

Chương 2

Mô tả dữ liệu

2.1 Kiểm tra dữ liệu định lượng

2.2 Xem xét dữ liệu định tính

2.3 Phân tích tình huống: Vắc-xin sốt rét

Chương này tập trung vào các nguyên lý và phương pháp xây dựng thống kê mô tả và biểu đồ. Chúng tôi sử dụng phần mềm thống kê để tạo ra các bảng tổng hợp và biểu đồ được trình bày trong chương cũng như toàn bộ cuốn sách. Tuy nhiên, do đây có thể là lần đầu tiên bạn tiếp cận những khái niệm này, chúng tôi sẽ trình bày chi tiết phương pháp xây dựng và diễn giải. Việc nắm vững nội dung của chương này là tiền đề quan trọng để hiểu và ứng dụng các phương pháp phân tích thống kê được giới thiệu trong các phần sau của cuốn sách.



Để biết video, trang trình bày và các tài nguyên khác, vui lòng truy cập

www.openintro.org/os

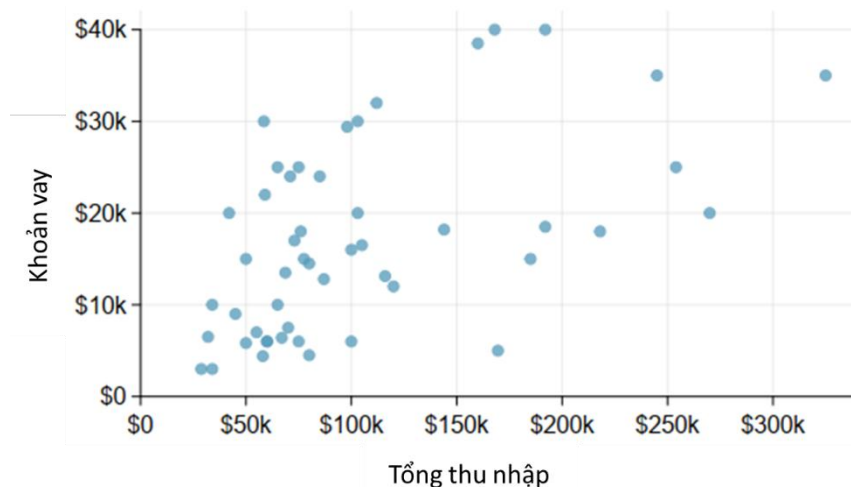
2.1 Kiểm tra dữ liệu định lượng

Trong chương này, chúng ta sẽ khám phá các kỹ thuật mô tả biến số định lượng. Ví dụ, xem xét biến số về số tiền vay trong tập dữ liệu loan50, biểu thị cho quy mô của 50 khoản vay trong tập dữ liệu. Đây là một biến số định lượng, vì chúng ta có thể đo lường và so sánh sự khác biệt về giá trị giữa hai khoản vay. Ngược lại, mã vùng và mã bưu chính không phải là biến số định lượng mà thuộc nhóm biến phân loại.

Trong suốt phần này và phần tiếp theo, chúng ta sẽ áp dụng các phương pháp trên tập dữ liệu loan50 và county, vốn đã được giới thiệu trong Mục 1.2. Nếu bạn muốn xem lại các biến trong từng tập dữ liệu, vui lòng tham khảo Hình 1.3 và Hình 1.5.

2.1.1 Biểu đồ phân tán cho dữ liệu có hai biến

Biểu đồ phân tán cung cấp góc nhìn trực quan về dữ liệu theo từng trường hợp đối với hai biến số định lượng. Trong Hình 1.8 ở trang 16, biểu đồ phân tán được sử dụng để phân tích mối quan hệ giữa tỷ lệ sở hữu nhà và tỷ lệ đơn vị nhà ở thuộc các khu phức hợp (ví dụ: căn hộ) trong tập dữ liệu county. Một biểu đồ phân tán khác được trình bày trong Hình 2.1, so sánh tổng thu nhập của người vay (total_income) và số tiền vay (loan_amount) trong tập dữ liệu loan50. Trong bất kỳ biểu đồ phân tán nào, mỗi điểm dữ liệu đại diện cho một trường hợp quan sát. Vì loan50 có 50 trường hợp, nên Biểu đồ 2.1 chứa 50 điểm dữ liệu.



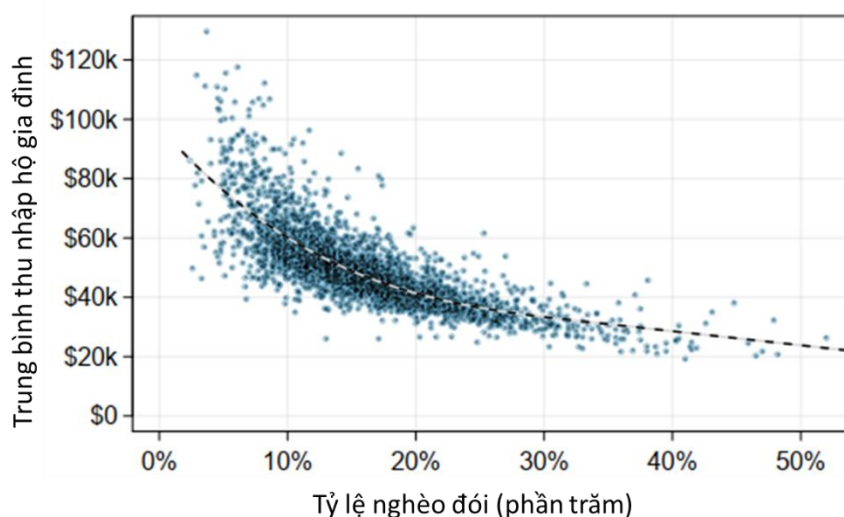
Hình 2.1: Biểu đồ phân tán giữa tổng thu nhập và giá trị khoản vay trong tập dữ liệu loan50.

Quan sát Biểu đồ 2.1, chúng ta thấy rằng có nhiều người vay có tổng thu nhập dưới 100 000 USD ở phía bên trái của biểu đồ, trong khi chỉ có một số ít người vay có thu nhập trên 250 000 USD.

VÍ DỤ 2.1

Hình 2.2 biểu diễn mối quan hệ giữa trung bình thu nhập hộ gia đình và tỷ lệ nghèo đói tại 3142 quốc gia. Dựa trên biểu đồ, có thể đưa ra những kết luận nào về mối quan hệ giữa hai biến số này?

Mối quan hệ giữa hai biến số rõ ràng là phi tuyến tính, được thể hiện qua đường nét đứt. Điều này khác biệt so với các biểu đồ phân tán trước đó, vốn không cho thấy xu hướng có độ cong đáng kể, nếu có.



Hình 2.2: Biểu đồ phân tán thể hiện mối quan hệ giữa trung bình thu nhập hộ gia đình và tỷ lệ nghèo đói trong tập dữ liệu county. Một mô hình thống kê đã được hiệu chỉnh và biểu diễn bằng đường nét đứt.

G

BÀI TẬP THỰC HÀNH 2.2

Biểu đồ phân tán tiết lộ điều gì về dữ liệu và đóng vai trò gì trong phân tích? ¹

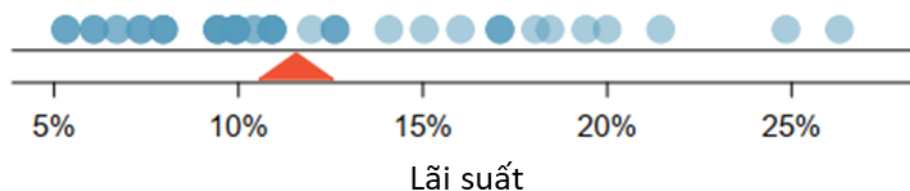
G

BÀI TẬP THỰC HÀNH 2.3

Mô tả hai biến số có mối quan hệ hình dạng móng ngựa trong biểu đồ phân tán ($\cap$ hoặc $\cup$).²

2.1.2 Biểu đồ chấm và giá trị trung bình

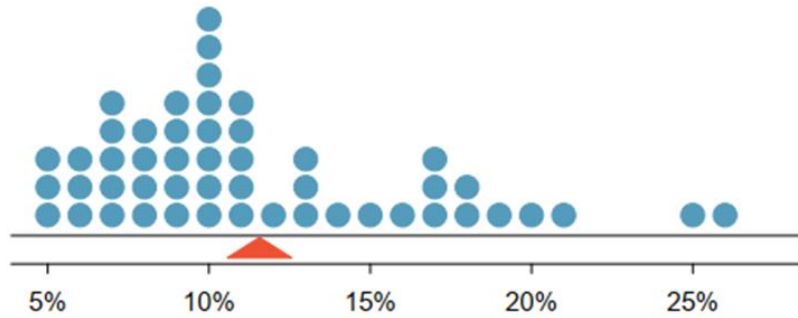
Đôi khi hai biến số có thể là quá nhiều khi chỉ một biến thực sự được quan tâm. Trong trường hợp này, biểu đồ chấm cung cấp cách trực quan hóa đơn giản nhất. **Biểu đồ chấm** là một dạng biểu đồ phân tán một biến, ví dụ về biểu đồ chấm thể hiện lãi suất của 50 khoản vay được trình bày trong Hình 2.3. Phiên bản xếp chồng của biểu đồ này được trình bày trong Hình 2.4.



Hình 2.3: Biểu đồ chấm biểu diễn biến `interest_rate` trong tập dữ liệu `loan50`. Giá trị trung bình của phân phối được biểu diễn bằng tam giác đỏ.

¹Câu trả lời có thể khác nhau. Biểu đồ phân tán hữu ích trong việc nhanh chóng xác định mối liên hệ giữa các biến, dù đó là những xu hướng đơn giản hay những mối quan hệ phức tạp hơn.

²Xem xét trường hợp trực tung đại diện cho một yếu tố “tốt” và trực hoành biểu thị một yếu tố chỉ có lợi khi ở mức độ vừa phải. Ví dụ mối quan hệ giữa sức khỏe và lượng nước tiêu thụ: cơ thể cần một lượng nước nhất định để duy trì sự sống, nhưng nếu tiêu thụ quá mức, nước có thể trở nên độc hại và gây nguy hiểm đến tính mạng.



Lãi suất, làm tròn đến hàng phần trăm

Hình 2.4: Biểu đồ chấm xếp chồng biểu diễn biến `interest_rate` trong tập dữ liệu `loan50`. Các mức lãi suất đã được làm tròn đến hàng phần trăm, và giá trị trung bình của phân phối được biểu diễn bằng tam giác đỏ.

Giá trị trung bình, thường được gọi là **trung bình cộng**, là một phương pháp phổ biến để xác định giá trị trung tâm của phân phối dữ liệu. Ví dụ để tính lãi suất trung bình, chúng ta cộng tất cả các mức lãi suất và chia cho số lượng quan sát:

$$\bar{x} = \frac{10,90\% + 9,92\% + 26,30\% + \dots + 6,08\%}{50} = 11,57\%$$

Giá trị trung bình thường được ký hiệu là $\bar{x}$. Ký hiệu x đóng vai trò như một biến đại diện cho đại lượng quan tâm, trong trường hợp này là biến `interest_rate` và dấu gạch ngang bên trên x biểu thị rằng chúng ta đang xét giá trị trung bình của lãi suất, với mức trung bình của 50 khoản vay là 11,57%. Giá trị trung bình có thể được xem như điểm cân bằng của phân phối dữ liệu, và trong Hình 2.3 và Hình 2.4 và nó được biểu diễn bằng một tam giác trong Hình 2.3 và Hình 2.4.

TRUNG BÌNH

Trung bình cộng của mẫu được tính bằng tổng các giá trị quan sát được chia cho số lượng quan sát:

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n}$$

trong đó $x_1, x_2, \dots, x_n$ là n giá trị quan sát được.

BÀI TẬP THỰC HÀNH 2.4

Xem xét phương trình tính trung bình và cho biết x_1 tương ứng với gì? x_2 thì sao? Từ đó, có thể suy luận ra ý nghĩa tổng quát của x_i là gì không? ³

BÀI TẬP THỰC HÀNH 2.5

Trong mẫu dữ liệu các khoản vay này, giá trị n là bao nhiêu? ⁴

Tập dữ liệu `loan50` là một mẫu được rút ra từ một tổng thể lớn hơn gồm các khoản vay được thực hiện thông qua Lending Club. Giá trị trung bình của tổng thể có thể được tính theo cách tương tự như trung bình của mẫu, nhưng nó có một ký hiệu đặc biệt: μ . Ký hiệu μ (chữ *mu* trong tiếng Hy Lạp) biểu thị giá trị trung bình của toàn bộ quần thể. Đôi khi, một chỉ số dưới, chẳng hạn như x , được thêm vào để xác định biến số mà trung bình quần thể đề cập đến, ví dụ μ_x . Trong thực tế, việc đo lường chính xác giá trị trung bình của tổng thể thường rất tốn kém, vì vậy trung bình mẫu $\bar{x}$ thường được sử dụng để ước lượng μ .

³ x_1 tương ứng với lãi suất của khoản vay đầu tiên trong mẫu (10,90%), x_2 là lãi suất của khoản vay thứ hai (9,92%) và x_i đại diện cho lãi suất của khoản vay thứ i trong tập dữ liệu. Ví dụ, nếu $i = 4$, thì ta đang xem xét x_4 tức là lãi suất của quan sát thứ tư trong tập dữ liệu.

⁴ Cỡ mẫu là $n=50$.

VÍ DỤ 2.6

Lãi suất trung bình của toàn bộ quần thể các khoản vay có thể được ước lượng từ dữ liệu mẫu. Dựa trên mẫu gồm 50 khoản vay, một ước lượng hợp lý cho μ_x tức là lãi suất trung bình của toàn bộ tập dữ liệu, sẽ là bao nhiêu?

- E Giá trị trung bình mẫu, 11,57%, là một ước lượng sơ bộ cho μ_x . Mặc dù không hoàn toàn chính xác, đây vẫn là ước lượng tốt nhất của chúng ta về lãi suất trung bình của toàn bộ tổng thể khoản vay đang được nghiên cứu. Trong Chương 5 and và các phần tiếp theo, chúng ta sẽ phát triển các công cụ để đánh giá độ chính xác của các *ước lượng điểm*, diễn hình như trung bình mẫu. Như có thể dự đoán, các ước lượng điểm từ mẫu lớn thường chính xác hơn so với những ước lượng từ mẫu nhỏ.

VÍ DỤ 2.7

Giá trị trung bình có ý nghĩa quan trọng trong việc chuẩn hóa hoặc hiệu chỉnh một thước đo, giúp dữ liệu trở nên dễ diễn giải và so sánh hơn. Hãy đưa ra hai ví dụ về cách sử dụng trung bình để hỗ trợ phân tích so sánh.

1. Chúng tôi muốn đánh giá liệu một loại thuốc mới có hiệu quả hơn trong điều trị cơn hen suyễn cấp so với thuốc hiện hành hay không. Một thử nghiệm lâm sàng được tiến hành trên 1500 người trưởng thành, trong đó 500 người thuộc nhóm can thiệp được sử dụng thuốc mới, trong khi 1000 người thuộc nhóm đối chứng sử dụng thuốc hiện hành.

	Thuốc mới	Thuốc hiện hành
Tổng số bệnh nhân	500	1000
Tổng số cơn hen suyễn	200	300

E Việc so sánh số cơn hen suyễn là 200 và 300 sẽ cho thấy thuốc mới có vẻ hiệu quả hơn, nhưng đây thực chất là một sai lệch do sự chênh lệch về số lượng bệnh nhân trong các nhóm. Thay vào đó, chúng ta nên xem xét số cơn hen suyễn trung bình trên từng nhóm bệnh nhân:

Nhóm thuốc mới: $200/500 = 0.4$

Nhóm thuốc hiện hành: $300/1000 = 0,3$

Nhóm bệnh nhân sử dụng thuốc hiện hành có số cơn hen suyễn trung bình trên mỗi bệnh nhân thấp hơn so với số cơn hen suyễn trung bình trong nhóm can thiệp sử dụng thuốc mới.

2. Emilio đã mở một xe bán thức ăn vào năm ngoái, nơi anh bán món burrito, và công việc kinh doanh của anh đã ổn định trong suốt 3 tháng qua. Trong khoảng thời gian này, anh đã kiếm được \$11 000 với 625 giờ làm việc. Thu nhập trung bình theo giờ của Emilio là một chỉ số quan trọng để đánh giá hiệu quả hoạt động kinh doanh của anh có hiệu quả, ít nhất là từ góc độ tài chính hay không?

$$\frac{\$11000}{625 \text{ giờ}} = \$17,60 \text{ mỗi giờ}$$

Khi biết được mức thu nhập trung bình theo giờ, Emilio đã chuyển thu nhập của mình thành chỉ số dễ dàng so sánh với các công việc khác mà anh có thể xem xét.

VÍ DỤ 2.8

Giả sử chúng ta muốn tính toán thu nhập bình quân người tại Hoa Kỳ. Để thực hiện điều này, chúng ta có thể đầu tiên nghĩ đến việc tính giá trị trung bình thu nhập đầu người tại 3 142 quận trong tập dữ liệu count. Tuy nhiên, cách tiếp cận nào sẽ hợp lý hơn?

E Tập dữ liệu county có đặc điểm đặc biệt là mỗi quận trong tập dữ liệu sẽ đại diện cho một số lượng lớn người dân. Nếu chúng ta chỉ tính giá trị trung bình biến income, chúng ta sẽ không phân biệt giữa các quận có dân số 5 000 người và 5 000 000 người, điều này có thể dẫn đến sai lệch trong phân tích. Thay vào đó, phương pháp tốt nhất là tính tổng thu nhập của mỗi quận, sau đó cộng dồn tổng thu nhập của tất cả các quận và chia cho tổng số người trong toàn bộ các quận. Thực hiện theo phương pháp này sẽ cho ra thu nhập bình quân đầu người của Hoa Kỳ là \$30 861. Ngược lại, nếu chỉ tính giá trị trung bình thu nhập bình quân đầu người của các quận, kết quả sẽ chỉ là \$26 093, làm giảm tính chính xác của ước tính.

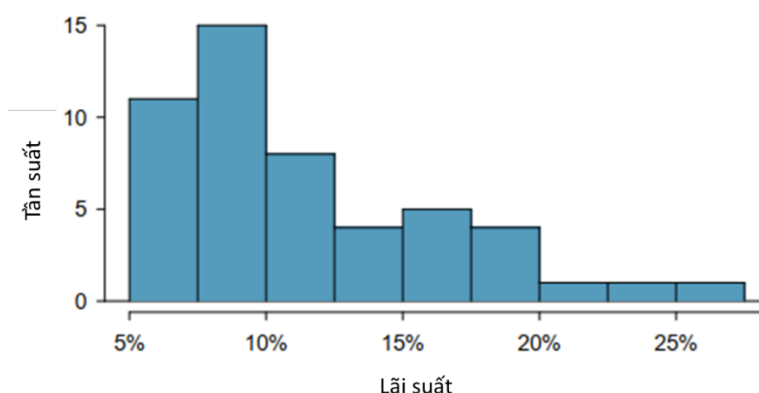
Ví dụ này sử dụng một phương pháp được gọi là **trung bình có trọng số**. Để tìm hiểu thêm về chủ đề này, bạn có thể tham khảo tài liệu trực tuyến tại openintro.org/d?file=stat wtd mean.

2.1.3 Biểu đồ tần suất và hình dạng phân phối

Biểu đồ chấm thể hiện giá trị chính xác của từng quan sát, điều này rất có ích khi làm việc với các bộ dữ liệu nhỏ nhưng chúng có thể trở nên khó đọc khi kích thước mẫu dữ liệu trở nên lớn hơn. Thay vì biểu diễn tần số của mỗi quan sát, chúng ta thường phân chia giá trị thành các *khoảng*. Ví dụ, trong bộ dữ liệu loan50, chúng ta tạo một bảng đếm số lượng khoản vay có lãi suất trong các khoảng như từ 5,0% đến 7,5%, từ 7,5% đến 10,0% và tiếp tục như vậy. Các quan sát nằm tại ranh giới của một khoảng (ví dụ: 10,00%) sẽ được phân vào khoảng thấp hơn. Bảng phân khoảng giá trị này được thể hiện trong Hình 2.5. Các dữ liệu đã phân khoảng này sau đó được biểu diễn dưới dạng các cột trong Hình 2.6, tạo thành một **biểu đồ tần suất**, đây có thể được coi là một phiên bản với mức độ phân tổ chi tiết hơn của biểu đồ chấm xếp chồng đã được trình bày trong Hình 2.4.

Lãi suất	5,0% - 7,5%	7,5% - 10,0%	10,0% - 12,5%	12,5% - 15,0%	...	25,0% - 27,5%
Tần số	11	15	8	4	...	1

Hình 2.5: Số lượng các khoản vay trong các khoảng lãi suất đã phân khoảng



Hình 2.6: Biểu đồ tần suất của lãi suất. Phân phối này có sự lệch nhiều về phía bên phải

Biểu đồ tần suất cung cấp cái nhìn về **mật độ dữ liệu**. Những cột cao hơn biểu thị nơi dữ liệu có mật độ tương đối cao hơn. Ví dụ, trong bộ dữ liệu, có rất nhiều khoản vay có lãi suất từ 5% đến 10%, so với số khoản vay có lãi suất từ 20% đến 25%. Các cột này giúp dễ dàng nhận thấy sự thay đổi của mật độ dữ liệu theo từng mức lãi suất.

Biểu đồ tần suất đặc biệt hữu ích trong việc hiểu rõ hình dạng phân phối dữ liệu. Hình 2.6 cho thấy phần lớn các khoản vay có lãi suất dưới 15%, trong khi chỉ có một số ít khoản vay có lãi suất trên 20%. Khi dữ liệu có xu hướng giảm dần về phía bên phải và có đuôi phải dài hơn, hình dạng phân phối được gọi là **phân phối lệch phải**.⁵

Các bộ dữ liệu có đặc điểm ngược lại với đuôi phân phối kéo dài về phía bên trái được gọi là **phân phối lệch trái**. Chúng ta cũng có thể nói rằng phân phối này có một đuôi trái dài. Các bộ dữ liệu có sự giảm dần tương đối đều ở cả hai phía được gọi là **phân phối đối xứng**.

ĐUÔI DÀI ĐỂ XÁC ĐỊNH ĐỘ LỆCH PHÂN PHỐI

Khi dữ liệu kéo dài về một hướng, phân phối sẽ có một **đuôi dài**. Nếu phân phối có đuôi dài bên trái, nó được gọi là lệch trái. Nếu phân phối có đuôi dài bên phải, nó được gọi là lệch phải.

⁵ Các cách diễn đạt khác để mô tả dữ liệu có độ lệch phải bao gồm: **lệch về phía bên phải**, **lệch về phía giá trị cao**, hoặc **lệch về phía các giá trị dương**.

HƯỚNG DẪN THỰC HÀNH 2.9

Hãy quan sát các biểu đồ chấm trong Hình 2.3 và 2.4. Bạn có thể nhận thấy độ lệch trong dữ liệu không? Liệu độ lệch này có dễ nhận thấy hơn trong biểu đồ tần suất hay trong biểu đồ chấm? ⁶

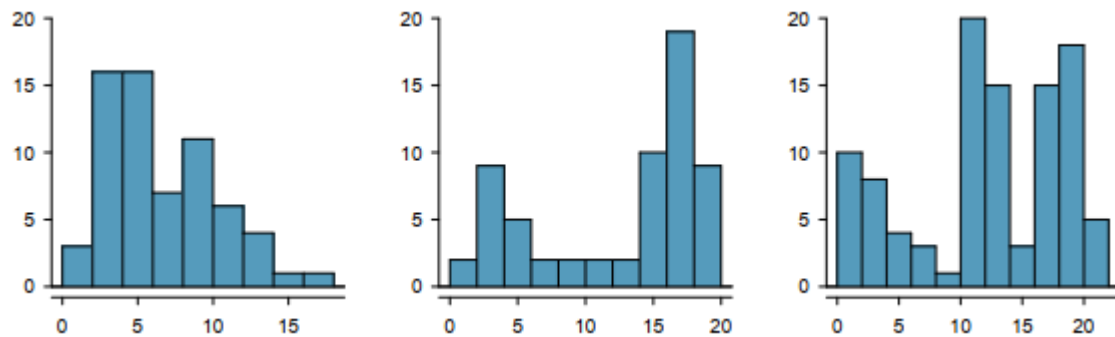
HƯỚNG DẪN THỰC HÀNH 2.10

Ngoài giá trị trung bình (đã được ghi nhận), bạn có thể quan sát được những gì trong biểu đồ chấm mà không thể quan sát thấy trong biểu đồ tần suất? ⁷

Ngoài việc quán sát được độ lệch hay tính đối xứng của phân phối, biểu đồ tần suất cũng có thể được sử dụng để xác định các mode. Một **mode** được đặc trưng bởi sự xuất hiện của một đỉnh nổi bật. Trong biểu đồ tần suất của biến `loan_amount`, chỉ có một đỉnh nổi bật duy nhất, điều này cho thấy phân phối chỉ có một yếu vị duy nhất.

Một định nghĩa về *mode* thường được giảng dạy trong các lớp toán học là giá trị có số lần xuất hiện nhiều nhất trong tập dữ liệu. Tuy nhiên, đối với nhiều bộ dữ liệu trong thực tế, thường không có bất kỳ quan sát nào có cùng giá trị trong một tập dữ liệu, điều này khiến định nghĩa này trở nên không thực tế trong phân tích dữ liệu.

Hình 2.7 trình bày các biểu đồ tần suất với một, hai hoặc ba đỉnh nổi bật. Các phân phối này lần lượt được gọi là **đơn đỉnh**, **hai đỉnh**, và **đa đỉnh**. Bất kỳ phân phối nào có hơn hai đỉnh rõ ràng sẽ được gọi là phân phối đa đỉnh. Cần lưu ý rằng trong phân phối đơn đỉnh, mặc dù có một đỉnh rõ ràng duy nhất, một đỉnh phụ thứ hai vẫn tồn tại nhưng không được tính đến, do sự khác biệt giữa nó và các khoảng lân cận chỉ là một số lượng quan sát không đáng kể.



Hình 2.7: Khi chỉ tính các đỉnh nổi bật, các phân phối lần lượt là (từ trái sang phải) đơn đỉnh, hai đỉnh và đa đỉnh. Lưu ý rằng việc cố ý phân loại biểu đồ bên trái là đơn đỉnh. Điều này là bởi vì chúng ta chỉ tính các đỉnh nổi bật, chứ không phải bất kỳ đỉnh nào.

VÍ DỤ 2.11

Hình 2.6 cho thấy một yếu vị nổi bật duy nhất trong tỷ lệ lãi suất. Vậy phân phối này phải là đơn đỉnh, hai đỉnh hay đa đỉnh?

Đơn đỉnh - Unimodal. Hãy nhớ rằng *uni* đại diện cho số 1 (hãy nghĩ đến xe đạp một bánh - unicycles). Tương tự, "bi" đại diện cho số 2 (hãy nghĩ đến xe đạp hai bánh - "bicycles"). Hy vọng một ngày nào đó sẽ có một thuật ngữ cho *phân phối nhiều đỉnh* - *multicycle* để hoàn thiện khái niệm này.

BÀI TẬP THỰC HÀNH 2.12

Các phép đo chiều cao của học sinh và giáo viên tại một trường tiểu học K-3 đã được thu thập. Trong bộ dữ liệu chiều cao này, bạn sẽ kỳ vọng có bao nhiêu yếu vị? ⁸

Việc xác định số lượng các yếu vị trong một phân phối không phải nhằm đưa ra một câu trả lời rõ ràng và chính xác về số lượng các yếu vị trong một phân phối, đó là lý do tại sao khái niệm "*prominent*" không được định nghĩa chặt chẽ trong cuốn sách này. Phần quan trọng nhất của quá trình khảo sát này là giúp bạn hiểu rõ hơn về dữ liệu của mình.

⁶ Sự lệch phân phối có thể quan sát rõ ràng trong cả ba biểu đồ, tuy nhiên, biểu đồ chấm phẳng ít có giá trị phân tích hơn. Biểu đồ chấm xếp chồng và biểu đồ tần suất là những công cụ trực quan hữu ích giúp nhận diện sự lệch phân phối một cách rõ ràng hơn.

⁷ Lãi suất cho vay cá nhân.

⁸ Có thể quan sát thấy hai nhóm chiều cao trong bộ dữ liệu: một nhóm của học sinh và một nhóm của người lớn. Tức là, dữ liệu có thể là phân phối nhị thức.

2.1.4 Phương sai và độ lệch chuẩn

Trung bình được giới thiệu như một phương pháp để mô tả đặc trưng độ tập trung dữ liệu, tuy nhiên, sự phân tán trong dữ liệu cũng đóng vai trò quan trọng không kém. Tại đây, chúng ta sẽ giới thiệu hai chỉ số đo lường sự phân tán là: phương sai và độ lệch chuẩn. Cả hai chỉ số này đều rất hữu ích trong phân tích dữ liệu, mặc dù công thức tính toán của chúng có thể khá phức tạp khi thực hiện thủ công. Độ lệch chuẩn là chỉ số dễ hiểu hơn trong hai chỉ số này, và nó mô tả một cách tổng quát mức độ phân tán của các quan sát so với giá trị trung bình.

Chúng ta gọi khoảng cách giữa một quan sát và giá trị trung bình của nó là **độ lệch**. Dưới đây là các độ lệch của quan sát thứ 1st, 2nd, 3rd, and 50th trong biến lãi suất:

$$\begin{aligned}x_1 - \bar{x} &= 10,90 - 11,57 = -0,67 \\x_2 - \bar{x} &= 9,92 - 11,57 = -1,65 \\x_3 - \bar{x} &= 26,30 - 11,57 = 14,73 \\&\vdots \\x_{50} - \bar{x} &= 6,08 - 11,57 = -5,49\end{aligned}$$

Nếu chúng ta bình phương các độ lệch này và sau đó tính trung bình, kết quả sẽ bằng **phương sai** mẫu, ký hiệu là s^2 :

$$\begin{aligned}s^2 &= \frac{(-0,67)^2 + (-1,65)^2 + (14,73)^2 + \dots + (-5,49)^2}{50 - 1} \\&= \frac{0,45 + 2,72 + 216,97 + \dots + 30,14}{49} \\&= 25,52\end{aligned}$$

Khi tính phương sai mẫu, chúng ta chia cho $n - 1$ thay vì chia cho n . Mặc dù có những lý do toán học phức tạp đằng sau, nhưng kết quả cuối cùng là việc làm này giúp chỉ số trở nên đáng tin cậy hơn và hữu ích hơn.

Lưu ý rằng việc bình phương các độ lệch mang lại hai tác động chính. Thứ nhất, nó làm tăng đáng kể các giá trị lớn, điều này có thể quan sát qua việc so sánh $(-0,67)^2$, $(-1,65)^2$, $(14,73)^2$, và $(-5,49)^2$. Thứ hai, việc này giúp loại bỏ các dấu âm.

Độ lệch chuẩn được định nghĩa là căn bậc hai của phương sai:

$$s = \sqrt{25,52} = 5,05$$

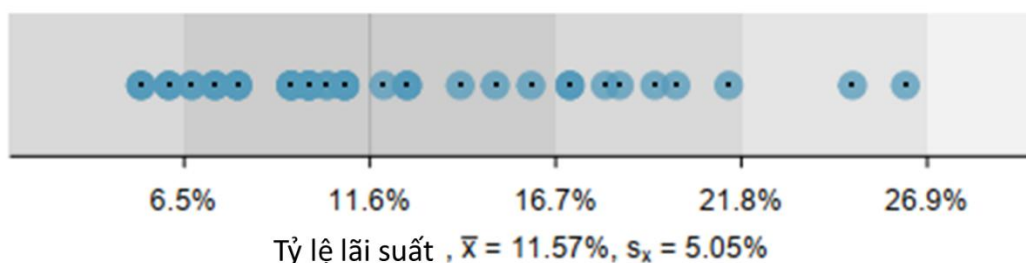
Mặc dù thường bị bỏ qua, một chỉ số phụ x có thể được thêm vào phương sai và độ lệch chuẩn, tức là s_x^2 và s_x , nếu điều này giúp nhấn mạnh rằng đây là phương sai và độ lệch chuẩn của các quan sát được biểu diễn bởi $x_1, x_2, \dots, x_n$.

PHƯƠNG SAI VÀ ĐỘ LỆCH CHUẨN

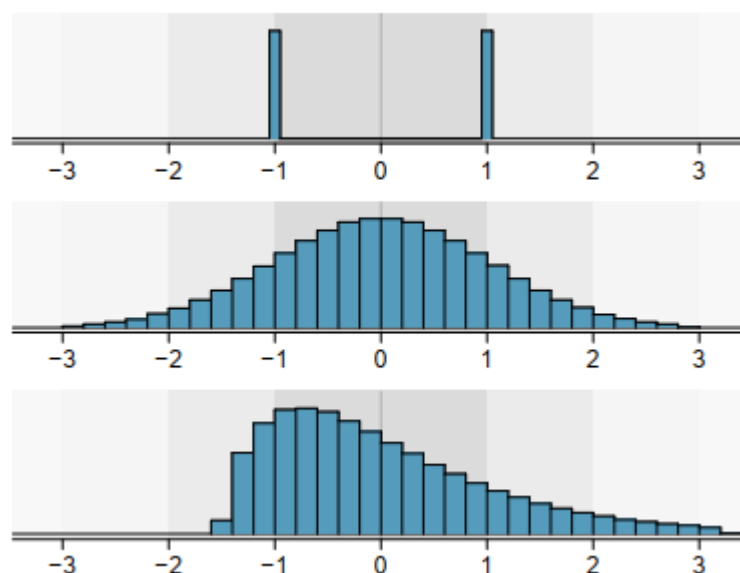
Phương sai là giá trị trung bình cộng của tổng bình phương độ lệch so với giá trị trung bình. Độ lệch chuẩn là căn bậc hai của phương sai. Độ lệch chuẩn rất hữu ích trong việc đánh giá mức độ phân tán của dữ liệu xung quanh giá trị trung bình.

Độ lệch chuẩn phản ánh sự lệch điển hình của các quan sát so với giá trị trung bình. Thông thường, khoảng 70% dữ liệu sẽ nằm trong một độ lệch chuẩn xung quanh giá trị trung bình, và khoảng 95% sẽ nằm trong hai độ lệch chuẩn. Tuy nhiên, như minh họa trong Hình 2.8 và 2.9 các tỷ lệ này không phải là quy tắc tuyệt đối.

Giống như giá trị trung bình, các giá trị của phương sai và độ lệch chuẩn của tổng thể có các ký hiệu đặc biệt: σ^2 cho phương sai và σ cho độ lệch chuẩn. Ký hiệu σ là chữ cái Hy Lạp *sigma*.



Hình 2.8: Đối với biến `interest_rate`, 34 trong số 50 khoản vay (68%) có lãi suất nằm trong 1 độ lệch chuẩn so với giá trị trung bình, và 48 trong số 50 khoản vay (96%) có lãi suất nằm trong 2 độ lệch chuẩn. Thông thường, khoảng 70% dữ liệu sẽ nằm trong 1 độ lệch chuẩn so với giá trị trung bình và 95% nằm trong 2 độ lệch chuẩn. Tuy nhiên, đây không phải là một quy tắc tuyệt đối.



Hình 2.9: Ba phân phối tổng thể rất khác nhau với cùng giá trị trung bình $\mu = 0$ và độ lệch chuẩn $\sigma = 1$.

BÀI TẬP THỰC HÀNH 2.13

- G** Trang 45 đã giới thiệu khái niệm về hình dạng của một phân phối. Trong đó, một mô tả đầy đủ về hình dạng của phân phối cần bao gồm tính chất mô hình và việc phân phối có đối xứng hay bị lệch sang một phía. Sử dụng Hình 2.9 làm ví dụ, giải thích lý do tại sao mô tả như vậy lại quan trọng.⁹

VÍ DỤ 2.14

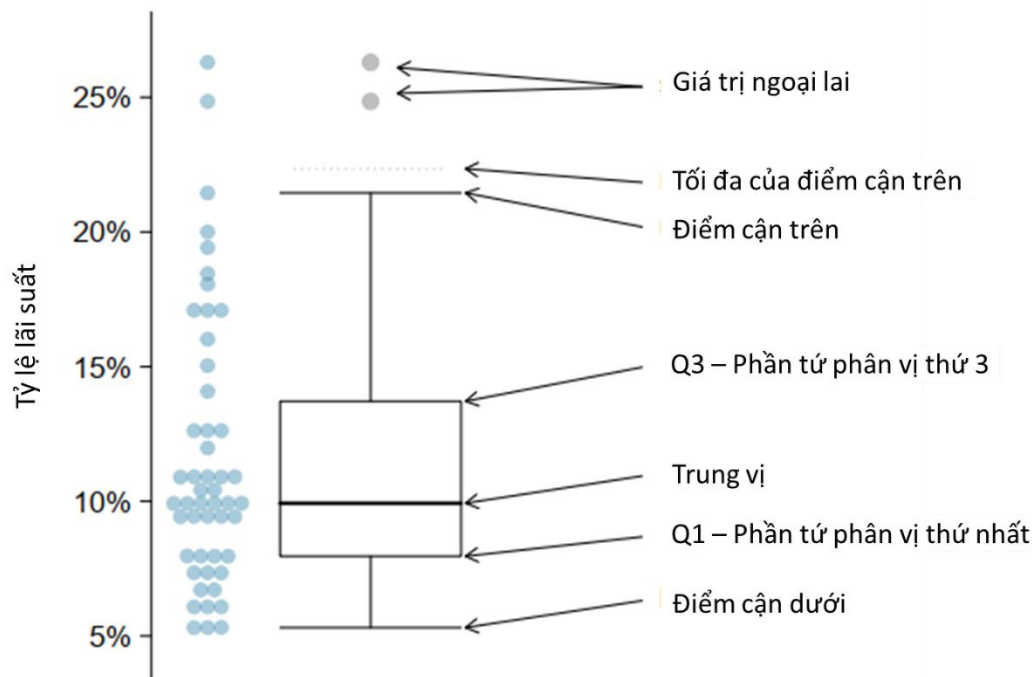
- E** Mô tả phân phối của biến lãi suất dựa trên biểu đồ tần suất trong Hình 2.6. Mô tả này cần bao gồm: **giá trị trung tâm, độ phân tán và hình dạng của phân phối** được đặt trong bối cảnh cụ thể. Ngoài ra, cần lưu ý những trường hợp đặc biệt hoặc bất thường, nếu có.
- Phân bố lãi suất có hình dạng đơn đỉnh và lệch phải, với đa số các giá trị lãi suất tập trung gần giá trị trung bình (11,57%). Phần lớn các giá trị nằm trong khoảng một độ lệch chuẩn (5,05%) tính từ giá trị trung bình. Tuy nhiên, trong mẫu còn tồn tại một số giá trị lãi suất cực đại, vượt quá 20%.

Trong thực tế, phương sai và độ lệch chuẩn đôi khi được sử dụng như một công cụ nhằm đạt được mục tiêu, trong đó "mục tiêu" là khả năng ước tính chính xác sự không chắc chắn liên quan đến tham số mẫu. Ví dụ, trong Chương 5, độ lệch chuẩn được sử dụng trong các phép tính giúp chúng ta hiểu rõ hơn về mức độ biến động của trung bình mẫu giữa các mẫu khác nhau.

⁹ Hình 2.9 hiển thị ba phân phối có hình dạng khá khác nhau, nhưng tất cả đều có cùng giá trị trung bình, phương sai và độ lệch chuẩn. Bằng cách sử dụng tính chất của mô hình, chúng ta có thể phân biệt đồ thị đầu tiên (hai đỉnh) với hai đồ thị còn lại (một đỉnh). Dựa trên độ lệch, chúng ta có thể phân biệt đồ thị cuối cùng (độ lệch phải) với hai đồ thị đầu tiên. Mặc dù các biểu đồ như biểu đồ tần suất giúp cung cấp một cái nhìn tổng quát hơn về phân phối, nhưng chúng ta vẫn có thể sử dụng tính chất của mô hình và hình dạng (đối xứng/độ lệch) để mô tả những đặc điểm cơ bản của phân phối.

2.1.5 Biểu đồ hộp, phân vị và trung vị

Biểu đồ hộp mô tả bộ số liệu thông qua năm chỉ số thống kê chính, đồng thời hiển thị các quan sát bất thường. Hình 2.10 cung cấp một biểu đồ chấm dọc theo chiều dọc cùng với một biểu đồ hộp của biến tỷ lệ lãi suất từ bộ dữ liệu loan50.



Hình 2.10: Một biểu đồ chấm dọc theo chiều dọc, trong đó các chấm được xếp chồng lên nhau theo chiều ngang, bên cạnh một biểu đồ hộp có nhãn thể hiện tỷ lệ lãi suất của 50 khoản vay.

Bước đầu tiên trong việc xây dựng biểu đồ hộp là vẽ một đường đậm biểu thị trung vị, chia bộ dữ liệu thành hai nửa. Hình 2.10 cho thấy 50% dữ liệu nằm dưới trung vị và 50% còn lại nằm trên trung vị. Bộ dữ liệu bao gồm 50 khoản vay (một số chẵn), do đó, dữ liệu được phân chia một cách hoàn hảo thành hai nhóm mỗi nhóm có 25 quan sát. Trung vị trong trường hợp này được tính là trung bình của hai quan sát gần nhất với phân vị 50, các giá trị này là giống nhau: $(9,93\% + 9,93\%)/2 = 9,93\%$. Khi số lượng quan sát là lẻ, sẽ có một quan sát duy nhất phân chia dữ liệu thành hai nửa, và trong trường hợp đó, quan sát đó chính là trung vị (không cần tính trung bình).

TRUNG VỊ: GIÁ TRỊ Ở GIỮA

Khi dữ liệu được sắp xếp theo thứ tự tăng dần, trung vị là quan sát nằm ở vị trí chính giữa. Nếu số lượng quan sát là chẵn, sẽ có hai giá trị nằm ở giữa, và trung vị được tính là trung bình của hai giá trị này.

Bước thứ hai trong việc xây dựng biểu đồ hộp là vẽ một hình chữ nhật để đại diện cho 50% dữ liệu ở giữa. Tổng chiều dài của hộp, được thể hiện theo chiều dọc trong Hình 2.10, được gọi là **khoảng tứ phân vị (IQR)**. Giống như độ lệch chuẩn, IQR là một chỉ số đo lường độ phân tán của dữ liệu. Dữ liệu có sự phân tán càng lớn thì độ lệch chuẩn và khoảng tứ phân vị cũng có xu hướng tăng. Hai đường biên của hộp, tương ứng với **phần vị thứ nhất** (Q1, tương ứng với phân vị 25%, tức là 25% dữ liệu nằm dưới giá trị này) và **phần vị thứ ba** (Q3, tương ứng với phân vị 75%) và chúng thường được ký hiệu là Q1 và Q3.

KHOẢNG TỨ PHẦN VỊ - INTERQUARTILE RANGE (IQR)

Khoảng tứ phân vị (IQR) là chiều dài của hộp trong biểu đồ hộp. Nó được tính bằng cách:

$$IQR = Q_3 - Q_1$$

trong đó Q_1 và Q_3 tương ứng với phân vị thứ 25 và phân vị thứ 75.

BÀI TẬP THỰC HÀNH 2.15

G Tỷ lệ phần trăm dữ liệu nằm trong khoảng Q_1 đến trung vị là bao nhiêu? Tỷ lệ phần trăm dữ liệu nằm trong khoảng từ trung vị đến Q_3 là bao nhiêu? ¹⁰

Các điểm cận (whiskers) kéo dài từ hộp nhằm bao quát các quan sát nằm ngoài phạm vi của hộp. Tuy nhiên, độ mở rộng của chúng bị giới hạn ở mức tối đa $1,5 \times IQR$. Các râu bao gồm toàn bộ các quan sát nằm trong phạm vi xác định. Trong Hình 2.10, râu trên không kéo dài đến hai điểm cực trị cuối cùng do các quan sát này vượt quá ngưỡng $Q_3 + 1,5 \times IQR$. Thay vào đó, râu chỉ mở rộng đến quan sát lớn nhất nằm trong giới hạn này. Râu dưới kết thúc tại giá trị nhỏ nhất, 5.31%, do không có quan sát nào nhỏ hơn để tiếp tục mở rộng. Giới hạn dưới của râu không được thể hiện trong hình vì phạm vi của biểu đồ không kéo dài đến $Q_1 - 1,5 \times IQR$. Về trực quan, hộp có thể được xem như phần trung tâm của biểu đồ hộp, trong khi các râu đóng vai trò như những phần mở rộng nhằm bao quát các quan sát còn lại trong tập dữ liệu.

Bất kỳ quan sát nào nằm ngoài phạm vi của các râu đều được biểu thị bằng một dấu chấm. Việc gán nhãn cho các điểm này, thay vì kéo dài râu đến giá trị nhỏ nhất và lớn nhất quan sát được nhằm mục đích xác định các quan sát có khoảng cách bất thường so với phần còn lại của tập dữ liệu. Những quan sát có khoảng cách bất thường này được gọi là **giá trị ngoại lai**. Trong trường hợp này, lãi suất 24,85% và 26,30% có thể được xem là các giá trị ngoại lai do chúng nằm cách biệt đáng kể so với phần lớn dữ liệu.

NGOẠI LAI LÀ CÁC GIÁ TRỊ CỰC ĐOAN

Một **giá trị ngoại lai** là một quan sát có giá trị cực đoan so với phần còn lại của tập dữ liệu. Việc phân tích dữ liệu để xác định ngoại lai có nhiều ý nghĩa quan trọng, bao gồm

1. Xác định độ lệch của phân phối.
2. Xác định các sai lệch tiềm ẩn phát sinh từ quy trình thu thập dữ liệu hoặc sai sót trong quá trình nhập liệu.
3. Cung cấp hiểu biết sâu hơn về các đặc điểm đáng chú ý của tập dữ liệu.

BÀI TẬP THỰC HÀNH 2.16

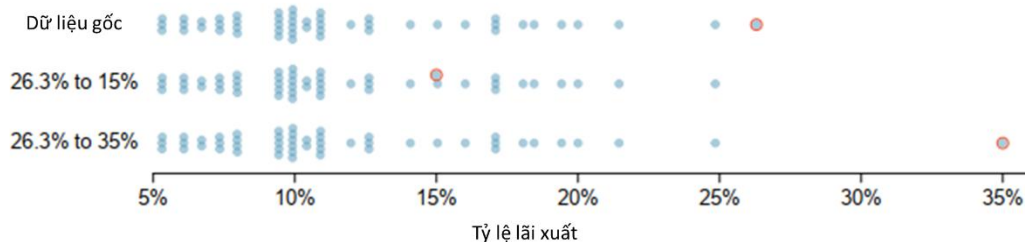
G Dựa vào Hình 2.10, ước tính các giá trị sau cho biến `interest_rate` trong tập dữ liệu `loan50`:
(a) Q_1 , (b) Q_3 và (c) IQR. ¹¹

¹⁰ Vì Q_1 và Q_3 bao quát 50% dữ liệu tập trung ở giữa, trong khi trung vị chia dữ liệu thành hai phần bằng nhau, nên 25% dữ liệu nằm trong khoảng từ Q_1 đến trung vị và 25% còn lại nằm trong khoảng từ trung vị đến Q_3 .

¹¹ Các ước tính trực quan này có thể có sai lệch nhỏ định giữa các quan sát: $Q_1 = 8\%$, $Q_3 = 14\%$, $IQR = Q_3 - Q_1 = 6\%$. (Giá trị thực: $Q_1 = 7,96\%$, $Q_3 = 13,72\%$, $IQR = 5,76\%$.)

2.1.6 Thống kê Robust

Các tham số mẫu của tập dữ liệu lãi suất bị tác động như thế nào bởi giá trị quan sát 26,3%? Điều gì sẽ xảy ra nếu khoản vay này chỉ là 15%? Điều gì sẽ xảy ra đối với các thống kê tóm tắt nếu giá trị quan sát tại mức 26,3% tăng lên 35%? Các trường hợp này đã được minh họa cùng dữ liệu gốc trong hình 2.11 và các tham số mẫu tương ứng được tính toán cho từng trường hợp trong Hình 2.12.



Hình 2.11: Biểu đồ chấm của tập dữ liệu lãi suất gốc và hai tập dữ liệu đã được điều chỉnh.

Trường hợp	chuẩn mạnh		không chuẩn mạnh	
	Trung vị	IQR	$\bar{x}$	s
Dữ liệu <code>interest_rate</code> gốc	9,93%	5,76%	11,57%	5,05%
Điều chỉnh 26.3% → 15%	9,93%	5,76%	11,34%	4,61%
Điều chỉnh 26.3% → 35%	9,93%	5,76%	11,74%	5,68%

Hình 2.12: So sánh sự thay đổi của trung vị, IQR, trung bình ($\bar{x}$) và độ lệch chuẩn (s) khi có các quan sát ngoại lai trong biến `interest_rate`.

BÀI TẬP THỰC HÀNH 2.17

- (a) Giá trị trung bình hay trung vị bị ảnh hưởng nhiều hơn bởi các quan sát ngoại lai? Hình 2.12 có thể sẽ hữu ích trong việc trả lời câu hỏi này (b) Độ lệch chuẩn hay khoảng tứ phân vị (IQR) bị ảnh hưởng nhiều hơn bởi các quan sát ngoại lai? ¹²

Trung vị và khoảng tứ phân vị (IQR) được gọi là các **tham số bền vững** do ít bị ảnh hưởng bởi các quan sát ngoại lai. Việc điều chỉnh giá trị cực đoạn thường không tác động đáng kể đến các tham số này. Ngược lại, trung bình và độ lệch chuẩn nhạy cảm hơn với các quan sát ngoại lai, điều này có thể quan trọng trong một số tình huống phân tích dữ liệu.

VÍ DỤ 2.18

Trung vị và khoảng tứ phân vị (IQR) không thay đổi trong cả ba trường hợp được trình bày trong Hình 2.12. Tại sao lại có trường hợp này?

Trung vị và khoảng tứ phân vị (IQR) chỉ nhạy cảm với các giá trị nằm xung quanh Q_1 , trung vị và Q_3 . Do các giá trị trong các vùng này không thay đổi giữa ba tập dữ liệu, nên ước lượng trung vị và IQR cũng giữ nguyên.

BÀI TẬP THỰC HÀNH 2.19

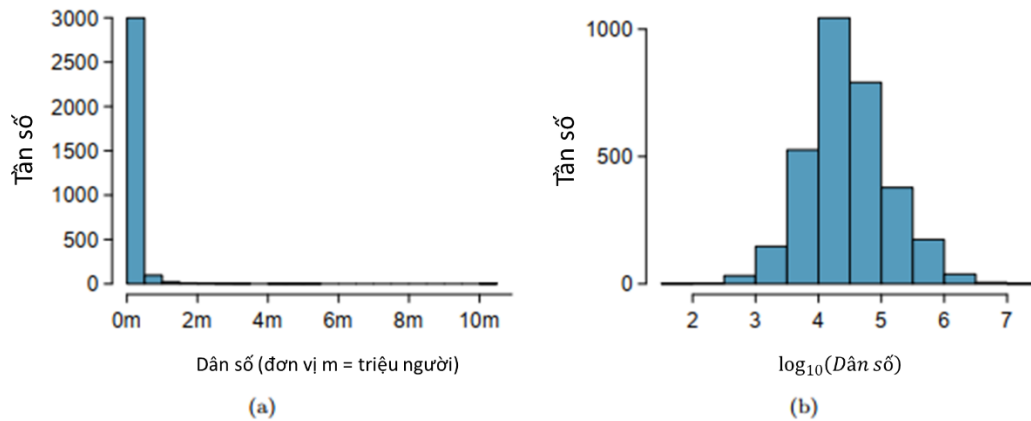
Phân phối của số tiền vay trong tập dữ liệu `loan50` có độ lệch phải, với một số khoản vay lớn nằm ở phần đuôi bên phải. Nếu bạn muốn hiểu về quy mô của khoản vay điển hình, bạn nên quan tâm đến trung bình hay trung vị? ¹³

¹² (a) Giá trị trung bình chịu ảnh hưởng nhiều hơn. (b) Độ lệch chuẩn chịu ảnh hưởng nhiều hơn. Giải thích chi tiết được trình bày trong phần hướng dẫn thực hành 2.17.

¹³ Câu trả lời có thể khác nhau! Nếu chúng ta chỉ muốn tìm hiểu một khoản vay điển hình trông như thế nào, thì trung vị có lẽ hữu ích hơn. Tuy nhiên nếu mục tiêu là hiểu một yếu tố có khả năng mở rộng tốt, chẳng hạn như tổng số tiền chúng ta có thể cần phải chuẩn bị nếu chúng ta cung cấp 1 000 khoản vay, thì giá trị trung bình sẽ hữu ích hơn.

2.1.7 Chuyển hóa dữ liệu (chủ đề đặc biệt)

Khi dữ liệu có độ lệch rất lớn, đôi khi chúng ta thực hiện chuyển đổi để giúp việc mô hình hóa trở nên dễ dàng hơn.



Hình 2.13: (a) Biểu đồ tần suất của dân số tất cả các quận tại Hoa Kỳ. (b) Biểu đồ tần suất của dân số các quận sau khi được biến đổi theo $\log_{10}$. Trong biểu đồ này, giá trị trên trục hoành tương ứng với lũy thừa của 10, ví dụ “4” trên trục hoành tương ứng với $10^4 = 10\,000$.

VÍ DỤ 2.20

Xem xét biểu đồ tần suất dân số các quận trong Hình 2.13(a), ta thấy phân phối có độ lệch lớn.

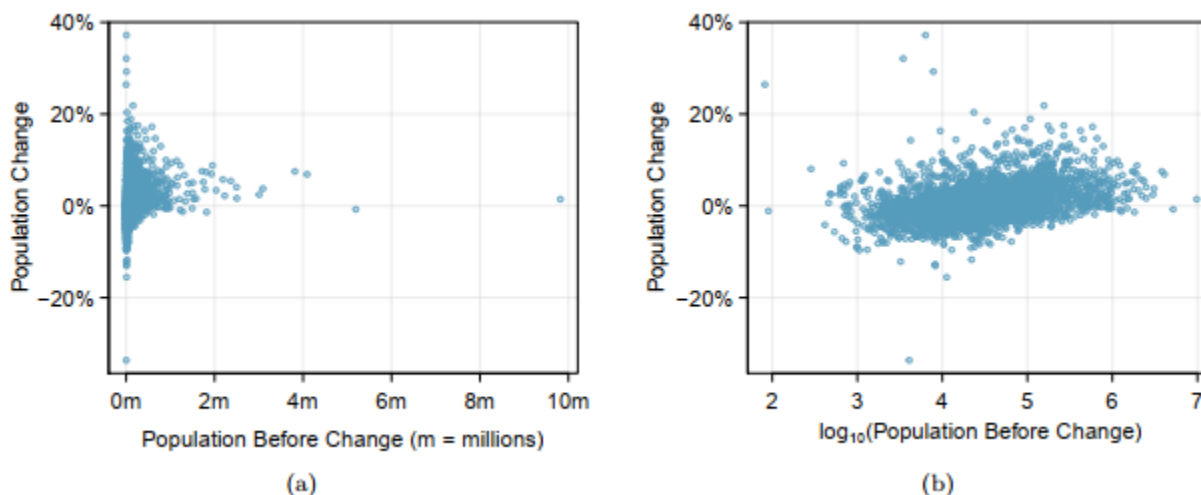
Điều gì khiến biểu đồ này kém hữu ích?

Phần lớn dữ liệu tập trung vào khoảng giá trị bên trái và độ lệch cực lớn làm che khuất nhiều chi tiết quan trọng tiềm ẩn trong tập dữ liệu.

Có một số phép chuyển đổi tiêu chuẩn có thể hữu ích đối với dữ liệu có độ lệch phải mạnh khi phần lớn giá trị dương tập trung gần 0. Một **phép chuyển đổi** thực chất là tái tỷ lệ dữ liệu bằng một hàm số. Ví dụ, khi áp dụng phép chuyển đổi logarit cơ số 10 cho dân số các quận, ta thu được biểu đồ tần suất mới như trong Hình 2.13(b). Dữ liệu này có tính đối xứng, và bất kỳ giá trị ngoại lai tiềm ẩn nào xuất hiện đều ít cực đoan hơn so với trong bộ dữ liệu gốc. Bằng cách kiểm soát các giá trị ngoại lai và độ lệch phân phối, các phép chuyển đổi như thế này thường giúp việc xây dựng các mô hình thống kê đối với dữ liệu trở nên dễ dàng hơn.

Các phép biến đổi cũng có thể được áp dụng cho một hoặc cả hai biến trong biểu đồ phân tán. Biểu đồ phân tán thay đổi dân số từ năm 2010 đến 2017 so với dân số năm 2010 được trình bày trong Hình 2.14(a). Trong biểu đồ phân tán đầu tiên này, thật khó để nhận diện bất kỳ mẫu hình quan trọng nào vì biến dân số có sự lệch phân phối mạnh. Tuy nhiên, nếu chúng ta áp dụng phép biến đổi $\log_{10}$ đối với biến dân số như được trình bày trong Hình 2.14(b) thì một mối quan hệ tích cực giữa các biến sẽ được làm rõ. Thực tế, chúng ta có thể quan tâm đến việc điều chỉnh một đường xu hướng cho dữ liệu khi nghiên cứu các phương pháp liên quan đến việc điều chỉnh các đường hồi quy trong Chương 8.

Các phép biến đổi khác ngoài phép logarit cũng có thể mang lại hiệu quả. Ví dụ, căn bậc hai ($\sqrt{\text{quan sát gốc}}$) và nghịch đảo ($\frac{1}{\text{quan sát gốc}}$) là những phương pháp thường được các nhà khoa học dữ liệu sử dụng. Các mục tiêu phổ biến trong việc chuyển đổi dữ liệu là nhìn nhận cấu trúc dữ liệu theo một cách khác, giảm độ lệch phân phối, hỗ trợ trong việc mô hình hóa, hoặc làm thẳng mối quan hệ phi tuyến trong biểu đồ phân tán.



Hình 2.14: (a) Biểu đồ phân tán của biến động dân số so với dân số trước biến động. (b) Biểu đồ phân tán của cùng một dữ liệu nhưng với kích thước dân số đã được chuyển đổi bằng log.

2.1.8 Lập bản đồ dữ liệu (chủ đề đặc biệt)

Bộ dữ liệu **country** cung cấp nhiều biến số định lượng mà chúng ta có thể biểu diễn thông qua các biểu đồ chấm, biểu đồ phân tán hoặc biểu đồ hộp nhưng những phương pháp này chưa phản ánh đúng bản chất của dữ liệu. Hơn nữa khi gặp các dữ liệu địa lý chúng ta lên tạo các **bản đồ mật độ**, trong đó màu sắc được sử dụng để thể hiện các giá trị cao và thấp của một biến. Hình 2.15 và 2.16 trình bày bản đồ mật độ cho tỷ lệ nghèo đói tính theo phần trăm (biến **poverty**), tỷ lệ thất nghiệp (biến **unemployment_rate**), tỷ lệ sở hữu nhà tính theo phần trăm (biến **homeownership**) và trung bình thu nhập hộ gia đình (biến **median_hh_income**). Bảng chú giải màu sắc thể hiện sự tương ứng giữa màu sắc và giá trị. Các bản đồ mật độ thường không thực sự hữu ích cho việc xác định bất kỳ quận nào nhưng chúng rất hữu ích trong việc quan sát các xu hướng địa lý và tạo ra các câu hỏi nghiên cứu hoặc giả thuyết thú vị.

VÍ DỤ 2.21

Điều thú vị nào có thể nhận thấy trong các bản đồ mật độ theo biến số **poverty** và **unemployment_rate**?

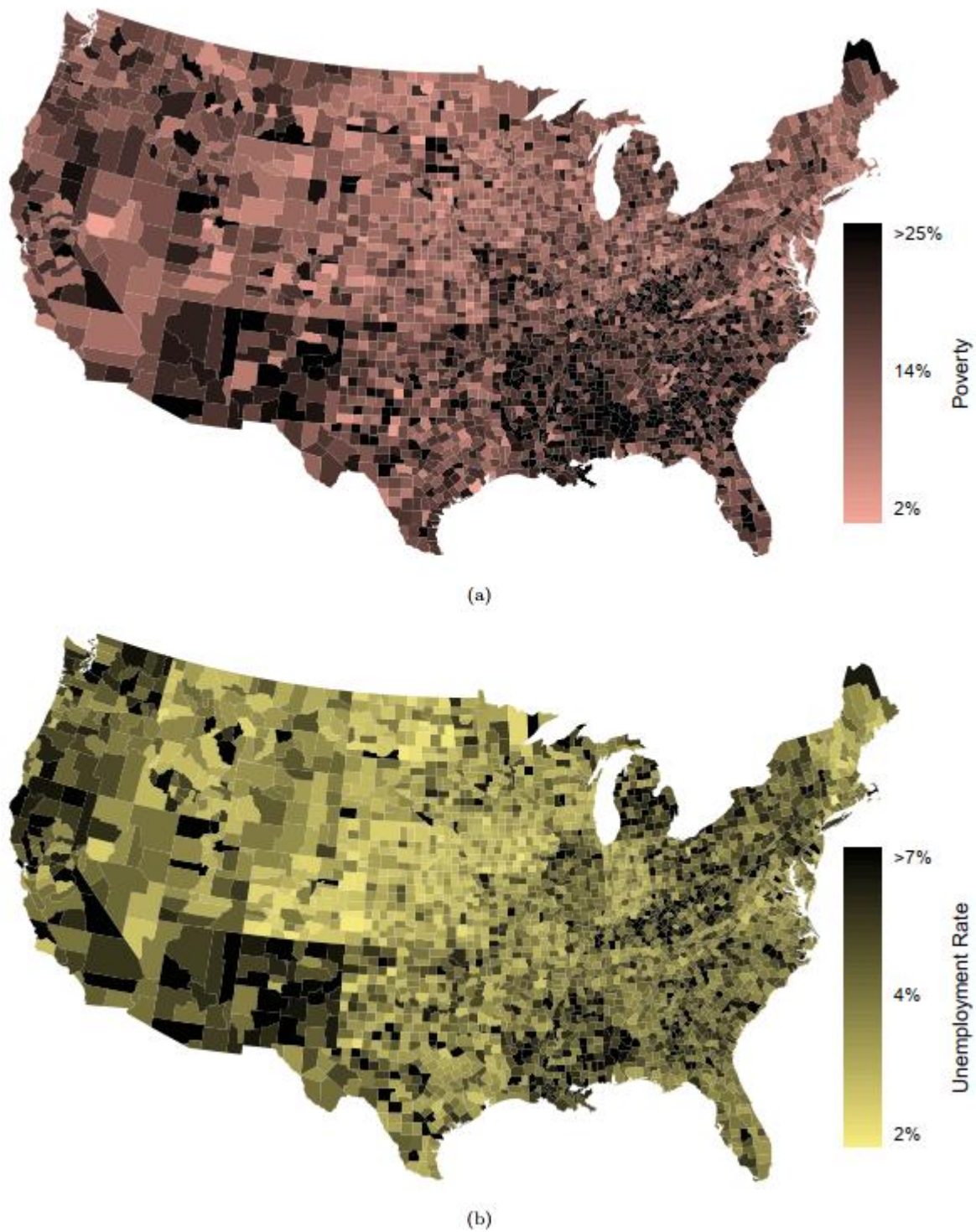
Tỷ lệ nghèo đói rõ ràng cao hơn ở một số khu vực. Đáng chú ý là các khu vực sau về phía Nam có tỷ lệ nghèo đói cao, tương tự như phần lớn Arizona và New Mexico. Tỷ lệ nghèo đói cao cũng được ghi nhận rõ ràng ở bãi bồi sông Mississippi cách New Orleans một khoảng về phía bắc, cũng như ở một khu vực rộng lớn thuộc Kentucky.

Tỷ lệ thất nghiệp cũng tuân theo xu hướng tương tự, và chúng ta có thể thấy sự tương quan giữa hai biến số này. Trên thực tế, tỷ lệ thất nghiệp cao có mối liên hệ chặt chẽ với tỷ lệ nghèo đói. Một quan sát nổi bật khi so sánh hai bản đồ là tỷ lệ nghèo đói cao hơn nhiều so với tỷ lệ thất nghiệp điều này có nghĩa là mặc dù nhiều người có thể đang làm việc, nhưng họ không kiếm đủ thu nhập để thoát khỏi cảnh nghèo.

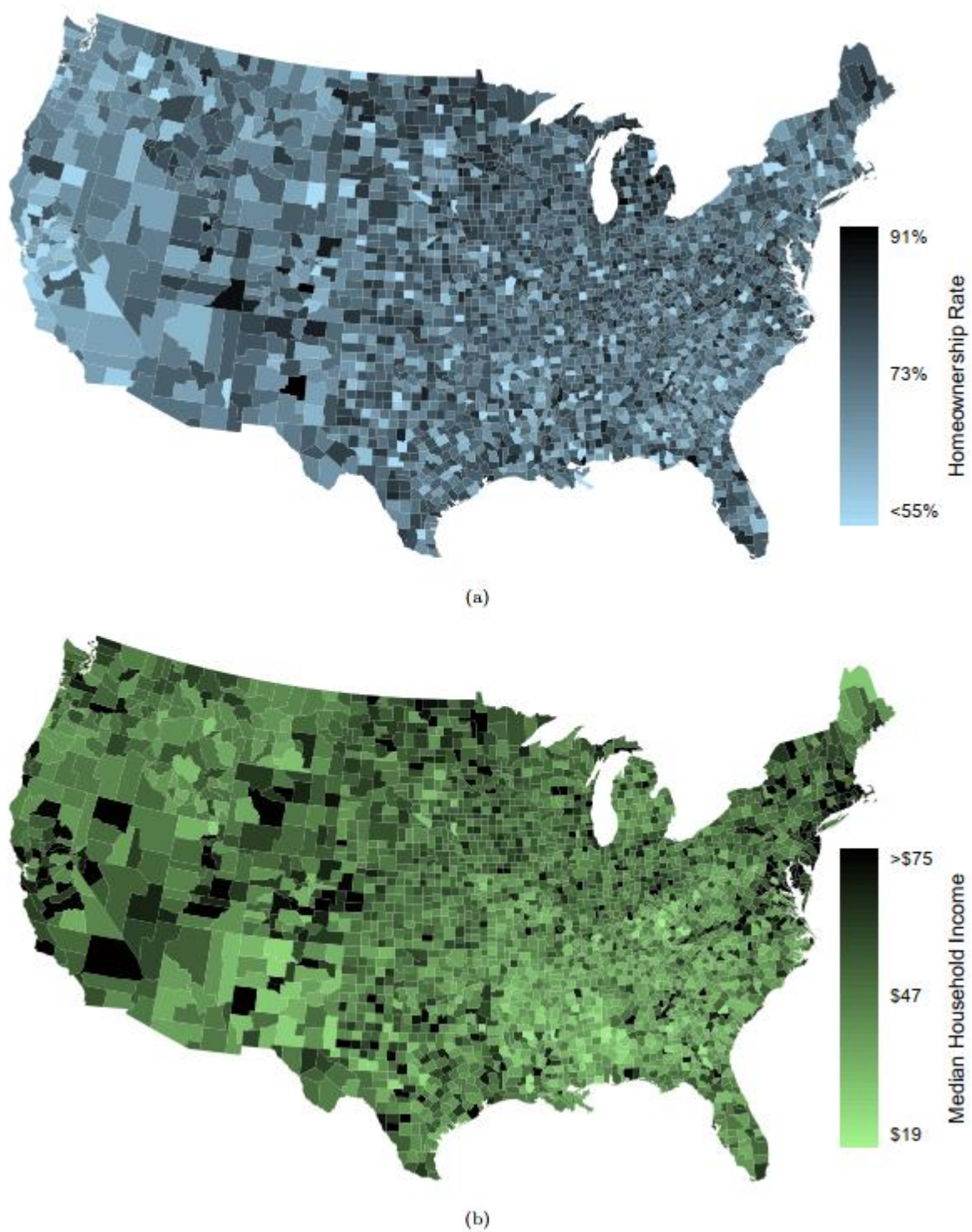
BÀI TẬP THỰC HÀNH 2.22

Những đặc điểm thú vị nào được thể hiện rõ trên bản đồ mật độ thu nhập trung bình hộ gia đình trong Hình 2.16(b)?¹⁴

¹⁴ Lưu ý: Câu trả lời có thể khác nhau. Có một số sự tương quan giữa thu nhập cao và các khu vực đô thị, nơi chúng ta có thể thấy các điểm tối hơn (thu nhập trung bình hộ gia đình cao hơn) mặc dù có một số ngoại lệ. Bạn có thể tìm kiếm các thành phố lớn mà bạn quen thuộc và thử xác định chúng trên bản đồ dưới dạng các điểm tối.



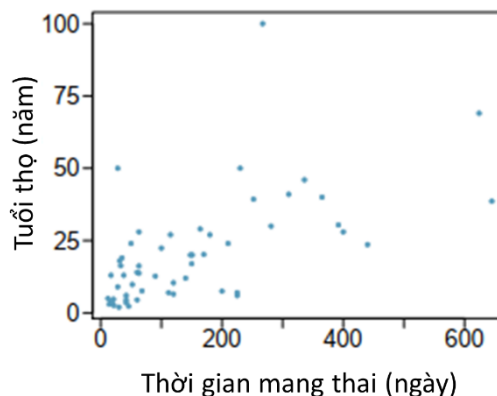
Hình 2.15: (a) Bản đồ mật độ tỷ lệ nghèo đói (phần trăm). (b) Bản đồ tỷ lệ thất nghiệp (phần trăm).



Hình 2.16: (a) Bản đồ mật độ tỷ lệ sở hữu nhà ở (phần trăm). (b) Bản đồ mật độ thu nhập trung bình hộ gia đình (\$1000s).

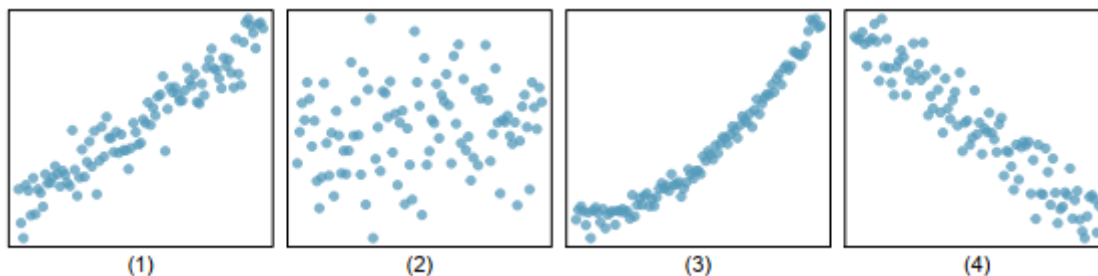
Bài tập

2.1 Tuổi thọ của động vật có vú. Dữ liệu thu thập về tuổi thọ (năm) và thời gian mang thai (ngày) của 62 loài động vật có vú. Biểu đồ phân tán của tuổi thọ so với thời gian mang thai được thể hiện bên dưới.¹⁵



- Loại mối quan hệ giữa tuổi thọ và thời gian mang thai là gì?
- Nếu đảo ngược các trục của biểu đồ, tức là vẽ thời gian mang thai theo tuổi thọ, bạn kỳ vọng sẽ thấy mối quan hệ nào?
- Tuổi thọ và thời gian mang thai có độc lập không? Giải thích câu trả lời của bạn?

2.2 Các mối quan hệ. Chỉ ra biểu đồ nào cho thấy (a) tương quan dương, (b) tương quan âm hoặc (c) không tương quan. Đồng thời, xác định xem các mối tương quan dương và âm có phải là tuyến tính hay phi tuyến. Mỗi phần có thể áp dụng cho nhiều hơn một biểu đồ.



2.3 Sự sinh sản của vi khuẩn. Sự sinh sản của vi khuẩn. Giả sử rằng chỉ có đủ không gian và chất dinh dưỡng để hỗ trợ một triệu tế bào vi khuẩn trong đĩa petri. Bạn đặt một vài tế bào vi khuẩn vào đĩa petri này, cho phép chúng sinh sản tự do và ghi lại số lượng tế bào vi khuẩn trong đĩa theo thời gian. Hãy phác thảo một biểu đồ thể hiện mối quan hệ giữa số lượng tế bào vi khuẩn và thời gian.

2.4 Năng suất làm việc tại văn phòng. Năng suất làm việc thường thấp khi nhân viên không cảm thấy áp lực về công việc hoặc sự ổn định công việc. Tuy nhiên, mức độ căng thẳng quá cao cũng có thể dẫn đến giảm năng suất của nhân viên. Hãy phác thảo một biểu đồ để thể hiện mối quan hệ giữa mức độ căng thẳng và năng suất làm việc.

2.5 Tham số và thống kê. Xác định giá trị nào thể hiện trung bình mẫu và giá trị nào thể hiện trung bình tổng thể qua các tình huống sau:

- Các hộ gia đình Mỹ đã chi tiêu trung bình khoảng 52 đô la vào năm 2007 cho các sản phẩm liên quan đến Halloween như trang phục, đồ trang trí và kẹo. Để kiểm tra xem mức chi tiêu này có thay đổi hay không, các nhà nghiên cứu đã thực hiện một khảo sát mới vào năm 2008 trước khi có số liệu chính thức từ ngành công nghiệp. Cuộc khảo sát này bao gồm 1.500 hộ gia đình và kết quả cho thấy mức chi tiêu trung bình cho Halloween là 58 đô la mỗi hộ.
- Điểm trung bình (GPA) của sinh viên tại một trường đại học tư thục vào năm 2001 là 3,37. Một cuộc khảo sát trên mẫu gồm 203 sinh viên từ trường đại học này đã cho thấy điểm trung bình GPA là 3,59 sau một thập kỷ.

2.6 Giấc ngủ của sinh viên đại học. Một bài báo gần đây trên tờ báo của trường đại học nói rằng sinh viên đại học ngủ trung bình 5,5 giờ mỗi đêm. Một sinh viên tỏ ra hoài nghi về giá trị này đã quyết định tiến hành một cuộc khảo sát bằng cách chọn ngẫu nhiên 25 sinh viên. Kết quả cho thấy, trung bình các sinh viên được khảo sát ngủ 6,25 giờ mỗi đêm. Hãy xác định giá trị nào đại diện cho trung bình mẫu và giá trị nào đại diện cho trung bình tổng thể được tuyên bố.

¹⁵T. Allison and D.V. Cicchetti. "Sleep in mammals: ecological and constitutional correlates". In: *Arch. Hydrobiol* 75 (1975), p. 442.

2.7 Ngày nghỉ tại một nhà máy khai thác mỏ. Công nhân tại một khu khai thác mỏ nhận được trung bình 35 ngày nghỉ phép có lương, thấp hơn so với mức trung bình quốc gia. Người quản lý của nhà máy này đang chịu áp lực từ công đoàn địa phương để tăng số ngày nghỉ phép có lương. Tuy nhiên, ông ta không muốn cung cấp thêm ngày nghỉ cho công nhân vì điều đó sẽ tốn kém. Thay vào đó, ông quyết định sa thải 10 nhân viên sao cho có thể làm tăng số ngày nghỉ phép trung bình được báo cáo bởi các nhân viên của mình. Để đạt được mục tiêu này, ông ta nên sa thải những nhân viên có nhiều ngày nghỉ nhất, ít ngày nghỉ nhất, hay những người có số ngày nghỉ gần bằng mức trung bình?

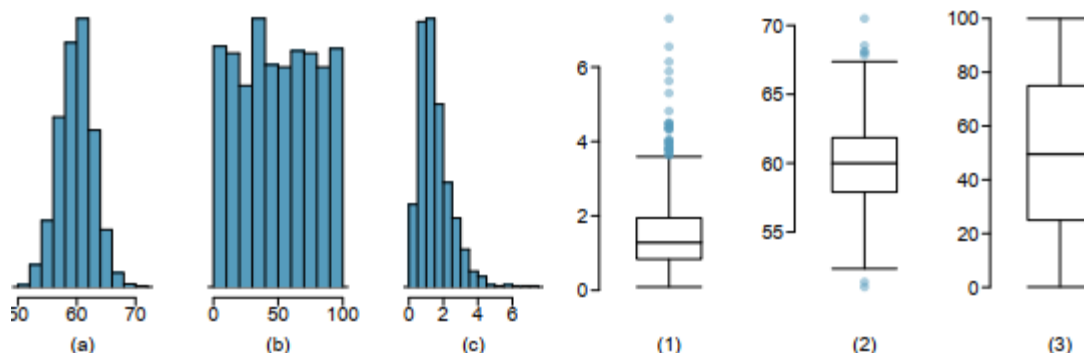
2.8 Trung vị và khoảng tứ phân vị. Với mỗi phần, hãy so sánh phân phối (1) và (2) dựa trên trung vị và khoảng tứ phân vị của chúng. Bạn không cần tính toán các tham số này, chỉ cần nêu rõ cách so sánh giữa các trung vị và IQRs. Đảm bảo giải thích rõ lý luận của bạn.

- | | | | |
|-----|--------------------|-----|----------------------------|
| (a) | (1) 3, 5, 6, 7, 9 | (c) | (1) 1, 2, 3, 4, 5 |
| | (2) 3, 5, 6, 7, 20 | | (2) 6, 7, 8, 9, 10 |
| (b) | (1) 3, 5, 6, 7, 9 | (d) | (1) 0, 10, 50, 60, 100 |
| | (2) 3, 5, 7, 8, 9 | | (2) 0, 100, 500, 600, 1000 |

2.9 Giá trị trung bình và độ lệch chuẩn. Với mỗi phần, hãy so sánh phân phối (1) và (2) dựa trên giá trị trung bình và độ lệch chuẩn của chúng. Bạn không cần tính toán các tham số này, chỉ cần nêu rõ cách so sánh giữa các giá trị trung bình và độ lệch chuẩn. Đảm bảo giải thích rõ lý luận của bạn. *Gợi ý:* Có thể hữu ích nếu vẽ biểu đồ chấm của các phân phối.

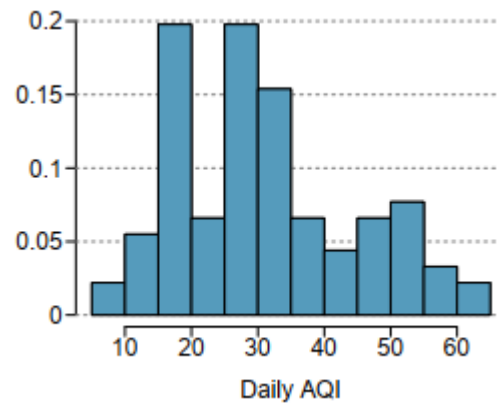
- | | | | |
|-----|-----------------------------------|-----|-----------------------------|
| (a) | (1) 3, 5, 5, 5, 8, 11, 11, 11, 13 | (c) | (1) 0, 2, 4, 6, 8, 10 |
| | (2) 3, 5, 5, 5, 8, 11, 11, 11, 20 | | (2) 20, 22, 24, 26, 28, 30 |
| (b) | (1) -20, 0, 0, 0, 15, 25, 30, 30 | (d) | (1) 100, 200, 300, 400, 500 |
| | (2) -40, 0, 0, 0, 15, 25, 30, 30 | | (2) 0, 50, 300, 550, 600 |

2.10 Kết hợp và ghép nối. Mô tả phân phối trong các biểu đồ tần suất dưới đây và ghép chúng với các biểu đồ hộp tương ứng.

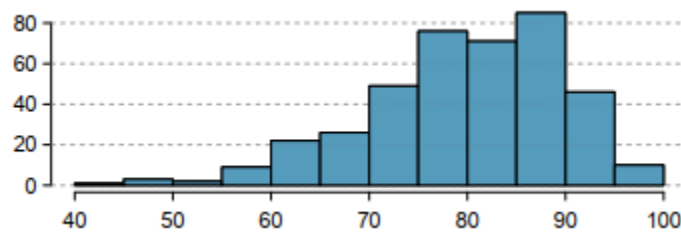


2.11 Chất lượng không khí. Chất lượng không khí hàng ngày được đo bằng chỉ số chất lượng không khí do Cơ quan Bảo vệ Môi trường công bố. Chỉ số này phản ánh mức độ ô nhiễm và các tác có thể đáng lo ngại liên quan đến sức khỏe. Chỉ số được tính toán dựa trên năm loại ô nhiễm không khí chính được quy định bởi Đạo luật Không khí Sạch và nhận giá trị từ 0 đến 300, trong đó giá trị càng cao cho thấy chất lượng không khí càng thấp. AQI đã được báo cáo cho một mẫu gồm 91 ngày trong năm 2011 tại Durham, NC. Biểu đồ tần suất tương đối dưới đây thể hiện phân phối của các giá trị AQI trong những ngày này.¹⁶

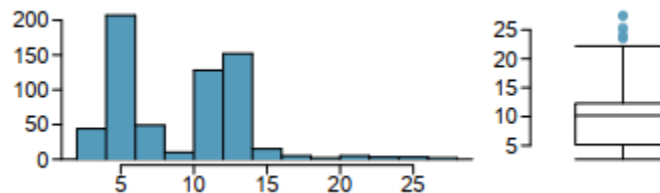
- Ước lượng giá trị trung vị của AQI trong mẫu này.
- Hãy dự đoán giá trị trung bình của AQI trong mẫu này có sẽ cao hơn hay thấp hơn giá trị trung vị? Giải thích lý do.
- Ước tính Q1, Q3 và IQR cho phân phối này.
- Có ngày nào trong mẫu này được cho là có AQI thấp hoặc quá cao bất thường không? Giải thích.



2.12 Trung vị so với giá trị trung bình. Hãy tính toán trung vị cho 400 quan sát được thể hiện trong biểu đồ tần suất và cho biết bạn kỳ vọng giá trị trung bình sẽ cao hơn hay thấp hơn trung vị.



2.13 Biểu đồ tần suất so với biểu đồ hộp. So sánh hai biểu đồ dưới đây. Những đặc điểm nào của phân phối được thể hiện rõ trong biểu đồ tần suất mà không có trong biểu đồ hộp? Những đặc điểm nào được thể hiện rõ trong biểu đồ hộp mà không có trong biểu đồ tần suất?



2.14 Bạn bè trên Facebook. Dữ liệu từ Facebook cho thấy 50% người dùng Facebook có 100 bạn trở lên và số lượng bạn bè trung bình của người dùng là 190. Những phát hiện này gợi ý gì về hình dạng phân phối của số lượng bạn bè trên Facebook?¹⁷

2.15 Phân phối và các tham số phù hợp, Phần I. Với mỗi trường hợp sau, hãy cho biết bạn kỳ vọng phân phối sẽ đối xứng, lệch phải, hay lệch trái. Đồng thời, chỉ ra liệu giá trị trung bình hay trung vị sẽ đại diện tốt hơn cho một quan sát điển hình trong dữ liệu và liệu sự phân tán của các quan sát sẽ được biểu diễn tốt nhất bằng độ lệch chuẩn (standard deviation) hay khoảng tứ phân vị (IQR). Giải thích lý luận của bạn.

- Số lượng thú cưng trong mỗi hộ gia đình
- Khoảng cách đến nơi làm việc, tức số dặm giữa nhà và nơi làm việc
- Chiều cao của nam giới trưởng thành.

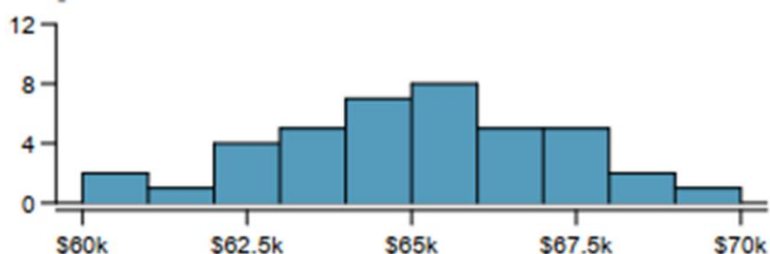
¹⁶ US Environmental Protection Agency, AirData, 2011.

¹⁷ Lars Backstrom. "Anatomy of Facebook". In: *Facebook Data Team's Notes* (2011).

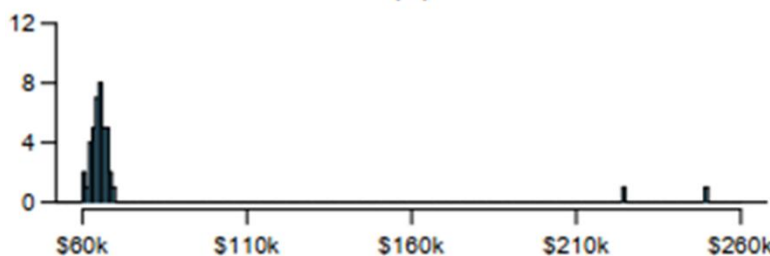
2.16 Phân phối và các tham số phù hợp, Phần II. Với mỗi trường hợp sau, hãy cho biết bạn kỳ vọng phân phối sẽ đối xứng, lệch phải, hay lệch trái. Đồng thời, chỉ ra liệu giá trị trung bình hay trung vị sẽ đại diện tốt hơn cho một quan sát điển hình trong dữ liệu, và liệu sự phân tán các quan sát sẽ được biểu diễn tốt nhất bằng độ lệch chuẩn hay khoảng tứ phân vị. Giải thích lý luận của bạn.

- Giá nhà ở một quốc gia mà 25% số nhà có giá dưới \$350 000, 50% số nhà có giá dưới \$450 000, 75% số nhà có giá dưới \$1 000 000 và một số lượng đáng kể nhà có giá trên \$6 000 000.
- Giá nhà ở một quốc gia mà 25% số nhà có giá dưới \$300 000, 50% số nhà có giá dưới \$600 000, 75% số nhà có giá dưới \$900 000 và rất ít nhà có giá trên \$1 200 000.
- Số lượng đồ uống có cồn mà sinh viên đại học tiêu thụ trong một tuần. Giả sử rằng hầu hết sinh viên không uống vì họ dưới 21 tuổi và chỉ một số ít uống quá nhiều.
- Mức lương hàng năm của nhân viên tại một công ty Fortune 500, nơi mà chỉ một số ít lãnh đạo cấp cao có mức lương cao hơn nhiều so với tất cả các nhân viên khác.

2.17 Thu nhập tại quán cà phê. Biểu đồ tần suất đầu tiên dưới đây thể hiện phân phối thu nhập hàng năm của 40 khách hàng tại một quán cà phê trong trường đại học. Giả sử có hai người mới bước vào quán cà phê: một người có thu nhập \$225 000 và người kia là \$250 000. Biểu đồ tần suất thứ hai thể hiện phân phối thu nhập mới. Các tham số thống kê cũng được cung cấp.



(1)



(2)

	(1)	(2)
n	40	42
Nhỏ nhất	60680	60680
Q1	63620	63710
Trung vị	65240	65350
Trung bình	65090	73300
Q3	66160	66540
Lớn nhất	69890	250000
Độ lệch chuẩn	2122	37321

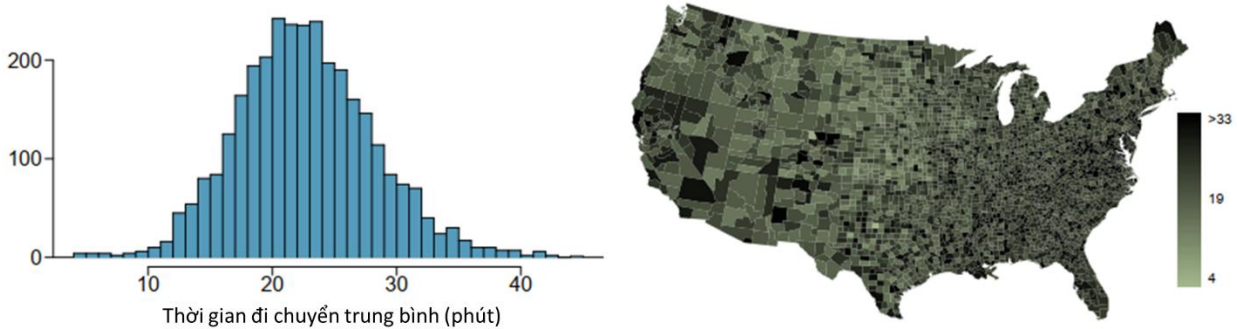
- Giá trị trung bình hay trung vị sẽ đại diện tốt hơn cho mức thu nhập điển hình của 42 khách hàng tại quán cafe này? Điều này nói lên điều gì về tính chuẩn mạnh của hai đại lượng thống kê?
- Độ lệch chuẩn hay khoảng tứ phân vị (IQR) sẽ đại diện tốt hơn cho mức độ biến thiên trong thu nhập của 42 khách hàng tại quán cà phê này? Điều này phản ánh gì về tính chuẩn mạnh của hai thước đo này?

2.18 Trung bình khoảng. Trung bình khoảng của một phân phối được định nghĩa là trung bình cộng của giá trị lớn nhất và giá trị nhỏ nhất trong phân phối đó. Liệu thống kê này có mạnh đối với các giá trị ngoại lai và độ lệch cực đoan hay không? Giải thích lý luận của bạn.

2.1 KIỂM TRA DỮ LIỆU ĐỊNH LƯỢNG

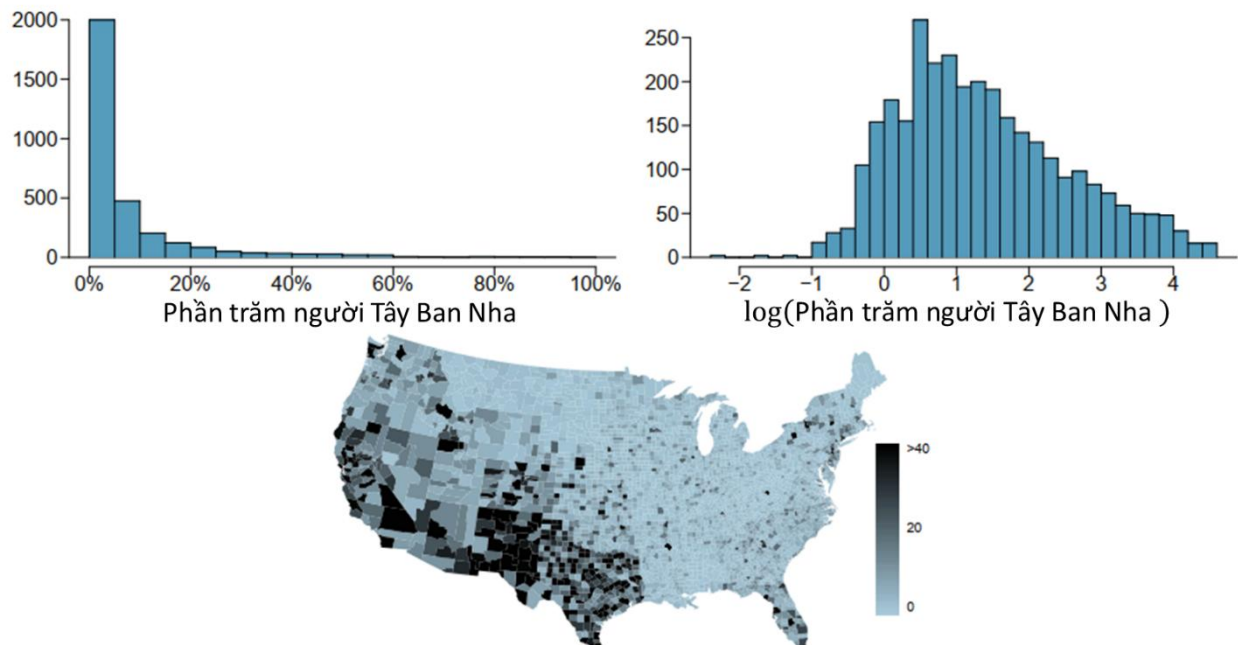
60

2.19 Thời gian di chuyển. Điều tra dân số Hoa Kỳ thu thập dữ liệu về thời gian di chuyển đi làm của người Mỹ, cùng với nhiều biến số khác. Biểu đồ tần suất dưới đây thể hiện phân phối thời gian di chuyển trung bình tại 3142 quận của Hoa Kỳ vào năm 2010. Bên cạnh đó, một bản đồ mật độ không gian của cùng dữ liệu này cũng được trình bày.



- (a) Mô tả đặc điểm phân phối và nhận xét liệu phép biến đổi logarit có phù hợp để áp dụng cho tập dữ liệu này hay không.
(b) Mô tả sự phân phối không gian của thời gian di chuyển dựa trên bản đồ trên.

2.20 Dân số người Mỹ gốc Tây Ban Nha. Theo cuộc Tổng điều tra dân số Hoa Kỳ thu thập dữ liệu về chủng tộc và sắc tộc của người Mỹ, cùng nhiều biến số khác. Biểu đồ tần suất dưới đây thể hiện phân phối của tỷ lệ phần trăm dân số là người gốc Mỹ gốc Tây Ban Nha tại 3142 quận của Hoa Kỳ vào năm 2010. Ngoài ra, biểu đồ tần suất của logarit các giá trị này cũng được thể hiện.



- (a) Mô tả đặc điểm phân phối và luận giải tại sao phép biến đổi logarithm có thể được cân nhắc sử dụng trong phân tích hoặc mô hình hóa dữ liệu này.
(b) Những đặc trưng phân phối nào của dân số Hispanic tại các county Hoa Kỳ được biểu thị rõ qua bản đồ nhưng không thể hiện trên biểu đồ tần suất? những đặc trưng nào chỉ quan sát được trên biểu đồ tần suất mà không nhận diện được qua bản đồ?
(c) Có phải một trong hai cách trực quan hóa dữ liệu này thích hợp hoặc hữu ích hơn cách còn lại? Hãy giải thích lập luận của bạn.

2.2 Xem xét dữ liệu định tính

Trong phần này, chúng tôi sẽ giới thiệu các bảng và các công cụ khác để phân tích dữ liệu định tính được sử dụng xuyên suốt cuốn sách này. Bộ số liệu **loan50** đại diện cho một mẫu được trích xuất từ bộ dữ liệu lớn hơn có tên **loans**. Bộ dữ liệu lớn này chứa thông tin của 10 000 khoản vay được thực hiện thông qua Lending Club. Chúng tôi sẽ kiểm tra mối quan hệ giữa tình trạng sở hữu nhà trong tập dữ liệu khoản vay có thể lấy các giá trị như thuê (**rent**), thế chấp (**mortgage**) (sở hữu nhưng còn nợ thế chấp) hoặc **own** (sở hữu hoàn toàn) và **app_type** (loại hình áp dụng), xác định liệu đơn vay đồng sở hữu hay là đơn vay cá nhân.

2.2.1 Bảng tần suất chéo và biểu đồ cột

Hình 2.17 mô tả hai biến số: **app_type** loại hình áp dụng và **homeownership** tình trạng sở hữu nhà. Một bảng mô tả dữ liệu cho hai biến phân loại theo cách này được gọi là **bảng tần suất chéo**. Mỗi giá trị trong bảng đại diện cho số lần một tổ hợp kết quả biến cụ thể xuất hiện. Ví dụ, giá trị 3496 tương ứng với số lượng khoản vay trong tập dữ liệu mà người vay thuê nhà và loại hình áp dụng là cá nhân. Tổng hàng và tổng cột cũng được bao gồm. **Tổng hàng** cung cấp tổng số lượng trên mỗi hàng (ví dụ: $3496 + 3839 + 1170 = 8505$) và **tổng cột** là tổng số lượng theo mỗi cột. Chúng ta cũng có thể tạo một bảng chỉ hiển thị tỷ lệ phần trăm hoặc tỷ lệ tổng thể cho mỗi tổ hợp hoặc chúng ta có thể tạo một bảng cho một biến đơn lẻ, chẳng hạn như bảng được hiển thị trong Hình 2.18 cho **homeownership** - biến tình trạng sở hữu nhà.

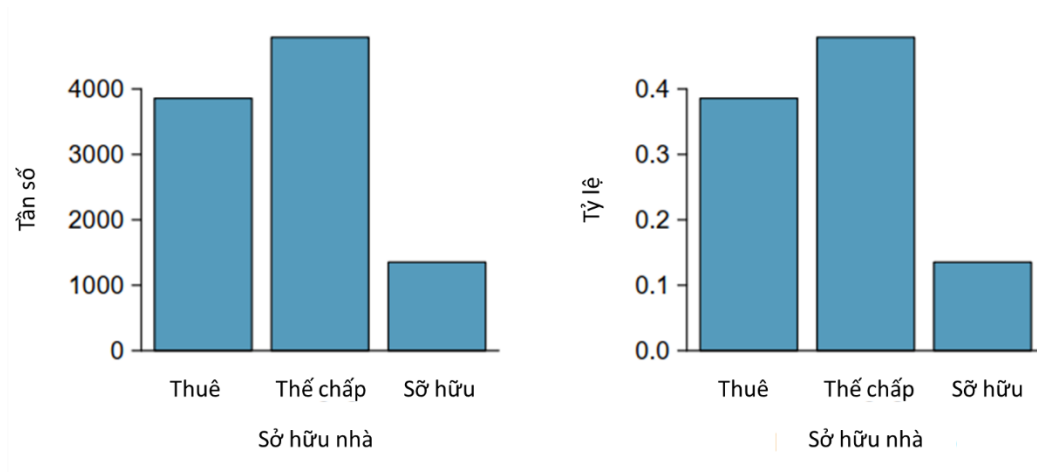
		homeownership			
		Thuê	Thế chấp	Sở hữu	Tổng
app_type	Cá nhân	3496	3839	1170	8505
	Đồng sở hữu	362	950	183	1495
	Tổng	3858	4789	1353	10000

Hình 2.17: Bảng tần suất chéo cho biến **app_type** và biến **homeownership**.

homeownership	Tần số
Thuê	3858
Thế chấp	4789
Sở hữu	1353
Tổng	10000

Hình 2.18: Bảng mô tả tần suất của mỗi giá trị cho biến **homeownership**.

Biểu đồ cột là một phương pháp phổ biến để hiển thị một biến định tính đơn lẻ. Bảng bên trái của Hình 2.19 trình bày **biểu đồ cột** cho biến **homeownership**. Trong bảng bên phải, số lượng được chuyển đổi thành tỷ lệ, cho thấy tỷ lệ quan sát thuộc mỗi cấp độ (ví dụ: $3858/10000 = 0,3858$ cho thuê).



Hình 2.19: Hai biểu đồ cột biểu diễn số lượng (number). Bảng bên trái hiển thị số lượng và bảng bên phải hiển thị tỷ lệ trong mỗi nhóm.

2.2.2 Tỷ Lệ Hàng và Cột

Đôi khi, việc hiểu rõ sự phân chia theo tỷ lệ của một biến trong một biến khác là hữu ích và chúng ta có thể điều chỉnh bảng tần suất chéo để cung cấp cái nhìn như vậy. Hình 2.20 trình bày **tỷ lệ hàng** cho Hình 2.17 được tính bằng cách lấy số lượng chia cho tổng hàng của chúng. Giá trị 3496 tại giao điểm của cá nhân và thuê được thay thế bằng $3496/8505 = 0,411$ tức là 3496 chia cho tổng hàng của nó là 8505. Vậy 0,411 đại diện cho điều gì? Nó tương ứng với tỷ lệ người nộp đơn cá nhân đang thuê nhà.

	Thuê	Thế chấp	Sở hữu	Tổng
Cá nhân	0,411	0,451	0,138	1,000
Đồng sở hữu	0,242	0,635	0,122	1,000
Tổng	0,386	0,479	0,135	1,000

Hình 2.20: Bảng tần suất chéo với tỷ lệ hàng cho các biến **app_type** và **homeownership**. Tổng hàng bị sai lệch 0,001 đối với hàng đồng sở hữu do sai số làm tròn.

Bảng tần suất chéo của tỷ lệ cột được tính toán theo cách tương tự, trong đó **mỗi tỷ lệ cột** được tính bằng cách lấy số lượng chia cho tổng cột tương ứng. Hình 2.21 hiển thị một bảng như vậy và ở đây giá trị 0,906 cho thấy 90,6% người thuê nhà nộp đơn vay với tư cách cá nhân. Tỷ lệ này cao hơn so với các khoản vay từ những người có thể chấp (80,2%) hoặc sở hữu nhà (86,5%). Vì các tỷ lệ này khác nhau giữa ba cấp độ của tình trạng sở hữu nhà (thuê, thế chấp, sở hữu) - **homeownership** (rent, mortgage, own), điều này cung cấp bằng chứng rằng các biến **app_type** và **homeownership**.

	Thuê	Thế chấp	Sở hữu	Tổng
Cá nhân	0,906	0,802	0,865	0,851
Đồng sở hữu	0,094	0,198	0,135	0,150
Tổng	1,000	1,000	1,000	1,000

Hình 2.21: Bảng tần suất chéo với tỷ lệ cột cho các biến **app_type** và **homeownership**. Tổng của cột cuối cùng bị sai lệch 0,001 do sai số làm tròn.

Chúng ta cũng có thể kiểm tra sự tương quan giữa biến **app_type** và **homeownership** trong Hình 2.20 bằng cách sử dụng tỷ lệ hàng. Khi so sánh các tỷ lệ hàng này, chúng ta sẽ xem xét theo cột để xác định liệu tỷ lệ khoản vay mà người vay thuê, thế chấp hoặc sở hữu nhà có sự khác biệt giữa các loại hình áp dụng cá nhân (individual) và đồng sở hữu (joint) hay không.

BÀI TẬP THỰC HÀNH 2.23

G

- (a) Giá trị 0,451 đại diện cho điều gì trong Hình 2.20?
 (b) Giá trị 0,802 đại diện cho điều gì trong Hình 2.21? ¹⁸

BÀI TẬP THỰC HÀNH 2.24

G

- (a) Giá trị 0,122 tại giao điểm của đồng sở hữu và sở hữu đại diện cho điều gì trong Hình 2.20?
 (b) Giá trị 0,135 đại diện cho điều gì trong Hình 2.21? ¹⁹

VÍ DỤ 2.25

E

Các nhà khoa học dữ liệu sử dụng thống kê để lọc thư rác khỏi các tin nhắn email đến. Bằng cách ghi nhận các đặc điểm cụ thể của một email, một nhà khoa học dữ liệu có thể phân loại một số email là thư rác hoặc không phải thư rác với độ chính xác cao. Một trong những đặc điểm như vậy là liệu email có chứa số hay không, số nhỏ hay số lớn. Một đặc điểm khác là định dạng email cho biết liệu email có chứa nội dung HTML hay không, chẳng hạn như văn bản in đậm. Chúng ta sẽ tập trung vào định dạng email và trạng thái thư rác bằng cách sử dụng tập dữ liệu email và các biến này được mô tả trong bảng tần suất chéo trong Hình 2.22. Đối với bảng này, tỷ lệ hàng hay tỷ lệ cột sẽ hữu ích hơn cho người muốn phân loại email là thư rác hay email thông thường?

Một nhà khoa học dữ liệu sẽ quan tâm đến việc tỷ lệ thư rác thay đổi như thế nào trong từng định dạng email. Điều này tương ứng với tỷ lệ cột: tỷ lệ thư rác trong email văn bản thuần túy và tỷ lệ thư rác trong email HTML.

Nếu chúng ta tính toán tỷ lệ cột, chúng ta có thể thấy rằng một phần lớn hơn của email văn bản thuần túy là thư rác ($209/1195 = 17,5\%$) so với email HTML ($158/2726 = 5,8\%$). Bản thân thông tin này không đủ để phân loại một email là thư rác hay không phải thư rác, vì hơn 80% email văn bản thuần túy không phải là thư rác. Tuy nhiên, khi chúng ta kết hợp cẩn thận thông tin này với nhiều đặc điểm khác, chúng ta có thể phân loại một số email là thư rác hoặc không phải thư rác một cách tự tin.

	Văn bản thuần	HTML	Tổng
Thư rác	209	158	367
Không phải thư rác	986	2568	3554
Tổng	1195	2726	3921

Hình 2.22: Bảng tần suất chéo cho thư rác (spam) và định dạng (format).

Ví dụ 2.25 chỉ ra rằng tỷ lệ hàng và tỷ lệ cột không tương đương. Trước khi quyết định một dạng cho bảng, điều quan trọng là phải xem xét từng dạng để đảm bảo rằng bảng hữu ích nhất được xây dựng. Tuy nhiên, đôi khi đơn giản là không rõ dạng nào nếu có hữu ích hơn.

VÍ DỤ 2.26

E

Hãy xem lại bảng 2.20 và 2.21. Liệu có trường hợp nào rõ ràng mà một bảng hữu ích hơn bảng kia không?

Không có trường hợp nào mà chúng tôi thấy rõ ràng! Điều khác biệt giữa biến `app_type` và `homeownership` so với ví dụ email là hai biến này không có mối quan hệ biến giải thích-biến phản hồi rõ ràng mà chúng ta có thể giả định (xem Phần 1.2.4 để biết các thuật ngữ này). Thông thường, việc "điều kiện hóa" trên biến giải thích là hữu ích nhất. Ví dụ, trong ví dụ email định dạng email được xem là một biến giải thích tiềm năng cho việc tin nhắn có phải là thư rác hay không, vì vậy chúng ta sẽ quan tâm hơn đến việc tính toán tần suất tương đối (tỷ lệ) cho mỗi định dạng email.

¹⁸ (a) 0,451 đại diện cho tỷ lệ người nộp đơn cá nhân có thể chấp. (b) 0,802 đại diện cho tỷ lệ người nộp đơn có thể chấp nộp đơn với tư cách cá nhân.

¹⁹ (a) 0,122 đại diện cho tỷ lệ người nộp đơn vay theo hình thức đồng sở hữu có sở hữu nhà trong tổng số người vay. (b) 0,135 đại diện cho tỷ lệ người vay sở hữu nhà nộp đơn vay theo hình thức đồng sở hữu trong tổng số người vay sở hữu nhà

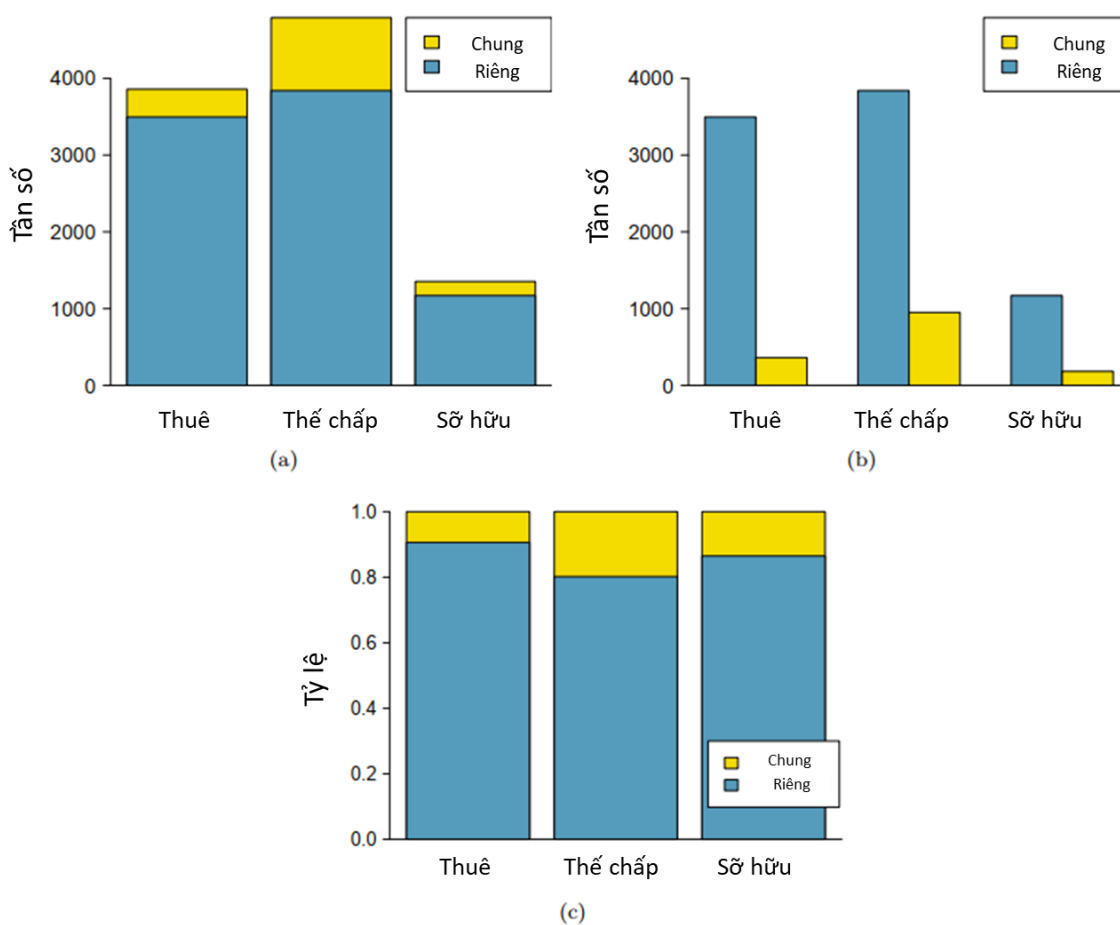
2.2.3 Sử dụng biểu đồ cột với hai biến

Bảng tần suất chéo sử dụng tỷ lệ hàng hoặc cột đặc biệt hữu ích để kiểm tra mối quan hệ giữa hai biến phân loại. Biểu đồ cột chồng cung cấp một cách để trực quan hóa thông tin trong các bảng này.

Biểu đồ cột chồng là một hình thức hiển thị đồ họa thông tin bảng tần suất chéo. Ví dụ, biểu đồ cột chồng đại diện cho Hình 2.21 được hiển thị trong Hình 2.23(a), trong đó chúng ta đã tạo biểu đồ cột bằng biến tình trạng sở hữu nhà và sau đó chia mỗi nhóm theo các cấp độ của loại hình áp dụng (**app_type**).

Một hình thức trực quan hóa liên quan đến biểu đồ cột xếp chồng là **biểu đồ cột kép**, được minh họa trong Hình 2.23(b).

Đối với loại biểu đồ cột cuối cùng mà chúng ta giới thiệu, tỷ lệ cột cho bảng tần suất chéo giữa biến **app_type** và biến **homeownership** đã được chuyển thành biểu đồ cột chồng chuẩn hóa trong Hình 2.23(c). Loại hình trực quan hóa này hữu ích trong việc hiểu tỷ lệ đơn xin vay cá nhân hoặc đồng sở hữu cho người vay ở mỗi cấp độ biến **homeownership**. Ngoài ra, vì tỷ lệ người vay đồng sở hữu và cá nhân thay đổi giữa các nhóm, chúng ta có thể kết luận rằng hai biến có mối liên hệ.



Hình 2.23: (a) Biểu đồ cột chồng cho biến tình trạng sở hữu nhà (**homeownership**), trong đó số lượng đã được phân chia thêm theo biến loại hình áp dụng (**app_type**). (b) Biểu đồ cột kép cho biến tình trạng sở hữu nhà (**homeownership**) và biến loại hình áp dụng (**app_type**). (c) Phiên bản chuẩn hóa của biểu đồ cột chồng.

VÍ DỤ 2.27

Hãy xem xét ba biểu đồ cột trong Hình 2.23. Khi nào thì biểu đồ cột chồng, biểu đồ cột kép hoặc biểu đồ cột chồng chuẩn hóa hữu ích nhất?

Biểu đồ cột chồng hữu ích nhất khi có thể gán một biến làm biến giải thích và biến còn lại làm biến phản hồi, vì chúng ta đang nhóm theo một biến trước và sau đó phân tách theo các biến còn lại.

Biểu đồ cột kép thể hiện dữ liệu một cách trung lập hơn, không mặc định biến nào là biến giải thích hay biến phản hồi. Hình thức này cũng giúp dễ dàng nhận biết số lượng quan sát trong sáu tổ hợp nhóm khác nhau. Tuy nhiên, một nhược điểm là nó thường yêu cầu nhiều không gian theo chiều ngang hơn, độ hẹp của Hình 2.23(b) làm cho biểu đồ cảm thấy hơi chật chội. Ngoài ra, khi hai nhóm có kích thước rất khác nhau, như chúng ta thấy trong nhóm sở hữu so với một trong hai nhóm còn lại, rất khó để nhận biết liệu có tồn tại mối liên hệ giữa các biến hay không.

Biểu đồ cột chồng chuẩn hóa hữu ích nếu biến chính trong biểu đồ cột chồng tương đối mất cân bằng, ví dụ: nhóm sở hữu chỉ có một phần ba số quan sát trong nhóm thể chấp, làm cho biểu đồ cột xếp chồng đơn giản kém hiệu quả hơn trong việc kiểm tra mối quan hệ giữa các biến. Nhược điểm chính của phiên bản chuẩn hóa là chúng ta mất đi cảm giác về số lượng trường quan sát mà mỗi cột biểu diễn.

2.2.4 Biểu đồ khám

Biểu đồ khám là một kỹ thuật trực quan hóa phù hợp cho bảng tần suất chéo, tương tự như biểu đồ cột chồng chuẩn hóa, nhưng có ưu điểm là chúng ta vẫn thấy được quy mô tương đối của các nhóm trong biến chính.

Để bắt đầu tạo biểu đồ khám đầu tiên của chúng ta, chúng ta sẽ chia một hình vuông thành các cột tương ứng với từng nhóm của biến **homeownership**, như minh họa trong Hình 2.24(a). Mỗi cột biểu diễn một mức của biến **homeownership** và chiều rộng của cột tỷ lệ với tỷ lệ khoản vay trong mỗi nhóm. Ví dụ, số khoản vay mà người vay là chủ sở hữu nhà ít hơn so với số khoản vay mà người vay có thể chấp. Nhìn chung, biểu đồ khám sử dụng diện tích của các ô để biểu diễn số trường hợp trong từng nhóm.

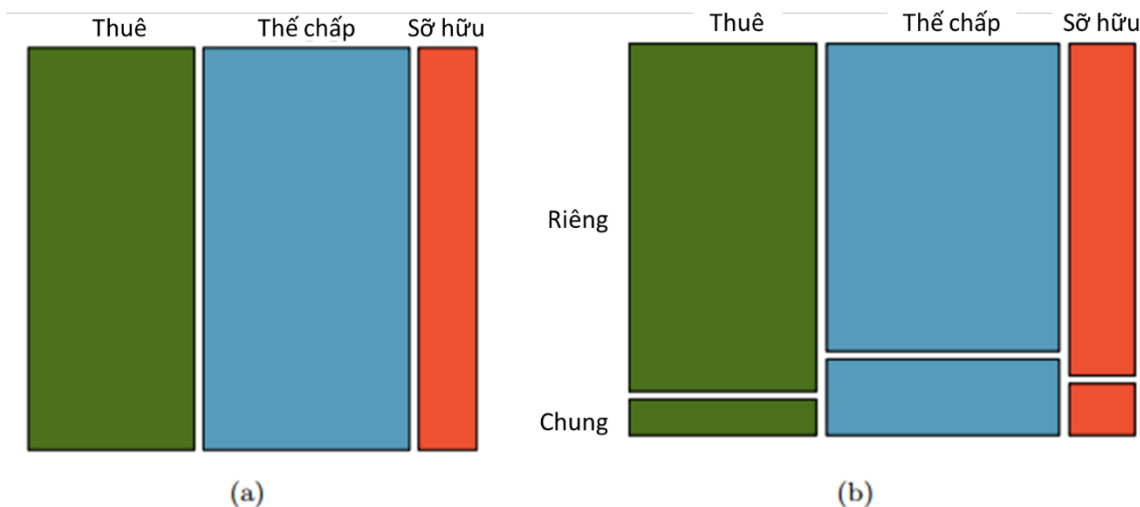
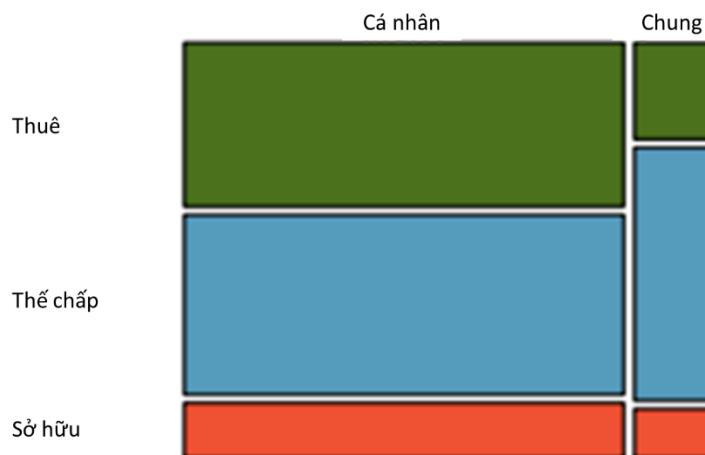


Figure 2.24: (a) Biểu đồ khám một biến cho biến **homeownership**. (b) Biểu đồ khám hai biến cho cả tình trạng **homeownership** và **app_type**.

Để tạo một biểu đồ khám hoàn chỉnh, biểu đồ khám đơn biến được tiếp tục chia nhỏ trong Hình 2.24(b) bằng cách sử dụng biến **app_type**. Mỗi cột được chia theo tỷ lệ số khoản vay từ người vay cá nhân và đồng vay. Ví dụ, cột thứ hai biểu diễn các khoản vay mà người vay có thể chấp và nó được chia thành khoản vay cá nhân (phần trên) và khoản vay đồng vay (phần dưới). Tương tự, phần dưới của cột thứ ba biểu diễn các khoản vay mà người vay sở hữu nhà và đăng ký đồng vay, trong khi phần trên biểu diễn các khoản vay mà người sở hữu nhà đăng ký vay cá nhân. Chúng ta có thể sử dụng biểu đồ này để nhận thấy rằng biến **homeownership** và biến **app_type** có mối liên hệ, vì một số cột được chia theo các tỷ lệ khác nhau.

Vị trí theo chiều dọc của các đoạn trong cột khác nhau, đây cũng chính là kỹ thuật được sử dụng để kiểm tra mối liên hệ trong biểu đồ cột chồng chuẩn hóa.

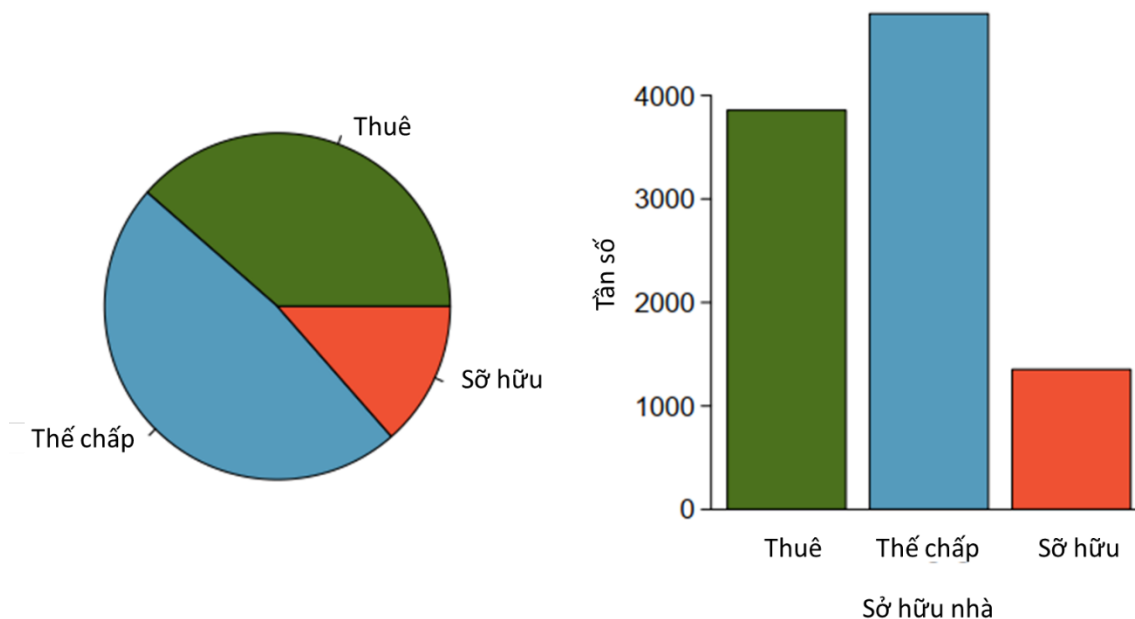
Trong Hình 2.24, chúng ta chọn cách chia đầu tiên theo tình trạng sở hữu nhà của người vay. Tuy nhiên, ta cũng có thể chia trước theo loại đơn đăng ký, như minh họa trong Hình 2.25. Tương tự như biểu đồ cột, thông thường trong biểu đồ mosaic, biến giải thích được sử dụng để thực hiện phép chia đầu tiên, sau đó biến phản hồi sẽ chia nhỏ từng mức của biến giải thích, nếu việc gán nhãn này phù hợp với các biến đang được xem xét.



Hình 2.25: Biểu đồ khám trong đó các khoản vay được nhóm theo biến **homeownership** khi đã được chia thành các loại đơn đăng ký cá nhân (**individual**) và đồng vay (**joint**).

2.2.5 Biểu đồ tròn duy nhất bạn sẽ thấy trong cuốn sách này

Biểu đồ tròn được hiển thị trong Hình 2.26 cùng với biểu đồ cột đại diện cho cùng một thông tin. Biểu đồ tròn có thể hữu ích để cung cấp cái nhìn tổng quan cấp cao nhằm cho thấy một tập hợp các trường hợp được phân chia như thế nào. Tuy nhiên, cũng khó giải mã các chi tiết trong biểu đồ tròn. Ví dụ, mất vài giây lâu hơn để nhận ra rằng có nhiều khoản vay hơn khi người vay có thể chấp so với thuê khi nhìn vào biểu đồ tròn, trong khi chi tiết này rất rõ ràng trong biểu đồ cột. Mặc dù biểu đồ tròn có thể hữu ích, chúng tôi thích biểu đồ cột hơn vì chúng dễ dàng so sánh các nhóm.



Hình 2.26: Biểu đồ tròn và biểu đồ cột về biến **homeownership**.

2.2.6 So sánh dữ liệu định lượng giữa các nhóm

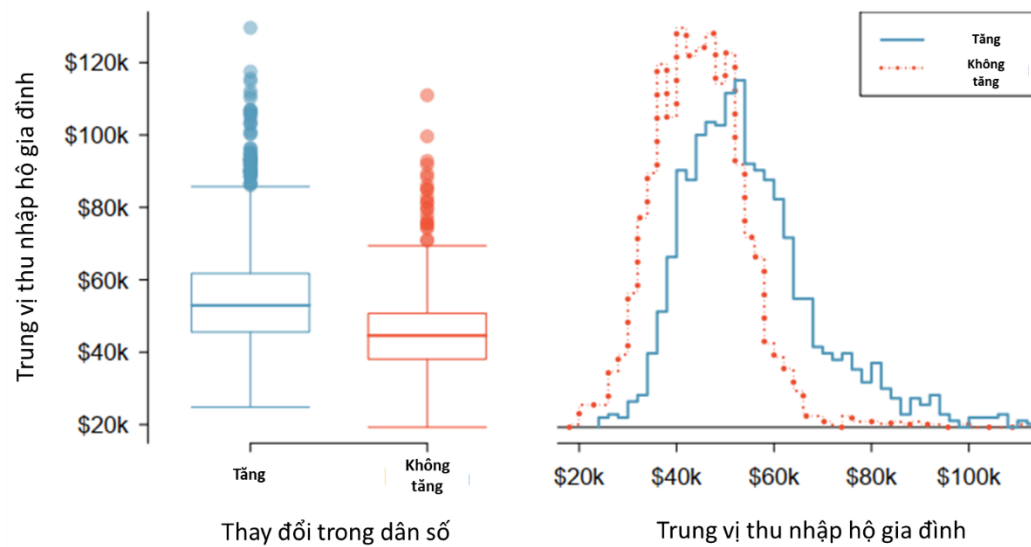
Một trong những phân tích thú vị nhất là so sánh dữ liệu định lượng giữa các nhóm khác nhau. Phương pháp thực hiện không có gì mới về bản chất, chúng ta chỉ cần vẽ đồ thị biểu diễn dữ liệu định lượng riêng cho từng nhóm trên cùng một biểu đồ. Hai phương pháp phổ biến được giới thiệu ở đây là: biểu đồ hộp cạnh nhau và biểu đồ tần suất rỗng.

Chúng ta sẽ tiếp tục sử dụng bộ dữ liệu về các quận (county) và so sánh thu nhập hộ gia đình trung vị giữa các quận có dân số tăng từ năm 2010 đến 2017 với các quận không có sự gia tăng dân số. Mặc dù có thể mong muốn thiết lập một mối quan hệ nhân quả ở đây, nhưng cần nhớ rằng đây là dữ liệu quan sát, do đó bất kỳ suy luận nhân quả nào cũng sẽ chỉ mang tính suy đoán và chưa được kiểm chứng.

Có 1454 quận có dân số tăng từ năm 2010 đến 2017 và 1672 quận không có sự gia tăng (trong đó tất cả trừ một quận đều có dân số giảm). Hình 2.27 hiển thị một mẫu ngẫu nhiên gồm 100 quận từ nhóm đầu tiên và 50 quận từ nhóm thứ hai để giúp minh họa rõ hơn về dữ liệu thô của thu nhập trung vị.

Thu nhập trung bình của 150 quận, đơn vị \$1000s										
Tăng dân số						Không tăng dân số				
38,2	43,6	42,2	61,5	51,1	45,7	48,3	60,3	50,7		
44,6	51,8	40,7	48,1	56,4	41,9	39,3	40,4	40,3		
40,6	63,3	52,1	60,3	49,8	51,7	57	47,2	45,9		
51,1	34,1	45,5	52,8	49,1	51	42,3	41,5	46,1		
80,8	46,3	82,2	43,6	39,7	49,4	44,9	51,7	46,4		
75,2	40,6	46,3	62,4	44,1	51,3	29,1	51,8	50,5		
51,9	34,7	54	42,9	52,2	45,1	27	30,9	34,9		
61	51,4	56,5	62	46	46,4	40,7	51,8	61,1		
53,8	57,6	69,2	48,4	40,5	48,6	43,4	34,7	45,7		
53,1	54,6	55	46,4	39,9	56,7	33,1	21	37		
63	49,1	57,2	44,1	50	38,9	52	31,9	45,7		
46,6	46,5	38,9	50,9	56	34,6	56,3	38,7	45,7		
74,2	63	49,6	53,7	77,5	60	56,2	43	21,7		
63,2	47,6	55,9	39,1	57,8	42,6	44,5	34,5	48,9		
50,4	49	45,6	39	38,8	37,1	50,9	42,1	43,2		
57,2	44,7	71,7	35,3	100,2		35,4	41,3	33,6		
42,6	55,5	38,6	52,7	63		43,4	56,5			

Hình 2.27: Bảng này hiển thị thu nhập hộ gia đình trung vị (đơn vị: nghìn USD) từ một mẫu ngẫu nhiên gồm 100 quận có tăng dân số bên trái và 50 quận không có tăng dân số bên phải.



Hình 2.28: Biểu đồ hộp cạnh nhau (bên trái) và biểu đồ tần suất rỗng (bên phải) của biến `med_hh_income`, trong đó các quận được phân nhóm theo tình trạng tăng hoặc không tăng dân số.

Đồ thị **hộp hai bên (side-by-side box plot)** là một công cụ truyền thống để so sánh giữa các nhóm. Một ví dụ được hiển thị trong bảng bên trái của Hình 2.28, nơi có hai đồ thị hộp, mỗi đồ thị cho một nhóm, được đặt vào cùng một cửa sổ vẽ và được vẽ trên cùng một tỷ lệ.

Một phương pháp vẽ hữu ích khác sử dụng **biểu đồ histogram rỗng (hollow histograms)** để so sánh dữ liệu số giữa các nhóm. Đây chỉ là các đường viền của histogram của mỗi nhóm được đặt trên cùng một biểu đồ, như được thể hiện trong bảng bên phải của Hình 2.28.

BÀI TẬP THỰC HÀNH 2.28

G

Sử dụng các biểu đồ trong Hình 2.28 để so sánh thu nhập của các quận giữa hai nhóm. Bạn nhận thấy điều gì về trung tâm xấp xỉ của mỗi nhóm? Bạn nhận thấy điều gì về độ biến động giữa các nhóm? Hình dạng có tương đối nhất quán giữa các nhóm không? Có bao nhiêu *đỉnh nổi bật* cho mỗi nhóm? ²⁰

BÀI TẬP THỰC HÀNH 2.29

G

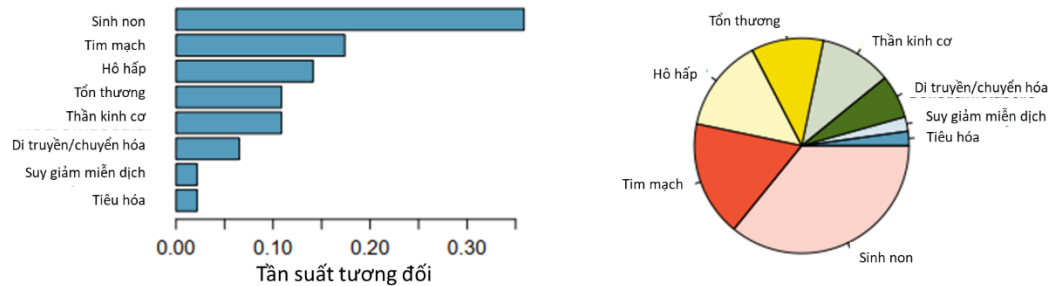
Những thành phần nào của mỗi biểu đồ trong Hình 2.28 mà bạn thấy hữu ích nhất? ²¹

²⁰ Các câu trả lời có thể khác nhau một chút. Các quận có mức tăng dân số có xu hướng có thu nhập cao hơn (trung bình khoảng \$45,000) so với các quận không có mức tăng (trung bình khoảng \$40,000). Độ biến động cũng lớn hơn một chút ở nhóm có mức tăng dân số. Điều này thể hiện rõ trong IQR, lớn hơn khoảng 50% trong *nhóm có mức tăng*. Cả hai phân bố đều cho thấy sự lệch bên phải nhẹ đến vừa và đều là đơn cực. Các biểu đồ hộp chỉ ra rằng có nhiều quan sát nằm xa so với trung bình trong mỗi nhóm, mặc dù chúng ta nên dự đoán rằng nhiều quan sát sẽ nằm ngoài các sợi dây whiskers khi xem xét bất kỳ tập dữ liệu nào chứa hơn vài trăm điểm dữ liệu.

²¹ Các câu trả lời sẽ khác nhau. Các đồ thị hộp hai bên đặc biệt hữu ích để so sánh trung tâm và độ phân tán, trong khi các biểu đồ histogram rỗng thì hữu ích hơn để thấy hình dạng phân bố, độ lệch và các bất thường tiềm tàng.

Bài tập

1.1 Sử dụng kháng sinh ở trẻ em. Biểu đồ cột và biểu đồ hình tròn bên dưới cho thấy sự phân bố các tình trạng y tế có sẵn trước của trẻ em tham gia vào một nghiên cứu về thời gian tối ưu sử dụng kháng sinh trong điều trị viêm khí quản, một loại nhiễm trùng đường hô hấp trên.



- Những đặc điểm nào rõ ràng trong biểu đồ cột nhưng không có trong biểu đồ hình tròn?
- Những đặc điểm nào rõ ràng trong biểu đồ hình tròn nhưng không có trong biểu đồ cột?
- Bạn thích sử dụng biểu đồ nào để hiển thị dữ liệu phân loại này?

1.2 Quan điểm về di cư. 910 cử tri đã đăng ký được chọn ngẫu nhiên từ Tampa, FL được hỏi liệu họ có nghĩ rằng những công nhân đã nhập cảnh trái phép vào Mỹ nên (i) được phép giữ công việc của mình và nộp đơn xin nhập tịch Mỹ, (ii) được phép giữ công việc của mình như những công nhân khách tạm thời nhưng không được phép nộp đơn xin nhập tịch Mỹ, hoặc (iii) mất việc và phải rời khỏi đất nước. Kết quả của cuộc khảo sát theo khuynh hướng chính trị được trình bày bên dưới.²²

		Khuynh hướng chính trị			
		Bảo thủ	Ôn hòa	Tự do	Tổng
Phản hồi	(i) Nộp đơn xin nhập tịch	57	120	101	278
	(ii) Công nhân khách	121	113	28	262
	(iii) Rời khỏi đất nước	179	126	45	350
	(iv) Không chắc	15	4	1	20
	Tổng	372	363	175	910

- Bao nhiêu phần trăm các cử tri ở Tampa, FL tự xác định mình là những người bảo thủ?
- Bao nhiêu phần trăm các cử tri ở Tampa, FL ủng hộ lựa chọn nhập tịch?
- Bao nhiêu phần trăm các cử tri ở Tampa, FL tự xác định mình là những người bảo thủ và ủng hộ lựa chọn nhập tịch?
- Bao nhiêu phần trăm các cử tri ở Tampa, FL tự xác định mình là những người bảo thủ cũng ủng hộ lựa chọn nhập tịch? Bao nhiêu phần trăm những người trung dung chia sẻ quan điểm này? Bao nhiêu phần trăm những người tự do chia sẻ quan điểm này?
- Liệu khuynh hướng chính trị và quan điểm về di cư có vẻ như độc lập với nhau không? Giải thích lý do của bạn.

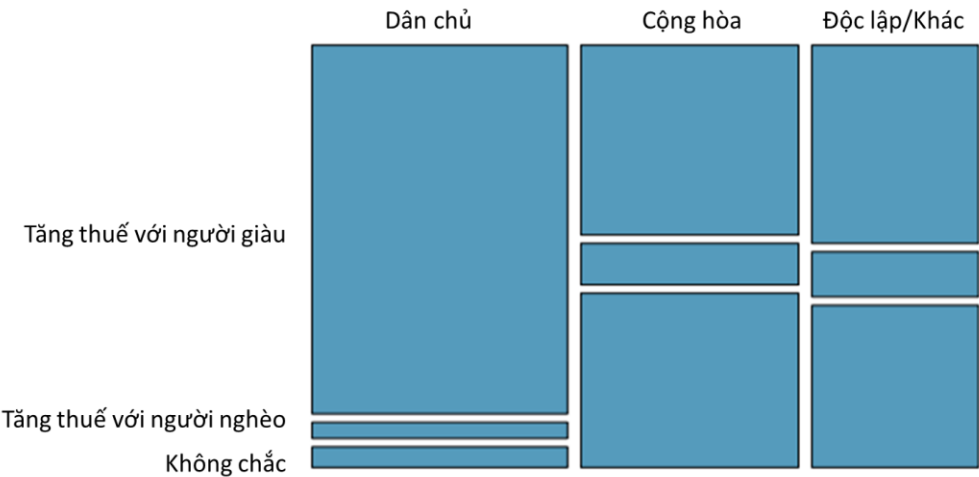
²² SurveyUSA, News Poll #18927, data collected Jan 27-29, 2012.

2.2. XEM XÉT DỮ LIỆU ĐỊNH TÍNH

1.3 Quan điểm về Đạo luật DREAM. Một mẫu ngẫu nhiên các cử tri đã đăng ký từ Tampa, FL được hỏi liệu họ có ủng hộ Đạo luật DREAM, một luật được đề xuất nhằm cung cấp một con đường nhập tịch cho những người bị đưa vào Mỹ trái phép khi còn nhỏ. Cuộc khảo sát cũng thu thập thông tin về khuynh hướng chính trị của những người tham gia. Dựa trên biểu đồ mosaic được hiển thị bên dưới, liệu quan điểm về Đạo luật DREAM và khuynh hướng chính trị có vẻ như độc lập với nhau không? Giải thích lý do của bạn. ²³



1.4 Tăng thuế. Một mẫu ngẫu nhiên các cử tri đã đăng ký trên toàn quốc được hỏi liệu họ nghĩ rằng tốt hơn là tăng thuế đối với người giàu hay tăng thuế đối với người nghèo. Cuộc khảo sát cũng thu thập thông tin về đảng phái chính trị của những người tham gia. Dựa trên biểu đồ mosaic được hiển thị bên dưới, liệu quan điểm về việc tăng thuế và đảng phái chính trị có vẻ như độc lập với nhau không? Giải thích lý do của bạn. ²⁴



²³ SurveyUSA, News Poll #18927, data collected Jan 27-29, 2012.

²⁴ Public Policy Polling, Americans on College Degrees, Classic Literature, the Seasons, and More, data collected Feb 20-22, 2015.

2.3. Case study: vaccine sốt rét

VÍ DỤ 2.30

Giả sử giáo sư của bạn chia sinh viên trong lớp thành hai nhóm: sinh viên ngồi bên trái và sinh viên ngồi bên phải. Nếu $\hat{p}_L$ và $\hat{p}_R$ lần lượt đại diện cho tỷ lệ sinh viên sở hữu sản phẩm Apple ở bên trái và bên phải, bạn có ngạc nhiên không nếu $\hat{p}_L$ không bằng chính xác $\hat{p}_R$?

Mặc dù các tỷ lệ này có thể gần nhau, nhưng rất hiếm khi chúng chính xác như nhau. Chúng ta có thể quan sát thấy một sự khác biệt nhỏ do yếu tố ngẫu nhiên.

BÀI TẬP THỰC HÀNH 2.31

Nếu chúng ta không cho rằng vị trí ngồi trong lớp có liên quan đến việc một người có sở hữu sản phẩm Apple hay không, giả định nào đang được đưa ra về mối quan hệ giữa hai biến này? ²⁵

2.3.1 Biến thiên trong dữ liệu

Chúng ta xem xét một nghiên cứu về một loại vaccine sốt rét mới có tên là PfSPZ. Trong nghiên cứu này, các bệnh nhân tình nguyện được phân ngẫu nhiên vào một trong hai nhóm thử nghiệm: 14 bệnh nhân được dùng vaccine thử nghiệm và 6 bệnh nhân được dùng vaccine giả dược. Mười chín tuần sau, tất cả 20 bệnh nhân được tiếp xúc với một chủng ký sinh trùng sốt rét nhạy cảm với thuốc; mục đích sử dụng chủng ký sinh trùng nhạy cảm với thuốc ở đây là vì lý do đạo đức, cho phép điều trị hiệu quả mọi trường hợp nhiễm bệnh. Kết quả được tóm tắt trong Bảng 2.29, trong đó 9 trong số 14 bệnh nhân được điều trị không có dấu hiệu nhiễm trùng, trong khi cả 6 bệnh nhân trong nhóm đối chứng đều có một số dấu hiệu nhiễm trùng cơ bản.

		outcome		
		Nhiễm bệnh	Không nhiễm bệnh	Tổng cộng
treatment	Nhóm vaccine	5	9	14
	Nhóm giả dược	6	0	6
	Tổng cộng	11	9	20

Bảng 2.29: Tóm tắt kết quả của thí nghiệm vaccine sốt rét.

BÀI TẬP THỰC HÀNH 2.32

Đây là một nghiên cứu quan sát hay một thí nghiệm? Loại nghiên cứu này có ý nghĩa gì đối với những gì có thể suy ra từ kết quả? ²⁶

Trong nghiên cứu này, bệnh nhân được dùng vaccine có tỷ lệ có dấu hiệu nhiễm trùng nhỏ hơn (35,7% so với 100%). Tuy nhiên, kích thước mẫu rất nhỏ, và chúng ta chưa đánh giá được liệu sự khác biệt này có cung cấp bằng chứng thuyết phục rằng vaccine có hiệu quả hay không.

²⁵ Chúng ta đang giả định rằng hai biến này độc lập với nhau.

²⁶ Đây là một thí nghiệm, vì bệnh nhân được phân ngẫu nhiên vào nhóm thử nghiệm. Vì đây là một thí nghiệm, kết quả có thể được sử dụng để đánh giá mối quan hệ nhân quả giữa vaccine sốt rét và việc bệnh nhân có biểu hiện dấu hiệu nhiễm trùng hay không.

VÍ DỤ 2.30

Các nhà khoa học dữ liệu đôi khi được yêu cầu đánh giá độ mạnh của bằng chứng. Khi xem xét tỷ lệ nhiễm bệnh ở bệnh nhân trong hai nhóm của nghiên cứu này, chúng ta cần làm gì để đánh giá liệu dữ liệu có cung cấp bằng chứng thuyết phục về sự khác biệt thực sự hay không?

- E** Tỷ lệ nhiễm bệnh quan sát được (35,7% ở nhóm điều trị so với 100% ở nhóm đối chứng) gợi ý rằng vaccine có thể có hiệu quả. Tuy nhiên, chúng ta không thể chắc chắn liệu sự khác biệt quan sát được này có thực sự phản ánh hiệu quả của vaccine hay chỉ đơn thuần là kết quả của sự ngẫu nhiên. Thông thường, luôn có một số biến thiên trong dữ liệu mẫu, và chúng ta sẽ không mong đợi tỷ lệ giữa 2 nhóm trong mẫu phải chính xác bằng nhau, ngay cả khi thực tế là tỷ lệ nhiễm bệnh không bị ảnh hưởng bởi việc tiêm vaccine. Hơn nữa, với cỡ mẫu nhỏ như vậy, có thể là do may rủi mà chúng ta quan sát thấy sự khác biệt lớn như vậy khi chia nhóm ngẫu nhiên

Ví dụ 2.33 nhắc nhở chúng ta rằng kết quả quan sát được trong mẫu dữ liệu có thể không phản ánh hoàn toàn mối quan hệ thực sự giữa các biến do **yếu tố nhiễu ngẫu nhiên (random noise)**. Mặc dù chúng ta quan sát thấy sự khác biệt lớn về tỷ lệ nhiễm bệnh, nhưng cỡ mẫu của nghiên cứu này nhỏ, khiến cho việc xác định liệu sự khác biệt này có phải là do hiệu quả của vaccine hay chỉ do may rủi trở nên không rõ ràng. Chúng ta đặt tên cho hai giả thuyết cạnh tranh này là H_0 và H_A được đọc là “H-nought” và “H-A”:

H_0 : **Giả thuyết độc lập.** Hai biến (điều trị và kết quả) là độc lập, không có mối quan hệ nào giữa chúng. Sự khác biệt quan sát được giữa tỷ lệ nhiễm bệnh trong hai nhóm (64,3%) chỉ là kết quả của sự ngẫu nhiên.

H_A : **Giả thuyết thay thế.** Hai biến này không độc lập. Sự khác biệt 64,3% trong tỷ lệ nhiễm bệnh không phải do may rủi mà là do tác động của vaccine

Nếu giả thuyết độc lập là đúng, nghĩa là vaccine không ảnh hưởng đến tỷ lệ nhiễm bệnh, thì điều đó có nghĩa là *11 bệnh nhân sẽ bị nhiễm bệnh dù họ thuộc nhóm nào, và 9 bệnh nhân sẽ không nhiễm bệnh bất kể họ được chỉ định vào nhóm nào*. Nói cách khác, nếu vaccine không có tác dụng, thì sự khác biệt quan sát được trong tỷ lệ nhiễm bệnh chỉ là kết quả của quá trình phân nhóm ngẫu nhiên.

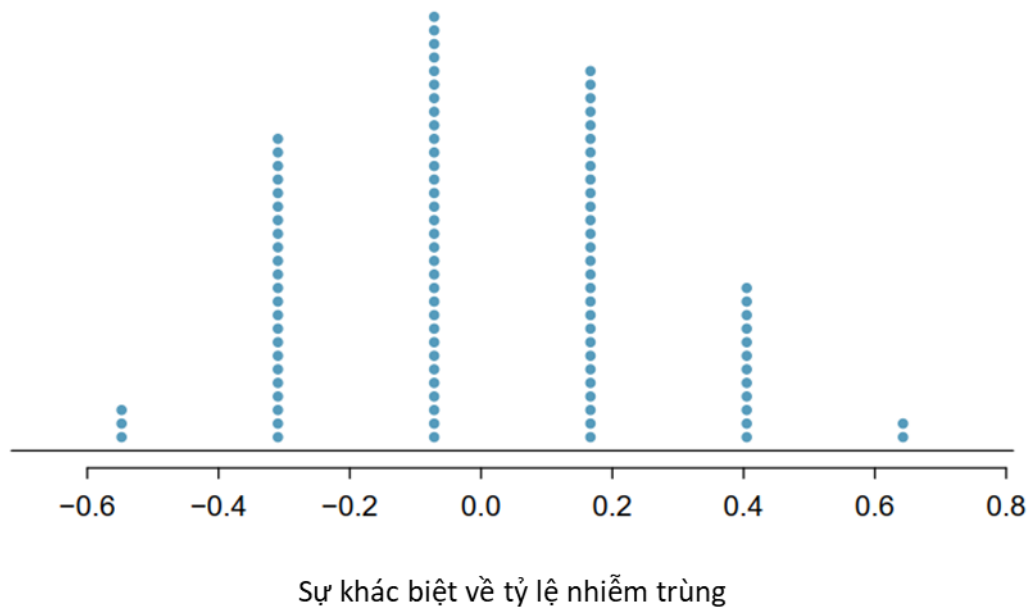
Bây giờ, hãy xem xét giả thuyết thay thế: tỷ lệ nhiễm bệnh bị ảnh hưởng bởi việc một bệnh nhân có được tiêm vaccine hay không. Nếu điều này đúng, và đặc biệt nếu ảnh hưởng của vaccine là đáng kể, thì chúng ta mong đợi sẽ thấy một sự khác biệt trong tỷ lệ nhiễm bệnh giữa các nhóm.

Chúng ta lựa chọn giữa hai giả thuyết cạnh tranh này bằng cách đánh giá xem dữ liệu mâu thuẫn đến mức nào với H_0 . Nếu dữ liệu mâu thuẫn quá nhiều với giả thuyết độc lập, chúng ta sẽ bác bỏ H_0 và kết luận rằng vaccine có hiệu quả.

2.3.2 Mô phỏng nghiên cứu

Chúng ta sẽ thực hiện các **mô phỏng**, trong đó chúng ta giả định rằng vaccine sốt rét đang được thử nghiệm *không* có tác dụng. Mục tiêu cuối cùng của chúng ta là hiểu liệu sự khác biệt lớn mà chúng ta quan sát được có phải là điều phổ biến trong các mô phỏng này hay không. Nếu sự khác biệt đó xuất hiện thường xuyên, có lẽ sự khác biệt quan sát được chỉ là kết quả của ngẫu nhiên. Nếu sự khác biệt này hiếm khi xảy ra, thì khả năng vaccine có tác dụng trở nên đáng tin cậy hơn.

Bảng 2.29 cho thấy có 11 bệnh nhân bị nhiễm bệnh và 9 bệnh nhân không bị nhiễm bệnh. Trong mô phỏng của chúng ta, chúng ta giả định rằng việc bị nhiễm bệnh là độc lập với việc tiêm vaccine, và chúng ta sẽ *quay ngược* thời gian về lúc các nhà nghiên cứu phân nhóm bệnh nhân. Nếu chúng ta tình cờ phân nhóm lại các bệnh nhân theo một cách khác, chúng ta có thể thu được một kết quả khác trong thế giới giả định này, nơi vaccine không ảnh hưởng đến nguy cơ nhiễm bệnh. Hãy thực hiện một lần **phân nhóm ngẫu nhiên** khác bằng cách mô phỏng.



Hình 2.31: Biểu đồ chấm xếp chồng thể hiện sự khác biệt từ 100 lần mô phỏng được thực hiện theo mô hình độc lập H_0 , trong đó giả định rằng vaccine không ảnh hưởng đến tình trạng nhiễm bệnh. Trong số 100 lần mô phỏng, chỉ có 2 lần xuất hiện sự khác biệt ít nhất 64,3%, đúng bằng mức quan sát được trong nghiên cứu.

Việc sự khác biệt 64,3% là một sự kiện hiếm gặp cho thấy hai cách lý giải kết quả của nghiên cứu:

H_0 **Giả thuyết độc lập.** Vaccine không có tác động đến tỷ lệ nhiễm bệnh, và sự khác biệt quan sát được chỉ là một hiện tượng ngẫu nhiên hiếm khi xảy ra.

H_A **Giả thuyết thay thế.** Vaccine có ảnh hưởng đến tỷ lệ nhiễm bệnh, và sự khác biệt quan sát được thực sự là do vaccine có hiệu quả trong việc phòng chống sốt rét, điều này giải thích tại sao mức chênh lệch đạt đến 64,3%.

Dựa trên các mô phỏng, chúng ta có hai lựa chọn: (1) Kết luận rằng kết quả nghiên cứu không cung cấp bằng chứng đủ mạnh để bác bỏ giả thuyết độc lập. Điều này có nghĩa là chúng ta chưa có đủ bằng chứng để khẳng định rằng vaccine có tác dụng trong bối cảnh lâm sàng này. (2) Kết luận rằng bằng chứng là đủ mạnh để bác bỏ H_0 và khẳng định vaccine có hiệu quả. Trong các nghiên cứu chính thống, chúng ta thường bác bỏ giả thuyết rằng sự khác biệt quan sát được chỉ là một sự kiện hiếm gặp. Vì vậy, trong trường hợp này, chúng ta bác bỏ giả thuyết độc lập và ủng hộ giả thuyết thay thế.²⁸ Nói cách khác, chúng ta kết luận rằng dữ liệu cung cấp bằng chứng đủ mạnh về tác dụng bảo vệ của vaccine với sốt rét trong bối cảnh lâm sàng này.

Một lĩnh vực trong thống kê học gọi là suy luận thống kê tập trung vào việc đánh giá xem một sự khác biệt nào đó có phải chỉ do ngẫu nhiên hay không. Trong suy luận thống kê, các nhà khoa học dữ liệu xác định mô hình nào là hợp lý nhất dựa trên dữ liệu quan sát được. Sai số vẫn có thể xảy ra, giống như các sự kiện hiếm gặp, và đôi khi chúng ta có thể chọn sai giả thuyết. Mặc dù không phải lúc nào cũng có thể đưa ra quyết định chính xác, suy luận thống kê cung cấp các công cụ để kiểm soát và đánh giá tần suất của các sai sót này. Trong Chương 5, chúng ta sẽ có một phần giới thiệu chính thức về vấn đề lựa chọn giả thuyết. Hai chương tiếp theo sẽ giúp xây dựng nền tảng về xác suất và lý thuyết cần thiết để giúp cho quá trình suy luận này chặt chẽ hơn.

²⁸ Lập luận này không thể áp dụng rộng rãi cho các quan sát mang tính giai thoại (anecdotal overstatement). Mỗi ngày, chúng ta đều chứng kiến những sự kiện cực kỳ hiếm gặp mà không thể đoán trước. Tuy nhiên, trong bối cảnh bằng chứng không chặt chẽ như các câu chuyện giai thoại, hầu như bất kỳ điều gì cũng có thể trông như một sự kiện hiếm gặp. Do đó, việc cố gắng tìm kiếm các sự kiện hiếm trong cuộc sống hàng ngày có thể dẫn đến những hiểu lầm nguy hiểm. Ví dụ, hãy xét đến xổ số: Xác suất để dãy số chiến thắng trong giải Powerball lớn nhất lịch sử (ngày 13/01/2016) là (04, 08, 19, 27, 34) với số Powerball là (10) chỉ là 1 trên 292 triệu. Nhưng dù vậy, những con số này vẫn được chọn! Tuy nhiên, nếu bất kỳ bộ số nào khác xuất hiện, xác suất của chúng cũng hiếm như vậy. Điều này cũng đúng trong cuộc sống hàng ngày: mỗi sự kiện có vẻ như cực kỳ hiếm, nhưng nếu xét đến tất cả các khả năng có thể xảy ra, thì bất kỳ kết quả nào cũng đều hiếm như nhau. Vì vậy, chúng ta nên cẩn trọng để không hiểu sai những bằng chứng mang tính giai thoại.

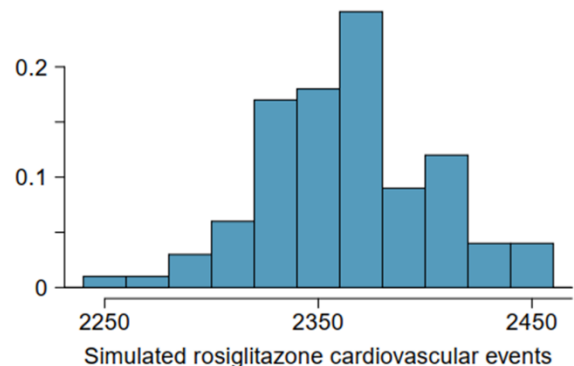
Bài tập

2.25 Tác dụng phụ của Avandia. Rosiglitazone là hoạt chất chính trong loại thuốc tiểu đường loại 2 gây tranh cãi mang tên Avandia và được cho là có liên quan đến với tăng nguy cơ mắc các vấn đề tim mạch nghiêm trọng như đột quỵ, suy tim và tử vong. Một lựa chọn điều trị thay thế phổ biến là pioglitazone, hoạt chất chính trong một loại thuốc tiểu đường có tên Actos. Trong một nghiên cứu quan sát hồi cứu trên toàn quốc đối với 227.571 bệnh nhân Medicare từ 65 tuổi trở lên, người ta nhận thấy rằng 2.593 trên 67.593 bệnh nhân sử dụng rosiglitazone và 5.386 trên 159.978 bệnh nhân sử dụng pioglitazone đã gặp phải các vấn đề tim mạch nghiêm trọng. Dữ liệu này được tóm tắt trong bảng phân tầng dưới đây.²⁹

Điều trị		Vấn đề tim mạch		
		Có	Không	Tổng cộng
	Rosiglitazone	2.593	65	67.593
	Pioglitazone	5.386	154.592	159.978
	Tổng cộng	7.979	219.592	227.571

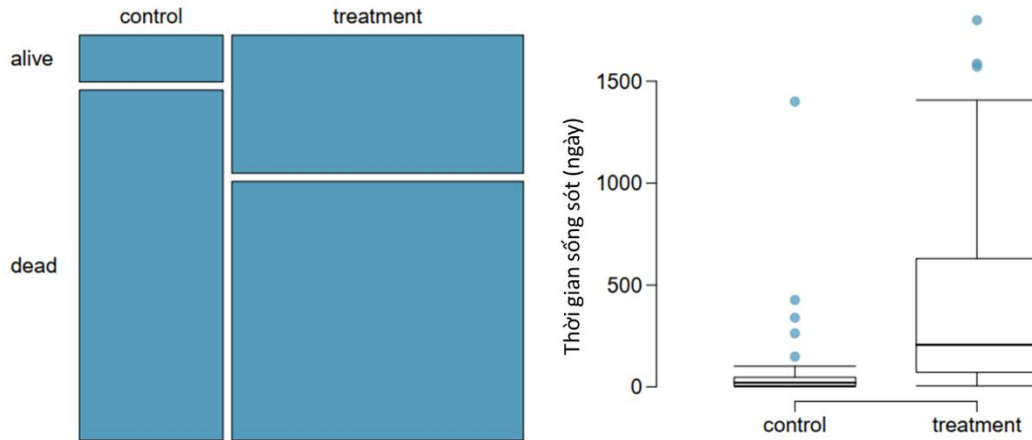
- (a) Hãy xác định xem mỗi câu phát biểu sau là đúng hay sai. Nếu sai, hãy giải thích lý do. *Cẩn thận:* Lập luận có thể sai ngay cả khi kết luận của phát biểu đúng. Trong những trường hợp như vậy, phát biểu được coi là sai.
- Vì số lượng bệnh nhân dùng pioglitazone gặp vấn đề tim mạch cao hơn (5.386 so với 2.593), chúng ta có thể kết luận rằng tỷ lệ bệnh nhân gặp vấn đề tim mạch khi dùng pioglitazone cao hơn.
 - Dữ liệu cho thấy bệnh nhân tiểu đường dùng rosiglitazone có khả năng gặp vấn đề tim mạch cao hơn, vì tỷ lệ mắc bệnh là $(2.593 / 67.593 = 0,038)$ 3,8% đối với nhóm dùng rosiglitazone, trong khi chỉ là $(5.386 / 159.978 = 0,034)$ 3,4% đối với nhóm dùng pioglitazone.
 - Việc tỷ lệ mắc bệnh cao hơn ở nhóm dùng rosiglitazone chứng minh rằng rosiglitazone gây ra các vấn đề tim mạch nghiêm trọng.
 - Dựa trên thông tin hiện có, chúng ta không thể xác định liệu sự khác biệt giữa hai tỷ lệ mắc bệnh là do mối quan hệ thực sự giữa hai biến số hay chỉ do ngẫu nhiên.
- (b) Tỷ lệ bệnh nhân gặp vấn đề tim mạch trên tổng số bệnh nhân là bao nhiêu?
- (c) Nếu loại thuốc điều trị và nguy cơ mắc bệnh tim mạch không liên quan (độc lập với nhau), thì chúng ta kỳ vọng bao nhiêu bệnh nhân trong nhóm rosiglitazone gặp phải vấn đề tim mạch?
- (d) Chúng ta có thể đánh giá mối quan hệ giữa kết quả và phương pháp điều trị trong nghiên cứu này bằng cách sử dụng kỹ thuật ngẫu nhiên hoá. Trên thực tế, chúng ta sẽ sử dụng phần mềm thống kê để thực hiện các mô phỏng cần thiết cho quá trình ngẫu nhiên hoá. Tuy nhiên, giả sử chúng ta thực hiện mô phỏng bằng cách sử dụng các thẻ ghi chỉ số. Để mô phỏng theo giả thuyết độc lập – trong đó giả định rằng kết quả sức khỏe không bị ảnh hưởng bởi loại thuốc điều trị – chúng ta sẽ ghi lại tình trạng tim mạch của từng bệnh nhân (có hoặc không gặp vấn đề) lên các thẻ. Sau đó, tất cả các thẻ này sẽ được xáo trộn ngẫu nhiên rồi chia vào hai nhóm có kích thước tương ứng với nghiên cứu: 67.593 thẻ cho nhóm dùng rosiglitazone và 159.978 thẻ cho nhóm dùng pioglitazone. Quá trình này được lặp lại 1.000 lần, và mỗi lần chúng ta ghi lại số lượng bệnh nhân trong nhóm rosiglitazone có vấn đề tim mạch. Dữ liệu thu được từ các lần mô phỏng này sẽ được sử dụng để tạo biểu đồ histogram. Dùng biểu đồ này để trả lời các câu hỏi (i)-(iii).

- Những giả thuyết nào đang được kiểm định?
- So với con số tính được ở phần (c), điều gì sẽ hỗ trợ mạnh mẽ hơn cho giả thuyết thay thế: Số bệnh nhân gặp vấn đề tim mạch trong nhóm rosiglitazone cao hơn hay thấp hơn?
- Kết quả mô phỏng gợi ý điều gì về mối quan hệ giữa việc sử dụng rosiglitazone và nguy cơ mắc bệnh tim mạch ở bệnh nhân tiểu đường?



²⁹ D.J. Graham et al. "Risk of acute myocardial infarction, stroke, heart failure, and death in elderly Medicare patients treated with rosiglitazone or pioglitazone". In: *JAMA* 304.4 (2010), p. 411. ISSN: 0098-7484.

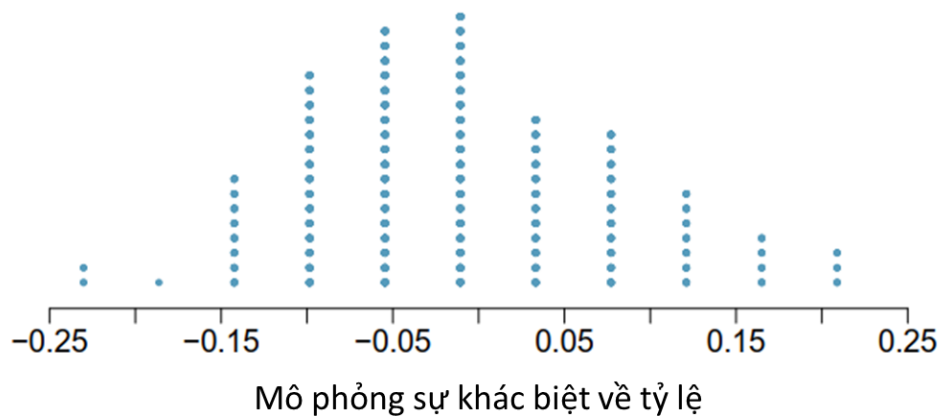
2.26 Ghép tim. Nghiên cứu Cấy ghép Tim của Đại học Stanford được thực hiện nhằm xác định liệu liệu pháp cấy ghép tim thử nghiệm có giúp kéo dài tuổi thọ hay không. Mỗi bệnh nhân tham gia chương trình đều được chỉ định chính thức cấy ghép tim, nghĩa là họ mắc bệnh nghiêm trọng và việc thay tim có thể cải thiện sức khỏe của họ. Một số bệnh nhân đã được cấy ghép, trong khi những người khác thì không. Biến "transplant" (ghép tim) cho biết bệnh nhân thuộc nhóm nào; những bệnh nhân thuộc nhóm điều trị (treatment) đã được cấy ghép tim, trong khi những bệnh nhân thuộc nhóm đối chứng (control) thì không. Trong số 34 bệnh nhân thuộc nhóm đối chứng, 30 người đã tử vong (dead). Trong số 69 bệnh nhân thuộc nhóm điều trị, 45 người đã tử vong. Một biến khác có tên "survived" được sử dụng để chỉ ra liệu bệnh nhân còn sống (alive) vào cuối nghiên cứu hay không.³⁰



- Dựa trên biểu đồ mosaic, liệu khả năng sống sót có độc lập với việc bệnh nhân có được cấy ghép tim hay không? Giải thích lập luận của bạn.
- Các biểu đồ hộp dưới đây gợi ý điều gì về hiệu quả của phương pháp điều trị bằng cấy ghép tim?
- Tỷ lệ bệnh nhân tử vong trong nhóm điều trị và nhóm đối chứng lần lượt là bao nhiêu?
- Một cách tiếp cận để kiểm tra xem phương pháp điều trị có hiệu quả hay không là sử dụng kỹ thuật ngẫu nhiên hoá.
 - Các giả thuyết đang được kiểm định là gì?
 - Đoạn văn dưới đây mô tả cách thiết lập phương pháp này nếu không sử dụng phần mềm thống kê. Điền vào chỗ trống bằng một con số hoặc cụm từ thích hợp.

Chúng ta viết *sống* lên_ thẻ đại diện cho những bệnh nhân còn sống vào cuối nghiên cứu và *tử vong* lên _____ thẻ đại diện cho những bệnh nhân đã tử vong. Sau đó, chúng ta xáo trộn các thẻ này và chia chúng thành hai nhóm: một nhóm có kích thước _____ đại diện cho nhóm điều trị, và một nhóm có kích thước _____ đại diện cho nhóm đối chứng. Chúng ta tính toán sự khác biệt giữa tỷ lệ thẻ *tử vong* trong nhóm điều trị và nhóm đối chứng (điều trị - đối chứng) và ghi lại giá trị này. Chúng ta lặp lại quá trình này 100 lần để xây dựng một phân phối tập trung tại _____. Cuối cùng, chúng ta tính toán tỷ lệ số lần mà sự khác biệt trong tỷ lệ tử vong được mô phỏng là _____ . Nếu tỷ lệ này thấp, chúng ta kết luận rằng khó có thể quan sát được kết quả như vậy chỉ do ngẫu nhiên, và giả thuyết không (H_0) nên bị bác bỏ để ủng hộ giả thuyết thay thế (H_A).

- Kết quả của các mô phỏng được hiển thị dưới đây gợi ý điều gì về hiệu quả của liệu pháp cấy ghép tim?



³⁰ B. Turnbull et al. "Survivorship of Heart Transplant Data". In: *Journal of the American Statistical Association* 69 (1974), pp. 74–80.

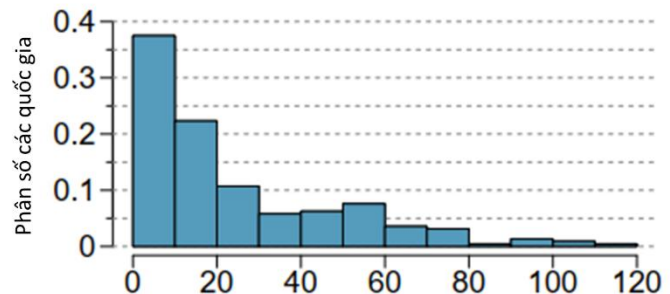
Bài tập chương

2.27 Bài kiểm tra bù. Trong một lớp học có 25 sinh viên, 24 người trong số đó đã làm bài kiểm tra trên lớp, và 1 sinh viên làm bài kiểm tra bù vào ngày hôm sau. Giáo sư đã chấm điểm đợt đầu tiên gồm 24 bài kiểm tra và thấy rằng điểm trung bình là 74 điểm với độ lệch chuẩn 8,9 điểm. Sinh viên làm bài kiểm tra bù vào ngày hôm sau đạt 64 điểm.

- Điểm của sinh viên mới này làm tăng hay giảm điểm trung bình?
- Điểm trung bình mới là bao nhiêu?
- Điểm của sinh viên mới này làm tăng hay giảm độ lệch chuẩn của tập điểm số?

2.28 Tỷ lệ tử vong ở trẻ sơ sinh. Tỷ lệ tử vong ở trẻ sơ sinh được định nghĩa là số ca tử vong của trẻ sơ sinh trên 1.000 ca sinh sống. Chỉ số này thường được sử dụng để đánh giá mức độ y tế của một quốc gia. Biểu đồ tần suất tương đối (relative frequency histogram) dưới đây cho thấy phân bố ước tính tỷ lệ tử vong ở trẻ sơ sinh của 224 quốc gia có dữ liệu vào năm 2014.³¹

- Ước lượng Q1, trung vị và Q3 từ biểu đồ tần suất.
- Bạn nghĩ giá trị trung bình của tập dữ liệu này sẽ nhỏ hơn hay lớn hơn trung vị? Giải thích lập luận của bạn.



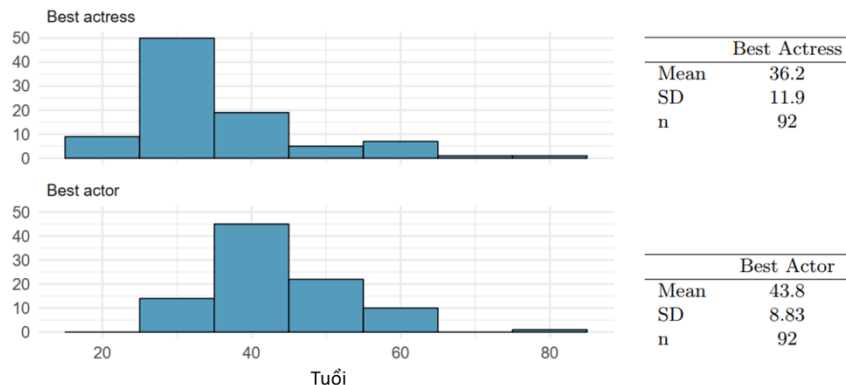
Tỷ lệ tử vong ở trẻ sơ sinh (trên 1000 ca sinh sống)

2.29 Người xem TV. Học sinh trong một lớp Thống kê AP được hỏi về số giờ xem TV mỗi tuần (bao gồm cả phát trực tuyến). Mẫu thu được có giá trị trung bình là 4,71 giờ, với độ lệch chuẩn là 4,18 giờ. Phân phối của số giờ xem TV hàng tuần có đối xứng không? Nếu không, bạn mong đợi phân phối này có hình dạng như thế nào? Giải thích lý do của bạn.

2.39 A new statistic. Thống kê $\frac{\bar{x}}{\text{median}}$ có thể được sử dụng như một thước đo độ lệch (skewness). Giả sử chúng ta có một phân phối trong đó tất cả các quan sát đều lớn hơn 0, tức là $x_i > 0$. Bạn mong đợi hình dạng của phân phối trong các điều kiện sau đây như thế nào? Giải thích lý do của bạn.

- $\frac{\bar{x}}{\text{median}} > 1$
- $\frac{\bar{x}}{\text{median}} < 1$
- $\frac{\bar{x}}{\text{median}} = 1$

2.31 Oscar winners. Giải thưởng Oscar đầu tiên cho Nam diễn viên chính xuất sắc nhất (Best actor) và Nữ diễn viên chính xuất sắc (Best actress) nhất được trao vào năm 1929. Các biểu đồ tần suất dưới đây thể hiện phân bố độ tuổi của tất cả những người chiến thắng ở hai hạng mục này từ năm 1929 đến 2018. Các thống kê tóm tắt của các phân phối này cũng được cung cấp (Trung bình – Mean, Độ lệch chuẩn – SD và tần số - n). Hãy so sánh phân bố độ tuổi của những người đoạt giải Nam diễn viên chính xuất sắc nhất và Nữ diễn viên chính xuất sắc nhất.³²



³² Oscar winners from 1929 – 2012, data up to 2009 from the Journal of Statistics Education data archive and more current data from wikipedia.org.

³¹ CIA Factbook, Country Comparisons, 2014.

2.3. CASE STUDY: VACCINE SỐT RẾT

78

2.32 Điểm thi. Điểm trung bình của một bài kiểm tra lịch sử (tính trên thang điểm 100) là 85, với độ lệch chuẩn là 15. Phân bố điểm của bài kiểm tra này có đối xứng không? Nếu không, bạn kỳ vọng phân bố này có hình dạng như thế nào? Giải thích lý do của bạn.

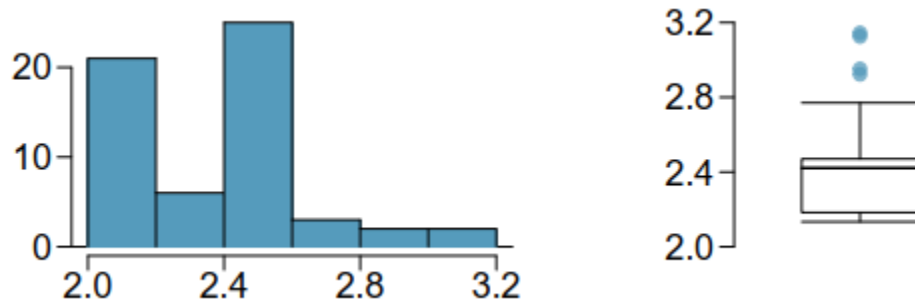
2.33 Stats scores. Dưới đây là điểm thi cuối kỳ của hai mươi sinh viên học môn thống kê nhập môn:

57, 66, 69, 71, 72, 73, 74, 77, 78, 78, 79, 79, 81, 81, 82, 83, 83, 88, 89, 94

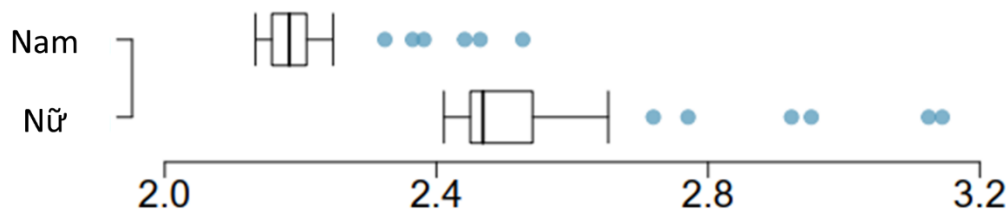
Hãy tạo một biểu đồ hộp (box plot) thể hiện phân bố điểm số này. Tóm tắt năm số liệu dưới đây có thể hữu ích.

Min	Q1	Q2 (trung vị)	Q3	Max
57	72,5	78,5	82,5	94

2.34 Người chiến thắng cuộc thi marathon. Biểu đồ tần suất (histogram) và biểu đồ hộp (box plot) dưới đây thể hiện phân phối thời gian hoàn thành cuộc thi marathon (tính theo giờ) của những người chiến thắng nam và nữ tại giải New York Marathon từ năm 1970 đến 1999.



- Những đặc điểm nào của phân phối có thể thấy rõ trong biểu đồ tần suất nhưng không thể thấy trong biểu đồ hộp? Những đặc điểm nào có thể thấy rõ trong biểu đồ hộp nhưng không thể thấy trong biểu đồ tần suất?
- Điều gì có thể là nguyên nhân dẫn đến phân phối hai đỉnh (bimodal distribution)? Giải thích.
- So sánh phân phối thời gian chạy marathon của nam và nữ dựa trên biểu đồ hộp được hiển thị bên dưới.



- Biểu đồ chuỗi thời gian (time series plot) dưới đây là một cách khác để quan sát dữ liệu này. Mô tả những gì có thể thấy trong biểu đồ này mà không thể thấy trong các biểu đồ khác.

