TRƯỜNG ĐẠI HỌC Y DƯỢC TP HỒ CHÍ MINH
KHOA Y TẾ CÔNG CỘNG
Bộ môn Thống kê Y Học và Tin Học

Căn bản thống kê y học

Betty Kirwood
(London School of Hygiene and Tropical Medicine)
Dịch thuật: Đỗ Văn Dũng

TP Hồ Chí Minh
Tháng 1/2001

MỤC LỤC

LỜI NÓI ĐẦU

Mục đích của việc viết cuốn sách này là đưa những phương pháp thống kê y học đa dạng áp dụng trong nghiên cứu y khoa vào trong thực hành, và trong khi làm việc đó, tôi hi vọng là tôi đã kết hợp được sự đơn giản với tính sâu sắc. Tôi đã sử dụng một các sắp xếp các chủ đề khác hơn với hầu hết các sách giáo khoa khác, dựa trên tiến trình logic những khái niệm thực hành, hơn là dựa trên các bước phát triển của toán học hình thức. Ý tưởng thống kê được đưa vào khi cần thiết, và tất cả các phương pháp được mô tả trong bối cảnh của những ví dụ phù hợp được rút ra từ những tình huống thực sự. Có nhiều tham khảo qua lại để liên kết và đối chiếu những cách tiếp cận khác nhau có thể áp dụng trong những tình huống tương tự. Theo cách này, người đó sẽ được dẫn dắt mau hơn đến việc phân tích những vấn đề thực hành và sẽ dễ dàng nắm bắt được những thủ tục gì có thể được áp dụng khi nào.

Cuốn sách này là thích hợp để tự học, là bạn đồng hành cho những khóa giảng về thống kê y học hay là một tài liệu tham khảo. Nó bao gồm tất cả các chủ đề mà một nhà nghiên cứu y khoa hay một sinh viên có thể gặp phải. Một số những phương pháp cao cấp (hay hiếm) chỉ được mô tả ngắn gọn,và người đọc được đề nghị tham khảo những sách chuyên môn hơn. Dù vậy, chúng tôi hi vọng rằng, ít có trường hợp phải tìm kiếm một chủ đề trong chỉ mục và tìm không tìm được một lưu ý nào. Tất cả các công thức đề được nhấn mạnh một cách rõ ràng để dễ dàng tham khảo và có những tóm tắt hữu dụng của những phương pháp ở bìa sách.

Cuốn sách này là sự giới thiệu ngắn gọn và trực tiếp những phương pháp và ý tưởng cơ bản của thống kê y khoa. Dù vậy, nó không dừng ở đó. Nó có mục đích là một hướng dẫn viên toàn diện về chủ đề. Đối với ai thực sự quan tâm đến áp dụng thống kê, sẽ là không đủ nếu chỉ có thể tiến hành, thí dụ như, kiểm định t. Nó cũng quan trọng để đánh giá những hạn chế của phương pháp đơn giản và biết chúng có thể được mở rộng khi nào và như thế nào. Vì lí do này, có những chương như phân tích phương sai và hồi quy đa biến đã được đưa vào. Khi giải quyết với những phương pháp cao cấp, giải pháp chú trọng đến những nguyên lí có liên quan và việc lí giải kết quả, bởi vì sự có mặt rộng rãi những phương tiện tính toán, do đó việc làm quen với những chi tiết của tính toán không còn cần thiết nữa. Những phần cao cấp hơn có thể được bỏ qua trong lần đọc đầu, như đã chỉ ra trong những phần thích hợp của bài. Dù vậy, chúng tôi đề nghị phần mở đầu của tất cả các chương cần được đọc bởi vì nó cho phép đưa các phương pháp khác nhau vào bối cảnh.

Người đọc cũng sẽ tìm thấy những chủ đề như test khuynh hướng cho bảng nhiều chiều, phương pháp chuẩn hóa, sử dụng phép biến đổi, phân tích sống còn và nghiên cứu bệnh chứng. Phần tư cuối của cuốn sách để dành cho những chủ đề liên quan đến việc thiết kế và tiến hành nghiên cứu. Phần này không tách rời khỏi phần phướngphap phân tích và phản ánh tầm quan trọng của nhận thức thống kê thông qua thực hiện nghiên cứu. Có một tóm tắt chi tiết làm thế nào để quyết định cỡ mẫu thích hợp và việc đưa vào sử dụng máy vi tính, trong đó có giải thích nhiều từ chuyên môn.

Cuốn sách này là sự kết hợp của nhiều năm kinh nghiệm giảng dạy thống kê cho nhiều người chuyên môn ngành y và kinh nghiệm cộng tác nghiên cứu. Tôi hi vọng cách tiếp cận đã được chọn lựa sẽ hấp dẫn cho bất kì ai làm việc trong hay liên quan đến lãnh vực và sẽ làm hài lòng cả những người chuyên môn y khoa cũng như những nhà thống kê. Đặc biệt, tôi hi vọng kết quả sẽ trả lời những nhu cầu của nhiều người cho rằng vấn đề tiến hành công việc thống kê không phải là cơ chế của một kiểm định đặc hiệu, mà là biết được phương pháp nào được áp dụng khi nào.

Tôi muốn bày tỏ lòng biết ơn đến những đồng nghiệp, sinh viên và bạn bè đã hỗ trợ tôi trong nhiệm vụ này. Đặc biệt, tôi muốn cám ơn David Ross và Cesar Victoria đã sẵn sàng độc bản thảo và đã góp ý hết sức chi tiết, Richard Hayes cho nhiều lần thảo luận về giảng dạy trong nhiều năm, Laura Rodrigues đã chia xẻ sự hiểu biết sâu sắc về phương pháp dịch tễ cho tôi, Peter Smith đã góp ý và nâng đỡ chung, Helen Edwards cho sự giúp đỡ kiên nhẫn và lành

nghề trong công tác đánh máy và Jacqui Wright cho việc giúp đỡ trong soạn thảo những bảng phụ lục. Tôi cũng muốn cám ơn chồng tôi là Tom Kirkwood không những chỉ góp ý cho những bản thảo, vô vàn cuộc thảo luận và những giúp đỡ thực tế, mà còn bởi vì sự hỗ trợ và khuyến khích không ngừng. Tôi muốn đề tặng cuốn sách này cho Tom. Cuối cùng tôi muốn nhắc đến Daisy và Sam Kirkwood, mặc dù sự ra đời của hai cháu đã làm chậm trễ việc kết thúc của bản thảo gần hoàn tất, nhưng đã cho tôi một cơ hội để có một cách nhìn mới mẻ vào những gì tôi đã viết và thực hiện những cải tiến quan trọng.

Betty Kirwood

London School of Hygiene and Tropical Medicine

CĂN BẢN

Thống kê là gì?

Thống kê là khoa học thu thập, tổng kết, trình bày và lí giải số liệu, và dùng chúng để kiểm định giả thuyết. Trong vài thập niên qua, thống kê đã đóng vai trò trung tâm ngày càng tăng trong các điều tra y khoa. Có nhiều lí do và 3 lí do chính như sau. Đầu tiên, thống kê cho phép tổ chức các thông tin trên cơ sở rộng hơn và căn bản hơn sự trao đổi các giai thoại và kinh nghiệm cá nhân. Thứ nhì, ngày càng nhiều các thứ có thể đo lường định lượng được trong y khoa. Thứ ba, có sự biến thiên rất lớn trong hầu hết các quá trình sinh học. Thí dụ, huyết áp không chỉ khác nhau từ người này đến người khác, mà trong cùng một người, nó cũng thay đổi từ ngày này sang ngày khác và từ giờ này sang giờ khác. Sự lí giải những số liệu khi có những biến thiên nằm ở trọng tâm của thống kê. Do đó, trong việc điều tra tỉ lệ bệnh tật liên hệ với một nghề nghiệp nhất định có nhiều kích xúc, phương pháp thống kê cần thiết để đánh giá có phải huyết áp trung bình quan sát được cao hơn huyết áp của dân số chung chỉ đơn giản là do sự biến thiên tình cờ hay nó phản ánh một nguy cơ sức khỏe nghề nghiệp thực sự.

Sự biến thiên có thể bắt nguồn từ các tác động ngẫu nhiên của sự tình cờ trong dân số. Cá nhân không phản ứng như nhau đối với cùng một kích thích. Do đó mặc dù, hút thuốc lá và uống rượu nói chung là có hại cho sức khỏe, người ta không hiếm khi nghe thấy một người hút thuốc lá và uống rượu nhiều sống khỏe mạnh tới già, trong khi một người chống rượu và không hút thuốc lại chết trẻ. Một thí dụ khác, đánh giá một vaccine mới. Cá nhân có thể thay đổi về sự đáp ứng với vaccine và sự nhậy cảm và tiếp xúc với bệnh. Không chỉ một số người nào đó không tiêm vaccine không bị bệnh mà một số người có tiêm vaccin có thể bị bệnh. Có thể kết luận được gì nếu phần trăm người không có bệnh cao hơn trong nhóm tiêm vaccine so với nhóm không tiêm vaccine? có phải vaccine có hiệu quả thực sự hay không? có thể kết quả chỉ do tình cờ? hay, có một số các sai lệch trong cách chọn cá nhân được tiêm chủng, thí dụ có phải họ khác nhau về tuổi tác hay giai cấp xã hội khiến cho nguy cơ mắc bệnh thấp hơn? phương pháp phân tích thống kê để phân biệt giữa hai khả năng đầu, trong khi việc lựa chọn thiết kế đúng sẽ loại trừ khả năng thứ ba. Thí dụ này minh họa sự hữu dụng của thống kê không chỉ nằm trong việc phân tích kết quả. Nó cũng có vai trò trong việc thiết kế và tiến hành nghiên cứu.

Dân số và mẫu

Có liên hệ với vấn đề cơ bản của sự biến thiên là một điểm quan trọng: trừ khi một cuộc tổng điều tra được tiến hành, số liệu chỉ là của một mẫu (sample) trong một nhóm lớn hơn được gọi là dân số (population). Mẫu được quan tâm không phải bởi vì chính nó mà bởi vì cái mà nó cho người điều tra biết về dân số. Bởi vì sự tình cờ, những mẫu khác nhau sẽ cho những kết quả khác nhau và điều này phải được xét đến khi dùng các mẫu để kết luận về dân số. Hiện tượng này được gọi là sự biến **thiên lấy mẫu (sampling variation), nằm ở trọng tâm của thống kê. Nó được** trình bày chi tiết ở Chương 3.

Từ 'dân số' được dùng trong thống kê có nghĩa rộng lớn hơn bình thường. Nó không chỉ gồm dân số người mà có thể dùng cho bất kì một tập hợp các đối tượng. Thí dụ, số liệu có thể là mẫu của 20 bệnh viện trong một dân số các bệnh viện của quốc gia. Trong trường hợp đó, dễ dàng có thể thấy rằng có thể liệt kê toàn bộ dân số và có thể chọn mẫu trực tiếp từ đó. Dù vậy trong nhiều trường hợp, dân số và giới hạn của nó không được chỉ rõ một cách chính xác và phải cẩn thận để đảm bảo rằng mẫu thực sự đại diện cho dân số cần lấy thông tin. Dân số này đôi khi được gọi là dân số mục tiêu (target population). Thí dụ, xem một cuộc thử nghiệm vaccine được tiến hành trong các sinh viên tự nguyện. Giả sử rằng đáp ứng với vaccine và tiếp xúc với bệnh tật của sinh viên là điển hình cho cộng đồng nói chung, kết quả có tính áp dụng tổng quát. Mặt khác nếu sinh viên khác về bất kì phương diện nào mà có thể tác động sự đáp ứng với vaccine và tiếp xúc với bệnh tật, kết luận về thử nghiệm chỉ giới hạn cho dân số

sinh viên và không có tính áp dụng tổng quát. Trong trường hợp này, dân số mục tiêu bao gồm không chỉ những người sống hiện nay mà cả những người sống trong tương lai. Hiển nhiên rằng không thể đếm các dân số như vậy.

Xác định dân số

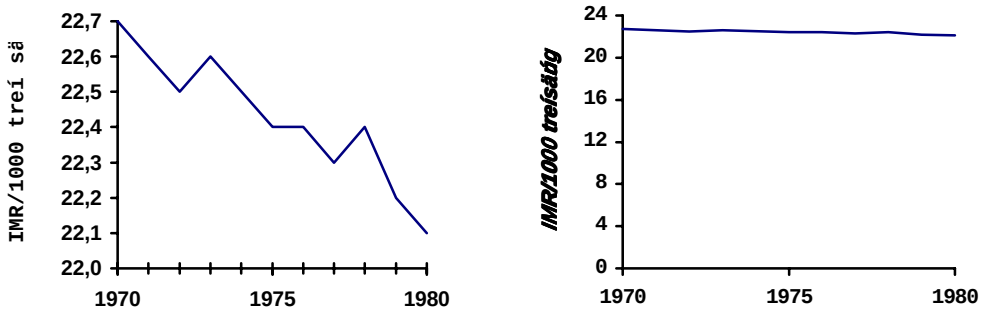
Các số liệu thô của điều tra bao gồm các quan sát (observations) trên các cá nhân. Trong nhiều trường hợp cá nhân là con người nhưng không nhất thiết như vậy. Thí dụ, cá nhân có thể là hồng cầu, mẫu nước tiểu, chuột, hay bệnh viện. Số các cá nhân được gọi la cỡ mẫu (sample size). Bất kì khía cạnh nào của cá nhân được đo lường, như huyết áp, hay được ghi nhận, như tuổi và giới tính, được gọi là biến **số** (variable). Có thể có một hay nhiều biến số trong một nghiên cứu.

Chia các biến số thành các loại khác nhau có ích bởi vì có thể áp dụng các phương pháp thống kê khác nhau cho mỗi loại. Cách chia tổng quát là chia thành biến **định tính - qualitative (biến phạm trù - catergorical) hay biến định lượng -** quantitative (biến số - numerical). Biến định tính là biến không phải là số như nơi sinh, nhóm dân tộc hay loại thuốc. Một loại đặc biệt là biến nhị phân (binary), trong đó đáp ứng chỉ là một trong hai khả năng. Thí dụ, giới tính là nam hay nữ, bệnh nhân còn sống hay chết. Biến định lượng là biến số và hoặc là rời rạc (discrete) hay liên tục (continous). Giá trị của biến rời rạc thường là số nguyên, như số các trường hợp bạch hầu trong một tuần. Một biến liên tục là sự đo lường trên thang liên tục. Thí dụ là chiều cao, cân nặng, huyết áp và tuổi.

Phân tích số liệu và trình bày kết quả

Phương pháp tổng kết và phân tích số liệu để lí giải kết quả của một nghiên cứu là căn bản của cuốn sách này. Có ba điểm chính cần nhấn mạnh ở đây. Thứ nhất là cần tránh áp dụng các phương pháp phức tạp chỉ vì để đạt được sự phức tạp. Điều quan trọng là bắt đầu bằng việc sử dụng các tổng kết căn bản và kĩ thuật đồ thị để thăm dò số liệu. Việc phân tích phải đi từ đơn giản đến phức tạp. Phải chọn phương pháp đơn giản nhất phù hợp với yêu cầu của số liệu.

Điểm thứ nhì có liên quan là phải ứng dụng các lí luận thống kê cùng với lí trí. Điều quan trọng là không để mất nhận thức vào con số, các yếu tố tác động đến chúng và chúng đại diện cho cái gì trong khi thao tác con số trong quá trình phân tích. Bradford Hill (1977), Colton (1974), và Oldham (1968) đã có những chương rất hay minh họa các ngụy biện phổ biến và các khó khăn xuất phát trong việc lí giải số liệu.



Hình 1.1 Giảm tỉ suất tử vong trẻ em từ 1970 đến 1980 (a) chọn thang đo không phù hợp làm khuếch đại sai lầm mức giảm (b) dùng thang đo đúng.

Điểm thứ ba là nên dùng các kĩ thuật đồ thị (graphical techniques) cả trong giai đoạn thăm dò phân tích và trình bày kết quả, bởi vì sự quan hệ, khuynh hướng, và sự tương phản thường dễ nhận biết trong các giản đồ hơn từ trong bảng. Giản đồ (và bảng) phải luôn luôn được ghi tựa đề rõ ràng và dễ hiểu: không cần thiết phải đọc lại văn bản để hiểu chúng. Đồng thời chúng

không được lộn xộn với quá nhiều chi tiết và chúng không được gây mơ hồ. Các điểm gẫy và không liên tục trong thang đo phải được đánh dấu rõ ràng và, nếu được, cần phải tránh. Hình 1.1 (a) cho thấy dạng thể hiện sai thường gặp do sử dụng thang đo không phù hợp. Giảm tỉ suất chết trẻ em được làm thấy nhiều lên bằng cách mở rộng trục tung, trong khi thực tế sự giảm trong 10 năm chỉ rất ít (từ 22,7 đến 22,1/1000 sinh sống/năm). Một cách trình bày chân thực hơn ở trong Hình 1.1(b) với trục đứng bắt đầu từ 0.

Chọn máy tính cầm tay

Một máy tính cầm tay (calculator) cần thiết cho các ứng dụng thống kê dù là đơn giản nhất. Có một số loại máy khác nhau với nhiều giá cả khác nhau. Các phương tiện dưới đây được coi là tối thiểu

1. Các hàm toán học như lấy căn, logarithm và giai thừa

2. Tối thiểu có một bộ nhớ

3. Tính tự động trung bình và độ lệch chuẩn

4. Tính tự động tương quan và hồi quy tuyến tính

5. Phương tiện lập trình, với khả năng giữ tối thiểu 100 bước lập trình, còn giữ lại khi đã tắt máy tính. Khả năng này phải đủ để cho phép 2 chương trình thường trú trong máy tính đảm bảo sử dụng hai kiểm định thống kê phổ biến nhất, kiểm định t để so sánh 2 trung bình (xem Chương 7) và kiểm định chi bình phương để so sánh hai tỉ lệ (xem chương 13).

Có thể tìm được một máy tính tương đối rẻ tiền (khoảng 30 Bảng Anh) thỏa mãn các điều kiện trên. Các máy mắc tiền hơn có thể có lợi ích là tăng số bước lập trình và khả năng giữ các chương trình trong các vật thể kí tin bên ngoài như thẻ từ tính hoặc băng cassette.

TẦN SUẤT, PHÂN PHỐI TẦN SUẤT VÀ TỔ CHỨC ĐỒ

Giới thiệu

Bước đầu tiên trong phân tích là tổng kết số liệu, bởi vì số liệu không được tổ chức sẽ rất khó hiểu. Minh họa số liệu bằng một giản đồ có ghi tựa đề rõ ràng và dễ hiểu sẽ hữu ích.

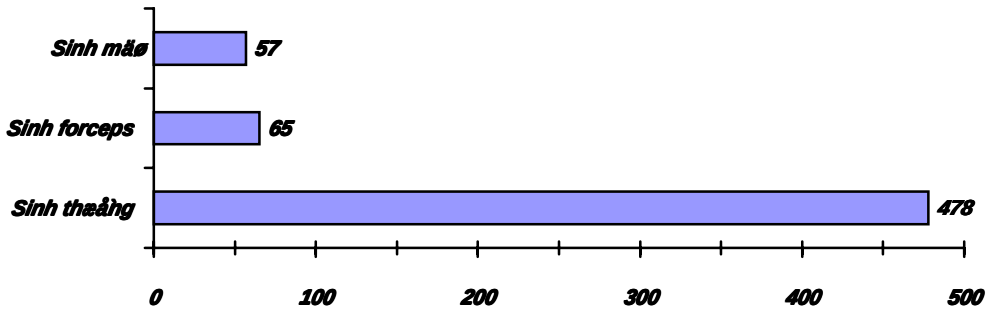
Tần suất (số liệu định tính)

Tổng kết số liệu định tính rất dễ dàng, nhiệm vụ đầu tiên là đếm các quan sát trong mỗi phạm trù. Số quan sát đếm được được gọi là tần suất (frequencies). Chúng thường được trình bày thành tần suất tương đối (relative frequencies), là phần trăm so với tổng số các cá nhân. Thí dụ, bảng 2.4 tổng kết phương pháp đỡ đẻ được ghi nhận trong 600 trường hợp sinh trong bệnh viện. Biến số cần quan tâm là phương pháp đỡ đẻ, một biến định tính có 3 phạm trù, sinh thường, sinh forceps và sinh mổ

Bảng 2.1. Phương pháp đỡ đẻ 600 em bé sinh trong bệnh viện

Phương pháp đỡ đẻ	Số sinh	phần trăm
Sinh thường	478	79,7
Sinh forceps	65	10,8
Sinh mổ	57	9,5
Tổng số	600	100,0

Tần suất và tần suất tương đối thường được minh họa bằng giản đồ thanh (bar diagram) (xem hình 2.1) hay đồ thị hình bánh (pie chart) (xem hình 2.2). Trong giản đồ thanh, chiều dài của thanh được vẽ tỉ lệ với tần suất và trong đồ thị hình bánh, vòng tròn được chia sao cho diện tích của mỗi phần tỉ lệ với tần suất



Hình 2.1 Giản đồ thanh trình bày phương pháp đỡ đẻ 600 trẻ sinh trong bệnh viện.

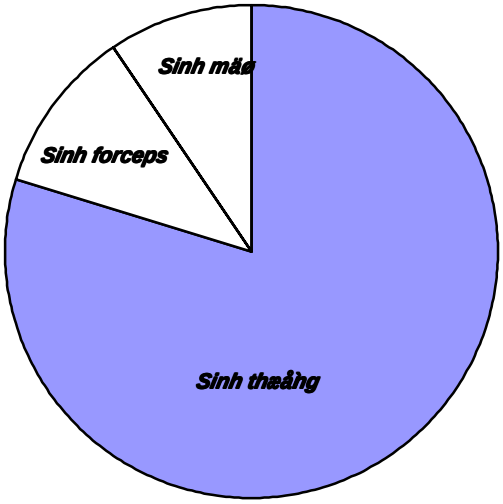
Phân phối tần suất (số liệu định lượng)

Nếu có nhiều hơn 20 quan sát, bước đầu tiên có ích trong việc tổng kết số liệu định lượng là thành lập phân phối tần suất (frequency distribution). Đó là bảng trình bày số các quan sát ở các giá trị khác nhau hay trong các khoảng giá trị nhất định. Đối với biến rời rạc, tần suất có thể lập bảng hoặc là cho mỗi giá trị của biến hoặc là cho một nhóm các giá trị. Với biến liên tục, phải thành lập nhóm. Hình 2.2. trình bày một thí dụ, trong đó hemoglobin được đo lường tới 0,1g/100 ml vvvà nhóm 11- gồm tất cả các đo lường ở giữa 11,0 và 11,9g/100 ml.

Khi thành lập phân phối tần suất, điều đầu tiên cần làm là đếm số các quan sat và xác định giá trị lớn nhất và nhỏ nhất. Sau đó quyết định số liệu có cần phân nhóm hay không và nếu có phải dùng khoảng phân nhóm nào. Nói chung người ta chia thành 5-20 nhóm tùy theo số các

quan sát. Nếu khoảng được chọn cho việc phân nhóm quá rộng, nhiều chi tiết sẽ bị mất đi, trong khi nếu khoảng quá nhỏ, bảng sẽ khó sử dụng. Điểm đầu tiên của nhóm phải là số chẵn và chiều rộng của các khoảng phải bằng nhau nếu có thể. Bảng phải được kí hiệu sao cho có thể quyết định khi quan sát nằm ở ranh giới.

Thí dụ, trong bảng 2.2, có 70 đo lường hemoglobin. Giá trị nhỏ nhất là 8,8 và lớn nhất là 15,1 g/100ml. Chọn chiều rộng khoảng là 1g/100ml sẽ cho 8 nhóm trong phân phối tần suất. Đặt tên nhóm 8-, 9- là rõ ràng. Có thể đặt tên là 8,0-8,9, 9,0-9,9 v.v... Lưu ý rằng đặt tên 8- 9, 9-10 là không rõ bởi vì người ta không biết đo lường 9,0g/100ml thuộc nhóm nào.



Hình 2.2 Đồ thị hình bánh trình bày phương pháp đỡ đẻ 600 trẻ sinh trong bệnh viện.

Bảng 2.2 Nồng độ hemoglobin ở 70 phụ nữ (đơn vị g/100 ml)

(a) Số liệu thô (gạch dưới giá trị lớn nhất và nhỏ nhất)

10.2	13.7	10.4	14.9	11.5	12.0	11.0
13.3	12.9	12.1	9.4	13.2	10.8	11.7
10.6	10.5	13.7	11.8	14.1	10.3	13.6
12.1	12.9	11.4	12.7	10.6	11.4	11.9
9.3	13.5	14.6	11.2	11.7	10.9	10.4
12.0	12.9	11.1	8.8	10.2	11.6	12.5
13.4	12.1	10.9	11.3	14.7	10.8	13.3
11.9	11.4	12.5	13.0	11.6	13.1	9.7
11.2	15.1	10.7	12.9	13.4	12.3	11.0
14.6	11.1	13.5	10.9	13.1	11.8	12.2

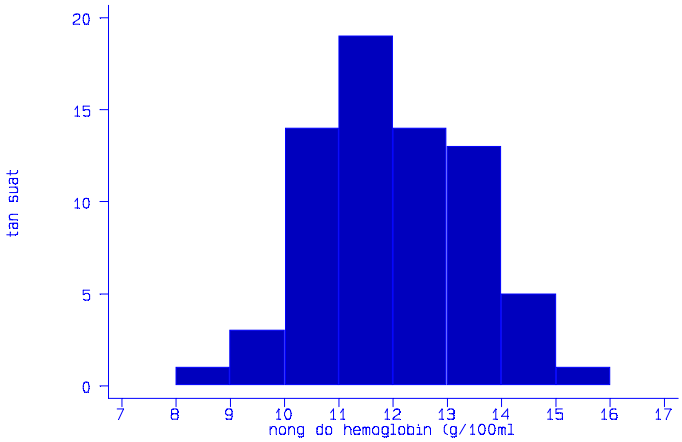
(b) phân phối tần suất

Hemoglobin (g/100ml)	đánh dấu	số phụ nữ	phần trăm
8-	1	1	1.4
9-	111	3	4.3
10-	~~1111~~ ~~1111~~ 1111	14	20.0
11-	~~1111~~ ~~1111~~ ~~1111~~ 1111	19	27.1
12-	~~1111~~ ~~1111~~ 1111	14	20.0
13-	~~1111~~ ~~1111~~ 111	13	18.6
14	~~1111~~	5	7.1
15-15.9	1	1	1.4
Tổng số		70	100.0

Khi đã quyết định dạng thức của bảng, có thể đếm các số trong mỗi nhóm. Có thể tránh được sai lầm bằng cách tiến hành số liệu theo thứ tự. Đối với một giá trị, đánh dấu vào nhóm thích hợp. Để dễ đếm, những đánh dấu này được xếp thành nhóm năm bằng cách gạch dấu thứ năm nằm ngang qua bốn dấu trước đó. Chúng được gọi là cổng năm thanh (five-bar gates). Quá trình này được gọi là đánh dấu (tallying) và được minh họa trong bảng 2.2(b).

Tổ chức đồ

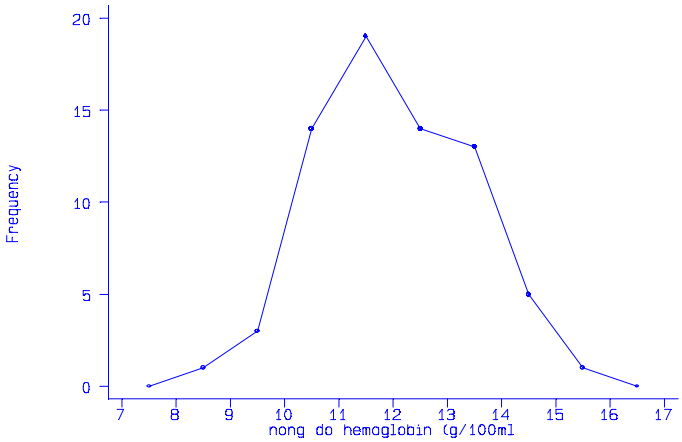
Phân phối tần suất thường được minh họa bằng tổ chức đồ (histogram) như được trình bày trong hình 2.3 về số liệu hemoglobin. Dù là dùng tần suất hay phần trăm, hình dạng của tổ chức đồ cũng như nhau.



Hình 2.3 Tổ chức đồ của nồng độ hemoglobin của 70 phụ nữ

Dễ dàng xây dựng tổ chức đồ khi các khoảng cách nhóm của phân phối tần suất bằng nhau như trong trường hợp hình 2.3. Nếu khoảng có chiều rộng khác nhau, cần phải lưu ý khi vẽ tổ chức đồ nếu không sẽ bị sai lệch. Thí dụ, giả sử hai nhóm hemoglobin cao nhất được kết hợp lại. Tần suất của nhóm kết hợp này (14,0-15,9 g/100ml) sẽ là 6, nhưng rõ ràng sẽ sai lầm nếu vẽ hình chữ nhật có chiều cao 6 từ 14- 16g/100ml. Bởi vì khoảng này lớn gấp đôi chiều rộng khác khoảng khác, chiều cao của đường sẽ là 3, phân nửa của tần suất tổng cộng của nhóm này. Điều này được minh họa trong hình 2.3. Quy tắc chung để vẽ tổ chức đồ khi các khoảng không cùng chiều rộng là để chiều cao của hình chữ nhật tỉ lệ với tần suất chia cho chiều rộng, để cho diện tích của hình chữ nhật trong tổ chức đồ tỉ lệ với tần suất.

Đa giác tần suất



Hình 2.4 Đa giác tần suất của nồng độ hemoglobin của 70 phụ nữ.

Một cách khác để minh họa phân phối tần suất nhưng kém phổ biến hơn là đa giác tần suất, được minh họa trong Hình 2.4. Nó đặc biệt có ích khi so sánh hai hay nhiều hơn các phân phối tần suất bằng cách cùng vẽ trên một giản đồ. Đa giác được vẽ bằng cách tưởng tượng (hay vẽ phác bằng chì) tổ chức đồ và nối các trung điểm của cạnh trên hình chữ nhật. Điểm cuối của đường vừa vẽ được nối với trục hoành ở điểm giữa của nhóm sát trên nhóm lớn nhất và điểm giữa của nhóm sát dưới nhóm nhỏ nhất. Đối với số liệu của hemoglobin đó là nhóm 7,0-7,9 và 16,0- 16,9g/100ml. Do đó trên hình 2.4 đa giác tần suất được nối với trục hoành ở 7,5 và 16,5g/100ml.

Phân phối tần suất của dân số

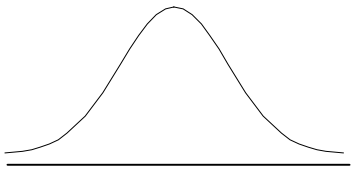
Hình 2.3 và 2.4 minh họa phân phối tần suất của hemoglobin của mẫu 70 phụ nữ. Chúng ta dùng số liệu này để cho thông tin về phân phối nồng độ hemoglobin trong phụ nữ nói chung.

Thí dụ, dường như rất ít khi phụ nữ có mức dưới 9,0g/100ml hay trên 15,0g/100ml. Sự tin cậy khi rút ra các kết luận tổng quát từ số liệu phụ thuộc vào có bao nhiêu cá nhân được đo lường. Mẫu được đo càng lớn, các khoảng cách nhóm được chọn càng nhỏ thì tổ chức đồ và đa giác tần suất trở nên mịn hơn và càng giống phân phối của dân số tổng quát. Nếu có thể biết được nồng độ hemoglobin của toàn dân số phụ nữ, giản đồ tạo được sẽ trở thành một đường cong trơn.

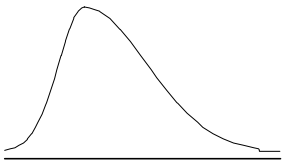
Hình dạng của phân phối tần suất

Hình 2.5 trình bày 3 hình dạng phân phối tần suất phổ biến nhất. Chúng có tần suất cao ở trung tâm và tần suất thấp ở 2 đầu, được gọi là đuôi trên hay đuôi dưới (upper and lower tails) của phân phối. Phân phối của hình 2.5(a) được gọi là đối **xứng (symmetrical) qua tâm; dạng đường cong này thường được gọi là 'hình** chuông'. Hai phân bố kia được gọi là bất đối xứng hay lệch (skewed). Đuôi trên của phân phối trên Hình 2.5(b) dài hơn đuôi dưới; nó được gọi là lệch dương hay lệch về phía phải. Phân phối của hình 2.5(c) là lệch âm hay lệch về phía trái.

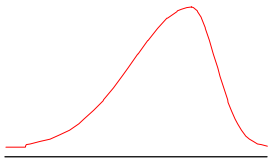
Tất cả phân phối trong hình 2.5 là một yếu vị (unimodal) bởi vì chúng chỉ có một đỉnh. Hình 2.6(a) trình bày phân phối tần suất hai yếu vị (bimodal), đó là phân phối có 2 đỉnh. Đôi khi ta thấy được phân phối này và nó cho thấy số liệu là hỗn hợp của hai phân phối riêng biệt. Hình 2.6 trình bày hai phân phối khác ít gặp khác; đó là phân phối hình J ngược (reverse J-shaped) và đồng nhất (uniform).



(a) đối xứng và hình chuông
td: chiều cao

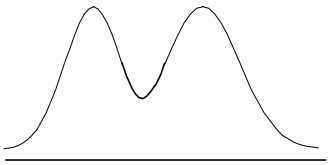


(b) lệch dương hay lệch phải
td: bề dày lớp mỡ dưới da

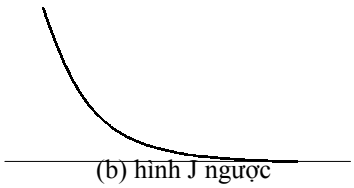


(c) lệch âmhay lệch trái
td: thời gian thai kì

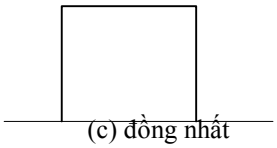
Hình 2.5 Ba dạng phân phối phổ biến và ví dụ của mỗi loại



(a) hai yếu vị
td: nồng độ hormone ở nam và nữ



(b) hình J ngược
td: thời gian sống sau khi chẩn đoán ung thư phổi



(c) đồng nhất
td: sự xuất hiện bệnh không theo mùa

Hình 2.6 Ba dạng phân phối ít phổ biến và ví dụ của mỗi loại

TRUNG BÌNH, ĐỘ LỆCH CHUẨN VÀ SAI SỐ CHUẨN

Giới thiệu

Phân phối tần suất cho một bức tranh tổng quát về giá trị của các biến số. Dù vậy sẽ tiện lợi hơn nếu tổng kết các biến định lượng bằng cách chỉ cho 2 số đo: giá trị trung bình và sự trải rộng của giá trị.

Trung bình, trung vị và yếu vị

Giá trị trung bình thường được thể hiện bằng trung bình cộng (arithmetic mean), thường được gọi là trung bình. Đó là tổng số các giá trị chia cho số các giá trị

$$\text{Trung bình, } \bar{x} = \frac{\sum x}{n}$$

Trong đó x biểu thị giá trị của biến số, S (mẫu tự tiếng Hy lạp sigma hoa) có nghĩa là tổng của và n là số các quan sát. Trung bình được kí hiệu là $\bar{x}$ (đọc là 'x gạch').

Một số đo khác của giá trị trung bình là trung vị (median) và yếu vị (mode). Trung vị là giá trị chia phân phối làm đôi. Nếu các giá trị được sắp theo thứ tự tăng dần, trung vị là quan sát ở chính giữa.

$$\text{Trung vị = giá trị ở vị trí } \frac{(n+1)}{2} \text{ trong các quan sát được sắp thứ tự}$$

Nếu có một số chẵn các quan sát, không có quan sát ở chính giữa thì người ta lấy trung bình của 2 quan sát ở giữa. Yếu vị (mode) là giá trị xảy ra thường xuyên nhất

Thí dụ 3.1

Số liệu sau là thể tích huyết tương của 8 người đàn ông khỏe mạnh

2,75 2,86 3,37 2,76 2,62 3,49 3,05 3,12 lít

(a) n = 8

$\Sigma x = 2,75 + 2,86 + 3,37 + 2,76 + 2,62 + 3,49 + 3,05 + 3,12 = 24,021$

Trung bình, $\bar{x} = \Sigma x/n = 24,02/8 = 3,001$

(b) sắp xếp lại các số đo theo thứ tự tăng dần

2,62; 2,75; 2,76; 2,86; 3,05; 3,12; 3,37; 3,49

Trung vị = giá trị thứ (n+1)/2 = 9/2 = giá trị thứ 4,5

= trung bình của giá trị thứ 4 và thứ 5 = (2,86+3,05)/2 = 2,96

(c) không có ước lượng của yếúu vị bởi vì các giá trị đều khác nhau

Trung bình thường là số đo được chọn lựa bởi vì nó tính đến mỗi quan sát cá nhân và có thể được xử lí bằng kĩ thuật toán và thống kê. Trung vị là số đo mô tả hữu ích nếu có một hoặc hai giá trị quá cao hoặc quá thấp, làm cho trung bình không đại diện được đa số số liệu. Yếu vị ít khi được dùng. Nếu mẫu nhỏ thì có thể không ước lượng được yếu vị (như trong ví dụ 3.1c) hay ước lượng bị sai lệch. Trung bình, trung vị và yếu trị, nói chung là bằng nhau khi phân phối đối xứng và có một yếu vị. Khi phân phối bị lệch dương, trung bình nhân (geomtric mean) thích hợp hơn trung bình cộng. Điều này được thảo luận ở Chương 19.

Số đo sự biến thiên

Số đo sự biến thiên đơn giản nhất là phạm vi (range), đó là hiệu số giữa giá trị lớn nhất và nhỏ nhất. Khuyết điểm của nó là chỉ dựa trên hai quan sát và không cho ý niệm về cách các quan sát khác sắp xếp ra sao. Tương tự, khi cỡ mẫu càng lớn thì phạm vi càng lớn.

Bởi vì sự biến thiên nhỏ khi các quan sát tập trung gần chung quanh trung bình và lớn khi các quan sát phân tán trên một phạm vi đáng kể, sự biến thiên thường được đo lường theo độ lệch (deviation) của các quan sát so với trung bình. Phương sai (variance) là trung bình của bình phương những hiệu số này. Khi tính phương sai của một mẫu, tổng của độ lệch bình phương được chia cho (n-1) chứ không phải cho n bởi vì như vậy sẽ cho một ước lượng tốt hơn của phương sai dân số toàn bộ.

$$\text{Phương sai, } s^2 = \frac{\sum (x-\bar{x})^2}{(n-1)}$$

Độ tự do

Mẫu số (n-1) được gọi là độ tự do (degrees of freedom) của phương sai. Con số này là (n-1) chứ không phải là n, bởi vì chỉ có (n-1) độ lệch (x- $\bar{x}$) độc lập với nhau. Độ lệch cuối cùng có thể được tính từ các độ lệch khác bởi vì tổng tất cả các độ lệch bằng zero.

Độ lệch chuẩn

Phương sai có các tính chất toán học thuận lợi và là số đo thích hợp khi nghiên cứu lí thuyết thống kê. Dù vậy, nó có một khuyết điểm là nó có đơn vị là bình phương đơn vị của quan sát. Thí dụ, nếu quan sát là trọng lượng tính bằng gram thì phương sai là gram bình phương. Trong nhiều trường hợp sẽ thuận lợi hơn khi biểu thị độ biến thiên theo đơn vị ban đầu bằng cách lấy căn của phương sai. Nó được gọi là **độ lệch chuẩn (standard deviation - SD).**

$$SD \text{ hay s} = \sqrt{\left[\frac{\sum (x-\bar{x})^2}{(n-1)}\right]}$$

Hay tương đương

$$SD \text{ hay s} = \sqrt{\left[\frac{\sum x^2 - (\sum x)^2 / n}{(n-1)}\right]}$$

Công thức sau tiện lợi hơn cho việc tính toán bởi vì không cần phải tính trung bình và sau đó trừ các giá trị quan sát cho trung bình. Tương đương của hai công thức trên được minh họa trong thí dụ 3.2 (lưu ý: nhiều máy tính cầm tay có những hàm để tính trung bình và độ lệch chuẩn. Các phím bấm thường được kí hiệu bằng $\bar{x}$ và σ_{n-1}, trong đó σ là mẫu tự Hi lạp sigma thường).

Thí dụ 3.2

Bảng 3.1 trình bày các bước của tính toán độ lệch chuẩn của 8 số đo thể tích huyết tương ở thí dụ 3.1. Lưu ý rằng

$$\Sigma x^2 - (\Sigma x)^2/n = 72{,}7980 - (242)^2/8 = 0{,}6780$$

Cho kết quả giống như $\Sigma (x- \bar{x})2$:

$$s = 0{,}6780/7 = 0{,}311$$

Bảng 3.1 Tính toán độ lệch chuẩn của thể tích huyết tương của 8 đàn ông khỏe mạnh (giống như trong thí dụ 3.1). Trung bình, $\bar{x}$=3,001

	Thể tích huyết tương x	Độ lệch khỏi trung bình $x-\bar{x}$	Bình phương độ lệch $(x-\bar{x})^2$	bình phương của quan sát x^2
	2.75	-0.25	0.0638	7.5625
	2.86	-0.14	0.0203	8.1796
	3.37	0.37	0.1351	11.3569
	2.76	-0.24	0.0588	7.6176
	2.62	-0.38	0.1463	6.8644
	3.49	0.49	0.2377	12.1801
	3.05	0.05	0.0023	9.3025
	3.12	0.12	0.0138	9.7344
Tổng	24.02	0.00	0.6780	72.7980

Lí giải

Thông thường 70% quan sát nằm trong phạm vi một độ lệch chuẩn so với kể từ trung bình và khoảng 95% nằm trong phạm vi hai độ lệch chuẩn. Các con số này dựa trên một phân phối tần suất lí thuyết được gọi là phân phối bình thường, được mô tả ở chương 4.

Hệ số biến thiên (Coefficient of variation)

$$\text{c.v.} = \frac{s}{\bar{x}} \times 100\%$$

Hệ số biến thiên là độ lệch chuẩn tính theo phần trăm của trung bình mẫu. Chúng hữu ích khi cần quan tâm đến độ lớn của sự biến thiên so với độ lớn của quan sát, và nó có ưu điểm là hệ số biến thiên độc lập với đơn vị của quan sát. Thí dụ giá trị của độ lệch chuẩn của các trọng lượng sẽ khác nhau tùy theo chúng được đo lường theo kilogram hay pound. Dù vậy, hệ số biến thiên sẽ giống như nhau.

Tính toán trung bình và độ lệch chuẩn từ phân phối tần suất

Bảng 3.2 trình bày phân phối của số các lần mang thai trước của một nhóm phụ nữ khám tiền sản. Mười tám trong 100 phụ nữ không có mang trước đó, 27 đã có mang một lần, 31 có mang hai lần, 19 có mang 3 lần và 5 phụ nữ có mang 4 lần. Vì cộng 2 ba mươi mốt lần cũng giống như tích của (2 x 31), tổng số của các lần có thai trước đó được tính bằng:

$$\Sigma x=(0\times18)+(1\times27)+(2\times31)+(3\times19)+(4\times5)=0+27+62+57+20=166$$

Do đó số trung bình của các lần mang thai trước đó là

$$\bar{x}= 166/100=1,66$$

Tương tự

$$\Sigma x^2= (0\times18)+(1\times27)+(2^2\times31)+(3^2\times19)+(4^2\times5) = 0+27+124+171+80=402$$

Do đó độ lệch chuẩn

$$s= \sqrt{}\ [(402 -1662/100)/99] = [126,44/99] = 1,13$$

Bảng 3.2 Phân phối của số các lần có thai trước của một nhóm phụ nữ tuổi từ 30 đến 34 đến khám tại phòng khám tiền sản

	Số lần có thai					
	0	1	2	3	4	Tổng số
Số phụ nữ	18	27	31	19	5	100

Nếu các biến số được phân nhóm để xây dựng phân phối tần suất, cần phải tính trung bình và độ lệch chuẩn từ các giá trị nguyên thủy chứ không dùng phân phối tần suất. Dù vậy, đôi khi chỉ có phân phối tần suất. Trong trường hợp đó, giá trị xấp xỉ của trung bình và phương sai có thể tính được bằng cách dùng giá trị trung điểm của nhóm và tiến hành như trên.

Thay đổi đơn vị

Cộng hay trừ quan sát cho một hằng số làm trung bình cũng cộng hay trừ hằng số đó nhưng không thay đổi độ lệch chuẩn. Nhân hay chia các quan sát cho một hằng số làm trung bình và độ lệch chuẩn cũng nhân hay chia cho hằng số đó.

Thí dụ, giả sự nhiệt độ được chuyển từ độ Fahrenheit thành Celsius bằng cách trừ cho 32 và nhân cho 5 và chia cho 9. Trung bình mới sẽ được tính từ trung bình cũ theo cách tương tự như vậy: trừ cho 32, nhân 5 và chia cho 9. Độ lệch chuẩn mới là độ lệch chuẩn cũ nhân 5 và chia cho 9 bởi vì phép trừ không tác động đến độ lệch chuẩn.

Sai số lấy mẫu và sai số chuẩn

Như đã nói ở Chương 1, mẫu được quan tâm không phải vì chính nó mà bởi vì nó nói cho người nghiên cứu về dân số mà nó đại diện. Trung bình mẫu, $\bar{x}$, và độ lệch chuẩn,s , được dùng để ước lượng trung bình và độ lệch chuẩn của dân số, kí hiệu bằng chữ Hi lạp μ (mu) và s (sigma). Trung bình mẫu không thể chính xác bằng trung bình dân số. Một mẫu khác sẽ cho ước lượng khác, sự khác biệt là do sự biến thiên lấy mẫu. Giả sử tiến hành lấy nhiều mẫu độc lập có cỡ bằng nhau và tính trung bình mẫu cho mỗi mẫu và tạo phân phối tần suất của các trung bình đó. Trung bình của phân phối sẽ bằng với trung bình của dân số và có thể chứng minh rằng độ lệch chuẩn sẽ bằng s /√ n. Nó được gọi là sai số chuẩn của trung **bình mẫu (standard error of the sample mean) và nó đo lường trung bình của dân số** được ước lượng bởi trung bình mẫu chính xác tới mức nào. Độ lớn của sai số chuẩn phụ thuộc vào sự biến thiên trong dân số và cỡ mẫu. Mẫu càng lớn thì sai số chuẩn càng nhỏ.

Chúng ta ít khi biết được độ lệch chuẩn của dân số, s, và vì vậy chúng ta dùng độ lệch chuẩn mẫu để tính sai số chuẩn

$$\text{s.e.} = \frac{s}{\sqrt{n}}$$

Thí dụ 3.3

Trung bình của 8 thể tích huyết tương được trình bày trong bảng 3.1 là 3,001 (thí dụ 3.1) và độ lệch chuẩn là 0,311 (thí dụ 2). Sai số chuẩn của trung bình được tính bằng

$$s/\sqrt{n}=0{,}31/\sqrt{8}=0{,}111$$

Thí dụ 3.4

Hình 3.1 trình bày kết quả của một trò chơi trong một lớp học có 30 sinh viên để minh họa khái niệm biến thiên lấy mẫu, phân phối lấy mẫu và sai số chuẩn. Người ta đo lường huyết áp của 250 phi công. Phân phối của đo lường này được trình bày trong hình 3.1(a). Trung bình dân số, μ là 78,2mmHg và độ lệch chuẩn dân số, s, là 9,4mmHg. Mỗi giá trị được viết trên một đĩa nhỏ và 250 đĩa được đặt trong một cái túi. Mỗi sinh viên được đề nghị lắc túi chọn 10 đĩa và viết 10 huyết áp tâm trương. Bằng cách này ta có 30 mẫu khác nhau và 30 trung bình mẫu khác nhau, mỗi trung bình đều ước lượng cùng một trung bình dân số. Trung bình của những trung bình mẫu này là 78,23 mmHg, gần với trung bình dân số. Phân phối được trình

bày trong hình 3.1(b). Độ lệch chuẩn của trung bình mẫu là 31 mmHg, phù hợp với giá trị lí thuyết, s /√n=9,4/√10=2,97 mmHg sai số chuẩn của trung bình có cỡ mẫu là 10.

Bài tập được lập lại với cỡ mẫu 20, kết quả được trình bày trong hình 3.1(c). Dễ dàng thấy sự giảm biến thiên của trung bình mẫu do việc tăng cỡ mẫu từ 10 lên 20. Trung bình của trung bình mẫu là 78,14 mmHg cũng gần với trung bình dân số. Độ lệch chuẩn là 2,07 mmHg, cũng phù hợp với giá trị lí thuyết 9,4/√ 20=2,10 mmHg

Lí giải

Lí giải sai số chuẩn của trung bình mẫu tương tự như sai số chuẩn. Khoảng 95% trung bình mẫu có được bởi sự lấy mẫu lập lại sẽ nằm trong phạm vi hai độ lệch chuẩn so với trung bình dân số. Điều này được dùng để xây dựng một phạm vi giá trị khả dĩ của trung bình dân số, dựa trên các trung bình mẫu quan sát được và sai số chuẩn của nó. Những phạm vi như vậy được gọi là khoảng tin cậy (confidence interval). Phương pháp xây dựng khoảng tin cậy được trình bày ở Chương 5 bởi vì nó sử dụng đến phân phối bình thường, được mô tả ở Chương 4.

Sự hiệu chỉnh dân số giới hạn

Nếu cỡ mẫu trong một dân số có giới hạn, thí dụ như các căn nhà trong một làng, sai số lấy mẫu có thể nhỏ hơn s /√ n khi phần lớn dân số được lấy mẫu. Nó sẽ bằng 0 nếu toàn thể dân số được lấy mẫu không phải là do không có sự biến thiên trong các cá nhân trong dân số, nhưng bởi vì trung bình mẫu chính là trung bình dân số. Một mẫu thứ hai có cỡ tương tự (toàn dân số) sẽ có kết quả tương tự. Khi đó người ta áp dụng sự hiệu chỉnh dân số giới hạn (finite population correction) cho sai số chuẩn. Công thức trở thành

$$\text{s.e.với hiệu chỉnh dân số giới hạn} = \frac{\sigma}{\sqrt{n}}\sqrt{\left(1-\frac{n}{N}\right)}$$

Trong đó N là kích thước của dân số và n/N là phân số lấy mẫu (sampling fraction).

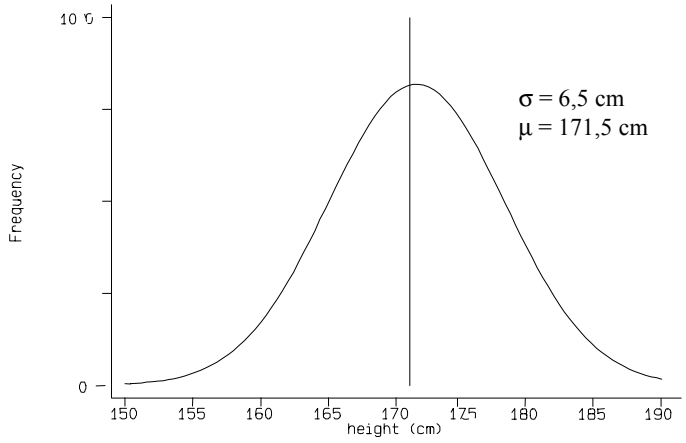
Bỏ qua sự hiệu chỉnh dân số giới hạn gây nên sự ước lượng thừa sai số chuẩn. Thí dụ, nếu 75% dân số được lấy mẫu, hiệu chỉnh dân số giới hạn sẽ bằng (1-0,75)=0,5. Nếu bỏ qua điều này, sai số chuẩn sẽ gấp đôi giá trị chính xác. Sự hiệu chỉnh ít có tác động và có thể bị bỏ qua khi phân số lấy mẫu nhỏ hơn 10%.

PHÂN PHỐI BÌNH THƯỜNG

Giới thiệu

Phân phối tần suất và các hình dạng của nó được thảo luận ở Chương 2. Trên thực tế người ta thấy rằng có thể mô tả hợp lí nhiều biến số bằng phân phối bình **thường (normal distribution), đôi khi còn được gọi là phân phối Gauss (Gaussian** distribution) theo tên của người phát hiện Gauss. Đường cong của phân phối bình thường đối xứng qua trung bình và có dạng hình chuông; hình chuông cao và hẹp đối khi độ lệch chuẩn nhỏ và thấp và rộng khi độ lệch chuẩn lớn. Hình 4.1 minh họa phân phối bình thường mô tả chiều cao của người lớn ở Anh. Một thí dụ khác về biến số được phân phối xấp xỉ bình thường là huyết áp, thân nhiệt và nồng độ hemoglobin. Thí dụ của các biến số không phân phối bình thường là bề dày lớp mỡ dưới da sau cánh tay và thu nhập, cả hai biến này đều bị lệch dương. Đôi khi biến đổi một biến, thí dụ như lấy logarithm sẽ làm phân phối trở thành bình thường. Điều này được mô tả ở Chương 19 và cách đánh giá xem một biến số có phân phối bình thường không được mô tả ở chương 18.

Phân phối bình thường quan trọng không chỉ bởi vì nó mô tả tốt các biến số mà còn bởi vì nó có một vai trò trọng tâm trong kĩ thuật phân tích thống kê. Thí dụ, nó là cơ sở lí luận cho việc tính toán khoảng tin cậy được trình bày ở chương 3 và được mô tả ở chương 5. Nó cũng là cơ sở cho phương pháp kiểm định mức ý nghĩa của trung bình được giới thiệu ở Chương 6. Vì những lí do này, điều quan trọng là mô tả việc ứng dụng phân phối bình thường một cách chi tiết trước khi trình bày tiếp mặc dù chúng ta không quan tâm đến phương trình toán học chính xác để định nghĩa bởi vì chúng ta đã có bảng.



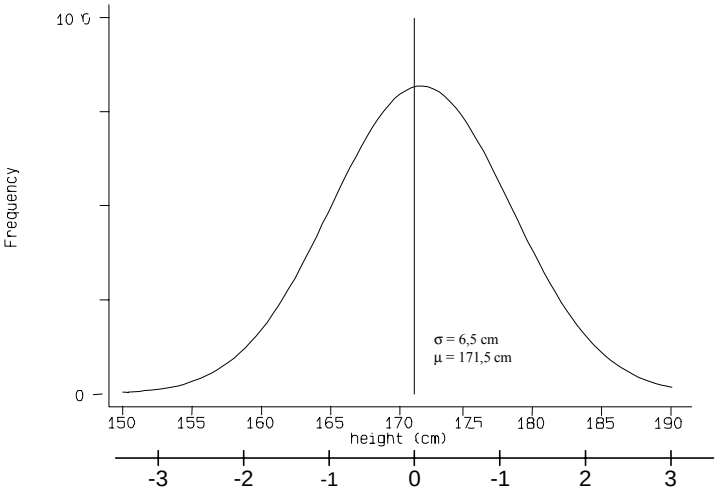
Hình 4.1 Giản đồ trình bày đường cong xấp xỉ bình thường mô tả chiều cao đàn ông trưởng thành

Phân phối bình thường chuẩn

Nếu một biến có phân phối bình thường thì việc đổi đơn vị không tác động đến chúng. Do đó dù chiều cao đo bằng centimetre hay bằng inch nó cũng phân phối bình thường. Thay đổi trung bình chỉ có nghĩa là chuyển đường cong qua lại trục trong khi thay đổi độ lệch chuẩn thay đổi chiều cao và chiều rộng của đường cong.

Đặc biệt, bằng cách thay đổi đơn vị, bất cứ một biến số có phân phối bình thường nào cũng có thể thành phân phối bình thường chuẩn (standard normal distribution - còn gọi là phân phối chuẩn) có trung bình bằng 0 và độ lệch chuẩn bằng 1. Có thể làm được điều này bằng cách trừ mỗi quan sát cho trung bình rồi chia cho độ lệch chuẩn. Quan hệ là

$$SND, z = \frac{x - \mu}{\sigma}$$



Hình 4.2 Quan hệ giữa phân phối bình thường theo đơn vị đo lường nguyên thủy và theo độ lệch bình thường chuẩn

SND = (chiều cao -171, 5)/6,5

Chiều cao = 171, 5 + (6, 5 x SND)

Trong đó x là biến nguyên thủy có trung bình m và độ lệch chuẩn s và z là độ **lệch bình thường chuẩn (standard normal deviate - SND). Điều này được minh họa** cho phân phối chiều cao đàn ông trong hình 4.2. Khả năng chuyển bất kì một biến có phân phối bình thường thành độ lệch bình thường chuẩn (SND) có nghĩa là chỉ cần một bảng cho phân phối bình thường chuẩn và không cần tất nhiều bảng cho tất cả các giá trị trung bình và độ lệch chuẩn khác nhau. Có hai cách phổ biến nhất để tạo thành bảng (i) diện tích dưới đường cong phân phối tần suất và (ii) điểm bách vị

Bảng tính diện tích dưới đường cong của phân phối bình thường

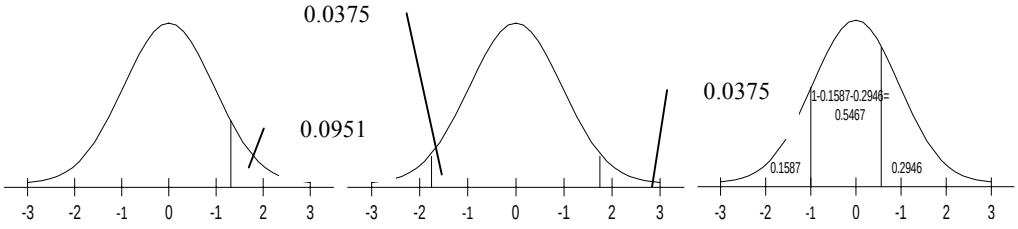
Bảng tính diện tích dưới đường cong phân phối bình thường của một phân phối bình thường hữu ích trong việc xác định tỉ lệ dân số có giá trị trong một phạm vi nhất định. Điều này đươc minh họa trong hình 4.1 và 4.2 về chiều cao của đàn ông ở Anh, có phân phối bình thường với trung bình μ =171,5 cm và độ lệch chuẩn s = 6,5 cm.

Diện tích ở đuôi trên của phân phối

Phân phối bình thường có thể được dùng để ước lượng tỉ lệ đàn ông cao hơn 180 cm. Tỉ lệ này được xem là phân số diện tích nằm dưới đường cong phân phối tần suất ở bên phải 180 cm. Độ lệch bình thường chuẩn tương ứng là

$$z = \frac{180 - 171.5}{6.5} = 1.31$$

Điều này tương đương với tỉ lệ diện tích của phân phối bình thường chuẩn ở bên phải 1,31. Diện tích này được minh họa trên hình 4.3 (a) và có thể tìm thấy từ bảng A1. Hàng của bảng chỉ giá trị z với một số lẻ và cột chỉ số lẻ thứ hai. Do đó diện tích trên 1,31 được ghi ở hành 1,3 và cột 0,01 và do đó là 0,0951. Chúng ta có thể kết luận 0,0951 hay 9,51% đàn ông cao hơn 180 cm.



(a) diãûn têch trãn z=1,31 tçm trong baíng A1

(b) diãûn têch dæåïi z=1,77 bàòng diãûn têch trãn z=1,77 do âäúi xæïng

(c) diãûn têch giæîa z=1 vaì z=0,54 tçm bàòng caïch træì

Hình 4.3 thí dụ tính toán các diện tích của phân phối bình thường chuẩn

Diện tích ở đuôi dưới của phân phối

Tỉ lệ đàn ông thấp hơn 160 cm có thể được ước tính tương tự

$$z = \frac{160 - 171.5}{6.5} = -1.77$$

Diện tích cần thiết được minh họa ở hình 4.3(b). Bởi vì phân phối bình thường chuẩn là đối xứng qua 0 nên diện tích bên trái z=-1,77 cũng bằng diện tích bên phải z=1,77 và là 0,0375. Do đó 3,75 đàn ông thấp hơn 160 cm.

Diện tích phân phối giữa hai giá trị

Tỉ lệ đàn ông có chiều cao giữa 165cm và 175 cm được ước tính bằng cách tìm tỉ lệ đàn ông thấp hơn 165cm và cao hơn 175 cm và lấy 1 trừ đi chúng. Điều này được minh họa bởi hình 4.3 (c)

(i) độ lệch bình thường chuẩn tương ứng với 165 cm là

$$z = \frac{165 - 171.5}{6.5} = -1$$

Tỉ lệ dưới chiều cao này là 0,1587

(ii) độ lệch bình thường chuẩn tương ứng với 175 cm là

$$z = \frac{175 - 171.5}{6.5} = 0.54$$

Tỉ lệ trên chiều cao này là 0,2946

(iii) Tỉ lệ đàn ông có chiều cao giữa 165 cm và 175 cm là

1 - tỉ lệ dưới 165 cm - tỉ lệ trên 175 cm

= 1 -0,1587 -0,2946 = 0,5467 hay 54,67%

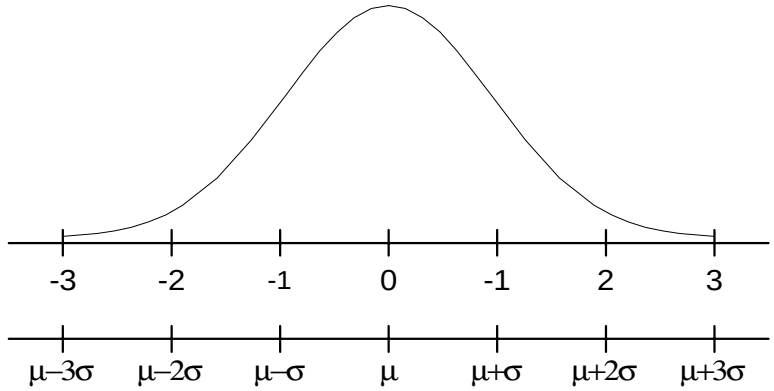
Giá trị tương ứng với một diện tích đuôi nhất định

Bảng A1 có thể dùng theo cách khác, đó là bắt đầu với diện tích và tìm điểm z tương ứng. Thí dụ, chiều cao nào thấp hơn 5% chiều cao của dân số? Hãy nhìn vào bảng tìm giá trị gần nhất với 0,05 ở hàng 1,6 và cột 0,04 vậy giá trị z cần thiết là 1,64. Chiều cao tương ứng được tìm thấy bằng cách chuyển đổi:

$$x = \mu + z\sigma = 171,5 + (1,64 \times 6,5) = 182,2 \text{ cm}$$

Các điểm phần trăm của phân phối bình thường

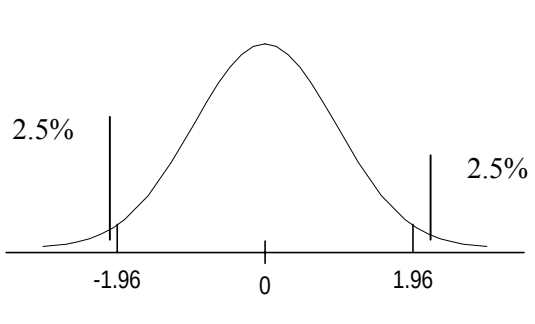
Một cách lí giải hữu dụng của độ lệch bình thường chuẩn là nó biểu thị giá trị của biến số cách số trung bình bao nhiêu độ lệch chuẩn. Điều này được trình bày trên thang đo của giá trị nguyên thủy trong hình 4.4. Do đó z=1 tương ứng với giá trị ở một độ lệch chuẩn trên trung bình và z=-1 là giá trị ở một độ lệch chuẩn dưới trung bình. Diện tích trên z=1 và dưới z=-1 đều là 0,1587 hay 15,87%. Do đó 31,74% phân phối cách trung bình hơn một độ lệch chuẩn hay nói cách khác 68,26% phân phối nằm trong phạm vi 1 độ lệch chuẩn so với trung bình. Tương tự, 4,55% phân phối cách trung bình hơn 2 độ lệch chuẩn hay nói cách khác 95,45% phân phối nằm trong phạm vi 2 độ lệch chuẩn so với trung bình. Điều này là cơ sở lí luận cho việc lí giải của độ lệch chuẩn ở Chương 3.



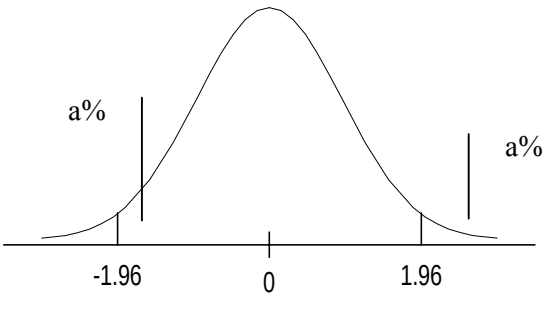
Hình 4.4 Lí giải SND bằng thang đo cho thấy số độ lệch chuẩn cách xa khỏi trung bình

Giá trị z bao gồm chính xác 95% phân phối giữa -z và z là 1,96 (hình 4.5a). 1,96 là **điểm 5 phần trăm (5% percentage point) của phân phối bình thường bởi vì 5%** phân phối cách trung bình hơn 1,96 lần độ lệch chuẩn (2,5% ở mỗi đuôi). Tương tự như vậy 2,58 là điểm phần trăm 1%. Các điểm phần trăm thường dùng được lập thành bảng A2. Lưu ý rằng các điểm phần trăm có thể tìm được từ bảng A1.

Điểm phần trăm được mô tả ở đây được gọi là điểm phần trăm hai đuôi (two- sided) bởi vì chúng bao gồm cả các quan sát ở đuôi trên và dưới của phân phối. Một vài cuốn sách lập bảng điểm phần trăm một đuôi (one- sided) chỉ xét đến một đuôi của phân phối (hình 4.5b). Thí dụ 1,96 là điểm 2,5% một đuôi bởi vì 2,5% phân phối bình thường chuẩn ở trên 1,96 và nó chính là điểm 5% hai đuôi. Sự khác biệt này được thảo luận lại ở Chương 6 trong phần kiểm định ý nghĩa.



(a) 1.96 laì âiãøm 2.5% mäüt bãn hay laì âiãøm 5% hai bãn



(b) z laì âiãøm a% mäüt bãn hay laì âiãøm 2a% hai bãn

KHOẢNG TIN CẬY CỦA TRUNG BÌNH

Giới thiệu

Sự biến thiên lấy mẫu và sai số chuẩn của trung bình mẫu được thảo luận ở Chương 3. Ở đây chúng ta xét bằng cách nào chúng ta có thể dùng trung bình mẫu và sai số chuẩn để biết được giá trị khả dĩ của trung bình dân số mà thường không biết được.

Trường hợp mẫu cỡ lớn (phân phối bình thường)

Chúng ta đã lưu ý ở Chương 3 rằng khoảng 95% của trung bình mẫu trong phân phối thu được bằng cách lấy mẫu lập lại sẽ nằm trong phạm vị hai sai số chuẩn trên hay dưới trung bình dân số. Điều này dựa trên giả thiết rằng tính phân phối bình thường của trung bình mẫu và trung bình của phân phối là trung bình của dân số, μ, và độ lệch chuẩn là sai số chuẩn của trung bình mẫu, s /√ n. Điều này có thể được biện minh khi cỡ mẫu lớn, thí dụ như n lớn hơn 60 bởi vì gần như luôn luôn phân phối của trung bình mẫu là bình thường (xem ở dưới); hơn nữa, độ lệch chuẩn mẫu, s, là một ước lượng đáng tin cậy của độ lệch chuẩn dân số, s, thường không biết. Từ Chương 4 chúng ta có thể khẳng định chính xác rằng 95% trung bình mẫu phải nằm trong phạm vi 1,96 sai số chuẩn so với trung bình dân số, 19,6 là điểm 5% của phân phối bình thường chuẩn. Do đó 95% là xác suất một trung bình mẫu nằm trong phạm vi 1,96 sai số chuẩn so với trung bình dân số

Trên thực tiễn, kết quả này được dùng để từ trung bình mẫu quan sát ($\bar{x}$) và sai số chuẩn (s.e.=s/√ n) ước lượng phạm vi trung bình dân số khả dĩ nằm trong đó. Bởi vì có xác suất 95% trung bình mẫu nằm trong 1,96 sai số chuẩn so với trung bình mẫu, có xác suất 95% khoảng nằm giữa $\bar{x}$ - 1,96 s.e. và $\bar{x}$ + 1,96 s.e. chứa trung bình dân số (chưa biết). Khoảng từ $\bar{x}$ - 1,96 s.e. đến $\bar{x}$ + 1,96 s.e. cho được xem là các giá trị khả dĩ của trung bình dân số. Nó được gọi là khoảng tin cậy 95% (95% confidence interval - c.i.) của trung bình dân số, và $\bar{x}$ + 1,96 s.e. và $\bar{x}$ - 1,96 s.e. là **giới hạn tin cậy 95% (95% confidence limits) của trung bình dân số.**

Khoảng tin cậy 95% cho mẫu lớn = $\bar{x} \pm (1{,}96 \times s/\sqrt{n})$

Khoảng tin cây cho các phần trăm khác được tính cũng giống như vậy và dùng điểm phần trăm tương ứng, z', của phân phối bình thường chuẩn thay vì 1,96. Thí dụ khoảng tin cậy 99% là $\bar{x} \pm (2{,}58 \times \text{s.e.})$

Khoảng tin cậy mẫu lớn = $\bar{x} \pm (z' \times s/\sqrt{n})$

Thí dụ 5.1

Trong chương trình khống chế sốt rét, người ta dự tính phun tất cả 10.000 nhà ở một vùng nông thôn với thuốc diệt côn trùng và cần thiết ước lượng số lượng cần thiết. Bởi vì không thể đo lường tất cả 10.000 nhà, người ta chọn một mẫu ngẫu nhiên 100 nhà và đo diện tích bề mặt có thể phun thuốc của từng căn nhà.

Diện tích bề mặt có thể phun thuốc trung bình của 100 nhà này là 23,2 m2 và độ lệch chuẩn là 5,9 m2. Diện tích trung bình của mẫu 100 căn nhà ($\bar{x}$) không thể bằng chính xác diện tích trung bình của 10.000 nhà (μ). Độ chính xác được đo lường bằng sai số chuẩn s/√ n gần bằng s/√ n=5,9/100=0,6 m2. Xác suất 95% trung bình mẫu (23,2m2) khác với trung bình dân số ít hơn 1,96 s.e. = 1,96 × 0,6 = 1,2 m2. Khoảng tin cậy 95% là

$\bar{x} \pm 1{,}96 \times s/\sqrt{n} = 23{,}2 \pm 1{,}2 = 22{,}0$ đến 24,4 m2

Người ta quyết định dùng giới hạn tin cậy 95% trên trong dự trù cho lượng thuốc trừ sâu cần thiết bởi vì người ta muốn tính thừa hơn là tính thiếu. Một lít thuốc trừ sâu đủ phun cho 50 m2 và lượng dự trù là

10.000 x 24,4/50 =4880 lít

Dầu vậy vẫn con có khả năng không có đủ thuốc trừ sâu. Khoảng 22,0 tới 44,4 m^2 cho phạm vi các giá trị khả dĩ của diện tích bề mặt trung bình của tất cả 10.000 nhà. Có xác suất 95% khoảng này chứa trung bình dân số nhưng có xác suất 5% là chúng không chứa trung bình dân số và có xác suất 2,5% ước lượng dựa trên giới hạn tin cậy trên là quá nhỏ. Ước lượng cẩn thận hơn lượng thuốc diệt côn trùng cần thiết cần dựa trên khoảng tin cậy rộng hơn, thí dụ như 99%, thì xác suất ước lượng thiếu sẽ nhỏ hơn (0,05%).

Mẫu nhỏ

Khi tính toán khoảng tin cậy chúng ta giả thuyết rằng cỡ mẫu là lớn (lớn hơn 60). Khi cỡ mẫu không lớn có hai khía cạnh sẽ thay đổi. Thứ nhất, độ lệch chuẩn mẫu, s, dễ bị tác động bởi các biến thiên lấy mẫu, có thể không phải là ước lượng đáng tin cậy của s . Thứ nhì, khi phân phối của dân số phi bình thường, phân phối của trung bình mẫu cũng có thể phi bình thường.

Tác động thứ nhì chỉ có ý nghĩa thực tiễn khi cỡ mẫu rất nhỏ (nhỏ hơn 15) và khi phân phối của dân số rất khác bình thường. Điều này là do một tính chất toán học rất hữu ích được gọi là định lí giới hạn trung tâm (central limit theorem) khẳng định rằng ngay cả biến không có phân phối bình thường, trung bình mẫu có khuynh hướng phân phối bình thường (có thể biểu diễn thực tế của định lí này bằng cách tiến hành một trò chơi lấy mẫu như trong thí dụ 3.4 nhưng thay thế 250 huyết áp bằng một dân số phân phối phi bình thường. Mẫu càng lớn thì trung bình mẫu càng gần với phân phối bình thường; cỡ mẫu cần thiết để cho xấp xỉ bình thường phụ thuộc vào dân số có phân phối bình thường nhiều hay ít nhưng trong phần lớn các trường hợp, cỡ mẫu 15 là đủ.

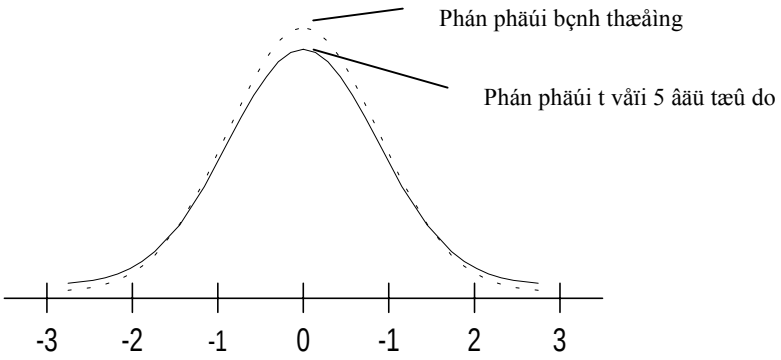
Bởi vì định lí giới hạn trung tâm, nên thường chỉ có vấn đề thứ nhất là sự biến thiên s do lấy mẫu mới khiến cho không thể sử dụng phân phối bình thường để tính khoảng tin cậy. Thay vào đó người ta dùng phân phối t. Nói chính xác, điều này chỉ đúng đắn khi dân số có phân phối bình thường. Dù vậy việc dùng phân phối t là xác đáng trừ khi dân số phân phối rất phi bình thường (tính chất ngày được gọi là **tính vững (robustness)). Sau này chúng ta sẽ trình bày phải làm gì trong trường hợp** cực kì phi bình thường.

Khoảng tin cậy dùng phân phối t

Các tính toán khoảng tin cậy ở trên dùng phân phối bình thường dựa trên sự kiện $(\bar{x}-\mu)/(s/\sqrt{n})$ là một giá trị của phân phối bình thường và chúng ta có thể dùng s tay cho s đối với mẫu lớn. Trên thực tế, $(\bar{x}-\mu)/(s/\sqrt{n})$ không phải là một giá trị của phân phối bình thường mà là của phân phối t với (n-1) độ tự do (t distribution with (n-1) degrees of freedom). Phân phối này được W.S. Gossett, có bút danh là Student, tìm ra và thường được gọi là phân phối t của Student. Giống như phân phối bình thường, phân phối t là một phân phối hình chuông đối xứng có trung bình là zero nhưng trải rộng hơn và có đuôi dài hơn (Hình 5.1)

Hình dạng chính xác của phân phối t phụ thuộc vào độ tự do (degrees of freedom - d.f.) n-1 của độ lệch chuẩn mẫu, s; độ tự do càng nhỏ phân phối t càng trải rộng. Người ta lập bảng A3 điểm phần trăm cho các độ tự do khác nhau. Thí dụ, nếu cỡ mẫu là 8, độ tự do là 7 và điểm 5% là 2,36. Trong trường hợp này khoảng tin cậy 95% dùng độ lệch chuẩn mẫu s là $\bar{x} \pm 2{,}36 \times s/\sqrt{n}$. Nói chung khoảng tin cậy được tính dùng t, điểm phần trăm thích hợp của phân phối t với (n-1) độ tự do.

Khoảng tin cậy cỡ mẫu nhỏ = $\bar{x} + (t \times s/\sqrt{n})$



Hình 5.1 Phân phối t có 5 độ tự do so sánh với phân phối bình thường

Đối với độ tự do nhỏ, điểm phần trăm của phân phối t lớn hơn đáng kể so với điểm phần trăm của phân phối bình thường. Điều này do độ lệch chuẩn mẫu s có thể là ước lượng không chính xác của s, và khi tính đến điều này khoảng tin cậy rộng hơn đáng kể so với t đã biết s. Đối với độ tự do lớn, phân phối t hầu như là giống như phân phối bình thường chuẩn, bởi vì s là ước lượng đúng của s. Đáy của Bảng A3 trình bày điểm phần trăm của phân phối t với vô hạn (∞) bậc tự do và ta có thể so sánh với Bảng A2 để thấy là chúng giống với phân phối bình thường chuẩn.

Thí dụ 5.2

Sau đây là số giờ không bị đau của 6 bệnh nhân viêm khớp sau khi sử dụng một loại thuốc mới

2,2 2,4 4,9 2,5 3,7 4,3 giờ

$\bar{x}$ = 3,3 giờ, s = 1,13 giờ, n = 6, d.f. = 6-1 = 5

s/√ n = 0,46 giờ

Điểm 5% của phân phối t với 5 bậc tự do là 2,57 và khoảng tin cậy 95% của số giờ trung bình không bị đau đối với bệnh nhân viêm khớp là

3,3 ± 2,57 × 0,46 = 3,3 ±1,2 = 2,1 tới 4,5 giờ

Tính phi bình thường nghiêm trọng

Khi phân phối của dân số rất phi bình thường, ta có thể biến đổi thang đo của biến số x để cho nó có phân phối bình thường theo thang mới (xem Chương 19). Một cách khác là tính khoảng tin cậy phi tham số (Chương 20) mặc dù cách này khá phức tạp.

Tóm tắt các trường hợp

Bảng 5.1 tóm tắt các cách xây dựng khoảng tin cậy. Không có ranh giới chính xác giữa tính bình thường và phi bình thường nhưng ta có thể cho phân phối hình J ngược (hình 2.6b) là cực kì phi bình thường nhưng phân phối bị lệch (Hình 2.5b và c) chỉ hơi phi bình thường.

Bảng 5.1 Các cách được đề nghị để xây dựng khoảng tin cậy

(a) Không biết độ lệch chuẩn dân số s

	Phân phối dân số	
Cỡ mẫu	xấp xỉ bình thường	rất phi bình thường
Trên 60	$\bar{x} \pm (z' \times s/\sqrt{n})$	$\bar{x} \pm (z' \times s/\sqrt{n})$
Nhỏ hơn 60	$\bar{x} \pm (t' \times s/\sqrt{n})$	phi tham số

(b) Đã biết độ lệch chuẩn dân số

	Phân phối dân số	
Cỡ mẫu	xấp xỉ bình thường	rất phi bình thường
Trên 15	$\bar{x} \pm (z' \times \sigma/\sqrt{n})$	$\bar{x} \pm (z' \times \sigma/\sqrt{n})$
Nhỏ hơn 15	$\bar{x} \pm (z' \times \sigma/\sqrt{n})$	phi tham số

Trong các trường hợp hiếm người ta biết độ lệch chuẩn dân số và do đó không ước lượng từ dân số. Khi đó, điểm phần trăm được dùng để tính khoảng tin cậy bất kể cỡ mẫu, với điều kiện phân phối của dân số không được quá phi bình thường (trong trường hợp này xem lại đoạn trên).

KIỂM ĐỊNH Ý NGHĨA CỦA MỘT TRUNG BÌNH

Giới thiệu

Trong Chương 5 chúng ta đã biết làm sao để sử dụng trung bình và độ lệch chuẩn mẫu để xây dựng khoảng tin cậy thể hiện các giá trị khả dĩ của trung bình dân số. Trong chương này chúng ta mô tả cách ngược lại để xem trung bình mẫu có phù hợp với một giá trị trung bình dân số giả thuyết. Phương pháp đánh giá này được gọi là kiểm định ý nghĩa. Nó thường dựa vào phân phối t hay phân phối bình thường, yếu tố quyết định sự lựa chọn cũng giống như yếu tố quyết định trong việc tính toán khoảng tin cậy. Một dạng đặc biệt nhưng phổ biến của kiểm định t một mẫu, kiểm định t cặp đôi, sẽ được mô tả trước tiên, sau đó là dạng tổng quát của kiểm định t và cuối cùng là kiểm định bình thường một mẫu.

Kiểm định t cặp đôi

Kiểm định t cặp đôi (paired t test) để kiểm định xem có sự khác nhau giữa một cặp biến số đo lường trên một cá nhân có bằng zero hay không, tính trên trung bình. Xem xét với một ví dụ cụ thể sẽ dễ dàng hơn. Xét- kết quả của một thử nghiệm lâm sàng kiểm tra hiệu quả của thuốc ngủ trong đó người ta quan sát giấc ngủ của 10 bệnh nhân trong một đêm không dùng thuốc và một đêm có dùng thuốc. Kết quả được trình bày trong Bảng 6.1. Đối với bệnh nhân. Người ta ghi nhận một cặp thời gian ngủ và tính hiệu số. Số trung bình của hiệu số số giờ ngủ có thuốc với số giờ ngủ dùng placebo là $\bar{x}$=1,78 giờ, và độ lệch chuẩn là s=1,77 giờ. Sai số chuẩn là s/√n=1,77/√ 10=0,56 giờ.

Ngay cả khi thuốc không có tác dụng, thời gian ngủ trung bình khi có thuốc và khi dùng placebo sẽ không giống hệt nhau. Do đó có hai giải thích khả dĩ cho sự gia tăng mà ta vừa thấy. Thức nhất đó chỉ là do sự biến thiên tình cờ và nói chung thuốc và placebo gây ngủ như nhau. Cách giải thích thứ nhì là sự gia tăng thời gian ngủ là do tác dụng có thực của thuốc. Người ta dùng kiểm định ý nghĩa (significant test) để quyết định xem lời giải thích nào khả dĩ hơn.

Kiểm định bắt đầu bằng cách giả thuyết rằng giải thích đầu tiên là đúng. Bước đầu tiên do đó là định nghĩa giả thuyết trung tính (null hypothesis) khẳng định rằng không có sự khác biệt thực sự giữa thời gian ngủ khi có không hoặc khi có placebo, hay nói cách khác sự trung bình thực sự (μ) của hiệu số là zero. Bước tiếp theo là tính toán xác suất có được giá trị trung bình quan sát ($\bar{x}$) do sự biến thiên tình cờ nếu giả thuyết này đúng. Tìm được bằng cách tính xác suất có được hiệu số giờ ngủ lớn hơn hoặc bằng giá trị trung bình quan sát được. Xác suất này được gọi là **mức ý nghĩa (significance level) của kết quả. Bác bỏ giả thuyết trung tính,** nghĩa là bác bỏ sự giải thích là do tình cờ, khi xác suất này rất nhỏ.

Bảng 6.1 Kết quả của thử nghiệm lâm sàng có đối chứng để xem hiệu quả của thuốc ngủ

	Giờ ngủ		
Bệnh nhân	thuốc	placebo	hiệu số
1	6,1	5,2	0,9
2	7,0	7,9	-0,9
3	8,2	3,9	4,3
4	7,6	4,7	2,9
5	6,5	5,3	1,2
6	8,4	5,5	3,0
7	6,9	4,2	2,7
8	6,7	6,1	0,6
9	7,4	3,8	3,6
10	5,8	6,3	-0,5
Trung bình	7,06	5,28	1,78

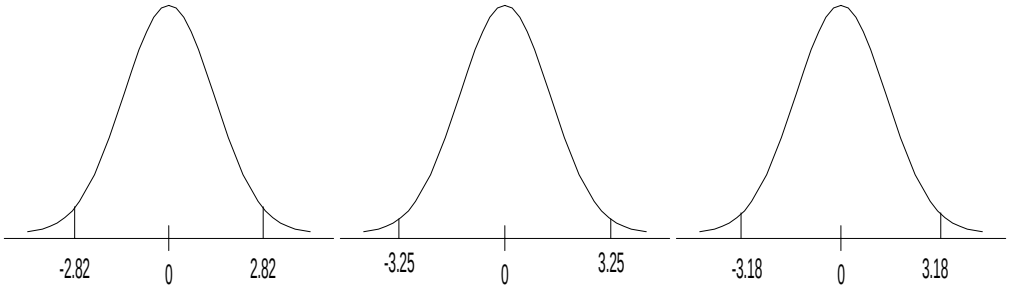
Chúng ta dùng phân phối t để tính mức ý nghĩa của kết quả. Từ Chương 5 chúng ta biết rằng với điều kiện là hiệu số giờ ngủ phân phối bình thường, $(\bar{x}-\mu)/(s/\sqrt{n})$ là giá trị của phân phối t có 9 bậc tự do. Giả thuyết không của hiệu số thời gian ngủ là zero tương đương với trung bình dân số, m, bằng zero và do đó.

$$t \text{ càõ âäi} = \frac{\bar{x}}{s/\sqrt{n}}, d.f. = n-1$$

Giá trị tuyệt đối của t càng lớn càng ít khả năng kết quả là do tình cờ (bỏ qua dấu của t). Trong thí dụ này, giá trị t tương ứng với hiệu số giờ ngủ là 1,78 là

t =1,78/0,56=3,18, độ tự do = 9

Điểm 5% (2 đuôi) của phân phối t với 9 độ tự do là 2,26 (Bảng A3), có nghĩa là có xác suất 5% giá trị t lớn hơn 2,26 về trị tuyệt đối. Điểm 2% là 2,82 và 1% là 3,25. Giá trị quan sát được ở giữa 2% và 1% (Hình 6.1). Hiệu số thời gian ngủ được xem là có mức ý nghĩa 2% (significant at the 2% level), bởi vì xác suất có sự khác biệt lớn đó là do tình cờ ít hơn 2%. Nếu viết xác suất theo thập phân người ta thường viết P<0,02. Bởi vì xác suất này nhỏ người ta kết luận rằng giả thuyết này là không phù hợp và thử nghiệm kết luận rằng thuốc có ảnh hưởng đến thời gian ngủ.



(a) 2.82 laì âiãøm 2%

(b) 3.25 laì âiãøm 1%

(c) 3.18 åí giæîa 2.82 vaì 3.25 do âoï täøng diãûn têch hai bãn khoaíng tæì 1% âãún 2%

Hình 6.1 Ýĩ nghĩa của giá trị 3,18 của phân phối t 9 độ tự do

Mức ý nghĩa càng nhỏ thì kết quả càng có ý nghĩa, bởi vì nó chứng minh chống lại giả thuyết trung tính mạnh mẽ hơn. Người ta thường xem xác suất dưới 5% là bằng chứng hữu lí để bác bỏ giả thuyết trung tính, xác suất dưới 1% là bằng chứng mạnh mẽ để bác bỏ giả thuyết trung tính ủng hộ sự kiện có tác động thực sự. Xác suất lớn hơn 5% được coi là không có ý nghĩa bởi vì nó thường được xem là tương đối phù hợp với giả thuyết không. Điều đó không có nghĩa là giả thuyết trung tính đúng mà chỉ là không có bằng chứng mạnh mẽ nào để bác bỏ chúng.

Trên lí thuyết có thể tính toán xác suất chính xác tương ứng với giá trị t = 3,18 và một vài chương trình máy tính đã làm như thế. Dù vậy, để làm bằng tay cần phải một bảng chi tiết (giống như Bảng A1 cho phân phối bình thường) cho mỗi độ tự do của phân phối t. Thông thường chỉ cần so sánh giá trị t với điểm phần trăm của độ tự do tương ứng.

Sau khi đã kết luận rằng thuốc có tác dụng thực sự chúng ta nên trình bày ước lượng của thời gian ngủ thêm do thuốc. Chúng ta thực hiện bằng cách tính 95% độ tin cậy. Đó là 1,78 ± (2,26 × 0,56) bằng 0,51 tới 3,05 giờ.

Quan hệ giữa khoảng tin cậy và kiểm định ý nghĩa

Có sự liên hệ chặt chẽ giữa ước lượng khoảng tin cậy và kiểm định ý nghĩa. Khoảng tin cậy cho phạm vi của giá trị trung bình dân số rút ra từ số liệu, trong khi kiểm định ý nghĩa đánh giá xem số liệu có phù hợp với một giá trị giả thuyết nào đó hay không. Do đó, nếu kết quả khác biệt có ý nghĩa với giá trị giả thuyết ở mức ý nghĩa đã chọn, khoảng tin cậy tương ứng sẽ không bao gồm giá trị này. Mặt khác, nếu giá trị không có ý nghĩa, khoảng tin cậy sẽ chứa giá trị giả thuyết. trong thí dụ trên, trung bình mẫu quan sát (1,78) khác biệt có ý nghĩa với zero ở mức 5%, và trong khoảng tin cậy 95% của trung bình dân số không chứa zero. Mặt khác, trung bình mẫu không khác biệt có ý nghĩa với zero ở mức 1% và khoảng tin cậy 95% của trung bình dân số (1,78 ± 3,25 × 0,56 = -0,04 tới 3,60) có chứa giá trị zero.

Kiểm định ý nghĩa 1 đuôi và 2 đuôi

Kiểm định ý nghĩa được mô tả ở trên là hai đuôi (two-sided), bởi vì chúng ta cho phép đi khỏi giả thuyết trung tính theo 2 hướng. Chúng ta chỉ quan tâm đến giá trị tuyệt đối của t và không tính dấu của nó. giải thuyết trung tính cho rằng không có sự khác biệt giữa thuốc và placebo và giả thuyết thay thế là có sự khác nhau theo hai hướng (thuốc gây ngủ ít hay nhiều hơn placebo) là có thể và đều đáng quan tâm. Xác suất của cả hai trường hợp đều được xét đến trong tính toán mức ý nghĩa; khi đó ta dùng điểm phần trăm 2 đuôi của phân phối t.

Kiểm định một đuôi (one-sided)thích hợp nếu người ta xem xét thuốc sẽ gây ngủ nhiều hơn hoặc bằng placebo, nhưng không gây ngủ ít hơn. Nếu trường hợp này đúng, kết quả cho thấy thuốc gây ngủ trung bình ít hơn placebo sẽ được cho là do sự biến thiên tình cờ chứ không phải là tác dụng thực, dù rằng thời gian ngủ trung bình có lớn tới cỡ nào. Kiểm định ý nghĩa dựa trên đuôi trên của phân phối t và ta dùng điểm phần trăm 1 đuôi. Trong thí dụ này thuốc gây ngủ nhiều hơn 1,78 giờ và giá trị t là 3,18 với 9 độ tự do. Số này vượt quá điểm 1% một chiều, 2,68, và vì vậy nếu dùng kiểm định một đuôi nó có ý nghĩa ở mức 1% (lưu lý rằng 2,82 là điểm 2% hai đuôi và kiểm định có ý nghĩa ở mức 2% chứ không phải ở mức 1%).

Kiểm định hai đuôi thích hợp trong đa số các trường hợp. Kiểm định một đuôi có vẻ cám dỗ bởi vì có nhiều khả năng cho kết quả có ý nghĩa hơn nhưng nó có nguy cơ bác bỏ giả thuyết trung tính và kết luận một cách sai lầm là có tác động thực sự bác bỏ. Kiểm định một đuôi chỉ nên dùng khi có lí do rõ ràng. Điều này phải được quyết định trước khi thu thập số liệu và phải được khẳng định rõ ràng trong phần trình bày kết quả.

Kiểm định t một mẫu

Kiểm định t bắt cặp là một trường hợp đặc biệt của kiểm định t một mẫu, kiểm định xem trung bình mẫu có khác với một giá trị cho trước μ hay không, μ không nhất thiết phải bằng zero. Công thức tổng quát là

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}, d.f. = n - 1$$

Thí dụ 6.1

Dưới đây là chiều cao tính theo centimetre của 24 trẻ trai 2 tuổi người Jamaica mắc bệnh hồng cầu hình liềm đồng hợp tử

84,4	89,9	89,0	91,9	87,0	78,5	84,1	86,3
80,6	80,0	81,3	86,8	83,4	89,8	85,4	80,6
85,0	82,5	80,7	84,3	85,4	85,0	85,5	81,9

Theo tiêu chuẩn về chiều cao và trọng lượng của Anh, chiều cao tham khảo của trẻ trai 2 tuổi là 86,5 cm. Mẫu ở trên có chứng minh rằng trẻ trai bị bệnh hồng cầu liềm có chiều cao khác tiêu chuẩn hay không?

$\bar{x}$=84,4 cm, s=3,11 cm , n=24, s/√ n=0,63 cm

Giả thuyết trung tính là chiều cao trung bình của trẻ trai 2 tuổi bị bệnh hồng cầu liềm là 86,5 cm.

$$t = \frac{\bar{x} - 86.5}{s/\sqrt{n}} = \frac{84.1 - 86.5}{0.63} = -3{,}81,\ d.f. = n - 1, \mathrm{P} < 0.001$$

Do đó chiều cao trung bình của các trẻ trai Jamaica bị bệnh hồng cầu liềm thấp hơn có ý nghĩa so với chiều cao tiêu chuẩn của Anh (P<0,001). Lưu ý rằng cần cẩn thận khi lí giải kết quả bởi vì sự phát triển của trẻ em bị hồng cầu liềm Jamaica được so sánh với tiêu chuẩn của Anh chứ không phải trẻ em Jamaica khỏe mạnh. Có thể rằng bệnh hồng cầu liềm làm còi sự phát triển nhưng cần phải so sánh với trẻ Jamaica không bệnh để khẳng định điều này. Phương pháp so sánh hai mẫu được trình bày ở Chương 7.

Kiểm định bình thường

Trong trường hợp ước lượng khoảng tin cậy, khi kiểm định trung bình của một mẫu lớn (60 hay hơn 60) hay trong trường hợp đã biết độ lệch chuẩn của dân số, người ta dùng phân phối bình thường chứ không phải dùng phân phối t. Dạng kiểm định ý nghĩa thì hoàn toàn giống nhau

(Mẫu lớn) (đã biết σ)

$$z = \frac{\bar{x} - \mu}{s/\sqrt{n}} \qquad t = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

Dường như không có sự thống nhất trong việc đặt tên kiểm định này. Đôi khi người ta gọi là kiểm định bình thường, hay kiểm định z hay kiểm định độ lệch bình thường chuẩn. Sách này dùng tên kiểm định bình thường.

Các loại sai lầm trong kiểm định giả thuyết

Như đã nói, kiểm định ý nghĩa là phương pháp đánh giá xe, kết quả có thể là do tình cờ hay do một hiệu quả có thực. Nó không thể chứng minh điều này hoặc điều kia và một trong hai loại sai lầm trên có thể xảy ra. Giả thuyết trung tính có thể bị bác bỏ khi nó thực sự đúng (sai lầm loại I) hay ngược lại chúng ta có thể không bác bỏ được trong khi nó sai (sai lầm loại II) (Bảng 6.2)

Bảng 6.2 Các loại sai lầm trong kiểm định thống kê

Kết luận của kiểm định ý nghĩa	Thực tế	
	giả thuyết trung tính đúng	giả thuyết trung tính sai
Bác bỏ giả thuyết trung tính	sai lầm loại I (xác suất=ý nghĩa)	kết luận đúng (xác suất=năng lực)
Không bác bỏ giả thuyết trung tính	kết luận đúng (xác suất = 1 - mức ý nghĩa)	sai lầm loại II (xác suất=1-năng lực)

Nhắc lại mức ý nghĩa của một kiểm định bằng với xác suất xảy ra kết quả bằng hoặc vượt quá giá trị quan sát được, nếu giả thuyết trung tính đúng. Xác suất chúng ta mắc sai lầm loại I và bác bỏ sai giả thuyết trung tính bằng mức ý nghĩa của kiểm định. Thí dụ, có xác suất 5% có kết quả có mức ý nghĩa 5% chỉ do sự biến thiên lấy mẫu đơn thuần và do đó nếu chúng ta đánh giá kết quả là bằng chứng bác bỏ giả thuyết trung tính, có xác suất 5% chúng ta sẽ mắc sai lầm (hình 6.2a).

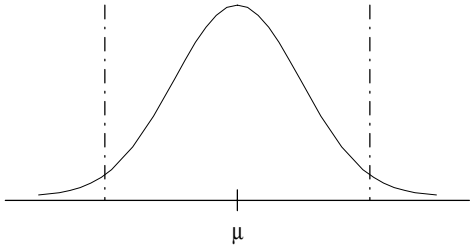
Loại sai lầm thứ hai là không bác bỏ giả thuyết trung tính khi nó sai. Điều này xảy ra là do sự chồng lắp giữa phân phối lấy mẫu thực của trung bình mẫu quanh trung bình dân số μ' ($\neq\mu$) và miền chấp nhận của giả thuyết trung tính dựa trên phân phối lấy mẫu giả thuyết về trung bình không đúng, μ . Điều này được minh họa trong Hình 6.2b. Miền gạch chéo cho thấy tỉ lệ b% của phân phối lấy mẫu thực sẽ rơi vào trong miền chấp nhận của giả thuyết trung tính, đó là chúng phù hợp với giả thuyết trung tính ở mức ý nghĩa 5%. Nếu chọn mức ý nghĩa thấp hơn, khiến xác suất sai lầm loại I nhỏ hơn, kích thước miền có gạch chéo tăng lên dẫn tới khả năng không bác bỏ được giả thuyết trung tính sẽ lớn hơn. Điều ngược lại cũng đúng. Xác suất ta không mắc sai lầm loại II là năng lực (power) của kiểm định. Điều này được thảo luận sâu hơn trong phần xác định cỡ mẫu ở Chương 26. Nói chunng tăng cỡ mẫu sẽ tăng năng lực bởi vì đường cong phân phối lấy mẫu trong hình 6.2b sẽ cao hơn và hẹp hơ và do đó sự chồng lắp sẽ giảm.

(a)

Bác bỏ giả thuyết trung tính nếu trung bình mẫu ở vùng này

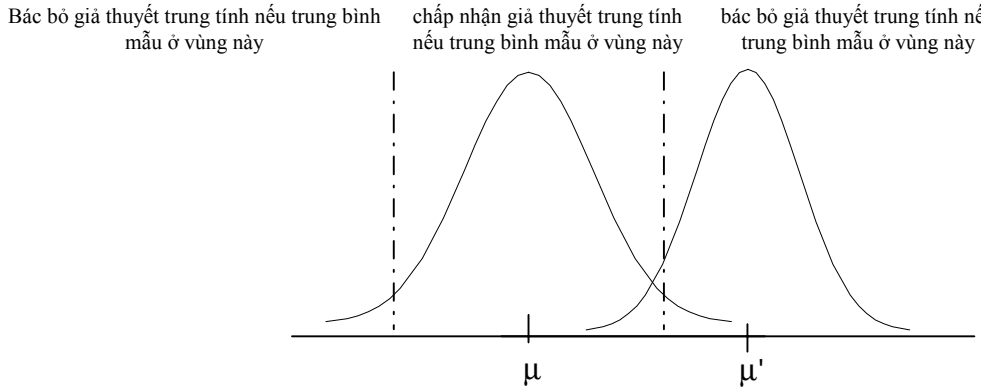
chấp nhận giả thuyết trung tính nếu trung bình mẫu ở vùng này

bác bỏ giả thuyết trung tính nếu trung bình mẫu ở vùng này



(a) Sai lầm loại I. Giả thuyết trung tính đúng. Trung bình dân số = m. Đường cong cho thấy phân phối lấy mẫu của trung bình mẫu. Vùng gạch chéo (tổng số = 5%) cho thấy xác suất giả thuyết trung tính bị loại bỏ một cách sai lầm.

(b)



(b) Sai lầm loại 2. Giả thuyết trung tính sai. Trung bình dân số = m' khác m. Đường cong liên tục cho thấy phân phối lấy mẫu thực của trung bình mẫu, trong đó đường chấm là phân phối lấy mẫu theo giả thuyết trung tính. Vùng gạch chéo là xác suất (b%) giả thuyết trung tính không bị bác bỏ.

Hình 6.2 Xác suất xuất hiện hai loại sai lầm kiểm định ý nghĩa, minh họa ở kiểm định ở mức 5%

SO SÁNH HAI TRUNG BÌNH

Giới thiệu

Việc sử dụng kiểm định t và bình thường để so sánh một trung bình mẫu với một giá trị giả thuyết của trung bình dân số được trình bày ở Chương 6. Trong chương này chúng ta mô tả các sử dụng những kiểm định này để so sánh trung bình của hai mẫu khác nhau thí dụ trong lượng trung bình của trẻ em con những người hút thuốc lá nhiều và của trẻ em con những người không hút thuốc. Kiểm định này sử dụng những quy luật rất giống những quy luật trong trường hợp một mẫu và dạng của kiểm định thì giống nhau, cũng là lượng cần kiểm định chia cho sai số chuẩn của nó. Dù vậy không giống như trường hợp một mẫu, công thức cho kiểm t và bình thường khác nhau, bởi vì việc tính toán sai số chuẩn khác nhau và kiểm định t cần thêm một giả thiết khác đó là hai độ lệch chuẩn dân số bằng nhau. Ở dưới sẽ trình bày chi tiết. Khoảng tin cậy được tính như bình thường.

Cần phải lưu ý rằng kiểm định 2 mẫu không thích hợp cho việc so sánh hai trung bình của hai biến số đo lường trên cùng một mẫu. Trong trường hợp này cần tính hiệu số của mỗi cặp quan sát và dùng kiểm định cặp một mẫu như đã trình bày trong chương 6.

Phân phối lấy mẫu của hiệu số hai trung bình

Hiệu số, (x1 - $\bar{x}$2, của trung bình hai mẫu độc lập với nhau có phân phối bình thường với điều kiện mỗi trung bình có phân phối bình thường. trung bình của phân phối này là hiệu số giữa hai trung bình dân số, μ1 - μ2. Giả thuyết trung tính của kiểm định ý nghĩa là hai trung bình này bằng nhau nghĩa là μ1 - μ2=0. Công thức tính sai số chuẩn dựa trên sự kết hợp giữa hai sai số chuẩn của mỗi trung bình.

$$\text{s.e.} = \sqrt{(\sigma_1^2/n_1 + \sigma_2^2/n_2)}$$

Các sai số chuẩn này được ước lượng từ độ lệch chuẩn mẫu s1 và s2.

Kiểm định bình thường (mẫu lớn hay biết độ lệch chuẩn)

Kiểm định bình thường được dùng khi mẫu lớn hoặc biết độ lệch chuẩn.

Mẫu lớn

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{(s_1^2/n + s_2^2/n)}}$$

biết s

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{(\sigma_1^2/n + \sigma_2^2/n)}}$$

Xem xét kết quả sau từ một chương trình sốt rét trong đó các lam máu của trẻ từ 1 đến 4 tuổi được xét nghiệm tìm kí sinh trùng Plasmodium falciparum. Kí sinh trùng được tìm thấy trong 70 lam máu, và nồng độ hemoglobin của những trẻ này là 10,6g% với độ lệch chuẩn là 1,4g%. Mức hemogloin của 80 trẻ khác là 11,5g% với độ lệch chuẩn là 1,3g%. điều này có phù hợp sự kiện nồng độ hemoglobin bị giảm trong lúc bị nhiễm P. falciparum? Bởi cả hai mẫu đều lớn nên có thể dùng kiểm định bình thường.

$n_1 = 70$, $\bar{x}_1 = 10{,}6$, $s_1 = 1{,}4$

$n_2 = 80$, $\bar{x}_2 = 11{,}5$, $s_2 = 1{,}3$

$$z = \frac{10.6 - 11.5}{\sqrt{1.4^2/70 + 1.3^2/80}} = -4.06, P < 0{,}001$$

Kết quả này có mức ý nghĩa ở 1% bởi vì 4,06 lớn hơn 3,29, điểm 0,1% của phân phối bình thường (Bảng A2). Do đó nồng độ hemoglobin trung bình của trẻ có P. Falciparum trong máu thấp hơn những trẻ không có P. falciparum một cách có ý nghĩa (P<0,001).

Khoảng tin cậy

Mẫu lớn	biết s
c.i.=($\bar{x}_1$- $\bar{x}_2$) ± (z' × s.e.)	c.i.=($\bar{x}_1$- $\bar{x}_2$) ± (z' × s.e.)
s.e.= √ ($s_1^2/n_1+s_2^2/n_2$)	s.e.= √ ($\sigma_1^2/n_1+\sigma_2^2/n_2$)

Khoảng tin cậy được tính sử dụng z, điểm phần trăm thích hợp của phân phối bình thường. Khoảng tin cậy 95% của hiệu số giữa các nồng độ hemoglobin trung bình ở thí dụ trên là:

- 0,9 ± (1,96 × 0,2216) = -1,33 tới - 0,47 g%

Có nghĩa là trẻ có kí sinh trùng trong máu có hemoglobin trung bình thấp hơn từ 0,47 tới 1,33 g%.

Kiểm định t (mẫu nhỏ, độ lệch chuẩn bằng nhau)

Kiểm định t được dùng cho mẫu nhỏ. Nó đòi hỏi phân phối dân số bình thường nhưng nó cũng vững đối với sự vi phạm giả thiết này, giống như trong trường hợp một mẫu. Khi so sánh hai trung bình, sự hợp lệ của test t cũng phụ thuộc vào hai độ lệch chuẩn dân số có bằng nhau hay không. Trong nhiều tình huống có thể giả thuyết chúng bằng nhau. Nếu độ lệch chuẩn mẫu rất khác nhau, thí dụ như lớn gấp đôi, cần phải sử dụng phương pháp khác. Điều này sẽ được nói ở dưới.

Công thức tính sai số chuẩn của sự khác biệt giữa hai trung bình được đơn giản thành

$$\text{s.e.} = \sqrt{(\sigma^2/n_1 + \sigma^2/n_2)} = \sigma\sqrt{(1/n_1 + 1/n_2)}$$

Trong đó s là độ lệch chuẩn chung. Có hai ước lượng mẫu của s từ 2 mẫu, s_1 và s_2, và chúng kết hợp để cho một ước lượng của độ lệch chuẩn dân số chung, s, với độ tự do bằng $(n_1-1) + (n_2-1) = n_1+n_2-2$

$$s = \sqrt{\left[\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{(n_1 - n_2 - 2)}\right]}$$

Công thức này gán ước lượng từ mẫu lớn hơn trọng số (weight) lớn hơn bởi vì nó đáng tin cậy hơn. Sai số chuẩn của sự khác biệt của trung bình được ước lượng bằng

$$\text{s.e.} = s\sqrt{(1/n_1 + 1/n_2)}$$

Và giá trị t được tính

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s\sqrt{1/n_1 + 1/n_2}}; d.f. = n_1 + n_2 - 2$$

Bảng 7.1 trình bày trọng lượng trẻ em của 15 người không hút thuốc lá và 14 hút thuốc lá nặng việc tính toán kiểm định t để so sánh hai nhóm như sau

$$s = \sqrt{\left[\frac{14\times 0.3707^2 + 13\times 0.4927^2}{(15+14-2)}\right]} = 0.4337kg$$

$$t = \frac{(3.5933 - 3.2029)}{0.4337\sqrt{1/15+1/14}} = \frac{0.3904}{0.1612} = 2.42; d.f. = 15+14-2 = 27$$

Từ bảng 3 có thể thấy 2,42 xấp xỉ bằng 2,47 điểm 2% của phân phối t với 27 độ tự do. Do đó trẻ trung bình con của người không hút thuốc nặng hơn của của người hút thuốc với mức ý nghĩa 2%

Khoảng tin cậy

$$\text{c.i} = (\bar{x}_1 - \bar{x}_2) \pm (t' \times \text{s.e.}),\ \text{s.e.} = s\sqrt{(1/n_1+1/n_2)}$$

Khoảng tin cậy được tính bằng cách dùng t, điểm phần trăm thích hợp của phân phối t với (n_1 + n_2 - 2) độ tự do. Trong thí dụ trên khoảng tin cậy 95% của sự khác biệt giữa trung bình trọng lượng lúc sinh là:

$$0{,}3904 \pm (2{,}05 \times 0{,}1612) = 0{,}06 \text{ tới } 0{,}72 \text{ kg}$$

Có nghĩa là trẻ sinh ra do người mẹ không hút thuốc lá nặng hơn từ 0,06 tới 0,72 kg so với trẻ sinh từ người mẹ hút thuốc lá.

Bảng 7.1 So sánh trọng lượng lúc sinh của con 15 người không hút thuốc lá và trọng lượng lúc sinh của con 14 người hút thuốc lá nặng

	Trọng lượng lúc sinh	
	Không hút thuốc	Hút thuốc lá nặng
	3,99	3,18
	3,79	2,84
	3,60	2,90
	3,73	3,27
	3,21	3,85
	3,60	3,52
	4,08	3,23
	3,61	2,76
	3,83	3,60
	3,31	3,75
	4,13	3,59
	3,26	3,63
	3,54	2,38
	3,51	2,34
	2,71	
$\bar{x}$	3,5933	3,2029
s	0,3707	0,4927
n	15	14

Cỡ mẫu nhỏ, độ lệch chuẩn không bằng nhau

Khi độ lệch chuẩn dân số của hai nhóm khác nhau, nên làm theo phương án thứ nhất nghĩa là tìm sự biến đổi thang đo để điều chỉnh (xem Chương 19), để cho có thể dùng kiểm định t. Thí dụ nếu dường như độ lệch chuẩn tỉ lệ với trung bình, có thể lấy logarithm của từng giá trị. Một phương án khác là dùng kiểm định phi tham số (xem Chương 20) hay dùng kiểm định Fisher-Behrens hay Welch (Armitage & Berry, 1987). ở đây không trình bày chi tiết những kiểm định này bởi vì chúng ít được sử dụng.

SO SÁNH NHIỀU TRUNG BÌNH - PHÂN TÍCH PHƯƠNG SAI

Giới thiệu

Thường có những tập hợp số liệu phức tạp chứa hơn hai nhóm và trong phân tích thường phải so sánh những trung bình của các nhóm thành phần. Thí dụ, người ta có thể muốn phân tích các số đo hemoglobin được thu thập trên một cuộc điều tra cộng đồng để xem nó có khác nhau theo tuổi và giới tính hay không và xem có phải là sự khác biệt giữa các nhóm tuổi là như nhau dù là nam hay nữ. Thoạt đầu, dường như có thể làm điều này bằng cách dùng một loạt các kiểm định t, so sánh từng 2 nhóm một. Điều này không chỉ rắc rối về mặt thực tiễn mà còn vô lí về mặt lí thuyết, bởi vì tiến hành một số lớn các kiểm định ý nghĩa có thể dẫn tới một kết quả có ý nghĩa sai lạc. Thí dụ có thể trông đợi 1 trong 20 (5%) các kiểm định được tiến hành sẽ có ý nghĩa ở mức 5% ngay cả khi không có sự khác biệt (xem lại thảo luận về sai lầm loại I trong chương 6).

Một phương pháp khác được gọi là phân tích phương sai (analysis of variance). Ý nghĩa của tên này được trình bày sau. Phương pháp khá phức tạp. Việc tính toán mất nhiều thời gian và thường được tiến hành nhờ các gói phần mềm máy tính chuẩn. Vì lí do này, chương này nhấn mạnh đến các nguyên lí với mục đích giúp người đọc có đủ kiến thức để chỉ định dạng phân tích cần thiết và lí giải kết quả. Dù vậy trong chương này cũng trình bày chi tiết của việc tính toán trong trường hợp đơn giản nhất, đó là phân tích phương sai một chiều, bởi vì nó sẽ giúp ích cho việc nắm vững căn bản của phương pháp và quan hệ của nó với kiểm định t.

Phân tích phương sai một chiều thích hợp khi các nhóm so sánh được xác bằng bởi một yếu tố (factor), thí dụ như so sánh trung bình giữa các giai cấp khác nhau hay giữa các dân tộc khác nhau. Phân tích phương sai hai chiều được mô tả và thích hợp khi việc chia nhóm dựa trên 2 yếu tố, thí dụ như tuổi và giới tính. Phương pháp dễ dàng được mở rộng để so sánh các nhóm đươc phân loại chéo bằng nhiều hai yếu tố.

Một yếu tố được phân tích phương sai bởi vì người ta muốn so sánh các mức khác nhau của nó hay bởi vì nó gây cho sự biến thiên cần loại trừ. Xem thí dụ sau. Sau khi khám phá tỉ suất bệnh mạch vành thay đổi đáng kể giữa các nhóm dân tộc khác nhau, người ta tiến hành một cuộc điều tra để xem điều này có phải là do nồng độ lipid trung bình khác nhau giữa các nhóm dân tộc khác nhau. Bởi vì nồng độ lipid thay đổi theo giới tính và tuổi, do đó cần phân tích phương sai của nhóm tuổi và giới tính cũng như nhóm dân tộc, mặc dù tuổi và giới tính không phải là mối quan tâm chính của nghiên cứu này. Việc đưa vào phân tích chúng có hai lợi ích. Thứ nhất, kiểm định ý nghĩa sự khác biệt giữa các nhóm chủng tộc trở nên mạnh mẽ (powerful) hơn, nghĩa là dễ khiến cho sự khác biệt thực sự trở thành có ý nghĩa. Thứ nhì, nó đảm bảo sự so sánh các nhóm chủng tộc không bị sai lệch do cơ cấu nhóm tuổi và giới tính (xem thảo luận về biến số gây nhiễu ở Chương 14. Xem thêm phương pháp chuẩn hóa ở Chương 15 được dùng để trình bày trung bình của các nhóm dân tộc được hiệu chỉnh theo tuổi và giới tính).

Cũng có thể phân tích số liệu được phân thành nhiều yếu tố bằng cách dùng một kĩ thuật tương tự nhưng tổng quát hơn gọi là hồi quy bội (multiple regression) được mô tả ở Chương 10. Cả hai phương pháp đều cho kết quả giống hệt nhau nhưng bởi vì hồi quy bội tổng quát hơn nên nó cần tính toán phức tạp hơn. Vì thế nó không hiệu quả trong các trường hợp đơn giản. Dù vậy, sự lựa chọn phụ thuộc vào chương trình máy tính có được và chúng có dễ sử dụng hay không.

Một vài độc giả có thể thấy nội dung của chương này hơi khó. Nó có thể bỏ qua trong lần đọc đầu.

Phân tích phương sai một chiều

Phân tích phương sai một chiều (one-way analysis of variance) được dùng để so sánh trung bình của một số nhóm, thí dụ nhưng nồng độ hemoglobin trung bình của bệnh nhân của các loại bệnh hồng cầu liềm khác nhau (bảng 8.1a). Phương pháp phân tích được gọi là một chiều bởi vì số liệu được phân tích theo một chiều, trong trường hợp này là loại bệnh hồng cầu liềm. Phương pháp dựa trên việc xác định thành phần trong toàn bộ biến thiên được quy cho sự khác biệt giữa các trung bình nhóm và so sánh chúng với thành phần có thể quy cho sự khác biệt giữa các cá nhân trong cùng nhóm. Do đó nó có tên là phân tích phương sai.

Chúng ta bắt đầu bằng cách tính phương sai của tất cả các quan sát, bỏ qua việc chia thành từng nhóm. Nhớ lại trong Chương 3 rằng phương sai là bình phương của độ lệch chuẩn và bằng tổng bình phương các hiệu số của quan sát và trung bình, chia cho độ tự do. Phương pháp phân tích phương sai một chiều chia tổng **các bình phương (sum of square - SS) thành 2 phần riêng biệt:**

(i) Tổng các bình phương do sự khác biệt giữa các trung bình nhóm

(ii) Tổng cách bình phương do sự khác biệt giữa các quan sát trong từng nhóm nó được gọi là tổng bình phương phần dư (residual sum of squares)

Tổng các độ tự do cũng được tách ra theo cách tương tự. Việc tính toán số liệu hồng cầu liềm được trình bày ở bảng 8.1(b) và kết quả trình bày của bảng phân tích phương sai ở trong bảng 8.1(c).

Cột thứ tư trong bảng trình bày lượng biến thiên cho mỗi độ tự do và được gọi là **trung bình bình phương (mean square - MS). Kiểm định ý nghĩa cho sự khác** biệt giữa các nhóm dựa trên trung bình bình phương giữa các nhóm (between groups) và trong nội bộ các nhóm (within groups). Nếu sự khác biệt quan sát được trong nồng độ hemoglobin của các loại bệnh hồng cầu liềm khác nhau chỉ là tình cờ, sự biến thiên giữa các nhóm cũng tương đương với sự biến thiên giữa các đối tượng trong cùng một loại bệnh. Ngược lại nếu chúng là do sự khác biệt thực sự thì sự biến thiên giữa các nhóm sẽ lớn hơn. Trung bình bình phương được so sánh bằng kiểm định F, đôi khi còn được gọi là kiểm định tỉ số phương sai (variance-ratio).

Trong đó N là tổng số các quan sát và k là số các nhóm.

F phải xấp xỉ bằng 1 nếu không có sự khác biệt thực sự giữa các nhóm và lớn hơn 1 nếu có sự khác biệt. Theo giả thuyết trung tính cho rằng sự khác biệt chỉ là do tình cờ, tỉ số này sẽ tuân theo phân phối F mà không giống với các phân phối khác, nó có một cặp độ tự do: (k-1) độ tự do ở tử số và (N-k) độ tự do ở mẫu số. Điểm phần trăm của phân phối F được lập bảng theo các cặp độ tự do ở Bảng A4. Cột của bảng chỉ độ tự do của tử số và các khối gồm nhiều hàng chỉ độ tự do của mẫu số. trong mỗi khối này có những hàng khác nhau cho mức phần trăm khác nhau. Điểm phần trăm là một đuôi bởi vì kiểm định dựa trên phân phối F lớn hơn một.

Trong bảng 8.1(c), F=50,26/0,95=52,9 với độ tự do (2,38). Bảng điểm phần trăm có hàng cho 30 và 40 độ tự do chứ không có hàng cho 38 độ tự do. Dù vậy chúng ta có thể nói rằng điểm 0,1% của F(2,38) ở giữa 8,77 và 8,25 (là điểm 0,1% của F(2,30) và F(2,40)). Rõ ràng 52,9 lớn hơn cả hai. Do đó nồng độ hemoglobin khác nhau một cách có ý nghĩa giữa các bệnh nhân mắc các loại bệnh hồng cầu liềm khác nhau (P<0,001). Nồng độ trung bình thấp nhất là bênh nhân có Hb SS, trung bình đối với bệnh nhân có Hb S/ß-thalassaemia và cao nhất đối với bệnh nhân có Hb SC.

Giả thiết

Có hai giả thiết cần cho kiểm định F. Thứ nhất là số liệu phải phân phối bình thường. Thứ nhì là độ lệch chuẩn giữa các cá thể trong cùng một nhóm phải giống nhau. Có thể ước lượng bằng căn bậc hai của trung bình bình phương (MS) trong các nhóm. Có thể bỏ qua sự phân

phối không bình thường nhưng các độ lệch chuẩn không bằng nhau có thể gây hậu quả nghiêm trọng. Trong trường hợp này có thể biến đổi số liệu.

Mối liên hệ với kiểm định t hai mẫu

Phân tích phương sai một chiều là sự mở rộng của kiểm định t hai mẫu. Khi chỉ có hai mẫu, nó cho kết quả y như là kiểm định t. Giá trị F bằng bình phương giá trị t tương ứng và điểm phần trăm của phân phối F với (1,N-2) độ tự do cũng bằng bình phương của điểm phần trăm của phân phối t với N-2 độ tự do.

Bảng 8.1 Phân tích phương sai một chiều: sự khác biệt trong nồng độ hemoglobin giữa các bệnh nhân bị các loại bệnh hồng cầu liềm khác nhau. Số liệu từ Anionwo et al. (1981) British Medical Journal, 282, 283-6

(a) Số liệu

Loại bệnh hồng cầu liềm	Số bệnh nhân (n_i)	Trung bình ($\bar{x}_i$)	s.d. (s_i)	Giá trị của các cá thể hemoglobin g% (x)
Hb SS	16	8,7125	0,8445	7,2, 7,7, 8,0, 8,1, 8,3, 8,4, 8,4, 8,5, 8,6, 8,7, 9,1, 9,1, 9,1, 9,8, 10,1, 10,3
Hb S/b-thalassaemia	10	10,6300	1,2841	8,1, 9,2, 10,0, 10,4, 10,6, 10,9, 11,1, 11,9, 12,0, 12,1
Hb SC	15	13,300	0,9419	10,7, 11,3, 11,5, 11,6, 11,7, 11,8, 12,0, 12,1, 12,3, 12,6, 12,6, 13,3, 13,8, 13,8, 13,9

(b) Tính toán

$N = \Sigma ni = 16 + 10 + 15 = 41$, số nhóm (k) = 3

$\Sigma x = 7,2 + 7,7 + ... + 13,8 + 13,9 = 430,2$

$\Sigma x^2 = 7,2^2 + 7,7^2 + ... + 13,8^2 + 13,9^2 = 4651,80$

Tổng cộng: $SS = \Sigma (x - \bar{x})2 = \Sigma x^2 - (\Sigma x)^2/N = 4651,80 - 430,22/41 = 137,85$

d.f.= N-1 = 40

Giữa các nhóm $SS = \Sigma n_i(\bar{x}_i - \bar{x})^2 = \Sigma n_i x_i^2 - (\Sigma x)^2/N$

$= 16 \times 8,7125^2 + 10 \times 10,6300^2 + 15 \times 12,300^2 - 430,2^2/41 = 99,89$

d.f. = k-1 = 2

Trong các nhóm $SS = \Sigma (n_i - 1)s_i^2 = 15 \times 0,84452 + 9 \times 1,28412 + 14 \times 0,9419 = 37,96$

d.f.= N - k = 41-3 = 38

(c) Phân tích phương sai

Nguồn biến thiên	SS	d.f.	MS=SS/d.f.	F= MS giữa các nhóm / MS bên trong nhóm
Giữa các nhóm	99,89	2	50,26	52,9,P<0,001
Trong các nhóm	37,96	38	0,95	
Tổng cộng	137,85	40		

Phân tích phương sai hai chiều

Người ta dùng phân tích phương sai hai chiều (two way analysis of variance) khi số liệu được phân loại theo hai chiều thí dụ như theo tuổi và giới tính. Số liệu là quy **hoạch cân đối (balanced design) nếu số các quan sát trong các nhóm là bằng** nhau và quy hoạch không cân đối (unbalanced design) nếu số các quan sát trong các nhóm không bằng nhau. Qui hoạch cân đối có hai loại có lặp (with replication) nếu có nhiều quan sát trong mỗi nhóm và không có lặp (without replication) nếu chỉ có một quan sát. Ba loại quy hoạch này sẽ được trình bày riêng.

Quy hoạch cân đối có lặp

Bảng 8.2 trình bày kết quả thực nghiệm trên 3 chủng chuột mỗi chủng gồm 5 chuột đực và 5 chuột cái được điều trị bằng hormone tăng trưởng. Mục đích là tìm xem các chủng chuột và giới tính chuột có đáp ứng với điều trị như nhau hay không. Số đo của đáp ứng là tăng trọng sau 7 ngày.

Những số liệu này được phân loại theo hai chiều, bởi chủng tộc và giới tính. Quy hoạch là cân đối có lặp (balanced with replication) bởi vì có 5 quan sát trong mỗi nhóm chủng-giới tính. Phân tích phương sai 2 chiều chia tổng bình phương thành 4 thành phần

(i) Tổng bình phương do sự khác biệt giữa các chủng. Điều này là tác động **chính** (main effect) của yếu tó, chủng. Độ tự do của nó là số các chủng chuột trừ một và bằng 2.

(ii) Tổng bình phương do sự khác biệt giới tính, đó là tác động chính của giới tính. Độ tự do của nó bằng 1, bằng số các giới tính trừ một.

(iii) Tổng bình phương do sự tương tác (interaction) giữa chủng và giới tính. Sự tương tác có nghĩa là sự khác biệt do chủng không giống nhau trên cả hai giới hay ngược lại sự khác biệt do giới tính không giống nhau trên 3 chủng chuột. Độ tự do bằng tích số độ tự do của 2 tác động chính bằng $2 \times 1=1$

(iv) tổng bình phương phần dư là sự khác biệt giữa các con chuột trong cùng nhóm chủng-giới tính. Độ tự do bằng 24, tích số của số chủng (3) số giới tính (2) và số quan sát trong mỗi nhóm trừ một (4).

Tác động chính và tương tác được kiểm định độ ý nghĩa bằng cách dùng kiểm định F để so sánh trung bình bình phương của nó với trung bình bình phương phần dư như được mô tả trong phân tích phương sai một chiều. Thực nghiệm này không thu được kết quả có ý nghĩa.

Bảng 8.2 Sự khác biệt đáp ứng với hormone sinh trưởng trên 3 chủng chuột khác nhau (mỗi chủng gồm 5 đực và 5 cái).

(a) Tăng trọng trung bình (tính theo gram) với độ lệch chuẩn ở trong ngoặc (n=5 trong mỗi nhóm),

Giới tính	chủng A	chủng B	chủng C
Nam	11,9(0,9)	12,1(0,7)	12,2(0,7)
Nữ	12,3(1,1)	11,8(0,6)	13,1(0,9)

(b)Phân tích phương sai hai chiều: quy hoạch cân bằng có lặp

Nguồn biến thiên	SS	d.f.	MS=SS/ d.f.	MS tác động F= ---------- MS phần dư
Tác động chính				
Chủng	2,63	2	1,32	1,9,P>0,1
Giới tính	1,16	1	1,16	1,7,P>0,1
Tương tác Chủng × Giới	1,65	2	0,83	1,2,>0,1
Phần dư	16,86	24	0,70	
Tổng cộng	22,30	29		

Quy hoạch cân đối không lặp

Năm phương pháp để xác định tuổi thai được so sánh trên 10 phụ nữ trong bảng 8.3. Không có tổng bình phương phần dư trong phân tích phương sai bởi vì chỉ có một quan sát cho một

phương pháp áp dụng trên một phụ nữ. Trong trường hợp như vậy, tương tác được giả thiết là do sự biến thiên tình cờ và trung bình bình phương được dùng làm ước lượng trung bình bình phương phần dư để tính giá trị F của tác động chính. Tác động chính do tuổi thai khác nhau giữa 10 phụ nữ hiển nhiên có ý nghĩa. Bản thân điều này không được quan tâm lắm nhưng nó là một nguồn biến thiên quan trọng cần phải tính đến trong khi so sánh các phương pháp. Tác động chính do sự khác biệt giữa các phương pháp là có ý nghĩa ở mức 5% (F=757,85/202,81= 3,74, d.f.=[4,36]).

Phân chia tổng bình phương

Cần xem xét chi tiết các hiệu số tạo nên tác động có ý nghĩa. Thí dụ, phương pháp dựa trên ngày thai máy cho con số trung bình cao hơn đáng kể so với các phương pháp khác. Có thể phân chia tổng bình phương của tác động chính đối với các phương pháp trong bảng 8.3 (c) thành:

(i) Tổng bình phương các hiệu số giữa phương pháp dựa trên ngày thai máy và các phương pháp khác. Tổng này có 1 độ tự do.

Bảng 8.3 Tuổi thai tính theo ngày của 10 phụ nữ được ước tính bằng 5 phương pháp - kì kinh cuối (last mentrual period - LMP), khám âm đạo (Vaginal examination - VE), ngày thai máy (date of quickening - DOQ), siêu âm (Ultra sound - US) và oxydase diamine máu (Diamine oxidase - DAO).

(a) số liệu

Đối tượng	LMP	VE	DOQ	US	DAO	
1	275	273	288	273	244	270,6
2	292	283	284	285	329	294,6
3	281	274	298	270	252	275,0
4	284	275	271	272	258	272,0
5	285	294	307	278	275	287,8
6	283	279	301	276	279	283,6
7	290	265	298	291	295	287,8
8	294	277	295	290	271	285,4
9	300	304	293	279	271	289,4
10	284	297	352	292	284	301,8
Trung bình	286,4	282,1	298,7	280,6	275,8	

(b) Phân tích phương sai hai chiều: quy hoạch cân đối không có lặp (trung bình bình phương tương tác được dùng làm ước lượng trung bình bình phương phần dư trong kiểm định F)

Nguồn biến thiên	SS	d.f.	MS=SS/d.f.	F= MS tác động / MS tương tác
Đối tượng	4437,6	9	493,07	2,43, P<0,05
Phương pháp	3031,4	4	757,85	3,74, P<0,05
Tương tác	7301,0	36	202,81	
Tổng cộng	14770,0	49		

(c) Phân chia tổng bình phương theo phương pháp

Nguồn biến thiên	SS	d.f.	MS=SS/d.f.	F= MS tác động / MS tương tác
DOQ so với các phương pháp khác	2415,1	1	2415,10	11,91,P<0,001
Khác biệt giữa LMP, VE, US và DAO	616,3	3	205,43	1,01,P>0,1
Kĩ thuật	3031,4	4		

(ii) Tổng bình phương còn lại có 3 độ tự do, thể hiện các hiệu số trong số 4 phương pháp khác (LMP, VE, US, DAO).

Mỗi thành phần được kiểm định bằng kiểm định F theo cách bình thường. Sự phân chia này cho thấy phương pháp dựa trên ngày thai máy khác đáng kể (P<0,001) với các phương pháp khác, nhưng không có sự khác biệt có ý nghĩa trong 4 phương pháp này.

Lưu ý rằng tổng bình phương đã được chia theo các phương pháp khác nhau, và thành các thành phần độc lập bằng độ tự do, trong trường hợp này là 4. Sự phân chia phụ thuộc vào sự so sánh quan tâm và tốt nhất phải được dựa trên nền tảng *tiên nghiệm (a priori) trước khi phân tích số liệu. Tiến hành phân chia nhờ* **phương pháp tương phản tuyến tính (method of linear contrasts). Người đọc có thể** tham khảo Armitage & Berry (1987) để biết rõ chi tiết.

Quan hệ với kiểm định t một mẫu

Phương pháp phân tích phương sai hai chiều quy hoạch cân đối không có lặp là mở rộng của kiểm định t bắt cặp một mẫu, so sánh các giá trị của nhiều biến được đo lường trên một cá thể. Trong trường hợp này, có 5 biến: tuổi thai được ước tính bằng các phương pháp khác nhau trên một phụ nữ. Hai cách tiếp cận cho kết quả tương tự khi chỉ có 2 biến và giá trị F bằng giá trị t bình phương.

Quy hoạch không cân đối

Bảng 8.4(a) tóm tắt số liệu về nhiễm giun móc và mức hemoglobin, được thu thập trong một cuộc điều tra về nhiễm kí sinh trùng ở Đông châu Phi. Số liệu được phân loại theo hai yếu tố, giới tính và mật độ nhiễm giun móc. Có thể thấy rằng đối với mỗi giới tính, nồng độ hemoglobin giảm khi nhiễm giun móc càng nhiều, và đối với một mức độ nhiễm giun móc, hemoglobin trung bình ở nữ thấp hơn ở nam. Dù vậy quy hoạch này là không cân đối bởi vì số người trong mỗi nhóm không bằng nhau. Điều này có nghĩa là không thể tách tác động của giới tính và mật độ nhiễm giun khiến cho việc lí giải số liệu không thể tiến hành trực tiếp.

Bảng 8.4 Nồng độ hemoglobin (g%) theo mật độ nhiễm giun móc ở nam và nữ

(a) Số liệu

	Nam			Nữ		
Mật độ nhiễm giun móc	Số	Hb trung bình	s.d.	Số	Hb trung bình	s.d.
Âm tính	22	12,3	1,8	35	11,1	1.1
Thấp	20	11,9	1,2	27	10,8	1,3
Trung bình	17	10,7	1,6	14	9,5	1,9
Cao	15	9,0	1,4	11	8,6	1,7

(b) Phân tích phương sai hai chiều: quy hoạch không cân đối

Nguồn biến thiên	SS	d.f.	MS=SS/d.f.	MS tác động F= ---------- MS phần dư
Giới tính	20,94	1	20,94	9,9,P<0,01
Mật độ giun móc điều chỉnh theo giới	176,68	3	58,89	27,8, P<0,001
Tương tác	3,24	3	1,08	0,5, P>0,1
Phần dư	324,28	153	2,12	
Tổng cộng	525,14	160		

Tổng bình phương không thể chia thành các thành phần quy về 2 yếu tố độc lập với nhau và trong bảng 8.4(b) trình bày phân tích phương sai được cải tiến. Đầu tiên tính tổng bình phương do sự khác biệt giới tính. Trừ khi hai giới tính có phân phối giun móc giống nhau, tổng bình phương sẽ gồm cả một số biến thiên do sự khác biệt mật độ giun. Sau đó tính tổng bình phương do mật độ nhiễm giun. Tổng này đánh giá quan hệ giữa nồng độ hemoglobin và mật độ nhiễm giun có điều chỉnh cho sự khác biệt giới tính giữa các nhóm mật độ nhiễm giun. Cả hai tác động chính đều có ý nghĩa, mức ý nghĩa 1% đối với giới tính (F=9,9, d.f.=[1,153]) và 0,1% đối với mật độ nhiễm giun (F=27,8, d.f=[3,153]). Sự tương tác không có ý nghĩa.

Theo phương án khác, tác động của nhiễm giun móc được phân tích phương sai trước, trong trường hợp đó nó gồm cả sự biến thiên do khác biệt nồng độ hemoglobin giữa nam và nữ. Sau đó tác động chính của giới tính sẽ là sự khác biệt còn lại sau khi điều chỉnh cho sự khác biệt mật độ giun giữa nam và nữ. Đối với quy hoạch không cân bằng cần tiến hành phân tích theo cả hai cách. Dù vậy, trong thí dụ này, sự xem xét đã dẫn đến rằng nên tính tới giới tính trước.

Số liệu không cân đối phổ biến và không thể tránh được trong cuộc nghiên cứu điều tra. Dù vậy, thử nghiệm lâm sàng và thực nghiệm labo nên dự trù để có quy hoạch cân đối. Không phải mọi dự trù đều thành công thí dụ như có người rời khỏi vùng trong khi thử nghiệm. Các chương trình phân tích phương sai của các phần mềm máy tính nhỏ có thể dùng cho các quy hoạch cân đối hay quy hoạch chỉ có một số nhỏ các giá trị bị khuyết (missing value); trong những trường hợp này các chương trình hồi quy bội có thể dùng cho thiết kế không cân đối (xem Chương 10).

Tác động cố định và ngẫu nhiên

Yếu tố có thể chia làm hai loại, tác động cố định (phổ biến hơn) và tác động ngẫu nhiên. Các yếu tó như giới tính, nhóm tuổi, và loại bệnh hồng cầu liềm là các tác **động cố định (fixed effects) bởi vì các mức riêng lẻ của nó có các giá trị nhất** định; giới tính luôn luôn là nam hay nữ. Ngược lại, các mức riêng lẻ của của tác **động ngẫu nhiên (random effects) không được sự quan tâm mà chỉ là một mẫu** đại diện cho sự biến thiên. Thí dụ, xét một nghiên cứu điều tra sự biến thiên natri và sucrose trong dung dịch ORS được pha ở nhà, trong đó có

10 người được đề nghị từng người pha 8 dung dịch. Trong trường hợp này 10 người chỉ là đại diện cho nguồn biến thiên giữa các dung dịch được pha bởi những người khác nhau. Con người là một tác động ngẫu nhiên. Trong thí dụ này và để xem tác động con người có ý nghĩa không, chúng ta sẽ quan tâm đến việc ước lượng độ lớn của sự biến thiên nồng độ giữa các dung dịch được pha bởi một người và sự biến thiên giữa các dung dịch được pha bởi các người khác nhau. Chúng được gọi lài thành **phần của sự biến thiên (components of variation). Xem Huitson để biết cách** ước lượng chúng.

Phương pháp kiểm định ý nghĩa giống nhau trong tác động cố định và ngẫu nhiên trong quy hoạch một chiều và trong quy hoạch hai chiều không có lặp, nhưng không giống nhau trong quy hoạch hai chiều (hay nhiều chiều hơn) có lặp. Trong quy hoạch hai chiều có lặp, nếu cả hai tác động đều cố định, trung bình bình phương được so sánh với trung bình bình phương phần dư như đã nói ở trên. Mặt khác nếu cả hai tác động đều là ngẫu nhiên, trung bình bình phương được so sánh với trung bình bình phương tương tác chứ không phải với trung bình bình phương phần dư. Nếu một tác động là ngẫu nhiên và một là cố định, nó sẽ là cách khác: tác động ngẫu nhiên được so sánh với trung bình bình phương phần dư, và tác động cố định sẽ đươc so sánh với trung bình bình phương tương tác. Đây là những điểm phức tạp. Người đọc quan tâm nhiều đến chi tiết nên tham khảo Huitson (1980).

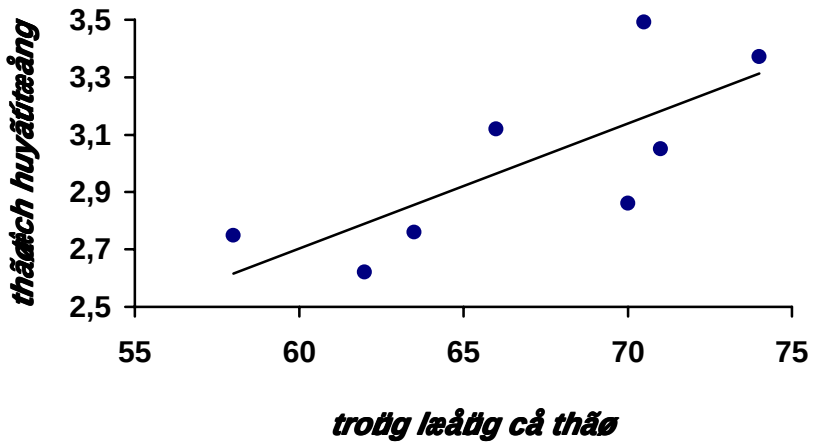
TƯƠNG QUAN VÀ HỒI QUY TUYẾN TÍNH

Giới thiệu

Các chương trước chú trọng đến phân tích và so sánh trung bình của một biến. Bây giờ đến lượt chúng ta chú ý đến quan hệ giữa các biến khác nhau và trong chương này chú ý đến kĩ thuật tương quan và hồi quy tuyến tính để tìm hiểu mối quan hệ tuyến tính (linear) giữa hai biến liên tục. Tương quan (correlation) đó lường sự chặt chẽ của mối liên hệ trong khi hồi quy tuyến tính (linear regression) cho biết phương trình đường thẳng mô tả sự liên hệ tốt nhất và cho phép tiên đoán biến số này từ biến số khác.

Bảng 9.1 Thể tích huyết tương và trọng lượng cơ thể của 8 người đàn ông khỏe mạnh

Đối tượng	trọng lượng cơ thể (kg)	Thể tích huyết tương (lít)
1	58,0	2,75
2	70,0	2,86
3	74,0	3,37
4	63,5	2,76
5	62,0	2,62
6	70,5	3,49
7	71,0	3,05
8	66,0	3,12

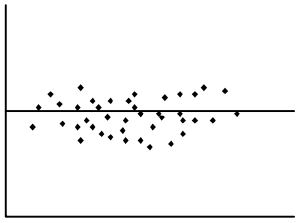


Hình 9.1 Phân tán đồ của thể tích huyết tương và trọng lượng cơ thể cùng với đường hồi quy tuyến tính

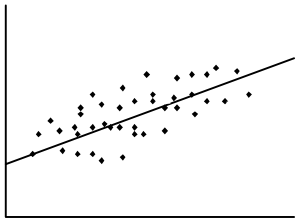
Tương quan

Bảng 9.1 trình bày trọng lượng cơ thể và thể tích huyết tương của 8 người đàn ông khỏe mạnh. Phân tán đồ (scatter diagram) của những số liệu này (hình 9.1) cho thấy thể tích huyết tương lớn có liên quan đến trọng lượng cơ thể cao và ngược lại. Sự liên quan này được đo lường bằng hệ số tương quan (correllation coefficient), r.

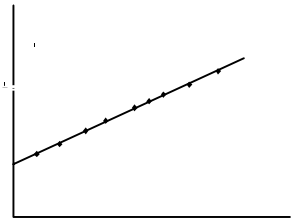
$$r = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\sqrt{\Sigma(x-\bar{x})^2\,\Sigma(y-\bar{y})^2}}$$



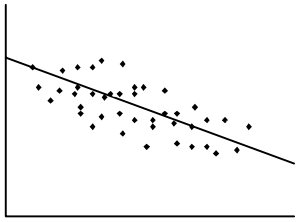
(a) Không tương quan



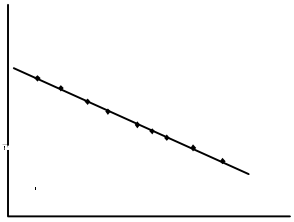
(b) Tương quan dương không hoàn toàn



(c) Tương quan dương hoàn toàn



(d) Tương quan âm không hoàn toàn



(e) Tương quan âm hoàn toàn

Hình 9.2 phân tán đồ minh họa các giá trị khác nhau của hê số tương quan. Trong đây cũng có các đường hồi quy.

Trong đó x là trọng lượng, y thể tích huyết tương, $\overline{x}$ và (y là trung bình tương ứng. Phân tán đồ minh họa những hệ số tương quan khác nhau đươc trình bày trong hình 9.2. Hệ số tương quan luôn luôn là một số giữa -1 và +1 và bằng zero nếu hai biến không liên hệ. Nó dương nếu x và y cùng cao hay cùng thấp và càng lớn khi sự liên quan càng chặt. Ta có giá trị 1 nếu các điểm trên phân tán đồ nằm ngay trên đường thẳng. Ngược lại, hệ số tương quan âm nếu giá trị y cao gắn liền với giá trị x thấp và ngược lại.

Điều quan trọng là sự tương quan giữa hai biến số cho thấy sự liên hệ nhưng không nhất thiết có nghĩa là cá quan hệ 'nhân quả'. Xem thảo luận đầy đủ hơn ở Bradford Hill (1977).

Hệ sối tương quan được tính dễ dàng hơn bằng cách lưu ý:

$$\Sigma (x- \overline{x})(y- \overline{y})=\Sigma xy- (\Sigma x)(\Sigma y)/n$$

$$\Sigma (x- \overline{x})^2=\Sigma x^2-(\Sigma x)^2/n$$

$$\Sigma (y-y)^2=\Sigma y^2-(\Sigma y)^2/n$$

Trong thí dụ này

N= 8

$\Sigma x =535$ $\overline{x} = 66,835$ $\Sigma x^2 = 35983,5$

$\Sigma y=24,2 \qquad \bar{y} = 3,0025 \qquad \Sigma y^2= 72,7980$

$\Sigma xy=58,0 \times 2,75 + 70,0 \times 2,86 +...+ 66,0 \times 3,12 =1615,295$

Do đó

$\Sigma (x- \bar{x})(y- \bar{y}) = 1616,295 - 535 \times 24,02/8=8,96$

$\Sigma (x-x)^2 = 35983,5 - 5352/8=205,38$

$\Sigma (y- \bar{y})^2 = 72,798 - 24,022/8=0,6780$

Và

$r= 8,96/(205,38 \times 0,6780)=0,76$

Kiểm định ý nghĩa

Kiểm định t được dùng để xem r có khác zero một cách có ý nghĩa hay không hay nói cách khác có phải sự tương quan quan sát được chỉ là do tình cờ hay không

$$t = r\sqrt{\left[\frac{n-2}{1-r^2}\right]}, d.f. = n-2$$

Trong thí dụ này:

$$t = 0.76\sqrt{\left[\frac{8-2}{1-0.76^2}\right]} = 2.86, d.f. = 6$$

Điều này có ý nghĩa ở mức 5% xác nhận ý nghĩa của sự liên hệ giữa thể tích huyết tương và trọng lượng cơ thể

Mức ý nghĩa phụ thuộc của cả vào độ lớn của mối tương quan và số các quan sát. Lưu ý rằng tương quan yếu có thể có ý nghĩa thống kê nếu nó dựa trên một số lớn quan sát, trong khi sự tương quan mạnh có thể không đạt được mức ý nghĩa nếu chỉ có một ít quan sát.

Hồi quy tuyến tính

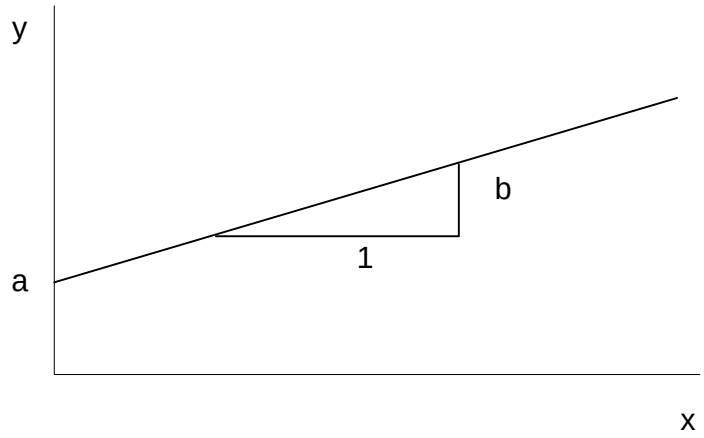
Hồi quy tuyến tính cho phương trình đường thẳng mô tả nếu biến x tăng thì biến y tăng như thế nào. Không giống như tương quan, việc lựa chọn biến nào để làm biến y là quan trọng bởi vì hai phương pháp không cùng cho một kết quả, y thường được gọi là biến số phụ thuộc (dependent variable) và x là biến số độc lập hay giải thích (independent or explanatory variable). Trong thí dụ này, rõ ràng chúng ta cần quan tâm sự phụ thuộc thể tích huyết tương và trọng lượng cơ thể.

Phương trình hồi quy là

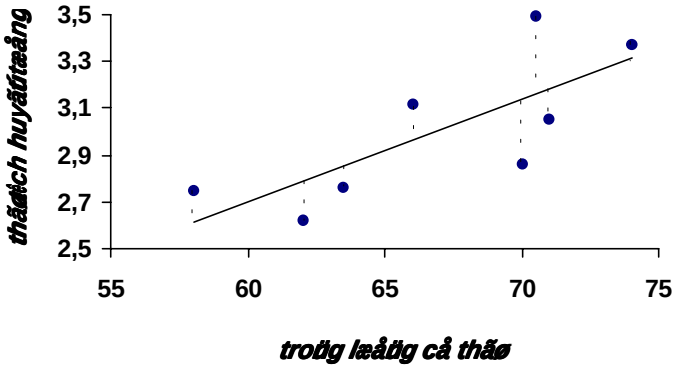
$$y = a + bx$$

Trong đó a là điểm chặn (intercept) và b là độ dốc (slope) của đường thẳng (Hình 9.3). Giá trị đối với a và b được tính sao cho cực tiểu hóa bình phương khoảng cách theo chiều đứng từ các điểm số liệu tới đường thẳng. Nó được gọi là phù hợp bình phương tối thiểu (least squares fit) (Hình 9.4). Độ dốc b đôi khi được gọi là hệ số hồi quy (regression coefficient). Nó có cùng dấu với hệ số tương quan. Khi không có sự tương quan, b bằng zero, tương ứng với một đường thẳng hồi quy nằm ngang đi qua điểm y.

ĉ và a = (y - b(x



Hình 9.3 Giao điểm và độ dốc của phương trình hồi quy y = a + bx. Giao điểm a là điểm mà đường thẳng cắt trục y và cho giá trị y ở x = 0. Độ dốc b là mức tăng của y tương ứng với sự gia tăng một đơn vị của x.



Hình 9.4 Đường thẳng hồi quy tuyến tính, y = a + bx, được làm phù hợp bằng bình phương tối thiểu, a và b được tính để cực tiểu hóa tổng bình phương của các độ lệch thẳng đứng (vẽ bằng các đường chấm) của các điểm đối với đường thẳng, mỗi độ lệch bằng hiệu số giữa số y quan sát và tiềm tương ứng trên đường thẳng a + bx

Trong thí dụ này

$$b = 8,96/205,38 = 0,043615$$

Và:

$$a = 3,0025 - 0,043615 \times 66,875 = 0,0857$$

Do đó sự phụ thuộc của thể tích huyết tương vào trọng lượng cơ thể được mô tả bằng

$$\text{Thể tích huyết tương} = 0,0857 + 0,0436 \times \text{trọng lượng}$$

và được vẽ trên Hình 9.1. Đường hồi quy được vẽ bằng cách tính tọa độ của hai điểm nằm trên nó. Thí dụ:

$$x = 60,\ y = 0,0857 + 0,0436 \times 60 = 2,7$$

Và

$$x = 70,\ y = 0,0857 + 0,0436 \times 70 = 3,1$$

Đường thẳng phải đi qua điểm $(\bar{x}, \bar{y})=(66,9\ ,\ 3,0)$

Giá trị a và b là các ước lượng mẫu của giá trị giao điểm và độ dốc của đường thẳng hồi quy mô tả mối liên hệ tuyến tính giữa x và y trong toàn bộ dân số. Do đó chúng bị các biến thiên lấy mẫu và độ chính xác của chúng có thể đo lường bằng sai số chuẩn.

và

trong đó

$$s = \sqrt{\left[\frac{\Sigma(y-\bar{y})^2 - b^2\,\Sigma(x-\bar{x})^2}{(n-2)}\right]}$$

s là độ lệch chuẩn của các điểm số liệu so với đường thẳng, có (n-2) độ tự do.

$$s = \sqrt{\frac{0.6780 - 0.0436^2 \times 205.38}{6}} = 0.2189$$

$$s.e.(a) = 0.2819\sqrt{\left[\frac{1}{8} + \frac{66.9^2}{205.38}\right]} = 1.3197$$

$$s.e.(b) = \frac{0.2189}{\sqrt{205.38}} = 0.0153$$

Kiểm định ý nghĩa

Kiểm định t được dùng để kiểm định xem b có khác biệt có ý nghĩa với một giá trị nhất định, được kí hiệu là ß (kí tự Hy lạp beta)

$$t = \frac{b-\beta}{s.e.(b)}, d.f. = n-2$$

Đặc biệt nó có thể dùng để kiểm định xem b có khác biệt có ý nghĩa với zero hay không. Và nó chính xác tương đương với kiểm định t của r = 0. Kiểm định này cho thể tích huyết tương và trọng lượng cơ thể cho kết quả

$$t = \frac{b-\beta}{s.e.(b)} = \frac{0.0436}{0.0153} = 2.85$$

Và giống như giá trị 2,86 (sự khác nhau là do sai số làm tròn) rút ra từ kiểm định cho hệ số tương quan. Do đó thể tích huyết tương tăng có ý nghĩa (P<0,05) đối với trọng lượng cơ thể.

Tiên đoán

Trong một số tình huống, có thể sử dụng phương trình hồi quy để tiên đoán giá trị y cho một giá trị đặc biệt của x được gọi là x'. Giá trị tiên đoán là:

y' = a + bx'

Và sai số chuẩn của nó là

$$s.e.(a) = s\sqrt{\left[\frac{1}{n} + \frac{\bar{x}^2}{\Sigma(x-\bar{x})^2}\right]} \qquad s.e.(b) = \frac{s}{\sqrt{\Sigma(x-\bar{x})^2}}$$

$$s.e.(y') = s\sqrt{\left[1 + \frac{1}{n} + \frac{(x' - \bar{x})^2}{\sum(x - \bar{x})^2}\right]}$$

Sai số chuẩn này tối thiểu khi x' gần với trung bình x. Nói chung phải thận trọng khi sử dụng đường hồi quy để tính các giá trị ngoài phạm vi của x trong số liệu gốc, bởi vì quan hệ tuyến tính không nhất thiết sẽ đúng ở ngoài phạm vi mà nó được làm phù hợp.

Trong thí dụ này, sự đo lường thể tích huyết tương tốn nhiều thời gian và do đó trong một số trường hợp, có thể tiên đoán từ trọng lượng cơ thể. Thí dụ thể tích plasma huyết tương của một người đàn ông nặng 66 kg là

$$0{,}0832 + 0{,}0436 \times 66 = 2{,}96 \text{ lít}$$

Và sai số chuẩn bằng

$$s.e.(y') = s\sqrt{\left[1 + \frac{1}{8} + \frac{(66 - 66.9)^2}{205.38}\right]} = 0.23l$$

Giả thiết

Có hai giả thiết nền tảng trong phương pháp hồi quy tuyến tính. Giả thiết thứ nhất là đối với bất cứ giá trị x nào, y có phân phối bình thường. Giả thiết thứ hai là độ phân tán của các điểm quanh đường thẳng là như nhau trong suốt đoạn thẳng. Độ phân tán được đo lường bằng độ lệch chuẩn s của các điểm số liệu so với đường thẳng như đã định nghĩa ở trên. Sự thay đổi thang đo có thể thích hợp nếu các giả thuyết trên không thỏa hay quan hệ dường như phi tuyến tính (xem Chương 19). Các quan hệ phi tuyến được thảo luận ở chương 10.

Sử dụng máy tính cầm tay

Một vài máy tính cầm tay có hàm số hồi quy tuyến tính và tương quan tự động (xem Chương 1), giúp cho tránh những tính toán tỉ mỉ mô tả ở trên.

HỒI QUY BỘI

Giới thiệu

Đôi khi xảy ra tình huống là chúng ta quan tâm sự phụ thuộc của một biến vào *nhiều các biến giải thích chứ không phải chỉ vào một biến. Thí dụ, 100 phụ nữ đi* khám tiền sản tham gia vào một cuộc nghiên cứu để xác định các biến số có liên quan với trọng lượng lúc sinh của trẻ, với mục tiêu cuối cùng là tiên đoán phụ nữ 'có nguy cơ' bị đẻ con nhẹ cân. Kết quả cho thấy rằng trọng lượng lúc sinh liên quan có ý nghĩa với tuổi của mẹ, chiều cao của mẹ, số con, thời gian mang thai, thu nhập gia đình nhưng những biến số này không độc lập. Thí dụ: chiều cao của mẹ và thời gian mang thai có liên quan với nhau. Sự tác động kết hợp của những biến số này, tính đến những mối liên hệ dương giữa chúng, có thể được tìm hiểu bằng cách dùng hồi quy bội (mutitple regression) được thảo luận trong bối cảnh của thí dụ này.

Để giới thiệu những phương pháp cần thiết cho hồi quy bội, một phương án hồi quy tuyến tính đơn khác, tương đương với phương án được mô tả ở Chương 9, nhưng dễ tổng quát hóa hơn sẽ được mô tả. Sau đó nó sẽ được mở rộng cho nhiều biến giải thích. Chi tiết tính toán không được trình bày bởi vì việc phân tích chắc chắn sẽ được tiến hành bằng cách dùng phần mềm máy tính. Phương pháp luận khá phức tạp và một số độc giả có thể bỏ qua chương này ở lần đọc đầu tiên.

Phương pháp phân tích phương sai dùng cho hồi quy tuyến tính đơn

Xem xét mối quan hệ giữa trọng lượng lúc sinh với chiều cao của mẹ. Phân tán đồ cho thấy rằng mối liên hệ là tuyến tính. Hệ số tương quan là 0,26 (P<0,01).

Hồi quy tuyến tính cho phương trình đường thẳng mô tả quan hệ:

Trọng lượng lúc sinh = a + b × chiều cao của mẹ

Trong đó a và b được tính sao cho cực tiểu hóa tổng các bình phương độ lệch của số liệu so với đường thẳng. Tổng các bình phương độ lệch đối với đường thẳng sau khi đã tiến hành cực tiểu hóa được gọi là tổng bình phương phần dư (residual sum of squares). Nó nhỏ hơn tổng bình phương tổng các biến thiên của trọng lượng lúc sinh và hiệu số đó được gọi là tổng các bình phương được giải **thích bởi hồi quy (sum of squares explained by the regression) của trọng lượng lúc sinh** theo chiều cao của mẹ, hay gọi đơn giản là tổng bình phương hồi quy (regression sum of squares). Tách đôi tổng bình phương của toàn thể biến thiên trọng lượng lúc sinh thành 2 phần có thể được trình bày trong bảng phân tích phương sai (Bảng 10.1). Có một độ tự do cho hồi quy và n - 2 = 98 độ tự do cho phần dư.

Bảng 10.1 Phân tích phương sai của hồi quy tuyến tính trọng lượng lúc sinh theo chiều cao của mẹ (n=100)

Nguồn biến thiên	Tổng bình phương (SS)	Độ tự do (d.f.)	Trung bình bình phương (MS=SS/d.f.)	MS hồi qui F=---------- MS phần dư
Hồi quy theo chiều cao của mẹ	1,48	1	1,4800	7,11,P<0,01
Phần dư	20,39	98	0,2081	
Tổng số	21,87	99		

Nếu thực sự không có sự liên hệ giữa các biến số, thì trung bình bình phương hồi quy (MS regression) sẽ có cùng độ lớn với trung bình bình phương phần dư (MS residual). Nếu có sự liên hệ nó sẽ lớn hơn. Kiểm định bằng cách dùng kiểm định F như được mô tả trong Chương 8.

$$F = \frac{MS\ \text{häöquy}}{MS\ \text{pháöi dæ}}, d.f. = (1, n-2)$$

F phải khoảng gần 1 nếu không có sự liên hệ và lớn hơn nếu có. Kiểm định F này tương đương với kiểm định t cho b=0 và r=0 mô tả ở chương 9, giá trị của F tương đương với t2. Trong thí dụ này F = 1,4800/0,2081= 7,11 với (1,98) độ tự do. Từ bảng A4, điểm 1% của F(1,60) là 7,08 và F(1,120) là 6,85. Điểm 1% cho F(1,98) do đó nằm giữa 7,08 và 6,85 và F=7,11 có ý nghĩa ở mức 1%.

Quan hệ giữa hệ số tương quan và bảng phân tích phương sai

Bảng phân tích phương sai cho phép lí giải hệ số tương quan theo cách khác: bình phương của hệ số tương quan, r2, bằng với tổng bình phương hồi quy chia cho tổng bình phương toàn bộ (0,262=0,0676=1,48/21,87) và do đó là tỉ lệ biến thiên được giải thích bằng hồi quy. Chúng ta có thể nói chiều cao người mẹ giải thích cho 6,76% tổng số biến thiến của trọng lượng lúc sinh.

Hồi quy bội với 2 biến số

Xét quan hệ của trọng lượng lúc sinh với hai biến: chiều cao của mẹ và tuổi thai. Phân tán đồ cho thấy rằng quan hệ giữa trọng lượng lúc sinh với hai biến kia là tuyến tính. Hệ số tương quan tương ứng là 0,26 (P<0,01) và 0,39 (P<0,001). Mối quan hệ kết hợp có thể được thể hiện bằng phương trình hồi quy bội (multiple regression equation),

$$y = a + b_1x_1 + b_2x_2$$

Và trong thí dụ này

Trọng lượng lúc sinh = a + b1 chiều cao của mẹ + b2tuổi thai

Nó có nghĩa là với bất cứ tuổi thai nào, trọng lượng lúc sinh quan hệ tuyến tính với chiều cao của mẹ và cũng vậy, với bất cứ chiều cao của mẹ nào, trọng lượng lúc sinh liên hệ tuyến tính với tuổi thai. b1và b2 được gọi là hệ số hồi quy riêng **phần** (partial regression coefficients) và sự tương quan tương ứng là tương quan **riêng phần (partial correlations). Lưu ý rằng b1 và b2 khác nhau hệ số hồi quy thông** thường (đối với một biến số) trừ khi hai biến số này không liên quan với nhau.

Mặc dù bây giờ khó có thể hình dung được đường hồi quy, ta cũng sử dụng những nguyên lí tương tự. Mỗi trọng lượng lúc sinh quan sát được được so sánh với (a + b1 chiều cao của mẹ + b2tuổi thai). a, b1 , b2 được chọn để cực tiểu hóa tổng bình phương các hiệu số này. Kết quả được trình bay trong bảng phân tích phương sai (Bảng 10.2). Có hai độ tự do của hồi quy bởi vì có hai biến số giải thích.

Bảng 10.2 Phân tích phương sai của hồi quy tuyến tính trọng lượng lúc sinh theo chiều cao của mẹ và tuổi thai (n=100)

Nguồn biến thiên	Tổng bình phương (SS)	Độ tự do (d.f.)	Trung bình bình phương (MS=SS/d.f.)	MS hồi qui F=---------- MS phần dư
Hồi quy theo chiều cao của mẹ và tuổi thai	4,43	2	2,2150	12,32,P<0,001
Phần dư	17,44	97	0,1798	
Tổng số	21,87	99		

Kiểm định F của hồi quy này là 12,32 với (2,97) độ tự do và có ý nghĩa ở mức 0,1%. Hồi quy đóng góp cho 20,26% (4,43/21,87) tổng số biến thiên. Tỉ lệ này bằng R2, R = √ 0,2026=0,45

được gọi là hệ số tương quan bội (multiple correlation coefficient). R luôn luôn dương tính bởi vì không thể xác định hướng của liên quan khi có nhiều biến số.

Tổng bình phương do hồi quy của trọng lượng lúc sinh do chiều cao của mẹ và tuổi thai gồm tổng bình phương được giải thích bởi chiều cao của mẹ (tính như trong hồi quy tuyến tính đơn) cộng với tổng bình phương được giải thích bởi tuổi thai sau khi đã được giải thích bằng chiều cao của mẹ (Bảng 10.3a). Tổng này có thể được kiểm định ý nghĩa bằng kiểm định F dùng trung bình bình phương phần dư

$$F = 2,95/0,1798=16,41 \quad d.f.=(1,97),\ P<0,001$$

Do đó mô hình của hai biến số là sự cải thiện so với mô hình chỉ dựa trên chiều cao của mẹ bởi vì tác động của tuổi thai có ý nghĩa ngay cả khi đã tính đến chiều cao của mẹ.

Bảng 10.3 Sự đóng góp của chiều cao của mẹ và tuổi thai vào hồi quy bội bao gồm cả hai biến

(a) chiều cao của mẹ đưa vào hồi quy bội trước

Nguồn biến thiên	Tổng bình phương (SS)	Độ tự do (d.f.)	Trung bình bình phương (MS=SS/d.f.)	MS hồi qui F=---------- MS phần dư
Chiều cao của mẹ	1,48	1	1,48	8,23,P<0,01
Tuổi thai sau khi đã điều chỉnh theo chiều cao	2,95	1	2,95	16,41, P<0,001
Chiều cao của mẹ và tuổi thai	4,43	2		

(b) tuổi thaiđưa vào hồi quy bộ trước

Nguồn biến thiên	Tổng bình phương (SS)	Độ tự do (d.f.)	Trung bình bình phương (MS=SS/d.f.)	MS hồi qui F=---------- MS phần dư
Tuổi thai	3,33	1	3,33	18,52,P<0,001
Chiều cao sau khi đã điều chỉnh theo tuổi thai	1,10	1	1,10	6,12, P<0,025
Chiều cao của mẹ và tuổi thai	4,43	2		

Theo cách khác, tổng bình phương được giải thích bởi hồi quy với cả hai biến bao gồm tổng bình phương được giải thích bởi tuổi thai (tính bằng hồi quy tuyến tính đơn) cộng với tổng bình phương được giải thích do chiều cao của mẹ sau khi đã điều chỉnh theo tuổi thai. Điều này trình bày trong Bảng 10.3(b). Một lần nữa chúng ta thấy rằng mô hình hai biến số có cải thiện so với mô hình chỉ có tuổi thai, bởi vì tác động của chiều cao của mẹ có ý nghĩa ngay cả khi tuổi thai đã được tính.

Hai cách tách tổng bình phương hồi quy kết hợp trong bàng 10.2 thành tổng bình phương riêng biệt không cho các bình phương kết quả thành phần giống nhau (Bảng 10.3a và b) bởi vì bản thân các biến giải thích (chiều cao của mẹ và tuổi thai) liên quan với nhau.

Hồi quy bội với nhiều biến

Hồi quy bội có thể mở rộng ra với một số các biến số, mặc dù cần nhắc rằng số các biến số nên tương đối nhỏ và với nhiều biến số việc lí giải trở nên phức tạp. Một số biến số, như tuổi và giới tính, có thể cần thiết đưa vào phương trình hồi quy bội bởi vì cần thiết điều chỉnh tác động của chúng trước khi nghiên cứu các quan hệ khác. Đưa vào những biến khác dựa trên chúng liên hệ nhiều hay ít với biến phụ thuộc. Mục tiêu là đạt được sự cân bằng giữa một mặt là bao gồm đủ các biến số để thu được sự phù hợp tốt nhất giữa phương trình hồi quy bội và

số liệu, mặt khác là không có quá nhiều biến số để quan hệ trở thành khó lí giải. Việc chọn lựa biến số có thể tiến hành theo một trong 3 cách:

1. Hồi quy bước tới (Step-up regression). Hồi quy tuyến tính đơn được tiến hành cho mỗi biến giải thích. Biến nào đóng góp phần trăm biến thiên lớn nhất được chọn và làm biến số đầu tiên. Sau đó tiến hành hồi quy bội hai biến bằng cách thêm vào từng biến số giải thích khác. Sau đó chọn hồi quy hai biến đóng góp phần trăm biến thiên lớn nhất. Quá trình này tiếp tục bằng cách chọn thêm một biến ở mỗi giai đoạn. Quá trình này ngừng khi (i) thêm vào bất kì biến số nào cũng không làm tăng có ý nghĩa phần đóng góp của nó hay (ii) Khi đã đạt được số các biến số tối đa đã định trước trong hồi quy bội.

2. Hồi quy bước lùi (Step down regression) Hồi quy bội được tiến hành bằng cách dùng tất cả các biến số. Sau đó các biến được loại bỏ từng biến một. Ở mỗi giai đoạn, biến được chọn để loại bỏ là biến đóng góp ít nhất vào việc giải thích các biến thiên. Quá trình này tiếp tục cho đến khi (i) tất cả các biến còn lại đều có ý nghĩa hay (ii) cho đến khi đã đạt được số các biến số tối đa đã định trước trong phương trình.

3. Hồi quy tổ hợp tối ưu (Optimal combination regression). Hồi quy từng bước theo cách 1 và 2 không nhất thiết đưa đến cùng một chọn lựa cuối cùng, ngay cả khi chúng cùng kết thúc ở một số các biến số giải thích nhất định. Không nhất thiết rằng chúng chọn được hồi quy tốt nhất cho một số các biến số giải thích. Cách tốt hơn là tìm một biến nào là tốt nhất, rồi từng cặp biến nào là tốt nhất, sau đó là từng cặp 3 biến nào là tốt nhất, bằng cách tiến hành hồi quy tất cả các tổ hợp có thể. Lưu ý rằng mặc dù cặp hồi quy tốt nhất thường chứa biến hồi quy đơn tốt nhất, nhưng điều đó không nhất thiết phải xảy ra.

Hồi quy bội với các biến giải thích rời rạc

Người ta thường muốn đưa vào các biến liên tục và rời rạc trong phân tích hồi quy bội. Thí dụ, trong nghiên cứu trọng lượng lúc sinh, sáu phụ nữ bị nhiễm mycoplasma trong lúc mang thai và trọng lượng trung bình của con họ sẽ nhỏ hơn. Yếu tố này có thể được đưa vào nhờ một biến số giả (dummy variable) của sự nhiễm trùng. Nó bằng 1 cho người phụ nữ bị nhiễm mycoplasma và bằng zero cho phụ nữ không bị. Phương trình hồi quy

Trọng lượng lúc sinh = a + b1 chiều cao của mẹ + b2 tuổi thai + b3 nhiễm trùng

Điều này tương đương với một cặp phương trình

(a) Trọng lượng lúc sinh = a + b1 chiều cao của mẹ + b2 tuổi thai + b3 cho người phụ nữ bị nhiễm mycoplasma

(b) Trọng lượng lúc sinh = a + b1 chiều cao của mẹ + b2 tuổi thai cho người phụ nữ không bị nhiễm mycoplasma

Hệ số b3 đo lường sự khác nhau trung bình của trọng lượng lúc sinh của con người mẹ bị nhiễm mycoplasma so với con người mẹ có cùng trọng lượng và tuổi thai và không bị nhiễm. Tổng bình phương gia số do nhiễm trùng mycoplasma được tìm bằng phương pháp đã được mô tả ở trên. Nó là hiệu số giữa tổng bình phương do hồi quy bội 3 biến trừ đi hồi quy chỉ dựa trên chiều cao của mẹ và tuổi thai. Nó có một độ tự do và được kiểm định ý nghĩa bằng kiểm định F.

Nhiễm trùng mycoplasma là một yếu tố có hai mức, có hay không. Yếu tố có hơn 2 mức, thí dụ như nhóm tuổi, được đưa vào bằng một loạt các biến giả để mô tả sự khác nhau. Nếu có k mức, sẽ cần k-1 biến giả và độ tự do bằng k-1. Xem chi tiết ở Armitage và Berry (1987).

Hồi quy bội với các biến giải thích phi tuyến tính

Người ta thường thấy quan hệ phi tuyến giữa biến phụ thuộc và biến giải thích. Có 3 cách sử dụng biến giải thích đó trong phương trình hồi quy bội. Phương pháp thứ nhất, phổ biến nhất là chia biến thành một số các nhóm nhỏ và xem nó như là một yếu tố với một mức tương ứng với một nhóm nhỏ, như được mô tả ở phần trên chứ không phải là biến liên tục. Thí dụ, tuổi

có thể chia thành nhóm 5 tuổi một. Quan hệ với tuổi được dựa trên sự so sánh trung bình trong mỗi nhóm tuổi và không cần giả thiết về dạng quan hệ với tuổi. Ở bước phân tích đầu tiên, người ta thường đưa biến giải thích vào dưới hai dạng liên tục và yếu tố. Hiệu số của tổng bìnhphương được dùng để đánh giá xem có thành phần phi tuyến trong mối quan hệ hay không. Trong phần lớn trường hợp, chia thành 3 tới 5 nhóm nhỏ là đủ để nghiên cứu tính phi tuyến của quan hệ.

Khả năng thứ nhì là tìm sự biến đổi thích hợp cho biến giải thích. Thí dụ, trong nghiên cứu trọng lượng lúc sinh, người ta thấy rằng trọng lượng lúc sinh có liên hệ tuyến tính với logarithm của thu nhập gia đình chứ không liên hệ tuyến tính với thu nhập gia đình. Dùng các phép biến đổi được thảo luận đầy đủ hơn ở Chương 19. Khả năng thứ ba là tìm cách mô tả đại số mối quan hệ. Thí dụ, nó có thể là dạng bình phương, trong trường hợp đó cả biến số (x) và bình phương của biến số (x2) sẽ được đưa vào phương trình.

Quan hệ giữa hồi quy bội và phân tích phương sai

Có nhiều sự trùng lắp giữa hồi quy bội và phân tích phương sai. Hồi quy bội trong đó các biến giải thích là rời rạc cũng giống như phân tích phương sai với nhiều yếu tố. Hai phương pháp cho cùng một kết quả giống nhau. Trong trường hợp này người ta khuyên nên dựa vào các chương trình máy tính có sẵn và tính dễ sử dụng của chúng để lựa chọn.

Một kĩ thuật khác không được mô tả ở đây là phân tích đồng phương sai (analysis of covariance) (Armitage & Berry 1987). Nó là một phương pháp khác nhưng tương đương để nghiên cứu sự khác biệt trong các nhóm khi có các biến giải thích liên tục. Một thí dụ của bài toán mô tả ở trên khi so sánh trọng lượng lúc sinh giữa các bà mẹ bị nhiễm mycoplasma trong thai kì và những bà mẹ không bị. Chiều cao của mẹ và tuổi thai là các biến giải thích bổ sung và sẽ được gọi là đồng biến số (covariate) trong trường hợp này. Lưu ý rằng đối khi phân tích đồng phương sai được gộp trong phân tích phương sai trong một số phần mềm máy tính.

Phân tích đa biến

Hồi quy bội, phân tích phương sai, hồi quy logistic (xem chương 13) và mô hình log tuyến tính (xem Chương 13) thường được gọi là những phương pháp đa biến (mulivariate methods), bởi vì chúng nghiên cứu biến phụ thuộc liên quan tới nhiều biến giải thích như thế nào. Theo nghĩa thống kê chặt chẽ, phân tích đa biến (multivariate analysis) có nghĩa là nghiên cứu những biến phụ thuộc thay đổi cùng với nhau như thế nào. Có bốn phương pháp thích hợp cho nghiên cứu y khoa được mô tả ngắn gọn ở đây. Xem chi tiết ở Armitage & Berry (1987).

Phân tích thành phần chính (pricipal component analysis) là phương pháp dùng để tìm một số các kết hợp các biến số, được gọi là thành phần để giải thích toàn bộ các biến thiên quan sát được và giảm tính phức tạp của số liệu. Phân **tích phân biệt (discriminat analysis) là phương pháp dùng để tìm một kết hợp duy** nhất các biến số, được gọi là hàm phân biệt (discriminant function) phân biệt tốt nhất các nhóm. Hàm số này được dùng để tiên đoán một cá nhân sẽ có thể thuộc vào nhóm nào. Thí dụ, nó đã được dùng để tiên đoán trẻ em có nguy cơ đột tử. **Phân tích yếu tố (factor analysis) là phương pháp thương được dùng trong các** trắc nghiệm tâm lí học. Nó tìm cách giải thích các trả lời cho các mục trắc nghiệm bị tác động bởi một số các yếu tố, như tình cảm, suy lý v.v. như thế nào. Cuối cùng, phân tích cụm (cluster analysis) la phương pháp xem xét các biến số sể xem một cá nhân có thể chia thành một hệ thống tự nhiên các nhóm. Các kĩ thuật được dùng bao gồm các kĩ thuật phân loại học số (numerical taxonomy), thành phần chính, và phân tích tương ứng (correspondance analysis).

XÁC SUẤT

Giới thiệu

Xác suất đã được dùng nhiều lần trong các chương trước, và nghĩa của nó rõ ràng trong bối cảnh của nó. Bây giờ ta sẽ trình bày nó chính thức hơn và đưa ra các quy tắc tính toán. Mặc dù xác suất là một khái niệm được dùng trong cuộc sống hàng ngày, chúng ta khó lòng định nghĩa nó chính xác được. Định nghĩa theo tần **suất** (frequentist definition) thường được dùng trong thống kê. Nó nói rằng xác suất xuất hiện của một biến cố bằng tỉ lệ số lần các biến cố đó xuất hiện trong một số lớn các lần thử giống nhau được lập lại. Nó có giá trị giữa 0 và 1, bằng 0 nếu sự kiện không thể xảy ra và bằng 1 nếu nó chắc chắn xảy ra. Xác suất có thể tính bằng phần trăm và có giá trị từ 0% đến 100%. Thí dụ, giả sử một đồng tiền được tung một ngàn lần và trong phân nửa các trường hợp nó nằm ngửa mặt và trong phân nửa trường hợp nó nằm xấp. Xác suất nó ngửa ở một lần tung sẽ là 50%.

Một cách khác là định nghĩa chủ quan (subjective definition) khi độ lớn của xác suất chỉ thể hiện mức độ tin tưởng của một người vào sự xuất hiện của biến cố. Định nghĩa này gần với cách dùng thường ngày và là cơ sở của phương pháp Bayes (Bayesian approach). Theo phương pháp này người nghiên cứu gắn một xác suất **tiền nghiệm (prior probability) cho biến cố được nghiên cứu. Sau đó người ta** tiến hành nghiên cứu, thu thập số liệu số liệu, và thay đổi xác suất tùy theo kết quả có được. Xác suất được biến cải này được gọi là xác suất hậu nghiệm (posterior probability). Phương pháp thống kê bắt nguồn từ phương pháp này hoàn toàn khác lạ và ít được sử dụng rộng rãi. Xem chi tiết ở Lindley (1965).

Tính toán xác suất

Có hai quy tắc căn bản của tính toán xác suất:

1. Quy tắc nhân (mulplicative rule) tính xác suất xuất hiện của cả hai biến cố A *và B*

2. Quy tắc cộng (additive rule) tính xác suất xuất hiện của biến cố (A) hay biến cố (B) hay cả hai

Quy tắc nhân

Xem một cặp vợ chồng dự định sẽ co 2 con. Có 4 tổ hợp giới tính của trẻ được trình bày trong bảng 11.1. Mỗi tổ hợp như nhau về mặt khả năng và có xác suất bằng 1/4.

Bảng 11.1 Các tổ hợp có thể của giới tính hai đứa trẻ với xác suất của chúng

		Trẻ thứ nhì	
Đứa thứ nhất		Trai	Gái
		1/2	1/2
Trai 1/2		1/4 (trai, trai)	1/4 (trai, gái)
Gái 1/2		1/4 (gái, trai)	1/4 (gái, gái)

Mỗi xác suất 1/4 được suy ra từ xác suất về giới tính của mỗi đứa trẻ. Xem chi tiết trường hợp 2 đứa trẻ là gái. Xác suất đứa trẻ đầu là gái là 1/2 và sau đó xác suất 1/2 của nó (1/2 của 1/2 là 1/4) rằng đứa trẻ thứ hai cũng là con gái. Do đó:

Xác suất (hai đứa trẻ là gái) = xác suất (đứa trẻ đầu là gái) × xác suất (đứa trẻ thứ nhì là gái) = 1/2 × 1/2 = 1/4

Quy tắc chung của xác suất xuất hiện hai biến cố là

Xác suất (A và B) = xác suất (A) × xác suất (B khi A đã xẩy ra).

Xác suất (B khi A đã xảy ra) được gọi là xác suất điều kiện (conditional probability) bởi vì nó là xác suất xảy ra biến cố B khi đã xảy ra biến cố A. Nếu khả năng của biến cố B không bị tác động bởi sự xuất hiện hay không xuất hiện của biến cố A và ngược lại, biến cố A và biến cố B được gọi là độc lập (independent) và qui tắc trở thành:

Xác suất (A và B) = xác suất (A) × xác suất (B), nếu A và B độc lập

Giới tính của đứa trẻ là biến cố độc lập bởi vì đứa trẻ sau là gái không bị tác động bởi giới tính của đứa bé trước. Một thí dụ của biến cố phụ thuộc (dependent) là xác suất một đứa trẻ ở Ấn độ bị thiếu máu và suy dinh dưỡng, bởi vì đứa trẻ sẽ dễ bị thiếu máu nếu nó bị suy dinh dưỡng.

Quy tắc cộng

Xác suất (A hay B hay cả hai) = xác suất (A) + xác suất (B) - xác suất (cả hai)

Quy tắc này được minh họa trong một thí dụ. Xem một vùng ở Nam Mỹ ở đó xác xuất bị nhiễm giun móc là 0,5 và xác suất bị sán máng là 0,6. Rõ ràng xác suất có hoặc là giun móc, hoặc là sán máng, hoặc là cả hai không phải là tổng số 0,5 + 0,6 = 1,1, bởi vì xác suất không thể lớn hơn 1. Vấn đề là xác suất bị nhiễm giun móc và sán máng được tính hai lần, một lần trong xác suất nhiễm giun móc, một lần trong xác suất nhiễm sán máng. Tính toán đúng như sau:

Xác suất (giun móc hay sáng máng hay cả hai) = xác suất (giun móc)+xác suất (sán máng) - xác suất (nhiễm hai loại)

Nếu hai bệnh độc lập với nhau

Xác suất (nhiễm hai loại) = 0,5 × 0,6 = 0,3

Xác suất (giun móc hay sáng máng hay cả hai) = 0,5 + 0,6 - 0,3 = 0,8

TỈ LỆ

Giới thiệu

Phương pháp được mô tả từ Chương 4 tới Chương 10 dành cho biến số liên tục. Chúng ta hãy xem xét phương pháp thích hợp cho tỉ lệ, đó là biến số rời rạc. Chúng ta bắt đầu bằng việc mô tả phân phối lấy mẫu của một tỉ lệ đựa trên phân phối nhị thức, và chúng ta sẽ mô tả phân phối nhị thức có thể được xấp xỉ bằng phân phối bình thường như thế nào.

Phân phối nhị thức

Tỉ lệ được dựa trên một biến số nhị phân (binary variable), trong đó giá trị cho mỗi cá thể trong một mẫu sẽ là một trong hai giá trị mà ta gọi là A hay B. Thí dụ, bệnh nhân sống sót (A) hay chết (B), một bệnh phẩm là dương tính (A) hay âm tính (B), hay một đứa trẻ được chích ngừa (A) hay không chích ngừa (B). Tỉ lệû (proportion) của A là số (r) các biến cố A chia cho tổng số trong mẫu

$$P = r/n$$

Số (và tỉ lệ) của biến cố A được quan sát trong mẫu dĩ nhiên sẽ bị biến thiên lấy mẫu (xem Chương 3), các mẫu khác nhau có thể có những giá trị khác nhau. Phân phối lấy mẫu được gọi là phân phối nhị thức (binomial distribution) và có thể được tính từ cỡ mẫu, n, và tỉ lệ dân số, p , như trong thí dụ 12.1. p (kí tự Hi lạp pi; không liên quan gì đến hằng số toán học 3,1416) là xác suất đáp ứng của một cá nhân là A.

Thí dụ 12.1

Một người đàn ông và đàn bà có tính trạng lặn của bệnh hồng cầu liềm (AS; nghĩa là dị hợp tử của gen hemoglobin hồng cầu liềm [S] và bình thường [A]) có 4 đứa con. Tính xác suất có không đứa trẻ nào, một, hai, hai, ba hay bốn đứa trẻ bị bệnh hồng cầu liềm (SS)?

Đối với mỗi đứa trẻ xác suất bị bệnh hồng cầu liềm (SS) là xác suất bị di truyền gen S từ cha và mẹ, đó là 0,5 × 0,5 = 0,25 theo quy tắc nhân xác suất (Chương 11). Xác suất không bị SS (đó là AS hay AA) do đó là 0,75. Chúng ta gọi SS là A và không phải SS là B do đó p = 0,25

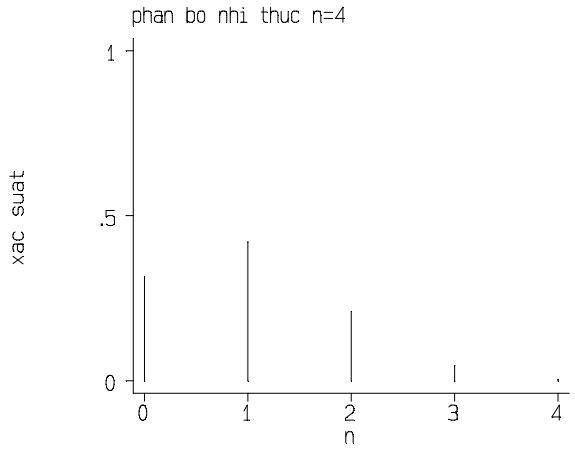
Xác suất của không đứa trẻ nào bị bênh là SS (r=0) là 0,75 × 0,75 × 0,75 × 0,75 = 0,754 = 0,3164, trong đó $0,75^4$ là 0,75 lũy thừa 4 nghĩa là 0,75 nhân với nhau 4 lần. Điều này là do quy tắc nhân xác suất.

Xác suất cho một đứa trẻ bị SS là xác suất (đứa trẻ thứ nhất bị SS và đứa trẻ thứ hai, thứ ba, thứ tư không bị SS) hay (đứa trẻ thứ hai bị SS và đứa trẻ thứ nhất, thứ ba, thứ tư không bị SS) hay (đứa trẻ thứ ba bị SS và đứa trẻ thứ nhất, thứ hai, thứ tư không bị SS) hay (đứa trẻ thứ tư bị SS và đứa trẻ thứ nhất, thứ hai, thứ ba không bị SS). Mỗi khả năng này có xác suất là 0,25 × 0,753 (quy tắc nhân xác suất) và bởi vì chúng không thể xảy ra đồng thời với nhau xác suất xảy ra một trong 4 khả năng trên là $4 \times 0,25 \times 0,75^3 = 0,4219$, theo quy tắc cộng xác suất.

Theo cách tương tự ta có thể tính xác suất có hai, ba, hay bốn đứa trẻ bị bệnh hồng cầu liềm bằng cách tính trong từng trường hợp, những chỉnh hợp có thể trong gia đình và cộng xác suất của chúng. Có được xác suất chỉ ra trong bảng 12.1. Lưu ý rằng tổng xác suất là 1, đó là phải có 1 trong 4 khả năng phải xảy ra. Xác suất cũng được dùng minh họa phân phối xác suất trong hình 12.1. Đây là phân phối nhị thức cho p = 0,25 và n = 4.

Bảng 12.1 Tính toán xác suất số trẻ em bị di truyền bịnh hồng cầu liềm (SS) trong một gia đình có 4 đứa trẻ trong đó cha mẹ có tính trạng hồng cầu liềm lặn (xác suất một cá nhân bị di truyền bệnh hồng cầu liềm là 0,25)

Bị hồng cầu liềm	Không bị hồng cầu liềm	số cách các tổ hợp có thể xảy ra	xác suất
0	4	1	$1 \times 1 \times 0{,}75^4$ = 0,3164
1	3	4	$4 \times 0{,}25 \times 0{,}75^3$ = 0,3164
2	2	6	$6 \times 0{,}25^2 \times 0{,}75^2$ =0,2109
3	1	4	$4 \times 0{,}25^3 \times 0{,}75$ =0,0469
4	0	1	$1 \times 0{,}25^4 \times 1$ =0,0039
			Tổng số = 1,000



Hình 12.1 phân phối xác suất của số trẻ trong gia đình 4 người bị bệnh hồng cầu liềm khi cả bố mẹ mang tính trạng lặn hồng cầu liềm. Xác suất đứa trẻ bị hồng cầu liềm là 0,25.

Công thức tổng quát cho phân phối nhị thức

Công thức tổng quát cho xác xuất có r biến cố A trong mẫu gồm n đối tượng khi xác suất xảy ra biến cố A cho mỗi đối tượng là p:

$$P(\text{r lần biến cố A}) = \frac{n!}{r!(n\text{-}r)!} n^r (1-r)^{n-r}$$

Trong phần thứ nhất của công thức là số các cách trong đó r biến cố A có thể có trong một mẫu n, phần thứ hai bằng xác suất của một trong các cách đó. Dấu chấm than kí hiệu giai thừa (factorial) của một số và có nghĩa là nhân tất cả với nhau các số nguyên từ số đó đến số 1. (0! được định nghĩa là bằng 1), pr có nghĩa là p lũy thừa r hay p nhân với nhau r lần. Một số bất kì lũy thừa zero bằng 1. Thí dụ áp dụng công thức trên trong thí dụ trên để tính xác suất có hai trong bốn đứa trẻ bị bệnh hồng cầu liềm là:

$$P(\text{2 trong số 4 đứa trẻ}) = \frac{4!}{2!(4\text{-}2)!} 0.25^2 (1-0.25)^{4-2} = 0.2109$$

Nhiều máy tính cầm tay có phím để tính giai thừa và lũy thừa. Nếu không có thể tính dễ dàng bằng cách dùng biểu thức sau, trong đó (r- n)! được triệt tiêu trong n!

$$\frac{n!}{r!(n-r)!} = \frac{n\times(n-1)\times(n-2)\times...\times(n-r-1)}{r\times(r-1)\times...\times3\times2\times1}$$

Thí dụ nếu n=18 r=5, (n-r+1)=18-5+1=14 và biểu thức trở thành

$$\frac{18\times17\times16\times15\times14}{5\times4\times3\times2\times1} = \frac{1028160}{120} = 8568$$

Độc giả quan tâm có thể thực tập công thức trên bằng cách tính lại các công thức trình bày ở bảng 12.1 và 12.3. Lưu ý rằng khi p = 0,5 thì (1-p) cũng bằng 0,5 và phần thứ hai của công thức bằng 0,5n

Tính chất của phân phối nhị thức

Hình 12.2 trình bày thí dụ của phân phối nhị thức cho những giá trị khác nhau của p và n. Những phân phối này đã được trình bày theo số lần xuất hiện, r, biến cố A trong mẫu, mặc dù chúng có thể áp dụng p, tỉ lệ của biến cố A. Thí dụ, khi cỡ mẫu bằng 5 các giá trị có thể của r là 0, 1, 2, 3, 4, 5 và trục hoành được kí hiệu theo đó. Các tỉ lệ tương ứng là 0, 0,2, 0,4, 0,6, 0,8, 1. Kí hiệu lại trục hoành theo giá trị các tỉ lệ sẽ có phân phối bình thường của p. Lưu ý rằng, mặc dù p là phân số, phân phối của nó là rời rạc và không liên tục, bởi vì nó chỉ gồm một số các giá trị giới hạn cho một cỡ mẫu nhất định.

Bởi vì phân phối nhị thức là phân phối lấy mẫu cho số (hay tỷ lệ) các biến cố A, trung bình của nó bằng trung bình dân số và độ lệch chuẩn của nó thể hiện sai số chuẩn, đo lường giá trị mẫu ước lượng giá trị dân số chính xác tới đâu. Trung bình dân số và sai số chuẩn được cho ở Bảng 12.2 cho số lần xảy ra, tỉ lệ và phần trăm biến cố A. Dĩ nhiên phần trăm là tỉ lệ nhân cho 100.

Bảng 12.2 Trung bình dân số và sai số chuẩn của số lần xuất hiện, tỉ lệ và phần trăm của biến cố A trong mẫu

	giá trị quan sát	trung bình dân số	sai số chuẩn
Số lần xuất hiện biến cố A	r	nπ	√ {Ûn π(1-p)}
Tỉ lệ của A	p=r/n	π	√□π (1-π)/n}
Phần trăm biến cố A	100p	100π	100 √ {π(1-π)/n}

Kiểm định ý nghĩa cho tỉ lệ đơn dùng phân phối nhị thức

Xem kết quả của thử nghiệm lâm sàng để so sánh 2 thuốc giảm đau, A và

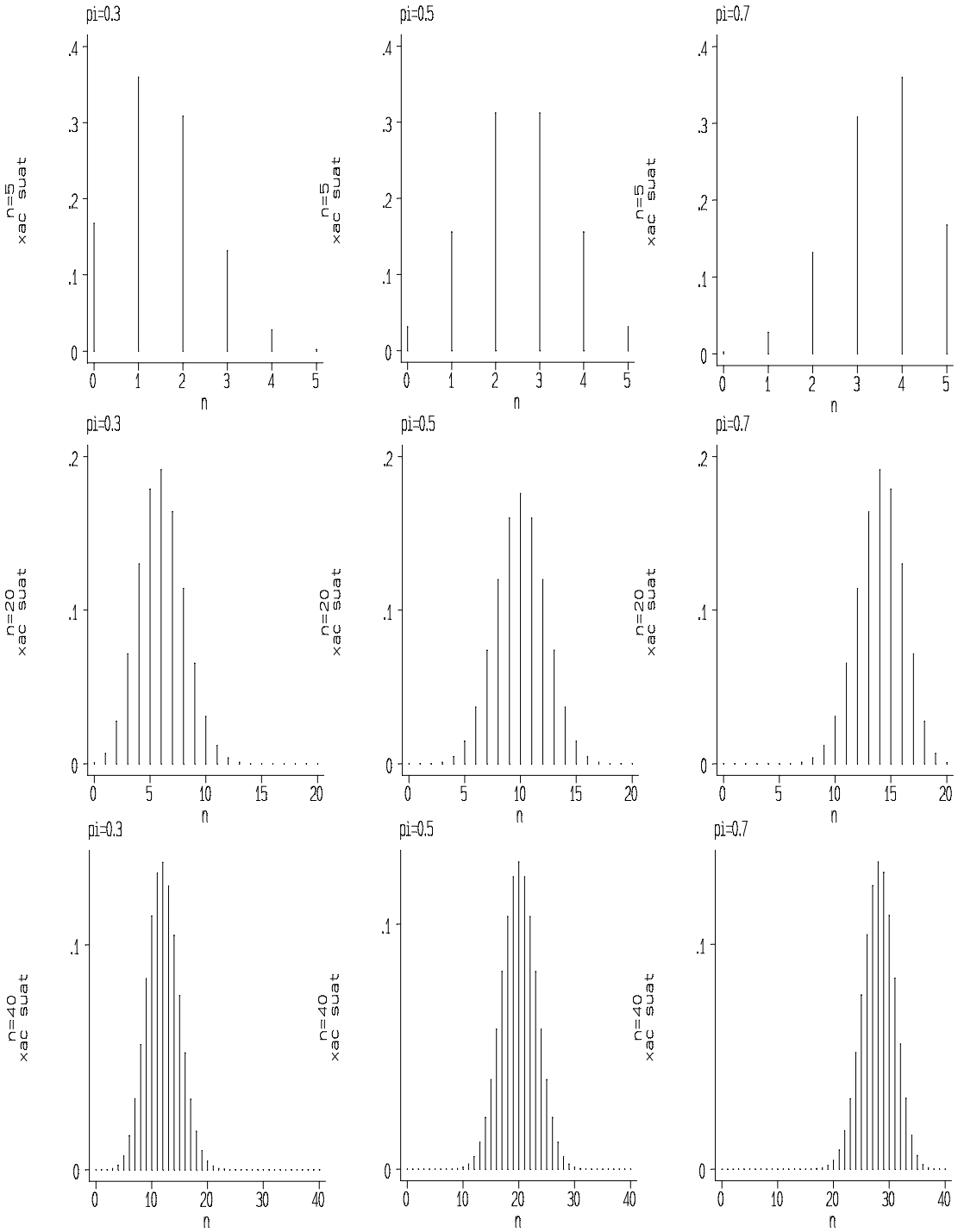
b. Mười hai người bị migrain được cho thuốc A vào lúc này và thuốc B vào lúc khác, để cho mỗi bệnh nhân được cho mỗi loại thuốc theo thứ tự nghẫu nhiên. Sau đó bệnh nhân được hỏi loại thuốc nào giảm đau tốt hơn. Chín bệnh nhân đáp ứng với thuốc A và 3 bệnh nhân đáp ứng với thuốc B. Điều này có nghĩa là thuốc A thực sự hiệu quả hơn hay không hay sự khác nhau chỉ do tình cờ?

Như trong Chương 6 kiểm định ý nghĩa được dùng để quyết định xem cách lý giải nào là khả dĩ hơn. Giả thuyết trung tính là hai thuốc các tác dụng bằng nhau. Nếu điều này đúng, bệnh nhân phải có cơ hội đáp ứng với A và B như nhau (giả thiết rằng có một vài lựa chọn). Phân phối lấy mẫu đối với số lần xuất hiện của đáp ứng với thuốc A phải là phân phối nhị thức với p =0,5 và n = 12.

Chương 6 cho biết mức ý nghĩa là xác suất có kết quả bằng hay cực đoan hơn kết quả quan sát được. Như trong ví dụ này, đó là xác suất xuất hiện đáp ứng với thuốc A là chín, mười, mười một hay mười hai cộng với xác suất xuất hiện đáp ứng với A là ba, hai, một hay không. Trong một kiểm định ý nghĩa 2 đuôi, chúng ta phải tính cả hai đuôi của phân phối. Phân phối nhị thức là phân phối đối xứng với p = 0,5. Do đó trong ví dụ này, xác suất xuất hiện đáp ứng với thuốc A 3 lần cũng bằng với xác suất xuất hiện đáp ứng với thuốc A 9 lần, cũng như

trong 12 tung một đồng tiền thì có khả năng có 3 lần ngửa và 9 lần xấp cũng bằng 9 lần ngửa và 3 lần xấp.

Xác suất xuất hiện chín, mười, mười một hay mười hai đáp ứng với A được ghi ở bảng 12.3 và tổng số là 0,0729. Xác suất xuất hiện (3,2,1,0) đáp ứng với thuốc A theo hướng khác cũng là 0,0729. Do đó xác suất có kết quả bằng hay lớn hơn xuất hiện 9 đáp ứng với thuốc A từ 12 bệnh nhân do tình cờ là 0,1458 hay 14,48%. Bởi vì xác suất này quá lớn chúng ta kết luận rằng thử nghiệm không xác định sự khác biệt có ý nghĩa về hiệu quả các thuốc.



Hình 12.3 Phân phối nhị thức cho các giá trị các nhau của p và n. Các giá trị trên trục hoành là r.

Tính toán mức ý nghĩa khi p ≠ 0,5

Nói chung, chúng ta có thể dùng phương pháp trên để kiểm định ý nghĩa của sự khác biệt giữa tỉ lệ của mẫu, p với bất kì giá trị giả thuyết trung tính của tỉ lệ dân số, p. Dù vậy khi p không phải là 0,5 thì phân phối nhị thức không đối xứng và xác suất của đuôi trên và đuôi dưới không giống nhau. Ta có thể tính mức ý nghĩa bằng cách cộng vào (i) xác suất của giá trị mẫu quan sát được và các giá trị cực đoan hơn theo hướng đó và (ii) xác suất của mỗi giá trị theo hướng ngược lại mà những xác suất này nhỏ hơn xác suất của mẫu quan sát được. Theo cách khác là ta có thể tính tổng các xác suất theo hướng quan sát, đó là xác suất (i) và

nhân đôi nó. Kết quả khác nhau nhưng trên thực tế nó không tác động đến việc đánh giá sự khác biệt quan sát được là do tình cờ hay do tác động có thực. Không có phương pháp nào hay hơn phương pháp nào nhưng phương pháp thứ nhì dễ tiến hành hơn.

Bảng 12.3 Tính toán mức ý nghĩa cho thử nghiệm lâm sàng trong đó 9 trong 12 bệnh nhân thích thuốc A hơn thuốc B

Số bệnh nhân thích thuốc A		Xác suất
Giá trị quan sát	9	$220 \times 0{,}512=0{,}0537$
Các giá trị cực đoan hơn	10	$66 \times 0{,}512=0{,}0161$
	11	$12 \times 0{,}512=0{,}0029$
	12	$1 \times 0{,}512=0{,}0002$
Tổng cộng: 9,10,11,12		0,0729
Đuôi ngược lại: 3,2,1,0		0,0729
Tổng số (mức ý nghĩa ở hai đuôi)		0,1458

Xấp xỉ phân phối bình thường của phân phối nhị thức

Khi cỡ mẫu,n , lớn tính toán xác suất cho mỗi giá trị của một đuôi của phân phối nhị thức để tính kiểm định ý nghĩa có thể rất mất công. May mắn thay khi n tăng phân phối nhị thức trở thành rất gần với phân phối bình thường (xem Hình 12.2). Trong thực tế, phân phối bình thường là xấp xỉ hợp lí của phân phối nhị thức nếu cả np và n-np lớn hơn hoặc bằng 5. Phân phối xấp xỉ bình thường có trung bình và độ lệch chuẩn như trong phân phối nhị thức (xem bảng 12.2)

Kiểm định ý nghĩa và khoảng tin cậy dùng xấp xỉ bình thường

Phân phối bình thường có thể dùng để kiểm định ý nghĩa tỉ lệ trong cách giống y như chúng ta đã làm để kiểm định ý nghĩa trung bình (xem Chương 6 và 7) và có thể dùng để xây dựng khoảng tin cậy

Kiểm định ý nghĩa dùng tỉ lệ đơn

Kiểm định bình thường cho một tỉ lệ đơn dùng công thức

$$Z = \frac{p-\pi}{s.e.(p)} = \frac{p-\pi}{\sqrt{\pi(1-\pi)/n}}$$

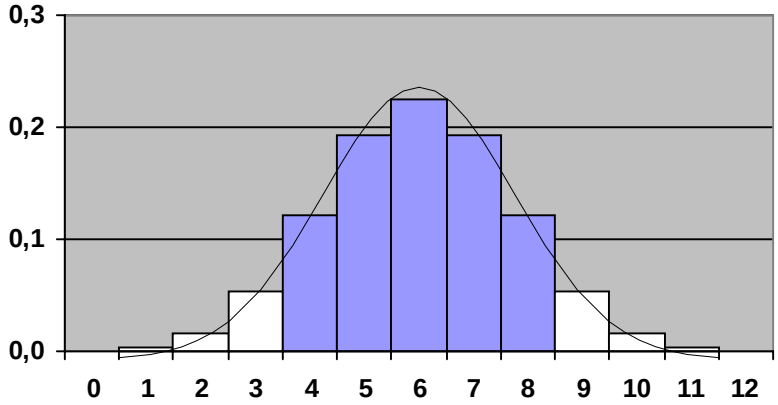
Thí dụ trong thử nghiệm lâm sàng để so sánh hay loại thuốc đã mô tả ở trên, tỉ lệ của bệnh nhân thích thuốc A là 9/12 = 0,75 so sánh với giá trị giả thuyết trung tính là p = 0,5

$$Z = \frac{0.75-0.5}{\sqrt{0.5\times 0.5/12}}$$

Giá trị 1,73 nằm ở giữa điểm 5 và 10% cua phân phối bình thường chuẩn (Bảng A2). Mức ý nghĩa chính xác của nó có thể được tính từ bảng Bảng A1 và là 2 (0,0418=0,0836 hay 8,36%. Số này nhỏ hơn đáng kể so với giá trị 14,58% đươc tính từ xác suất nhị thức. Lí do là chúng ta đã dùng phân phối liên tục, phân phối bình thường, để xấp xỉ một phân phối rời rạc, phân phối nhị thức. Tình huống này được hiệu chỉnh bằng cách đưa cái gọi là hiệu chỉnh tính liên tục (continuity correction) vào công thức kiểm định bình thường

$$z = \frac{|p-\pi| - 1/(2n)}{\sqrt{\pi(1-\pi)/n}}$$

Trong đó |p-p| có nghĩa là giá trị tuyệt đối của p - p hay nói cách khác đó là p - p không tính đến dấu trừ nếu có.



Hình 12.3 Phân phối nhị thức (n=12, p =0,5) ứng dụng để tính số bệnh nhân thích thuốc A, cùng với xấp xỉ bình thường của nó. Các vùng gạch chéo cho thấy xác suất số bệnh nhân thích thuốc A lớn hơn hay bằng 9 và minh họa việc sử dụng hiệu chỉnh tính liên tục.

Tác động của hiệu chỉnh tính liên tục là giảm giá trị z và cải thiện cho nó phù hợp với phân phối nhị thức. Trong thí dụ này, giá trị z có hiệu chỉnh tính liên tục là

$$z = \frac{|0{,}75 - 0{,}5| - 1/24}{\sqrt{\{0{,}5 \times 0{,}5/12\}}} = 1{,}44$$

Từ bảng A1, mức ý nghĩa của z = 1,44 là 2 × 0,0749=0,1498 hay 14,98% khá phù hợp giá trị 14,58% từ kiểm định nhị thức.

Lí luận của hiệu chỉnh liên tục có thể được hiểu rất rõ nhờ hình 12.3 trong đó phân phối bình thường và nhị thức được đặt chồng lên nhau. Cái mà ta đã làm khi hiệu chỉnh tính liên tục trên thực tế chính là tính toán vùng ở dưới phân phối bình thường bên trái r = 3,5 và bên phải r = 8,5, thay vì bên trái r = 3 và bên phải r = 9. Có thể thấy rằng nó phù hợp hơn với vùng gạch chéo thể hiện đuôi của phân phối nhị thức.

Khoảng tin cậy cho tỉ lệ đơn

Dùng phân phối nhị thức để rút ra khoảng tin cậy cho tỉ lệ rất phức tạp. Dù vậy người ta đã lập bảng các khoảng trong Geigy Scientific Tables (1982) cho các tỉ lệ khác nhau và cho cỡ mẫu tới 100. Một cách khác là dùng công thức xấp xỉ bằng cách sử dụng phân phối bình thường và phương pháp tương tự như đã mô tả ở Chương 5. Nó cho công thức

$$\text{c.i.} = p \pm z' \times \text{s.e.}, \qquad \text{s.e.} = \sqrt{[\hat{U}p(1-p)/n]}$$

Trong đó z' là điểm phần trăm tương ứng của phân phối bình thường chuẩn. Thí dụ đối với khoảng tin cậy 95% z'=1,96. Công thức này không được hiệu chỉnh tính liên tục và chỉ thích hợp khi cả np và n-np lớn hơn hoặc bằng 10.

Kiểm định ý nghĩa so sánh hai tỉ lệ

Kiểm định bình thường để so sánh hai tỉ lệ mẫu, p1=r1/n1 và p2=r2/n2 dựa trên

$$z = \frac{p_1 - p_2}{s.e.(p_1 - p_2)}$$

Trong só sai số chuẩn của p1-p2 dưới giả thuyết trung tính nghĩa là tỉ lệ dân số bằng nhau (p1=p2=p) là:

$$s.e.(p_1 - p_2) = \sqrt{\pi(1-\pi)(1/n_1 + 1/n_2)}$$

p là ước lượng tỉ lệ toàn bộ của cả hai mẫu đó là:

$$p = \frac{r_1 + r_2}{n_1 + n_2}$$

Nhưng trong trường hợp tỉ lệ đơn, kiểm định được cải thiện nhờ việc đưa vào hiệu chỉnh tính liên tục. Công thức là

$$z = \frac{|p_1 - p_2| - \{1/(2n_1) + 1/(2n_2)\}}{\sqrt{p(1-p)(1/n_1 + 1/n_2)}}, \qquad p = \frac{r_1 + r_2}{n_1 + n_2}$$

(Như đã trình bày ở trên, |p1-p2| là kích thước của hiệu số, bỏ qua dấu của nó). Kiểm định bình thường này là một xấp xỉ hợp lí cho hoặc là (i) n1+n2 lớn hơn 40 hay (ii) n1p, n1-n1p, n2p, n2-n2p đều lớn hơn hoặc bằng 5. Kiểm định chính xác được mô tả ở Chương 4 khi điều kiện không được thỏa mãn.

Khoảng tin cậy cho hiệu số giữa hai tỉ lệ

Công thức tính khoảng tin cậy cho hiệu số giữa 2 tỉ lệ là

$$c.i. = (p_1 - p_2) \pm (z \times s.e.), \quad s.e. = \sqrt{p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2}$$

Trong đó z là điểm phần trăm thích hợp của phân phối bình thường. Xấp xỉ này hợp lí nếu n1p, n1-n1p, n2p, n2-n2p đều lớn hơn hoặc bằng 10 và sẽ tốt hơn nếu những số này lớn hơn.

Thí dụ 12.2

Xét kết quả của một thử nghiệm vaccin cúm được tiến hành trong một vụ dịch. Hai mươi trong 240 (p1=0,083 hay 8,3%) người được tiêm chủng vaccin thực mắc bệnh cúm so với 80 trong số 220 người(p2=0,364 hay 36,4%) người được chích placebo. Đây có phải là bằng chứng thuyết phục rằng vaccin có hiệu quả.

Tỉ lệ toàn bộ mắc cúm là

$$p = \frac{20 + 80}{240 + 220} = \frac{100}{460} = 0{,}217 \quad \text{hay} \quad 21{,}7\%$$

Và như vậy

$$z = \frac{|0{,}083 - 0{,}364| - (1/480 + 1/440)}{\sqrt{0{,}217(1-0{,}217)(1/240 + 1/220)}}$$
$$= \frac{0{,}281 - 0{,}0044}{0{,}0385} = 7{,}18$$

Giá trị 7,18 lớn hơn 3,29 là điểm 0,1% cho phân phối bình thường chuẩn. Do đó chúng ta kết luận rằng giảm nguy cơ nhiễm cúm sau khi tiêm chủng vaccin thật với mức ý nghĩa cao (P<0,001). Khoảng tin cậy của mức giảm là:

$$(p_1 - p_2) \pm 1{,}96\sqrt{p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2}$$
$$= -0{,}281 \pm (1{,}96 \times 0{,}037) = -0{,}354; -0{,}208$$

Nghĩa làm xác suất mắc cúm được ước lượng thấp hơn 20,8% đến 35,4% sau khi tiêm chủng vaccin cúm so với tiêm placebo.

Thí dụ này được minh họa một lần nữa ở Chương tiếp theo và trong phần thử nghiệm vaccin ở Chương 25, ở đó đưa vào khái niệm hiệu quả vaccin.

KIỂM ĐỊNH CHI BÌNH PHƯƠNG CHO BẢNG DỰ TRÙ

Giới thiệu

Trình bày số liệu của các biến định tính được mô tả ở Chương 2. Khi có hai biến định tính, số liệu được sắp xếp trong bảng dự trù (contigency table). Các phạm trù cho một biến số tạo thành hàng và các phạm trù cho biến số khác tạo thành cột. Cá nhân được đưa vào một ô thích hợp của bảng dự trù tùy theo giá trị của hai biến số. Bảng dự trù cũng được dùng cho các biến số định lượng rời rạ hay biến số định lượng liên tục khi các giá trị được phân nhóm.

Kiểm định chi bình phương (χ^2) được dùng để kiểm định xem có sự liên hệ giữa các biến số hàng và biến số cột hay không hay nói cách khác, sự phân phối của các cá nhân trong các phạm trù của một biến số có phụ thuộc vào sự phân phối trong các phạm trù của biến kia hay không. Khi bảng chỉ có hai hàng và hai cột điều này có nghĩa là so sánh 2 tỉ lệ.

Bảng 2 × 2 (so sánh hai tỉ lệ)

Kết quả của thử nghiệm cúm được mô tả ở Thí dụ 12.2 được sắp xếp trong theo dạng bảng trong Bảng 13.1(a). Trong 460 người lớn, 240 người được tiêm chủng vaccin cúm và 220 người được tiêm placebo. Tổng số 100 người bị cúm trong đó có 20 người ở trong nhóm vaccin và 80 người ở trong nhóm placebo. Bảng đó được gọi là bảng dự trù 2 × 2 (2 × 2 contigency table). Bởi vì hai biến số đó, loại vaccin và sự xuất hiện cúm, đều có 2 giá trị khả dĩ. Đôi khi nó được gọi là bảng **gấp 4 (fourfold table) bởi vì các đối tượng sẽ ở một bốn phạm trù khả dĩ.**

Bước đầu tiên trong việc lí giải số liệu bảng dự trù là tính toán tỉ lệ hay phần trăm thích hợp. Do đó tỉ lệ mắc cúm là 8,3% trong nhóm vaccin, 36,4% trong nhóm placebo và 21,7% toàn bộ. Sau đó chúng ta cần quyết định như vậy có đủ chứng cứ để xem vaccin có hiệu quả hay sự khác biệt là do tình cờ.

Điều này được tiến hành bằng kiểm định chi bình phương (chi square test) nhằm so sánh số quan sát trong một trong bốn phạm trù trong bảng dự trù với số kì vọng nếu không có sự khác biệt về hiệu quả giữa vaccine và placebo. Tổng số 100/460 người bị cúm và nếu vaccine và placebo có hiệu quả bằng nhau, tỉ lệ bị cúm trong hai nhóm cũng như vậy và có 100/460×240=52,2 người trong nhóm vaccine và 100/460 × 220=47,8 người trong nhóm placebo sẽ bị cúm. Tương tư như vậy sẽ có 360/460 × 240 =187,8 người và 360/480 × 220 người không bị cúm. Những số kì vọng này đươc trình bày trong bảng 13.1(b). Chúng cũng tạo tổng số hàng và tổng số cột tương tự như trị số quan sát. Giá trị chi bình phương có được bằng cách tính (quan sát - kì vọng)2/kì vọng cho mỗi ô trong bảng dự trù và cộng chúng lại.

$$\chi^2 = \Sigma \frac{(O-E)^2}{E}, d.f. = 1 \text{ âäú våïi baíng } 2 \text{ x } 2$$

Hiệu số giữa số quan sát được và số kì vọng càng lớn, giá trị χ^2 càng lớn và ít có thể sự khác biệt này là do tình cờ. Điểm phần trăm của phân phối χ^2 được trình bày trong bảng A5. Giá trị này phụ thuộc vào độ tự do và trong bảng 2 × 2 độ tự do bằng 1.

Trong thí dụ này

$$\chi^2 = \frac{(20-52,2)^2}{52,2} + \frac{(80-47,8)^2}{47,8} + \frac{(220-187,8)^2}{187,8} + \frac{(140-172,2)^2}{172,2}$$
$$= 19,86 + 21,69 + 5,52 + 6,02 = 53,09$$

53,09 lớn hơn 10,85, điểm 0,1% của phân phối χ^2 một độ tự do. Do đó xác suất của sự khác biệt quan sát được về bệnh cúm do tình cờ nhỏ hơn 0,1%, nếu không có sự khác biệt thực sự giữa vaccine và placebo. Do đó có thể kết luận rằng vaccin có hiệu quả.

Bảng 13.1 Kết quả thử nghiệm vaccine cúm

(a) Số quan sát

Cúm	vaccine	placebo	tổng số
có	20 (8,3%)	80 (36,4%)	100 (21,7%)
không	220	140	360
tổng số	240	220	460

(a) Số kì vọng

cúm	vaccine	placebo	tổng số
có	52,2	47,8	100
không	187,8	172,2	360
tổng số	240	220	460

Sự hiệu chỉnh tính liên tục

Giống như kiểm định bình thường, kiểm định chi bình phương đối với bảng 2 × 2 có thể được cải tiến nhờ hiệu chỉnh tính liên tục, thường được gọi là hiệu chỉnh **tính liên tục của Yates (Yates' continuity correction). Công thức như sau**

$$\chi^2 = \Sigma \frac{(|O-E|-\frac{1}{2}}{E}, \quad d.f. = 1$$

cho giá trị χ^2 nhỏ hơn, |O-E| có nghĩa là giá trị tuyệt đối của O-E hay nói cách khác, giá trị của O-E bỏ qua dấu của nó.

Trong thí dụ này giá trị của χ^2 lài

$$\chi^2 = \frac{(32-\frac{1}{2})^2}{52,2} + \frac{(32,2-\frac{1}{2})^2}{47,8} + \frac{(32,2-\frac{1}{2})^2}{187,8} + \frac{(32,2-\frac{1}{2})^2}{172,2}$$
$$= 19,25 + 21,02 + 5,35 + 5,84 = 51,46 \quad P < 0,001$$

So sánh với kiểm định bình thường

Kiểm định bình thường để so sánh hai tỉ lệ và kiểm định chi bình phương cho bảng dự trù 2 × 2 thực chất là tương đương với nhau và $\chi^2 = z^2$. Điều này đúng với cả khi có hay không có hiệu chỉnh tính liên tục, với điều kiện là nó cùng hiệu chỉnh hoặc không cùng hiệu chỉnh. Từ thí dụ 12.2, z^2 với (hiệu chỉnh tính liên tục) = 7,182 = 51,55 mà ngoại trừ sai số làm tròn nó giống hệt như giá trị χ^2 = 51,46 đã được tính ở trên. Kiểm định bình thường có ưu điểm là dễ tính khoảng tin cậy hơn. Dù vậy kiểm định χ^2 dễ áp dụng hơn và có thể mở rộng để so sánh nhiều tỉ lệ và dùng cho bảng dự trù lớn hơn.

Lưu ý rằng điểm phần trăm trong Bảng A5 cho kiểm định chi bình phương một độ tự do tương ứng với điểm phần trăm hai đuôi trong bảng A2 của phân phối bình thường. (Khái niệm kiểm định một đuôi hay hai đuôi không dùng đối với kiểm định chi bình phương có độ tự do lớn hơn bởi vì chúng bao gồm việc so sánh nhiều tỉ lệ (multiple comparison).)

Tính hợp lệ (validity)

Nên luôn luôn sử dụng hiệu chỉnh tính liên tục mặc dù chúng có tác động nhiều nhất khi số kì vọng nhỏ. Khi chúng rất nhỏ kiểm định chi bình phương (và kiểm định bình thường) không phải là xấp xỉ tốt, ngay cả khi có hiệu chỉnh tính liên tục và khi đó nên dùng kiểm định chính xác (exact test) cho bảng 2 × 2 (xem Chương 14). Cochran (1954) đề nghị sử dụng kiểm định

chính xác khi tổng số của bảng nhỏ hơn 20 hay khi nó ở giữa 20 và 40 và số nhỏ nhất trong bốn giá trị kì vọng nhỏ hơn 5. Do đó kiểm định chi bình phương hợp lệ khi tổng số phải lớn hơn 40 bất kể các giá trị kì vọng hay khi tổng số kì vọng ở giữa 20 và 40 với điều kiện tất cả các giá trị kì vọng phải lớn hơn hoặc bằng 5.

Bảng 13.2 Kí hiệu tổng quát cho bảng dự trù 2 × 2

Cúm	vaccin	placebo	tổng số
Có	a	b	e
Không	c	d	f
tổng số	g	h	n

Công thức tính nhanh

Nếu các số trong bảng dự trù được kí hiệu bằng các kí tự như trong bảng 13.2 thì công thức để tính chi bình phương nhanh hơn cho bảng 2 × 2 như sau:

$$\chi^2 = \frac{n(ad-bc)^2}{efgh} = \frac{460\times(20\times140-80\times220)^2}{100\times360\times240\times220} = 53{,}01$$

Mà ngoại trừ sai số làm tròn sẽ giống hệt như giá trị đã tính ở trên. Với hiệu chỉnh tính liên tục, công thức sẽ trở thành

$$\chi^2 = \frac{n(|\,ad-bc\,|-n/2)^2}{efgh}, \quad d.f.=1$$

Trong thí dụ này

$$\chi^2 = \frac{460\times(14800-230)^2}{100\times360\times240\times220} = 51{,}37$$

Kết quả này tương tự như như giá trị đã tính ở trên, nếu không xét đến sai số làm tròn.

Bảng lớn

Kiểm định chi bình phương có thể được áp dụng cho bảng lớn hơn, nói chung là bảng r × c, trong đó r kí hiệu số hàng trong bảng và c là số cột.

$$\chi^2 = \Sigma\frac{(O-E)^2}{E}, \quad d.f.=(r-1)\times(c-1)$$

Và không có hiệu chỉnh tính liên tục hay kiểm định chính xác cho bảng dự trù ngoại trừ bảng 2 × 2 . Cochran (1954) đã đề nghị rằng xấp xỉ của kiểm định chi bình phương sẽ hợp lệ nếu có ít hơn 20% số các giá trị kì vọng dưới 5 và không có giá trị kì vọng nào nhỏ hơn một. Có thể vượt qua hạn chế này bằng cách kết hợp các hàng (hay các cột) có giá trị kì vọng thấp.

Không có công thức tính nhanh cho bảng r × c (trường hợp đặc biệt 2 × c hay r × 2 sẽ được xét ở phần sau). Phải tính số kì vọng cho mỗi ô. Sử dụng các lí luận y như trong trường hợp bảng 2 × 2 . Qui tắc chung để tính số kì vọng là:

$$E = \frac{\text{tổng của cột} \times \text{tổng của hàng}}{\text{tổng số chung}}$$

Cần lưu ý rằng kiểm định chi bình phương chỉ hợp lệ nếu được áp dụng cho số thực tế trong các phạm trù khác nhau. Không bao giờ được áp dụng nó cho bảng chỉ có tỉ lệ hay phần trăm mà thôi.

Bảng 13.3 So sánh các nguồn nước chính được sử dụng bởi gia đình trong 3 làng ở Tây phi

(a) Số quan sát

NGUỒN NƯỚC	LÀNG A	LÀNG B	LÀNG C	TỔNG SỐ
Sông	20(40,0%)	32(53,3%)	18(45,0%)	70(46,7%)
Ao hồ	18(36,0%)	20(33,3%)	12(30,0%)	70(33,3%)
Suối	12(24,0%)	8(13,3%)	10(25,0%)	30(20,0%)
Tổng số	50(100,0%)	60(100,0%)	40(100,0%)	150(100,0%)

(b) Số kì vọng

NGUỒN NƯỚC	LÀNG A	LÀNG B	LÀNG C	TỔNG SỐ
Sông	23,3	28,0	18,7	70
Ao hồ	16,7	20,0	13,3	50
Suối	10,0	12,0	8,0	30
Tổng số	50	60	40	150

Thí dụ 13.1

Bảng 13.3(a) trình bày kết quả của cuộc điều tra so sánh nguồn nước chính trong 3 xã ở Tây châu Phi. Trong bảng trình bày số và phần trăm các gia đình dùng, nước sông, nước ao, hay suối. Thí dụ trong làng A, 40% sử dụng nước sông chủ yếu, 36% nước ao hồ, 24,0% sử dụng giếng. Việc tính toán các phần trăm là cần thiết trong việc lí giải số liệu của bảng dự trù. Nói chung, 70 trong 150 hộ dùng nước giếng. Nếu không có sự khác biệt giữa các làng, người ta có thể cho rằng tỉ lệ dùng nước sông là giống nhau trong mỗi làng. Do đó số kì vọng của số hộ dùng nước sống là

$$70 \times 50/150 = 23,3 \qquad 70 \times 60/150 = 28,0 \qquad 70 \times 40/150 = 18,7$$

Số kì vọng có thể được tính bằng cách áp dụng quy tắc chung. Thí dụ số kì vọng của hộ dùng nước sống trong làng B là:

$$\frac{\text{tæng cuéa hµng (sµng)} \times \text{tæng cuéa cét (B)}}{\text{tæng sè chung}} = \frac{70 \times 60}{150} = 28,0$$

Số kì vọng của toàn bộ bảng được trình bày trong bảng 13.3(b)

$$\chi^2 = \Sigma \frac{(O-E)^2}{E}$$

$$= (20\text{-}23,3)^2/23,3 + (32\text{-}28,0)^2/28,0 + (18\text{-}18,7)^2/18,7 +$$
$$(18\text{-}16,7)^2/16,7 + (20\text{-}20,0)^2/20,0 + (12\text{-}12,3)^2/13,3 +$$
$$(12\text{-}10,0)^2/10,0 + (8\text{-}12,0)^2/12,0 + (10\text{-}8,0)^2/8,0$$
$$= 3,53$$

$$df = (r-1) \times (c-1) = 2 \times 2 = 4$$

Bởi vì 3,53 nhỏ hơn 5,39 (điểm 25% của χ^2 4 độ tự do), có thể kết luận rằng không có sự khác biệt ý nghĩa giữa các làng về phần trăm số hộ dùng các nguồn nước khá nhau (P>0,25)

Công thức ngắn gọn cho bảng 2 × c

Kiểm định chi bình phương được áp dụng cho bảng 2 × c , đó là bảng chỉ có 2 hàng trình bày sự khác biệt giữa c tỉ lệ thể hiện bởi c cột trong bảng. Công thức cô đọng hơn trong trường hợp này

$$\chi^2 = \frac{N^2[\Sigma(r^2/n) - R^2/N]}{R(N-R)}, \quad d.f. = c-1$$

Trong đó n thể hiện tổng số cho cột và r là giá trị của ô trên trong cột đó. r^2/n được tính cho mỗi cột trong bảng và tổng của chúng là $\Sigma (r^2/n)$. N là tổng số toàn bộ và R là tổng số cả hàng trên. (đối với bảng có 2 cột chứ không phải hai hàng, từ 'cột' và 'hàng' sẽ đổi chỗ cho nhau trong phần trình bày trên.)

Bảng 13.4 Suất mắc toàn bộ của Schistosoma mansoni theo nghề nghiệp

	Nghề nghiệp				
S. Manosi	Ngư dân	Nông dân	Buôn bán	thợ thủ công	tổng số
Dương tính	22(62,9%)	21 (48,8%)	17 (29,3%)	15 (51,7%)	75 (45,5%)
Âm tính	13	22	41	14	90
Tổng số	35	43	58	29	165

Thí dụ 13.2

Bảng 13.4 trình bày kết quả cuộc điều tra ở một vùng nông thôn ở Trung Phi để so sánh suất mắc toàn bộ Schistosoma mansoni trong các nghề nghiệp khác nhau. Áp dụng công thức ngắn gọn cho χ^2:

$$\Sigma (r^2/n) = 222/35 + 212/43 + 172/58 + 152/29$$
$$= 13{,}83 + 10{,}26 + 4{,}98 + 7{,}76 = 36{,}83$$
$$R^2/N = 75^2/165 = 34{,}09$$
$$\chi^2 = \frac{1652(36{,}83\text{-}34{,}09)}{75\times 90} = 11{,}05 \quad d.f. = 3$$

Điều này có ý nghĩa ở mức 2,5%, gợi ý rằng có thể có sự liên hệ giữa nguy cơ nhiễm bệnh và nghề nghiệp. Suất mắc toàn bộ của S. mansoni cao ở người ngư dân, thấp ở người buôn bán so với nông dân và thợ thủ công.

BỔ SUNG MỘT SỐ PHƯƠNG PHÁP CHO BẢNG DỰ TRÙ

Giới thiệu

Phân tích bảng dự trù được giới thiệu và kiểm định chi bình phương được mô tả ở Chương 13. Trong chương này chúng ta sẽ mô tả phương pháp cần dùng khi kiểm định chi bình phương đơn giản không thích hợp hay không đủ phân tích số liệu. Phạm trù đầu tiên bao gồm các trường hợp khi mẫu quá nhỏ nên việc xấp xỉ không đúng hay khi so sánh tỉ lệ bắt cặp (xem thí dụ ở bảng 14.3). Phạm trù thứ nhì là việc phân tích kết hợp một số bảng dự trù và tìm kiếm khuynh hướng tăng (hay giảm) đối với bảng tỉ lệ $2 \times c$.

Kiểm định chính xác cho bảng 2 2

Như đã trình bày trong Chương 13, xấp xỉ của kiểm định chi bình phương và kiểm định bình thường đối với bảng 2×2 có thể không hợp lệ nếu cỡ mẫu nhỏ. Cochran (1954) đề nghị sử dụng kiểm định chính xác (exact test), đôi khi đươc gọi là kiểm định chính xác của Fisher khi:

(i) Tổng số chung của bảng nhỏ hơn 20

(ii) Tổng sổ chung từ 20 đến40 và số nhỏ nhất trong 4 số kì vọng nhỏ hơn 5.

Kiểm định được mô tả dễ hiểu nhất trong trường hợp một thí dụ cụ thể. Bảng 14.1 trình bày kết quả từ một nghiên cứu để so sánh hai chế độ điều trị để kiểm soát chảy máu ở người bị hemophili trong khi phẫu thuật. Chỉ có 1 trong số 13 (8%) số người bị hemophili được điều trị bằng chế độ A bị chảy máu, so với 3 trong 12 (25%) người được điều trị bởi chế độ B. Những con số này nhỏ nên kiểm định chi bình phương không thể hợp lệ; tổng số chung là 25, nhỏ hơn 40 và số kì vọng nhỏ nhất là 1,9 (biến chứng trong chế độ điều trị B) nhỏ hơn 5. Do đó cần dùng kiểm định chính xác.

Kiểm định chính xác được dựa vào sự tính toán xác suất chính xác của bảng quan sát được và bảng 'cực đoan' hơn (more extreme table) với cùng tổng số hàng và cột, dùng công thức sau

Bảng 14.1 So sánh hai chế độ điều trị để kiểm soát chảy máu ở người bị bệnh hemophili trong phẫu thuật

	chế độ điều trị		
Biến chứng chảy máu	A	B	Tổng số
có	1	3	4
không	12	9	21
Tổng số	13	12	25

$$\text{Xaùc suaát chính xaùc cuûa baíng } 2 = \frac{e!f!g!h!}{n!a!b!c!d!}$$

Trong đó kí hiệu đã được định nghĩa ở Bảng 13.2. Dấu chấm than kí hiệu giai thừa của một số và có nghĩa là nhân tất cả các số nguyên từ số đó tới 1 với nhau. (0! được định nghĩa là 1). Nhiều máy tính cầm tay có phím bấm giai thừa, mặc dù biểu thức này có thể dễ dàng tính toán bằng cách đơn giản các thừa số ở tử và mẫu số. Xác suất chính xác cho bảng 14.1 là

$$\frac{4!\,21!\,13!\,12!}{25!\,1!\,3!\,12!\,9!} = \frac{4\times13\times12\times11\times10}{25\times24\times23\times22} = 0{,}2261$$

(21! được đơn giản với 25! thành $25 \times 24 \times 23 \times 22$).

X

Để kiểm định giả thuyết trung tính không có sự khác biệt giữa 2 chế độ điều trị, chúng ta cần tính không những xác suất của bảng quan sát được mà cả xác suất của những bảng cực đoan hơn có thể xảy ra tình cờ. Như vậy có 5 bảng có thể có mà có cùng tổng số hàng và cột như trong số liệu. Chúng được trình bày ở bảng 14.2 cùng với xác suất của nó. Trường hợp quan sát được ở bảng 14.2(b) với xác suất 0.2261. Với nhận định rằng cực đoan hơn nghĩa là kém khả năng hơn, bảng 14.2(a) và 14.2(e) cực đoan hơn với xác suất tương ứng là 0,0391 và 0,0565. Tổng các xác suất cần cho mức ý nghĩa là 0,2261 + 0,0391 + 0,0565 = 0,3217 và do đó sự khác biệt giữa các chế độ là không có ý nghĩa.

$$\text{Mức ý nghĩa (cách I)} = \text{xác suất của bảng quan sát được} + \text{xác suất của các bảng còn xác suất ít hơn}$$

Các thứ nhì hạn chế việc tính toán cho các bảng cực đoan cùng hướng với hướng đã quan sát và sau đó nhân đôi xác suất để có cả sự khác biệt theo hướng ngược lại. Trong thí dụ này, mức ý nghĩa thu được sẽ thu được sẽ gấp đôi tổng xác suất của Bảng 14.2(a) và 14.2(b) đó là $2 \times (0{,}0391 + 0{,}2261) = 0{,}534$. Không có phương pháp nào hay hơn phương pháp kia nhưng phương pháp dễ tiến hành hơn. Mặc dù hai phương pháp cho kết quả khác nhau nhưng trên thực tế cách lựa chọn ít khi tác động đến việc đánh giá xem sự khác biệt là do tình cờ hay là hiệu quả có thực.

$$\text{Mức ý nghĩa (cách II)} = 2 \times \left[\text{xác suất của bảng quan sát được} + \text{xác suất của các bảng cực đoan hơn theo hướng tương tự} \right]$$

Bảng 14.2 Những bảng có cùng tổng số hàng và cột như bảng 14.1 và xác suất của chúng

(a)

0	4	4
13	8	21
13	12	25

p = 0,0391

(b)

1	3	4
12	9	21
13	12	25

p = 0,2261

(c)

2	2	4
11	10	21
13	12	25

p = 0,4070

(d)

3	1	4
10	11	21
13	12	25

p = 0,2713

(e)

4	0	4
9	12	21
13	12	25

p = 0,0565

So sánh 2 tỉ lệ - trường hợp cặp đôi

Trong một số nghiên cứu chúng ta quan tâm đến việc so sánh hai tỉ lệ được tính từ các quan sát được cặp đôi (paired). Điều này xảy ra khi hai tỉ lệ cùng được đo lường trên cùng một cá

nhân (xem thí dụ dưới hay từ nghiên cứu bệnh chứng và thử nghiệm lâm sàng trong đó sử dụng quy hoạch bắt cặp đôi (matched paired design) (xem Chương 24 và 25)

Thí dụ, xét kết quả một thử nghiệm so sánh phương pháp Bell và Kato- Katz để phát hiện trứng Schistosoma masoni trong phân gồm 315 bệnh phẩm được phân tích bằng hai phương pháp. Kết quả được trình bày trong bảng dự trù 2 × 2 trong bảng 14.3(a), cho thấy sự phù hợp giữa hai phương pháp. Dù vậy chúng ta không quan tâm đến sự phù hợp ở đạy mà chúng ta quan tâm chủ yếu đến tỉ lệ bệnh phẩm dương tính của hai phương pháp. Đó là 238/315 (75,6%) khi dùng phương pháp Bell và 198/315 (62,9%) khi dùng phương pháp Kato-Katz. Lưu ýy rằng nếu trình bày số liệu như trong bảng 14.3(b) và áp dụng kiểm định chi bình phương là sai, bởi vì nó không tính đến bản chất cặp đôi của số liệu, đó là cùng 315 bệnh phẩm được xét nghiệm bởi cả hai phương pháp, chứ không phải là 630 bệnh phẩm khác nhau.

Bảng 14.3 So sánh phương pháp Bell và Kato-Katz để phát hiện trứng Schistosoma *mansoni trong phâm. 315 bệnh phẩm được xét nghiệm bằng cả hai phương pháp. Số liệu* từ Sleigh et *al*. (1982) *Transaction of the Royal Society of Tropical Medicine and Hygiene* **76**, 403-6 (đã xin phép)

(a) Trình bày đúng

		Kato-Katz		
		+	-	Tổng số
Bell	+	184	54	238
	-	14	63	77
Tổng số		198	117	315

(b) Trình bày sai

Kết quả	Bell	Kato-Katz	Tổng số
+	238	198	436
-	77	117	194
Tổng số	315	215	630

Phương pháp đúng như sau. Một trăm tám mươi bốn bệnh phẩm dương tính với cả 2 phương pháp và 63 âm tính với cả hai phương pháp. 274 bệnh phẩm này không cho thông tin nào về phương pháp tốt hơn trong việc phát hiện trứng S. mansoni. Thông tin mà chúng ta cần đến hoàn toàn nằm trong 69 trường hợp mà hai phương pháp không phù hợp. Trong số này có 54 trường hợp chỉ dương tính với phương pháp Bell so với 14 chỉ dương tính với phương pháp Kato-Katz.

Nếu không có sự khác biệt về khả năng phát hiện trứng S. Masoni, chúng ta sẽ không kì vọng sự phù hợp hoàn toàn bởi vì đó là 2 mẫu khác nhau, nhưng chúng ta kì vọng rằng có một phân nửa trường hợp không phù hợp sẽ chỉ dương tính với phương pháp Bell mà thôi và phân nửa sẽ chỉ dương tính với phương pháp Kato-Katz. Do đó kiểm định ý nghĩa thích hợp là so sánh tỉ lệ chỉ dương tính với phương pháp Bell (54/68) với giá trị giả thuyết là 0,5. Điều này có thể làm được nhờ kiểm định bình thường (với hiệu chỉnh tính liên tục) như đã được trình bày trong Chương 12. Nó tính ra

$$z = \frac{|54/68 - 0{,}5| - 1/(2\times 68)}{\sqrt{0{,}5\times 0{,}5/68}} = 4{,}73$$

Nói lên rằng phương pháp Bell tốt hơn có ý nghĩa (P<0,001) hơn phương pháp Kato-Katz trong việc phát hiện trứng S. masoni. (Lưu ý rằng sẽ thu được kết quả tương tự nếu so sánh tỉ lệ chỉ dương tính với Kato-Katz, đó là 14/68, được so sánh với 0,5.)

Kiểm định chi bình phương McNemar

Có kiểm định chi bình phương, được gọi là kiểm định chi bình phương McNemar (McNemar's chi squared test), dựa trên số các cặp không phù hợp (discordant pairs), r và s. Công thức có hiệu chỉnh tính liên tục là:

$$\chi^2_{\text{bắtcặp}} = \frac{(|r-s|-1)^2}{r+s}, \quad d.f. = 1$$

Trong thí dụ r = 54 và s = 14 cho

$$\chi^2_{\text{hiệu chỉnh}} = \frac{(|54-14|-1)^2}{54+14} = \frac{39^2}{68} = 22{,}36 \quad d.f. = 1, \quad P < 0{,}001$$

Ngoại trừ sai số làm tròn, giá trị χ^2 này giống hệt như với bình phương giá trị z có được ở trên (4,732 = 22,37), hai kiểm định tương đương về mặt toán học với nhau.

Tính hợp lệ

Sử dụng kiểm định chi bình phương McNemar hay kiểm định bình thường hợp lệ nếu tổng số các cặp không phù hợp phải lớn hơn hoặc bằng 10. Nếu có ít hơn 10 cặp không phù hợp, phải tính xác suất nhị thức chính xác như đã trình bày ở Chương 12.

Khoảng tin cậy

Tỉ lệ mẫu dương tính với phương pháp Bell là 0,7556 so với 0,6286 của phương pháp Kato-Katz. Hiệu số của hai tỉ lệ này là 0,1270. Hiệu số này cũng như sai số chuẩn của nó có thể được tính từ số các cặp không phù hợp, r và s và tổng số các cặp n.

$$\text{Hiệu số của hai tỉ lệ hiệu chỉnh} = \frac{r-s}{n}, \quad s.e. = \frac{\sqrt{r+s}}{n}$$

Ở đây sự khác biệt, (r-s)/n= 40/315=0.1270 giống như khi được tính trực tiếp từ tỉ lệ, với s.e. = (√ 68)/315=0,0262. Khoảng tin cậy xấp xỉ có thể được tính dựa trên phân phối bình thường.

$$c.i. = \frac{r-s}{n} \pm z' \frac{\sqrt{r+s}}{n}$$

Khoảng tin cậy 95% của hiệu số giữa tỉ lệ của bệnh phẩm dương tính với phương pháp Bell và Kato-Katz là:

$$0{,}1270 \pm 1{,}96 \times 0{,}0262 = 0{,}0756 \text{ đến } 0{,}1784$$

Điều này có nghĩa là tỉ suất dương tính cao hơn từ 7,6% và 17,8% khi dùng phương pháp Bell để phát hiện trứng S. mansoni so với khi dùng phương pháp Kato-Katz.

Phân tích nhiều bảng 2 × 2

Một điểm tổng quát rất quan trọng cần phải nhận thức khi tiến hành phân tích là khi tập hợp (pool) nhiều tập hợp số liệu có thể dẫn đến kết quả sai lạc. Thí dụ, xét một kết quả được trình bày ở bảng 14.4 từ một cuộc điều tra được tiến hành để so sánh tỷ suất mắc toàn bộ của kháng thể chống leptospirosis ở nông thôn và thành thị ở vùng West Indies. Điểm chính cần lưu ý ở đây là đối với nam hay nữ, tỷ suất mắc toàn bộ cao ở vùng nông thôn nhưng khi kết hợp hai giới tính thì không có sự khác biệt. Điều này là do sự kết hợp của hai yếu tố

(i) tỷ suất mắc toàn bộ của kháng thể không giống nhau ở nam và nữ. Nó thấp ở nữ, dù rằng ở nông thôn hay thành thị.

(ii) Mẫu từ vùng nông thôn và thành thị có cấu trúc giới tính khác nhau. Tỉ lệ nam là 100/200 (50%) ở mẫu thành thị nhưng chỉ có 50/200 (25%) ở mẫu nông thôn.

Bảng □1 So sánh tỉ suất mắc toàn bộ của kháng thế chống leptospiro ở giữa vùng thành thị và nông thôn.

(a) Đàn ông

Kháng thể	Nông thôn	Thành thị	Tổng số
Có	36 (72%)	50 (50%)	86
Không	14	50	64
Tổng số	50	100	150

$\chi^2 = 5,73$ d.f.=1 P<0,025

(b) Đàn bà

Kháng thể	Nông thôn	Thành thị	Tổng số
Có	24 (16%)	10 (10	34
Không	126	90	216
Tổng số	150	100	250

$\chi^2 = 1,36$ d.f.=1 P ≈ 0,25

(c) Đàn ông và đàn bà

Kháng thể	Nông thôn	Thành thị	Tổng số
Có	60 (30%)	60 (30%)	120
Không	140	140	280
Tổng số	200	200	400

Giới tính được gọi là biến số gây nhiễu (confounding variable) bởi vì nó có quan hệ của với biến số cần quan tâm (tỷ suất mắc toàn bộ của kháng thể) và nhóm được so sánh (thành thị và nông thôn). Nếu trong khi phân tích bỏ qua giới tính sẽ dẫn đến kết quả sai lệch. Do đó điều quan trọng là phân tích nam riêng, nữ riêng. Áp dụng kiểm định chi bình phương cho sự khác biệt giữa nông thôn và thành thị có ý nghĩa ở nam nhưng không có ý nghĩa ở nữ giới (Bảng 14.4). Kiểm định tổng kết (summary test) nhằm kết hợp hai mảng thông tin này được mô tả ở dưới và phương pháp chuẩn hóa để điều chỉnh tỉ lệ toàn bộ để phản ánh sự khác biệt về giới tính được trình bày ở Chương 15.

Trong trường hợp này, sự kết hợp số liệu của 2 giới tính sẽ che dấu sự khác biệt có thực. Trong một số tình huống khác, sự kết hợp những nhóm số liệu khác nhau có thể cho thấy sự khác biệt trong khi nó không tồn tại, thậm chí cho thấy sự khác biệt ngược với sự khác biệt có thực. Một thí dụ khác là việc đánh giá xem người bị nhiễm schistosoma có tỉ suất tử vong cao hơn người không bị nhiễm hay không. Trong trường hợp này cần phải tính đến tuổi bởi vì cả nguy cơ chết và nguy cơ bị nhiễm schistosoma đều gia tăng với tuổi. Nếu không xét đến tuổi, nhiễm schistosoma sẽ dường như sẽ có tỉ suất tử vong cao hơn, ngay cả khi nếu điều này không xảy ra trong thực tế, bởi vì những người bị nhiễm schistosoma thường lớn tuổi hơn và do đó có nhiều khả năng bị chết hơn so với những người trẻ không bị nhiễm schistosoma.

Kiểm định chi bình phương Mantel-Haenszel

Khi có biến số gây nhiễu, điều quan trọng là phải phân tích riêng biệt từng nhóm nhỏ các số liệu thích hợp. Dù vậy, cũng cần áp dụng một kiểm định có thể kết hợp tất cả các chứng cứ từ các nhóm riêng biệt nhưng có tính đến yếu tố gây nhiễu. Điều này đặc biệt đúng khi có sự khác biệt nhưng không có ý nghĩa ở trong từng nhóm nhỏ, bởi vì có thể là do cỡ mẫu nhỏ. Kiểm định chi bình phương Mantel Haenszel được dùng cho mục đích này khi số liệu chỉ gồm vài bảng 2×2 . Tính ba giá trị sau đây từ mỗi bảng và lấy tổng trên các bảng (kí hiệu được định nghĩa trong bảng 13.2):

(i) giá trị quan sát a

(ii) giá trị kì vọng của a, Ea=eg/n và

(iii) phương sai của a, Va=efgh/[n2(n-1)]

Giá trị chi bình phương (có hiệu chỉnh tính liên tục) là:

$$\chi^2_{MH} = \frac{(|\Sigma a - \Sigma E_a| - 0{,}5)^2}{\Sigma V_a}, \quad d.f. = 1$$

Lưu ý rằng nó chỉ có một độ tự do dù rằng nó tổng kết bao nhiêu bảng

Việc tính toán số liệu ở bảng 14.4 được trình bày trong bảng 14.5(a). Tổng số 60 người trong vùng nông thôn có kháng thể chống lại Leptospira so với số kì vọng là 49,1, dựa trên giả định không có sự khác biệt về tần suất mắc toàn bộ giữa nông thôn và thành thị. Giá trị χ^2 Mantel-Haenszel bằng 7,08 và có ý nghĩa ở mức 1%.

$$\chi^2 = \frac{(|60 - 49{,}1| - 0{,}5)^2}{15{,}2874} = \frac{10{,}4^2}{15{,}2874} = 7{,}08 \quad d.f. = 1, \quad P < 0{,}01$$

Có vẻ kì lạ là dường như kiểm định chỉ dựa hoàn toàn vào giá trị quan sát và giá trị kì vọng của a và không dựa vào các ô khác trong bảng. Dù vậy điều này không hẳn như vậy, bởi vì một khi có được giá trị a có thể tính được giá trị b, c, d từ các tổng số trong bảng. Nếu kiểm định Mantel-Haenszel được áp dụng cho bảng 2 × 2 thì giá trị χ^2 thu được sẽ gần với nhưng không chính xác bằng với giá trị χ^2 chuẩn. Nó hơn nhỏ hơn và bằng (n-1)/n nhân với giá trị chuẩn. Sự khác biệt này có thể bỏ qua cho n lớn hơn hoặc bằng 20, mà điều này cũng cần thiết cho tính hợp lệ của kiểm định chi bình phương.

Bảng 14.5 Tính toán kiểm định χ^2 Mantel-Haenszel, ứng dụng cho số liệu ở Bảng 14.4

(a) Kiểm định

Nhóm nhỏ	a	E_a	Va
Nam	36	28,7	$86 \times 64 \times 50 \times 100/(150^2 \times 149) = 8{,}2088$
Nữ	24	20,4	$30 \times 216 \times 150 \times 100/(250^2 \times 249) = 7{,}0786$
Tổng số	60	49,1	15,2874

(b) Quy tắc 5 để kiểm tra tính hợp lệ

Nhóm nhỏ	Tối thiểu của (e,g)	g - f	Tối đa của (0,g-f)
Nam	50	-14	0
Nữ	34	-66	0
Tổng số	84		0

Tính hợp lệ

Kiểm định χ^2 Mantel-Haenszel chỉ là xấp xỉ. Độ đầy đủ của nó được đánh giá bằng quy tắc số 5 (rule of 5). Ta tính thêm hai giá trị sau cho mỗi bảng và lấy tổng của các bảng

(i) Tối thiểu của (e,g), đó là số nhỏ nhất của e và g

(ii) Tối đại của (0,g-f), bằng không nếu g nhỏ hơn hay bằng f, bằng g- f nếu g lớn hơn.

Cả hai tổng số phải khác giá trị kì vọng, Σ Ea, ít nhất là 5 để cho kiểm định hợp lệ. Chi tiết của các tính toán này cho số liệu leptospira được trình bày ở bảng 14.5(b). Cả hai tổng số 84 và zero khác với 49,1 một số lớn hơn 5, do đó kiểm định χ^2 Mantel-Haenszel có giá trị.

Kiểm định chi bình phương định hướng

Kiểm định chi bình phương chuẩn cho bảng 2 × c là một kiểm định tổng quát để đánh giá có sự khác nhau trong c tỉ lệ haykhông. Khi các phạm trù trong một cột có các thứ tự tự nhiên thì một kiểm định hợp lí sẽ là tìm xem có định hướng tăng hay giảm của những tỉ lệ theo cột hay không. Thí dụ được đưa ra ở Bảng 14.6, cho thấy số liệu béo phì ở phụ nữ. Có thể thấy rằng tỉ lệ phụ nữ dậy thì sớm có bề dầy lớp mỡ dưới da ở vùng cơ tam đầu tăng. Định hướng này có thể được kiểmđ ịnh bằng kiểm định chi bình phương định hướng (chi squared test for trend), có một độ tự do.

Bước đầu tiên tiên là gán các điểm (score) cho các cột để mô tả định hướng. Thông thường là chọn số cột 1, 2, 3 v.v. Như ở đây (hay 0, 1, 2, 3 v.v.). Nó đại diện cho định hướng tăng (hay giảm) đều từ cột này sang cột khác và thích hợp trong hầu hết các trường hợp. Một trả năng khác là dùng bề dày lớp mỡ dưới da trung bình trong mỗi nhóm, phản ánh định hướng tuyến tính với các giá trị tuyệt đối của bề dầy lớp mỡ dưới da. (Lưu ý rằng hiệu số giữa kiểm định χ^2 chuẩn và χ^2 định hướng là giá trị χ^2 với (c-2) độ tự do để kiểm định sự sai lệch khỏi định hướng phù hợp (departures from the fitted trend).)

Bước tiếp theo là tính ba giá trị cho mỗi cột và cộng các kết quả.

Bảng 14.6 Quan hệ giữa bề dày lớp mỡ dưới da ở vùng cơ tam đầu và dậy thì sớm. Số liệu từ một nghiên cứu béo phì ở phụ nữ - Beckles et al. (1985), International Journal of *Obesity* **9**, 127-135.

	Bề dày lớp mỡ dưới da vùng cơ tam đầu			
Tuổi dậy thì	Mỏng	Trung bình	Dày	Tổng số
<12 tuổi	15(8,8%)	29(12,8%)	36(19,4%)	80
>=12 tuổi	156	197	150	503
Tổng số	171	226	186	583
Điểm (score) cho kiểm định định hướng	1	2	3	

(i) rx, tích số của r, ô trong hàng đầu tiên của cột và điểm x

(ii) nx, tích số của tổng số, n, của cột và điểm x

(iii) nx2, tích số của tổng số, n, của cột và bình phương của điểm, x2

Dùng N để kí hiệu tổng số chung và R là tổng hàng đầu tiên, công thức tính kiểm định chi bình phương định hướng (có hiệu chỉnh tính liên tục) như sau:

$$\chi^2_{\text{coihæåïng}} = \frac{(|A|-0{,}5)^2}{B}, \quad d.f. = 1$$

$$\text{trong âoï } A = \Sigma(rx) - \frac{R}{N}\Sigma(nx) \quad \text{vaì} \quad B = \frac{R(N-R)}{N^2(N-1)}[N\Sigma(nx^2) - (\Sigma nx)^2]$$

Có những dạng kiểm định khác nhau nhưng hầu hết đều tương đương về mặt đại số. Sự khác biệt duy nhất là một số dạng dùng (N-1) thay chỗ cho N trong khi tính toán B. Sự khác biệt này không quan trọng. Kiểm định được chọn ở đây tương đối đơn giản và có thể mở rộng cho một số bảng 2 × c . Trừ 0,5 ở hàng đầu tiên là *hiệu chỉnh tính liên tục. Nếu điểm (score) không phải là số cột thì phải loại bỏ chúng.*

Việc tính toán cho số liệu ở Bảng 14.6 như sau

Σ (rx) = $15 \times 1 + 29 \times 2 + 36 \times 3 = 181$

Σ (nx)= $171 \times 1 + 226 \times 2 + 186 \times 3 = 1181$

Σ (nx^2)= $171 \times 1 + 226 \times 4 + 186 \times 9 = 2749$

R = 80; N = 583; N-R = 503

$$A = 181 - \frac{80}{583} \times 1181 = 18{,}9417$$

$$B = \frac{80 \times 503}{583^2 \times 582} \times (583 \times 2749 - 1181^2) = 42{,}2927$$

$$\chi^2_{\text{coïhæång}} = \frac{(18{,}9417 - 0{,}5)^2}{42{,}2927} = 8{,}04 \quad d.f. = 1, \quad P < 0{,}005$$

Do đó định hướng gia tăng tỉ lệ phụ nữ dậy thì sớm với sự gia tăng bề dày lớp mỡ dưới da cơ tam đầu là có ý nghĩa cao (P<0,005)

Mở rộng cho một số bảng 2 × c

Khi số liệu có nhiều nhóm nhỏ mà không thể gộp lại bởi vì biến số gây nhiễu, phải tiến hành cho mỗi nhóm một kiểm định định hướng. Dù vậy có kiểm định tổng kết kết hợp các chứng cứ từ các kiểm định định hướng riêng biệt và có tính đến yếu tố gây nhiễu. Đó là kiểm định chi bình phương có một độ tự do, bất kể số nhóm nhỏ. A và B được tính cho mỗi bảng như đã mô tả ở trên và được lấy tổng. Giá trị χ^2 là:

$$\text{Toïm taät}\chi^2_{\text{coïhæång}} = \frac{(|\Sigma A| - 0{,}5)^2}{\Sigma B}, \quad d.f. = 1$$

Kĩ thuật phức tạp hơn

Hai kĩ thuật phức tạp được nhắc nhở ngắn ở đây bởi vì chúng là những phương pháp mạnh mẽ để phân tích tỉ lệ và bảng dự trù. Phương pháp đầu tiên là hồi **quy logistic (logistic regression), phương pháp để phân tích tỉ lệ tương tự như hồi** quy bội đối với biến liên tục. Phân tích một số bảng 2 × 2 (hay 2 × c) được xem là trường hợp đặc biệt của phương pháp này. Thí dụ, số liệu trong bảng 14.4 cho thấy tỉ suất mắc toàn bộ kháng thể leptospira thay đổi theo hai yếu tố giới tính và khu vực. Hồi quy logistic tổng quát hơn phương pháp chi bình phương mô tả ở trên bởi vì chúng cho phép bao gồm các biến liên tục và đánh giá sự tương tác giữa các biến (xem chương 8). Hồi quy logistic được gọi là logistic bởi vì nó nghiên cứu sự phụ thuộc của biến đổi logistic (logistic tranformation) của tỉ lệ theo theo một số biến giải thích. Biến đổi logistic hay logit, được định nghĩa như sau:

$$Logit = \log_e\left[\frac{\text{tè lãû}}{1 - \text{tè lãû}}\right]$$

Mô hình được cân chỉnh bằng kĩ thuật toán học gọi là độ khả dĩ cực đại (maximum likelihood) và cũng tính đến yếu tố là sự biến thiên của tỉ lệ có phân phối nhị thức

Kĩ thuật thứ hai tương tự là sử dụng mô hình tuyến tính log (log linear model). Cả hai phương pháp này có thể được tiến hành bằng cách sử dụng phần mềm GLIM (1978), mà trên của nó là các chữ đầu của mô hình tương tác tuyến tính tổng quát hóa (Generalized Linear Interactive Modelling). Một loạt các mô hình khác nhau được cân chỉnh giống như trong hồi quy bội. Ý nghĩa của một biến được đánh giá bằng cách chân chỉnh hai mô hình: một có biến số đó và một không. Kiểm định dựa trên sự khác biệt deviance (giống như tổng bình phương phần dư, xem Chương 10) của hai mô hình. Sự khác biệt này có phân phối χ^2 với số độ tự do bằng số tham số thêm vào. Xem thêm chi tiết ở Anderson et al. (1980) hay xem cách tiếp cận toán học ở Dobson (1984).

Có thể tìm thấy phần trình bày hồi quy logistic ở trong sách của Breslow và Day (1980) và Schlesselman (1982). Mặc dù chúng là được viết trong bối cảnh một nghiên cứu bệnh chứng (xem chương 24), phương pháp có thể áp dụng cho phân tích tỉ lệ. Những cuốn sách này cũng

trình bày hồi quy logistic có điều kiện (conditioned logistic regression). Đây là loại hồi quy logisitic thích hợp cho phân tích số liệu bắt cặp (matched data), có thể xem như là mở rộng của kiểm định cặp đôi χ^2 McNemar.

ĐO LƯỜNG BỆNH TẬT VÀ TỬ VONG

Giới thiệu

Ở chương này chúng ta quan tâm đến việc đo lường tử vong và bệnh tật và đặc biệt xem xét đến việc sử dụng tỉ suất (rate). Tỉ suất tử vong hay bệnh tật là số người chết hay mắc bệnh trong một khoảng thời gian nhất định chia cho tổng số người được quan sát để tìm biến cố đó. Trước tiên sẽ trình bày những tỉ suất thường dùng. Sau đó sẽ xem xét các vấn đề về tính toán và so sánh các tỉ suất trong các dân số nghiên cứu có cấu trúc tuổi giới tính khác nhau và mô tả các kĩ thuật thống kê để chuẩn hóa tỉ suất. Cuối cùng sẽ tổng kết ngắn gọn phương pháp phân tích

Tỉ suất sinh và chết

Tỉ suất sinh và chết của dân số được dựa trên các số thống kê thu thập được trong một năm lịch. Để tiện lợi các tỉ suất này thường được nhân cho 1000 để tránh các số lẻ và được trình bày chẳng hạn như số sinh cho mỗi 1000 thành viên dân số hàng năm. Các đo lường phổ biến là:

1. $\text{Tỉ suất sinh} = \dfrac{\text{Số sinh trong năm}}{\text{dân số giữa năm}} \times 1000$

2. $\text{Tỉ suất mẹo âeí} = \dfrac{\text{Số sinh sống trong năm}}{\text{dân số phuï nöõ tuoåi 15-44 åí giữa năm}} \times 1000$

3. $\text{Tỉ suất tæí vong thä} = \dfrac{\text{Số chết trong năm}}{\text{dân số giữa năm}} \times 1000$

4. $\dfrac{\text{Tỉ suất tæí vong}}{\text{âặc hiãu theo tuöíi (giåíi)}} = \dfrac{\text{Số chết trong năm trong mäüt nhoïm tuöíi (giåíi)}}{\text{dân số giữa năm của nhoïm tuöíi (giåíi) âoï}} \times 1000$

5. $\text{Tỉ suất tæí vong treí em} = \dfrac{\text{Số chết åí treí dæåïi 1 tuöíi xaíy ra trong năm}}{\text{số sinh sống trong năm}} \times 1000$

6. $\text{Tỉ suất tæí vong så sinh} = \dfrac{\text{Số chết åí treí dæåïi 28 ngaìy tuöíi xaíy ra trong năm}}{\text{số sinh sống trong năm}} \times 1000$

7. $\dfrac{\text{Tỉ suất tæí vong}}{\text{chu sinh}} = \dfrac{\text{Số chết tæí sanh} + \text{chết dæåïi 7 ngaìy tuöíi trong năm}}{\text{số sinh (sống} + \text{tæí sanh) trong năm}} \times 1000$

8. $\dfrac{\text{Tỉ suất tæí vong}}{\text{âặc hiãu nguyãn nhán}} = \dfrac{\text{Số chết do mäüt nguyãn nhán âặc biãût trong năm}}{\text{dân số giữa năm}} \times 1000$

9. $\dfrac{\text{Tỉ suất tæí vong}}{\text{tỉ lãû}} = \dfrac{\text{Số chết do mäüt nguyãn nhán âặc biãût trong năm}}{\text{täøng số chết trong năm}} \times 1000$

10. $\text{tỉ suất bãûnh - tæí vong} = \dfrac{\text{Số chết do mäüt nguyãn nhán âặc biãût}}{\text{täøng số của của bãûnh âoï}} \times 1000$

Lưu ý rằng mặc dù những số này được gọi là tỉ suất, trên thực tế chúng thuộc hai loại khác nhau. Những số được đánh dấu sao (*), số 1, 2, 3, 4, 8, là những số thực sự là tỉ suất theo nghĩa toán học đo lường các thuộc tính trong dân số thay đổi theo thời gian nhanh chậm thế nào. Do đó tỉ suất sinh đo lường kích thước dân số được tăng do số sinh mới nhanh như thế

nào và tỉ suất tử vong thô đo lường sự giảm dân số do tử vong nhanh như thế nào. Các số khá, được đánh dấu thập (†), số 5, 6, 7, 9, 10, là những số đúng hơn nên gọi là nguy cơ (risk) bởi vì chúng là xác suất xuất hiện một biến cố nhất định. Do đó, tỉ suất ác tính đo lường nguy cơ một người vì một bệnh nào đó khi mắc bệnh đó. Mặc dù sự phân biệt giữa nguy cơ và tỷ suất đã được mô tả rõ ràng bởi William Farr trong đầu thế kỉ19, chỉ gần đây nó mới được chú ý trở lại (Vandenbrouke 1985) và các thuật ngữ hiện đại vẫn còn lẫn lộn. Sự khác biệt sẽ được thảo luận thêm trong chương về số đo mới mắc bệnh.

Lưu ý rằng người ta dùng một số phương pháp tính toán gián tiếp để đo lường nguy cơ. Thí dụ, tỉ suất tử vong trẻ em là nguy cơ đứa trẻ chết trước kì sinh nhật đầu tiên của nó. Phương pháp trực tiếp để đo lường là theo dõi một đoàn hệ các trẻ mới sinh và ghi nhận tỉ lệ trẻ trong đoàn hệ chết trước khi kì sinh nhật thứ nhát của nó. Dù vậy, để tiện lợi, tỉ suất tử vong trẻ em được ước tính bằng tỷ số chết dưới một năm tuổi và số sinh sống trong một năm. Do đó tử số sẽ gồm một số chết trong những đứa trẻ được sinh trong năm trước và như thế cũng bỏ qua một số tử vong trong số trẻ đã sinh trong năm mà không được theo dõi tới năm sau. Trong một dân số ổn định người ta giả sử rằng hai yếu tố này sẽ triệt tiêu lẫn nhau. Một quan điểm tương tự sẽ được áp dụng để ước tính tỉ suất tử vong sơ sinh và tỉ suất tử vong chu sinh.

Đo lường tử vong trong một nghiên cứu

Định nghĩa tỉ suất tử vong được trình bày ở trên dựa trên số chết xảy ra trong một năm lịch chia cho dân số giữa năm. Dù vậy trong một nghiên cứu, không những tử vong được ghi nhận trong một giai đoạn bất kì (chứ không phải một năm) mà các cá nhân còn có thể được theo dõi trong những khoảng thời gian khác nhau do những lí do khác nhau. Những lí do này bao gồm sự chiêu mộ dần dần các đối tượng nghiên cứu để trải đều công việc theo thời gian, sự chiêu mộ các trẻ mới sinh hay các case bệnh mới khi chúng xuất hiện, sự chiêu mộ các người dân mới nhập cư vào khu vực và không theo dõi được do họ di cư khỏi khu vực, không còn muốn tham gia hay chết. Do đó, số tử vong quan sát được liên quan với số người **năm quan sát (person-years of observation), chứ không phải với dân số giữa** năm. Lưu ý rằng một người năm quan sát có thể là một người được quan sát trong trọn một năm hay 12 người được quan sát trong một tháng

$$\text{Tỉ suất tử vong thô} = \frac{\text{Số chết trong năm}}{\text{người - năm quan sát}} \times 1000$$

Tỉ suất tử vong thường được nhân với 1000 và kết quả được tính theo 100 người năm theo dõi. Khi báo cáo tỉ suất cần phân biệt rõ giữa môi 1000 người mỗi năm , có nghĩa là theo dõi trong cùng một khoảng thời gian và mỗi 1000 người-năm.

Xét một tình huống phổ biến khi tử vong được ghi nhận bằng cuộc điều tra cộng đồng hai lần và đếm số người chết trong khoảng giữa. Một người có mặt ở cả hai cuộc điều tra sẽ tham gia tổng số người năm quan sát. Một người chết hay di chuyển trong giai đoạn đó tham gia một số năm bằng số năm giữa lần khám thứ nhất và ngày chết hay ngày rời khỏi, trong khi đó một người sinh ra hay di chuyển vào cộng đồng giữa hai cuộc điều tra góp vào một số năm bằng số năm giữa ngày sinh hay ngày gia nhập cộng đồng và cuộc điều tra thứ hai. Nếu không thể xác định chính xác ngày chết hay ngày rời khỏi, người ta thường giả sử rằng sự kiện đó xảy ra ở giữa hai cuộc điều tra.

Đo lường tử vong

Có hai phương pháp chính để đo lường sự mắc bệnh của cộng đồng. Số mới mắc (incidence) là số các trường hợp bệnh mới trong một giai đoạn thời gian nhất định so với số người có nguy cơ (persons at risk) mắc bệnh đó; trong khi mắc toàn bộ (prevalence) dựa trên tổng số các trường hợp bệnh hiện có trong toàn dân số. Độ lớn của mắc toàn bộ phụ thuộc vào độ lớn

của mắc toàn bộ và khoảng thời gian mắc bệnh; đối với mới mắc cố định, mắc toàn bộ sẽ lớn hơn đối với bệnh kéo dài trong nhiều năm so với bệnh chỉ kéo dài trong vài ngày.

Số mắc toàn bộ

Số mắc toàn bộ có thể đo lường ở một thời điểm (mắc toàn bộ thời điểm - point prevalence) hay trong một khoảng thời gian (số mắc toàn bộ thời khoảng - period prevalence). Nó thường được tính theo phần trăm hay nếu nhỏ theo phần ngàn của dân số

$$\text{1. Tỉ suất hiện mắc (thời điểm)} = \frac{\text{Số người có bệnh tại một thời điểm}}{\text{toàn bộ dân số}}$$

$$\text{2. Tỉ suất hiện mắc (thời khoảng)} = \frac{\text{Số người có bệnh ấy một thời gian nào đó trong thời khoảng cho trước}}{\text{toàn bộ dân số ấy giữa thời khoảng}}$$

Số mắc toàn bộ thời điểm phổ biến hơn và hữu ích hơn và thường được gọi đơn giản là mắc toàn bộ

Số mới mắc

Có hai cách khác nhau để định nghĩa mới mắc; nó có thể được đo lường như một nguy cơ hay như một tỉ suất. Nguy cơ mới mắc (incidence risk) là xác suất một người lúc đầu không mắc bệnh sau đó có bệnh ở một thời gian nào đó trong khoảng thời gian quan sát. Nó thường được tính bằng phần trăm và nếu nhỏ bằng phần ngàn người. Mặt khác tỉ suất mới mắc (incidence rate) là tốc độ mắc bệnh trong số những người vẫn còn nguy cơ (khi một người mắc bệnh họ không còn nguy cơ). Số case mới không phụ thuộc vào số người nguy cơ lúc đầu mà với số người nguy cơ trung bình trong thời kì đó, nhân với chiều dài của thời kì, tương đương với số người năm nguy cơ (person-years at risk) trong thời kì. Đối với bệnh phổ biến, như là tiêu chảy, có thể tấn công nhiều lần, tỉ suất mới mắc đo lường số lần mắc cho mỗi người cho mỗi năm.

$$\text{1. Nguy cơ mới mắc} = \frac{\text{Số trường hợp bệnh mới trong một khoảng thời gian nhất định}}{\text{toàn bộ dân số}}$$

$$\text{2. Tỉ suất mới mắc} = \frac{\text{Số trường hợp bệnh mới trong một khoảng thời gian nhất định}}{\text{số người năm nguy cơ trong suất thời khoảng}}$$

$$(\text{nguy cơ trung bình} \times \text{khoảng thời gian})$$

Người năm nguy cơ được tính tương tự như cách tính người năm quan sát. Sự khác biệt duy nhất là đối với bệnh có thể xảy ra nhiều lần, một người tham gia vào tổng số thời gian không mắc bệnh trong nghiên cứu, chứ không phải chỉ thời gian từ lúc bắt đầu nghiên cứu tới lần mắc bệnh đầu tiên

Đối với bệnh hiếm như ung thư, hai mới mắc này giống nhau về cơ bản, bởi vì số người nguy cơ sẽ gần bằng với tổng dân số ở mọi lúc. Vì lí do này, chúng thường không phân biệt rõ ràng. Dù vậy, đối với bệnh phổ biến nguy cơ và tỉ suất mới mắc khác nhau.

Thí dụ, xét bệnh sởi trong năm tuổi thứ hai ở một nước mà ít có tiêm chủng sởi. Giả sử rằng theo dõi 1000 trẻ trước đây chưa bị nhiễm sởi từ sau sinh nhật lần thứ nhất của nó trong vòng một năm và có 140 trẻ bị sởi và 860 trẻ vẫn còn có nguy cơ ở cuối năm. Nguy cơ mới mắc là

xác suất trẻ không mắc bệnh sởi trong năm đầu tiên mắc bệnh trong năm thứ hai và bằng 140/1000 = 0,1400 hay 14,0%. Mặt khác tỉ suất nguy cơ là tỉ suất (tốc độ) mắc bệnh sởi trong năm thứ hai. Nếu chúng ta giả sử rằng số trường hợp mắc sởi trải đều trong năm, số nguy cơ trung bình sẽ bằng trung bình cộng ở lúc đầu năm và cuối năm bằng (1000 + 860)/2 = 930. Tỉ suất nguy cơ bằng 140/930 = 0,1505 hay 150,5 trên 1000 trẻ-năm nguy cơ (child-years at risk)

Nguy cơ tương đối

Đôi khi người ta tóm tắt sự so sánh hai mới mắc bằng tỉ số của chúng chứ không phải bằng hiệu số của chúng. Thí dụ, con người hút thuốc lá có nguy cơ mắc suyễn gấp 3 lần con những người không hút. Tỉ số này được gọi là nguy cơ tương đối (relative risk) và được nói rõ trong chương 24.

Tỉ suất chuẩn hóa

Bởi vì cả nguy cơ chết và nguy cơ mắc bệnh liên quan đến tuổi, và bởi vì chúng khác nhau tùy theo giới tính, tỉ suất tử vong thô và tỉ suất mới mắc và mắc toàn bộ chung phụ thuộc vào cấu trúc tuổi giới tính của dân số. Thí dụ, một dân số tương đối già sẽ có tỉ suất tử vong thô cao hơn dân số trẻ, ngay cả khi tử vong là như nhau. Do đó sẽ sai lầm nếu dùng tỉ suất chung khi so sánh hai dân số khác nhau trừ khi chúng có cùng một cấu trúc tuổi giới tính. Thay vì vậy cần dựa trên tỉ suất đặc hiệu tuổi giới tính và kiểm định chung sự khác biệt bằng kiểm định χ^2 Mantel-Haenszel áp dụng cho từng nhóm tuổi giới tính (xem Chương 14).

Dù vậy cũng cần có tỉ suất chung tổng kết đã điều chỉnh cho sự khác biệt tuổi giới tính. Có thể đạt được điều này nhờ việc sử dụng một dân số chuẩn (standard population) với phương pháp chuẩn hóa (standardization) trực tiếp hay gián tiếp. Trong phương pháp chuẩn hóa trực tiếp (direct method), tỉ suất đặc hiệu tuổi giới tính từ mỗi dân số nghiên cứu đươc áp dụng cho một dân số chuẩn để có tỉ suất **tử vong được điều chỉnh theo tuổi giới tính (age-sex adjusted mortality rate),** trong phương pháp gián tiếp (indirect method) tỉ suất đặc hiệu tuổi giới tính của dân số chuẩn được áp dụng cho mỗi dân số quan tâm để có tỉ số tử vong chuẩn hóa (standardized mortality ratio). Tỉ số này có thể dùng để tính tỉ suất điều chỉnh. So sánh được trình bày trong bảng 15.1 cho thấy hai phương pháp đòi hỏi những số liệu khác nhau.

Việc chọn lựa phương pháp phụ thuộc vào số liệu có sẵn và sự chính xác của chúng. Thí dụ, trong khi so sánh tỉ suất tử vong của các quốc gia khác nhau, phương pháp gián tiếp hay hơn bởi vì rất khó có được số liệu quốc gia về tỉ suất tử vong đặc hiệu tuổi giới tính. Nói chung, chuẩn hóa gián tiếp thường được dùng phổ biến cho tử vong là số mới mắc trong khi chuẩn hóa gián tiếp dùng cho mắc toàn bộ.

Bảng 15.1 So sánh phương pháp chuẩn hóa trực tiếp và gián tiếp

	Chuẩn hóa trực tiếp	Chuẩn hóa gián tiếp
Số liệu cần thiết		
Dân số nghiên cứu	tỉ suất đặc hiệu tuổi giới tính	cấu trúc dân số + tỉ suất chung
Dân số chuẩn	cấu trúc dân số	tỉ suất đặc hiệu tuổi giới tính + tỉ suất chung
Phương pháp	tỉ suất nghiên cứu đươc áp dụng vào dân số chuẩn	tỉ suất chuẩn được áp dụng vào dân số nghiên cứu
Kết quả	tỉ suất điều chỉnh tuổi giới tính	tỉ suất tử vong chuẩn hóa (+ tỉ suất điều chỉnh giới tính)

Cả hai phương pháp có thể được mở rộng để điều chỉnh các yếu tố khác ngoài tuổi và giới tính, như cấu trúc chủng tộc trong các nhóm nghiên cứu. Phương pháp trực tiếp có thể dùng

để tính trung bình chuẩn hóa (standardized mean) như huyết áp trung bình được điều chỉnh theo tuổi giới tính cho các nhóm nghề nghiệp khác nhau.

Chuẩn hóa trực tiếp

Bảng 15.2 trình bày tỉ suất mắc toàn bộ đặc hiệu và tỉ suất mắc toàn bộ chung thô của bệnh nhiễm onchocerca trong hai làng A và B. Lưu ý rằng tỉ suất mắc toàn bộ của onchocerca gia tăng theo tuổi và cao ở nam hơn ở nữ và cấu trúc tuổi giới tính của các mẫu từ 2 làng khác nhau. Tỉ suất mắc toàn bộ chung phải được điều chỉnh theo tuổi giới tính trước khi chúng được so sánh. Điều này có thể tiến hành bằng phương pháp chuẩn hóa trực tiếp.

Trước tiên, cần phải xác định dân số chuẩn. Thường theo một trong các cách sau: hoặc là một dân số nghiên cứu, hay tổng số hai dân số nghiên cứu (ở dưới sẽ làm theo cách này) và dân số điều tra từ địa phương hay quốc gia. Việc chọn lựa này có phần nào tùy tiện. Dù vậy cần lưu ý rằng mặc dù những lựa chọn khác nhau sẽ cho các tỉ suất tổng kết khác nhau nhưng điều này ít ảnh hưởng đến việc lí giải kết quả.

Bảng 15.2 Sử dụng chuẩn hóa trực tiếp để so sánh tỉ suất mắc toàn bộ của onchocerca trong hai làng A và B. Dân số chung của hai làng được dùng làm dân số chuẩn.

	Làng A			Làng B			Dân số chuẩn (làng A + B)		
Tuổi	Số được khám	số bị nhiễm	% nhiễm	Số được khám	số bị nhiễm	% nhiễm	Số được khám	Làng A	Làng B
Nam									
0-4	36	2	5,6	14	1	7,1	50	2,8	3,6
5-9	42	4	9,5	28	4	14,3	70	6,7	10,0
10-14	12	3	25,0	10	9	90,0	22	5,5	19,8
15-29	6	4	66,7	11	11	10,0	16	10,7	16,0
30-49	11	9	81,8	11	11	100,0	22	18,0	22,0
50+	21	17	81,0	8	8	100,0	29	23,5	29,0
Nữ									
0-4	35	1	2,9	17	1	5,9	52	1,5	3,1
5-9	52	4	7,7	15	2	13,3	67	5,2	8,9
10-14	15	3	20,0	9	4	44,4	24	4,8	10,7
15-29	24	14	58,3	12	11	91,7	36	21,0	33,0
30-49	25	19	76,0	17	17	100,0	42	31,9	42,0
50+	8	6	75,0	7	7	100,0	15	11,3	15,0
Tổng số	287	86	30,0	158	85	53,8	445	142,9	213,1
Tỉ suất mắc toàn bộ điều chỉnh tuổi giới tính								32,1	47,9

Thứ nhì, sẽ tính toán số người trong dân số chuẩn hóa bị nhiễm bệnh nếu tỉ suất mắc toàn bộ đặc hiệu tuổi giới tính sẽ giống như trong làng A. Thí dụ, xét phái nam tuổi từ 0-4 tuổi. Hai trong 36 người trong nhóm tuổi này bị nhiễm bệnh ở làng A vì vậy tỉ suất mắc toàn bộ là 5,6%. Có 50 phái nam thuộc lứa tuổi này trong dân số chuẩn và nếu tỉ suất mắc toàn bộ cũng là 5,6% thì có nghĩa là 5,6% × 50 = 2,8 người trong số đó bị nhiễm bệnh. Các số khác trong các nhóm tuổi giới tính khác được trình bày trong Bảng 15.2. Nói chung sẽ có 142,9 trong 445 người bị nhiễm onchocerca trong dân số chuẩn, tính ra tỉ suất mắc toàn bộ chung là 32,1% (142,9/445 × 100). Số này được gọi là tỉ suất mắc toàn bộ điều chỉnh tuổi giới tính của làng A.

$$\text{Tỉ suất mắc toàn bộ điều chỉnh theo tuổi-giới tính} = \frac{\text{Toàn bộ tỉ suất mắc trong dân số chuẩn} \times \text{tỉ suất mắc đặc hiệu tuổi giới tính trong dân số quan tâm}}{\text{tổng tỉ suất mắc đặc hiệu tuổi giới tính trong dân số chuẩn}}$$

Tỉ suất mắc toàn bộ điều chỉnh theo tuổi-giới tính cho làng B cũng được tính tương tự. Nó bằng 47,9% cao hơn đáng kể so với làng A. Kiểm định χ^2 Mantel-Haenszel so sánh tỉ suất mắc toàn bộ giữa hai làng này có ý nghĩa (χ^2=21,7; d.f.=1; P<0,001).

Chuẩn hóa gián tiếp

Bảng 15.3 trình bày kết quả của một cuộc nghiên cứu lớn trong một vùng lưu hành onchocerca, trong đó có 12816 người từ trên 30 tuổi được theo dõi trong 1 năm. Một điểm cần quan tâm là đánh giá xem có phải mù lòa, hậu quả của nhiễm onchocerca, dẫn đêm nguy cơ tử vong gia tăng hay không. Điều quan trọng là xem xét các tỉ suất tử vong tuổi giới tính riêng bởi vì tử vong gia tăng theo tuổi và khác nhu giữa giới tính mà còn bởi vì mắc toàn bộ của mù lòa gia tăng theo tuổi và cao ở nam hơn ở nữ. Dân số bị mù do đó dễ già hơn và có tỉ lệ nam cao hơn và như vậy tỉ suất chết thô sẽ cao hơn so với dân số không mù ngay cả khi tỉ suất đặc hiệu tuổi giới tính giống nhau. So sánh chung giữa người mù và không mù phải dùng phương pháp chuẩn hóa gián tiếp.

Giống như trong chuẩn hóa trực tiếp, bước đầu tiên là xác định dân số chuẩn. Việc lựa chọn cũng giống như trên nhưng có hạn chế là cần tỉ suất tử vong đặc hiệu tuổi giới tính cho dân số chuẩn và do đó dân số được chọn cho mục đích này phải đủ lớn để có ước lượng tin cậy của những tỉ suất này.

Những tỉ suất chuẩn sau đó được áp dụng cho dân số quan tâm để tính số tử vong kì vọng trong dân số nếu tử vong giống như trong dân số chuẩn. Thí dụ, người ta kì vọng tỉ lệ 19/2400 của 120 người mù tuổi từ 30 đến 39 sẽ bị chết nếu nguy cơ chết giống như những người đàn ông không mù cùng tuổi. Tính chung, có 22,55 tử vong kì vong trong dân số nười mù so với số quan sát là 69. Tk số giữa số tử vong quan sát và số tử vong kì vọng được gọi là tỉ số tử vong chuẩn hóa (standardize mortality ratio - SMR). Trong trường hợp này nó bằng 3,1 (69/22,55).

$$\text{Tỉ số tử vong chuẩn hóa (SMR)} = \frac{\text{số chết quan sát được}}{\text{Số chết kỳ vọng nếu tỉ suất đặc hiệu tuổi giới tính trong dân số chuẩn}}$$

SMR đánh giá một người trong dân số nghiên cứu so với một người trong dân số chuẩn dễ chết hơn (hay ít hơn) mấy lần. Giá trị bằng 1 có nghĩa là khả năng chết bằng nhau, giá trị lớn hơn một có nghĩa là dễ chết hơn, nhỏ hơn một nghĩa là khó chết hơn. SMR đôi khi được nhân cho 100 và tính bằng phần trăm.

Bởi vì dân số không bị mù được dùng làm chuẩn, số kì vọng và số quan sát của nó bằng nhau do đó SMR = 1. So sánh hai SMR cho thấy nhìn chung, người mù dễ bị chết gấp 3,1 lần so với người không bị mù trong khoảng thời gian một năm. Kiểm định χ^2 Mantel-Hanenzel có ý nghĩa cao (χ^2=55,7, độ tự do = 1, P <0,001).

Bảng 15.3 Sử dụng chuẩn hóa gián tiếp để so sánh tỉ suất tử von giữa người mù và người không mù, được thu thập trong một cuộc nghiên cứu một năm trong vùng lưu hành onchocerca. Tỉ suất tử vong của dân số không bị mù được dùng làm tỉ suất chuẩn

	dân số không mù			dân số mù				Số chết kì vọng trong dân số người mù nếu tỉ suất
Tuổi	Số theo dõi	Số chết	Chết/1000/năm	Số theo dõi	phần trăm bị mù	Số chết	Chết/1000/nàm	giống như trong dân số không bị mù
Nam								
30-39	2400	19	7,9	120	4,8	3	25,0	0,95
40-49	1590	21	13,2	171	9,7	7	40,9	2,26
50-59	1120	20	17,9	244	17,9	13	53,3	4,36
60+	610	20	32,8	237	28,0	24	101,3	7,77
Nữ								
30-39	3100	23	7,4	84	2,6	2	23,8	0,62
40-49	1610	22	13,7	69	4,1	3	43,5	0,94
50-59	930	16	17,2	198	15,3	8	47,6	2,89
60+	270	8	29,6	93	25,6	9	96,8	2,76
Tổng số	11630	149	12,8	1196	9,3	69	58,2	22,55
SMR			1,0				3,1(69/22,5)	
Tỉ suất tử vong điều chỉnh tuổi giới tính			12,8				39,7(3,1× 12,8)	

Tỉ suất tử vong được điều chỉnh theo tuổi giới tính có thể tính được bằng cách nhân SMR cho tỉ suất tử vong thô của dân số chuẩn. Tính ra được tỉ suất tử vong điều chỉnh tuổi giới tính là 12,8 và 39,7/1000/năm cho dân số không bị mù và dân số bị mù

$$\text{Tỉ suất tử vong điều chỉnh tuổi giới tính} = SMR \times \text{tỉ suất tử vong thô dân số chuẩn.}$$

Phân tích tỉ suất

Phương pháp được mô tả ở Chương 12, 13 và 14 liên quan đến tỉ lệ được áp dụng cho phân tích tỉ suất mắc toàn bộ, tử vong và mới mắc nguy cơ bởi vì chúng là các tỉ lệ. Những phương pháp này có thể được áp dụng cho phân tích những tử vong và mới mắc nguy cơ, khi số trường hợp tử vong (hay trường hợp bệnh mới) nhỏ so với số người nguy cơ trung bình, bởi vì trong trường hợp này tỉ suất thường được xem như là tỉ lệ. Một cách khác tỉ suất tử vong và mới mắc có thể được phân tích nhờ phân phối Poisson, được mô tả ở Chương 17 và mô hình tuyến tính log được nhắc đến ở Chương 14, với điều kiện mỗi trường hợp xảy ra độc lập với nhau và ngẫu nhiên về thời gian.