

Chương 3

Xác suất - Probability

3.1 Định nghĩa xác suất

3.2 Xác suất có điều kiện

3.3 Lấy mẫu từ một quần thể nhỏ

3.4 Biến ngẫu nhiên

3.5 Phân phối liên tục

Xác suất là nền tảng của thống kê, và có lẽ bạn đã quen thuộc với nhiều khái niệm được trình bày trong chương này. Tuy nhiên, việc chính thức hóa các khái niệm về xác suất có thể là điều mới mẻ đối với hầu hết người đọc.

Mặc dù chương này cung cấp nền tảng lý thuyết cho các ý tưởng trong những chương sau và giúp bạn hiểu sâu hơn, nhưng không bắt buộc phải nắm vững toàn bộ các khái niệm trong chương này để có thể áp dụng các phương pháp được giới thiệu trong phần còn lại của cuốn sách.



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

www.openintro.org/os

3.1 Định nghĩa xác suất

Thống kê được xây dựng dựa trên xác suất, và mặc dù không bắt buộc phải hiểu xác suất để áp dụng các kỹ thuật trong cuốn sách này, nhưng nó có thể giúp bạn hiểu sâu hơn về các phương pháp và tạo nền tảng vững chắc hơn cho các khóa học sau này.

3.1.1 Ví dụ

Trước khi đi vào các khái niệm kỹ thuật, hãy cùng xem qua một số ví dụ cơ bản có thể quen thuộc với bạn.

VÍ DỤ 3.1

Một "xúc xắc" là một khối lập phương có sáu mặt, được đánh số 1, 2, 3, 4, 5 và 6. Xác suất tung được mặt số 1 là bao nhiêu?

E

Nếu viên xúc xắc là công bằng, thì xác suất ra mặt 1 sẽ bằng với xác suất ra bất kỳ số nào khác. Vì có tổng cộng sáu kết quả có thể xảy ra, nên xác suất ra mặt 1 là 1 trên 6, hay tương đương $1/6$.

VÍ DỤ 3.2

Xác suất tung được mặt 1 hoặc 2 trong lần tung tiếp theo là bao nhiêu?

E

Mặt 1 và mặt 2 chiếm hai trong số sáu kết quả có thể xảy ra với xác suất bằng nhau, do đó, xác suất để tung được một trong hai mặt này là $2/6 = 1/3$.

VÍ DỤ 3.3

Xác suất tung được một trong các mặt 1, 2, 3, 4, 5 hoặc 6 trong lần tung tiếp theo là bao nhiêu?

E

100%. Kết quả chắc chắn sẽ là một trong những số này.

VÍ DỤ 3.4

Xác suất không tung được mặt 2 là bao nhiêu?

E

Vì xác suất tung được mặt 2 là $1/6$ hoặc 16,6%, nên xác suất không tung được mặt 2 sẽ là $100\% - 16,6\% = 83,3\%$ hoặc $5/6$.

Một cách khác để tính là nhận thấy rằng không tung được mặt 2 đồng nghĩa với việc tung được một trong các mặt 1, 3, 4, 5 hoặc 6, chiếm 5 trong 6 kết quả có thể xảy ra với xác suất bằng nhau, do đó xác suất là $5/6$.

VÍ DỤ 3.5

Xét trường hợp tung hai viên xúc xắc. Nếu $1/6$ lần tung, viên xúc xắc thứ nhất ra mặt 1, và trong $1/6$ số lần đó, viên xúc xắc thứ hai cũng ra mặt 1, thì xác suất để cả hai viên xúc xắc đều ra mặt 1 là bao nhiêu?

E

Vì $1/6$ số lần viên xúc xắc thứ nhất ra mặt 1, và trong $1/6$ số lần đó, viên xúc xắc thứ hai cũng ra mặt 1, nên xác suất để cả hai viên xúc xắc cùng ra mặt 1 là $(1/6) \times (1/6) = 1/36$.

3.1.2 Xác suất

Chúng ta sử dụng xác suất để xây dựng các công cụ mô tả và hiểu rõ sự ngẫu nhiên có vẻ như xuất hiện trong thực tế. Thông thường, xác suất được đặt trong bối cảnh một quá trình **ngẫu nhiên** tạo ra một **kết quả**.

Tung một viên xúc xắc $\rightarrow$ 1, 2, 3, 4, 5, or 6

Tung một đồng xu $\rightarrow$ Mặt ngửa hoặc Mặt sấp

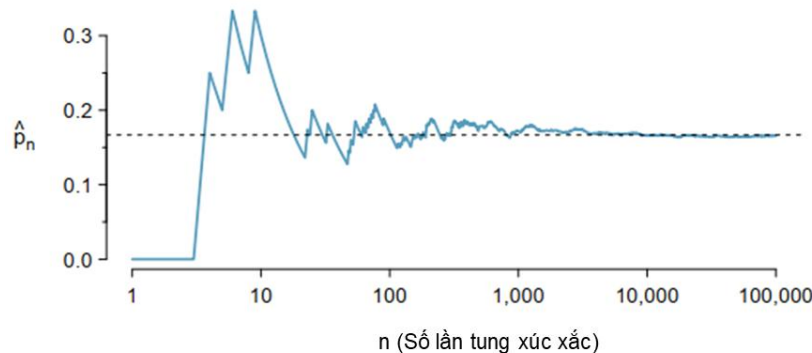
Tung một viên xúc xắc hoặc một đồng xu là một quá trình có vẻ ngẫu nhiên, và mỗi lần thực hiện sẽ tạo ra một kết quả.

XÁC XUẤT

Xác suất của một kết quả là tỷ lệ số lần kết quả đó xảy ra nếu chúng ta quan sát quá trình ngẫu nhiên một số lần vô hạn.

Xác suất được định nghĩa dưới dạng một tỷ lệ và luôn có giá trị trong khoảng từ 0 đến 1 (bao gồm cả hai giá trị này). Nó cũng có thể được biểu diễn dưới dạng phần trăm, trong khoảng từ 0% đến 100%.

Xác suất có thể được minh họa bằng cách tung một viên xúc xắc nhiều lần. Gọi $\hat{p}_n$ là tỷ lệ số lần xuất hiện mặt 1 sau n lần tung đầu tiên. Khi số lần tung tăng lên, $\hat{p}_n$ sẽ hội tụ về xác suất thực tế của việc tung ra mặt 1, tức là $p = 1/6$. Hình 3.1 minh họa sự hội tụ này với 100.000 lần tung xúc xắc. Xu hướng của $\hat{p}_n$ ổn định quanh giá trị p được mô tả bởi **Luật số lớn (Law of Large Numbers)**.



Hình 3.1: Tỷ lệ số lần tung xúc xắc ra mặt 1 tại mỗi giai đoạn trong một mô phỏng. Tỷ lệ này có xu hướng tiến gần hơn đến xác suất $1/6 \approx 0.167$ khi số lần tung tăng lên.

LUẬT SỐ LỚN

Khi thu thập nhiều quan sát hơn, tỷ lệ $\hat{p}_n$ của một kết quả cụ thể sẽ hội tụ về xác suất p của kết quả đó.

Đôi khi, tỷ lệ $\hat{p}_n$ sẽ lệch khỏi xác suất thực tế và dường như đi ngược lại Luật số lớn, như đã xảy ra nhiều lần trong Hình 3.1. Tuy nhiên, những sai lệch này sẽ dần nhỏ lại khi số lần tung xúc xắc tăng lên.

Ở trên, chúng ta viết p là xác suất tung được mặt 1. Ta cũng có thể biểu diễn xác suất này dưới dạng:

$$P(\text{tung được mặt 1})$$

Khi đã quen với ký hiệu này, chúng ta có thể rút gọn hơn nữa. Ví dụ, nếu ngữ cảnh rõ ràng là quá trình “tung xúc xắc”, ta có thể viết gọn $P(\text{tung được mặt 1})$ thành $P(1)$.

BÀI TẬP THỰC HÀNH 3.6

G

Các quá trình ngẫu nhiên bao gồm tung xúc xắc và tung đồng xu. (a) Hãy nghĩ về một quá trình ngẫu nhiên khác. (b) Mô tả tất cả các kết quả có thể xảy ra của quá trình đó. Ví dụ, tung xúc xắc là một quá trình ngẫu nhiên với các kết quả có thể xảy ra là 1, 2, ..., 6.¹

Điều chúng ta xem là quá trình ngẫu nhiên không nhất thiết phải thực sự ngẫu nhiên, mà có thể chỉ quá phức tạp để hiểu chính xác. Ví dụ thứ tư trong phần lời giải chú thích của Thực hành hướng dẫn 3.6 cho rằng hành vi của bạn cùng phòng là một quá trình ngẫu nhiên. Tuy nhiên, ngay cả khi hành vi đó không thực sự ngẫu nhiên, việc mô hình hóa nó như một quá trình ngẫu nhiên vẫn có thể hữu ích.

3.1.3 Các kết quả rời nhau hoặc loại trừ lẫn nhau

Hai kết quả được gọi là **rời nhau (disjoint)** hoặc **loại trừ lẫn nhau (mutually exclusive)** nếu chúng không thể xảy ra đồng thời. Ví dụ, khi tung một viên xúc xắc, các kết quả 1 và 2 là rời nhau vì chúng không thể cùng xuất hiện. Ngược lại, các kết quả 1 và "tung được số lẻ" không phải là rời nhau, vì nếu kết quả tung là 1, thì cả hai đều đúng. Thuật ngữ *rời nhau* và *loại trừ lẫn nhau* có ý nghĩa tương đương và có thể dùng thay thế cho nhau.

Việc tính xác suất của các kết quả rời nhau rất đơn giản. Khi tung một viên xúc xắc, các kết quả 1 và 2 là rời nhau, và xác suất để một trong hai kết quả này xảy ra được tính bằng cách cộng các xác suất riêng lẻ của chúng:

$$P(1 \text{ or } 2) = P(1) + P(2) = 1/6 + 1/6 = 1/3$$

Vậy còn xác suất tung được 1, 2, 3, 4, 5 hoặc 6 thì sao? Trong trường hợp này, tất cả các kết quả đều rời nhau, vì vậy chúng ta tính xác suất bằng cách cộng các xác suất riêng lẻ:

$$\begin{aligned} P(1 \text{ hoặc } 2 \text{ hoặc } 3 \text{ hoặc } 4 \text{ hoặc } 5 \text{ hoặc } 6) \\ &= P(1) + P(2) + P(3) + P(4) + P(5) + P(6) \\ &= 1/6 + 1/6 + 1/6 + 1/6 + 1/6 + 1/6 = 1 \end{aligned}$$

QUY TẮC CỘNG CHO CÁC KẾT QUẢ RỜI NHAU

Quy tắc Cộng (Addition Rule) đảm bảo tính chính xác của phương pháp này khi các kết quả là rời nhau. Nếu A_1 và A_2 là hai kết quả rời nhau, thì xác suất để một trong hai kết quả này xảy ra được tính bằng:

$$P(A_1 \text{ hoặc } A_2) = P(A_1) + P(A_2)$$

Nếu có nhiều kết quả rời nhau $A_1, \dots, A_k$, thì xác suất để một trong những kết quả này xảy ra được tính bằng:

$$P(A_1) + P(A_2) + \dots + P(A_k)$$

¹ Dưới đây là bốn ví dụ: (i) Việc một người có bị ốm trong tháng tới hay không là một quá trình có vẻ ngẫu nhiên với các kết quả có thể xảy ra là **sick** và **not**. (ii) Chúng ta có thể tạo ra một quá trình ngẫu nhiên bằng cách chọn ngẫu nhiên một người và đo chiều cao của họ. Kết quả của quá trình này sẽ là một số dương. (iii) Việc thị trường chứng khoán tăng hay giảm trong tuần tới là một quá trình có vẻ ngẫu nhiên với các kết quả có thể xảy ra là **up**, **down**, và **no_change**. Ngoài ra, chúng ta cũng có thể sử dụng phần trăm thay đổi của thị trường chứng khoán làm kết quả số. (iv) Việc bạn cùng phòng của bạn có rửa bát tối nay hay không có lẽ cũng là một quá trình ngẫu nhiên với các kết quả có thể xảy ra là **cleans_dishes** và **leaves_dishes**.

3.1. ĐỊNH NGHĨA XÁC XUẤT

BÀI TẬP THỰC HÀNH 3.7

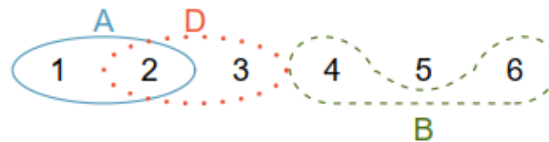
Chúng ta quan tâm đến xác suất tung được 1, 4 hoặc 5. (a) Giải thích vì sao các kết quả 1, 4 và 5 là rời nhau. (b) Áp dụng Quy tắc Cộng cho các kết quả rời nhau để xác định $P(1 \text{ hoặc } 4 \text{ hoặc } 5)$.²

BÀI TẬP THỰC HÀNH 3.8

Trong tập dữ liệu khoản vay ở Chương 2, biến homeownership mô tả việc người vay thuê nhà (rent), có thể chấp (mortgage), hoặc sở hữu nhà (own). Trong số 10.000 người vay, 3.858 người thuê nhà, 4.789 người có thể chấp, 1.353 người sở hữu nhà.³

- (a) Các kết quả thuê nhà, thế chấp và sở hữu nhà có rời nhau không?
- (b) Xác định tỷ lệ các khoản vay thuộc về những người có thể chấp và sở hữu nhà riêng biệt.
- (c) Áp dụng Quy tắc Cộng cho các kết quả rời nhau để tính xác suất một khoản vay được chọn ngẫu nhiên thuộc về người có thể chấp hoặc sở hữu nhà.

Các nhà khoa học dữ liệu hiếm khi làm việc với từng kết quả riêng lẻ mà thay vào đó xem xét tập hợp hoặc nhóm kết quả. Gọi A là sự kiện xúc xắc ra 1 hoặc 2, tức là $A = \{1, 2\}$ và gọi B là sự kiện xúc xắc ra 4 hoặc 6, tức là $B = \{4, 6\}$. Những tập hợp này thường được gọi là **sự kiện (events)**. Vì A và B không có phần tử chung, chúng là sự kiện rời nhau (disjoint events). Hình 3.2. minh họa hai sự kiện này.



Hình 3.2: Ba sự kiện A, B và D bao gồm các kết quả từ việc tung một viên xúc xắc. A và B là rời nhau vì chúng không có kết quả chung.

Quy tắc Cộng áp dụng cho cả các kết quả rời nhau và các sự kiện rời nhau. Xác suất để một trong hai sự kiện A hoặc B xảy ra được tính bằng tổng các xác suất riêng lẻ:

$$P(A \text{ hoặc } B) = P(A) + P(B) = 1/3 + 1/3 = 2/3$$

BÀI TẬP THỰC HÀNH 3.9

- (a) Xác minh rằng xác suất của sự kiện A, $P(A)$, bằng $1/3$ bằng cách sử dụng Quy tắc Cộng. (b) Thực hiện tương tự cho sự kiện B.⁴

BÀI TẬP THỰC HÀNH 3.10

- (a) Dựa vào Hình 3.2, các kết quả nào thuộc sự kiện D? (b) Sự kiện B và sự kiện D có rời nhau không?
- (c) Sự kiện A và sự kiện D có rời nhau không? ⁵

BÀI TẬP THỰC HÀNH 3.11

Trong BÀI TẬP THỰC HÀNH 3.10, bạn đã xác nhận rằng B và D trong Hình 3.2 là rời nhau. Hãy tính xác suất để sự kiện B hoặc sự kiện D xảy ra. ⁶

² (a) Quá trình ngẫu nhiên ở đây là tung xúc xắc, và tại một lần tung, chỉ có thể xuất hiện một trong các kết quả này. Điều này có nghĩa là chúng là các kết quả rời nhau. (b) $P(1 \text{ hoặc } 4 \text{ hoặc } 5) = P(1) + P(4) + P(5) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{3}{6} = \frac{1}{2}$

³ (a) Đúng, vì mỗi khoản vay chỉ thuộc về một trong ba loại hình thức sở hữu nhà (homeownership). (b) Thế chấp (Mortgage): $\frac{4789}{10000} = 0,479$. Sở hữu nhà (Own): $\frac{1353}{10000} = 0,135$. (c) $P(\text{Thế chấp hoặc sở hữu}) = P(\text{Thế chấp}) + P(\text{Sở hữu}) = 0,479 + 0,135 = 0,614$

⁴ (a) $P(A) = P(1 \text{ hoặc } 2) = P(1) + P(2) = \frac{1}{6} + \frac{1}{6} = \frac{2}{6} = \frac{1}{3}$. (b) Tương tự, $P(B) = \frac{1}{3}$

⁵ (a) Sự kiện D gồm các kết quả 2 và 3. (b) Có, sự kiện B và D là rời nhau vì chúng không có kết quả chung. (c) Sự kiện A và D có chung kết quả 2, vì vậy chúng không rời nhau.

⁶ Vì B và D là các sự kiện rời nhau, áp dụng Quy tắc Cộng: $P(B \text{ hoặc } D) = P(B) + P(D) = \frac{1}{3} + \frac{1}{3} = \frac{2}{3}$.

3.1.4 Xác suất khi các sự kiện không rời nhau

Hãy xem xét cách tính xác suất cho hai sự kiện không rời nhau trong bối cảnh bộ bài 52 lá tiêu chuẩn, được minh họa trong Hình 3.3. Nếu bạn chưa quen với các lá bài trong một bộ bài tiêu chuẩn, vui lòng tham khảo chú thích.⁷



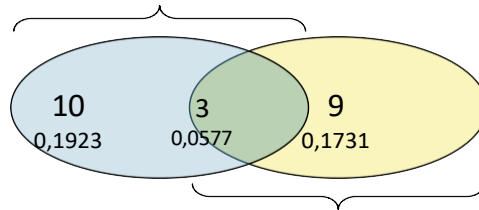
Hình 3.3: Minh họa 52 lá bài độc nhất trong một bộ bài tiêu chuẩn.

BÀI TẬP THỰC HÀNH 3.12

(a) Xác suất để một lá bài được chọn ngẫu nhiên là lá bài rô là bao nhiêu? (b) Xác suất để một lá bài được chọn ngẫu nhiên là lá bài hình (J, Q, K) là bao nhiêu?⁸

Biểu đồ Venn rất hữu ích khi các kết quả có thể được phân loại là "thuộc" hoặc "không thuộc" đối với hai hoặc ba biến số, thuộc tính, hoặc quá trình ngẫu nhiên. Biểu đồ Venn trong Hình 3.4 sử dụng một vòng tròn để biểu diễn các lá bài rô (diamonds) và một vòng tròn khác để biểu diễn các lá bài hình (face cards). Nếu một lá bài vừa là rô vừa là bài hình, nó sẽ nằm trong giao nhau của hai vòng tròn. Nếu một lá bài là rô nhưng không phải bài hình, nó sẽ nằm trong phần còn lại của vòng tròn bên trái, không giao với vòng tròn bên phải (và tương tự cho các trường hợp khác). Tổng số lá bài rô được xác định bằng tổng số lá bài nằm trong vòng tròn "bài rô": $10 + 3 = 13$. Các xác suất cũng được hiển thị (ví dụ: $10/52 = 0,1923$).

Bài rô: 0,2500



Bài hình: 0,2308

Có 30 lá bài không phải bài rô cũng không phải bài hình

Hình 3.4: Biểu đồ Venn cho bài rô và bài hình.

Gọi A là sự kiện lá bài được chọn ngẫu nhiên là bài rô, và B là sự kiện lá bài được chọn là bài hình. Làm thế nào để tính $P(A \text{ hoặc } B)$? Vì A và B không phải là hai sự kiện rời nhau – các lá bài $J♦️, Q♦️, K♦️$ thuộc cả hai nhóm – nên chúng ta không thể sử dụng Quy tắc Cộng cho các sự kiện rời nhau. Thay vào đó, ta sử dụng biểu đồ Venn. Bước đầu tiên là cộng các xác suất của hai sự kiện:

$$P(A) + P(B) = P(♦️) + P(\text{bài hình}) = 13/52 + 12/52$$

⁷ Bộ bài 52 lá được chia thành bốn chất: ♣️ (chuồn), ♦️ (rô), ♥️ (cơ), ♠️ (bích). Mỗi chất có 13 lá bài được đánh số: 2, 3, ..., 10, J (jack), Q (queen), K (king), và A (ace). Như vậy, mỗi lá bài là một sự kết hợp duy nhất giữa một chất và một giá trị, ví dụ: $4♥️$ và $J♠️$. 12 lá bài J, Q, K được gọi là lá bài hình. Các lá bài thuộc chất ♦️ (rô) và ♥️ (cơ) thường có màu **đỏ**, trong khi hai chất còn lại có màu **đen**.

⁸ (a) Có tổng cộng 52 lá bài và 13 lá thuộc chất rô. Nếu bộ bài được xáo trộn kỹ, mỗi lá bài có cơ hội được chọn ngang nhau. Do đó, xác suất để một lá bài ngẫu nhiên là chất rô là: $P(♦️) = \frac{13}{52} = 0,250$. (b) Tương tự, có 12 lá bài hình, nên xác suất để chọn một lá bài hình là: $P(\text{bài hình}) = \frac{12}{52} = \frac{3}{13} = 0,231$.

3.1. ĐỊNH NGHĨA XÁC SUẤT

Tuy nhiên, ba lá bài thuộc cả hai sự kiện đã bị đếm hai lần, một lần trong mỗi xác suất. Chúng ta cần điều chỉnh sự đếm trùng này:

$$\begin{aligned} P(A \text{ hoặc } B) &= P(\diamond \text{ hoặc bài hình}) \\ &= P(\diamond) + P(\text{bài hình}) - P(\diamond \text{ và bài hình}) \\ &= 13/52 + 12/52 - 3/52 \\ &= 22/52 = 11/26 \end{aligned}$$

Phương trình này là một ví dụ về **Quy tắc Cộng Tổng quát (General Addition Rule)**.

QUY TẮC CỘNG TỔNG QUÁT

Nếu A và B là hai sự kiện bất kỳ, dù có rời nhau hay không, thì xác suất ít nhất một trong hai sự kiện xảy ra được tính theo công thức:

$$P(A \text{ hoặc } B) = P(A) + P(B) - P(A \text{ và } B)$$

trong đó $P(A \text{ và } B)$ là xác suất cả hai sự kiện cùng xảy ra.

MẸO: "hoặc" có tính bao hàm

Khi chúng ta sử dụng từ "hoặc" trong thống kê, nó có nghĩa là "và/hoặc" trừ khi được chỉ định rõ ràng khác đi.

BÀI TẬP THỰC HÀNH 3.13

(a) Nếu A và B là rời nhau, hãy giải thích tại sao điều này có nghĩa là $P(A \text{ và } B) = 0$. (b) Sử dụng phần (a), hãy xác minh rằng Quy tắc Cộng Tổng quát được đơn giản hóa thành Quy tắc Cộng cho các sự kiện rời nhau nếu A và B là rời nhau.⁹

BÀI TẬP THỰC HÀNH 3.14

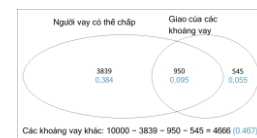
Trong tập dữ liệu khoản vay mô tả 10.000 khoản vay, 1495 khoản vay đến từ các đơn đăng ký chung (ví dụ: một cặp vợ chồng cùng nộp đơn), 4789 người vay có thể chấp, và 950 người có cả hai đặc điểm này. Hãy tạo một biểu đồ Venn cho tình huống này.¹⁰

BÀI TẬP THỰC HÀNH 3.15

(a) Sử dụng biểu đồ Venn từ BÀI TẬP THỰC HÀNH 3.14 để xác định xác suất một khoản vay được chọn ngẫu nhiên từ tập dữ liệu khoản vay đến từ một đơn đăng ký chung, trong đó cặp vợ chồng có thể chấp. (b) Xác suất để khoản vay có ít nhất một trong hai đặc điểm này là bao nhiêu?¹¹

⁹ (a) Nếu A và B là hai sự kiện rời nhau, thì chúng không bao giờ có thể xảy ra đồng thời. (b) Nếu A và B là hai sự kiện rời nhau, thì xác suất $P(A \text{ và } B)$ trong công thức Quy tắc Cộng tổng quát sẽ bằng 0 (theo phần (a)), và khi đó công thức chỉ còn lại Quy tắc Cộng cho các sự kiện rời nhau.

¹⁰ Cả số lượng và xác suất tương ứng (ví dụ: $3839/10000 = 0,384$) đều được hiển thị. Lưu ý rằng số vay được biểu diễn trong vòng tròn bên trái là $3839 + 950 = 4789$, và số lượng trong vòng tròn bên phải là $545 = 1495$.



lượng khoản
phải là $950 +$

¹¹ (a) Giải pháp được biểu diễn bởi phần giao của hai vòng tròn: 0,095. (b) Đây là tổng của ba xác được hiển thị trong các vòng tròn: $0,384 + 0,095 + 0,055 = 0,534$ (chênh lệch 0,001 do làm tròn số).

suất rời nhau

3.1.5 Phân phối xác suất

Một **phân phối xác suất** là một bảng liệt kê tất cả các kết quả rời nhau và xác suất tương ứng của chúng. Hình 3.5 hiển thị phân phối xác suất cho tổng của hai viên xúc xắc.

Tổng xúc xắc	2	3	4	5	6	7	8	9	10	11	12
Xác suất	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

Hình 3.5: Phân phối xác suất cho tổng của hai viên xúc xắc.

QUY TẮC CHO PHÂN PHỐI XÁC SUẤT

Một phân phối xác suất là danh sách các kết quả có thể xảy ra cùng với xác suất tương ứng, thỏa mãn ba quy tắc sau:

1. Các kết quả liệt kê phải rời nhau.
2. Mỗi xác suất phải nằm trong khoảng từ 0 đến 1.
3. Tổng tất cả các xác suất phải bằng 1.

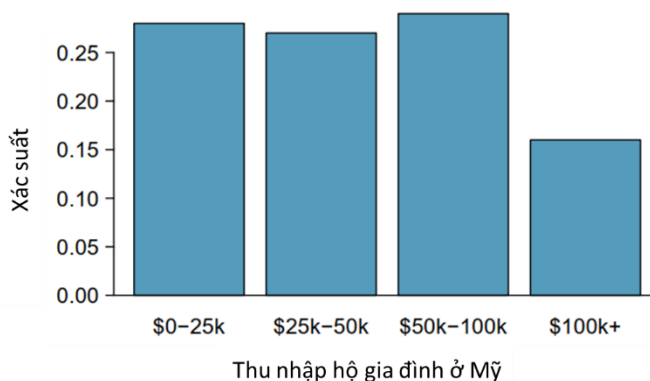
BÀI TẬP THỰC HÀNH 3.16

Hình 3.6 đề xuất ba phân phối thu nhập hộ gia đình ở Hoa Kỳ. Chỉ một trong số đó là đúng. Đó là phân phối nào? Hai phân phối còn lại sai ở điểm nào?¹²

Mức thu nhập	\$0-25k	\$25k-50k	\$50k-100k	\$100k+
(a)	0,18	0,39	0,33	0,16
(b)	0,38	0,27	0,52	0,37
(c)	0,28	0,27	0,29	0,16

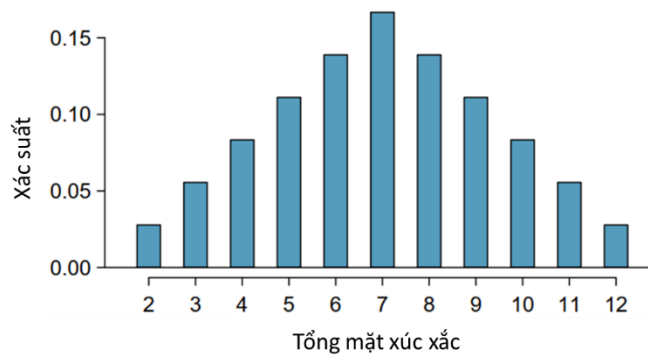
Hình 3.6: Các phân phối được đề xuất về thu nhập hộ gia đình ở Hoa Kỳ (Bài thực hành hướng dẫn 3.16).

Chương 1 nhấn mạnh tầm quan trọng của việc vẽ biểu đồ dữ liệu để cung cấp bản tóm tắt nhanh. Phân phối xác suất cũng có thể được tóm tắt bằng biểu đồ cột. Ví dụ, phân phối thu nhập hộ gia đình ở Hoa Kỳ được hiển thị trong Hình 3.7 dưới dạng biểu đồ cột. Phân phối xác suất cho tổng của hai con xúc xắc được thể hiện trong Hình 3.5 và được vẽ trong Hình 3.8.



Hình 3.7: Phân phối xác suất của thu nhập hộ gia đình ở Hoa Kỳ.

¹² Xác suất trong (a) không có tổng bằng 1. Xác suất thứ hai trong (b) là một số âm. Do đó, chỉ còn (c), và đúng là nó đáp ứng các yêu cầu của một phân phối. Một trong ba phân phối được cho là phân phối thực tế của thu nhập hộ gia đình ở Hoa Kỳ, vì vậy đó phải là (c).



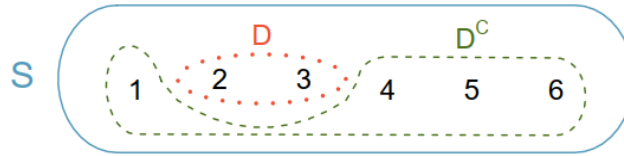
Hình 3.8: Phân phối xác suất của tổng hai con xúc xắc.

Trong các biểu đồ cột này, chiều cao của các cột biểu diễn xác suất của các kết quả. Nếu các kết quả là số và rời rạc, thì thường thuận tiện (về mặt trực quan) để vẽ biểu đồ cột giống như biểu đồ tần suất, như trong trường hợp tổng của hai con xúc xắc. Một ví dụ khác về việc vẽ các cột tại các vị trí tương ứng của chúng được hiển thị trong Hình 3.18 trên trang 115.

3.1.6 Bổ sung của một biến cố

Khi tung một con xúc xắc, ta nhận được một giá trị trong tập hợp $\{1, 2, 3, 4, 5, 6\}$. Tập hợp tất cả các kết quả có thể này được gọi là **không gian mẫu** (S) cho phép thử tung xúc xắc. Chúng ta thường sử dụng không gian mẫu để xem xét tình huống khi một biến cố không xảy ra.

Gọi $D = \{2, 3\}$ là biến cố xảy ra khi kết quả tung xúc xắc là 2 hoặc 3. Khi đó, **bổ sung** của D là tập hợp tất cả các kết quả trong không gian mẫu không thuộc D , được ký hiệu là $D^c = \{1, 4, 5, 6\}$. Nói cách khác, D^c bao gồm tất cả các kết quả có thể mà không thuộc biến cố D . Hình 3.9 minh họa mối quan hệ giữa D , D^c và không gian mẫu S .

Hình 3.9: Biến cố $D = \{2, 3\}$ và bổ sung của nó, $D^c = \{1, 4, 5, 6\}$. S đại diện cho không gian mẫu, tức là tập hợp tất cả các kết quả có thể.

BÀI TẬP THỰC HÀNH 3.17

G

(a) Tính $P(D^c) = P(\text{xuất hiện mặt } 1, 4, 5 \text{ hoặc } 6)$. (b) Giá trị của $P(D) + P(D^c)$ là bao nhiêu? ¹³

BÀI TẬP THỰC HÀNH 3.18

G

Các sự kiện $A = \{1, 2\}$ và $B = \{4, 6\}$ được hiển thị trong Hình 3.2 trên trang 84. (a) Viết ra tập hợp mà A^c và B^c biểu diễn. (b) Tính $P(A^c)$ và $P(B^c)$. (c) Tính $P(A) + P(A^c)$ và $P(B) + P(B^c)$. ¹⁴

¹³ (a) Các kết quả là rời nhau và mỗi kết quả có xác suất $1/6$, nên tổng xác suất là $4/6 = 2/3$. (b) Ta cũng có thể thấy rằng $P(D) = 1/6 + 1/6 = 1/3$. Vì D và D^c là rời nhau, nên $P(D) + P(D^c) = 1$.

¹⁴ Giải pháp tóm tắt: (a) $A^c = \{3, 4, 5, 6\}$ và $B^c = \{1, 2, 3, 5\}$. (b) Vì mỗi kết quả là rời nhau, ta cộng các xác suất riêng lẻ để có $P(A^c) = 2/3$ và $P(B^c) = 2/3$. (c) A và A^c là rời nhau, và điều tương tự cũng đúng với B và B^c . Do đó, $P(A) + P(A^c) = 1$ và $P(B) + P(B^c) = 1$.

3.1. ĐỊNH NGHĨA XÁC XUẤT

Phần bù của một biến cố A được xây dựng để có hai tính chất quan trọng sau: (i) Mọi kết quả có thể xảy ra không thuộc A đều thuộc A^c . (ii) A và A^c là rời nhau. Tính chất (i) suy ra:

$$P(A \text{ hoặc } A^c) = 1$$

Nghĩa là, nếu kết quả không thuộc A, thì nó phải thuộc A^c . Ta sử dụng Quy tắc Cộng cho các biến cố rời nhau để áp dụng tính chất (ii):

$$P(A \text{ hoặc } A^c) = P(A) + P(A^c)$$

Kết hợp hai phương trình trên, ta có một mối quan hệ rất hữu ích giữa xác suất của một biến cố và xác suất của phần bù của nó.

PHẦN BÙ (COMPLEMENT)

Phần bù của biến cố A được ký hiệu là A^c , và A^c bao gồm tất cả các kết quả không thuộc A. A và A^c có mối quan hệ toán học như sau:

$$P(A) + P(A^c) = 1, \quad \text{Ví dụ } P(A) = 1 - P(A^c)$$

Trong những ví dụ đơn giản, tính toán A hoặc A^c có thể thực hiện trong vài bước. Tuy nhiên, sử dụng phần bù có thể tiết kiệm nhiều thời gian khi bài toán trở nên phức tạp hơn.

BÀI TẬP THỰC HÀNH 3.19

Gọi A là sự kiện khi chúng ta tung hai con xúc xắc và tổng của chúng nhỏ hơn 12. (a) Sự kiện A^c biểu thị điều gì? (b) Xác định $P(A^c)$ từ Hình 3.5 trên trang 87. (c) Xác định $P(A)$.¹⁵

BÀI TẬP THỰC HÀNH 3.20

Tìm các xác suất sau khi tung hai con xúc xắc:¹⁶

- (a) Tổng của hai xúc xắc *không* bằng 6.
- (b) Tổng ít nhất là 4. Tức là, xác định xác suất của sự kiện $B = \{4, 5, \dots, 12\}$.
- (c) Tổng không lớn hơn 10. Tức là, xác định xác suất của sự kiện $D = \{2, 3, \dots, 10\}$.

3.1.7 Tính độc lập (Independence)

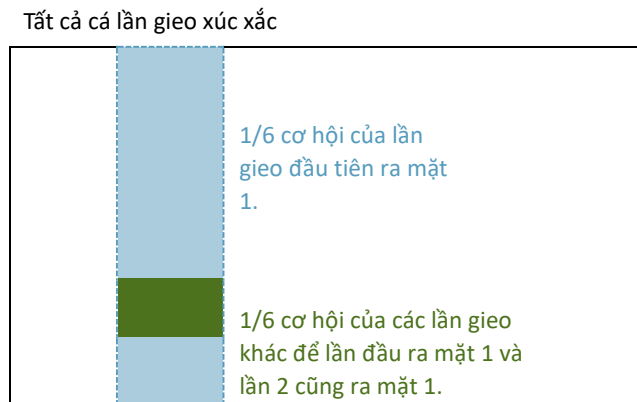
Cũng như các biến và quan sát có thể độc lập, các quá trình ngẫu nhiên cũng có thể độc lập. Hai quá trình được coi là **độc lập** nếu biết kết quả của một quá trình không cung cấp bất kỳ thông tin hữu ích nào về kết quả của quá trình kia. Ví dụ, tung một đồng xu và lần một con xúc xắc là hai quá trình độc lập – biết rằng đồng xu ra mặt ngửa không giúp xác định kết quả của lần lăn xúc xắc. Ngược lại, giá cổ phiếu thường biến động cùng nhau, vì vậy chúng không độc lập.

Ví dụ 3.5 cung cấp một minh họa cơ bản về hai quá trình độc lập: lần hai con xúc xắc. Chúng ta muốn xác định xác suất cả hai con xúc xắc đều ra mặt 1. Giả sử một con xúc xắc màu đỏ và con còn lại màu trắng. Nếu kết quả của xúc xắc đỏ là 1, điều này không cung cấp bất kỳ thông tin nào về kết quả của xúc xắc trắng. Chúng ta đã gặp câu hỏi này trong Ví dụ 3.5 (trang 81), nơi xác suất được tính theo cách sau: Xác suất xúc xắc đỏ ra 1 là $\frac{1}{6}$, và trong số những lần đó, xác suất xúc xắc trắng cũng ra 1 là $\frac{1}{6}$.

¹⁵ (a) Bổ sung của sự kiện A: khi tổng hai xúc xắc bằng 12. (b) $P(A^c) = 1/36$. (c) Sử dụng xác suất của bổ sung từ phần (b), $P(A^c) = 1/36$, và công thức bổ sung: $P(\text{tổng nhỏ hơn 12}) = 1 - P(12) = 1 - 1/36 = 35/36$.

¹⁶ (a) Trước tiên, tìm $P(6) = 5/36$, sau đó sử dụng bổ sung: $P(\text{không phải 6}) = 1 - P(6) = 31/36$. (b) Trước tiên, tìm bổ sung vì nó dễ tính hơn nhiều: $P(2 \text{ hoặc } 3) = 1/36 + 2/36 = 1/12$. Sau đó, tính $P(B)$: $P(B) = 1 - P(B^c) = 1 - 1/12 = 11/12$. (c) Tương tự, tìm bổ sung là cách thông minh để xác định $P(D)$. Trước tiên, tìm $P(D^c) = P(11 \text{ hoặc } 12) = 2/36 + 1/36 = 1/12$. Sau đó, tính: $P(D) = 1 - P(D^c) = 11/12$.

Điều này được minh họa trong Hình 3.10. Vì các lần gieo xúc xắc là độc lập, ta có thể nhân các xác suất của từng kết quả để tìm đáp án cuối cùng: $(1/6) \cdot (1/6) = 1/36$. Quy tắc này có thể được mở rộng cho nhiều quá trình độc lập khác.



Hình 3.10: Trong 1/6 số lần gieo, lượt đầu tiên ra mặt 1. Trong 1/6 số lần tiếp đó, lượt gieo thứ hai cũng ra mặt 1.

VÍ DỤ 3.21

Điều gì sẽ xảy ra nếu có thêm một con xúc xắc màu xanh, độc lập với hai con còn lại? Xác suất để gieo cả ba con xúc xắc và nhận được toàn bộ mặt 1 là bao nhiêu?

Logic tương tự như trong Ví dụ 3.5. Nếu 1/36 số lần gieo, cả hai con xúc xắc trắng và đỏ đều ra mặt 1, thì trong 1/6 số lần đó, con xúc xắc xanh cũng sẽ ra mặt 1. Do đó, ta nhân các xác suất lại với nhau:

$$\begin{aligned} P(\text{trắng} = 1 \text{ và } \text{đỏ} = 1 \text{ và } \text{xanh dương} = 1) &= P(\text{trắng} = 1) \times P(\text{đỏ} = 1) \times P(\text{xanh dương} = 1) \\ &= (1/6) \times (1/6) \times (1/6) = 1/216 \end{aligned}$$

Ví dụ 3.21 minh họa quy tắc nhân (Multiplication Rule) cho các quá trình độc lập.

QUY TẮC NHÂN CHO CÁC QUÁ TRÌNH ĐỘC LẬP

Nếu A và B là các biến cố từ hai quá trình khác nhau và độc lập, thì xác suất để cả A và B xảy ra có thể được tính bằng tích của xác suất riêng lẻ của chúng:

$$P(A \text{ và } B) = P(A) \times P(B)$$

Tương tự, nếu có k biến cố $A_1, \dots, A_k$ từ k quá trình độc lập, thì xác suất để tất cả chúng xảy ra là:

$$P(A_1) \times P(A_2) \times \dots \times P(A_k)$$

VÍ DỤ 3.22

Khoảng 9% dân số thuận tay trái. Giả sử chọn ngẫu nhiên 2 người từ dân số Hoa Kỳ. Vì cỡ mẫu 2 là rất nhỏ so với tổng dân số, nên có thể giả định rằng hai người này là độc lập. (a) Xác suất để cả hai người đều thuận tay trái là bao nhiêu? (b) Xác suất để cả hai người đều thuận tay phải là bao nhiêu? ¹⁷

¹⁷(a) Xác suất để người đầu tiên thuận tay trái là 0,09, và xác suất này cũng tương tự đối với người thứ hai. Áp dụng Quy tắc Nhân cho các quá trình độc lập, xác suất để cả hai người đều thuận tay trái là: $0,09 \times 0,09 = 0,0081$.

(b) Có thể giả định hợp lý rằng tỷ lệ người thuận cả hai tay (cả phải và trái) gần bằng 0, do đó xác suất để một người thuận tay phải là: $P(\text{thuận tay phải}) = 1 - 0,09 = 0,91$. Sử dụng cùng lập luận như ở phần (a), xác suất để cả hai người đều thuận tay phải là: $0,91 \times 0,91 = 0,8281$.

BÀI TẬP THỰC HÀNH 3.23

Giả sử chọn ngẫu nhiên 5 người.¹⁸

- (a) Xác suất để tất cả đều thuận tay phải là bao nhiêu?
 (b) Xác suất để tất cả đều thuận tay trái là bao nhiêu?
 (c) Xác suất để không phải tất cả mọi người đều thuận tay phải là bao nhiêu?

Giả sử các biến thuận tay và giới tính là độc lập, tức là biết giới tính của một người không cung cấp thông tin hữu ích về việc họ thuận tay nào và ngược lại. Khi đó, ta có thể tính xác suất để một người được chọn ngẫu nhiên vừa thuận tay phải vừa là nữ bằng Quy tắc Nhân:

$$\begin{aligned} P(\text{thuận tay phải và là nữ}) &= P(\text{thuận tay phải}) \times P(\text{là nữ}) \\ &= 0,91 \times 0,50 = 0,455 \end{aligned}$$

BÀI TẬP THỰC HÀNH 3.24

Ba người được chọn ngẫu nhiên.²⁰

- (a) Xác suất người đầu tiên là nam và thuận tay phải là bao nhiêu?
 (b) Xác suất hai người đầu tiên đều là nam và thuận tay phải là bao nhiêu?
 (c) Xác suất người thứ ba là nữ và thuận tay trái là bao nhiêu?
 (d) Xác suất hai người đầu tiên đều là nam và thuận tay phải, còn người thứ ba là nữ và thuận tay trái là bao nhiêu?

Đôi khi chúng ta tự hỏi liệu một kết quả có cung cấp thông tin hữu ích về một kết quả khác hay không. Câu hỏi đặt ra là liệu hai sự kiện có độc lập với nhau hay không? Chúng ta nói rằng hai sự kiện A và B là độc lập nếu chúng thỏa mãn: $P(A \text{ và } B) = P(A) \times P(B)$.

VÍ DỤ 3.25

Nếu chúng ta xáo trộn một bộ bài và rút một lá, liệu việc rút lá bài là cơ có độc lập với sự kiện lá bài là át không?

Xác suất để lá bài là cơ là $1/4$ và xác suất để lá bài là át là $1/13$. Xác suất để lá bài là át cơ là $1/52$. Chúng ta kiểm tra xem phương trình $P(A \text{ và } B) = P(A) \times P(B)$ có thỏa mãn không:

$$P(\heartsuit) \times P(\text{át}) = \frac{1}{4} \times \frac{1}{13} = \frac{1}{52} = P(\heartsuit \text{ và át})$$

Vì phương trình đúng, nên sự kiện lá bài là cơ và sự kiện lá bài là át là hai sự kiện độc lập.

¹⁸ (a) Các ký hiệu RH và LH lần lượt được sử dụng để chỉ người thuận tay phải và người thuận tay trái. Vì mỗi người được chọn là độc lập, chúng ta áp dụng Quy tắc Nhân cho các quá trình độc lập:

$$\begin{aligned} P(\text{cả năm người đều RH}) &= P(\text{người thứ nhất RH, người thứ hai RH, ..., người thứ năm RH}) \\ &= P(\text{người thứ nhất RH}) \times P(\text{người thứ hai RH}) \times \dots \times P(\text{người thứ năm RH}) \\ &= 0,91 \times 0,91 \times 0,91 \times 0,91 \times 0,91 = 0,624 \end{aligned}$$

(b) Áp dụng cùng phương pháp như (a): $0,09 \times 0,09 \times 0,09 \times 0,09 \times 0,09 = 0,0000059$

(c) Sử dụng bổ sung của xác suất $P(\text{cả năm đều RH})$ để trả lời câu hỏi:

$$P(\text{không phải tất cả đều RH}) = 1 - P(\text{tất cả đều RH}) = 1 - 0,624 = 0,376$$

¹⁹ Tỷ lệ thực tế của dân số Hoa Kỳ là nữ vào khoảng 50%, vì vậy chúng ta sử dụng 0.5 làm xác suất chọn một phụ nữ ngẫu nhiên. Tuy nhiên, xác suất này có thể khác ở các quốc gia khác.

²⁰ Các câu trả lời ngắn gọn: (a) Có thể viết dưới dạng ký hiệu xác suất: $P(\text{một người được chọn ngẫu nhiên là nam và thuận tay phải}) = 0,455$ (b) 0,207 (c) 0,045 (d) 0,0093

Bài tập

3.1 Đúng hay sai. Xác định các phát biểu sau là đúng hay sai và giải thích lý do của bạn.

- (a) Nếu một đồng xu công bằng được tung nhiều lần và tám lần tung gần nhất đều ra mặt ngửa, thì xác suất lần tung tiếp theo sẽ ra mặt ngửa nhỏ hơn 50%.
- (b) Rút một lá bài hình (J, Q, K) và rút một lá bài đỏ từ một bộ bài đầy đủ là hai sự kiện loại trừ lẫn nhau.
- (c) Rút một lá bài hình và rút một lá bài Át từ một bộ bài đầy đủ là hai sự kiện loại trừ lẫn nhau.

3.2 Vòng quay Roulette. Trò chơi roulette bao gồm việc quay một bánh xe với 38 ô: 18 ô đỏ, 18 ô đen và 2 ô xanh. Một viên bi được quay vào bánh xe và cuối cùng sẽ rơi vào một ô, với mỗi ô có xác suất như nhau để bắt viên bi.

- (a) Bạn quan sát bánh xe quay 3 lần liên tiếp và viên bi rơi vào ô đỏ mỗi lần. Xác suất viên bi rơi vào ô đỏ ở lần quay tiếp theo là bao nhiêu?
- (b) Bạn quan sát bánh xe quay 300 lần liên tiếp và viên bi rơi vào ô đỏ mỗi lần. Xác suất viên bi rơi vào ô đỏ ở lần quay tiếp theo là bao nhiêu?
- (c) Bạn có tự tin như nhau về câu trả lời của mình trong (a) và (b) không? Tại sao hoặc tại sao không?



Photo by Håkan Dahlström
(<http://flic.kr/p/93fEzp>) CC BY 2.0 license

3.3 Bốn trò chơi, một người thắng. Dưới đây là bốn phiên bản của cùng một trò chơi. Đối thủ của bạn sẽ chọn phiên bản trò chơi, sau đó bạn có thể chọn số lần tung đồng xu: 10 lần hoặc 100 lần. Xác định số lần tung đồng xu bạn nên chọn cho mỗi phiên bản trò chơi. Mỗi lần chơi trò chơi tốn 1 đô la. Giải thích lý do của bạn.

- (a) Nếu tỷ lệ mặt ngửa lớn hơn 0.60, bạn thắng 1 đô la.
- (b) Nếu tỷ lệ mặt ngửa lớn hơn 0.40, bạn thắng 1 đô la.
- (c) Nếu tỷ lệ mặt ngửa nằm trong khoảng 0.40 và 0.60, bạn thắng 1 đô la.
- (d) Nếu tỷ lệ mặt ngửa nhỏ hơn 0.30, bạn thắng 1 đô la.

3.4 Backgammon. Backgammon là một trò chơi bàn cờ dành cho hai người, trong đó các quân cờ di chuyển theo kết quả của hai viên xúc xắc. Người chơi thắng khi loại bỏ tất cả quân cờ của mình khỏi bàn cờ, vì vậy việc tung được số cao thường có lợi. Bạn đang chơi backgammon với một người bạn và bạn tung được hai con 6 ở lượt đầu tiên, sau đó lại tung hai con 6 ở lượt thứ hai. Người bạn của bạn tung hai con 3 ở lượt đầu tiên và lại tung hai con 3 ở lượt thứ hai. Người bạn của bạn cho rằng bạn gian lận, vì tung hai con 6 hai lần liên tiếp là rất khó xảy ra. Sử dụng xác suất để chứng minh rằng kết quả của bạn cũng có khả năng xảy ra như kết quả của anh ấy.

3.5 Tung đồng xu. Nếu bạn tung một đồng xu công bằng 10 lần, xác suất:

- (a) Ra toàn bộ mặt sấp là bao nhiêu?
- (b) Ra toàn bộ mặt ngửa là bao nhiêu?
- (c) Có ít nhất một lần ra mặt sấp là bao nhiêu?

3.6 Tung xúc xắc. Nếu bạn tung hai viên xúc xắc công bằng, xác suất

- (a) Được tổng bằng 1 là bao nhiêu?
- (b) Được tổng bằng 5 là bao nhiêu?
- (c) Được tổng bằng 12 là bao nhiêu?

3.7 Cử tri dao động. Một cuộc khảo sát của Pew Research đã hỏi 2.373 cử tri đăng ký được chọn ngẫu nhiên về đảng phái chính trị của họ (Cộng hòa, Dân chủ hoặc Độc lập) và liệu họ có xác định mình là cử tri dao động hay không. 35% số người được hỏi xác định là Độc lập, 23% xác định là cử tri dao động và 11% thuộc cả hai nhóm.²¹

- (a) Việc là Độc lập và là cử tri dao động có loại trừ lẫn nhau không?
- (b) Vẽ sơ đồ Venn tóm tắt các biến và xác suất tương ứng của chúng.
- (c) Bao nhiêu phần trăm cử tri là Độc lập nhưng không phải là cử tri dao động?
- (d) Bao nhiêu phần trăm cử tri là Độc lập hoặc cử tri dao động?
- (e) Bao nhiêu phần trăm cử tri không phải là Độc lập cũng không phải là cử tri dao động?
- (f) Sự kiện một người là cử tri dao động có độc lập với sự kiện một người là Độc lập về chính trị không?

3.8 Nghèo đói và ngôn ngữ. Cuộc Khảo sát Cộng đồng Mỹ là một khảo sát đang diễn ra nhằm cung cấp dữ liệu hàng năm để hỗ trợ các cộng đồng trong việc lập kế hoạch đầu tư và dịch vụ. Cuộc khảo sát năm 2010 ước tính rằng 14,6% người Mỹ sống dưới mức nghèo khổ, 20,7% nói một ngôn ngữ khác ngoài tiếng Anh tại nhà, và 4,2% thuộc cả hai nhóm.²²

- (a) Việc sống dưới mức nghèo khổ và nói một ngôn ngữ nước ngoài tại nhà có loại trừ lẫn nhau không?
- (b) Vẽ sơ đồ Venn tóm tắt các biến và xác suất tương ứng của chúng.
- (c) Bao nhiêu phần trăm người Mỹ sống dưới mức nghèo khổ nhưng chỉ nói tiếng Anh tại nhà?
- (d) Bao nhiêu phần trăm người Mỹ sống dưới mức nghèo khổ hoặc nói một ngôn ngữ nước ngoài tại nhà?
- (e) Bao nhiêu phần trăm người Mỹ sống trên mức nghèo khổ và chỉ nói tiếng Anh tại nhà?
- (f) Sự kiện một người sống dưới mức nghèo khổ có độc lập với sự kiện người đó nói một ngôn ngữ nước ngoài tại nhà không?

3.9 Sự kiện loại trừ và độc lập. Trong các phần (a) và (b), xác định xem các sự kiện có loại trừ nhau, độc lập hay không thuộc loại nào (vì các sự kiện không thể vừa loại trừ vừa độc lập).

- (a) Bạn và một sinh viên được chọn ngẫu nhiên từ lớp của bạn đều đạt điểm A trong khóa học này.
- (b) Bạn và bạn học nhóm của bạn đều đạt điểm A trong khóa học này.
- (c) Nếu hai sự kiện có thể xảy ra cùng lúc, liệu chúng có nhất thiết phải phụ thuộc không?

3.10 Đoán câu trả lời trong bài kiểm tra. Trong một bài kiểm tra trắc nghiệm, có 5 câu hỏi và mỗi câu có 4 lựa chọn (a, b, c, d). Nancy chưa học gì cho bài kiểm tra và quyết định đoán câu trả lời một cách ngẫu nhiên. Xác suất:

- (a) Câu hỏi đầu tiên mà cô ấy trả lời đúng là câu thứ 5?
- (b) Cô ấy trả lời đúng tất cả các câu hỏi?
- (c) Cô ấy trả lời đúng ít nhất một câu hỏi?

²¹ Trung tâm Nghiên cứu Pew, Với việc cử tri tập trung vào kinh tế, khoảng cách dẫn đầu của Obama thu hẹp lại, dữ liệu thu thập từ ngày 4-15 tháng 4 năm 2012.

²² Cục Điều tra Dân số Hoa Kỳ, Khảo sát Cộng đồng Hoa Kỳ năm 2010 - Ước tính trong 1 năm, Đặc điểm của người dân theo ngôn ngữ sử dụng tại nhà.

3.1. ĐỊNH NGHĨA XÁC XUẤT

3.11 Trình độ học vấn của các cặp vợ chồng. Bảng dưới đây cho thấy phân bố trình độ học vấn của cư dân Hoa Kỳ theo giới tính, dựa trên dữ liệu thu thập từ Khảo sát Cộng đồng Hoa Kỳ năm 2010.²³

		Giới tính	
		Nam	Nữ
<i>Trình độ học vấn cao nhất</i>	Dưới lớp 9	0,07	0,13
	Từ lớp 9 đến 12, không có bằng tốt nghiệp	0,10	0,09
	Tốt nghiệp THPT (hoặc tương đương)	0,30	0,20
	Một số tín chỉ đại học, không có bằng	0,22	0,24
	Bằng cao đẳng	0,06	0,08
	Bằng cử nhân	0,16	0,17
	Bằng sau đại học hoặc chuyên nghiệp	0,09	0,09
	Tổng	1,00	1,00

- Xác suất để một người đàn ông được chọn ngẫu nhiên có ít nhất bằng Cử nhân là bao nhiêu?
- Xác suất để một người phụ nữ được chọn ngẫu nhiên có ít nhất bằng Cử nhân là bao nhiêu?
- Xác suất để cả một người đàn ông và một người phụ nữ khi kết hôn đều có ít nhất bằng Cử nhân là bao nhiêu? Hãy nêu các giả định cần thiết để trả lời câu hỏi này.
- Nếu bạn đã đưa ra giả định trong phần (c), liệu giả định đó có hợp lý không? Nếu bạn không đưa ra giả định nào, hãy kiểm tra lại câu trả lời trước đó rồi quay lại phần này.

3.12 Vắng mặt ở trường học. Dữ liệu thu thập tại các trường tiểu học ở Quận DeKalb, bang Georgia cho thấy mỗi năm, tỷ lệ học sinh nghỉ học do ốm như sau: 25% học sinh nghỉ đúng 1 ngày, 15% học sinh nghỉ 2 ngày, 28% học sinh nghỉ từ 3 ngày trở lên.²⁴

- Xác suất để một học sinh được chọn ngẫu nhiên không nghỉ ngày nào do ốm trong năm nay là bao nhiêu?
- Xác suất để một học sinh được chọn ngẫu nhiên nghỉ không quá một ngày là bao nhiêu?
- Xác suất để một học sinh được chọn ngẫu nhiên nghỉ ít nhất một ngày là bao nhiêu?
- Nếu một phụ huynh có hai con học tại một trường tiểu học ở Quận DeKalb, xác suất để cả hai con đều không nghỉ ngày nào là bao nhiêu? Hãy nêu các giả định cần thiết để trả lời câu hỏi này.
- Nếu một phụ huynh có hai con học tại một trường tiểu học ở Quận DeKalb, xác suất để cả hai con đều nghỉ ít nhất một ngày là bao nhiêu? Hãy nêu các giả định bạn đưa ra.
- Nếu bạn đã đưa ra giả định trong phần (d) hoặc (e), liệu giả định đó có hợp lý không? Nếu bạn không đưa ra giả định nào, hãy kiểm tra lại câu trả lời trước đó.

²³ Cục Điều tra Dân số Hoa Kỳ, Khảo sát Cộng đồng Hoa Kỳ năm 2010 - Ước tính trong 1 năm, Trình độ Học vấn Đạt được.

²⁴ S.S. Mizan và cộng sự. “Vắng mặt, vắng mặt kéo dài và đi học muộn nhiều lần liên quan đến tình trạng hen suyễn ở trẻ em tiểu học”. Xuất bản trong: Tạp chí Hen suyễn 48.3 (2011), trang 228–234.

3.2 Xác suất có điều kiện - Conditional probability

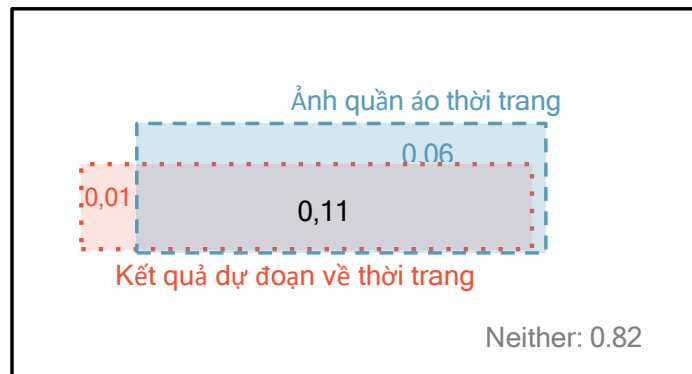
Có thể có mối quan hệ phức tạp giữa hai hoặc nhiều biến mà việc hiểu chúng rất hữu ích. Chẳng hạn, một công ty bảo hiểm ô tô sẽ xem xét thông tin về lịch sử lái xe của một người để đánh giá rủi ro mà họ sẽ gây ra tai nạn. Những mối quan hệ như vậy thuộc phạm vi của xác suất có điều kiện.

3.2.1 Khám phá xác suất bằng bảng đối chiếu

Bộ dữ liệu photo classify đại diện cho một bộ phân loại trên mẫu gồm 1822 bức ảnh từ một trang web chia sẻ ảnh. Các nhà khoa học dữ liệu đã và đang nỗ lực cải thiện một bộ phân loại nhằm xác định liệu một bức ảnh có liên quan đến thời trang hay không. 1822 bức ảnh này được sử dụng để kiểm tra hiệu suất của bộ phân loại đó. Mỗi bức ảnh nhận hai phân loại: Mach learn là kết quả phân loại từ hệ thống học máy (ML), với hai giá trị có thể có: pred fashion (dự đoán là thời trang) hoặc pred not (dự đoán không phải thời trang). Truth là nhãn thực tế, được gán bởi một nhóm chuyên gia, đóng vai trò như nguồn sự thật, với hai giá trị: fashion (thời trang) và not (không phải thời trang). Hình 3.11 tóm tắt kết quả phân loại này.

	Kết quả		
	Ảnh thời trang	Không phải	Tổng
Máy đoán			
Đoán là ảnh thời trang	197	22	219
Đoán không là ảnh thời trang	112	1491	1603
Tổng	309	1513	1822

Hình 3.11: Bảng đối chiếu tóm tắt bộ dữ liệu phân loại ảnh.



Hình 3.12: Biểu đồ Venn sử dụng hộp cho tập dữ liệu phân loại ảnh.

VÍ DỤ 3.26

Nếu một bức ảnh thực sự thuộc về thời trang, xác suất thuật toán máy học phân loại đúng bức ảnh đó là về thời trang là bao nhiêu?

Chúng ta có thể ước lượng xác suất này dựa trên dữ liệu. Trong tổng số 309 bức ảnh thời trang, thuật toán máy học đã phân loại chính xác 197 bức ảnh:

$$P(\text{Máy học dự đoán kết quả thời trang} | \text{Kết quả thời trang thực tế}) = \frac{197}{309} = 0,638$$

VÍ DỤ 3.27

Chúng ta lấy ngẫu nhiên một bức ảnh từ tập dữ liệu và biết rằng thuật toán máy học đã dự đoán bức ảnh này không thuộc về thời trang. Xác suất thuật toán dự đoán sai, tức là bức ảnh thực sự thuộc về thời trang, là bao nhiêu?

Nếu bộ phân loại ML dự đoán một bức ảnh không thuộc về thời trang, thì bức ảnh đó nằm trong hàng thứ hai của tập dữ liệu. Trong số 1603 bức ảnh này, có 112 bức ảnh thực sự thuộc về thời trang:

$$P(\text{Hình ảnh thực tế là thời trang} | \text{Kết quả máy dự đoán là không}) = \frac{112}{1603} = 0,070$$

3.2.2 Xác suất biên và xác suất kết hợp

Hình 3.11 bao gồm tổng hàng và tổng cột cho từng biến riêng biệt trong tập dữ liệu phân loại ảnh. Các tổng này đại diện cho **xác suất biên** của mẫu, tức là xác suất dựa trên một biến duy nhất mà không xét đến bất kỳ biến nào khác. Ví dụ, một xác suất chỉ dựa trên biến máy học là một xác suất biên:

$$P(\text{mach_learn dự đoán là pred_fashion}) = \frac{219}{1822} = 0,12$$

Xác suất của kết quả liên quan đến hai hoặc nhiều biến hoặc quá trình được gọi là **xác suất kết hợp**:

$$P(\text{mach_learn dự đoán là pred_fashion và kết quả là đúng}) = \frac{197}{1822} = 0,11$$

Thông thường, dấu phẩy được sử dụng thay cho từ “và” trong xác suất kết hợp, mặc dù cả hai cách viết đều có thể chấp nhận được:

$$P(\text{mach_learn dự đoán là pred_fashion, kết quả là đúng})$$

có cùng ý nghĩa với

$$P(\text{mach_learn dự đoán là pred_fashion và kết quả là đúng})$$

XÁC SUẤT BIÊN VÀ XÁC SUẤT KẾT HỢP

Nếu một xác suất được tính dựa trên một biến duy nhất, đó là *xác suất biên*. Xác suất của kết quả liên quan đến hai hoặc nhiều biến hoặc quá trình được gọi là *xác suất kết hợp*.

Chúng tôi sử dụng **bảng tỷ lệ** để tóm tắt xác suất kết hợp cho mẫu phân loại ảnh. Các tỷ lệ này được tính bằng cách chia từng số đếm trong Hình 3.11 cho tổng số của bảng, 1822, để thu được các tỷ lệ trong Hình 3.13. Phân phối xác suất kết hợp của các biến máy học và thực tế được hiển thị trong Hình 3.14.

	truth: fashion (Đoán đúng)	truth: not (Đoán sai)	Tổng
mach_learn: pred_fashion (Máy đoán ảnh là ảnh thời trang)	0,1081	0,0121	0,1202
mach_learn: pred_not (Máy đoán ảnh không phải ảnh thời trang)	0,0615	0,8183	0,8798
Tổng	0,1696	0,8304	1,00

Hình 3.13: Bảng xác suất tóm tắt tập dữ liệu phân loại ảnh.

Kết hợp kết quả	Xác suất
Máy đoán ảnh là ảnh thời trang và kết quả là đúng	0,1081
Máy đoán ảnh là ảnh thời trang và kết quả là sai	0,0121
Máy đoán ảnh không là ảnh thời trang và kết quả là đúng	0,0615
Máy đoán ảnh không là ảnh thời trang và kết quả là sai	0,8183
Tổng	1,0000

Hình 3.14: Phân phối xác suất kết hợp cho tập dữ liệu phân loại ảnh.

BÀI TẬP THỰC HÀNH 3.28

Xác minh rằng Hình 3.14 đại diện cho một phân phối xác suất: các sự kiện là rời rạc, tất cả các xác suất đều không âm, và tổng các xác suất bằng 1.²⁵

Chúng ta có thể tính xác suất biên từ xác suất kết hợp trong các trường hợp đơn giản. Ví dụ, xác suất để một bức ảnh được chọn ngẫu nhiên từ tập dữ liệu thuộc về thời trang có thể được tính bằng cách cộng các kết quả mà biến thực tế có giá trị là thời trang:

$$\begin{aligned}
 P(\text{Đoán là ảnh thời trang}) &= P(\text{Máy đoán là ảnh thời trang và kết quả là đúng}) + P(\text{máy đoán không là ảnh thời trang và kết quả là ảnh thời trang}) \\
 &= 0,1081 + 0,0615 \\
 &= 0,1696
 \end{aligned}$$

3.2.3 Xác định xác suất có điều kiện

Bộ phân loại máy học (ML classifier) dự đoán liệu một bức ảnh có liên quan đến thời trang hay không, mặc dù nó không hoàn hảo. Chúng ta muốn hiểu rõ hơn cách sử dụng thông tin từ một biến như mach learn để cải thiện ước lượng xác suất của một biến thứ hai, trong trường hợp này là truth.

Xác suất để một bức ảnh ngẫu nhiên từ tập dữ liệu thuộc về thời trang là khoảng 0.17. Nếu chúng ta biết bộ phân loại máy học dự đoán rằng bức ảnh thuộc về thời trang, liệu có thể có một ước lượng tốt hơn về xác suất thực sự của bức ảnh thuộc về thời trang không? Chắc chắn có. Để làm điều này, chúng ta chỉ xem xét 219 trường hợp mà bộ phân loại máy học dự đoán bức ảnh thuộc về thời trang và tính tỷ lệ trong số đó mà bức ảnh thực sự thuộc về thời trang:

$$P(\text{Kết quả là đúng} \mid \text{Máy đoán là ảnh thời trang}) = \frac{197}{219} = 0,900$$

Chúng ta gọi đây là **xác suất có điều kiện** vì nó được tính toán dựa trên một điều kiện: bộ phân loại máy học đã dự đoán rằng bức ảnh thuộc về thời trang.

Xác suất có điều kiện bao gồm hai phần: **kết quả quan tâm** và **điều kiện**. Có thể hiểu điều kiện như là thông tin mà chúng ta biết chắc chắn là đúng, và thông tin này thường được mô tả như một kết quả hoặc sự kiện đã biết. Thông thường, trong ký hiệu xác suất, chúng ta sử dụng dấu gạch dọc (|) để phân tách kết quả quan tâm và điều kiện, như sau:

$$\begin{aligned}
 &P(\text{Bức ảnh thực sự là thời trang, với điều kiện máy học dự đoán nó là thời trang}) \\
 &= P(\text{Bức ảnh thực sự là thời trang} \mid \text{máy dự đoán nó là thời trang}) = \frac{197}{219} = 0,900
 \end{aligned}$$

Dấu gạch dọc “|” được đọc là *với điều kiện*.

²⁵ Mỗi trong bốn tổ hợp kết quả đều rời rạc (không trùng lặp), tất cả các xác suất đều không âm, và tổng các xác suất là: 0,1081 + 0,0121 + 0,0615 + 0,8183 = 1,00.

Trong phương trình cuối cùng, chúng ta đã tính xác suất một bức ảnh thuộc về thời trang, với điều kiện thuật toán máy học đã dự đoán bức ảnh đó thuộc về thời trang, dưới dạng một phân số:

$$\begin{aligned} & P(\text{Bức ảnh thực sự là thời trang} \mid \text{máy dự đoán nó là thời trang}) \\ &= \frac{\# \text{ trường hợp kết quả đúng và máy đoán là ảnh thời trang}}{\# \text{ trường hợp máy đoán là ảnh thời trang}} \\ &= \frac{197}{219} = 0,900 \end{aligned}$$

Chúng ta chỉ xem xét các trường hợp thỏa mãn điều kiện máy dự đoán ảnh thuộc về thời trang, sau đó tính tỷ lệ trong số đó mà bức ảnh thực sự thuộc về thời trang.

Thông thường, xác suất biên và xác suất kết hợp được cung cấp thay vì dữ liệu dạng đếm. Ví dụ, tỷ lệ mắc bệnh thường được biểu thị dưới dạng phần trăm thay vì số lượng ca mắc. Chúng ta muốn có thể tính toán xác suất có điều kiện ngay cả khi không có dữ liệu dạng đếm, và phương trình cuối cùng đóng vai trò là mẫu để hiểu kỹ thuật này.

Chúng ta đã chỉ xem xét những trường hợp thỏa mãn điều kiện, tức là thuật toán máy học dự đoán bức ảnh thuộc về thời trang. Trong số những trường hợp này, xác suất có điều kiện chính là tỷ lệ phản ánh kết quả quan tâm, tức là bức ảnh thực sự thuộc về thời trang. Giả sử chúng ta chỉ có thông tin từ Hình 3.13, tức là chỉ có dữ liệu xác suất. Khi đó, nếu lấy mẫu 1000 bức ảnh, ta kỳ vọng khoảng 12.0%, tức là: $0.120 \times 1000 = 120$ bức ảnh sẽ được dự đoán là thời trang (`mach_learn` is `pred_fashion`). Tương tự, ta kỳ vọng khoảng 10.8%, tức là: $0.108 \times 1000 = 108$ bức ảnh vừa thỏa mãn điều kiện dự đoán của máy học vừa thực sự thuộc về thời trang. Khi đó, xác suất có điều kiện có thể được tính toán như sau:

$$\begin{aligned} & P(\text{Bức ảnh thực sự là thời trang} \mid \text{máy dự đoán nó là thời trang}) \\ &= \frac{\# \text{ trường hợp kết quả đúng và máy đoán là ảnh thời trang}}{\# \text{ trường hợp máy đoán là ảnh thời trang}} \\ &= \frac{108}{120} = \frac{0,108}{0,120} = 0,90 \end{aligned}$$

Ở đây, chúng ta đang xem xét chính xác tỷ lệ giữa hai xác suất, 0.108 và 0.120, có thể viết dưới dạng:

$P(\text{Bức ảnh là thời trang và máy dự đoán ảnh là ảnh thời trang})$ và $P(\text{máy dự đoán ảnh là thời trang})$

Tỷ lệ giữa hai xác suất này là một ví dụ về công thức tổng quát của xác suất có điều kiện.

XÁC SUẤT CÓ ĐIỀU KIỆN

Xác suất có điều kiện của kết quả A với điều kiện B được tính theo công thức sau:

$$P(A|B) = \frac{P(A \text{ và } B)}{P(B)}$$

BÀI TẬP THỰC HÀNH 3.29

(a) Viết lại phát biểu sau bằng ký hiệu xác suất có điều kiện: “Xác suất dự đoán của máy học là đúng, nếu bức ảnh thực sự thuộc về thời trang.” Ở đây, điều kiện được xác định dựa trên trạng thái thực tế của bức ảnh, không phải theo thuật toán máy học.

(b) Xác định xác suất từ phần (a). Bảng 3.13 ở trang 96 có thể hữu ích cho việc tính toán này.²⁶

²⁶ (a) Nếu bức ảnh thực sự thuộc về thời trang và thuật toán máy học dự đoán đúng, thì giá trị dự đoán của thuật toán máy học sẽ là ảnh thời trang: $P(\text{Máy đoán là ảnh thời trang} \mid \text{Kết quả đúng})$

(b) Công thức xác suất có điều kiện cho thấy chúng ta cần xác định trước:

$P(\text{Máy đoán là ảnh thời trang và kết quả đúng}) = 0,1081$ và $P(\text{kết quả đúng}) = 0,1696$. Sau đó, tỷ lệ giữa hai xác suất này chính là xác suất có điều kiện: $0,1081/0,1696 = 0,6374$

BÀI TẬP THỰC HÀNH 3.30

- (a) Xác định xác suất thuật toán dự đoán sai nếu biết rằng bức ảnh là về thời trang.
 (b) Sử dụng kết quả từ phần (a) và BÀI TẬP THỰC HÀNH 3.29 (b), tính:
 $P(\text{Máy đoán là ảnh thời trang} \mid \text{Kết quả đúng}) + P(\text{Máy đoán sai ảnh thời trang} \mid \text{Kết quả đúng})$
 (c) Đưa ra một lập luận trực quan để giải thích tại sao tổng trong phần (b) bằng 1.²⁷

3.2.4 Đậu mùa ở Boston, năm 1721

Bộ dữ liệu về bệnh đậu mùa cung cấp một mẫu gồm 6.224 cá nhân từ năm 1721, những người đã bị phơi nhiễm với bệnh đậu mùa ở Boston. Các bác sĩ thời đó tin rằng tiêm chủng – tức là việc đưa mầm bệnh vào cơ thể theo một cách có kiểm soát – có thể làm giảm nguy cơ tử vong.

Mỗi trường hợp trong dữ liệu đại diện cho một cá nhân với hai biến số: tiêm chủng (inoculated) và kết quả (result). Biến tiêm chủng có hai mức: có hoặc không, cho biết liệu người đó có được tiêm chủng hay không. Biến kết quả có hai kết cục: sống sót hoặc tử vong. Các dữ liệu này được tổng hợp trong Bảng 3.15 và Bảng 3.16.

		Tiêm chủng		Tổng
		Có	Không	
Kết cục	Sống sót	238	5136	5374
	Tử vong	6	844	850
	Tổng	244	5980	6224

Hình 3.15: Bảng đối chiếu cho bộ dữ liệu về bệnh đậu mùa.

		Tiêm chủng		Tổng
		Có	Không	
Kết cục	Sống sót	0,0382	0,8252	0,8634
	Tử vong	0,0010	0,1356	0,1366
	Tổng	0,0392	0,9608	1,0000

Hình 3.16: Bảng tỷ lệ cho dữ liệu về bệnh đậu mùa, được tính bằng cách chia từng số liệu cho tổng số mẫu, 6224.

BÀI TẬP THỰC HÀNH 3.31

Viết xác suất dưới dạng ký hiệu chính thức cho xác suất một người được chọn ngẫu nhiên, không được tiêm chủng, tử vong do bệnh đậu mùa, và tính xác suất này.²⁸

BÀI TẬP THỰC HÀNH 3.32

Xác định xác suất một người đã được tiêm chủng tử vong do bệnh đậu mùa. Kết quả này so sánh như thế nào với kết quả của BÀI TẬP THỰC HÀNH 3.31?²⁹

²⁷ (a) (a) Xác suất này là: $\frac{P(\text{Máy đoán là sai nhưng kết quả là đúng})}{P(\text{Đoán kết quả đúng})} = \frac{0,0615}{0,1696} = 0,3626$. (b) Tổng bằng 1. (c) Với điều kiện bức ảnh thực sự là về thời trang, thuật toán của máy tính hoặc đã dự đoán ảnh là về thời trang hoặc đã dự đoán ảnh không phải về thời trang. Nguyên tắc bổ sung vẫn áp dụng cho xác suất có điều kiện, miễn là tất cả các xác suất được tính dựa trên cùng một thông tin điều kiện.

²⁸ $P(\text{Tử vong} \mid \text{Không tiêm chủng}) = P(\text{Tử vong và không tiêm chủng}) = \frac{0,1356}{0,9584} = 0,1411$.

²⁹ $P(\text{Tử vong} \mid \text{Có tiêm chủng}) = P(\text{Tử vong và có tiêm chủng}) = \frac{0,0010}{0,0392} = 0,0255$. (nếu tránh làm tròn, ta sẽ có $6/244 = 0,0246$). Tỷ lệ tử vong của những người đã được tiêm chủng chỉ khoảng 1 trên 40, trong khi tỷ lệ tử vong của những người không được tiêm chủng là khoảng 1 trên 7.

BÀI TẬP THỰC HÀNH 3.33

Người dân Boston tự quyết định có tiêm chủng hay không.

- (a) Đây là một nghiên cứu quan sát hay một thí nghiệm?
 (b) Chúng ta có thể suy luận bất kỳ mối quan hệ nhân quả nào từ dữ liệu này không?
 (c) Những biến nhiễu tiềm ẩn nào có thể ảnh hưởng đến việc một người sống hay chết và đồng thời ảnh hưởng đến việc người đó có được tiêm chủng hay không? ³⁰

3.2.5 Quy tắc nhân tổng quát - General multiplication rule

Mục 3.1.7 đã giới thiệu Quy tắc nhân đối với các quá trình độc lập. Ở đây, chúng tôi cung cấp **Quy tắc nhân tổng quát** cho các sự kiện có thể không độc lập.

QUY TẮC NHÂN TỔNG QUÁT

Nếu A và B đại diện cho hai kết quả hoặc sự kiện, thì:

$$P(A \text{ và } B) = P(A | B) \times P(B)$$

Quy Tắc Nhân Tổng Quát này thực chất chỉ là một sự sắp xếp lại của phương trình xác suất có điều kiện.

VÍ DỤ 3.34

Xét bộ dữ liệu bệnh đậu mùa. Giả sử chúng ta chỉ có hai thông tin sau: 96,08% cư dân không được tiêm chủng, và 85,88% cư dân không được tiêm chủng đã sống sót. Làm thế nào chúng ta có thể tính xác suất một cư dân không được tiêm chủng và sống sót?

Chúng ta sẽ tính toán kết quả bằng Quy tắc Nhân Tổng Quát, sau đó xác minh bằng Hình 3.16. Chúng ta cần xác định:

$$P(\text{Sống sót và không tiêm chủng})$$

và chúng ta có

$$P(\text{Sống sót} | \text{không tiêm chủng}) = 0,8588, P(\text{Không tiêm chủng}) = 0,9608$$

Trong số 96,08% số người không được tiêm chủng, 85,88% sống sót:

$$P(\text{Sống sót và không tiêm chủng}) = 0,8588 \times 0,9608 = 0,82$$

Điều này tương đương với Quy tắc Nhân Tổng Quát. Chúng ta có thể xác nhận xác suất này trong Hình 3.16 tại giao điểm của "không tiêm chủng" và "sống sót" (có một sai số làm tròn nhỏ).

BÀI TẬP THỰC HÀNH 3.35

Sử dụng $P(\text{có tiêm chủng}) = 0,0392$ và $P(\text{sống sót} | \text{có tiêm chủng}) = 0,9754$ để xác định xác suất một người vừa tiêm chủng vừa sống sót. ³¹

BÀI TẬP THỰC HÀNH 3.36

Nếu 97.54% số người được tiêm chủng sống sót, thì tỷ lệ số người được tiêm chủng đã tử vong là bao nhiêu? ³²

³⁰ Câu trả lời ngắn gọn: (a) Nghiên cứu quan sát. (b) Không, chúng ta không thể suy ra mối quan hệ nhân quả từ nghiên cứu quan sát này. (c) Khả năng tiếp cận với dịch vụ y tế tiên tiến và tốt nhất. Ngoài ra, còn có các câu trả lời hợp lý khác cho phần (c).

³¹ Đáp án là 0.0382, có thể được xác minh bằng Hình 3.16.

³² Chỉ có hai kết quả có thể xảy ra: sống sót hoặc tử vong. Điều này có nghĩa là $100\% - 97.54\% = 2.46\%$ số người được tiêm chủng đã tử vong.

TỔNG CÁC XÁC SUẤT CÓ ĐIỀU KIỆN

Giả sử $A_1, \dots, A_k$ là tất cả các kết quả rời rạc của một biến hoặc một quá trình. Khi đó, nếu B là một sự kiện thì:

$$P(A_1|B) + \dots + P(A_k|B) = 1$$

Quy tắc bù cũng được áp dụng khi một sự kiện và phần bù của nó được điều kiện hóa trên cùng một thông tin:

$$P(A|B) = 1 - P(A^c|B)$$

BÀI TẬP THỰC HÀNH 3.37

- G Dựa trên các xác suất đã được tính toán ở trên, có vẻ như việc tiêm chủng có hiệu quả trong việc giảm nguy cơ tử vong do bệnh đậu mùa hay không? ³³

3.2.6 Xem xét tính độc lập trong xác suất có điều kiện

Nếu hai sự kiện là độc lập, thì việc biết kết quả của một sự kiện sẽ không cung cấp thông tin về sự kiện còn lại. Chúng ta có thể chứng minh điều này một cách toán học bằng cách sử dụng xác suất có điều kiện.

BÀI TẬP THỰC HÀNH 3.38

Gọi X và Y là kết quả của hai con xúc xắc. ³⁴

- G (a) Xác suất để con xúc xắc thứ nhất, X , bằng 1 là bao nhiêu?
 (b) Xác suất để cả hai xúc xắc, X và Y , đều bằng 1 là bao nhiêu?
 (c) Sử dụng công thức xác suất có điều kiện để tính $P(Y=1|X=1)$.
 (d) Xác suất $P(Y=1)$ là bao nhiêu? Kết quả này có khác với kết quả ở phần (c) không? Giải thích.

Chúng ta có thể chứng minh trong BÀI TẬP THỰC HÀNH 3.38(c) rằng thông tin điều kiện không có ảnh hưởng bằng cách sử dụng Quy tắc Nhân cho các quá trình độc lập:

$$\begin{aligned} P(Y=1|X=1) &= \frac{P(Y=1 \text{ và } X=1)}{P(X=1)} \\ &= \frac{P(Y=1) \times P(X=1)}{P(X=1)} \\ &= P(Y=1) \end{aligned}$$

BÀI TẬP THỰC HÀNH 3.39

G Ron đang quan sát một bàn roulette trong sòng bạc và nhận thấy rằng năm kết quả gần nhất đều là màu đen. Anh ta cho rằng khả năng ra màu đen sáu lần liên tiếp là rất nhỏ (khoảng 1/64) và đặt toàn bộ tiền lương của mình vào màu đỏ. Sai lầm trong suy luận của anh ta là gì? ³⁵

³⁰ Các mẫu dữ liệu có kích thước lớn so với sự khác biệt về tỷ lệ tử vong giữa nhóm "được tiêm chủng" và nhóm "không được tiêm chủng", do đó có vẻ như có mối liên hệ giữa việc tiêm chủng và kết quả. Tuy nhiên, như đã đề cập trong lời giải của BÀI TẬP THỰC HÀNH 3.33, đây là một nghiên cứu quan sát và chúng ta không thể chắc chắn liệu có mối quan hệ nhân quả hay không. (Các nghiên cứu sau này đã chứng minh rằng việc tiêm chủng có hiệu quả trong việc giảm tỷ lệ tử vong.)

³⁴ Giải pháp ngắn gọn: (a) Xác suất để xúc xắc thứ nhất ra 1 là 1/6. (b) Xác suất để cả X và Y đều ra 1 là 1/36. (c)

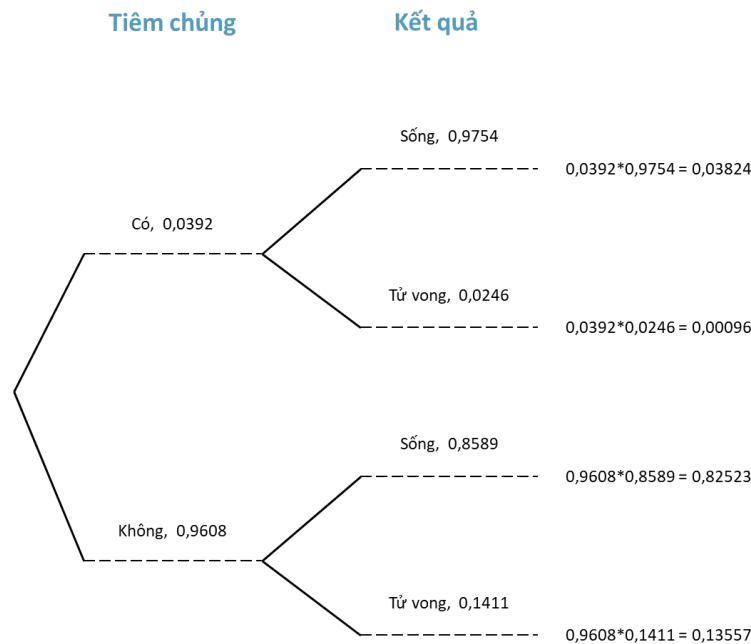
$\frac{P(Y=1 \text{ và } X=1)}{P(X=1)} = \frac{1/36}{1/6} = 1/6$. (d) Xác suất $P(Y=1)$ cũng bằng 1/6, giống như trong phần (c). Điều này có nghĩa là biết X không ảnh hưởng đến xác suất của Y , điều này hợp lý vì X và Y là độc lập.

³⁵ Ron đã quên rằng mỗi lần quay roulette là độc lập với các lần quay trước đó. Sòng bạc thường hiển thị kết quả của các lần quay gần nhất để đánh lừa người chơi rằng họ có thể dự đoán kết quả tiếp theo. Đây được gọi là **ngụy biện của người chơi (gambler's fallacy)**—một sai lầm khi tin rằng các sự kiện trước đó ảnh hưởng đến xác suất của một sự kiện ngẫu nhiên sắp tới.

3.2.7 Tree diagrams

Sơ đồ cây là một công cụ để tổ chức các kết quả và xác suất dựa trên cấu trúc của dữ liệu. Chúng đặc biệt hữu ích khi có hai hoặc nhiều quá trình xảy ra theo trình tự, và mỗi quá trình phụ thuộc vào các quá trình trước đó.

Dữ liệu về bệnh đậu mùa phù hợp với mô tả này. Chúng ta có thể thấy rằng dân số được chia thành hai nhóm theo tình trạng tiêm chủng: có tiêm chủng và không tiêm chủng. Sau sự phân chia này, tỷ lệ sống sót được quan sát cho từng nhóm. Cấu trúc này được thể hiện trong **sơ đồ cây** ở Hình 3.17. Nhánh đầu tiên liên quan đến tình trạng tiêm chủng được gọi là **nhánh chính**, trong khi các nhánh còn lại là **nhánh phụ**.



Hình 3.17: Sơ đồ cây của tập dữ liệu bệnh đậu mùa.

Sơ đồ cây được chú thích với các xác suất biên và xác suất có điều kiện, như thể hiện trong Hình 3.17. Sơ đồ cây này chia dữ liệu bệnh đậu mùa theo tình trạng tiêm chủng thành hai nhóm: có tiêm chủng và không tiêm chủng, với các xác suất biên lần lượt là 0.0392 và 0.9608. Các nhánh thứ cấp được điều kiện hóa dựa trên nhánh đầu tiên, vì vậy chúng được gán các xác suất có điều kiện. Ví dụ, nhánh trên cùng trong Hình 3.17 thể hiện xác suất kết quả = sống sót với điều kiện đã tiêm chủng. Chúng ta có thể (và thường làm) tính toán xác suất kết hợp tại cuối mỗi nhánh trong sơ đồ cây bằng cách nhân các giá trị gặp phải khi di chuyển từ trái sang phải. Các xác suất kết hợp này được tính toán bằng Quy tắc Nhân tổng quát.

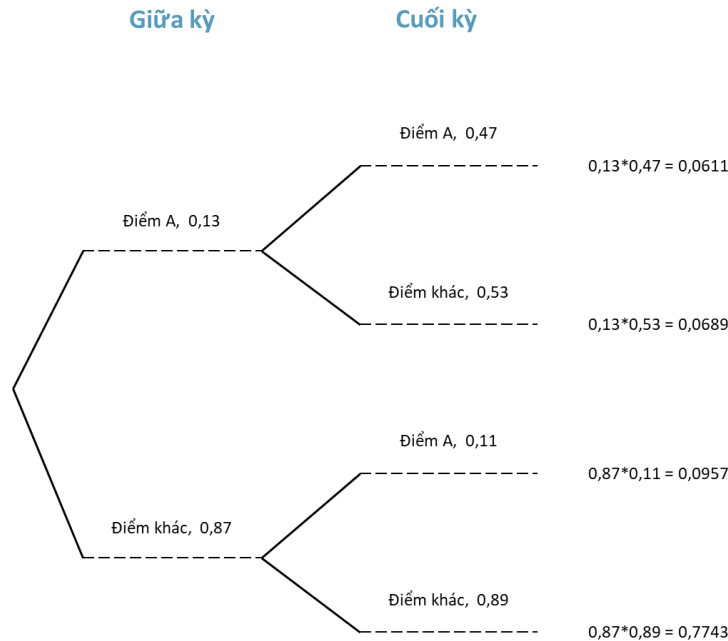
$$\begin{aligned}
 P(\text{Có tiêm chủng và sống sót}) &= P(\text{Có tiêm chủng}) \times P(\text{Sống sót} \mid \text{Có tiêm chủng}) \\
 &= 0,0392 \times 0,9754 = 0,0382
 \end{aligned}$$

VÍ DỤ 3.40

Xét bài kiểm tra giữa kỳ và cuối kỳ của một lớp học thống kê. Giả sử rằng 13% sinh viên đạt điểm A trong bài kiểm tra giữa kỳ. Trong số những sinh viên đạt A ở giữa kỳ, 47% tiếp tục nhận được A ở bài kiểm tra cuối kỳ. Ngoài ra, 11% sinh viên không đạt A ở giữa kỳ vẫn nhận được A ở cuối kỳ. Bạn chọn ngẫu nhiên một bài kiểm tra cuối kỳ và thấy rằng sinh viên này đạt A. Hỏi xác suất để sinh viên này cũng đã đạt A ở bài kiểm tra giữa kỳ là bao nhiêu??

Mục tiêu cuối cùng là tìm $P(\text{điểm giữa kỳ} = A \mid \text{điểm cuối kỳ} = A)$. Để tính xác suất có điều kiện này, chúng ta cần các xác suất sau: $P(\text{điểm giữa kỳ} = A \text{ và điểm cuối kỳ} = A)$ và $P(\text{điểm cuối kỳ} = A)$

Tuy nhiên, những thông tin này không được cung cấp trực tiếp và không rõ ràng cách tính chúng. Vì chưa chắc chắn về cách tiếp cận, việc tổ chức thông tin thành một sơ đồ cây sẽ hữu ích:



Khi xây dựng một sơ đồ cây, các biến có xác suất biên thường được sử dụng để tạo các nhánh chính của cây; trong trường hợp này, xác suất biên được cung cấp cho điểm thi giữa kỳ. Điểm thi cuối kỳ, tương ứng với các xác suất có điều kiện được cung cấp, sẽ được hiển thị trên các nhánh phụ.

Sau khi sơ đồ cây được xây dựng, chúng ta có thể tính các xác suất cần thiết:

$$P(\text{Điểm giữa kỳ} = A \text{ và Điểm cuối kỳ} = A) = 0,0611$$

$$\begin{aligned} P(\text{Điểm cuối kỳ} = A) &= P(\text{Điểm giữa kỳ} = \text{khác và điểm cuối kỳ} = A) + \\ &\quad P(\text{Điểm giữa kỳ} = A \text{ và Điểm cuối kỳ} = A) \\ &= 0,0957 + 0,0611 = 0,1568 \end{aligned}$$

Xác suất biên, $P(\text{Điểm cuối kỳ} = A)$, được tính bằng cách cộng tất cả các xác suất kết hợp ở phía bên phải của cây, tương ứng với Điểm cuối kỳ = A. Cuối cùng, chúng ta có thể lấy tỷ lệ của hai xác suất này:

$$\begin{aligned} P(\text{Điểm giữa kỳ} = A \text{ và Điểm cuối kỳ} = A) &= \frac{P(\text{Điểm giữa kỳ} = A \text{ và Điểm cuối kỳ} = A)}{P(\text{Điểm cuối kỳ} = A)} \\ &= \frac{0,0611}{0,1568} = 0,3897 \end{aligned}$$

Xác suất sinh viên đó cũng đạt điểm A trong kỳ thi giữa kỳ là khoảng 0.39.

BÀI TẬP THỰC HÀNH 3.39

G

Sau một khóa học thống kê nhập môn, 78% sinh viên có thể vẽ được sơ đồ cây. Trong số những sinh viên có thể vẽ sơ đồ cây, 97% đã vượt qua (đạt yêu cầu), trong khi chỉ có 57% sinh viên không thể vẽ sơ đồ cây đạt yêu cầu. (a) Tổ chức thông tin này thành một sơ đồ cây. (b) Xác suất một sinh viên được chọn ngẫu nhiên đạt yêu cầu là bao nhiêu? (c) Tính xác suất một sinh viên có thể vẽ sơ đồ cây nếu biết rằng sinh viên đó đã đạt yêu cầu.³⁶

3.2.8 Định lý Bayes - Bayes' Theorem

Trong nhiều trường hợp, chúng ta được cung cấp xác suất có điều kiện dưới dạng:

$P(\text{mệnh đề về biến 1} | \text{mệnh đề về biến 2})$, nhưng điều chúng ta thực sự muốn biết là xác suất có điều kiện đảo ngược: $P(\text{mệnh đề về biến 2} | \text{mệnh đề về biến 1})$.

Sơ đồ cây có thể được sử dụng để tìm xác suất có điều kiện thứ hai khi đã biết xác suất có điều kiện thứ nhất. Tuy nhiên, đôi khi không thể biểu diễn kịch bản bằng sơ đồ cây. Trong những trường hợp như vậy, chúng ta có thể áp dụng một công thức rất hữu ích và tổng quát: Định lý Bayes.

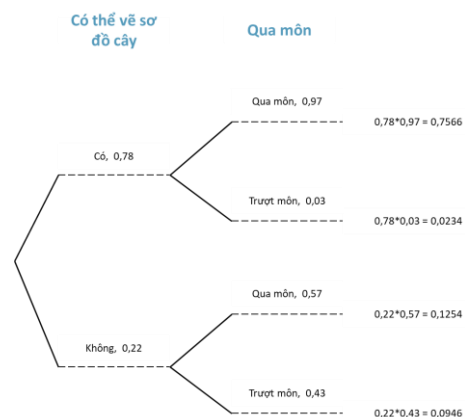
Trước tiên, chúng ta sẽ xem xét kỹ một ví dụ về việc đảo ngược xác suất có điều kiện trong đó chúng ta vẫn sử dụng sơ đồ cây.

Trước tiên, chúng ta sẽ xem xét kỹ một ví dụ về việc đảo ngược xác suất có điều kiện trong đó chúng ta vẫn sử dụng sơ đồ cây.

³⁶ a) Sơ đồ cây được hiển thị ở bên phải.

(b) Xác định hai xác suất kết hợp biểu diễn các sinh viên đã đạt yêu cầu, rồi cộng chúng lại: $P(\text{đạt yêu cầu}) = 0,7566 + 0,1254 = 0,8820$.

(c) $P(\text{vẽ được sơ đồ cây} | \text{đạt yêu cầu}) = \frac{0,7566}{0,8820} = 0,8578$.



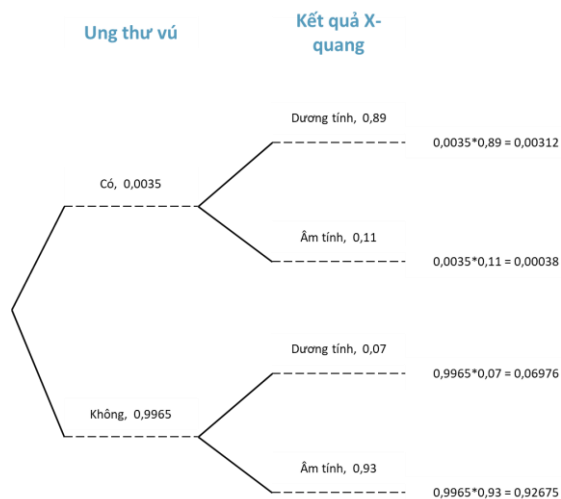
VÍ DỤ 3.42

Ở Canada, khoảng 0,35% phụ nữ trên 40 tuổi sẽ mắc ung thư vú trong bất kỳ năm nào. Một xét nghiệm tầm soát ung thư phổ biến là chụp X-quang tuyến vú (mammogram), nhưng xét nghiệm này không hoàn hảo. Khoảng 11% bệnh nhân mắc ung thư vú nhận kết quả **âm tính giả**: xét nghiệm cho thấy họ không mắc ung thư trong khi thực tế họ có bệnh. Tương tự, xét nghiệm cho kết quả **dương tính giả** ở 7% bệnh nhân không mắc ung thư: xét nghiệm cho thấy họ mắc ung thư trong khi thực tế họ không có bệnh. Nếu chúng ta kiểm tra ngẫu nhiên một phụ nữ trên 40 tuổi bằng phương pháp chụp X-quang tuyến vú và xét nghiệm cho kết quả dương tính – nghĩa là xét nghiệm cho thấy bệnh nhân có ung thư – thì xác suất thực sự bệnh nhân mắc ung thư là bao nhiêu?

Lưu ý rằng chúng ta được cung cấp đủ thông tin để nhanh chóng tính xác suất xét nghiệm cho kết quả dương tính nếu một phụ nữ mắc ung thư vú ($1,00 - 0,11 = 0,89$). Tuy nhiên, điều chúng ta cần tìm là xác suất đảo ngược: xác suất một người mắc ung thư khi đã có kết quả xét nghiệm dương tính. (Hãy cẩn thận với ngôn ngữ y khoa không trực quan: một kết quả xét nghiệm "dương tính" trong tầm soát bằng chụp X-quang tuyến vú cho thấy khả năng có ung thư.) Xác suất đảo ngược này có thể được chia thành hai phần:

$$P(\text{mắc ung thư vú} \mid \text{kết quả X-quang dương tính}) = \frac{P(\text{mắc ung thư vú và kết quả X-quang dương tính})}{P(\text{kết quả X-quang dương tính})}$$

Chúng ta có thể xây dựng một sơ đồ cây cho các xác suất này:



Xác suất bệnh nhân mắc ung thư vú và có kết quả chụp X-quang dương tính là:

$$\begin{aligned} P(\text{Mắc ung thư vú} \mid \text{Kết quả X-quang dương tính}) &= \frac{P(\text{Kết quả X-quang dương tính} \mid \text{Mắc ung thư vú}) \times P(\text{Mắc ung thư vú})}{P(\text{Kết quả X-quang dương tính})} \\ &= 0,89 \times 0,0035 = 0,00312 \end{aligned}$$

Xác suất có kết quả xét nghiệm dương tính là tổng của hai trường hợp tương ứng:

$$\begin{aligned} P(\text{Kết quả X-quang dương tính}) &= P(\text{Kết quả X-quang dương tính và Mắc ung thư vú}) + P(\text{Kết quả X-quang dương tính và Không mắc ung thư vú}) \\ &= P(\text{Mắc ung thư vú}) \times P(\text{Kết quả X-quang dương tính} \mid \text{Mắc ung thư vú}) + P(\text{Không mắc ung thư vú}) \times P(\text{Kết quả X-quang dương tính} \mid \text{Không mắc ung thư vú}) \\ &= 0,0035 \times 0,89 + 0,9965 \times 0,07 = 0,07288 \end{aligned}$$

Do đó, nếu một bệnh nhân có kết quả chụp X-quang dương tính, xác suất bệnh nhân thực sự mắc ung thư vú là:

$$\begin{aligned} P(\text{Mắc ung thư vú} \mid \text{Kết quả X-quang dương tính}) &= \frac{P(\text{mắc ung thư vú và kết quả X-quang dương tính})}{P(\text{kết quả X-quang dương tính})} \\ &= \frac{0,00312}{0,07288} \approx 0,0428 \end{aligned}$$

Nói cách khác, ngay cả khi một bệnh nhân có kết quả chụp X-quang tuyến vú dương tính, xác suất thực sự mắc ung thư vú chỉ khoảng 4%.

Ví dụ 3.42 nêu bật lý do tại sao các bác sĩ thường thực hiện thêm nhiều xét nghiệm bất kể kết quả dương tính ban đầu. Khi một tình trạng y tế hiếm gặp, một kết quả xét nghiệm dương tính duy nhất thường không mang tính quyết định.

Xét lại phương trình cuối cùng trong Ví dụ 3.42. Sử dụng sơ đồ cây, ta có thể thấy rằng tử số (phần trên của phân số) bằng tích sau:

$$P(\text{Mắc ung thư vú và Kết quả } X - \text{quang dương tính}) = P(\text{Kết quả } X - \text{quang dương tính} | \text{Mắc ung thư vú}) \times P(\text{Mắc ung thư vú})$$

Mẫu số – xác suất có kết quả sàng lọc dương tính – bằng tổng các xác suất cho từng trường hợp sàng lọc dương tính:

$$P(\text{Kết quả } X - \text{quang dương tính}) = P(\text{Kết quả } X - \text{quang dương tính và Không mắc ung thư vú}) + P(\text{Kết quả } X - \text{quang dương tính và Mắc ung thư vú})$$

Trong ví dụ này, từng xác suất ở vế phải được phân tích thành tích của một xác suất có điều kiện và một xác suất biên bằng cách sử dụng sơ đồ cây.

$$\begin{aligned} P(\text{Kết quả } X - \text{quang dương tính}) &= P(\text{Kết quả } X - \text{quang dương tính và Không mắc ung thư vú}) + \\ &\quad P(\text{Kết quả } X - \text{quang dương tính và Mắc ung thư vú}) \\ &= P(\text{Kết quả } X - \text{quang dương tính} | \text{Không mắc ung thư vú}) \times P(\text{Không mắc ung thư vú}) + \\ &\quad P(\text{Kết quả } X - \text{quang dương tính} | \text{Mắc ung thư vú}) \times P(\text{Mắc ung thư vú}) \end{aligned}$$

Ta có thể thấy ứng dụng của Định lý Bayes bằng cách thay thế các biểu thức xác suất thu được vào tử số và mẫu số của xác suất có điều kiện ban đầu.

$$\begin{aligned} &P(\text{Mắc ung thư vú} | \text{Kết quả } X - \text{quang dương tính}) \\ &= \frac{P(\text{Kết quả } X - \text{quang dương tính} | \text{Mắc ung thư vú}) \times P(\text{Mắc ung thư vú})}{P(\text{Kết quả } X - \text{quang dương tính} | \text{Không mắc ung thư vú}) \times P(\text{Không mắc ung thư vú}) + P(\text{Kết quả } X - \text{quang dương tính} | \text{Mắc ung thư vú}) \times P(\text{Mắc ung thư vú})} \end{aligned}$$

ĐỊNH LÝ BAYES: ĐẢO NGƯỢC XÁC SUẤT - INVERTING PROBABILITIES

Xét xác suất có điều kiện sau đây cho biến 1 và biến 2:

$$P(\text{Kết quả } A_1 \text{ của biến 1} | \text{Kết quả } B \text{ của biến 2})$$

Định lý Bayes phát biểu rằng xác suất có điều kiện này có thể được xác định dưới dạng phân số sau:

$$\frac{P(B|A_1) \times P(A_1)}{P(B|A_1) \times P(A_1) + P(B|A_2) \times P(A_2) + \dots + P(B|A_k) \times P(A_k)}$$

Định lý Bayes là một sự tổng quát hóa những gì chúng ta đã thực hiện bằng cách sử dụng sơ đồ cây. Tử số xác định xác suất của việc xảy ra cả A_1 và B . Mẫu số là xác suất biên của việc xảy ra B . Thành phần dưới cùng của phân số này trông có vẻ dài và phức tạp vì chúng ta phải cộng các xác suất từ tất cả các cách khác nhau dẫn đến B . Khi sử dụng sơ đồ cây, chúng ta luôn thực hiện bước này. Tuy nhiên, thường thì chúng ta làm nó như một bước riêng lẻ, nên nó không có vẻ quá phức tạp.

Để áp dụng đúng Định lý Bayes, có hai bước chuẩn bị:

- (1) Trước tiên, xác định xác suất biên của từng kết quả có thể có của biến thứ nhất: $P(A_1)$, $P(A_2)$, ..., $P(A_k)$.
- (2) Sau đó, xác định xác suất của kết quả B , có điều kiện theo từng kịch bản có thể có của biến thứ nhất: $P(B|A_1)$, $P(B|A_2)$, ..., $P(B|A_k)$.

Sau khi xác định được các xác suất này, chúng ta có thể được áp dụng trực tiếp vào công thức. Định lý Bayes thường là một lựa chọn tốt khi có quá nhiều kịch bản khiến việc vẽ sơ đồ cây trở nên phức tạp.

BÀI TẬP THỰC HÀNH 3.43

G Jose đến khuôn viên trường vào mỗi tối thứ Năm. Tuy nhiên, một số ngày bãi đậu xe bị đầy, thường là do các sự kiện của trường đại học. Các sự kiện học thuật diễn ra vào 35% số buổi tối, sự kiện thể thao diễn ra vào 20% số buổi tối, và không có sự kiện nào vào 45% số buổi tối. Khi có sự kiện học thuật, bãi đậu xe đầy khoảng 25% thời gian. Khi có sự kiện thể thao, bãi đậu xe đầy 70% thời gian. Khi không có sự kiện nào, bãi đậu xe chỉ đầy khoảng 5% thời gian. Nếu Jose đến trường và thấy bãi đậu xe đầy, xác suất để có một sự kiện thể thao đang diễn ra là bao nhiêu? Hãy sử dụng sơ đồ cây để giải bài toán này.³⁷

VÍ DỤ 3.44

Ở đây, chúng ta giải bài toán tương tự đã được trình bày trong BÀI TẬP THỰC HÀNH 3.43, nhưng lần này chúng ta sử dụng Định lý Bayes.

Kết quả mà chúng ta quan tâm là liệu có một sự kiện thể thao đang diễn ra hay không (gọi là A_1), với điều kiện rằng bãi đậu xe bị đầy (B). Gọi A_2 là sự kiện học thuật và A_3 là trường hợp không có sự kiện nào diễn ra trong khuôn viên trường. Khi đó, các xác suất đã cho có thể được viết như sau:

$$P(A_1) = 0,2$$

$$P(A_2) = 0,35$$

$$P(A_3) = 0,45$$

$$P(B|A_1) = 0,7$$

$$P(B|A_2) = 0,25$$

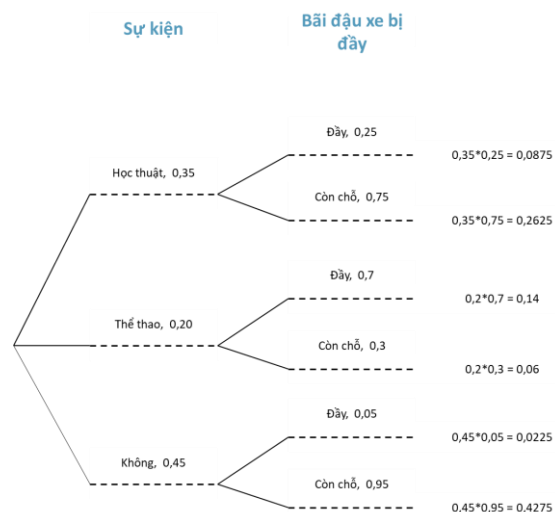
$$P(B|A_3) = 0,05$$

Sử dụng Định lý Bayes để tính xác suất có một sự kiện thể thao (A_1) khi biết rằng bãi đậu xe bị đầy (B):

$$\begin{aligned} P(A_1|B) &= \frac{P(B|A_1) \times P(A_1)}{P(B|A_1) \times P(A_1) + P(B|A_2) \times P(A_2) + P(B|A_3) \times P(A_3)} \\ &= \frac{0,7 \times 0,2}{0,7 \times 0,2 + 0,25 \times 0,35 + 0,05 \times 0,45} \\ &= 0,56 \end{aligned}$$

Dựa trên thông tin rằng bãi đậu xe bị đầy, xác suất có một sự kiện thể thao đang diễn ra vào buổi tối đó là 56%.

³⁷ Sơ đồ cây với ba nhánh chính được hiển thị bên phải. Tiếp theo, chúng ta xác định hai xác suất từ sơ đồ cây: (1) Xác suất có một sự kiện thể thao và bãi đậu xe bị đầy: 0,14. (2) Xác suất bãi đậu xe bị đầy: $0,0875 + 0,14 + 0,0225 = 0,25$. Sau đó, lời giải được tính bằng tỷ số của hai xác suất này: $\frac{0,14}{0,25} = 0,56$. Nếu bãi đậu xe bị đầy, thì có 56% xác suất rằng một sự kiện thể thao đang diễn ra.



BÀI TẬP THỰC HÀNH 3.45

- G Sử dụng thông tin từ bài tập và ví dụ trước để xác minh rằng xác suất có một sự kiện học thuật với điều kiện bãi đỗ xe đầy là 0,35.³⁸

BÀI TẬP THỰC HÀNH 3.46

- G Trong BÀI TẬP THỰC HÀNH 3.43 và 3.45, bạn đã tìm thấy rằng nếu bãi đỗ xe đầy, xác suất có một sự kiện thể thao là 0.56 và xác suất có một sự kiện học thuật là 0,35. Sử dụng thông tin này, hãy tính $P(\text{Không có sự kiện} | \text{Bãi đỗ xe đầy})$.³⁹

Những bài tập gần đây đã cung cấp một cách để cập nhật niềm tin của chúng ta về việc có một sự kiện thể thao, sự kiện học thuật hoặc không có sự kiện nào đang diễn ra tại trường, dựa trên thông tin rằng bãi đỗ xe đầy. Chiến lược cập nhật niềm tin này bằng cách sử dụng Định lý Bayes thực chất là nền tảng của một lĩnh vực thống kê rộng lớn gọi là **thống kê Bayes**. Mặc dù thống kê Bayes rất quan trọng và hữu ích, nhưng chúng ta sẽ không có đủ thời gian để đề cập nhiều hơn về nó trong cuốn sách này.

³⁸ Câu trả lời ngắn:

$$\begin{aligned}
 P(A_1|B) &= \frac{P(B|A_2) \times P(A_2)}{P(B|A_1) \times P(A_1) + P(B|A_2) \times P(A_2) + P(B|A_3) \times P(A_3)} \\
 &= \frac{0,25 \times 0,35}{0,7 \times 0,2 + 0,25 \times 0,35 + 0,05 \times 0,45} \\
 &= 0,35
 \end{aligned}$$

³⁹ Mỗi xác suất đều được tính có điều kiện dựa trên cùng một thông tin rằng gara đã đầy, do đó, có thể sử dụng phần bù: $1,00 - 0,56 - 0,35 = 0,09$.

Bài tập

3.1 Xác suất kết hợp và xác suất có điều kiện. $P(A) = 0,3$, $P(B) = 0,7$

- (a) Bạn có thể tính $P(A \text{ và } B)$ chỉ với thông tin $P(A)$ và $P(B)$?
- (b) Giả sử các sự kiện A và B đến từ các quá trình ngẫu nhiên độc lập:
- Tính $P(A \text{ và } B)$?
 - Tính $P(A \text{ hoặc } B)$?
 - Tính $P(A | B)$?
- (c) Nếu biết rằng $P(A \text{ và } B) = 0,1$, liệu các biến ngẫu nhiên sinh ra các sự kiện A và B có độc lập không?
- (d) Nếu biết rằng $P(A \text{ và } B) = 0,1$, tính $P(A | B)$?

3.2 Bơ đậu phộng và mứt. Giả sử 80% số người thích bơ đậu phộng, 89% thích mứt, và 78% thích cả hai. Biết rằng một người được chọn ngẫu nhiên thích bơ đậu phộng, xác suất để người đó cũng thích mứt là bao nhiêu?

3.3 Biến đổi khí hậu toàn cầu. Một cuộc khảo sát của Pew Research đã hỏi 1.306 người Mỹ: "Dựa trên những gì bạn đã đọc và nghe, liệu có bằng chứng vững chắc rằng nhiệt độ trung bình trên Trái Đất đã tăng lên trong vài thập kỷ qua hay không?" Bảng sau cho thấy phân bố câu trả lời theo đảng phái và tư tưởng chính trị, với các số liệu được thay bằng tần suất tương đối.⁴⁰

	Phản hồi			Tổng	
	Trái đất nóng lên	Không nóng lên	Không biết/Từ chối		
Đảng phái và tư tưởng chính trị	Đảng Cộng hòa bảo thủ	0,11	0,20	0,02	0,33
	Đảng Cộng hòa ôn hòa/tự do	0,06	0,06	0,01	0,13
	Đảng Dân chủ ôn hòa/bảo thủ	0,25	0,07	0,02	0,34
	Đảng Dân chủ tự do	0,18	0,01	0,01	0,20
	Tổng	0,60	0,34	0,06	1,00

- (a) Liệu việc tin rằng Trái Đất đang nóng lên và việc là một đảng viên Dân chủ tự do có loại trừ lẫn nhau không?
- (b) Xác suất để một người được chọn ngẫu nhiên tin rằng Trái Đất đang nóng lên hoặc là một đảng viên Dân chủ tự do là bao nhiêu?
- (c) Xác suất để một người được chọn ngẫu nhiên tin rằng Trái Đất đang nóng lên, với điều kiện người đó là một đảng viên Dân chủ tự do, là bao nhiêu?
- (d) Xác suất để một người được chọn ngẫu nhiên tin rằng Trái Đất đang nóng lên, với điều kiện người đó là một đảng viên Cộng hòa bảo thủ, là bao nhiêu?
- (e) Có vẻ như niềm tin vào việc Trái Đất đang nóng lên có độc lập với đảng phái và tư tưởng chính trị của người trả lời không? Giải thích lý do.
- (f) Xác suất để một người được chọn ngẫu nhiên là một đảng viên Cộng hòa ôn hòa/tự do, với điều kiện người đó không tin rằng Trái Đất đang nóng lên, là bao nhiêu?

⁴⁰ Pew Research Center, Majority of Republicans No Longer See Evidence of Global Warming, data collected on October 27, 2010.

3.4 Bảo hiểm y tế, tần suất tương đối. Hệ thống Giám sát Yếu tố Nguy cơ Hành vi (BRFSS) là một cuộc khảo sát qua điện thoại hàng năm nhằm xác định các yếu tố nguy cơ trong dân số trưởng thành và báo cáo các xu hướng sức khỏe mới nổi. Bảng sau đây hiển thị phân bố tình trạng sức khỏe của những người tham gia khảo sát (xuất sắc, rất tốt, tốt, trung bình, kém) và liệu họ có bảo hiểm y tế hay không.

		Tình trạng sức khỏe					
		Xuất sắc	Rất tốt	Tốt	Trung bình	Kém	Tổng
Bảo hiểm y tế	Không	0,0230	0,0364	0,0427	0,0192	0,0050	0,1262
	Có	0,2099	0,3123	0,2410	0,0817	0,0289	0,8738
	Tổng	0,2329	0,3486	0,2838	0,1009	0,0338	1,0000

- (a) Việc có sức khỏe xuất sắc và có bảo hiểm y tế có loại trừ lẫn nhau không?
 (b) Xác suất để một cá nhân được chọn ngẫu nhiên có sức khỏe xuất sắc là bao nhiêu?
 (c) Xác suất để một cá nhân có sức khỏe xuất sắc, với điều kiện họ có bảo hiểm y tế, là bao nhiêu?
 (d) Xác suất để một cá nhân có sức khỏe xuất sắc, với điều kiện họ không có bảo hiểm y tế, là bao nhiêu?
 (e) Liệu có vẻ như sức khỏe xuất sắc và có bảo hiểm y tế là hai sự kiện độc lập không?

3.5 Sở thích về bánh hamburger. Một cuộc khảo sát của SurveyUSA vào năm 2010 đã hỏi 500 cư dân Los Angeles: "Theo bạn, đâu là nơi bán bánh hamburger ngon nhất ở miền Nam California? Five Guys Burgers? In-N-Out Burger? Fat Burger? Tommy's Hamburgers? Umami Burger? Hay một nơi khác?" Bảng dưới đây hiển thị phân bố câu trả lời theo giới tính.⁴¹

		Giới tính		
		Nam	Nữ	Tổng
Quán hamburger yêu thích	Five Guys Burgers	5	6	11
	In-N-Out Burger	162	181	343
	Fat Burger	10	12	22
	Tommy's Hamburgers	27	27	54
	Umami Burger	5	1	6
	Khác	26	20	46
	Không chắc	13	5	18
Tổng		248	252	500

- (a) Việc là nữ và thích Five Guys Burgers có loại trừ lẫn nhau không?
 (b) Xác suất để một người nam được chọn ngẫu nhiên thích In-N-Out nhất là bao nhiêu?
 (c) Xác suất để một người nữ được chọn ngẫu nhiên thích In-N-Out nhất là bao nhiêu?
 (d) Xác suất để một cặp đôi nam và nữ đều thích In-N-Out nhất là bao nhiêu? Ghi chú bất kỳ giả định nào bạn đưa ra và đánh giá liệu giả định đó có hợp lý không.
 (e) Xác suất để một người được chọn ngẫu nhiên thích Umami nhất hoặc là nữ là bao nhiêu?

⁴¹SurveyUSA, Results of SurveyUSA News Poll #17718, data collected on December 2, 2010.

3.6 Chọn bạn đời theo đặc điểm tương đồng. Chọn bạn đời theo đặc điểm tương đồng (assortative mating) là một mô hình chọn bạn đời không ngẫu nhiên, trong đó các cá nhân có kiểu gen và/hoặc kiểu hình tương đồng có xu hướng kết đôi với nhau thường xuyên hơn so với mô hình chọn bạn đời ngẫu nhiên. Các nhà nghiên cứu về chủ đề này đã thu thập dữ liệu về màu mắt của 204 nam giới Scandinavia và đối tác nữ của họ. Bảng dưới đây tổng hợp kết quả.⁴²

		Đối tác (nữ)			
		Xanh dương	Nâu	Xanh lá	Tổng
Bản thân (Nam)	Xanh dương	78	23	13	114
	Nâu	19	23	12	54
	Xanh lá	11	9	16	36
	Tổng	108	55	41	204

- (a) Xác suất để một nam giới được chọn ngẫu nhiên hoặc bạn đời của anh ta có mắt màu xanh là bao nhiêu?
 (b) Xác suất để một nam giới có mắt màu xanh có bạn đời cũng có mắt màu xanh là bao nhiêu?
 (c) Xác suất để một nam giới có mắt nâu có bạn đời có mắt màu xanh là bao nhiêu? Còn đối với nam giới có mắt xanh lá cây thì sao?
 (d) Có vẻ như màu mắt của nam giới và đối tác của họ là độc lập không? Giải thích lý do.

3.7 Vẽ biểu đồ hộp. Sau một khóa học thống kê nhập môn, 80% sinh viên có thể vẽ biểu đồ hộp (box plot) thành công. Trong số những sinh viên có thể vẽ biểu đồ hộp, 86% đỗ môn học, trong khi chỉ 65% sinh viên không thể vẽ biểu đồ hộp đạt yêu cầu..

- (a) Vẽ biểu đồ cây thể hiện kịch bản này.
 (b) Tính xác suất một sinh viên có thể vẽ biểu đồ hộp nếu biết rằng họ đã đỗ môn học.

3.8 Khuynh hướng bị huyết khối - thrombosis. Một xét nghiệm di truyền được sử dụng để xác định liệu một người có khuynh hướng bị huyết khối hay không. Huyết khối là sự hình thành cục máu đông bên trong mạch máu, gây cản trở dòng chảy của máu trong hệ tuần hoàn. Người ta tin rằng 3% dân số thực sự có khuynh hướng này. Xét nghiệm có độ chính xác 99% nếu một người thực sự có khuynh hướng huyết khối, tức là xác suất có kết quả dương tính khi một người thực sự có khuynh hướng này là 0.99. Xét nghiệm có độ chính xác 98% nếu một người không có khuynh hướng huyết khối. Xác suất để một người được chọn ngẫu nhiên, có kết quả xét nghiệm dương tính, thực sự có khuynh hướng huyết khối là bao nhiêu??

3.9 “Không bao giờ là lupus”. Lupus là một bệnh tự miễn, trong đó hệ miễn dịch tấn công nhầm các protein huyết tương, làm tăng nguy cơ hình thành cục máu đông. Người ta tin rằng 2% dân số mắc bệnh này. Xét nghiệm có độ chính xác 98% nếu một người thực sự mắc bệnh lupus. Xét nghiệm có độ chính xác 74% nếu một người không mắc bệnh lupus. Trong chương trình truyền hình House, nhân vật chính thường nói: “Không bao giờ là lupus”, ngay cả khi xét nghiệm của bệnh nhân cho kết quả dương tính. Bạn có nghĩ câu nói này có phần đúng không? Hãy sử dụng các xác suất thích hợp để đưa ra kết luận.

3.10 Khảo sát cử tri sau bỏ phiếu. Edison Research đã thu thập kết quả khảo sát cử tri trong cuộc bầu cử bãi nhiệm thống đốc Wisconsin, Scott Walker. Kết quả cho thấy: 53% cử tri bỏ phiếu ủng hộ Scott Walker. Trong số những người bỏ phiếu ủng hộ Scott Walker, 37% có bằng đại học. Trong số những người bỏ phiếu chống lại Scott Walker, 44% có bằng đại học. Câu hỏi: Nếu chọn ngẫu nhiên một người tham gia khảo sát và biết rằng người đó có bằng đại học, xác suất để người đó đã bỏ phiếu ủng hộ Scott Walker là bao nhiêu?⁴³

⁴²B. Laeng et al. “Why do blue-eyed men prefer women with the same eye color?” In: *Behavioral Ecology and Sociobiology* 61.3 (2007), pp. 371–384.

⁴³New York Times, Wisconsin recall exit polls.

3.3 Lấy mẫu từ một quần thể nhỏ

Khi chúng ta lấy mẫu các quan sát từ một quần thể, thường thì chỉ lấy một phần nhỏ trong tổng số cá thể hoặc trường hợp có thể có. Tuy nhiên, đôi khi kích thước mẫu của chúng ta đủ lớn hoặc quần thể đủ nhỏ đến mức chúng ta lấy mẫu hơn 10% quần thể⁴⁴ mà *không hoàn lại* (tức là không có khả năng lấy cùng một trường hợp hai lần). Việc lấy mẫu một tỷ lệ đáng kể của quần thể có thể ảnh hưởng quan trọng đến cách chúng ta phân tích mẫu.

VÍ DỤ 3.47

Giảng viên đôi khi chọn ngẫu nhiên một sinh viên để trả lời câu hỏi. Nếu mỗi sinh viên có cơ hội được chọn như nhau và lớp bạn có 15 người, thì xác suất để giảng viên chọn bạn cho câu hỏi tiếp theo là bao nhiêu?

Nếu có 15 người trong lớp và không ai vắng mặt, thì xác suất là $1/15$, tức khoảng 0.067.

VÍ DỤ 3.48

Nếu giảng viên đặt 3 câu hỏi, xác suất bạn không bị chọn là bao nhiêu? Giả sử rằng giảng viên sẽ không chọn cùng một người hai lần trong cùng một buổi học.

Đối với câu hỏi đầu tiên, xác suất giảng viên chọn người khác là $14/15$. Khi đặt câu hỏi thứ hai, chỉ còn lại 14 người chưa được chọn, do đó, nếu bạn không bị chọn ở câu hỏi đầu tiên, thì xác suất bạn tiếp tục không bị chọn là $13/14$. Tương tự, xác suất bạn không bị chọn ở câu hỏi thứ ba là $12/13$. Vì vậy, xác suất bạn không bị chọn cho bất kỳ câu hỏi nào trong ba câu hỏi là:

$$\begin{aligned} P(\text{Không bị chọn trong 3 câu hỏi}) &= P(Q1 = \text{không chọn}, Q2 = \text{Không chọn}, Q3 = \text{Không chọn}) \\ &= \frac{14}{15} \times \frac{13}{14} \times \frac{12}{13} = \frac{12}{15} = 0,80 \end{aligned}$$

VÍ DỤ 3.48

Quy tắc nào cho phép chúng ta nhân các xác suất trong Ví dụ 3.48?⁴⁵

⁴⁴ Quy tắc 10% là một nguyên tắc kinh nghiệm để xác định khi nào những cân nhắc này trở nên quan trọng hơn.

⁴⁵ Ba xác suất mà chúng ta tính thực chất bao gồm một xác suất biên, $P(Q1 = \text{không chọn})$, và hai xác suất có điều kiện: $P(Q2 = \text{không chọn} | Q1 = \text{không chọn})$, $P(Q3 = \text{không chọn} | Q1 = \text{không chọn}, Q2 = \text{không chọn})$

Sử dụng Quy tắc Nhân Tổng quát (General Multiplication Rule), tích của ba xác suất này chính là xác suất không bị chọn trong cả 3 câu hỏi.

VÍ DỤ 3.50

Giả sử giảng viên chọn ngẫu nhiên mà không quan tâm đến việc ai đã được chọn trước đó, tức là sinh viên có thể được chọn nhiều lần. Xác suất bạn không bị chọn trong cả ba câu hỏi là bao nhiêu?

Mỗi lần chọn là độc lập, và xác suất không bị chọn cho một câu hỏi riêng lẻ là $14/15$. Do đó, chúng ta có thể sử dụng Quy tắc Nhân cho các quá trình độc lập:

E

$$\begin{aligned} P(\text{Không bị chọn trong 3 câu hỏi}) &= P(Q1 = \text{không chọn}, Q2 = \text{Không chọn}, Q3 = \text{Không chọn}) \\ &= \frac{14}{15} \times \frac{14}{15} \times \frac{14}{15} = 0,813 \end{aligned}$$

Bạn có xác suất không bị chọn hơi cao hơn so với trường hợp giảng viên chọn một người mới cho mỗi câu hỏi. Tuy nhiên, trong trường hợp này, một sinh viên có thể bị chọn nhiều lần.

BÀI TẬP THỰC HÀNH 3.51

G

Dưới cách chọn trong Ví dụ 3.50, xác suất bạn được chọn để trả lời cả ba câu hỏi là bao nhiêu? ⁴⁶

Nếu chúng ta lấy mẫu từ một quần thể nhỏ mà **không hoàn lại**, thì các quan sát không còn độc lập với nhau. Trong Ví dụ 3.48, xác suất không bị chọn cho câu hỏi thứ hai phụ thuộc vào việc bạn không bị chọn cho câu hỏi thứ nhất. Trong Ví dụ 3.50, giảng viên lấy mẫu sinh viên **có hoàn lại**: cô ấy chọn ngẫu nhiên từ toàn bộ lớp nhiều lần mà không quan tâm đến việc ai đã được chọn trước đó.

BÀI TẬP THỰC HÀNH 3.52

G

Khoa của bạn đang tổ chức một cuộc rút thăm trúng thưởng. Họ bán 30 vé và trao bảy giải thưởng.

(a) Họ đặt tất cả vé vào một chiếc mũ và rút một vé cho mỗi giải thưởng. Các vé được rút mà không hoàn lại, tức là vé đã chọn sẽ không được đặt lại vào mũ. Xác suất trúng một giải thưởng nếu bạn mua một vé là bao nhiêu?

(b) Nếu các vé được rút có hoàn lại thì sao? ⁴⁷

BÀI TẬP THỰC HÀNH 3.46

G

So sánh câu trả lời của bạn trong BÀI TẬP THỰC HÀNH 3.52. Phương pháp lấy mẫu ảnh hưởng bao nhiêu đến cơ hội trúng giải của bạn? ⁴⁸

Nếu chúng ta lặp lại BÀI TẬP THỰC HÀNH 3.52 với 300 vé thay vì 30, chúng ta sẽ nhận thấy một điều thú vị: kết quả gần như giống hệt nhau. Xác suất trúng thưởng sẽ là 0,0233 nếu lấy mẫu không hoàn lại và 0,0231 nếu lấy mẫu có hoàn lại. Khi kích thước mẫu chỉ chiếm một phần nhỏ của quần thể (dưới 10%), các quan sát gần như độc lập ngay cả khi lấy mẫu không hoàn lại.

⁴⁶ $P(\text{được chọn để trả lời cả ba câu hỏi}) = \left(\frac{1}{15}\right)^3 = 0,00030$.

⁴⁷ (a) Xác suất bạn không trúng giải trong lần rút thăm đầu tiên là $29/30$, trong lần thứ hai là $28/29$, ..., và trong lần thứ bảy là $23/24$. Xác suất bạn không trúng bất kỳ giải nào là tích của các xác suất riêng lẻ này: $23/30$. Do đó, xác suất trúng ít nhất một giải là: $1 - 23/30 = 7/30 = 0,233$. (b) Khi vé được rút có hoàn lại, có bảy lần rút độc lập. Trước tiên, ta tìm xác suất không trúng giải: $(29/30)^7 = 0,789$. Vậy xác suất trúng ít nhất một giải khi rút có hoàn lại là: $1 - 0,789 = 0,211$.

⁴⁸ Xác suất trúng giải khi lấy mẫu không hoàn lại cao hơn khoảng 10% so với khi lấy mẫu có hoàn lại. Tuy nhiên, khi lấy mẫu không hoàn lại, mỗi người chỉ có thể trúng tối đa một giải.

Bài tập

3.1 Bi ve trong một chiếc bình. Hãy tưởng tượng bạn có một chiếc bình chứa 5 viên bi đỏ, 3 viên bi xanh và 2 viên bi cam.

- Xác suất viên bi đầu tiên bạn rút ra là màu xanh là bao nhiêu?
- Giả sử bạn đã rút được một viên bi xanh trong lần rút đầu tiên. Nếu rút có hoàn lại, xác suất rút được một viên bi xanh trong lần rút thứ hai là bao nhiêu?
- Giả sử thay vào đó, bạn rút được một viên bi cam trong lần rút đầu tiên. Nếu rút có hoàn lại, xác suất rút được một viên bi xanh trong lần rút thứ hai là bao nhiêu?
- Nếu rút có hoàn lại, xác suất rút được hai viên bi xanh liên tiếp là bao nhiêu?
- Khi rút có hoàn lại, các lần rút có độc lập không? Giải thích.

3.2 Tất trong ngăn kéo. Trong ngăn kéo tất của bạn có 4 đôi màu xanh, 5 đôi màu xám và 3 đôi màu đen. Một buổi sáng khi còn ngủ, bạn lấy ngẫu nhiên 2 chiếc tất và đi chúng. Hãy tìm xác suất rằng bạn sẽ:

- Đi 2 chiếc tất màu xanh.
- Không đi chiếc tất màu xám nào.
- Đi ít nhất một chiếc tất màu đen.
- Đi một chiếc tất màu xanh lá cây.
- Đi hai chiếc tất giống nhau.

3.3 Quân cờ trong túi. Hãy tưởng tượng bạn có một túi chứa 5 quân cờ đỏ, 3 quân cờ xanh và 2 quân cờ cam.

- Giả sử bạn rút một quân cờ và đó là màu xanh. Nếu rút không hoàn lại, xác suất rút được quân cờ xanh trong lần tiếp theo là bao nhiêu?
- Giả sử bạn rút được một quân cờ cam, sau đó rút tiếp một quân cờ khác mà không hoàn lại. Xác suất quân cờ thứ hai là màu xanh là bao nhiêu?
- Nếu rút không hoàn lại, xác suất rút được hai quân cờ xanh liên tiếp là bao nhiêu?
- Khi rút không hoàn lại, các lần rút có độc lập không? Giải thích.

3.4 Sách trên kệ sách. Bảng dưới đây cho thấy phân bố sách trên kệ dựa trên hai tiêu chí: sách hư cấu hay phi hư cấu và bìa cứng hay bìa mềm.

		Bìa		Tổng
		Bìa cứng	Bìa mềm	
Loại sách	Sách hư cấu	13	59	72
	Sách phi hư cấu	15	8	23
	Tổng	28	67	95

- Tìm xác suất rút được một cuốn sách bìa cứng trước, sau đó là một cuốn sách hư cấu bìa mềm, khi rút không hoàn lại.
- Xác định xác suất rút được một cuốn sách hư cấu trước, sau đó là một cuốn sách bìa cứng, khi rút không hoàn lại.
- Tính xác suất của tình huống trong phần (b), nhưng lần này giả sử cuốn sách đầu tiên được đặt lại trên kệ trước khi rút ngẫu nhiên cuốn sách thứ hai.
- Kết quả cuối cùng của phần (b) và (c) rất giống nhau. Giải thích lý do tại sao.

3.5 Trang phục của học sinh. Trong một lớp học có 24 học sinh: 7 học sinh mặc quần jean, 4 học sinh mặc quần short, 8 học sinh mặc váy, Số còn lại mặc quần legging. Nếu chúng ta chọn ngẫu nhiên 3 học sinh không hoàn lại, xác suất rằng một học sinh mặc legging và hai học sinh mặc quần jean là bao nhiêu? Lưu ý rằng các loại trang phục này loại trừ lẫn nhau (mỗi học sinh chỉ mặc một loại).

3.6 Bài toán sinh nhật. Giả sử chúng ta chọn ngẫu nhiên ba người. Đối với các câu hỏi sau, bỏ qua trường hợp đặc biệt khi ai đó có thể sinh vào ngày 29 tháng 2, và giả sử rằng ngày sinh phân bố đều trong năm.

- Xác suất hai người đầu tiên có cùng ngày sinh là bao nhiêu?
- Xác suất có ít nhất hai người có cùng ngày sinh là bao nhiêu?

3.4 Biến ngẫu nhiên

Việc mô hình hóa một quá trình bằng cái gọi là biến ngẫu nhiên thường rất hữu ích. Mô hình này cho phép chúng ta áp dụng một khung toán học và các nguyên tắc thống kê để hiểu rõ hơn và dự đoán kết quả trong thế giới thực.

VÍ DỤ 3.54

Hai cuốn sách được yêu cầu cho một lớp thống kê: một sách giáo khoa và một sách hướng dẫn học tập đi kèm. Hiệu sách của trường xác định rằng: 20% sinh viên đăng ký học không mua cuốn nào. 55% sinh viên chỉ mua sách giáo khoa. 25% sinh viên mua cả hai cuốn. Tỷ lệ này tương đối ổn định qua các kỳ học. Nếu có 100 sinh viên đăng ký, hiệu sách nên dự đoán sẽ bán được bao nhiêu cuốn sách?

Khoảng 20 sinh viên sẽ không mua sách nào (0 cuốn). Khoảng 55 sinh viên sẽ mua một cuốn (55 cuốn). Khoảng 25 sinh viên sẽ mua hai cuốn (50 cuốn). Tổng số sách mà hiệu sách có thể kỳ vọng bán cho lớp học này là 105 cuốn.

BÀI TẬP THỰC HÀNH 3.55

Bạn có ngạc nhiên không nếu hiệu sách bán nhiều hơn hoặc ít hơn một chút so với 105 cuốn? ⁴⁹

VÍ DỤ 3.56

Giá sách giáo khoa là \$137, và sách hướng dẫn học tập có giá \$33. Vậy hiệu sách có thể kỳ vọng bao nhiêu doanh thu từ lớp học gồm 100 sinh viên này?

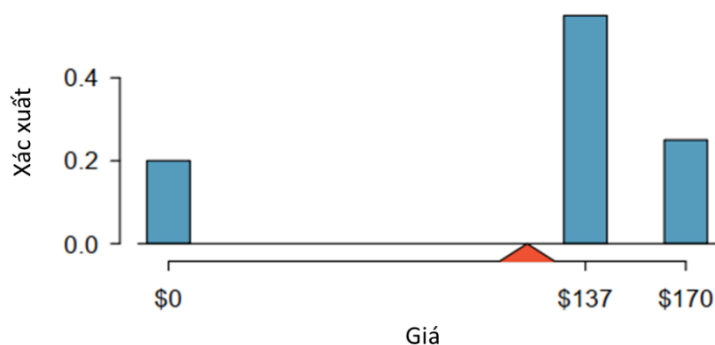
Khoảng 55 sinh viên chỉ mua sách giáo khoa, mang lại doanh thu:

$$\$137 \times 55 = \$7535$$

Khoảng 25 sinh viên mua cả sách giáo khoa và sách hướng dẫn học tập, với tổng chi phí:

$$(\$137 + \$33) \times 25 = \$170 \times 25 = \$4250$$

Vậy tổng doanh thu mà hiệu sách có thể kỳ vọng từ lớp học này là: $\$7535 + \$4250 = \$11785$ Tuy nhiên, do sự biến động ngẫu nhiên trong số lượng sinh viên mua sách, doanh thu thực tế có thể hơi khác một chút.



Hình 3.18: Phân phối xác suất của doanh thu mà hiệu sách thu được từ một sinh viên. Hình tam giác biểu thị doanh thu trung bình trên mỗi sinh viên.

⁴⁹ Nếu họ bán nhiều hơn hoặc ít hơn một chút, điều đó không có gì đáng ngạc nhiên. Hy vọng rằng Chương 1 đã giúp làm rõ rằng dữ liệu quan sát được luôn có sự biến động tự nhiên. Ví dụ, nếu chúng ta tung một đồng xu 100 lần, nó thường sẽ không cho kết quả mặt ngửa đúng chính xác 50 lần, nhưng sẽ xấp xỉ con số đó.

VÍ DỤ 3.57

Trung bình doanh thu trên mỗi sinh viên cho khóa học này là bao nhiêu?

Tổng doanh thu kỳ vọng là \$11785, và có 100 sinh viên. Do đó, doanh thu kỳ vọng trên mỗi sinh viên là: $\$11,785/100 = \$117,85$.

3.4.1 Kỳ vọng (Expectation)

Chúng ta gọi một biến hoặc quá trình có kết quả số học là **biến ngẫu nhiên (random variable)**, và thường ký hiệu biến ngẫu nhiên này bằng một chữ in hoa như X , Y hoặc Z . Số tiền mà một sinh viên sẽ chi cho sách thống kê của mình là một biến ngẫu nhiên, và chúng ta ký hiệu nó là X .

BIẾN NGẪU NHIÊN (RANDOM VARIABLE)

Một quá trình ngẫu nhiên hoặc một biến có kết quả là một giá trị số.

Các kết quả có thể có của X được ký hiệu bằng chữ thường tương ứng x cùng với chỉ số dưới. Ví dụ, ta viết $x_1 = 0$, $x_2 = 137$, và $x_3 = 170$, với xác suất xảy ra lần lượt là 0,20, 0,55 và 0,25. Phân phối của X được tóm tắt trong Hình 3.18 và Hình 3.19.

i	1	2	3	Tổng
x_i	\$0	\$137	\$170	-
$P(X = x_i)$	0,20	0,55	0,25	1,00

Hình 3.19: Phân phối xác suất của biến ngẫu nhiên X , đại diện cho doanh thu của nhà sách từ một sinh viên.

Chúng tôi đã tính toán giá trị trung bình của X là \$117.85 trong Ví dụ 3.57. Giá trị trung bình này được gọi là **giá trị kỳ vọng** của X , ký hiệu là $E(X)$. Kỳ vọng của một biến ngẫu nhiên được tính bằng cách cộng các kết quả có thể xảy ra, mỗi kết quả được nhân với xác suất tương ứng của nó:

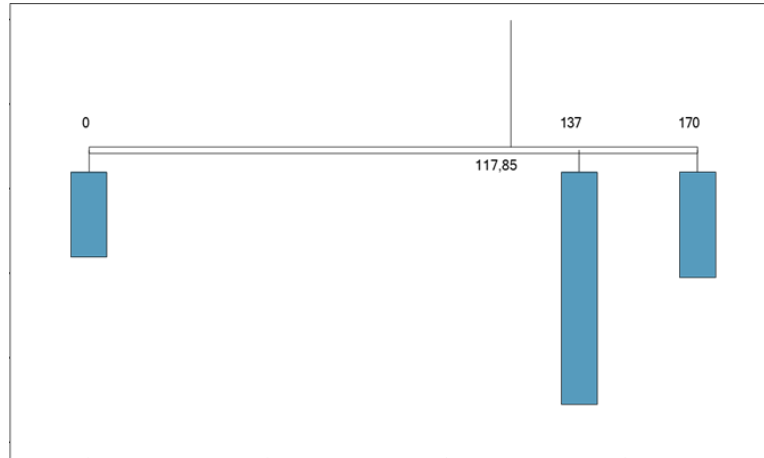
$$\begin{aligned} E(X) &= 0 \times P(X = 0) + 137 \times P(X = 137) + 170 \times P(X = 170) \\ &= 0 \times 0,20 + 137 \times 0,55 + 170 \times 0,25 = 117,85 \end{aligned}$$

GIÁ TRỊ KỲ VỌNG CỦA MỘT BIẾN NGẪU NHIÊN RỜI RẠC

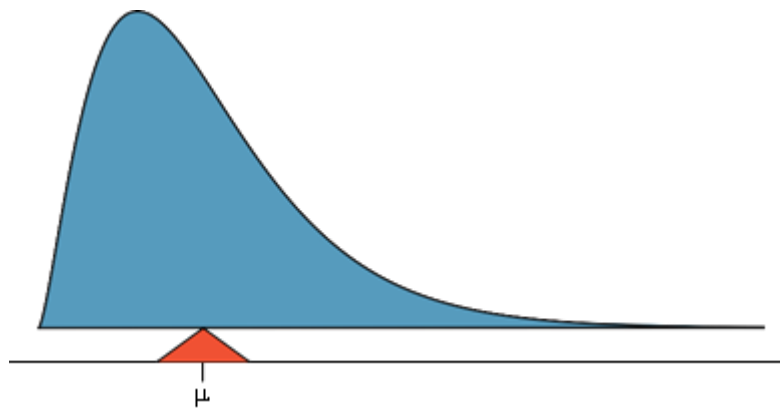
Nếu X nhận các giá trị $x_1, \dots, x_k$ với các xác suất tương ứng $P(X = x_1), \dots, P(X = x_k)$, thì giá trị kỳ vọng của X được tính bằng tổng của mỗi kết quả nhân với xác suất tương ứng của nó:

$$\begin{aligned} E(X) &= x_1 \times P(X = x_1) + \dots + x_k \times P(X = x_k) \\ &= \sum_{i=1}^k x_i P(X = x_i) \end{aligned}$$

Chữ cái Hy Lạp μ có thể được sử dụng thay cho ký hiệu $E(X)$.



Hình 3.20: Một hệ thống trọng số biểu diễn phân phối xác suất của X . Sơ dây giữ phân phối tại giá trị trung bình để duy trì trạng thái cân bằng.



Hình 3.21: Một phân phối liên tục cũng có thể được cân bằng tại giá trị trung bình của nó.

Giá trị kỳ vọng của một biến ngẫu nhiên thể hiện kết quả trung bình. Ví dụ, $E(X) = 117,85$ đại diện cho số tiền trung bình mà nhà sách kỳ vọng thu được từ một sinh viên, và ta cũng có thể viết dưới dạng $\mu = 117,85$.

Cũng có thể tính giá trị kỳ vọng của một biến ngẫu nhiên liên tục (xem Mục 3.5). Tuy nhiên, điều này đòi hỏi một chút kiến thức về giải tích, vì vậy chúng ta sẽ để dành cho một bài học sau.⁵⁰

Trong vật lý, giá trị kỳ vọng mang ý nghĩa tương tự như trọng tâm. Phân phối có thể được biểu diễn bằng một loạt các khối lượng đặt tại từng giá trị kết quả, và giá trị trung bình chính là điểm cân bằng. Điều này được minh họa trong Hình 3.18 và Hình 3.20.

Ý tưởng về trọng tâm cũng được mở rộng cho các phân phối xác suất liên tục. Hình 3.21 cho thấy một phân phối xác suất liên tục được cân bằng trên một điểm tựa đặt tại giá trị trung bình.

⁵⁰ $\mu = \int x f(x) dx$ trong đó $f(x)$ là hàm mật độ xác suất của phân phối.

3.4.2 Sự biến thiên trong các biến ngẫu nhiên

Giả sử bạn điều hành một hiệu sách đại học. Bên cạnh việc quan tâm đến doanh thu dự kiến, bạn cũng có thể muốn biết mức độ biến động (sự thay đổi) của doanh thu.

Phương sai và độ lệch chuẩn có thể được sử dụng để mô tả mức độ biến thiên của một biến ngẫu nhiên. Mục 2.1.4 đã giới thiệu một phương pháp để tìm phương sai và độ lệch chuẩn cho một tập dữ liệu. Trước tiên, chúng ta tính độ lệch so với giá trị trung bình ($x_i - \mu$), bình phương các độ lệch đó và lấy giá trị trung bình để thu được phương sai. Trong trường hợp của một biến ngẫu nhiên, chúng ta cũng tính các độ lệch bình phương. Tuy nhiên, thay vì chỉ lấy trung bình, chúng ta tính tổng của chúng với trọng số là xác suất tương ứng của từng giá trị, giống như cách tính kỳ vọng. Tổng có trọng số của các độ lệch bình phương này chính là phương sai, và chúng ta tính độ lệch chuẩn bằng cách lấy căn bậc hai của phương sai, tương tự như cách đã làm trong Mục 2.1.4.

CÔNG THỨC TỔNG QUÁT CỦA PHƯƠNG SAI GIÁ

Nếu biến ngẫu nhiên X có các giá trị đầu ra $x_1, \dots, x_k$ với các xác suất tương ứng $P(X = x_1), \dots, P(X = x_k)$, à giá trị kỳ vọng: $\mu = E(X)$, thì phương sai của X , được ký hiệu là $\text{Var}(X)$ hay là σ^2 , được tính theo công thức:

$$\begin{aligned}\sigma^2 &= (x_1 - \mu)^2 \times P(X = x_1) + \dots + (x_k - \mu)^2 \times P(X = x_k) \\ &= \sum_{i=1}^k (x_i - \mu)^2 P(X = x_i)\end{aligned}$$

Độ lệch chuẩn của X , ký hiệu là σ , là căn bậc hai của phương sai.

VÍ DỤ 3.58

Tính giá trị kỳ vọng, phương sai và độ lệch chuẩn của X , tức là doanh thu từ một sinh viên thống kê đối với nhà sách.

Việc lập một bảng chứa các phép tính cho từng kết quả riêng lẻ, sau đó cộng tổng các kết quả lại, sẽ rất hữu ích.

i	1	2	3	Tổng
x_i	\$0	\$137	\$170	
$P(X = x_i)$	0,20	0,55	0,25	
$x_i \times P(X = x_i)$	0	75,35	42,50	117,85

Do đó, giá trị kỳ vọng là $\mu = 117,85$, mà chúng ta đã tính trước đó. Phương sai có thể được xác định bằng cách mở rộng bảng này:

i	1	2	3	Tổng
x_i	\$0	\$137	\$170	
$P(X = x_i)$	0,20	0,55	0,25	
$x_i \times P(X = x_i)$	0	75,35	42,50	117,85
$x_i - \mu$	-117,85	19,15	52,15	
$(x_i - \mu)^2$	13888,62	366,72	2719,62	
$(x_i - \mu)^2 \times P(X = x_i)$	2777,7	201,7	679,9	3659,3

Phương sai của X là $\sigma^2 = 3659,3$, có nghĩa là độ lệch chuẩn là $\sigma = \sqrt{3659,3} = \$60,49$.

BÀI TẬP THỰC HÀNH 3.59

Hiệu sách cũng bán một cuốn sách giáo khoa hóa học với giá 159 đô la và một tài liệu bổ trợ với giá 41 đô la. Từ kinh nghiệm trước đây, họ biết rằng khoảng 25% sinh viên hóa học chỉ mua sách giáo khoa, trong khi 60% mua cả sách giáo khoa và tài liệu bổ trợ.⁵¹

- (a) Tỷ lệ sinh viên không mua bất kỳ cuốn sách nào là bao nhiêu? Giả sử không có sinh viên nào mua tài liệu bổ trợ mà không mua sách giáo khoa.
- (b) Gọi Y là doanh thu từ một sinh viên. Viết phân phối xác suất của Y , tức là lập một bảng cho từng kết quả có thể xảy ra và xác suất tương ứng.
- (c) Tính toán doanh thu kỳ vọng từ một sinh viên hóa học.
- (d) Tìm độ lệch chuẩn để mô tả mức độ biến động của doanh thu từ một sinh viên.

3.4.3 Kết hợp tuyến tính của các biến ngẫu nhiên

Cho đến nay, chúng ta đã xem xét mỗi biến như một câu chuyện hoàn chỉnh và độc lập. Tuy nhiên, đôi khi việc sử dụng sự kết hợp của các biến sẽ hợp lý hơn. Ví dụ, tổng thời gian một người dành để đi làm mỗi tuần có thể được chia nhỏ thành nhiều lần đi lại hằng ngày. Tương tự, tổng lợi nhuận hoặc thua lỗ trong một danh mục đầu tư chứng khoán chính là tổng của các khoản lợi nhuận và thua lỗ từ từng thành phần trong danh mục đó.

VÍ DỤ 3.60

John đi làm năm ngày một tuần. Chúng ta sẽ sử dụng X_1 để biểu thị thời gian di chuyển của anh ấy vào thứ Hai, X_2 cho thứ Ba, và tiếp tục như vậy. Hãy viết một phương trình sử dụng $X_1, \dots, X_5$ để biểu thị tổng thời gian di chuyển trong tuần của anh ấy, được ký hiệu là W .

Tổng thời gian di chuyển hàng tuần của anh ấy là tổng của năm giá trị di chuyển hằng ngày:

$$W = X_1 + X_2 + X_3 + X_4 + X_5$$

Việc chia nhỏ tổng thời gian di chuyển hàng tuần W thành từng phần giúp hiểu rõ hơn từng nguồn biến động và rất hữu ích trong việc mô hình hóa W .

⁵¹ (a) $100\% - 25\% - 60\% = 15\%$ số sinh viên không mua bất kỳ cuốn sách nào cho lớp học. Phần (b) được thể hiện bởi hai dòng đầu tiên trong bảng dưới đây. Kỳ vọng cho phần (c) được tính bằng tổng của dòng $y_i \times P(Y = y_i)$. Kết quả của phần (d) là căn bậc hai của phương sai được liệt kê trong tổng ở dòng cuối cùng: $\sigma = \sqrt{\text{Var}(Y)} = \$69,28$.

i (tình huống)	1 (không mua sách)	2 (mua sách giáo khoa)	3 (mua cả hai)	Tổng
y_i	0,00	159,00	200,00	
$P(Y = y_i)$	0,15	0,25	0,60	
$y_i \times P(Y = y_i)$	0,00	39,75	120,0	$E(Y) = 159,75$
$y_i - E(Y)$	-159,75	-0,75	40,25	
$(y_i - E(Y))^2$	25520,06	0,56	1620,06	
$(y_i - E(Y))^2 \times P(Y)$	3828,0	0,1	972,0	$\text{Var}(Y) \approx 4800$

VÍ DỤ 3.61

John mất trung bình 18 phút mỗi ngày để đi làm. Vậy thời gian đi làm trung bình của anh ấy trong cả tuần sẽ là bao nhiêu?

Chúng ta biết rằng thời gian đi làm trung bình mỗi ngày là 18 phút, tức là: $E(X_i) = 18$. Để tính thời gian đi làm trung bình trong 5 ngày, ta cộng trung bình của từng ngày lại:

$$\begin{aligned} E(W) &= E(X_1 + X_2 + X_3 + X_4 + X_5) \\ &= E(X_1) + E(X_2) + E(X_3) + E(X_4) + E(X_5) \\ &= 18 + 18 + 18 + 18 + 18 = 90 \text{ phút} \end{aligned}$$

Kết quả này cho thấy tổng thời gian kỳ vọng cho cả tuần chính là tổng của kỳ vọng từng ngày. Nói chung, kỳ vọng của tổng các biến ngẫu nhiên luôn bằng tổng kỳ vọng của từng biến ngẫu nhiên đó.

BÀI TẬP THỰC HÀNH 3.63

Elena đang bán một chiếc TV trong một cuộc đấu giá tiền mặt và cũng có ý định mua một lò nướng bánh mì trong cùng cuộc đấu giá. Nếu X biểu thị lợi nhuận từ việc bán TV và Y biểu thị chi phí mua lò nướng bánh mì, hãy viết một phương trình biểu thị sự thay đổi ròng trong số tiền mặt của Elena.⁵³

BÀI TẬP THỰC HÀNH 3.64

Bạn có ngạc nhiên không nếu thời gian di chuyển hàng tuần của John không chính xác là 90 phút hoặc nếu Elena không kiếm được chính xác 152 đô la? Giải thích.⁵⁴

Hai khái niệm quan trọng liên quan đến tổ hợp các biến ngẫu nhiên đã được giới thiệu cho đến nay. Thứ nhất, một giá trị cuối cùng đôi khi có thể được mô tả như tổng của các phần trong một phương trình. Thứ hai, trực giác cho thấy rằng khi đưa các giá trị trung bình riêng lẻ vào phương trình này, chúng ta sẽ thu được giá trị trung bình tổng thể mong đợi. Điểm thứ hai này cần được làm rõ – nó được đảm bảo đúng trong cái gọi là tổ hợp tuyến tính của các biến ngẫu nhiên.

Tổ hợp tuyến tính của hai biến ngẫu nhiên X và Y là một cụm từ chuyên môn để mô tả một tổ hợp:

$$aX + bY$$

trong đó a và b là các số cố định và đã biết. Đối với thời gian đi làm của John, có năm biến ngẫu nhiên – mỗi biến tương ứng với một ngày làm việc – và mỗi biến ngẫu nhiên có thể được viết với một hệ số cố định là 1:

$$1X_1 + 1X_2 + 1X_3 + 1X_4 + 1X_5$$

Đối với khoản lãi hoặc lỗ ròng của Elena, biến ngẫu nhiên X có hệ số là +1, trong khi biến ngẫu nhiên Y có hệ số là -1.

⁵² Cô ấy sẽ kiếm được X đô la từ việc bán TV nhưng sẽ chi Y đô la để mua lò nướng bánh mì: $X - Y$.

⁵³ Kỳ vọng toán học của tổng lợi nhuận là:

$E(X - Y) = E(X) - E(Y) = 175 - 23 = \152 . Do đó, cô ấy có thể mong đợi thu về khoảng \$152.

⁵⁴ Không, vì có thể sẽ có sự biến động. Ví dụ, tình trạng giao thông có thể thay đổi theo từng ngày, và giá đấu giá cũng có thể thay đổi tùy thuộc vào chất lượng hàng hóa cũng như mức độ quan tâm của người tham dự.

Khi xét đến giá trị trung bình của một tổ hợp tuyến tính của các biến ngẫu nhiên, ta có thể thay thế mỗi biến

ngẫu nhiên bằng kỳ vọng (giá trị trung bình) của nó và sau đó tính toán kết quả cuối cùng một cách an toàn. Đối với một số trường hợp tổ hợp phi tuyến của các biến ngẫu nhiên – khi ta không thể đơn giản thay thế bằng giá trị trung bình – vui lòng tham khảo chú thích bên dưới.⁵⁵

TỔ HỢP TUYẾN TÍNH CỦA CÁC BIẾN NGẪU NHIÊN VÀ GIÁ TRỊ TRUNG BÌNH

Nếu X và Y là các biến ngẫu nhiên, thì một tổ hợp tuyến tính của chúng được biểu diễn dưới dạng:

$$aX + bY$$

trong đó a và b là các hằng số cố định. Để tính giá trị trung bình của một tổ hợp tuyến tính của các biến ngẫu nhiên, ta thay thế mỗi biến ngẫu nhiên bằng kỳ vọng của nó và thực hiện phép tính như sau:

$$a \times E(X) + b \times E(Y)$$

Nhớ rằng kỳ vọng toán $E(X)$ chính là giá trị trung bình, tức là $E(X) = \mu_X$.

VÍ DỤ 3.65

Leonard đã đầu tư 6000 đô la vào Caterpillar Inc (mã chứng khoán: CAT) và 2000 đô la vào Exxon Mobil Corp (XOM). Nếu X đại diện cho mức thay đổi của cổ phiếu Caterpillar trong tháng tới và Y đại diện cho mức thay đổi của cổ phiếu Exxon Mobil trong tháng tới, hãy viết một phương trình mô tả số tiền mà Leonard sẽ kiếm được hoặc mất đi từ danh mục đầu tư của mình trong tháng đó.

E Để đơn giản hóa, giả sử X và Y không được biểu diễn dưới dạng phần trăm mà ở dạng thập phân (ví dụ: nếu cổ phiếu Caterpillar tăng 1%, thì $X = 0,01$; hoặc nếu giảm 1%, thì $X = -0,01$). Khi đó, phương trình mô tả lợi nhuận của Leonard có thể viết như sau:

$$\$6000 \times X + \$2000 \times Y$$

Nếu chúng ta thay giá trị thay đổi của cổ phiếu vào X và Y , phương trình này sẽ cho ra mức thay đổi trong giá trị danh mục đầu tư của Leonard trong tháng. Một giá trị dương biểu thị lợi nhuận, trong khi một giá trị âm biểu thị khoản lỗ.

BÀI TẬP THỰC HÀNH 3.66

G Cổ phiếu Caterpillar gần đây đã tăng trung bình 2,0% mỗi tháng, trong khi cổ phiếu Exxon Mobil tăng 0,2% mỗi tháng. Hãy tính mức thay đổi kỳ vọng trong danh mục đầu tư của Leonard cho tháng tới.⁵⁶

BÀI TẬP THỰC HÀNH 3.67

G Bạn sẽ thấy rằng Leonard dự kiến sẽ có lợi nhuận dương trong BÀI TẬP THỰC HÀNH 3.66. Tuy nhiên, liệu bạn có ngạc nhiên nếu anh ấy thực sự bị lỗ trong tháng này không?⁵⁷

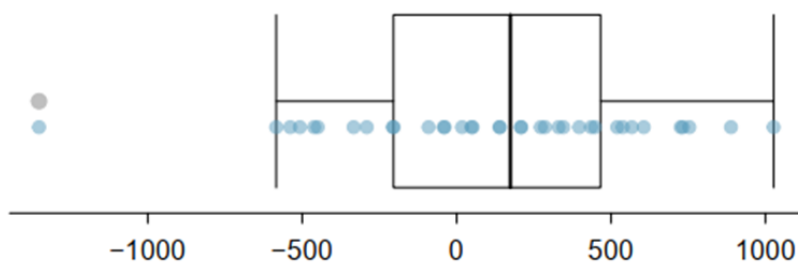
⁵⁵ Nếu X và Y là các biến ngẫu nhiên, hãy xem xét các tổ hợp sau: X^{1+Y} , $X \cdot Y$, X / Y . Trong những trường hợp này, việc thay thế giá trị trung bình của từng biến ngẫu nhiên rồi tính toán kết quả thường sẽ không dẫn đến giá trị trung bình chính xác cho kết quả cuối cùng.

⁵⁶ $E(\$6000 \times X + \$2000 \times Y) = \$6000 \times 0,020 + \$2000 \times 0,002 = \$124$.

⁵⁷ Không. Mặc dù cổ phiếu có xu hướng tăng giá theo thời gian, nhưng chúng thường biến động mạnh trong ngắn hạn.

3.4.4 Độ biến động trong các tổ hợp tuyến tính của biến ngẫu nhiên

Việc định lượng kết quả trung bình từ một tổ hợp tuyến tính của các biến ngẫu nhiên là hữu ích, nhưng cũng quan trọng để có một cái nhìn về sự không chắc chắn liên quan đến tổng kết quả của tổ hợp các biến ngẫu nhiên đó. Lợi nhuận hoặc thua lỗ ròng kỳ vọng từ danh mục đầu tư cổ phiếu của Leonard đã được xem xét trong BÀI TẬP THỰC HÀNH 3.66. Tuy nhiên, không có thảo luận định lượng nào về sự biến động của danh mục đầu tư này. Ví dụ, trong khi lợi nhuận trung bình hàng tháng có thể khoảng \$124 theo số liệu, thì khoản lợi nhuận đó không được đảm bảo. Hình 3.22 cho thấy những thay đổi hàng tháng trong một danh mục đầu tư giống như của Leonard trong khoảng thời gian ba năm. Các khoản lãi và lỗ biến động rất nhiều, và việc định lượng những biến động này là quan trọng khi đầu tư vào cổ phiếu.



Lợi suất hàng tháng trong 3 năm

Hình 3.22: Sự thay đổi trong một danh mục đầu tư giống như của Leonard trong 36 tháng, trong đó \$6000 đầu tư vào cổ phiếu của Caterpillar và \$2000 vào cổ phiếu của Exxon Mobil.

Cũng như chúng ta đã làm trong nhiều trường hợp trước đây, chúng ta sử dụng phương sai và độ lệch chuẩn để mô tả sự không chắc chắn liên quan đến lợi suất hàng tháng của Leonard. Để thực hiện việc này, phương sai của lợi suất hàng tháng của từng cổ phiếu sẽ rất hữu ích, và những điều này được hiển thị trong Hình 3.23. Lợi suất của các cổ phiếu gần như là độc lập với nhau.

Tại đây, chúng ta sử dụng một phương trình từ lý thuyết xác suất để mô tả sự không chắc chắn của lợi suất hàng tháng của Leonard; chúng tôi sẽ để phần chứng minh phương pháp này cho một khóa học xác suất chuyên biệt. Phương sai của một tổ hợp tuyến tính của các biến ngẫu nhiên có thể được tính bằng cách thay thế các phương sai của các biến ngẫu nhiên riêng lẻ và bình phương các hệ số của các biến ngẫu nhiên:

$$\text{Var}(aX + bY) = a^2 \times \text{Var}(X) + b^2 \times \text{Var}(Y)$$

Điều quan trọng cần lưu ý là sự bình đẳng này giả định rằng các biến ngẫu nhiên là độc lập; nếu không có tính độc lập, thì sẽ cần một sự điều chỉnh cho phương trình này, mà chúng tôi để lại như một chủ đề cho một khóa học trong tương lai. Phương trình này có thể được sử dụng để tính toán phương sai của lợi suất hàng tháng của Leonard:

$$\begin{aligned} \text{Var}(6000 \times X + 2000 \times Y) &= 6000^2 \times \text{Var}(X) + 2000^2 \times \text{Var}(Y) \\ &= 36.000.000 \times 0,0057 + 4.000.000 \times 0,0021 \\ &\approx 213.600 \end{aligned}$$

Độ lệch chuẩn được tính toán là căn bậc hai của phương sai: $\sqrt{213.600} = \$463$. Mặc dù lợi suất hàng tháng trung bình là \$124 trên một khoản đầu tư \$8000 là điều không thể coi thường, nhưng lợi suất hàng tháng lại rất biến động đến nỗi Leonard nên mong đợi thu nhập này sẽ ổn định.

	Trung bình ($\bar{x}$)	Độ lệch chuẩn (s)	Phương sai (s^2)
CAT	0,0204	0,0757	0,0057
XOM	0,0025	0,0455	0,0021

Hình 3.23: Trung bình, độ lệch chuẩn và phương sai của cổ phiếu CAT và XOM. Các số liệu thống kê này được ước lượng từ dữ liệu cổ phiếu lịch sử, vì vậy ký hiệu được sử dụng cho các thống kê mẫu đã được áp dụng.

ĐỘ BIẾN ĐỘNG CỦA CÁC TỔ HỢP TUYẾN TÍNH CỦA CÁC BIẾN NGẪU NHIÊN

Phương sai của một tổ hợp tuyến tính của các biến ngẫu nhiên có thể được tính bằng cách bình phương các hằng số, thay thế các phương sai của các biến ngẫu nhiên và tính toán kết quả:

$$Var(aX + bY) = a^2 \times Var(X) + b^2 \times Var(Y)$$

Phương trình này có giá trị miễn là các biến ngẫu nhiên là độc lập với nhau. Độ lệch chuẩn của tổ hợp tuyến tính có thể được tìm bằng cách lấy căn bậc hai của phương sai.

VÍ DỤ 3.68

Giả sử thời gian đi làm hàng ngày của John có độ lệch chuẩn là 4 phút. Vậy sự không chắc chắn trong tổng thời gian đi làm của anh ấy trong tuần là bao nhiêu?

Biểu thức cho thời gian đi làm của John là:

$$X_1 + X_2 + X_3 + X_4 + X_5$$

Mỗi hệ số là 1, và phương sai của thời gian mỗi ngày là $4^2 = 16$. Do đó, phương sai của tổng thời gian đi làm trong tuần là:

$$\text{phương sai} = 1^2 \times 16 + 1^2 \times 16 + 1^2 \times 16 + 1^2 \times 16 + 1^2 \times 16 = 5 \times 16 = 80$$

$$\text{độ lệch chuẩn} = \sqrt{\text{phương sai}} = \sqrt{80} = 8,94$$

Độ lệch chuẩn cho thời gian đi làm hàng tuần của John khoảng 9 phút.

BÀI TẬP THỰC HÀNH 3.69

Phép tính trong Ví dụ 3.68 dựa trên một giả định quan trọng: thời gian đi làm của mỗi ngày là độc lập với thời gian của những ngày khác trong tuần đó. Bạn có nghĩ rằng điều này là hợp lý không? Giải thích.⁵⁸

BÀI TẬP THỰC HÀNH 3.70

Xét hai cuộc đấu giá của Elena từ BÀI TẬP THỰC HÀNH 3.62 trên trang 120. Giả sử các cuộc đấu giá này gần như độc lập và độ biến động trong giá đấu giá liên quan đến chiếc TV và lò nướng có thể được mô tả bằng độ lệch chuẩn là \$25 và \$8. Tính toán độ lệch chuẩn của lợi nhuận ròng của Elena.⁵⁹

Xét lại BÀI TẬP THỰC HÀNH 3.70. Hệ số âm cho Y trong tổ hợp tuyến tính đã bị loại bỏ khi chúng ta bình phương các hệ số. Điều này thường đúng: các hệ số âm trong một tổ hợp tuyến tính sẽ không ảnh hưởng đến độ biến động được tính toán cho một tổ hợp tuyến tính, nhưng chúng có ảnh hưởng đến các phép tính giá trị kỳ vọng.

⁵⁸ Một mối quan tâm là liệu các mẫu giao thông có xu hướng có chu kỳ tuần (ví dụ: thứ Sáu có thể tồi tệ hơn các ngày khác) hay không. Nếu đúng như vậy, và John lái xe, thì giả định này có thể không hợp lý. Tuy nhiên, nếu John đi bộ đến nơi làm việc, thì thời gian đi làm của anh ấy có thể không bị ảnh hưởng bởi bất kỳ chu kỳ giao thông nào trong tuần.

⁵⁹ Phương trình cho Elena có thể được viết như sau: $(1) \times X + (-1) \times Y$

Các phương sai của X và Y lần lượt là 625 và 64. Chúng ta bình phương các hệ số và thay thế các phương sai:

$(1)^2 \times Var(X) + (-1)^2 \times Var(Y) = 1 \times 625 + 1 \times 64 = 689$

Phương sai của tổ hợp tuyến tính là 689, và độ lệch chuẩn là căn bậc hai của 689: khoảng \$26,25.

Bài tập

3.1 Những người hút thuốc ở trường đại học. Tại một trường đại học, 13% sinh viên hút thuốc.

- Tính số lượng dự kiến những người hút thuốc trong một mẫu ngẫu nhiên gồm 100 sinh viên từ trường đại học này.
- Phòng tập thể dục của trường đại học mở cửa vào lúc 9 giờ sáng vào các sáng thứ Bảy. Một sáng thứ Bảy lúc 8 giờ 55 có 27 sinh viên đứng bên ngoài phòng gym chờ đợi nó mở cửa. Bạn có nên sử dụng cách tiếp cận giống như ở phần (a) để tính số lượng dự kiến những người hút thuốc trong số 27 sinh viên này không?

3.2 Lá bài át cơ chiến thắng. Cần nhắc trò chơi bài sau với một bộ bài được xáo trộn kỹ. Nếu bạn rút một lá bài đỏ, bạn sẽ không thắng gì. Nếu bạn rút một lá bài bích, bạn thắng \$5. Đối với bất kỳ lá bài cơ nào, bạn thắng \$10 cộng thêm \$20 nếu đó là át cơ.

- Tạo một mô hình xác suất cho số tiền bạn thắng trong trò chơi này. Đồng thời, tìm số tiền dự kiến bạn thắng trong một trò chơi và độ lệch chuẩn của số tiền thắng.
- Số tiền tối đa bạn sẵn sàng trả để chơi trò này là bao nhiêu? Giải thích lý do của bạn.

3.3 Lá bài cơ chiến thắng. Trong một trò chơi bài mới, bạn bắt đầu với một bộ bài đầy đủ đã được xáo trộn và rút 3 lá bài mà không thay thế. Nếu bạn rút 3 lá bài trái tim, bạn thắng \$50. Nếu bạn rút 3 lá bài đen, bạn thắng \$25. Đối với bất kỳ lượt rút nào khác, bạn không thắng gì.

- Tạo một mô hình xác suất cho số tiền bạn thắng trong trò chơi này và tìm số tiền thắng dự kiến. Tính cả độ lệch chuẩn của phân phối này.
- Nếu trò chơi này có giá \$5 để chơi, giá trị và độ lệch chuẩn dự kiến của lợi nhuận (hoặc thua lỗ) ròng sẽ là gì? (Gợi ý: $\text{lợi nhuận} = \text{số tiền thắng} - \text{chi phí}$; $X - 5$)
- Nếu trò chơi này có giá \$5 để chơi, bạn có nên tham gia không? Giải thích.

3.4 Có xứng đáng không? Andy luôn tìm kiếm cách để kiếm tiền nhanh chóng. Gần đây, anh đã cố gắng kiếm tiền bằng cách đánh bạc. Đây là trò chơi mà anh đang cân nhắc chơi: Trò chơi này có giá \$2 để chơi. Anh rút một lá bài từ một bộ bài. Nếu anh rút được một lá bài số (2-10), anh không thắng gì. Đối với bất kỳ lá bài mặt (jack, queen hoặc king), anh thắng \$3. Đối với bất kỳ lá át, anh thắng \$5 và anh thắng thêm \$20 nếu anh rút được át cơ.

- Tạo một mô hình xác suất và tìm lợi nhuận dự kiến của Andy cho mỗi trò chơi.
- Bạn có đề xuất trò chơi này cho Andy như một cách tốt để kiếm tiền không? Giải thích.

3.5 Lợi tức danh mục đầu tư. Giá trị của một danh mục đầu tư tăng 18% trong thời kỳ kinh tế bùng nổ và 9% trong thời kỳ bình thường. Nó giảm 12% trong thời kỳ suy thoái. Lợi tức dự kiến của danh mục đầu tư này sẽ là gì nếu mỗi kịch bản có xác suất như nhau?

3.6 Phí hành lý. Một hãng hàng không tính phí hành lý như sau: \$25 cho túi đầu tiên và \$35 cho túi thứ hai. Giả sử 54% hành khách không có hành lý ký gửi, 34% có một kiện hành lý ký gửi và 12% có hai kiện. Chúng tôi giả định rằng một tỷ lệ không đáng kể hành khách ký gửi hơn hai kiện.

- Xây dựng một mô hình xác suất, tính doanh thu trung bình trên mỗi hành khách và tính độ lệch chuẩn tương ứng.
- Hãng hàng không nên mong đợi doanh thu khoảng bao nhiêu cho một chuyến bay có 120 hành khách? Độ lệch chuẩn là bao nhiêu? Lưu ý các giả định bạn đưa ra và xem bạn có nghĩ rằng chúng là hợp lý không.

3.7 Roulette kiểu Mỹ. Trò chơi roulette Mỹ liên quan đến việc quay một bánh xe có 38 khe: 18 khe màu đỏ, 18 khe màu đen và 2 khe màu xanh lá. Một viên bi được quay vào bánh xe và cuối cùng sẽ rơi vào một khe, với mỗi khe có xác suất như nhau để giữ viên bi. Người chơi có thể đặt cược vào màu đỏ hoặc màu đen. Nếu viên bi rơi vào màu của họ, họ gấp đôi số tiền. Nếu rơi vào màu khác, họ mất số tiền của mình. Giả sử bạn đặt cược \$1 vào màu đỏ. Giá trị kỳ vọng và độ lệch chuẩn của số tiền thắng của bạn là gì?

3.8 Roulette kiểu Châu Âu. Trò chơi roulette châu Âu liên quan đến việc quay một bánh xe có 37 khe: 18 khe màu đỏ, 18 khe màu đen và 1 khe màu xanh lá. Một viên bi được quay vào bánh xe và cuối cùng sẽ rơi vào một khe, với mỗi khe có xác suất như nhau để giữ viên bi. Người chơi có thể đặt cược vào màu đỏ hoặc màu đen. Nếu viên bi rơi vào màu của họ, họ gấp đôi số tiền. Nếu rơi vào màu khác, họ mất số tiền của mình.

- Giả sử bạn chơi roulette và đặt cược \$3 cho một vòng chơi đơn. Giá trị kỳ vọng và độ lệch chuẩn của tổng số tiền thắng của bạn là gì?
- Giả sử bạn đặt cược \$1 cho ba vòng chơi khác nhau. Giá trị kỳ vọng và độ lệch chuẩn của tổng số tiền thắng của bạn là gì?
- Các câu trả lời của bạn cho phần (a) và (b) so sánh như thế nào? Điều này nói lên điều gì về độ rủi ro của hai trò chơi?

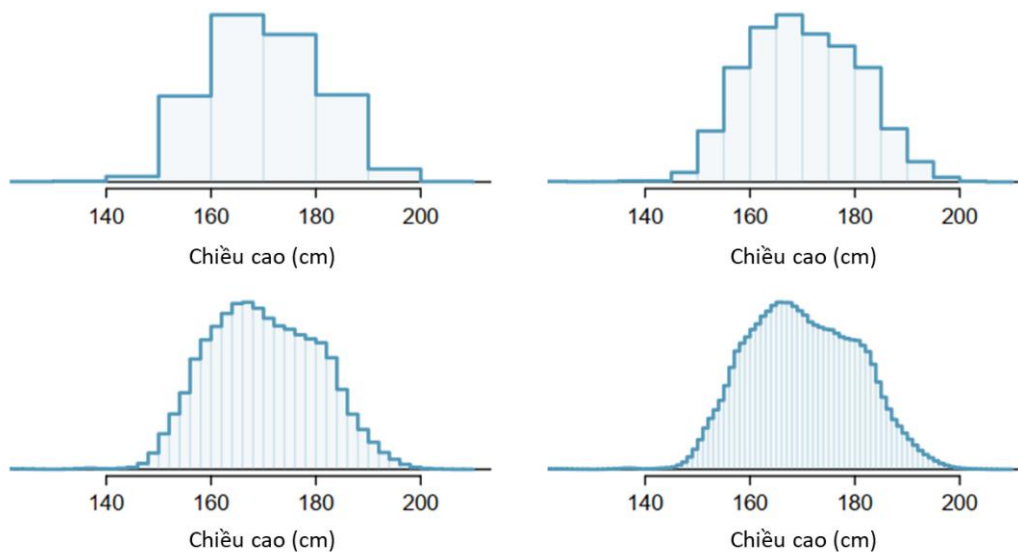
3.5 Phân phối liên tục - Continuous distributions

Cho đến nay trong chương này, chúng ta đã thảo luận về những trường hợp mà kết quả của một biến là rời rạc. Trong phần này, chúng ta sẽ xem xét một bối cảnh mà kết quả là một biến số liên tục.

VÍ DỤ 3.71

Hình 3.24 cho thấy một vài biểu đồ hình chóp rỗng khác nhau về chiều cao của người lớn ở Hoa Kỳ. Việc thay đổi số lượng các khoảng (bins) cho phép bạn đưa ra những cách diễn giải khác nhau về dữ liệu như thế nào?

Việc thêm nhiều khoảng hơn cung cấp chi tiết hơn. Mẫu dữ liệu này cực kỳ lớn, đó là lý do tại sao các khoảng nhỏ hơn vẫn hoạt động tốt. Thông thường, chúng ta không sử dụng nhiều khoảng như vậy với các kích thước mẫu nhỏ hơn vì số lượng nhỏ trên mỗi khoảng có nghĩa là chiều cao của các khoảng rất biến động.



Hình 3.24: Bốn biểu đồ hình chóp rỗng thể hiện chiều cao của người lớn ở Hoa Kỳ với các độ rộng khoảng khác nhau.

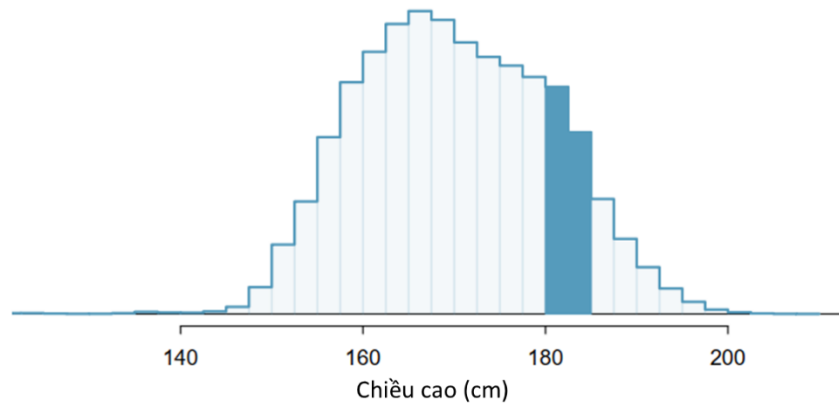
VÍ DỤ 3.72

Tỉ lệ của mẫu nằm trong khoảng chiều cao từ 180 cm đến 185 cm (khoảng 5'11" đến 6'1") là bao nhiêu?

Chúng ta có thể cộng chiều cao của các khoảng trong khoảng từ 180 cm đến 185 cm và chia cho kích thước mẫu. Ví dụ, điều này có thể được thực hiện với hai khoảng được tô bóng như được hiển thị trong Hình 3.25. Hai khoảng trong khu vực này có số lượng lần lượt là 195,307 và 156,239 người, dẫn đến ước tính xác suất như sau:

$$\frac{195.307 + 156.239}{3.000.000} = 0,1172$$

Phân số này cũng chính là tỉ lệ của diện tích biểu đồ hình chóp nằm trong khoảng từ 180 đến 185 cm

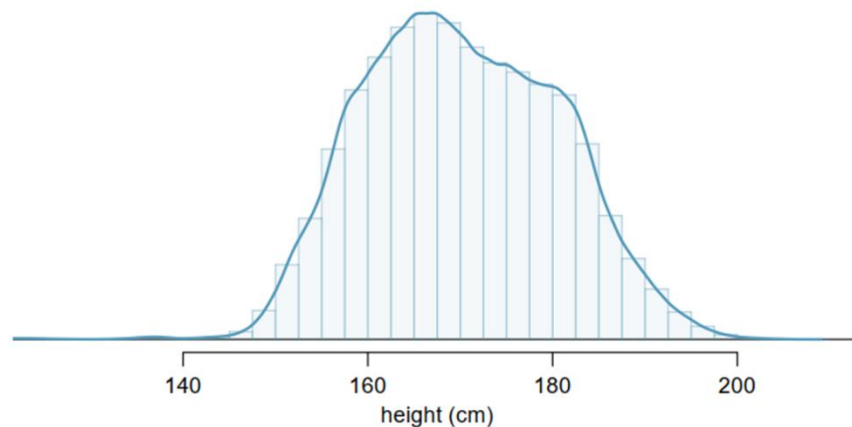


Hình 3.25: Một biểu đồ lịch sử với kích thước ô là 2.5 cm. Khu vực đã được tô đậm đại diện cho những cá nhân có chiều cao nằm trong khoảng từ 180 đến 185 cm.

3.5.1 Từ biểu đồ lịch sử đến phân phối liên tục

Xem xét sự chuyển tiếp từ một biểu đồ lịch sử rỗng với hình dạng khối ở phía trên bên trái của Hình 3.24 đến biểu đồ mượt mà hơn ở phía dưới bên phải. Trong biểu đồ cuối cùng này, các ô rất mảnh đến mức biểu đồ rỗng bắt đầu giống như một đường cong mượt mà. Điều này gợi ý rằng chiều cao của dân số như một biến số *liên tục* có thể được giải thích tốt nhất bởi một đường cong đại diện cho hình dạng của các ô cực kỳ mảnh.

Đường cong mượt mà này đại diện cho một **hàm mật độ xác suất - probability density function** (còn được gọi là **mật độ** hoặc **phân phối**), và đường cong như vậy được hiển thị trong Hình 3.26 phủ lên một biểu đồ histogram của mẫu. Một hàm mật độ có một đặc điểm đặc biệt: tổng diện tích dưới đường cong của hàm mật độ là 1.



Hình 3.26: Phân phối xác suất liên tục của chiều cao cho người lớn ở Mỹ.

3.5.2 Xác suất từ các phân phối liên tục

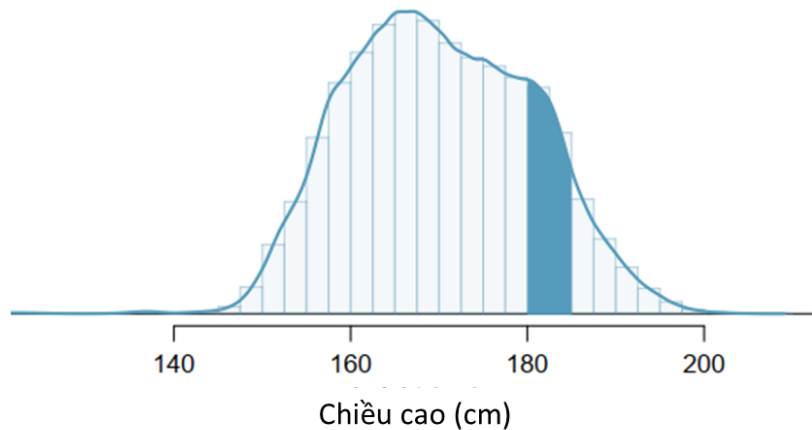
Chúng tôi đã tính toán tỷ lệ cá nhân có chiều cao từ 180 đến 185 cm trong Ví dụ 3.72 dưới dạng một phân số:

$$\frac{\text{số người có chiều cao từ 180 đến 185}}{\text{tổng kích thước mẫu}}$$

Chúng tôi đã xác định số người có chiều cao trong khoảng từ 180 đến 185 cm bằng cách xác định phân số diện tích của biểu đồ histogram trong khu vực này. Tương tự, chúng tôi có thể sử dụng diện tích trong khu vực đã được tô đậm dưới đường cong để tìm xác suất (với sự hỗ trợ của máy tính):

$$P(\text{chiều cao giữa 180 và 185}) = \text{diện tích giữa 180 và 185} = 0,1157$$

Xác suất rằng một người được chọn ngẫu nhiên có chiều cao giữa 180 và 185 cm là 0,1157. Điều này rất gần với ước tính từ Ví dụ 3.72: 0,1172.



Hình 3.27: Mật độ chiều cao trong dân số người lớn ở Mỹ với diện tích giữa 180 và 185 cm được tô đậm. So sánh biểu đồ này với Hình 3.25.

BÀI TẬP THỰC HÀNH 3.73

G

Ba người lớn ở Mỹ được chọn ngẫu nhiên. Xác suất một người lớn là giữa 180 và 185 cm là 0,1157.⁶⁰

- Xác suất rằng cả ba người đều có chiều cao từ 180 đến 185 cm là gì?
- Xác suất rằng không ai trong số họ có chiều cao từ 180 đến 185 cm là gì?

VÍ DỤ 3.74

Xác suất rằng một người được chọn ngẫu nhiên có chiều cao **chính xác** 180 cm là gì? Giả sử bạn có thể đo một cách chính xác?

E

Xác suất này là 0. Một người có thể gần bằng 180 cm, nhưng không thể có chiều cao chính xác 180 cm. Điều này cũng phù hợp với định nghĩa về xác suất dưới dạng diện tích; không có diện tích nào được xác định giữa 180 cm và 180 cm..

BÀI TẬP THỰC HÀNH 3.75

G

Giả sử chiều cao của một người được làm tròn đến cm gần nhất. Có khả năng nào rằng chiều cao đo được của một người ngẫu nhiên sẽ là 180 cm không?

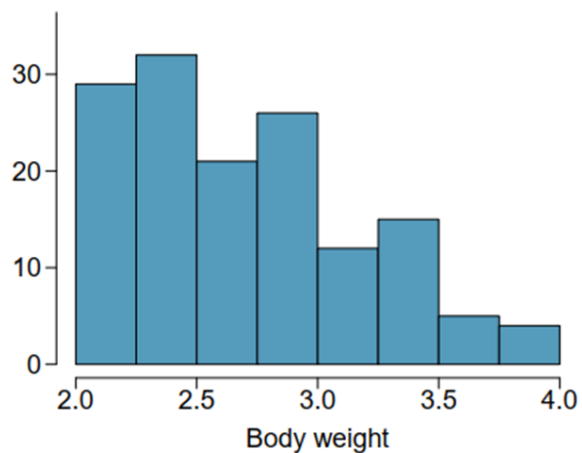
⁶⁰ Câu trả lời ngắn gọn: (a) $0,1157 \times 0,1157 \times 0,1157 = 0,0015$. (b) $(1 - 0,1157)^3 = 0,692$.

⁶¹ Điều này có xác suất dương. Bất kỳ ai có chiều cao từ 179,5 cm đến 180,5 cm đều sẽ có chiều cao đo được là 180 cm. Đây có lẽ là một kịch bản thực tế hơn để gặp phải trong thực hành so với Ví dụ 3.74.

Bài tập

3.1 Cân nặng của mèo. Biểu đồ histogram bên dưới thể hiện cân nặng (tính bằng kg) của 47 con mèo cái và 97 con mèo đực.⁶²

- (a) Tỷ lệ của những con mèo này nặng dưới 2.5 kg là bao nhiêu?
 (b) Tỷ lệ của những con mèo này nặng giữa 2.5 và 2.75 kg là bao nhiêu?
 (c) Tỷ lệ của những con mèo này nặng giữa 2.75 và 3.5 kg là bao nhiêu?



3.2 Thu nhập và giới tính. Bảng tần suất tương đối bên dưới hiển thị phân phối thu nhập cá nhân tổng cộng hàng năm (tính theo đô la đã điều chỉnh theo lạm phát năm 2009) cho một mẫu đại diện gồm 96,420,486 người Mỹ. Dữ liệu này được lấy từ Khảo sát Cộng đồng Mỹ cho giai đoạn 2005-2009. Mẫu này gồm 59% nam và 41% nữ.⁶³

- (a) Mô tả phân phối tổng thu nhập cá nhân.
 (b) Xác suất rằng một cư dân Mỹ được chọn ngẫu nhiên có thu nhập dưới \$50.000 mỗi năm là bao nhiêu?
 (c) Xác suất rằng một cư dân Mỹ được chọn ngẫu nhiên có thu nhập dưới \$50.000 mỗi năm và là nữ là bao nhiêu? Ghi chú bất kỳ giả định nào mà bạn đưa ra.
 (d) Nguồn dữ liệu tương tự cho biết rằng 71,8% phụ nữ có thu nhập dưới 50.000 đô la mỗi năm. Sử dụng giá trị này để xác định xem giả định bạn đã đưa ra trong phần (c) có hợp lệ hay không.

Thu nhập	Tổng
\$1 đến \$9.999 hoặc nhỏ hơn	2,2%
\$10.000 đến \$14.999	4,7%
\$15.000 đến \$24.999	15,8%
\$25.000 đến \$34.999	18,3%
\$35.000 đến \$49.999	21,2%
\$50.000 đến \$64.999	13,9%
\$65.000 đến \$74.999	5,8%
\$75.000 đến \$99.999	8,4%
Trên \$100.000	9,7%

⁶² W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Fourth Edition. www.stats.ox.ac.uk/pub/MASS4. New York: Springer, 2002.

⁶³ U.S. Census Bureau, 2005-2009 American Community Survey.

Bài tập chương

3.3 Phân phối điểm số. Mỗi hàng trong bảng dưới đây là một phân phối điểm được đề xuất cho một lớp học. Hãy xác định xem mỗi phân phối có phải là phân phối xác suất hợp lệ hay không và giải thích lý do của bạn.

	Điểm				
	A	B	C	D	F
(a)	0,3	0,3	0,3	0,2	0,1
(b)	0	0	1	0	0
(c)	0,3	0,3	0,3	0	0
(d)	0,3	0,5	0,2	0,1	-0,1
(e)	0,2	0,4	0,2	0,1	0,1
(f)	0	-0,1	1,1	0	0

3.4 Bảo hiểm sức khỏe, tần suất. Hệ thống Giám sát Yếu tố Rủi ro Hành vi (BRFSS) là một cuộc khảo sát điện thoại hàng năm được thiết kế để xác định các yếu tố rủi ro trong dân số trưởng thành và báo cáo các xu hướng sức khỏe mới nổi. Bảng sau tóm tắt hai biến đối với người được khảo sát: tình trạng sức khỏe và bảo hiểm sức khỏe, mô tả xem mỗi người được khảo sát có bảo hiểm y tế hay không.⁶⁴

		Tình trạng sức khỏe				
		Tuyệt vời	Rất tốt	Tốt	Bình thường	Tệ
Health Coverage	Không	459	727	854	385	99
	Có	4198	6245	4821	1634	578
	Tổng	4657	6972	5675	2019	677
						20000

- (a) Nếu chúng ta chọn ngẫu nhiên một người, xác suất người được khảo sát có sức khỏe tuyệt vời và không có bảo hiểm y tế là bao nhiêu?
- (b) Nếu chúng ta chọn ngẫu nhiên một người, xác suất người được khảo sát có sức khỏe tuyệt vời hoặc không có bảo hiểm y tế là bao nhiêu?

3.5 HIV ở Swaziland. Swaziland có tỷ lệ nhiễm HIV cao nhất thế giới: 25,9% dân số nước này bị nhiễm HIV.⁶⁵ Xét nghiệm ELISA là một trong những xét nghiệm đầu tiên và chính xác nhất đối với HIV. Đối với những người mang HIV, xét nghiệm ELISA có độ chính xác 99,7%. Đối với những người không mang HIV, xét nghiệm có độ chính xác 92,6%. Nếu một cá nhân từ Swaziland có kết quả xét nghiệm dương tính, xác suất người đó mang HIV là bao nhiêu?

3.6 Sinh đôi. Khoảng 30% cặp song sinh ở người là sinh đôi cùng trứng, và phần còn lại là sinh đôi khác trứng. Sinh đôi cùng trứng chắc chắn cùng giới tính - một nửa là nam và một nửa là nữ. Một phần tư cặp song sinh khác trứng đều là nam, một phần tư đều là nữ, và một nửa là hỗn hợp: một nam, một nữ. Bạn vừa mới sinh đôi và được biết cả hai đều là bé gái. Với thông tin này, xác suất chúng là sinh đôi cùng trứng là bao nhiêu?

3.7 Chi phí bữa sáng. Sally mỗi ngày đều mua một tách cà phê và một chiếc bánh mì nướng làm bữa sáng từ một trong nhiều quán cà phê trong khu phố của cô ấy. Cô ấy chọn một quán cà phê mỗi buổi sáng một cách ngẫu nhiên và độc lập với những ngày trước. Giá trung bình của một tách cà phê là \$1,40 với độ lệch chuẩn là 30 xu (\$0,30), giá trung bình của một chiếc bánh mì nướng là \$2,50 với độ lệch chuẩn là 15 xu, và hai giá này độc lập với nhau.

- (a) Giá trị trung bình và độ lệch chuẩn của số tiền cô ấy chi tiêu cho bữa sáng hàng ngày là bao nhiêu?
- (b) Giá trị trung bình và độ lệch chuẩn của số tiền cô ấy chi tiêu cho bữa sáng hàng tuần (7 ngày) là bao nhiêu?

⁶⁴ Office of Surveillance, Epidemiology, and Laboratory Services Behavioral Risk Factor Surveillance System, BRFSS 2010 Survey Data.

⁶⁵ Source: CIA Factbook, Country Comparison: HIV/AIDS - Adult Prevalence Rate.

3.5. PHÂN PHỐI LIÊN TỤC

3.8 Múc kem. Kem thường được đóng hộp 1,5 lít (48 fl oz), và muỗng múc kem chứa khoảng 2 fl oz. Tuy nhiên, có một số biến đổi về lượng kem trong hộp cũng như lượng kem được múc ra. Chúng ta biểu diễn lượng kem trong hộp là X và lượng kem được múc ra là Y . Giả sử các biến ngẫu nhiên này có các giá trị trung bình, độ lệch chuẩn và phương sai sau:

	Trung bình	Độ lệch chuẩn	Phương sai
X	48	1	1
Y	2	0,25	0,0625

- (a) Toàn bộ một hộp kem, cộng thêm 3 muỗng từ hộp thứ hai được phục vụ tại một bữa tiệc. Bạn mong đợi đã có bao nhiêu kem được phục vụ tại bữa tiệc này? Độ lệch chuẩn của lượng kem được phục vụ là bao nhiêu?
- (b) Bạn mong đợi còn lại bao nhiêu kem trong hộp sau khi múc ra một muỗng kem? Tức là, hãy tìm giá trị kỳ vọng của $X - Y$. Độ lệch chuẩn của lượng kem còn lại trong hộp là bao nhiêu?
- (c) Sử dụng ngữ cảnh của bài tập này, hãy giải thích tại sao chúng ta cộng các phương sai khi chúng ta trừ một biến ngẫu nhiên khỏi biến ngẫu nhiên khác.

3.9 Phương sai của trung bình, Phần I. Giả sử chúng ta có các quan sát độc lập X_1 và X_2 từ một phân phối có trung bình μ và độ lệch chuẩn σ . Phương sai của trung bình của hai giá trị: $\frac{X_1 + X_2}{2}$ là bao nhiêu?

3.10 Phương sai của trung bình, Phần II. Giả sử chúng ta có 3 quan sát độc lập X_1, X_2, X_3 từ một phân phối có trung bình μ và độ lệch chuẩn σ . Phương sai của trung bình của 3 giá trị này: $\frac{X_1 + X_2 + X_3}{3}$ là bao nhiêu?

3.11 Phương sai của trung bình, Phần III. Giả sử chúng ta có n quan sát độc lập $X_1, X_2, \dots, X_n$ từ một phân phối có trung bình μ và độ lệch chuẩn σ . Phương sai của trung bình của n giá trị này: $\frac{X_1 + X_2 + \dots + X_n}{n}$ là bao nhiêu?