung quanh số tổng trung bình M bằng một biến ngẫu nhiên ε_i . Nói cách khác:

$$x_i = m_i + e_i$$

Trong đó:

$$m_i = M + \varepsilon_i$$

Thành ra:

$$x_i = M + \varepsilon_i + e_i$$

Và phương sai của x_i bây giờ có hai thành phần:

$$\text{var}[x_i] = \text{var}[M] + \text{var}[\varepsilon_i] + \text{var}[e_i] = 0 + s_\varepsilon^2 + s_e^2$$

Như ta thấy qua công thức này, s_e^2 phản ánh độ dao động giữa các nghiên cứu (between-study variation), còn s_ϵ^2 phản ánh độ dao động trong mỗi nghiên cứu (within-study variation). Mục đích của phân tích tổng hợp ảnh hưởng biến thiên là ước tính M , s_e^2 và s_ϵ^2 .

Nói tóm lại, Phân tích tổng hợp ảnh hưởng bất biến và Phân tích tổng hợp ảnh hưởng biến thiên chỉ khác nhau ở phương sai. Trong khi phân tích tổng hợp bất biến xem $s_\epsilon^2 = 0$, thì phân tích tổng hợp biến thiên đặt yêu cầu phải ước tính s_ϵ^2 . Tất nhiên, nếu $s_\epsilon^2 = 0$ thì kết quả của hai phân tích này giống nhau. Trong bài này tôi sẽ tập trung vào cách phân tích tổng hợp ảnh hưởng bất biến.

14.3 Qui trình của một phân tích tổng hợp

Cũng như bất cứ nghiên cứu nào, một phân tích tổng hợp được tiến hành qua các công đoạn như: thu thập dữ liệu, kiểm tra dữ liệu, phân tích dữ liệu, và kiểm tra kết quả phân tích.

- Bước thứ nhất: sử dụng hệ thống thư viện y khoa PubMed hay một hệ thống thư viện khoa học của chuyên ngành để tìm những bài báo liên quan đến vấn đề cần nghiên cứu. Bởi vì có nhiều nghiên cứu, vì lí do nào đó (như kết quả “tiêu cực” chẳng hạn), không được công bố, cho nên nhà nghiên cứu có khi cũng cần phải thêm vào các nghiên cứu đó. Việc làm này tuy nói thì dễ, nhưng trong thực tế không dễ dàng chút nào!
- Bước thứ hai: rà soát xem trong số các nghiên cứu được truy tìm đó, có bao nhiêu đạt các tiêu chuẩn đã được đề ra. Các tiêu chuẩn này có thể là đối tượng bệnh nhân, tình trạng bệnh, độ tuổi, giới tính, tiêu chí, v.... Chẳng hạn như trong số hàng trăm nghiên cứu về ảnh hưởng của viatmin D đến loãng xương, có thể chỉ vài chục nghiên cứu đạt tiêu chuẩn như đối tượng phải là phụ nữ sau thời mãn kinh, mật độ xương thấp, phải là nghiên cứu lâm sàng đối chứng ngẫu nhiên (randomized controlled clinical trials - RCT), tiêu chí phải là gãy xương đùi, v.v... (Những tiêu chuẩn này phải được đề ra trước khi tiến hành nghiên cứu).
- Bước thứ ba: chiết số liệu và dữ kiện (data extraction). Sau khi đã xác định được đối tượng nghiên cứu, bước kế tiếp là phải lên kế hoạch chiết số liệu từ các nghiên cứu đó. Chẳng hạn như nếu là các nghiên cứu RCT, chúng ta phải tìm cho được số liệu cho hai nhóm can thiệp và đối chứng. Có khi các số liệu này không được công bố hay trình bày trong bài báo, và trong trường hợp đó, nhà nghiên cứu phải trực tiếp liên lạc với tác giả để tìm số liệu. Một bảng tóm lược kết quả nghiên cứu có thể tương tự như **Bảng 1** dưới đây.

- Bước thứ tư: tiến hành phân tích thống kê. Trong bước này, mục đích là ước tính mức độ ảnh hưởng chung cho tất cả nghiên cứu và độ dao động của ảnh hưởng đó. Trong bài này, tôi sẽ giải thích cụ thể cách làm.
- Bước thứ năm: xem xét các kết quả phân tích, và tính toán thêm một số chỉ tiêu khác để đánh giá độ tin cậy của kết quả phân tích.

Cũng như phân tích thống kê cho từng nghiên cứu riêng lẻ tùy thuộc vào loại tiêu chí (như là biến số liên tục – continuous variables – hay biến số nhị phân – dichotomous variables), phương pháp phân tích tổng hợp cũng tùy thuộc vào các tiêu chí của nghiên cứu. Tôi sẽ lần lược mô tả hai phương pháp chính cho hai loại biến số liên tục và nhị phân.

14.4 Phân tích tổng hợp ảnh hưởng bắt biến cho một tiêu chí liên tục (Fixed-effects meta-analysis for a continuous outcome).

14.4.1 Phân tích tổng hợp bằng tính toán “thủ công”

Ví dụ 1. Thời gian nằm viện để điều trị ở các bệnh nhân đột quỵ là một tiêu chí quan trọng trong việc xác định chính sách tài chính. Các nhà nghiên cứu muốn biết sự khác biệt về thời gian nằm viện giữa hai nhóm bệnh viện chuyên khoa và bệnh viện đa khoa. Các nhà nghiên cứu ra soát và thu thập số liệu từ 9 nghiên cứu như sau (xem **Bảng 1**). Một số nghiên cứu cho thấy thời gian nằm viện trong các bệnh viện chuyên khoa ngắn hơn các bệnh viện đa khoa (như nghiên cứu 1, 2, 3, 4, 5, 8), một số nghiên cứu khác cho thấy ngược lại (như nghiên cứu 7 và 9). Vấn đề đặt ra là các số liệu này có phù hợp với giả thiết bệnh nhân các bệnh viện đa khoa thường có thời gian nằm viện ngắn hơn các bệnh viện đa khoa hay không. Chúng ta có thể trả lời câu hỏi này qua các bước sau đây:

Bước 1: tóm lược dữ liệu trong một bảng thống kê như sau:

Bảng 1. Thời gian nằm bệnh viện của các bệnh nhân đột quỵ trong hai nhóm bệnh viện chuyên khoa và đa khoa

Nghiên cứu (<i>i</i>)	Bệnh viện chuyên khoa			Bệnh viện đa khoa		
	N _{1i}	LOS _{1i}	SD _{1i}	N _{2i}	LOS _{2i}	SD _{2i}
1	155	55	47	156	75	64
2	31	27	7	32	29	4
3	75	64	17	71	119	29
4	18	66	20	18	137	48
5	8	14	8	13	18	11
6	57	19	7	52	18	4
7	34	52	45	33	41	34
8	110	21	16	183	31	27

9	60	30	27	52	23	20
Tổng cộng	548			610		

Chú thích: Trong bảng này, i là chỉ số chỉ mỗi nghiên cứu, $i=1,2,\dots,9$. N_1 và N_2 là số bệnh nhân nghiên cứu cho từng nhóm bệnh viện; LOS_1 và LOS_2 (length of stay): thời gian trung bình nằm viện (tính bằng ngày); SD_1 và SD_2 : độ lệch chuẩn (standard deviation) của thời gian nằm viện.

Bước 2: ước tính mức độ khác biệt trung bình và phương sai (variance) cho từng nghiên cứu. Mỗi nghiên cứu ước tính một độ ảnh hưởng, hay nói chính xác hơn là khác biệt về thời gian nằm viện kí hiệu, và tôi sẽ đặt kí hiệu là d_i giữa hai nhóm bệnh viện. Chỉ số ảnh hưởng này chỉ đơn giản là:

$$d_i = LOS_{1i} - LOS_{2i}$$

Phương sai của d_i (tôi sẽ kí hiệu là s_i^2) được ước tính bằng một công thức chuẩn dựa vào độ lệch chuẩn và số đối tượng trong từng nghiên cứu. Với mỗi nghiên cứu i ($i = 1, 2, 3, \dots, 9$), chúng ta có:

$$s_i^2 = \frac{(N_{1i}-1)SD_{1i}^2 + (N_{2i}-1)SD_{2i}^2}{N_{1i}+N_{2i}-2} \left(\frac{1}{N_{1i}} + \frac{1}{N_{2i}} \right)$$

Chẳng hạn như với nghiên cứu 1, chúng ta có:

$$d_1 = 75 - 55 = 20$$

và phương sai của d_1 :

$$s_1^2 = \frac{(155-1)(47)^2 + (156-1)(64)^2}{155+156-2} \left(\frac{1}{155} + \frac{1}{156} \right) = 40.59$$

hay độ lệch chuẩn: $s_1 = \sqrt{40.59} = 6.37$

Với độ lệch chuẩn s_i chúng ta có thể ước tính khoảng tin cậy 95% (95% confidence interval hay 95%CI) cho d_i bằng lí thuyết phân phối chuẩn (Normal distribution). Cần nhắc lại rằng, nếu một biến số tuân theo định luật phân phối chuẩn thì 95% các giá trị của biến số sẽ nằm trong khoảng ± 1.96 lần độ lệch chuẩn. Do đó, khoảng tin cậy 95% cho mức độ khác biệt của nghiên cứu 1 là:

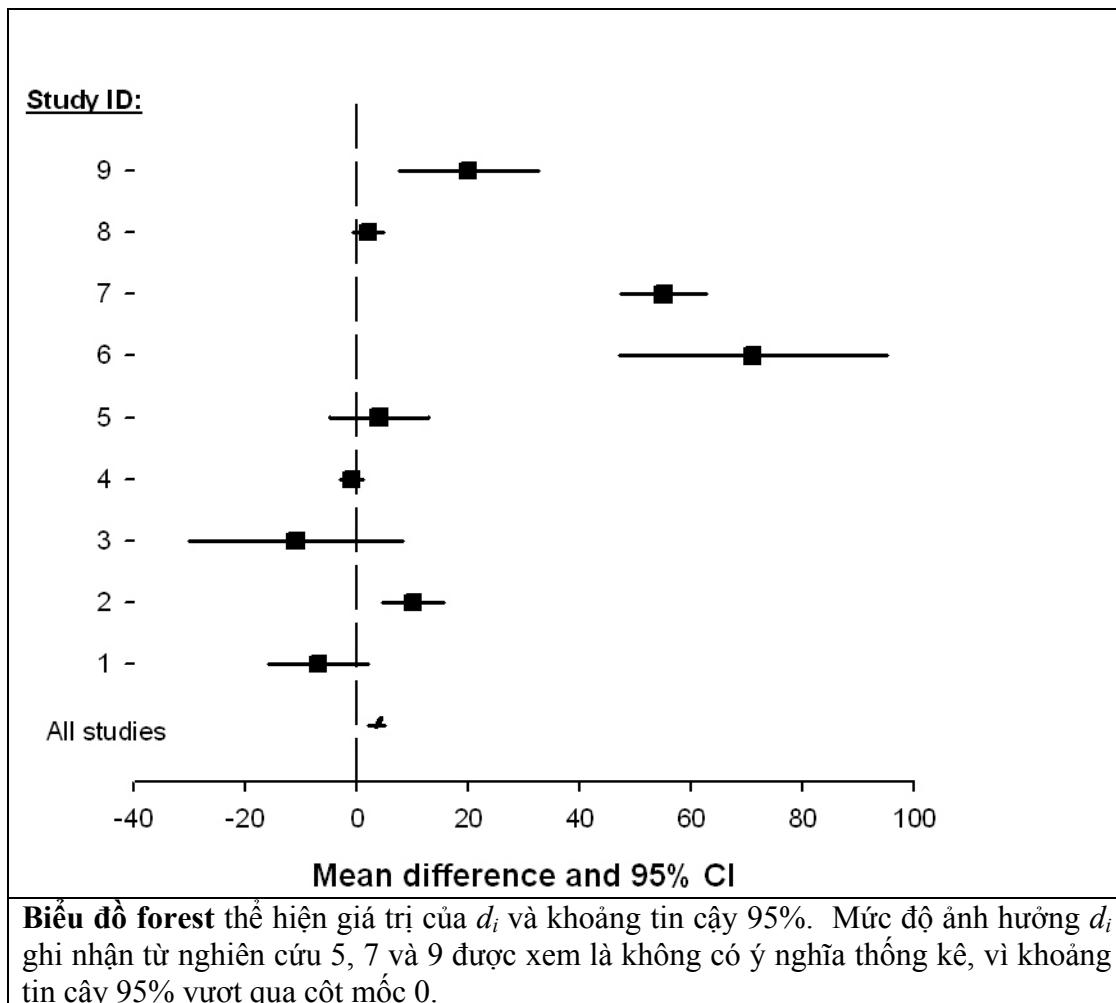
$$\begin{aligned} d_1 - 1.96 * s_1 &= 20 - 1.96 * 6.37 = 7.71 \text{ ngày} \\ \text{đến} \\ d_1 + 1.96 * s_1 &= 20 + 1.96 * 6.37 = 32.49 \text{ ngày} \end{aligned}$$

Tiếp tục tính như thế cho các nghiên cứu khác, chúng ta sẽ có thêm bốn cột trong bảng sau đây:

Bảng 1a. Độ khác biệt về thời gian giữa hai nhóm và khoảng tin cậy 95%

Nghiên cứu (i)	d_i	s_i^2	s_i	$d_i - 1.96 * s_i$	$d_i + 1.96 * s_i$
1	20	40.6	6.37	7.51	32.49
2	2	2.0	1.43	-0.80	4.80
3	55	15.3	3.91	47.34	62.66
4	71	150.2	12.26	46.98	95.02
5	4	20.2	4.49	-4.81	12.81
6	-1	1.2	1.11	-3.17	1.17
7	-11	95.4	9.77	-30.14	8.14
8	10	8.0	2.83	4.45	15.55
9	-7	20.7	4.55	-15.92	1.92

Đến đây chúng ta có thể thể hiện mức độ ảnh hưởng d_i và khoảng tin cậy 95% trong một biểu đồ có tên là “**forest plot**” như sau:



Bước 3: ước tính “trọng số” (weight) cho mỗi nghiên cứu. Trọng số (W_i) thực ra chỉ là số đảo của phuong sai s_i^2 ,

$$W_i = 1/s_i^2$$

Chẳng hạn như với nghiên cứu 1, chúng ta có: $W_1 = \frac{1}{40.59} = 0.0246$

Và chúng ta có thêm một cột mới cho bảng trên như sau:

Bảng 1b. Trọng số (weight) cho từng nghiên cứu

Nghiên cứu	d_i	s_i^2	W_i
1	20	40.6	0.0246
2	2	2.0	0.4886
3	55	15.3	0.0654
4	71	150.2	0.0067
5	4	20.2	0.0495
6	-1	1.2	0.8173
7	-11	95.4	0.0105
8	10	8.0	0.1245
9	-7	20.7	0.0483
Tổng số			1.6354

Bước 4: ước tính trị số trung bình của d cho tất cả các nghiên cứu. Chúng ta có thể đơn giản tính trung bình d bằng cách cộng tất cả d_i và chia cho 9, nhưng cách tính như thế không khách quan, bởi vì mỗi giá trị d_i có một phuong sai và trọng số (W_i) cá biệt. Chẳng hạn như nghiên cứu 4, vì phuong sai cao nhất (150.2), chúng tố rằng nghiên cứu này có số đối tượng ít hay độ dao động rất cao, và độ dao động cao có nghĩa là chúng ta không đặt “niềm tin cậy” vào đó cao được. Chính vì thế mà trọng số cho nghiên cứu này rất thấp, chỉ 0.0067. Ngược lại, nghiên cứu 6 có trọng số cao vì độ dao động thấp (phuong sai thấp) và ước tính ảnh hưởng của nghiên cứu này có “trọng lượng” hơn các nghiên cứu khác trong nhóm.

Do đó, để tính trung bình d cho tổng số nghiên cứu, chúng ta phải xem xét đến trọng số W_i . Với mỗi d_i và W_i chúng ta có thể tính trị số **trung bình trọng số** (weighted mean) theo phương pháp chuẩn như sau:

$$d = \frac{\sum_{i=1}^9 W_i d_i}{\sum_{i=1}^9 W_i}$$

Bất cứ một ước tính thông kê (estimate) nào cũng phải có một phuơng sai. Và trong trường hợp d , phuơng sai (tôi sē kí hiệu là s_d^2) chỉ đơn giản là số đảo của tòng trọng số W_i :

$$s_d^2 = \frac{1}{\sum_{i=1}^9 W_i}$$

Sai số chuẩn (standard error, SE) của d , do đó là: $SE(d) = s_d$. Theo lí thuyết phân phối chuẩn (Normal distribution), khoảng tin cậy 95% (95% confidence interval, 95%CI) có thể được ước tính như sau:

$$95\%CI \text{ của } d = d \pm 1.96(s_d)$$

Để tính d chúng ta cần thêm một cột nữa: đó là cột $W_i d_i$. Chẳng hạn như với nghiên cứu 1, chúng ta có $W_1 d_1 = 0,0246 \times 20 = 0,4928$. Tiếp tục như thế, chúng ta có thêm một cột.

Bảng 1c. Tính toán trị số trung bình

Nghiên cứu	d_i	s_i^2	W_i	$W_i d_i$
1	20	40.6	0.0246	0.4928
2	2	2.0	0.4886	0.9771
3	55	15.3	0.0654	3.5993
4	71	150.2	0.0067	0.4726
5	4	20.2	0.0495	0.1981
6	-1	1.2	0.8173	-0.8173
7	-11	95.4	0.0105	-0.1153
8	10	8.0	0.1245	1.2450
9	-7	20.7	0.0483	-0.3383
Tổng số			1.6354	5.7140

Sau đó, cộng tất cả W_i và $W_i d_i$ (trong hàng “Tổng số” của bảng trên). Như vậy, trị số trung bình trọng số của d là:

$$d = \frac{\sum_{i=1}^9 W_i d_i}{\sum_{i=1}^9 W_i} = \frac{0.4928 + 0.9771 + \dots - 0.3383}{0.0246 + 0.4886 + \dots + 0.0483} = \frac{5.7140}{1.6354} = 3.49.$$

Và phuơng sai của d là: $s_d^2 = \frac{1}{1.6345} = 0.61$.

Nói cách khác, sai số chuẩn (standard error) của d là: $s_d = \sqrt{0.61} = 0.782$.

Khoảng tin cậy 95% (95% confidence interval hay 95%CI) có thể được ước tính như sau:

$$3.49 \pm 1.96 * 0.782 = 1.96 \text{ đến } 5.02.$$

Đến đây, chúng ta có thể nói rằng, tính trung bình, thời gian nằm viện tại các bệnh viện đa khoa dài hơn các bệnh viện chuyên khoa 3.49 ngày và 95% khoảng tin cậy là từ 1.96 ngày đến 5.02 ngày.

Bước 5: ước tính chỉ số đồng nhất (homogeneity) và bất đồng nhất (heterogeneity) giữa các nghiên cứu [3]. Trong thực tế, đây là chỉ số đo lường độ khác biệt giữa mỗi nghiên cứu và trị số trung bình trọng số. Chỉ số đồng nhất (index of homogeneity) được tính theo công thức sau đây:

$$Q = \sum_{i=1}^k W_i (d_i - d)^2$$

Ở đây, k là số nghiên cứu (trong ví dụ trên $k = 9$). Theo lí thuyết xác suất, Q có độ phân phối theo luật *Chi-square* với **bậc tự do (degrees of freedom – df)** là $k-1$ (tức là χ_{k-1}^2). Nói cách khác, nếu Q lớn hơn χ_{k-1}^2 thì đó là tín hiệu cho thấy sự bất đồng nhất giữa các nghiên cứu “có ý nghĩa thống kê” (significant).

Nhiều nghiên cứu trong thời gian qua chỉ ra rằng Q thường không phát hiện được sự bất đồng nhất một cách nhất quán, cho nên ngày nay ít ai dùng chỉ số này trong phân tích tổng hợp. Một chỉ số khác thay thế Q có tên là **index of heterogeneity (I^2)** mà tôi tạm dịch là **chỉ số bất đồng nhất**, nhưng sẽ giữ cách viết I^2 . Chỉ số này được định nghĩa như sau:

$$I^2 = \frac{Q - (k - 1)}{Q}$$

I^2 có giá trị từ âm đến 1. Nếu $I^2 < 0$, thì chúng ta sẽ cho nó là 0; nếu I^2 gần bằng 1 thì đó là dấu hiệu cho thấy có sự bất đồng nhất giữa các nghiên cứu.

Trong ví dụ trên, để ước tính Q và I^2 , chúng ta cần tính $W_i (d_i - d)^2$ cho từng nghiên cứu. Chẳng hạn như, với nghiên cứu 1:

$$W_1 (d_1 - d)^2 = 0,0246 * (20 - 3.49)^2 = 6,7129$$

Bảng 1d. Tính toán các chỉ số đồng nhất và bất đồng nhất

Nghiên cứu	d_i	s_i^2	W_i	$W_i d_i$	$W_i (d_i - d)^2$
1	20	40.6	0.0246	0.4928	6.7129
2	2	2.0	0.4886	0.9771	1.0903

3	55	15.3	0.0654	3.5993	173.6080
4	71	150.2	0.0067	0.4726	30.3356
5	4	20.2	0.0495	0.1981	0.0127
6	-1	1.2	0.8173	-0.8173	16.5054
7	-11	95.4	0.0105	-0.1153	2.2026
8	10	8.0	0.1245	1.2450	5.2701
9	-7	20.7	0.0483	-0.3383	5.3215
Tổng số			1.6354	5.7140	241.05

Sau khi đã ước tính $W_i(d_i - d)^2$ cho từng nghiên cứu, chúng ta cộng lại số này (xem cột sau cùng) và đó chính là Q :

$$Q = \sum_{i=1}^k W_i(d_i - d)^2 = 241.05$$

Từ đó, I^2 có thể ước tính như sau:

$$I^2 = \frac{241.05 - 8}{241.05} = 0.966$$

Chỉ số bất đồng nhất I^2 rất cao, cho thấy độ dao động về d_i giữa các nghiên cứu rất cao. Điều này chúng ta có thể thấy được chỉ qua nhìn vào cột số 2 trong bảng thống kê trên.

Bước 6: đánh giá khả năng publication bias [4]. Publication bias (tạm dịch: **trong thiên vị**) là một khái niệm tương đối mới có thể giải thích bằng tình huống thực tế sau đây. Chúng ta biết rằng khi một nghiên cứu cho ra kết quả “negative” (kết quả tiêu cực, tức là không phát hiện một ảnh hưởng hay một mối liên hệ có ý nghĩa thống kê) công trình nghiên cứu đó rất khó có cơ hội được công bố trên các tạp san, bởi vì giới chủ bút tạp san nói chung không thích in những bài như thế. Ngược lại, một nghiên cứu với một kết quả “tích cực” (tức có ý nghĩa thống kê) thì nghiên cứu có khả năng xuất hiện trên các tạp san khoa học cao hơn là các nghiên cứu với kết quả “tiêu cực”. Thế nhưng phần lớn những phân tích tổng hợp lại dựa vào các kết quả đã công bố trên các tạp san khoa học. Do đó, ước tính của một phân tích tổng hợp có khả năng thiêu khách quan, vì chưa xem xét đầy đủ đến các nghiên cứu tiêu cực chưa bao giờ công bố.

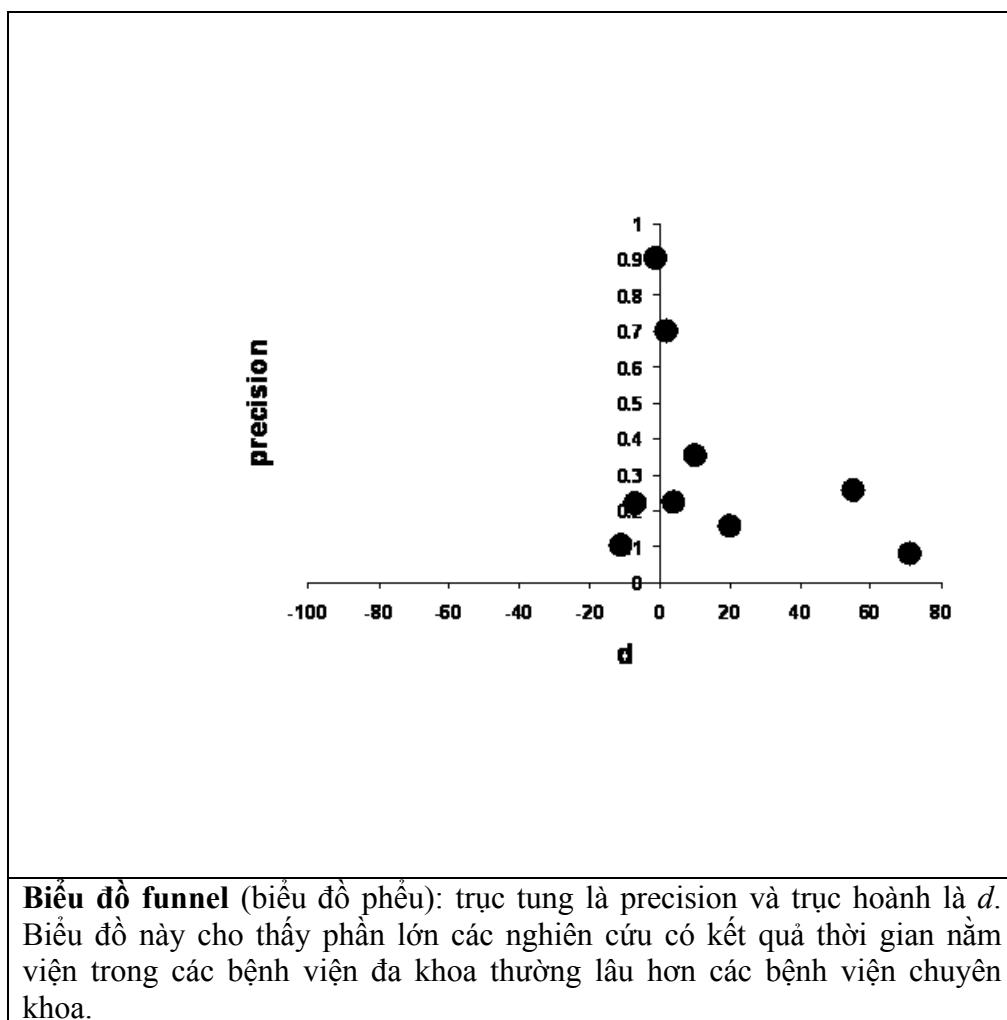
Một số nhà nghiên cứu đề nghị dùng biểu đồ funnel (còn gọi là **funnel plot**) để kiểm tra khả năng publication bias. Biểu đồ funnel được thể hiện bằng cách vẽ **độ chính xác – precision** (trục tung, y-axis) với ước tính mức độ ảnh hưởng cho từng nghiên cứu. Ở đây precision được định nghĩa là số đảo của sai số chuẩn (standard error):

$$\text{precision} = \frac{1}{s_{di}}$$

Nói cách khác, biểu đồ funnel biếu diễn *precision* với d_i . Chẳng hạn như với nghiên cứu 1, chúng ta có: $\text{precision} = 1/\sqrt{40,6} = 0,157$. Tính cho từng nghiên cứu, chúng ta có dùng bảng thống kê sau để vẽ biểu đồ funnel như sau:

Bảng 1e. Uớc tính publication bias

Nghiên cứu	d_i	s_i^2	$1/s_i$
1	20	40.6	0.1570
2	2	2.0	0.6990
3	55	15.3	0.2558
4	71	150.2	0.0816
5	4	20.2	0.2225
6	-1	1.2	0.9041
7	-11	95.4	0.1024
8	10	8.0	0.3528
9	-7	20.7	0.2198



Cái logic đằng sau biểu đồ funnel là nếu các công trình nghiên cứu lớn (tức có độ precision cao) có khả năng được công bố cao, thì số lượng nghiên cứu với kết quả tích cực sẽ nhiều hơn số lượng nghiên cứu nhỏ hay với kết quả tiêu cực trong các tập san. Và nếu điều này xảy ra, thì biểu đồ funnel sẽ thể hiện một sự thiếu cân đối (asymmetry). Nói cách khác, sự thiếu cân đối của một biểu đồ funnel là dấu hiệu cho thấy có vấn đề về publication bias. Nhưng vấn đề đặt ra là publication bias đó có ý nghĩa thống kê hay không? Biểu đồ funnel không thể trả lời câu hỏi này, chúng ta cần đến các phương pháp phân tích định lượng nghiêm chỉnh hơn.

Nghiệm toán Egger

Vài năm gần đây có ý kiến cho rằng biểu đồ funnel rất khó diễn dịch, và có thể gây nên ngộ nhận về publication bias [5-6]. Thật vậy, một số tạp san y học có chính sách khuyến khích các nhà nghiên cứu tìm một phương pháp khác để đánh giá publication bias thay vì dùng biểu đồ funnel.

Một trong những phương pháp đó là nghiệm toán Egger (còn gọi là **Egger's test**). Với phương pháp này, chúng ta mô hình rằng $SND = a + b \times precision$, trong đó SND được ước tính bằng cách lấy d chia cho sai số chuẩn của d , tức là: $SND_i = \frac{d_i}{S_{di}}$, a và b là hai thông số phải ước tính từ mô hình hồi qui đường thẳng đó. Ở đây, a cung cấp cho chúng ta một ước số về tình trạng thiếu cân đối của biểu đồ funnel: $a > 0$ có nghĩa là xu hướng nghiên cứu càng có qui mô lớn càng có ước số về độ ảnh hưởng với sự chính xác cao.

Trong ví dụ trên, chúng ta có thể dùng một phần mềm phân tích thống kê (như SAS hay R) để ước tính a và b như sau:

$$SND_i = 4.20 + -4.17084 * precision_i$$

Kết quả ước số $a = 4.20$ tuy là > 0 nhưng không có ý nghĩa thống kê, cho nên ở đây bằng chứng cho thấy không có sự publication bias.

Tuy nhiên, như đã thấy trong thực tế, nghiệm toán Egger này cũng chỉ là một cách thể hiện biểu đồ funnel mà thôi, chứ cũng không có thay đổi gì lớn. Có một cách đánh giá publication bias, cho đến nay, được xem là đáng tin cậy nhất: đó là phương pháp phân tích hồi qui đường thẳng (linear regression) giữa d_i và tổng số mẫu (N_i). Nói cách khác, chúng ta tìm a và b trong mô hình [7]:

$$d_i = a + b * N_i$$

Nếu không có publication bias thì giá trị của b sẽ rất gần với 0 hay không có ý nghĩa thống kê. Nếu trị số b khác với 0 thì đó là một tín hiệu của publication bias. Trong ví dụ vừa nêu với dữ liệu sau đây,

Nghiên cứu	d_i	N_i
------------	-------	-------

1	20	311
2	2	63
3	55	146
4	71	36
5	4	21
6	-1	109
7	-11	67
8	10	293
9	-7	112

chúng ta có phương trình:

$$d_i = 16.0 - 0.0009*/N_i$$

và quả thật giá trị của b quá thấp (cũng như không có ý nghĩa thống kê), cho nên đến đây chúng ta có thể kết luận rằng không có vấn đề publication bias trong nghiên cứu vừa đề cập đến.

Nói tóm lại, qua phân tích tổng hợp này, chúng ta có bằng chứng đáng tin cậy để kết luận rằng thời gian nằm viện của bệnh nhân trong các bệnh viện đa khoa dài hơn các bệnh viện chuyên khoa khoảng 3 ngày rưỡi, hoặc trong 95% trường hợp thời gian khác biệt khoảng từ 2 ngày đến 5 ngày. Kết quả này cũng cho thấy không có thiên vị xuất bản (publication bias) trong phân tích.

14.4.2 Phân tích tổng hợp bằng R

R có hai package được viết và thiết kế cho phân tích tổng hợp. Package được sử dụng khá thông dụng là meta. Bạn đọc có thể tải miễn phí từ trang web của R (trong phần packages): <http://cran.R-project.org>.

Để phân tích tổng hợp bằng R chúng ta phải nhập package meta vào môi trường vận hành của R (với điều kiện, tất nhiên, là bạn đọc đã tải và cài đặt meta vào R).

```
> library(meta)
```

Sau đó, chúng ta sẽ nhập số liệu trong ví dụ 1 vào R biến như sau:

- Nhập dữ liệu cho từng cột trong **Bảng 1** và cho vào một dataframe gọi là los:

```
> n1 <- c(155,31,75,18,8,57,34,110,60)
> los1 <- c(55,27,64,66,14,19,52,21,30)
> sd1 <- c(47,7,17,20,8,7,45,16,27)

> n2 <- c(156,32,71,18,13,52,33,183,52)
> los2 <- c(75,29,119,137,18,18,41,31,23)
> sd2 <- c(64,4,29,48,11,4,34,27,20)
```

```
> los <- data.frame(n1,los1,sd1,n2,los2,sd2)
```

- Sử dụng hàm metacont (dùng để phân tích các biến liên tục – do đó cont=continuous variable) và cho kết quả vào đối tượng res:

```
> res <- metacont(n1,los1,sd1,n2,los2,sd2,data=los)
> res
> res
      WMD          95%-CI %W(fixed) %W(random)
1 -20 [-32.4744; -7.5256]    1.44     10.69
2 -2 [-4.8271; 0.8271]    28.11    12.67
3 -55 [-62.7656; -47.2344]   3.73    11.89
4 -71 [-95.0223; -46.9777]   0.39     7.39
5 -4 [-12.1539; 4.1539]    3.38    11.80
6  1 [-1.1176; 3.1176]    50.11    12.72
7 11 [-8.0620; 30.0620]   0.62     8.76
8 -10 [-14.9237; -5.0763]   9.27    12.41
9  7 [-1.7306; 15.7306]   2.95    11.67
```

Number of trials combined: 9

	WMD	95%-CI	z	p.value
Fixed effects model	-3.4636	[-4.9626; -1.9646]	-4.5286	< 0.0001
Random effects model	-13.9817	[-24.0299; -3.9336]	-2.7272	0.0064

Quantifying heterogeneity:

$\tau^2 = 205.4094$; $H = 5.46$ [4.54; 6.58]; $I^2 = 96.7\%$ [95.2%; 97.7%]

Test of heterogeneity:

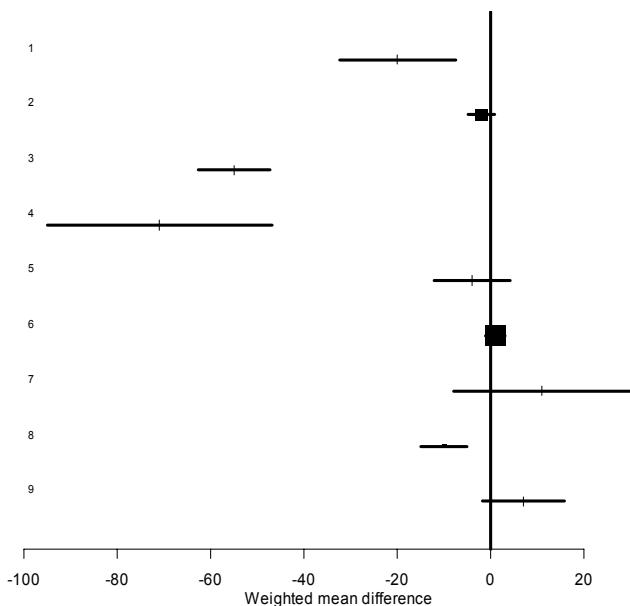
Q	d.f.	p.value
238.92	8	< 0.0001

Method: Inverse variance method

meta cung cấp cho chúng ta hai kết quả: một kết quả dựa vào mô hình fixed-effects và một dựa vào mô hình random-effects. Như thấy qua kết quả trên, mức độ khác biệt giữa hai mô hình khá lớn, nhưng kết quả chung thì giống nhau, tức kết quả của cả hai mô hình đều có ý nghĩa thống kê.

Chúng ta tất nhiên cũng có thể sử dụng hàm plot để thể hiện kết quả trên bằng biểu đồ forest như sau:

```
> plot(res, lwd=3)
```



14.5. Phân tích tổng hợp ảnh hưởng bất biến cho một tiêu chí nhị phân (Fixed-effects meta-analysis for a dichotomous outcome).

Trong phần trên, tôi vừa mô tả những bước chính trong một phân tích tổng hợp những nghiên cứu mà tiêu chí là một biến liên tục (continuous variable). Đối với các biến liên tục, trị số trung bình và độ lệch chuẩn là hai chỉ số thống kê thường được sử dụng để tóm lược. Nhưng hai chỉ số này không thể ứng dụng cho những tiêu chí mang tính thể loại hay thứ bậc như tử vong, gãy xương, v.v... vì những tiêu chí này chỉ có hai giá trị: hoặc là có, hoặc là không. Một người hoặc là còn sống hay chết, bị gãy xương hay không gãy xương, mắc bệnh suy tim hay không mắc bệnh suy tim, v.v... Đối với những biến này, chúng ta cần một phương pháp phân tích khác với phương pháp dành cho các biến liên tục.

14.5.1 Mô hình phân tích

Đối với những tiêu chí nhị phân (chỉ có hai giá trị), chỉ số thống kê tương đương với trị số trung bình là **tỉ lệ** hay **proportion**, có thể tính phần trăm); và chỉ số tương đương với độ lệch chuẩn là sai số chuẩn (**standard error**). Chẳng hạn như nếu một nghiên cứu theo dõi 25 bệnh nhân trong một thời gian, và trong thời gian đó có 5 bệnh nhân mắc bệnh, thì tỉ lệ (kí hiệu là p) đơn giản là: $p = 5/25 = 0,20$ (hay 20%). Theo lí thuyết xác suất, phương sai của p (kí hiệu là $\text{var}[p]$) là: $\text{var}[p] = p(1-p)/n = 0,2*(1 - 0,8)/25 = 0,0064$. Theo đó, sai số chuẩn của p (kí hiệu $\text{SE}[p]$) là:

$SE[p] = \sqrt{\text{var}[p]} = \sqrt{0,0064} = 0,08$. Chúng ta còn có thể ước tính khoảng tin cậy 95% của tỉ lệ như sau: $p \pm 1,96 \times SE[p] = 0,2 \pm 1,96 \times 0,08 = 0,04$ đến 0,36.

Vì cách tính của các tiêu chí nhị phân khá đặc thù, cho nên phương pháp phân tích tổng hợp các nghiên cứu với biến nhị phân cũng khác. Để minh họa cách phân tích tổng hợp dạng này, tôi sẽ lấy một ví dụ (phỏng theo một nghiên cứu có thật).

Ví dụ 2: Beta-blocker (sẽ viết tắt là BB) là một loại thuốc có chức năng điều trị và phòng chống cao huyết áp. Có giả thiết cho rằng BB cũng có thể phòng chống bệnh suy tim, hay ít ra là làm giảm nguy cơ suy tim. Để thử nghiệm giả thiết này, hàng loạt nghiên cứu lâm sàng đối chứng ngẫu nhiên đã được tiến hành trong thời gian 20 năm qua. Mỗi nghiên cứu có 2 nhóm bệnh nhân: nhóm được điều trị bằng BB, và một nhóm không được điều trị (còn gọi là placebo hay giả dược). Trong thời gian 2 năm theo dõi, các nhà nghiên cứu xem xét tần số tử vong cho từng nhóm. **Bảng 2** sau đây tóm lược 13 nghiên cứu trong quá khứ:

Bảng 2. Beta-blocker và bệnh suy tim (congestive heart failure)

Nghiên cứu (i)	Beta-blocker		Placebo	
	N ₁	Tử vong (d ₁)	N ₂	Tử vong (d ₂)
1	25	5	25	6
2	9	1	16	2
3	194	23	189	21
4	25	1	25	2
5	105	4	34	2
6	320	53	321	67
7	33	3	16	2
8	261	12	84	13
9	133	6	145	11
10	232	2	134	5
11	1327	156	1320	228
12	1990	145	2001	217
13	214	8	212	17
Tổng cộng	4879	420	4516	612

N: số bệnh nhân nghiên cứu; Tử vong: số bệnh nhân chết trong thời gian theo dõi.

Như chúng ta thấy, một số nghiên cứu có số mẫu khá nhỏ, lại có những nghiên cứu với số mẫu gần 4000 người! Câu hỏi đặt ra là tổng hợp các nghiên cứu này, kết quả có nhất quán hay phù hợp với giả thiết BB làm giảm nguy cơ suy tim hay không? Để trả lời câu hỏi này, chúng ta tiến hành những bước sau đây:

Bước 1: ước tính mức độ ảnh hưởng cho từng nghiên cứu. Mỗi nghiên cứu có hai tỉ lệ: một cho nhóm BB và một cho nhóm placebo. Tôi sẽ gọi hai tỉ lệ này là p_1 và p_2 . Chỉ số để đánh giá mức độ ảnh hưởng của thuốc BB là tỉ số nguy cơ tương đối (relative risk – RR), và RR có thể được ước tính như sau:

$$RR = \frac{p_1}{p_2}$$

Chẳng hạn như, trong nghiên cứu 1, chúng ta có: $p_1 = \frac{5}{25} = 0,20$ và $p_2 = \frac{8}{25} = 0,24$.

Như vậy tỉ số nguy cơ cho nghiên cứu 1 là: $RR = \frac{0,20}{0,24} = 0,833$. Tính toán tương tự cho các nghiên cứu còn lại, chúng ta sẽ có một bảng như sau:

Bảng 2a. Ước tính tỉ lệ tử vong và tỉ số nguy cơ tương đối

Nghiên cứu (i)	Tỉ lệ tử vong nhóm BB (p_1)	Tỉ lệ tử vong nhóm placebo (p_2)	Tỉ số nguy cơ (RR)
1	0.200	0.240	0.833
2	0.111	0.125	0.889
3	0.119	0.111	1.067
4	0.040	0.080	0.500
5	0.038	0.059	0.648
6	0.166	0.209	0.794
7	0.091	0.125	0.727
8	0.046	0.155	0.297
9	0.045	0.076	0.595
10	0.009	0.037	0.231
11	0.118	0.173	0.681
12	0.073	0.108	0.672
13	0.037	0.080	0.466

Bước 2: biến đổi RR thành đơn vị logarithm và tính phương sai, sai số chuẩn. Mỗi ước số thống kê, như có lần nói, đều có một luật phân phối, và luật phân phối có thể phản ánh bằng phân sai (hay sai số chuẩn). Cách tính phương sai của RR khá phức tạp, cho nên chúng ta sẽ tính bằng một phương pháp gián tiếp. Theo phương pháp này, chúng ta sẽ biến đổi RR thành $\log[RR]$ (chú ý “log” ở đây có nghĩa là loga tự nhiên, tức là \log_e hay có khi còn viết tắt là \ln – natural logarithm), và sau đó sẽ tính phương sai của $\log[RR]$.

Nếu N_1 và N_2 là lần lược tổng số mẫu của nhóm 1 và nhóm 2; và d_1 và d_2 là số tử vong của nhóm 1 và nhóm 2 của một nghiên cứu, thì phương sai của $\log[RR]$ có thể ước tính bằng công thức sau đây:

$$\text{Var}[\log RR] = \frac{1}{d_1} - \frac{1}{N_1 - d_1} + \frac{1}{d_2} - \frac{1}{N_2 - d_2}$$

Và sai số chuẩn của $\log[RR]$ là:

$$SE[\log RR] = \sqrt{\frac{1}{d_1} - \frac{1}{N_1 - d_1} + \frac{1}{d_2} - \frac{1}{N_2 - d_2}}$$

Trong ví dụ trên, với nghiên cứu 1, chúng ta có:

$$\log[RR] = \log_e(0.833) = -0.182$$

Với phương sai:

$$\text{var}[\log RR] = \frac{1}{5} - \frac{1}{25-5} + \frac{1}{6} - \frac{1}{25-6} = 0.264$$

Và sai số chuẩn:

$$SE[\log RR] = \sqrt{0.264} = 0.514$$

Dựa vào luật phân phối chuẩn, chúng ta cũng có thể tính toán khoảng tin cậy 95% của RR cho từng nghiên cứu bằng cách biến đổi ngược lại theo đơn vị RR. Chẳng hạn như với nghiên cứu 1, chúng ta có khoảng tin cậy 95% của $\log[RR]$ là:

$$\log RR \pm 1.96 * SE[\log RR] = -0.182 \pm 1.96 * 0.514 = -1.19 \text{ đến } 0.82$$

hay biến đổi thành đơn vị nguyên thủy của RR là:

$$\exp(-1.19) = 0.30 \text{ đến } \exp(0.82) = 2.28$$

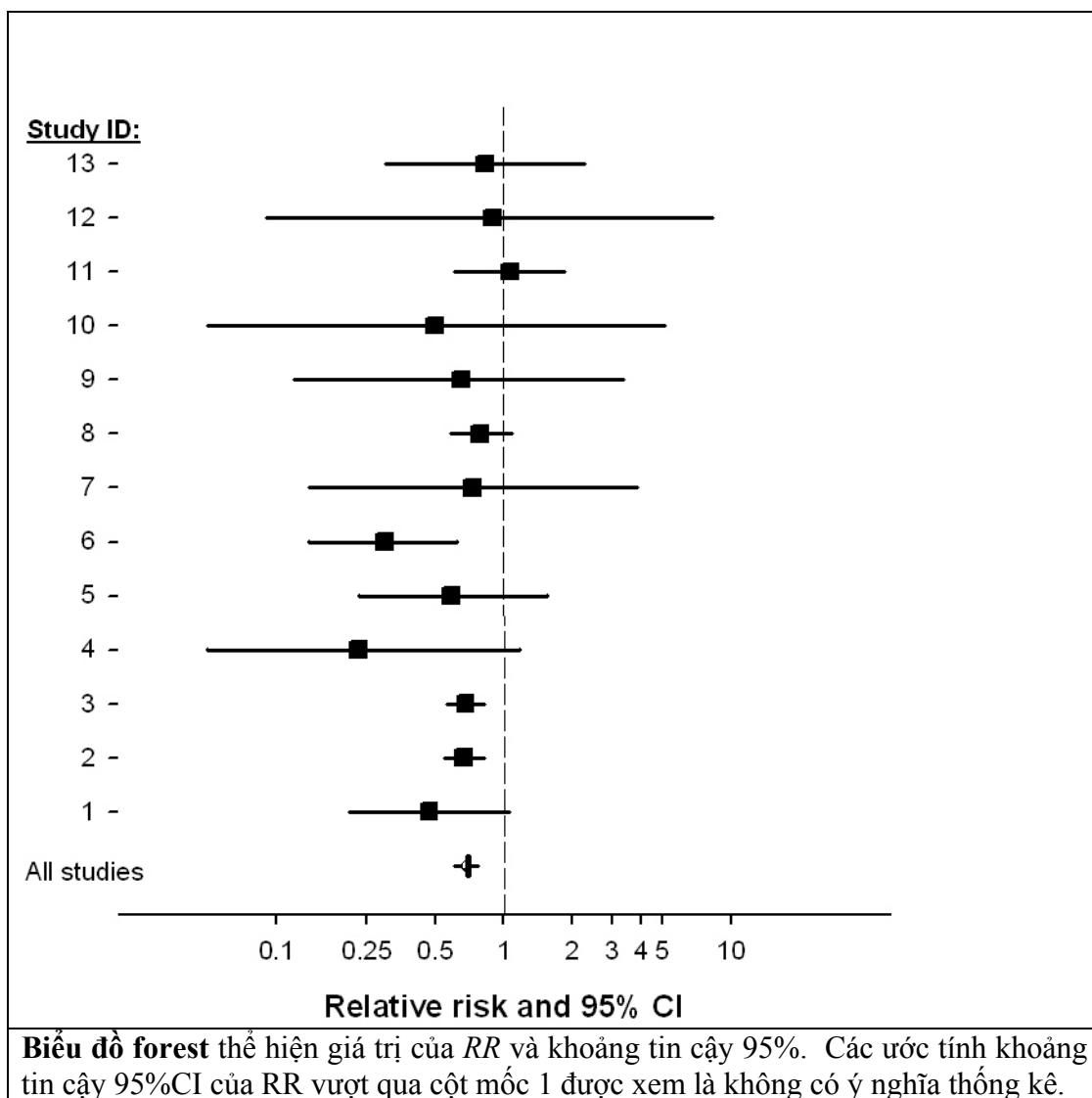
Tính toán tương tự cho các nghiên cứu khác, chúng ta có thêm một bảng mới như sau:

Bảng 2b. Ước tính tỉ số nguy cơ tương đối, phương sai, sai số chuẩn và khoảng tin cậy 95% cho từng nghiên cứu

Nghiên cứu (<i>i</i>)	Tỉ số nguy cơ (<i>RR</i>)	$\log[RR]$	$\text{Var}[\log RR]$	$SE[\log RR]$	Phản thấp 95%CI của RR	Phản cao 95% CI của RR
1	0.200	-0.182	0.264	0.514	0.30	2.28
2	0.111	-0.118	1.304	1.142	0.09	8.33
3	0.119	0.065	0.079	0.282	0.61	1.85
4	0.040	-0.693	1.415	1.189	0.05	5.15
5	0.038	-0.434	0.709	0.842	0.12	3.37
6	0.166	-0.231	0.026	0.162	0.58	1.09
7	0.091	-0.318	0.729	0.854	0.14	3.87
8	0.046	-1.214	0.142	0.377	0.14	0.62

9	0.045	-0.520	0.242	0.492	0.23	1.56
10	0.009	-1.465	0.688	0.829	0.05	1.17
11	0.118	-0.385	0.009	0.095	0.56	0.82
12	0.073	-0.398	0.010	0.102	0.55	0.82
13	0.037	-0.763	0.174	0.417	0.21	1.06

Chúng ta có thể thể hiện RR và khoảng tin cậy 95% bằng biểu đồ forest như sau:



Bước 3: ước tính trọng số (weight) cho từng nghiên cứu và RR cho toàn bộ nghiên cứu. Biểu đồ trên cho thấy một số nghiên cứu có độ dao động RR rất lớn (chứng tỏ các nghiên cứu này có số mẫu nhỏ hay ước số RR không ổn định), và ngược lại, một số nghiên cứu lớn có ước số RR ổn định hơn. Trọng số cho mỗi nghiên cứu (W_i – tôi sẽ cho vào kí hiệu i) để đo lường độ ổn định này là số đảo của phương sai:

$$W_i = \frac{1}{\text{var}[\log RR_i]}$$

Và số trung bình trọng số của $\log[RR]$ (kí hiệu là $\log wRR$) có thể ước tính từ tổng của tích $W_i \times \log[RR_i]$:

$$\log wRR = \frac{\sum W_i \times \log[RR_i]}{\sum W_i}$$

Với phương sai:

$$\text{Var}[\log wRR] = \frac{1}{\sum W_i}$$

và sai số chuẩn:

$$SE[\log wRR] = \sqrt{\frac{1}{\sum W_i}}$$

Ngoài ra, khoảng tin cậy 95% có thể ước tính bằng:

$$\log wRR \pm SE[\log wRR]$$

Để tính trung bình trọng số logRR, chúng ta cần một cột $W_i \times \log[RR_i]$. Chẳng hạn như với nghiên cứu 1, chúng ta có:

$$W_1 = \frac{1}{0,264} = 3.79$$

và

$$W_1 \times \log[RR_1] = 3.79 \times (-0.182) = -0.69$$

Tương tự cho các nghiên cứu khác:

Bảng 2c. Ước tính tỉ trọng số (W_i)

Nghiên cứu (i)	Log[RR]	Var[logRR]	W_i	$W_i \times \log[RR_i]$
1	-0.182	0.264	3.79	-0.69
2	-0.118	1.304	0.77	-0.09
3	0.065	0.079	12.61	0.82
4	-0.693	1.415	0.71	-0.49
5	-0.434	0.709	1.41	-0.61
6	-0.231	0.026	38.30	-8.86

7	-0.318	0.729	1.37	-0.44
8	-1.214	0.142	7.03	-8.54
9	-0.520	0.242	4.13	-2.15
10	-1.465	0.688	1.45	-2.13
11	-0.385	0.009	110.78	-42.63
12	-0.398	0.010	96.13	-38.23
13	-0.763	0.174	5.75	-4.39
Tổng số			284.24	-108.42

Chúng ta có:

$$\sum W_i = 3.79 + 0.77 + \dots + 5.75 = 284.24$$

$$\sum W_i \times \log[RR_i] = -0.69 - 0.09 + \dots - 4.39 = -108.42$$

Do đó, trung bình trọng số của $\log[RR]$ có thể ước tính bằng:

$$\log wRR = \frac{\sum W_i \times \log[RR_i]}{\sum W_i} = \frac{-108,42}{284,24} = -0.38$$

Với phương sai:

$$Var[\log wRR] = \frac{1}{\sum W_i} = \frac{1}{284.24} = 0.0035$$

và sai số chuẩn:

$$SE[\log wRR] = \sqrt{\frac{1}{\sum W_i}} = \sqrt{0.0035} = 0.06$$

Do đó, khoảng tin cậy 95% của $\log wRR$ có thể ước tính bằng:

$$\log wRR \pm SE[\log wRR] = -0.38 \pm 1.96 \times 0.06 = 0.498 \text{ đến } -0.265$$

Nhưng chúng ta muốn thể hiện bằng đơn vị gốc (tức tỉ số); do đó, các ước số trên phải được biến chuyển về đơn vị gốc:

$$RR = \exp(\log wRR) = \log(-0.38) = 0.68$$

Và khoảng tin cậy 95%:

$$\exp(-0.498) = 0.61 \text{ đến } \exp(-0.265) = 0.77.$$

Đến đây chúng ta có thể nói rằng tỉ lệ tử vong trong các bệnh nhân được điều trị bằng BB bằng 0.68 (hay thấp hơn 32%) so với các bệnh nhân giả dược (placebo). Ngoài ra, vì khoảng tin cậy 95% không bao gồm 1, chúng ta cũng có thể nói, mức độ khác biệt này có ý nghĩa thống kê.

Bước 4: ước tính chỉ số đồng nhất và bất đồng nhất. Như đã nói trong phần (1) liên quan đến phân tích biến liên tục, sau khi đã ước tính tỉ số nguy cơ trung bình, chúng ta cần phải xem xét chỉ số I^2 .

Để ước tính chỉ số I^2 , chúng ta cần tính $W_i(\log RR_i - \log wRR)^2$ cho mỗi nghiên cứu. Chẳng hạn như với nghiên cứu 1, chúng ta có:

$$W_1(\log RR_1 - \log wRR)^2 = 3.79 \times (-0.182 + 0.38)^2 = 0.1502$$

và cho các nghiên cứu khác:

Bảng 2d. Ước tính chỉ số heterogeneity (I^2)

Nghiên cứu (i)	Log[RR $_i$]	W_i	$W_i(\log RR_i - \log wRR)^2$
1	-0.182	3.79	0.1502
2	-0.118	0.77	0.0533
3	0.065	12.61	2.5118
4	-0.693	0.71	0.0687
5	-0.434	1.41	0.0040
6	-0.231	38.30	0.8635
7	-0.318	1.37	0.0054
8	-1.214	7.03	4.8731
9	-0.520	4.13	0.0790
10	-1.465	1.45	1.7074
11	-0.385	110.78	0.0012
12	-0.398	96.13	0.0253
13	-0.763	5.75	0.8382
Tổng số		284.24	11.1811

Ví dụ 2 có $k = 13$ nghiên cứu. Do đó,

$$Q = \sum_{i=1}^k W_i (\log RR_i - \log wRR)^2 = 11.1811$$

Và,

$$I^2 = \frac{Q - (k-1)}{Q} = \frac{11.18 - 12}{11.18} = -0.16$$

Vì $I^2 < 0$, nên chúng ta có thể cho $I^2 = 0$. Nói cách khác, mức độ khác biệt về RR giữa các nghiên cứu không có ý nghĩa thống kê.

Bước 5: đánh giá khả năng publication bias. Như đã giải thích trong phần 1f, cách đánh giá khả năng publication bias có ý nghĩa nhất là phân tích hồi qui đường thẳng $\log[RR]$ và tổng số mẫu (N):

$$\log[RR_i] = a + b \times N_i$$

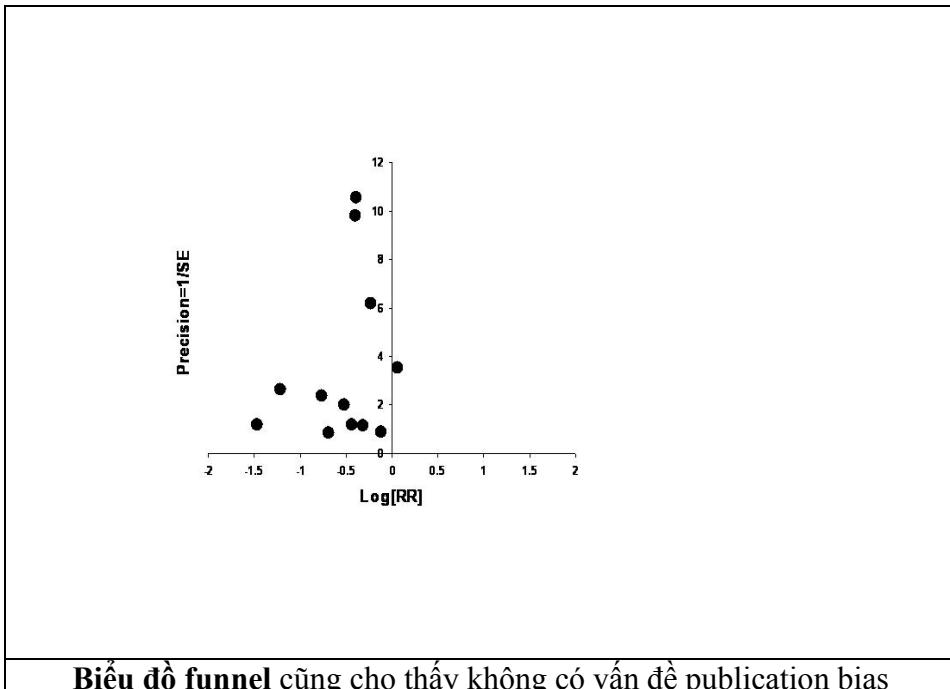
Dựa vào bảng thống kê sau.

Nghiên cứu (i)	$\log[RR_i]$	N_i
1	-0.182	50
2	-0.118	25
3	0.065	383
4	-0.693	50
5	-0.434	139
6	-0.231	641
7	-0.318	49
8	-1.214	345
9	-0.520	278
10	-1.465	366
11	-0.385	2647
12	-0.398	3991
13	-0.763	426

Chúng ta có thể ước tính a và b như sau:

$$\log[RR_i] = -0.534 + 0.00003 \times N_i$$

Ước tính $b = 0.00003$ không có ý nghĩa thống kê ($p = 0.782$). Do đó, chúng ta có thể phát biểu rằng mức độ thiên lệch về xuất bản không đáng kể trong phân tích tổng hợp này.



Biểu đồ funnel cũng cho thấy không có vấn đề publication bias

14.5.2 Phân tích bằng R

Package meta có hàm `metabin` có thể sử dụng để tiến hành phân tích tổng hợp cho các biến nhị phân như số liệu trong ví dụ 2 trên đây. Khoi đầu, chúng ta “nạp” package meta (nếu chưa làm) vào môi trường vận hành, và sau đó thu nhập số liệu vào một data frame:

```
library(meta)

# Số liệu từ ví dụ 2
n1 <- c(25.9.194.25.105.320.33.261.133.232.1327.1990.214)
d1 <- c(5.1.23.1.4.53.3.12.6.2.156.145.8)
n2 <- c(25.16.189.25.34.321.16.84.145.134.1320.2001.212)
d2 <- c(6.2.21.2.2.67.2.13.11.5.228.217.17)

# Tạo một dataframe lấy tên là bb
bb <- data.frame(n1,d1,n2,d2)

# Phân tích bằng hàm metabin và kết quả trong res
> res <- metabin(d1,n1,d2,n2,data=bb,sm="RR",meth="I")
> res
> res
      RR          95%-CI %W(fixed) %W(random)
1  0.8333 [0.2918; 2.3799]     1.26      1.26
2  0.8889 [0.0930; 8.4951]     0.27      0.27
3  1.0670 [0.6116; 1.8617]     4.47      4.47
4  0.5000 [0.0484; 5.1677]     0.25      0.25
5  0.6476 [0.1240; 3.3814]     0.51      0.51
6  0.7935 [0.5731; 1.0986]    13.08     13.08
```

7	0.7273	[0.1346; 3.9282]	0.49	0.49
8	0.2971	[0.1410; 0.6258]	2.49	2.49
9	0.5947	[0.2262; 1.5632]	1.48	1.48
10	0.2310	[0.0454; 1.1744]	0.52	0.52
11	0.6806	[0.5635; 0.8221]	38.81	38.81
12	0.6719	[0.5496; 0.8214]	34.31	34.31
13	0.4662	[0.2056; 1.0570]	2.07	2.07

Number of trials combined: 13

	RR	95%-CI	z	p.value
Fixed effects model	0.6821	[0.6064; 0.7672]	-6.3741	< 0.0001
Random effects model	0.6821	[0.6064; 0.7672]	-6.3741	< 0.0001

Quantifying heterogeneity:

tau² = 0; H = 1 [1; 1.45]; I² = 0% [0%; 52.6%]

Test of heterogeneity:

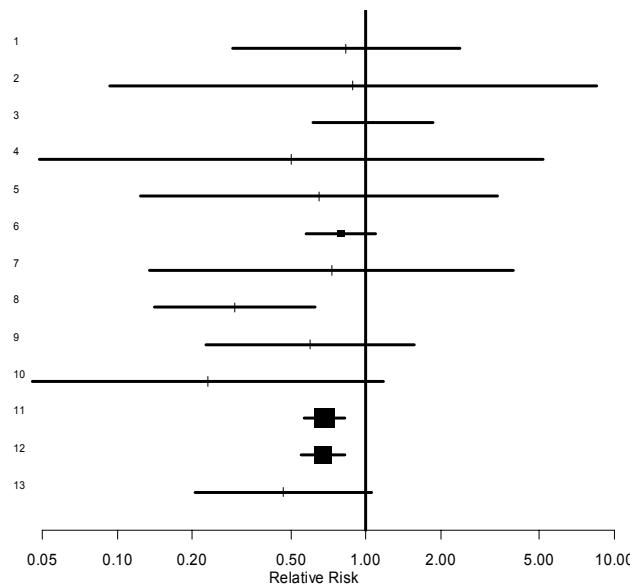
Q	d.f.	p.value
11	12	0.5292

Method: Inverse variance method

Kết quả từ mô hình fixed-effects và random-effects một lần nữa cho chúng ta bằng chứng để kết luận rằng beta-blocker có hiệu nghiệm trong việc làm giảm nguy cơ tử vong.

Biểu đồ forest

```
> plot(res, lwd=3)
```



Thực ra, trong khoa học nói chung, chúng ta đã có một truyền thống lâu đời về việc duyệt xét bằng chứng nghiên cứu (review), duyệt xét kiến thức hiện hành. Nhưng các duyệt xét như thế thường mang tính định chất (qualitative review), và vì tính định chất, chúng ta khó mà biết chính xác được những khác biệt mang tính định lượng giữa các nghiên cứu. Phân tích tổng hợp cung cấp cho chúng ta một phương tiện định lượng để hệ thống bằng chứng. Với phân tích tổng hợp, chúng ta có cơ hội để:

- xem xét những nghiên cứu nào đã được tiến hành để giải quyết vấn đề;
- kết quả của các nghiên cứu đó như thế nào;
- hệ thống các tiêu chí lâm sàng đáng quan tâm;
- rà soát những khác biệt về đặc tính giữa các nghiên cứu;
- cách thức để tổng hợp kết quả; và
- truyền đạt kết quả một cách khoa học.

Mục đích của phân tích tổng hợp, xin nhắc lại một lần nữa, là ước tính một chỉ số ảnh hưởng trung bình sau khi đã xem xét tất cả kết quả nghiên cứu hiện hành. Một kết quả chung như thế giúp cho chúng ta đi đến một kết luận chính xác và đáng tin cậy hơn.

Hai ví dụ trên đây hi vọng đã giúp ích cho bạn đọc hiểu được “cơ chế” và ý nghĩa của một phân tích tổng hợp. Hi vọng bạn đọc có thể tự mình làm một phân tích như thế khi đã có dữ liệu. Thực ra, tất cả các tính toán trên có thể thực hiện bằng một phần mềm như Microsoft Excel. Ngoài ra, một số phần mềm chuyên môn khác (như SAS chẳng hạn) cũng có thể tiến hành những phân tích trên. Tôi sẽ giải thích phần này trong phần kế tiếp). Các phép tính thật đơn giản. Vấn đề của phân tích tổng hợp không phải là tính toán, mà là dữ liệu đãng sau tính toán.

Phân tích tổng hợp cũng không phải là không có những khuyết điểm. Trong nghiên cứu người ta có câu “rác vào, rác ra”, tức là nếu các dữ liệu được sử dụng trong phân tích không có chất lượng cao thì kết quả của phân tích tổng hợp cũng chẳng có giá trị khoa học gì. Do đó, vấn đề quan trọng nhất trong phân tích tổng hợp là chọn lựa dữ liệu và nghiên cứu để phân tích. Vấn đề này cần phải được cân nhắc cực kì cẩn thận để đảm bảo tính hợp lý và khoa học của kết quả.

Tài liệu tham khảo và chú thích

- [1] Glass GV. Primary, secondary, and meta-analysis of research. Educational Researcher 1976; 5:3-8.
- [2] Normand SL. Meta-analysis: formulating, evaluating, combining, and reporting. Stat Med. 1999;18(3):321-59.
- [3] Higgins JPT, Thompson SG. Quantifying heterogeneity in a meta-analysis. Stat Med. 2002;21:1539-1558

- [4] Egger M. Davey Smith G. Schneider M. Minder C. Bias in meta-analysis detected by a simple graphical test. *Br Med J* 1997;315:629–34.
- [5] Tang JL. Liu JL. Misleading funnel plot for detection of bias in meta-analysis. *J Clin Epidemiol*. 2000;53(5):477-84.
- [6] Peters JL. Sutton AJ. Jones DR. Abrams KR. Rushton L. Comparison of two methods to detect publication bias in meta-analysis. *JAMA*. 2006;295(6):676-80.
- [7] Macaskill P. Walter SD. Irwig L. A comparison of methods to detect publication bias in meta-analysis. *Stat Med*. 2001;20:641-654.

Tóm tắt phân tích tổng hợp

Đối với các biến số liên tục	Đối với các biến số nhị phân
Nhóm 1 (số mẫu. trung bình. độ lệch chuẩn): $n_{1i} \cdot x_{1i} \cdot s_{1i}$; $i = 1, 2, 3, \dots, k$	Nhóm 1 (số mẫu. số sự kiện): $n_{1i} \cdot x_{1i}$; $i = 1, 2, 3, \dots, k$
Nhóm 2 (số mẫu. trung bình. độ lệch chuẩn): $n_{2i} \cdot x_{2i} \cdot s_{2i}$	Nhóm 2 (số mẫu. số sự kiện): $n_{2i} \cdot x_{2i}$; $i = 1, 2, 3, \dots, k$
Độ ảnh hưởng (effect size. ES): $d_i = x_{2i} - x_{1i}$	Độ ảnh hưởng (effect size. ES) tính bằng tỉ số nguy cơ RR: $RR_i = \left(\frac{x_{2i}}{n_{2i}} \right) \div \left(\frac{x_{1i}}{n_{1i}} \right)$ Biến chuyển sang logarithm: $\theta_i = \log(RR_i)$
Phương sai của d_i : $s_{di}^2 = \frac{(n_{1i}-1)s_{1i}^2 + (n_{2i}-1)s_{2i}^2}{n_{1i} + n_{2i} - 2} \times \left(\frac{1}{n_{1i}} + \frac{1}{n_{2i}} \right)$	Phương sai của θ_i : $s_{\theta i}^2 = \frac{1}{x_{1i}} - \frac{1}{n_{1i} - x_{1i}} + \frac{1}{x_{2i}} - \frac{1}{n_{2i} - x_{2i}}$
Sai số (standard error) của d_i : $s_{di} = \sqrt{s_{di}^2}$	Sai số của θ_i : $s_{\theta i} = \sqrt{\frac{1}{x_{1i}} - \frac{1}{n_{1i} - x_{1i}} + \frac{1}{x_{2i}} - \frac{1}{n_{2i} - x_{2i}}}$
Trọng số: $W_i = \frac{1}{s_{di}^2}$	Trọng số: $W_i = \frac{1}{s_{\theta i}^2}$
Uớc số ảnh hưởng chung: $d = \sum_{i=1}^k W_i d_i / \sum_{i=1}^k W_i$	Uớc số ảnh hưởng chung: $\theta = \sum_{i=1}^k W_i \theta_i / \sum_{i=1}^k W_i$
Phương sai của d : $s^2 = 1 / \sum_{i=1}^k W_i$	Phương sai của θ : $s^2 = 1 / \sum_{i=1}^k W_i$
Khoảng tin cậy 95%: $d \pm 1,96 \times \sqrt{s^2}$	Khoảng tin cậy 95%: $\theta \pm 1,96 \times \sqrt{s^2}$
Index of homogeneity: $Q = \sum_{i=1}^k W_i (d_i - d)^2$	Index of homogeneity: $Q = \sum_{i=1}^k W_i (\theta_i - \theta)^2$
Index of heterogeneity: $I^2 = \frac{Q - (k-1)}{Q}$	Index of heterogeneity: $I^2 = \frac{Q - (k-1)}{Q}$
Xem xét publication bias: Phân tích hồi qui tuyến tính: $d_i = a + b * N_i$. (N_i là tổng số mẫu của nghiên cứu i). Xem ý nghĩa thống kê của b .	Xem xét publication bias: Phân tích hồi qui tuyến tính: $\theta_i = a + b * N_i$. (N_i là tổng số mẫu của nghiên cứu i). Xem ý nghĩa thống kê của b .

Ước tính phương sai giữa các nghiên cứu
(between-study variance):

$$\tau^2 = \max \left[0, \frac{Q - (k - 1)}{\sum_{i=1}^k W_i - \frac{\sum_{i=1}^k W_i^2}{\sum_{i=1}^k W_i}} \right]$$

Ước tính phương sai giữa các nghiên cứu
(between-study variance):

$$\tau^2 = \max \left[0, \frac{Q - (k - 1)}{\sum_{i=1}^k W_i - \frac{\sum_{i=1}^k W_i^2}{\sum_{i=1}^k W_i}} \right]$$

15

Ước tính cỡ mẫu (Sample size estimation)

Một công trình nghiên cứu thường dựa vào một mẫu (sample). Một trong những câu hỏi quan trọng nhất trước khi tiến hành nghiên cứu là cần bao nhiêu mẫu hay bao nhiêu đối tượng cho nghiên cứu. “Đối tượng” ở đây là đơn vị căn bản của một nghiên cứu, là số bệnh nhân, số tình nguyện viên, số mẫu ruộng, cây trồng, thiết bị, v.v... Ước tính số lượng đối tượng cần thiết cho một công trình nghiên cứu đóng vai trò cực kì quan trọng, vì nó có thể là yếu tố quyết định sự thành công hay thất bại của nghiên cứu. Nếu số lượng đối tượng không đủ thì kết luận rút ra từ công trình nghiên cứu không có độ chính xác cao, thậm chí không thể kết luận gì được. Ngược lại, nếu số lượng đối tượng quá nhiều hơn số cần thiết thì tài nguyên, tiền bạc và thời gian sẽ bị hao phí. Do đó, vấn đề then chốt trước khi nghiên cứu là phải ước tính cho được một số đối tượng vừa đủ cho mục tiêu của nghiên cứu. Số lượng đối tượng “vừa đủ” tùy thuộc vào ba yếu tố chính:

- Sai sót mà nhà nghiên cứu chấp nhận, cụ thể là sai sót loại I và II;
- Độ dao động (variability) của đo lường, mà cụ thể là độ lệch chuẩn; và
- Mức độ khác biệt hay ảnh hưởng mà nhà nghiên cứu muốn phát hiện.

Không có số liệu về ba yếu tố này thì không thể nào ước tính cỡ mẫu. Kinh nghiệm của người viết cho thấy rất nhiều người khi tiến hành nghiên cứu thường không có ý niệm gì về các số liệu này, cho nên khi đến tham vấn các chuyên gia về thống kê học, họ chỉ nhận câu trả lời: “không thể tính được”! Trong chương này tôi sẽ bàn qua ba yếu tố trên.

15.1 Khái niệm về “power”

Thống kê học là một phương pháp khoa học có mục đích phát hiện, hay đi tìm những cái có thể gộp chung lại bằng cụm từ “chưa được biết” (unknown). Cái chưa được biết ở đây là những hiện tượng chúng ta không quan sát được, hay quan sát được nhưng không đầy đủ. “Cái chưa biết” có thể là một ẩn số (như chiều cao trung bình ở người Việt Nam, hay trọng lượng một phần tử), hiệu quả của một thuật điều trị, gen có chức năng làm cho cây lá có màu xanh, sở thích của con người, v.v... Chúng ta có thể đo chiều cao, hay tiến hành xét nghiệm để biết hiệu quả của thuốc, nhưng các nghiên cứu như thế chỉ được tiến hành trên một nhóm đối tượng, chứ không phải toàn bộ quần thể của dân số.

Ở mức độ đơn giản nhất, những cái chưa biết này có thể xuất hiện dưới hai hình thức: hoặc là có, hoặc là không. Chẳng hạn như một thuật điều trị có hay không có hiệu quả chống gãy xương, khách hàng thích hay không thích một loại nước giải khát. Bởi vì không ai biết hiện tượng một cách đầy đủ, chúng ta phải đặt ra giả thiết. Giả thiết đơn

giản nhất là *giả thiết đảo* (hiện tượng không tồn tại, kí hiệu H-) và *giả thiết chính* (hiện tượng tồn tại, kí hiệu H+).

Chúng ta sử dụng các phương pháp kiểm định thống kê (statistical test) như kiểm định t , F , z , χ^2 , v.v... để đánh giá khả năng của giả thiết. Kết quả của một kiểm định thống kê có thể đơn giản chia thành hai giá trị: hoặc là *có ý nghĩa thống kê* (statistical significance), hoặc là *không có ý nghĩa thống kê* (non-significance). Có ý nghĩa thống kê ở đây, như đề cập trong Chương 7, thường dựa vào trị số P: nếu $P < 0.05$, chúng ta phát biểu kết quả có ý nghĩa thống kê; nếu $P > 0.05$ chúng ta nói kết quả không có ý nghĩa thống kê. Cũng có thể xem có ý nghĩa thống kê hay không có ý nghĩa thống kê như là có tín hiệu hay không có tín hiệu. Hãy tạm đặt kí hiệu T+ là kết quả có ý nghĩa thống kê, và T- là kết quả kiểm định không có ý nghĩa thống kê.

Hãy xem xét một ví dụ cụ thể: để biết thuốc risedronate có hiệu quả hay không trong việc điều trị loãng xương, chúng ta tiến hành một nghiên cứu gồm 2 nhóm bệnh nhân (một nhóm được điều trị bằng risedronate và một nhóm chỉ sử dụng giả dược placebo). Chúng ta theo dõi và thu thập số liệu gãy xương, ước tính tỉ lệ gãy xương cho từng nhóm, và so sánh hai tỉ lệ bằng một kiểm định thống kê. Kết quả kiểm định thống kê hoặc là *có ý nghĩa thống kê* ($P < 0.05$) hay không có ý nghĩa thống kê ($P > 0.05$). Xin nhắc lại rằng chúng ta không biết risedronate thật sự có hiệu nghiệm chống gãy xương hay không; chúng ta chỉ có thể đặt giả thiết H. Do đó, khi xem xét một giả thiết và kết quả kiểm định thống kê, chúng ta có bốn tình huống:

- (a) Giả thuyết H đúng (thuốc risedronate có hiệu nghiệm) và kết quả kiểm định thống kê $P < 0.05$.
- (b) Giả thuyết H đúng, nhưng kết quả kiểm định thống kê không có ý nghĩa thống kê;
- (c) Giả thuyết H sai (thuốc risedronate không có hiệu nghiệm) nhưng kết quả kiểm định thống kê có ý nghĩa thống kê;
- (d) Giả thuyết H sai và kết quả kiểm định thống kê không có ý nghĩa thống kê.

Ở đây, trường hợp (a) và (d) không có vấn đề, vì kết quả kiểm định thống kê nhất quán với thực tế của hiện tượng. Nhưng trong trường hợp (b) và (c), chúng ta phạm sai lầm, vì kết quả kiểm định thống kê không phù hợp với giả thiết. Trong ngôn ngữ thống kê học, chúng ta có vài thuật ngữ:

- xác suất của tình huống (b) xảy ra được gọi là *sai sót loại II* (type II error), và thường kí hiệu bằng β .
- xác suất của tình huống (a) được gọi là *Power*. Nói cách khác, *power* chính là xác suất mà kết quả kiểm định thống kê ra kết quả $p < 0.05$ với điều kiện giả thiết H là thật. Nói cách khác: $power = 1 - \beta$;

- xác suất của tình huống (c) được gọi là *sai sót loại I* (type I error, hay significance level), và thường kí hiệu bằng α . Nói cách khác, α chính là xác suất mà kết quả kiểm định thống cho ra kết quả $p<0.05$ với điều kiện giả thiết H sai;
- xác suất tình huống (d) không phải là vấn đề càn quan tâm, nên không có thuật ngữ, dù có thể gọi đó là kết quả *âm tính thật* (hay true negative).

Có thể tóm lược 4 tình huống đó trong một Bảng 1 sau đây:

Bảng 1. Các tình huống trong việc thử nghiệm một giả thiết khoa học

Kết quả kiểm định thống kê	Giả thuyết H	
	Đúng (thuộc có hiệu nghiệm)	Sai (thuộc không có hiệu nghiệm)
Có ý nghĩa thống kê ($p<0,05$)	Dương tính thật (power), $1-\beta = P(s H+)$	Sai sót loại I (type I error) $\alpha = P(s H-)$
Không có ý nghĩa thống kê ($p>0,05$)	Sai sót loại II (type II error) $\beta = P(ns H+)$	Âm tính thật (true negative) $1-\alpha = P(ns H-)$

Chú thích: s trong biểu đồ này có nghĩa là significant; ns non-significant; $H+$ là giả thuyết đúng; và $H-$ là giả thuyết sai. Do đó, có thể mô tả 4 tình huống trên bằng ngôn ngữ xác suất có điều kiện như sau: Power = $1 - \beta = P(s | H+)$; $\beta = P(ns | H+)$; và $\alpha = P(s | H-)$.

15.2 Thử nghiệm giả thiết thống kê và chẩn đoán y khoa

Có lẽ những lí giải trên đây, đối với một số bạn đọc, vẫn còn khá trừu tượng. Một cách dễ minh họa các khái niệm *power* và trị số P là qua chẩn đoán y khoa. Thật vậy, có thể ví nghiên cứu khoa học và suy luận thống kê như là một qui trình chẩn đoán bệnh. Trong chẩn đoán, thoát đầu chúng ta không biết bệnh nhân mắc bệnh hay không, và phải thu thập thông tin (như tìm hiểu tiền sử bệnh, cách sống, thói quen, v.v...) và làm xét nghiệm (như quang tuyến X, như siêu âm, phân tích máu, nước tiểu, v.v...) để đi đến kết luận.

Có hai giả thiết: bệnh nhân không có bệnh (kí hiệu $H-$) và bệnh nhân mắc bệnh ($H+$). Ở mức độ đơn giản nhất, kết quả xét nghiệm có thể là *dương tính* (+ve) hay *âm tính* (-ve). Trong chẩn đoán cũng có 4 tình huống và tôi sẽ bàn trong phần dưới đây, nhưng để vấn đề rõ ràng hơn, chúng ta hãy xem qua một ví dụ cụ thể như sau:

Trong chẩn đoán ung thư, để biết chắc chắn có ung thư hay không, phương pháp chuẩn là dùng sinh thiết (tức giải phẫu để xem xét mô dưới ống kính hiển vi để xác định xem *có ung thư* hay *không có ung thư*). Nhưng sinh thiết là một phẫu thuật có tính cách

xâm phạm vào cơ thể bệnh nhân, nên không thể áp dụng phẫu thuật này một cách đại trà cho mọi người. Thay vào đó, y khoa phát triển những phương pháp xét nghiệm không mang tính xâm phạm để thử nghiệm ung thư. Các phương pháp này bao gồm quang tuyến X hay thử máu. Kết quả của một xét nghiệm bằng quang tuyến X hay thử máu có thể tóm tắt bằng hai giá trị: hoặc là dương tính (+ve), hoặc là âm tính (-ve).

Nhưng không có một phương pháp gián tiếp thử nghiệm nào, dù tinh vi đến đâu đi nữa, là hoàn hảo và chính xác tuyệt đối. Một số người có kết quả dương tính, nhưng thực sự không có ung thư. Và một số người có kết quả âm tính, nhưng trong thực tế lại có ung thư. Đến đây thì chúng ta có bốn khả năng:

- Bệnh nhân có ung thư, và kết quả thử nghiệm là dương tính. Đây là trường hợp **dương tính thật** (danh từ chuyên môn là *độ nhạy*, tiếng Anh gọi là *sensitivity*);
- bệnh nhân không có ung thư, nhưng kết quả thử nghiệm là dương tính. Đây là trường hợp **dương tính giả** (*false positive*);
- bệnh nhân không có ung thư, nhưng kết quả thử nghiệm là âm tính. Đây là trường hợp của **âm tính thật** (*specificity*); và,
- bệnh nhân có ung thư, và kết quả thử nghiệm là âm tính. Đây là trường hợp **âm tính giả** hay **độ đặc hiệu** (*false negative*).

Có thể tóm lược 4 tình huống đó trong Bảng 2 sau đây:

Bảng 2. Các tình huống trong việc chẩn đoán y khoa: kết quả xét nghiệm và bệnh trạng

Kết quả xét nghiệm	Bệnh trạng	
	Có bệnh	Không có bệnh
+ve (dương tính)	Độ nhạy (<i>sensitivity</i>),	Dương tính giả (<i>false positive</i>)
-ve (âm tính)	Âm tính giả (<i>false negative</i>),	Độ đặc hiệu (<i>Specificity</i>),

Đến đây, chúng ta có thể thấy qua mối tương quan song song giữa chẩn đoán y khoa và thử nghiệm thống kê. Trong chẩn đoán y khoa có chỉ số dương tính thật, tương đương với khái niệm “power” trong nghiên cứu. Trong chẩn đoán y khoa có xác suất dương tính giả, và xác suất này chính là trị số p trong suy luận khoa học. Bảng sau đây sẽ cho thấy mối tương quan đó:

Bảng 3. Tương quan giữa chẩn đoán y khoa và suy luận trong khoa học

Chẩn đoán y khoa	Thử nghiệm giả thiết khoa học
Chẩn đoán bệnh	Thử nghiệm một giả thiết khoa học
Bệnh trạng (có hay không)	Giả thiết khoa học ($H+$ hay $H-$)
Phương pháp xét nghiệm	Kiểm định thống kê
Kết quả xét nghiệm +ve	Trị số $p < 0.05$ hay “có ý nghĩa thống kê”
Kết quả xét nghiệm -ve	Trị số $p > 0.05$ hay “không có ý nghĩa thống kê”
Dương tính thật (sensitivity)	Power; $1-\beta$; $P(s H+)$
Dương tính giả (false positive)	Sai sót loại I; trị số p ; α ; $P(s H-)$
Âm tính giả (false negative)	Sai sót loại II; β ; $\beta = P(ns H+)$
Âm tính thật (đặc hiệu, hay specificity)	Âm tính thật; $1-\alpha = P(ns H-)$

Cũng như các phương pháp xét nghiệm y khoa không bao giờ hoàn hảo, các phương pháp kiểm định thống kê cũng có sai sót. Và do đó, kết quả nghiên cứu lúc nào cũng có độ bất định (như sự bất định trong một chẩn đoán y khoa vậy). Vấn đề là chúng ta phải thiết kế nghiên cứu sao cho *sai sót* loại I và II thấp nhất.

15.3 Số liệu để ước tính cỡ mẫu

Như đã đề cập trong phần đầu của chương này, để ước tính số đối tượng cần thiết cho một công trình nghiên cứu, chúng ta cần phải có 3 số liệu: xác suất sai sót loại I và II, độ dao động của đo lường, và độ ảnh hưởng.

- Về xác suất sai sót, thông thường một nghiên cứu chấp nhận sai sót loại I khoảng 1% hay 5% (tức $\alpha = 0.01$ hay 0.05), và xác suất sai sót loại II khoảng $\beta = 0.1$ đến $\beta = 0.2$ (tức power phải từ 0.8 đến 0.9).
- Độ dao động chính là độ lệch chuẩn (standard deviation) của đo lường mà công trình nghiên cứu dựa vào để phân tích. Chẳng hạn như nếu nghiên cứu về cao huyết áp, thì nhà nghiên cứu cần phải có độ lệch chuẩn của áp suất máu. Chúng ta tạm gọi độ dao động là σ .
- Độ ảnh hưởng, nếu là công trình nghiên cứu so sánh hai nhóm, là độ khác biệt trung bình giữa hai nhóm mà nhà nghiên cứu muốn phát hiện. Chẳng hạn như nhà nghiên cứu có thể giả thiết rằng bệnh nhân được điều trị bằng thuốc A có áp suất máu giảm 10 mmHg so với nhóm giả được. Ở đây, 10 mmHg được xem là độ ảnh hưởng. Chúng ta tạm gọi độ ảnh hưởng là Δ .

Một nghiên cứu có thể có một nhóm đối tượng hay hai (và có khi hơn 2) nhóm đối tượng. Và ước tính cỡ mẫu cũng tùy thuộc vào các trường hợp này.

Trong trường hợp một nhóm đối tượng, số lượng đối tượng (n) cần thiết cho nghiên cứu có thể tính toán một cách “thủ công” như sau:

$$n = \frac{C}{(\Delta/\sigma)^2} \quad [1]$$

Trong trường hợp có hai nhóm đối tượng, số lượng đối tượng (n) cần thiết cho nghiên cứu có thể tính toán như sau:

$$n = 2 \times \frac{C}{(\Delta/\sigma)^2} \quad [2]$$

Trong đó, hằng số C được xác định từ xác suất sai sót loại I và II (hay power) như sau:

Bảng 3: Hằng số C liên quan đến sai sót loại I và II

$\alpha =$	$\beta = 0.20$ (Power = 0.80)	$\beta = 0.10$ (Power = 0.90)	$\beta = 0.05$ (Power = 0.95)
0.10	6.15	8.53	10.79
0.05	7.85	10.51	13.00
0.01	13.33	16.74	19.84

15.4 Ước tính cỡ mẫu

15.4.1 Ước tính cỡ mẫu cho một chỉ số trung bình

Ví dụ 1: Chúng ta muốn ước tính chiều cao ở dàn ông người Việt, và chấp nhận sai số trong vòng 1 cm ($d = 1$) với khoảng tin cậy 0.95 (tức $\alpha=0.05$) và power = 0.8 (hay $\beta = 0.2$). Các nghiên cứu trước cho biết độ lệch chuẩn chiều cao ở người Việt khoảng 4.6 cm. Chúng ta có thể áp dụng công thức [1] để ước tính cỡ mẫu cần thiết cho nghiên cứu:

$$n = \frac{C}{(\Delta/\sigma)^2} = \frac{7.85}{(1/4.6)^2} = 166$$

Nói cách khác, chúng ta cần phải đo chiều cao ở 166 đối tượng để ước tính chiều cao dàn ông Việt với sai số trong vòng 1 cm.

Nếu sai số chấp nhận là 0.5 cm (thay vì 1 cm), số lượng đối tượng cần thiết là:
 $n = \frac{7.85}{(0.5/4.6)^2} = 664$. Nếu độ sai số mà chúng ta chấp nhận là 0.1 cm thì số lượng đối tượng nghiên cứu lên đến 16610 người! Qua các ước tính này, chúng ta dễ dàng thấy cỡ mẫu tùy thuộc rất lớn vào độ sai số mà chúng ta chấp nhận. Muốn có ước tính càng chính xác, chúng ta cần càng nhiều đối tượng nghiên cứu.

Trong R có hàm power.t.test có thể áp dụng để ước tính cỡ mẫu cho ví dụ trên như sau. Chú ý chúng ta cho R biết vẫn đề là một nhóm tức type="one.sample":

```
# sai số 1 cm, độc lệch chuẩn 4.6, a=0.05, power=0.8
> power.t.test(delta=1, sd=4.6, sig.level=.05, power=.80,
   type='one.sample')

One-sample t test power calculation

      n = 168.0131
      delta = 1
      sd = 4.6
      sig.level = 0.05
      power = 0.8
      alternative = two.sided
```

kết quả tính toán từ R là 168, khác với cách tính thủ công 2 đổi tượng, vì cố nhiên R sử dụng nhiều số lẻ hơn và chính xác hơn cách tính thủ công. Với sai số 0.5 cm:

```
# sai số 0.5 cm, độc lệch chuẩn 4.6, a=0.05, power=0.8
> power.t.test(delta=0.5, sd=4.6, sig.level=.05, power=.80,
   type='one.sample')

One-sample t test power calculation

      n = 666.2525
      delta = 0.5
      sd = 4.6
      sig.level = 0.05
      power = 0.8
      alternative = two.sided
```

Ví dụ 2: Một loại thuốc điều trị có khả năng tăng độ alkaline phosphatase ở bệnh nhân loãng xương. Độ lệch chuẩn của alkaline phosphatase là 15 U/l. Một nghiên cứu mới sẽ tiến hành trong một quần thể bệnh nhân ở Việt Nam, và các nhà nghiên cứu muốn biết bao nhiêu bệnh nhân cần tuyển để chứng minh rằng thuốc có thể alkaline phosphatase từ 60 đến 65 U/l sau 3 tháng điều trị, với sai số $\alpha = 0.05$ và power = 0.8.

Đây là một loại nghiên cứu “trước – sau” (before-after study); có nghĩa là trước và sau khi điều trị. Ở đây, chúng ta chỉ có một nhóm bệnh nhân, nhưng được đo hai lần (trước khi dùng thuốc và sau khi dùng thuốc). Chỉ tiêu lâm sàng để đánh giá hiệu nghiệm của thuốc là độ thay đổi về alkaline phosphatase. Trong trường hợp này, chúng ta có trị số tăng trung bình là 5 U/l và độ lệch chuẩn là 15 U/l, hay nói theo ngôn ngữ R, delta=5, sd=15, sig.level=.05, power=.80, và lệnh:

```
> power.t.test(delta=3, sd=15, sig.level=.05, power=.80,
   type='one.sample')

One-sample t test power calculation
```

```

n = 198.1513
delta = 3
sd = 15
sig.level = 0.05
power = 0.8
alternative = two.sided

```

Như vậy, chúng ta cần phải có 198 bệnh nhân để đạt các mục tiêu trên.

15.4.2 Ước tính cỡ mẫu cho so sánh hai số trung bình

Trong thực tế, rất nhiều nghiên cứu nhằm so sánh hai nhóm với nhau. Cách ước tính cỡ mẫu cho các nghiên cứu này chủ yếu dựa vào công thức [2] như trình bày phần 15.3.1.

Ví dụ 3: Một nghiên cứu được thiết kế để thử nghiệm thuốc alendronate trong việc điều trị loãng xương ở phụ nữ sau thời kỳ mãn kinh. Có hai nhóm bệnh nhân được tuyển: nhóm 1 là nhóm can thiệp (được điều trị bằng alendronate), và nhóm 2 là nhóm đối chứng (tức không được điều trị). Tiêu chí để đánh giá hiệu quả của thuốc là mật độ xương (bone mineral density – BMD). Số liệu từ nghiên cứu dịch tễ học cho thấy giá trị trung bình của BMD trong phụ nữ sau thời kỳ mãn kinh là 0.80 g/cm^2 , với độ lệch chuẩn là 0.12 g/cm^2 . Vấn đề đặt ra là chúng ta cần phải nghiên cứu ở bao nhiêu đối tượng để “chứng minh” rằng sau 12 tháng điều trị BMD của nhóm 1 tăng khoảng 5% so với nhóm 2?

Trong ví dụ trên, tạm gọi trị số trung bình của nhóm 2 là μ_2 và nhóm 1 là μ_1 , chúng ta có: $\mu_1 = 0.8 * 1.05 = 0.84 \text{ g/cm}^2$ (tức tăng 5% so với nhóm 1), và do đó, $\Delta = 0.84 - 0.80 = 0.04 \text{ g/cm}^2$. Độ lệch chuẩn là $\sigma = 0.12 \text{ g/cm}^2$. Với power = 0.90 và $\alpha = 0.05$, cỡ mẫu cần thiết là:

$$n = \frac{2C}{(\Delta/\sigma)^2} = \frac{2 \times 10.51}{(0.04/0.12)^2} = 189$$

Và lời giải từ R qua hàm power.t.test như sau:

```
> power.t.test(delta=0.04, sd=0.12, sig.level=0.05, power=0.90,
  type="two.sample")
```

```
Two-sample t test power calculation
```

```

n = 190.0991
delta = 0.04
sd = 0.12
sig.level = 0.05
power = 0.9
alternative = two.sided

```

NOTE: n is number in *each* group

Chú ý trong hàm power.t.test, ngoài các thông số thông thường như delta (độ ảnh hưởng hay khác biệt theo giả thiết), sd (độ lệch chuẩn), sig.level xác suất sai sót loại I, và power, chúng ta còn phải cụ thể chỉ ra rằng đây là nghiên cứu gồm có hai nhóm với thông số type="two.sample".

Kết quả trên cho biết chúng ta cần 190 bệnh nhân **cho mỗi nhóm** (hay 380 bệnh nhân cho công trình nghiên cứu). Trong trường hợp này, power = 0.90 và $\alpha = 0.05$ có nghĩa là gì? Trả lời: hai thông số đó có nghĩa là nếu chúng ta tiến hành thật nhiều nghiên cứu (ví dụ 1000) và mỗi nghiên cứu với 380 bệnh nhân, sẽ có 90% (hay 900) nghiên cứu sẽ cho ra kết quả trên với trị số $p < 0.05$.

15.4.3 Ước tính cỡ mẫu cho phân tích phương sai

Phương pháp ước tính cỡ mẫu cho so sánh giữa hai nhóm cũng có thể khai triển thêm để ước tính cỡ mẫu cho trường hợp so sánh hơn hai nhóm. Trong trường hợp có nhiều nhóm, như đề cập trong Chương 11, phương pháp so sánh là phân tích phương sai. Theo phương pháp này, số trung bình bình phương phần dư (residual mean square, RMS) chính là ước tính của độ dao động của đo lường trong mỗi nhóm, và chỉ số này rất quan trọng trong việc ước tính cỡ mẫu.

Chi tiết về lí thuyết đằng sau cách ước tính cỡ mẫu cho phân tích phương sai khá phức tạp, và không nằm trong phạm vi của chương này. Nhưng nguyên lý chủ yếu vẫn không khác so với lí thuyết so sánh giữa hai nhóm. Gọi số trung bình của k nhóm là $\mu_1, \mu_2, \mu_3, \dots, \mu_k$, chúng ta có thể tính tổng bình phương giữa các nhóm bằng $SS_{\text{SS}} = \sum_{i=1}^k (\mu_i - \bar{\mu})^2$, trong đó, $\bar{\mu} = \sum_{i=1}^k \mu_i / k$. Cho $\lambda = \frac{SS}{(k-1)RMS}$, vẫn đề đặt ra là tìm cỡ lượng cỡ mẫu n sao cho z_β đáp ứng yêu cầu power = 0.80 hay 0.9, mà

$$z_\beta = \frac{1}{\sqrt{(k-1)(1+n\lambda)F + k(n-1)(1+2n\lambda)}} \times \\ \left(\sqrt{k(n-1)[2(k-1)(1+n\lambda)^2 - (1+2n\lambda)]} - \sqrt{F(k-1)(1+n\lambda)(2k(n-1)-1)} \right)$$

Trong đó F là kiểm định F . (Xem J. Fleiss, "The Design and Analysis of Clinical Experiments", John Wiley & Sons, New York 1986, trang 373).

Ví dụ 4. Để so sánh độ ngọt của một loại nước uống giữa 4 nhóm đối tượng khác nhau về giới tính và độ tuổi (tạm gọi 4 nhóm là A, B, C và D), các nhà nghiên cứu giả thiết rằng độ ngọt trong nhóm A, B, C và D lần lượt là 4.5, 3.0, 5.6, và 1.3. Qua xem xét nhiều nghiên cứu trước, các nhà nghiên cứu còn biết rằng RMS về độ ngọt trong mỗi

nhóm là khoảng 8.7. Vấn đề đặt ra là bao nhiêu đối tượng cần nghiên cứu để phát hiện sự khác biệt có ý nghĩa thống kê ở mức độ $\alpha = 0.05$ và power = 0.9.

Hàm `power.anova.test` trong R có thể ứng dụng để giải quyết vấn đề. Chúng ta chỉ cần đơn giản cung cấp 4 số trung bình theo giả thiết và số RMS như sau:

```
# trước hết cho 4 số trung bình vào một vector
> groupmeans <- c(4.5, 3.0, 5.6, 1.3)

# sau đó, "gọi" hàm power.anova.test:
> power.anova.test(groups = length(groupmeans),
                     between.var=var(groupmeans),
                     within.var=8.7, power=0.90, sig.level=0.05)

Balanced one-way analysis of variance power calculation

groups = 4
n = 12.81152
between.var = 3.486667
within.var = 8.7
sig.level = 0.05
power = 0.9

NOTE: n is number in each group
```

Kết quả cho thấy các nhà nghiên cứu cần khoảng 13 đối tượng cho mỗi nhóm (tức 52 đối tượng cho toàn bộ nghiên cứu).

15.4.4 Ước tính cỡ mẫu để ước tính một tỉ lệ

Nhiều nghiên cứu mô tả có mục đích khá đơn giản là ước tính một tỉ lệ. Chẳng hạn như giới y tế thường hay tìm hiểu tỉ lệ một bệnh trong cộng đồng, hay giới thăm dò ý kiến và thị trường thường tìm hiểu tỉ lệ dân số ưa thích một sản phẩm. Trong các trường hợp này, chúng ta không có những đo lường mang tính liên tục, nhưng kết quả chỉ là những giá trị nhị như có / không, thích / không thích, v.v... Và cách ước tính cỡ mẫu cũng khác với ba ví dụ trên đây.

Năm 1991, một cuộc thăm dò ý kiến ở Mĩ cho thấy 45% người được hỏi sẵn sàng khuyến khích con họ nên hiến một quả thận cho những bệnh nhân cần thiết. Khoảng tin cậy 95% của tỉ lệ này là 42% đến 48%, tức một khoảng cách đến 6%! Kết quả này [tương đối] thiếu chính xác, dù số lượng đối tượng tham gia lên đến 1000 người. Tại sao? Để trả lời câu hỏi này, chúng ta thử xem qua một vài lí thuyết về ước tính cỡ mẫu cho một tỉ lệ.

Chúng ta biết qua Chương 6 và 9 rằng nếu \hat{p} được ước tính từ n đối tượng, thì khoảng tin cậy 95% của một tỉ lệ p [trong dân số] là: $\hat{p} \pm 1.96 \times SE(\hat{p})$, trong đó $SE(\hat{p}) = \sqrt{\hat{p}(1-\hat{p})/n}$.

Bây giờ thử lật ngược vấn đề: chúng ta muốn ước tính p sao khoảng rộng $2 \times 1.96 \times SE(\hat{p})$ không quá một hằng số m . Nói cách khác, chúng ta muốn:

$$1.96 \times \sqrt{\hat{p}(1-\hat{p})/n} \leq m$$

Chúng ta muốn tìm số lượng đối tượng n để đạt yêu cầu trên. Qua cách diễn đạt trên, dễ dàng thấy rằng:

$$n \geq \left(\frac{1.96}{m} \right)^2 \hat{p}(1-\hat{p})$$

Do đó, số lượng cỡ mẫu tùy thuộc vào độ sai số m và tỉ lệ p mà chúng ta muốn ước tính. Độ sai số càng thấp, số lượng cỡ mẫu càng cao.

Ví dụ 5: Chúng ta muốn ước tính tỉ lệ đàn ông hút thuốc ở Việt Nam, sao cho ước số không cao hơn hay thấp hơn 2% so với tỉ lệ thật trong toàn dân số. Một nghiên cứu trước cho thấy tỉ lệ hút thuốc trong đàn ông người Việt có thể lên đến 70%. Câu hỏi đặt ra là chúng ta cần nghiên cứu trên bao nhiêu đàn ông để đạt yêu cầu trên.

Trong ví dụ này, chúng ta có sai số $m = 0.02$, $\hat{p} = 0.70$, và số lượng cỡ mẫu cần thiết cho nghiên cứu là:

$$n \geq \left(\frac{1.96}{0.02} \right)^2 0.7 \times 0.3$$

Nói cách khác, chúng ta cần nghiên cứu ít nhất là 2017.

Nếu chúng ta muốn giảm sai số từ 2% xuống 1% (tức $m = 0.01$) thì số lượng đối tượng sẽ là 8067! Chỉ cần thêm độ chính xác 1%, số lượng mẫu có thể thêm hơn 6000 người. Do đó, vấn đề ước tính cỡ mẫu phải rất thận trọng, xem xét cân bằng giữa độ chính xác thông tin cần thu thập và chi phí.

R không có hàm cho ước tính cỡ mẫu cho một tỉ lệ, nhưng với công thức trên, bạn đọc có thể viết một hàm để tính rất dễ dàng.

15.4.5 Ước tính cỡ mẫu cho so sánh hai tỉ lệ

Nhiều nghiên cứu mang tính suy luận thường có hai [hay nhiều hơn hai] nhóm để so sánh. Trong phần 15.4.2 chúng ta đã làm quen với phương pháp ước tính cỡ mẫu để so sánh hai số trung bình bằng kiểm định t. Đó là những người có tiêu chí là những biến số liên tục. Nhưng có nghiên cứu biến số không liên tục mà mang tính nhị phân như tôi vừa bàn trong phần 15.4.3. Để so sánh hai tỉ lệ, phương pháp kiểm định thông dụng

nhất là kiểm định nhị phân (binomial test) hay Chi bình phương (χ^2 test). Trong phần này, tôi sẽ bàn qua cách tính cỡ mẫu cho hai loại kiểm định thống kê này.

Gọi hai tỉ lệ [mà chúng ta không biết nhưng muốn tìm hiểu] là p_1 và p_2 , và gọi $\Delta = p_1 - p_2$. Giả thiết mà chúng ta muốn kiểm định là $\Delta = 0$. Lí thuyết đãng sau để ước tính cỡ mẫu cho kiểm định giả thiết này khá rườm rà, nhưng có thể tóm gọn bằng công thức sau đây:

$$n = \frac{\left(z_{\alpha/2} \sqrt{2\bar{p}(1-\bar{p})} + z_\beta \sqrt{p_1(1-p_1) + p_2(1-p_2)} \right)^2}{\Delta^2}$$

Trong đó, $\bar{p} = (p_1 + p_2)/2$, $z_{\alpha/2}$ là trị số z của phân phối chuẩn cho xác suất $\alpha/2$ (chẳng hạn như khi $\alpha = 0.05$, thì $z_{\alpha/2} = 1.96$; khi $\alpha = 0.01$, thì $z_{\alpha/2} = 2.57$), và z_β là trị số z của phân phối chuẩn cho xác suất β (chẳng hạn như khi $\beta = 0.10$, thì $z_\beta = 1.28$; khi $\beta = 0.20$, thì $z_\beta = 0.84$).

Ví dụ 6: Một thử nghiệm lâm sàng đối chứng ngẫu nhiên được thiết kế để đánh giá hiệu quả của một loại thuốc chống gãy xương sống. Hai nhóm bệnh nhân sẽ được tuyển. Nhóm 1 được điều trị bằng thuốc, và nhóm 2 là nhóm đối chứng (không được điều trị). Các nhà nghiên cứu giả thiết rằng tỉ lệ gãy xương trong nhóm 2 là khoảng 10%, và thuốc có thể làm giảm tỉ lệ này xuống khoảng 6%. Nếu các nhà nghiên cứu muốn thử nghiệm giả thiết này với sai sót I là $\alpha = 0.01$ và power = 0.90, bao nhiêu bệnh nhân cần phải được tuyển mộ cho nghiên cứu?

Ở đây, chúng ta có $\Delta = 0.10 - 0.06 = 0.04$, và $\bar{p} = (0.10 + 0.06)/2 = 0.08$. Với $\alpha = 0.01$, $z_{\alpha/2} = 2.57$ và với power = 0.90, $z_\beta = 1.28$. Do đó, số lượng bệnh nhân cần thiết cho mỗi nhóm là:

$$n = \frac{\left(2.57 \sqrt{2 \times 0.08 \times 0.92} + 1.28 \sqrt{0.1 \times 0.90 + 0.06 \times 0.94} \right)^2}{(0.04)^2} = 1361$$

Như vậy, công trình nghiên cứu này cần phải tuyển ít nhất là 2722 bệnh nhân để kiểm định giả thiết trên.

Hàm `power.prop.test` R có thể ứng dụng để tính cỡ mẫu cho trường hợp trên. Hàm `power.prop.test` cần những thông tin như `power`, `sig.level`, `p1`, và `p2`. Trong ví dụ trên, chúng ta có thể viết:

```
> power.prop.test(p1=0.10, p2=0.06, power=0.90, sig.level=0.01)
```

```
Two-sample comparison of proportions power calculation
```

```
n = 1366.430
p1 = 0.1
p2 = 0.06
sig.level = 0.01
power = 0.9
alternative = two.sided

NOTE: n is number in *each* group
```

Chú ý kết quả từ R có phần chính xác hơn (1366 đối tượng cho mỗi nhóm) vì R dùng nhiều số lẽ cho tính toán hơn là tính “thủ công”.

Trước khi rời chương này, tôi muốn nhân cơ hội này để nhấn mạnh một lần nữa, ước tính cỡ mẫu cho nghiên cứu là một bước cực kì quan trọng trong việc thiết kế một nghiên cứu cho có ý nghĩa khoa học, vì nó có thể quyết định thành bại của nghiên cứu. Trước khi ước tính cỡ mẫu nhà nghiên cứu cần phải biết trước (hay ít ra là có vài giả thiết *cụ thể*) về vấn đề mình quan tâm. Ước tính cỡ mẫu cần một số thông số như đề cập đến trong phần đầu của chương, và nếu các thông số này không có thì không thể ước tính được. Trong trường hợp một nghiên cứu hoàn toàn mới, tức chưa ai từng làm trước đó, có thể các thông số về độ ảnh hưởng và độ dao động đo lường sẽ không có, và nhà nghiên cứu cần phải tiến hành một số mô phỏng (simulation) hay một nghiên cứu sơ khởi để có những thông số cần thiết. Cách ước tính cỡ mẫu bằng mô phỏng là một lĩnh vực nghiên cứu khá chuyên sâu, không nằm trong đề tài của sách này, nhưng bạn đọc có thể tìm hiểu thêm phương pháp này trong các sách giáo khoa về thống kê học cấp cao hơn.

16

Phụ lục 1: Lập trình và hàm với R

R được phát triển sao cho người sử dụng có thể phát triển những hàm thích hợp cho mục đích phân tích và tính toán của mình. Thật vậy, như đã đề cập trong phần đầu của sách, có thể xem R là một ngôn ngữ thống kê, và chúng ta có thể sử dụng ngôn ngữ để giải quyết các vấn đề không thường thấy trong sách giáo khoa. Trong phần này, tôi chỉ trình bày một vài hàm đơn giản để bạn đọc có thể hiểu cách vận hành của R và hi vọng giúp bạn đọc tự phát triển các hàm sau đó.

Hàm (hay có khi còn gọi là “macro” trong các phần mềm khác) thực chất là tập hợp một số lệnh được lưu trữ dưới một cái tên. Ở mức độ đơn giản nhất, hàm là “tốc kí” cho một nhóm lệnh.

Ví dụ 1. Trong các lệnh sau đây, chúng ta tạo hai dữ liệu (data1 và data2). Mỗi dữ liệu có hai cột số liệu được tạo ra bằng mô phỏng từ phân phối chuẩn. Sau đó, vẽ biểu đồ cho hai dữ liệu với ghi chú.

```
data1 <- cbind(rnorm(100,1), rnorm(100,0))
data2 <- cbind(rnorm(100,-1), rnorm(100,0))
xr <- range(rbind(data1,data2)[,1])
yr <- range(rbind(data1,data2)[,2])
plot(data1, xlim=xr, ylim=yr, col=1, xlab="", ylab="")
par(new=T)
plot(data2, xlim=xr, ylim=yr, col=2, xlab="", ylab="")
title(main="My simulated data", xlab="Weight", ylab="Yield")
legend(-3.0, -1.5, c("Big", "Small"), col=1:2, pch=1)
```

Một cách để nhớ tất cả các lệnh này là lưu trữ chúng trong một text file chẵng hạn. Mỗi lần muốn sử dụng, chúng ta chỉ đơn giản cắt và dán các lệnh này vào R. Một cách khác tốt hơn là tạo ra một hàm gồm các lệnh trên để có thể sử dụng nhiều lần.

Mỗi hàm R phải có tên. Tất cả các lệnh được chứa trong khu vực được giới hạn bằng hai kí hiệu { và }. Kí hiệu { cho biết tất cả các lệnh sau đó là nằm trong hàm; và kí hiệu } cho biết chấm dứt hàm. Trong ví dụ trên, chúng ta gọi hàm là `plotfigure`:

```
plotfigure <- function()
{
  data1 <- cbind(rnorm(100,1), rnorm(100,0))
  data2 <- cbind(rnorm(100,-1), rnorm(100,0))
  xr <- range(rbind(data1,data2)[,1])
  yr <- range(rbind(data1,data2)[,2])
  plot(data1, xlim=xr, ylim=yr, col=1, xlab="", ylab="")
  par(new=T)
  plot(data2, xlim=xr, ylim=yr, col=2, xlab="", ylab="")
  title(main="My simulated data", xlab="Weight", ylab="Yield")
```

```

        legend(-3.0, -1.5, c("Big", "Small"), col=1:2, pch=1)
    }

```

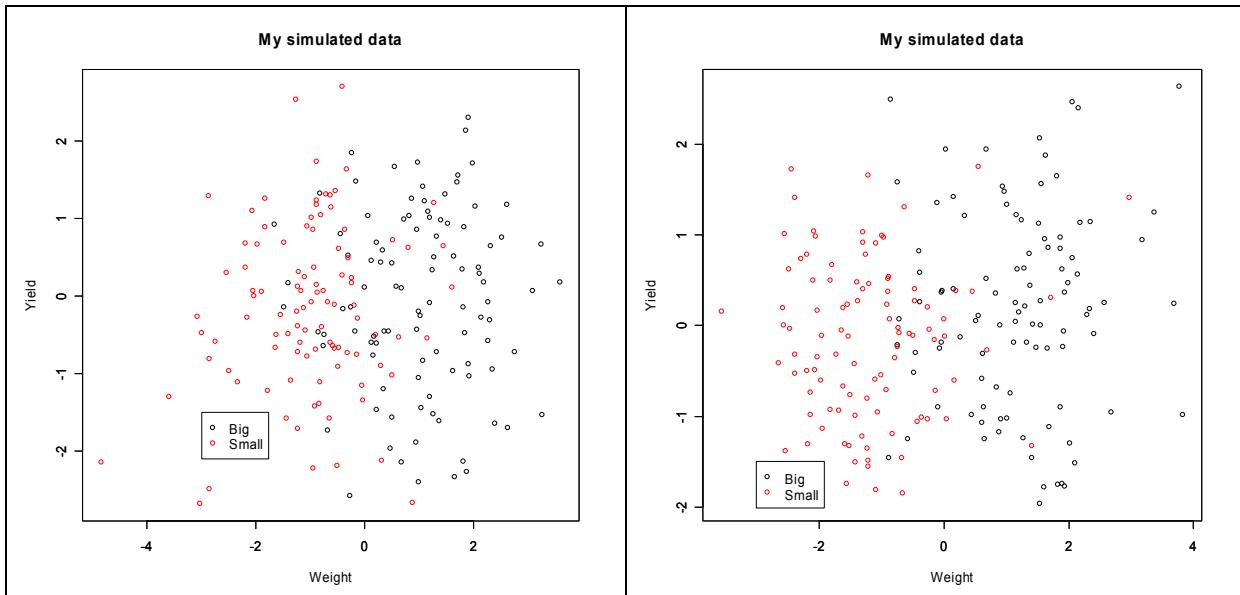
Sau khi đã cho vào R, chúng ta chỉ đơn giản gọi hàm nhiều lần như sau:

```

> plotfigure()
> plotfigure()

```

và kết quả sẽ như sau:



Trong hàm `plotfigure` trên, chúng ta mô phỏng 100 số liệu từ phân phối chuẩn. Và cứ mỗi lần ứng dụng, hàm chỉ tạo ra 100 số liệu, chứ chúng ta không thay đổi được (ngoại trừ phải thay đổi từ lúc biên tập, hay lập hàm). Nói cách khác, hàm trên không có thông số.

Khía cạnh tiện lợi của hàm là chúng ta có thể làm cho thông số thay đổi theo ý muốn của người sử dụng. Chẳng hạn như chúng ta muốn thay đổi số số liệu mô phỏng và trung bình từ luật phân phối chuẩn, chúng ta chỉ cần cho hai con số này là hai thông số (parameters) để người sử dụng có thể thay đổi. Tạm gọi đó là thông số `n`, `mean1`, và `mean2`, thì hàm sẽ như sau:

```

plotfigure <- function(n, mean1, mean2)
{
  data1 <- cbind(rnorm(n,mean1), rnorm(n,0))
  data2 <- cbind(rnorm(n,mean2), rnorm(n,0))
  xr <- range(rbind(data1,data2)[,1])
  yr <- range(rbind(data1,data2)[,2])
  plot(data1, xlim=xr, ylim=yr, col=1, xlab="", ylab="")
  par(new=T)
  plot(data2, xlim=xr, ylim=yr, col=2, xlab="", ylab="")
}

```

```

    title(main="My simulated data", xlab="Weight", ylab="Yield")
    legend(-3.0, -1.5, c("Big", "Small"), col=1:2, pch=1)
}

```

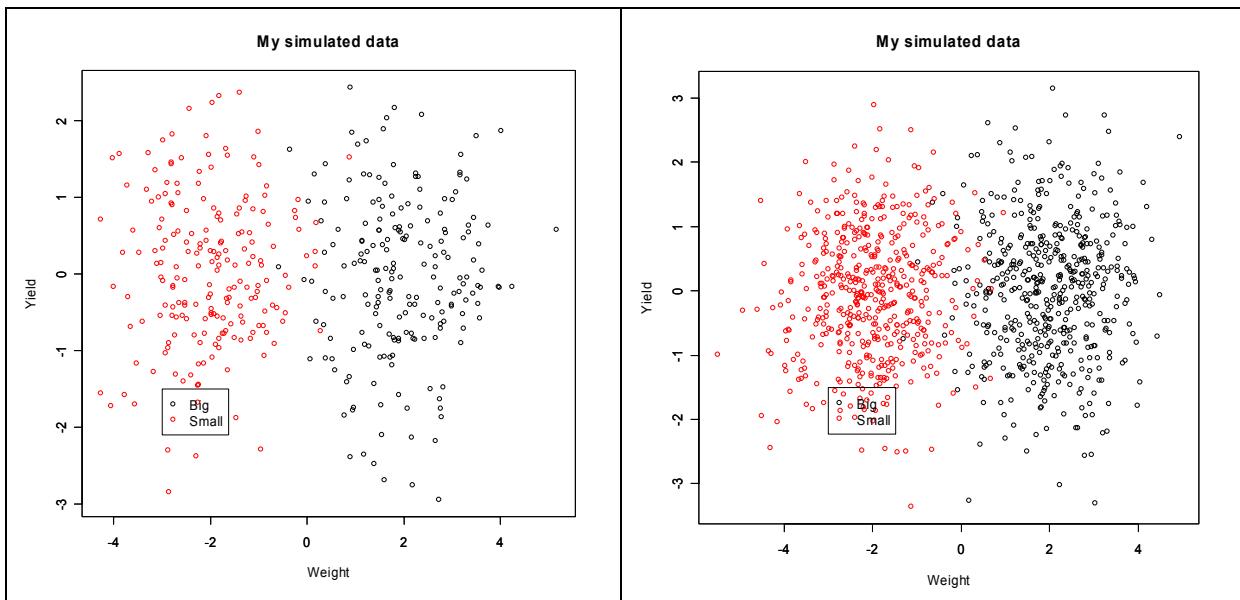
Khi ứng dụng hàm, chúng ta chỉ đơn giản thay đổi `n` và `mean`. Trong hai lệnh sau đây, chúng ta đầu tiên vẽ một biểu đồ tán xạ với 200 số liệu, và số trung bình -2 và 2. Trong lệnh hai, chúng ta nâng số liệu lên 200, nhưng trung bình vẫn như lần mô phỏng trước:

```

> plotfigure(200, 2, -2)
> plotfigure(500, 2, -2)

```

Và kết quả sẽ khác trên:



Ví dụ 2. Chúng ta muốn viết một hàm để cộng hai số. (Tất nhiên R có khả năng làm “việc” này, nhưng vì lí do minh họa, tôi sẽ giả thiết đơn giản như thế). Gọi hàm đó là `add`. Hai thông số `a` và `b` là “arguments”. Cách viết như sau:

```

add <- function(a, b)
{
  sum = a+b
  ans <- "Answer = "
  cat(ans, sum, "\n")
}

```

Thế là xong! Như thấy, bước đầu tiên, chúng ta cho tên hàm là `add` và định nghĩa thông số `a` và `b`. Một hàm phải được mở đầu bằng kí hiệu `{` và chấm dứt bằng `}`. `sum` là một biến số cộng `a` và `b`. `ans <- "Answer = "` định nghĩa trả lời (có thể không cần). `cat(ans, sum, "\n")` có chức năng thu thập số liệu và trình bày kết quả

cho người sử dụng hàm, trong đó “\” có nghĩa là sau khi trình bày, cho người sử dụng một prompt khác. Bạn đọc có thể dán các lệnh trên vào R và thử cho lệnh:

```
> add(3, 9)
Answer = 12

> add(sqrt(5), exp(10))
Answer = 22028.7
```

Ví dụ 3. Hàm sau đây tiến hành nhiều tính toán hơn hàm trong ví dụ 1. Nếu chúng ta có một biến số gồm n phần tử $x_1, x_2, x_3, \dots, x_n$ tuân theo luật phân phối chuẩn với trung bình μ và phương sai σ^2 . Viết theo kí hiệu toán:

$$x_i \sim N(\mu, \sigma^2)$$

Nếu chúng ta có thông tin trước cho biết μ có luật phân phối chuẩn với trung bình θ và phương sai τ^2 , hay:

$$\mu \sim N(\theta, \tau^2)$$

Qua định lí Bayes, chúng ta có thể ước tính trung bình $\mu_p = \frac{\frac{\theta}{\tau^2} + \frac{\bar{x}}{\sigma^2}}{\frac{1}{\tau^2} + \frac{n}{\sigma^2}}$ và phương sai

$\sigma_p^2 = \left(\frac{1}{\tau^2} + \frac{n}{\sigma^2} \right)^{-1}$. Trong đó, \bar{x} là số trung bình của mẫu n . μ_p và σ_p^2 được gọi là

“posterior”. Chúng ta có thể viết một hàm bằng R để tính hai số này như sau. Gọi tên hàm là bayes.

```
bayes <- function(x, prior.mean, prior.var)
{
  n <- length(x)
  sample.mean <- mean(x)
  sample.var <- var(x)
  numerator <- (prior.mean/prior.var) + (n*sample.mean/sample.var)
  denominator <- 1/prior.var + n/sample.var
  posterior.mean = numerator/denominator
  posterior.var = 1/denominator
  a <- "Posterior mean = "
  b <- "Posterior variance = "
  cat("Sample size = ", n, "\n")
  cat("Sample mean = ", sample.mean, "\n")
  cat("Sample var = ", sample.var, "\n")
  cat("Prior mean = ", prior.mean, "\n")
  cat("Prior var = ", prior.var, "\n")
  cat(a, posterior.mean, "\n")
```

```
    cat(b, posterior.var, "\n")
}
```

Ví dụ 4. Mật độ chất khoáng trong xương (bone mineral density - bmd) trong một quần thể thường phân phối theo luật phân phối chuẩn, với giá trị trung bình khoáng 1.0 g/cm^2 và phương sai 0.0144 g/cm^4 . Giả dụ chúng ta đo mật độ xương của một nhóm bệnh nhân như sau: $1.0, 1.5, 2.1, 1.7, 1.8, 0.9, 0.7$. Chúng ta muốn biết giá trị trung bình và phương sai của mẫu này sau khi “điều chỉnh” cho trung bình và phương sai đã biết trước. Trước hết, chúng ta gọi nhóm số liệu này là bmd:

```
> bmd <- c(1.0, 1.5, 2.1, 1.7, 1.8, 0.9, 0.7)
```

và sau đó “gọi” hàm bayes như sau:

```
> bayes(bmd, 1.0, 0.0144)
Sample size = 7
Sample mean = 1.385714
Sample var = 0.2747619
Prior mean = 1
Prior var = 0.0144
Posterior mean = 1.103525
Posterior variance = 0.01053507
```

Trên đây chỉ là một vài hàng giới thiệu cách lập trình và viết hàm bằng ngôn ngữ R. Trong thực tế, tất cả các hàm như survival, BMA, meta, Hmisc, v.v... đều được phát triển bằng ngôn ngữ R. Bạn đọc có thể tham khảo tài liệu “Introduction to R” của W. Venables và B. Ripley (phần cuối của sách) để biết thêm chi tiết kỹ thuật.

17

Phục lục 2

Một số lệnh thông dụng trong R

Lệnh về môi trường vận hành của R

getwd()	Cho biết directory hiện hành là gì
setwd(c:/works)	Chuyển directory vận hành về c:\works (chú ý R dùng "/")
options(prompt="R>")	Đổi prompt thành R>
options(width=100)	Đổi chiều rộng cửa sổ R thành 100 characters
options(scipen=3)	Đổi số thành 3 số thập phân (thay vì kiểu 1.2E-04)
options()	Cho biết các thông số về môi trường hiện nay của R

Lệnh cơ bản

ls()	Liệt kê các đối tượng (objects) trong bộ nhớ
rm(object)	Xóa bỏ đối tượng
search()	Tìm hướng

Kí hiệu tính toán

+	Cộng
-	Trừ
*	Nhân
/	Chia
^	Lũy thừa
%/%	Chia số nguyên
%%	Số dư từ chia hai số nguyên

Kí hiệu logic

==	Bằng
!=	Không bằng
<	Nhỏ hơn
>	Lớn hơn
<=	Nhỏ hơn hoặc bằng
>=	Lớn hơn hoặc bằng
is.na(x)	Có phải x là biến số missing
&	Và (AND)

	Hoặc (OR)
!	Không là (NOT)

Phát số

numeric(n)	Cho ra n số 0
character(n)	Cho ra n ký tự “”
logical(n)	Cho ra n FALSE
seq(-4, 3, 0.5)	Dãy số -4.0, -3.5, -3.0, ..., 3.0
1:10	Giống như lệnh seq(1, 10, 1)
c(5, 7, 9, 1)	Nhập số 5, 7, 8 và 1
rep(1, 5)	Cho ra 5 số 1: 1, 1, 1, 1, 1.
G1(3, 2, 12)	Yêu tố 3 bậc, lặp lại 2 lần, tổng cộng 12 số: 1 1 2 2 3 3 1 1 2 2 3 3

Tạo nên số ngẫu nhiên bằng mô phỏng theo các luật phân phối (simulation)

rnorm(n, mean=0, sd=1)	Phân phối chuẩn (normal distribution) với trung bình = 0 và độ lệch chuẩn = 1.
rexp(n, rate=1)	Phân phối mũ (exponential distribution)
rgamma(n, shape, scale=1)	Phân phối gamma
rpois(n, lambda)	Phân phối Poisson
rweibull(n, shape, scale=1)	Phân phối Weibull
rcauchy(n, location=0, scale=1)	Phân phối Cauchy
rbeta(n, shape1, shape2)	Phân phối beta
rt(n, df)	Phân phối t
rchisq(n, df)	Phân phối Chi bình phương
rbinom(n, size, prob)	Phân phối nhị phân (binomial)
rgeom(n, prob)	Phân phối geometric
rhyper(nn, m, n, k)	hypergeometric
rlnorm(n, meanlog=0, sdlog=1)	Phân phối log normal
rlogis(n, location=0, scale=1)	Phân phối logistic
rnbnom(n, size, prob)	Phân phối negative Binomial
runif(n, min=0, max=1)	Phân phối uniform

Biến đổi số thành ký tự và ngược lại

as.numeric(x)	Biến đổi x thành biến số số học để có thể tính toán
as.character(x)	Biến đổi x thành biến số chữ (character) để phân loại
as.logical(x)	Biến đổi x thành biến số logic
factor(x)	Biến đổi x thành biến số yếu tố

Data frames

data.frame(x, y)	Nhập x và y thành một data frame
tuan\$age	Chọn biến số age từ dataframe tuan.
attach(tuan)	Đưa dataframe tuan vào hệ thống R
detach(tuan)	Xóa bỏ dataframe tuan khỏi hệ thống R

Hàm số toán

log(x)	Logarít bậc e
log10(x)	Logarít bậc 10
exp(x)	Số mũ
sin(x)	Sin
cos(x)	Cosin
tan(x)	Tangent
asin(x)	Arcsin (hàm sin đảo)
acos(x)	Arccosin (hàm cosin đảo)
atan(x)	Arctang(hàm tan đảo)

Hàm số thống kê

min(x)	Số nhỏ nhất của biến số x
max(x)	Số lớn nhất của biến số x
which.max(x)	Tìm dòng nào có giá trị lớn nhất của biến số x
which.min(x)	Tìm dòng nào có giá trị nhỏ nhất của biến số x
length(x)	Tổng số yếu tố (elements) trong một biến số (hay số mẫu)
sum(x)	Số tổng của biến số x
range(x)	Khác biệt giữa max(x) và min(x)
mean(x)	Số trung bình của biến số x
median(x)	Số trung vị (median) của biến số x
sd(x)	Độ lệch chuẩn (standard deviation) của biến số x
var(x)	Phương sai (variance) của biến số x
cov(x, y)	Hiệp biến (covariance) giữa hai biến số x và y
cor(x, y)	Hệ số tương quan (coefficient of correlation) giữa biến số x và y.
quantile(x)	Chỉ số của biến số x
cor(x, y)	Hệ số tương quan (correlation coefficient) giữa biến số x và y
is.na(x)	Kiểm tra xem x có phải là số trống không (missing value)
complete.cases(x1, x2, ...)	Kiểm tra nếu tất cả x1, x2, ... đều không có số trống.

Chỉ số ma trận

<code>x[1]</code>	Số đầu tiên của biến số x
<code>x[1:5]</code>	Năm số đầu tiên của biến số x
<code>x[y<=30]</code>	Chọn x sao cho y nhỏ hơn hoặc bằng 30
<code>x[sex=="male"]</code>	Chọn x sao cho sex bằng male

Nhập dữ liệu

<code>data(name)</code>	Xây dựng một kho dữ liệu
<code>read.table("name")</code>	Đọc / nhập số liệu từ file name
<code>read.csv("name")</code>	Đọc / nhập số liệu dạng excel (cách nhau bằng ",") từ file name
<code>read.delim("name")</code>	Đọc / nhập số liệu dạng tab delimited
<code>read.delim2("name")</code>	Đọc / nhập số liệu dạng tab delimited, cách nhau bằng ";" và số thập phân là ","
<code>read.csv2("name")</code>	Đọc / nhập số liệu dạng csv, cách nhau bằng ";" và số thập phân là ","

Phần phụ trong `read.table`

<code>header=TRUE</code>	Hàng đầu tiên của dữ liệu là tên của biến số
<code>sep=","</code>	Số liệu ngăn cách bằng dấu hiệu ","
<code>dec=","</code>	Số thập phân là "," (để phân biệt với ".")
<code>na.strings=".,"</code>	Số liệu trống (missing value) là ".,"

Phân phối thống kê

<code>pnorm(x, mean, sd)</code>	Phân phối chuẩn
<code>plnorm(x, mean, sd)</code>	Phân phối chuẩn logarit
<code>pt(x, df)</code>	Phân phối t
<code>pf(x, n1, n2)</code>	Phân phối F
<code>pchisq(x, df)</code>	Phân phối Chi bình phương
<code>ppois(x, lambda)</code>	Phân phối Poisson
<code>punif(x, min, max)</code>	Phân phối uniform (đồng dạng)
<code>pexp(x, rate)</code>	Phân phối hàm mũ
<code>pgamma(x, shape, scale)</code>	Phân phối gamma
<code>pbeta(x, a, b)</code>	Phân phối beta

Phân tích thống kê

<code>t.test</code>	Kiểm định t
<code>pairwise.t.test</code>	Kiểm định t cho paired design
<code>cor.test</code>	Kiểm định hệ số tương quan

	method = "kendall" method = "spearman"
var.test bartlett.test	Kiểm định phuong sai Kiểm định nhiều phuong sai
wilcoxon.test kruskal.test friedman.test	Kiểm định Wilcoxon Kiểm định Kruskal Kiểm định Friedman
lm(y ~ x) lm(y ~ factor) lm(y ~ factor+x) lm(y ~ x1+x2+x3)	Phân tích hồi qui tuyến tính (linear regression) Phân tích phuong sai 1 chiều (1-way analysis of variance) Phân tích hiệp biến (analysis of covariance) Phân tích hồi qui tuyến tính đa biến số (multiple linear regression)
binom.test prop.test prop.trend.test fisher.test chisq.test glm(y~x1+x2+x+x3)	Kiểm định nhị phân (Binomial test) Kiểm định so sánh nhiều tỉ số Kiểm định so sánh nhiều tỉ số theo xu hướng Kiểm định Fisher Kiểm định Chi bình phuong Phân tích hồi qui logistic
s<-Surv(time, event) survfit(s) survdiff(s~g) coxph(s ~ x`+x2)	Phân tích survival Biểu đồ Kaplan-Meier Kiểm định Log-rank giữa hai nhóm g Phân tích hồi qui Cox

Đồ thị

plot(y~x)	Vẽ đồ thị y và x (scatter plot)
hist(x)	Vẽ đồ thị y và x (scatter plot)
plot(y ~ x + z)	Vẽ hai biểu đồ x và y theo từng nhóm của z
pie(x)	Vẽ đồ thị tròn
boxplot(x)	Vẽ đồ thị theo dạng hình hộp
qqnorm(x)	Vẽ phân phối quantile của biến số x
qqplot(x, y)	Vẽ phân phối quantile của biến số y theo x
barplot(x)	Vẽ biểu đồ hình khói cho biến số x
hist(x)	Vẽ histogram cho biến số x
stars(x)	Vẽ biểu đồ sao cho biến số x
abline(a, b)	Vẽ đường thẳng với intercept=a và slope=b
abline(h=y)	Vẽ đường thẳng ngang
abline(v=x)	Vẽ đường thẳng đứng
abline(lm.object)	Vẽ đồ thị theo mô hình tuyến tính

Một số thông số cho đồ thị

pch	Kí hiệu để vẽ đồ thị (pch = <i>plotting characters</i>)
mfrow, mfcoll	Tạo ra nhiều cửa sổ để vẽ nhiều đồ thị cùng một lúc (<i>multiframe</i>)
xlim, ylim	Cho giới hạn của trục hoành và trục tung
xlab, ylab	Viết tên trục hoành và trục tung
lty, lwd	Dạng và kích thước của đường biểu diễn
cex, mex	Kích thước và khoảng cách giữa các kí tự.
col	Màu sắc

18

Phục lục 3

Thuật ngữ dùng trong sách

Tiếng Anh

95% confidence interval
Akaike Information criterion (AIC)
Analysis of covariance
Analysis of variance (ANOVA)
Bar chart
Binomial distribution
Box plot
Categorical variable
Clock chart
Coefficient of correlation
Coefficient of determination
Coefficient of heterogeneity
Combination
Continuous variable
Correlation
Covariance
Cross-over experiment
Cumulative probability distribution
Degree of freedom
Determinant
Discrete variable
Dot chart
Estimate
Estimator
Factorial analysis of variance
Fixed effects
Frequency
Function
Heterogeneity
Histogram
Homogeneity
Hypothesis test
Inverse matrix
Latin square experiment

Tiếng Việt

Khoảng tin cậy 95%
Tiêu chuẩn thông tin Akaike
Phân tích hiệp biến
Phân tích phương sai
Biểu đồ thanh
Phân phối nhị phân
Biểu đồ hình hộp
Biến thứ bậc
Biểu đồ đồng hồ
Hệ số tương quan
Hệ số xác định bội
Hệ số bất đồng nhất
Tổ hợp
Biến liên tục
Tương quan
Hợp biến
Thí nghiệm giao chéo
Hàm phân phối tích lũy
Bậc tự do
Định thức
Biến rời rạc
Biểu đồ điểm
Ước số
Hàm ước lượng thống kê
Phân tích phương sai cho thí nghiệm gai thừa
Ảnh hưởng bất biến
Tần số
Hàm
Bất đồng nhất
Biểu đồ tần số
Đồng nhất
Kiểm định giả thiết
Ma trận nghịch đảo
Thí nghiệm hình vuông Latin

Least squares method	Phương pháp bình phương nhỏ nhất
Linear Logistic regression analysis	Phân tích hồi qui tuyến tính logistic
Linear regression analysis	Phân tích hồi qui tuyến tính
Matrix	Ma trận
Maximum likelihood method	Phương pháp hợp lí cực đại
Mean	Số trung bình
Median	Số trung vị
Meta-analysis	Phân tích tổng hợp
Missing value	Giá trị không
Model	Mô hình
Multiple linear regression analysis	Phân tích hồi qui tuyến tính đa biến
Normal distribution	Phân phối chuẩn
Object	Đối tượng
Parameter	Thông số
Permutation	Hoán vị
Pie chart	Biểu đồ hình tròn
Poisson distribution	Phân phối Poisson
Polynomial regression	Hồi qui đa thức
Probability	Xác suất
Probability density distribution	Hàm mật độ xác suất
P-value	Trị số P
Quantile	Hàm định bậc
Random effects	Ảnh hưởng ngẫu nhiên
Random variable	Biến ngẫu nhiên
Relative risk	Tỉ số nguy cơ tương đối
Repeated measure experiment	Thí nghiệm tái đo lường
Residual	Phần dư
Residual mean square	Trung bình bình phương phần dư
Residual sum of squares	Tổng bình phương phần dư
Scalar matrix	Ma trận vô hướng
Scatter plot	Biểu đồ tán xạ
Significance	Có ý nghĩa thống kê
Simulation	Mô phỏng
Standard deviation	Độ lệch chuẩn
Standard error	Sai số chuẩn
Standardized normal distribution	Phân phối chuẩn hóa
Survival analysis	Phân tích biến cố
Transposed matrix	Ma trận chuyển vị
Variable	Biến (biến số)
Variance	Phương sai
Weight	Trọng số

Weighted mean

Trung bình trọng số

19

Lời bạt

(tài liệu tham khảo và đọc thêm)

Qua 15 chương sách và 3 phụ lục bạn đọc đã cùng tôi đi một hành trình khá dài trong phân tích thống kê và biểu đồ. Thiết tưởng trước khi “chia tay” bạn đọc, tôi cũng nên có đôi lời tạm biệt.

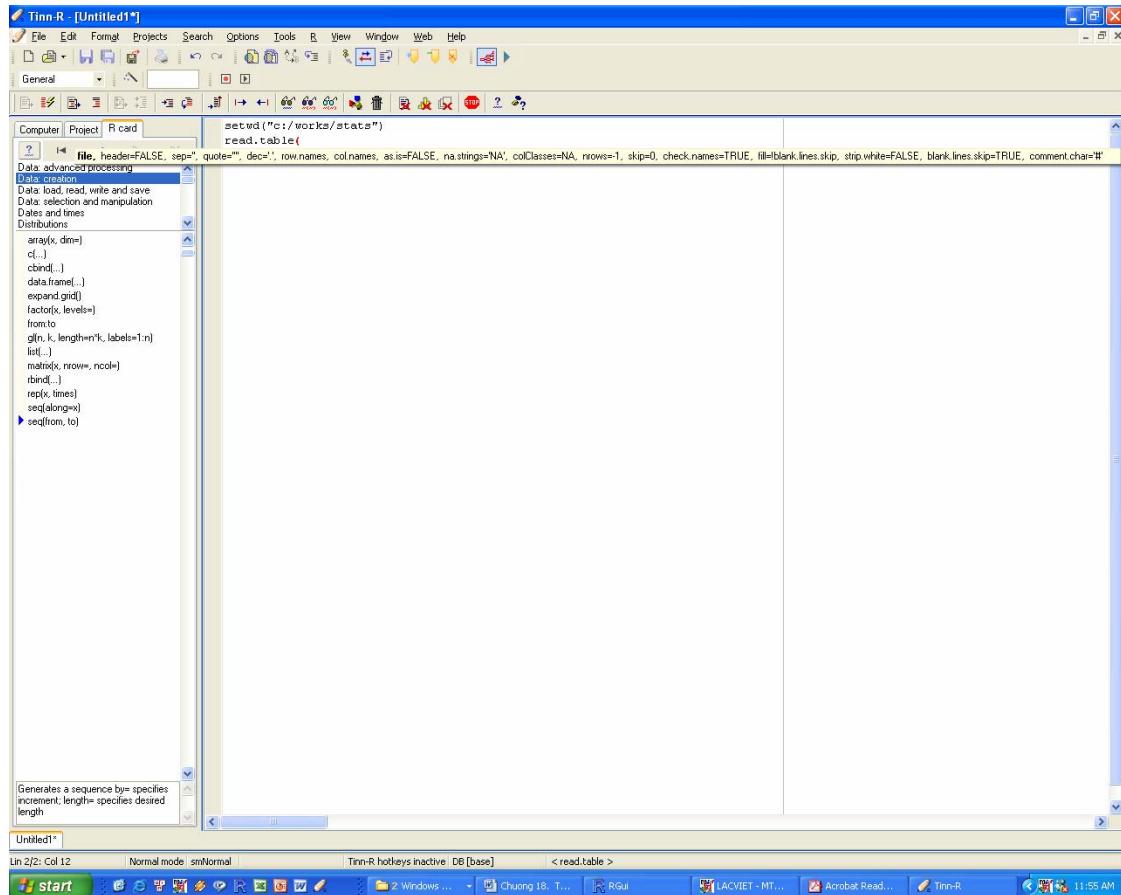
Kinh nghiệm giảng dạy và nghiên cứu cá nhân cho thấy phần lớn sinh viên khi tiếp cận với khoa học thống kê lần đầu là một kinh nghiệm chẳng mấy gì hào hứng, nếu không muốn nói là khó khăn, chỉ vì sách giáo khoa soạn cho môn học này rất xa rời thực tế, hay có khi dính dáng đến thực tế nhưng với những ví dụ vô bổ, nhạt nhẽo. Những khái niệm trừu tượng, những công thức rắc rối, những phép tính phức tạp và rườm rà làm cho người học cảm thấy chao đảo và từ đó cảm thấy thiếu hứng thú theo đuổi môn học. Thật vậy, có khi đọc sách giáo khoa, đọc các bài báo nghiên cứu khoa học, chúng ta bắt gặp những phương pháp hay và những mô hình thích hợp cho nghiên cứu của chính mình, nhưng không biết làm sao tính toán các mô hình đó. Trong cuốn sách này, tôi muốn cung cấp cho bạn đọc một phương tiện phân tích thực tế để lấp vào cái khoảng trống phương pháp đó.

Học phải đi đôi với hành. Cách học về phương pháp hay nhất, theo tôi, là [nói một cách nôm na] ... bắt chước. R cung cấp cho bạn đọc cách học mô phỏng đó rất ư là tiện lợi. Trong khi đọc những chương sách này cùng với những ví dụ, bạn đọc có thể gõ những lệnh vào máy tính và xem kết quả có nhất quán với những gì mình đọc hay không. Sau khi đã biết được cách sử dụng một hàm hay một lệnh nào đó, bạn đọc có thể thêm vào (hay bớt ra) những thông số của hàm để xem kết quả ra sao. Chỉ có học như thế thì bạn đọc mới nắm vững được các khái niệm và cách sử dụng R.

Chúng ta học từ sai sót. Trong sách này, tôi muốn bạn đọc đi một quãng đường khá ... gập ghềnh, tức là bạn đọc phải tương tác với máy tính bằng những lệnh của R. Trong quá trình tương tác đó, có thể một số lệnh sẽ không chạy, vì gõ sai tên biến số hay sai chính tả, vì không để ý đến kí tự viết hoa và viết thường, vì số liệu không đầy đủ hay sai sót, v.v... Tất cả những lần sai sót đó sẽ làm cho bạn đọc rút ra kinh nghiệm và trở nên thuần thạo hơn. Đó là cách học mà người Anh hay gọi là “trial and error”, học từ sai lầm và thử nghiệm.

Một công trình phân tích số liệu cần nhiều lệnh và hàm R. Tuy nhiên, vì tính tương tác mà bạn đọc theo dõi, các lệnh này sẽ biến mất khi ngưng R. Vấn đề đặt ra là có cách nào lưu trữ các lệnh này trong một hồ sơ để sau này sử dụng lại. Phần mềm cực kì có ích cho mục đích này là Tinn-R (cũng có thể tải xuống và cài đặt vào máy hoàn toàn miễn phí). Website để tải Tinn-R và tài liệu sử dụng là: <http://www.sciviews.org/Tinn-R>.

Tinn-R thực chất là một editor cho R (và nhiều phần mềm khác). Tinn-R cho phép chúng ta lưu trữ tất cả các lệnh cho một công trình phân tích trong một hồ sơ. Với Tinn-R, chúng ta có sẵn một chỉ dẫn trực tuyến về cách sử dụng các lệnh hay hàm trong R. Trong khi lệnh gõ sai “văn phạm” R, Tinn-R sẽ báo ngay và đề nghị cách sửa! Giao diện Tinn-R có thể giống như sau:



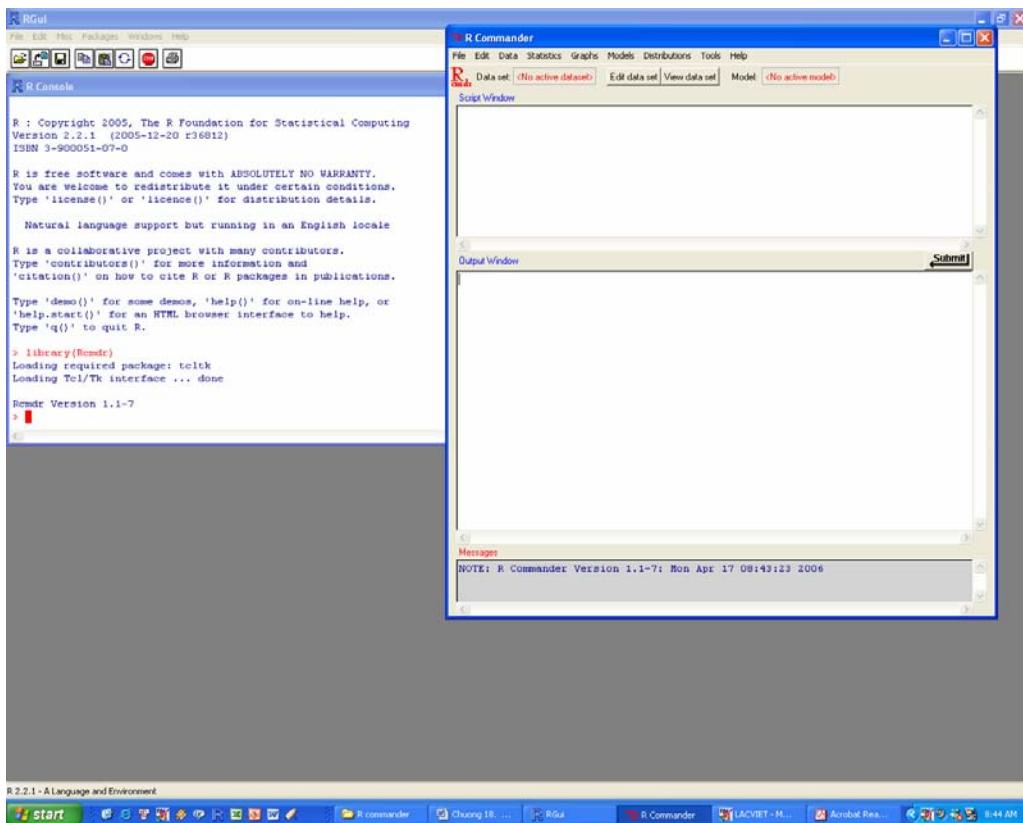
Chẳng hạn như trong giao diện trên, khi chúng ta gõ `read.table` (thì một chỉ dẫn ngay phía dưới hiện ra, với tất cả thông số của hàm `read.table`). Với Tinn-R chúng ta ít khi phạm phải những sai sót nhỏ trong khi chạy R. Sau khi đã xong một số lệnh, chúng ta có thể dùng chuột để tô đậm (highlight) những lệnh cần chạy và gửi sang R. Chú ý chúng ta không cần phải rời Tinn-R trong khi R chạy.

Đến đây, có lẽ bạn đọc sẽ hỏi: có cách nào sử dụng R dễ dàng hơn mà không cần phải gõ các lệnh? Câu trả lời là ... có. Tại sao tôi không giới thiệu trước, ngay từ chương đầu? Tại vì tôi muốn bạn đọc đi con đường khó trước khi đi con đường dễ, nên đến bây giờ mới nói đến một phần mềm phụ khác có khả năng giúp cho bạn đọc sử dụng R một cách nhanh chóng hơn, dễ dàng hơn, và tiện lợi hơn bằng chuột thay vì bằng bàn phím.

Phần mềm để “tự động hóa” R có tên là `Rcmdr` (viết tắt từ `R commander`). Trong thực tế, `Rcmdr` là một package, mà bạn đọc có thể tải từ website chính thức của R

(<http://cran.au.r-project.org/src/contrib/Descriptions/Rcmdr.html>) hay website của tác giả của Rcmdr sau đây: <http://socserv.socsci.mcmaster.ca/jfox/Misc/Rcmdr>. Chú ý, khi Rcmdr vận hành tốt khi có những package sau đây trong máy: `relimp`, `multcomp`, `lmtest`, `effects`, `car`, và `abind`. Nếu chưa có những package này, bạn đọc nên tải chúng về máy. Tài liệu chỉ dẫn Rcmdr cũng có thể tải từ website <http://cran.R-project.org/doc/packages/Rcmdr.pdf>.

Khi đã tải Rcmdr xuống và cài đặt vào máy tính, bạn đọc chỉ đơn giản lệnh: `library(Rcmdr)`, và một giao diện như sau sẽ xuất hiện. Với phần “menu” (như File, Edit, Data, Statistics, Graphs, Models, Distributions, Tools, Help) bạn đọc có thể tự mình khám phá cách vận hành của Rcmdr bằng chuột.



Về nội dung lần in thứ nhất này, tôi không có ý định bàn về những mô hình phân tích đa biến (multivariate analysis model) như phân tích yếu tố (factor analysis), phân tích tập hợp (cluster analysis), phân tích tương quan đa biến (correspondence analysis), phân tích phương sai đa biến (multivariate analysis of variance), v.v... vì đây là những phương pháp tương đối cao cấp, đòi hỏi người sử dụng phải thông thạo chẳng những về lý thuyết thống kê, mà còn phải hiểu rất rõ những phương pháp phân tích căn bản như trình bày trong sách này. Tuy nhiên, bạn đọc có nhu cầu cho các phương pháp phân tích này cũng có thể tìm hiểu trong trang web của R để biết thêm các package chuyên dụng cho phân tích đa biến.

Tài liệu tham khảo

Hiện nay, thư viện sách về R còn tương đối khiêm tốn so với thư viện cho các phần mềm thương mại như SAS và SPSS. Tuy nhiên, trong thời đại tiến bộ phi thường về thông tin internet và toàn cầu hóa như hiện nay, sách in và sách xuất bản trên website không còn là những khác nhau bao xa. Phần lớn chỉ dẫn về cách sử dụng R có thể tìm thấy rải rác đây đó trên các website từ các trường đại học và website cá nhân trên khắp thế giới. Trong phần này tôi chỉ liệt kê một số sách mà bạn đọc, nếu cần tham khảo thêm, nên tìm đọc. Trong quá trình viết cuốn sách mà bạn đọc đang cầm trên tay, tôi cũng tham khảo một số sách và trang web mà tôi sẽ liệt kê sau đây với vài lời nhận xét cá nhân.

Tài liệu tham khảo chính về R là bài báo của hai người sáng tạo ra R: Ihaka R, Gentleman R. R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics* 1996; 5:299-314.

18.1 Sách tham khảo về R

- “**Data Analysis and Graphics Using R – An Example Approach**” (Nhà xuất bản Cambridge University Press, 2003) của John Maindonald nay đã xuất in lại lần thứ 2 với thêm một tác giả mới John Braun. Đây là cuốn sách rất có ích cho những ai muốn tìm hiểu và học về R. Năm chương đầu của sách viết cho bạn đọc chưa từng biết về R, còn các chương sau thì viết cho các bạn đọc đã biết cách sử dụng R thành thạo.
- “**Introductory Statistics With R**” (Nhà xuất bản Springer, 2004) của Peter Dalgaard là một cuốn sách loại căn bản cho R nhắm vào bạn đọc chưa biết gì về R. Sách tương đối ngắn (chỉ khoảng 200 trang) nhưng khá đắt giá!
- “**Linear Models with R**” (Nhà xuất bản Chapman & Hall/CRC, 2004) của Julian Faraway. Sách hiện có thể tải từ internet xuống miễn phí tại website sau đây: <http://www.stat.lsa.umich.edu/~faraway/book/prab.pdf> hay <http://cran.r-project.org/doc/contrib/Faraway-PRA.pdf>. Tài liệu dài 213 trang.
- “**R Graphics (Computer Science and Data Analysis)**” (Nhà xuất bản Chapman & Hall/CRC, 2005) của Paul Murrell. Đây là cuốn sách chuyên về phân tích biểu đồ bằng R. Sách có rất nhiều mã để bạn đọc có thể tự mình thiết kế các biểu đồ phức tạp và ... màu mè.
- “**Modern Applied Statistics with S-Plus**” (Nhà xuất bản Springer, 4th Edition, 2003) của W. N. Venables và B. D. Ripley được viết cho ngôn ngữ S-Plus nhưng tất cả các lệnh và mã trong sách này đều có thể áp dụng cho R mà không cần thay đổi. (S-Plus là tiền thân của R, nhưng S-Plus là một phần mềm thương mại, còn R thì hoàn toàn miễn phí!) Đây là cuốn sách có thể nói là cuốn sách tham khảo cho tất cả ai muốn phát triển thêm về R. Hai tác giả cũng là những chuyên gia có thâm

quyền về ngôn ngữ R. Sách dành cho bạn đọc với trình độ cao về máy tính và thống kê học.

18.2 Các website quan trọng hay có ích về R

- Rất nhiều tài liệu tham khảo có thể tải từ website chính thức của R sau đây:
<http://cran.R-project.org/other-docs.html>

Trong đó có một số tài liệu quan trọng như “**An Introduction to R**” của W. N. Venables và B. D. Ripley.

Địa chỉ internet: <http://cran.r-project.org/doc/manuals/R-intro.pdf>.

- Vài tài liệu hướng dẫn cách sử dụng R có thể tải (miễn phí) và tham khảo như sau:

“**R for Beginners**” (57 trang) của Emmanuel Paradis. Tài liệu được soạn cho bạn đọc mới làm quen với R.

Địa chỉ internet: http://cran.r-project.org/doc/contrib/Paradis-rdebut_en.pdf.

“**Using R for Data Analysis and Graphics: Introduction, Code and Commentary**” (35 trang) của John Maindonald là một tóm lược các lệnh và hàm căn bản của R cho phân tích số liệu và biểu đồ. Chủ đề của tài liệu này rất gần với cuốn sách mà bạn đang đọc.

Địa chỉ internet: <http://cran.r-project.org/doc/contrib/usingR.pdf>

“**Statistical Analysis with R – a quick start**” (46 trang) của Oleg Nenadic và Walter Zucchini. Web. Tài liệu hướng dẫn cách ứng dụng R cho phân tích thống kê và biểu đồ.

Địa chỉ internet: http://www.statoek.wiso.uni-goettingen.de/mitarbeiter/ogi/pub/r_workshop.pdf

“**A Brief Guide to R for Beginners in Econometrics**” (31 trang) của M. Arai. Tài liệu chủ yếu soạn cho giới phân tích thống kê kinh tế.

Địa chỉ internet: http://people.su.se/~ma/R_intro

“**Notes on the use of R for psychology experiments and questionnaires**” (39 trang) của Jonathan Baron và Yuelin Li. Web. Tài liệu được soạn cho giới nghiên cứu tâm lí học và xã hội học. Có ví dụ về log-linear model và một số mô hình phân tích phương sai trong tâm lí học.

Địa chỉ internet: <http://www.psych.upenn.edu/~baron/rpsych/rpsych.html>

- StatsRus gồm một sưu tập về các mẹo để sử dụng R hữu hiệu hơn (dài khoảng 80 trang). Địa chỉ internet: <http://lark.cc.ukans.edu/pauljohn/R/statsRus.html>
- Và sau cùng là một tài liệu “**Hướng dẫn sử dụng R cho phân tích số liệu và biểu đồ**” (khoảng 50 trang – thường xuyên cập nhật hóa) do chính tôi viết bằng tiếng Việt. Website: www.R.ykhoa.net thực chất là tóm lược một số chương chính của

cuốn sách này. Trang web này còn có tất cả các dữ liệu (datasets) và các mã sử trong trong sách để bạn đọc có thể tải xuống máy tính cá nhân để sử dụng.

Phân tích số liệu và biểu đồ bằng



Nguyễn Văn Tuấn
Garvan Institute of Medical Research
Sydney, Australia

Mục lục

1	Tải R xuống và cài đặt vào máy tính	4
2	Tải R package và cài đặt vào máy tính	6
3	“Văn phạm” R	7
3.1	Cách đặt tên trong R	9
3.2	Hỗ trợ trong R	9
4	Cách nhập dữ liệu vào R	10
4.1	Nhập số liệu trực tiếp: <code>c()</code>	10
4.2	Nhập số liệu trực tiếp: <code>edit(data.frame())</code>	12
4.3	Nhập số liệu từ một <i>text file</i> : <code>read.table</code>	13
4.4	Nhập số liệu từ Excel	14
4.5	Nhập số liệu từ SPSS	15
4.6	Thông tin về số liệu	16
4.7	Tạo dãy số bằng hàm <code>seq</code> , <code>rep</code> và <code>gl</code>	17
5	Biên tập số liệu	19
5.1	Tách rời số liệu: <code>subset</code>	19
5.2	Chiết số liệu từ một <code>data.frame</code>	20
5.3	Nhập hai <code>data.frame</code> thành một: <code>merge</code>	21
5.4	Biến đổi số liệu (data coding)	22
5.5	Biến đổi số liệu bằng cách dùng <code>replace</code>	23
5.6	Biến đổi thành yếu tố (<i>factor</i>)	23
5.7	Phân nhóm số liệu bằng <code>cut2</code> (Hmisc)	24
6	Sử dụng R cho tính toán đơn giản	24
6.1	Tính toán đơn giản	24
6.2	Sử dụng R cho các phép tính ma trận	26
7	Sử dụng R cho tính toán xác suất	31
7.1	Phép hoán vị (permutation)	31
7.2	Biến số ngẫu nhiên và hàm phân phối	32
7.3	Biến số ngẫu nhiên và hàm phân phối	32
7.3.1	Hàm phân phối nhị phân (Binomial distribution)	33
7.3.2	Hàm phân phối Poisson (Poisson distribution)	35
7.3.3	Hàm phân phối chuẩn (Normal distribution)	36
7.3.4	Hàm phân phối chuẩn hóa (Standardized Normal distribution)	38
7.4	Chọn mẫu ngẫu nhiên (random sampling)	41
8	Biểu đồ	42
8.1	Số liệu cho phân tích biểu đồ	42
8.2	Biểu đồ cho một biến số rời rạc (discrete variable): <code>barplot</code>	44
8.3	Biểu đồ cho hai biến số rời rạc (discrete variable): <code>barplot</code>	45
8.4	Biểu đồ hình tròn	46
8.5	Biểu đồ cho một biến số liên tục: <code>stripchart</code> và <code>hist</code>	47
8.5.1	<code>Stripchart</code>	47
8.5.2	<code>Histogram</code>	48
8.6	Biểu đồ hộp (<code>boxplot</code>)	49
8.7	Phân tích biểu đồ cho hai biến liên tục	50
8.7.1	Biểu đồ tán xạ (scatter plot)	50
8.8	Phân tích Biểu đồ cho nhiều biến: <code>pairs</code>	53

8.9	Biểu đồ với sai số chuẩn (standard error)	54
9	Phân tích thống kê mô tả	55
9.1	Thống kê mô tả (descriptive statistics, summary)	55
9.2	Thống kê mô tả theo từng nhóm	60
9.3	Kiểm định t (<i>t.test</i>)	61
9.3.1	Kiểm định t một mẫu	61
9.3.2	Kiểm định t hai mẫu	62
9.4	Kiểm định Wilcoxon cho hai mẫu (<i>wilcox.test</i>)	63
9.5	Kiểm định t cho các biến số theo cặp (paired t-test, <i>t.test</i>)	64
9.6	Kiểm định Wilcoxon cho các biến số theo cặp (<i>wilcox.test</i>)	65
9.7	Tần số (frequency)	66
9.8	Kiểm định tỉ lệ (proportion test, <i>prop.test</i> , <i>binom.test</i>)	67
9.9	So sánh hai tỉ lệ (<i>prop.test</i> , <i>binom.test</i>)	68
9.10	So sánh nhiều tỉ lệ (<i>prop.test</i> , <i>chisq.test</i>)	69
9.10.1	Kiểm định Chi bình phương (Chi squared test, <i>chisq.test</i>)	70
9.10.2	Kiểm định Fisher (Fisher's exact test, <i>fisher.test</i>)	71
10	Phân tích hồi qui tuyến tính	71
10.1	Hệ số tương quan	73
10.1.1	Hệ số tương quan Pearson	73
10.1.2	Hệ số tương quan Spearman	74
10.1.3	Hệ số tương quan Kendall	74
10.2	Mô hình của hồi qui tuyến tính đơn giản	75
10.3	Mô hình hồi qui tuyến tính đa biến (multiple linear regression)	82
11	Phân tích phương sai	85
11.1	Phân tích phương sai đơn giản (one-way analysis of variance)	85
11.2	So sánh nhiều nhóm và điều chỉnh trị số p	87
11.3	Phân tích bằng phương pháp phi tham số	90
11.4	Phân tích phương sai hai chiều (two-way ANOVA)	91
12	Phân tích hồi qui logistic	94
12.1	Mô hình hồi qui logistic	95
12.2	Phân tích hồi qui logistic bằng R	97
12.3	Ước tính xác suất bằng R	101
13	Ước tính cỡ mẫu (sample size estimation)	103
13.1	Khái niệm về "power"	104
13.2	Số liệu để ước tính cỡ mẫu	106
13.4	Ước tính cỡ mẫu	107
13.4.1	Ước tính cỡ mẫu cho một chỉ số trung bình	107
13.4.2	Ước tính cỡ mẫu cho so sánh hai số trung bình	108
13.4.3	Ước tính cỡ mẫu cho phân tích phương sai	110
13.4.4	Ước tính cỡ mẫu để ước tính một tỉ lệ	111
13.4.5	Ước tính cỡ mẫu cho so sánh hai tỉ lệ	112
14	Tài liệu tham khảo	115
15	Thuật ngữ dùng trong sách	117

Giới thiệu R

Phân tích số liệu và biểu đồ thường được tiến hành bằng các phần mềm thông dụng như SAS, SPSS, Stata, Statistica, và S-Plus. Đây là những phần mềm được các công ty phần mềm phát triển và giới thiệu trên thị trường khoảng ba thập niên qua, và đã được các trường đại học, các trung tâm nghiên cứu và công ty kĩ nghệ trên toàn thế giới sử dụng cho giảng dạy và nghiên cứu. Nhưng vì chi phí để sử dụng các phần mềm này tương đối đắt tiền (có khi lên đến hàng trăm ngàn đô-la mỗi năm), một số trường đại học ở các nước đang phát triển (và ngay cả ở một số nước đã phát triển) không có khả năng tài chính để sử dụng chúng một cách lâu dài. Do đó, các nhà nghiên cứu thống kê trên thế giới đã hợp tác với nhau để phát triển một phần mềm mới, với chủ trương mã nguồn mở, sao cho tất cả các thành viên trong ngành thống kê học và toán học trên thế giới có thể sử dụng một cách thống nhất và **hoàn toàn miễn phí**.

Năm 1996, trong một bài báo quan trọng về tính toán thống kê, hai nhà thống kê học Ross Ihaka và Robert Gentleman [lúc đó] thuộc Trường đại học Auckland, New Zealand phát hoa một ngôn ngữ mới cho phân tích thống kê mà họ đặt tên là R [1]. Sáng kiến này được rất nhiều nhà thống kê học trên thế giới tán thành và tham gia vào việc phát triển R.

Cho đến nay, qua chưa đầy 10 năm phát triển, càng ngày càng có nhiều nhà thống kê học, toán học, nghiên cứu trong mọi lĩnh vực đã chuyển sang sử dụng R để phân tích dữ liệu khoa học. Trên toàn cầu, đã có một mạng lưới hơn một triệu người sử dụng R, và con số này đang tăng rất nhanh. Có thể nói trong vòng 10 năm nữa, vai trò của các phần mềm thống kê thương mại sẽ không còn lớn như trong thời gian qua nữa.

Vậy R là gì? Nói một cách ngắn gọn, R là một phần mềm sử dụng cho phân tích thống kê và vẽ biểu đồ. Thật ra, về bản chất, R là ngôn ngữ máy tính đa năng, có thể sử dụng cho nhiều mục tiêu khác nhau, từ tính toán đơn giản, toán học giải trí (recreational mathematics), tính toán ma trận (matrix), đến các phân tích thống kê phức tạp. Vì là một ngôn ngữ, cho nên người ta có thể sử dụng R để phát triển thành các phần mềm chuyên môn cho một vấn đề tính toán cá biệt.

Vì thế, những ai làm nghiên cứu khoa học, nhất là ở các nước còn nghèo khó như nước ta, cần phải học cách sử dụng R cho phân tích thống kê và đồ thị. Bài viết ngắn này sẽ hướng dẫn bạn đọc cách sử dụng R. Tôi giả định rằng bạn đọc không biết gì về R, nhưng tôi kì vọng bạn đọc biết qua về cách sử dụng máy tính.

1. Tải R xuống và cài đặt vào máy tính

Để sử dụng R, việc đầu tiên là chúng ta phải cài đặt R trong máy tính của mình. Để làm việc này, ta phải truy nhập vào mạng và vào website có tên là “Comprehensive R Archive Network” (CRAN) sau đây:

<http://cran.R-project.org>

Tài liệu cần tải về, tùy theo phiên bản, nhưng thường có tên bắt đầu bằng mẫu tự R và số phiên bản (version). Chẳng hạn như phiên bản tôi sử dụng vào cuối năm 2005 là 2.2.1, nên tên của tài liệu cần tải là:

R-2.2.1-win32.zip

Tài liệu này khoảng 26 MB, và địa chỉ cụ thể để tải là:

<http://cran.r-project.org/bin/windows/base/R-2.2.1-win32.exe>

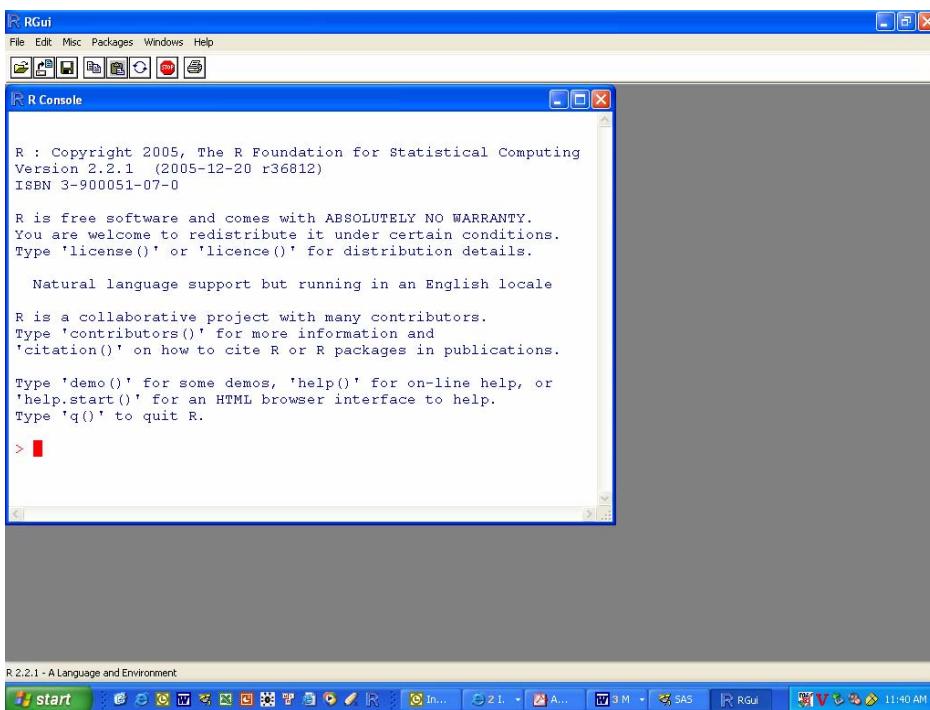
Tại website này, chúng ta có thể tìm thấy rất nhiều tài liệu chỉ dẫn cách sử dụng R, đủ trình độ, từ sơ đẳng đến cao cấp. Nếu chưa quen với tiếng Anh, tài liệu này của tôi có thể cung cấp những thông tin cần thiết để sử dụng mà không cần phải đọc các tài liệu khác.

Khi đã tải R xuống máy tính, bước kế tiếp là cài đặt (set-up) vào máy tính. Để làm việc này, chúng ta chỉ đơn giản nhấn chuột vào tài liệu trên và làm theo hướng dẫn cách cài đặt trên màn hình. Đây là một bước rất đơn giản, chỉ cần 1 phút là việc cài đặt R có thể hoàn tất.

Sau khi hoàn tất việc cài đặt, một *icon*



sẽ xuất hiện trên *desktop* của máy tính. Đến đây thì chúng ta đã sẵn sàng sử dụng R. Có thể nhấp chuột vào icon này và chúng ta sẽ có một *window* như sau:



2. Tải R package và cài đặt vào máy tính

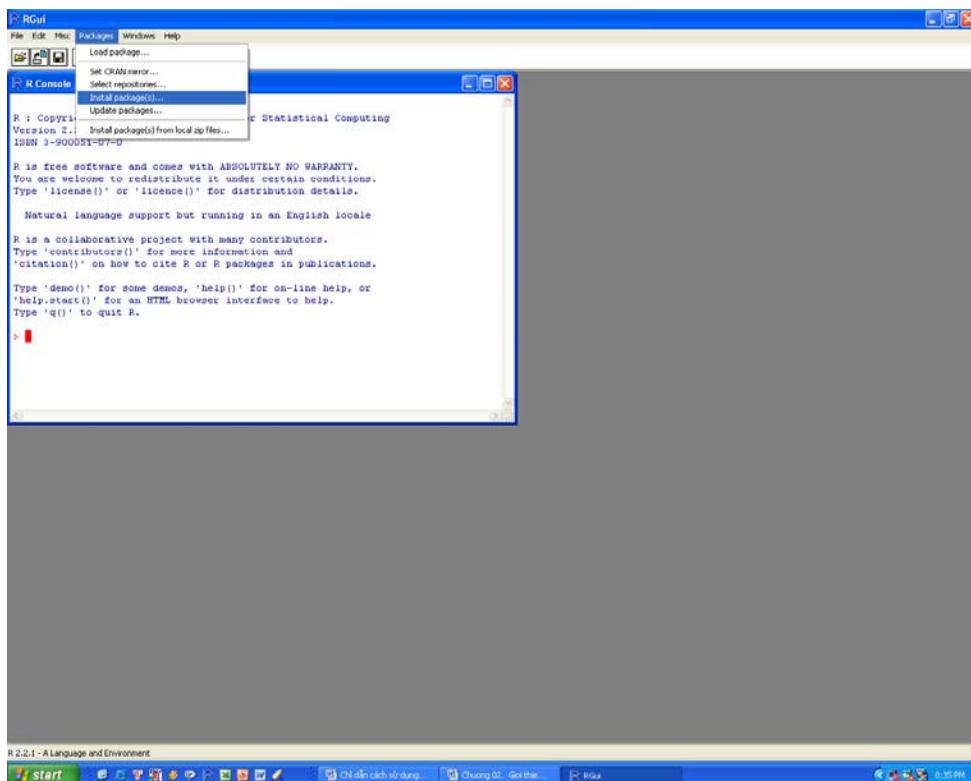
R cung cấp cho chúng ta một “ngôn ngữ” máy tính và một số *function* để làm các phân tích căn bản và đơn giản. Nếu muốn làm những phân tích phức tạp hơn, chúng ta cần phải tải về máy tính một số *package* khác. Package là một phần mềm nhỏ được các nhà thống kê phát triển để giải quyết một vấn đề cụ thể, và có thể chạy trong hệ thống R. Chẳng hạn như để phân tích hồi qui tuyến tính, R có function `lm` để sử dụng cho mục đích này, nhưng để làm các phân tích sâu hơn và phức tạp hơn, chúng ta cần đến các package như `lme4`. Các package này cần phải được tải về và cài đặt vào máy tính.

Địa chỉ để tải các package vẫn là: <http://cran.r-project.org>, rồi bấm vào phần “Packages” xuất hiện bên trái của mục lục trang web. Theo tôi, một số package cần tải về máy tính để sử dụng cho các phân tích dịch tễ học là:

Tên package	Chức năng
trellis	Dùng để vẽ đồ thị và làm cho đồ thị đẹp hơn
lattice	Dùng để vẽ đồ thị và làm cho đồ thị đẹp hơn
Hmisc	Một số phương pháp mô hình dữ liệu của F. Harrell
Design	Một số mô hình thiết kế nghiên cứu của F. Harrell
Epi	Dùng cho các phân tích dịch tễ học
epitools	Một package khác chuyên cho các phân tích dịch tễ học
Foreign	Dùng để nhập dữ liệu từ các phần mềm khác như SPSS, Stata, SAS, v.v...
Rmeta	Dùng cho phân tích tổng hợp (meta-analysis)
meta	Một package khác cho phân tích tổng hợp

survival	Chuyên dùng cho phân tích theo mô hình Cox (Cox's proportional hazard model)
Zelig	Package dùng cho các phân tích thống kê trong lĩnh vực xã hội học
Genetics	Package dùng cho phân tích số liệu di truyền học
BMA	Bayesian Model Average

Các package này có thể cài đặt trực tuyến bằng cách chọn **Install packages** trong phần **packages** của R như hình dưới đây. Ngoài ra, nếu package đã được tải xuống máy tính cá nhân, việc cài đặt có thể nhanh hơn bằng cách chọn **Install package(s) from local zip file** cũng trong phần **packages** (xem hình dưới đây).



3. “Văn phạm” R

R là một ngôn ngữ tương tác (interactive language), có nghĩa là khi chúng ta ra lệnh, và nếu lệnh theo đúng “văn phạm”, R sẽ “đáp” lại bằng một kết quả. Và, sự tương tác tiếp tục cho đến khi chúng ta đạt được yêu cầu. “Văn phạm” chung của R là một lệnh (command) hay function (tôi sẽ thỉnh thoảng đề cập đến là “hàm”). Mà đã là hàm thì phải có thông số; cho nên sau hàm là những thông số mà chúng ta phải cung cấp. Cú pháp chung của R là như sau:

đối tượng <- hàm(thông số 1, thông số 2, ..., thông số n)

Chẳng hạn như:

```
> reg <- lm(y ~ x)
```

thì `reg` là một đối tượng (object), còn `lm` là một hàm, và `y ~ x` là thông số của hàm. Hay:

```
> setwd("c:/works/stats")
```

thì `setwd` là một hàm, còn “`c:/works/stats`” là thông số của hàm.

Để biết một hàm cần có những thông số nào, chúng ta dùng lệnh `args(x)`, (`args` viết tắt chữ `arguments`) mà trong đó `x` là một hàm chúng ta cần biết:

```
> args(lm)
function (formula, data, subset, weights, na.action, method = "qr",
model = TRUE, x = FALSE, y = FALSE, qr = TRUE, singular.ok = TRUE,
contrasts = NULL, offset, ...)
NULL
```

R là một ngôn ngữ “đối tượng” (object oriented language). Điều này có nghĩa là các dữ liệu trong **R** được chứa trong object. Định hướng này cũng có vài ảnh hưởng đến cách viết của **R**. Chẳng hạn như thay vì viết `x = 5` như thông thường chúng ta vẫn viết, thì **R** yêu cầu viết là `x == 5`.

Đối với **R**, `x = 5` tương đương với `x <- 5`. Cách viết sau (dùng kí hiệu `<-`) được khuyến khích hơn là cách viết trước (`=`). Chẳng hạn như:

```
> x <- rnorm(10)
```

có nghĩa là mô phỏng 10 số liệu và chứa trong object `x`. Chúng ta cũng có thể viết `x = rnorm(10)`.

Một số kí hiệu hay dùng trong **R** là:

<code>x == 5</code>	<code>x</code> bằng 5
<code>x != 5</code>	<code>x</code> không bằng 5
<code>y < x</code>	<code>y</code> nhỏ hơn <code>x</code>
<code>x > y</code>	<code>x</code> lớn hơn <code>y</code>
<code>z <= 7</code>	<code>z</code> nhỏ hơn hoặc bằng 7
<code>p >= 1</code>	<code>p</code> lớn hơn hoặc bằng 1
<code>is.na(x)</code>	Có phải <code>x</code> là biến số trống không (missing value)
<code>A & B</code>	<code>A</code> và <code>B</code> (AND)
<code>A B</code>	<code>A</code> hoặc <code>B</code> (OR)
<code>!</code>	Không là (NOT)

Với R, tất cả các câu chữ hay lệnh sau kí hiệu # đều không có hiệu ứng, vì # là kí hiệu dành cho người sử dụng thêm vào các ghi chú, ví dụ:

```
> # lệnh sau đây sẽ mô phỏng 10 giá trị normal
> x <- rnorm(10)
```

3.1 Cách đặt tên trong R

Đặt tên một đối tượng (object) hay một biến số (variable) trong R khá linh hoạt, vì R không có nhiều giới hạn như các phần mềm khác. Tên một object phải được viết liền nhau (tức không được cách rời bằng một khoảng trống). Chẳng hạn như R chấp nhận myobject nhưng không chấp nhận my object.

```
> myobject <- rnorm(10)
> my object <- rnorm(10)
Error: syntax error in "my object"
```

Nhưng đôi khi tên myobject khó đọc, cho nên chúng ta nên tác rời bằng “.” Như my.object.

```
> my.object <- rnorm(10)
```

Một điều quan trọng cần lưu ý là R phân biệt mẫu tự viết hoa và viết thường. Cho nên My.object khác với my.object. Ví dụ:

```
> My.object.u <- 15
> my.object.L <- 5
> My.object.u + my.object.L
[1] 20
```

Một vài điều cần lưu ý khi đặt tên trong R là:

- Không nên đặt tên một biến số hay variable bằng kí hiệu “_” (underscore) như my_object hay my-object.
- Không nên đặt tên một object giống như một biến số trong một dữ liệu. Ví dụ, nếu chúng ta có một data.frame (dữ liệu hay dataset) với biến số age trong đó, thì không nên có một object trùng tên age, tức là không nên viết: age <- age. Tuy nhiên, nếu data.frame tên là data thì chúng ta có thể đề cập đến biến số age với một kí tự \$ như sau: data\$age. (Tức là biến số age trong data.frame data), và trong trường hợp đó, age <- data\$age có thể chấp nhận được.

3.2 Hỗ trợ trong R

Ngoài lệnh `args()` R còn cung cấp lệnh `help()` để người sử dụng có thể hiểu “văn phạm” của từng hàm. Chẳng hạn như muốn biết hàm `lm` có những thông số (arguments) nào, chúng ta chỉ đơn giản lệnh:

```
> help(lm)
```

hay

```
> ?lm
```

Một cửa sổ sẽ hiện ra bên phải của màn hình chỉ rõ cách sử dụng ra sao và thậm chí có cả ví dụ. Bạn đọc có thể đơn giản copy và dán ví dụ vào R để xem cách vận hành.

Trước khi sử dụng R, ngoài sách này nếu cần bạn đọc có thể đọc qua phần chỉ dẫn có sẵn trong R bằng cách chọn mục `help` và sau đó chọn `Html help` như hình dưới đây để biết thêm chi tiết. Bạn đọc cũng có thể copy và dán các lệnh trong mục này vào R để xem cho biết cách vận hành của R.

4. Cách nhập dữ liệu vào R

Muốn làm phân tích dữ liệu bằng R, chúng ta phải có sẵn dữ liệu ở dạng mà R có thể hiểu được để xử lí. Dữ liệu mà R hiểu được phải là dữ liệu trong một `data.frame`. Có nhiều cách để nhập số liệu vào một `data.frame` trong R, từ nhập trực tiếp đến nhập từ các nguồn khác nhau. Sau đây là những cách thông dụng nhất:

4.1 Nhập số liệu trực tiếp: `c()`

Ví dụ 1: chúng ta có số liệu về độ tuổi và insulin cho 10 bệnh nhân như sau, và muốn nhập vào R.

50	16.5
62	10.8
60	32.3
40	19.3
48	14.2
47	11.3
57	15.5
70	15.8
48	16.2
67	11.2

Chúng ta có thể sử dụng function có tên `c` như sau:

```
> age <- c(50, 62, 60, 40, 48, 47, 57, 70, 48, 67)
> insulin <- c(16.5, 10.8, 32.3, 19.3, 14.2, 11.3, 15.5, 15.8, 16.2, 11.2)
```

Lệnh thứ nhất cho R biết rằng chúng ta muốn tạo ra một cột dữ liệu (từ nay tôi sẽ gọi là *biến số*, tức *variable*) có tên là *age*, và lệnh thứ hai là tạo ra một cột khác có tên là *insulin*. Tất nhiên, chúng ta có thể lấy một tên khác mà mình thích.

Chúng ta dùng function *c* (viết tắt của chữ *concatenation* – có nghĩa là “móc nối vào nhau”) để nhập dữ liệu. Chú ý rằng mỗi số liệu cho mỗi bệnh nhân được cách nhau bằng một dấu phẩy.

Kí hiệu *insulin* *<-* (cũng có thể viết là *insulin =*) có nghĩa là các số liệu theo sau sẽ có nằm trong biến số *insulin*. Chúng ta sẽ gặp kí hiệu này rất nhiều lần trong khi sử dụng R.

R là một ngôn ngữ cấu trúc theo dạng đối tượng (thuật ngữ chuyên môn là “object-oriented language”), vì mỗi cột số liệu hay mỗi một *data.frame* là một đối tượng (object) đối với R. Vì thế, *age* và *insulin* là hai đối tượng riêng lẻ. Nay giờ chúng ta cần phải nhập hai đối tượng này thành một *data.frame* để R có thể xử lí sau này. Để làm việc này chúng ta cần đến function *data.frame*:

```
> tuan <- data.frame(age, insulin)
```

Trong lệnh này, chúng ta muốn cho R biết rằng nhập hai cột (hay hai đối tượng) *age* và *insulin* vào một đối tượng có tên là *tuan*.

Đến đây thì chúng ta đã có một đối tượng hoàn chỉnh để tiến hành phân tích thống kê. Để kiểm tra xem trong *tuan* có gì, chúng ta chỉ cần đơn giản gõ:

```
> tuan
```

Và R sẽ báo cáo:

	<i>age</i>	<i>insulin</i>
1	50	16.5
2	62	10.8
3	60	32.3
4	40	19.3
5	48	14.2
6	47	11.3
7	57	15.5
8	70	15.8
9	48	16.2
10	67	11.2

Nếu chúng ta muốn lưu lại các số liệu này trong một file theo dạng R, chúng ta cần dùng lệnh *save*. Giả dụ như chúng ta muốn lưu số liệu trong directory có tên là “*c:\works\insulin*”, chúng ta cần gõ như sau:

```
> setwd("c:/works/insulin")
> save(tuan, file="tuan.rda")
```

Lệnh đầu tiên (`setwd` – chữ `wd` có nghĩa là *working directory*) cho R biết rằng chúng ta muốn lưu các số liệu trong directory có tên là “`c:\works\insulin`”. Lưu ý rằng thông thường Windows dùng dấu backward slash “`/`”, nhưng trong R chúng ta dùng dấu forward slash “`/`”.

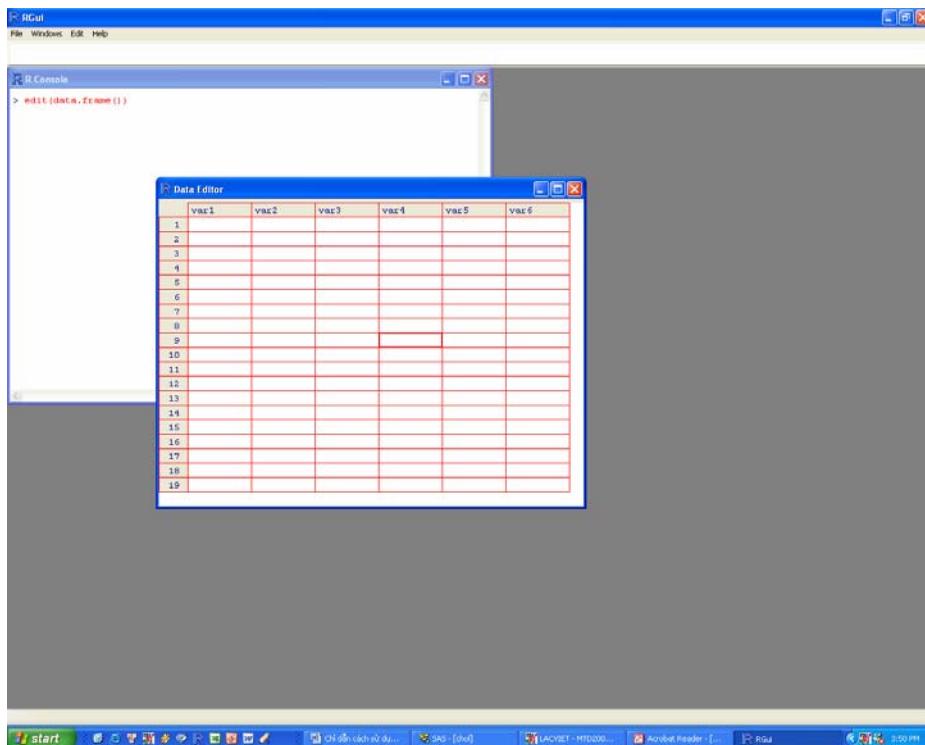
Lệnh thứ hai (`save`) cho R biết rằng các số liệu trong đối tượng `tuan` sẽ lưu trong file có tên là “`tuan.rda`”). Sau khi gõ xong hai lệnh trên, một file có tên `tuan.rda` sẽ có mặt trong directory đó.

4.2 Nhập số liệu trực tiếp: `edit(data.frame())`

Ví dụ 1 (tiếp tục): chúng ta có thể nhập số liệu về độ tuổi và insulin cho 10 bệnh nhân bằng một function rất có ích, đó là: `edit(data.frame())`. Với function này, R sẽ cung cấp cho chúng ta một window mới với một dãy cột và dòng giống như Excel, và chúng ta có thể nhập số liệu trong bảng đó. Ví dụ:

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

Chúng ta sẽ có một cửa sổ như sau:



Ở đây, R không biết chúng ta có biến số nào, cho nên R liệt kê các biến số `var1`, `var2`, v.v... Nhấp chuột vào cột `var1` và thay đổi bằng cách gõ vào đó `age`. Nhấp chuột vào cột `var2` và thay đổi bằng cách gõ vào đó `insulin`. Sau đó gõ số liệu cho

từng cột. Sau khi xong, bấm nút chéo X ở góc phải của spreadsheet, chúng ta sẽ có một data.frame tên `ins` với hai biến số `age` và `insulin`.

4.3 Nhập số liệu từ một text file: `read.table`

Ví dụ 2: Chúng ta thu thập số liệu về độ tuổi và cholesterol từ một nghiên cứu ở 50 bệnh nhân mắc bệnh cao huyết áp. Các số liệu này được lưu trong một text file có tên là `chol.txt` tại directory `c:\works\insulin`. Số liệu này như sau: cột 1 là mã số của bệnh nhân, cột 2 là giới tính, cột 3 là body mass index (bmi), cột 4 là HDL cholesterol (viết tắt là hdl), kế đến là LDL cholesterol, total cholesterol (tc) và triglycerides (tg).

id	sex	age	bmi	hdl	ldl	tc	tg
1	Nam	57	17	5.000	2.0	4.0	1.1
2	Nữ	64	18	4.380	3.0	3.5	2.1
3	Nữ	60	18	3.360	3.0	4.7	0.8
4	Nam	65	18	5.920	4.0	7.7	1.1
5	Nam	47	18	6.250	2.1	5.0	2.1
6	Nữ	65	18	4.150	3.0	4.2	1.5
7	Nam	76	19	0.737	3.0	5.9	2.6
8	Nam	61	19	7.170	3.0	6.1	1.5
9	Nam	59	19	6.942	3.0	5.9	5.4
10	Nữ	57	19	5.000	2.0	4.0	1.9
...							
46	Nữ	52	24	3.360	2.0	3.7	1.2
47	Nam	64	24	7.170	1.0	6.1	1.9
48	Nam	45	24	7.880	4.0	6.7	3.3
49	Nữ	64	25	7.360	4.6	8.1	4.0
50	Nữ	62	25	7.750	4.0	6.2	2.5

Chúng ta muốn nhập các dữ liệu này vào R để tiện việc phân tích sau này. Chúng ta sẽ sử dụng lệnh `read.table` như sau:

```
> setwd("c:/works/insulin")
> chol <- read.table("chol.txt", header=TRUE)
```

Lệnh thứ nhất chúng ta muốn đảm bảo R truy nhập đúng directory mà số liệu đang được lưu giữ. Lệnh thứ hai yêu cầu R nhập số liệu từ file có tên là “`chol.txt`” (trong directory `c:\works\insulin`) và cho vào đối tượng `chol`. Trong lệnh này, `header=TRUE` có nghĩa là yêu cầu R đọc dòng đầu tiên trong file đó như là tên của từng cột dữ kiện.

Chúng ta có thể kiểm tra xem R đã đọc hết các dữ liệu hay chưa bằng cách ra lệnh:

```
> chol
```

Hay

```
> names(chol)
```

R sẽ cho biết có các cột như sau trong dữ liệu (names là lệnh hỏi trong dữ liệu có những cột nào và tên gì):

```
[1] "id"   "sex"  "age"  "bmi"  "hdl"  "ldl"  "tc"   "tg"
```

Bây giờ chúng ta có thể lưu dữ liệu dưới dạng R để xử lí sau này bằng cách ra lệnh:

```
> save(chol, file="chol.rda")
```

4.4 Nhập số liệu từ Excel: `read.csv`

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

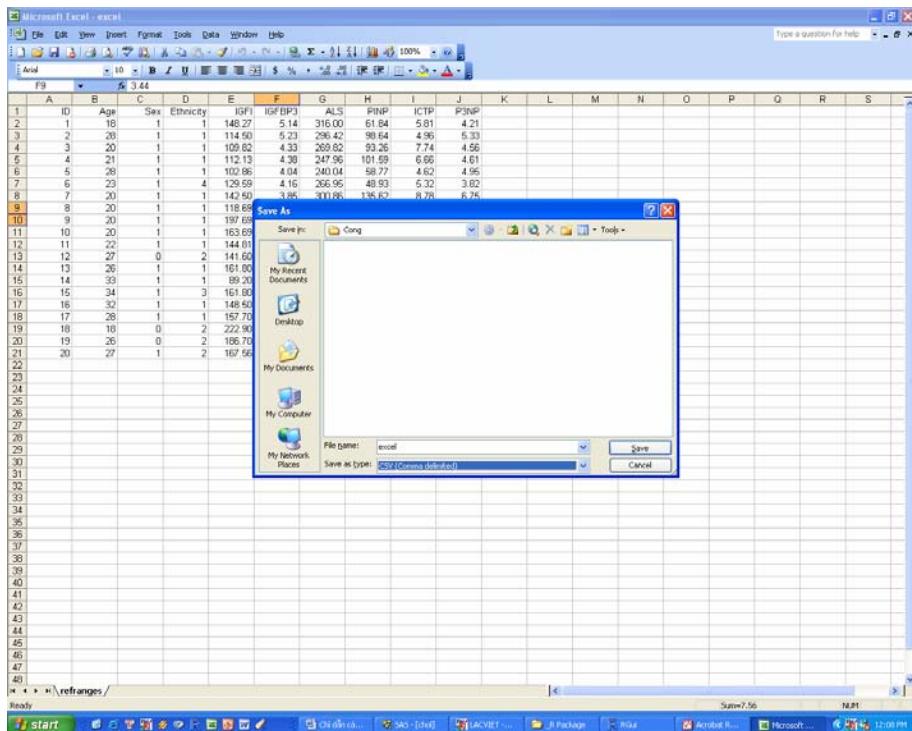
- Bước 1: Dùng lệnh “Save as” trong Excel và lưu số liệu dưới dạng “csv”;
- Bước 2: Dùng R (lệnh `read.csv`) để nhập dữ liệu dạng csv.

Ví dụ 3: Một dữ liệu gồm các cột sau đây đang được lưu trong Excel, và chúng ta muốn chuyển vào R để phân tích. Dữ liệu này có tên là `excel.xls`.

ID	Age	Sex	Ethnicity	IGFI	IGFBP3	ALS	PINP	ICTP	P3NP
1	18	1	1	148.27	5.14	316.00	61.84	5.81	4.21
2	28	1	1	114.50	5.23	296.42	98.64	4.96	5.33
3	20	1	1	109.82	4.33	269.82	93.26	7.74	4.56
4	21	1	1	112.13	4.38	247.96	101.59	6.66	4.61
5	28	1	1	102.86	4.04	240.04	58.77	4.62	4.95
6	23	1	4	129.59	4.16	266.95	48.93	5.32	3.82
7	20	1	1	142.50	3.85	300.86	135.62	8.78	6.75
8	20	1	1	118.69	3.44	277.46	79.51	7.19	5.11
9	20	1	1	197.69	4.12	335.23	57.25	6.21	4.44
10	20	1	1	163.69	3.96	306.83	74.03	4.95	4.84
11	22	1	1	144.81	3.63	295.46	68.26	4.54	3.70
12	27	0	2	141.60	3.48	231.20	56.78	4.47	4.07
13	26	1	1	161.80	4.10	244.80	75.75	6.27	5.26
14	33	1	1	89.20	2.82	177.20	48.57	3.58	3.68
15	34	1	3	161.80	3.80	243.60	50.68	3.52	3.35
16	32	1	1	148.50	3.72	234.80	83.98	4.85	3.80
17	28	1	1	157.70	3.98	224.80	60.42	4.89	4.09
18	18	0	2	222.90	3.98	281.40	74.17	6.43	5.84
19	26	0	2	186.70	4.64	340.80	38.05	5.12	5.77
20	27	1	2	167.56	3.56	321.12	30.18	4.78	6.12

Việc đầu tiên là chúng ta cần làm, như nói trên, là vào Excel để lưu dưới dạng csv:

- Vào Excel, chọn File → Save as
- Chọn Save as type “CSV (Comma delimited)”



Sau khi xong, chúng ta sẽ có một file với tên “excel.csv” trong directory “c:\works\insulin”.

Việc thứ hai là vào R và ra những lệnh sau đây:

```
> setwd("c:/works/insulin")
> gh <- read.csv ("excel.txt", header=TRUE)
```

Lệnh thứ hai `read.csv` yêu cầu R đọc số liệu từ “excel.csv”, dùng dòng thứ nhất là tên cột, và lưu các số liệu này trong một object có tên là `gh`.

Bây giờ chúng ta có thể lưu `gh` dưới dạng R để xử lí sau này bằng lệnh sau đây:

```
> save(gh, file="gh.rda")
```

4.5 Nhập số liệu từ một SPSS: `read.spss`

Phần mềm thống kê SPSS lưu dữ liệu dưới dạng “sav”. Chẳng hạn như nếu chúng ta đã có một dữ liệu có tên là `testo.sav` trong directory `c:\works\insulin`, và muốn chuyển dữ liệu này sang dạng R có thể hiểu được, chúng ta cần sử dụng lệnh `read.spss` trong package có tên là `foreign`. Các lệnh sau đây sẽ hoàn tất dễ dàng việc này:

Việc đầu tiên chúng ta cho truy nhập `foreign` bằng lệnh `library`:

```
> library(foreign)
```

Việc thứ hai là lệnh `read.spss`:

```
> setwd("c:/works/insulin")
> testo <- read.spss("testo.sav", to.data.frame=TRUE)
```

Lệnh thứ hai `read.spss` yêu cầu R đọc số liệu từ “`testo.sav`”, và cho vào một `data.frame` có tên là `testo`.

Bây giờ chúng ta có thể lưu `testo` dưới dạng R để xử lí sau này bằng lệnh sau đây:

```
> save(testo, file="testo.rda")
```

4.6 Thông tin về dữ liệu

Giả dụ như chúng ta đã nhập số liệu vào một `data.frame` có tên là `chol` như trong ví dụ 1. Để tìm hiểu xem trong dữ liệu này có gì, chúng ta có thể nhập vào R như sau:

- Dẫn cho R biết chúng ta muốn xử lí `chol` bằng cách dùng lệnh `attach(arg)` với `arg` là tên của dữ liệu..

```
> attach(chol)
```

- Chúng ta có thể kiểm tra xem `chol` có phải là một `data.frame` không bằng lệnh `is.data.frame(arg)` với `arg` là tên của dữ liệu. Ví dụ:

```
> is.data.frame(chol)
[1] TRUE
```

R cho biết `chol` quả là một `data.frame`.

- Có bao nhiêu cột (hay *variable* = *biến số*) và dòng số liệu (*observations*) trong dữ liệu này? Chúng ta dùng lệnh `dim(arg)` với `arg` là tên của dữ liệu. (`dim` viết tắt chữ `dimension`). Ví dụ (kết quả của R trình bày ngay sau khi chúng ta gõ lệnh):

```
> dim(chol)
[1] 50   8
```

- Như vậy, chúng ta có 50 dòng và 8 cột (hay biến số). Vậy những biến số này tên gì? Chúng ta dùng lệnh `names(arg)` với `arg` là tên của dữ liệu. Ví dụ:

```
> names(chol)
[1] "id"   "sex"  "age"  "bmi"  "hdl"  "ldl"  "tc"   "tg"
```

- Trong biến số `sex`, chúng ta có bao nhiêu nam và nữ? Để trả lời câu hỏi này, chúng ta có thể dùng lệnh `table(arg)` với `arg` là tên của biến số. Ví dụ:

```
> table(sex)
sex
nam Nam Nu
 1   21 28
```

Kết quả cho thấy dữ liệu này có 21 nam và 28 nữ.

4.7 Tạo dãy số bằng hàm `seq`, `rep` và `gl`

R còn có công dụng tạo ra những dãy số rất tiện cho việc mô phỏng và thiết kế thí nghiệm. Những hàm thông thường cho dãy số là `seq` (sequence), `rep` (repetition) và `gl` (generating levels):

Áp dụng `seq`

- Tạo ra một vector số từ 1 đến 12:

```
> x <- (1:12)
> x
[1] 1 2 3 4 5 6 7 8 9 10 11 12

> seq(12)
[1] 1 2 3 4 5 6 7 8 9 10 11 12
```

- Tạo ra một vector số từ 12 đến 5:

```
> x <- (12:5)
> x
[1] 12 11 10 9 8 7 6 5

> seq(12, 7)
[1] 12 11 10 9 8 7
```

Công thức chung của hàm `seq` là `seq(from, to, by=)` hay `seq(from, to, length.out=)`. Cách sử dụng sẽ được minh họa bằng vài ví dụ sau đây:

- Tạo ra một vector số từ 4 đến 6 với khoảng cách bằng 0.25:

```
> seq(4, 6, 0.25)
[1] 4.00 4.25 4.50 4.75 5.00 5.25 5.50 5.75 6.00
```

- Tạo ra một vector 10 số, với số nhỏ nhất là 2 và số lớn nhất là 15

```
> seq(length=10, from=2, to=15)
[1] 2.000000 3.444444 4.888889 6.333333 7.777778 9.222222
10.666667 12.111111 13.555556 15.000000
```

Áp dụng rep

Công thức của hàm rep là `rep(x, times, ...)`, trong đó, `x` là một biến số và `times` là số lần lặp lại. Ví dụ:

- Tạo ra số 10, 3 lần:

```
> rep(10, 3)
[1] 10 10 10
```

- Tạo ra số 1 đến 4, 3 lần:

```
> rep(c(1:4), 3)
[1] 1 2 3 4 1 2 3 4 1 2 3 4
```

- Tạo ra số 1.2, 2.7, 4.8, 5 lần:

```
> rep(c(1.2, 2.7, 4.8), 5)
[1] 1.2 2.7 4.8 1.2 2.7 4.8 1.2 2.7 4.8 1.2 2.7 4.8 1.2 2.7 4.8
```

- Tạo ra số 1.2, 2.7, 4.8, 5 lần:

```
> rep(c(1.2, 2.7, 4.8), 5)
[1] 1.2 2.7 4.8 1.2 2.7 4.8 1.2 2.7 4.8 1.2 2.7 4.8 1.2 2.7 4.8
```

Áp dụng gl

`gl` được áp dụng để tạo ra một biến thứ bậc (categorical variable), tức biến không để tính toán, mà là đếm. Công thức chung của hàm `gl` là `gl(n, k, length = n*k, labels = 1:n, ordered = FALSE)` và cách sử dụng sẽ được minh họa bằng vài ví dụ sau đây:

- Tạo ra biến gồm bậc 1 và 2; mỗi bậc được lặp lại 8 lần:

```
> gl(2, 8)
[1] 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2
Levels: 1 2
```

Hay một biến gồm bậc 1, 2 và 3; mỗi bậc được lặp lại 5 lần:

```
> gl(3, 5)
[1] 1 1 1 1 1 2 2 2 2 3 3 3 3 3
Levels: 1 2 3
```

- Tạo ra biến gồm bậc 1 và 2; mỗi bậc được lặp lại 10 lần (do đó `length=20`):

```
> gl(2, 10, length=20)
[1] 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2
Levels: 1 2
```

Hay:

```
> gl(2, 2, length=20)
[1] 1 1 2 2 1 1 2 2 1 1 2 2 1 1 2 2
Levels: 1 2
```

- Cho thêm kí hiệu:

```
> gl(2, 5, label=c("C", "T"))
[1] C C C C C T T T T T
Levels: C T
```

- Tạo một biến gồm 4 bậc 1, 2, 3, 4. Mỗi bậc lặp lại 2 lần.

```
> rep(1:4, c(2,2,2,2))
[1] 1 1 2 2 3 3 4 4
```

Cũng tương đương với:

```
> rep(1:4, each = 2)
[1] 1 1 2 2 3 3 4 4
```

- Với ngày giờ tháng:

```
> x <- .leap.seconds[1:3]
> rep(x, 2)
[1] "1972-06-30 17:00:00 Pacific Standard Time" "1972-12-31 16:00:00
Pacific Standard Time"
[3] "1973-12-31 16:00:00 Pacific Standard Time" "1972-06-30 17:00:00
Pacific Standard Time"
[5] "1972-12-31 16:00:00 Pacific Standard Time" "1973-12-31 16:00:00
Pacific Standard Time"

> rep(as.POSIXlt(x), rep(2, 3))
[1] "1972-06-30 17:00:00 Pacific Standard Time" "1972-06-30 17:00:00
Pacific Standard Time"
[3] "1972-12-31 16:00:00 Pacific Standard Time" "1972-12-31 16:00:00
Pacific Standard Time"
[5] "1973-12-31 16:00:00 Pacific Standard Time" "1973-12-31 16:00:00
Pacific Standard Time"
```

5. Biên tập số liệu

5.1 Tách rời dữ liệu: subset

Chúng ta sẽ quay lại với dữ liệu chol trong ví dụ 1. Để tiện việc theo dõi và hiểu “câu chuyện”, tôi xin nhắc lại rằng chúng ta đã nhập số liệu vào trong một dữ liệu R có tên là chol từ một text file có tên là chol.txt:

```
> setwd("c:/works/insulin")
> chol <- read.table("chol.txt", header=TRUE)
> attach(chol)
```

Nếu chúng ta, vì một lí do nào đó, chỉ muốn phân tích riêng cho nam giới, chúng ta có thể tách chol ra thành hai data.frame, tạm gọi là nam và nu. Để làm chuyện này, chúng ta dùng lệnh subset(data, cond), trong đó data là data.frame mà chúng ta muốn tách rời, và cond là điều kiện. Ví dụ:

```
> nam <- subset(chol, sex=="Nam")
> nu <- subset(chol, sex=="Nu")
```

Sau khi ra hai lệnh này, chúng ta đã có 2 dữ liệu (hai data.frame) mới tên là nam và nu. Chú ý điều kiện `sex == "Nam"` và `sex == "Nu"` chúng ta dùng == thay vì = để chỉ điều kiện chính xác.

Tất nhiên, chúng ta cũng có thể tách dữ liệu thành nhiều data.frame khác nhau với những điều kiện dựa vào các biến số khác. Chẳng hạn như lệnh sau đây tạo ra một data.frame mới tên là `old` với những bệnh nhân trên 60 tuổi:

```
> old <- subset(chol, age>=60)
> dim(old)
[1] 25 8
```

Hay một data.frame mới với những bệnh nhân trên 60 tuổi và nam giới:

```
> n60 <- subset(chol, age>=60 & sex=="Nam")
> dim(n60)
[1] 9 8
```

5.2 Chiết số liệu từ một data .frame

Trong `chol` có 8 biến số. Chúng ta có thể chiết dữ liệu `chol` và chỉ giữ lại những biến số cần thiết như mã số (`id`), độ tuổi (`age`) và total cholesterol (`tc`). Để ý từ lệnh `names(chol)` rằng biến số `id` là cột số 1, `age` là cột số 3, và biến số `tc` là cột số 7. Chúng ta có thể dùng lệnh sau đây:

```
> data2 <- chol[, c(1,3,7)]
```

Ở đây, chúng ta lệnh cho R biết rằng chúng ta muốn chọn cột số 1, 3 và 7, và đưa tất cả số liệu của hai cột này vào data.frame mới có tên là `data2`. Chú ý chúng ta sử dụng ngoặc kép vuông [] chứ không phải ngoặc kép vòng (), vì `chol` không phải là một function. Dấu phẩy phía trước `c`, có nghĩa là chúng ta chọn tất cả các dòng số liệu trong data.frame `chol`.

Nhưng nếu chúng ta chỉ muốn chọn 10 dòng số liệu đầu tiên, thì lệnh sẽ là:

```
> data3 <- chol[1:10, c(1,3,7)]
> print(data3)
  id sex  tc
1   1  Nam 4.0
2   2   Nu 3.5
3   3   Nu 4.7
4   4  Nam 7.7
5   5  Nam 5.0
6   6   Nu 4.2
7   7  Nam 5.9
8   8  Nam 6.1
```

```
9   9  Nam 5.9
10 10  Nu 4.0
```

Chú ý lệnh `print(arg)` đơn giản liệt kê tất cả số liệu trong `data.frame arg`. Thật ra, chúng ta chỉ cần đơn giản gõ `data3`, kết quả cũng giống y như `print(data3)`.

5.3 Nhập hai `data.frame` thành một: `merge`

Giả dụ như chúng ta có dữ liệu chứa trong hai `data.frame`. Dữ liệu thứ nhất tên là `d1` gồm 3 cột: `id`, `sex`, `tc` như sau:

```
id sex tc
1  Nam 4.0
2  Nu 3.5
3  Nu 4.7
4  Nam 7.7
5  Nam 5.0
6  Nu 4.2
7  Nam 5.9
8  Nam 6.1
9  Nam 5.9
10 Nu 4.0
```

Dữ liệu thứ hai tên là `d2` gồm 3 cột: `id`, `sex`, `tg` như sau:

```
id sex tg
1  Nam 1.1
2  Nu 2.1
3  Nu 0.8
4  Nam 1.1
5  Nam 2.1
6  Nu 1.5
7  Nam 2.6
8  Nam 1.5
9  Nam 5.4
10 Nu 1.9
11 Nu 1.7
```

Hai dữ liệu này có chung hai biến số `id` và `sex`. Nhưng dữ liệu `d1` có 10 dòng, còn dữ liệu `d2` có 11 dòng. Chúng ta có thể nhập hai dữ liệu thành một `data.frame` bằng cách dùng lệnh `merge` như sau:

```
> d <- merge(d1, d2, by="id", all=TRUE)
> d
  id sex.x  tc sex.y  tg
```

1	1	Nam	4.0	Nam	1.1	
2	2	Nu	3.5	Nu	2.1	
3	3	Nu	4.7	Nu	0.8	
4	4	Nam	7.7	Nam	1.1	
5	5	Nam	5.0	Nam	2.1	
6	6	Nu	4.2	Nu	1.5	
7	7	Nam	5.9	Nam	2.6	
8	8	Nam	6.1	Nam	1.5	
9	9	Nam	5.9	Nam	5.4	
10	10		Nu	4.0	Nu	1.9
11	11	<NA>	NA		Nu	1.7

Trong lệnh `merge`, chúng ta yêu cầu R nhập 2 dữ liệu `d1` và `d2` thành một và đưa vào `data.frame` mới tên là `d`, và dùng biến số `i` để làm chuẩn. Chúng ta để ý thấy bệnh nhân số 11 không có số liệu cho `tc`, cho nên R cho là NA (một dạng “not available”).

5.4 Biến đổi số liệu (data coding)

Trong việc xử lí số liệu dịch tễ học, nhiều khi chúng ta cần phải biến đổi số liệu từ biến liên tục sang biến mang tính cách phân loại. Chẳng hạn như trong chẩn đoán loãng xương, những phụ nữ có chỉ số T của mật độ chất khoáng trong xương (bone mineral density hay BMD) bằng hay thấp hơn -2.5 được xem là “loãng xương”, những ai có BMD giữa -2.5 và -1.0 là “xốp xương” (osteopenia), và trên -1.0 là “bình thường”. Ví dụ, chúng ta có số liệu BMD từ 10 bệnh nhân như sau:

```
-0.92, 0.21, 0.17, -3.21, -1.80, -2.60, -2.00, 1.71, 2.12, -2.11
```

Để nhập các số liệu này vào R chúng ta có thể sử dụng *function* `c` như sau:

```
bmd <- c(-0.92, 0.21, 0.17, -3.21, -1.80, -2.60, -2.00, 1.71, 2.12, -2.11)
```

Để phân loại 3 nhóm loãng xương, xốp xương, và bình thường, chúng ta có thể dùng mã số 1, 2 và 3. Nói cách khác, chúng ta muốn tạo nên một biến số khác (hãy gọi là `diagnosis`) gồm 3 giá trị trên dựa vào giá trị của `bmd`. Để làm việc này, chúng ta sử dụng lệnh:

```
# tạm thời cho biến số diagnosis bằng bmd
> diagnosis <- bmd

# biến đổi bmd thành diagnosis
> diagnosis[bmd <= -2.5] <- 1
> diagnosis[bmd > -2.5 & bmd <= 1.0] <- 2
> diagnosis[bmd > -1.0] <- 3

# tạo thành một data frame
> data <- data.frame(bmd, diagnosis)

# liệt kê để kiểm tra xem lệnh có hiệu quả không
> data
```

```
bmd diagnosis
1 -0.92      3
2  0.21      3
3  0.17      3
4 -3.21      1
5 -1.80      2
6 -2.60      1
7 -2.00      2
8  1.71      3
9  2.12      3
10 -2.11     2
```

5.5 Biến đổi số liệu bằng cách dùng *replace*

Một cách biến đổi số liệu khác là dùng *replace*, dù cách này có vẻ rườm rà chút ít. Tiếp tục ví dụ trên, chúng ta biến đổi từ *bmd* sang *diagnosis* như sau:

```
> diagnosis <- bmd
> diagnosis <- replace(diagnosis, bmd <= -2.5, 1)
> diagnosis <- replace(diagnosis, bmd > -2.5 & bmd <= 1.0, 2)
> diagnosis <- replace(diagnosis, bmd > 1.0, 3)
```

5.6 Biến đổi thành yếu tố (*factor*)

Trong phân tích thống kê, chúng ta phân biệt một biến số mang tính *yếu tố* (*factor*) và biến số liên tục bình thường. Biến số yếu tố không thể dùng để tính toán như cộng trừ nhân chia, nhưng biến số số học có thể sử dụng để tính toán. Chẳng hạn như trong ví dụ *bmd* và *diagnosis* trên, *diagnosis* là yếu tố vì giá trị trung bình giữa 1 và 2 chẳng có ý nghĩa thực tế gì cả; còn *bmd* là biến số số học.

Nhưng hiện nay, *diagnosis* được xem là một biến số số học. Để biến thành biến số yếu tố, chúng ta cần sử dụng *function factor* như sau:

```
> diag <- factor(diagnosis)
> diag
[1] 3 3 3 1 2 1 2 3 3 2
Levels: 1 2 3
```

Chú ý R bây giờ thông báo cho chúng ta biết *diag* có 3 bậc: 1, 2 và 3. Nếu chúng ta yêu cầu R tính số trung bình của *diag*, R sẽ không làm theo yêu cầu này, vì đó không phải là một biến số số học:

```
> mean(diag)
[1] NA
Warning message:
argument is not numeric or logical: returning NA in: mean.default(diag)
```

Dĩ nhiên, chúng ta có thể tính giá trị trung bình của *diagnosis*:

```
> mean(diagnosis)
[1] 2.3
```

nhưng kết quả 2.3 này không có ý nghĩa gì trong thực tế cả.

5.7 Phân nhóm số liệu bằng `cut2` (Hmisc)

Trong phân tích thống kê, có khi chúng ta cần phải phân chia một biến số liên tục thành nhiều nhóm dựa vào phân phối của biến số. Chẳng hạn như đối với biến số `bmd` chúng ta có thể “cắt” dãy số thành 3 nhóm tương đương nhau bằng cách dùng function `cut2` (trong thư viện `Hmisc`) như sau:

```
> # nhập thư viện Hmisc để có thể dùng function cut2
> library(Hmisc)
> bmd <- c(-0.92, 0.21, 0.17, -3.21, -1.80, -2.60, -2.00, 1.71, 2.12, -2.11)
> # chia biến số bmd thành 2 nhóm và để trong đối tượng group
> group <- cut2(bmd, g=2)
> table(group)
group
[-3.21, -0.92) [-0.92, 2.12]
      5           5
```

Như thấy qua ví dụ trên, `g = 2` có nghĩa là chia thành 2 nhóm (`g = group`). R tự động chia thành nhóm 1 gồm giá trị `bmd` từ -3.21 đến -0.92, và nhóm 2 từ -0.92 đến 2.12. Mỗi nhóm gồm có 5 số.

Tất nhiên, chúng ta cũng có thể chia thành 3 nhóm bằng lệnh:

```
> group <- cut2(bmd, g=3)
```

Và với lệnh `table` chúng ta sẽ biết có 3 nhóm, nhóm 1 gồm 4 số, nhóm 2 và 3 mỗi nhóm có 3 số:

```
> table(group)
group
[-3.21, -1.80) [-1.80, 0.21) [ 0.21, 2.12]
      4           3           3
```

6. Sử dụng R cho tính toán đơn giản

Một trong những lợi thế của R là có thể sử dụng như một ... máy tính cầm tay. Thật ra, hơn thế nữa, R có thể sử dụng cho các phép tính ma trận và lập chương. Trong chương này tôi chỉ trình bày một số phép tính đơn giản mà học sinh hay sinh viên có thể sử dụng lập tức trong khi đọc những dòng chữ này.

6.1 Tính toán đơn giản

Cộng hai số hay nhiều số với nhau: <pre>> 15+2997 [1] 3012</pre>	Cộng và trừ: <pre>> 15+2997-9768 [1] -6756</pre>
Nhân và chia <pre>> -27*12/21 [1] -15.42857</pre>	Số lũy thừa: $(25 - 5)^3$ <pre>> (25 - 5)^3 [1] 8000</pre>
Căn số bậc hai: $\sqrt{10}$ <pre>> sqrt(10) [1] 3.162278</pre>	Số pi (π) <pre>> pi [1] 3.141593 > 2+3*pi [1] 11.42478</pre>
Logarit: \log_e <pre>> log(10) [1] 2.302585</pre>	Logarit: \log_{10} <pre>> log10(100) [1] 2</pre>
Số mũ: $e^{2.7689}$ <pre>> exp(2.7689) [1] 15.94109</pre> <pre>> log10(2+3*pi) [1] 1.057848</pre>	Hàm số lượng giác <pre>> cos(pi) [1] -1</pre>
Vector <pre>> x <- c(2,3,1,5,4,6,7,6,8) > x [1] 2 3 1 5 4 6 7 6 8 > sum(x) [1] 42 > x^2 [1] 4 6 2 10 8 12 14 12 16</pre>	<pre>> exp(x/10) [1] 1.221403 1.349859 1.105171 1.648 1.491825 1.822119 2.013753 1.822119 [9] 2.225541 > exp(cos(x/10)) [1] 2.664634 2.599545 2.704736 2.405 2.511954 2.282647 2.148655 2.282647 [9] 2.007132</pre>
Tính tổng bình phương (sum of squares): $1^2 + 2^2 + 3^2 + 4^2 + 5^2 = ?$ <pre>> x <- c(1,2,3,4,5) > sum(x^2) [1] 55</pre>	Tính tổng bình phương điều chỉnh (adjusted sum of squares): $\sum_{i=1}^n (x_i - \bar{x})^2 = ?$ <pre>> x <- c(1,2,3,4,5) > sum((x-mean(x))^2) [1] 10</pre> <p>Trong công thức trên <code>mean(x)</code> là số trung bình của vector <code>x</code>.</p>
Tính sai số bình phương (mean square):	Tính phương sai (variance) và độ lệch chuẩn (standard deviation):

$\sum_{i=1}^n (x_i - \bar{x})^2 / n = ?$ <pre>> x <- c(1, 2, 3, 4, 5) > sum((x - mean(x))^2) / length(x) [1] 2</pre> <p>Trong công thức trên, <code>length(x)</code> có nghĩa là tổng số phần tử (elements) trong vector <code>x</code>.</p>	Phương sai: $s^2 = \sum_{i=1}^n (x_i - \bar{x})^2 / (n-1) = ?$ <pre>> x <- c(1, 2, 3, 4, 5) > var(x) [1] 2.5</pre> <p>Độ lệch chuẩn: $\sqrt{s^2}$:</p> <pre>> sd(x) [1] 1.581139</pre>
--	--

6.2 Sử dụng R cho các phép tính ma trận

Như chúng ta biết ma trận (matrix), nói đơn giản, gồm có dòng (row) và cột (column). Khi viết `A[m, n]`, chúng ta hiểu rằng ma trận A có m dòng và n cột. Trong R, chúng ta cũng có thể thể hiện như thế. Ví dụ: chúng ta muốn tạo một ma trận vuông A gồm 3 dòng và 3 cột, với các phần tử (element) 1, 2, 3, 4, 5, 6, 7, 8, 9, chúng ta viết:

$$A = \begin{pmatrix} 1 & 4 & 7 \\ 2 & 5 & 8 \\ 3 & 6 & 9 \end{pmatrix}$$

Và với R:

```
> y <- c(1, 2, 3, 4, 5, 6, 7, 8, 9)
> A <- matrix(y, nrow=3)
> A
[,1] [,2] [,3]
[1,]    1    4    7
[2,]    2    5    8
[3,]    3    6    9
```

Nhưng nếu chúng ta lệnh:

```
> A <- matrix(y, nrow=3, byrow=TRUE)
> A
```

thì kết quả sẽ là:

```
[,1] [,2] [,3]
[1,]    1    2    3
[2,]    4    5    6
[3,]    7    8    9
```

Tức là một **ma trận chuyển vị (transposed matrix)**. Một cách khác để tạo một ma trận hoán vị là dùng `t()`. Ví dụ:

```
> y <- c(1,2,3,4,5,6,7,8,9)
> A <- matrix(y, nrow=3)
> A
[,1] [,2] [,3]
[1,]    1    4    7
[2,]    2    5    8
[3,]    3    6    9
```

và $B = A'$ có thể diễn tả bằng R như sau:

```
> B <- t(A)
> B
[,1] [,2] [,3]
[1,]    1    2    3
[2,]    4    5    6
[3,]    7    8    9
```

Ma trận vô hướng (scalar matrix) là một ma trận vuông (tức số dòng bằng số cột), và tất cả các phần tử ngoài đường chéo (off-diagonal elements) là 0, và phần tử đường chéo là 1. Chúng ta có thể tạo một ma trận như thế bằng R như sau:

```
> # tạo ra một ma trận 3 x 3 với tất cả phần tử là 0.
> A <- matrix(0, 3, 3)

> # cho các phần tử đường chéo bằng 1
> diag(A) <- 1
> diag(A)
[1] 1 1 1

> # bây giờ ma trận A sẽ là:
> A
[,1] [,2] [,3]
[1,]    1    0    0
[2,]    0    1    0
[3,]    0    0    1
```

6.2.1 Chiết phần tử từ ma trận

```
> y <- c(1,2,3,4,5,6,7,8,9)
> A <- matrix(y, nrow=3)
> A
[,1] [,2] [,3]
[1,]    1    4    7
[2,]    2    5    8
[3,]    3    6    9

> # cột 1 của ma trận A
> A[,1]
```

```
[1] 1 4 7

> # cột 3 của ma trận A
> A[3,]
[1] 7 8 9

> # dòng 1 của ma trận A
> A[1,]
[1] 1 2 3

> # dòng 2, cột 3 của ma trận A
> A[2,3]
[1] 6

> # tất cả các dòng của ma trận A, ngoại trừ dòng 2
> A[-2,]
[,1] [,2] [,3]
[1,]     1     4     7
[2,]     3     6     9

> # tất cả các cột của ma trận A, ngoại trừ cột 1
> A[,-1]
[,1] [,2]
[1,]     4     7
[2,]     5     8
[3,]     6     9

> # xem phần tử nào cao hơn 3.
> A>3
[,1] [,2] [,3]
[1,] FALSE TRUE TRUE
[2,] FALSE TRUE TRUE
[3,] FALSE TRUE TRUE
```

6.2.2 Tính toán với ma trận

Cộng và trừ hai ma trận. Cho hai ma trận A và B như sau:

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

> B <- matrix(-1:-12, 3, 4)
> B
[,1] [,2] [,3] [,4]
[1,]   -1   -4   -7  -10
```

```
[2,]   -2    -5    -8   -11
[3,]   -3    -6    -9   -12
```

Chúng ta có thể cộng A+B:

```
> C <- A+B
> C
[,1] [,2] [,3] [,4]
[1,]    0    0    0    0
[2,]    0    0    0    0
[3,]    0    0    0    0
```

Hay A-B:

```
> D <- A-B
> D
[,1] [,2] [,3] [,4]
[1,]    2     8    14    20
[2,]    4    10    16    22
[3,]    6    12    18    24
```

Nhân hai ma trận. Cho hai ma trận:

$$A = \begin{pmatrix} 1 & 4 & 7 \\ 2 & 5 & 8 \\ 3 & 6 & 9 \end{pmatrix} \quad \text{và} \quad B = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix}$$

Chúng ta muốn tính AB , và có thể triển khai bằng R bằng cách sử dụng `%*%` như sau:

```
> y <- c(1,2,3,4,5,6,7,8,9)
> A <- matrix(y, nrow=3)
> B <- t(A)
> AB <- A %*% B
> AB
[,1] [,2] [,3]
[1,] 66   78   90
[2,] 78   93   108
[3,] 90   108  126
```

Hay tính BA , và có thể triển khai bằng R bằng cách sử dụng `%*%` như sau:

```
> BA <- B %*% A
> BA
[,1] [,2] [,3]
[1,] 14   32   50
[2,] 32   77  122
[3,] 50  122  194
```

Nghịch đảo ma trận và giải hệ phương trình. Ví dụ chúng ta có hệ phương trình sau đây:

$$3x_1 + 4x_2 = 4$$

$$x_1 + 6x_2 = 2$$

Hệ phương trình này có thể viết bằng kí hiệu ma trận: $AX = Y$, trong đó:

$$A = \begin{pmatrix} 3 & 4 \\ 1 & 6 \end{pmatrix}, \quad X = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \quad \text{và} \quad Y = \begin{pmatrix} 4 \\ 2 \end{pmatrix}$$

Nghiệm của hệ phương trình này là: $X = A^{-1}Y$, hay trong R:

```
> A <- matrix(c(3,1,4,6), nrow=2)
> Y <- matrix(c(4,2), nrow=2)
> X <- solve(A) %*% Y
> X
      [,1]
[1,] 1.1428571
[2,] 0.1428571
```

Chúng ta có thể kiểm tra:

```
> 3*X[1,1]+4*X[2,1]
[1] 4
```

Trị số eigen cũng có thể tính toán bằng function `eigen` như sau:

```
> eigen(A)
$values
[1] 7 2

$vectors
      [,1]      [,2]
[1,] -0.7071068 -0.9701425
[2,] -0.7071068  0.2425356
```

Định thức (determinant). Làm sao chúng ta xác định một ma trận có thể đảo nghịch hay không? Ma trận mà định thức bằng 0 là **ma trận suy biến (singular matrix)** và không thể đảo nghịch. Để kiểm tra định thức, R dùng lệnh `det()`:

```
> E <- matrix((1:9), 3, 3)
> E
      [,1] [,2] [,3]
[1,]    1    4    7
[2,]    2    5    8
[3,]    3    6    9
```

```
> det(E)
[1] 0
```

Nhưng ma trận F sau đây thì có thể đảo nghịch:

```
> F <- matrix((1:9)^2, 3, 3)
> F
[,1] [,2] [,3]
[1,]    1   16   49
[2,]    4   25   64
[3,]    9   36   81
> det(F)
[1] -216
```

Và nghịch đảo của ma trận F (F^{-1}) có thể tính bằng function `solve()` như sau:

```
> solve(F)
[,1]      [,2]      [,3]
[1,] 1.291667 -2.166667 0.9305556
[2,] -1.166667  1.666667 -0.6111111
[3,]  0.375000 -0.500000  0.1805556
```

Ngoài những phép tính đơn giản này, R còn có thể sử dụng cho các phép tính phức tạp khác. Một lợi thế đáng kể của R là phần mềm cung cấp cho người sử dụng tự do tạo ra những phép tính phù hợp cho từng vấn đề cụ thể. R có một package Matrix chuyên thiết kế cho tính toán ma trận. Bạn đọc có thể tải package xuống, cài vào máy, và sử dụng, nếu cần. Địa chỉ để tải là: http://cran.au.r-project.org/bin/windows/contrib/r-release/Matrix_0.995-8.zip cùng với tài liệu chỉ dẫn cách sử dụng (dài khoảng 80 trang): <http://cran.au.r-project.org/doc/packages/Matrix.pdf>.

7. Sử dụng R cho tính toán xác suất

7.1 Phép hoán vị (permutation)

Chúng ta biết rằng $3! = 3 \cdot 2 \cdot 1 = 6$, và $0!=1$. Nói chung, công thức tính hoán vị cho một số n là: $n! = n(n-1)(n-2)(n-3) \times \dots \times 1$. Trong R cách tính này rất đơn giản với lệnh `prod()` như sau:

- Tìm $3!$
- ```
> prod(3:1)
[1] 6
```
- Tìm  $10!$
- ```
> prod(10:1)
[1] 3628800
```

- Tìm 10.9.8.7.6.5.4

```
> prod(10:4)
[1] 604800
```
- Tìm $(10.9.8.7.6.5.4) / (40.39.38.37.36)$

```
> prod(10:4) / prod(40:36)
[1] 0.007659481
```

7.2 Tổ hợp (combination)

Số lần chọn k người từ n phần tử là: $\binom{n}{k} = \frac{n!}{k!(n-k)!}$. Công thức này cũng có khi viết là C_k^n thay vì $\binom{n}{k}$. Với R, phép tính này rất đơn giản bằng hàm `choose(n, k)`. Sau đây là vài ví dụ minh họa:

- Tìm $\binom{5}{2}$

```
> choose(5, 2)
[1] 10
```
- Tìm xác suất cặp A và B trong số 5 người được đắc cử vào hai chức vụ:

```
> 1/choose(5, 2)
[1] 0.1
```

7.3 Biến số ngẫu nhiên và hàm phân phối

Khi nói đến “phân phối” (hay distribution) là đề cập đến các giá trị mà biến số có thể có. Các *hàm phân phối* (distribution function) là hàm nhằm mô tả các biến số đó một cách có hệ thống. “Có hệ thống” ở đây có nghĩa là theo một mô hình toán học cụ thể với những thông số cho trước. Trong xác suất thống kê có khá nhiều hàm phân phối, và ở đây chúng ta sẽ xem xét qua một số hàm quan trọng nhất và thông dụng nhất: đó là phân phối nhị phân, phân phối Poisson, và phân phối chuẩn. Trong mỗi luật phân phối, có 4 loại hàm quan trọng mà chúng ta cần biết:

- hàm mật độ xác suất (probability density distribution);
- hàm phân phối tích lũy (cumulative probability distribution);
- hàm định bậc (quantile); và
- hàm mô phỏng (simulation).

R có những hàm sẵn trên có thể ứng dụng cho tính toán xác suất. Tên mỗi hàm được gọi bằng một tiếp đầu ngữ để chỉ loại hàm phân phối, và viết tắt tên của hàm đó. Các tiếp đầu ngữ là `d` (chỉ distribution hay xác suất), `p` (chỉ cumulative probability, xác suất tích lũy), `q` (chỉ định bậc hay quantile), và `r` (chỉ random hay số ngẫu nhiên). Các

tên viết tắt là `norm` (normal, phân phối chuẩn), `binom` (binomial , phân phối nhị phân), `pois` (Poisson, phân phối Poisson), v.v... Bảng sau đây tóm tắt các hàm và thông số cho từng hàm:

Hàm phân phối	Mật độ	Tích lũy	Định bậc	Mô phỏng
Chuẩn	<code>dnorm(x, mean, sd)</code>	<code>pnorm(q, mean, sd)</code>	<code>qnorm(p, mean, sd)</code>	<code>rnorm(n, mean, sd)</code>
Nhị phân	<code>dbinom(k, n, p)</code>	<code>pbinom(q, n, p)</code>	<code>qbinom(p, n, p)</code>	<code>rbinom(k, n, prob)</code>
Poisson	<code>dpois(k, lambda)</code>	<code>ppois(q, lambda)</code>	<code>qpois(p, lambda)</code>	<code>rpois(n, lambda)</code>
Uniform	<code>dunif(x, min, max)</code>	<code>runif(q, min, max)</code>	<code>qunif(p, min, max)</code>	<code>rrunif(n, min, max)</code>
Negative binomial	<code>dnbinom(x, k, p)</code>	<code>pnbinom(q, k, p)</code>	<code>qnbinom(p, k, prob)</code>	<code>rbinom(n, n, prob)</code>
Beta	<code>dbeta(x, shape1, shape2)</code>	<code>pbeta(q, shape1, shape2)</code>	<code>qbeta(p, shape1, shape2)</code>	<code>rbeta(n, shape1, shape2)</code>
Gamma	<code>dgamma(x, shape, rate, scale)</code>	<code>gamma(q, shape, rate, scale)</code>	<code>qgamma(p, shape, rate, scale)</code>	<code>rgamma(n, shape, rate, scale)</code>
Geometric	<code>dgeom(x, p)</code>	<code>pgeom(q, p)</code>	<code>qgeom(p, prob)</code>	<code>rgeom(n, prob)</code>
Exponential	<code>dexp(x, rate)</code>	<code>pexp(q, rate)</code>	<code>qexp(p, rate)</code>	<code>rexp(n, rate)</code>
Weibull	<code>dnorm(x, mean, sd)</code>	<code>pnorm(q, mean, sd)</code>	<code>qnorm(p, mean, sd)</code>	<code>rnorm(n, mean, sd)</code>
Cauchy	<code>dcauchy(x, location, scale)</code>	<code>pcauchy(q, location, scale)</code>	<code>qcauchy(p, location, scale)</code>	<code>rcauchy(n, location, scale)</code>
F	<code>df(x, df1, df2)</code>	<code>pf(q, df1, df2)</code>	<code>qf(p, df1, df2)</code>	<code>rf(n, df1, df2)</code>
T	<code>dt(x, df)</code>	<code>pt(q, df)</code>	<code>qt(p, df)</code>	<code>rt(n, df)</code>
Chi-squared	<code>dchisq(x, df)</code>	<code>pchi(q, df)</code>	<code>qchisq(p, df)</code>	<code>rchisq(n, df)</code>

Chú thích: Trong bảng trên, df = degrees of freedom (bậc tự do); $prob$ = probability (xác suất); n = sample size (số lượng mẫu). Các thông số khác có thể tham khảo thêm cho từng luật phân phối. Riêng các luật phân phối F, t, Chi-squared còn có một thông số khác nữa là non-centrality parameter (ncp) được cho số 0. Tuy nhiên người sử dụng có thể cho một thông số khác thích hợp, nếu cần.

7.3.1 Hàm phân phối nhị phân (Binomial distribution)

Như tên gọi, hàm phân phối nhị phân chỉ có hai giá trị: nam / nữ, sống / chết, có / không, v.v... Hàm nhị phân được phát biểu bằng định lí như sau: Nếu một thử nghiệm được tiến hành n lần, mỗi lần cho ra kết quả hoặc là thành công hoặc là thất bại, và gồm xác suất thành công được biết trước là p , thì xác suất có k lần thử nghiệm thành công là: $P(k|n,p) = C_k^n p^k (1-p)^{n-k}$, trong đó $k = 0, 1, 2, \dots, n$. Trong R, có hàm `dbinom(k, n, p)` có thể giúp chúng ta tính công thức $P(k|n,p) = C_k^n p^k (1-p)^{n-k}$ một cách nhanh chóng. Trong trường hợp trên, chúng ta chỉ cần đơn giản lệnh:

```
> dbinom(2, 3, 0.60)
[1] 0.432
```

Ví dụ 2: Hàm nhị phân tích lũy (Cumulative Binomial probability distribution). Xác suất thuộc chủng loãng xương có hiệu nghiệm là khoảng 70% (tức là $p = 0.70$). Nếu chúng ta điều trị 10 bệnh nhân, xác suất có tối thiểu 8 bệnh nhân với kết quả tích cực là bao nhiêu? Nói cách khác, nếu gọi X là số bệnh nhân được điều trị thành công, chúng ta cần tìm $P(X \geq 8) = ?$ Để trả lời câu hỏi này, chúng ta sử dụng hàm

`pbinom(k, n, p)`. Xin nhắc lại rằng hàm `pbinom(k, n, p)` cho chúng ta $P(X \leq k)$. Do đó, $P(X \geq 8) = 1 - P(X \leq 7)$. Thành ra, đáp số bằng R cho câu hỏi là:

```
> 1-pbinom(7, 10, 0.70)
[1] 0.3827828
```

Ví dụ 3: Mô phỏng hàm nhị phân: Biết rằng trong một quần thể dân số có khoảng 20% người mắc bệnh cao huyết áp; nếu chúng ta tiến hành chọn mẫu 1000 lần, mỗi lần chọn 20 người trong quần thể đó một cách ngẫu nhiên, sự phân phối số bệnh nhân cao huyết áp sẽ như thế nào? Để trả lời câu hỏi này, chúng ta có thể ứng dụng hàm `rbinom(n, k, p)` trong R với những thông số như sau:

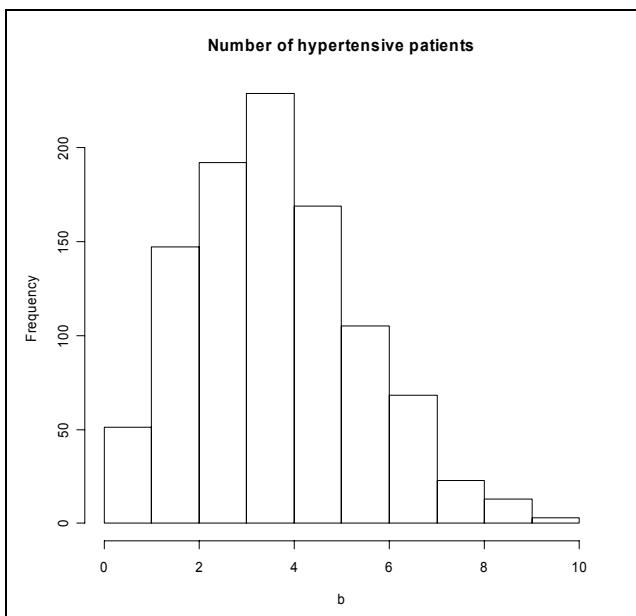
```
> b <- rbinom(1000, 20, 0.20)
```

Trong lệnh trên, kết quả mô phỏng được tạm thời chứa trong đối tượng tên là `b`. Để biết `b` có gì, chúng ta đếm bằng lệnh `table`:

```
> table(b)
b
 0   1   2   3   4   5   6   7   8   9   10
 6  45 147 192 229 169 105  68  23  13    3
```

Dòng số liệu thứ nhất (`0, 1, 2, ..., 10`) là số bệnh nhân mắc bệnh cao huyết áp trong số 20 người mà chúng ta chọn. Dòng số liệu thứ hai cho chúng ta biết số lần chọn mẫu trong 1000 lần xảy ra. Do đó, có 6 mẫu không có bệnh nhân cao huyết áp nào, 45 mẫu với chỉ 1 bệnh nhân cao huyết áp, v.v... Có lẽ cách để hiểu là vẽ đồ thị các tần số trên bằng lệnh `hist` như sau:

```
> hist(b, main="Number of hypertensive patients")
```



Biểu đồ 1. Phân phối số bệnh nhân cao huyết áp trong số 20 người được chọn ngẫu nhiên trong một quần thể gồm 20% bệnh nhân cao huyết áp, và chọn mẫu được lặp lại 1000 lần.

Qua biểu đồ trên, chúng ta thấy xác suất có 4 bệnh nhân cao huyết áp (trong mỗi lần chọn mẫu 20 người) là cao nhất (22.9%). Điều này cũng có thể hiểu được, bởi vì tỉ lệ cao huyết áp là 20%, cho nên chúng ta kì vọng rằng trung bình 4 người trong số 20 người được chọn phải là cao huyết áp. Tuy nhiên, điều quan trọng mà biểu đồ trên thể hiện là có khi chúng ta quan sát đến 10 bệnh nhân cao huyết áp dù xác suất cho mẫu này rất thấp (chỉ 3/1000).

7.3.2 Hàm phân phối Poisson (Poisson distribution)

Hàm phân phối Poisson, nói chung, rất giống với hàm nhị phân, ngoại trừ thông số p thường rất nhỏ và n thường rất lớn. Vì thế, hàm Poisson thường được sử dụng để mô tả các biến số rủi ro xảy ra (như số người mắc ung thư trong một dân số chẵng hạn). Hàm Poisson còn được ứng dụng khá nhiều và thành công trong các nghiên cứu kỹ thuật và thị trường như số lượng khách hàng đến một nhà hàng mỗi giờ.

Ví dụ 4: Hàm mật độ Poisson (Poisson density probability function). Qua theo dõi nhiều tháng, người ta biết được tỉ lệ đánh sai chính tả của một thư ký đánh máy. Tính trung bình cứ khoảng 2.000 chữ thì thư ký đánh sai 1 chữ. Hỏi xác suất mà thư ký đánh sai chính tả 2 chữ, hơn 2 chữ là bao nhiêu?

Vì tần số khá thấp, chúng ta có thể giả định rằng biến số “sai chính tả” (tạm đặt tên là biến số X) là một hàm ngẫu nhiên theo luật phân phối Poisson. Ở đây, chúng ta có tỉ lệ sai chính tả trung bình là 1 ($\lambda = 1$). Luật phân phối Poisson phát biểu rằng xác suất mà $X = k$, với điều kiện tỉ lệ trung bình λ :

$$P(X = k | \lambda) = \frac{e^{-\lambda} \lambda^k}{k!}$$

Do đó, đáp số cho câu hỏi trên là: $P(X = 2 | \lambda = 1) = \frac{e^{-2} 1^2}{2!} = 0.1839$. Đáp số này có thể tính bằng R một cách nhanh chóng hơn bằng hàm `dpois` như sau:

```
> dpois(2, 1)
[1] 0.1839397
```

Chúng ta cũng có thể tính xác suất sai 1 chữ, và xác suất không sai chữ nào:

```
> dpois(1, 1)
[1] 0.3678794
```

```
> dpois(0, 1)
```

```
[1] 0.3678794
```

Chú ý trong hàm trên, chúng ta chỉ đơn giản cung cấp thông số $k = 2$ và $(\lambda = 1)$. Trên đây là xác suất mà thư kí đánh sai chính tả đúng 2 chữ. Nhưng xác suất mà thư kí đánh sai chính tả hơn 2 chữ (tức 3, 4, 5, ... chữ) có thể ước tính bằng:

$$\begin{aligned} P(X > 2) &= P(X = 3) + P(X = 4) + P(X = 5) + \dots \\ &= 1 - P(X \leq 2) \\ &= 1 - 0.3678 - 0.3678 - 0.1839 \\ &= 0.08 \end{aligned}$$

Bằng R, chúng ta có thể tính như sau:

```
# P(X ≤ 2)
> ppois(2, 1)
[1] 0.9196986
```

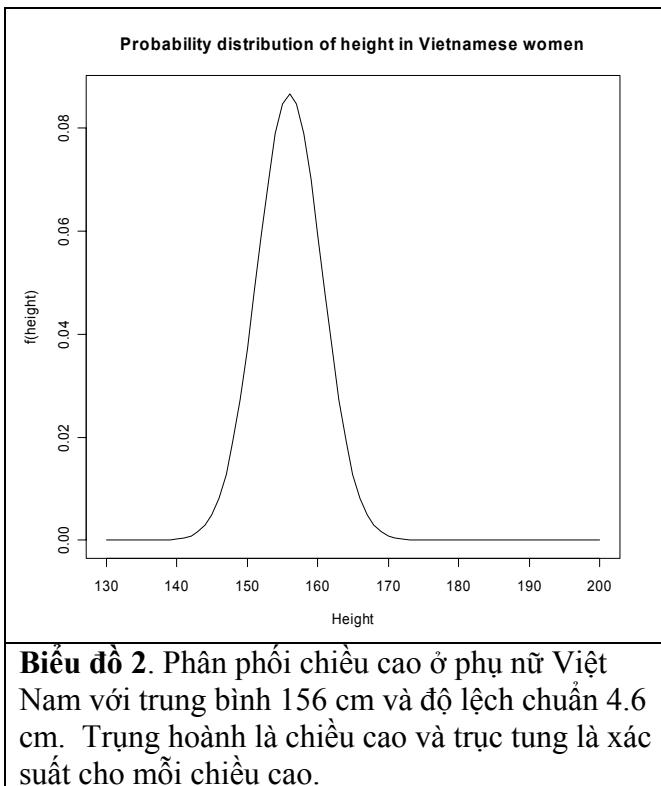
```
# 1-P(X ≤ 2)
> 1-ppois(2, 1)
[1] 0.0803014
```

7.3.3 Hàm phân phối chuẩn (Normal distribution)

Hai luật phân phối mà chúng ta vừa xem xét trên đây thuộc vào nhóm phân phối áp dụng cho các biến số phi liên tục (discrete distributions), mà trong đó biến số có những giá trị theo bậc thứ hay thể loại. Đối với các biến số liên tục, có vài luật phân phối thích hợp khác, mà quan trọng nhất là phân phối chuẩn. Phân phối chuẩn là nền tảng quan trọng nhất của phân tích thống kê. Có thể nói không ngoa rằng hầu hết lí thuyết thống kê được xây dựng trên nền tảng của phân phối chuẩn. Hàm mật độ phân phối chuẩn có hai thông số: trung bình μ và phương sai σ^2 (hay độ lệch chuẩn σ). Gọi X là một biến số (như chiều cao chẳng hạn), hàm mật độ phân phối chuẩn phát biểu rằng xác suất mà $X = x$ là:

$$P(X = x | \mu, \sigma^2) = f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$

Ví dụ 5: Hàm mật độ phân phối chuẩn (Normal density probability function). Chiều cao trung bình hiện nay ở phụ nữ Việt Nam là 156 cm, với độ lệch chuẩn là 4.6 cm. Cũng biết rằng chiều cao này tuân theo luật phân phối chuẩn. Với hai thông số $\mu=156$, $\sigma=4.6$, chúng ta có thể xây dựng một hàm phân phối chiều cao cho toàn bộ quần thể phụ nữ Việt Nam, và hàm này có hình dạng như sau:



Biểu đồ trên được vẽ bằng hai lệnh sau đây. Lệnh đầu tiên nhằm tạo ra một biến số height có giá trị 130, 131, 132, ..., 200 cm. Lệnh thứ hai là vẽ biểu đồ với điều kiện trung bình là 156 cm và độ lệch chuẩn là 4.6 cm.

```
> height <- seq(130, 200, 1)
> plot(height, dnorm(height, 156, 4.6),
  type="l",
  ylab="f(height)",
  xlab="Height",
  main="Probability distribution of height in Vietnamese women")
```

Với hai thông số trên (và biểu đồ), chúng ta có thể ước tính xác suất cho bất cứ chiều cao nào. Chẳng hạn như xác suất một phụ nữ Việt Nam có chiều cao 160 cm là:

$$P(X = 160 | \mu=156, \sigma=4.6) = \frac{1}{4.6\sqrt{2 \times 3.1416}} \exp\left[-\frac{(160-156)^2}{2(4.6)^2}\right] \\ = 0.0594$$

Hàm `dnorm(x, mean, sd)` trong R có thể tính toán xác suất này cho chúng ta một cách gọn nhẹ:

```
> dnorm(160, mean=156, sd=4.6)
[1] 0.05942343
```

Hàm xác suất chuẩn tích lũy (cumulative normal probability function). Vì chiều cao là một biến số liên tục, trong thực tế chúng ta ít khi nào muốn tìm xác suất cho một giá trị cụ thể x , mà thường tìm xác suất cho một khoảng giá trị a đến b . Chẳng hạn như chúng ta muốn biết xác suất chiều cao từ 150 đến 160 cm (tức là $P(160 \leq X \leq 150)$, hay xác suất chiều cao thấp hơn 145 cm, tức $P(X < 145)$). Để tìm đáp số các câu hỏi như thế, chúng ta cần đến hàm xác suất chuẩn tích lũy, được định nghĩa như sau:

$$P(a \leq X \leq b) = \int_a^b f(x) dx$$

Thành ra, $P(160 \leq X \leq 150)$ chính là diện tích tính từ trục hoành = 150 đến 160 của **biểu đồ 2**. Trong R có hàm pnorm(x, mean, sd) dùng để tính xác suất tích lũy cho một phân phối chuẩn rất có ích.

$$\text{pnorm}(a, \text{mean}, \text{sd}) = \int_{-\infty}^a f(x) dx = P(X \leq a | \text{mean}, \text{sd})$$

Chẳng hạn như xác suất chiều cao phụ nữ Việt Nam bằng hoặc thấp hơn 150 cm là 9.6%:

```
> pnorm(150, 156, 4.6)
[1] 0.0960575
```

Hay xác suất chiều cao phụ nữ Việt Nam bằng hoặc cao hơn 165 cm là:

```
> 1-pnorm(164, 156, 4.6)
[1] 0.04100591
```

Nói cách khác, chỉ có khoảng 4.1% phụ nữ Việt Nam có chiều cao bằng hay cao hơn 165 cm.

Ví dụ 6: Ứng dụng luật phân phối chuẩn: Trong một quần thể, chúng ta biết rằng áp suất máu trung bình là 100 mmHg và độ lệch chuẩn là 13 mmHg, hỏi: có bao nhiêu người trong quần thể này có áp suất máu bằng hoặc cao hơn 120 mmHg? Câu trả lời bằng R là:

```
> 1-pnorm(120, mean=100, sd=13)
[1] 0.0619679
```

Tức khoảng 6.2% người trong quần thể này có áp suất máu bằng hoặc cao hơn 120 mmHg.

7.3.4 Hàm phân phối chuẩn chuẩn hóa (Standardized Normal distribution)

Một biến X tuân theo luật phân phối chuẩn với trung bình μ và phương sai σ^2 thường được viết tắt là:

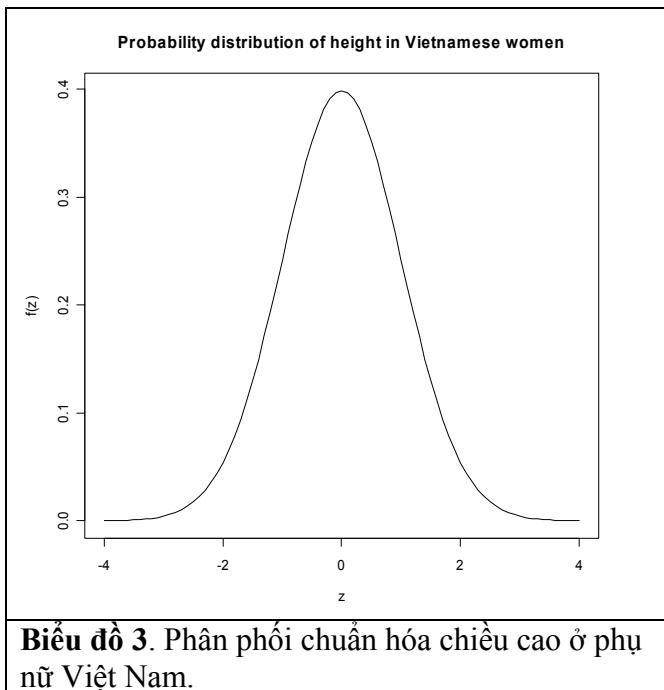
$$X \sim N(\mu, \sigma^2)$$

Ở đây μ và σ^2 tùy thuộc vào đơn vị đo lường của biến số. Chẳng hạn như chiều cao được tính bằng cm (hay m), huyết áp được đo bằng mmHg, tuổi được đo bằng năm, v.v... cho nên đôi khi mô tả một biến số bằng đơn vị gốc rất khó so sánh. Một cách đơn giản hơn là chuẩn hóa (standardized) X sao cho số trung bình là 0 và phương sai là 1. Sau vài thao tác số học, có thể chứng minh dễ dàng rằng, cách biến đổi X để đáp ứng điều kiện trên là:

$$Z = \frac{X - \mu}{\sigma}$$

Nói theo ngôn ngữ toán: nếu $X \sim N(\mu, \sigma^2)$, thì $(X - \mu)/\sigma \sim N(0, 1)$. Như vậy qua công thức trên, Z thực chất là độ khác biệt giữa một số và trung bình tính bằng số độ lệch chuẩn. Nếu $Z = 0$, chúng ta biết rằng X bằng số trung bình μ . Nếu $Z = -1$, chúng ta biết rằng X thấp hơn μ đúng 1 độ lệch chuẩn. Tương tự, $Z = 2.5$, chúng ta biết rằng X cao hơn μ đúng 2.5 độ lệch chuẩn. v.v...

Biểu đồ phân phối chiều cao của phụ nữ Việt Nam có thể mô tả bằng một đơn vị mới, đó là chỉ số z như sau:



Biểu đồ trên được vẽ bằng hai lệnh sau đây:

```
> height <- seq(-4, 4, 0.1)
> plot(height, dnorm(height, 0, 1),
       type="l",
       ylab="f(z)",
       xlab="z",
       main="Probability distribution of height in Vietnamese women")
```

Với phân phối chuẩn hóa, chúng ta có một tiện lợi là có thể dùng nó để mô tả và so sánh mật độ phân phối của bất cứ biến nào, vì tất cả đều được chuyển sang chỉ số z.

Trong biểu đồ trên, trục tung là xác suất z và trục hoành là biến số z. Chúng ta có thể tính toán xác suất z nhỏ hơn một hằng số (constant) nào đó dễ dàng bằng R. Ví dụ, chúng ta muốn tìm $P(z \leq -1.96) = ?$ cho một phân phối mà trung bình là 0 và độ lệch chuẩn là 1.

```
> pnorm(-1.96, mean=0, sd=1)
[1] 0.02499790
```

Hay $P(z \leq 1.96) = ?$

```
> pnorm(1.96, mean=0, sd=1)
[1] 0.9750021
```

Do đó, $P(-1.96 < z < 1.96)$ chính là:

```
> pnorm(1.96) - pnorm(-1.96)
[1] 0.9500042
```

Nói cách khác, xác suất 95% là z nằm giữa -1.96 và 1.96. (Chú ý trong lệnh trên tôi không cung cấp `mean=0, sd=1`, bởi vì trong thực tế, `pnorm` giá trị mặc định (default value) của thông số `mean` là 0 và `sd` là 1).

Ví dụ 5 (tiếp tục). Xin nhắc lại để tiện việc theo dõi, chiều cao trung bình ở phụ nữ Việt Nam là 156 cm và độ lệch chuẩn là 4.6 cm. Do đó, một phụ nữ có chiều cao 170 cm cũng có nghĩa là $z = (170 - 156) / 4.6 = 3.04$ độ lệch chuẩn, và tỉ lệ các phụ nữ Việt Nam có chiều cao cao hơn 170 cm là rất thấp, chỉ khoảng 0.1%.

```
> 1-pnorm(3.04)
[1] 0.001182891
```

Tìm định lượng (quantile) của một phân phối chuẩn. Đôi khi chúng ta cần làm một tính toán đảo ngược. Chẳng hạn như chúng ta muốn biết: nếu xác suất Z nhỏ hơn một hằng số z nào đó cho trước bằng p , thì z là bao nhiêu? Diễn tả theo kí hiệu xác suất, chúng ta muốn tìm z trong nếu:

$$P(Z < z) = p$$

Để trả lời câu hỏi này, chúng ta sử dụng hàm `qnorm(p, mean=, sd=)`.

Ví dụ 7: Biết rằng $Z \sim N(0, 1)$ và nếu $P(Z < z) = 0.95$, chúng ta muốn tìm z .

```
> qnorm(0.95, mean=0, sd=1)
[1] 1.644854
```

Hay $P(Z < z) = 0.975$ cho phân phối chuẩn với trung bình 0 và độ lệch chuẩn 1:

```
> qnorm(0.975, mean=0, sd=1)
[1] 1.959964
```

7.4 Chọn mẫu ngẫu nhiên (random sampling)

Trong xác suất và thống kê, lấy mẫu ngẫu nhiên rất quan trọng, vì nó đảm bảo tính hợp lí của các phương pháp phân tích và suy luận thống kê. Với R, chúng ta có thể lấy mẫu một mẫu ngẫu nhiên bằng cách sử dụng hàm `sample`.

Ví dụ 8: Chúng ta có một quân thê gồm 40 người (mã số 1, 2, 3, ..., 40). Nếu chúng ta muốn chọn 5 đối tượng quân thê đó, ai sẽ là người được chọn? Chúng ta có thể dùng lệnh `sample()` để trả lời câu hỏi đó như sau:

```
> sample(1:40, 5)
[1] 32 26 6 18 9
```

Kết quả trên cho biết đối tượng 32, 26, 8, 18 và 9 được chọn. Mỗi lần ra lệnh này, R sẽ chọn một mẫu khác, chứ không hoàn toàn giống như mẫu trên. Ví dụ:

```
> sample(1:40, 5)
[1] 5 22 35 19 4
```

```
> sample(1:40, 5)
[1] 24 26 12 6 22
```

```
> sample(1:40, 5)
[1] 22 38 11 6 18
```

v.v...

Trên đây là lệnh để chúng ta chọn mẫu ngẫu nhiên mà không thay thế (random sampling without replacement), tức là mỗi lần chọn mẫu, chúng ta không bỏ lại các mẫu đã chọn vào quân thê.

Nhưng nếu chúng ta muốn chọn mẫu thay thế (tức mỗi lần chọn ra một số đối tượng, chúng ta bỏ vào lại trong quân thê để chọn tiếp lần sau). Ví dụ, chúng ta muốn chọn 10 người từ một quân thê 50 người, bằng cách lấy mẫu với thay thế (random sampling with replacement), chúng ta chỉ cần thêm tham số `replace = TRUE`:

```
> sample(1:50, 10, replace=T)
```

```
[1] 31 44 6 8 47 50 10 16 29 23
```

Hay ném một đồng xu 10 lần; mỗi lần, dĩ nhiên đồng xu có 2 kết quả H và T; và kết quả 10 lần có thể là:

```
> sample(c("H", "T"), 10, replace=T)
[1] "H" "T" "H" "H" "H" "T" "H" "H" "T" "T"
```

Cũng có thể tưởng tượng chúng ta có 5 quả banh màu xanh (X) và 5 quả banh màu đỏ (D) trong một bao. Nếu chúng ta chọn 1 quả banh, ghi nhận màu, rồi để lại vào bao; rồi lại chọn 1 quả banh khác, ghi nhận màu, và bỏ vào bao lại. Cứ như thế, chúng ta chọn 20 lần, kết quả có thể là:

```
> sample(c("X", "D"), 20, replace=T)
[1] "X" "D" "D" "D" "D" "D" "X" "X" "X" "X" "X" "X" "D" "X" "X" "D" "X" "X" "X" "X"
[20] "D"
```

Ngoài ra, chúng ta còn có thể lấy mẫu với một xác suất cho trước. Trong hàm sau đây, chúng ta chọn 10 đối tượng từ dãy số 1 đến 5, nhưng xác suất không bằng nhau:

```
> sample(5, 10, prob=c(0.3, 0.4, 0.1, 0.1, 0.1), replace=T)
[1] 3 1 3 2 2 2 2 5 1
```

Đối tượng 1 được chọn 2 lần, đối tượng 2 được chọn 5 lần, đối tượng 3 được chọn 2 lần, v.v... Tuy không hoàn toàn phù hợp với xác suất 0.3, 0.4, 0.1 như cung cấp vì số mẫu còn nhỏ, nhưng cũng không quá xa với kì vọng.

8. Biểu đồ

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* hay *trellis* có thể tải từ website của R. Trong chương này tôi sẽ chỉ cách vẽ các biểu đồ thông dụng bằng cách sử dụng các hàm phổ biến trong R.

8.1 Số liệu cho phân tích biểu đồ

Sau khi đã biết qua môi trường và những lựa chọn để thiết kế một biểu đồ, bây giờ chúng ta có thể sử dụng một số hàm thông dụng để vẽ các biểu đồ cho số liệu. Theo tôi, biểu đồ có thể chia thành 2 loại chính: biểu đồ dùng để mô tả một biến số và biểu đồ về mối liên hệ giữa hai hay nhiều biến số. Tất nhiên, biến số có thể là liên tục hay không liên tục, cho nên, trong thực tế, chúng ta có 4 loại biểu đồ. Trong phần sau đây, tôi sẽ điểm qua các loại biểu đồ, từ đơn giản đến phức tạp.

Có lẽ cách tốt nhất để tìm hiểu cách vẽ đồ thị bằng R là bằng một dữ liệu thực tế. Tôi sẽ quay lại **ví dụ 2** (phần 4.2). Trong ví dụ đó, chúng ta có dữ liệu gồm 8 cột (hay

biến số): `id`, `sex`, `age`, `bmi`, `hdl`, `ldl`, `tc`, và `tg`. (Chú ý, `id` là mã số của 50 đối tượng nghiên cứu; `sex` là giới tính (nam hay nữ); `age` là độ tuổi; `bmi` là tỉ số trọng lượng; `hdl` là high density cholesterol; `ldl` là low density cholesterol; `tc` là tổng số - total - cholesterol; và `tg` triglycerides). Dữ liệu được chứa trong directory directory `c:\works\insulin` dưới tên `chol.txt`. Trước khi vẽ đồ thị, chúng ta bắt đầu bằng cách nhập dữ liệu này vào R.

```
> setwd("c:/works/stats")
> cong <- read.table("chol.txt", header=TRUE, na.strings=".")
> attach(cong)
```

Hay để tiện việc theo dõi tôi sẽ nhập các dữ liệu đó bằng các lệnh sau đây:

```
sex <- c("Nam", "Nu", "Nu", "Nam", "Nam", "Nu", "Nam", "Nam", "Nu",
       "Nu", "Nam", "Nu", "Nam", "Nu", "Nu", "Nu", "Nu", "Nu",
       "Nu", "Nu", "Nu", "Nam", "Nu", "Nam", "Nu", "Nu", "Nu",
       "Nu", "Nam", "Nu", "Nu", "Nu", "Nu", "Nu", "Nu", "Nu",
       "Nu", "Nu", "Nu", "Nu", "Nu", "Nu", "Nu", "Nu", "Nu")

age <- c(57, 64, 60, 65, 47, 65, 76, 61, 59, 57,
       63, 51, 60, 42, 64, 49, 44, 45, 80, 48,
       61, 45, 70, 51, 63, 54, 57, 70, 47, 60,
       60, 50, 60, 55, 74, 48, 46, 49, 69, 72,
       51, 58, 60, 45, 63, 52, 64, 45, 64, 62)

bmi <- c( 17, 18, 18, 18, 18, 19, 19, 19, 19, 20, 20, 20, 20, 20,
         20, 21, 21, 21, 21, 21, 21, 22, 22, 22, 22, 22, 22,
         22, 22, 22, 23, 23, 23, 23, 23, 23, 23, 23, 24, 24, 24,
         24, 24, 24, 25, 25)

hdl <- c(5.000,4.380,3.360,5.920,6.250,4.150,0.737,7.170,6.942,5.000,
        4.217,4.823,3.750,1.904,6.900,0.633,5.530,6.625,5.960,3.800,
        5.375,3.360,5.000,2.608,4.130,5.000,6.235,3.600,5.625,5.360,
        6.580,7.545,6.440,6.170,5.270,3.220,5.400,6.300,9.110,7.750,
        6.200,7.050,6.300,5.450,5.000,3.360,7.170,7.880,7.360,7.750)

ldl <- c(2.0, 3.0, 4.0, 2.1, 3.0, 3.0, 3.0, 3.0, 3.0, 2.0,
        5.0, 1.3, 1.2, 0.7, 4.0, 4.1, 4.3, 4.0, 4.3, 4.0,
        3.1, 3.0, 1.7, 2.0, 2.1, 4.0, 4.1, 4.0, 4.2, 4.2,
        4.4, 4.3, 2.3, 6.0, 3.0, 3.0, 2.6, 4.4, 4.3, 4.0,
        3.0, 4.1, 4.4, 2.8, 3.0, 2.0, 1.0, 4.0, 4.6, 4.0)

tc <- c(4.0, 3.5, 4.7, 7.7, 5.0, 4.2, 5.9, 6.1, 5.9, 4.0,
       6.2, 4.1, 3.0, 4.0, 6.9, 5.7, 5.7, 5.3, 7.1, 3.8,
       4.3, 4.8, 4.0, 3.0, 3.1, 5.3, 5.3, 5.4, 4.5, 5.9,
       5.6, 8.3, 5.8, 7.6, 5.8, 3.1, 5.4, 6.3, 8.2, 6.2,
       6.2, 6.7, 6.3, 6.0, 4.0, 3.7, 6.1, 6.7, 8.1, 6.2)

tg <- c(1.1, 2.1, 0.8, 1.1, 2.1, 1.5, 2.6, 1.5, 5.4, 1.9,
       1.7, 1.0, 1.6, 1.1, 1.5, 1.0, 2.7, 3.9, 3.0, 3.1,
       2.2, 2.7, 1.1, 0.7, 1.0, 1.7, 2.9, 2.5, 6.2, 1.3,
       3.3, 3.0, 1.0, 1.4, 2.5, 0.7, 2.4, 2.4, 1.4, 2.7,
       2.4, 3.3, 2.0, 2.6, 1.8, 1.2, 1.9, 3.3, 4.0, 2.5)

cong <- data.frame(sex, age, bmi, hdl, ldl, tc, tg)
```

8.2 Biểu đồ cho một biến số rời rạc (discrete variable): barplot

Biến `sex` trong dữ liệu trên có hai giá trị (nam và nữ), tức là một biến không liên tục. Chúng ta muốn biết tần số của giới tính (bao nhiêu nam và bao nhiêu nữ) và vẽ một biểu đồ đơn giản. Để thực hiện ý định này, trước hết, chúng ta cần dùng hàm `table` để biết tần số:

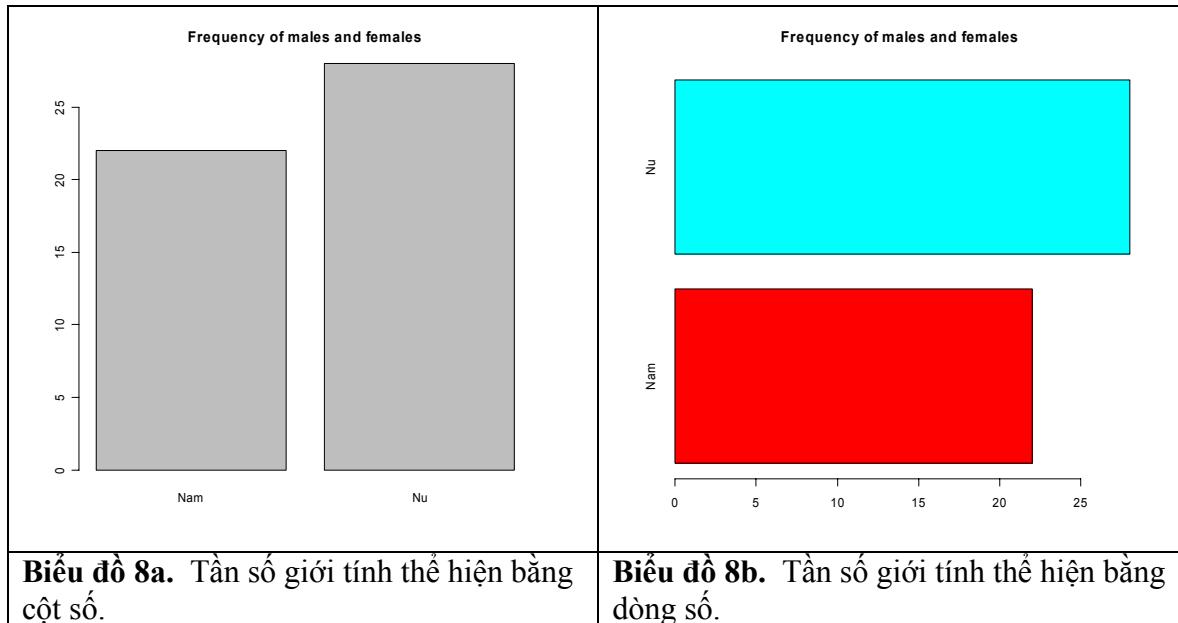
```
> sex.freq <- table(sex)
> sex.freq
sex
Nam   Nu
22    28
```

Có 22 nam và 28 nữa trong nghiên cứu. Sau đó dùng hàm `barplot` để thể hiện tần số này như sau:

```
> barplot(sex.freq, main="Frequency of males and females")
```

Biểu trên cũng có thể có được bằng một lệnh đơn giản hơn (**Biểu đồ 8a**):

```
> barplot(table(sex), main="Frequency of males and females")
```



Thay vì thể hiện tần số nam và nữ bằng 2 cột, chúng ta có thể thể hiện bằng hai dòng bằng thông số `horiz = TRUE`, như sau (xem kết quả trong **Biểu đồ 6b**):

```
> barplot(sex.freq,
          horiz = TRUE,
          col = rainbow(length(sex.freq)),
          main="Frequency of males and females")
```

8.3 Biểu đồ cho hai biến số rời rạc (discrete variable): barplot

Age là một biến số liên tục. Chúng ta có thể chia bệnh nhân thành nhiều nhóm dựa vào độ tuổi. Hàm `cut` có chức năng “cắt” một biến liên tục thành nhiều nhóm rời rạc. Chẳng hạn như:

```
> ageg <- cut(age, 3)
> table(ageg)
ageg
(42,54.7] (54.7,67.3] (67.3,80]
    19          24          7
```

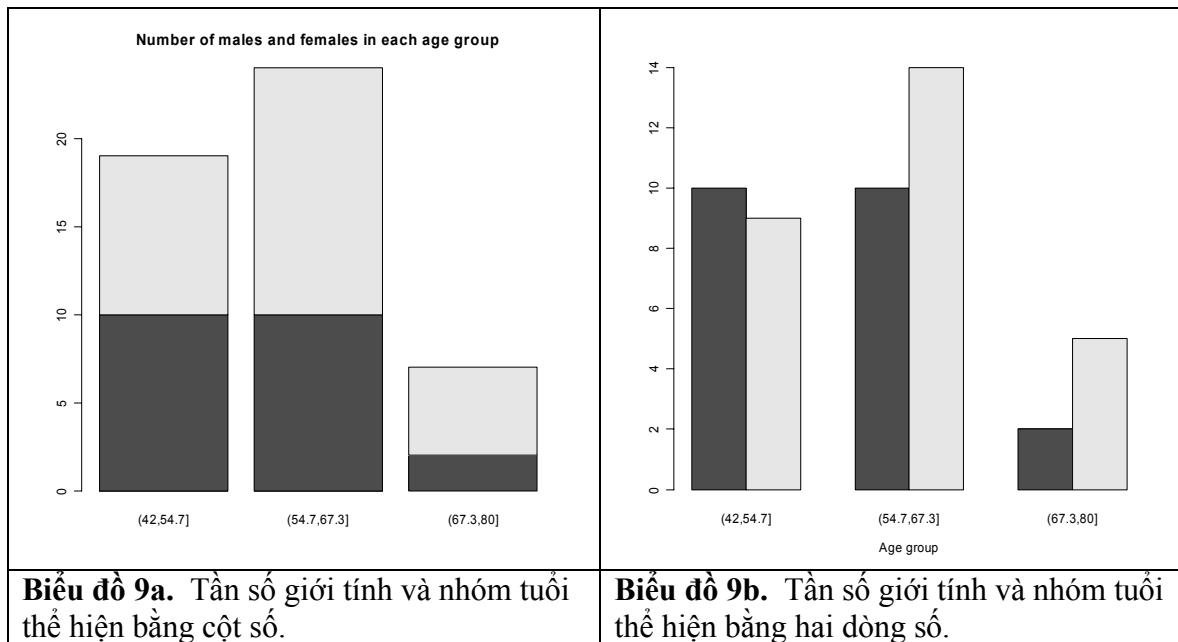
Có hiệu quả chia biến `age` thành 3 nhóm. Tần số của ba nhóm này là: 42 tuổi đến 54.7 tuổi thành nhóm 1, 54.7 đến 67.3 thành nhóm 2, và 67.3 đến 80 tuổi thành nhóm 3. Nhóm 1 có 19 bệnh nhân, nhóm 2 và 3 có 24 và 7 bệnh nhân.

Bây giờ chúng ta muốn biết có bao nhiêu bệnh nhân trong từng độ tuổi và từng giới tính bằng lệnh `table`:

```
> age.sex <- table(sex, ageg)
> age.sex
ageg
sex  (42,54.7] (54.7,67.3] (67.3,80]
  Nam      10      10      2
  Nu       9      14      5
```

Kết quả trên cho thấy chúng ta có 10 bệnh nhân nam và 9 nữ trong nhóm tuổi thứ nhất, 10 nam và 14 nữ trong nhóm tuổi thứ hai, v.v... Để thể hiện tần số của hai biến này, chúng ta vẫn dùng `barplot`:

```
> barplot(age.sex, main="Number of males and females in each age group")
```



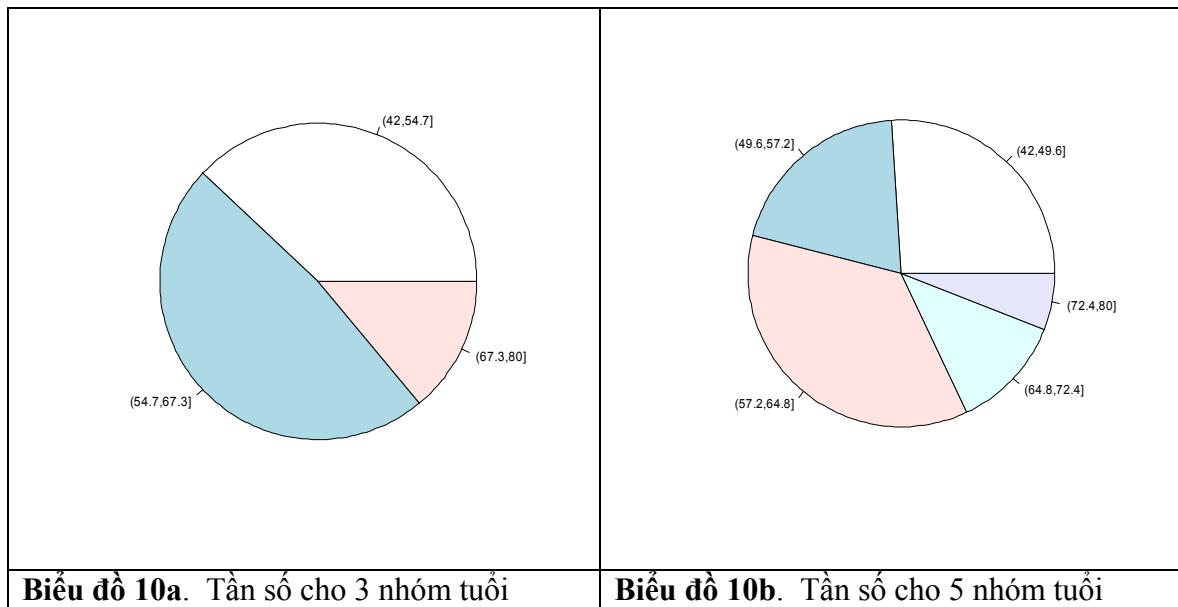
Trong **Biểu đồ 9a**, mỗi cột là cho một độ tuổi, và phần đậm của cột là nữ, và phần màu nhạt là tần số của nam giới. Thay vì thể hiện tần số nam nữ trong một cột, chúng ta cũng có thể thể hiện bằng 2 cột với `beside=T` như sau (**Biểu đồ 9b**):

```
barplot(age.sex, beside=TRUE, xlab="Age group")
```

8.4 Biểu đồ hình tròn

Tần số một biến rời rạc cũng có thể thể hiện bằng biểu đồ hình tròn. Ví dụ sau đây vẽ biểu đồ tần số của độ tuổi. **Biểu đồ 10a** là 3 nhóm độ tuổi, và **Biểu đồ 10b** là biểu đồ tần số cho 5 nhóm tuổi:

```
> pie(table(ageg))
pie(table(cut(age, 5)))
```

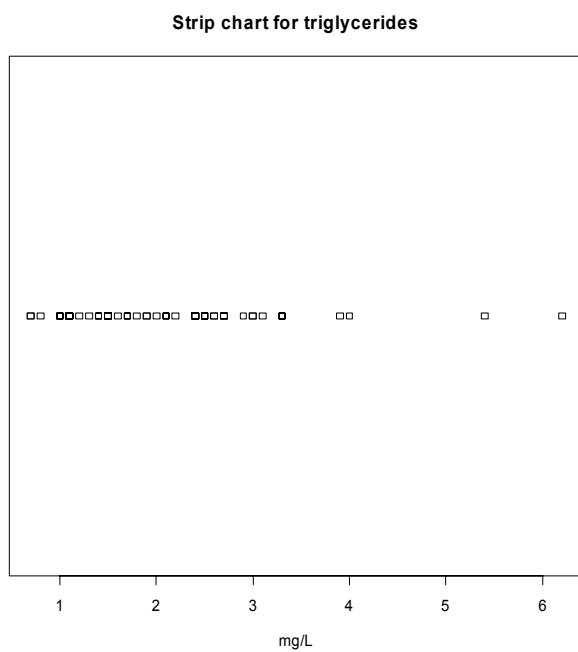


8.5 Biểu đồ cho một biến số liên tục: `stripchart` và `hist`

8.5.1 Stripchart

Biểu đồ strip cho chúng ta thấy tính liên tục của một biến số. Chẳng hạn như chúng ta muốn tìm hiểu tính liên tục của triglyceride (tg), hàm `stripchart()` sẽ giúp trong mục tiêu này:

```
> stripchart(tg,
             main="Strip chart for triglycerides", xlab="mg/L")
```

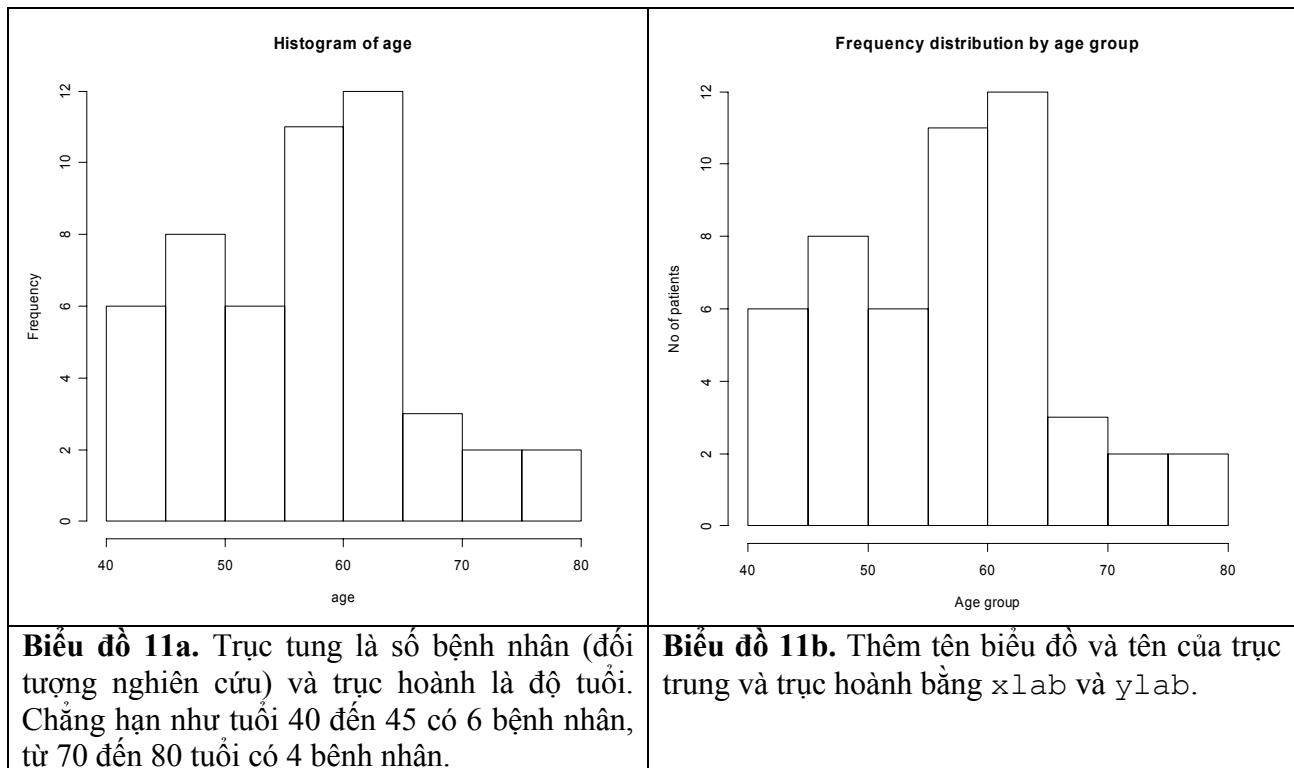


Chúng ta thấy biến số tg có sự bất liên tục, nhất là các đối tượng có tg cao. Trong khi phần lớn đối tượng có độ tg thấp hơn 5, thì có 2 đối tượng với tg rất cao (>5).

8.5.2 Histogram

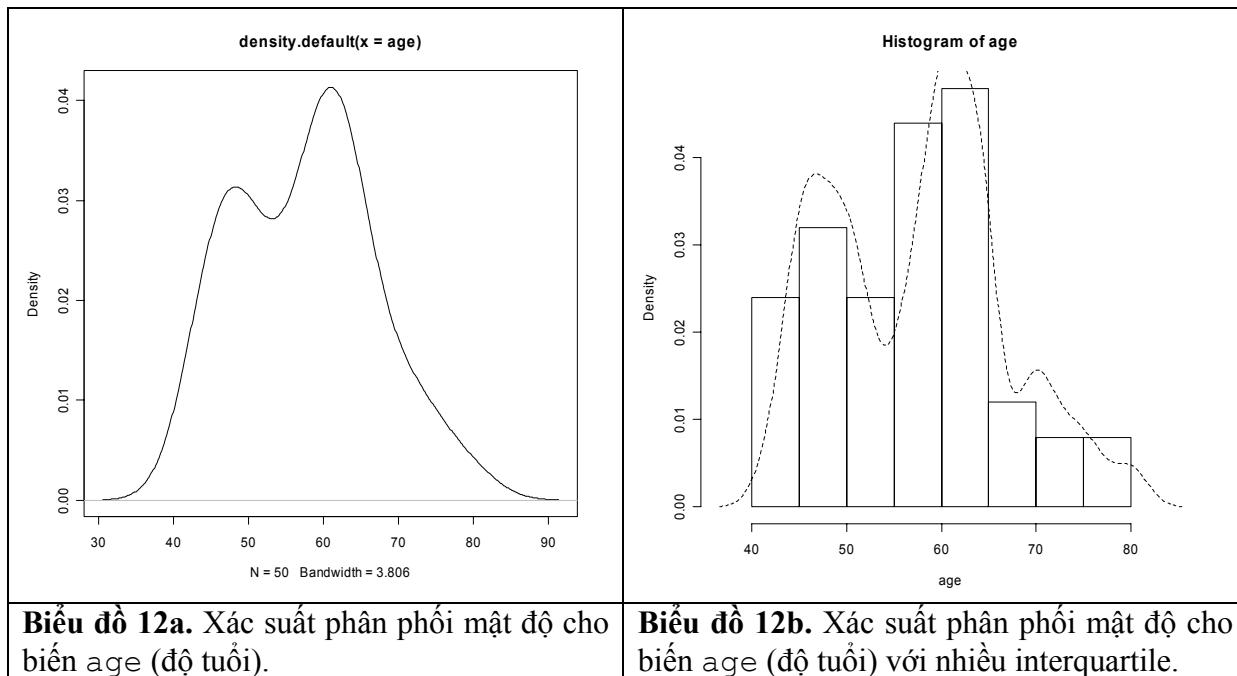
Age là một biến số liên tục. Để vẽ biểu đồ tần số của biến số age, chúng ta chỉ đơn giản lệnh `hist(age)`. Như đã đề cập trên, chúng ta có thể cải tiến đồ thị này bằng cách cho thêm tựa đề chính (main) và tựa đề của trục hoành (xlab) và trục tung (ylab):

```
> hist(age)
> hist(age, main="Frequency distribution by age group", xlab="Age
group", ylab="No of patients")
```



Chúng ta cũng có thể biến đổi biểu đồ thành một đồ thị phân phối xác suất bằng hàm `plot(density)` như sau (kết quả trong **Biểu đồ 12a**):

```
> plot(density(age), add=TRUE)
```

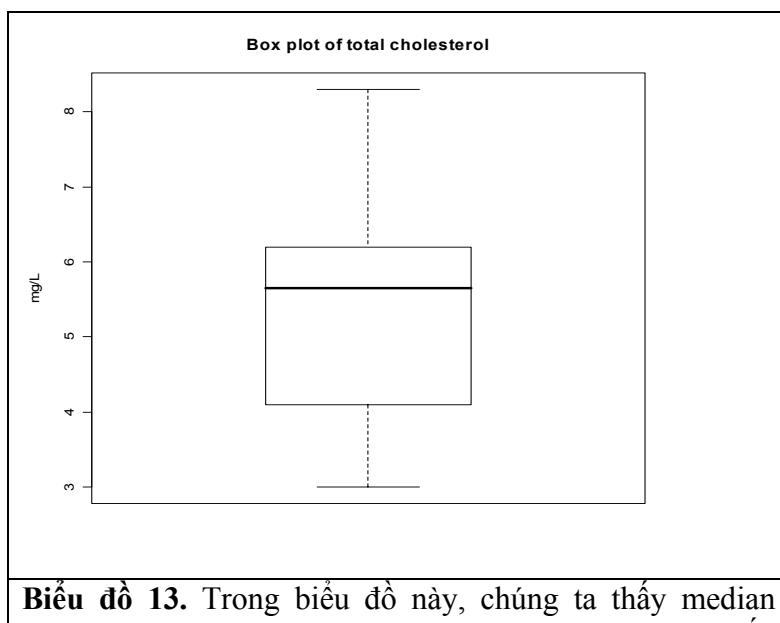


Chúng ta có thể vẽ hai đồ thị chồng lên bằng cách dùng hàm `interquartile` như sau (kết quả xem **Biểu đồ 12b**):

8.6 Biểu đồ hộp (boxplot)

Để vẽ biểu đồ hộp của biến số `tc`, chúng ta chỉ đơn giản lệnh:

```
> boxplot(tc, main="Box plot of total cholesterol", ylab="mg/L")
```



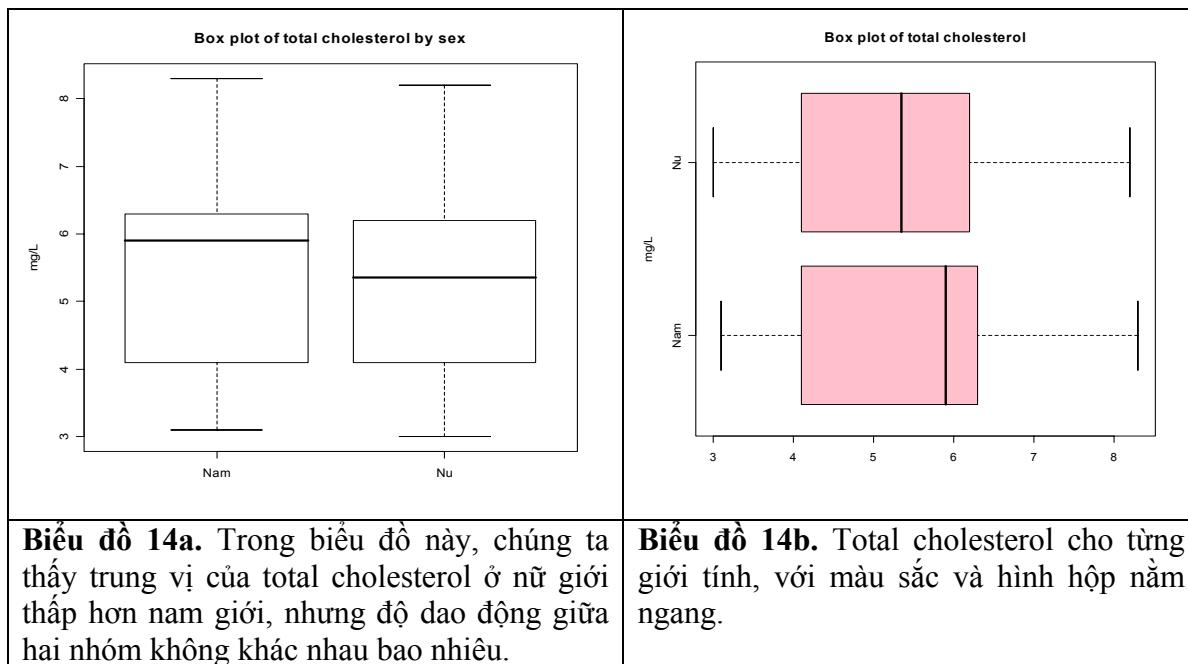
là khoang 3, và cao nhất là trên 8 mg/L.

Trong biểu đồ sau đây, chúng ta so sánh tc giữa hai nhóm nam và nữ:

```
> boxplot(tc ~ sex, main="Box plot of total cholesterol by sex",
  ylab="mg/L")
```

Kết quả trình bày trong **Biểu đồ 14a**. Chúng ta có thể biến đổi giao diện của đồ thị bằng cách dùng thông số horizontal=TRUE và thay đổi màu bằng thông số col như sau (**Biểu đồ 14b**):

```
> boxplot(tc~sex, horizontal=TRUE, main="Box plot of total
cholesterol", ylab="mg/L", col = "pink")
```

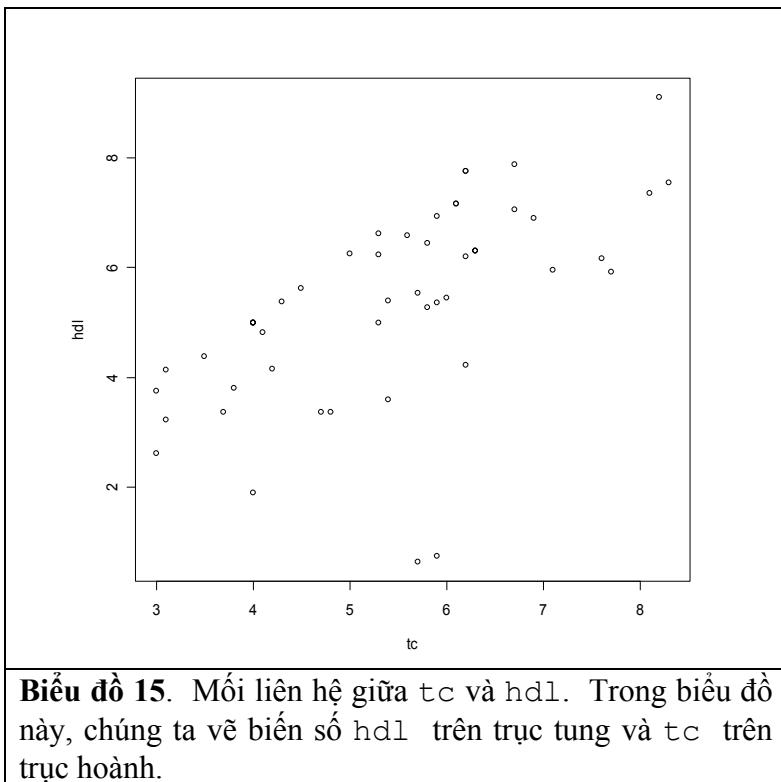


8.7 Phân tích biểu đồ cho hai biến liên tục

8.7.1 Biểu đồ tán xạ (scatter plot)

Để tìm hiểu mối liên hệ giữa hai biến, chúng ta dùng biểu đồ tán xạ. Để vẽ biểu đồ tán xạ về mối liên hệ giữa biến số tc và hdl , chúng ta sử dụng hàm `plot`. Thông số thứ nhất của hàm `plot` là trục hoành (x-axis) và thông số thứ 2 là trục tung. Để tìm hiểu mối liên hệ giữa tc và hdl chúng ta đơn giản lệnh:

```
> plot(tc, hdl)
```

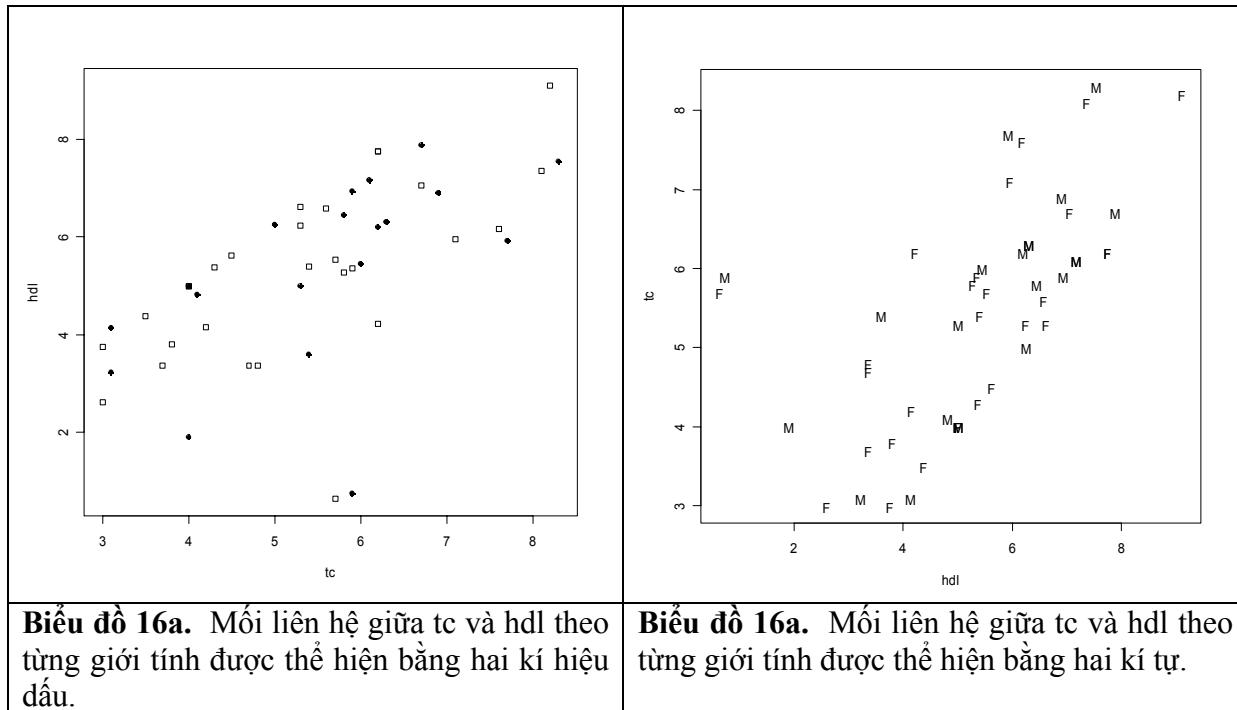


Chúng ta muốn phân biệt giới tính (nam và nữ) trong biểu đồ trên. Để vẽ biểu đồ đó, chúng ta phải dùng đến hàm `ifelse`. Trong lệnh sau đây, nếu `sex=="Nam"` thì vẽ kí tự số 16 (ô tròn), nếu không nam thì vẽ kí tự số 22 (tức ô vuông):

```
> plot(hdl, tc, pch=ifelse(sex=="Nam", 16, 22))
```

Kết quả là **Biểu đồ 16a**. Chúng ta cũng có thể thay kí tự thành “M” (nam) và “F” (nữ) (xem **Biểu đồ 16b**):

```
> plot(hdl, tc, pch=ifelse(sex=="Nam", "M", "F"))
```



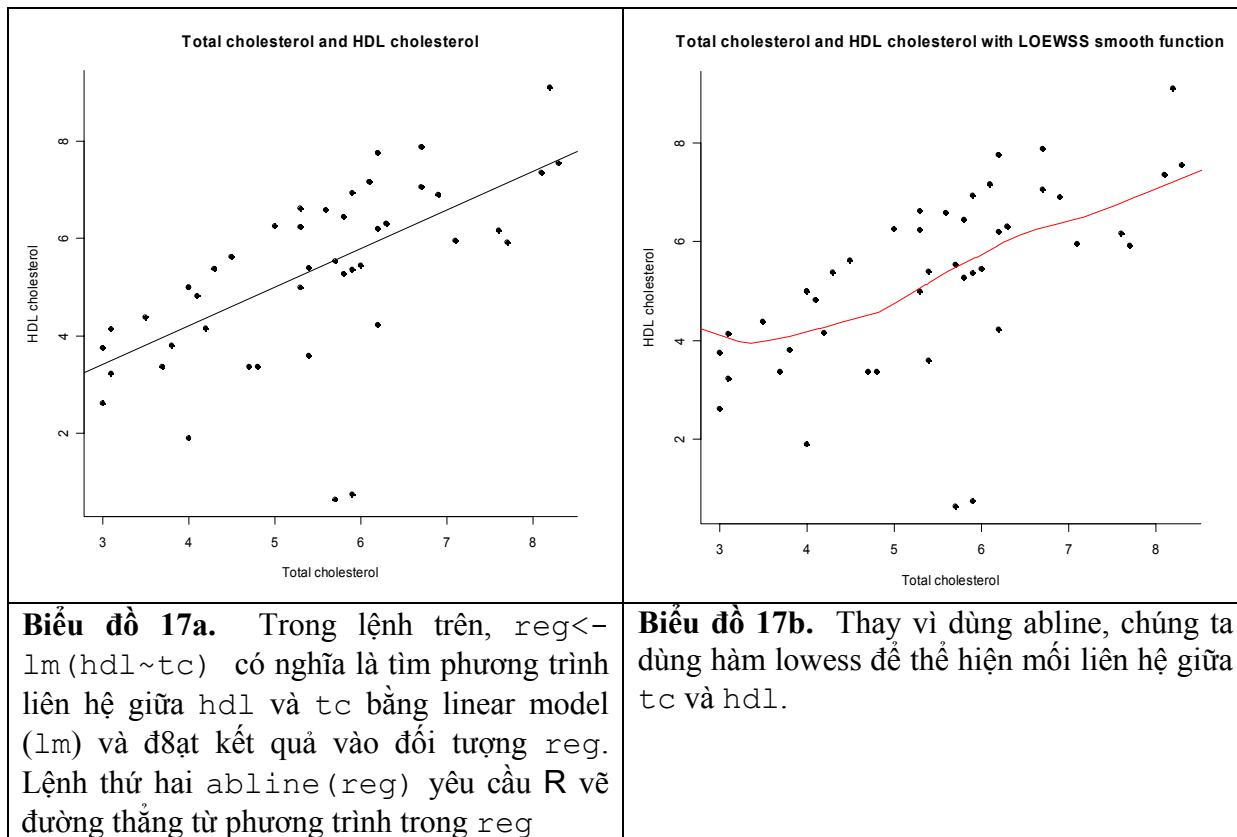
Chúng ta cũng có thể vẽ một đường biểu diễn hồi qui tuyến tính (regression line) qua các điểm trên bằng cách tiếp tục ra các lệnh sau đây:

```
> plot(hdl ~ tc, pch=16, main="Total cholesterol and HDL cholesterol",
      xlab="Total cholesterol", ylab="HDL cholesterol", bty="l")
> reg <- lm(hdl ~ tc)
> abline(reg)
```

Kết quả là **Biểu đồ 17a** dưới đây. Chúng ta cũng có thể dùng hàm trơn (smooth function) để biểu diễn mối liên hệ giữa hai biến số. Đồ thị sau đây sử dụng lowess (một hàm thông thường nhất) trong việc “làm trơn” số liệu tc và hdl (**Biểu đồ 17b**).

```
> plot(hdl ~ tc, pch=16,
      main="Total cholesterol and HDL cholesterol with LOESS smooth
      function",
      xlab="Total cholesterol", ylab="HDL cholesterol", bty="l")

> lines(lowess(hdl, tc, f=2/3, iter=3), col="red")
```



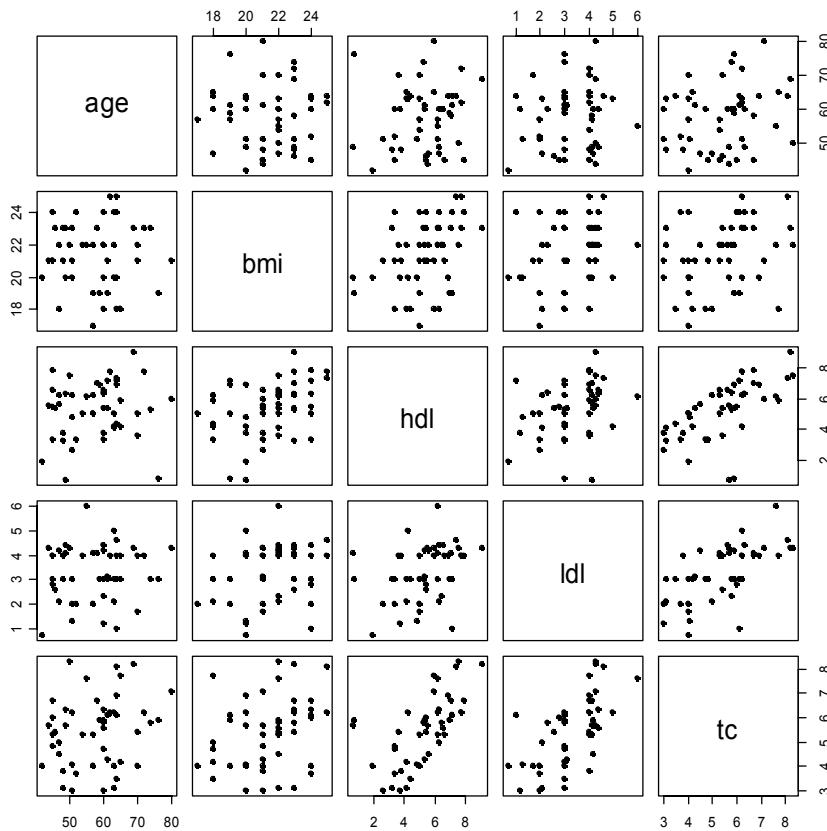
Bạn đọc có thể thí nghiệm với nhiều thông số $f=1/2$, $f=2/5$, hay thậm chí $f=1/10$ sẽ thấy đồ thị biến đổi một cách “thú vị”.

8.8 Phân tích Biểu đồ cho nhiều biến: `pairs`

Chúng ta có thể tìm hiểu mối liên hệ giữa các biến số như `age`, `bmi`, `hdl`, `ldl` và `tc` bằng cách dùng lệnh `pairs`. Nhưng trước hết, chúng ta phải đưa các biến số này vào một `data.frame` chỉ gồm những biến số có thể vẽ được, và sau đó sử dụng hàm `pairs` trong R.

```
> lipid <- data.frame(age,bmi,hdl,ldl,tc)
> pairs(lipid, pch=16)
```

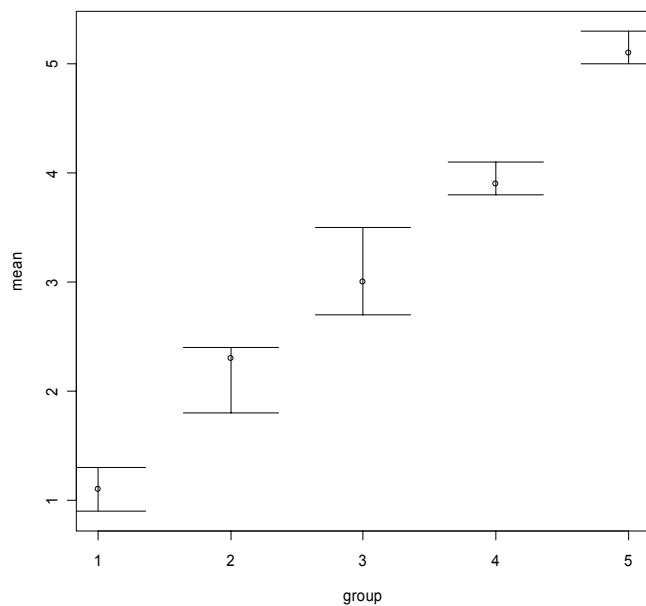
Kết quả sẽ là:



8.9 Biểu đồ với sai số chuẩn (standard error)

Trong biểu đồ sau đây, chúng ta có 5 nhóm (biến số x được mô phỏng chứ không phải số liệu thật), và mỗi nhóm có giá trị trung bình mean , và độ tin cậy 95% (lcl và ucl). Thông thường $\text{lcl} = \text{mean} - 1.96 * \text{SE}$ và $\text{ucl} = \text{mean} + 1.96 * \text{SE}$ (SE là sai số chuẩn). Chúng ta muốn vẽ biểu đồ cho 5 nhóm với sai số chuẩn đó. Các lệnh và hàm sau đây sẽ cần thiết:

```
> group <- c(1,2,3,4,5)
> mean <- c(1.1, 2.3, 3.0, 3.9, 5.1)
> lcl <- c(0.9, 1.8, 2.7, 3.8, 5.0)
> ucl <- c(1.3, 2.4, 3.5, 4.1, 5.3)
> plot(group, mean, ylim=range(c(lcl, ucl)))
> arrows(group, ucl, group, lcl, length=0.5, angle=90, code=3)
```



9. Phân tích thống kê mô tả

9.1 Thống kê mô tả (descriptive statistics, summary)

Để minh họa cho việc áp dụng R vào thống kê mô tả, tôi sẽ sử dụng một dữ liệu nghiên cứu có tên là `igfdata`. Trong nghiên cứu này, ngoài các chỉ số liên quan đến giới tính, độ tuổi, trọng lượng và chiều cao, chúng tôi đo lường các hormone liên quan đến tình trạng tăng trưởng như `igfi`, `igfbp3`, `als`, và các markers liên quan đến sự chuyển hóa của xương `pinp`, `ictp` và `pinp`. Có 100 đối tượng nghiên cứu. Dữ liệu này được chứa trong directory `c:\works\stats`. Trước hết, chúng ta cần phải nhập dữ liệu vào R với những lệnh sau đây (các câu chữ theo sau dấu # là những chú thích để bạn đọc theo dõi):

```
> options(width=100)
# chuyển directory
> setwd("c:/works/stats")

# đọc dữ liệu vào R
> igfdata <- read.table("igf.txt", header=TRUE, na.strings=".")
> attach(igfdata)

# xem xét các cột số trong dữ liệu
> names(igfdata)
[1] "id"          "sex"         "age"         "weight"       "height"       "ethnicity"
[7] "igfi"        "igfbp3"      "als"         "pinp"        "ictp"        "p3np"

> igfdata
   id    sex age weight height ethnicity    igfi    igfbp3      als     pinp     ictp     p3np

```

1	1	Female	15	42	162	Asian	189.000	4.00000	323.667	353.970	11.2867	8.3367
2	2	Male	16	44	160	Caucasian	160.000	3.75000	333.750	375.885	10.4300	6.7450
3	3	Female	15	43	157	Asian	146.833	3.43333	248.333	199.507	8.3633	12.5000
4	4	Female	15	42	155	Asian	185.500	3.40000	251.000	483.607	13.3300	14.2767
5	5	Female	16	47	167	Asian	192.333	4.23333	322.000	105.430	7.9233	4.5033
6	6	Female	25	45	160	Asian	110.000	3.50000	284.667	76.487	4.9833	4.9367
7	7	Female	19	45	161	Asian	157.000	3.20000	274.000	75.880	6.3500	5.3200
8	8	Female	18	43	153	Asian	146.000	3.40000	303.000	86.360	7.3700	4.6700
9	9	Female	15	41	149	Asian	197.667	3.56667	308.500	254.803	11.8700	6.8200
10	10	Female	24	45	157	African	148.000	3.40000	273.000	44.720	3.7400	6.1600
...												
...												
97	97	Female	17	54	168	Caucasian	204.667	4.96667	441.333	64.130	5.1600	4.4367
98	98	Male	18	55	169	Asian	178.667	3.86667	273.000	185.913	7.5267	8.8333
99	99	Female	18	48	151	Asian	237.000	3.46667	324.333	105.127	5.9867	5.6600
100	100	Male	15	54	168	Asian	130.000	2.70000	259.333	325.840	10.2767	6.5933

Trên đây chỉ là một phần số liệu trong số 100 đối tượng.

Cho một biến số $x_1, x_2, x_3, \dots, x_n$ chúng ta có thể tính toán một số chỉ số thống kê mô tả như sau:

Lí thuyết	Hàm R
Số trung bình: $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$	mean(x)
Phương sai: $s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$	var(x)
Độ lệch chuẩn: $s = \sqrt{s^2}$	sd(x)
Sai số chuẩn (standard error): $SE = \frac{s}{\sqrt{n}}$	Không có
Trị số thấp nhất	min(x)
Trị số cao nhất	max(x)
Toàn cự (range)	range(x)

Ví dụ 9: Để tìm giá trị trung bình của độ tuổi, chúng ta chỉ đơn giản lệnh:

```
> mean(age)
[1] 19.17
```

Hay phương sai và độ lệch chuẩn của tuổi:

```
> var(age)
[1] 15.33444
```

```
> sd(age)
[1] 3.915922
```

Tuy nhiên, R có lệnh `summary` có thể cho chúng ta tất cả thông tin thống kê về một biến số:

```
> summary(age)
   Min. 1st Qu. Median     Mean 3rd Qu.     Max.
   13.00    16.00   19.00   19.17   21.25   34.00
```

Nói chung, kết quả này đơn giản và các viết tắt cũng có thể dễ hiểu. Chú ý, trong kết quả trên, có hai chỉ số “1st Qu” và “3rd Qu” có nghĩa là first quartile (tương đương với vị trí 25%) và third quartile (tương đương với vị trí 75%) của một biến số. First quartile = 16 có nghĩa là 25% đối tượng nghiên cứu có độ tuổi bằng hoặc nhỏ hơn 16 tuổi. Tương tự, Third quartile = 34 có nghĩa là 75% đối tượng có độ tuổi bằng hoặc thấp hơn 34 tuổi. Tất nhiên số trung vị (median) 19 cũng có nghĩa là 50% đối tượng có độ tuổi 19 trở xuống (hay 19 tuổi trở lên).

R không có hàm tính sai số chuẩn, và trong hàm `summary`, R cũng không cung cấp độ lệch chuẩn. Để có các số này, chúng ta có thể tự viết một hàm đơn giản (hãy gọi là `desc`) như sau:

```
desc <- function(x)
{
  av <- mean(x)
  sd <- sd(x)
  se <- sd/sqrt(length(x))
  c(MEAN=av, SD=sd, SE=se)
}
```

Và có thể gọi hàm này để tính bất cứ biến nào chúng ta muốn, như tính biến `als` sau đây:

```
> desc(als)
      MEAN           SD           SE
  301.841120  58.987189  5.898719
```

Để có một “quang cảnh” chung về dữ liệu `igfdata` chúng ta chỉ đơn giản lệnh `summary` như sau:

```
> summary(igfdata)
      id          sex         age        weight       height      ethnicity
Min. : 1.00  Female:69  Min. :13.00  Min. :41.00  Min. :149.0  African : 8
1st Qu.: 25.75 Male   :31   1st Qu.:16.00  1st Qu.:47.00  1st Qu.:157.0  Asian   :60
Median : 50.50                    Median :19.00  Median :50.00  Median :162.0  Caucasian:30
Mean   : 50.50                    Mean   :19.17  Mean   :49.91  Mean   :163.1  Others   : 2
3rd Qu.: 75.25                   3rd Qu.:21.25  3rd Qu.:53.00  3rd Qu.:168.0
Max.   :100.00                   Max.   :34.00  Max.   :60.00  Max.   :196.0

      igfi        igfbp3        als        pinp        ictp
```

```

Min.    : 85.71   Min.    :2.000   Min.    :192.7   Min.    : 26.74   Min.    : 2.697
1st Qu.:137.17   1st Qu.:3.292   1st Qu.:256.8   1st Qu.: 68.10   1st Qu.: 4.878
Median  :161.50   Median  :3.550   Median  :292.5   Median  :103.26   Median  : 6.338
Mean    :165.59   Mean    :3.617   Mean    :301.8   Mean    :167.17   Mean    : 7.420
3rd Qu.:186.46   3rd Qu.:3.875   3rd Qu.:331.2   3rd Qu.:196.45   3rd Qu.: 8.423
Max.    :427.00   Max.    :5.233   Max.    :471.7   Max.    :742.68   Max.    :21.237

```

```

p3np
Min.    : 2.343
1st Qu.: 4.433
Median  : 5.445
Mean    : 6.341
3rd Qu.: 7.150
Max.    :16.303

```

R tính toán tất cả các biến số nào có thể tính toán được! Thành ra, ngay cả cột id (tức mã số của đối tượng nghiên cứu) R cũng tính luôn! (và chúng ta biết kết quả của cột id chẳng có ý nghĩa thống kê gì). Đối với các biến số mang tính phân loại như sex và ethnicity (sắc tộc) thì R chỉ báo cáo tần số cho mỗi nhóm.

Kết quả trên cho tất cả đối tượng nghiên cứu. Nếu chúng ta muốn kết quả cho từng nhóm nam và nữ riêng biệt, hàm by trong R rất hữu dụng. Trong lệnh sau đây, chúng ta yêu cầu R tóm lược dữ liệu igfdata theo sex.

```
> by(igfdata, sex, summary)
```

sex: Female

	id	sex	age	weight	height
Min.	: 1.0	Female:69	Min. :13.00	Min. :41.00	Min. :149.0
1st Qu.	:21.0	Male : 0	1st Qu.:17.00	1st Qu.:47.00	1st Qu.:156.0
Median	:47.0		Median :19.00	Median :50.00	Median :162.0
Mean	:48.2		Mean :19.59	Mean :49.35	Mean :161.9
3rd Qu.	:75.0		3rd Qu.:22.00	3rd Qu.:52.00	3rd Qu.:166.0
Max.	:99.0		Max. :34.00	Max. :60.00	Max. :196.0
		ethnicity	igfi	igfbp3	als
African	: 4	Min. : 85.71	Min. : 2.767	Min. :204.3	
Asian	:43	1st Qu.:136.67	1st Qu.:3.333	1st Qu.:263.8	
Caucasian	:22	Median :163.33	Median :3.567	Median :302.7	
Others	: 0	Mean :167.97	Mean :3.695	Mean :311.5	
		3rd Qu.:186.17	3rd Qu.:3.933	3rd Qu.:361.7	
		Max. :427.00	Max. :5.233	Max. :471.7	
		pinp	ictp	p3np	
Min.	: 26.74	Min. : 2.697	Min. : 2.343		
1st Qu.	:62.75	1st Qu.: 4.717	1st Qu.: 4.337		
Median	: 78.50	Median : 5.537	Median : 5.143		
Mean	:108.74	Mean : 6.183	Mean : 5.643		
3rd Qu.	:115.26	3rd Qu.: 7.320	3rd Qu.: 6.143		
Max.	:502.05	Max. :13.633	Max. :14.420		

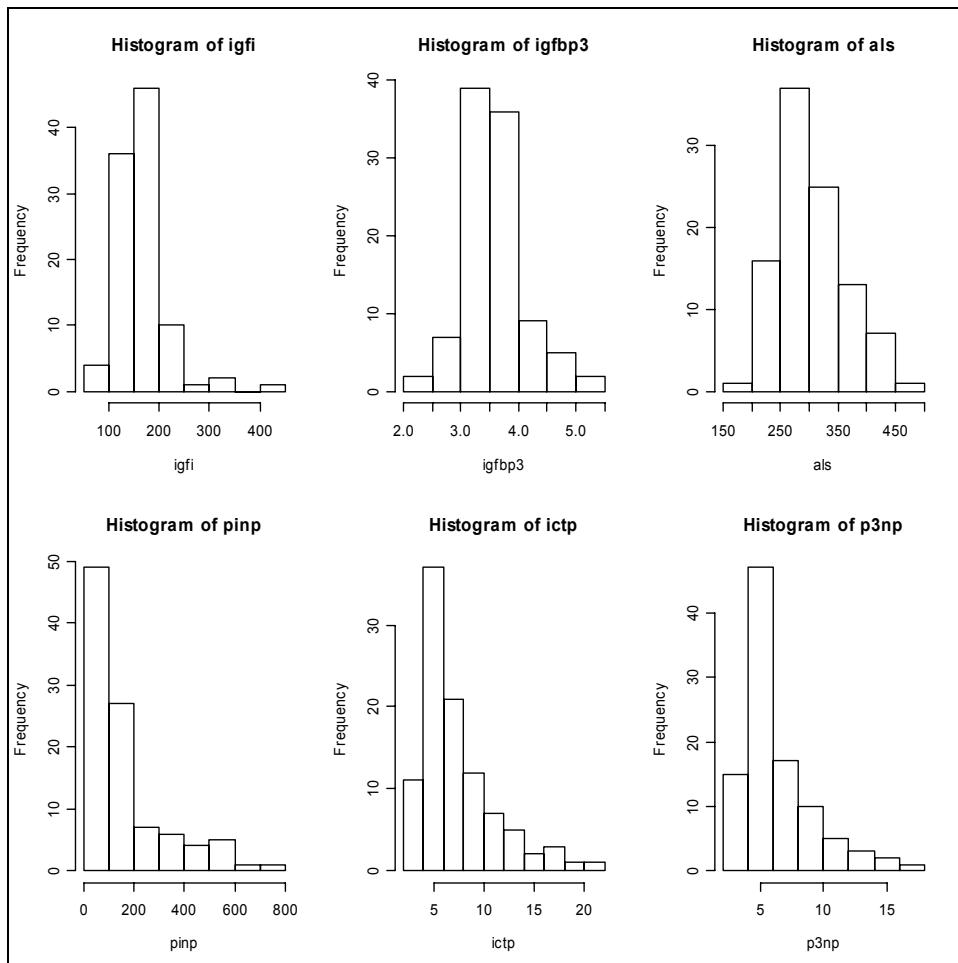
sex: Male

	id	sex	age	weight	height
Min.	: 2.00	Female: 0	Min. :14.00	Min. :44.00	Min. :155.0
1st Qu.	:34.50	Male :31	1st Qu.:15.00	1st Qu.:48.50	1st Qu.:161.5
Median	: 56.00		Median :17.00	Median :51.00	Median :164.0
Mean	: 55.61		Mean :18.23	Mean :51.16	Mean :165.6
3rd Qu.	: 75.00		3rd Qu.:20.00	3rd Qu.:53.50	3rd Qu.:169.0
Max.	:100.00		Max. :27.00	Max. :59.00	Max. :191.0
		ethnicity	igfi	igfbp3	als

African : 4	Min. : 94.67	Min. : 2.000	Min. : 192.7
Asian : 17	1st Qu.: 138.67	1st Qu.: 3.183	1st Qu.: 249.8
Caucasian: 8	Median : 160.00	Median : 3.500	Median : 276.0
Others : 2	Mean : 160.29	Mean : 3.443	Mean : 280.2
	3rd Qu.: 183.00	3rd Qu.: 3.775	3rd Qu.: 311.3
	Max. : 274.00	Max. : 4.500	Max. : 388.7
pinp	ictp	p3np	
Min. : 56.28	Min. : 3.650	Min. : 3.390	
1st Qu.: 135.07	1st Qu.: 6.900	1st Qu.: 5.375	
Median : 245.92	Median : 9.513	Median : 7.140	
Mean : 297.21	Mean : 10.173	Mean : 7.895	
3rd Qu.: 450.38	3rd Qu.: 13.517	3rd Qu.: 10.010	
Max. : 742.68	Max. : 21.237	Max. : 16.303	

Để xem qua phân phối của các hormones và chỉ số sinh hóa cùng một lúc, chúng ta có thể vẽ đồ thị cho tất cả 6 biến số. Trước hết, chia màn ảnh thành 6 cửa sổ (với 2 dòng và 3 cột); sau đó lần lượt vẽ:

```
> op <- par(mfrow=c(2, 3))
> hist(igfi)
> hist(igfbp3)
> hist(als)
> hist(pinp)
> hist(ictp)
> hist(p3np)
```



9.2 Thống kê mô tả theo từng nhóm

Nếu chúng ta muốn tính trung bình của một biến số như igfi cho mỗi nhóm nam và nữ giới, hàm tapply trong R có thể dùng cho việc này:

```
> tapply(igfi, list(sex), mean)
  Female      Male
  167.9741  160.2903
```

Trong lệnh trên, igfi là biến số chúng ta cần tính, biến số phân nhóm là sex, và chỉ số thống kê chúng ta muốn là trung bình (mean). Qua kết quả trên, chúng ta thấy số trung bình của igfi cho nữ giới (167.97) cao hơn nam giới (160.29).

Nhưng nếu chúng ta muốn tính cho từng giới tính và sắc tộc, chúng ta chỉ cần thêm một biến số trong hàm list:

```
> tapply(igfi, list(ethnicity, sex), mean)
  Female      Male
  African  145.1252 120.9168
```

Asian	165.6589	160.4999
Caucasian	176.6536	169.4790
Others	NA	200.5000

Trong kết quả trên, NA có nghĩa là “not available”, tức không có số liệu cho phụ nữ trong các sắc tộc “others”.

9.3 Kiểm định t (t.test)

Kiểm định t dựa vào giả thiết phân phối chuẩn. Có hai loại kiểm định t: kiểm định t cho một mẫu (one-sample t-test), và kiểm định t cho hai mẫu (two-sample t-test). Kiểm định t một mẫu nhằm trả lời câu hỏi dữ liệu từ một mẫu có phải thật sự bằng một thông số nào đó hay không. Còn kiểm định t hai mẫu thì nhằm trả lời câu hỏi hai mẫu có cùng một luật phân phối, hay cụ thể hơn là hai mẫu có thật sự có cùng trị số trung bình hay không. Tôi sẽ lần lượt minh họa hai kiểm định này qua số liệu igfdata trên.

9.3.1 Kiểm định t một mẫu

Ví dụ 10. Qua phân tích trên, chúng ta thấy tuổi trung bình của 100 đối tượng trong nghiên cứu này là 19.17 tuổi. Chẳng hạn như trong quân thể này, trước đây chúng ta biết rằng tuổi trung bình là 30 tuổi. Vấn đề đặt ra là có phải mẫu mà chúng ta có được có đại diện cho quân thể hay không. Nói cách khác, chúng ta muốn biết giá trị trung bình 19.17 có thật sự khác với giá trị trung bình 30 hay không.

Để trả lời câu hỏi này, chúng ta sử dụng kiểm định t. Theo lí thuyết thống kê, kiểm định t được định nghĩa bằng công thức sau đây:

$$t = \frac{\bar{x} - \mu}{s / \sqrt{n}}$$

Trong đó, \bar{x} là giá trị trung bình của mẫu, μ là trung bình theo giả thiết (trong trường hợp này, 30), s là độ lệch chuẩn, và n là số lượng mẫu (100). Nếu giá trị t cao hơn giá trị lí thuyết theo phân phối t ở một tiêu chuẩn có ý nghĩa như 5% chẳng hạn thì chúng ta có lí do để phát biểu khác biệt có ý nghĩa thống kê. Giá trị này cho mẫu 100 có thể tính toán bằng hàm qt của R như sau:

```
> qt(0.95, 100)
[1] 1.660234
```

Nhưng có một cách tính toán nhanh gọn hơn để trả lời câu hỏi trên, bằng cách dùng hàm t.test như sau:

```
> t.test(age, mu=30)
```

```
One Sample t-test
```

```

data: age
t = -27.6563, df = 99, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 30
95 percent confidence interval:
18.39300 19.94700
sample estimates:
mean of x
19.17

```

Trong lệnh trên `age` là biến số chúng ta cần kiểm định, và $\mu=30$ là giá trị giả thiết. R trình bày trị số $t = -27.66$, với 99 bậc tự do, và trị số $p < 2.2e-16$ (tức rất thấp). R cũng cho biết độ tin cậy 95% của `age` là từ 18.4 tuổi đến 19.9 tuổi (30 tuổi nằm quá ngoài khoảng tin cậy này). Nói cách khác, chúng ta có lí do để phát biểu rằng độ tuổi trung bình trong mẫu này thật sự thấp hơn độ tuổi trung bình của quần thể.

9.3.2 Kiểm định t hai mẫu

Ví dụ 11. Qua phân tích mô tả trên (phím `summary`) chúng ta thấy phụ nữ có độ hormone `igfi` cao hơn nam giới (167.97 và 160.29). Câu hỏi đặt ra là có phải thật sự đó là một khác biệt có hệ thống hay do các yếu tố ngẫu nhiên gây nên. Trả lời câu hỏi này, chúng ta cần xem xét mức độ khác biệt trung bình giữa hai nhóm và độ lệch chuẩn của độ khác biệt.

$$t = \frac{\bar{x}_2 - \bar{x}_1}{SED}$$

Trong đó \bar{x}_1 và \bar{x}_2 là số trung bình của hai nhóm nam và nữ, và SED là độ lệch chuẩn của $(\bar{x}_1 - \bar{x}_2)$. Thực ra, SED có thể ước tính bằng công thức:

$$SED = \sqrt{SE_1^2 + SE_2^2}$$

Trong đó SE_1 và SE_2 là sai số chuẩn (standard error) của hai nhóm nam và nữ. Theo lí thuyết xác suất, t tuân theo luật phân phối t với bậc tự do $n_1 + n_2 - 2$, trong đó n_1 và n_2 là số mẫu của hai nhóm. Chúng ta có thể dùng R để trả lời câu hỏi trên bằng hàm `t.test` như sau:

```

> t.test(igfi ~ sex)

Welch Two Sample t-test

data: igfi by sex
t = 0.8412, df = 88.329, p-value = 0.4025
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-10.46855 25.83627
sample estimates:
mean in group Female mean in group Male
167.9741      160.2903

```

R trình bày các giá trị quan trọng trước hết:

```
t = 0.8412, df = 88.329, p-value = 0.4025
```

df là bậc tự do. Trị số p = 0.4025 cho thấy mức độ khác biệt giữa hai nhóm nam và nữ không có ý nghĩa thống kê (vì cao hơn 0.05 hay 5%).

```
95 percent confidence interval:  
-10.46855 25.83627
```

là khoảng tin cậy 95% về độ khác biệt giữa hai nhóm. Kết quả tính toán trên cho biết độ igf ở nữ giới có thể thấp hơn nam giới 10.5 ng/L hoặc cao hơn nam giới khoảng 25.8 ng/L. Vì độ khác biệt quá lớn và đó là thêm bằng chứng cho thấy không có khác biệt có ý nghĩa thống kê giữa hai nhóm.

Kiểm định trên dựa vào giả thiết hai nhóm nam và nữ có khác phương sai. Nếu chúng ta có lí do để cho rằng hai nhóm có cùng phương sai, chúng ta chỉ thay đổi một thông số trong hàm t với var.equal=TRUE như sau:

```
> t.test(igfi ~ sex, var.equal=TRUE)

Two Sample t-test

data: igfi by sex
t = 0.7071, df = 98, p-value = 0.4812
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-13.88137 29.24909
sample estimates:
mean in group Female    mean in group Male
167.9741                  160.2903
```

Về mặc số, kết quả phân tích trên có khác chút ít so với kết quả phân tích dựa vào giả định hai phương sai khác nhau, nhưng trị số p cũng đi đến một kết luận rằng độ khác biệt giữa hai nhóm không có ý nghĩa thống kê.

9.4 Kiểm định Wilcoxon cho hai mẫu (wilcox.test)

Kiểm định t dựa vào giả thiết là phân phối của một biến phải tuân theo luật phân phối chuẩn. Nếu giả định này không đúng, kết quả của kiểm định t có thể không hợp lý (valid). Để kiểm định phân phối của igfi, chúng ta có thể dùng hàm shapiro.test như sau:

```
> shapiro.test(igfi)

Shapiro-Wilk normality test
```

```
data: igfi
W = 0.8528, p-value = 1.504e-08
```

Trị số p nhỏ hơn 0.05 rất nhiều, cho nên chúng ta có thể nói rằng phân phối của igfi không tuân theo luật phân phối chuẩn. Trong trường hợp này, việc so sánh giữa hai nhóm có thể dựa vào phương pháp phi tham số (non-parametric) có tên là kiểm định Wilcoxon, vì kiểm định này (không như kiểm định t) không tùy thuộc vào giả định phân phối chuẩn.

```
> wilcox.test(igfi ~ sex)

Wilcoxon rank sum test with continuity correction

data: igfi by sex
W = 1125, p-value = 0.6819
alternative hypothesis: true mu is not equal to 0
```

Trị số p = 0.682 cho thấy quả thật độ khác biệt về igfi giữa hai nhóm nam và nữ không có ý nghĩa thống kê. Kết luận này cũng không khác với kết quả phân tích bằng kiểm định t.

9.5 Kiểm định t cho các biến số theo cặp (paired t-test, t.test)

Kiểm định t vừa trình bày trên là cho các nghiên cứu gồm hai nhóm độc lập nhau (như giữa hai nhóm nam và nữ), nhưng không thể ứng dụng cho các nghiên cứu mà một nhóm đối tượng được theo dõi theo thời gian. Tôi tạm gọi các nghiên cứu này là nghiên cứu theo cặp. Trong các nghiên cứu này, chúng ta cần sử dụng một kiểm định t có tên là paired t-test.

Ví dụ 12. Một nhóm bệnh nhân gồm 10 người được điều trị bằng một thuốc nhằm giảm huyết áp. Huyết áp của bệnh nhân được đo lúc khởi đầu nghiên cứu (lúc chưa điều trị), và sau khi điều trị. Số liệu huyết áp của 10 bệnh nhân như sau:

Trước khi điều trị (x_0)	180, 140, 160, 160, 220, 185, 145, 160, 160, 170
Sau khi điều trị (x_1)	170, 145, 145, 125, 205, 185, 150, 150, 145, 155

Câu hỏi đặt ra là độ biến chuyển huyết áp trên có đủ để kết luận rằng thuốc điều trị có hiệu quả giảm áp huyết. Để trả lời câu hỏi này, chúng ta dùng kiểm định t cho từng cặp như sau:

```
> # nhập dữ kiện
> before <- c(180, 140, 160, 160, 220, 185, 145, 160, 160, 170)
> after <- c(170, 145, 145, 125, 205, 185, 150, 150, 145, 155)
> bp <- data.frame(before, after)

> # kiểm định t
> t.test(before, after, paired=TRUE)
```

Paired t-test

```

data: before and after
t = 2.7924, df = 9, p-value = 0.02097
alternative hypothesis: true difference in means is not equal to
0
95 percent confidence interval:
 1.993901 19.006099
sample estimates:
mean of the differences
 10.5

```

Kết quả trên cho thấy sau khi điều trị áp suất máu giảm 10.5 mmHg, và khoảng tin cậy 95% là từ 2.0 mmHg đến 19 mmHg, với trị số $p = 0.0209$. Như vậy, chúng ta có bằng chứng để phát biểu rằng mức độ giảm huyết áp có ý nghĩa thống kê.

Chú ý nếu chúng ta phân tích sai bằng kiểm định thống kê cho hai nhóm độc lập dưới đây thì trị số $p = 0.32$ cho biết mức độ giảm áp suất không có ý nghĩa thống kê!

```
> t.test(before, after)
```

Welch Two Sample t-test

```

data: before and after
t = 1.0208, df = 17.998, p-value = 0.3209
alternative hypothesis: true difference in means is not equal to
0
95 percent confidence interval:
 -11.11065 32.11065
sample estimates:
mean of x mean of y
 168.0      157.5

```

9.6 Kiểm định Wilcoxon cho các biến số theo cặp (wilcox.test)

Thay vì dùng kiểm định t cho từng cặp, chúng ta cũng có thể sử dụng hàm `wilcox.test` cho cùng mục đích:

```
> wilcox.test(before, after, paired=TRUE)
```

Wilcoxon signed rank test with continuity correction

```

data: before and after
V = 42, p-value = 0.02291
alternative hypothesis: true mu is not equal to 0

```

Kết quả trên một lần nữa khẳng định rằng độ giảm áp suất máu có ý nghĩa thống kê với trị số ($p=0.023$) chẳng khác mấy so với kiểm định t cho từng cặp.

9.7 Tần số (frequency)

Hàm `table` trong R có chức năng cho chúng ta biết về tần số của một biến số mang tính phân loại như `sex` và `ethnicity`.

```
> table(sex)
sex
Female    Male
      69      31

> table(ethnicity)
ethnicity
  African     Asian Caucasian   Others
      8         60        30         2
```

Một bảng thống kê 2 chiều:

```
> table(sex, ethnicity)
  ethnicity
sex      African Asian Caucasian Others
Female      4     43       22       0
Male        4     17        8       2
```

Chú ý trong các bảng thống kê trên, hàm `table` không cung cấp cho chúng ta số phần trăm. Để tính số phần trăm, chúng ta cần đến hàm `prop.table` và cách sử dụng có thể minh họa như sau:

```
# tạo ra một object tên là freq để chứa kết quả tần số
> freq <- table(sex, ethnicity)

# kiểm tra kết quả
> freq
  ethnicity
sex      African Asian Caucasian Others
Female      4     43       22       0
Male        4     17        8       2

# dùng hàm margin.table để xem kết quả
> margin.table(freq, 1)
sex
Female    Male
      69      31

> margin.table(freq, 2)
ethnicity
  African     Asian Caucasian   Others
```

```
8          60         30          2
```

```
# tính phần trăm bằng hàm prop.table
> prop.table(freq, 1)
      ethnicity
sex      African      Asian Caucasian   Others
Female  0.05797101 0.62318841 0.31884058 0.00000000
Male    0.12903226 0.54838710 0.25806452 0.06451613
```

Trong bảng thống kê trên, prop.table tính tỉ lệ sắc tộc cho từng giới tính. Chẳng hạn như ở nữ giới (female), 5.8% là người Phi châu, 62.3% là người Á châu, 31.8% là người Tây phương da trắng . Tổng cộng là 100%. Tương tự, ở nam giới tỉ lệ người Phi châu là 12.9%, Á châu là 54.8%, v.v...

```
# tính phần trăm bằng hàm prop.table
> prop.table(freq, 2)
      ethnicity
sex      African      Asian Caucasian   Others
Female  0.5000000 0.7166667 0.7333333 0.0000000
Male    0.5000000 0.2833333 0.2666667 1.0000000
```

Trong bảng thống kê trên, prop.table tính tỉ lệ giới tính cho từng sắc tộc. Chẳng hạn như trong nhóm người Á châu, 71.7% là nữ và 28.3% là nam.

```
# tính phần trăm cho toàn bộ bảng
> freq/sum(freq)
      ethnicity
sex      African Asian Caucasian   Others
Female  0.04    0.43     0.22    0.00
Male    0.04    0.17     0.08    0.02
```

9.8 Kiểm định tỉ lệ (proportion test, prop.test, binom.test)

Kiểm định một tỉ lệ thường dựa vào giả định phân phối nhị phân (binomial distribution). Với một số mẫu n và tỉ lệ p , và nếu n lớn (tức hơn 50 chẳng hạn), thì phân phối nhị phân có thể tương đương với phân phối chuẩn với số trung bình np và phuơng sai $np(1 - p)$. Gọi x là số biến cố mà chúng ta quan tâm, kiểm định giả thiết $p = \pi$ có thể sử dụng thống kê sau đây:

$$z = \frac{x - n\pi}{\sqrt{n\pi(1-\pi)}}$$

Ở đây, z tuân theo luật phân phối chuẩn với trung bình 0 và phuơng sai 1. Cũng có thể nói z^2 tuân theo luật phân phối Chi bình phuơng với bậc tự do bằng 1.

Ví dụ 13. Trong nghiên cứu trên, chúng ta thấy có 69 nữ và 31 nam. Như vậy tỉ lệ nữ là 0.69 (hay 69%). Để kiểm định xem tỉ lệ này có thật sự khác với tỉ lệ 0.5 hay không, chúng ta có thể sử dụng hàm `prop.test(x, n, pi)` như sau:

```
> prop.test(69, 100, 0.50)

 1-sample proportions test with continuity correction

data: 69 out of 100, null probability 0.5
X-squared = 13.69, df = 1, p-value = 0.0002156
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.5885509 0.7766330
sample estimates:
      p
 0.69
```

Trong kết quả trên, `prop.test` ước tính tỉ lệ nữ giới là 0.69, và khoảng tin cậy 95% là 0.588 đến 0.776. Giá trị Chi bình phương là 13.69, với trị số $p = 0.00216$. Như vậy, nghiên cứu này có tỉ lệ nữ cao hơn 50%.

Một cách tính chính xác hơn kiểm định tỉ lệ là kiểm định nhị phân `binom.test(x, n, pi)` như sau:

```
> binom.test(69, 100, 0.50)

  Exact binomial test

data: 69 and 100
number of successes = 69, number of trials = 100, p-value = 0.0001831
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.5896854 0.7787112
sample estimates:
probability of success
 0.69
```

Nói chung, kết quả của kiểm định nhị phân không khác gì so với kiểm định Chi bình phương, với trị số $p = 0.00018$, chúng ta càng có bằng chứng để kết luận rằng tỉ lệ nữ giới trong nghiên cứu này thật sự cao hơn 50%.

9.9 So sánh hai tỉ lệ (`prop.test`, `binom.test`)

Phương pháp so sánh hai tỉ lệ có thể khai triển trực tiếp từ lí thuyết kiểm định một tỉ lệ vừa trình bày trên. Cho hai mẫu với số đối tượng n_1 và n_2 , và số biến cố là x_1 và x_2 . Do đó, chúng ta có thể ước tính hai tỉ lệ p_1 và p_2 . Lí thuyết xác suất cho phép chúng ta phát biểu rằng độ khác biệt giữa hai mẫu $d = p_1 - p_2$ tuân theo luật phân phối chuẩn với số trung bình 0 và phương sai bằng:

$$V_d = \left(\frac{1}{n_1} + \frac{1}{n_2} \right) p(1-p)$$

Trong đó:

$$p = \frac{x_1 + x_2}{n_1 + n_2}$$

Thành ra, $z = d/V_d$ tuân theo luật phân phối chuẩn với trung bình 0 và phương sai 1. Nói cách khác, z^2 tuân theo luật phân phối Chi bình phương với bậc tự do bằng 1. Do đó, chúng ta cũng có thể sử dụng `prop.test` để kiểm định hai tỉ lệ.

Ví dụ 14. Một nghiên cứu được tiến hành so sánh hiệu quả của thuốc chống gãy xương. Bệnh nhân được chia thành hai nhóm: nhóm A được điều trị gồm có 100 bệnh nhân, và nhóm B không được điều trị gồm 110 bệnh nhân. Sau thời gian 12 tháng theo dõi, nhóm A có 7 người bị gãy xương, và nhóm B có 20 người gãy xương. Vấn đề đặt ra là tỉ lệ gãy xương trong hai nhóm này bằng nhau (tức thuốc không có hiệu quả)? Để kiểm định xem hai tỉ lệ này có thật sự khác nhau, chúng ta có thể sử dụng hàm `prop.test(x, n, π)` như sau:

```
> fracture <- c(7, 20)
> total <- c(100, 110)
> prop.test(fracture, total)

 2-sample test for equality of proportions with continuity
correction

data: fracture out of total
X-squared = 4.8901, df = 1, p-value = 0.02701
alternative hypothesis: two.sided
95 percent confidence interval:
-0.20908963 -0.01454673
sample estimates:
prop 1    prop 2
0.0700000 0.1818182
```

Kết quả phân tích trên cho thấy tỉ lệ gãy xương trong nhóm 1 là 0.07 và nhóm 2 là 0.18. Phân tích trên còn cho thấy xác suất 95% rằng độ khác biệt giữa hai nhóm có thể 0.01 đến 0.20 (tức 1 đến 20%). Với trị số $p = 0.027$, chúng ta có thể nói rằng tỉ lệ gãy xương trong nhóm A quả thật thấp hơn nhóm B.

9.10 So sánh nhiều tỉ lệ (`prop.test`, `chisq.test`)

Kiểm định `prop.test` còn có thể sử dụng để kiểm định nhiều tỉ lệ cùng một lúc. Trong nghiên cứu trên, chúng ta có 4 nhóm sắc tộc và tần số cho từng giới tính như sau:

```
> table(sex, ethnicity)
```

	ethnicity			
sex	African	Asian	Caucasian	Others
Female	4	43	22	0
Male	4	17	8	2

Chúng ta muốn biết tỉ lệ nữ giới giữa 4 nhóm sắc tộc có khác nhau hay không, và để trả lời câu hỏi này, chúng ta lại dùng `prop.test` như sau:

```
> female <- c( 4, 43, 22, 0)
> total <- c(8, 60, 30, 2)
> prop.test(female, total)

 4-sample test for equality of proportions without continuity
 correction

data: female out of total
X-squared = 6.2646, df = 3, p-value = 0.09942
alternative hypothesis: two.sided
sample estimates:
prop 1    prop 2    prop 3    prop 4 
0.5000000 0.7166667 0.7333333 0.0000000

Warning message:
Chi-squared approximation may be incorrect in: prop.test(female, total)
```

Tuy tỉ lệ nữ giới giữa các nhóm có vẻ khác nhau lớn (73% trong nhóm 3 (người da trắng) so với 50% trong nhóm 1 (Phi châu) và 71.7% trong nhóm Á châu, nhưng kiểm định Chi bình phương cho biết trên phương diện thống kê, các tỉ lệ này không khác nhau, vì trị số $p = 0.099$.

9.10.1 Kiểm định Chi bình phương (Chi squared test, `chisq.test`)

Thật ra, kiểm định Chi bình phương còn có thể tính toán bằng hàm `chisq.test` như sau:

```
> chisq.test(sex, ethnicity)

Pearson's Chi-squared test

data: sex and ethnicity
X-squared = 6.2646, df = 3, p-value = 0.09942

Warning message:
Chi-squared approximation may be incorrect in: chisq.test(sex,
ethnicity)
```

Kết quả này hoàn toàn giống với kết quả từ hàm `prop.test`.

9.10.2 Kiểm định Fisher (Fisher's exact test, fisher.test)

Trong kiểm định Chi bình phương trên, chúng ta chú ý cảnh báo:

"Warning message:

Chi-squared approximation may be incorrect in: prop.test(female, total)"

Vì trong nhóm 4, không có nữ giới cho nên tỉ lệ là 0%. Hơn nữa, trong nhóm này chỉ có 2 đối tượng. Vì số lượng đối tượng quá nhỏ, cho nên các ước tính thống kê có thể không đáng tin cậy. Một phương pháp khác có thể áp dụng cho các nghiên cứu với tần số thấp như trên là kiểm định fisher (còn gọi là Fisher's exact test). Bạn đọc có thể tham khảo lí thuyết đăng sau kiểm định fisher để hiểu rõ hơn về logic của phương pháp này, nhưng ở đây, chúng ta chỉ quan tâm đến cách dùng R để tính toán kiểm định này. Chúng ta chỉ đơn giản lệnh:

```
> fisher.test(sex, ethnicity)

Fisher's Exact Test for Count Data

data: sex and ethnicity
p-value = 0.1048
alternative hypothesis: two.sided
```

Chú ý trị số p từ kiểm định Fisher là 0.1048, tức rất gần với trị số p của kiểm định Chi bình phương. Cho nên, chúng ta có thêm bằng chứng để khẳng định rằng tỉ lệ nữ giới giữa các sắc tộc không khác nhau một cách đáng kể.

10. Phân tích hồi qui tuyến tính

Ví dụ 15. Để minh họa cho vấn đề, chúng ta thử xem xét nghiên cứu sau đây, mà trong đó nhà nghiên cứu đo lường độ cholesterol trong máu của 18 đối tượng nam. Tỉ trọng cơ thể (body mass index) cũng được ước tính cho mỗi đối tượng bằng công thức tính BMI là lấy trọng lượng (tính bằng kg) chia cho chiều cao bình phương (m^2). Kết quả đo lường như sau:

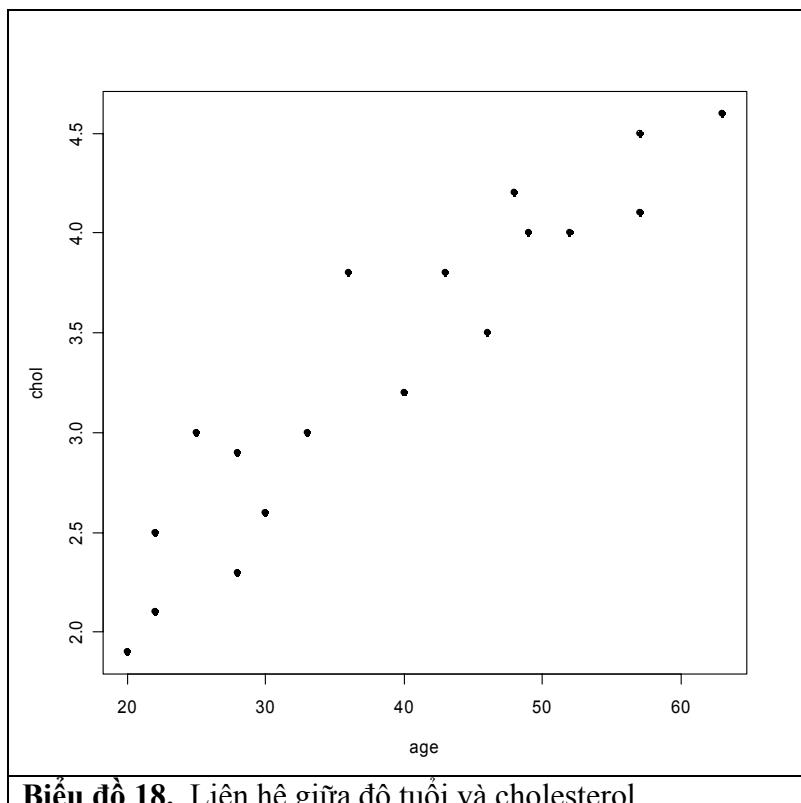
Độ tuổi, tỉ trọng cơ thể và cholesterol

Mã số ID (id)	Độ tuổi (age)	BMI (bmi)	Cholesterol (cho1)
1	46	25.4	3.5
2	20	20.6	1.9
3	52	26.2	4.0
4	30	22.6	2.6
5	57	25.4	4.5
6	25	23.1	3.0
7	28	22.7	2.9
8	36	24.9	3.8

9	22	19.8	2.1
10	43	25.3	3.8
11	57	23.2	4.1
12	33	21.8	3.0
13	22	20.9	2.5
14	63	26.7	4.6
15	40	26.4	3.2
16	48	21.2	4.2
17	28	21.2	2.3
18	49	22.8	4.0

Nhìn sơ qua số liệu chúng ta thấy người có độ tuổi càng cao độ cholesterol cũng càng cao. Chúng ta thử nhập số liệu này vào R và vẽ một biểu đồ tán xạ như sau:

```
> age <- c(46,20,52,30,57,25,28,36,22,43,57,33,22,63,40,48,28,49)
> bmi <- c(25.4,20.6,26.2,22.6,25.4,23.1,22.7,24.9,19.8,25.3,23.2,
   21.8,20.9,26.7,26.4,21.2,21.2,22.8)
> chol <- c(3.5,1.9,4.0,2.6,4.5,3.0,2.9,3.8,2.1,3.8,4.1,3.0,
   2.5,4.6,3.2, 4.2,2.3,4.0)
> data <- data.frame(age, bmi, chol)
> plot(chol ~ age, pch=16)
```



Biểu đồ 18 trên đây gợi ý cho thấy mối liên hệ giữa độ tuổi (age) và cholesterol là một đường thẳng (tuyến tính). Để “đo lường” mối liên hệ này, chúng ta có thể sử dụng hệ số tương quan (coefficient of correlation).

10.1 Hệ số tương quan

Hệ số tương quan (r) là một chỉ số thống kê đo lường mối liên hệ tương quan giữa hai biến số, như giữa độ tuổi (x) và cholesterol (y). Hệ số tương quan có giá trị từ -1 đến 1. Hệ số tương quan bằng 0 (hay gần 0) có nghĩa là hai biến số không có liên hệ gì với nhau; ngược lại nếu hệ số bằng -1 hay 1 có nghĩa là hai biến số có một mối liên hệ tuyệt đối. Nếu giá trị của hệ số tương quan là âm ($r < 0$) có nghĩa là khi x tăng cao thì y giảm (và ngược lại, khi x giảm thì y tăng); nếu giá trị hệ số tương quan là dương ($r > 0$) có nghĩa là khi x tăng cao thì y cũng tăng, và khi x tăng cao thì y cũng giảm theo.

Thực ra có nhiều hệ số tương quan trong thống kê, nhưng ở đây tôi sẽ trình bày 3 hệ số tương quan thông dụng nhất: hệ số tương quan Pearson r , Spearman ρ , và Kendall τ .

10.1.1 Hệ số tương quan Pearson

Cho hai biến số x và y từ n mẫu, hệ số tương quan Pearson được ước tính bằng

$$\text{công thức sau đây: } r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}. \text{ Trong đó, như định nghĩa phần trên, } \bar{x}$$

và \bar{y} là giá trị trung bình của biến số x và y . Để ước tính hệ số tương quan giữa độ tuổi age và cholesterol, chúng ta có thể sử dụng hàm `cor(x, y)` như sau:

```
> cor(age, chol)
[1] 0.936726
```

Chúng ta có thể kiểm định giả thiết hệ số tương quan bằng 0 (tức hai biến x và y không có liên hệ). Phương pháp kiểm định này thường dựa vào phép biến đổi Fisher mà R đã có sẵn một hàm `cor.test` để tiến hành việc tính toán.

```
> cor.test(age, chol)

Pearson's product-moment correlation

data: age and chol
t = 10.7035, df = 16, p-value = 1.058e-08
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
0.8350463 0.9765306
sample estimates:
cor
0.936726
```

10.1.2 Hệ số tương quan Spearman ρ

Hệ số tương quan Pearson chỉ hợp lí nếu biến số x và y tuân theo luật phân phối chuẩn. Nếu x và y không tuân theo luật phân phối chuẩn, chúng ta phải sử dụng một hệ số tương quan khác tên là Spearman, một phương pháp phân tích phi tham số. Hệ số này được ước tính bằng cách biến đổi hai biến số x và y thành thứ bậc (rank), và xem độ tương quan giữa hai dãy số bậc. Do đó, hệ số còn có tên tiếng Anh là Spearman's Rank correlation. R ước tính hệ số tương quan Spearman bằng hàm `cor.test` với thông số `method="spearman"` như sau:

```
> cor.test(age, chol, method="spearman")

  Spearman's rank correlation rho

data: age and chol
S = 51.1584, p-value = 2.57e-09
alternative hypothesis: true rho is not equal to 0
sample estimates:
rho
0.947205

Warning message:
Cannot compute exact p-values with ties in: cor.test.default(age,
chol, method = "spearman")
```

10.1.3 Hệ số tương quan Kendall τ

Hệ số tương quan Kendall (cũng là một phương pháp phân tích phi tham số) được ước tính bằng cách tìm các cặp số (x, y) “song hành” với nhau. Một cặp (x, y) song hành ở đây được định nghĩa là hiệu (độ khác biệt) trên trực hoành có cùng dấu hiệu (dương hay âm) với hiệu trên trực tung. Nếu hai biến số x và y không có liên hệ với nhau, thì số cặp song hành bằng hay tương đương với số cặp không song hành.

Bởi vì có nhiều cặp phải kiểm định, phương pháp tính toán hệ số tương quan Kendall đòi hỏi thời gian của máy tính khá cao. Tuy nhiên, nếu một dữ liệu dưới 5000 đối tượng thì một máy vi tính có thể tính toán khá dễ dàng. R dùng hàm `cor.test` với thông số `method="kendall"` để ước tính hệ số tương quan Kendall:

```
> cor.test(age, chol, method="kendall")

  Kendall's rank correlation tau

data: age and chol
z = 4.755, p-value = 1.984e-06
alternative hypothesis: true tau is not equal to 0
sample estimates:
tau
0.8333333
```

Warning message:

Cannot compute exact p-value with ties in: cor.test.default(age, chol, method = "kendall")

10.2 Mô hình của hồi qui tuyến tính đơn giản

Để tiện việc theo dõi và mô tả mô hình, gọi độ tuổi cho cá nhân i là x_i và cholesterol là y_i . Ở đây $i = 1, 2, 3, \dots, 18$. Mô hình hồi tuyến tính phát biểu rằng:

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

Nói cách khác, phương trình trên giả định rằng độ cholesterol của một cá nhân bằng một hằng số α cộng với một hệ số β liên quan đến độ tuổi, và một sai số ε_i . Trong phương trình trên, α là *chặn* (intercept, tức giá trị lúc $x_i = 0$), và β là độ dốc (slope hay gradient). Trong thực tế, α và β là hai thông số (parameter, còn gọi là *regression coefficient* hay hệ số hồi qui), và ε_i là một biến số theo luật phân phối chuẩn với trung bình 0 và phương sai σ^2 .

Các thông số α , β và σ^2 phải được ước tính từ dữ liệu. Phương pháp để ước tính các thông số này là phương pháp *bình phương nhỏ nhất* (least squares method). Như tên gọi, phương pháp bình phương nhỏ nhất tìm giá trị α , β sao cho $\sum_{i=1}^n [y_i - (\alpha + \beta x_i)]^2$ nhỏ nhất. Sau vài thao tác toán, có thể chứng minh dễ dàng rằng, ước số cho α và β đáp ứng điều kiện đó là:

$$\hat{\beta} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad \text{và} \quad \hat{\alpha} = \bar{y} - \hat{\beta} \bar{x}$$

Ở đây, \bar{x} và \bar{y} là giá trị trung bình của biến số x và y . Chú ý, tôi viết $\hat{\alpha}$ và $\hat{\beta}$ (với dấu mũ phía trên) là để nhắc nhở rằng đây là hai ước số (estimates) của α và β , chứ không phải α và β (chúng ta không biết chính xác α và β , nhưng chỉ có thể ước tính mà thôi). Sau khi đã có ước số $\hat{\alpha}$ và $\hat{\beta}$, chúng ta có thể ước tính độ cholesterol trung bình cho từng độ tuổi như sau:

$$\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$$

Tất nhiên, \hat{y}_i ở đây chỉ là số trung bình cho độ tuổi x_i , và phần còn lại (tức $y_i - \hat{y}_i$) gọi là *phần dư* (residual). Và phương sai của phần dư có thể ước tính như sau:

$$s^2 = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n-2} . \quad \text{Ở đây, } s^2 \text{ chính là ước số của } \sigma^2.$$

Hàm `lm` (viết tắt từ **linear model**) trong R có thể tính toán các giá trị của $\hat{\alpha}$ và $\hat{\beta}$, cũng như s^2 một cách nhanh gọn. Chúng ta tiếp tục với ví dụ bằng R như sau:

```
> lm(chol ~ age)

Call:
lm(formula = chol ~ age)

Coefficients:
(Intercept)      age
1.08922       0.05779
```

Trong lệnh trên, “`chol ~ age`” có nghĩa là mô tả `chol` là một hàm số của `age`. Kết quả tính toán của `lm` cho thấy $\hat{\alpha} = 1.0892$ và $\hat{\beta} = 0.05779$. Nói cách khác, với hai thông số này, chúng ta có thể ước tính độ cholesterol cho bất cứ độ tuổi nào trong khoảng tuổi của mẫu bằng phương trình tuyến tính:

$$\hat{y}_i = 1.08922 + 0.05779 \times \text{age}$$

Phương trình này có nghĩa là khi độ tuổi tăng 1 năm thì độ cholesterol tăng khoảng 0.058 mmol/L.

Thật ra, hàm `lm` còn cung cấp cho chúng ta nhiều thông tin khác, nhưng chúng ta phải đưa các thông tin này vào một object. Gọi object đó là `reg`, thì lệnh sẽ là:

```
> reg <- lm(chol ~ age)
> summary(reg)

Call:
lm(formula = chol ~ age)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.40729 -0.24133 -0.04522  0.17939  0.63040 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 1.089218   0.221466   4.918 0.000154 ***
age         0.057788   0.005399  10.704 1.06e-08 ***
---
Signif. codes:  0 '****' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3027 on 16 degrees of freedom
Multiple R-Squared:  0.8775,    Adjusted R-squared:  0.8698 
F-statistic: 114.6 on 1 and 16 DF,  p-value: 1.058e-08
```

Lệnh thứ hai, `summary(reg)`, yêu cầu R liệt kê các thông tin tính toán trong `reg`. Phần kết quả chia làm 3 phần:

(a) Phần 1 mô tả phần dư (residuals) của mô hình hồi qui:

Residuals:

	Min	1Q	Median	3Q	Max
	-0.40729	-0.24133	-0.04522	0.17939	0.63040

Chúng ta biết rằng trung bình phần dư phải là 0, và ở đây, số trung vị là -0.04, cũng không xa 0 bao nhiêu. Các số quantiles 25% (1Q) và 75% (3Q) cũng khá cân đối chung quan số trung vị, cho thấy phần dư của phương trình này tương đối cân đối.

(b) Phần hai trình bày ước số của $\hat{\alpha}$ và $\hat{\beta}$ cùng với sai số chuẩn và giá trị của kiểm định t. Giá trị kiểm định t cho $\hat{\beta}$ là 10.74 với trị số p = 1.06e-08, cho thấy β không phải bằng 0. Nói cách khác, chúng ta có bằng chứng để cho rằng có một mối liên hệ giữa cholesterol và độ tuổi, và mối liên hệ này có ý nghĩa thống kê.

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	1.089218	0.221466	4.918	0.000154 ***
age	0.057788	0.005399	10.704	1.06e-08 ***

Signif. codes:	0 '***'	0.001 '**'	0.01 '*'	0.05 '.'
	0.1	'	'	1

(c) Phần ba của kết quả cho chúng ta thông tin về phương sai của phần dư (residual mean square). Ở đây, $s^2 = 0.3027$. Trong kết quả này còn có kiểm định F, cũng chỉ là một kiểm định xem có quả thật β bằng 0, tức có ý nghĩa tương tự như kiểm định t trong phần trên. Nói chung, trong trường hợp phân tích hồi qui tuyến tính đơn giản (với một yếu tố) chúng ta không cần phải quan tâm đến kiểm định F.

Residual standard error: 0.3027 on 16 degrees of freedom
Multiple R-Squared: 0.8775, Adjusted R-squared: 0.8698
F-statistic: 114.6 on 1 and 16 DF, p-value: 1.058e-08

Ngoài ra, phần 3 còn cho chúng ta một thông tin quan trọng, đó là trị số R^2 hay *hệ số xác định bởi* (coefficient of determination). Tức là bằng tổng bình phương giữa số ước tính và trung bình chia cho tổng bình phương số quan sát và trung bình. Trị số R^2 trong ví dụ này là 0.8775, có nghĩa là phương trình tuyến tính (với độ tuổi là một yếu tố) giải thích khoảng 88% các khác biệt về độ cholesterol giữa các cá nhân. Tất nhiên trị số R^2 có giá trị từ 0 đến 100% (hay 1). Giá trị R^2 càng cao là một dấu hiệu cho thấy mối liên hệ giữa hai biến số độ tuổi và cholesterol càng chặt chẽ.

Một hệ số cũng cần đề cập ở đây là *hệ số điều chỉnh xác định bởi* (mà trong kết quả trên R gọi là “Adjusted R-squared”). Đây là hệ số cho chúng ta biết mức độ cải tiến của phương sai phần dư (residual variance) do yếu tố độ tuổi có mặt trong mô hình tuyến tính. Nói chung, hệ số này không khác mấy so với hệ số xác định bởi, và chúng ta cũng không cần chú tâm quá mức.

Giả định của phân tích hồi qui tuyến tính

Tất cả các phân tích trên dựa vào một số giả định quan trọng như sau:

- (a) x là một biến số cố định hay fixed, (“cố định” ở đây có nghĩa là không có sai sót ngẫu nhiên trong đo lường);
- (b) ε_i phân phối theo luật phân phối chuẩn;
- (c) ε_i có giá trị trung bình (mean) là 0;
- (d) ε_i có phương sai σ^2 cố định cho tất cả x_i ; và
- (e) các giá trị liên tục của ε_i không có liên hệ tương quan với nhau (nói cách khác, ε_1 và ε_2 không có liên hệ với nhau).

Nếu các giả định này không được đáp ứng thì phương trình mà chúng ta ước tính có vấn đề hợp lý (validity). Do đó, trước khi trình bày và diễn dịch mô hình trên, chúng ta cần phải kiểm tra xem các giả định trên có đáp ứng được hay không. Trong trường hợp này, giả định (a) không phải là vấn đề, vì độ tuổi không phải là một biến số ngẫu nhiên, và không có sai số khi tính độ tuổi của một cá nhân.

Đối với các giả định (b) đến (e), cách kiểm tra đơn giản nhưng hữu hiệu nhất là bằng cách xem xét mối liên hệ giữa \hat{y}_i , x_i , và phần dư e_i ($e_i = y_i - \hat{y}_i$) bằng những đồ thị tán xạ.

Với lệnh `fitted()` chúng ta có thể tính toán \hat{y}_i cho từng cá nhân như sau (ví dụ đối với cá nhân 1, 46 tuổi, độ cholesterol có thể tiên đoán như sau: $1.08922 + 0.05779 \times 46 = 3.747$).

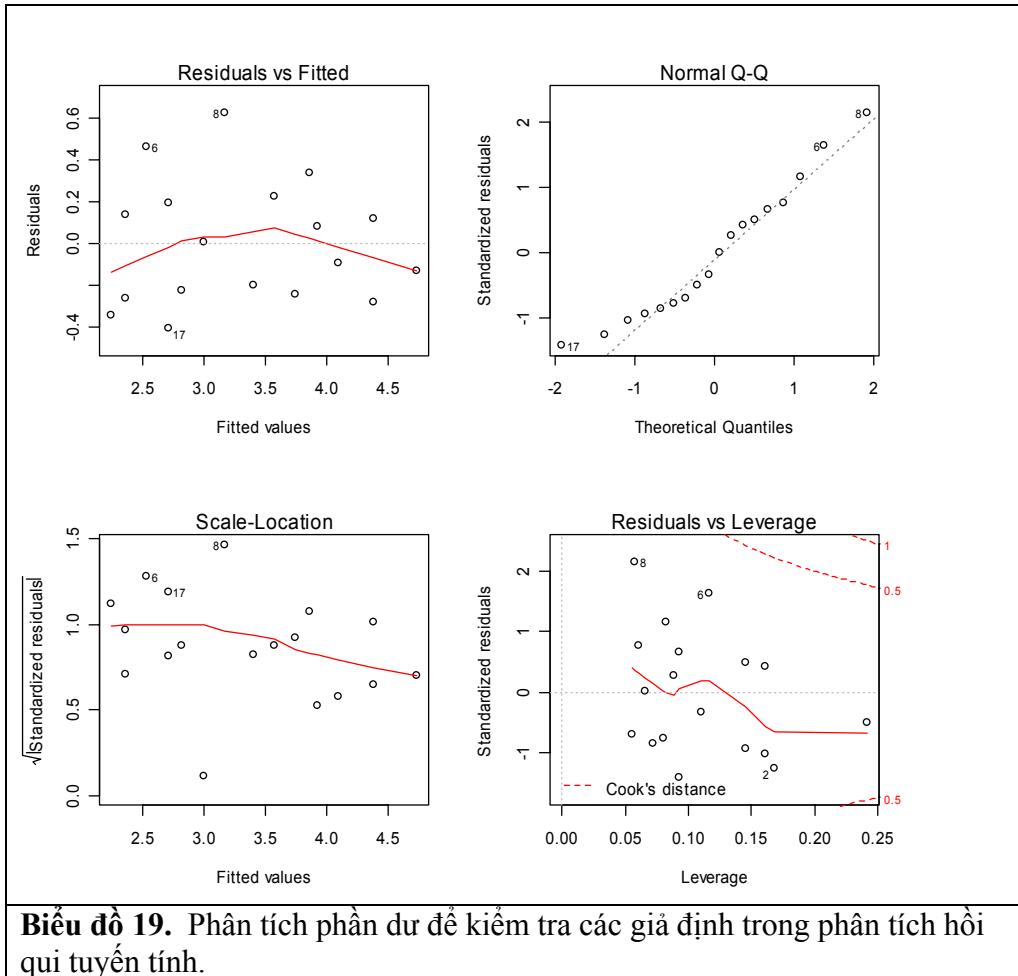
```
> fitted(reg)
   1          2          3          4          5          6          7          8
3.747483 2.244985 4.094214 2.822869 4.383156 2.533927 2.707292 3.169600
   9          10         11         12         13         14         15         16
2.360562 3.574118 4.383156 2.996234 2.360562 4.729886 3.400753 3.863060
  17         18
2.707292 3.920849
```

Với lệnh `resid()` chúng ta có thể tính toán phần dư e_i cho từng cá nhân như sau (với đối tượng 1, $e_1 = 3.5 - 3.74748 = -0.24748$):

```
> resid(reg)
   1          2          3          4          5          6
-0.247483426 -0.344985415 -0.094213736 -0.222869265 0.116844338 0.466072660
   7          8          9          10         11         12
  0.192707505  0.630400424 -0.260562185  0.225881729 -0.283155662  0.003765579
  13         14         15         16         17         18
  0.139437815 -0.129885972 -0.200753116  0.336939804 -0.407292495  0.079151419
```

Để kiểm tra các giả định trên, chúng ta có thể vẽ một loạt 4 đồ thị mà tôi sẽ giải thích sau đây:

```
> op <- par(mfrow=c(2, 2)) # yêu cầu R dành ra 4 cửa sổ
> plot(reg) # vẽ các đồ thị trong reg
```



(a) Đồ thị bên trái dòng 1 vẽ phần dư e_i và giá trị tiên đoán cholesterol \hat{y}_i . Đồ thị này cho thấy các giá trị phần dư tập chung quanh đường $y = 0$, cho nên giả định (c), hay e_i có giá trị trung bình 0, là có thể chấp nhận được.

(b) Đồ thị bên phải dòng 1 vẽ giá trị phần dư và giá trị kì vọng dựa vào phân phối chuẩn. Chúng ta thấy các số phần dư tập trung rất gần các giá trị trên đường chuẩn, và do đó, giả định (b), tức e_i phân phối theo luật phân phối chuẩn, cũng có thể đáp ứng.

(c) Đồ thị bên trái dòng 2 vẽ căn số phần dư chuẩn (standardized residual) và giá trị của \hat{y}_i . Đồ thị này cho thấy không có gì khác nhau giữa các số phần dư chuẩn cho các giá trị

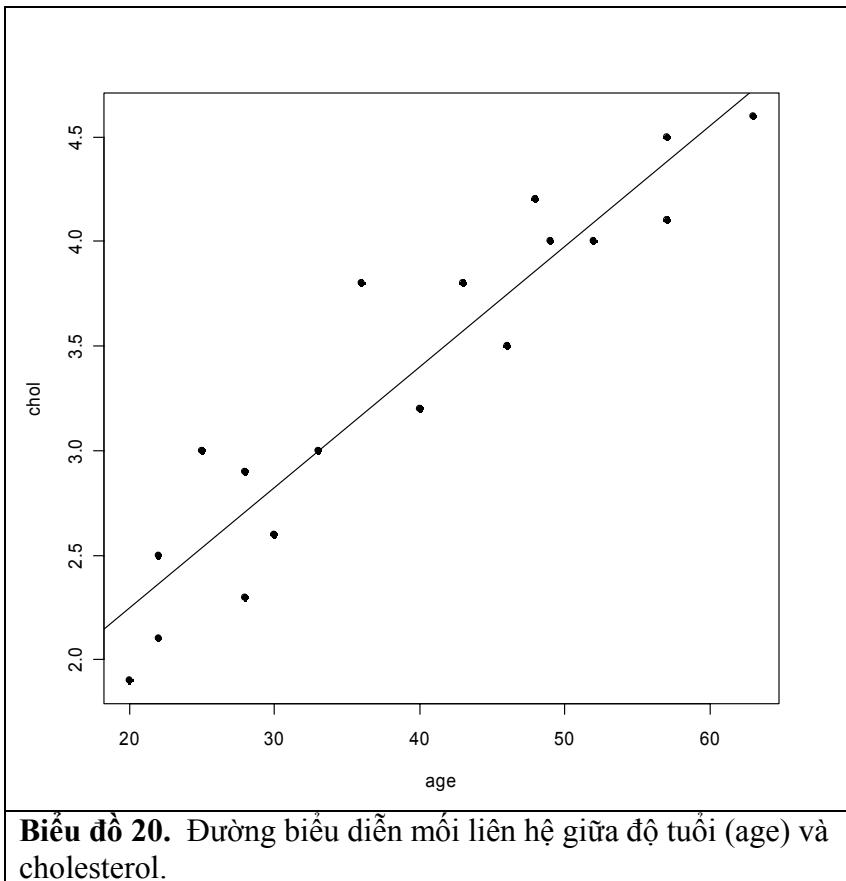
của \hat{y}_i , và do đó, giả định (d), tức ε_i có phương sai σ^2 cố định cho tất cả x_i , cũng có thể đáp ứng.

Nói chung qua phân tích phần dư, chúng ta có thể kết luận rằng mô hình hồi qui tuyến tính mô tả mối liên hệ giữa độ tuổi và cholesterol một cách khá đầy đủ và hợp lí.

Mô hình tiên đoán

Sau khi mô hình tiên đoán cholesterol đã được kiểm tra và tính hợp lí đã được thiết lập, chúng ta có thể vẽ đường biểu diễn của mối liên hệ giữa độ tuổi và cholesterol bằng lệnh `abline` như sau (xin nhắc lại object của phân tích là `reg`):

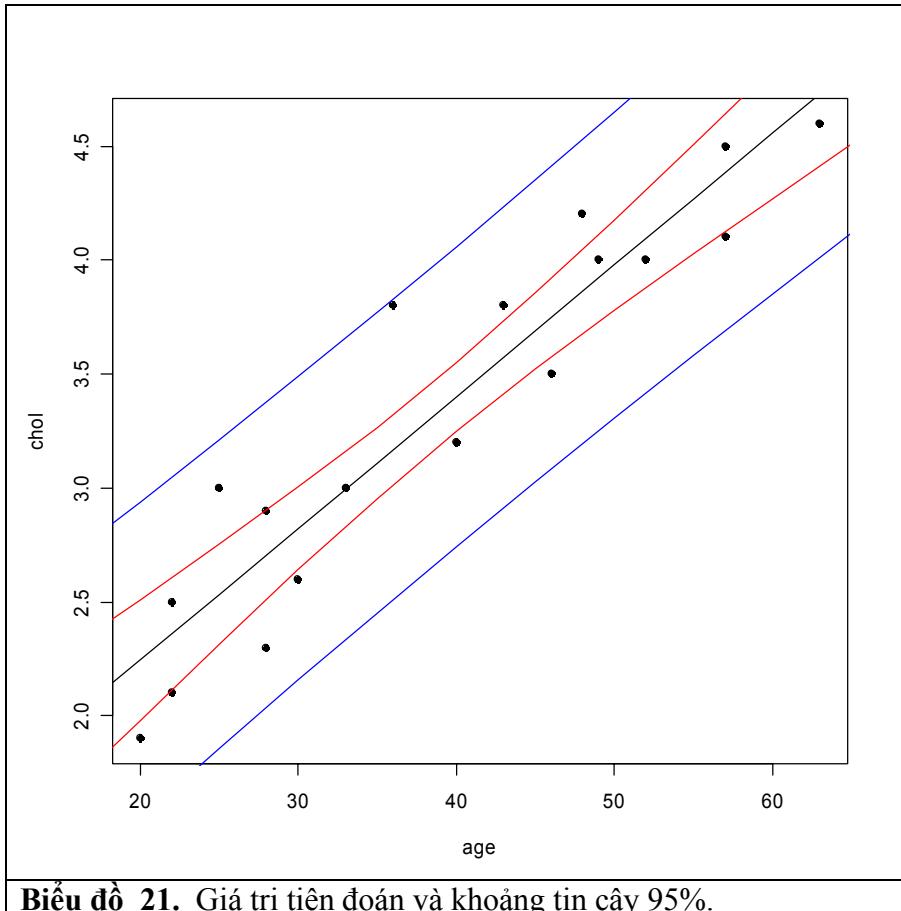
```
> plot(chol ~ age, pch=16)
> abline(reg)
```



Nhưng mỗi giá trị \hat{y}_i được tính từ ước số $\hat{\alpha}$ và $\hat{\beta}$, mà các ước số này đều có sai số chuẩn, cho nên giá trị tiên đoán \hat{y}_i cũng có sai số. Nói cách khác, \hat{y}_i chỉ là trung bình,

nhưng trong thực tế có thể cao hơn hay thấp hơn tùy theo chọn mẫu. Khoảng tin cậy 95% này có thể ước tính qua R bằng các lệnh sau đây:

```
> reg <- lm(chol ~ age)
> new <- data.frame(age = seq(15, 70, 5))
> pred.w.plim <- predict.lm(reg, new, interval="prediction")
> pred.w.clim <- predict.lm(reg, new, interval="confidence")
> resc <- cbind(pred.w.clim, new)
> resp <- cbind(pred.w.plim, new)
> plot(chol ~ age, pch=16)
> lines(resc$fit ~ resc$age)
> lines(resc$lwr ~ resc$age, col=2)
> lines(resc$upr ~ resc$age, col=2)
> lines(resp$lwr ~ resp$age, col=4)
> lines(resp$upr ~ resp$age, col=4)
```



Biểu đồ 21. Giá trị tiên đoán và khoảng tin cậy 95%.

Biểu đồ trên vẽ giá trị tiên đoán trung bình \hat{y}_i (đường thẳng màu đen), và khoảng tin cậy 95% của giá trị này là đường màu đỏ. Ngoài ra, đường màu xanh là khoảng tin cậy của giá trị tiên đoán cholesterol cho một độ tuổi mới trong quần thể.

10.3 Mô hình hồi qui tuyến tính đa biến (multiple linear regression)

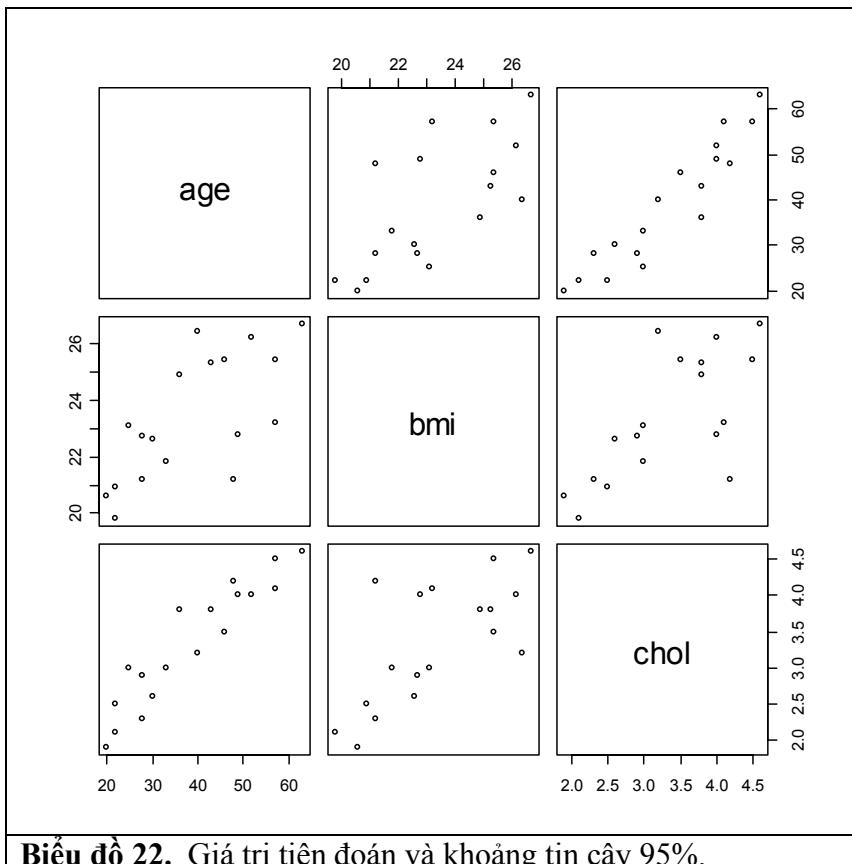
Mô hình được diễn đạt qua phương trình $y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ki} + \varepsilon_i$ có một yếu tố duy nhất (đó là x), và vì thế thường được gọi là mô hình hồi qui tuyến tính đơn giản (simple linear regression model). Trong thực tế, chúng ta có thể phát triển mô hình này thành nhiều biến, chứ không chỉ giới hạn một biến như trên, chẳng hạn như:

$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ki} + \varepsilon_i$$

Chú ý trong phương trình trên, chúng ta có nhiều biến x (x_1, x_2, \dots đến x_k), và mỗi biến có một thông số β_j ($j = 1, 2, \dots, k$) cần phải ước tính. Vì thế mô hình này còn được gọi là mô hình hồi qui tuyến tính đa biến.

Ví dụ 16. Chúng ta quay lại nghiên cứu về mối liên hệ giữa độ tuổi, bmi và cholesterol. Trong ví dụ, chúng ta chỉ mới xét mối liên hệ giữa độ tuổi và cholesterol, mà chưa xem đến mối liên hệ giữa cả hai yếu tố độ tuổi và bmi và cholesterol. Biểu đồ sau đây cho chúng ta thấy mối liên hệ giữa ba biến số này:

```
> pairs(data)
```



Biểu đồ 22. Giá trị tiên đoán và khoảng tin cậy 95%.

Cũng như giữa độ tuổi và cholesterol, mối liên hệ giữa bmi và cholesterol cũng gần tuân theo một đường thẳng. Biểu đồ trên còn cho chúng ta thấy độ tuổi và bmi có liên hệ với

nhau. Thật vậy, phân tích hồi qui tuyến tính đơn giản giữa bmi và cholesterol cho thấy như mối liên hệ này có ý nghĩa thống kê:

```
> summary(lm(chol ~ bmi))

Call:
lm(formula = chol ~ bmi)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.9403 -0.3565 -0.1376  0.3040  1.4330 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -2.83187   1.60841  -1.761  0.09739 .  
bmi         0.26410   0.06861   3.849  0.00142 ** 
---
Signif. codes:  0 '****' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.623 on 16 degrees of freedom
Multiple R-Squared: 0.4808,    Adjusted R-squared: 0.4483 
F-statistic: 14.82 on 1 and 16 DF,  p-value: 0.001418
```

BMI giải thích khoảng 48% độ dao động về cholesterol giữa các cá nhân. Nhưng vì BMI cũng có liên hệ với độ tuổi, chúng ta muốn biết nếu hai yếu tố này được phân tích cùng một lúc thì yếu tố nào quan trọng hơn. Để biết ảnh hưởng của cả hai yếu tố age (x_1) và bmi (tạm gọi là x_2) đến cholesterol (y) qua một mô hình hồi qui tuyến tính đa biến, và mô hình đó là:

$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$$

hay phương trình cũng có thể mô tả bằng kí hiệu ma trận: $\mathbf{Y} = \mathbf{X}\beta + \boldsymbol{\varepsilon}$ mà tôi vừa trình bày trên. Ở đây, \mathbf{Y} là một vector vector 18×1 , \mathbf{X} là một matrix 18×2 phần tử, β và một vector 2×1 , và $\boldsymbol{\varepsilon}$ là vector gồm 18×1 phần tử. Để ước tính hai hệ số hồi qui, β_1 và β_2 chúng ta cũng ứng dụng hàm `lm()` trong R như sau:

```
> mreg <- lm(chol ~ age + bmi)
> summary(mreg)

Call:
lm(formula = chol ~ age + bmi)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.3762 -0.2259 -0.0534  0.1698  0.5679 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 0.455458   0.918230   0.496   0.627
```

```

age          0.054052   0.007591   7.120 3.50e-06 ***
bmi         0.033364   0.046866   0.712      0.487
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

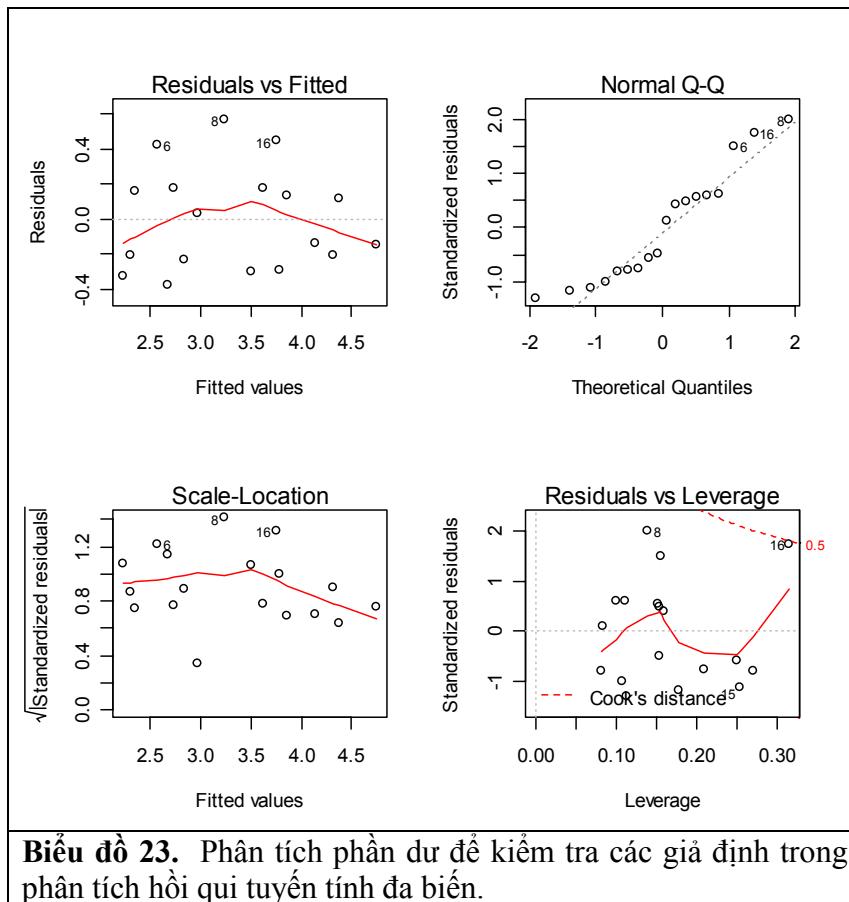
Residual standard error: 0.3074 on 15 degrees of freedom
 Multiple R-Squared: 0.8815, Adjusted R-squared: 0.8657
 F-statistic: 55.77 on 2 and 15 DF, p-value: 1.132e-07

Kết quả phân tích trên cho thấy ước số $\hat{\alpha} = 0.455$, $\hat{\beta}_1 = 0.054$ và $\hat{\beta}_2 = 0.0333$. Nói cách khác, chúng ta có phương trình ước đoán độ cholesterol dựa vào hai biến số độ tuổi và bmi như sau:

$$\text{Cholesterol} = 0.455 + 0.054(\text{age}) + 0.0333(\text{bmi})$$

Phương trình cho biết khi độ tuổi tăng 1 năm thì cholesterol tăng 0.054 mg/L (ước số này không khác mấy so với 0.0578 trong phương trình chỉ có độ tuổi), và mỗi 1 kg/m² tăng BMI thì cholesterol tăng 0.0333 mg/L. Hai yếu tố này “giải thích” khoảng 88.2% ($R^2 = 0.8815$) độ dao động của cholesterol giữa các cá nhân.

Chúng ta chú ý phương trình với độ tuổi (trong phân tích phần trước) giải thích khoảng 87.7% độ dao động cholesterol giữa các cá nhân. Khi chúng ta thêm yếu tố BMI, hệ số này tăng lên 88.2%, tức chỉ 0.5%. Câu hỏi đặt ra là 0.5% tăng trưởng này có ý nghĩa thống kê hay không. Câu trả lời có thể xem qua kết quả kiểm định yếu tố bmi với trị số p = 0.487. Như vậy, bmi không cung cấp cho chúng thêm thông tin hay tiên đoán cholesterol hơn những gì chúng ta đã có từ độ tuổi. Nói cách khác, khi độ tuổi đã được xem xét, thì ảnh hưởng của bmi không còn ý nghĩa thống kê. Điều này có thể hiểu được, bởi vì qua Biểu đồ 10.5 chúng ta thấy độ tuổi và bmi có một mối liên hệ khá cao. Vì hai biến này có tương quan với nhau, chúng ta không cần cả hai trong phương trình. (Tuy nhiên, ví dụ này chỉ có tính cách minh họa cho việc tiến hành phân tích hồi qui tuyến tính đa biến bằng R, chứ không có ý định mô phỏng dữ liệu theo định hướng sinh học).



Tuy BMI không có ý nghĩa thống kê trong trường hợp này, **Biểu đồ 10.6** cho thấy các giả định về mô hình hồi qui tuyến tính có thể đáp ứng.

11. Phân tích phương sai

11.1 Phân tích phương sai đơn giản (one-way analysis of variance - ANOVA)

Ví dụ 17. Bảng dưới đây so sánh độ galactose trong 3 nhóm bệnh nhân: nhóm 1 gồm 9 bệnh nhân với bệnh Crohn; nhóm 2 gồm 11 bệnh nhân với bệnh viêm ruột kết (colitis); và nhóm 3 gồm 20 đối tượng không có bệnh (gọi là nhóm đối chứng). Câu hỏi đặt ra là độ galactose giữa 3 nhóm bệnh nhân có khác nhau hay không?

Độ galactose cho 3 nhóm bệnh nhân Crohn, viêm ruột kết và đối chứng

Nhóm 1: bệnh Crohn	Nhóm 2: bệnh viêm ruột kết	Nhóm 3: đối chứng (control)
--------------------	----------------------------	-----------------------------

1343	1264	1809 2850
1393	1314	1926 2964
1420	1399	2283 2973
1641	1605	2384 3171
1897	2385	2447 3257
2160	2511	2479 3271
2169	2514	2495 3288
2279	2767	2525 3358
2890	2827	2541 3643
	2895	2769 3657
	3011	
$n=9$ Trung bình: 1910 SD: 516	$n=11$ Trung bình: 2226 SD: 727	$n=20$ Trung bình: 2804 SD: 527

Chú thích: SD là độ lệch chuẩn (standard deviation).

Gọi giá trị trung bình của ba nhóm là μ_1 , μ_2 , và μ_3 , và nói theo ngôn ngữ của kiểm định giả thiết thì giả thiết đảo là:

$$H_0: \mu_1 = \mu_2 = \mu_3$$

Và giả thiết chính là:

$$H_A: \text{có một khác biệt giữa } 3 \mu_j \ (j = 1, 2, 3)$$

Thoạt đầu có lẽ bạn đọc, sau khi đã học qua phương pháp so sánh hai nhóm bằng kiểm định t, sẽ nghĩ rằng chúng ta cần làm 3 so sánh bằng kiểm định t: giữa nhóm 1 và 2, nhóm 2 và 3, và nhóm 1 và 3. Nhưng phương pháp này không hợp lý, vì có ba phương sai khác nhau. Phương pháp thích hợp cho so sánh là phân tích phương sai. Phân tích phương sai có thể ứng dụng để so sánh nhiều nhóm cùng một lúc (simultaneous comparisons).

Để minh họa cho phương pháp phân tích phương sai, chúng ta phải dùng kí hiệu. Gọi độ galactose của bệnh nhân i thuộc nhóm j ($j = 1, 2, 3$) là x_{ij} . Mô hình phân tích phương sai phát biểu rằng:

$$x_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

Hay cụ thể hơn:

$$x_{i1} = \mu + \alpha_1 + \varepsilon_{i1}$$

$$x_{i2} = \mu + \alpha_2 + \varepsilon_{i2}$$

$$x_{i3} = \mu + \alpha_3 + \varepsilon_{i3}$$

Trước hết, chúng ta cần phải nhập dữ liệu vào R. Bước thứ nhất là báo cho R biết rằng chúng ta có ba nhóm bệnh nhân (1, 2 và 3), nhóm 1 gồm 9 người, nhóm 2 có 11 người, và nhóm 3 có 20 người:

```
> group <- c(1,1,1,1,1,1,1,1,1, 2,2,2,2,2,2,2,2,2,2,2,  
3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3)
```

Để phân tích phương sai, chúng ta phải định nghĩa biến group là một yếu tố - factor.

```
> group <- as.factor(group)
```

Bước kế tiếp, chúng ta nạp số liệu galactose cho từng nhóm như định nghĩa trên (gọi object là galactose):

```
> galactose <- c(1343,1393,1420,1641,1897,2160,2169,2279,2890,
1264,1314,1399,1605,2385,2511,2514,2767,2827,2895,3011,
1809,2850,1926,2964,2283,2973,2384,3171,2447,3257,2479,3271,2495,3288,
2525,3358,2541,3643,2769,3657)
```

Đưa hai biến group và galactose vào một dataframe và gọi là data:

```
> data <- data.frame(group, galactose)
> attach(data)
```

Sau khi đã có dữ liệu sẵn sàng, chúng ta dùng hàm lm() để phân tích phương sai như sau:

```
> analysis <- lm(galactose ~ group)
```

Trong hàm trên chúng ta cho R biết biến galactose là một hàm số của group. Gọi kết quả phân tích là analysis.

Kết quả phân tích phương sai. Nay giờ chúng ta dùng lệnh anova để biết kết quả phân tích:

```
> anova(analysis)
Analysis of Variance Table

Response: galactose
          Df  Sum Sq Mean Sq F value    Pr(>F)
group      2  5683620  2841810   8.6655 0.0008191 ***
Residuals 37 12133923   327944
---
Signif. codes:  0 '****' 0.001 '***' 0.01 '**' 0.05 '*' 0.1 '.' 1
```

Trong kết quả trên, có ba cột: Df (degrees of freedom) là bậc tự do; Sum Sq là tổng bình phương (sum of squares), Mean Sq là trung bình bình phương (mean square); F value là giá trị F; và Pr (>F) là trị số P liên quan đến kiểm định F.

11.2 So sánh nhiều nhóm (multiple comparisons) và điều chỉnh trị số p

Cho k nhóm, chúng ta có ít nhất là $k(k-1)/2$ so sánh. Ví dụ trên có 3 nhóm, cho nên tổng số so sánh khả dĩ là 3 (giữa nhóm 1 và 2, nhóm 1 và 3, và nhóm 2 và 3). Khi $k=10$, số lần so sánh có thể lên rất cao. Như đã đề cập trong chương 7, khi có nhiều so sánh, trị số p tính toán từ các kiểm định thống kê không còn ý nghĩa ban đầu nữa, bởi vì các kiểm định này có thể cho ra kết quả dương tính giả (tức kết quả với $p<0.05$ nhưng

trong thực tế không có khác nhau hay ảnh hưởng). Do đó, trong trường hợp có nhiều so sánh, chúng ta cần phải điều chỉnh trị số p sao cho hợp lý.

Có khá nhiều phương pháp điều chỉnh trị số p, và 4 phương pháp thông dụng nhất là: Bonferroni, Scheffé, Holm và Tukey (tên của 4 nhà thống kê học danh tiếng). Phương pháp nào thích hợp nhất? Không có câu trả lời dứt khoát cho câu hỏi này, nhưng hai điểm sau đây có thể giúp bạn đọc quyết định tốt hơn:

- (a) Nếu $k < 10$, chúng ta có thể áp dụng bất cứ phương pháp nào để điều chỉnh trị số p. Riêng cá nhân tôi thì thấy phương pháp Tukey thường rất hữu ích trong so sánh.
- (b) Nếu $k > 10$, phương pháp Bonferroni có thể trở nên rất “bảo thủ”. Bảo thủ ở đây có nghĩa là phương pháp này rất ít khi nào tuyên bố một so sánh có ý nghĩa thống kê, dù trong thực tế là có thật! Trong trường hợp này, hai phương pháp Tukey, Holm và Scheffé có thể áp dụng.

Quay lại ví dụ trên, các trị số p trên đây là những trị số chưa được điều chỉnh cho so sánh nhiều lần. Trong chương về trị số p, tôi đã nói các trị số này phóng đại ý nghĩa thống kê, không phản ánh trị số p lúc ban đầu (tức 0.05). Để điều chỉnh cho nhiều so sánh, chúng ta phải sử dụng đến phương pháp điều chỉnh Bonferroni.

Chúng ta có thể dùng lệnh `pairwise.t.test` để có được tất cả các trị số p so sánh giữa ba nhóm như sau:

```
> pairwise.t.test(galactose, group, p.adj="bonferroni")
Pairwise comparisons using t tests with pooled SD

data: galactose and group

  1      2
2 0.6805 -
3 0.0012 0.0321

P value adjustment method: bonferroni
```

Kết quả trên cho thấy trị số p giữa nhóm 1 (Crohn) và viêm ruột kết là 0.6805 (tức không có ý nghĩa thống kê); giữa nhóm Crohn và đối chứng là 0.0012 (có ý nghĩa thống kê), và giữa nhóm viêm ruột kết và đối chứng là 0.0321 (tức cũng có ý nghĩa thống kê).

Một phương pháp điều chỉnh trị số p khác có tên là phương pháp Holm:

```
> pairwise.t.test(galactose, group)
Pairwise comparisons using t tests with pooled SD

data: galactose and group
```

```

1      2
2 0.2268 -
3 0.0012 0.0214

```

P value adjustment method: holm

Kết quả này cũng không khác so với phương pháp Bonferroni.

Tất cả các phương pháp so sánh trên sử dụng một sai số chuẩn chung cho cả ba nhóm. Nếu chúng ta muốn sử dụng cho từng nhóm thì lệnh sau đây (pool.sd=F) sẽ đáp ứng yêu cầu đó:

```
> pairwise.t.test(galactose, group, pool.sd=FALSE)
```

Pairwise comparisons using t tests with non-pooled SD

data: galactose and group

```

1      2
2 0.2557 -
3 0.0017 0.0544

```

P value adjustment method: holm

Một lần nữa, kết quả này cũng không làm thay đổi kết luận.

Trong các phương pháp trên, chúng ta chỉ biết trị số p so sánh giữa các nhóm, nhưng không biết mức độ khác biệt cũng như khoảng tin cậy 95% giữa các nhóm. Để có những ước số này, chúng ta cần đến một hàm khác có tên là aov (viết tắt từ analysis of variance) và hàm TukeyHSD (HSD là viết tắt từ Honest Significant Difference, tạm dịch nôm na là “Khác biệt có ý nghĩa thành thật”) như sau:

```

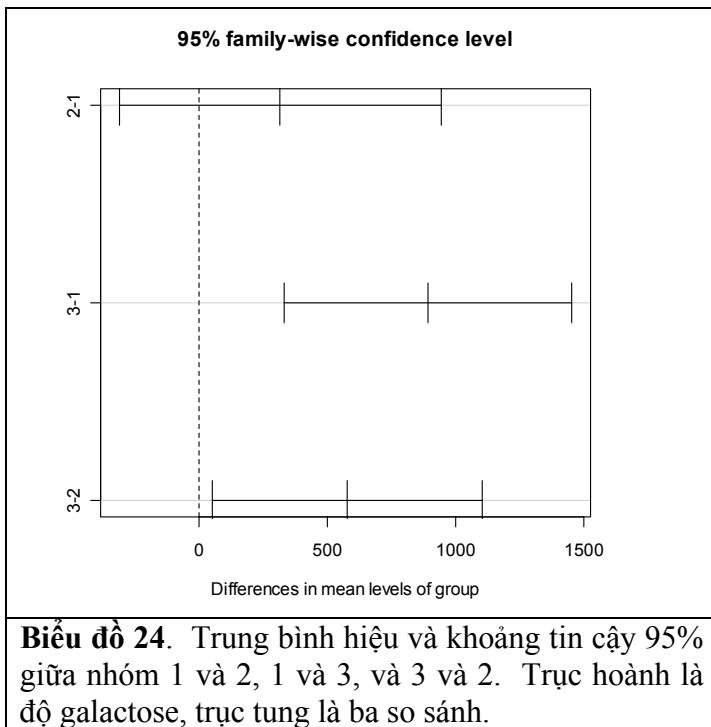
> res <- aov(galactose ~ group)
> TukeyHSD(res)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = galactose ~ group)

$group
     diff      lwr      upr      p adj
2-1 316.3232 -312.09857 944.745 0.4439821
3-1 894.2778  333.07916 1455.476 0.0011445
3-2 577.9545   53.11886 1102.790 0.0281768

```

Kết quả trên cho chúng ta thấy nhóm 3 và 1 khác nhau khoảng 894 đơn vị, và khoảng tin cậy 95% từ 333 đến 1455 đơn vị. Tương tự, galactose trong nhóm bệnh nhân viêm ruột kết thấp hơn nhóm đối chứng (nhóm 3) khoảng 578 đơn vị, và khoảng tin cậy 95% từ 53 đến 1103.



Biểu đồ 24. Trung bình hiệu và khoảng tin cậy 95% giữa nhóm 1 và 2, 1 và 3, và 3 và 2. Trục hoành là độ galactose, trục tung là ba so sánh.

11.3 Phân tích bằng phương pháp phi tham số

Phương pháp so sánh nhiều nhóm phi tham số (non-parametric statistics) tương đương với phương pháp phân tích phương sai là Kruskal-Wallis. Cũng như phương pháp Wilcoxon so sánh hai nhóm theo phương pháp phi tham số, phương pháp Kruskal-Wallis cũng biến đổi số liệu thành thứ bậc (ranks) và phân tích độ khác biệt thứ bậc này giữa các nhóm. Hàm `kruskal.test` trong R có thể giúp chúng ta trong kiểm định này:

```
> kruskal.test(galactose ~ group)

Kruskal-Wallis rank sum test

data: galactose by group
Kruskal-Wallis chi-squared = 12.1381, df = 2, p-value = 0.002313
```

Trị số p từ kiểm định này khá thấp ($p = 0.002313$) cho thấy có sự khác biệt giữa ba nhóm như phân tích phương sai qua hàm `lm` trên đây. Tuy nhiên, một bất tiện của kiểm định phi tham số Kruskal-Wallis là phương pháp này không cho chúng ta biết hai nhóm nào khác nhau, mà chỉ cho một trị số p chung. Trong nhiều trường hợp, phân tích phi tham số như kiểm định Kruskal-Wallis thường không có hiệu quả như các phương pháp thống kê tham số (parametric statistics).

11.4 Phân tích phương sai hai chiều (two-way analysis of variance - ANOVA)

Phân tích phương sai đơn giản hay một chiều chỉ có một yếu tố (factor). Nhưng phân tích phương sai hai chiều (two-way ANOVA), như tên gọi, có hai yếu tố. Phương pháp phân tích phương sai hai chiều chỉ đơn giản khai triển từ phương pháp phân tích phương sai đơn giản. Thay vì ước tính phương sai của một yếu tố, phương pháp phân sai hai chiều ước tính phương sai của hai yếu tố.

Ví dụ 18. Trong ví dụ sau đây, để đánh giá hiệu quả của một kỹ thuật sơn mới, các nhà nghiên cứu áp dụng sơn trên 3 loại vật liệu (1, 2 và 3) trong hai điều kiện (1, 2). Mỗi điều kiện và loại vật liệu, nghiên cứu được lặp lại 3 lần. Độ bền được đo là chỉ số bền bỉ (tạm gọi là score). Tổng cộng, có 18 số liệu như sau:

Độ bền bỉ của sơn cho 2 điều kiện và 3 vật liệu

Điều kiện (i)	Vật liệu (j)		
	1	2	3
1	4.1, 3.9, 4.3	3.1, 2.8, 3.3	3.5, 3.2, 3.6
2	2.7, 3.1, 2.6	1.9, 2.2, 2.3	2.7, 2.3, 2.5

Gọi x_{ij} là score của điều kiện i ($i = 1, 2$) cho vật liệu j ($j = 1, 2, 3$). (Để đơn giản hóa vấn đề, chúng ta tạm thời bỏ qua k đối tượng). Mô hình phân tích phương sai hai chiều phát biểu rằng:

$$x_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$$

μ là số trung bình cho toàn quần thể, các hệ số α_i (ảnh hưởng của điều kiện i) và β_j (ảnh hưởng của vật liệu j) cần phải ước tính từ số liệu thực tế. ε_{ij} được giả định tuân theo luật phân phối chuẩn với trung bình 0 và phương sai σ^2 .

Để phân tích bằng R, chúng ta cần phải tổ chức dữ liệu sao cho có 4 biến như sau:

Condition (điều kiện)	Material (vật liệu)	Đối tượng	Score
1	1	1	4.1
1	1	2	3.9
1	1	3	4.3
1	2	4	3.1
1	2	5	2.8
1	2	6	3.3
1	3	7	3.5
1	3	8	3.2
1	3	9	3.6
2	1	10	2.7
2	1	11	3.1
2	1	12	2.6
2	2	13	1.9
2	2	14	2.2
2	2	15	2.3

2	3	16	2.7
2	3	17	2.3
2	3	18	2.5

Chúng ta có thể tạo ra một dãy số bằng cách sử dụng hàm `gl` (generating levels).

```
> condition <- gl(2, 9, 18)
> material <- gl(3, 3, 18)
```

Và tạo nên 18 mã số (từ 1 đến 18):

```
> id <- 1:18
```

Sau cùng là số liệu cho `score`:

```
> score <- c(4.1, 3.9, 4.3, 3.1, 2.8, 3.3, 3.5, 3.2, 3.6,
  2.7, 3.1, 2.6, 1.9, 2.2, 2.3, 2.7, 2.3, 2.5)
```

Tất cả cho vào một dataframe tên là `data`:

```
> data <- data.frame(condition, material, id, score)
> attach(data)
```

Bây giờ số liệu đã sẵn sàng cho phân tích. Để phân tích phương sai hai chiều, chúng ta vẫn sử dụng lệnh `lm` với các thông số như sau:

```
> twoway <- lm(score ~ condition + material)
> anova(twoway)
Analysis of Variance Table

Response: score
          Df Sum Sq Mean Sq F value    Pr(>F)
condition   1 5.0139  5.0139  95.575 1.235e-07 ***
material   2 2.1811  1.0906  20.788 6.437e-05 ***
Residuals 14 0.7344  0.0525
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Ba nguồn dao động (variation) của `score` được phân tích trong bảng trên. Qua trung bình bình phương (mean square), chúng ta thấy ảnh hưởng của điều kiện có vẻ quan trọng hơn là ảnh hưởng của vật liệu thí nghiệm. Tuy nhiên, cả hai ảnh hưởng đều có ý nghĩa thống kê, vì trị số p rất thấp cho hai yếu tố. Chúng ta yêu cầu R tóm lược các ước số phân tích bằng lệnh `summary`:

```
> summary(twoway)

Call:
lm(formula = score ~ condition + material)

Residuals:
      Min       1Q   Median       3Q      Max 
-0.32778 -0.16389  0.03333  0.16111  0.32222
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	3.9778	0.1080	36.841	2.43e-15 ***
condition2	-1.0556	0.1080	-9.776	1.24e-07 ***
material2	-0.8500	0.1322	-6.428	1.58e-05 ***
material3	-0.4833	0.1322	-3.655	0.0026 **

Signif. codes:	0 '***'	0.001 '**'	0.01 '*'	0.05 '.'
	0.1	'	'	1

Residual standard error: 0.229 on 14 degrees of freedom
 Multiple R-Squared: 0.9074, Adjusted R-squared: 0.8875
 F-statistic: 45.72 on 3 and 14 DF, p-value: 1.761e-07

Kết quả trên cho thấy so với điều kiện 1, điều kiện 2 có score thấp hơn khoảng 1.056 và sai số chuẩn là 0.108, với trị số $p = 1.24e-07$, tức có ý nghĩa thống kê. Ngoài ra, so với vật liệu 1, score cho vật liệu 2 và 3 cũng thấp hơn đáng kể với độ thấp nhất ghi nhận ở vật liệu 2, và ảnh hưởng của vật liệu thí nghiệm cũng có ý nghĩa thống kê.

Giá trị có tên là “Residual standard error” được ước tính từ trung bình bình phương phần dư trong phần (a), tức là $\sqrt{0.0525} = 0.229$, tức là ước số của $\hat{\sigma}$.

Hệ số xác định bội (R^2) cho biết hai yếu tố điều kiện và vật liệu giải thích khoảng 91% độ dao động của toàn bộ mẫu. Hệ số này được tính từ tổng bình phương trong kết quả phần (a) như sau:

$$R^2 = \frac{5.0139 + 2.1811}{5.0139 + 2.1811 + 0.7344} = 0.9074$$

Và sau cùng, hệ số R^2 điều chỉnh phản ánh độ “cái tiến” của mô hình. Để hiểu hệ số này tốt hơn, chúng ta thấy phương sai của toàn bộ mẫu là $s^2 = (5.0139 + 2.1811 + 0.7344) / 17 = 0.4644$. Sau khi điều chỉnh cho ảnh hưởng của điều kiện và vật liệu, phương sai này còn 0.0525 (tức là residual mean square). Như vậy hai yếu tố này làm giảm phương sai khoảng $0.4644 - 0.0525 = 0.4119$. Và hệ số R^2 điều chỉnh là:

$$\text{Adj } R^2 = 0.4119 / 0.4644 = 0.88$$

Tức là sau khi điều chỉnh cho hai yếu tố điều kiện và vật liệu phương sai của score giảm khoảng 88%.

So sánh giữa các nhóm. Chúng ta sẽ ước tính độ khác biệt giữa hai điều kiện và ba vật liệu bằng hàm TukeyHSD với aov:

```
> res <- aov(score ~ condition + material + condition)
> TukeyHSD(res)
Tukey multiple comparisons of means
 95% family-wise confidence level
```

```
Fit: aov(formula = score ~ condition + material + condition)
```

```
$condition
```

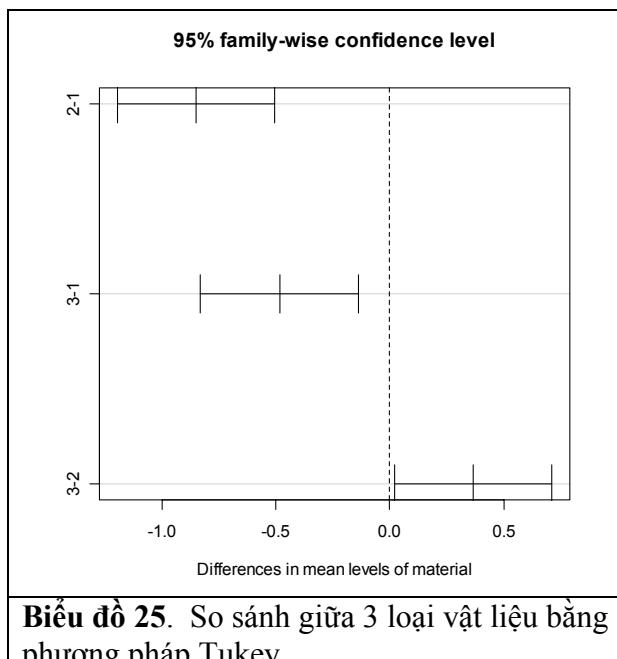
	diff	lwr	upr	p	adj
2-1	-1.055556	-1.287131	-0.8239797	1e-07	

```
$material
```

	diff	lwr	upr	p	adj
2-1	-0.8500000	-1.19610279	-0.5038972	0.0000442	
3-1	-0.4833333	-0.82943612	-0.1372305	0.0068648	
3-2	0.3666667	0.02056388	0.7127695	0.0374069	

Biểu đồ sau đây sẽ minh họa cho các kết quả trên:

```
> plot(TukeyHSD(res), ordered=TRUE)
There were 16 warnings (use warnings() to see them)
```



12. Phân tích hồi qui logistic

Trong các phần trước về phân tích hồi qui tuyến tính và phân tích phương sai, chúng ta tìm mô hình và mối liên hệ giữa một biến phụ thuộc liên tục (continuous dependent variable) và một hay nhiều biến độc lập (independent variable) hoặc là liên tục hoặc là không liên tục. Nhưng trong nhiều trường hợp, biến phụ thuộc không phải là biến liên tục mà là biến mang tính đo lường nhị phân: có/không, mắc bệnh/không mắc bệnh, chết/sống, xảy ra/không xảy ra, v.v..., còn các biến độc lập có thể là liên tục hay không liên tục. Chúng ta cũng muốn tìm hiểu mối liên hệ giữa các biến độc lập và biến phụ thuộc.

Ví dụ 19. Trong một nghiên cứu do tôi tiến hành để tìm hiểu mối liên hệ giữa nguy cơ gãy xương (fracture, viết tắt là fx) và mật độ xương cùng một số chỉ số sinh hóa khác, 139 bệnh nhân nam (hay nói đúng hơn là đối tượng nghiên cứu) tuổi từ 60 trở lên. Năm 1990, các số liệu sau đây được thu thập cho mỗi đối tượng: độ tuổi (age), tỉ trọng cơ thể (body mass index hay BMI), mật độ chất khoáng trong xương (bone mineral density hay BMD), chỉ số hủy xương ICTP, chỉ số tạo xương PINP. Các đối tượng nghiên cứu được theo dõi trong vòng 15 năm. Trong thời gian theo dõi, các bệnh nhân bị gãy xương hay không gãy xương được ghi nhận. Câu hỏi đặt ra ban đầu là có một mối liên hệ gì giữa BMD và nguy cơ gãy xương hay không. Số liệu của nghiên cứu này được trình bày trong phần cuối của chương này, và sẽ trình bày một phần dưới đây để bạn đọc nắm được vấn đề.

Một phần số liệu nghiên cứu về các yếu tố nguy cơ cho gãy xương

id	fx	age	bmi	bmd	ictp	pinp
1	1	79	24.7252	0.818	9.170	37.383
2	1	89	25.9909	0.871	7.561	24.685
3	1	70	25.3934	1.358	5.347	40.620
4	1	88	23.2254	0.714	7.354	56.782
5	1	85	24.6097	0.748	6.760	58.358
6	0	68	25.0762	0.935	4.939	67.123
7	0	70	19.8839	1.040	4.321	26.399
8	0	69	25.0593	1.002	4.212	47.515
9	0	74	25.6544	0.987	5.605	26.132
10	0	79	19.9594	0.863	5.204	60.267
...						
137	0	64	38.0762	1.086	5.043	32.835
138	1	80	23.3887	0.875	4.086	23.837
139	0	67	25.9455	0.983	4.328	71.334

Ở đây, vì biến phụ thuộc (gãy xương) không được đo lường theo tính liên tục (mà chỉ là *có* hay *không*), cho nên phương pháp phân tích hồi qui tuyến tính để phân tích mối liên hệ giữa biến phụ thuộc và biến độc lập. Một phương pháp phân tích được phát triển tương đối gần đây (vào thập niên 1970s) có tên là logistic regression analysis (hay phân tích hồi qui logistic) có thể áp dụng cho trường hợp trên.

Trong nghiên cứu này, sau 15 năm theo dõi, có 38 bệnh nhân bị gãy xương. Tính theo phần trăm, tỉ lệ gãy xương là $38 / 139 = 0.273$ (hay 27.3%).

12.1 Mô hình hồi qui logistic

Cho một tần số biến cỗ x ghi nhận từ n đối tượng, chúng ta có thể tính xác suất của biến cỗ đó là:

$$p = \frac{x}{n}$$

p có thể xem là một chỉ số đo lường nguy cơ của một biến cỗ. Một cách thể hiện nguy cơ khác là *odds* (một danh từ, nếu tôi không làm, chỉ có trong tiếng Anh – ngay cả tiếng Pháp, Đức, Tây Ban Nha ... cũng không có danh từ tương đương với *odds*). Tôi tạm dịch

odds là *khả năng*. *Khả năng* của một biến cõi được định nghĩa đơn giản bằng tỉ số xác suất biến cõi xảy ra trên xác suất biến cõi không xảy ra:

$$\text{odds} = \frac{p}{1-p}$$

Hàm *logit* của *odds* được định nghĩa như sau:

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right)$$

Cho một biến độc lập x (x có thể là liên tục hay không liên tục), mô hình hồi qui logistic phát biểu rằng:

$$\text{logit}(p) = \alpha + \beta x$$

Tương tự như mô hình hồi qui tuyến tính, α và β là hai thông số tuyến tính cần phải ước tính từ dữ liệu nghiên cứu. Nhưng ý nghĩa của thông số này, đặc biệt là thông số β , rất khác với ý nghĩa mà ta đã quen với mô hình hồi qui tuyến tính. Để hiểu ý nghĩa của hai thông số này, tôi sẽ quay lại với ví dụ 19.

Vấn đề mà chúng ta muốn biết là mối liên hệ giữa mật độ xương bmd và nguy cơ gãy xương (f_x). Để tiện cho việc minh họa, gọi bmd là x , vấn đề mà chúng ta cần biết có thể viết bằng ngôn ngữ mô hình như sau

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right)\alpha + \beta x$$

Nói cách khác:

$$\text{odds}(p) = \frac{p}{1-p} = e^{\alpha+\beta x}$$

Nói cách khác, mô hình hồi qui logistic vừa trình bày trên phát biểu rằng mối liên hệ giữa xác suất gãy xương (p) và mật độ xương bmd là một mối liên hệ theo hình chữ S. Mô hình trên còn cho thấy xác suất gãy xương p tùy thuộc vào giá trị của x . Thành ra, mô hình trên có thể viết một cách chính xác hơn rằng *khả năng* gãy xương với điều kiện x là:

$$\text{odds}(p | x) = e^{\alpha+\beta x}$$

Khi $x = x_0$, khả năng gãy xương là: $\text{odds}(p | x = x_0) = e^{\alpha+\beta x_0}$

Khi $x = x_0 + 1$ (tức tăng 1 đơn vị từ x_0), khả năng gãy xương là:

$$\text{odds}(p | x = x_0 + 1) = e^{\alpha+\beta(x_0+1)}$$

Và, tỉ số của hai xác suất gãy xương:

$$\frac{odds(p | x = x_0 + 1)}{odds(p | x = x_0)} = \frac{e^{\alpha + \beta(x_0 + 1)}}{e^{\alpha + \beta x_0}} = e^\beta$$

Trong dịch tễ học, e^β được gọi là *odds ratio*. *Odds ratio*, như tên gọi là, *tỉ số khả năng* hay *tỉ số khả dĩ*. Nói cách khác, hệ số β trong mô hình hồi qui logistic chính là tỉ số khả dĩ.

Phương pháp để ước tính thông số trong mô hình [3] khá phức tạp (dùng phương pháp maximum likelihood – tức phương pháp *Hợp lí cực đại*) và không nằm trong phạm vi của cuốn sách này, nên tôi sẽ không trình bày ở đây (bạn đọc có thể tham khảo sách giáo khoa để biết thêm, nếu cần thiết). Tuy nhiên, tôi muốn đề cập ngắn gọn là phương pháp hợp lí cực đại cung cấp cho chúng ta một hệ phương trình như sau:

$$\begin{cases} \sum_{i=1}^n y_i = \sum_{i=1}^n \left(1 + e^{-(\hat{\alpha} + \hat{\beta}x_i)}\right)^{-1} \\ \sum_{i=1}^n x_i y_i = \sum_{i=1}^n x_i \left(1 + e^{-(\hat{\alpha} + \hat{\beta}x_i)}\right) \end{cases}$$

Trong đó, Trong đó, y_i là biến phụ thuộc (gãy xương với giá trị 0 hay 1), và x_i là biến độc lập (mật độ xương), và n là số mẫu. Để tìm ước số $\hat{\alpha}$ và $\hat{\beta}$, một trong những phép tính hay sử dụng là iterative weighted least square hay Newton-Raphson. R sử dụng phép tính Newton-Raphson để tìm hai ước số đó.

Sau khi đã có ước số $\hat{\alpha}$ và $\hat{\beta}$ chúng ta có thể ước tính xác suất p cho bất cứ giá trị nào của x như sau (sau vài thao tác đại số):

$$\hat{p} = \frac{e^{\hat{\alpha} + \hat{\beta}x}}{1 + e^{\hat{\alpha} + \hat{\beta}x}} = \frac{1}{1 + e^{-(\hat{\alpha} + \hat{\beta}x)}}$$

Chú ý tôi dùng dấu mũ \hat{p} để chỉ số ước tính (predicted value), chứ không phải p là xác suất quan sát. Nếu mô hình mô tả dữ liệu tốt và đầy đủ, độ khác biệt giữa p và \hat{p} nhỏ; nếu mô hình không thích hợp hay không tốt, độ khác biệt đó có thể sẽ cao. Độ khác biệt giữa p và \hat{p} được gọi là *deviance*. Phương pháp tính deviance khá phức tạp, nhưng đó không phải là chủ đề ở đây, cho nên tôi chỉ nói qua khái niệm mà thôi. Khi chúng ta có nhiều mô hình để mô phỏng một hay nhiều mối liên hệ, deviance có thể được sử dụng để đánh giá sự thích hợp của một mô hình, hay để chọn một mô hình “tối ưu”.

12.2 Phân tích hồi qui logistic bằng R

Bây giờ, chúng ta quay lại với ví dụ 1, dùng số liệu trong Bảng 12.1 để ước tính hai thông số α và β bằng R. Trước hết chúng ta phải nhập toàn bộ số liệu vào một data

frame, và cho một cái tên, chẳng hạn như `fracture`. Trong trường hợp của tôi, dữ liệu được chứa trong directory c:\works\stats dưới tên `fracture.txt`, do đó, các lệnh sau đây cần thiết để nhập số liệu:

```
# báo cho R biết nơi chứa số liệu
> setwd("c:/works/stats")

# nhập số liệu và cho vào một data frame tên fracture
> fracture <- read.table("fracture.txt", header=TRUE, na.string=".")

# kiểm tra xem có bao nhiêu biến trong dữ liệu fracture
> names(fracture)
[1] "id"    "fx"    "age"   "bmi"   "bmd"   "ictp"  "pinp"

# Chọn những bệnh nhân có đầy đủ số liệu cho phân tích
> fulldata <- na.omit(fracture)
> attach(fulldata)
```

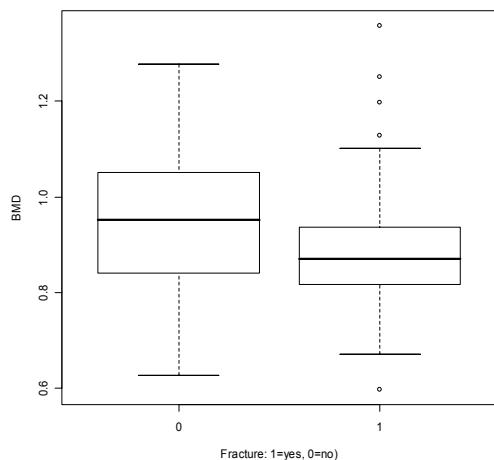
Hai biến mà chúng ta quan tâm trong ví dụ này là: `fx` (gãy xương) và `bmd` (mật độ xương). Chúng ta kiểm tra xem có bao nhiêu bệnh nhân gãy xương:

```
> table(fx)
fx
 0    1
101  38
```

Kế đến, xem mật độ xương trong nhóm gãy xương và không gãy xương ra sao:

```
> tapply(bmd, fx, mean)
0          1
0.9444851 0.9016667

> boxplot(bmd ~ fx,
  xlab="Fracture: 1=yes, 0=no",
  ylab="BMD")
```



Kết quả trên cho thấy, bmd trong nhóm bệnh nhân bị gãy xương thấp hơn so với nhóm không bị gãy xương (0.90 và 0.94). Và, kiểm định t sau đây cho thấy mức độ khác biệt này không có ý nghĩa thống kê ($p = 0.15$).

```
> t.test(bmd~fx)
```

```
Welch Two Sample t-test
```

```
data: bmd by fx
t = 1.4572, df = 53.952, p-value = 0.1508
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-0.01609226 0.10172922
sample estimates:
mean in group 0 mean in group 1
0.9444851 0.9016667
```

Để ước tính thông số trong mô hình [4], hàm số `glm` (viết tắt từ *generalized linear model*) trong R có thể áp dụng, với “cú pháp” như sau:

```
> logistic <- glm(fx ~ bmd, family="binomial")
> summary(logistic)
```

```
Call:
glm(formula = fx ~ bmd, family = "binomial")
```

```
Deviance Residuals:
```

Min	1Q	Median	3Q	Max
-1.0287	-0.8242	-0.7020	1.3780	2.0709

```
Coefficients:
```

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	1.063	1.342	0.792	0.428
bmd	-2.270	1.455	-1.560	0.119

```
(Dispersion parameter for binomial family taken to be 1)
```

```
Null deviance: 157.81 on 136 degrees of freedom
Residual deviance: 155.27 on 135 degrees of freedom
AIC: 159.27
```

Number of Fisher Scoring iterations: 4

Tôi sẽ lần lượt giải thích các kết quả trên:

- (a) Trong lệnh `logistic <- glm(fx ~ bmd, family="binomial")` chúng ta yêu cầu R phân tích theo mô hình `fx` là một hàm số với `bmd` như mô hình [4]. Trong `glm` có nhiều luật phân phối, mà trong đó phân phối nhị phân (`binomial`) là một luật phân phối chuẩn cho hồi qui logistic. Do đó, `family="binomial"` cần thiết cho R.
- (b) Deviance: phần thứ nhất của kết quả cho biết qua về deviance.

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.0287	-0.8242	-0.7020	1.3780	2.0709

Deviance như giải thích trên phản ánh độ khác biệt giữa mô hình và dữ liệu (cũng tương tự như mean square residual trong phân tích hồi qui tuyến tính vậy). Đối với một mô hình đơn lẻ như ví dụ này thì giá trị của deviance không có ý nghĩa gì nhiều.

- (c) Phần kế tiếp cung cấp ước số của $\hat{\alpha}$ (mà R đặt tên là `intercept`) và $\hat{\beta}$ (`bmd`) và sai số chuẩn (standard error).

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	1.063	1.342	0.792	0.428
bmd	-2.270	1.455	-1.560	0.119

Qua kết quả này, chúng ta có $\hat{\alpha} = 1.063$ và $\hat{\beta} = -2.27$. Ước số $\hat{\beta}$ là số âm cho thấy mối liên hệ giữa nguy cơ gãy xương và `bmd` là mối liên hệ nghịch đảo: xác suất gãy xương tăng khi giá trị của `bmd` giảm. Tuy nhiên, kiểm định z (tính bằng cách lấy ước số chia cho sai số chuẩn) cho chúng ta thấy ảnh hưởng của `bmd` không có ý nghĩa thống kê, vì trị số p = 0.119.

Nhớ rằng tỉ số khả dĩ (odds ratio hay viết tắt là OR) chính là $e^{-2.27} = 0.1033$. Nói cách khác, khi `bmd` tăng 1 g/cm^2 (đơn vị đo lường của `bmd` là g/cm^2) thì tỉ số OR giảm 0.9067 hay 90.67%. Nhưng tăng 1 g/cm^2 là mật độ rất cao trong xương và không thực tế. Cho nên một cách tính khác là tính trên độ lệch chuẩn (standard deviation) của `bmd`. Chúng ta sẽ tìm hiểu độ lệch chuẩn của `bmd`:

```
> sd(bmd)
[1] 0.1406543
```

Do đó, OR sẽ tính trên mỗi 0.14 g/cm^2 . Và OR cho mỗi độ lệch chuẩn, do đó, là:

$$e^{-2.27*0.1406} = 0.7267$$

Tức là, khi bmd tăng một độ lệch chuẩn thì tỉ số khả dĩ gãy xương giảm khoảng 28%. Cũng có thể nói cách khác, là khi bmd giảm một độ lệch chuẩn thì tỉ số khả dĩ tăng $e^{2.27*0.1406} = 1.376$ hay khoảng 38%.

Một cách khác để biết ảnh hưởng của bmd là ước tính xác suất gãy xương qua phương trình:

$$\hat{p} = \frac{e^{1.063 - 2.27(bmd)}}{1 + e^{1.063 - 2.27(bmd)}}$$

Theo đó, khi $bmd = 1.00$, $p = 0.23$. Khi $bmd = 0.86$ (tức giảm 1 độ lệch chuẩn), $p = 0.291$. Tức là, nếu BMD giảm 1 độ lệch chuẩn thì xác suất gãy xương tăng $0.291/0.23 = 1.265$ hay 26%5.

(d) Phần cuối của kết quả cung cấp deviance cho hai mô hình: mô hình không có biến độc lập (null deviance), và mô hình với biến độc lập, tức là bmd trong ví dụ (residual deviance).

```
Null deviance: 157.81 on 136 degrees of freedom
Residual deviance: 155.27 on 135 degrees of freedom
AIC: 159.27
```

Qua hai số này, chúng ta thấy bmd ảnh hưởng rất thấp đến việc tiên đoán gãy xương, chỉ làm giảm deviance từ 157.8 xuống còn 155.27, và mức độ giảm này không có ý nghĩa thống kê.

Ngoài ra, R còn cung cấp giá trị của AIC (Akaike Information Criterion) được tính từ deviance và bậc tự do. Tôi sẽ quay lại ý nghĩa của AIC trong phần sắp đến khi so sánh các mô hình.

12.3 Ước tính xác suất bằng R

Xin nhắc lại trong phân tích trên, chúng ta cho các kết quả vào đối tượng `logistic`. Trong đối tượng này có nhiều thông tin có ích, nhưng nếu muốn xem các thông tin này chúng ta phải dùng đến các lệnh như `summary` chẳng hạn. Trong phần này, tôi sẽ trình bày một vài hàm để xem xét các thông tin liên quan đến việc tiên đoán xác suất.

- predict dùng để liệt kê các giá trị ước tính (predicted values) của mô hình
 $\log\left(\frac{p}{1-p}\right) = \alpha + \beta x$ cho từng bệnh nhân.

```
> predict(logistic)
```

```

2.377576584 1.085694014 -2.141117756 1.492824115 0.965379946 -0.941253280
    7          8          9          10         11         12
-1.733686514 -1.675645430 -0.665282957 -0.507046129 -0.941854868 -0.648740461
...

```

Các số trên là $\log(p / (1 - p))$, tức *log odds*, không có ý nghĩa hực tế bao nhiêu. Chúng ta muốn biết giá trị tiên đoán xác suất p tính từ phương trình $\hat{p} = \frac{e^{1.063-2.27(bmd)}}{1+e^{1.063-2.27(bmd)}}$. Để có giá trị này cho từng bệnh nhân, chúng ta cho thông số `type="response"` vào hàm `predict` như sau:

```

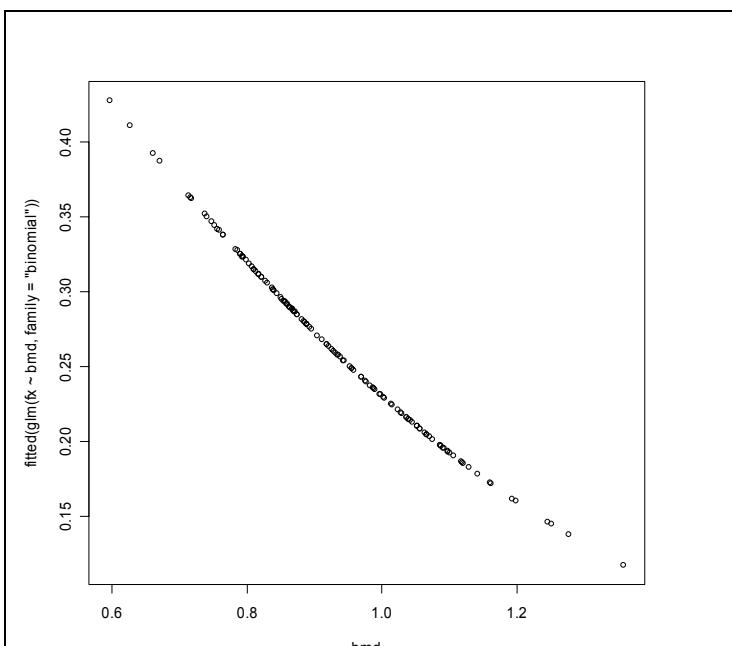
> predict(logistic, type="response")
      1        2        3        4        5        6        7
0.91510135 0.74757001 0.10516416 0.81650178 0.72419767 0.28064726 0.15011664
      8        9       10       11       12       13       14
0.15767295 0.33955387 0.37588624 0.28052582 0.34327343 0.44305196 0.23830776
...

```

Trong kết quả trên (chỉ in một phần) ước tính xác suất gãy xương cho bệnh nhân 1 là 0.915, cho bệnh nhân 2 là 0.747, v.v...

- Chúng ta có thể xem xét các giá trị tiên đoán này với độ bmd bằng cách dùng hàm `plot` thông thường:

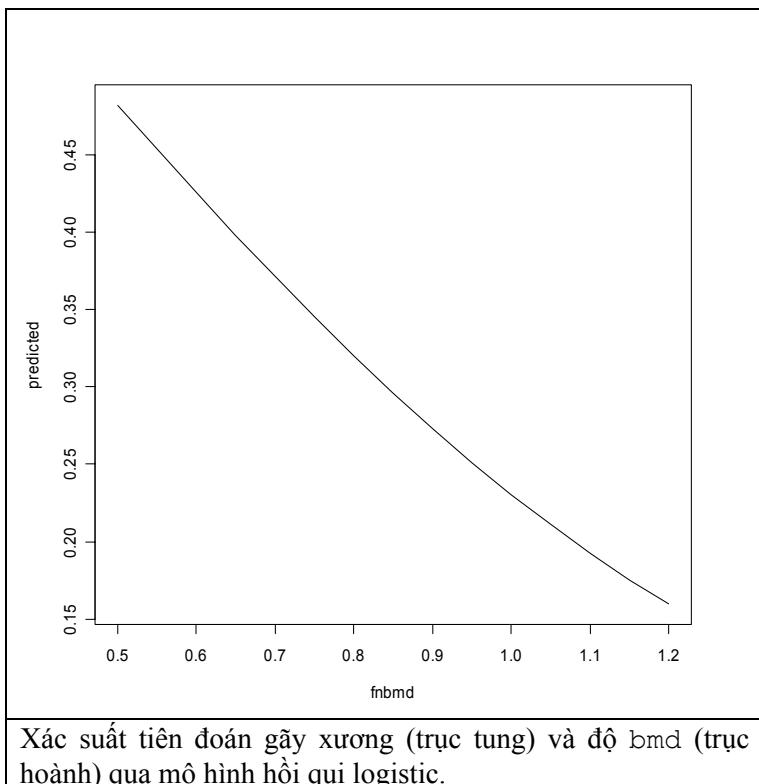
```
> plot(bmd, fitted(glm(fx ~ bmd, family="binomial")))
```



Xác suất tiên đoán gãy xương (trục tung) và độ bmd (trục hoành) qua mô hình hồi qui logistic.

Biểu đồ trên có thể cải tiến bằng cách cho các khoảng cách giá trị bmd gần nhau hơn (như 0.50, 0.55, 0.60, ..., 1.20 chẳng hạn), và dùng đường biểu diễn thay vì dùng dấu chấm. Các lệnh sau đây sẽ cải tiến biểu đồ.

```
> logistic <- glm(fx ~ bmd, family="binomial")
> fnbmd <- seq(0.5, 1.2, 0.05) #cho fnbmd từ > 0.50, 0.55, 0.6,...,1.2
> new.data <- data.frame(bmd = fnbmd) #cho vào một dataframe mới
> predicted <- predict(logistic, new.data, type="response")
> plot(predicted ~ fnbmd, type="l")
```



13. Ước tính cỡ mẫu (sample size estimation)

Một công trình nghiên cứu thường dựa vào một mẫu (sample). Một trong những câu hỏi quan trọng nhất trước khi tiến hành nghiên cứu là cần bao nhiêu mẫu hay bao nhiêu đối tượng cho nghiên cứu. “Đối tượng” ở đây là đơn vị căn bản của một nghiên cứu, là số bệnh nhân, số tình nguyện viên, số mẫu ruộng, cây trồng, thiết bị, v.v... Ước tính số lượng đối tượng cần thiết cho một công trình nghiên cứu đóng vai trò cực kì quan trọng, vì nó có thể là yếu tố quyết định sự thành công hay thất bại của nghiên cứu. Nếu số lượng đối tượng không đủ thì kết luận rút ra từ công trình nghiên cứu không có độ chính xác cao, thậm chí không thể kết luận gì được. Ngược lại, nếu số lượng đối tượng quá nhiều hơn số cần thiết thì tài nguyên, tiền bạc và thời gian sẽ bị hao phí. Do đó, vấn đề then chốt trước khi nghiên cứu là phải ước tính cho được một số đối tượng vừa đủ cho mục tiêu của nghiên cứu. Số lượng đối tượng “vừa đủ” tùy thuộc vào ba yếu tố chính:

- Sai sót mà nhà nghiên cứu chấp nhận, cụ thể là sai sót loại I và II;
- Độ dao động (variability) của đo lường, mà cụ thể là độ lệch chuẩn; và
- Mức độ khác biệt hay ảnh hưởng mà nhà nghiên cứu muốn phát hiện.

Không có số liệu về ba yếu tố này thì không thể nào ước tính cỡ mẫu. Kinh nghiệm của người viết cho thấy rất nhiều người khi tiến hành nghiên cứu thường không có ý niệm gì về các số liệu này, cho nên khi đến tham vấn các chuyên gia về thống kê học, họ chỉ nhận câu trả lời: “không thể tính được”! Trong chương này tôi sẽ bàn qua ba yếu tố trên.

13.1 Khái niệm về “power”

Thống kê học là một phương pháp khoa học có mục đích phát hiện, hay đi tìm những cái có thể gộp chung lại bằng cụm từ “chưa được biết” (unknown). Cái chưa được biết ở đây là những hiện tượng chúng ta không quan sát được, hay quan sát được nhưng không đầy đủ. “Cái chưa biết” có thể là một ẩn số (như chiều cao trung bình ở người Việt Nam, hay trọng lượng một phần tử), hiệu quả của một thuật điều trị, gen có chức năng làm cho cây lá có màu xanh, sở thích của con người, v.v... Chúng ta có thể đo chiều cao, hay tiến hành xét nghiệm để biết hiệu quả của thuốc, nhưng các nghiên cứu như thế chỉ được tiến hành trên một nhóm đối tượng, chứ không phải toàn bộ quần thể của dân số.

Ở mức độ đơn giản nhất, những cái chưa biết này có thể xuất hiện dưới hai hình thức: hoặc là có, hoặc là không. Chẳng hạn như một thuật điều trị có hay không có hiệu quả chống gãy xương, khách hàng thích hay không thích một loại nước giải khát. Bởi vì không ai biết hiện tượng một cách đầy đủ, chúng ta phải đặt ra giả thiết. Giả thiết đơn giản nhất là *giả thiết đảo* (hiện tượng không tồn tại, kí hiệu H-) và *giả thiết chính* (hiện tượng tồn tại, kí hiệu H+).

Chúng ta sử dụng các phương pháp kiểm định thống kê (statistical test) như kiểm định t , F , z , χ^2 , v.v... để đánh giá khả năng của giả thiết. Kết quả của một kiểm định thống kê có thể đơn giản chia thành hai giá trị: hoặc là *có ý nghĩa thống kê* (statistical significance), hoặc là *không có ý nghĩa thống kê* (non-significance). Có ý nghĩa thống kê ở đây, như đề cập trong Chương 7, thường dựa vào trị số P: nếu $P < 0.05$, chúng ta phát biểu kết quả có ý nghĩa thống kê; nếu $P > 0.05$ chúng ta nói kết quả không có ý nghĩa thống kê. Cũng có thể xem có ý nghĩa thống kê hay không có ý nghĩa thống kê như là có tín hiệu hay không có tín hiệu. Hãy tạm đặt kí hiệu T+ là kết quả có ý nghĩa thống kê, và T- là kết quả kiểm định không có ý nghĩa thống kê.

Hãy xem xét một ví dụ cụ thể: để biết thuốc risedronate có hiệu quả hay không trong việc điều trị loãng xương, chúng ta tiến hành một nghiên cứu gồm 2 nhóm bệnh nhân (một nhóm được điều trị bằng risedronate và một nhóm chỉ sử dụng giả dược placebo). Chúng ta theo dõi và thu thập số liệu gãy xương, ước tính tỉ lệ gãy xương cho từng nhóm, và so sánh hai tỉ lệ bằng một kiểm định thống kê. Kết quả kiểm định thống kê hoặc là *có ý nghĩa thống kê* ($P < 0.05$) hay không có ý nghĩa thống kê ($P > 0.05$). Xin nhắc lại rằng chúng ta không biết risedronate thật sự có hiệu nghiệm chống gãy xương

hay không; chúng ta chỉ có thể đặt giả thiết H. Do đó, khi xem xét một giả thiết và kết quả kiểm định thống kê, chúng ta có bốn tình huống:

- (a) Giả thuyết H đúng (thuốc risedronate có hiệu nghiệm) và kết quả kiểm định thống kê $P<0.05$.
- (b) Giả thuyết H đúng, nhưng kết quả kiểm định thống kê không có ý nghĩa thống kê;
- (c) Giả thuyết H sai (thuốc risedronate không có hiệu nghiệm) nhưng kết quả kiểm định thống kê có ý nghĩa thống kê;
- (d) Giả thuyết H sai và kết quả kiểm định thống kê không có ý nghĩa thống kê.

Ở đây, trường hợp (a) và (d) không có vấn đề, vì kết quả kiểm định thống kê nhất quán với thực tế của hiện tượng. Nhưng trong trường hợp (b) và (c), chúng ta phạm sai lầm, vì kết quả kiểm định thống kê không phù hợp với giả thiết. Trong ngôn ngữ thống kê học, chúng ta có vài thuật ngữ:

- xác suất của tình huống (b) xảy ra được gọi là *sai sót loại II* (type II error), và thường kí hiệu bằng β .
- xác suất của tình huống (a) được gọi là *Power*. Nói cách khác, *power* chính là xác suất mà kết quả kiểm định thống cho ra kết quả $p<0.05$ với điều kiện giả thiết H là thật. Nói cách khác: $power = 1-\beta$;
- xác suất của tình huống (c) được gọi là *sai sót loại I* (type I error, hay significance level), và thường kí hiệu bằng α . Nói cách khác, α chính là xác suất mà kết quả kiểm định thống cho ra kết quả $p<0.05$ với điều kiện giả thiết H sai;
- xác suất tình hống (d) không phải là vấn đề cần quan tâm, nên không có thuật ngữ, dù có thể gọi đó là kết quả *âm tính thật* (hay true negative).

Có thể tóm lược 4 tình huống đó trong một Bảng 1 sau đây:

Các tình huống trong việc thử nghiệm một giả thiết khoa học

Kết quả kiểm định thống kê	Giả thuyết H	
	Đúng (thuốc có hiệu nghiệm)	Sai (thuốc không có hiệu nghiệm)
Có ý nghĩa thống kê ($p<0,05$)	Dương tính thật (power), $1-\beta = P(s H+)$	Sai sót loại I (type I error) $\alpha = P(s H-)$
Không có ý nghĩa thống kê ($p>0,05$)	Sai sót loại II (type II error) $\beta = P(ns H+)$	Âm tính thật (true negative) $1-\alpha = P(ns H-)$

Chú thích: s trong biểu đồ này có nghĩa là significant; ns non-significant; $H+$ là giả thuyết đúng; và $H-$ là giả thuyết sai. Do đó, có thể mô tả 4 tình huống trên bằng ngôn ngữ xác suất có điều kiện như sau: $\text{Power} = 1 - \beta = P(s | H+)$; $\beta = P(ns | H+)$; và $\alpha = P(s | H-)$.

13.2 Số liệu để ước tính cỡ mẫu

Như đã đề cập trong phần đầu của chương này, để ước tính số đối tượng cần thiết cho một công trình nghiên cứu, chúng ta cần phải có 3 số liệu: xác suất sai sót loại I và II, độ dao động của đo lường, và độ ảnh hưởng.

- Về xác suất sai sót, thông thường một nghiên cứu chấp nhận sai sót loại I khoảng 1% hay 5% (tức $\alpha = 0.01$ hay 0.05), và xác suất sai sót loại II khoảng $\beta = 0.1$ đến $\beta = 0.2$ (tức power phải từ 0.8 đến 0.9).
- Độ dao động chính là độ lệch chuẩn (standard deviation) của đo lường mà công trình nghiên cứu dựa vào để phân tích. Chẳng hạn như nếu nghiên cứu về cao huyết áp, thì nhà nghiên cứu cần phải có độ lệch chuẩn của áp suất máu. Chúng ta tạm gọi độ dao động là σ .
- Độ ảnh hưởng, nếu là công trình nghiên cứu so sánh hai nhóm, là độ khác biệt trung bình giữa hai nhóm mà nhà nghiên cứu muốn phát hiện. Chẳng hạn như nhà nghiên cứu có thể giả thiết rằng bệnh nhân được điều trị bằng thuốc A có áp suất máu giảm 10 mmHg so với nhóm giả được. Ở đây, 10 mmHg được xem là độ ảnh hưởng. Chúng ta tạm gọi độ ảnh hưởng là Δ .

Một nghiên cứu có thể có một nhóm đối tượng hay hai (và có khi hơn 2) nhóm đối tượng. Và ước tính cỡ mẫu cũng tùy thuộc vào các trường hợp này.

Trong trường hợp một nhóm đối tượng, số lượng đối tượng (n) cần thiết cho nghiên cứu có thể tính toán một cách “thủ công” như sau:

$$n = \frac{C}{(\Delta/\sigma)^2}$$

Trong trường hợp có hai nhóm đối tượng, số lượng đối tượng (n) cần thiết cho nghiên cứu có thể tính toán như sau:

$$n = 2 \times \frac{C}{(\Delta/\sigma)^2}$$

Trong đó, hằng số C được xác định từ xác suất sai sót loại I và II (hay power) như sau:

Hàng số C liên quan đến sai sót loại I và II

$\alpha =$	$\beta = 0.20$ (Power = 0.80)	$\beta = 0.10$ (Power = 0.90)	$\beta = 0.05$ (Power = 0.95)
0.10	6.15	8.53	10.79
0.05	7.85	10.51	13.00
0.01	13.33	16.74	19.84

13.4 Ước tính cỡ mẫu

13.4.1 Ước tính cỡ mẫu cho một chỉ số trung bình

Ví dụ 20: Chúng ta muốn ước tính chiều cao ở đàn ông người Việt, và chấp nhận sai số trong vòng 1 cm ($d = 1$) với khoảng tin cậy 0.95 (tức $\alpha=0.05$) và power = 0.8 (hay $\beta = 0.2$). Các nghiên cứu trước cho biết độ lệch chuẩn chiều cao ở người Việt khoảng 4.6 cm. Chúng ta có thể áp dụng công thức [1] để ước tính cỡ mẫu cần thiết cho nghiên cứu:

$$n = \frac{C}{(\Delta/\sigma)^2} = \frac{7.85}{(1/4.6)^2} = 166$$

Nói cách khác, chúng ta cần phải đo chiều cao ở 166 đối tượng để ước tính chiều cao đàn ông Việt với sai số trong vòng 1 cm.

Nếu sai số chấp nhận là 0.5 cm (thay vì 1 cm), số lượng đối tượng cần thiết là:
 $n = \frac{7.85}{(0.5/4.6)^2} = 664$. Nếu độ sai số mà chúng ta chấp nhận là 0.1 cm thì số lượng đối tượng nghiên cứu lên đến 16610 người! Qua các ước tính này, chúng ta dễ dàng thấy cỡ mẫu tùy thuộc rất lớn vào độ sai số mà chúng ta chấp nhận. Muốn có ước tính càng chính xác, chúng ta cần càng nhiều đối tượng nghiên cứu.

Trong R có hàm power.t.test có thể áp dụng để ước tính cỡ mẫu cho ví dụ trên như sau. Chú ý chúng ta cho R biết vấn đề là một nhóm tức type="one.sample":

```
# sai số 1 cm, độc lệch chuẩn 4.6, a=0.05, power=0.8
> power.t.test(delta=1, sd=4.6, sig.level=.05, power=.80,
   type='one.sample')
```

One-sample t test power calculation

```
n = 168.0131
delta = 1
sd = 4.6
sig.level = 0.05
power = 0.8
alternative = two.sided
```

kết quả tính toán từ R là 168, khác với cách tính thủ công 2 đổi tượng, vì có nhiên R sử dụng nhiều số lẻ hơn và chính xác hơn cách tính thủ công. Với sai số 0.5 cm:

```
# sai số 0.5 cm, độc lệch chuẩn 4.6, a=0.05, power=0.8
> power.t.test(delta=0.5, sd=4.6, sig.level=.05, power=.80,
                 type='one.sample')

One-sample t test power calculation

n = 666.2525
delta = 0.5
sd = 4.6
sig.level = 0.05
power = 0.8
alternative = two.sided
```

Ví dụ 21: Một loại thuốc điều trị có khả năng tăng độ alkaline phosphatase ở bệnh nhân loãng xương. Độ lệch chuẩn của alkaline phosphatase là 15 U/l. Một nghiên cứu mới sẽ tiến hành trong một quần thể bệnh nhân ở Việt Nam, và các nhà nghiên cứu muốn biết bao nhiêu bệnh nhân cần tuyển để chứng minh rằng thuốc có thể alkaline phosphatase từ 60 đến 65 U/l sau 3 tháng điều trị, với sai số $\alpha = 0.05$ và power = 0.8.

Đây là một loại nghiên cứu “trước – sau” (before-after study); có nghĩa là trước và sau khi điều trị. Ở đây, chúng ta chỉ có một nhóm bệnh nhân, nhưng được đo hai lần (trước khi dùng thuốc và sau khi dùng thuốc). Chỉ tiêu lâm sàng để đánh giá hiệu nghiệm của thuốc là độ thay đổi về alkaline phosphatase. Trong trường hợp này, chúng ta có trị số tăng trung bình là 5 U/l và độ lệch chuẩn là 15 U/l, hay nói theo ngôn ngữ R, $delta=5$, $sd=15$, $sig.level=.05$, $power=.80$, và lệnh:

```
> power.t.test(delta=3, sd=15, sig.level=.05, power=.80,
                 type='one.sample')

One-sample t test power calculation

n = 198.1513
delta = 3
sd = 15
sig.level = 0.05
power = 0.8
alternative = two.sided
```

Như vậy, chúng ta cần phải có 198 bệnh nhân để đạt các mục tiêu trên.

13.4.2 Ước tính cỡ mẫu cho so sánh hai số trung bình

Trong thực tế, rất nhiều nghiên cứu nhằm so sánh hai nhóm với nhau. Cách ước tính cỡ mẫu cho các nghiên cứu này chủ yếu dựa vào công thức [2] như trình bày phần 15.3.1.

Ví dụ 22: Một nghiên cứu được thiết kế để thử nghiệm thuốc alendronate trong việc điều trị loãng xương ở phụ nữ sau thời kỳ mãn kinh. Có hai nhóm bệnh nhân được tuyển: nhóm 1 là nhóm can thiệp (được điều trị bằng alendronate), và nhóm 2 là nhóm đối chứng (tức không được điều trị). Tiêu chí để đánh giá hiệu quả của thuốc là mật độ xương (bone mineral density – BMD). Số liệu từ nghiên cứu dịch tỦ học cho thấy giá trị trung bình của BMD trong phụ nữ sau thời kỳ mãn kinh là 0.80 g/cm^2 , với độ lệch chuẩn là 0.12 g/cm^2 . Vấn đề đặt ra là chúng ta cần phải nghiên cứu ở bao nhiêu đối tượng để “chứng minh” rằng sau 12 tháng điều trị BMD của nhóm 1 tăng khoảng 5% so với nhóm 2?

Trong ví dụ trên, tạm gọi trị số trung bình của nhóm 2 là μ_2 và nhóm 1 là μ_1 , chúng ta có: $\mu_1 = 0.8 * 1.05 = 0.84 \text{ g/cm}^2$ (tức tăng 5% so với nhóm 1), và do đó, $\Delta = 0.84 - 0.80 = 0.04 \text{ g/cm}^2$. Độ lệch chuẩn là $\sigma = 0.12 \text{ g/cm}^2$. Với power = 0.90 và $\alpha = 0.05$, cỡ mẫu cần thiết là:

$$n = \frac{2C}{(\Delta/\sigma)^2} = \frac{2 \times 10.51}{(0.04/0.12)^2} = 189$$

Và lời giải từ R qua hàm power.t.test như sau:

```
> power.t.test(delta=0.04, sd=0.12, sig.level=0.05, power=0.90,
  type="two.sample")
```

```
Two-sample t test power calculation

      n = 190.0991
    delta = 0.04
      sd = 0.12
 sig.level = 0.05
    power = 0.9
  alternative = two.sided
```

NOTE: n is number in *each* group

Chú ý trong hàm power.t.test, ngoài các thông số thông thường như delta (độ ảnh hưởng hay khác biệt theo giả thiết), sd (độ lệch chuẩn), sig.level xác suất sai sót loại I, và power, chúng ta còn phải cụ thể chỉ ra rằng đây là nghiên cứu gồm có hai nhóm với thông số type="two.sample".

Kết quả trên cho biết chúng ta cần 190 bệnh nhân **cho mỗi nhóm** (hay 380 bệnh nhân cho công trình nghiên cứu). Trong trường hợp này, power = 0.90 và $\alpha = 0.05$ có nghĩa là gì? Trả lời: hai thông số đó có nghĩa là nếu chúng ta tiến hành thật nhiều nghiên cứu (ví dụ 1000) và mỗi nghiên cứu với 380 bệnh nhân, sẽ có 90% (hay 900) nghiên cứu sẽ cho ra kết quả trên với trị số $p < 0.05$.

13.4.3 Ước tính cỡ mẫu cho phân tích phương sai

Phương pháp ước tính cỡ mẫu cho so sánh giữa hai nhóm cũng có thể khai triển thêm để ước tính cỡ mẫu cho trường hợp so sánh hơn hai nhóm. Trong trường hợp có nhiều nhóm, như đề cập trong Chương 11, phương pháp so sánh là phân tích phương sai. Theo phương pháp này, số trung bình bình phương phần dư (residual mean square, RMS) chính là ước tính của độ dao động của đo lường trong mỗi nhóm, và chỉ số này rất quan trọng trong việc ước tính cỡ mẫu.

Chi tiết về lí thuyết đằng sau cách ước tính cỡ mẫu cho phân tích phương sai khá phức tạp, và không nằm trong phạm vi của chương này. Nhưng nguyên lý chủ yếu vẫn không khác so với lí thuyết so sánh giữa hai nhóm. Gọi số trung bình của k nhóm là $\mu_1, \mu_2, \mu_3, \dots, \mu_k$, chúng ta có thể tính tổng bình phương giữa các nhóm bằng $SS = \sum_{i=1}^k (\mu_i - \bar{\mu})^2$, trong đó, $\bar{\mu} = \sum_{i=1}^k \mu_i / k$. Cho $\lambda = \frac{SS}{(k-1)RMS}$, vẫn đề đặt ra là tìm cỡ lượng cỡ mẫu n sao cho z_β đáp ứng yêu cầu power = 0.80 hay 0.9, mà

$$z_\beta = \frac{1}{\sqrt{(k-1)(1+n\lambda)F + k(n-1)(1+2n\lambda)}} \times \\ \left(\sqrt{k(n-1)[2(k-1)(1+n\lambda)^2 - (1+2n\lambda)]} - \sqrt{F(k-1)(1+n\lambda)(2k(n-1)-1)} \right)$$

Trong đó F là kiểm định F . (Xem J. Fleiss, “The Design and Analysis of Clinical Experiments”, John Wiley & Sons, New York 1986, trang 373).

Ví dụ 23. Để so sánh độ ngọt của một loại nước uống giữa 4 nhóm đối tượng khác nhau về giới tính và độ tuổi (tạm gọi 4 nhóm là A, B, C và D), các nhà nghiên cứu giả thiết rằng độ ngọt trong nhóm A, B, C và D lần lượt là 4.5, 3.0, 5.6, và 1.3. Qua xem xét nhiều nghiên cứu trước, các nhà nghiên cứu còn biết rằng RMS về độ ngọt trong mỗi nhóm là khoảng 8.7. Vấn đề đặt ra là bao nhiêu đối tượng cần nghiên cứu để phát hiện sự khác biệt có ý nghĩa thống kê ở mức độ $\alpha = 0.05$ và power = 0.9.

Hàm `power.anova.test` trong R có thể ứng dụng để giải quyết vấn đề. Chúng ta chỉ cần đơn giản cung cấp 4 số trung bình theo giả thiết và số RMS như sau:

```
# trước hết cho 4 số trung bình vào một vector
> groupmeans <- c(4.5, 3.0, 5.6, 1.3)

# sau đó, "gọi" hàm power.anova.test:
> power.anova.test(groups = length(groupmeans),
                     between.var=var(groupmeans),
                     within.var=8.7, power=0.90, sig.level=0.05)

Balanced one-way analysis of variance power calculation

groups = 4
```

```

n = 12.81152
between.var = 3.486667
within.var = 8.7
sig.level = 0.05
power = 0.9

```

NOTE: n is number in each group

Kết quả cho thấy các nhà nghiên cứu cần khoảng 13 đối tượng cho mỗi nhóm (tức 52 đối tượng cho toàn bộ nghiên cứu).

13.4.4 Ước tính cỡ mẫu để ước tính một tỉ lệ

Nhiều nghiên cứu mô tả có mục đích khá đơn giản là ước tính một tỉ lệ. Chẳng hạn như giới y tế thường hay tìm hiểu tỉ lệ một bệnh trong cộng đồng, hay giới thăm dò ý kiến và thị trường thường tìm hiểu tỉ lệ dân số ưa thích một sản phẩm. Trong các trường hợp này, chúng ta không có những đo lường mang tính liên tục, nhưng kết quả chỉ là những giá trị nhị phân có / không, thích / không thích, v.v... Và cách ước tính cỡ mẫu cũng khác với ba ví dụ trên đây.

Năm 1991, một cuộc thăm dò ý kiến ở Mĩ cho thấy 45% người được hỏi sẵn sàng khuyến khích con họ nên hiến một quả thận cho những bệnh nhân cần thiết. Khoảng tin cậy 95% của tỉ lệ này là 42% đến 48%, tức một khoảng cách đến 6%! Kết quả này [tương đối] thiếu chính xác, dù số lượng đối tượng tham gia lên đến 1000 người. Tại sao? Để trả lời câu hỏi này, chúng ta thử xem qua một vài lí thuyết về ước tính cỡ mẫu cho một tỉ lệ.

Chúng ta biết qua Chương 6 và 9 rằng nếu \hat{p} được ước tính từ n đối tượng, thì khoảng tin cậy 95% của một tỉ lệ p [trong dân số] là: $\hat{p} \pm 1.96 \times SE(\hat{p})$, trong đó $SE(\hat{p}) = \sqrt{\hat{p}(1-\hat{p})/n}$.

Bây giờ thử lật ngược vấn đề: chúng ta muốn ước tính p sao khoảng rộng $2 \times 1.96 \times SE(\hat{p})$ không quá một hằng số m . Nói cách khác, chúng ta muốn:

$$1.96 \times \sqrt{\hat{p}(1-\hat{p})/n} \leq m$$

Chúng ta muốn tìm số lượng đối tượng n để đạt yêu cầu trên. Qua cách diễn đạt trên, dễ dàng thấy rằng:

$$n \geq \left(\frac{1.96}{m} \right)^2 \hat{p}(1-\hat{p})$$

Do đó, số lượng cỡ mẫu tùy thuộc vào độ sai số m và tỉ lệ p mà chúng ta muốn ước tính. Độ sai số càng thấp, số lượng cỡ mẫu càng cao.

Ví dụ 24: Chúng ta muốn ước tính tỉ lệ đàn ông hút thuốc ở Việt Nam, sao cho ước số không cao hơn hay thấp hơn 2% so với tỉ lệ thật trong toàn dân số. Một nghiên cứu trước cho thấy tỉ lệ hút thuốc trong đàn ông người Việt có thể lên đến 70%. Câu hỏi đặt ra là chúng ta cần nghiên cứu trên bao nhiêu đàn ông để đạt yêu cầu trên.

Trong ví dụ này, chúng ta có sai số $m = 0.02$, $\hat{p} = 0.70$, và số lượng cỡ mẫu cần thiết cho nghiên cứu là:

$$n \geq \left(\frac{1.96}{0.02} \right)^2 0.7 \times 0.3$$

Nói cách khác, chúng ta cần nghiên cứu ít nhất là 2017.

Nếu chúng ta muốn giảm sai số từ 2% xuống 1% (tức $m = 0.01$) thì số lượng đôi tượng sẽ là 8067! Chỉ cần thêm độ chính xác 1%, số lượng mẫu có thể thêm hơn 6000 người. Do đó, vẫn đề ước tính cỡ mẫu phải rất thận trọng, xem xét cân bằng giữa độ chính xác thông tin cần thu thập và chi phí.

R không có hàm cho ước tính cỡ mẫu cho một tỉ lệ, nhưng với công thức trên, bạn đọc có thể viết một hàm để tính rất dễ dàng.

13.4.5 Ước tính cỡ mẫu cho so sánh hai tỉ lệ

Nhiều nghiên cứu mang tính suy luận thường có hai [hay nhiều hơn hai] nhóm để so sánh. Trong phần 15.4.2 chúng ta đã làm quen với phương pháp ước tính cỡ mẫu để so sánh hai số trung bình bằng kiểm định t. Đó là những người có tiêu chí là những biến số liên tục. Nhưng có nghiên cứu biến số không liên tục mà mang tính nhị phân như tôi vừa bàn trong phần 15.4.3. Để so sánh hai tỉ lệ, phương pháp kiểm định thông dụng nhất là kiểm định nhị phân (binomial test) hay Chi bình phương (χ^2 test). Trong phần này, tôi sẽ bàn qua cách tính cỡ mẫu cho hai loại kiểm định thống kê này.

Gọi hai tỉ lệ [mà chúng ta không biết nhưng muốn tìm hiểu] là p_1 và p_2 , và gọi $\Delta = p_1 - p_2$. Giả thiết mà chúng ta muốn kiểm định là $\Delta = 0$. Lí thuyết đằng sau để ước tính cỡ mẫu cho kiểm định giả thiết này khá rườm rà, nhưng có thể tóm gọn bằng công thức sau đây:

$$n = \frac{\left(z_{\alpha/2} \sqrt{2\bar{p}(1-\bar{p})} + z_\beta \sqrt{p_1(1-p_1) + p_2(1-p_2)} \right)^2}{\Delta^2}$$

Trong đó, $\bar{p} = (p_1 + p_2)/2$, $z_{\alpha/2}$ là trị số z của phân phối chuẩn cho xác suất $\alpha/2$ (chẳng hạn như khi $\alpha = 0.05$, thì $z_{\alpha/2} = 1.96$; khi $\alpha = 0.01$, thì $z_{\alpha/2} = 2.57$), và z_β là trị số z của

phân phối chuẩn cho xác suất β (chẳng hạn như khi $\beta = 0.10$, thì $z_\beta = 1.28$; khi $\beta = 0.20$, thì $z_\beta = 0.84$).

Ví dụ 25: Một thử nghiệm lâm sàng đối chứng ngẫu nhiên được thiết kế để đánh giá hiệu quả của một loại thuốc chống gãy xương sống. Hai nhóm bệnh nhân sẽ được tuyển. Nhóm 1 được điều trị bằng thuốc, và nhóm 2 là nhóm đối chứng (không được điều trị). Các nhà nghiên cứu giả thiết rằng tỉ lệ gãy xương trong nhóm 2 là khoảng 10%, và thuốc có thể làm giảm tỉ lệ này xuống khoảng 6%. Nếu các nhà nghiên cứu muốn thử nghiệm giả thiết này với sai sót I là $\alpha = 0.01$ và power = 0.90, bao nhiêu bệnh nhân cần phải được tuyển mộ cho nghiên cứu?

Ở đây, chúng ta có $\Delta = 0.10 - 0.06 = 0.04$, và $\bar{p} = (0.10 + 0.06)/2 = 0.08$. Với $\alpha = 0.01$, $z_{\alpha/2} = 2.57$ và với power = 0.90, $z_\beta = 1.28$. Do đó, số lượng bệnh nhân cần thiết cho mỗi nhóm là:

$$n = \frac{(2.57\sqrt{2 \times 0.08 \times 0.92} + 1.28\sqrt{0.1 \times 0.90 + 0.06 \times 0.94})^2}{(0.04)^2} = 1361$$

Như vậy, công trình nghiên cứu này cần phải tuyển ít nhất là 2722 bệnh nhân để kiểm định giả thiết trên.

Hàm `power.prop.test` R có thể ứng dụng để tính cỡ mẫu cho trường hợp trên. Hàm `power.prop.test` cần những thông tin như `power`, `sig.level`, `p1`, và `p2`. Trong ví dụ trên, chúng ta có thể viết:

```
> power.prop.test(p1=0.10, p2=0.06, power=0.90, sig.level=0.01)
```

```
Two-sample comparison of proportions power calculation

      n = 1366.430
      p1 = 0.1
      p2 = 0.06
      sig.level = 0.01
      power = 0.9
      alternative = two.sided
```

NOTE: n is number in *each* group

Chú ý kết quả từ R có phần chính xác hơn (1366 đối tượng cho mỗi nhóm) vì R dùng nhiều số lẽ cho tính toán hơn là tính “thủ công”.

Trước khi rời chương này, tôi muốn nhân cơ hội này để nhấn mạnh một lần nữa, ước tính cỡ mẫu cho nghiên cứu là một bước cực kì quan trọng trong việc thiết kế một nghiên cứu cho có ý nghĩa khoa học, vì nó có thể quyết định thành bại của nghiên cứu. Trước khi ước tính cỡ mẫu nhà nghiên cứu cần phải biết trước (hay ít ra là có vài giả thiết cụ thể) về vấn đề mình quan tâm. Ước tính cỡ mẫu cần một số thông số như đê cập đến

trong phần đầu của chương, và nếu các thông số này không có thì không thể ước tính được. Trong trường hợp một nghiên cứu hoàn toàn mới, tức chưa ai từng làm trước đó, có thể các thông số về độ ảnh hưởng và độ dao động đo lường sẽ không có, và nhà nghiên cứu cần phải tiến hành một số mô phỏng (simulation) hay một nghiên cứu sơ khởi để có những thông số cần thiết. Cách ước tính cỡ mẫu bằng mô phỏng là một lĩnh vực nghiên cứu khá chuyên sâu, không nằm trong đề tài của sách này, nhưng bạn đọc có thể tìm hiểu thêm phương pháp này trong các sách giáo khoa về thống kê học cấp cao hơn.

Trên đây là vài hướng dẫn nhanh để bạn đọc có thể sử dụng R cho phân tích số liệu và tạo biểu đồ. Bài viết này thực chất là tóm lược từ cuốn *Phân tích số liệu và tạo biểu đồ bằng R: hướng dẫn và thực hành*, do Nhà xuất bản Đại học Quốc gia Thành phố Hồ Chí Minh ấn hành vào năm 2006. Chi tiết về lí thuyết và một số phương pháp khác như phân tích sự kiện, xây dựng mô hình thống kê, mô phỏng, lập chương, v.v... có thể tìm trong sách trên.

14. Tài liệu tham khảo

Hiện nay, thư viện sách về R còn tương đối khiêm tốn so với thư viện cho các phần mềm thương mại như SAS và SPSS. Tuy nhiên, trong thời đại tiến bộ phi thường về thông tin internet và toàn cầu hóa như hiện nay, sách in và sách xuất bản trên website không còn là những khác nhau bao xa. Phần lớn chỉ dẫn về cách sử dụng R có thể tìm thấy rải rác đây đó trên các website từ các trường đại học và website cá nhân trên khắp thế giới. Trong phần này tôi chỉ liệt kê một số sách mà bạn đọc, nếu cần tham khảo thêm, nên tìm đọc. Trong quá trình viết cuốn sách mà bạn đọc đang cầm trên tay, tôi cũng tham khảo một số sách và trang web mà tôi sẽ liệt kê sau đây với vài lời nhận xét cá nhân.

Tài liệu tham khảo chính về R là bài báo của hai người sáng tạo ra R: Ihaka R, Gentleman R. R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics* 1996; 5:299-314.

- “**Data Analysis and Graphics Using R – An Example Approach**” (Nhà xuất bản Cambridge University Press, 2003) của John Maindonald nay đã xuất in lại lần thứ 2 với thêm một tác giả mới John Braun. Đây là cuốn sách rất có ích cho những ai muốn tìm hiểu và học về R. Năm chương đầu của sách viết cho bạn đọc chưa từng biết về R, còn các chương sau thì viết cho các bạn đọc đã biết cách sử dụng R thành thạo.
- “**Introductory Statistics With R**” (Nhà xuất bản Springer, 2004) của Peter Dalgaard là một cuốn sách loại căn bản cho R nhắm vào bạn đọc chưa biết gì về R. Sách tương đối ngắn (chỉ khoảng 200 trang) nhưng khá đắt giá!
- “**Linear Models with R**” (Nhà xuất bản Chapman & Hall/CRC, 2004) của Julian Faraway. Sách hiện có thể tải từ internet xuống miễn phí tại website sau đây: <http://www.stat.lsa.umich.edu/~faraway/book/prab.pdf> hay <http://cran.r-project.org/doc/contrib/Faraway-PRA.pdf>. Tài liệu dài 213 trang.
- “**R Graphics (Computer Science and Data Analysis)**” (Nhà xuất bản Chapman & Hall/CRC, 2005) của Paul Murrell. Đây là cuốn sách chuyên về phân tích biểu đồ bằng R. Sách có rất nhiều mã để bạn đọc có thể tự mình thiết kế các biểu đồ phức tạp và ... màu mè.
- “**Modern Applied Statistics with S-Plus**” (Nhà xuất bản Springer, 4th Edition, 2003) của W. N. Venables và B. D. Ripley được viết cho ngôn ngữ S-Plus nhưng tất cả các lệnh và mã trong sách này đều có thể áp dụng cho R mà không cần thay đổi. (S-Plus là tiền thân của R, nhưng S-Plus là một phần mềm thương mại, còn R thì hoàn toàn miễn phí!) Đây là cuốn sách có thể nói là cuốn sách tham khảo cho tất cả ai muốn phát triển thêm về R. Hai tác giả cũng là những chuyên gia có thẩm quyền về ngôn ngữ R. Sách dành cho bạn đọc với trình độ cao về máy tính và thống kê học.

Các website quan trọng hay có ích về R

- Rất nhiều tài liệu tham khảo có thể tải từ website chính thức của R sau đây: <http://cran.R-project.org/other-docs.html>

Trong đó có một số tài liệu quan trọng như “**An Introduction to R**” của W. N. Venables và B. D. Ripley.

Địa chỉ internet: <http://cran.r-project.org/doc/manuals/R-intro.pdf>.

- Vài tài liệu hướng dẫn cách sử dụng R có thể tải (miễn phí) và tham khảo như sau:

“**R for Beginners**” (57 trang) của Emmanuel Paradis. Tài liệu được soạn cho bạn đọc mới làm quen với R.

Địa chỉ internet: http://cran.r-project.org/doc/contrib/Paradis-rdebuts_en.pdf.

“**Using R for Data Analysis and Graphics: Introduction, Code and Commentary**” (35 trang) của John Maindonald là một tóm lược các lệnh và hàm căn bản của R cho phân tích số liệu và biểu đồ. Chủ đề của tài liệu này rất gần với cuốn sách mà bạn đang đọc.

Địa chỉ internet: <http://cran.r-project.org/doc/contrib/usingR.pdf>

“**Statistical Analysis with R – a quick start**” (46 trang) của Oleg Nenadic và Walter Zucchini. Web. Tài liệu hướng dẫn cách ứng dụng R cho phân tích thống kê và biểu đồ.

Địa chỉ internet: http://www.statoek.wiso.uni-goettingen.de/mitarbeiter/ogi/pub/r_workshop.pdf

“**A Brief Guide to R for Beginners in Econometrics**” (31 trang) của M. Arai. Tài liệu chủ yếu soạn cho giới phân tích thống kê kinh tế.

Địa chỉ internet: http://people.su.se/~ma/R_intro

“**Notes on the use of R for psychology experiments and questionnaires**” (39 trang) của Jonathan Baron và Yuelin Li. Web. Tài liệu được soạn cho giới nghiên cứu tâm lí học và xã hội học. Có ví dụ về log-linear model và một số mô hình phân tích phương sai trong tâm lí học.

Địa chỉ internet: <http://www.psych.upenn.edu/~baron/rpsych/rpsych.html>

- StatsRus gồm một sưu tập về các mẹo để sử dụng R hữu hiệu hơn (dài khoảng 80 trang). Địa chỉ internet: <http://lark.cc.ukans.edu/pauljohn/R/statsRus.html>
- Và sau cùng là một tài liệu “**Hướng dẫn sử dụng R cho phân tích số liệu và biểu đồ**” (khoảng 50 trang – thường xuyên cập nhật hóa) do chính tôi viết bằng tiếng Việt. Website: www.R.ykhoa.net thực chất là tóm lược một số chương chính của cuốn sách này. Trang web này còn có tất cả các dữ liệu (datasets) và các mã sử dụng trong sách để bạn đọc có thể tải xuống máy tính cá nhân để sử dụng.

15. Thuật ngữ dùng trong sách

Tiếng Anh

95% confidence interval	Khoảng tin cậy 95%
Akaike Information criterion (AIC)	Tiêu chuẩn thông tin Akaike
Analysis of covariance	Phân tích hiệp biến
Analysis of variance (ANOVA)	Phân tích phương sai
Bar chart	Biểu đồ thanh
Binomial distribution	Phân phối nhị phân
Box plot	Biểu đồ hình hộp
Categorical variable	Biến thứ bậc
Clock chart	Biểu đồ đồng hồ
Coefficient of correlation	Hệ số tương quan
Coefficient of determination	Hệ số xác định bội
Coefficient of heterogeneity	Hệ số bất đồng nhất
Combination	Tổ hợp
Continuous variable	Biến liên tục
Correlation	Tương quan
Covariance	Hợp biến
Cross-over experiment	Thí nghiệm giao chéo
Cumulative probability distribution	Hàm phân phối tích lũy
Degree of freedom	Bậc tự do
Determinant	Định thức
Discrete variable	Biến rời rạc
Dot chart	Biểu đồ điểm
Estimate	Ước số
Estimator	Hàm ước lượng thống kê
Factorial analysis of variance	Phân tích phương sai cho thí nghiệm gai thừa
Fixed effects	Ảnh hưởng bất biến
Frequency	Tần số
Function	Hàm
Heterogeneity	Bất đồng nhất
Histogram	Biểu đồ tần số
Homogeneity	Đồng nhất
Hypothesis test	Kiểm định giả thiết
Inverse matrix	Ma trận nghịch đảo
Latin square experiment	Thí nghiệm hình vuông Latin
Least squares method	Phương pháp bình phương nhỏ nhất
Linear Logistic regression analysis	Phân tích hồi qui tuyến tính logistic
Linear regression analysis	Phân tích hồi qui tuyến tính

Matrix	Ma trận
Maximum likelihood method	Phương pháp hợp lí cực đại
Mean	Số trung bình
Median	Số trung vị
Meta-analysis	Phân tích tổng hợp
Missing value	Giá trị không
Model	Mô hình
Multiple linear regression analysis	Phân tích hồi qui tuyến tính đa biến
Normal distribution	Phân phối chuẩn
Object	Đối tượng
Parameter	Thông số
Permutation	Hoán vị
Pie chart	Biểu đồ hình tròn
Poisson distribution	Phân phối Poisson
Polynomial regression	Hồi qui đa thức
Probability	Xác suất
Probability density distribution	Hàm mật độ xác suất
P-value	Trị số P
Quantile	Hàm định bậc
Random effects	Ảnh hưởng ngẫu nhiên
Random variable	Biến ngẫu nhiên
Relative risk	Tỉ số nguy cơ tương đối
Repeated measure experiment	Thí nghiệm tái đo lường
Residual	Phần dư
Residual mean square	Trung bình bình phương phần dư
Residual sum of squares	Tổng bình phương phần dư
Scalar matrix	Ma trận vô hướng
Scatter plot	Biểu đồ tán xạ
Significance	Có ý nghĩa thống kê
Simulation	Mô phỏng
Standard deviation	Độ lệch chuẩn
Standard error	Sai số chuẩn
Standardized normal distribution	Phân phối chuẩn hóa
Survival analysis	Phân tích biến cố
Trposed matrix	Ma trận chuyển vị
Variable	Biến (biến số)
Variance	Phương sai
Weight	Trọng số
Weighted mean	Trung bình trọng số