California Polytechnic State University, San Luis Obispo

JAY L. DEVORE

Tài liệu môn học

Probability and Statistics for Engineering and Sciences

XÁC SUẤT THỐNG KÊ ỨNG DỤNG

Biên soạn:
Nguyễn Hồng Nhung - Chương 1, 6, 12.
Hoàng Thị Minh Thảo - Chương 7, 8.
Lê Thị Mai Trang - Chương 5,9.
Nguyễn Ngọc Tứ - Chương 2, 3, 4.

Bộ môn Toán - ĐH SPKT, Tp. Hồ Chí Minh - Năm 2018

Mục lục

MỞ ĐẦU

Mục đích

Việc sử dụng các mô hình xác suất và phương pháp thống kê để phân tích dữ liệu đã trở thành phổ biến trong hầu hết các ngành khoa học. Cuốn sách này cố gắng để cung cấp một giới thiệu một cách toàn diện những mô hình và phương pháp có thể gặp nhất và được sử dụng bởi các sinh viên trong công việc của mình trong kỹ thuật và khoa học tự nhiên. Mặc dù các ví dụ và bài tập đã được thiết kế với các nhà khoa học và kỹ sư, nhưng hầu hết các phương pháp là cơ bản để phân tích thống kê trong nhiều lĩnh vực khác vì vậy cuốn sách cũng sẽ giúp ích cho sinh viên kinh tế và các ngành khoa học xã hội.

Cách tiếp cận

Người học trong một khóa học thống kê, được thiết kế cho các chuyên ngành khác, có thể là bước đầu hoài nghi về giá trị và sự liên quan của đối tượng, nhưng kinh nghiệm của tôi là học sinh có thể nắm bắt được thống kê qua việc sử dụng các ví dụ tốt và các bài tập pha trộn kinh nghiệm hàng ngày của họ với lợi ích khoa học của mình. Do vậy, trong cuốn sách này tôi sử dụng các ví dụ thực, chứ không phải là ví dụ, dữ liệu nhân tạo, dữ liệu. Nhiều phương pháp trình bày, đặc biệt là trong các chương sau này suy luận thống kê, được minh họa bằng cách phân tích dữ liệu lấy từ nguồn xuất bản và rất nhiều các bài tập cũng liên quan đến làm việc với các dữ liệu đó. Dôi khi người đọc có thể chưa quen với bối cảnh của một vấn đề cụ thể, nhưng tôi đã tìm thấy rằng các sinh viên đang thu hút nhiều hơn bởi các vấn đề thực sự với một bối cảnh hơi lạ hơn bởi vấn đề nhân tạo trong một khung cảnh quen thuộc.

Mức độ Toán học

Sự giải thích sử dụng các kiến thúc toán học khá khiêm tốn. Kiến thức về giải tích sử dụng đáng kể trong chương 4 và một phần của chương 5 và 6. Ma trận đại số không được sử dụng ở tất cả các chương. Do đó người đọc cần có nền toán học bao

gồm đạo hàm và tích phân.

Nội dung

Chương 1 bắt đầu với một số khái niệm cơ bản và các thuật ngữ-tổng thể, mẫu, thống kê mô tả và suy luận. Một sự phát triển của xác suất truyền thống được đưa ra trong chương 2, tiếp theo là phân bố xác suất của biến ngẫu nhiên rời rạc và liên tục trong chương 3 và 4, tương ứng. Phân phối đồng thời và các tính chất sẽ được thảo luận trong phần đầu tiên của chương 5. Các phần sau của chương này giới thiệu các thống kê và phân phối mẫu, tạo thành cầu nối giữa xác suất và suy luận thống kê. Ba chương tiếp theo bao gồm ước lượng điểm, khoảng tin cậy thống kê, và kiểm định giả thuyết dựa trên một mẫu duy nhất. Các phương pháp suy luận liên quan đến hai mẫu độc lập và dữ liệu kết hợp được thể hiện trong chương 9. Các phân tích phương sai là chủ đề của các chương 10 và 11 (một nhân tố và đa nhân tố). Chương 12 trình bày về mô hình hồi quy (các mô hình hồi quy tuyến tính đơn giản và tương quan) và chương 13 sẽ quay lại trình bày các mô hình hồi quy một cách sâu rộng hơn. Ba chương cuối cùng phát triển các phương pháp Khi bình phương, quá trình phân phối phi tham số và các kỹ thuật kiểm tra chất lượng thống kê.

Hỗ trợ người học

Mặc dù mức độ toán học của cuốn sách sẽ phù hợp cho hầu hết các sinh viên khoa học và kỹ thuật với mức độ khó một chút, tuy nhiên để có một sự hiểu biết về các khái niệm và nắm được các phương pháp đôi khi có thể đòi hỏi những nỗ lực đáng kể. Để giúp người học đạt được như vậy, tôi đã cung cấp nhiều bài tập khác nhau có liên quan đến ứng dụng thường xuyên của tài liệu đến các tình huống hơi mới. Có rất nhiều bài tập nhiều hơn hầu hết các giảng viên sẽ muốn gán trong bất kỳ khóa học đặc biệt, nhưng tôi khuyên các sinh viên được yêu cầu làm việc một số lượng đáng kể của họ; trong một kỷ luật giải quyết vấn đề, sự tham gia tích cực của loại này là cách chắc chắn nhất để xác định và thu hẹp khoảng cách trong sự hiểu biết rằng chắc chắn phát sinh. Câu trả lời cho bài tập số lẻ xuất hiện trong phần câu trả lời ở phía sau của văn bản. Ngoài ra, một tay giải pháp cho sinh viên, bao gồm các giải pháp làm việc hết công suất để hầu như tất cả các bài tập số lẻ, có sẵn. Để truy cập vào tài liệu và các nguồn tài nguyên đồng hành bổ sung, vui lòng truy cập www.cengagebrain.com. Tại trang chủ CengageBrain.com, tìm kiếm ISBN

của tiêu đề của bạn (từ trang bìa sau của cuốn sách của bạn) bằng cách sử dụng hộp tìm kiếm ở trên cùng của trang. Điều này sẽ đưa bạn đến trang sản phẩm mà các nguồn miễn phí có thể được tìm thấy.

Jay Devore

Chương 1

TỔNG QUAN VÀ THỐNG KÊ MÔ TẢ

Giới thiệu

Thống kê học là một ngành khoa học nghiên cứu các quy luật của đám đông, tổng thể. Các phương pháp thống kê là tập hợp các lập luận, tư duy hợp lý nhằm tìm hiểu và đưa ra các quy luật của đám đông, tổng thể. Thống kê không là một ngành của Toán học nhưng Thống kê sử dụng các công cụ của Toán học để chứng minh các lập luận, phương pháp Thống kê là hợp lý.

Thống kê bao gồm Thống kê mô tả và suy luận Thống kê.

Thống kê mô tả được sử dụng để trình bày một cách có hệ thống dữ liệu thu thập được từ quá nghiên cứu thực nghiệm thông qua các phương pháp lấy mẫu khác nhau. Trong Thống kê mô tả ngoài hệ thống bảng số liệu các đại lượng đặc trưng cho số liệu như trung vị, kỳ vọng, phương sai,...; cũng như các biểu đồ được sử dụng nhằm giúp người đọc nắm được thông tin ban đầu về số liệu nhanh nhất.

Sau khi đã thu được một mẫu từ một tổng thể, một điều tra viên sẽ thường xuyên muốn sử dụng thông tin mẫu để vẽ một số loại kết luận (làm suy luận của một số loại) về tổng thể. Kỹ thuật cho việc khái quát hóa tổng thể từ một mẫu gọi là **thống kê suy luận**. Suy luận Thống kê là quá trình phân tích số liệu bằng các mô hình toán học nhằm rút ra các kết luận về mục tiêu nghiên cứu với mức độ tin cậy nào đó. Ví dụ như một chủ ao cá muốn biết trong ao cá có số lượn cá là bao

nhiêu mà không cần rút hết nước trong ao, khi đó chúng ta có thể sử dụng khoảng ước lượng trong thống kê để đưa ra khoảng ước lượng về số lượng cá trong ao với độ tin cậy cho trước. Hay để dự đoán mức tiêu thụ điện trung bình của người dân ở vùng A khi biết nhiệt độ tại vùng đó trong khoảng thời gian nhất định ta có thể sử dụng mô hình tương quan hồi quy để trả lời câu hỏi này.

Thống kê được sử dụng trong nghiên cứu của nhiều ngành khoa học khác nhau như Kỹ thuật, Kinh tế, Y tế, Sinh học, Xã hội,... Với sự phổ biến của máy tính các phương pháp thống kê hiện đại mặc dù có khối lượng tính toán lớn đã trở lên khả thi và được các chuyên gia phân tích thống kê áp dụng rộng rãi cho nhiều lĩnh vực nghiên cứu.

1.1 Tổng thể, mẫu và qui trình

Hàng ngày chúng ta thường xuyên tiếp xúc với dữ liệu, sự kiện liên quan đến lĩnh vực chuyên môn cũng như các hoạt động đời sống thường ngày. Thống kê cung cấp phương pháp để tổ chức và tổng kết dữ liệu và rút ra kết luận dựa trên các thông tin chứa trong dữ liệu. Một nghiên cứu Thống kê sẽ tập trung vào các đối tượng chứa các đặc tính, thống tin mà nghiên cứu đang hướng tới. Tập hợp các đối tượng này cấu thành một tổng thể xác định. Ví dụ như muốn nghiên cứu chất lượng mũ bảo hiểm của công ty Y, thì tổng thể nghiên cứu sẽ là toàn bộ mũ bảo hiểm do công ty Y sản xuất ra. Điều tra khác có liên quan đến tổng thể bao gồm tất cả các cá nhân nhận bằng kỹ sư trong năm học gần đây nhất.

Khi các thông tin mong muốn có sẵn cho tất cả các đối tượng trong tổng thể, chúng ta có thể điều tra toàn bộ tổng thể. Tuy nhiên những hạn chế về thời gian, tiền bạc và các nguồn lực khan hiếm khác thường làm cho một điều tra tổng thể không thực tế hoặc không khả thi. Ví dụ như nếu ta mang toàn mang toàn bộ mũ bảo hiểm của công ty Y sản xuất ra kiểm tra chất lượng bằng cách tác động lực vào mũ. Khi đó toàn bộ mũ của công ty Y sản xuất ra sau kiểm định sẽ không sử dụng được nữa. Vì vậy trong nghiên cứu Thống kê thay vì nghiên cứu toàn bộ tổng thể, một tập hợp con của tổng thể, gọi là mẫu, được chọn theo một số cách thức nhất định sẽ được dùng để nghiên cứu.

Như vậy, chúng ta có thể có được một mẫu các mũ bảo hiểm do công ty Y sản xuất để kiểm tra xem chất lượng mũ có đảm bảo an toàn cho người sử dụng không, hoặc

chúng ta có thể chọn một mẫu các sinh viên tốt nghiệp trường kỹ thuật năm ngoái để có được thông tin phản hồi về chất lượng của các chương trình đào tạo kỹ thuật.

Chúng ta thường chỉ quan tâm đến một số đặc điểm của các đối tượng trong một tổng thể: độ dày của mỗi bức tường, đường kính của các trục máy, tuổi thọ của thiết bị, giới tính của một kỹ sư tốt nghiệp, độ tuổi mà các cá nhân đã tốt nghiệp, ... Các đặc tính này có thể được phân loại, chẳng hạn như giới tính hoặc loại sự cố,... Hoặc thông tin của các đặc tính này biểu diễn bởi các số ví dụ như tuổi = 23 hoặc chiều dài = 0.502cm.
Một biến là bất cứ đặc điểm mà giá trị có thể thay đổi từ một đối tượng khác trong tổng thể. Ví dụ như
x = điểm thi đại học môn Toán của máy tính của một sinh viên.
y = số lần truy cập vào một trang web cụ thể trong một khoảng thời gian quy định.
z = khoảng cách của một ô tô với các phương tiện khác để đảm bảo an toàn theo các điều kiện quy định về phanh xe.
Dữ liệu là kết quả từ việc quan sát hoặc một biến duy nhất hoặc đồng thời trên hai hay nhiều biến. Một tập hợp dữ liệu đơn biến bao gồm các quan sát trên một biến duy nhất. Ví dụ, chúng ta khảo sát thương hiệu điện thoại của một sinh viên đang sử dụng: Samsung (S); Oppo (O); Iphone (I); Nokia (N); Acesus (A); Biphone (B) hoặc loại khác (K), trên một số sinh viên của trường Dại học X, kết quả được tập dữ liệu

A O S S I B K S I S O S I N I B N O S K

Các mẫu sau đây của tuổi thọ (giờ) của pin thương hiệu D đưa vào khai thác nhất định là một tập hợp dữ liệu đơn biến số:

5.5 5.4 6.2 6.1 5.8 6.5 5.8 5.5

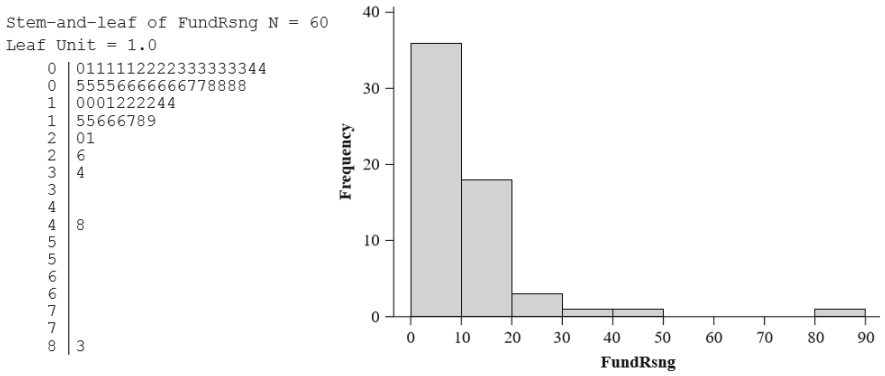
Chúng ta có dữ liệu hai biến khi mỗi đối tượng ta quan sát hai đặc tính. Ví dụ như bộ dữ liệu bao gồm một cặp (cân nặng (đơn vị: kg); chiều cao (đơn vị: cm)) cho mỗi cầu thủ bóng rổ trên một nhóm, với quan sát đầu tiên là (72, 168), thứ hai là (75, 212) và ... Dữ liệu đa biến phát sinh khi với mỗi đối tượng ta quan sát nhiều hơn n=một đặc tính. Ví dụ như, một bác sĩ có thể xác định huyết áp tâm thu, huyết áp tâm trương và mức cholesterol trong huyết thanh của từng bệnh nhân tham gia nghiên cứu. Mỗi quan sát sẽ là một bộ ba con số, chẳng hạn như (120, 80, 146).
Trong nhiều bộ dữ liệu đa biến, một số biến là số và những biến khác là biến. Chẳng hạn như, đánh giá các loại ô tô dựa trên các biến như: loại xe (nhỏ, thể thao, nhỏ

gọn, kích cỡ trung bình, lớn), tiết kiệm nhiên liệu chạy trong thành phố (mpg), hiệu quả nhiên liệu chạy trên đường cao tốc (mpg), loại hệ thống truyền lực (bánh sau , bánh xe phía trước, bốn bánh xe), và ...

Một điều tra viên đã thu thập dữ liệu có thể chỉ đơn giản muốn tóm tắt và mô tả các tính năng quan trọng của dữ liệu. Điều này đòi hỏi phải sử dụng các phương pháp thống kê mô tả. Một trong số những phương pháp này là vẽ đồ thị; xây dựng các biểu đồ, boxplots, và các điểm phân tán.

Ví dụ 1.1 Từ thiện là một loại hình kinh doanh tại Hoa Kỳ. Các trang web charity-navigator.com cung cấp thông tin về khoảng 5500 tổ chức từ thiện. Một số tổ chức từ thiện hoạt động rất hiệu quả, với chi phí gây quỹ và chi phí quản lý hành chính chỉ chiếm một tỷ lệ nhỏ trong tổng chi phí, trong khi những tổ chức khác dành một tỷ lệ phần trăm cao của quỹ cho các hoạt động như vậy. Dưới đây là số liệu về chi phí huy động vốn là một tỷ lệ phần trăm của tổng chi phí cho một mẫu ngẫu nhiên gồm 60 tổ chức từ thiện:

6.1	12.6	34.7	1.6	18.8	2.2	3	2.2	5.6	3.8
2.2	3.1	1.3	1.1	14.1	4	21	6.1	1.3	20.4
7.5	3.9	10.1	8.1	19.5	5.2	12	15.8	10.4	5.2
6.4	10.8	83.1	3.6	6.2	6.3	16.3	12.7	1.3	0.8
8.8	5.1	3.7	26.3	6	48	8.2	11.7	7.2	3.9
15.3	16.6	8.8	12	4.7	14.7	6.4	17	2.5	16.2



Hình 1.1: Một Minitab cuống và lá (phần mười chữ số cắt ngắn) và biểu đồ cho các dữ liệu phần trăm từ thiện gây quỹ.

Hình 1.1 mô tả cách các tỷ lệ phần trăm được phân bố trên các vùng giá trị từ 0

đến 100. Rõ ràng là một phần đáng kể của các tổ chức từ thiện trong mẫu chi tiêu ít hơn 20% vào việc gây quỹ và chỉ có một vài tỷ lệ có thể được xem như là vượt ra ngoài giới hạn của chi phí hợp lý.

Ví dụ 1.2

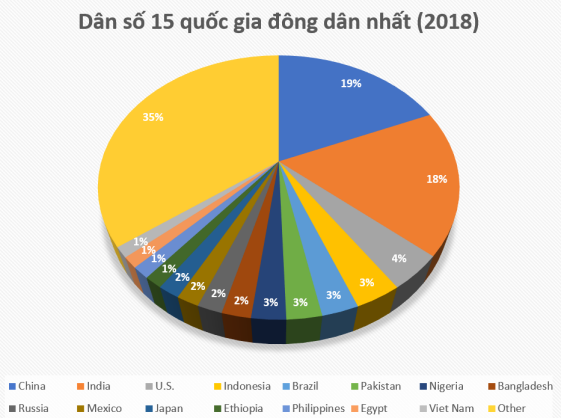
Theo thống kê của Liên hiệp quốc, dân số Thế giới đến ngày 22 tháng 3 năm 2017 là 7,49 tỷ người. Ước tính dân số Thế giới dự kiến đạt 8 tỷ người vào năm 2023 và đạt 10 tỷ người vào năm 2056. Theo cục điều tra dân số Hoa Kỳ ta có số liệu về dân số của 15 quốc gia đông dân nhất thế giới. Số liệu được mô tả dưới dạng bảng sau:

Bảng 1.1: Dân số của 15 quốc gia đông dân nhất Thế giới.

Quốc Gia	**Số dân**	**Chiếm tỷ lệ**
Trung Quốc	1415045928	18.54%
Ấn Độ	1354051854	17.74%
Hoa Kỳ	326766748	4.28%
Indonesia	266794980	3.50%
Brazil	210867954	2.76%
Pakistan	200813818	2.63%
Nigeria	195875237	2.57%
Banglades	166368149	2.18%
Nga	143964709	1.89%
Mexico	130759074	1.71%
Nhật Bản	127185332	1.67%
Ethiopia	107534882	1.41%
Philippin	106512074	1.40%
Hy Lạp	99375741	1.30%
Việt Nam	96491146	1.26%
Các quốc gia khác	2684411699	35.16%

Bảng số liệu cung cấp cho chúng ta thông tin về dân số của 15 quốc gia đông dân nhất. Bảng số liệu cung cấp các con số chính xác về dân số của từng nước trong nhóm các quốc gia đông dân, tuy nhiên biểu đồ hình quạt dưới đây giúp người đọc dễ hình dung về thông tin được cung cấp hơn.

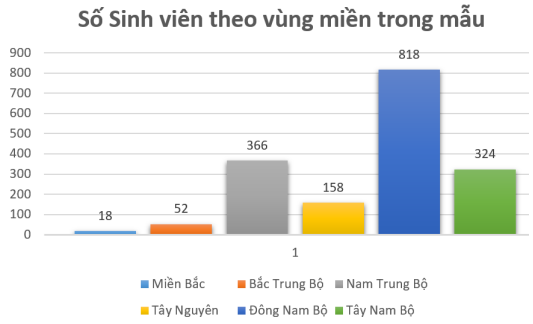
Nhìn vào biểu đồ ta thấy Trung Quốc và Ấn Độ là hai quốc gia đông dân nhất Thế Giới. Tiếp theo đó theo chiều kim đồng hồ là các quốc gia khác có số dân đông tiếp theo. Cuối cùng là phần biểu diễn số dân của 218 quốc gia còn chiếm khoảng 35% dân số Thế Giới.



Hình 1.2: Biểu đồ hình quạt biểu diễn dân số của các quốc gia đông dân nhất Thế giới.

Việt Nam với 96.621.188 dân, dựa trên ước lượng của Cục Điều tra dân số Hoa Kỳ ngày 22 tháng 8 năm 2018, chiếm 1,26% dân số Thế Giới và đứng thứ 15 trong số các nước đông dân nhất. Tổng diện đất của Việt Nam là 310.070 km^2, mật độ dân của Việt Nam là 311 người trên km^2. Cũng theo Cục Điều tra dân số Hoa Kỳ độ tuổi trung bình của người Việt Nam là 30,9.

Ví dụ 1.3 Trong một cuộc khảo sát các sinh viên đang học các môn Khoa học Cơ bản: Toán, Lý, Hóa trong học kì 2 năm học 2015-2016 tại trường Đại học Sư phạm Kỹ thuật Thành phố Hồ Chí Minh; điều tra viên có thu thông tin về nơi cư trú của sinh viên. Biểu đồ cột sau mô tả dữ liệu về sinh viên trong mẫu đến từ các vùng miền trong nước.

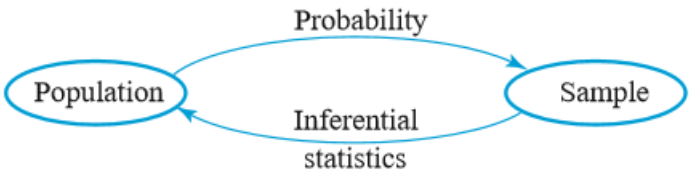


Hình 1.3: Biểu đồ cột biểu diễn số sinh viên mẫu khóa 2015 trường Đại học Sư phạm Kỹ thuật Tp.HCM theo vùng miền.

Nhìn biểu đồ ta có thể thấy số sinh viên trong mẫu đến từ Đông nam bộ là nhiều nhất và đến từ miền Bắc là ít nhất. Số lượng sinh viên đến từ Nam trung bộ và Tây nam bộ khá tương đồng với số lượng đứng sau Đông nam bộ. Sinh viên viên đến từ Bắc trung bộ cũng không nhiều chỉ hơn số lượng sinh viên đến từ miền Bắc một chút.

So với các bảng số liệu các biểu đồ giúp người xem dễ ghi nhận thông tin hơn là các bảng đầy các số và chữ.

Trong bài toán xác suất, tính chất của tổng thể mà ta đang nghiên cứu được giả định là đã biết (ví dụ số lượng của tổng thể, phân phối của một số tham số của tổng thể có thể được giả định), khi đó các câu hỏi liên quan đến một mẫu lấy từ tổng thể được đặt ra và trả lời. Trong bài toán thống kê, đặc trưng tính chất nghiên cứu có trong mẫu và các thông tin này được sử dụng để rút ra kết luận về tổng thể. Mối quan hệ giữa xác suất và thống kê: theo các xác suất từ tổng thể ta có các suy luận đối với các mẫu (suy luận suy diễn), trong khi theo thống kê từ mẫu suy ra các kết luận về tổng thể (suy luận quy nạp). Mối quan hệ này được minh họa trong hình 1.4.



Hình 1.4: Mối quan hệ giữa xác suất và thống kê suy luận.

Ví dụ 1.4 Ví dụ như điều tra việc người lái xe sử dụng thắt lưng buộc qua vạt áo được trong bị trong xe với hệ thống đai vai tự động. (Bài báo "Thắt lưng ghế ô tô: Sử dụng hệ thống thắt lưng tự động," tạp chí *Human Factors*, năm 1998, trang 126-135, tóm tắt dữ liệu sử dụng). Trong bài toán xác suất, chúng ta có thể giả định rằng 50% tất cả các lái xe trong một khu vực đô thị nhất định được trang bị thắt lưng thường xuyên sử dụng đai lưng của họ (một giả định về tổng thể). Vì vậy chúng ta có thể hỏi, "Làm thế nào có khả năng là có một mẫu của 100 lái xe y sẽ bao gồm ít nhất 70 người thường xuyên sử dụng đai lưng của họ?" hoặc "có bao nhiêu lái xe trong một mẫu kích thước 100 chúng ta có thể mong đợi là thường xuyên sử dụng đai lưng của họ?". Mặt khác, trong bài toán thống kê suy luận, chúng ta có thông

tin mẫu có sẵn. Ví dụ, một mẫu của 100 lái xe của chiếc xe đó tiết lộ rằng 65 thường xuyên sử dụng đai lưng của họ. Sau đó chúng ta có thể hỏi, "Có phải điều này cung cấp bằng chứng đáng kể để kết luận rằng hơn 50% của tất cả các trình điều khiển như vậy trong khu vực này thường xuyên sử dụng đai lưng của họ?" Trong trường hợp thứ hai này, chúng ta đang cố gắng sử dụng thông tin của mẫu để trả lời một câu hỏi về cấu trúc của toàn bộ tổng thể mà từ đó mẫu được chọn.

Trước khi chúng ta có thể hiểu những gì một mẫu cụ thể cho chúng ta biết về tổng thể, đầu tiên chúng ta nên hiểu sự không chắc chắn liên quan đến việc lấy mẫu từ một tổng thể số nhất định. Đây là lý do tại sao chúng ta nghiên cứu Xác suất trước Thống kê.

Thu thập dữ liệu

Thống kê không chỉ bao gồm tổ chức và phân tích dữ liệu mà cồn bao gồm quá trình thu thập dữ liệu. Nếu dữ liệu không được thu thập đúng cách, một điều tra viên có thể không có khả năng trả lời các câu hỏi được xem xét với một mức độ tin cậy hợp lý. Một vấn đề phổ biến là các phần tử của tổng thể có thể khác nhau trong quá trình lấy mẫu ra từ tổng thể. Ví dụ, các nhà quảng cáo muốn phân loại thông tin khác nhau về những thói quen xem truyền hình của các khách hàng tiềm năng. Các thông tin này thu được từ việc đặt thiết bị giám sát trong một số ít các gia đình trên khắp Hoa Kỳ. Nó cũng được phỏng đoán là bản thân vị trí của các thiết bị giám sát làm thay đổi hành vi của người xem, do đó đặc điểm của mẫu có thể khác nhau từ những phần tử khác nhau của tổng thể.

Ví dụ 1.5 Tạp chí *New York Times* (ngày 27 tháng 1 năm 1987) đưa tin aspirin có thể giảm nguy cơ đau tim. Kết luận này được dựa trên một thí nghiệm được thiết kế gồm hai nhóm, một nhóm đối chứng sử dụng giả dược có sự xuất hiện của aspirin nhưng được biết đến là trơ và một nhóm điều trị mà dùng aspirin theo một chế độ quy định. Các đối tượng được phân ngẫu nhiên vào các nhóm để tránh bất kỳ thành kiến nào và do đó có thể sử dụng phương pháp xác suất để phân tích dữ liệu.
Trong số 11,034 cá nhân trong nhóm đối chứng, 189 người sau đó trải qua các cơn đau tim, trong khi chỉ có 104 người trong 11,037 người thuộc nhóm dùng aspirin đã có một cơn đau tim. Tỷ lệ mắc bệnh đau tim ở nhóm điều trị chỉ khoảng một nửa

trong nhóm đối chứng. Một trong các lời giải thích có thể đưa ra cho kết quả này là sự thay đổi là tình cờ, tức là dùng aspirin thực sự không có hiệu quả như mong muốn và các quan sát khác biệt có cùng sự thay đổi như tung hai đồng xu giống hệt nhau sẽ thường tạo ra các số khác nhau về mặt có hình. Tuy nhiên, trong trường hợp này, phương pháp suy luận cho rằng sự thay đổi ngẫu nhiên tự nó không thể giải thích đầy đủ cho sự khác biệt trong một số lượng lớn các quan sát được.

Thống kê là một ngành khoa học dễ bị lợi dụng bởi các cá nhân hay tổ chức muốn lợi dụng Thống kê để bóp méo sự thật bằng nhiều cách như bịa các con số không có thật, lựa chọn các số liệu theo chủ quan của người nghiên cứu. Ví dụ như để chứng minh một loại sữa A, là sản phẩm của công ty X, giúp tăng tầm vóc của trẻ em. Trước khi cho trẻ uống sữa A, công ty X lấy mẫu là các trẻ em ở vùng nông thôn hay chọn các có thể trạng thấp bé. Sau khi cho trẻ uống sữa A, công ty X lại cố tình lấy mẫu là các trẻ em ở vùng thành thị hay có tầm vóc cao lớn. Như vậy hai mẫu này không phù hợp để đưa ra kết luận về hiệu quả của việc uống sữa A đối với tầm vóc của trẻ em. Vì vậy để các kết luận Thống kê là đáng tin cậy thì ngoài việc sử dụng các phương pháp Thống kê phù hợp với số liệu thu được thì việc lấy mẫu thực nghiệm cũng hết sức quan trọng.

Có rất nhiều phần mềm Thống kê được sử dụng như: SPPS, SAS, S-Plus, Stata, Eview, R, Minitab, Mathematical, Matlab, Excel,... SPSS là phần mềm Thống kê được sử dụng khá phổ biến ở việt Nam với khá là nhiều sách hướng dẫn sử dụng và nhiều trung tâm đào tạo ngắn hạn chọn dạy phần mềm này cho học viên. R có thể được tải về mà không mất phí từ trang http://www.r-project.org. Bạn cũng có thể sử dụng Excel một phần mềm văn phòng có sẵn để phân tích Thống kê vì sử dụng Excel để vẽ các biểu đồ rất đơn giản và đẹp, cũng như Excel đủ mạnh để giải quyết các vấn đề Thống kê thường gặp. Tuy nhiên khi cần phân tích các Thống kê chuyên sâu bạn cần sử dụng các phần mềm Thống kê chuyên dụng.

Bài tập 1.1

1.1.1 Đưa ra một mẫu có cỡ 5 từ mỗi tổng thể sau:

a. Tất cả các tờ báo hàng này xuất bản tại Việt Nam.

b. Các chương trình truyền hình phát sóng trên các kênh Tiếng Việt của đài truyền hình Việt Nam.

c. Tổng thể các ngành học tại trường đại học Sư Phạm Kỹ Thuật Tp. HCM.
d. Tổng thể các sinh viên trường đại học Sư Phạm Kỹ Thuật Tp. HCM.
e. Tổng thể các điểm trung bình môn Toán của sinh viên trường đại học Sư Phạm Kỹ Thuật Tp. HCM.

1.1.2 Cho 3 ví dụ về 3 tổng thể cụ thể và 3 ví dụ về giả thuyết thống kê tương ứng. Với mỗi tổng thể và giả thuyết thống kê cụ thể đặt một câu hỏi xác suất và một một câu hỏi suy luận thống kê.

1.2 Phương pháp trực quan và biểu đồ trong Thống kê mô tả

1.2.1 Ký hiệu

Số lượng quan sát trong một mẫu đơn gọi kích thước mẫu, hay cỡ mẫu, được ký hiệu là n. Ví dụ như $n = 4$ là cỡ mẫu gồm các trường đại học Đại học Sư phạm Kỹ thuật Tp. HCM, Đại học Kinh tế Luật, Đại học Tôn Đức Thắng, Đại học Ngân hàng và cũng là cỡ của mẫu các phép đo pH 6.3 , 6.2, 5.9, 6.5. Nếu xét đồng thời hai mẫu, ta ký hiệu m và n hay n_1 và n_2 biểu thị số lượng các quan sát của hai mẫu. Giả sử như đo hiệu suất nhiệt của hai loại động cơ diesel khác nhau đươc hai bộ dữ liệu 29.7, 31.6, 30.9 và 28.7, 29.5, 29.4, 30.3, thì ta có 2 cỡ mẫu $m = 3$ và $n = 4$.

Cho một tập dữ liệu gồm n quan sát trên một số biến x, những quan sát riêng lẻ sẽ được ký hiệu là $x_1, x_2, ..., x_n$. Chỉ số dưới không liên quan đến độ lớn của một quan sát cụ thể. Do đó x_1 nói chung không là quan sát nhỏ nhất trong bộ dữ liệu này, cũng như x_n thường không là giá trị lớn nhất. Trong nhiều ứng dụng, x_1 sẽ được quan sát đầu tiên được thu thập bởi các thí nghiệm, x_2 lần thứ hai, và tiếp tục như vậy. Các quan sát thứ i trong tập dữ liệu sẽ được ký hiệu bằng x_i.

1.2.2 Biểu đồ Gốc-Lá

Xét một tập hợp các dữ liệu số $x_1, x_2, ..., x_n$ mà mỗi x_i bao gồm ít nhất hai chữ số. Một cách nhanh chóng để có được một biểu diễn trực quan thông tin của tập dữ liệu là xây dựng một biểu đồ gốc và lá.

Hướng dẫn tạo biểu đồ gốc và lá
1. Chọn một hoặc nhiều chữ số đầu cho các giá trị gốc, các chữ số sau là lá.
2. Liệt kê các giá trị gốc có thể có trong một cột dọc.
3. Ghi lá cho mỗi quan sát bên cạnh giá trị gốc tương ứng.
4. Chỉ ra các đơn vị cho cành và lá.

Nếu bộ dữ liệu bao gồm điểm thi, từ 0 đến 100, 83 điểm sẽ có một gốc là 8 và một lá 3. Đối với một tập dữ liệu hiệu quả nhiên liệu ô tô (mpg), dữ liệu nằm giữa 8.1 và 47.8, ta có thể sử dụng các chữ số hàng chục là gốc, do đó 32,6 sau đó sẽ có một lá 2,6. Nói chung, các biểu đồ gốc và lá dựa số gốc từ 5 đến 20 được khuyến khích sử dụng.

Ví dụ 1.6 Khảo sát chiều cao (đơn vị: m) của 45 sinh viên trường Đại học Sư phạm Kỹ thuật Tp. HCM ta có bảng số liệu:

1,55	1,73	1,65	1,71	1,55	1,51	1,68	1,56	1,63
1,55	1,57	1,63	1,59	1,54	1,80	1,45	1,60	1,55
1,60	1,65	1,70	1,68	1,70	1,65	1,56	1,54	1,52
1,64	1,67	1,50	1,80	1,48	1,65	1,70	1,82	1,79
1,70	1,40	1,67	1,70	1,67	1,58	1,60	1,62	1,52

Biểu đồ gốc lá của mẫu về chiều cao của 45 sinh viên có dạng: Biểu đồ cho thấy

Tần số	Gốc	& Lá	
1	1,4	0	
2	1,4	58	
6	1,5	012244	
9	1,5	555566789	Độ rộng của gốc: 0.10
7	1,6	0002334	Mỗi lá: 1 trường hợp
9	1,6	555577788	Đơn vị: m
7	1,7	0000013	
1	1,7	9	
3	1,8	002	

Bảng 1.2: Biểu đồ gốc lá của chiều cao 45 sinh viên trường Đại học Sư phạm Kỹ thuật Tp. HCM.

số lượng sinh viên có chiều cao từ 1,55m đến 1,59m và từ 1,65m đến 1,69m là đông

nhất, mỗi nhóm có 9 người. Phần lớn các sinh viên trong mẫu có chiều cao thuộc vào khoảng từ 1,50m đến 1,75m. Có một giá trị có khả năng bất thường trong mẫu là chiều cao của một sinh viên là 1,40m; tuy nhiên cỡ mẫu khá nhỏ 45 sinh viên nên cũng chưa thể kết luận về giá trị bất thường này.

Biểu đồ gốc và lá có ưu điểm là biểu điễn được dữ liệu gốc trên biểu đồ. Bên cạnh đó một biểu đồ gốc và lá truyền tải thông tin về các khía cạnh sau của dữ liệu:

- Xác định các giá trị tiêu biểu hoặc đại diện.
- Mức độ lan truyền của các giá trị tiêu biểu.
- Hiển diện của bất kỳ khoảng trống trong dữ liệu.
- Mức độ đối xứng trong phân phối giá trị.
- Số lượng và vị trí của đỉnh.
- Hiển diện của bất kỳ giá trị ngoại lai.

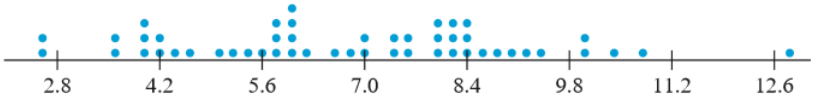
1.2.3 Biểu đồ Chấm

Một biểu đồ chấm là một bản tóm tắt hấp dẫn của các dữ liệu số khi tập dữ liệu là nhỏ hoặc có rất ít giá trị dữ liệu riêng biệt. Mỗi quan sát được đại diện bởi một dấu chấm ở trên các vị trí tương ứng trên thang điểm đo nằm ngang. Khi một giá trị xuất hiện nhiều hơn một lần, có một dấu chấm cho mỗi lần xuất hiện, và những dấu chấm được xếp chồng lên nhau theo chiều dọc. Giống như biểu đồ gốc và lá, một biểu đồ chấm dotplots cũng cung cấp thông tin về vị trí, mức độ lan truyền, giá trị ngoại lai và những khoảng trống.

Ví dụ 1.7 Dưới đây là số liệu về tỷ lệ phần trăm của doanh thu thuế của tiểu bang và địa phương trong năm tài chính 2006-2007 phân bổ cho giáo dục đại học (từ Tóm tắt thống kê của Hoa Kỳ); giá trị được liệt kê theo thứ tự chữ viết tắt của các bang (AL đầu tiên, WY cuối cùng):

10,8	6,9	8,0	8,8	7,3	3,6	4,1	6,0	4,4	8,3
8,1	8,0	5,9	5,9	7,6	8,9	8,5	8,1	4,2	5,7
4,0	6,7	5,8	9,9	5,6	5,8	9,3	6,2	2,5	4,5
12,8	3,5	10,0	9,1	5,0	8,1	5,3	3,9	4,0	8,0
7,4	7,5	8,4	8,3	2,6	5,1	6,0	7,0	6,5	10,3

Hình 1.5 cho thấy một biểu đồ chấm của dữ liệu, các giá trị thay đổi đáng kể. Giá trị lớn nhất (New Mexico) và hai giá trị nhỏ nhất (New Hampshire và Vermont) đều được phần nào tách khỏi phần lớn các dữ liệu, mặc dù chưa đủ để được coi là giá trị ngoại lai.



Hình 1.5: Biểu đồ chấm cho dữ liệu trong ví dụ trên.

1.2.4 Biểu đồ Histogram

Một số dữ liệu số thu được bằng cách đếm để xác định giá trị của một biến (số lượng giao thông trích dẫn một người nhận được trong năm qua, số lượng khách hàng đến với dịch vụ trong một khoảng thời gian cụ thể), trong khi các dữ liệu khác thu được bằng cách lấy số đo (trọng lượng của một cá nhân, thời gian phản ứng với một kích thích đặc biệt). Việc quy định cho việc vẽ một biểu đồ nói chung là khác nhau đối với hai trường hợp này.

Định nghĩa 1.2.1. Biến số là rời rạc nếu tập các giá trị có thể có hoặc là hữu hạn hoặc có thể liệt kê được trong một dãy vô hạn. Biến số liên tục nếu các giá trị có thể có của nó bao gồm toàn bộ khoảng trên trục số.

Một biến rời rạc x gần như luôn luôn là kết quả từ một quá trình đếm nào đó. Trong trường hợp đó có thể giá trị là 0, 1, 2, 3, . . . hoặc một số tập hợp con của các số nguyên. Các biến liên tục phát sinh từ việc đo lường. Ví dụ, nếu x là độ pH của một chất hóa học, trong khi đó theo lý thuyết x có thể là bất kỳ số giữa 0 và 14: 7.0, 7.03, 7.032, ... Tất nhiên, trong thực tế có những hạn chế về mức độ chính xác của bất kỳ dụng cụ đo lường, vì vậy chúng ta không thể biểu thị độ pH, thời gian phản ứng, chiều cao bởi một lượng lớn các chữ số thập phân tùy ý. Tuy nhiên, từ quan điểm của việc tạo ra mô hình toán học về phân phối dữ liệu, nó là hữu ích để hình dung toàn bộ một chuỗi các giá trị có thể.

Xét dữ liệu bao gồm các quan sát của một biến rời rạc x. **Tần số** của một giá trị của x bất kỳ cụ thể là số lần giá trị xảy ra trong tập dữ liệu. **Tần số tương đối**

của một giá trị là tỷ lệ giữa tần số của giá trị đó và số lượng quan sát của tập số liệu:

$$\text{Tần số tương đối} = \frac{\text{số lần giá trị xuất hiện}}{\text{số lượng quan sát của tập số liệu}}$$

Giả sử bộ dữ liệu của chúng ta bao gồm 200 quan sát về x= số lượng môn học trong một học kì một sinh viên đại học đang theo học. Nếu có 70 giá trị x là 3, thì

Tần số giá trị 3 của x là: 70

Tần số tương đối giá trị 3 của x là $\frac{70}{120} = 0.35$

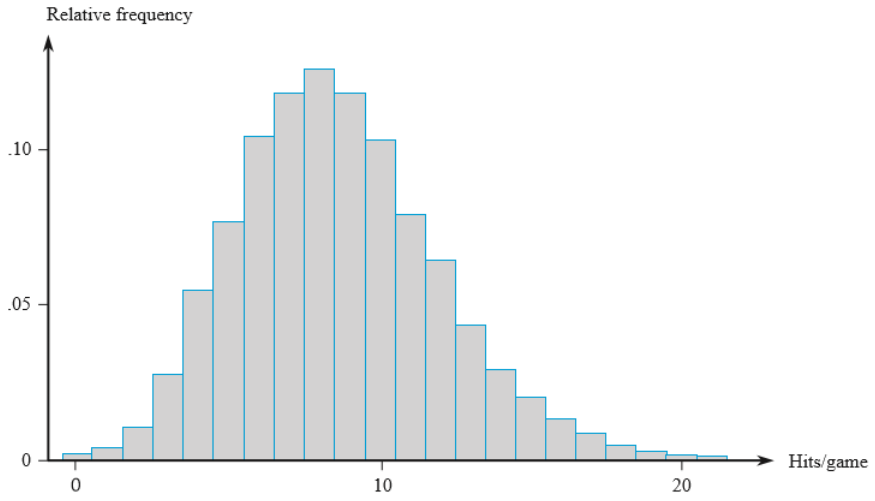
Hướng dẫn tạo biểu đồ cột
* Đầu tiên, xác định tần số và tần số tương đối của mỗi giá trị x.
* Sau đó đánh dấu các giá trị x có thể có trên một trục nằm ngang.
* Phía trên mỗi giá trị, vẽ một hình chữ nhật có chiều cao là tần số tương đối (hoặc cách khác là tỷ số) của giá trị đó.

Diện tích của mỗi hình chữ nhật tỉ lệ với tần số tương đối của các giá trị. Vì vậy, nếu các tần số tương đối của $x = 1$ và $x = 5$ tương ứng là 0.35 và 0.07, thì diện tích của hình chữ nhật trên 1 bằng là năm lần diện tích của hình chữ nhật trên 5.

Ví dụ 1.8 Theo dõi số lượt đánh trúng của các đội chơi trong 19383 trận đấu gồm 9 hiệp trong một giải bóng chày từ năm 1989 đến năm 1993 ta có bảng số liệu dưới đây. Tổng các tần số lớn hơn 1 một chút là do có sự làm tròn các tần số trong bảng. Tỷ lệ số trận có số lượt đánh trúng không quá 2 lượt và tỷ lệ đánh trúng được nhiều hơn 10, 15, 20 cú đánh là bao nhiêu?

Số lượt trúng/Trận	Số trận	Tỷ lệ	Số lượt trúng/Trận	Số trận	Tỷ lệ
0	20	0,0010	14	569	0,0294
1	72	0,0037	15	393	0,0203
2	209	0,0108	16	253	0,0131
3	527	0,0272	17	171	0,0088
4	1048	0,541	18	97	0,0050
5	1457	0,0752	19	53	0,0027
6	1988	0,1026	20	31	0,0016
7	2256	0,1164	21	19	0,0010
8	2403	0,1240	22	13	0,0007
9	2256	0,1164	23	5	0,0003
10	1967	0,1015	24	1	0,0001
11	1509	0,0779	25	0	0,0000
12	1230	0,0635	26	1	0,0001
13	834	0,0430	27	1	0,0001

Biểu đồ histgram của dữ liệu trong hình 1.6 tăng khá suôn sẻ đến một đỉnh cao



Hình 1.6: Biểu đồ cột các cú đánh trong một trận 9 hiệp.

duy nhất và sau đó thì giảm. Biểu đồ kéo dài hơn một chút về bên phải (đối với giá trị lớn) hơn là so với bên trái, ta nói biểu đồ "lệch dương."

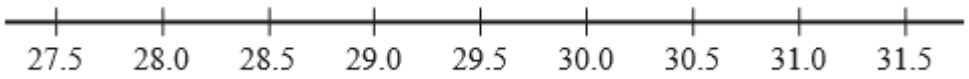
		tần suất	tần suất	tần suất
Tỷ lệ của các trận đấu với	=	tương đối +	tương đối +	tương đối
nhiều nhất là hai cú đánh		với $x = 0$	với $x = 1$	với $x = 2$
	=	0,0010 +	0,0037+	0,0108=0,155

Tương tự ta có

Tỷ lệ của các trận đấu với	=	0,0010 +	0,1026+	...+0,1015=0,6361
số cú đánh từ 5 đến 10				

Bạn đọc có thể tử trả lời câu hỏi còn lại tỷ lệ trận đấu đánh trúng được nhiều hơn 10, 15, 20 cú đánh là bao nhiêu bằng cách tính tương tự.

Xây dựng một biểu đồ cho dữ liệu liên tục đòi hỏi phải phân chia các trục đo thành một số lớp (khoảng) thích hợp sao cho mỗi quan sát được chứa trong chính xác một khoảng. Giả sử, ví dụ chúng ta có 50 quan sát về $x =$ hiệu quả nhiên liệu của một ô tô (mpg), các số đó nhỏ nhất là 27,8 và trong đó lớn nhất là 31,4. Sau đó, chúng ta có thể sử dụng các ranh giới giữa các khoảng là 27,5; 28,0; 28,5;. . . và 31,5 như ở đây:



Để tránh trường hợp một quan sát nằm trên ranh giới giữa các lớp (khoảng) như vậy khó xác định nó thuộc chính xác một khoảng nào, ví dụ như 29,0 ta thường sử dụng các lớp 27,5 đến <28,0; 28,0 đến <28,5; ...; 31,0 đến <31,5. Khi đó 29,0 rơi trong lớp 29,0 đến <29,5 chứ không phải trong lớp 28,5 đến 29,0. Nói cách khác, với quy ước này, một quan sát trên một ranh giới được đặt trong khoảng bên phải của ranh giới.

Hướng dẫn tạo biểu đồ cột cho dữ liệu liên tục:

Các khoảng có độ rộng như nhau

* Xác định tần số và tần số tương đối cho mỗi lớp.
* Đánh dấu ranh giới lớp trên một trục đo nằm ngang.
* Phía trên mỗi lớp (khoảng), vẽ một hình chữ nhật có chiều cao là tần số tương đối (hoặc tỷ lệ) tương ứng.

Ví dụ 1.9 Công ty điện lực cần thông tin về cách sử dụng điện của khách hàng để có được dự báo về nhu cầu sử dụng của khách hàng. Các nhà điều tra từ công ty

điện và ánh sáng Wisconsin xác định lượng năng lương tiêu thụ (BTU) trong một giai đoạn cụ thể cho một mẫu của 90 nhà. Giá trị tiêu thụ hiệu chỉnh được tính toán như sau:

$$\text{giá trị tiêu thụ hiệu chỉnh} = \frac{\text{tiêu thụ}}{(\text{thời tiết, nhiệt độ trong ngày})(\text{diện tích nhà})}$$

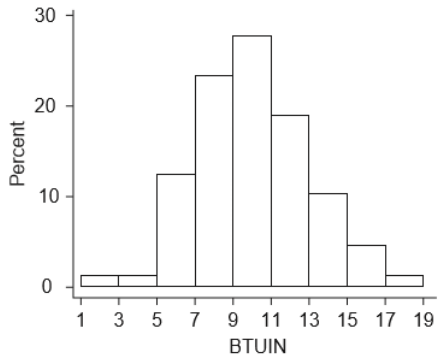
Dữ liệu được sắp xếp từ nhỏ đến lớn như sau

2.97	4.00	5.20	5.56	5.94	5.98	6.35	6.62	6.72	6.78
6.80	6.85	6.94	7.15	7.16	7.23	7.29	7.62	7.62	7.69
7.73	7.87	7.93	8.00	8.26	8.29	8.37	8.47	8.54	8.58
8.61	8.67	8.69	8.81	9.07	9.27	9.37	9.43	9.52	9.58
9.60	9.76	9.82	9.83	9.83	9.84	9.96	10.04	10.21	10.28
10.28	10.30	10.35	10.36	10.40	10.49	10.50	10.64	10.95	11.09
11.12	11.21	11.29	11.43	11.62	11.70	11.70	12.16	12.19	12.28
12.31	12.62	12.69	12.71	12.91	12.92	13.11	13.38	13.42	13.43
13.47	13.60	13.96	14.24	14.35	15.12	15.24	16.06	16.90	18.26

Dữ liệu biểu diễn dưới dạng các lớp (khoảng) có dạng

Khoảng	1-3	3-5	5-7	7-9	9-11	11-13	13-15	15-17	17-19
Tần số	1	1	11	21	25	17	9	4	1
Tỷ lệ	0,011	0,011	0,122	0,233	0,278	0,189	0,100	0,044	0,011

Biểu đồ Histogram của dữ liệu như sau



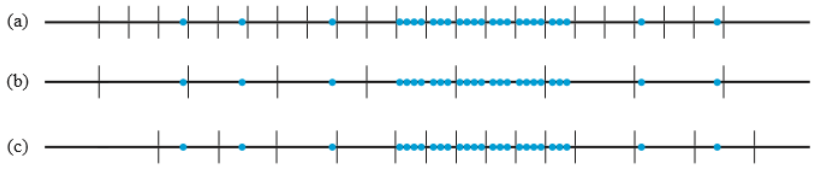
Hình 1.7: Biểu đồ cột cho dữ liệu mức tiêu thụ năng lượng trong ví dụ trên.

Tỷ lệ hộ được quan sát có giá trị tiêu thụ hiệu chỉnh nhỏ hơn 9 là $\frac{34}{90} \approx 0,0378$

Không có quy tắc cố định nào vè việc chọn số lớp (khoảng). Từ 5 đến 20 lớp (khoảng) hầu hết là hợp lý cho các bộ dữ liệu. Nói chung, số lượng lớp (khoảng) nhiều nên được sử dụng. Một quy tắc hợp lý có thể sử dụng được là quy tắc ngón tay cái

$$\text{số lượng lớp (khoảng)} \approx \sqrt{\text{số lượng quan sát}}$$

Độ rộng các khoảng bằng nhau có thể không hợp lý trong trường hợp quy mô đo lường mà có sự tập trung cao các giá trị dữ liệu ở một số nơi và ở các nơi khác dữ liệu xuất hiện khá thưa thớt. Hình 1.8 cho thấy một biểu đồ chấm của một tập dữ liệu có các quan sát tập trung nhiều ở giữa và tương đối ít các quan sát kéo dài ra hai bên. Sử dụng một số lượng nhỏ các khoảng có độ rộng bằng nhau thì kết quả hầu hết các quan sát rơi vào chỉ một hoặc hai trong số các khoảng. Nếu một số lượng lớn các lớp có độ rộng bằng nhau được sử dụng thì nhiều lớp sẽ có tần số zero. Một cách chia khoảng hợp lý hơn là sử dụng một vài khoảng rộng gần các quan sát phía hai bên và khoảng hẹp hơn trong các khu vực có các quan sát tập trung cao.



Hình 1.8: Chọn khoảng lớp học cho dữ liệu có "mật độ khác nhau" dữ liệu: (a) nhiều khoảng có chiều rộng khoảng bằng nhau và độ rộng khoảng là ngắn; (b) số ít khoảng có chiều rộng bằng nhau và độ rộng khoảng là lớn; (C) các khoảng có chiều rộng không bằng nhau.

Hướng dẫn tạo biểu đồ cột cho dữ liệu liên tục:

Các khoảng có độ rộng không như nhau

Sau khi xác định tần số và tần số tương đối, tính chiều cao của mỗi hình chữ nhật theo công thức

$$\text{chiều cao hình chữ nhật} = \frac{\text{tần số tương đối của lớp}}{\text{độ rộng của lớp}}$$

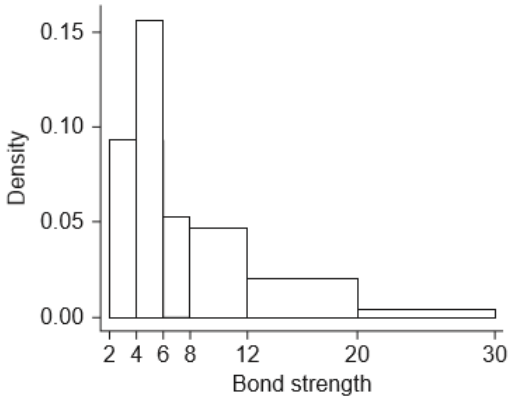
Kết quả là chiều cao hình chữ nhật thường được gọi là mật độ và thang đo theo chiều dọc là thang đo mật độ. Công thức này cũng đúng khi độ rộng các lớp bằng nhau.

Ví dụ 1.10 Sự ăn mòn của cốt thép là một vấn đề nghiêm trọng trong kết cấu bê tông trong những môi trường bị ảnh hưởng bởi điều kiện thời tiết khắc nghiệt. Vì lý do này, các nhà nghiên cứu đã tiến hành điều tra việc sử dụng cốt thép làm bằng vật liệu composite. Một nghiên cứu đã được thực hiện để phát triển các hướng dẫn liên kết sợi thủy tinh gia cố cốt thép bê tông nhựa ("Kiến nghị thiết kế cho độ dính bám của thép GFRP với bê tông" J. of Structural Engr, 1996: trang 247-254). Hãy xem xét về cường độ dính bám được đo trong 48 quan sát sau:

11.5	12.1	9.9	9.3	7.8	6.2	6.6	7.0	13.4	17.1	9.3	5.6
5.7	5.4	5.2	5.1	4.9	10.7	15.2	8.5	4.2	4.0	3.9	3.8
3.6	3.4	20.6	25.5	13.8	12.6	13.1	8.9	8.2	10.7	14.2	7.6
5.2	5.5	5.1	5.0	5.2	4.8	4.1	3.8	3.7	3.6	3.6	3.6

Class	2−<4	4−<6	6−<8	8−<12	12−<20	20−<30
Frequency	9	15	5	9	8	2
Relative frequency	.1875	.3125	.1042	.1875	.1667	.0417
Density	.094	.156	.052	.047	.021	.004

Biểu đồ hình cột của dữ liệu được trình bày trong hình 1.9.



Hình 1.9: Biểu đồ mật độ cho dữ liệu cường độ bám dính.

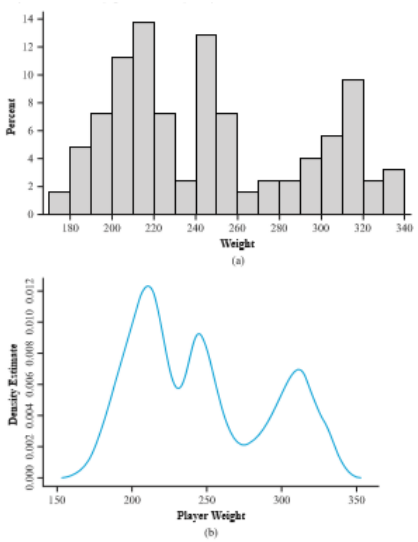
tần số tương đối = (độ rộng lớp).(mật độ)
=(chiều rộng hình chữ nhật).(chiều cao hình chữ nhật)
= diện tích hình chữ nhật

Diện tích mỗi hình chữ nhật là tần số tương đối của các lớp tương ứng. Hơn nữa, tổng của tần số tương đối phải 1 nên tổng diện tích của tất cả các hình chữ nhật

trong một biểu đồ mật độ là 1.

Hình dạng biểu đồ

Ví dụ 1.11 Hình 1.10 (a) cho thấy một biểu đồ của trọng lượng (đơn vị: lb) của 124 người chơi được liệt kê trên rosters của San Francisco 49ers và The New England Patriots. Hình 1.10 (b) là một biểu đồ được làm mượt (thực tế được gọi là một ước lượng của hàm mật độ) của dữ liệu từ các gói phần mềm R. Cả hai biểu đồ đều có ba đỉnh riêng biệt.

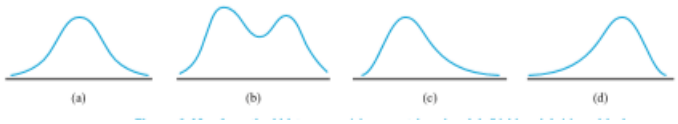


Hình 1.10: Trọng lượng của các người chơi NFL (a) biểu đồ cột; (b) Biểu đồ được làm mượt.

Một biểu đồ Histogram được gọi là đối xứng nếu nửa bên trên giống với nửa bên phải. Một biểu đồ Histogram không đối xứng gọi là lệch phải nếu phần bên phải hay phần đuôi tương ứng với phần giá trị lớn trải dải hơn so với phần bên trái. Hình 1.11 biểu diễn một số biểu đồ Histogram được làm mượt.

Dữ liệu định tính

Phân bố tần số và biểu đồ Histogram có thể được xây dựng khi tập dữ liệu là định tính (biến có giá trị phân loại). Trong một số trường hợp, sẽ có một thứ tự tự nhiên của các lớp, ví dụ như sinh viên năm nhất, năm thứ hai; đàn em, người cao tuổi; ... Trong khi ở các trường hợp khác theo thứ tự sẽ được xếp tùy ý, ví dụ như Công



Hình 1.11: Biểu đồ được làm mượt: (a) một đỉnh đối xứng; (b) hai đỉnh; (c) nghiêng dương; (d) nghiêng âm.

giáo, Do thái giáo, Tin Lành, ... Với dữ liệu chủng loại như vậy, các khoảng mà trên đó các hình chữ nhật được xây dựng nên có độ rộng mỗi khoảng bằng nhau.

Ví dụ 1.12 Viện Chính sách công cộng California đã tiến hành một cuộc khảo sát qua điện thoại của 2501 cư dân California là người lớn trong tháng 4 năm 2006 để xác định họ cảm thấy như thế nào về các khía cạnh khác nhau của K-12 giáo dục công. Một câu hỏi chung là "Bạn đánh giá chất lượng của các trường công lập ở khu phố của bạn hiện nay như thế nào?" Bảng sau hiển thị các tần số và tần số tương đối của dữ liệu thu được và hình 1.12 là biểu đồ Histogram tương ứng của dữ liệu.

Đánh giá	Tần số	Tỷ lệ
A	478	0,191
B	893	0,367
C	680	0,272
D	178	0,071
F	100	0,040
Không rõ	172	0,069

Hơn một nửa số người được hỏi cho điểm đánh giá A hoặc B và chỉ có khoảng hơn 10% đã đưa ra một D hoặc F giá..

Dữ liệu đa biến

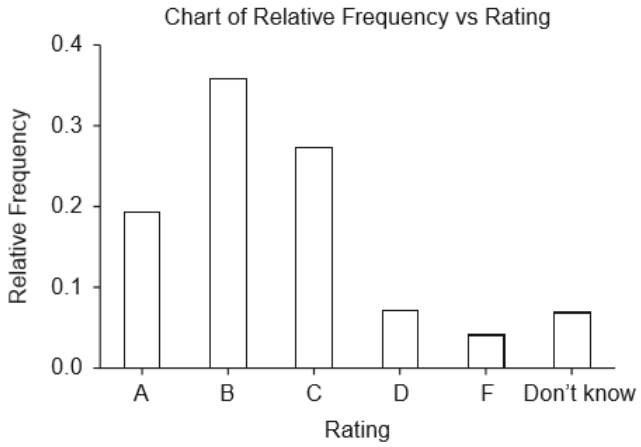
Dữ liệu đa biến nói chung là khá khó khăn để mô tả trực quan. Một trong các phương pháp để mô tả là dùng các ô phân tán cho dữ liệu số hai biến.

Bài tập 1.2

1.2.1 Sử dụng dữ liệu cường độ bê tông trong ví dụ 1.2,xây dựng biểu đồ gốc lá cho dữ liệu. Tính tỷ lệ bê tông trong mẫu có cường độ nén vượt quá 10 MPa?

1.2.2 Xây dựng biểu đồ lá và cột cho bộ dữ liệu sau:

31	35	36	36	37	38	40	40	40
41	41	42	42	42	42	42	43	44



Hình 1.12: Biểu đồ các đánh giá về trường học từ Minitab.

45	46	46	47	48	48	48	51	54
54	55	58	62	66	66	67	68	75

1.3 Các số đo đặc trưng vị trí

1.3.1 Trung Bình

Cho tập hợp $x_1, x_2, ..., x_n$, các biện pháp quen thuộc và hữu ích nhất để đo trung tâm là trung bình hoặc trung bình số học của tập hợp này. Chúng ta ký hiệu trung bình mẫu là $\bar{x}$.

Định nghĩa 1.3.1. Trung bình mẫu $\bar{x}$ của các quan sát $x_1, x_2, ..., x_n$ cho bởi

$$\bar{x} = \frac{x_1 + x_2 + ... + x_n}{n} = \frac{\Sigma_{i=1}^{n} x_i}{n}$$

Tử số của $\bar{x}$ có thể được viết là Σx_i, là tổng của tất cả các quan sát của mẫu.

Ví dụ 1.13 Sự ăn mòn, nứt sắt và thép của chất kiềm đã được nghiên cứu vì những lỗi của các đinh tán trong nồi thép và lỗi của các rotor hơi. Xét các quan sát $x =$ chiều dài vết nứt (μm), là kết quả của các bài kiểm tra ăn mòn liên tục trên các mẫu là các thanh căng mịn trong một thời gian cố định. (Các dữ liệu phù hợp với một biểu đồ và bản tóm tắt số lượng từ bài báo "Vai trò của Phospho trong chất kiềm gây ăn mòn, nứt Thép hợp kim" Corrosion Science, 1989: 53–68.)

$$x_1 = 16.1\ x_2 = 9.6\ \ x_3 = 24.9\ x_4 = 20.4\ x_5 = 12.7\ x_6 = 21.2\ x_7 = 30.2$$
$$x_8 = 25.8\ x_9 = 18.5\ x_{10} = 10.3\ x_{11} = 25.3\ x_{12} = 14.0\ x_{13} = 27.1\ x_{14} = 45.0$$
$$x_{15} = 23.3\ x_{16} = 24.2\ x_{17} = 14.6\ x_{18} = 8.9\ \ x_{19} = 32.4\ x_{20} = 11.8\ x_{21} = 28.5$$

Hình 1.13 biểu diễn biểu đồ cuống và lá của dữ liệu; chiều dài vết nứt ở độ tuổi 20 thấp dường như là "điển hình".

0H	96	89			
1L	27	03	40	46	18
1H	61	85			
2L	49	04	12	33	42
2H	58	53	71	85	
3L	02	24			
3H					
4L					
4H	50				

Stem: tens digit
Leaf: one and tenths digit

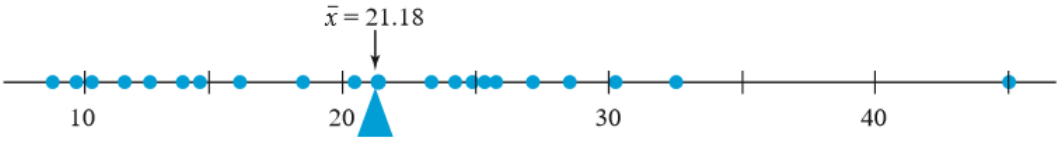
Hình 1.13: Biểu đồ cuống và lá cho dữ liệu chiều dài các vết nứt.

Với $\Sigma x_i = 444,8$, trung bình mẫu là

$$\bar{x} = \frac{444,8}{21} = 21,18$$

một giá trị phù hợp với thông tin từ biểu đồ gốc và lá.

Một giải thích vật lý của $\bar{x}$ sẽ giải thích cách nó đo vị trí (trung tâm) của một mẫu. Vẽ và mở rộng một trục đo ngang, sau đó đại diện cho mỗi quan sát của mẫu bằng một trọng lượng 1-lb đặt tại các điểm tương ứng trên trục. Điểm duy nhất mà tại đó một điểm tựa có thể được đặt để cân bằng trọng lượng của hệ thống là điểm tương ứng với giá trị của $\bar{x}$ (xem Hình 1.14).



Hình 1.14: Giá trị trung bình như là điểm cân bằng cho hệ thống trọng lượng.

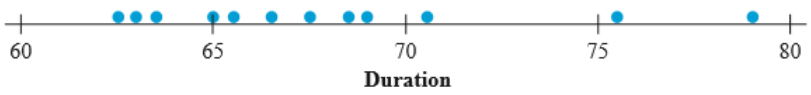
$\bar{x}$ đại diện cho các giá trị trung bình của các quan sát trong một mẫu, trung bình của tất cả các giá trị trong tổng thể có thể được tính được. Trung bình này được gọi là trung bình tổng thể và được ký hiệu bằng các chữ cái Hy Lạp μ. Khi có N giá trị trong tổng thể (tổng thể hữu hạn), thì μ = (tổng các giá trị trong tổng thể N) / N. Trong chương 3 và 4, chúng ta sẽ đưa ra một định nghĩa chung cho rằng áp dụng cho cả tổng thể hữu hạn và (theo lý thuyết) vô hạn. Cũng như $\bar{x}$ là một giá trị quan trọng của vị trí mẫu, μ thường là một giá trị vị trị quan trọng của một tổng thể.

1.3.2 Median-Trung vị

Đồng nghĩa với Median là "trung vị" (vị trí ở chính giữa). Trung vị thực sự là giá trị ở chính giữa các quan sát được sắp xếp từ nhỏ nhất đến lớn nhất. Khi quan sát được biểu thị bằng $x_1, x_2, ..., x_n$, chúng ta sẽ sử dụng các biểu tượng $\tilde{x}$ đại diện cho trung vị mẫu.

Định nghĩa 1.3.2. Sắp xếp các giá trị quan sát của mẫu từ nhỏ đến lớn (với bất kì giá trị quan sát nào lặp lại trong mẫu). Khi đó
* Nếu n là số lẻ thì $\tilde{x}$ là giá trị chính giữa thứ $\frac{n+1}{2}$.
* Nếu n là số chẵn thì $\tilde{x}$ là trung bình của hai giá trị chính ở giữa thứ $\frac{n}{2}$ và $\frac{n}{2}+1$.

Ví dụ 1.14 Những người không quen với âm nhạc cổ điển có thể có xu hướng tin rằng hướng dẫn của một nhà soạn nhạc để chơi một đoạn nhạc cụ thể sẽ không phụ thuộc vào tất cả những người biểu diễn. Tuy nhiên, thường có rất nhiều room để giải thích và người chỉ huy dàn nhạc và các nhạc sĩ tận dụng điều này. Các tác giả đã đi đến trang web ArkivMusic.com và chọn một mẫu của 12 bản ghi âm của Symphony # 9 ("hợp xướng", một tác phẩm tuyệt hay) của Beethoven, thì sự đàn hồi thời gian sau (phút) được liệt kê theo thứ tự tăng dần:
62,3 62,8 63,6 65,2 65,7 66,4 67,4 68,4 68,8 70,8 75,7 79,0
Biểu đồ chấm của dữ liệu $n = 12$ là số chẵn nên trung vị của mẫu là giá trị trung



Hình 1.15: Biểu đồ chấm cho dữ liệu cho ví dụ trên.

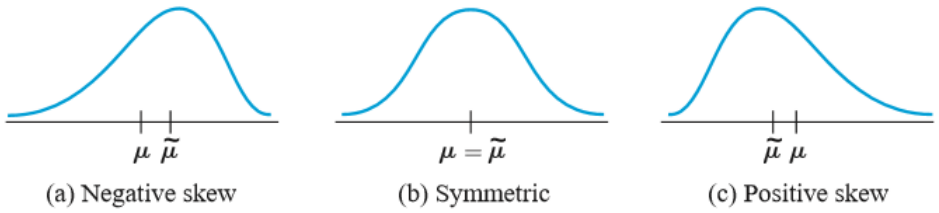
bình của hai giá trị thứ $n/2 = 6$ và $n/2 + 1 = 7$ trong dãy các giá trị đã sắp xếp

$$\tilde{x} = \frac{66,4 + 67,4}{2} = 66,90$$

Lưu ý rằng nếu quan sát lớn nhất 79,0 không được bao gồm trong mẫu, $n = 11$, kết quả trung vị mẫu cho các quan sát còn lại sẽ có được giá trị chính giữa (các giá trị được sắp thứ tự, tức là giá trị thứ 6 từ hai đầu củadanh sách) là 66,4 . Giá trị trung bình mẫu là $\bar{x} = \Sigma x_i = 816, 1/12 = 68, 01$, lớn hơn một phút so với trung vị. Giá trị trung bình được kéo ra một chút so với mức trung vị bởi vì mẫu "kéo dài ra" hơi nhiều về phần cuối phía trên hơn là phần cuối phía dưới.

Tương tự $\tilde{x}$ là giá trị trung vị trong mẫu, giá trị ở giữa tổng thể là trung vị của tổng thể ta kí hiệu bằng $\tilde{\mu}$. Ta sử dụng các trung vị mẫu $\tilde{x}$ như là một suy luận về trung vị của tông thể $\tilde{\mu}$.

Trung bình tổng thể μ và trung vị $\tilde{\mu}$ nói chung là không giống hệt nhau. Nếu phân bố tổng thể bị lệch trái hay lệch phải như ở hình 1.16. Khi đó để đưa ra kết luận về tổng thể, đầu tiên chúng ta quyết định quan tâm đến tham số nào của tổng thể hơn.



(a) Negative skew (b) Symmetric (c) Positive skew

Hình 1.16: Ba hình dạng khác nhau của phân phối của tổng thể.

1.3.3 Các tham số vị trí khác: Tứ phân vị, phân vị mức phần trăm và trung bình thu gọn.

Trung vị chia tập dữ liệu thành hai phần có kích thước bằng nhau. Để có những tham số vị trí tốt hơn, ta có thể phân chia dữ liệu thành nhiều hơn hai phần như vậy. Nói chung các tham số tứ phân vị chia dữ liệu thành bốn phần bằng nhau. Tương tự như vậy, một tập hợp dữ liệu (mẫu hoặc tổng thể) có thể còn phân chia tinh tế hơn bằng các tham số phần trăm.

Giá trị trung bình μ là khá cần thiết với một tham số đưa ra duy nhất, trong khi trung vị là không thể hiểu được với dữ liệu có nhiều giá trị ngoại lai. Một trung bình thu gọn là một sự thỏa hiệp giữa trung bình và trung vị. Ví dụ một trung bình thu gọn 10% sẽ được tính bằng cách loại bỏ 10% các giá trị nhỏ nhất và 10% các giá trị lớn nhất của dữ liệu và sau đó tính trung bình các giá trị còn lại.

1.3.4 Phân loại dữ liệu và Tỷ lệ mẫu.

Khi dữ liệu được phân loại, phân bố tần số hoặc phân phối tần số tương đối cung cấp một bản tóm tắt của dữ liệu. Ví dụ một cuộc khảo sát sở hữu máy ảnh kỹ thuật số của các cá nhân được thực hiện để nghiên cứu sự lựa chọn thương hiệu. Sau đó đếm số lượng người sở hữu Canon, Sony, Kodak , ... Xét việc lấy mẫu phân đôi mà tổng thể một chỉ gồm hai thương hiệu nào đó. Nếu chúng ta kí hiệu x là số cá thể trong mẫu rơi vào loại 1, thì số lượng loại 2 là $n - x$. Các tần số hoặc tỷ lệ mẫu tương đối loại 1 là x/n và tỷ lệ mẫu loại 2 là $1 - x/n$.

Tổng quát hơn, tập trung sự chú ý vào một thể loại đặc biệt ta đánh mã 1 cho các quan sát trong mẫu rôi vào loại này và mã 0 cho các quan sát không thuộc loại này. Sau đó, tỷ lệ mẫu của các quan sát trong các loại này là trung bình mẫu của các chuỗi 1 và 0. Vì vậy, một trung bình mẫu có thể được sử dụng để tóm tắt các kết quả của một mẫu phân loại.

Bài tập 1.3

1.3.1 Cho số liệu 590 815 575 608 350 1285 408 540 555 67

a. Tính giá trị trung bình và trung vị mẫu.

b. Giả sử quan sát thứ 6 là 985 thay vì 1285. Mức trung bình và trung vị thay đổi như thế nào?

c. Tính trung bình rút gọn 20% được xác định bằng cách bỏ đi hai quan sát nhỏ nhất và hai quan sát lớn nhất.

d. Tính trung bình rút gọn 30%.

1.3.2 Cho bộ số liệu:

U:	6,0	5,0	11,0	33,0	4,0	5,0	80,0	18,0	35,0	17,0	23,0
F:	4,0	14,0	11,0	9,0	9,0	8,0	4,0	20,0	5,0	8,9	21,0
	9,2	3,0	2,0	0,5							

a. Xác định trung bình mẫu cho mỗi mẫu và so sánh.
b. Xác định trung vị cho mỗi mẫu và so sánh. Tại sao trung bình mẫu và trung vị của mẫu đó khác nhau?
c. Tính trung bình rút gọn cho từng mẫu bằng cách xóa quan sát nhỏ nhất và lớn nhất. Giá trị này tương ứng với giá trị trung bình bao nhiêu phần trăm? So sánh các giá trị trung bình rút gọn và trung bình, trung vị mẫu tương ứng.

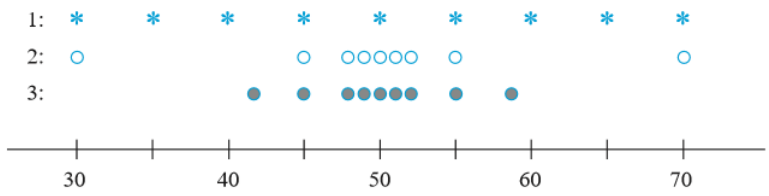
1.3.3 Cho bộ số liệu sau:

389	356	359	363	375	424	325	394	402
373	373	370	364	366	364	325	339	393
392	369	374	359	356	403	334	397	

a. Xây dựng biểu đồ gốc lá, cột cho bộ số liệu.
b. Tính giá trị trung bình và trung vị mẫu.
c. Quan sát lớn nhất hiện tại là 424, giá trị này có thể tăng nhiều nhất là bao nhiêu mà không ảnh hưởng đến giá trị trung vị mẫu.

1.4 Các số đo đặc trưng biến thiên

Các tham số trung tâm chỉ cung cấp một phần thông tin của bộ dữ liệu hoặc phân phối. Mẫu hoặc các tổng thể khác nhau có thể giống hệt nhau của các tham số trung tâm nhưng khác biệt nhau theo các cách quan trọng khác. Hình 1.17 biểu diễn biểu đồ chấm của ba mẫu với giá trị trung bình và trung vị tương tự nhau, nhưng mức độ lan truyền về trung tâm là khác nhau cho tất cả ba mẫu.



Hình 1.17: Các mẫu có độ đo trung tâm giống hệt hệt nhau nhưng khác nhau về độ biến thiên.

1.4.1 Độ đo độ biến thiên của mẫu

Đo độ biến thiên đơn giản nhất của một mẫu là phạm vi mẫu là sự khác biệt giữa các giá trị mẫu lớn nhất và nhỏ nhất. Phạm vi các giá trị mẫu 1 trong hình

1.17 là lớn hơn nhiều so với mẫu 3, phản ánh sự biến đổi ở mẫu đầu tiên là nhiều hơn trong mẫu thứ ba. Một khiếm khuyết trong phạm vi của mẫu 2 là nó phụ thuộc vào 2 giá trị ngoài cùng và không quan tâm đến vị trí của các $n-2$ giá trị còn lại. Mẫu 1 và 2 trong hình 1.17 có phạm vi giống hệt nhau, nhưng khi ta tập trung vào các quan sát giữa hai hai giá trị ngoài cùng, độ phân tán trong mẫu sẽ thay đổi ở mẫu thứ hai so với độ phân tán mẫu lúc đầu (bao gồm cả 2 giá trị ngoài cùng).

Độ đo độ biến thiên đơn giản nhất là đo đọ sai lệch so với giá trị trung bình. Đó là các sai lệch các giá trị của mẫu so với giá trị trung bình $x_1-\bar{x}, x_2-\bar{x}, ..., x_n-\bar{x}$. Một cách đơn giản để kết hợp các độ lệch vào một lượng duy nhất là trung bình của chúng. Thật không may, đây là một ý tưởng tồi vì

$$\text{Tổng các độ lệch} = \Sigma_{i=1}^{n}(x_i-\bar{x}) = 0$$

Một cách để khắc phục điều này là tính trung bình các trị tuyệt đối các độ lệch $\Sigma_{i=1}^{n}|x_i-\bar{x}|$. Bởi vì các giá trị tuyệt đối dẫn đến một số khó khăn về mặt lý thuyết. Ta xét bình phương các độ lệch $(x_1-\bar{x})^2, (x_2-\bar{x})^2, ..., (x_n-\bar{x})^2$, thay vì sử dụng các trị tuyệt đối ta sử dụng trung bình bình phương các độ lệch $\Sigma_{i=1}^{n}(x_i-\bar{x})^2$, vì nhiều lý do chúng ta chia tổng của bình phương sai độ lệch cho $n-1$ thay vì chia cho n.

Định nghĩa 1.4.1. Phương sai mẫu, kí hiệu s^2, xác định bởi

$$s^2 = \frac{\Sigma(x_i-\bar{x})^2}{n-1} = \frac{S_{xx}}{n-1}$$

Độ lệch chuẩn mẫu, kí hiệu s, là căn bậc hai dương của phương sai mẫu

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

Chú ý rằng cả phương sai mẫu s^2 và độ lệch chuẩn mẫu s đều không âm.

Ví dụ 1.15 Các trang web www.fueleconomy.gov chứa rất nhiều thông tin về đặc tính nhiên liệu của xe khác nhau. Ngoài ra EPA xếp hạng theo dặm đường, có rất nhiều loại xe mà người dùng đã báo cáo các giá trị về hiệu quả nhiên liệu (mpg) của riêng họ. Hãy xem xét mẫu cỡ $n=11$ sau đây về hiệu quả cho xe Ford Focus 2009 được trang bị hộp số tự động (đối với mô hình này, EPA báo cáo đánh giá tổng thể 27 mpg-24 mpg đối với xe lái thành phố và 33 mpg cho xe lái trên đường cao tốc):

Car	x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	
1	27.3	−5.96	35.522	
2	27.9	−5.36	28.730	
3	32.9	−0.36	0.130	
4	35.2	1.94	3.764	
5	44.9	11.64	135.490	
6	39.9	6.64	44.090	
7	30.0	−3.26	10.628	
8	29.7	−3.56	12.674	
9	28.5	−4.76	22.658	
10	32.0	−1.26	1.588	
11	37.6	4.34	18.836	
	$\sum x_i = 365.9$	$\sum(x_i - \bar{x}) = .04$	$\sum(x_i - \bar{x})^2 = 314.106$	$\bar{x} = 33.26$

Việc làm tròn dẫn đến tổng $\Sigma(x_i - \bar{x})$ không bằng 0. $S_{xx} = 314,106$ từ đó

$$s^2 = \frac{S_{xx}}{n-1} = \frac{314,106}{11-1} = 31,41; \quad s = 5,60$$

Động lực cho s^2

Để giải thích lý do cho số chia trong s^2 là $n-1$, trước tiên lưu ý là trong khi s^2 đo độ biến thiên mẫu, có một tham số đo độ biến đổi trong tổng thể gọi là phương sai tổng thể. Ta kí hiệu phơng sai tổng thể là σ^2 và độ lệch chuẩn tổng thể là σ. Khi tổng thể là hữu hạn và có N cá thể

$$\sigma^2 = \frac{\Sigma_{i=1}^{N}(x_i - \mu)^2}{N}$$

Cũng như $\bar{x}$ sẽ được sử dụng để suy luận về trung bình tổng thể μ, ta xác định các phương sai mẫu để sử dụng là suy luận về σ^2. Lưu ý là phương sai σ^2 liên quan đến trung bình tổng thể μ. Nếu chúng ta thực sự biết giá trị của μ, sau đó chúng ta có thể xác định phương sai mẫu là trung bình bình phương độ lệch của các x_i so với trung bình μ. Tuy nhiên, giá trị của μ hầu như không bao giờ biết đến, vì vậy tổng bình phương độ lệch so với $\bar{x}$ phải được sử dụng. Nhưng các x_i có xu hướng gần gũi trung bình mẫu $\bar{x}$ hơn với trung bình tổng thể μ, do đó, để bù đắp cho điều này mẫu số $n-1$ được sử dụng chứ không phải là n.

Ta xem s^2 có $n-1$ bậc tự do (df). Thuật ngữ này phản ánh thực tế là mặc dù s^2 được xác định dựa trên n giá trị $x_1 - \bar{x}, x_2 - \bar{x}, ..., x_n - \bar{x},$, nhưng các giá trị này có tổng bằng 0, do đó ta có thể xác định bất kì giá trị nào dựa trên $n-1$ giá trị còn lại.

Công thức tính s^2

Một biểu thức thay thế cho các tử số của s^2 là

$$S_{xx} = \Sigma(x_i - \bar{x})^2 = \Sigma x_i^2 - \frac{(\Sigma x_i)^2}{n}$$

Mệnh đề 1.4.2. *Cho mẫu $x_1, x_2, ..., x_n$ và c là một hằng số khác 0*
1. Nếu $y_1 = x_1 + c, y_2 = x_2 + c, ..., y_n = x_n + c$ thì $s_{xx}^2 = s_{yy}^2$ và
2. Nếu $y_1 = cx_1, y_2 = cx_2, ..., y_n = cx_n$ thì $s_{xx}^2 = c^2 s_{yy}^2, s_y = |c|s_x$
trong đó s_{xx}^2 là phương sai với mẫu các x_i và s_{yy}^2 là phương sai các mẫu y_i

Các tính chất này có thể được chứng minh bằng cách chú ý đối với kết quả 1 là $\bar{y} = \bar{x} + c$ và đối với kết quả 2 là $\bar{y} = c\bar{x}$.

1.4.2 Biểu đồ hộp

Biểu đồ gốc-lá và biểu đồ Histogram biểu diễn một tập hợp dữ liệu một cách khá chung chung, trong khi đó một bản tóm tắt đơn giản như là trung bình hoặc độ lệch chuẩn chỉ tập trung vào một khía cạnh của dữ liệu. Trong những năm gần đây, một bản tóm tắt bằng hình ảnh được gọi là biểu đồ hộp (boxplot) đã được sử dụng thành công để mô tả một số tính năng nổi bật nhất của một tập hợp dữ liệu của. Các tính năng này bao gồm: (1) trung tâm, (2) mức độ phân tán, (3) mức độ và tính chất đối xứng, và (4) xác định "giá trị ngoại lai" các quan sát nằm xa bất thường của tập hợp dữ liệu chính. Bởi vì ngay cả một giá trị ngoại lai duy nhất có thể ảnh hưởng đáng kể giá trị của $\bar{x}$ và s, một boxplot được dựa trên các độ đo "kháng" với sự có mặt của vài giá trị ngoại lai - trung vị và độ phân tán được gọi là tứ phân vị phân tán.

Định nghĩa 1.4.3. Sắp xếp n quan sát từ nhỏ nhất đến lớn nhất rồi tách riêng nửa nhỏ nhất và nửa lớn nhất; trung vị $\tilde{x}$ được bao gồm trong cả hai nửa nếu n là số lẻ. Sau đó, tứ phân vị thấp hơn là mức trung bình của nửa nhỏ nhất và tứ phân vị cao hơn là trung bình của nửa lớn nhất. Một độ đo độ phân tán kháng lại giá trị ngoại lai là tứ phân vị phân tán f_s, được đưa ra bởi

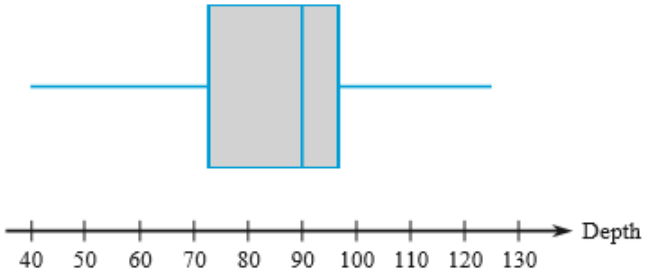
$$f_s = \text{tứ phân vị trên - tứ phân vị thấp}$$

Nói chung tứ phân vị phân tán không bị ảnh hưởng bởi vị trí của những quan sát trong 25% dữ liệu nhỏ nhất hoặc 25% dữ liệu lớn nhất. Do đó nó có khả năng chống ngoại lai. Các boxplot đơn giản nhất là dựa vào năm giá trị sau:
x_i nhỏ nhất tứ phân vị nhỏ trung vị tứ phân vị trên x_i lớn nhất
Đầu tiên, vẽ một thang đo ngang. Sau đó, đặt một hình chữ nhật ở trên trục này; cạnh trái của hình chữ nhật là vào tứ phân vị nhỏ và cạnh phải là tứ phân vị trên. Đặt một đoạn đường thẳng đứng hoặc một số biểu tượng khác bên trong hình chữ nhật tại địa điểm của trung vị; vị trí của các biểu tượng trung vị so với hai cạnh truyền tải thông tin về độ lệch giữa của 50% dữ liệu. Cuối cùng, vẽ "râu" ra từ hai đầu của hình chữ nhật tới và quan sát lớn nhất và nhỏ nhất. Một boxplot với một định hướng thẳng đứng cũng có thể được rút ra bằng cách làm thay đổi chiều thang đo trong quá trình xây dựng.

Ví dụ 1.16 Siêu âm được sử dụng để thu thập dữ liệu ăn mòn dựa vào độ dày của các tấm sàn của một bể chứa trên mặt đất được sử dụng để lưu trữ dầu thô ("Phân tích thống kê của UT về độ ăn mòn của dầu thô đối với tấm sàn bể chứa trên mặt đất Storage Tank," Materials Eval., 1994 : 846-849); mỗi quan sát là độ sâu hố lớn nhất trong đĩa, theo milli-in.

40 52 55 60 70 75 85 85 90 90 92 94 94 95 98 100 115 125 125

Ta có năm giá trị sau:
x_i nhỏ nhất=40; tứ phân vị nhỏ=72,5; trung vị=90; tứ phân vị trên=96,5; x_i lớn nhất=125



Hình 1.18: Biểu đồ hộp của dữ liệu trên.

Cạnh phải của hộp là gần trung vị hơn là cạnh bên trái, cho thấy sự lệch rất đáng kể giữa hai nửa của dữ liệu. Chiều rộng hộp (f_s) cũng là hợp lý lớn liên quan đến phạm vi của dữ liệu (khoảng cách giữa hai râu của hộp).
Hình 1.19 biểu diễn bảng kết quả của Minitab từ một yêu cầu về mô tả dữ liệu ăn mòn. Q1 và Q3 là các tứ phân vị thấp và trên; đây là tương tự như các tứ phân vị được tính toán dù hơi khác một chút. SE Mean là $s/\sqrt{n}$; đây sẽ là một số đại lượng quan trọng trong công việc tiếp theo của chúng ta liên quan đến suy luận về μ.

Variable	N	Mean	Median	TrMean	StDev	SE Mean
depth	19	86.32	90.00	86.76	23.32	5.35

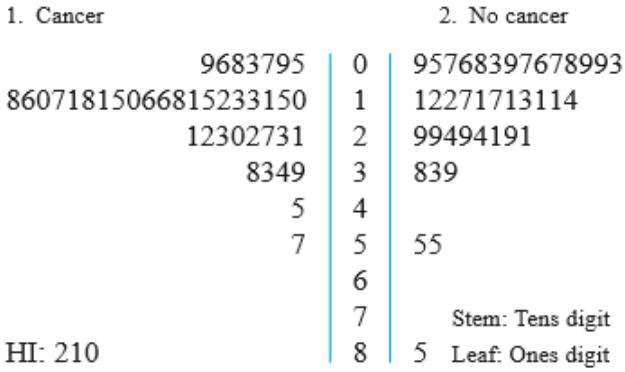
Variable	Minimum	Maximum	Q1	Q3
depth	40.00	125.00	70.00	98.00

Hình 1.19: Mô tả Minitab về dữ liệu độ sâu hố ăn mòn.

Biểu đồ hộp biểu diễn giá trị ngoại lai
Một boxplot có thể được lập lên để chỉ ra một cách rõ ràng sự hiện diện của giá trị ngoại lai. Nhiều thủ tục suy luận dựa trên giả định rằng tổng thể có phân bố chuẩn. Ngay cả một giá trị ngoại lai duy nhất trong mẫu cũng cảnh báo các nhà điều tra rằng các thủ tục như vậy có thể không đáng tin cậy và sự hiện diện của nhiều giá trị ngoại lai cũng truyền tải cùng một thông điệp.

Định nghĩa 1.4.4. Bất kỳ quan sát xa hơn $1,5f_s$ từ tứ phân vị gần nhất là một ngoại lại. Một ngoại lai là cực trị nếu nó là lớn hơn $3f_s$ tứ phân vị gần nhất và nó nhẹ (ôn hòa) hơn các giá trị khác.

So sánh các biểu đồ hộp
So sánh các biểu đồ hộp là một cách hiệu quả để tìmn sự tương đồng và khác biệt giữa hai hay nhiều bộ dữ liệu bao gồm các quan sát trên cùng một biến.
Ví dụ 1.17 Trong những năm gần đây, một số bằng chứng cho thấy rằng nồng độ radon trong nhà cao có thể liên quan đến sự phát triển của bệnh ung thư trẻ em, nhưng nhiều chuyên gia y tế vẫn còn hoài nghi. Một bài báo gần đây ("Nồng độ radon trong nhà và ung thư trẻ em," The Lancet, 1991: 1537-1538) trình bày các dữ liệu kèm theo nồng độ radon (Bq / m^3) trong hai mẫu khác nhau của các ngôi nhà. Mẫu đầu tiên bao gồm các nhà trong đó đứa trẻ đang cư trú được chẩn đoán ung thư. Các nhà trong mẫu thứ hai không có trường hợp ghi nhận ung thư trẻ em. Hình 1.20 trình bày biểu đồ gốc và lá của dữ liệu. Các giá trị tổng hợp như sau Các



Hình 1.20: Biểu đồ gốc lá của dữ liệu trên.

	$\bar{x}$	$\tilde{x}$	s	f_s
Cancer	22.8	16.0	31.7	11.0
No cancer	19.2	12.0	17.0	18.0

giá trị trung bình và trung vị cho rằng các mẫu ung thư tập trung một chút ở bên phải của mẫu không có ung thư trên thang đo. Tuy nhiên giá trị trung bình thổi phồng tầm quan trọng của sự thay đổi này, phần lớn là vì quan sát 210 trong mẫu bệnh ung thư. Các giá trị của s cho thấy sự biến đổi ở những mẫu ung thư hơn so với mẫu không có ung thư, nhưng ấn tượng này lại mâu thuẫn với tứ phân vị lây lan. Một lần nữa, quan sát thứ 210, một outlier cực trị, là thủ phạm. Hình 1.21 cho thấy một so sánh boxplot từ S-Plus. Hộp không có ung thư được kéo dài ra so với hộp ung thư và vị trí của các đường trung bình trong hai hộp cho thấy độ lệch nhiều hơn trong nửa giữa các mẫu không có ung thư hơn so với mẫu ung thư. Các giá trị ngoại lai được đại diện bởi các đoạn đường ngang và không có sự phân biệt giữa giá trị ngoại lai nhẹ và cực trị.

Bài tập 1.4

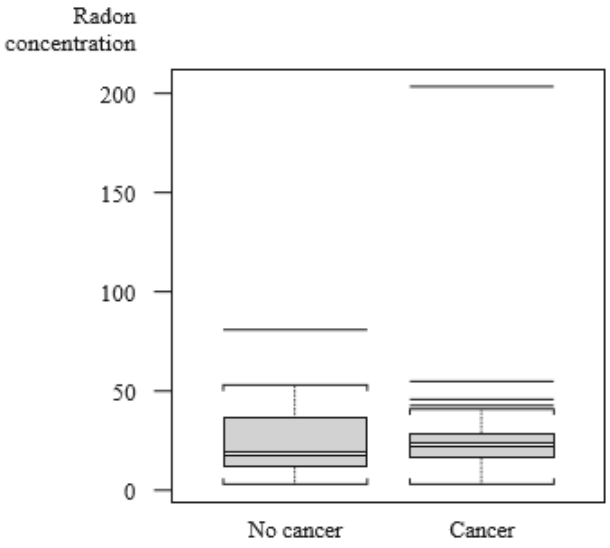
1.4.1 Cho bộ số liệu $29,5 \quad 49,3 \quad 30,6 \quad 28,2 \quad 28,0 \quad 26,3 \quad 33,9 \quad 29,4 \quad 23,5 \quad 31,6$
Tính các giá trị sau:

a. Cỡ mẫu.

b. Phương sai mẫu.

c. Độ lệch tiêu chuẩn mẫu.

1.4.2 Cho bộ số liệu:



Hình 1.21: Biểu đồ hộp của dữ liệu trên.

U:	6,0	5,0	11,0	33,0	4,0	5,0	80,0	18,0	35,0	17,.0	23,0
F:	4,0	14,0	11,0	9,0	9,0	8,.0	4,0	20,0	5,0	8,0	21,0
	9,5	3,5	2,5	0,5							

a. Xác định giá trị độ lệch tiêu chuẩn mẫu cho mỗi mẫu và so sánh.

b. Tính tứ phân vị cho mỗi mẫu và so sánh.

1.4.3 Cho bộ số liệu:

16	18	18	26	33	41	54	56	66	68	87	91	95	96
98	106	109	111	118	127	127	135	145	147	149	151	168	170
172	183	189	190	200	210	220	229	230	233	238	244	259	270

a. Xây dựng biểu đồ gốc và lá dựa trên lặp lại mỗi giá trị gốc hai lần, và nhận xét.

b. Xác định giá trị của phần tư và tứ phân vị.

c. Xây dựng một boxplot dựa trên tóm tắt năm số và nhận xét.

d. Một quan sát phải nhỏ hay lớn như thế nào sẽ được coi là giá trị ngoại lai, ngoại lai cực trị? Trong mẫu có giá trị ngoại lai hay không?

e. Quan sát hiện tại là 403 có thể giảm nhiều nhất là bao nhiêu để không ảnh hưởng đến f_s

BÀI TẬP TỔNG HỢP

TH 1.1 Cho bộ dữ liệu:

553	553	553	559	559	559	559	561	561	561	561
561	561	568	568	570	570	570	578	578	578	579
579	579	588	588	588	598	598	598	622	622	638
638	638	639	63.9	639	647	647	647	651	651	651
653	653	653	653	674	674	674	674	687	687	687
687	690	704	704	712	712	712	730	730	731	731
746	746	746	746	793	793	793	793	830	830	830

Tổng hợp dữ liệu rồi tổng kết và mô tả dữ liệu.

TH 1.2 Cho bộ số liệu sau:

Loại 1	350	350	350	358	370	370	370	371
	371	372	372	384	391	391	392	
Loại 2	350	354	359	363	365	368	369	371
	273	374	376	380	383	388	392	
Loại 3	350	361	362	364	364	365	366	371
	377	377	377	379	380	380	392	

a. Xây dựng một bảng so sánh boxplot, nhận xét về những điểm tương đồng và sự khác biệt.

b. Xây dựng một dotplot so sánh (một dotplot cho mỗi mẫu với một khoảng chia chung). Nhận xét về những điểm tương đồng và khác biệt.

TH 1.3 Cho bộ số liệu sau:

Khoảng	6-8	8-10	10-12	12-14	14-16	16-18	18-20
Tần số	6	23	30	35	32	48	42
Khoảng	20-22	22-24	24-26	26-28	28-30	30-35	35-40
Tần số	40	28	27	26	14	27	11

a. Vẽ biểu đồ tương ứng với các tần số này.

b. Tỷ lệ những các quan sát dưới 20 là bao nhiêu? Tính tỷ lệ những quan sát có độ dài ít nhất là 30?

c. Áng chừng giá trị của phần trăm thứ 90 của phân phối các quan sát của mẫu?

d. Áng chừng giá trị trung vị của mẫu.

Tài liệu tham khảo