

Chương 1

Giới thiệu về dữ liệu

1.1 Ví dụ tình huống: sử dụng stent để ngăn ngừa đột quỵ

1.2 Giới thiệu về dữ liệu

1.3 Nguyên tắc và cách lấy mẫu

1.4 Thử nghiệm

Các nhà khoa học tìm cách trả lời các câu hỏi bằng các phương pháp khoa học và quan sát cẩn thận. Những quan sát này – được thu thập từ các ghi chú thực địa, khảo sát và nghiên cứu – tạo thành tiền đề của một cuộc điều tra thống kê và được gọi là **dữ liệu**. Thống kê là nghiên cứu về cách tốt nhất để thu thập, phân tích và rút ra kết luận từ dữ liệu, và trong chương đầu tiên này, chúng tôi tập trung vào các thuộc tính của dữ liệu và về việc thu thập dữ liệu.



Để xem video, slides và các tài nguyên khác, vui lòng truy cập
www.openintro.org/os

1.1 Ví dụ tình huống: sử dụng stent để ngăn ngừa đột quy

Phần 1.1 giới thiệu một thách thức kinh điển trong thống kê: đánh giá hiệu quả của một phương pháp điều trị y tế. Các thuật ngữ trong phần này, và cũng như là phần lớn của chương này, tất cả sẽ được liệt kê trong chương này. Mục đích chỉ đơn giản là để hiểu được vai trò của số liệu thống kê trong thực tế.

Trong phần này, chúng ta sẽ xem xét một thử nghiệm nghiên cứu hiệu quả của stent trong điều trị bệnh nhân có nguy cơ đột quy. Stent là các thiết bị được đặt bên trong các mạch máu hỗ trợ phục hồi bệnh nhân sau các biến cố về tim và làm giảm nguy cơ đau tim hoặc tử vong thêm. Nhiều bác sĩ đã hy vọng rằng sẽ có những lợi ích tương tự cho bệnh nhân có nguy cơ đột quy. Chúng tôi bắt đầu bằng cách viết câu hỏi chính mà các nhà nghiên cứu hy vọng sẽ trả lời:

Việc sử dụng stent có làm giảm nguy cơ đột quy không?

Các nhà nghiên cứu đặt câu hỏi này đã tiến hành một thí nghiệm với 451 bệnh nhân có nguy cơ. Mỗi bệnh nhân tình nguyện được chỉ định ngẫu nhiên vào một trong hai nhóm:

Nhóm điều trị - Treatment group. Bệnh nhân trong nhóm điều trị được đặt stent và quản lý y tế. Quản lý y tế bao gồm thuốc, quản lý các yếu tố nguy cơ và giúp thay đổi lối sống.

Nhóm đối chứng - Control group. Bệnh nhân trong nhóm đối chứng được điều trị y tế tương tự như nhóm điều trị, nhưng họ không nhận được điều trị đặt stent.

Các nhà nghiên cứu đã chỉ định ngẫu nhiên 224 bệnh nhân vào nhóm điều trị và 227 bệnh nhân vào nhóm đối chứng. Trong nghiên cứu này, nhóm đối chứng cung cấp một điểm tham chiếu, từ đó chúng ta có thể đo lường tác động y tế của stent trong nhóm điều trị.

Các nhà nghiên cứu đã đánh giá tác dụng của stent tại hai thời điểm: 30 ngày sau khi tham gia nghiên cứu và 365 ngày sau tham gia nghiên cứu. Kết quả của 5 bệnh nhân được tóm tắt trong Hình 1.1. Kết quả của bệnh nhân được ghi lại là "đột quy" hoặc "không có biến cố", thể hiện liệu bệnh nhân có bị đột quy vào cuối khoảng thời gian hay không.

Bệnh nhân	Nhóm	0-30 ngày	0-365 ngày
1	Điều trị	Không có biến cố	Không có biến cố
2	Điều trị	Đột quy	Đột quy
3	Điều trị	Không có biến cố	Không có biến cố
...	...	...	...
450	Đối chứng	Không có biến cố	Không có biến cố
451	Đối chứng	Không có biến cố	Không có biến cố

Hình 1.1: Kết quả của năm bệnh nhân từ nghiên cứu stent.

Xem xét dữ liệu từ từng bệnh nhân riêng lẻ sẽ là một con đường dài và rườm rà để trả lời câu hỏi nghiên cứu ban đầu. Thay vào đó, thực hiện phân tích dữ liệu thống kê cho phép chúng ta xem xét tất cả dữ liệu cùng một lúc. Hình 1.2 mô tả dữ liệu thô theo cách hữu ích hơn. Trong bảng này, chúng ta có thể nhanh chóng thấy những gì đã xảy ra trong toàn bộ nghiên cứu. Ví dụ, để xác định số lượng bệnh nhân trong nhóm điều trị bị đột quy trong vòng 30 ngày, chúng tôi nhìn vào phía bên trái của bảng tại giao điểm của điều trị và đột quy: số 33.

	0-30 ngày		0-365 ngày	
	Đột quy	Không có biến cố	Đột quy	Không có biến cố
Điều trị	33	191	45	179
Đối chứng	13	214	28	199
Tổng cộng	46	405	73	378

Hình 1.2: Thống kê mô tả cho nghiên cứu stent.

BÀI TẬP THỰC HÀNH 1.1

Trong số 224 bệnh nhân trong nhóm điều trị, 45 người bị đột quỵ vào cuối năm đầu tiên. Sử dụng hai con số này, tính tỷ lệ bệnh nhân trong nhóm điều trị bị đột quỵ vào cuối năm đầu tiên. (Lưu ý: câu trả lời cho tất cả các bài tập thực hành có trong phần Chú thích cuối trang)¹

Chúng ta có thể tính toán số liệu thống kê từ bảng mô tả. **Thống kê mô tả** là cách tóm tắt một lượng lớn dữ liệu. Ví dụ, kết quả chính của nghiên cứu sau 1 năm có thể được mô tả bằng hai số trong thống kê mô tả: tỷ lệ người bị đột quỵ trong các nhóm điều trị và nhóm đối chứng.

Tỷ lệ bị đột quỵ ở nhóm điều trị (stent): $45/224 = 0.20 = 20\%$.

Tỷ lệ bị đột quỵ ở nhóm đối chứng: $28/227 = 0.12 = 12\%$.

Hai thống kê tóm tắt này rất hữu ích trong việc tìm kiếm sự khác biệt trong các nhóm và chúng tôi rất ngạc nhiên: thêm 8% bệnh nhân trong nhóm điều trị bị đột quỵ! Điều này rất quan trọng vì hai lý do chính: Đầu tiên, nó trái ngược với những gì các bác sĩ mong đợi, đó là stent sẽ làm giảm tỷ lệ đột quỵ. Thứ hai, nó dẫn đến một câu hỏi thống kê: dữ liệu có cho thấy sự khác biệt "thực sự" giữa các nhóm không?

Câu hỏi thứ hai này rất tinh tế. Giả sử bạn tung đồng xu 100 lần. Mặc dù cơ hội một đồng xu xuất hiện trong bất kỳ lần tung đồng xu nào là 50%, nhưng có lẽ chúng ta sẽ không quan sát được chính xác 50 đồng xu. Loại biến động này là một phần của hầu hết mọi loại quy trình tạo dữ liệu. Có thể sự khác biệt 8% trong nghiên cứu stent là do sự thay đổi tự nhiên này. Tuy nhiên, sự khác biệt chúng ta quan sát được càng lớn (đối với một cỡ mẫu cụ thể), thì càng khó để tin rằng sự khác biệt là do tình cờ. Vì vậy, những gì chúng tôi thực sự thắc mắc là như sau: liệu sự khác biệt quá lớn đến mức mà chúng ta nên bác bỏ ý tưởng rằng nó là do ngẫu nhiên?

Mặc dù chúng tôi chưa có các công cụ thống kê để tự giải quyết đầy đủ câu hỏi này, chúng tôi có thể hiểu được kết luận của phân tích được công bố: có bằng chứng thuyết phục về tác hại của việc đặt stent trong nghiên cứu này đối với bệnh nhân đột quỵ.

Hãy cẩn thận: Đừng khái quát hóa kết quả của nghiên cứu này cho tất cả bệnh nhân và tất cả các loại stent. Nghiên cứu này xem xét những bệnh nhân có đặc điểm rất cụ thể tình nguyện tham gia nghiên cứu này và những người có thể không đại diện cho tất cả các bệnh nhân đột quỵ. Ngoài ra, có nhiều loại stent và nghiên cứu này chỉ xem xét stent Wingspan tự mở rộng (Boston Scientific). Tuy nhiên, nghiên cứu này để lại cho chúng ta một bài học quan trọng: **chúng ta luôn sẵn sàng đón nhận những kết quả bất ngờ.**

¹ Tỷ lệ 224 bệnh nhân bị đột quỵ trong vòng 365 ngày: $45/224 = 0,20$.

Bài tập

1.1 Nghiên cứu về đau nửa đầu và châm cứu, Phần I. Đau nửa đầu là một loại đau đầu đặc biệt đau đớn, mà bệnh nhân đôi khi muốn điều trị bằng châm cứu. Để xác định xem châm cứu có làm giảm đau nửa đầu hay không, các nhà nghiên cứu đã tiến hành một nghiên cứu thử nghiệm ngẫu nhiên có đối chứng, trong đó 89 phụ nữ được chẩn đoán đau nửa đầu được chỉ định ngẫu nhiên vào một trong hai nhóm: nhóm điều trị hoặc nhóm đối chứng. 43 bệnh nhân trong nhóm điều trị được châm cứu để điều trị chứng đau nửa đầu. 46 bệnh nhân trong nhóm đối chứng được châm cứu giả (đưa kim vào các vị trí không phải huyệt vị). 24 giờ sau khi bệnh nhân được châm cứu, họ được hỏi liệu họ có hết đau hay không. Kết quả được tóm tắt trong bảng dự phòng dưới đây.²

Nhóm	Hết đau đầu		
	Có	Không	Tổng
	Điều trị	Đối chứng	Tổng
Điều trị	10	33	43
Đối chứng	2	44	46
Tổng	12	77	89



Hình từ tài liệu ban đầu hiển thị khu vực châm cứu đúng (M) so với khu vực châm cứu giả (S) được sử dụng trong điều trị chứng đau nửa đầu

- Bao nhiêu phần trăm bệnh nhân trong nhóm điều trị đã hết đau sau 24 giờ sau khi châm cứu?
- Bao nhiêu phần trăm đã hết đau trong nhóm đối chứng?
- Nhóm nào có tỷ lệ phần trăm cao hơn của bệnh nhân hết đau sau 24 giờ sau khi châm cứu?
- Những phát hiện của bạn cho đến nay cho thấy châm cứu có thể là một phương pháp điều trị hiệu quả cho chứng đau nửa đầu ở tất cả những người bị đau nửa đầu. Tuy nhiên, đây không phải là kết luận duy nhất. Một lời giải thích khác cho sự khác biệt về tỷ lệ phần trăm bệnh nhân hết đau sau 24 giờ sau khi được châm cứu ở hai nhóm có thể là gì?

1.2 Nghiên cứu về viêm xoang và kháng sinh, Phần I. Các nhà nghiên cứu đo lường tác dụng của điều trị kháng sinh đối với viêm xoang cấp tính so với các phương pháp điều trị triệu chứng đã chỉ định ngẫu nhiên 166 người bệnh được chẩn đoán viêm xoang cấp tính vào một trong hai nhóm: điều trị hoặc đối chứng. Những người tham gia nghiên cứu đã nhận được một liệu trình amoxicillin (một loại kháng sinh) trong 10 ngày hoặc một loại giả dược giống về hình thức và hương vị. Giả dược bao gồm các thành phần giúp điều trị triệu chứng như acetaminophen, thuốc thông mũi, v.v. Vào cuối khoảng thời gian 10 ngày, bệnh nhân được hỏi liệu họ có cải thiện các triệu chứng hay không. Các câu trả lời được tóm tắt dưới đây.³

Nhóm	Cải thiện các triệu chứng xoang		
	Có	Không	Tổng
	Điều trị	Đối chứng	Tổng
Điều trị	66	19	85
Đối chứng	65	16	81
Tổng	131	35	166

- Bao nhiêu phần trăm bệnh nhân trong nhóm điều trị có triệu chứng cải thiện?
- Bao nhiêu phần trăm đã cải thiện các triệu chứng trong nhóm đối chứng?
- Ở nhóm nào có tỷ lệ bệnh nhân cải thiện các triệu chứng cao hơn?
- Những phát hiện của bạn cho đến nay có thể cho thấy sự khác biệt thực sự về hiệu quả của các phương pháp điều trị bằng kháng sinh và giả dược để cải thiện các triệu chứng của viêm xoang. Tuy nhiên, đây không phải là kết luận duy nhất có thể được rút ra dựa trên những phát hiện của bạn cho đến nay. Một lời giải thích khác cho sự khác biệt quan sát được giữa tỷ lệ bệnh nhân trong nhóm điều trị kháng sinh và giả dược được cải thiện các triệu chứng viêm xoang là gì?

² G. Allais et al. "Ear acupuncture in the treatment of migraine attacks: a randomized trial on the efficacy of appropriate versus inappropriate acupoints". In: *Neurological Sci.* 32.1 (2011), pp. 173–175.

³ J.M. Garbutt et al. "Amoxicillin for Acute Rhinosinusitis: A Randomized Controlled Trial". In: *JAMA: The Journal of the American Medical Association* 307.7 (2012), pp. 685–692.

1.2 Thông tin cơ bản về dữ liệu

Tổ chức và mô tả dữ liệu một cách hiệu quả là bước đầu tiên trong hầu hết các phân tích. Phần này sẽ giới thiệu *ma trận dữ liệu* để giúp tổ chức thông tin, cũng như giải thích một số thuật ngữ liên quan đến các dạng dữ liệu khác nhau sẽ được sử dụng trong toàn bộ cuốn sách.

1.2.1 Quan sát, biến và ma trận dữ liệu

Hình 1.3 hiển thị các hàng 1, 2, 3 và 50 của bộ dữ liệu cho 50 khoản vay được lấy mẫu ngẫu nhiên thông qua Lending Club, một công ty cho vay ngang hàng. Những quan sát này sẽ được gọi là tập dữ liệu **loan50**.

Mỗi hàng trong bảng đại diện cho một khoản vay duy nhất. Tên chính thức cho một hàng là một **trường hợp** hoặc **đơn vị quan sát**. Các cột đại diện cho các đặc điểm, được gọi là **biến**, cho mỗi khoản vay. Ví dụ: hàng đầu tiên đại diện cho khoản vay 22.000 đô la với lãi suất 10.90%, trong đó người vay có trụ sở tại New Jersey (NJ) và có thu nhập là 59.000 đô la.

BÀI TẬP THỰC HÀNH 1.2

Mức của khoản vay đầu tiên trong Hình 1.3 là bao nhiêu? Và tình trạng sở hữu nhà của người vay cho khoản vay đầu tiên đó là gì? Đối với các Bài tập Thực hành có hướng dẫn này, bạn có thể kiểm tra câu trả lời của mình trong chú thích cuối trang.⁴

Trong thực tế, điều đặc biệt quan trọng là phải đặt câu hỏi làm rõ để đảm bảo các khía cạnh quan trọng của dữ liệu được hiểu. Ví dụ, điều quan trọng là phải chắc chắn rằng chúng ta biết ý nghĩa của mỗi biến và các đơn vị đo lường. Mô tả các biến **loan50** được đưa ra trong Hình 1.4.

loan_amount	interest_rate	term	grade	state	total_income	homeownership
22000	10.9	60	B	NJ	59000	Thuê
6000	9.92	36	B	CA	60000	Thuê
25000	26.3	36	E	SC	75000	Thế chấp
...	...	...	...	...	...	...
15000	6.08	36	A	TX	77500	Thế chấp

Hình 1.3: Bốn hàng từ ma trận dữ liệu **loan50**.

Biến số	Mô tả biến số
loan_amount	Số tiền vay nhận được, tính bằng đô la Mỹ.
interest_rate	Lãi suất khoản vay, tính theo tỷ lệ phần trăm hàng năm.
term	Thời gian vay, luôn được đặt ở đơn vị tháng.
grade	Xếp hạng khoản vay (A-G), biểu thị chất lượng khoản vay và khả năng hoàn trả.
state	Bang của Mỹ nơi người vay sinh sống.
total_income	Tổng thu nhập của người vay, bao gồm cả thu nhập phụ, tính bằng đô la Mỹ.
homeownership	Chỉ định tình trạng sở hữu nhà: sở hữu, sở hữu có thế chấp hoặc thuê.

Hình 1.4: Các biến và mô tả của chúng cho tập dữ liệu **loan50**.

Dữ liệu trong Hình 1.3 đại diện cho **ma trận dữ liệu**, đây là một cách thuận tiện và phổ biến để sắp xếp dữ liệu, đặc biệt nếu thu thập dữ liệu trong bảng tính. Mỗi hàng của ma trận dữ liệu tương ứng với một trường hợp duy nhất (đơn vị quan sát) và mỗi cột tương ứng với một biến.

⁴ Cấp khoản vay là B, và người vay thuê nơi ở của họ.

Khi ghi dữ liệu, hãy sử dụng ma trận dữ liệu trừ khi bạn có lý do chính đáng để sử dụng một cấu trúc khác. Cấu trúc này cho phép thêm các trường hợp mới dưới dạng hàng hoặc biến mới dưới dạng cột mới.

**BÀI TẬP THỰC HÀNH 1.3**

Các điểm số cho các bài tập, câu đố và bài kiểm tra trong một khóa học thường được ghi lại trong sổ điểm có dạng ma trận dữ liệu. Làm thế nào bạn có thể sắp xếp dữ liệu điểm số bằng ma trận dữ liệu?⁵

**BÀI TẬP THỰC HÀNH 1.4**

Chúng tôi xem xét dữ liệu của 3.142 quận tại Hoa Kỳ, bao gồm tên mỗi quận, tiểu bang nơi quận đó tọa lạc, dân số từng quận vào năm 2017, sự thay đổi dân số từ năm 2010 đến năm 2017, tỷ lệ nghèo đói, và sáu đặc điểm bổ sung khác. Những dữ liệu này có thể được tổ chức trong ma trận dữ liệu theo cách sau: mỗi hàng sẽ đại diện cho một quận, trong khi các cột sẽ chứa thông tin về tên quận, tiểu bang, dân số năm 2017, thay đổi dân số, tỷ lệ nghèo đói và các đặc điểm bổ sung?⁶

Dữ liệu được mô tả trong Thực hành có hướng dẫn 1.4 đại diện cho tập dữ liệu **county**, được hiển thị dưới dạng ma trận dữ liệu trong Hình 1.5. Các biến được tóm tắt trong Hình 1.6.

⁵ Có nhiều chiến lược có thể được thực hiện. Một chiến lược phổ biến là để mỗi học sinh được đại diện bằng một hàng, sau đó thêm một cột cho mỗi bài tập, bài kiểm tra hoặc bài kiểm tra. Theo thiết lập này, có thể dễ dàng xem lại một dòng duy nhất để hiểu lịch sử điểm của học sinh. Cũng nên có các cột để bao gồm thông tin học sinh, chẳng hạn như một cột để liệt kê tên học sinh.

⁶ Mỗi quận có thể được xem như một trường hợp, và có mười một mẫu thông tin được ghi lại cho mỗi trường hợp. Một bảng có 3.142 hàng và 11 cột có thể chứa các dữ liệu này, trong đó mỗi hàng đại diện cho một quận và mỗi cột đại diện cho một phần thông tin cụ thể.

name	state	pop	pop_change	poverty	homeownership	multi.unit	unemp_rate	metro	median_edu	median_hh_income
Autauga	Alabama	55504	1.48	13.7	77.5	7.2	3.86	yes	some_college	55317
Baldwin	Alabama	212628	9.19	11.8	76.7	22.6	3.99	yes	some_college	52562
Barbour	Alabama	25270	-6.22	27.2	68	11.1	5.9	no	hs_diploma	33368
Bibb	Alabama	22668	0.73	15.2	82.9	6.6	4.39	yes	hs_diploma	43404
Blount	Alabama	58013	0.68	15.6	82	3.7	4.02	yes	hs_diploma	47412
Bullock	Alabama	10309	-2.28	28.5	76.9	9.9	4.93	no	hs_diploma	29655
Butler	Alabama	19825	-2.69	24.4	69	13.7	5.49	no	hs_diploma	36326
Calhoun	Alabama	114728	-1.51	18.6	70.7	14.3	4.93	yes	some_college	43686
Chambers	Alabama	33713	-1.2	18.8	71.4	8.7	4.08	no	hs_diploma	37342
Cherokee	Alabama	25857	-0.6	16.1	77.5	4.3	4.05	no	hs_diploma	40041
...	...	...	...	...	...	...	...	...	...	...
Weston	Wyoming	6927	-2.93	14.4	77.9	6.5	3.98	no	some_college	59605

Hình 1.5: Mười một hàng từ tập dữ liệu quận.

Biến số	Mô tả biến số
name	Tên của quận.
state	Bang nơi quận thuộc về hoặc quận của thủ đô Washington, D.C.
pop	Dân số năm 2017.
pop_change	Tỷ lệ phần trăm thay đổi dân số từ năm 2010 đến 2017. Ví dụ, giá trị 1.48 trong hàng đầu tiên có nghĩa là dân số quận này tăng 1.48% từ năm 2010 đến 2017.
poverty	Tỷ lệ phần trăm dân số sống dưới mức nghèo.
homeownership	Tỷ lệ phần trăm dân số sở hữu nhà hoặc sống cùng chủ nhà (ví dụ: con cái sống với cha mẹ sở hữu nhà).
multi.unit	Tỷ lệ phần trăm các đơn vị nhà ở là dạng nhiều hộ gia đình (ví dụ: căn hộ chung cư).
unemp_rate	Tỷ lệ thất nghiệp, tính theo phần trăm.
metro	Chỉ định quận có thuộc khu vực đô thị hay không.
median_edu	Trình độ giáo dục trung bình, có thể là: below_hs (dưới trung học), hs_diploma (tốt nghiệp trung học), some_college (một phần đại học), hoặc bachelors (cử nhân).
median_hh_income	Thu nhập hộ gia đình trung bình của quận, tính theo tổng thu nhập của tất cả thành viên từ 15 tuổi trở lên.

Hình 1.6: Các biến và mô tả của chúng cho tập dữ liệu county.

1.2.2 Các loại biến

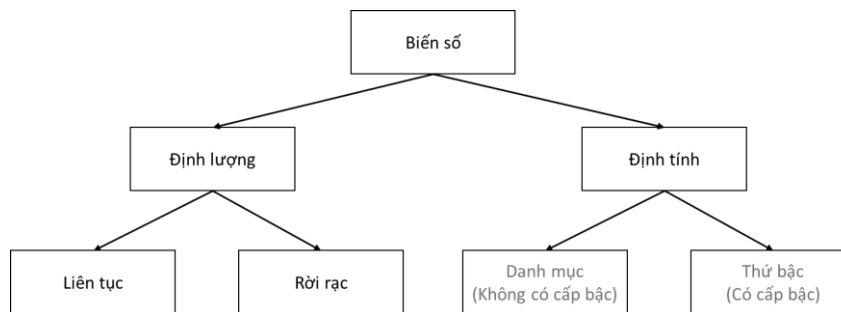
Kiểm tra các biến `unemp_rate`, `pop`, `state`, và `median_edu` trong tập dữ liệu `county`.

Trước tiên, hãy xem xét `unemp_rate`, được cho là một **biến định lượng** vì nó có thể lấy một loạt các giá trị số và có thể cộng, trừ hoặc lấy trung bình với các giá trị đó. Mặt khác, chúng ta sẽ không phân loại mã vùng điện thoại báo cáo thành biến định lượng vì trung bình, tổng và sự khác biệt của các mã vùng không có bất kỳ ý nghĩa rõ ràng nào.

Biến `pop` cũng là biến định lượng, mặc dù nó có vẻ hơi khác so với `unemp_rate`. Biến dân số này chỉ có thể lấy số không âm (0, 1, 2, ...). Vì lý do này, biến số dân được cho là **biến rời rạc** vì nó chỉ có thể lấy các giá trị số với các bước nhảy. Mặt khác, biến tỷ lệ thất nghiệp được cho là biến liên tục.

Biến `state` có thể chiếm tới 51 giá trị sau khi tính đến Washington, DC: AL, AK, ..., và WY. Bởi vì bản thân các câu trả lời là các danh mục, tiểu bang được gọi là **biến định tính** và các giá trị có thể được gọi là cấp độ của biến.

Cuối cùng, hãy xem xét biến `median_edu`, mô tả trình độ học vấn trung bình của cư dân trong các quận, bao gồm các giá trị như dưới cấp Trung học, Trung học, một số trường cao đẳng và cử nhân ở mỗi quận. Biến này dường như là một biến hỗn hợp: nó là biến phân loại nhưng có thứ tự tự nhiên giữa các mức. Biến có những thuộc tính này được gọi là **biến thứ bậc**, trong khi biến phân loại thông thường không có thứ tự đặc biệt được gọi là **biến định danh**. Để đơn giản hóa các phân tích, mọi biến thứ bậc trong cuốn sách này sẽ được coi như là biến định danh (không có thứ tự).



Hình 1.7: Phân tích các biến thành các loại tương ứng của chúng.

VÍ DỤ 1.5

Dữ liệu được thu thập về sinh viên trong một khóa học thống kê. Ba biến được ghi lại cho mỗi sinh viên: số lượng anh chị em, chiều cao của sinh viên và liệu trước đó sinh viên có tham gia khóa học thống kê hay không. Phân loại từng biến dưới dạng biến liên tục, biến rời rạc hoặc biến phân loại.

Số lượng anh chị em và chiều cao của học sinh đại diện cho các biến định lượng. Bởi vì số lượng anh chị em là một số, nó là biến rời rạc. Chiều cao thay đổi liên tục, vì vậy nó là một biến liên tục. Biến cuối cùng phân loại sinh viên thành hai loại – những người có và những người chưa tham gia khóa học thống kê – điều này làm cho biến này trở thành biến phân loại.

BÀI TẬP THỰC HÀNH 1.6

Một nghiên cứu đang đánh giá hiệu quả của một loại thuốc mới trong điều trị chứng đau nửa đầu. Một biến `group` được sử dụng để chỉ ra nhóm thử nghiệm cho mỗi bệnh nhân: điều trị hoặc đối chứng. Biến số chứng đau nửa đầu `num_migraines` đại diện cho số lượng chứng đau nửa đầu mà bệnh nhân trải qua trong khoảng thời gian 3 tháng.

Phân loại mỗi biến dưới dạng biến định lượng hoặc phân loại.⁷

⁷ Biến `group` chỉ có thể nhận một trong hai tên nhóm, do đó nó được coi là biến phân loại. Trong khi đó, biến `num_migraines` mô tả số lượng các lần bị chứng đau nửa đầu, là một kết quả hợp lý cho phép sử dụng số học cơ bản. Vì biến này đại diện cho một số lượng cụ thể, `num_migraines` là một biến rời rạc.

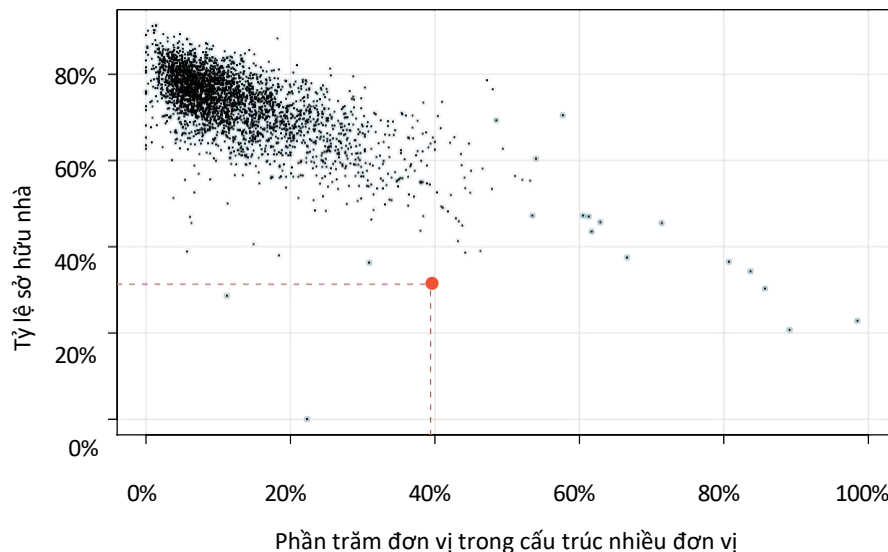
1.2.3 Mối quan hệ giữa các biến

Các nhà nghiên cứu thường tiến hành nhiều phân tích để tìm kiếm mối quan hệ giữa hai hoặc nhiều biến. Một nhà khoa học xã hội có thể muốn trả lời một số câu hỏi sau:

- (1) Nếu tỷ lệ sở hữu nhà ở một quận thấp hơn mức trung bình quốc gia, thì tỷ lệ phần trăm các cấu trúc nhiều căn hộ (ví dụ: chung cư) trong quận đó có xu hướng cao hơn hay thấp hơn mức trung bình quốc gia?
- (2) Liệu một mức tăng dân số cao hơn trung bình ở một quận có xu hướng liên quan đến các quận có thu nhập hộ gia đình trung bình cao hơn hay thấp hơn?
- (3) Trình độ học vấn trung bình có vai trò dự đoán mức thu nhập hộ gia đình trung bình của các quận ở Hoa Kỳ hữu ích đến mức độ nào?

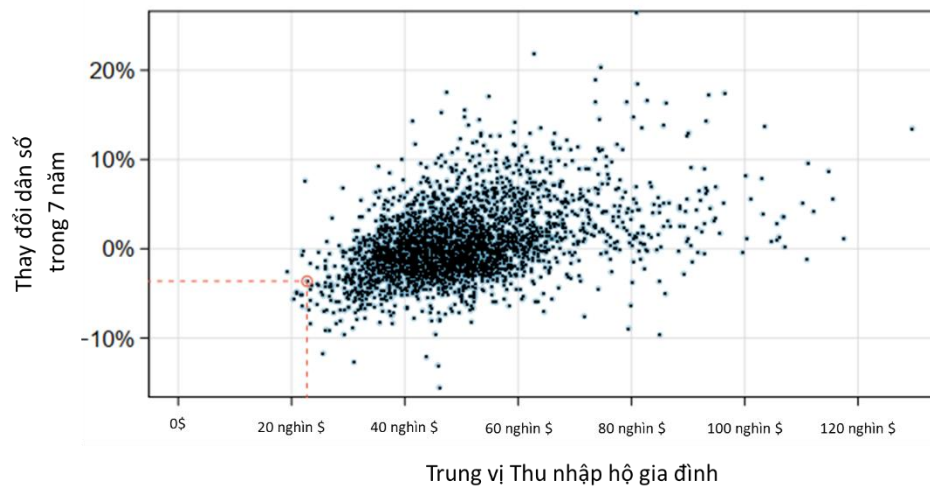
Để trả lời những câu hỏi này, dữ liệu phải được thu thập, chẳng hạn như bộ dữ liệu **county** được hiển thị trong Hình 1.5. Việc mô tả số liệu thành các bảng thống kê có thể cung cấp thông tin chi tiết cho mỗi câu hỏi trong ba câu hỏi về các quận. Ngoài ra, đồ thị có thể được sử dụng để khám phá dữ liệu một cách trực quan.

Biểu đồ phân tán là một loại biểu đồ được sử dụng để nghiên cứu mối quan hệ giữa hai biến định lượng. Hình 1.8 so sánh các biến **homeownership** và **multi_unit**, là tỷ lệ phần trăm của các đơn vị trong các cấu trúc nhiều căn hộ (ví dụ: chung cư). Mỗi điểm trên lô đất đại diện cho một quận. Ví dụ, điểm đỏ trên biểu đồ được đánh dấu tương ứng với Quận 413 trong bộ dữ liệu **county**: Quận Chat-tahoochee, Georgia, nơi có 39,4% đơn vị trong các cấu trúc nhiều đơn vị và tỷ lệ sở hữu nhà là 31,3%. Biểu đồ phân tán cho thấy mối quan hệ giữa hai biến: các quận có tỷ lệ nhiều đơn vị cao hơn có xu hướng có tỷ lệ sở hữu nhà thấp hơn. Chúng ta có thể suy nghĩ về lý do tại sao mối quan hệ này tồn tại và điều tra từng ý tưởng để xác định đâu là lời giải thích hợp lý nhất.



Hình 1.8: Biểu đồ phân tán quyền sở hữu nhà so với tỷ lệ phần trăm của các căn hộ trong các cấu trúc nhiều căn hộ cho các quận của Hoa Kỳ. Dấu chấm đỏ được đánh dấu đại diện cho Quận Chatta- hoochee, Georgia, nơi có tỷ lệ nhiều đơn vị là 39,4% và tỷ lệ tài là 31,3%.

Tỷ lệ cấu trúc nhiều căn hộ và sở hữu nhà được cho là có liên quan vì biểu đồ cho thấy một mô hình rõ ràng. Khi hai biến thể hiện một số kết nối với nhau, chúng được gọi là biến **liên kết**. Các biến liên kết cũng có thể được gọi là biến **phụ thuộc** và ngược lại.



Hình 1.9: Biểu đồ phân tán mối quan hệ giữa biến `pop_change` so với biến `median_hh_income`. Quận Owsley của bang Kentucky, được đánh dấu, đã mất 3,63% dân số từ năm 2010 đến năm 2017 và có trung vị thu nhập hộ gia đình là 22,736 đô la.

G

BÀI TẬP THỰC HÀNH 1.7

Kiểm tra các biến trong tập dữ liệu `loan50`, được mô tả trong Hình 1.4 trên trang 12. Tạo hai câu hỏi về mối quan hệ có thể có giữa các biến trong `loan50` mà bạn quan tâm.⁸

VÍ DỤ 1.8

Ví dụ này xem xét mối quan hệ giữa sự thay đổi dân số của một quận từ năm 2010 đến năm 2017 và trung vị thu nhập hộ gia đình, được hiển thị dưới dạng biểu đồ phân tán trong Hình 1.9. Đây có phải là những biến liên kết?

E

Trung vị thu nhập hộ gia đình của một quận càng lớn thì sự gia tăng dân số của quận đó càng cao. Mặc dù xu hướng này không đúng với mọi quận, nhưng xu hướng trong biểu đồ là hiển nhiên. Vì có một số mối quan hệ giữa các biến, chúng có tương quan với nhau.

Bởi vì có xu hướng giảm trong Hình 1.8 – các quận có nhiều đơn vị trong các cấu trúc nhiều căn hộ có liên quan đến quyền sở hữu nhà thấp hơn – các biến này được cho là biến **tương quan nghịch**. Một **tương quan thuận** được thể hiện trong mối quan hệ giữa biến `median_hh_income` và biến `pop_change` trong Hình 1.9, nơi các quận có trung vị thu nhập hộ gia đình cao thì có xu hướng có tỷ lệ tăng dân số cao.

Nếu hai biến không tương quan, thì chúng được gọi là **độc lập**. Nghĩa là, hai biến là độc lập nếu không có mối quan hệ rõ ràng giữa hai biến.

TƯƠNG QUAN HOẶC ĐỘC LẬP, KHÔNG PHẢI CẢ HAI

Một cặp biến có tương quan theo một cách nào đó (liên kết) hoặc không (độc lập). Không có cặp biến nào vừa tương quan vừa độc lập.

⁸ Hai câu hỏi ví dụ: (1) Mối quan hệ giữa số tiền vay và tổng thu nhập là gì? (2) Nếu thu nhập của ai đó cao hơn mức trung bình, lãi suất của họ có xu hướng cao hơn hay thấp hơn mức trung bình?

1.2.4 Biến giải thích và biến đáp ứng

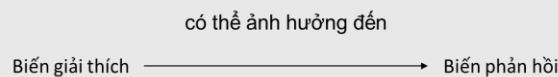
Khi chúng ta đặt câu hỏi về mối quan hệ giữa hai biến, đôi khi chúng ta cũng muốn xác định xem sự thay đổi trong một biến có gây ra sự thay đổi trong biến kia hay không. Hãy xem xét cách diễn đạt lại sau đây của một câu hỏi trước đó về bộ dữ liệu **county** :

Nếu có sự gia tăng trung vị thu nhập hộ gia đình trong một quận, điều này có thúc đẩy sự gia tăng dân số của quận đó không?

Trong câu hỏi này, chúng tôi đang hỏi liệu một biến có ảnh hưởng đến biến khác hay không. Nếu đây là niềm tin cơ bản của chúng tôi, thì trung vị thu nhập hộ gia đình là biến giải thích và thay đổi dân số là biến đáp ứng trong mối quan hệ giả định.⁹

CÁC BIẾN GIẢI THÍCH VÀ PHẢN HỒI

Khi chúng ta nghi ngờ một biến có mối quan hệ nhân quả đến một biến khác, chúng ta gọi biến đầu tiên là biến giải thích và biến thứ hai là biến phản hồi



Đối với nhiều cặp biến, không có mối quan hệ giả thuyết và các tên gọi này sẽ không được áp dụng cho một trong hai biến trong những trường hợp như vậy.

Hãy nhớ rằng hành động gọi tên các biến theo cách này không đảm bảo rằng mối quan hệ nhân quả tồn tại. Để đánh giá chính thức xem một biến có gây ra sự thay đổi ở biến khác hay không, cần phải thực hiện một thí nghiệm.

1.2.5 Giới thiệu các nghiên cứu qua sát và thử nghiệm

Có hai loại thu thập dữ liệu chính: nghiên cứu quan sát và thí nghiệm.

Các nhà nghiên cứu thực hiện một **nghiên cứu quan sát** khi họ thu thập dữ liệu theo cách không can thiệp trực tiếp vào cách dữ liệu phát sinh. Ví dụ, các nhà nghiên cứu có thể thu thập thông tin thông qua các cuộc khảo sát, xem xét hồ sơ y tế hoặc công ty, hoặc **theo dõi một nhóm** gồm nhiều cá nhân tương tự để hình thành các giả thuyết về lý do tại sao một số bệnh có thể phát triển. Trong mỗi tình huống này, các nhà nghiên cứu chỉ quan sát dữ liệu phát sinh. Nói chung, các nghiên cứu quan sát có thể cung cấp bằng chứng về mối liên hệ tự nhiên giữa các biến, nhưng bản thân chúng không thể cho thấy mối liên hệ nhân quả.

Khi các nhà nghiên cứu muốn kiểm tra khả năng tồn tại mối quan hệ nhân quả, họ sẽ tiến hành **thử nghiệm**. Thông thường, sẽ có một biến giải thích và một biến phản hồi. Ví dụ, chúng ta có thể nghi ngờ rằng việc sử dụng một loại thuốc có thể giúp giảm tỷ lệ tử vong ở bệnh nhân bị nhồi máu cơ tim trong vòng một năm. Để kiểm tra xem liệu có thực sự tồn tại mối quan hệ nhân quả giữa biến giải thích và biến phản hồi hay không, các nhà nghiên cứu sẽ thu thập một nhóm bệnh nhân và chia họ thành nhiều nhóm. Các cá nhân trong mỗi nhóm sẽ được chỉ định một phương pháp điều trị. Khi các cá nhân được phân nhóm một cách ngẫu nhiên, thí nghiệm đó được gọi là **thử nghiệm ngẫu nhiên**. Ví dụ, mỗi bệnh nhân đau tim trong thử nghiệm thuốc có thể được phân ngẫu nhiên, có lẽ bằng cách tung đồng xu, thành một trong hai nhóm: nhóm thứ nhất nhận **giả dược** (điều trị giả) và nhóm thứ hai nhận thuốc. Xem nghiên cứu tình huống trong Phần 1.1 để biết một ví dụ khác về thí nghiệm, mặc dù nghiên cứu đó không sử dụng giả dược.

MỐI TƯƠNG QUAN $\neq$ QUAN HỆ NHÂN QUẢ

Nhìn chung, mối tương quan giữa các biến không đồng nghĩa với quan hệ nhân quả, và nhân quả chỉ có thể được suy ra từ một thử nghiệm ngẫu nhiên.

⁹ Đôi khi biến giải thích được gọi là biến **độc lập** và biến đáp ứng được gọi là biến **phụ thuộc**. Tuy nhiên, điều này trở nên khó hiểu vì một *cặp* biến có thể độc lập hoặc phụ thuộc, vì vậy chúng tôi tránh ngôn ngữ này.

Bài tập

1.3 Nghiên cứu về ô nhiễm không khí và kết quả sinh non. Các nhà nghiên cứu đã thu thập dữ liệu để kiểm tra mối quan hệ giữa các chất gây ô nhiễm không khí và sinh non ở Nam California. Trong quá trình nghiên cứu, mức độ ô nhiễm không khí được đo bằng các trạm quan trắc chất lượng không khí. Cụ thể, nồng độ cacbon monoxit được ghi lại theo đơn vị phần triệu, nitơ dioxit và ôzôn theo đơn vị phần trăm triệu, và vật chất hạt thô (PM10) theo đơn vị $\mu\text{g}/\text{m}^3$. Dữ liệu về thời gian mang thai được thu thập trên 143.196 ca sinh từ năm 1989 đến năm 1993, và mức độ tiếp xúc với ô nhiễm không khí trong quá trình mang thai được tính cho mỗi lần sinh. Phân tích cho thấy rằng sự gia tăng PM10 trong môi trường xung quanh và, ở mức độ thấp hơn, nồng độ CO có thể liên quan đến sự xuất hiện của sinh non.¹⁰

- Xác định câu hỏi nghiên cứu chính của nghiên cứu.
- Các đối tượng trong nghiên cứu này là ai và bao nhiêu đối tượng được đưa vào?
- Các biến trong nghiên cứu là gì? Xác định mỗi biến dưới dạng biến định lượng hoặc biến phân loại. Nếu là biến định lượng, hãy nêu rõ biến đó là biến rời rạc hay biến liên tục. Nếu là biến phân loại, hãy nêu rõ liệu biến đó có phải là biến thứ bậc hay không.

1.4 Nghiên cứu về phương pháp thở Buteyko. Phương pháp Buteyko là một kỹ thuật thở nông được phát triển bởi Konstantin Buteyko, một bác sĩ người Nga, vào năm 1952. Bằng chứng giai thoại cho thấy phương pháp Buteyko có thể làm giảm các triệu chứng hen suyễn và cải thiện chất lượng cuộc sống. Trong một nghiên cứu khoa học để xác định hiệu quả của phương pháp này, các nhà nghiên cứu đã tuyển dụng 600 bệnh nhân hen suyễn từ 18-69 tuổi dựa vào thuốc để điều trị hen suyễn. Những bệnh nhân này được chia ngẫu nhiên thành hai nhóm nghiên cứu: một nhóm thực hành phương pháp Buteyko và nhóm kia thì không. Bệnh nhân được chấm điểm về chất lượng cuộc sống, hoạt động, triệu chứng hen suyễn và giảm thuốc theo thang điểm từ 0 đến 10. Trung bình, những người tham gia trong nhóm Buteyko đã giảm đáng kể các triệu chứng hen suyễn và cải thiện chất lượng cuộc sống.¹¹

- Xác định câu hỏi nghiên cứu chính của nghiên cứu.
- Các đối tượng trong nghiên cứu này là ai và bao nhiêu đối tượng được đưa vào?
- Các biến trong nghiên cứu là gì? Xác định mỗi biến dưới dạng biến định lượng hoặc biến phân loại. Nếu là biến định lượng, hãy nêu rõ biến đó là biến rời rạc hay biến liên tục. Nếu là biến phân loại, hãy nêu rõ liệu biến đó có phải là biến thứ bậc hay không.

1.5 Nghiên cứu về gian lận ở trẻ em. Các nhà nghiên cứu nghiên cứu mối quan hệ giữa sự trung thực, tuổi tác và khả năng tự kiểm soát đã tiến hành một nghiên cứu trên 160 trẻ em trong độ tuổi từ 5 đến 15. Những người tham gia báo cáo tuổi tác, giới tính và liệu họ có phải là con một hay không. Các nhà nghiên cứu yêu cầu mỗi đứa trẻ ném một đồng xu ở nơi riêng tư và ghi lại kết quả (trắng hoặc đen) trên một tờ giấy, và nói rằng họ sẽ chỉ thưởng cho những đứa trẻ báo cáo màu trắng. Kết quả của nghiên cứu có thể được tóm tắt như sau: "Một nửa số sinh viên được yêu cầu không gian lận một cách rõ ràng và những người khác không có yêu cầu rõ ràng. Trong nhóm không có hướng dẫn, xác suất gian lận được tìm thấy là thống nhất giữa các nhóm dựa trên đặc điểm của trẻ. Trong nhóm được yêu cầu không gian lận một cách rõ ràng, các cô gái ít có khả năng gian lận hơn, và trong khi tỷ lệ gian lận không thay đổi theo độ tuổi đối với các chàng trai, nó giảm theo độ tuổi đối với các cô gái."¹²

- Xác định câu hỏi nghiên cứu chính của nghiên cứu.
- Các đối tượng trong nghiên cứu này là ai và bao nhiêu đối tượng được đưa vào?
- Có bao nhiêu biến được ghi lại cho mỗi đối tượng trong nghiên cứu để kết luận những phát hiện này? Nêu các biến và loại của chúng.

¹⁰ B. Ritz et al. "Effect of air pollution on preterm birth among children born in Southern California between 1989 and 1993". In: *Epidemiology* 11.5 (2000), pp. 502–511.

¹¹ J. McGowan. "Health Education: Does the Buteyko Institute Method make a difference?" In: *Thorax* 58 (2003).

¹² Alessandro Bucciol and Marco Piovesan. "Luck or cheating? A field experiment on honesty with children". In: *Journal of Economic Psychology* 32.1 (2011), pp. 73–78.

1.6 Nghiên cứu về ăn cắp. Trong một nghiên cứu về mối quan hệ giữa tầng lớp kinh tế xã hội và hành vi phi đạo đức, 129 sinh viên đại học California tại Berkeley được yêu cầu xác định bản thân có tầng lớp xã hội thấp hay cao bằng cách so sánh mình với những người khác có nhiều (hoặc ít) tiền nhất, trình độ học vấn cao (hoặc thấp) nhất và công việc nhận được nhiều (hoặc ít) sự tôn trọng nhất. Họ cũng được tặng một lọ kẹo và được thông báo rằng kẹo dành cho trẻ em ở phòng thí nghiệm gần đó, nhưng họ có thể lấy một ít nếu muốn. Sau khi hoàn thành một số nhiệm vụ không liên quan, những người tham gia sẽ báo cáo số lượng kẹo họ đã lấy.¹³

- Xác định câu hỏi nghiên cứu chính của nghiên cứu.
- Các đối tượng trong nghiên cứu này là ai và bao nhiêu đối tượng được đưa vào?
- Nghiên cứu cho thấy rằng những học sinh được xác định là tầng lớp thượng lưu đã lấy nhiều kẹo hơn những người khác. Có bao nhiêu biến được ghi lại cho mỗi đối tượng trong nghiên cứu để kết luận những phát hiện này? Nêu các biến và loại của chúng.

1.7 Nghiên cứu về đau nửa đầu và châm cứu, Phần II. Bài tập 1.1 giới thiệu một nghiên cứu khám phá liệu châm cứu có ảnh hưởng đến chứng đau nửa đầu hay không. Các nhà nghiên cứu đã tiến hành một thử nghiệm ngẫu nhiên có nhóm chứng trong đó bệnh nhân được phân ngẫu nhiên vào một trong hai nhóm: nhóm điều trị hoặc nhóm đối chứng. Những bệnh nhân trong nhóm điều trị nhận được châm cứu được thiết kế đặc biệt để điều trị cơn đau nửa đầu. Những bệnh nhân trong nhóm đối chứng nhận được châm cứu giả (châm kim vào các vị trí không phải huyệt). 24 giờ sau khi bệnh nhân nhận được châm cứu, họ được hỏi xem họ có không còn đau hay không. Các biến giải thích và biến phản hồi trong nghiên cứu này là gì?

1.8 Nghiên cứu về viêm xoang và kháng sinh, Phần II. Bài tập 1.2 giới thiệu một nghiên cứu khám phá tác dụng của điều trị kháng sinh đối với viêm xoang cấp tính. Những người tham gia nghiên cứu đã nhận được một liệu trình 10 ngày amoxicillin (kháng sinh) hoặc giả dược tương tự về ngoại hình và hương vị. Vào cuối giai đoạn 10 ngày, bệnh nhân được hỏi liệu các triệu chứng có cải thiện không. Các biến giải thích và phản hồi trong nghiên cứu này là gì?

1.9 Hoa diên vĩ của Fisher. Sir Ronald Aylmer Fisher là một nhà thống kê, nhà sinh vật học tiến hóa và nhà di truyền học người Anh, người đã làm việc trên một bộ dữ liệu có chứa chiều dài và chiều rộng của đài quang, chiều dài và chiều rộng cánh hoa của ba loài hoa diên vĩ (*setosa*, *versicolor* và *virginica*). Có 50 bông hoa từ mỗi loài trong bộ dữ liệu.¹⁴

- Có bao nhiêu trường hợp được đưa vào dữ liệu?
- Có bao nhiêu biến định lượng được đưa vào dữ liệu? Cho biết các biến là loại gì (biến liên tục hay biến rời rạc).
- Có bao nhiêu biến phân loại được đưa vào dữ liệu và chúng là gì? Liệt kê các mức độ tương quan (danh mục).



Ảnh của Ryan Claussen
(<http://flic.kr/p/6QTcuX>) giấy
phép CC BY-SA 2.0 license

1.10 Thói quen hút thuốc của cư dân Vương quốc Anh. Một cuộc khảo sát đã được tiến hành để nghiên cứu thói quen hút thuốc của cư dân Vương quốc Anh. Dưới đây là ma trận dữ liệu hiển thị một phần dữ liệu được thu thập trong cuộc khảo sát này. Lưu ý rằng “£” là viết tắt của Bảng Anh và “N/A” đề cập đến một thành phần bị thiếu của dữ liệu.¹⁵

TT	Giới tính	Tuổi	Tình trạng hôn nhân	Thu nhập	Hút thuốc	Lượng thuốc lá cuối tuần	Lượng thuốc lá ngày thường
1	Nữ	42	Độc thân	Dưới £2,600	Có	12 điếu/ngày	12 điếu/ngày
2	Nam	44	Độc thân	£10,400 - £15,600	Không	N/A	N/A
3	Nam	53	Đã kết hôn	Trên £36,400	Có	6 điếu/ngày	6 điếu/ngày
...	...	...	...	...	...	...	...
1691	Nam	40	Độc thân	£2,600 - £5,200	Có	8 điếu/ngày	8 điếu/ngày

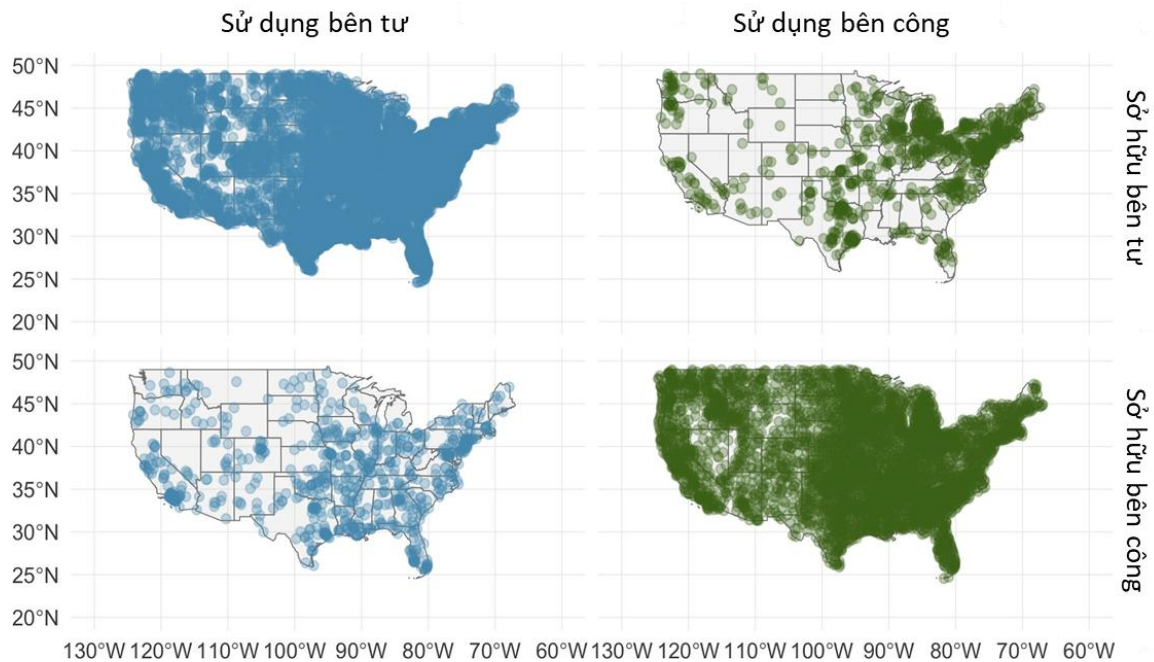
- Có bao nhiêu người tham gia khảo sát?
- Cho biết mỗi biến trong nghiên cứu là biến định lượng hay biến phân loại. Nếu là biến định lượng, hãy xác định là biến liên tục hay biến rời rạc. Nếu là biến phân loại, hãy cho biết biến đó có phải là biến thứ bậc hay không.

¹³ P.K. Piff et al. “Higher social class predicts increased unethical behavior”. In: *Proceedings of the National Academy of Sciences* (2012).

¹⁴ R.A Fisher. “The Use of Multiple Measurements in Taxonomic Problems”. In: *Annals of Eugenics* 7 (1936), pp. 179–188.

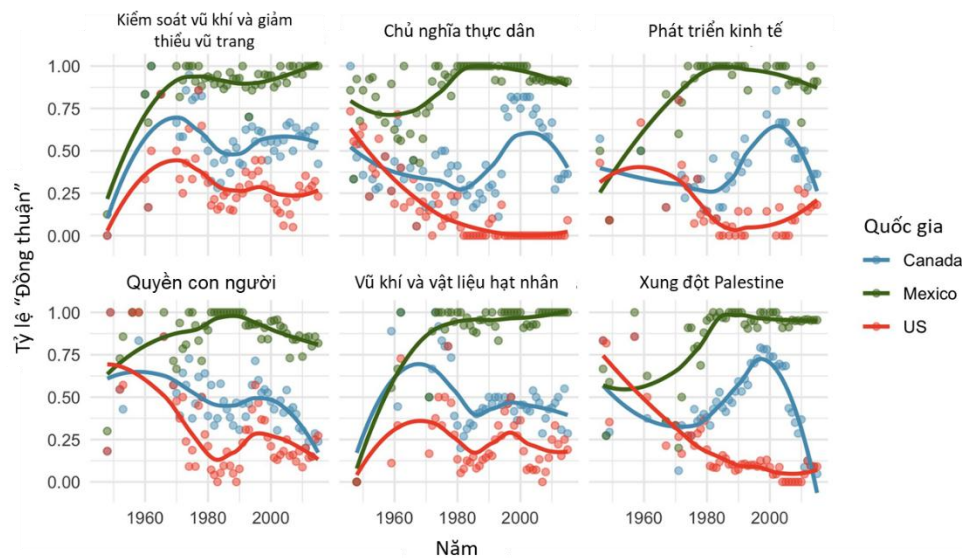
¹⁵ National STEM Centre, Large Datasets from stats4schools.

1.11 Sân bay Hoa Kỳ. Biểu đồ dưới đây minh họa sự phân bố địa lý của các sân bay ở Hoa Kỳ, bao gồm cả khu vực tiếp giáp với Washington, DC. Biểu đồ này được tạo ra dựa trên một tập dữ liệu, trong đó mỗi quan sát đại diện cho một sân bay cụ thể.¹⁶



- (a) Liệt kê các biến được sử dụng để tạo hình ảnh trực quan này.
 (b) Cho biết mỗi biến trong nghiên cứu là biến định lượng hay biến phân loại. Nếu là biến định lượng, hãy xác định là biến liên tục hoặc biến rời rạc. Nếu là biến phân loại, hãy cho biết biến đó có phải là biến thứ bậc hay không.

1.12 Phiếu bầu của Liên Hợp Quốc. Biểu đồ dưới đây minh họa các mô hình bỏ phiếu tại Đại hội đồng Liên hợp quốc ở Hoa Kỳ, Canada và Mexico, tập trung vào nhiều vấn đề khác nhau trong khoảng thời gian từ năm 1946 đến 2015. Cụ thể, biểu đồ thể hiện tỷ lệ phần trăm các phiếu bầu "Đồng thuận" của mỗi quốc gia cho từng vấn đề cụ thể. Dữ liệu trong biểu đồ được xây dựng từ một tập hợp các quan sát, trong đó mỗi quan sát là một cặp quốc gia và năm tương ứng.¹⁷



- (a) Liệt kê các biến được sử dụng để tạo hình ảnh trực quan này.
 (b) Cho biết mỗi biến trong nghiên cứu là biến định lượng hay biến phân loại. Nếu là biến định lượng, hãy xác định là biến liên tục hoặc biến rời rạc. Nếu là biến phân loại, hãy cho biết biến đó có phải là biến thứ bậc hay không.

¹⁶ Federal Aviation Administration, www.faa.gov/airports/airport_safety/airportdata 5010.

¹⁷ David Robinson. *unvotes: United Nations General Assembly Voting Data*. R package version 0.2.0. 2017. url: <https://CRAN.R-project.org/package=unvotes>.

1.3 Nguyên tắc và chiến lược lấy mẫu

Bước đầu tiên trong việc tiến hành nghiên cứu là xác định các chủ đề hoặc câu hỏi nghiên cứu. Một câu hỏi nghiên cứu được trình bày rõ ràng rất hữu ích trong việc xác định những chủ đề hoặc trường hợp nào nên được nghiên cứu và những biến số nào là quan trọng. Điều quan trọng nữa là phải xem xét *cách* dữ liệu được thu thập có đáng tin cậy và giúp đạt được các mục tiêu nghiên cứu.

1.3.1 Quần thể và mẫu

Hãy xem xét ba câu hỏi nghiên cứu sau đây:

1. Hàm lượng thủy ngân trung bình trong cá kiếm ở Đại Tây Dương là bao nhiêu?
2. Trong 5 năm qua, thời gian trung bình để hoàn thành bằng cấp cho sinh viên đại học Duke là bao nhiêu?
3. Một loại thuốc mới có làm giảm số ca tử vong ở bệnh nhân mắc bệnh tim không?

Mỗi câu hỏi nghiên cứu đều liên quan đến một **nhóm đối tượng** mục tiêu cụ thể. Trong câu hỏi đầu tiên, quần thể được nhắc đến là toàn bộ cá kiếm ở Đại Tây Dương, với mỗi con cá đại diện cho một đơn vị quan sát. Việc thu thập dữ liệu cho tất cả các đơn vị trong quần thể thường tốn kém và khó khăn. Do đó, thay vì khảo sát toàn bộ, một mẫu được chọn ra. **Mẫu** là một tập hợp con gồm các đơn vị quan sát, thường chỉ chiếm một phần nhỏ trong quần thể. Chẳng hạn, có thể chọn 60 con cá kiếm (hoặc một con số khác) trong quần thể để tạo thành mẫu. Dữ liệu thu thập từ mẫu này sẽ được sử dụng để ước lượng các thông số của quần thể và trả lời các câu hỏi nghiên cứu.

BÀI TẬP THỰC HÀNH 1.9

Đối với câu hỏi thứ hai và thứ ba ở trên, hãy xác định nhóm đối tượng mục tiêu và những gì đại diện cho một trường hợp cụ thể.¹⁸

1.3.2 Bảng chứng giai thoại

Hãy xem xét các câu trả lời có thể có sau đây cho ba câu hỏi nghiên cứu:

1. Một người đàn ông trên tin tức đã bị ngộ độc thủy ngân do ăn cá kiếm, vì vậy nồng độ thủy ngân trung bình trong cá kiếm phải cao một cách nguy hiểm.
2. Tôi đã gặp hai sinh viên mất hơn 7 năm để tốt nghiệp Duke, vì vậy phải mất nhiều thời gian hơn để tốt nghiệp tại Duke so với nhiều trường cao đẳng khác.
3. Bố của bạn tôi bị đau tim và qua đời sau khi họ cho ông uống thuốc điều trị tim mới, vì vậy thuốc không được tác dụng.

Mỗi kết luận trên đều dựa trên dữ liệu. Tuy nhiên, có hai vấn đề. Đầu tiên, dữ liệu chỉ đại diện cho một hoặc hai trường hợp. Thứ hai, và quan trọng hơn, không rõ liệu những trường hợp này có thực sự đại diện cho quần thể hay không. Dữ liệu được thu thập theo cách lộn xộn này được gọi là bảng chứng giai thoại.

BẢNG CHỨNG GIAI THOẠI

Hãy cẩn thận với dữ liệu được thu thập một cách bừa bãi. Bảng chứng như vậy có thể đúng và có thể kiểm chứng, nhưng nó chỉ có thể đại diện cho những trường hợp đặc biệt.

¹⁸ (2) Câu hỏi đầu tiên chỉ liên quan đến những sinh viên hoàn thành bằng cấp của họ; điểm trung bình không thể được tính bằng cách sử dụng một sinh viên chưa bao giờ hoàn thành bằng cấp của mình. Do đó, chỉ những sinh viên đại học Duke tốt nghiệp trong 5 năm qua đại diện cho các trường hợp trong quần thể đang được xem xét. Mỗi học sinh như vậy là một trường hợp riêng lẻ. (3) Một người mắc bệnh tim đại diện cho một trường hợp. Dân số bao gồm tất cả những người mắc bệnh tim.

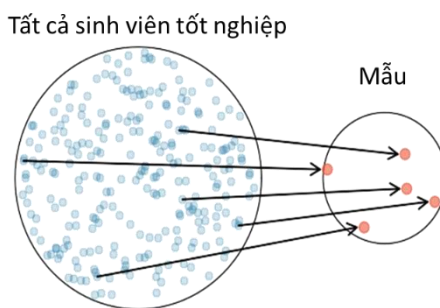


Hình 1.10: Vào tháng 2 năm 2010, một số chuyên gia truyền thông đã trích dẫn một cơn bão tuyết lớn là bằng chứng hợp lệ chống lại sự nóng lên toàn cầu. Như diễn viên hài Jon Stewart đã chỉ ra, "Đó là một cơn bão, trong một khu vực, của một quốc gia."

Bằng chứng giai thoại thường bao gồm các trường hợp bất thường mà chúng ta nhớ lại dựa trên các đặc điểm nổi bật của chúng. Ví dụ, chúng ta có nhiều khả năng nhớ hai người mà chúng ta đã gặp mà họ mất 7 năm để tốt nghiệp hơn sáu người khác đã tốt nghiệp trong bốn năm. Thay vì xem xét các trường hợp bất thường nhất, chúng ta nên kiểm tra một mẫu gồm nhiều trường hợp đại diện cho quần thể.

1.3.3 Lấy mẫu từ một quần thể

Chúng ta có thể ước tính thời gian tốt nghiệp của sinh viên đại học Duke trong 5 năm qua bằng cách thu thập một mẫu sinh viên. Tất cả những sinh viên tốt nghiệp trong 5 năm qua tạo thành quần thể, còn những sinh viên được chọn để xem xét thì được gọi là mẫu. Trong nghiên cứu, chúng ta thường hướng đến việc lựa chọn một mẫu một cách ngẫu nhiên từ quần thể. Một phương pháp cơ bản để chọn mẫu ngẫu nhiên là tương tự như việc tiến hành xổ số. Cụ thể, trong trường hợp lựa chọn sinh viên tốt nghiệp, chúng tôi có thể ghi tên của từng sinh viên trên các vé xổ số và rút ra 100 vé. Những cái tên được chọn sẽ đại diện cho một mẫu ngẫu nhiên gồm 100 sinh viên tốt nghiệp. Việc chọn mẫu ngẫu nhiên giúp giảm thiểu khả năng xảy ra thành kiến trong quá trình khảo sát.



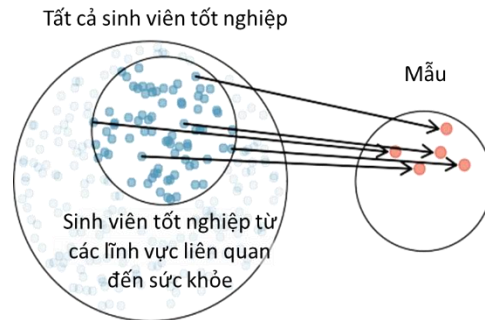
Hình 1.11: Trong hình này, 5 sinh viên tốt nghiệp được chọn ngẫu nhiên từ quần thể để đưa vào mẫu.

VÍ DỤ 1.10

Giả sử chúng ta yêu cầu một sinh viên tình cờ học chuyên ngành dinh dưỡng chọn một số sinh viên tốt nghiệp cho nghiên cứu. Bạn nghĩ cô ấy có thể thu thập loại học sinh nào? Bạn có nghĩ mẫu của cô ấy sẽ đại diện cho tất cả sinh viên tốt nghiệp?

E

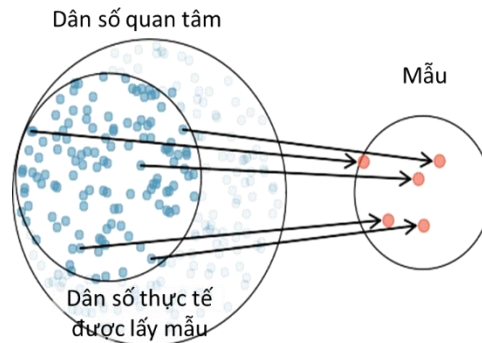
Có lẽ cô ấy sẽ chọn một số lượng không cân xứng sinh viên tốt nghiệp từ các lĩnh vực liên quan đến sức khỏe. Hoặc theo lựa chọn của cô ấy sẽ là một đại diện tốt cho quần thể. Khi chọn mẫu một cách thủ công, chúng ta có nguy cơ chọn một mẫu **thiên lệch (bias)**, ngay cả khi sự thiên vị không có chủ ý.



Hình 1.12: Khi được yêu cầu chọn một mẫu sinh viên tốt nghiệp, một sinh viên chuyên ngành dinh dưỡng có thể vô tình chọn một số lượng không ngẫu nhiên từ các ngành liên quan đến sức khỏe.

Nếu một người được phép lựa chọn và xác định chính xác sinh viên tốt nghiệp đưa vào mẫu, có khả năng mẫu đó sẽ bị sai lệch so với mục đích của họ, và điều này có thể hoàn toàn không cố ý. Sự **thiên lệch** này có thể ảnh hưởng đến tính chính xác của mẫu. Việc lấy mẫu ngẫu nhiên là biện pháp hiệu quả để giải quyết vấn đề này. Mẫu ngẫu nhiên cơ bản nhất được gọi là **mẫu ngẫu nhiên đơn giản**, tương đương với việc sử dụng xổ số để chọn các đơn vị quan sát. Điều này có nghĩa là mỗi đơn vị trong quần thể có cơ hội như nhau để được chọn, và không có mối liên hệ nào giữa các đơn vị trong mẫu.

Hành động lấy mẫu ngẫu nhiên đơn giản giúp giảm thiểu sự thiên lệch. Tuy nhiên, thiên lệch vẫn có thể xuất hiện qua các hình thức khác. Ngay cả khi việc chọn người được thực hiện ngẫu nhiên, như trong các cuộc khảo sát, cần đặc biệt chú ý đến tỷ lệ **không phản hồi** cao. Ví dụ, nếu chỉ 30% số người trong mẫu ngẫu nhiên thực sự phản hồi cho một cuộc khảo sát, thì không thể rõ ràng rằng kết quả có **đại diện** cho toàn bộ quần thể hay không. Thiên lệch không phản hồi này có thể làm lệch hướng kết quả.



Hình 1.13: Do khả năng không phản hồi, các nghiên cứu khảo sát có thể chỉ tiếp cận một nhóm nhất định trong quần thể. Việc khắc phục hoàn toàn vấn đề này là rất khó khăn, và đôi khi là không thể.

Một nhược điểm phổ biến khác là **mẫu thuận tiện**, trong đó những cá nhân dễ tiếp cận có nhiều khả năng được đưa vào mẫu. Chẳng hạn, nếu một cuộc khảo sát chính trị được thực hiện bằng cách chặn những người đi bộ ở Bronx, mẫu này sẽ không thể đại diện cho toàn bộ Thành phố New York. Việc xác định một mẫu thuận tiện có thực sự đại diện cho quần thể phụ nào đó thường rất khó khăn.

BÀI TẬP THỰC HÀNH 1.11

Chúng tôi có thể dễ dàng truy cập xếp hạng cho sản phẩm, người bán và công ty thông qua các trang web. Những xếp hạng này chỉ dựa trên những người cố gắng cung cấp xếp hạng. Nếu 50% đánh giá trực tuyến cho một sản phẩm là tiêu cực, bạn có nghĩ rằng điều này có nghĩa là 50% người mua không hài lòng với sản phẩm? ¹⁹

¹⁹ Câu trả lời sẽ khác nhau. Từ kinh nghiệm giai thoại của chính chúng tôi, chúng tôi tin rằng mọi người có xu hướng nói nhiều hơn về các sản phẩm thấp hơn mong đợi hơn là ca ngợi những sản phẩm hoạt động như mong đợi. Vì lý do này, chúng tôi nghi ngờ có sự thiên vị tiêu cực trong xếp hạng sản phẩm trên các trang web như Amazon. Tuy nhiên, vì kinh nghiệm của chúng ta có thể không đại diện, chúng ta cũng giữ một tâm trí cởi mở.

1.3.4 Nghiên cứu quan sát

Dữ liệu mà không có phương pháp điều trị nào được áp dụng một cách rõ ràng (hoặc bị giữ lại một cách rõ ràng) được gọi là **dữ liệu quan sát**. Ví dụ, dữ liệu cho vay và dữ liệu của quận được mô tả trong Phần 1.2 là cả hai ví dụ về dữ liệu quan sát. Đưa ra kết luận nhân quả dựa trên các thử nghiệm thường là hợp lý. Tuy nhiên, việc đưa ra kết luận nhân quả tương tự dựa trên dữ liệu quan sát có thể nguy hiểm và không được khuyến khích. Do đó, các nghiên cứu quan sát thường chỉ đủ để chỉ ra các mối liên hệ hoặc hình thành các giả thuyết mà sau này chúng tôi kiểm tra bằng cách sử dụng các thử nghiệm.

BÀI TẬP THỰC HÀNH 1.12

Giả sử một nghiên cứu quan sát theo dõi việc sử dụng kem chống nắng và ung thư da, và người ta thấy rằng càng sử dụng nhiều kem chống nắng, người đó càng có nhiều khả năng bị ung thư da. Điều này có nghĩa là kem chống nắng gây ung thư da? ²⁰

Một số nghiên cứu trước đây cho chúng ta biết rằng việc sử dụng kem chống nắng thực sự làm giảm nguy cơ ung thư da, vì vậy có thể có một biến số khác có thể giải thích mối liên hệ giả định này giữa việc sử dụng kem chống nắng và ung thư da. Một thông tin quan trọng không có mặt là tiếp xúc với ánh nắng mặt trời. Nếu ai đó ra ngoài nắng cả ngày, cô ấy có nhiều khả năng sử dụng kem chống nắng và có nhiều khả năng bị ung thư da. Tiếp xúc với ánh nắng mặt trời không được tính đến trong cuộc điều tra đơn giản.



Tiếp xúc với ánh nắng mặt trời là cái được gọi là **biến gây nhiễu**, ²¹ là một biến tương quan với cả biến giải thích và phản hồi. Mặc dù một phương pháp để biện minh cho việc đưa ra kết luận nhân quả từ các nghiên cứu quan sát là tìm kiếm hết các biến gây nhiễu, nhưng không có gì đảm bảo rằng tất cả các biến gây nhiễu có thể được kiểm tra hoặc đo lường.

BÀI TẬP THỰC HÀNH 1.13

Hình 1.8 cho thấy mối liên hệ nghịch giữa tỷ lệ sở hữu nhà và tỷ lệ phần trăm của các cấu trúc nhiều căn hộ trong một quận. Tuy nhiên, thật không hợp lý khi kết luận rằng có mối quan hệ nhân quả giữa hai biến. Đề xuất một biến có thể giải thích mối quan hệ nghịch. ²²

Các nghiên cứu quan sát có hai hình thức: nghiên cứu tiền cứu và nghiên cứu hồi cứu. **Một nghiên cứu tiền cứu** xác định các cá nhân và thu thập thông tin khi các sự kiện diễn ra. Ví dụ, các nhà nghiên cứu y tế có thể xác định và theo dõi một nhóm bệnh nhân trong nhiều năm để đánh giá ảnh hưởng có thể có của hành vi đối với nguy cơ ung thư. Một ví dụ về nghiên cứu như vậy là Nghiên cứu Sức khỏe của Y tá, bắt đầu vào năm 1976 và mở rộng vào năm 1989. Nghiên cứu tiền cứu này tuyển dụng các y tá đã đăng ký và sau đó thu thập dữ liệu từ họ bằng cách sử dụng bảng câu hỏi. Các **nghiên cứu hồi cứu** thu thập dữ liệu sau khi các sự kiện đã diễn ra, ví dụ: các nhà nghiên cứu có thể xem xét các sự kiện trong quá khứ trong hồ sơ y tế. Một số tập dữ liệu có thể chứa cả biến được thu thập tiền cứu và hồi cứu.

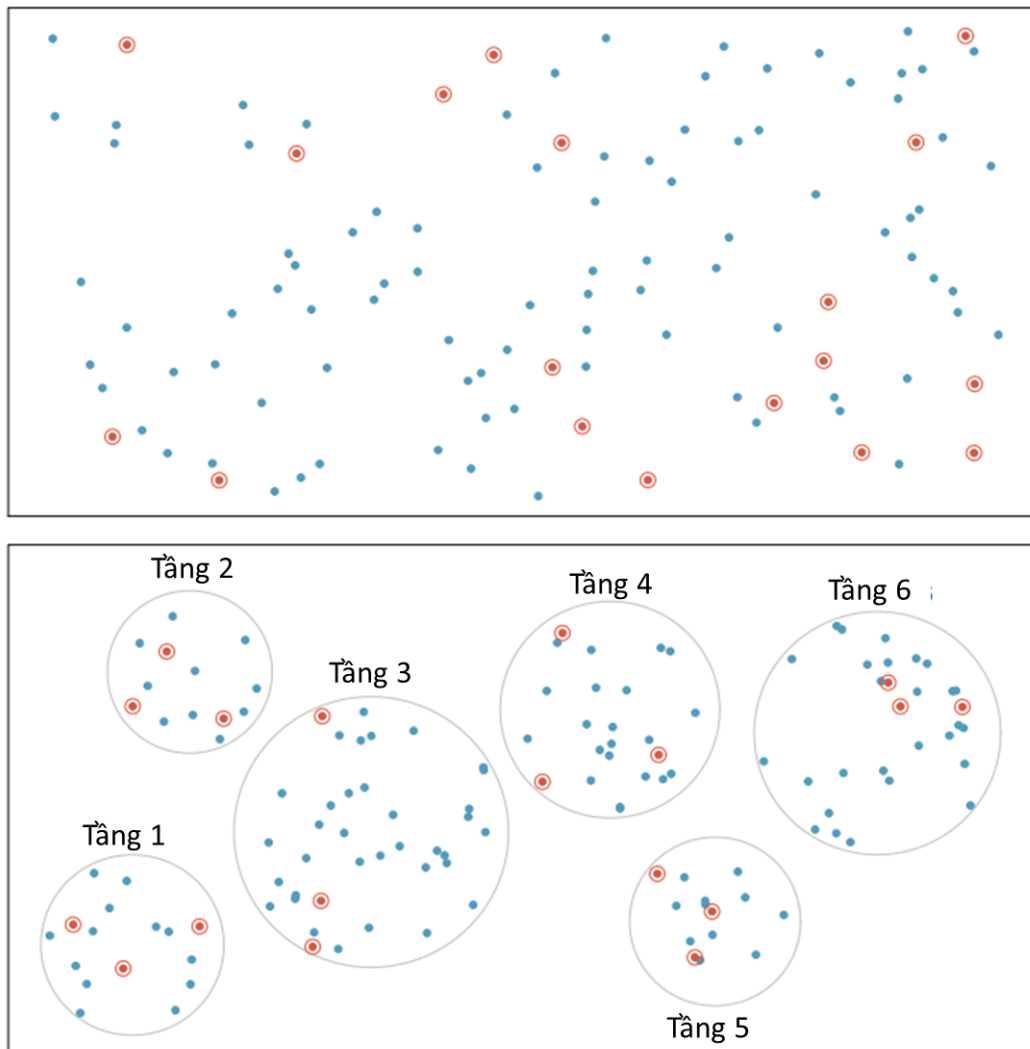
1.3.5 Bốn phương pháp lấy mẫu

Hầu như tất cả các phương pháp thống kê đều dựa trên khái niệm ngẫu nhiên giả định. Nếu dữ liệu quan sát không được thu thập trong một khuôn khổ ngẫu nhiên từ một quần thể, các phương pháp thống kê này – các ước tính và sai số liên quan – sẽ trở nên không đáng tin cậy. Ở đây chúng tôi xem xét bốn kỹ thuật lấy mẫu ngẫu nhiên: lấy mẫu ngẫu nhiên đơn, lấy mẫu phân tầng, lấy mẫu cụm và lấy mẫu nhiều giai đoạn. Hình 1.14 và 1.15 cung cấp các biểu diễn đồ họa của các kỹ thuật này.

²⁰ Không. Xem đoạn sau bài tập để được giải thích.

²¹ Còn được gọi là **biến tiềm ẩn**, **yếu tố gây nhiễu** hoặc **gây nhiễu**.

²² Câu trả lời sẽ khác nhau. Mật độ dân số có thể quan trọng. Nếu một quận rất đông đúc, thì điều này có thể yêu cầu một phần lớn cư dân sống trong các cấu trúc nhiều đơn vị. Ngoài ra, mật độ cao có thể góp phần làm tăng giá trị tài sản, khiến việc sở hữu nhà không khả thi đối với nhiều cư dân.



Hình 1.14: Ví dụ về lấy mẫu ngẫu nhiên và phân tầng đơn giản. Ở hình trên, lấy mẫu ngẫu nhiên đơn được sử dụng để chọn ngẫu nhiên 18 trường hợp. Ở hình dưới, lấy mẫu phân tầng được sử dụng: các trường hợp được nhóm thành các tầng, sau đó lấy mẫu ngẫu nhiên đơn được sử dụng trong mỗi tầng.

Lấy mẫu ngẫu nhiên đơn có lẽ là hình thức lấy mẫu ngẫu nhiên trực quan nhất. Hãy xem xét mức lương của các cầu thủ Bóng chày Major League (MLB), trong đó mỗi cầu thủ là thành viên của một trong 30 đội bóng của giải đấu. Để lấy một mẫu ngẫu nhiên đơn gồm 120 cầu thủ bóng chày và tiền lương của họ, chúng ta có thể viết tên của hàng trăm cầu thủ của mùa giải đó lên các mảnh giấy, thả các mảnh vào một cái xô, lắc cái xô cho đến khi chúng ta chắc chắn rằng tất cả các tên được trộn lẫn, sau đó rút ra các mảnh cho đến khi chúng ta có mẫu 120 cầu thủ. Nói chung, một mẫu được gọi là “ngẫu nhiên đơn” nếu mỗi đơn vị trong quần thể có cơ hội bằng nhau để được chọn vào mẫu cuối cùng, và việc biết một đơn vị được chọn sẽ không cung cấp thông tin hữu ích về các đơn vị khác đã được chọn trong mẫu.

Lấy mẫu phân tầng là một chiến lược lấy mẫu chia để trị. Quần thể được chia thành các nhóm gọi là tầng. Các tầng được chọn để các trường hợp tương tự được nhóm lại với nhau, sau đó một phương pháp lấy mẫu thứ hai, thường là lấy mẫu ngẫu nhiên đơn, được sử dụng trong mỗi tầng. Trong ví dụ về lương bóng chày, các đội có thể đại diện cho các tầng, vì một số đội có nhiều tiền hơn (có thể gấp hơn đến 4 lần!). Sau đó, chúng ta có thể lấy mẫu ngẫu nhiên 4 người chơi từ mỗi đội với tổng số 120 người chơi.

Lấy mẫu phân tầng đặc biệt hữu ích khi các trường hợp trong mỗi tầng rất giống nhau về kết quả quan tâm. Nhược điểm là phân tích dữ liệu từ một mẫu phân tầng sẽ phức tạp hơn so với phân tích dữ liệu từ một mẫu ngẫu nhiên đơn. Các phương pháp phân tích được giới thiệu trong cuốn sách này sẽ cần phải được mở rộng để phân tích dữ liệu được thu thập bằng cách sử dụng lấy mẫu phân tầng.

VÍ DỤ 1.4

Tại sao các trường hợp trong mỗi tầng nên có độ tương đồng cao?

Chúng ta có thể đạt được ước tính ổn định hơn cho quần thể phụ trong một tầng khi các trường hợp tương tự như nhau, dẫn đến độ chính xác cao hơn cho từng nhóm. Khi kết hợp những ước tính này thành một ước tính duy nhất cho toàn bộ quần thể, độ chính xác của ước tính tổng quan sẽ cao hơn, nhờ vào việc mỗi ước tính nhóm riêng lẻ đã chính xác hơn.

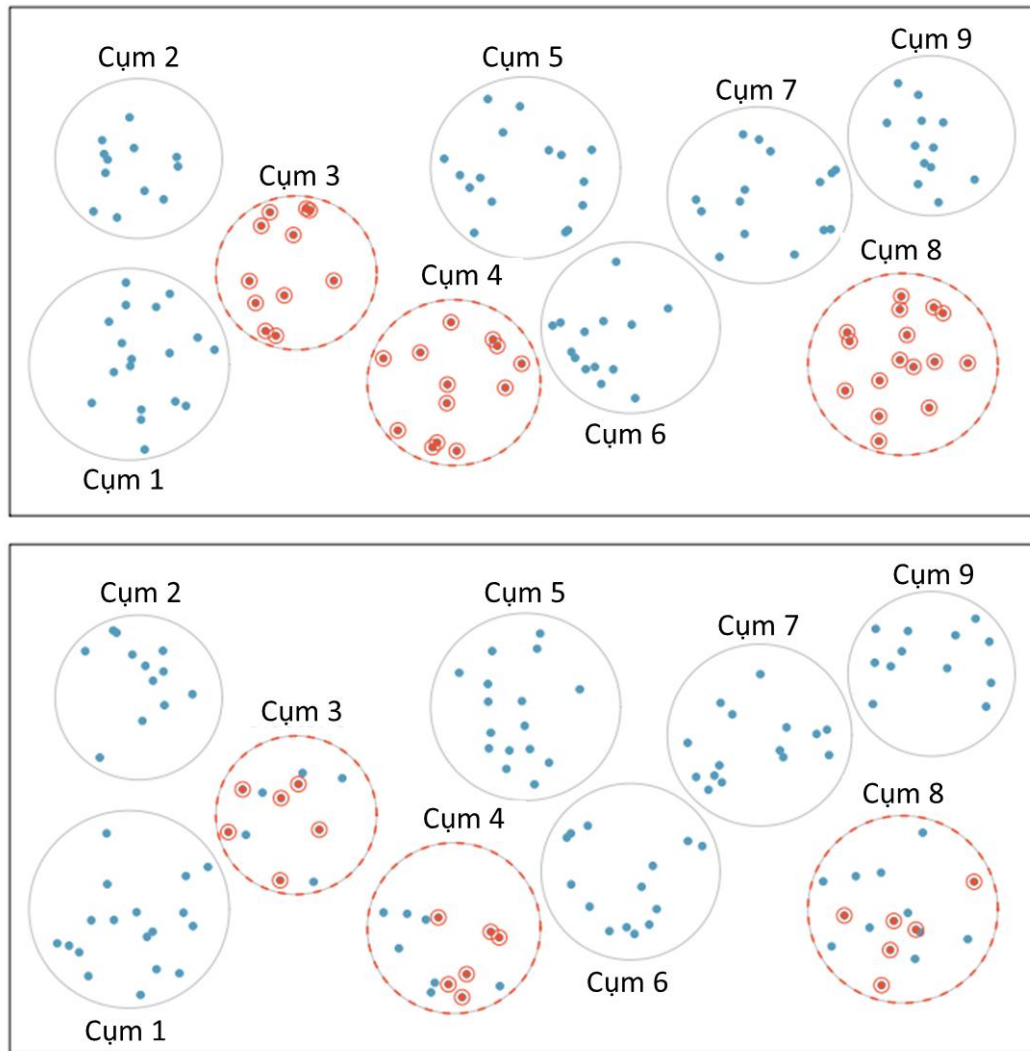
Trong **chọn mẫu cụm**, chúng ta chia quần thể thành nhiều nhóm, được gọi là các cụm. Sau đó, chúng ta lấy mẫu một số lượng cụm cố định và bao gồm tất cả các quan sát từ mỗi cụm đó trong mẫu. **Một mẫu nhiều bậc** giống như một mẫu cụm, nhưng thay vì giữ tất cả các quan sát trong mỗi cụm, chúng tôi thu thập một mẫu ngẫu nhiên trong mỗi cụm đã chọn.

Không giống như lấy mẫu phân tầng, các phương pháp này hữu ích nhất khi có nhiều sự thay đổi theo từng trường hợp trong một cụm nhưng bản thân các cụm không khác nhau lắm. Ví dụ: nếu các khu vực lân cận đại diện cho các cụm, thì việc lấy mẫu theo cụm hoặc chọn mẫu nhiều bậc sẽ hoạt động tốt nhất khi các khu vực lân cận rất đa dạng. Một nhược điểm của các phương pháp này là các kỹ thuật tiên tiến hơn thường được yêu cầu để phân tích dữ liệu, tuy nhiên các phương pháp trong cuốn sách này có thể được mở rộng để xử lý dữ liệu đó.

VÍ DỤ 1.5

Giả sử chúng ta quan tâm đến việc ước tính tỷ lệ sốt rét ở một khu vực nhiệt đới đông dân cư ở vùng nông thôn Indonesia. Chúng tôi biết rằng có 30 ngôi làng ở khu vực đó của rừng rậm Indonesia, các ngôi làng có nét tương đồng với nhau. Mục tiêu của chúng ta là xét nghiệm sốt rét cho 150 đối tượng. Nên sử dụng phương pháp chọn mẫu nào?

Một mẫu ngẫu nhiên đơn có thể sẽ chọn các đối tượng từ tất cả 30 ngôi làng, điều này có thể khiến việc thu thập dữ liệu trở nên cực kỳ tốn kém. Lấy mẫu phân tầng sẽ là một thách thức vì không rõ làm thế nào chúng ta sẽ xây dựng các tầng của các đối tượng tương tự. Tuy nhiên, lấy mẫu theo cụm hoặc lấy mẫu nhiều bậc có vẻ như là ý tưởng rất hay. Nếu chúng ta quyết định sử dụng mẫu nhiều bậc, chúng tôi có thể chọn ngẫu nhiên một nửa số làng, sau đó chọn ngẫu nhiên 10 người từ mỗi làng. Điều này có thể sẽ làm giảm đáng kể chi phí thu thập dữ liệu của chúng ta so với một mẫu ngẫu nhiên đơn và mẫu cụm vẫn sẽ cung cấp cho chúng ta thông tin đáng tin cậy, ngay cả khi chúng ta cần phân tích dữ liệu với các phương pháp nâng cao hơn một chút so với chúng ta thảo luận trong cuốn sách này.



Hình 1.15: Ví dụ về chọn mẫu cụm và chọn mẫu nhiều bậc như sau: Trong biểu đồ trên, phương pháp chọn mẫu cụm được áp dụng, trong đó dữ liệu được phân chia thành chín cụm. Từ đó, ba cụm được lựa chọn ngẫu nhiên, và tất cả các quan sát trong ba cụm này sẽ được đưa vào mẫu. Ngược lại, trong hình dưới, phương pháp chọn mẫu nhiều bậc đã được sử dụng. Điểm khác biệt so với chọn mẫu cụm là chúng tôi sẽ chọn ngẫu nhiên một tập hợp con từ mỗi cụm để đưa vào mẫu, thay vì lấy toàn bộ các trường hợp có trong mỗi cụm mẫu.

Bài tập

1.1 Nghiên cứu về ô nhiễm không khí và kết quả sinh. Bài tập 1.3 giới thiệu một nghiên cứu trong đó các nhà nghiên cứu thu thập dữ liệu để kiểm tra mối quan hệ giữa các chất gây ô nhiễm không khí và sinh non ở Nam California. Trong quá trình nghiên cứu, mức độ ô nhiễm không khí được đo bằng các trạm quan trắc chất lượng không khí. Dữ liệu về thời gian mang thai được thu thập trên 143.196 ca sinh từ năm 1989 đến năm 1993, và mức độ tiếp xúc với ô nhiễm không khí trong quá trình mang thai được tính cho mỗi lần sinh.

- (a) Xác định quần thể được quan tâm và mẫu trong nghiên cứu này.
- (b) Nhận xét về việc liệu kết quả của nghiên cứu có thể được khái quát hóa cho quần thể hay không, và liệu những phát hiện của nghiên cứu có thể được sử dụng để thiết lập các mối quan hệ nhân quả hay không.

1.2 Nghiên cứu về hiện tượng gian lận. Bài tập 1.5 giới thiệu một nghiên cứu trong đó các nhà nghiên cứu kiểm tra mối liên hệ giữa sự trung thực, tuổi tác và khả năng tự kiểm soát bản thân qua một thử nghiệm trên 160 trẻ em trong độ tuổi từ 5 đến 15. Các nhà nghiên cứu yêu cầu mỗi đứa trẻ ném một đồng xu ở nơi riêng tư và ghi lại kết quả (mặt trắng hoặc mặt đen) trên một tờ giấy, và nói rằng họ sẽ chỉ thưởng cho những đứa trẻ báo cáo mặt trắng. Một nửa số trẻ được yêu cầu không gian lận một cách rõ ràng và những người khác thì không có yêu cầu. Sự khác biệt đã được quan sát thấy trong tỷ lệ gian lận trong nhóm nhận được yêu cầu không gian lận và nhóm không có yêu cầu, cũng như một số khác biệt giữa các đặc điểm của trẻ em trong mỗi nhóm.

- (a) Xác định quần thể quan tâm và mẫu trong nghiên cứu này.
- (b) Nhận xét về việc liệu kết quả của nghiên cứu có thể được khái quát hóa cho quần thể hay không, và liệu những phát hiện của nghiên cứu có thể được sử dụng để thiết lập các mối quan hệ nhân quả hay không.

1.3 Nghiên cứu về phương pháp Buteyko. Bài tập 1.4 giới thiệu một nghiên cứu về việc sử dụng kỹ thuật thở nông Buteyko để giảm các triệu chứng hen suyễn và cải thiện chất lượng cuộc sống. Là một phần của nghiên cứu này, 600 bệnh nhân hen suyễn từ 18-69 tuổi dựa vào thuốc để điều trị hen suyễn đã được tuyển dụng và phân ngẫu nhiên vào hai nhóm: một nhóm thực hành phương pháp Buteyko và nhóm kia thì không. Những người trong nhóm Buteyko, nhìn chung, đã giảm đáng kể các triệu chứng hen suyễn và cải thiện chất lượng cuộc sống.

- (a) Xác định quần thể quan tâm và mẫu trong nghiên cứu này.
- (b) Nhận xét về việc liệu kết quả của nghiên cứu có thể được khái quát hóa cho quần thể hay không, và liệu những phát hiện của nghiên cứu có thể được sử dụng để thiết lập các mối quan hệ nhân quả hay không.

1.4 Nghiên cứu về hiện tượng ăn cắp. Bài tập 1.6 giới thiệu một nghiên cứu về mối quan hệ giữa tầng lớp kinh tế xã hội và hành vi phi đạo đức. Trong nghiên cứu này, 129 sinh viên tốt nghiệp Đại học California Berkeley được yêu cầu xác định bản thân có tầng lớp xã hội thấp hay cao bằng cách so sánh mình với những người khác có nhiều (hoặc ít) tiền nhất, trình độ học vấn cao (hoặc thấp) nhất và công việc nhận được nhiều (hoặc ít) sự tôn trọng nhất. Họ cũng được tặng một lọ kẹo riêng và được thông báo rằng kẹo dành cho trẻ em ở phòng thí nghiệm gần đó, nhưng họ có thể lấy một ít nếu muốn. Sau khi hoàn thành một số nhiệm vụ không liên quan, những người tham gia đã báo cáo số lượng kẹo họ đã lấy. Người ta phát hiện ra rằng những người được xác định là tầng lớp thượng lưu đã lấy nhiều kẹo hơn những người khác.

- (a) Xác định quần thể quan tâm và mẫu trong nghiên cứu này.
- (b) Nhận xét về việc liệu kết quả của nghiên cứu có thể được khái quát hóa cho quần thể hay không, và liệu những phát hiện của nghiên cứu có thể được sử dụng để thiết lập các mối quan hệ nhân quả hay không.

1.5 Khảo sát về thư giãn sau giờ làm việc. Khảo sát xã hội chung đặt ra câu hỏi: “Sau một ngày làm việc trung bình, bạn có bao nhiêu giờ để thư giãn hoặc theo đuổi các hoạt động mà bạn thích?” cho một mẫu ngẫu nhiên gồm 1.155 người Mỹ. Thời gian thư giãn trung bình được tìm thấy là 1,65 giờ. Xác định điều nào sau đây là quan sát, biến số, thống kê mẫu (giá trị được tính toán dựa trên mẫu quan sát) hoặc tham số quần thể.

- (a) Một người Mỹ trong mẫu.
- (b) Số giờ dành cho việc thư giãn sau một ngày làm việc trung bình.
- (c) 1.65.
- (d) Số giờ trung bình mà tất cả người Mỹ dành cho thư giãn sau một ngày làm việc trung bình.

1.6 Mèo trên YouTube. Giả sử bạn muốn ước tính tỷ lệ phần trăm video trên YouTube là video về mèo. Bạn không thể xem tất cả các video trên YouTube nên bạn sử dụng bộ chọn video ngẫu nhiên để chọn 1000 video cho mình. Bạn thấy rằng 2% trong số những video này là video về mèo. Xác định cái nào sau đây là quan sát, biến, thống kê mẫu (giá trị được tính dựa trên mẫu quan sát) hoặc tham số quần thể.

- (a) Tỷ lệ phần trăm của tất cả các video trên YouTube là video về mèo.
- (b) 2%.
- (c) Một video trong mẫu của bạn.
- (d) Video có phải là video mèo hay không.

1.7 Sự hài lòng của khóa học giữa các nhóm. Một lớp đại học lớn có 160 sinh viên. Tất cả 160 sinh viên cùng tham gia các buổi giảng, nhưng được chia thành 4 nhóm, mỗi nhóm gồm 40 sinh viên, cho các buổi thực hành được quản lý bởi những trợ giảng khác nhau. Giảng viên muốn tiến hành một cuộc khảo sát để tìm hiểu mức độ hài lòng của sinh viên với khóa học, và ông tin rằng nhóm thực hành mà một sinh viên tham gia có thể ảnh hưởng đến mức độ hài lòng tổng thể của sinh viên đối với khóa học này.

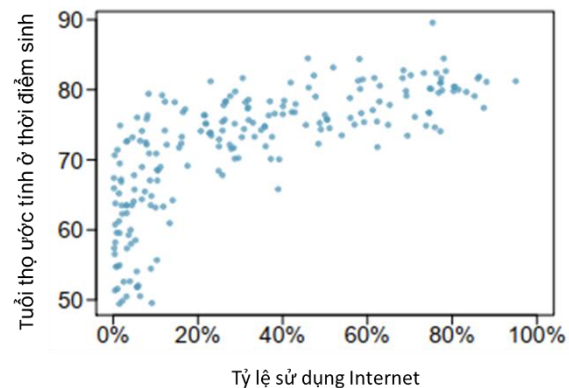
- (a) Đây là loại nghiên cứu nào?
- (b) Đề xuất một chiến lược lấy mẫu để thực hiện nghiên cứu này.

1.8 Đề xuất chỗ ở giữa các tòa ký túc xá. Tại một trường đại học, sinh viên năm nhất và năm hai sống trong các ký túc xá nằm ở khu vực phía đông của trường, trong khi sinh viên năm ba và năm tư sống trong các ký túc xá nằm ở khu vực phía tây. Giả sử bạn muốn thu thập ý kiến của sinh viên về một cấu trúc nhà ở mới mà ban quản lý trường đại học đang đề xuất, và bạn muốn đảm bảo rằng cuộc khảo sát của bạn đại diện công bằng cho ý kiến của sinh viên từ tất cả các năm học.

- (a) Đây là loại nghiên cứu nào?
- (b) Đề xuất một chiến lược lấy mẫu để thực hiện nghiên cứu này.

1.9 Sử dụng Internet và tuổi thọ. Đồ thị phân tán dưới đây được tạo ra như một phần của một nghiên cứu đánh giá mối quan hệ giữa tuổi thọ ước tính ở thời điểm sinh (tính đến năm 2014) và phần trăm người sử dụng Internet (tính đến năm 2009) ở 208 quốc gia có dữ liệu tương ứng.²³

- (a) Mô tả mối quan hệ giữa tuổi thọ và tỷ lệ người sử dụng Internet.
- (b) Đây là loại hình nghiên cứu gì?
- (c) Nêu một biến gây nhiễu có thể giải thích mối quan hệ này và mô tả tác động tiềm năng của nó.



1.10 Nghiên cứu về căng thẳng. Một nghiên cứu khảo sát một mẫu ngẫu nhiên gồm các học sinh trung học khỏe mạnh cho thấy họ có xu hướng gặp phải chuột rút cơ bắp nhiều hơn khi chịu áp lực. Nghiên cứu cũng nhận thấy rằng những học sinh này tiêu thụ nhiều cà phê hơn và ngủ ít hơn trong thời gian căng thẳng.

- (a) Đây là loại nghiên cứu nào?
- (b) Nghiên cứu này có thể được sử dụng để kết luận mối quan hệ nhân quả giữa căng thẳng gia tăng và chuột rút cơ bắp không?
- (c) Nêu các biến gây nhiễu có thể giải thích mối quan hệ quan sát được giữa căng thẳng gia tăng và chuột rút cơ.

1.11 Đánh giá các phương pháp lấy mẫu. Một trường đại học muốn xác định tỷ lệ sinh viên đại học ủng hộ mức phí mới 25 đô la mỗi năm để cải thiện hội sinh viên. Đối với từng phương pháp được đề xuất dưới đây, hãy cho biết liệu phương pháp đó có hợp lý hay không.

- (a) Khảo sát một mẫu ngẫu nhiên đơn gồm 500 sinh viên.
- (b) Phân tầng sinh viên theo lĩnh vực nghiên cứu của họ, sau đó lấy mẫu 10% học sinh từ mỗi tầng.
- (c) Phân nhóm sinh viên theo độ tuổi của họ (ví dụ: 18 tuổi trong một nhóm, 19 tuổi trong một nhóm, v.v.), sau đó ngẫu nhiên chọn ba nhóm và tiến hành khảo sát tất cả sinh viên trong các nhóm đó.

²³ CIA Factbook, Country Comparisons, 2014

1.12 Quay số ngẫu nhiên. Cuộc khảo sát Gallup sử dụng một quy trình gọi là gọi điện ngẫu nhiên bằng số, quy trình này tạo ra các số điện thoại dựa trên danh sách tất cả các mã vùng ở Mỹ, kết hợp với số lượng hộ gia đình dân cư tương ứng trong mỗi mã vùng. Tại sao cuộc khảo sát Gallup chọn phương pháp gọi điện ngẫu nhiên bằng số thay vì lấy số điện thoại từ danh bạ điện thoại.

1.13 Những người ghét sẽ ghét, nghiên cứu xác nhận. Một nghiên cứu được công bố trên Tạp chí *Nhân cách và Tâm lý học Xã hội* đã yêu cầu 200 đàn ông và phụ nữ được chọn ngẫu nhiên đánh giá cảm nhận của họ về các chủ đề khác nhau như cảm trại, chăm sóc sức khỏe, kiến trúc, mô hình, trò chơi ô chữ và Nhật Bản, nhằm đo lường thái độ của họ đối với các kích thích độc lập. Sau đó, các nhà nghiên cứu trình bày thông tin về một sản phẩm mới: lò vi sóng. Sản phẩm này không thực sự tồn tại, nhưng những người tham gia không biết điều đó, và họ được cung cấp ba đánh giá tích cực cùng ba đánh giá tiêu cực (giả tạo) về lò vi sóng. Kết quả cho thấy những người có phản ứng tích cực với các chủ đề về thái độ cũng có xu hướng đánh giá tích cực đối với lò vi sóng, trong khi những người phản ứng tiêu cực lại có xu hướng đánh giá tiêu cực với sản phẩm này. Các nhà nghiên cứu kết luận rằng có những người có xu hướng thích mọi thứ, trong khi một số khác thì không thích gì cả, và việc hiểu rõ hơn về những xu hướng này sẽ giúp nâng cao hiểu biết về tâm lý thái độ.²⁴

- (a) Các trường hợp trong nghiên cứu này là gì?
- (b) Các biến phản ứng được xác định trong nghiên cứu này là gì?
- (c) Các biến giải thích trong nghiên cứu này là gì?
- (d) Nghiên cứu có sử dụng phương pháp lấy mẫu ngẫu nhiên không?
- (e) Đây là một nghiên cứu quan sát hay một thí nghiệm? Vui lòng giải thích lý do của bạn.
- (f) Chúng ta có thể thiết lập mối liên hệ nhân quả giữa các biến giải thích và biến phản ứng không?
- (g) Kết quả của nghiên cứu có thể được khái quát hóa cho toàn bộ dân số hay không?

1.14 Quy mô gia đình. Giả sử chúng ta muốn ước tính quy mô hộ gia đình, trong đó "hộ gia đình" được định nghĩa là những người sống cùng nhau trong một ngôi nhà và chia sẻ chỗ ở. Nếu chúng ta chọn ngẫu nhiên học sinh tại một trường tiểu học và hỏi họ về quy mô hộ gia đình của mình, liệu điều này có phải là một thước đo chính xác cho quy mô hộ gia đình không? Hay mức trung bình mà chúng ta thu được từ các câu trả lời có thể bị thiên lệch? Nếu có, liệu nó sẽ đánh giá quá cao hay thấp giá trị thực của quy mô hộ gia đình?

1.15 Chiến lược lấy mẫu. Một học sinh có niềm đam mê với thống kê muốn khám phá mối quan hệ giữa thời gian mà học sinh dành cho các trang mạng xã hội và thành tích học tập của họ, vì vậy anh quyết định thực hiện một cuộc khảo sát. Dưới đây là những chiến lược nghiên cứu khác nhau mà anh sử dụng để thu thập dữ liệu. Trong mỗi trường hợp, hãy xác định phương pháp lấy mẫu đã được sử dụng và những thiên lệch có thể xảy ra.

- (a) Anh ta lấy mẫu ngẫu nhiên 40 sinh viên từ danh sách nghiên cứu, phát cho họ cuộc khảo sát và yêu cầu họ điền vào và mang trở lại vào ngày hôm sau.
- (b) Anh ấy chỉ phát khảo sát cho bạn bè của mình và yêu cầu mỗi người trong số họ hoàn thành khảo sát.
- (c) Anh ấy đăng một liên kết đến cuộc khảo sát trực tuyến trên Facebook và khuyến khích bạn bè của mình điền vào.
- (d) Anh ấy lấy mẫu ngẫu nhiên 5 lớp học và sau đó yêu cầu một mẫu ngẫu nhiên học sinh từ các lớp đó tham gia khảo sát.

1.16 Đọc báo. Dưới đây là trích đoạn từ hai bài báo được đăng trên *NY Times*:

- (a) Một bài báo có tựa đề *Rủi ro: Người hút thuốc dễ mắc chứng sa sút trí tuệ* hơn nói như sau:²⁵

"Các nhà nghiên cứu đã phân tích dữ liệu từ 23.123 thành viên chương trình bảo hiểm y tế đã tham gia vào một cuộc kiểm tra tự nguyện và khảo sát hành vi sức khỏe từ năm 1978 đến năm 1985, khi họ 50-60 tuổi. 23 năm sau, khoảng 25% nhóm bị sa sút trí tuệ, bao gồm 1.136 người mắc bệnh Alzheimer và 416 người mắc chứng sa sút trí tuệ mạch máu. Sau khi điều chỉnh các yếu tố khác, các nhà nghiên cứu kết luận rằng những người hút thuốc mỗi ngày có nguy cơ phát triển chứng mất trí nhớ cao hơn 37% so với những người không hút thuốc và nguy cơ tăng lên khi hút thuốc tăng lên; 44% cho một đến hai gói mỗi ngày; và gấp đôi rủi ro đối với hơn hai gói."

Dựa trên nghiên cứu này, liệu chúng ta có thể kết luận rằng việc hút thuốc gây ra chứng sa sút trí tuệ trong giai đoạn về già hay không? Giải thích lý do của bạn.

- (b) Một bài báo khác có tựa đề *The School Bully Is Sleepy* nói như sau:²⁶

"Nghiên cứu của Đại học Michigan, đã thu thập dữ liệu khảo sát từ phụ huynh về thói quen ngủ của từng đứa trẻ và yêu cầu cả cha mẹ và giáo viên đánh giá các mối quan tâm về hành vi. Khoảng một phần ba số học sinh được nghiên cứu được phụ huynh hoặc giáo viên xác định là có vấn đề với hành vi gây rối hoặc bất nết. Các nhà nghiên cứu phát hiện ra rằng những đứa trẻ có vấn đề về hành vi và những đứa trẻ được xác định là kẻ bất nết có khả năng xuất hiện các triệu chứng rối loạn giấc ngủ cao gấp đôi. Một người bạn của bạn vừa đọc một bài báo và nói, "Nghiên cứu cho thấy rằng rối loạn giấc ngủ dẫn đến việc trẻ em bị bất nết ở trường." Liệu tuyên bố này có hợp lý không? Nếu không, bạn sẽ mô tả kết luận có thể rút ra từ nghiên cứu này như thế nào?

²⁴ Justin Hepler and Dolores Albarracín. "Attitudes without objects - Evidence for a dispositional attitude, its measurement, and its consequences". In: *Journal of personality and social psychology* 104.6 (2013), p. 1060.

²⁵ R.C. Rabin. "Risks: Smokers Found More Prone to Dementia". In: *New York Times* (2010).

²⁶ T. Parker-Pope. "The School Bully Is Sleepy". In: *New York Times* (2011).

1.4 Thử nghiệm

Các nghiên cứu mà các nhà nghiên cứu chỉ định phương pháp điều trị cho các trường hợp được gọi là **thử nghiệm**. Khi nhiệm vụ này bao gồm ngẫu nhiên, ví dụ: sử dụng tung đồng xu để quyết định bệnh nhân nhận được phương pháp điều trị, nó được gọi là **thử nghiệm ngẫu nhiên**. Các thử nghiệm ngẫu nhiên về cơ bản rất quan trọng khi cố gắng chỉ ra mối liên hệ nhân quả giữa hai biến.

1.4.1 Nguyên tắc thiết kế thử nghiệm

Các thử nghiệm ngẫu nhiên thường được xây dựng dựa trên bốn nguyên tắc.

Kiểm soát. Các nhà nghiên cứu chỉ định phương pháp điều trị cho các trường hợp và họ cố gắng hết sức để **kiểm soát** bất kỳ sự khác biệt nào khác trong các nhóm.²⁷ Ví dụ, khi bệnh nhân uống thuốc ở dạng viên thuốc, một số bệnh nhân uống thuốc chỉ với một ngụm nước trong khi những người khác có thể uống thuốc với cả một cốc nước. Để kiểm soát ảnh hưởng của việc tiêu thụ nước, bác sĩ có thể yêu cầu tất cả bệnh nhân cùng uống một cốc nước khoảng 355 ml (12 ounce) với viên thuốc.

Ngẫu nhiên. Các nhà nghiên cứu ngẫu nhiên phân chia bệnh nhân vào các nhóm điều trị để tính đến các biến số không thể kiểm soát. Ví dụ, một số bệnh nhân có thể dễ mắc bệnh hơn những người khác do thói quen ăn uống của họ. Ngẫu nhiên hóa bệnh nhân vào nhóm điều trị hoặc nhóm đối chứng giúp loại bỏ sự khác biệt như vậy và nó cũng ngăn chặn sự thiên vị ngẫu nhiên tham gia vào nghiên cứu.

Sao chép. Các nhà nghiên cứu càng quan sát nhiều trường hợp, họ càng có thể ước tính chính xác ảnh hưởng của biến giải thích đối với phản ứng. Trong một nghiên cứu duy nhất, chúng tôi **tái tạo** bằng cách thu thập một mẫu đủ lớn. Ngoài ra, một nhóm các nhà khoa học có thể sao chép toàn bộ nghiên cứu để xác minh một phát hiện trước đó.

Chặn. Các nhà nghiên cứu đôi khi biết hoặc nghi ngờ rằng các biến số, ngoài điều trị, ảnh hưởng đến phản ứng. Trong những trường hợp này, trước tiên họ có thể nhóm các cá nhân dựa trên biến này thành các **nhóm** và sau đó chọn ngẫu nhiên các trường hợp trong mỗi khối thành các nhóm điều trị. Chiến lược này thường được gọi là **chặn**. Ví dụ, nếu chúng ta đang xem xét tác dụng của một loại thuốc đối với các cơn đau tim, trước tiên chúng ta có thể chia bệnh nhân trong nghiên cứu thành các nhóm có nguy cơ thấp và nguy cơ cao, sau đó chỉ định ngẫu nhiên một nửa bệnh nhân từ mỗi nhóm cho nhóm đối chứng và nửa còn lại cho nhóm điều trị, như thể hiện trong Hình 1.16. Chiến lược này đảm bảo mỗi nhóm điều trị có số lượng bệnh nhân nguy cơ thấp và nguy cơ cao bằng nhau.

Điều quan trọng là phải kết hợp ba nguyên tắc thiết kế thử nghiệm đầu tiên vào bất kỳ nghiên cứu nào và cuốn sách này mô tả các phương pháp áp dụng để phân tích dữ liệu từ các thử nghiệm như vậy. Chặn là một kỹ thuật tiên tiến hơn một chút và các phương pháp thống kê trong cuốn sách này có thể được mở rộng để phân tích dữ liệu được thu thập bằng cách chặn.

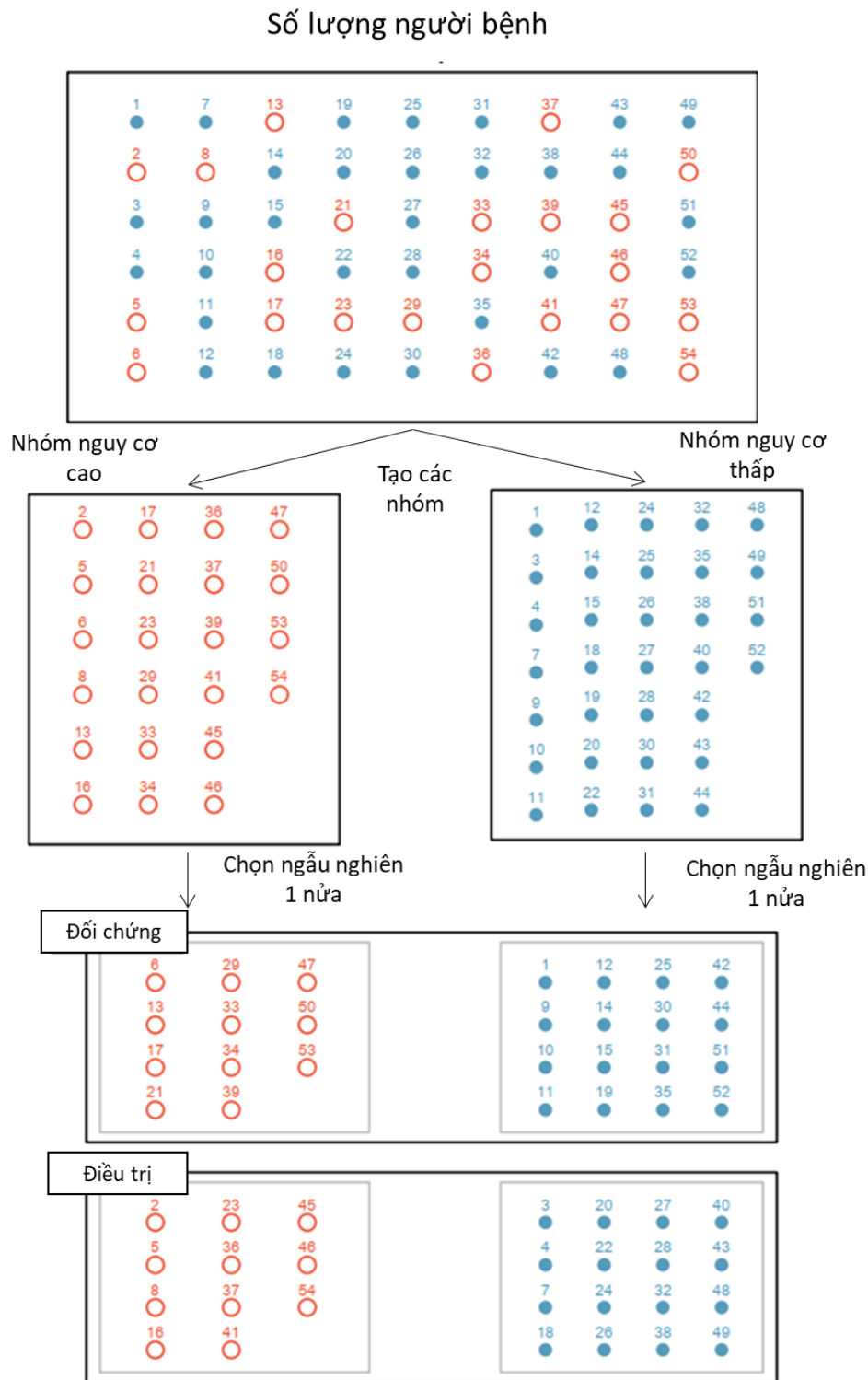
1.4.2 Giảm thiên vị (bias) trong các thử nghiệm trên người

Các thử nghiệm ngẫu nhiên là tiêu chuẩn vàng để thu thập dữ liệu, nhưng chúng không đảm bảo một quan điểm khách quan về mối quan hệ nhân quả trong mọi trường hợp. Các nghiên cứu trên con người là những ví dụ hoàn hảo mà sự thiên vị có thể vô tình phát sinh. Ở đây chúng tôi xem xét lại một nghiên cứu trong đó một loại thuốc mới được sử dụng để điều trị bệnh nhân đau tim. Đặc biệt, các nhà nghiên cứu muốn biết liệu thuốc có làm giảm tử vong ở bệnh nhân hay không.

Các nhà nghiên cứu này đã thiết kế một thử nghiệm ngẫu nhiên vì họ muốn rút ra kết luận nhân quả về tác dụng của thuốc. Tình nguyện viên nghiên cứu²⁸ được xếp ngẫu nhiên vào hai nhóm nghiên cứu. Một nhóm, **nhóm điều trị**, đã nhận được thuốc. Nhóm còn lại, được gọi là **nhóm đối chứng**, không được điều trị bằng thuốc.

²⁷Đây là một khái niệm khác với nhóm *đối chứng*, mà chúng ta thảo luận trong nguyên tắc thứ hai và trong Phần 1.4.2.

²⁸Đối tượng con người thường được gọi là **bệnh nhân**, **tình nguyện viên** hoặc **người tham gia nghiên cứu**.



Hình 1.16: Chặn bằng cách sử dụng một biến mô tả nguy cơ bệnh nhân. Trước tiên, bệnh nhân được chia thành các nhóm có nguy cơ thấp và nguy cơ cao, sau đó mỗi nhóm được chia đều thành các nhóm điều trị bằng cách chọn mẫu ngẫu nhiên. Chiến lược này đảm bảo đại diện bình đẳng của bệnh nhân trong mỗi nhóm điều trị từ cả nhóm nguy cơ thấp và nguy cơ cao.

Hãy tưởng tượng bạn là một người tham gia trong một nghiên cứu. Nếu bạn thuộc nhóm điều trị, bạn sẽ được cung cấp một loại thuốc mới mà bạn tin rằng sẽ mang lại lợi ích cho sức khỏe của mình. Ngược lại, một người trong nhóm đối chứng lại không nhận được thuốc và chỉ ngồi chờ đợi, hy vọng rằng sự tham gia của mình không làm tăng nguy cơ tử vong. Từ góc độ này, có thể nhận thấy hai tác động quan trọng: một là hiệu quả cụ thể của loại thuốc, và hai là tác động cảm xúc, mặc dù khó đo đếm.

Các nhà nghiên cứu thường không chú ý đến hiệu ứng cảm xúc, điều này có thể làm sai lệch kết quả nghiên cứu. Để khắc phục vấn đề này, họ thường giữ bí mật về việc bệnh nhân thuộc nhóm nào. Khi bệnh nhân không biết mình được điều trị hay không, họ được coi là bị **"làm mù"**. Tuy nhiên, một vấn đề phát sinh là nếu bệnh nhân không nhận được điều trị, họ sẽ biết mình thuộc nhóm đối chứng. Giải pháp cho vấn đề này là cung cấp cho bệnh nhân trong nhóm đối chứng một phương pháp điều trị giả, gọi là **giả dược**. Một giả dược hiệu quả là yếu tố quan trọng để đảm bảo tính làm mù của nghiên cứu. Ví dụ điển hình về giả dược là một viên thuốc đường được thiết kế giống như thuốc thực. Thông thường, việc sử dụng giả dược có thể dẫn đến cải thiện nhẹ ở bệnh nhân, hiện tượng này được gọi là **hiệu ứng giả dược**.

Không chỉ bệnh nhân cần được làm mù, mà các bác sĩ và nhà nghiên cứu cũng có thể bị thiên vị trong quá trình nghiên cứu. Nếu bác sĩ biết rằng một bệnh nhân đang nhận điều trị thực tế, họ có thể vô tình dành nhiều sự chú ý và chăm sóc hơn cho bệnh nhân đó so với những bệnh nhân sử dụng giả dược. Để giảm thiểu sự thiên vị này, nhiều nghiên cứu hiện đại áp dụng phương pháp **làm mù đôi**, trong đó cả bác sĩ và nhà nghiên cứu, cũng như bệnh nhân, đều không biết ai đang nhận điều trị thực sự và ai đang sử dụng giả dược. Phương pháp này đã cho thấy có hiệu quả trong việc đảm bảo tính khách quan của nghiên cứu.²⁹

BÀI TẬP THỰC HÀNH 1.16

Nhìn lại nghiên cứu trong Phần 1.1, nơi các nhà nghiên cứu đang kiểm tra xem liệu stent có hiệu quả trong việc giảm đột quỵ ở những bệnh nhân có nguy cơ hay không. Đây có phải là một thử nghiệm không? Nghiên cứu có bị làm mù không? Nó có bị làm mù đôi không?³⁰

BÀI TẬP THỰC HÀNH 1.17

Đối với việc học trong Phần 1.1, các nhà nghiên cứu có thể sử dụng giả dược không? Nếu vậy, giả dược đó sẽ trông như thế nào?³¹

Bạn có thể có nhiều câu hỏi về đạo đức của việc thực hiện phẫu thuật giả để tạo giả dược sau khi đọc Bài tập Thực hành có hướng dẫn 1.17. Những câu hỏi này có thể xuất hiện trong bối cảnh thử nghiệm chung, nơi một phương pháp điều trị tiềm năng hữu ích không được triển khai trên người bệnh trong nhóm đối chứng. Điểm khác biệt lớn là phẫu thuật giả thường mang lại rủi ro lớn, trong khi việc từ chối điều trị chỉ duy trì nguy cơ vốn có của bệnh nhân.

Luôn có nhiều quan điểm về thử nghiệm và giả dược, và hiếm khi rõ ràng là "đúng" về mặt đạo đức. Ví dụ, sử dụng phẫu thuật giả khi nó tạo ra rủi ro cho bệnh nhân có đạo đức không? Tuy nhiên, nếu chúng ta không sử dụng phẫu thuật giả, chúng ta có thể thúc đẩy việc sử dụng một phương pháp điều trị tốn kém mà không có tác dụng thực sự; Nếu điều này xảy ra, tiền và các nguồn lực khác sẽ được chuyển hướng khỏi các phương pháp điều trị khác được biết là hữu ích. Cuối cùng, đây là một tình huống khó khăn mà chúng ta không thể bảo vệ hoàn hảo cả những bệnh nhân đã tình nguyện tham gia nghiên cứu và những bệnh nhân có thể được hưởng lợi (hoặc không) từ việc điều trị trong tương lai.

²⁹ Luôn có một số nhà nghiên cứu tham gia vào nghiên cứu biết bệnh nhân nào đang được điều trị nào. Tuy nhiên, họ không tương tác với bệnh nhân của nghiên cứu và không cho các chuyên gia chăm sóc sức khỏe bị mù biết ai đang được điều trị phương pháp gì.

³⁰ Các nhà nghiên cứu đã chỉ định bệnh nhân vào các nhóm điều trị của họ, vì vậy nghiên cứu này là một thử nghiệm. Tuy nhiên, bệnh nhân có thể phân biệt được phương pháp điều trị mà họ nhận được, vì vậy nghiên cứu này không mù. Nghiên cứu không thể mù đôi vì nó không bị mù.

³¹ Cuối cùng, chúng ta có thể khiến bệnh nhân nghĩ rằng họ đã được điều trị sau một cuộc phẫu thuật không? Trên thực tế, chúng ta có thể, và một số thử nghiệm sử dụng cái gọi là **phẫu thuật giả**. Trong phẫu thuật giả, bệnh nhân trải qua phẫu thuật, nhưng bệnh nhân không được điều trị đầy đủ, mặc dù họ vẫn sẽ nhận được hiệu ứng giả dược.

Bài tập

1.1 Ánh sáng và hiệu suất kiểm tra. Một nghiên cứu được thiết kế để kiểm tra ảnh hưởng của mức độ ánh sáng đối với kết quả thi của học sinh. Nhà nghiên cứu tin rằng mức độ ánh sáng có thể có tác động khác nhau đối với nam và nữ, vì vậy muốn đảm bảo cả hai đều được đại diện như nhau trong mỗi lần điều trị. Các phương pháp xử lý là đèn huỳnh quang trên cao, đèn màu vàng trên cao, không chiếu sáng trên cao (chỉ có đèn bàn).

- (a) Biến phản hồi là gì?
- (b) Biến giải thích là gì? Các cấp độ của nó là gì?
- (c) Biến chặn là gì? Các cấp độ của nó là gì?

1.2 Bổ sung vitamin. Để đánh giá hiệu quả của việc dùng liều lượng lớn vitamin C trong việc giảm thời gian cảm lạnh thông thường, các nhà nghiên cứu đã tuyển dụng 400 tình nguyện viên khỏe mạnh từ nhân viên và sinh viên tại một trường đại học. Một phần tư số bệnh nhân được chỉ định giả dược, và số còn lại được chia đều giữa 1g Vitamin C, 3g Vitamin C hoặc 3g Vitamin C cộng với các chất phụ gia được dùng khi khởi phát cảm lạnh trong hai ngày tiếp theo. Tất cả các viên nén đều có hình thức và bao bì giống hệt nhau. Các y tá đưa thuốc kê đơn cho bệnh nhân biết bệnh nhân nào được điều trị nào, nhưng các nhà nghiên cứu đánh giá bệnh nhân khi họ bị bệnh thì không. Không có sự khác biệt đáng kể nào được quan sát thấy trong bất kỳ thước đo nào về thời gian lạnh hoặc mức độ nghiêm trọng giữa bốn nhóm và nhóm giả dược có thời gian triệu chứng ngắn nhất.³²

- (a) Đây là một thí nghiệm hay một nghiên cứu quan sát? Tại sao?
- (b) Các biến giải thích và phản ứng trong nghiên cứu này là gì?
- (c) Các bệnh nhân có bị mù quáng trước việc điều trị của họ không?
- (d) Nghiên cứu này có bị mù đôi không?
- (e) Những người tham gia cuối cùng có thể chọn có sử dụng thuốc được kê đơn cho họ hay không. Chúng ta có thể mong đợi rằng không phải tất cả họ sẽ tuân thủ và uống thuốc của họ. Điều này có đưa ra một biến gây nhiễu cho nghiên cứu không? Giải thích lý do của bạn.

1.3 Ánh sáng, tiếng ồn và hiệu suất thi. Một nghiên cứu được thiết kế để kiểm tra ảnh hưởng của mức độ ánh sáng và độ ồn đối với kết quả thi của học sinh. Nhà nghiên cứu tin rằng mức độ ánh sáng và tiếng ồn có thể có tác động khác nhau đối với nam và nữ, vì vậy muốn đảm bảo rằng cả hai đều được đại diện như nhau trong mỗi lần điều trị. Các phương pháp xử lý ánh sáng được xem xét là đèn huỳnh quang trên cao, đèn màu vàng, không chiếu sáng trên cao (chỉ đèn bàn). Các phương pháp xử lý tiếng ồn được xem xét là không có tiếng ồn, tiếng ồn xây dựng và tiếng ồn của con người.

- (a) Đây là loại nghiên cứu nào?
- (b) Có bao nhiêu yếu tố được xem xét trong nghiên cứu này? Xác định họ và mô tả cấp độ của họ.
- (c) Vai trò của biến giới tính trong nghiên cứu này là gì?

1.4 Âm nhạc và học tập. Bạn muốn tiến hành một thí nghiệm trong lớp để xem liệu học sinh có học tốt hơn nếu họ học không có bất kỳ bản nhạc nào, với âm nhạc không có lời (nhạc cụ) hoặc với nhạc có lời bài hát. Phác thảo ngắn gọn một thiết kế cho nghiên cứu này.

1.5 Sở thích soda. Bạn muốn tiến hành một thí nghiệm trong lớp để xem các bạn cùng lớp của bạn thích hương vị của Coke thông thường hay Diet Coke. Phác thảo ngắn gọn một thiết kế cho nghiên cứu này.

1.6 Tập thể dục và sức khỏe tâm thần. Một nhà nghiên cứu quan tâm đến tác động của tập thể dục đối với sức khỏe tâm thần và ông đề xuất nghiên cứu sau: Sử dụng lấy mẫu ngẫu nhiên phân tầng để đảm bảo tỷ lệ đại diện cho những người từ 18-30, 31-40 và 41-55 tuổi từ dân số. Tiếp theo, chỉ định ngẫu nhiên một nửa số đối tượng từ mỗi nhóm tuổi tập thể dục hai lần một tuần và hướng dẫn những người còn lại không tập thể dục. Tiến hành kiểm tra sức khỏe tâm thần vào đầu và cuối nghiên cứu, và so sánh kết quả.

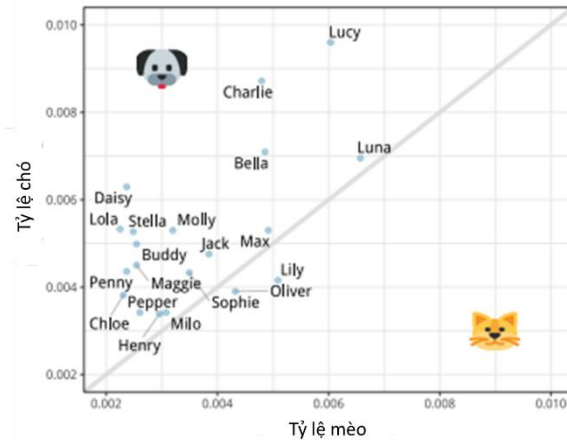
- (a) Đây là loại nghiên cứu nào?
- (b) Các nhóm điều trị và đối chứng trong nghiên cứu này là gì?
- (c) Nghiên cứu này có sử dụng chặn không? Nếu vậy, biến chặn là gì?
- (d) Nghiên cứu này có sử dụng tính năng làm mù không?
- (e) Nhận xét về việc liệu kết quả của nghiên cứu có thể được sử dụng để thiết lập mối quan hệ nhân quả giữa tập thể dục và sức khỏe tâm thần hay không, và chỉ ra liệu kết luận có thể được khái quát hóa cho dân số nói chung hay không.
- (f) Giả sử bạn được giao nhiệm vụ xác định xem nghiên cứu được đề xuất này có nên nhận được tài trợ hay không. Bạn có bất kỳ sự dè dặt nào về đề xuất nghiên cứu không?

³² C. Aulner et al. "Mega-dose vitamin C in treatment of the common cold: a randomised controlled trial". In: *Medical Journal of Australia* 175.7 (2001), pp. 359–362.

Bài tập chương

1.7 Tên thú cưng. Thành phố Seattle, WA có một cổng dữ liệu mở bao gồm các vật nuôi được đăng ký trong thành phố. Đối với mỗi vật nuôi đã đăng ký, chúng tôi có thông tin về tên và loài của thú cưng. Biểu đồ sau đây vẽ tỷ lệ chó có tên riêng so với tỷ lệ mèo có cùng tên. 20 tên mèo và chó phổ biến nhất được hiển thị. Đường chéo trên biểu đồ là đường $x = y$; nếu một cái tên xuất hiện trên dòng này, mức độ phổ biến của cái tên sẽ giống hệt nhau đối với chó và mèo.

- Những dữ liệu này được thu thập như một phần của thử nghiệm hay một nghiên cứu quan sát?
- Tên chó phổ biến nhất là gì? Tên mèo phổ biến nhất là gì?
- Những tên nào phổ biến hơn cho mèo so với chó?
- Mối quan hệ giữa hai biến là tích cực hay tiêu cực? Điều này có nghĩa là gì trong bối cảnh của dữ liệu?



1.8 Nghiên cứu về căng thẳng. Trong một nghiên cứu đánh giá mối quan hệ giữa căng thẳng và chuột rút cơ, các đối tượng được chia thành hai nhóm. Một nửa trong số họ được chỉ định ngẫu nhiên để trải nghiệm căng thẳng gia tăng bằng cách cho vào một thang máy rơi nhanh chóng và dừng đột ngột. Nhóm còn lại không chịu tác động của căng thẳng, hoặc chỉ trải qua căng thẳng cơ bản.

- Đây là loại nghiên cứu nào?
- Nghiên cứu này có thể được sử dụng để kết luận mối quan hệ nhân quả giữa căng thẳng gia tăng và chuột rút cơ không?

1.9 Hạt Chia và giảm cân. Chia Pets – những bức tượng nhỏ bằng đất nung có mái tóc xanh lông xù – đã giúp cây chia trở thành một cái tên quen thuộc. Tuy nhiên, chia đã có được một danh tiếng hoàn toàn mới như một chất bổ sung trong chế độ ăn kiêng. Trong một nghiên cứu năm 2009, một nhóm các nhà nghiên cứu đã tuyển dụng 38 nam giới và chia họ thành hai nhóm ngẫu nhiên: nhóm điều trị và nhóm đối chứng. Họ cũng tuyển thêm 38 phụ nữ và phân ngẫu nhiên một nửa vào nhóm điều trị và nửa còn lại vào nhóm đối chứng. Một nhóm được cho ăn 25 gram hạt chia hai lần mỗi ngày, trong khi nhóm còn lại nhận giả dược. Các đối tượng tham gia nghiên cứu là những người tình nguyện viên. Sau 12 tuần, các nhà khoa học không phát hiện thấy sự khác biệt đáng kể giữa các nhóm về cảm giác thèm ăn hay giảm cân.³³

- Đây là loại nghiên cứu nào?
- Các phương pháp điều trị thử nghiệm và kiểm soát trong nghiên cứu này là gì?
- Phương pháp chặn đã được sử dụng trong nghiên cứu này chưa? Nếu vậy, biến chặn là gì?
- Quy trình làm mù đã được sử dụng trong nghiên cứu này chưa?
- Nhận xét về việc liệu chúng ta có thể đưa ra một tuyên bố nhân quả hay không, và chỉ ra liệu chúng ta có thể khái quát hóa kết luận cho dân số nói chung hay không.

1.10 Khảo sát hội đồng thành phố. Một hội đồng thành phố đã yêu cầu thực hiện một cuộc khảo sát hộ gia đình ở một khu vực ngoại ô của thành phố. Khu vực này được chia thành nhiều khu phố khác nhau, mỗi khu phố có đặc điểm riêng: một số có những ngôi nhà lớn, một số chỉ có căn hộ, trong khi những khu phố khác là sự kết hợp đa dạng của các loại hình nhà ở. Đối với mỗi phần dưới đây, hãy xác định các phương pháp lấy mẫu được mô tả và phân tích ưu, nhược điểm thống kê của từng phương pháp trong bối cảnh của thành phố.

- Lấy mẫu ngẫu nhiên 200 hộ gia đình từ thành phố.
- Chia thành phố thành 20 khu phố và lấy mẫu 10 hộ gia đình từ mỗi khu phố.
- Chia thành phố thành 20 khu phố, lấy mẫu ngẫu nhiên 3 khu phố, sau đó lấy mẫu tất cả các hộ gia đình từ 3 khu phố đó.
- Chia thành phố thành 20 khu phố, lấy mẫu ngẫu nhiên 8 khu phố, sau đó lấy mẫu ngẫu nhiên 50 hộ gia đình từ các khu vực đó.
- Lấy mẫu 200 hộ gia đình gần văn phòng hội đồng thành phố nhất.

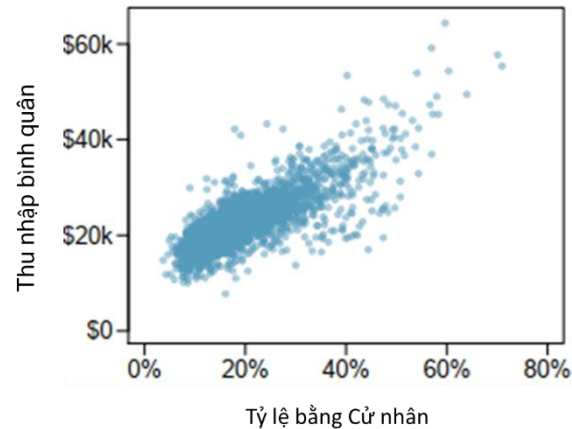
³³ D.C. Nieman et al. "Chia seed does not promote weight loss or alter disease risk factors in overweight adults". In: *Nutrition Research* 29.6 (2009), pp. 414–418.

1.11 Lý luận sai sót. Xác định (các) lỗi hổng trong lý luận trong các tình huống sau. Giải thích những gì các cá nhân trong nghiên cứu nên làm khác nếu họ muốn đưa ra kết luận mạnh mẽ như vậy.

- Học sinh tại một trường tiểu học được cung cấp một bảng câu hỏi mà các em được yêu cầu trả lại sau khi cha mẹ của chúng đã hoàn thành nó. Một trong những câu hỏi được đặt ra là, "Bạn có thấy rằng lịch trình làm việc của bạn khiến bạn khó dành thời gian cho con cái sau giờ học không?" Trong số các phụ huynh trả lời, 85% nói "không". Dựa trên những kết quả này, các quan chức nhà trường kết luận rằng đại đa số phụ huynh không gặp khó khăn khi dành thời gian cho con cái của họ sau giờ học.
- Một cuộc khảo sát được thực hiện trên một mẫu ngẫu nhiên đơn gồm 1.000 phụ nữ gần đây đã sinh con, hỏi họ về việc họ có hút thuốc khi mang thai hay không. Một cuộc khảo sát tiếp theo hỏi xem trẻ có vấn đề về hô hấp hay không được thực hiện 3 năm sau đó. Tuy nhiên, chỉ có 567 người trong số những phụ nữ này được liên lạc tại cùng một địa chỉ. Nhà nghiên cứu báo cáo rằng 567 phụ nữ này là đại diện cho tất cả các bà mẹ.
- Một bác sĩ chỉnh hình thực hiện một bảng câu hỏi cho 30 bệnh nhân của mình không có bất kỳ vấn đề nào về khớp và thấy rằng 20 người trong số họ thường xuyên chạy bộ. Ông kết luận rằng chạy bộ làm giảm nguy cơ mắc các vấn đề về khớp.

1.12 Thu nhập và giáo dục ở các quận của Hoa Kỳ. Biểu đồ phân tán dưới đây cho thấy mối quan hệ giữa thu nhập bình quân đầu người (tính bằng hàng nghìn đô la) và tỷ lệ phần trăm dân số có bằng cử nhân tại 3.143 quận ở Hoa Kỳ vào năm 2010.

- Các biến giải thích và phản hồi ở đây là gì?
- Mô tả mối quan hệ giữa hai biến. Thảo luận về các quan sát bất thường, nếu có.
- Chúng ta có thể kết luận rằng có bằng cử nhân làm tăng thu nhập của một người không?



1.13 Ăn ngon hơn, cảm thấy tốt hơn? Trong một nghiên cứu sức khỏe cộng đồng nhằm tìm hiểu ảnh hưởng của việc tiêu thụ trái cây và rau quả đối với sức khỏe tâm lý của thanh niên, các tham gia được chia ngẫu nhiên thành ba nhóm: (1) nhóm có chế độ ăn uống bình thường, (2) nhóm tham gia vào can thiệp sinh thái tạm thời, nhận nhắc nhở qua tin nhắn văn bản để tăng cường lượng tiêu thụ trái cây và rau quả, kèm theo phiếu mua hàng, và (3) nhóm can thiệp trái cây và rau quả, trong đó những người tham gia được cung cấp thêm hai khẩu phần trái cây và rau quả tươi mỗi ngày bên cạnh chế độ ăn uống bình thường của họ.³⁴

- Đây là loại nghiên cứu nào?
- Xác định các biến giải thích và phản hồi.
- Nhận xét về việc liệu kết quả của nghiên cứu có thể được khái quát hóa cho dân số hay không.
- Nhận xét về việc liệu kết quả của nghiên cứu có thể được sử dụng để thiết lập các mối quan hệ nhân quả hay không.
- Một bài báo đưa tin về nghiên cứu cho biết, "Kết quả của nghiên cứu này cung cấp bằng chứng cho thấy cho thanh niên ăn trái cây và rau quả tươi có thể mang lại lợi ích tâm lý, ngay cả trong một khoảng thời gian ngắn". Bạn đề nghị sửa đổi tuyên bố này như thế nào để nó có thể được hỗ trợ bởi nghiên cứu?

³⁴ Tamlin S Conner et al. "Let them eat fruit! The effect of fruit and vegetable consumption on psychological well-being in young adults: A randomized controlled trial". In: *PloS one* 12.2 (2017), e0171206.

1.14 Màn hình, thanh thiếu niên và sức khỏe tâm lý. Trong một nghiên cứu sử dụng ba bộ dữ liệu quy mô lớn đại diện cho toàn quốc từ Ireland, Hoa Kỳ và Vương quốc Anh ($n = 17.247$), các thanh thiếu niên từ 12 đến 15 tuổi được yêu cầu ghi lại thời gian sử dụng màn hình và trả lời các câu hỏi về cảm xúc và hành vi của họ. Các câu trả lời này sau đó được sử dụng để tính toán một chỉ số sức khỏe tâm lý. Dữ liệu bổ sung cũng được thu thập và đưa vào phân tích, bao gồm giới tính và độ tuổi của mỗi trẻ, cũng như trình độ học vấn của mẹ, sắc tộc, tình trạng lo âu tâm lý và tình trạng việc làm. Nghiên cứu kết luận rằng không có nhiều bằng chứng rõ ràng cho thấy thời gian sử dụng màn hình làm giảm sức khỏe của thanh thiếu niên.³⁵

- Đây là loại nghiên cứu nào?
- Xác định các biến giải thích.
- Xác định biến đáp ứng.
- Nhận xét về việc liệu kết quả của nghiên cứu có thể được khái quát hóa cho dân số hay không và tại sao.
- Nhận xét về việc liệu kết quả của nghiên cứu có thể được sử dụng để thiết lập các mối quan hệ nhân quả hay không.

1.15 Dự án Cảnh sát mở Stanford. Dự án Cảnh sát Mở Stanford thu thập, phân tích và công bố các hồ sơ về việc dừng xe của các cơ quan thực thi pháp luật trên khắp Hoa Kỳ. Mục tiêu của họ là hỗ trợ các nhà nghiên cứu, nhà báo và chính sách điều tra và cải thiện mối quan hệ giữa cảnh sát và cộng đồng. Dưới đây là một đoạn trích từ bảng tóm tắt được tạo ra dựa trên dữ liệu thu thập được trong khuôn khổ dự án này.

Quận	Bang	Chủng tộc của tài xế	Số lần dừng xe mỗi năm	% xe bị lục soát	% tài xế bị bắt
Quận Apaice	Arizona	Da đen	266	0.08	0.02
Quận Apaice	Arizona	Gốc Tây Ban Nha	1008	0.05	0.02
Quận Apaice	Arizona	Da trắng	6322	0.02	0.01
Quận Cochise	Arizona	Da đen	1169	0.05	0.01
Quận Cochise	Arizona	Gốc Tây Ban Nha	9453	0.04	0.01
Quận Cochise	Arizona	Da trắng	10826	0.02	0.01
...	...	...	...	...	...
Quận Wood	Wisconsin	Da đen	16	0.24	0.1
Quận Wood	Wisconsin	Gốc Tây Ban Nha	27	0.04	0.03
Quận Wood	Wisconsin	Da trắng	1157	0.03	0.03

- Những biến nào đã được thu thập trên mỗi điểm dừng giao thông riêng lẻ để tạo cho bảng tóm tắt ở trên?
- Nêu rõ mỗi biến là biến định lượng hay biến định tính. Nếu là biến định lượng, hãy cho biết nó là liên tục hay rời rạc. Nếu là biến định tính, hãy cho biết nó có phải là biến thứ hạng hay không.
- Giả sử chúng ta muốn đánh giá xem tỷ lệ tìm kiếm xe có khác nhau đối với các tay đua thuộc các chủng tộc khác nhau hay không. Trong phân tích này, biến nào sẽ là biến đáp ứng và biến nào sẽ là biến giải thích?

1.16 Phóng vũ trụ. Bảng tổng hợp sau đây cho thấy số lần phóng vũ trụ ở Mỹ theo loại cơ quan phóng và kết quả của lần phóng (thành công hoặc thất bại).³⁷

	1957 - 1999		2000 - 2018	
	Thành công	Thất bại	Thành công	Thất bại
Tư nhân	13	295	10	562
Nhà nước	281	3751	33	711
Khởi nghiệp	-	-	5	65

- Những biến nào đã được thu thập trong mỗi lần ra mắt để tạo vào bảng tóm tắt ở trên?
- Nêu rõ mỗi biến là số hay phân loại. Nếu là số, hãy cho biết nó là liên tục hay rời rạc. Nếu phân loại, hãy cho biết nó có phải là thứ tự hay không.
- Giả sử chúng ta muốn đánh giá xem tỷ lệ tìm kiếm phương tiện có khác nhau giữa các lái xe thuộc các chủng tộc khác nhau hay không. Trong phân tích này, biến nào sẽ là biến phản ứng và biến nào sẽ là biến giải thích?

³⁵ Amy Orben and AK Baukney-Przybylski. "Screens, Teens and Psychological Well-Being: Evidence from three time-use diary studies". In: *Psychological Science* (2018).

³⁶ Emma Pierson et al. "A large-scale analysis of racial disparities in police stops across the United States". In: *arXiv preprint arXiv:1706.05678* (2017).

³⁷ JSR Launch Vehicle Database, A comprehensive list of suborbital space launches, 2019 Feb 10 Edition.