

Chương 4

Phân bố của biến số ngẫu nhiên

4.1 Phân phối chuẩn

4.2 Phân phối hình học - Geometric distribution

4.3 Phân phối nhị thức

4.4 Phân phối nhị thức âm

4.5 Phân phối Poisson

Trong chương này, chúng ta sẽ thảo luận về các phân phối thống kê thường xuyên xuất hiện trong bối cảnh phân tích dữ liệu hoặc suy luận thống kê. Chúng tôi bắt đầu với phân phối chuẩn trong phần đầu tiên, phân phối này được sử dụng nhiều trong các chương sau của cuốn sách. Các phần còn lại sẽ được tham khảo thỉnh thoảng nhưng có thể coi là không bắt buộc đối với nội dung của cuốn sách.



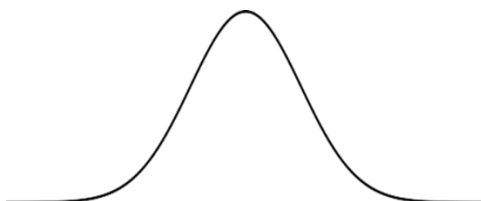
Để xem video, slides và các tài nguyên khác, vui lòng truy cập

www.openintro.org/os

4.1. PHÂN PHỐI CHUẨN

4.1. Phân phối chuẩn

Trong các loại phân bố ta thường gặp trong thực tế, có một loại phân phối xuất hiện với tần suất vượt trội hơn cả. Đó là đường cong đối xứng, đơn đỉnh và có hình chuông – một dạng phân phối rất phổ biến và người ta thường gọi nó là **phân phối chuẩn**,¹ được minh họa trong Hình 4.1. Các biến số như điểm SAT hoặc chiều cao của nam giới trưởng thành tại Hoa Kỳ thường tuân theo phân phối chuẩn.



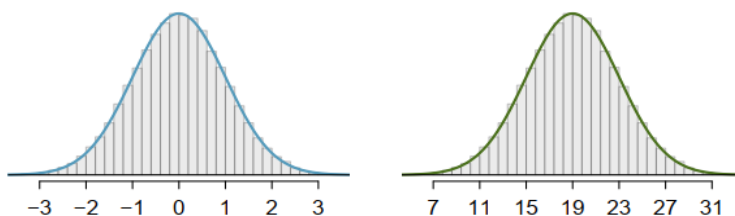
Hình 4.1: Phân phối chuẩn.

CÁC ĐẶC ĐIỂM PHÂN PHỐI CHUẨN

Nhiều biến số trong thực tế xấp xỉ tuân theo phân phối chuẩn, nhưng không có biến số nào hoàn toàn tuân theo phân phối này một cách tuyệt đối. Do đó, mặc dù phân phối chuẩn không hoàn hảo đối với bất kỳ vấn đề cụ thể nào, nhưng nó vẫn rất hữu ích trong nhiều tình huống khác nhau. Chúng ta sẽ sử dụng phân phối chuẩn trong quá trình khám phá dữ liệu và để giải quyết các bài toán quan trọng trong thống kê.

4.1.1. Mô hình phân phối chuẩn

Phân phối chuẩn luôn mô tả một đường cong đối xứng, đơn đỉnh và có dạng hình chuông. Tuy nhiên, hình dạng của đường cong này có thể thay đổi tùy theo các tham số của mô hình. Cụ thể, phân phối chuẩn được điều chỉnh bởi hai tham số chính: trung bình và độ lệch chuẩn. Như bạn có thể đoán, việc thay đổi trung bình sẽ làm dịch chuyển đường cong sang trái hoặc phải, trong khi thay đổi độ lệch chuẩn sẽ làm cho đường cong giãn rộng ra hoặc thu hẹp lại. Hình 4.2 minh họa phân phối chuẩn với trung bình bằng 0 và độ lệch chuẩn bằng 1 ở hình bên trái, và phân phối chuẩn với trung bình bằng 19 và độ lệch chuẩn bằng 4 ở hình bên phải. Hình 4.3 trình bày hai phân phối này trên cùng một trục để dễ so sánh.



Hình 4.2: Cả hai đường cong đều biểu diễn phân phối chuẩn. Tuy nhiên, chúng khác nhau về trung tâm và độ phân tán.

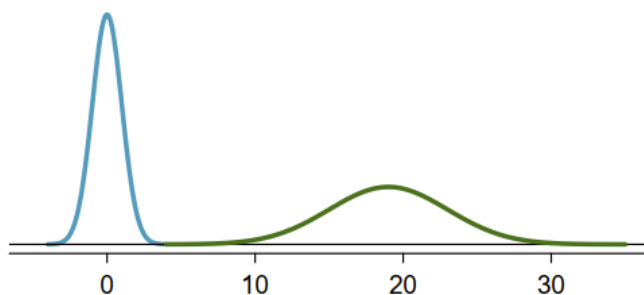
Nếu một phân phối chuẩn có trung bình là μ và độ lệch chuẩn là σ , ta có thể ký hiệu phân phối đó là $N(\mu, \sigma)$. Hai phân phối ở hình 4.3 có thể được ký hiệu như sau

$$N(\mu = 0, \sigma = 1) \text{ và } N(\mu = 19, \sigma = 4)$$

Vì trung bình và độ lệch chuẩn mô tả chính xác một phân phối chuẩn, chúng được gọi là các **tham số** của phân phối đó. Phân phối chuẩn có trung bình bằng 0 và độ lệch chuẩn bằng 1 được gọi là **phân phối chuẩn tắc**.

¹ Phân phối này cũng được gọi là phân phối Gaussian, đặt theo tên của Carl Friedrich Gauss, người đầu tiên mô tả biểu thức toán học của nó một cách chính thức.

4.1. PHÂN PHỐI CHUẨN



Hình 4.3: Hai phân phối chuẩn được trình bày trong Hình 4.2 nhưng được biểu diễn trên cùng một trục tọa độ và cùng một thang đo.

BÀI TẬP THỰC HÀNH 4.1

Viết ký hiệu rút gọn cho một phân phối chuẩn với ²

- Trung bình là 5 và độ lệch chuẩn là 3,
- Trung bình là -100 và độ lệch chuẩn là 10, và
- Trung bình là 2 và độ lệch chuẩn là 9.

4.1.2. Chuẩn hóa với điểm Z

Chúng ta thường muốn chuyển dữ liệu về cùng một thang đo chuẩn hóa, điều này giúp việc so sánh trở nên hợp lý hơn.

VÍ DỤ 4.2

Bảng 4.4 trình bày giá trị trung bình và độ lệch chuẩn của tổng điểm SAT và ACT. Phân phối điểm SAT và ACT đều gần như tuân theo phân phối chuẩn. Giả sử Ann đạt 1300 điểm SAT và Tom đạt 24 điểm ACT. Vậy ai có kết quả tốt hơn?

Chúng ta sử dụng độ lệch chuẩn làm thước đo để so sánh. Ann đạt cao hơn 1 độ lệch chuẩn so với trung bình điểm SAT: $1100 + 1 \times 200 = 1300$. Tom đạt cao hơn 0.5 độ lệch chuẩn so với trung bình điểm ACT: $21 + 0.5 \times 6 = 24$. Trong Hình 4.5, ta thấy rằng Ann có xu hướng làm bài tốt hơn so với số đồng học là Tom, nên điểm của Ann được đánh giá cao hơn.

	Điểm SAT	Điểm ACT
Trung bình	1100	21
Độ lệch chuẩn	200	6

Hình 4.4: Trung bình và độ lệch chuẩn của điểm SAT và ACT.

Ví dụ 4.2 sử dụng một kỹ thuật chuẩn hóa gọi là Điểm Z — phương pháp này thường được áp dụng cho các quan sát gần như phân phối chuẩn, nhưng cũng có thể sử dụng cho bất kỳ phân phối nào. **Điểm Z** của một quan sát được định nghĩa là số độ lệch chuẩn mà quan sát đó nằm phía trên hoặc phía dưới giá trị trung bình. Nếu một quan sát nằm cao hơn trung bình 1 độ lệch chuẩn, Z-score của nó là 1. Nếu một quan sát nằm *thấp* hơn trung bình 1.5 độ lệch chuẩn, Z-score của nó là -1.5. Giả sử x là một quan sát từ phân phối $N(\mu, \sigma)$, công thức toán học của Z-score được xác định như sau:

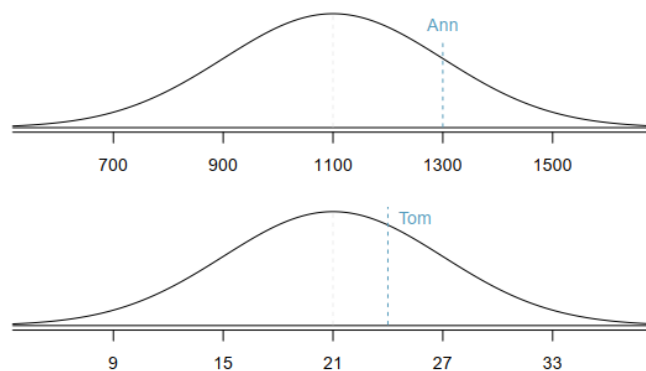
$$Z = \frac{x - \mu}{\sigma}$$

Sử dụng các giá trị $\mu_{\text{SAT}} = 1100$, $\sigma_{\text{SAT}} = 200$, và $x_{\text{Ann}} = 1300$, ta tính được **điểm Z của Ann** như sau:

$$Z_{\text{Ann}} = \frac{x_{\text{Ann}} - \mu_{\text{SAT}}}{\sigma_{\text{SAT}}} = \frac{1300 - 1100}{200} = 1$$

²(a) $N(\mu = 5, \sigma = 3)$. (b) $N(\mu = -100, \sigma = 10)$. (c) $N(\mu = 2, \sigma = 9)$.

4.1. PHÂN PHỐI CHUẨN



Hình 4.5: Điểm số của Ann và Tom được biểu diễn trên biểu đồ phân phối SAT và ACT.

ĐIỂM Z

Điểm Z của một quan sát là số độ lệch chuẩn mà quan sát đó nằm trên hoặc dưới trung bình. Chúng ta tính điểm Z cho một quan sát x từ phân phối có trung bình μ và độ lệch chuẩn σ theo công thức:

$$Z = \frac{x - \mu}{\sigma}$$

BÀI TẬP THỰC HÀNH 4.3

Sử dụng điểm ACT của Tom là 24, cùng với trung bình ACT là 21 và độ lệch chuẩn là 6, hãy tính điểm Z cho Tom.³

Các quan sát nằm phía trên trung bình luôn có điểm Z dương, trong khi các quan sát dưới trung bình có điểm Z âm. Nếu một quan sát bằng với giá trị trung bình (ví dụ như điểm SAT = 1100), thì điểm Z sẽ bằng 0.

BÀI TẬP THỰC HÀNH 4.4

Gọi X là biến ngẫu nhiên từ phân phối $N(\mu = 3, \sigma = 2)$, và giả sử chúng ta quan sát được $x = 5.19$. Tính điểm Z của x .

Sử dụng điểm Z để xác định x nằm trên hay dưới trung bình bao nhiêu độ lệch chuẩn.⁴

BÀI TẬP THỰC HÀNH 4.5

Chiều dài đầu của loài cáo túi đuôi chổi tuân theo phân phối chuẩn với trung bình $\mu = 92.6$ mm và độ lệch chuẩn $\sigma = 3.6$ mm. Tính điểm Z cho loài động vật này với chiều dài đầu bằng 95.4 mm and 85.8 mm.⁵

Chúng ta có thể sử dụng điểm Z để ước lượng mức độ bất thường của các quan sát. Một quan sát x_1 được cho là bất thường hơn một quan sát khác x_2 nếu giá trị tuyệt đối của điểm Z của nó lớn hơn giá trị tuyệt đối của điểm Z của quan sát còn lại: $|Z_1| > |Z_2|$. Phương pháp này đặc biệt hiệu quả khi phân phối là đối xứng, chẳng hạn như phân phối chuẩn.

BÀI TẬP THỰC HÀNH 4.6

Trong hai quan sát ở Bài tập 4.5, quan sát nào là **bất thường hơn**?⁶

³ $Z_{Tom} = \frac{x_{Tom} - \mu_{ACT}}{\sigma_{ACT}} = \frac{24 - 21}{6} = 0.5$

⁴ (a) Điểm Z của nó được tính theo công thức $Z = \frac{x - \mu}{\sigma} = \frac{5.19 - 3}{2} = \frac{2.19}{2} = 1.095$. (b) Quan sát x nằm *cao* hơn trung bình 1.095 độ lệch chuẩn. Ta biết điều đó vì Z là số dương.

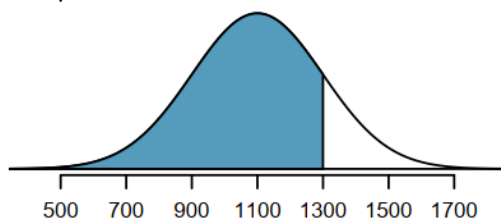
⁵ Với $x_1 = 95.4$ mm: $Z_1 = \frac{x_1 - \mu}{\sigma} = \frac{95.4 - 92.6}{3.6} = 0.78$. Với $x_2 = 85.8$ mm: $Z_2 = \frac{85.8 - 92.6}{3.6} = -1.89$.

⁶ Vì giá trị tuyệt đối của điểm Z trong quan sát thứ hai lớn hơn giá trị tuyệt đối của quan sát thứ nhất, nên quan sát thứ hai có chiều dài đầu bất thường hơn.

4.1. PHÂN PHỐI CHUẨN

4.1.3. Tìm diện tích phần đuôi

Khả năng xác định diện tích phần đuôi của một phân phối là một kỹ năng rất hữu ích trong thống kê. Ví dụ, bao nhiêu phần trăm người có điểm SAT thấp hơn điểm của Ann là 1300? Câu hỏi này cũng chính là tìm **bách phân vị** của Ann - tức là tỷ lệ phần trăm những người có điểm thấp hơn cô ấy. Ta có thể hình dung diện tích phần đuôi này như trong Hình 4.6, với vùng dưới đường cong được tô đậm.



Hình 4.6: Diện tích bên trái của Z biểu diễn tỷ lệ người có điểm thấp hơn Ann.

Có ba cách phổ biến để tìm diện tích phần đuôi.

1. Cách phổ biến nhất trong thực hành là dùng phần mềm thống kê. Ví dụ, trong ngôn ngữ lập trình **R**, ta có thể tính vùng tô bóng trong Hình 4.6 bằng lệnh:

```
> pnorm(1)
[1] 0.8413447
```

Theo tính toán này, vùng tô bóng phía dưới điểm 1300 tương ứng với 84.1% số người thi SAT có điểm $Z < 1$. Tổng quát hơn, ta cũng có thể nhập cụ thể giá trị trung bình và độ lệch chuẩn:

```
> pnorm(1300, mean = 1100, sd = 200)
[1] 0.8413447
```

Ngoài R, còn có nhiều phần mềm khác như Python, SAS, thậm chí Excel và Google Sheets cũng hỗ trợ tính toán này.

2. Một chiến lược phổ biến khác trong lớp học là dùng máy tính đồ thị, chẳng hạn như TI hoặc Casio. Các thao tác trên máy này khá cụ thể và dài dòng nên không trình bày chi tiết ở đây. Bạn có thể tìm hướng dẫn sử dụng máy tính để tìm diện tích phần đuôi qua thư viện video của OpenIntro:

www.openintro.org/videos

3. Một lựa chọn cuối cùng là dùng bảng xác suất, cách này đôi khi còn được dùng trong lớp học nhưng hiếm khi áp dụng trong thực tế. Phụ lục C.1 cung cấp một bảng như vậy cùng với hướng dẫn cách sử dụng.

Trong chương này, mỗi khi giải bài toán liên quan đến phân phối chuẩn, ta sẽ luôn bắt đầu bằng việc tính điểm Z. Lý do là vì trong Chương 5, bạn sẽ gặp những khái niệm tương tự gọi là thống kê kiểm định (test statistics) — trong nhiều trường hợp, chúng tương đương với điểm Z.

4.1. PHÂN PHỐI CHUẨN

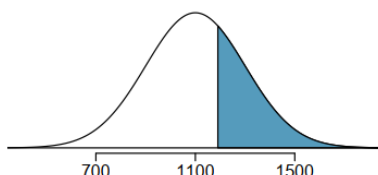
4.1.4. Ví dụ về xác suất trong phân phối chuẩn

Tổng điểm SAT được xấp xỉ rất tốt bởi một mô hình phân phối chuẩn, $N(\mu = 1100, \sigma = 200)$.

VÍ DỤ 4.7

Shannon là một thí sinh SAT được chọn ngẫu nhiên, và ta không biết gì về khả năng học tập của cô ấy. Vậy xác suất để Shannon đạt ít nhất 1190 điểm trong kỳ thi SAT là bao nhiêu?

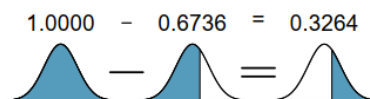
Bước đầu tiên, luôn luôn vẽ và ghi chú một biểu đồ của phân phối chuẩn. (Hình vẽ không cần chính xác tuyệt đối, chỉ cần hỗ trợ trực quan là đủ). Trong bài toán này, ta quan tâm đến xác suất Shannon đạt điểm cao hơn 1190, vì vậy ta tô bóng phần đuôi phía bên phải (đuôi trên):



E Hình minh họa cho thấy giá trị trung bình và các mốc cách trung bình 2 độ lệch chuẩn. Cách đơn giản nhất để tìm diện tích được tô bóng dưới đường cong là tính điểm Z tương ứng với giá trị cắt (cutoff). Với $\mu = 1100$, $\sigma = 200$, và $x = 1190$, ta tính

$$Z = \frac{x - \mu}{\sigma} = \frac{1190 - 1100}{200} = \frac{90}{200} = 0.45$$

Sử dụng phần mềm thống kê (hoặc bất kỳ phương pháp nào khác), ta tìm được diện tích bên trái $Z = 0.45$ là 0.6736. Để tính diện tích bên phải $Z = 0.45$, tức là xác suất Shannon đạt ít nhất 1190 điểm, ta lấy $1.0000 - 0.6736 = 0.3264$.



Vậy, xác suất Shannon đạt ít nhất 1190 điểm SAT là 0.3264.

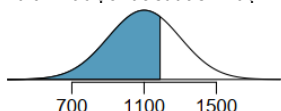
LUÔN LUÔN VẼ HÌNH TRƯỚC, SAU ĐÓ MỚI TÍNH ĐIỂM Z

Đối với bất kỳ bài toán xác suất nào liên quan đến phân phối chuẩn, *luôn luôn và luôn luôn* vẽ và ghi chú đường cong chuẩn trước tiên, đồng thời tô bóng vùng diện tích liên quan. Hình vẽ sẽ giúp bạn ước lượng sơ bộ xác suất. Sau khi hoàn thành hình minh họa thể hiện tình huống, hãy xác định điểm Z cho giá trị cần quan tâm.

BÀI TẬP THỰC HÀNH 4.8

G Nếu xác suất Shannon đạt ít nhất 1190 điểm là 0.3264, thì xác suất cô ấy đạt ít hơn 1190 điểm là bao nhiêu? Vẽ đường cong chuẩn, tô bóng vùng phía bên trái của 1190 thay vì bên phải như ví dụ trên.⁷

⁷ Ta tính được xác suất ở Ví dụ 4.7: 0.6736.



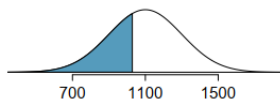
4.1. PHÂN PHỐI CHUẨN

VÍ DỤ 4.9

Edward đạt 1030 điểm SAT. Vậy thứ bậc phần trăm (percentile) của cậu ấy là bao nhiêu?

Trước hết, cần có hình minh họa. Thứ bậc phần trăm của Edward chính là tỷ lệ phần trăm những người có điểm thấp hơn 1030 – tức là diện tích bên trái điểm 1030.

E



Xác định: trung bình $\mu = 1100$, độ lệch chuẩn $\sigma = 200$, và điểm cần xét $x = 1030$. Ta tính được điểm Z:

$$Z = \frac{x - \mu}{\sigma} = \frac{1030 - 1100}{200} = -0.35$$

Sử dụng phần mềm thống kê, ta tìm được diện tích phía bên trái (tail area) là 0.3632. Vậy Edward nằm ở bậc 36 phần trăm.

BÀI TẬP THỰC HÀNH 4.10

G

Sử dụng kết quả từ Ví dụ 4.9 hãy tính tỷ lệ phần trăm người thi SAT có điểm cao hơn Edward.

Hãy vẽ thêm một hình mới.⁸

TÌM DIỆN TÍCH BÊN PHẢI

Nhiều phần mềm thống kê mặc định trả về diện tích bên trái của điểm Z. Nếu bạn cần tìm diện tích bên phải, hãy lấy 1 trừ đi diện tích bên trái.

BÀI TẬP THỰC HÀNH 4.11

G

Stuart đạt 1500 điểm SAT. Hãy vẽ hình cho từng phần sau.

- Thứ bậc phần trăm của Stuart là bao nhiêu?
- Tỷ lệ phần trăm người thi SAT đạt điểm cao hơn Stuart?⁹

Dựa trên mẫu gồm 100 người đàn ông, chiều cao trung bình của nam giới trưởng thành ở Mỹ gần như tuân theo phân phối chuẩn với trung bình 70.0-inch và độ lệch chuẩn 3.3 inch.

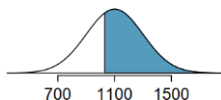
BÀI TẬP THỰC HÀNH 4.12

G

Mike cao 5'7" và Jose cao 6'4". Cả hai đều sống tại Mỹ.

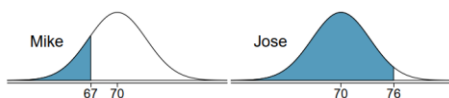
- Thứ bậc phần trăm chiều cao của Mike là bao nhiêu?
- Thứ bậc phần trăm chiều cao của Jose là bao nhiêu? Cũng vẽ hình minh họa cho từng phần.¹⁰

⁸ Nếu Edward làm bài tốt hơn 36% số người thi SAT, thì khoảng 64% người còn lại đã làm tốt hơn cậu ấy.



⁹ Chúng tôi để phần vẽ hình cho bạn tự thực hiện. (a) $Z = \frac{1500 - 1100}{200} = 2 \Rightarrow$ Diện tích bên trái = 0.9772. (b) $1 - 0.9772 = 0.0228 \Rightarrow$ Tức là chỉ 2.28% số người thi có điểm cao hơn Stuart.

¹⁰ Đầu tiên, chuyển đổi chiều cao sang inch: Mike = 67 inch, Jose = 76 inch. Hình minh họa được trình bày bên dưới. (a) $Z_{Mike} = \frac{67 - 70}{3.3} = -0.91 \Rightarrow$ Phần trăm $\approx 0.1814 \Rightarrow$ Mike ở bậc 18.14. (b) $Z_{Jose} = \frac{76 - 70}{3.3} = 1.82 \Rightarrow$ Phần trăm $\approx 0.9656 \Rightarrow$ Jose ở bậc 96.56.



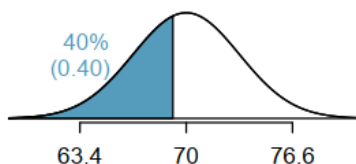
4.1. PHÂN PHỐI CHUẨN

Các bài tập trước tập trung vào việc tìm phần trăm vị trí (phần đuôi dưới) hoặc phần đuôi trên tương ứng với một giá trị quan sát cụ thể. Vậy nếu bạn muốn biết giá trị quan sát tương ứng với một phần trăm vị trí nhất định thì sao?

VÍ DỤ 4.13

Chiều cao của Erik nằm ở bậc phần trăm thứ 40. Vậy Erik cao bao nhiêu?

Như mọi khi, trước tiên hãy vẽ hình minh họa.



- E** Trong trường hợp này, xác suất đuôi dưới đã được biết (0.40), có thể được tô bóng trên đồ thị. Ta cần tìm giá trị quan sát tương ứng với phần diện tích này. Bước đầu tiên là xác định Z-score tương ứng với bậc 40. Dùng phần mềm thống kê, ta thu được $Z \approx -0.25$.

Biết rằng $Z_{\text{Erik}} = -0.25$ và các tham số dân số là $\mu = 70$ và $\sigma = 3.3$ inch, ta thiết lập công thức Z để tìm chiều cao chưa biết của Erik, ký hiệu là x_{Erik} :

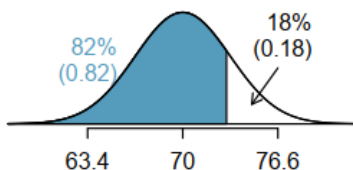
$$Z = \frac{x_{\text{Erik}} - \mu}{\sigma} = \frac{x_{\text{Erik}} - 70}{3.3} = -0.25$$

Giải phương trình trên, ta tìm được $x_{\text{Erik}} = 69.18$ inch. Tức là Erik cao khoảng 5 feet 9 inch.

VÍ DỤ 4.14

Chiều cao nam giới trưởng thành ở bậc phần trăm thứ 82 là bao nhiêu? Nội dung 1

Tiếp tục, vẽ hình trước.



E

Tiếp theo, tìm Z-score tương ứng với bậc 82, giá trị này là dương và có thể tìm bằng phần mềm: $Z = 0.92$.

Cuối cùng, sử dụng công thức Z với các giá trị $\mu = 70$, $\sigma = 3.3$, và $Z = 0.92$:

$$Z = \frac{x - \mu}{\sigma} = \frac{x - 70}{3.3} = 0.92$$

Giải phương trình, ta được $x = 73.04$ inch, tức khoảng 6 feet 1 inch — đây là chiều cao tương ứng với bậc 82.

BÀI TẬP THỰC HÀNH 4.15

G

Điểm SAT tuân theo phân phối chuẩn $N(1100, 200)$.¹¹

- Điểm SAT ở bậc 95 là bao nhiêu?
- Điểm SAT ở bậc 97.5 là bao nhiêu?

¹¹ Đáp án: (a) $Z_{.95} = 1.6449 \Rightarrow \text{Điểm SAT} = 1100 + 1.6449 \times 200 = 1429$. (b) $Z_{.97.5} = 1.96 \Rightarrow \text{Điểm SAT} = 1100 + 1.96 \times 200 = 1492$.

4.1. PHÂN PHỐI CHUẨN

BÀI TẬP THỰC HÀNH 4.16

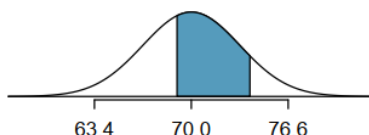
- (G) Chiều cao của nam giới trưởng thành tuân theo phân phối chuẩn $N(70.0", 3.3")$.¹²
- (a) Xác suất một người nam trưởng thành được chọn ngẫu nhiên có chiều cao ít nhất là 6'2" (74 inches) là bao nhiêu?
- (b) Xác suất một người nam trưởng thành có chiều cao thấp hơn 5'9" (69 inches) là bao nhiêu?

VÍ DỤ 4.17

Xác suất một người nam trưởng thành ngẫu nhiên có chiều cao nằm giữa 5'9" và 6'2" là bao nhiêu?

Chiều cao này tương ứng với 69 inch và 74 inch. Trước tiên, vẽ đồ thị minh họa. Vùng quan tâm không còn là phần đuôi trên hay dưới, mà là vùng nằm giữa hai giá trị.

(E)



Tổng diện tích dưới đường cong là 1. Nếu ta tìm diện tích của hai phần đuôi không được tô bóng (từ phần THỰC HÀNH CÓ HƯỚNG DẪN 4.16, các giá trị tương ứng là 0.3821 và 0.1131), thì ta có thể tính phần diện tích giữa:

$$1.0000 - 0.3821 - 0.1131 = 0.5048$$

Tức là, xác suất chiều cao nằm trong khoảng từ 5'9" đến 6'2" là 0.5048 (50.48%).

BÀI TẬP THỰC HÀNH 4.18

- (G) Điểm SAT tuân theo phân phối chuẩn $N(1100, 200)$. Bao nhiêu phần trăm thí sinh có điểm SAT từ 1100 đến 1400?¹³

BÀI TẬP THỰC HÀNH 4.19

- (G) Chiều cao nam giới trưởng thành tuân theo $N(70.0", 3.3")$. Bao nhiêu phần trăm nam giới trưởng thành có chiều cao từ 5'5" đến 5'7"?¹⁴

¹² (a) $Z = 1.21 \rightarrow 0.8869$, sau đó lấy 1 trừ đi giá trị này để được: $1 - 0.8869 = 0.1131$. (b) $Z = -0.30 \rightarrow 0.3821$.

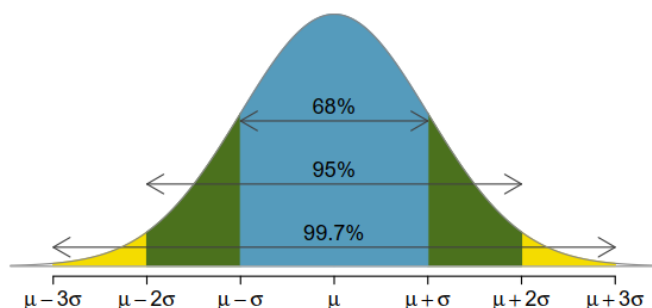
¹³ Đây là cách giải rút gọn. (Nhớ vẽ hình minh họa!) Trước tiên, tìm phần trăm số người có điểm dưới 1100 và phần trăm số người có điểm trên 1400: $Z_{1100} = 0.00 \rightarrow 0.5000$ (diện tích bên trái), $Z_{1400} = 1.5 \rightarrow 0.0668$ (diện tích bên phải). Kết quả cuối cùng: $1.0000 - 0.5000 - 0.0668 = 0.4332$

¹⁴ Chiều cao 5'5" là 65 inches ($Z = -1.52$). Chiều cao 5'7" là 67 inches ($Z = -0.91$). Giải pháp số học: $1.000 - 0.0643 - 0.8186 = 0.1171$, tức là 11.71%.

4.1. PHÂN PHỐI CHUẨN

4.1.5. Quy tắc 68-95-99.7

Đây là một quy tắc ước lượng nhanh hữu ích để xác định xác suất một giá trị rơi vào khoảng 1, 2 hoặc 3 độ lệch chuẩn so với trung bình trong phân phối chuẩn. Quy tắc này rất hữu ích trong nhiều tình huống thực tế, đặc biệt khi bạn cần ước lượng nhanh mà không dùng máy tính hay bảng Z.



Hình 4.7: Xác suất để rơi vào vùng cách trung bình 1, 2, và 3 độ lệch chuẩn trong phân phối chuẩn.

BÀI TẬP THỰC HÀNH 4.20

- G** Dùng phần mềm, máy tính hoặc bảng xác suất để xác nhận rằng khoảng 68%, 95%, và 99.7% các giá trị sẽ rơi vào vùng cách trung bình lần lượt là 1, 2 và 3 độ lệch chuẩn trong phân phối chuẩn. Ví dụ: Tìm diện tích giữa $Z = -1$ và $Z = 1$, sẽ xấp xỉ 0.68. Tương tự, diện tích giữa $Z = -2$ và $Z = 2$ xấp xỉ 0.95.¹⁵

Giá trị có thể rơi cách trung bình 4, 5 hoặc nhiều độ lệch chuẩn hơn, nhưng điều này rất hiếm nếu dữ liệu gần chuẩn. Xác suất cách hơn 4 độ lệch chuẩn là khoảng 1/15,000. Với 5 và 6 độ lệch chuẩn, xác suất là khoảng 1/2 triệu và 1/500 triệu.

BÀI TẬP THỰC HÀNH 4.21

- G** Điểm SAT tuân theo phân phối chuẩn với trung bình $\mu = 1100$ và độ lệch chuẩn $\sigma = 200$.¹⁶
- Khoảng bao nhiêu phần trăm người thi đạt điểm từ 700 đến 1500?
 - Bao nhiêu phần trăm người thi đạt điểm từ 1100 đến 1500?

¹⁵ Đầu tiên hãy vẽ hình minh họa. Sử dụng phần mềm, chúng ta thu được 0.6827 nằm trong khoảng 1 độ lệch chuẩn so với trung bình, 0.9545 nằm trong khoảng 2 độ lệch chuẩn, và 0.9973 nằm trong khoảng 3 độ lệch chuẩn.

¹⁶ (a) 700 và 1500 tương ứng với hai độ lệch chuẩn dưới và trên trung bình, điều đó có nghĩa là khoảng 95% người thi sẽ đạt điểm trong khoảng từ 700 đến 1500. (b) Chúng ta đã biết rằng khoảng từ 700 đến 1500 chiếm khoảng 95% số người thi. Vì phân phối chuẩn đối xứng, nên nhóm người thi này được chia đều qua trung tâm phân phối là 1100, vì vậy 95% chia đôi thành 47.5% số người thi đạt điểm từ 1100 đến 1500.

4.1. PHÂN PHỐI CHUẨN

Bài tập

4.1 Diện tích dưới đường cong – Phần I. Tìm phần trăm trong phân phối chuẩn $N(\mu = 0, \sigma = 1)$ cho mỗi vùng sau. Nhớ vẽ đồ thị.

- (a) $Z < -1.35$ (b) $Z > 1.48$ (c) $-0.4 < Z < 1.5$ (d) $|Z| > 2$

4.2 Diện tích dưới đường cong - Phần II. Tìm phần trăm trong phân phối chuẩn $N(\mu = 0, \sigma = 1)$ cho mỗi vùng sau. Nhớ vẽ đồ thị.

- (a) $Z > -1.13$ (b) $Z < 0.18$ (c) $Z > 8$ (d) $|Z| < 0.5$

4.3 Điểm GRE - Phần I. Sophia đạt 160 điểm phần Ngôn ngữ và 157 điểm phần Toán. Trung bình phần Ngôn ngữ là 151 ($\sigma = 7$), phần Toán là 153 ($\sigma = 7.67$), các phân phối gần chuẩn.

- Viết ký hiệu tắt cho hai phân phối chuẩn này.
- Tính Z-score của Sophia ở cả hai phần. Vẽ đường cong chuẩn và đánh dấu các điểm Z này.
- Các điểm Z này nói lên điều gì?
- Sophia làm tốt hơn ở phần nào?
- Tính phần trăm xếp hạng (percentile) của cô ấy.
- Có bao nhiêu phần trăm người làm tốt hơn cô ấy?
- Vì sao không thể chỉ so sánh điểm số gốc?
- Nếu phân phối không gần chuẩn, kết quả từ (b) đến (f) có thay đổi không?

4.4 Thời gian thi Triathlon - Phần I. Trong các cuộc thi ba môn phối hợp (triathlon), người thi thường được chia theo nhóm tuổi và giới tính. Hai người bạn là Leo và Mary đều đã hoàn thành cuộc thi Hermosa Beach Triathlon, trong đó Leo thi ở nhóm Nam, 30-34 tuổi, còn Mary thi ở nhóm Nữ, 25-29 tuổi. Leo hoàn thành cuộc thi trong 1 giờ 22 phút 28 giây (tức 4948 giây), Mary hoàn thành trong 1 giờ 31 phút 53 giây (tức 5513 giây). Rõ ràng là Leo hoàn thành nhanh hơn, nhưng họ tò mò về xếp hạng của mình trong nhóm tương ứng. Bạn có thể giúp họ không?

Dưới đây là thông tin về kết quả của các nhóm:

Thời gian hoàn thành của nhóm Nam, 30–34 tuổi có trung bình là 4313 giây và độ lệch chuẩn là 583 giây.

Thời gian hoàn thành của nhóm Nữ, 25–29 tuổi có trung bình là 5261 giây và độ lệch chuẩn là 807 giây.

Các phân phối thời gian hoàn thành của cả hai nhóm đều xấp xỉ chuẩn.

Lưu ý: Thành tích tốt hơn nghĩa là thời gian hoàn thành ngắn hơn.

- Viết ký hiệu rút gọn cho hai phân phối chuẩn này.
- Z-score của Leo và Mary là bao nhiêu? Những Z-score này cho ta biết điều gì?
- Ai trong hai người, Leo hay Mary, xếp hạng cao hơn trong nhóm của mình? Giải thích.
- Leo hoàn thành nhanh hơn bao nhiêu phần trăm vận động viên trong nhóm của anh ấy?
- Mary hoàn thành nhanh hơn bao nhiêu phần trăm vận động viên trong nhóm của cô ấy?
- Nếu phân phối thời gian hoàn thành không xấp xỉ chuẩn, các câu trả lời từ (b) đến (e) có thay đổi không? Giải thích.

4.5 Điểm GRE - Phần II. Trong Bài tập 4.3 ta đã thấy hai phân phối điểm GRE: $N(\mu = 151, \sigma = 7)$ cho phần Ngôn ngữ và $N(\mu = 153, \sigma = 7.67)$ cho phần Toán học. Hãy sử dụng thông tin này để tính:

- Mức điểm của một học sinh nằm ở phân vị thứ 80 (80th percentile) trong phần Toán học.
- Mức điểm của một học sinh có kết quả tệ hơn 70% người thi khác trong phần Ngôn ngữ.

4.1. PHÂN PHỐI CHUẨN

4.6 Thời gian thi Triathlon - Phần II. Trong Bài tập 4.4 ta đã có hai phân phối thời gian: $N(\mu = 4313, \sigma = 583)$ cho nam 30–34 tuổi và $N(\mu = 5261, \sigma = 807)$ cho nữ 25–29 tuổi. Thời gian được tính theo đơn vị giây. Dựa trên thông tin này, hãy tính:

- (a) Thời gian giới hạn cho 5% người có thời gian hoàn thành nhanh nhất trong nhóm nam.
- (b) Thời gian giới hạn cho 10% người có thời gian hoàn thành chậm nhất trong nhóm nữ.

4.7 Thời tiết ở Los Angeles – Phần I. Nhiệt độ cao trung bình hằng ngày trong tháng 6 tại Los Angeles là 77°F , với độ lệch chuẩn 5°F . Giả sử nhiệt độ trong tháng 6 xấp xỉ phân phối chuẩn.

- (a) Xác suất để một ngày ngẫu nhiên trong tháng 6 có nhiệt độ từ 83°F trở lên là bao nhiêu?
- (b) Những ngày lạnh nhất 10% trong tháng 6 có nhiệt độ cao nhất vào ban ngày là bao nhiêu?

4.8 CAPM. Mô hình định giá tài sản vốn (CAPM - Capital Asset Pricing Model) là một mô hình tài chính giả định rằng lợi nhuận danh mục đầu tư phân phối chuẩn. Giả sử một danh mục đầu tư có lợi nhuận trung bình hằng năm là 14.7%, với độ lệch chuẩn 33%. Lợi nhuận = 0% nghĩa là giá trị không đổi. Lợi nhuận âm nghĩa là lỗ vốn. Lợi nhuận dương nghĩa là có lời.

- (a) Bao nhiêu phần trăm số năm danh mục bị lỗ, tức lợi nhuận $< 0\%$?
- (b) Mức lợi nhuận giới hạn để nằm trong nhóm cao nhất 15% hàng năm là bao nhiêu?

4.9 Thời tiết ở Los Angeles - Phần II. Bài tập 4.7 cho biết nhiệt độ cao trung bình hằng ngày trong tháng 6 ở Los Angeles là 77°F , độ lệch chuẩn 5°F , và phân phối gần như chuẩn. Sử dụng công thức chuyển đổi từ Fahrenheit ($^\circ\text{F}$) sang Celsius ($^\circ\text{C}$):

$$C = (F - 32) \times 5/9$$

- (a) Viết mô hình phân phối xác suất cho nhiệt độ ($^\circ\text{C}$) trong tháng 6 ở LA.
- (b) Xác suất để có một ngày $\geq 28^\circ\text{C}$ (xấp xỉ 83°F) trong tháng 6 ở LA là bao nhiêu? Tính bằng mô hình $^\circ\text{C}$ từ (a).
- (c) Kết quả trong câu (b) và câu (a) của Bài tập 4.7 có giống nhau không? Bạn có bất ngờ không? Giải thích.
- (d) Ước tính khoảng tứ phân vị (IQR) của nhiệt độ ($^\circ\text{C}$) trong tháng 6 ở LA.

4.10 Tìm độ lệch chuẩn. Mức cholesterol ở phụ nữ 20–34 tuổi phân phối gần chuẩn, với trung bình là 185 mg/dL. Phụ nữ có cholesterol trên 220 mg/dL được xem là cao và khoảng 18.5% phụ nữ rơi vào nhóm này. Vậy độ lệch chuẩn của phân phối mức cholesterol ở nhóm phụ nữ 20–34 tuổi này là bao nhiêu?

4.2. PHÂN PHỐI HÌNH HỌC

4.2. Phân phối hình học - Geometric distribution

Chúng ta nên mong đợi bao lâu (hoặc bao nhiêu lần thử) để đồng xu xuất hiện mặt ngửa? Hoặc cần bao nhiêu lần gieo xúc xắc cho đến khi ra mặt số 1? Những câu hỏi này có thể được trả lời bằng cách sử dụng phân phối hình học (geometric distribution). Trước tiên, chúng ta sẽ mô hình hóa từng lần thử – chẳng hạn như một lần tung đồng xu hoặc gieo xúc xắc – bằng phân phối Bernoulli, sau đó kết hợp chúng với các công cụ xác suất đã học (Chương 3) để xây dựng phân phối hình học.

4.2.1. Phân phối Bernoulli

Nhiều gói bảo hiểm y tế ở Hoa Kỳ có khoản khấu trừ (deductible), trong đó người được bảo hiểm phải tự chi trả chi phí cho đến khi đạt mức khấu trừ, và sau đó các chi phí vượt quá mức này sẽ được chia sẻ giữa cá nhân và công ty bảo hiểm trong phần còn lại của năm.

Giả sử một công ty bảo hiểm y tế phát hiện rằng 70% số người mà họ bảo hiểm không vượt quá mức khấu trừ trong một năm bất kỳ. Mỗi người trong số này có thể được coi là **một lần thử** (trial). Chúng ta gọi một người là **“thành công”** (success) nếu chi phí chăm sóc sức khỏe của họ không vượt quá mức khấu trừ. Ngược lại, nếu họ vượt quá mức này trong năm thì được gọi là **“thất bại”** (failure). Vì 70% số người không vượt mức khấu trừ, nên ta ký hiệu **xác suất thành công** là $p = 0.7$. Xác suất thất bại đôi khi được ký hiệu là $q = 1 - p$, tức là 0.3 trong ví dụ về bảo hiểm này.

Khi một lần thử chỉ có hai kết quả có thể xảy ra – thường được gọi là thành công hoặc thất bại – thì nó được gọi là **biến ngẫu nhiên Bernoulli**. Chúng ta đã chọn gọi người không vượt mức khấu trừ là “thành công” và những người khác là “thất bại”. Tuy nhiên, chúng ta cũng có thể hoán đổi hai nhãn này. Khung toán học mà chúng ta xây dựng không phụ thuộc vào việc nhãn nào được gọi là thành công hay thất bại, miễn là chúng ta nhất quán trong suốt quá trình.

Các biến ngẫu nhiên Bernoulli thường được ký hiệu là 1 cho thành công và 0 cho thất bại. Cách biểu diễn này không chỉ thuận tiện khi nhập dữ liệu mà còn rất hữu ích về mặt toán học. Giả sử chúng ta quan sát 10 lần thử sau:

1 1 1 0 1 0 0 1 1 0

Khi đó, **tỷ lệ mẫu** - sample proportion (ký hiệu là $\hat{p}$) chính là trung bình mẫu của các quan sát này:

$$\hat{p} = \frac{\# \text{ thành công}}{\# \text{ lần thử}} = \frac{1 + 1 + 1 + 0 + 1 + 0 + 0 + 1 + 1 + 0}{10} = 0,6$$

Việc khảo sát toán học về các biến ngẫu nhiên Bernoulli có thể được mở rộng hơn nữa. Vì các kết quả 0 và 1 là những giá trị số, nên chúng ta có thể định nghĩa giá trị trung bình và độ lệch của một biến ngẫu nhiên Bernoulli (Xem Bài tập 4.15 và 4.16.)

Note

BIẾN NGẪU NHIÊN BERNOULLI

Nếu X là một biến ngẫu nhiên nhận giá trị 1 với xác suất thành công là p và giá trị 0 với xác suất $1 - p$, thì X là một biến ngẫu nhiên Bernoulli với giá trị trung bình và độ lệch chuẩn (standard deviation)

$$\mu = p$$

$$\sigma = \sqrt{p(1-p)}$$

Nhìn chung, sẽ hữu ích nếu xem một biến ngẫu nhiên Bernoulli như một quá trình ngẫu nhiên chỉ có hai kết quả: thành công hoặc thất bại. Sau đó, chúng ta xây dựng khung toán học bằng cách sử dụng các nhãn số 1 cho thành công và 0 cho thất bại.

4.2. PHÂN PHỐI HÌNH HỌC

4.2.2. Phân phối hình học

Phân phối hình học được sử dụng để mô tả cần bao nhiêu lần thử cho đến khi quan sát được một lần thành công đầu tiên.

Hãy cùng xem xét ví dụ đầu tiên.

VÍ DỤ 4.22

Giả sử chúng ta đang làm việc tại một công ty bảo hiểm và cần tìm một trường hợp mà người đó không vượt mức khấu trừ (deductible) để làm nghiên cứu tình huống. Nếu xác suất một người sẽ không vượt mức khấu trừ là 0.7 và chúng ta chọn người một cách ngẫu nhiên, thì khả năng người đầu tiên không vượt mức khấu trừ - tức là một thành công - là bao nhiêu? Người thứ hai thì sao? Người thứ ba? Vậy nếu chúng ta chọn ngẫu nhiên $n - 1$ trường hợp trước khi tìm được thành công đầu tiên - tức là người thứ n là người thành công đầu tiên - thì sao? (Nếu thành công đầu tiên là ở người thứ năm, thì $n = 5$.)

E

Xác suất dừng lại sau người đầu tiên chính là xác suất người đầu tiên không vượt mức khấu trừ: 0.7. Xác suất người thứ hai là người đầu tiên không vượt mức khấu trừ là:

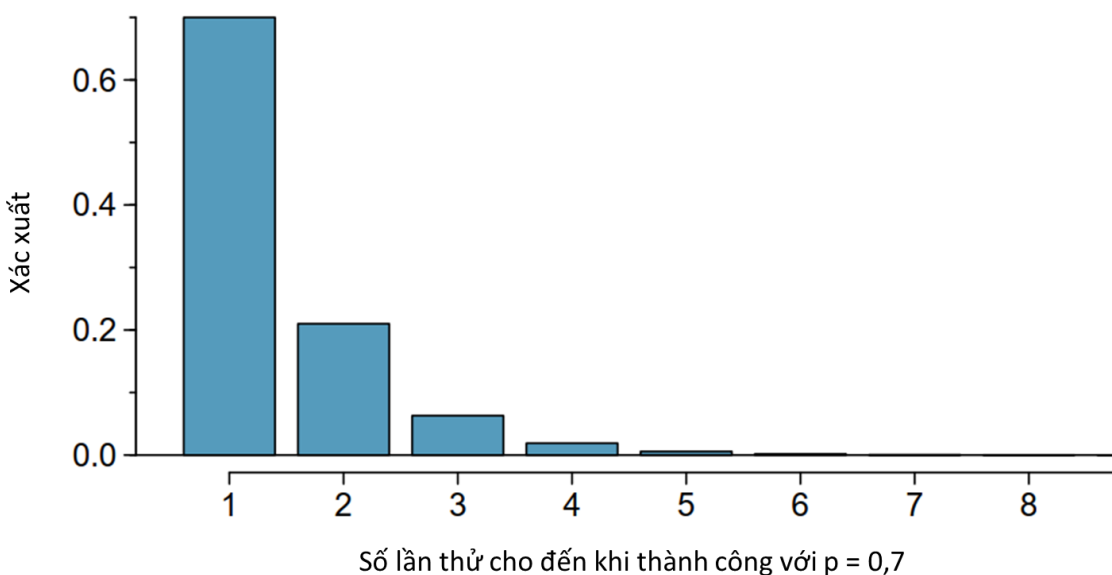
$$P(\text{người thứ hai là người đầu tiên không vượt mức khấu trừ}) \\ = P(\text{người đầu tiên vượt, người thứ hai không vượt}) = (0,3) * (0,7) = 0,21.$$

Tương tự, xác suất người thứ ba là người đầu tiên không vượt mức khấu trừ là: $(0,3)(0,3)(0,7) = 0,063$.

Nếu thành công đầu tiên là ở người thứ n , thì sẽ có $n - 1$ thất bại và cuối cùng là 1 thành công, tương ứng với xác suất: $(0,3)^{n-1} \times (0,7)$. Điều này cũng tương đương với: $(1 - 0,7)^{n-1}(0,7)$.

Ví dụ 4.22 minh họa cho **phân phối hình học**, mô tả thời gian cho đến khi xảy ra một lần thành công đối với các biến ngẫu nhiên Bernoulli **độc lập và phân phối giống hệt nhau (iid)**. Trong trường hợp này, tính độc lập có nghĩa là các cá nhân trong ví dụ không ảnh hưởng lẫn nhau, và giống hệt nhau nghĩa là mỗi người đều có cùng một xác suất thành công.

Phân phối hình học từ Ví dụ 4.22 được thể hiện trong Hình 4.8. Nhìn chung, các xác suất trong phân phối hình học **giảm theo cấp số nhân** một cách nhanh chóng.



Hình 4.8: Phân phối hình học khi xác suất thành công là $p = 0,7$.

Mặc dù tài liệu này sẽ không chứng minh các công thức cho giá trị trung bình (kỳ vọng) số lần thử cần thiết để tìm ra thành công đầu tiên, cũng như độ lệch chuẩn hoặc phương sai của phân phối này, chúng tôi sẽ trình bày các công thức tổng quát cho từng đại lượng trên.

4.2. PHÂN PHỐI HÌNH HỌC

PHÂN PHỐI HÌNH HỌC

Nếu xác suất thành công trong một lần thử là p và xác suất thất bại là $1 - p$, thì xác suất để tìm thấy thành công đầu tiên tại lần thử thứ n được cho bởi:

$$(1 - p)^{n-1}p$$

Trung bình (tức là giá trị kỳ vọng), phương sai và độ lệch chuẩn của thời gian chờ này được xác định bởi:

$$\mu = \frac{1}{p}$$

$$\sigma^2 = \frac{1-p}{p^2}$$

$$\sigma = \sqrt{\frac{1-p}{p^2}}$$

Không phải ngẫu nhiên mà chúng ta dùng ký hiệu μ cho cả giá trị trung bình và kỳ vọng. Giá trị trung bình và kỳ vọng là một và giống nhau.

Trung bình, phải mất $1/p$ lần thử để có được một lần thành công theo phân phối hình học. Kết quả toán học này phù hợp với trực giác: nếu xác suất thành công cao (ví dụ 0.8), thì ta thường không phải đợi lâu để có thành công: $1/0.8 = 1.25$ lần thử trung bình. Ngược lại, nếu xác suất thành công thấp (ví dụ 0.1), thì ta cần nhiều lần thử hơn để đạt được thành công: $1/0.1 = 10$ lần thử.

BÀI TẬP THỰC HÀNH 4.23

G Xác suất để một trường hợp cụ thể không vượt quá mức khấu trừ là 0.7. Nếu chúng ta kiểm tra từng trường hợp cho đến khi tìm được một người không vượt mức khấu trừ, thì chúng ta dự kiến phải kiểm tra bao nhiêu người? ¹⁷

VÍ DỤ 4.24

Xác suất để chúng ta có thành công đầu tiên trong 3 trường hợp đầu tiên là bao nhiêu?

E Đây là xác suất rằng thành công đầu tiên xảy ra ở trường hợp thứ nhất ($n = 1$), thứ hai ($n = 2$), hoặc thứ ba ($n = 3$) – tức là ba kết quả rời nhau. Vì các cá nhân được lấy mẫu ngẫu nhiên từ một quần thể lớn nên các trường hợp là độc lập. Ta tính xác suất của từng trường hợp rồi cộng lại:

$P(n = 1, 2, \text{ hoặc } 3)$
$= P(n = 1) + P(n = 2) + P(n = 3)$
$= (0,3)^{1-1}(0,7) + (0,3)^{3-1}(0,7) + (0,3)^{3-1}(0,7)$
$= 0,973$

Xác suất là 0,973 khả năng ta sẽ có một trường hợp thành công trong 3 thử đầu tiên.

BÀI TẬP THỰC HÀNH 4.25

G Hãy tìm cách thông minh hơn để giải Ví dụ 4.24. Chứng minh rằng bạn vẫn nhận được kết quả như cũ. ¹⁸

¹⁷ Chúng ta sẽ dự đoán sẽ có khoảng $1/0,7 \approx 1,43$ cá nhân tìm thấy thành công đầu tiên.

4.2. PHÂN PHỐI HÌNH HỌC

¹⁸ Đầu tiên, hãy tính xác suất của biến cố đối lập: $P(\text{không thành công trong 3 lần thử đầu tiên}) = 0,33 = 0,027$. Tiếp theo, tính một trừ đi xác suất này: $1 - P(\text{không thành công trong 3 lần thử}) = 1 - 0,027 = 0,973$.

VÍ DỤ 4.26

(E)

Giả sử một công ty bảo hiểm xe đã xác định rằng 88% tài xế sẽ không vượt mức khấu trừ trong một năm. Nếu một nhân viên của công ty chọn ngẫu nhiên hồ sơ của các tài xế cho đến khi tìm thấy người không vượt khấu trừ, thì số người trung bình cần kiểm tra là bao nhiêu? Độ lệch chuẩn của số hồ sơ cần kiểm tra là bao nhiêu?

Trong ví dụ này, một lần thành công lại là khi ai đó không vượt mức khấu trừ bảo hiểm, với xác suất là $p = 0,88$. Số người dự kiến cần kiểm tra là $1/p = 1/0,88 = 1,14$ và độ lệch chuẩn là $\sqrt{(1-p)/p^2} = 0,39$.

BÀI TẬP THỰC HÀNH 4.27

(G)

Sử dụng các kết quả từ Ví dụ 4.26, $\mu = 1.14$ và $\sigma = 0.39$, có phù hợp để sử dụng mô hình chuẩn (normal model) để tìm tỷ lệ các thí nghiệm kết thúc trong 3 lần thử hoặc ít hơn không? ¹⁹

Giả định về tính độc lập là rất quan trọng đối với việc mô tả bằng phân phối hình học. Về mặt toán học, chúng ta có thể thấy rằng để xây dựng xác suất thành công ở lần thử thứ n , chúng ta phải sử dụng Quy tắc nhân cho các quá trình độc lập. Rất khó để khái quát mô hình hình học cho các biến phụ thuộc.

4.2. PHÂN PHỐI HÌNH HỌC

¹⁹ Không. Phân phối hình học luôn lệch phải và không bao giờ có thể được ước lượng tốt bằng mô hình phân phối chuẩn.

Bài tập

4.11 Có phải phân phối Bernoulli không? Xác định xem mỗi lần thử sau đây có thể được xem là một phép thử Bernoulli độc lập hay không.

- Các lá bài được chia trong một ván bài poker.
- Kết quả mỗi lần gieo xúc xắc.

4.12 Có và không hoàn lại. Trong các tình huống sau, giả sử một nửa dân số được đề cập là nam và một nửa là nữ.

- Giả sử bạn đang lấy mẫu từ một căn phòng có 10 người. Xác suất chọn được hai người nữ liên tiếp là bao nhiêu nếu lấy mẫu có hoàn lại? Còn nếu lấy mẫu không hoàn lại thì sao?
- Bây giờ giả sử bạn đang lấy mẫu từ một sân vận động có 10.000 người. Xác suất chọn được hai người nữ liên tiếp là bao nhiêu nếu lấy mẫu có hoàn lại? Còn nếu lấy mẫu không hoàn lại thì sao?
- Chúng ta thường giả định rằng các cá thể được chọn từ một quần thể lớn là độc lập. Dựa trên kết quả từ phần (a) và (b), hãy giải thích xem giả định này có hợp lý không.

4.13 Màu mắt, phần I. Một cặp vợ chồng đều có mắt nâu nhưng mang gen khiến con họ có thể có mắt nâu (xác suất 0.75), mắt xanh (0.125) hoặc mắt lục (0.125).

- Xác suất để đứa trẻ mắt xanh đầu tiên của họ là đứa con thứ ba là bao nhiêu? Giả sử rằng màu mắt của các con là độc lập với nhau.
- Trung bình, cặp vợ chồng này sẽ có bao nhiêu con trước khi sinh ra một đứa trẻ có mắt xanh? Độ lệch chuẩn của số con mà họ dự kiến sẽ có cho đến khi sinh được một đứa mắt xanh là bao nhiêu?

4.14 Tỷ lệ lỗi. Một máy sản xuất loại transistor đặc biệt (một linh kiện trong máy tính) có tỷ lệ lỗi là 2%. Việc sản xuất được xem là một quá trình ngẫu nhiên, trong đó mỗi transistor là độc lập với các transistor khác.

- Xác suất để transistor thứ 10 được sản xuất là transistor bị lỗi đầu tiên là bao nhiêu?
- Xác suất để máy không sản xuất ra transistor bị lỗi nào trong một lô 100 cái là bao nhiêu?
- Trung bình, bạn mong đợi máy sẽ sản xuất bao nhiêu transistor trước khi có cái bị lỗi đầu tiên? Độ lệch chuẩn là bao nhiêu?
- Một máy khác cũng sản xuất transistor có tỷ lệ lỗi là 5%, trong đó mỗi transistor được tạo ra độc lập với các cái khác. Trung bình bạn mong đợi bao nhiêu transistor sẽ được sản xuất trước khi có cái bị lỗi đầu tiên? Độ lệch chuẩn là bao nhiêu?
- Dựa vào câu trả lời của bạn ở phần (c) và (d), việc tăng xác suất của một sự kiện sẽ ảnh hưởng như thế nào đến trung bình và độ lệch chuẩn của thời gian chờ đợi cho đến khi thành công?

4.15 Bernoulli, trung bình. Sử dụng các quy tắc xác suất từ Phần 3.4 để suy ra trung bình của một biến ngẫu nhiên Bernoulli, tức là biến ngẫu nhiên X nhận giá trị 1 với xác suất p và nhận giá trị 0 với xác suất $1 - p$. Tức là, hãy tính giá trị kỳ vọng của một biến ngẫu nhiên Bernoulli tổng quát

4.16 Bernoulli, độ lệch chuẩn. Sử dụng các quy tắc xác suất từ Phần 3.4 để suy ra độ lệch chuẩn của một biến ngẫu nhiên Bernoulli, tức là biến ngẫu nhiên X nhận giá trị 1 với xác suất p và nhận giá trị 0 với xác suất $1 - p$. Tức là, hãy tính căn bậc hai của phương sai của một biến ngẫu nhiên Bernoulli tổng quát.

4.3 PHÂN PHỐI NHỊ THỨC

4.3 Phân phối nhị thức

Phân phối nhị thức được sử dụng để mô tả số lần thành công trong một số lần thử cố định, khác với phân phối hình học, được sử dụng để mô tả số lần thử cần thực hiện trước khi quan sát được một lần thành công.

4.3.1 Phân phối nhị thức

Hãy tưởng tượng một lần nữa rằng chúng ta đang ở một công ty bảo hiểm, nơi 70% khách hàng không vượt quá mức khấu trừ bảo hiểm.

VÍ DỤ 4.28

Giả sử công ty bảo hiểm đang xem xét một mẫu ngẫu nhiên gồm 4 khách hàng đang sử dụng dịch vụ. Xác suất để đúng một trong số họ vượt quá mức khấu trừ và 3 người còn lại thì không là bao nhiêu? Để tiện lợi, ta đặt tên 4 người là Ariana (A), Brittany (B), Carlton (C) và Damian (D).

Ta xét một trường hợp trong đó chỉ một người vượt quá mức khấu trừ:

$$\begin{aligned} P(A = \text{vượt}, B = \text{không}, C = \text{không}, D = \text{không}) \\ &= P(A = \text{vượt})P(B = \text{không})P(C = \text{không})P(D = \text{không}) \\ &= (0,7)(0,7)(0,7)(0,7) \\ &= (0,7)^3(0,3)^1 \\ &= 0,103 \end{aligned}$$

Tuy nhiên còn 3 trường hợp khác: Brittany, Carlton hoặc Damian là người vượt quá mức khấu trừ. Mỗi trường hợp đều có xác suất là: $(0,7)^3(0,3)^1$. Tổng cộng có 4 trường hợp như vậy, nên xác suất tổng là $4 \times (0,7)^3(0,3)^1 = 0,412$

BÀI TẬP THỰC HÀNH 4.29

Chứng minh rằng xác suất để Brittany là người duy nhất vượt mức khấu trừ là $(0,7)^3(0,3)^1$.²⁰

Trường hợp ở ví dụ 4.28 là một ví dụ về phân phối nhị thức. Phân phối nhị thức mô tả xác suất có chính xác k lần thành công trong số n phép thử Bernoulli độc lập với xác suất thành công là p (trong ví dụ 4.28, $n = 4$, $k = 3$, $p = 0,7$). Ta muốn tìm công thức tổng quát cho xác suất trong phân phối nhị thức, tức là dùng n , k , và p để tính xác suất. Để làm điều này, chúng ta xem xét lại từng phần trong Ví dụ 4.28.

Có bốn cá nhân có thể là người vượt quá mức khấu trừ, và mỗi trường hợp trong bốn trường hợp này đều có cùng xác suất. Do đó, chúng ta có thể xác định xác suất cuối cùng như sau:

$$[\text{số trường hợp}] \times P(\text{mỗi trường hợp})$$

Thành phần đầu tiên của phương trình này là số cách để sắp xếp $k = 3$ lần thành công trong tổng số $n = 4$ lần thử. Thành phần thứ hai là xác suất của bất kỳ một trong bốn kịch bản (có xác suất bằng nhau).

²⁰ $P(A = \text{không}, B = \text{vượt}, C = \text{không}, D = \text{không}) = (0,7)(0,3)(0,7)(0,7) = (0,7)^3(0,3)^1$.

4.3 PHÂN PHỐI NHỊ THỨC

Xét xác suất của một kịch bản đơn lẻ trong trường hợp tổng quát gồm k lần thành công và $n - k$ lần thất bại trong số n lần thử. Trong bất kỳ trường hợp nào như vậy, ta áp dụng Quy tắc Nhân (Multiplication Rule) cho các biến cố độc lập:

$$p^k(1 - p)^{n-k}$$

Đây là công thức tổng quát của chúng ta cho xác suất của một trường hợp đơn lẻ - P (trường hợp đơn lẻ).

Tiếp theo, chúng ta giới thiệu một công thức tổng quát cho số cách chọn k lần thành công trong số n lần thử, tức là sắp xếp k lần thành công và $n - k$ lần thất bại:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

Kí hiệu $\binom{n}{k}$ được đọc là “ **n chọn k** ”.²¹ Dấu chấm than ($k!$) biểu thị một **biểu thức giai thừa**.

$$0! = 1$$

$$1! = 1$$

$$2! = 2 \times 1 = 2$$

$$3! = 3 \times 2 \times 1 = 6$$

$$4! = 4 \times 3 \times 2 \times 1 = 24$$

$$\vdots$$

$$n! = n \times (n - 1) \times \dots \times 3 \times 2 \times 1$$

Sử dụng công thức này, chúng ta có thể tính được số cách chọn $k = 3$ lần thành công trong $n = 4$ lần thử:

$$\binom{4}{3} = \frac{4!}{3!(4-3)!} = \frac{4!}{3!1!} = \frac{4 \times 3 \times 2 \times 1}{(3 \times 2 \times 1) \times (1)} = 4$$

Kết quả này chính là điều mà chúng ta đã tìm ra bằng cách xem xét cẩn thận từng trường hợp có thể xảy ra trong Ví dụ 4.28.

Thay thế n chọn k cho số trường hợp và $p^k(1 - p)^{n-k}$ cho xác suất của một trường hợp đơn lẻ, ta thu được công thức tổng quát của phân phối nhị thức.

PHÂN PHỐI NHỊ THỨC

Giả sử xác suất để một phép thử đơn lẻ là thành công là p . Khi đó, xác suất quan sát được chính xác k lần thành công trong số n phép thử độc lập là:

$$\binom{n}{k} p^k (1 - p)^{n-k} = \frac{n!}{k!(n-k)!} p^k (1 - p)^{n-k}$$

Giá trị trung bình, phương sai và độ lệch chuẩn của số lượt thành công quan sát được là:

$$\mu = np$$

$$\sigma^2 = np(1 - p)$$

$$\sigma = \sqrt{np(1 - p)}$$

CÓ PHẢI LÀ PHÂN PHỐI NHỊ THỨC KHÔNG? – BỐN ĐIỀU KIỆN CẦN KIỂM TRA.

- 1) Các phép thử là độc lập
- 2) Số lần thử n là cố định.
- 3) Kết quả của mỗi phép thử có thể phân loại thành *thành công* hoặc *thất bại*.
- 4) Xác suất thành công p là như nhau cho mỗi lần thử.

²¹ Những kí hiệu khác của n chọn k bao gồm ${}_nC_k$, C_n^k , và $C(n, k)$.

4.3 PHÂN PHỐI NHỊ THỨC

VÍ DỤ 4.30

Xác suất để 3 trong số 8 người được chọn ngẫu nhiên vượt quá mức khấu trừ bảo hiểm, tức là 5 trong số 8 người không vượt quá mức khấu trừ, là bao nhiêu? Nhớ rằng: 70% số người không vượt quá mức khấu trừ.

Chúng ta muốn áp dụng mô hình phân phối nhị thức, vì vậy cần kiểm tra các điều kiện. Số lần thử là cố định ($n = 8$) (điều kiện 2), kết quả mỗi phép thử có thể phân loại là thành công hoặc thất bại (điều kiện 3). Do mẫu được chọn ngẫu nhiên nên các phép thử là độc lập (điều kiện 1), xác suất thành công là như nhau trong mỗi lần thử (điều kiện 4).

Trong kết quả quan tâm, có $k = 5$ lần thành công trong $n = 8$ lần thử (Nhớ rằng: “thành công” ở đây được định nghĩa là người không vượt quá mức khấu trừ), và xác suất thành công là $p = 0.7$. Vậy xác suất để 5 người trong số 8 người không vượt mức khấu trừ (và 3 người vượt mức) được tính theo công thức:

$$\begin{aligned} \binom{8}{5} (0,7)^5 (1 - 0,7)^{8-5} &= \frac{8!}{5! (8-5)!} (0,7)^5 (1 - 0,7)^{8-5} \\ &= \frac{8!}{5! 3!} (0,7)^5 (1 - 0,7)^{8-5} \end{aligned}$$

Xử lý phần giai thừa:

$$\frac{8!}{5! 3!} = \frac{8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1}{(5 \times 4 \times 3 \times 2 \times 1)(3 \times 2 \times 1)} = \frac{8 \times 7 \times 6}{3 \times 2 \times 1} = 56$$

Biết rằng $(0,7)^5 (0,3)^3 \approx 0,00454$, xác suất cuối cùng sẽ là $56 \times 0,00454 \approx 0,254$

TÍNH TOÁN CÁC XÁC SUẤT NHỊ THỨC

Bước đầu tiên khi sử dụng mô hình nhị thức là kiểm tra xem mô hình có phù hợp không. Bước thứ hai là xác định các giá trị n , p , và k . Ở bước cuối cùng, sử dụng phần mềm hoặc các công thức để tính xác suất, sau đó diễn giải kết quả.

Nếu bạn phải thực hiện tính toán bằng tay, thì việc rút gọn càng nhiều số hạng càng tốt ở tử số và mẫu số của hệ số nhị thức thường rất hữu ích.

BÀI TẬP THỰC HÀNH 4.31

Nếu chúng ta chọn ngẫu nhiên 40 hồ sơ từ công ty bảo hiểm đã đề cập trước đó, thì ta kỳ vọng có bao nhiêu hồ sơ trong số đó không vượt quá mức khấu trừ trong một năm nhất định? Độ lệch chuẩn của số hồ sơ không vượt quá mức khấu trừ là bao nhiêu? ²²

BÀI TẬP THỰC HÀNH 4.32

Xác suất để một người hút thuốc ngẫu nhiên mắc một bệnh phổi nghiêm trọng trong suốt cuộc đời của họ là khoảng 0.3. Nếu bạn có 4 người bạn hút thuốc, thì các điều kiện để áp dụng mô hình phân phối nhị thức có được thỏa mãn không? ²³

²² Chúng ta được yêu cầu xác định con số kỳ vọng (giá trị trung bình) và độ lệch chuẩn, cả hai đều có thể được tính trực tiếp từ các công thức: $\mu = np = 40 \times 0,7 = 28$ và $\sigma = \sqrt{np(1-p)} = \sqrt{40 \times 0,7 \times 0,3} = 2,9$. Vì khoảng xấp xỉ 95% các quan sát nằm trong phạm vi ± 2 độ lệch chuẩn quanh giá trị trung bình (xem Mục 2.1.4), nên chúng ta có khả năng sẽ quan sát thấy ít nhất 22 nhưng ít hơn 34 người trong mẫu không vượt quá mức khấu trừ.

²³ Một câu trả lời khả thi: Nếu những người bạn đó quen biết nhau, thì giả định về tính độc lập có lẽ không được thỏa mãn. Ví dụ, những người quen có thể có thói quen hút thuốc tương tự nhau, hoặc họ có thể cùng nhau cam kết sẽ cai thuốc, điều này tạo ra sự phụ thuộc giữa các kết quả với nhau.

4.3 PHÂN PHỐI NHỊ THỨC

BÀI TẬP THỰC HÀNH 4.33

Giả sử bốn người bạn này không quen biết nhau và chúng ta có thể xem họ như một mẫu ngẫu nhiên từ quần thể. Khi đó, mô hình phân phối nhị thức có phù hợp không? Xác suất của các sự kiện sau là bao nhiêu: ²⁴

- a) Không ai trong số họ mắc bệnh phổi nghiêm trọng?
- b) Chỉ có một người mắc bệnh phổi nghiêm trọng?
- c) Tối đa một người mắc bệnh phổi nghiêm trọng?

BÀI TẬP THỰC HÀNH 4.34

Xác suất để có ít nhất 2 trong số 4 người bạn hút thuốc của bạn sẽ mắc bệnh phổi nghiêm trọng trong suốt cuộc đời của họ là bao nhiêu? ²⁵

BÀI TẬP THỰC HÀNH 4.35

Giả sử bạn có 7 người bạn hút thuốc và có thể xem họ như một mẫu ngẫu nhiên từ quần thể người hút thuốc. ²⁶

- a) Bạn kỳ vọng có bao nhiêu người trong số họ sẽ mắc bệnh phổi nghiêm trọng? (Tức là: giá trị trung bình là bao nhiêu?)
- b) Xác suất để tối đa 2 trong số 7 người bạn này sẽ mắc bệnh phổi nghiêm trọng là bao nhiêu?

Tiếp theo, chúng ta xét đến thành phần đầu tiên trong công thức xác suất nhị thức, đó là n chọn k trong một số trường hợp đặc biệt.

BÀI TẬP THỰC HÀNH 4.36

Tại sao $\binom{n}{0} = 1$ và $\binom{n}{n} = 1$ là đúng với mọi số nguyên n ? ²⁷

BÀI TẬP THỰC HÀNH 4.37

Có bao nhiêu cách để sắp xếp một lần thành công và $n - 1$ lần thất bại trong n lần thử? Có bao nhiêu cách để sắp xếp $n - 1$ lần thành công và một lần thất bại trong n lần thử? ²⁸

²⁴ Để kiểm tra xem mô hình nhị thức có phù hợp hay không, chúng ta phải xác minh các điều kiện sau. (i) Vì chúng ta giả định rằng có thể xem những người bạn này như một mẫu ngẫu nhiên, nên các phép thử là độc lập. (ii) Chúng ta có số lần thử cố định ($n = 4$). (iii) Mỗi kết quả là một lần thành công hoặc thất bại. (iv) Xác suất thành công là giống nhau trong mỗi phép thử vì các cá nhân này được coi như một mẫu ngẫu nhiên ($p = 0.3$ nếu chúng ta gọi “thành công” là một người mắc bệnh phổi, thì đó là một lựa chọn khá rùng rợn). Tính các phần (a) và (b) bằng cách sử dụng công thức phân phối nhị thức: $P(0) = \binom{4}{0} (0,3)^0 (0,7)^4 = 1 \times 1 \times 0,7^4 = 0,2401$, $P(1) = \binom{4}{1} (0,3)^1 (0,7)^3 = 0,4116$. Lưu ý: $0! = 1$. Phần (c) có thể được tính bằng cách cộng các phần (a) và (b): $P(0) + P(1) = 0,2401 + 0,4116 = 0,6517$. Điều đó có nghĩa là có khoảng 65% khả năng rằng không quá một trong bốn người bạn hút thuốc của bạn sẽ mắc bệnh phổi nghiêm trọng.

²⁵ Xác suất để “không quá một người sẽ mắc bệnh phổi nghiêm trọng” đã được tính trong Thực hành có hướng dẫn 4.33 là 0,6517, vì vậy chúng ta tính 1 trừ giá trị này: 0,3483.

²⁶ (a) $\mu = 0,3 \times 7 = 2,1$. (b) $P(0, 1, \text{ hoặc } 2 \text{ người sẽ mắc bệnh phổi nghiêm trọng}) = P(k = 0) + P(k = 1) + P(k = 2) = 0,6471$.

²⁷ Diễn giải các biểu thức này thành lời. Có bao nhiêu cách khác nhau để sắp xếp 0 lần thành công và n lần thất bại trong n lần thử? (1 cách) Có bao nhiêu cách khác nhau để sắp xếp n lần thành công và 0 lần thất bại trong n lần thử? (1 cách)

²⁸ Một lần thành công và $n-1$ lần thất bại: có chính xác n vị trí khác nhau để đặt một lần thành công, vì vậy có n cách để sắp xếp một lần thành công và $n-1$ lần thất bại. Lập luận tương tự cũng được áp dụng cho câu hỏi thứ hai. Về mặt toán học, chúng ta chứng minh những kết quả này bằng cách xác minh hai phương trình sau:

$$\binom{n}{1} = n \quad , \quad \binom{n}{n-1} = n$$

4.3 PHÂN PHỐI NHỊ THỨC

4.3.2 Xấp xỉ chuẩn cho phân phối nhị thức

Công thức phân phối nhị thức trở nên rườm rà khi cỡ mẫu (n) lớn, đặc biệt là khi chúng ta xét đến một khoảng giá trị quan sát. Trong một số trường hợp, chúng ta có thể sử dụng phân phối chuẩn như một cách ước lượng xác suất nhị thức đơn giản và nhanh hơn.

VÍ DỤ 4.38

Khoảng 15% dân số Hoa Kỳ hút thuốc lá. Một chính quyền địa phương tin rằng cộng đồng của họ có tỷ lệ người hút thuốc thấp hơn, nên đã ủy quyền thực hiện một cuộc khảo sát với 400 cá nhân được chọn ngẫu nhiên. Kết quả khảo sát cho thấy chỉ có 42 trong số 400 người tham gia hút thuốc. Nếu tỷ lệ thực sự của người hút thuốc trong cộng đồng này là 15%, thì xác suất quan sát được 42 người hút thuốc hoặc ít hơn trong một mẫu gồm 400 người là bao nhiêu?

Chúng ta để việc kiểm tra thông thường bốn điều kiện áp dụng mô hình nhị thức lại như một bài tập. Câu hỏi được đặt ra tương đương với việc hỏi: xác suất quan sát được $k = 0, 1, 2, \dots$, hoặc 42 người hút thuốc trong một mẫu có $n = 400$ khi $p = 0.15$? Chúng ta có thể tính 43 xác suất khác nhau này và cộng tất cả lại để tìm ra đáp án:

$$\begin{aligned} &P(k = 0 \text{ hoặc } k = 1 \text{ hoặc } \dots \text{ hoặc } k = 42) \\ &= P(k = 0) + P(k = 1) + \dots + P(k = 42) \\ &= 0,0054 \end{aligned}$$

Nếu tỷ lệ thực sự của người hút thuốc trong cộng đồng là $p = 0,15$, thì xác suất quan sát được 42 người hút thuốc hoặc ít hơn trong một mẫu gồm $n = 400$ là 0,0054.

Các phép tính trong Ví dụ 4.38 rất dài và tốn công. Nói chung, chúng ta nên tránh thực hiện những tính toán như vậy nếu tồn tại một phương pháp thay thế nhanh hơn, dễ hơn và vẫn chính xác. Hãy nhớ rằng việc tính xác suất cho một khoảng giá trị là dễ dàng hơn nhiều khi sử dụng mô hình chuẩn. Chúng ta có thể tự hỏi: liệu có hợp lý không khi sử dụng mô hình chuẩn thay thế cho phân phối nhị thức? Điều đáng ngạc nhiên là: có — nếu một số điều kiện nhất định được thỏa mãn.

BÀI TẬP THỰC HÀNH 4.29

Tại đây, chúng ta xét mô hình phân phối nhị thức khi xác suất thành công là $p = 0.10$. Hình 4.9 biểu diễn bốn biểu đồ cột rỗng (hollow histograms) cho các mẫu mô phỏng từ phân phối nhị thức, sử dụng bốn cỡ mẫu khác nhau: $n = 10, 30, 100, 300$. Điều gì xảy ra với hình dạng của các phân phối khi cỡ mẫu tăng lên? Biểu đồ rỗng cuối cùng giống với phân phối nào? ²⁹

XẤP XỈ CHUẨN CỦA PHÂN PHỐI NHỊ THỨC

Phân phối nhị thức với xác suất thành công p gần giống với phân phối chuẩn khi cỡ mẫu n đủ lớn sao cho cả np và $n(1 - p)$ đều lớn hơn hoặc bằng 10. Phân phối chuẩn xấp xỉ có các tham số tương ứng với giá trị trung bình và độ lệch chuẩn của phân phối nhị thức:

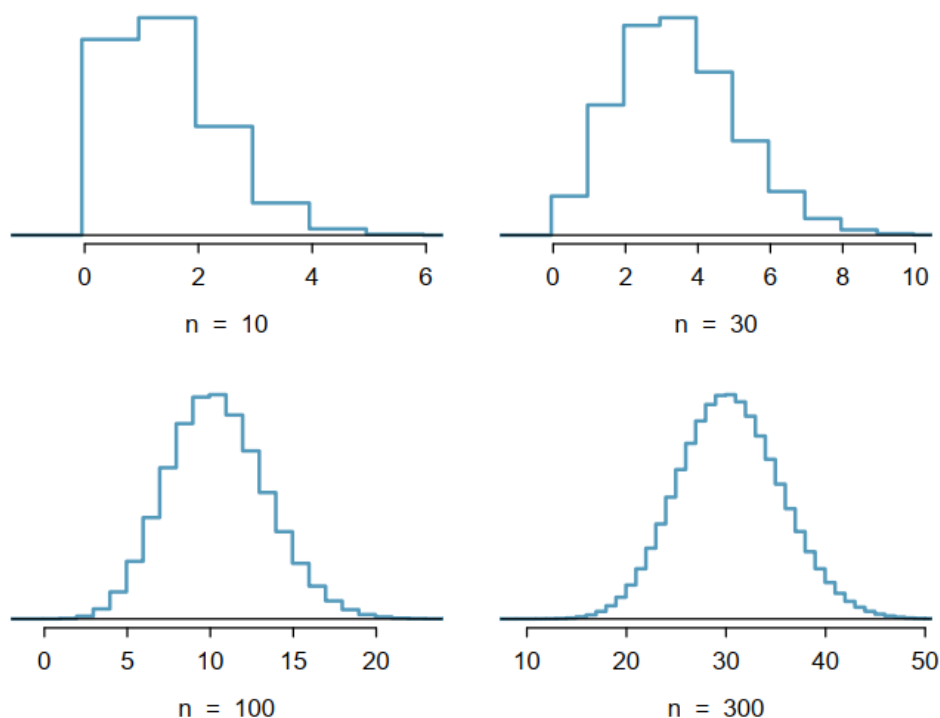
$$\mu = np$$

$$\sigma = \sqrt{np(1 - p)}$$

Xấp xỉ chuẩn có thể được sử dụng khi tính xác suất cho một khoảng gồm nhiều khả năng thành công. Chẳng hạn, chúng ta có thể áp dụng phân phối chuẩn cho tình huống được nêu trong Ví dụ 4.38.

²⁹ Phân phối được chuyển đổi từ một phân phối thô và lệch thành một phân phối có hình dạng khá giống với phân phối chuẩn trong biểu đồ rỗng cuối cùng.

4.3 PHÂN PHỐI NHỊ THỨC



Hình 4.9: Biểu đồ rỗng của các mẫu lấy từ mô hình phân phối nhị thức với $p = 0.10$. Cỡ mẫu tương ứng của bốn biểu đồ là $n = 10, 30, 100$, và 300 .

VÍ DỤ 4.40

Làm thế nào chúng ta có thể sử dụng xấp xỉ chuẩn để ước lượng xác suất quan sát được 42 người hút thuốc hoặc ít hơn trong một mẫu gồm 400 người, nếu tỷ lệ thực sự của người hút thuốc là $p = 0.15$?

Việc chứng minh rằng mô hình nhị thức là hợp lý đã được đưa ra như một bài tập trong Ví dụ 4.38. Chúng ta cũng kiểm tra rằng cả np và $n(1 - p)$ đều lớn hơn hoặc bằng 10:

$$np = 400 \times 0.15 = 60$$

$$n(1 - p) = 400 \times 0.85 = 340$$

Khi các điều kiện này đã được kiểm tra và thỏa mãn, chúng ta có thể sử dụng xấp xỉ chuẩn thay thế cho phân phối nhị thức, với việc sử dụng giá trị trung bình và độ lệch chuẩn lấy từ mô hình phân phối nhị thức:

$$\mu = np = 60$$

$$\sigma = \sqrt{np(1 - p)} = 7.14$$

Chúng ta muốn tìm xác suất quan sát được 42 người hút thuốc hoặc ít hơn bằng cách sử dụng mô hình này.

BÀI TẬP THỰC HÀNH 4.29

Sử dụng mô hình chuẩn $N(\mu = 60; \sigma = 7.14)$ để ước lượng xác suất quan sát được 42 người hút thuốc hoặc ít hơn. Kết quả của bạn nên xấp xỉ với lời giải trong Ví dụ 4.38: 0,0054.³⁰

³⁰ Tính Z-score trước: $Z = \frac{42 - 60}{7.14} = -2.52$. Diện tích phía bên trái tương ứng là 0,0059.

4.3 PHÂN PHỐI NHỊ THỨC

4.3.3 Xấp xỉ chuẩn trở nên không chính xác khi áp dụng cho các khoảng nhỏ

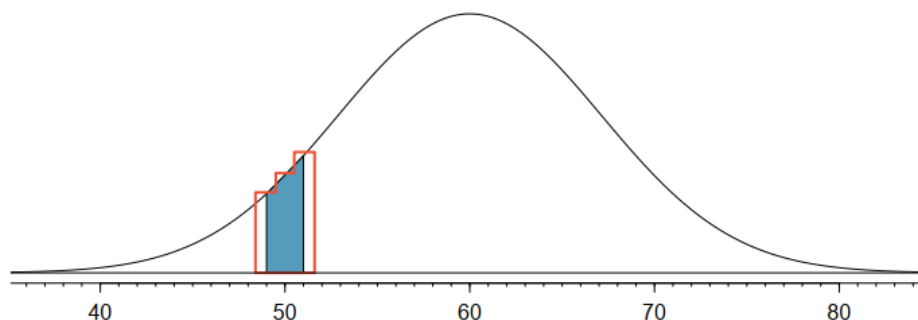
Xấp xỉ chuẩn cho phân phối nhị thức có xu hướng hoạt động kém hiệu quả khi ước lượng xác suất của một khoảng giá trị nhỏ, ngay cả khi các điều kiện áp dụng được thỏa mãn.

Giả sử chúng ta muốn tính xác suất quan sát được 49, 50 hoặc 51 người hút thuốc trong một mẫu gồm 400 người, khi $p = 0.15$. Với một cỡ mẫu lớn như vậy, chúng ta có thể bị cám dỗ áp dụng xấp xỉ chuẩn và sử dụng khoảng từ 49 đến 51. Tuy nhiên, chúng ta sẽ thấy rằng kết quả từ phân phối nhị thức và xấp xỉ chuẩn khác biệt một cách rõ rệt:

Phân phối nhị thức: 0,0649

Xấp xỉ chuẩn: 0,0421

Chúng ta có thể xác định nguyên nhân của sự sai lệch này bằng cách sử dụng Hình 4.10, trong đó hiển thị các diện tích biểu diễn xác suất theo phân phối nhị thức (viền nét) và xấp xỉ chuẩn (tô bóng). Lưu ý rằng bề rộng của vùng dưới đường cong phân phối chuẩn bị thiếu 0.5 đơn vị ở mỗi phía của khoảng.



Hình 4.10: Một đường cong chuẩn với vùng diện tích giữa 49 và 51 được tô bóng. Vùng được viền nét biểu thị xác suất chính xác theo phân phối nhị thức.

CẢI THIỆN XẤP XỈ CHUẨN CHO PHÂN PHỐI NHỊ THỨC

Xấp xỉ chuẩn cho phân phối nhị thức trong các khoảng giá trị thường được cải thiện nếu các giá trị ngưỡng (cutoff) được điều chỉnh nhẹ. Giá trị ngưỡng phía dưới của vùng tô bóng nên được giảm 0,5, và giá trị ngưỡng phía trên nên được tăng 0,5.

Mẹo thêm diện tích khi áp dụng xấp xỉ chuẩn thường hữu ích nhất khi xét một khoảng giá trị quan sát. Trong ví dụ ở trên, ước lượng đã điều chỉnh theo phân phối chuẩn là 0,0633 - gần hơn nhiều so với giá trị chính xác là 0,0649. Mặc dù cũng có thể áp dụng điều chỉnh này khi tính diện tích ở đuôi (tail area), nhưng lợi ích của việc điều chỉnh thường không còn rõ rệt, vì khoảng tổng thể trong những trường hợp đó thường khá rộng.

4.3 PHÂN PHỐI NHỊ THỨC

Bài tập

4.17 Uống rượu khi chưa đủ tuổi, Phần I. Dữ liệu thu thập bởi Cơ quan Quản lý Dịch vụ Sức khỏe Tâm thần và lạm dụng Chất (SAMSHA) cho thấy rằng 69,7% thanh niên từ 18 đến 20 tuổi đã tiêu thụ đồ uống có cồn trong bất kỳ năm nào.³¹

- Giả sử chọn ngẫu nhiên một mẫu gồm 10 người trong độ tuổi 18–20. Việc sử dụng phân phối nhị thức có phù hợp để tính xác suất chính xác 6 người đã tiêu thụ đồ uống có cồn không? Giải thích.
- Tính xác suất để chính xác 6 trong số 10 người được chọn ngẫu nhiên đã tiêu thụ đồ uống có cồn.
- Tính xác suất để chính xác 4 trong số 10 người 18–20 tuổi *không* tiêu thụ đồ uống có cồn.
- Tính xác suất để nhiều nhất 2 trong số 5 người 18–20 tuổi được chọn ngẫu nhiên đã tiêu thụ đồ uống có cồn.
- Tính xác suất để ít nhất 1 trong số 5 người 18–20 tuổi được chọn ngẫu nhiên đã tiêu thụ đồ uống có cồn.

4.18 Bệnh thủy đậu, Phần I. Bệnh viện Nhi Boston ước tính rằng 90% người Mỹ đã từng mắc bệnh thủy đậu trước khi trưởng thành.³²

- Giả sử chúng ta lấy một mẫu ngẫu nhiên gồm 100 người trưởng thành ở Mỹ. Việc sử dụng phân phối nhị thức có phù hợp để tính xác suất chính xác 97 trong số 100 người đó đã từng bị thủy đậu lúc nhỏ không? Giải thích.
- Tính xác suất để chính xác 97 trong số 100 người trưởng thành ngẫu nhiên đã từng bị thủy đậu khi còn nhỏ.
- Tính xác suất để chính xác 3 trong số 100 người trưởng thành *chưa từng* bị thủy đậu lúc nhỏ.
- Tính xác suất để ít nhất 1 trong số 10 người trưởng thành ngẫu nhiên đã từng bị thủy đậu.
- Tính xác suất để nhiều nhất 3 trong số 10 người trưởng thành ngẫu nhiên *chưa từng* bị thủy đậu.

4.19 Uống rượu khi chưa đủ tuổi, Phần II. Như đã tìm hiểu trong Bài tập 4.17, khoảng 70% người từ 18–20 tuổi tiêu thụ đồ uống có cồn mỗi năm. Bây giờ ta xét một mẫu ngẫu nhiên gồm 50 người trong độ tuổi này.

- Bạn kỳ vọng bao nhiêu người đã tiêu thụ đồ uống có cồn? Và độ lệch chuẩn là bao nhiêu?
- Bạn có ngạc nhiên nếu có 45 người trở lên đã tiêu thụ đồ uống có cồn không?
- Tính xác suất rằng 45 người trở lên trong mẫu này đã tiêu thụ đồ uống có cồn. Xác suất này liên quan thế nào đến câu trả lời phần (b)?

4.20 Bệnh thủy đậu, Phần II. Như đã nêu trong Bài tập 4.18 khoảng 90% người trưởng thành Mỹ đã từng bị thủy đậu trước tuổi trưởng thành. Giờ xét một mẫu ngẫu nhiên gồm 120 người trưởng thành Mỹ.

- Bạn kỳ vọng bao nhiêu người trong mẫu này đã từng bị thủy đậu khi nhỏ? Và độ lệch chuẩn là bao nhiêu?
- Bạn có ngạc nhiên nếu có tới 105 người đã từng bị thủy đậu khi còn nhỏ không?
- Tính xác suất rằng 105 người hoặc ít hơn trong mẫu này đã từng bị thủy đậu khi nhỏ. Xác suất này liên quan thế nào đến câu trả lời phần (b)?

4.21 Trò chơi con quay Dreidel. A Một con quay Dreidel có bốn mặt với các chữ cái Hebrew: *nun*, *gimel*, *hei* và *shin* - mỗi mặt có xác suất như nhau khi quay. Giả sử bạn quay con quay ba lần. Tính xác suất để:

- Có ít nhất một lần ra mặt *nun*?
- Có ít nhất một lần ra mặt *nun*?
- Chính xác 1 lần ra mặt *hei*?
- Nhiều nhất 2 lần ra mặt *gimel*?



Photo by Staccabees, cropped
(<http://flic.kr/p/7gLZTf>)
CC BY 2.0 license

³¹ SAMHSA, Office of Applied Studies, National Survey on Drug Use and Health, 2007 and 2008.

³² Boston Children's Hospital, Chickenpox summary page, referenced April 29, 2021.

4.3 PHÂN PHỐI NHỊ THỨC

4.22 Chứng sợ nhện. Một khảo sát Gallup cho thấy rằng 7% thanh thiếu niên (13-17 tuổi) mắc chứng sợ nhện nghiêm trọng. Ở một trại hè, có 10 thanh thiếu niên ngủ trong mỗi lều. Giả sử rằng 10 người này là độc lập với nhau.³³

- Tính xác suất để ít nhất 1 người mắc chứng sợ nhện.
- Tính xác suất để chính xác 2 người mắc chứng sợ nhện.
- Tính xác suất để nhiều nhất 1 người mắc chứng sợ nhện.
- Nếu người phụ trách muốn đảm bảo không quá 1 người trong mỗi lều sợ nhện, việc phân ngẫu nhiên các bạn vào lều có hợp lý không?

4.23 Màu mắt, Phần II. Bài tập 4.13 nêu rằng một cặp vợ chồng có mắt nâu có xác suất 0.75 sinh con mắt nâu, 0.125 mắt xanh và 0.125 mắt xanh lục.

- Tính xác suất để đứa con đầu có mắt xanh lục và đứa thứ hai thì không?
- Tính xác suất để chính xác một trong hai đứa trẻ có mắt xanh lục?
- Nếu họ có 6 đứa con, xác suất để chính xác hai trong số đó có mắt xanh lục là bao nhiêu?
- Nếu họ có 6 đứa con, xác suất để ít nhất một đứa có mắt xanh lục là bao nhiêu?
- Xác suất để đứa con mắt xanh lục đầu tiên là đứa con thứ tư là bao nhiêu?
- Liệu việc chỉ có 2 trong 6 đứa trẻ có mắt nâu có được xem là điều bất thường không?

4.24 Bệnh thiếu máu hồng cầu hình liềm. Đây là bệnh di truyền làm cho hồng cầu mất tính linh hoạt và có hình dạng bất thường, gây ra nhiều biến chứng. Nếu cả hai bố mẹ đều mang gen bệnh, thì con cái có: 25% khả năng mắc bệnh, 50% là người mang gen, và 25% không mang gen cũng không mắc bệnh. Nếu một cặp bố mẹ mang gen có 3 người con, tính xác suất để:

- Hai người con mắc bệnh?
- Không người con nào mắc bệnh?
- Ít nhất một người con không mắc bệnh và không mang gen bệnh?
- Đứa con đầu tiên mắc bệnh là đứa con thứ ba?

4.25 Khám phá hoán vị. Công thức số cách sắp xếp n đối tượng là $n! = n \times (n - 1) \times \dots \times 2 \times 1$. Bài tập này sẽ hướng dẫn bạn khai triển công thức cho một vài trường hợp đặc biệt. Một công ty nhỏ có 5 nhân viên: Anna, Ben, Carl, Damian và Eddy. Có 5 chỗ đậu xe xếp thành một hàng tại công ty, không ai được chỉ định chỗ cố định, và mỗi ngày các nhân viên sẽ đậu xe vào một vị trí ngẫu nhiên. Nói cách khác, mọi thứ tự sắp xếp xe trong hàng đều có xác suất như nhau.

- Vào một ngày cho trước, xác suất các nhân viên đậu xe theo đúng thứ tự bảng chữ cái là bao nhiêu?
- Nếu thứ tự theo bảng chữ cái có xác suất bằng nhau như các thứ tự khác, thì có bao nhiêu cách sắp xếp 5 chiếc xe?
- Giờ xét một mẫu gồm 8 nhân viên. Có bao nhiêu cách sắp xếp xe của 8 người?

4.26 Con trai. Dù người ta thường cho rằng xác suất sinh con trai và con gái bằng nhau, nhưng thực tế xác suất sinh con trai là 0.51. Giả sử một cặp đôi dự định có 3 con.

- Sử dụng mô hình nhị thức để tính xác suất hai trong số ba người con là con trai.
- Liệt kê tất cả cách sắp xếp có thể của 3 người con trong đó có 2 người là con trai. Sử dụng các trường hợp đó để tính xác suất giống như ở phần (a) nhưng áp dụng quy tắc cộng cho các biến cố rời nhau. Xác nhận rằng kết quả phần (a) và (b) giống nhau.
- Nếu muốn tính xác suất rằng một cặp đôi dự định có 8 người con và có đúng 3 người con trai, hãy mô tả ngắn gọn vì sao phương pháp ở phần (b) sẽ phức tạp hơn so với phần (a).

³³Gallup Poll, *What Frightens America's Youth?*, March 29, 2005.

4.4 PHÂN PHỐI NHỊ THỨC ÂM

4.4 Phân phối nhị thức âm

Phân phối hình học mô tả xác suất quan sát được lần thành công đầu tiên ở lần thử thứ n . Phân phối nhị thức âm thì tổng quát hơn: nó mô tả xác suất quan sát được lần thành công thứ k ở lần thử thứ n .

VÍ DỤ 4.42

Mỗi ngày, huấn luyện viên bóng bầu dục trung học nói với Brian – cầu thủ sút chính – rằng cậu ấy có thể về nhà sau khi sút thành công 4 quả phạt ở khoảng cách 35 yard. Giả sử mỗi lần sút có xác suất thành công là p . Nếu p nhỏ – ví dụ gần 0.1 – liệu chúng ta có mong đợi rằng Brian sẽ cần nhiều lần sút trước khi thành công 4 quả không?

E

Chúng ta đang chờ đợi **lần thành công thứ 4** ($k = 4$). Nếu xác suất thành công (p) nhỏ, thì số lần thử (n) có khả năng sẽ lớn. Điều này có nghĩa là Brian **nhiều khả năng sẽ cần nhiều lần sút** trước khi đạt được 4 lần thành công ($k = 4$). Nói cách khác, xác suất để n nhỏ là thấp.

Để xác định một trường hợp nhị thức âm, chúng ta kiểm tra 4 điều kiện. Ba điều kiện đầu cũng giống như trong phân phối nhị thức thông thường.

CÓ PHẢI LÀ NHỊ THỨC ÂM KHÔNG? BỐN ĐIỀU KIỆN CẦN KIỂM TRA:

- 1) Các lần thử là độc lập.
- 2) Mỗi lần thử chỉ có thể phân loại là thành công hoặc thất bại.
- 3) Xác suất thành công (p) là giống nhau ở mỗi lần thử.
- 4) Lần thử cuối cùng phải là một lần thành công.

BÀI TẬP THỰC HÀNH 4.43

G

Giả sử Brian rất chăm chỉ luyện tập và mỗi lần sút ở khoảng cách 35 yard, cậu có xác suất thành công là $p = 0.8$. Bạn hãy đoán xem Brian sẽ cần khoảng bao nhiêu lần sút trước khi thực hiện được cú đá thành công thứ 4.³⁴

VÍ DỤ 4.44

Ở buổi tập hôm qua, Brian chỉ mất 6 lần thử để hoàn thành cú đá thứ tư. Hãy viết ra tất cả các chuỗi cú đá có thể xảy ra.

E

Vì Brian cần 6 lần để đạt được lần thành công thứ tư, chúng ta biết rằng **cú đá cuối cùng chắc chắn là một lần thành công**. Điều này có nghĩa là **3 lần thành công và 2 lần thất bại** phải nằm trong **5 lần thử đầu tiên**. Có tất cả **10 chuỗi khả dĩ** gồm 3 lần thành công và 2 lần thất bại trong 5 lần thử đầu, được liệt kê trong Hình 4.11. Nếu Brian đạt được lần thành công thứ tư ($k = 4$) ở lần thử thứ sáu ($n = 6$), thì chuỗi các lần thành công và thất bại của cậu ấy phải là một trong 10 chuỗi khả dĩ này.

BÀI TẬP THỰC HÀNH 4.45

G

Mỗi chuỗi trong Hình 4.11 có chính xác 2 lần thất bại và 4 lần thành công, với lần cuối cùng luôn là một lần thành công. Nếu xác suất thành công là $p = 0.8$, hãy tính xác suất của chuỗi đầu tiên trong hình.³⁵

³⁴ Một câu trả lời khả dĩ: vì Brian có khả năng sút thành công mỗi quả phạt, nên cậu ấy sẽ cần ít nhất 4 lần thử, nhưng có thể không quá 6 hoặc 7 lần.

³⁵ Chuỗi đầu tiên: $0,2 \times 0,2 \times 0,8 \times 0,8 \times 0,8 \times 0,8 = 0,0164$.

4.4 PHÂN PHỐI NHỊ THỨC ÂM

Lượt đá bóng						
	1	2	3	4	5	6
1	F	F	1 S	2 S	3 S	4 S
2	F	1 S	F	2 S	3 S	4 S
3	F	1 S	2 S	F	3 S	4 S
4	F	1 S	2 S	3 S	F	4 S
5	1 S	F	F	2 S	3 S	4 S
6	1 S	F	2 S	F	3 S	4 S
7	1 S	F	2 S	3 S	F	4 S
8	1 S	2 S	F	F	3 S	4 S
9	1 S	2 S	F	3 S	F	4 S
10	1 S	2 S	3 S	F	F	4 S

Hình 4.11: Mười chuỗi khả dĩ khi cú sút thành công thứ tư xảy ra ở lần sút thứ sáu.

Nếu xác suất để Brian thực hiện thành công một cú sút phạt 35 yard là $p = 0.8$, thì xác suất để Brian mất chính xác sáu lần thử để đạt cú sút thành công thứ tư là bao nhiêu? Ta có thể viết như sau:

$$\begin{aligned}
 &P(\text{Brian mất 6 lần thử để thực hiện 4 cú sút thành công}) \\
 &= P(\text{Brian thực hiện 3 cú thành công trong 5 lần đầu, và thực hiện cú thứ 6 thành công}) \\
 &= P(\text{Chuỗi 1 HOẶC chuỗi 2 HOẶC ... HOẶC chuỗi 10})
 \end{aligned}$$

Trong đó các chuỗi là những chuỗi trong Hình 4.11. Ta có thể phân tách xác suất cuối cùng này thành tổng của mười khả năng rời nhau:

$$\begin{aligned}
 &P(\text{Chuỗi 1 HOẶC chuỗi 2 HOẶC ... HOẶC chuỗi 10}) \\
 &= P(\text{Chuỗi 1}) + P(\text{Chuỗi 2}) + \dots + P(\text{Chuỗi 10})
 \end{aligned}$$

Xác suất của chuỗi đầu tiên đã được xác định trong HƯỚNG DẪN THỰC HÀNH 4.45 là 0,0164, và mỗi chuỗi khác cũng có cùng xác suất. Vì mỗi chuỗi trong mười chuỗi đều có cùng xác suất, tổng xác suất là mười lần xác suất của một chuỗi bất kỳ.

Cách tính xác suất phân phối nhị thức âm này tương tự với cách giải các bài toán phân phối nhị thức trong Mục 4.3. Xác suất được chia thành hai phần:

$$\begin{aligned}
 &P(\text{Brian mất 6 lần thử để thực hiện 4 cú thành công}) \\
 &= [\text{Số chuỗi khả dĩ}] \times P(\text{Một chuỗi cụ thể})
 \end{aligned}$$

Ta xét riêng từng phần, rồi nhân lại để ra kết quả cuối cùng.

Trước tiên, ta xác định xác suất của một chuỗi cụ thể. Một trường hợp đơn giản là xảy ra tất cả các lần thất bại trước $(n-k)$ lần, sau đó là k lần thành công:

$$\begin{aligned}
 &P(\text{Một chuỗi cụ thể}) \\
 &= P(n - k \text{ lần thất bại và } k \text{ lần thành công}) \\
 &= (1 - p)^{n-k} p^k
 \end{aligned}$$

4.4 PHÂN PHỐI NHỊ THỨC ÂM

Chúng ta cũng cần xác định số chuỗi khả dĩ trong trường hợp tổng quát. Ở ví dụ trước, đã xác định có mười chuỗi mà trong đó cú sút thành công thứ tư xảy ra ở lần sút thứ sáu. Những chuỗi này được xác định bằng cách cố định lần thử cuối cùng là một lần thành công, rồi xem có bao nhiêu cách để sắp xếp các kết quả còn lại. Nói cách khác, có bao nhiêu cách để sắp xếp $k - 1$ lần thành công trong $n - 1$ lần thử? Điều này có thể được tìm ra bằng hệ số chọn chập k (n chọn k), nhưng áp dụng cho $n - 1$ và $k - 1$:

$$\binom{n-1}{k-1} = \frac{(n-1)!}{(k-1)!((n-1)-(k-1))!} = \frac{(n-1)!}{(k-1)!(n-k)!}$$

Đây là số cách khác nhau để sắp xếp $k - 1$ lần thành công và $n - k$ lần thất bại trong $n - 1$ lần thử. Nếu bạn chưa quen với ký hiệu giai thừa (dấu chấm than “!”), hãy xem lại phần trình bày về giai thừa ở trang 150.

PHÂN PHỐI NHỊ THỨC ÂM

Phân phối nhị thức âm mô tả xác suất để đạt được thành công thứ k tại lần thử thứ n , trong đó tất cả các lần thử là độc lập:

$$P(\text{thành công thứ } k \text{ ở lần thử thứ } n) = \binom{n-1}{k-1} p^k (1-p)^{n-k}$$

Trong đó p : là xác suất thành công trong một lần thử.

VÍ DỤ 4.46

Chứng minh rằng, sử dụng công thức của phân phối nhị thức âm, xác suất Brian thực hiện cú đá thành công thứ tư ở lần thử thứ sáu là 0,164.

(E)

Xác suất thành công trong một lần đá là $p = 0,8$, số lần thành công mong muốn là $k = 4$, và số lần thử cần thiết trong trường hợp này là $n = 6$.

$$\binom{n-1}{k-1} p^k (1-p)^{n-k} = \frac{5!}{3!2!} (0,8)^4 (0,2)^2 = 10 \times 0,0164 = 0,164$$

BÀI TẬP THỰC HÀNH 4.47

(G)

Phân phối nhị thức âm yêu cầu mỗi lần đá của Brian là độc lập. Bạn có nghĩ rằng điều này hợp lý không? ³⁶

BÀI TẬP THỰC HÀNH 4.48

(G)

Giả sử các cú đá của Brian là độc lập. Xác suất để Brian thực hiện thành công cú đá thứ tư trong vòng 5 lần đá là bao nhiêu? ³⁷

³⁶ Chưa thể kết luận dứt khoát là các lần đá của Brian có độc lập hay không. Tuy nhiên, nhiều đánh giá thống kê về hiệu suất thi đấu thể thao cho thấy những nỗ lực như vậy gần như độc lập với nhau.

³⁷ Nếu cú đá thành công thứ tư của Brian ($k = 4$) nằm trong 5 lần đá đầu tiên, thì cậu ấy hoặc mất 4 lần đá hoặc 5 lần đá ($n = 4$ hoặc $n = 5$). Ta có $p = 0,8$ từ trước. Sử dụng phân phối nhị thức âm để tính xác suất khi $n = 4$ và $n = 5$, sau đó cộng hai xác suất này lại:

$$\begin{aligned} P(n = 4 \text{ hoặc } n = 5) &= P(n = 4) + P(n = 5) \\ &= \binom{4-1}{4-1} (0,8)^4 + \binom{5-1}{4-1} (0,8)^4 (1-0,8) = 1 \times 0,41 + 4 \times 0,082 = 0,41 + 0,33 = 0,74 \end{aligned}$$

4.4 PHÂN PHỐI NHỊ THỨC ÂM

SO SÁNH GIỮA PHÂN PHỐI NHỊ THỨC VÀ NHỊ THỨC ÂM

Trong phân phối nhị thức, ta thường có một số lần thử cố định, và ta xem xét số lần thành công trong những lần thử đó. Trong phân phối nhị thức âm, ta xem xét cần bao nhiêu lần thử để quan sát được một số lần thành công cố định, và yêu cầu rằng lần thử cuối cùng phải là một lần thành công.

BÀI TẬP THỰC HÀNH 4.49

Vào 70% số ngày, bệnh viện tiếp nhận ít nhất một bệnh nhân đau tim. Vào 30% số ngày, không có bệnh nhân đau tim nào được tiếp nhận. Xác định từng trường hợp sau là thuộc phân phối nhị thức hay nhị thức âm, và tính xác suất tương ứng.³⁸

G

- Xác suất để bệnh viện tiếp nhận bệnh nhân đau tim chính xác trong ba ngày trong tuần này là bao nhiêu?
- Xác suất để ngày thứ tư trong tuần là ngày thứ hai mà bệnh viện tiếp nhận bệnh nhân đau tim là bao nhiêu?
- Xác suất để ngày thứ năm trong tháng tới là ngày đầu tiên bệnh viện tiếp nhận bệnh nhân đau tim là bao nhiêu?

³⁸ Trong mỗi phần, xác suất thành công là $p = 0.7$. (a) Số ngày là cố định, vì vậy đây là trường hợp phân phối nhị thức. Các tham số là $k = 3$ (số ngày có bệnh nhân đau tim) và $n = 7$ (7 ngày trong tuần). Xác suất = 0,097. (b) Thành công cuối cùng (tiếp nhận bệnh nhân đau tim) được cố định vào ngày cuối cùng (ngày thứ 4), vì vậy ta nên áp dụng phân phối nhị thức âm. Các tham số là $k = 2$ (lần tiếp nhận thứ hai), $n = 4$ (ngày thứ tư là tổng số lần thử). Xác suất = 0,132. (c) Đây là trường hợp phân phối nhị thức âm, với $k = 1$ (lần đầu tiên có bệnh nhân) và $n = 5$ (ngày thứ 5). Xác suất = 0,006. Lưu ý: Trường hợp phân phối nhị thức âm khi $k = 1$ chính là phân phối hình học.

4.4 PHÂN PHỐI NHỊ THỨC ÂM

Bài tập

4.27 Tung xúc xắc. Tính các xác suất sau và chỉ ra mô hình phân phối xác suất phù hợp trong từng trường hợp. Bạn tung một con xúc xắc công bằng 5 lần. Xác suất là bao nhiêu để:

- ra số 6 đầu tiên ở lần tung thứ năm?
- ra đúng ba lần số 6?
- lần ra số 6 thứ ba rơi vào lần tung thứ năm?

4.28 Chơi phi tiêu. Tính các xác suất sau và xác định mô hình phân phối xác suất phù hợp. Một người chơi phi tiêu rất giỏi có xác suất trúng hồng tâm (vòng tròn đỏ ở giữa bảng phi tiêu) là 65%. Xác suất là bao nhiêu để người đó:

- trúng hồng tâm lần thứ 10 vào lần ném thứ 15?
- trúng hồng tâm 10 lần trong 15 lần ném?
- lần trúng hồng tâm đầu tiên rơi vào lần ném thứ ba?

4.29 Lấy mẫu tại trường. Trong một dự án môn xã hội học, bạn được yêu cầu thực hiện khảo sát với 20 sinh viên tại trường. Bạn quyết định đứng ngoài cổng-tin ký túc xá và khảo sát một mẫu ngẫu nhiên gồm 20 sinh viên rời cổng-tin sau bữa tối. Ký túc xá của bạn có 45% nam và 55% nữ.

- Mô hình xác suất nào phù hợp nhất để tính xác suất rằng người được khảo sát thứ 4 là người nữ thứ 2? Giải thích.
- Tính xác suất trong phần (a).
- Ba kịch bản có thể dẫn đến việc người thứ 4 được khảo sát là nữ thứ 2 là

$$\{M, M, F, F\}, \{M, F, M, F\}, \{F, M, M, F\}$$

Một điểm chung giữa các kịch bản này là: lần khảo sát cuối cùng luôn là nữ. Trong 3 lần khảo sát đầu tiên có 2 nam và 1 nữ. Sử dụng hệ số nhị thức (binomial coefficient) để xác nhận rằng có 3 cách sắp xếp 2 nam và 1 nữ.

- Hãy sử dụng những phát hiện trình bày trong phần (c) để giải thích vì sao công thức hệ số trong phân phối nhị thức âm là $\binom{n-1}{k-1}$ trong khi công thức hệ số trong phân phối nhị thức là $\binom{n}{k}$.

4.30 Giao bóng trong bóng chuyền. Một vận động viên bóng chuyền không giỏi có 15% cơ hội thực hiện giao bóng thành công, tức là đánh quả bóng qua lưới và rơi vào sân đối phương. Giả sử rằng các lần giao bóng là độc lập với nhau.

- Xác suất để lần giao bóng thứ 10 là lần giao bóng thành công thứ 3 là bao nhiêu?
- Giả sử cô ấy đã giao bóng thành công 2 lần trong 9 lần giao bóng đầu. Xác suất để lần giao bóng thứ 10 thành công là bao nhiêu?
- Mặc dù phần (a) và (b) mô tả cùng một kịch bản, nhưng xác suất bạn tính được sẽ khác nhau. Bạn có thể giải thích lý do cho sự khác biệt này không?

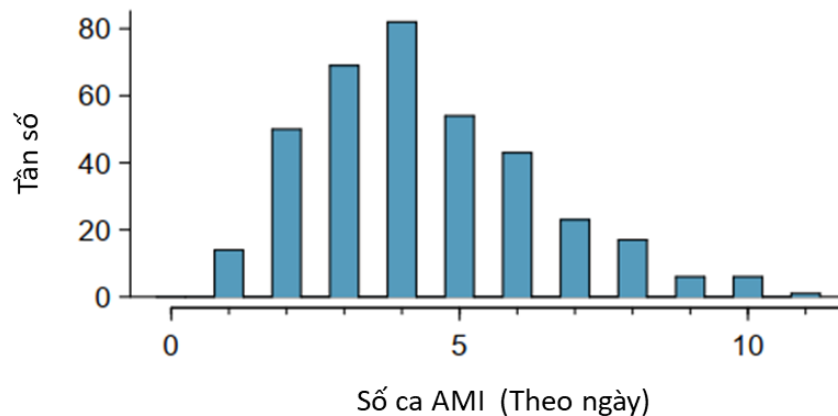
4.5 PHÂN PHỐI POISSON

4.5 Phân phối Poisson

VÍ DỤ 1.10

Có khoảng 8 triệu người sống tại thành phố New York. Vậy mỗi ngày, chúng ta có thể kỳ vọng có bao nhiêu người phải nhập viện vì nhồi máu cơ tim cấp (AMI) – tức là bị đau tim? Theo số liệu trong lịch sử, trung bình mỗi ngày có khoảng 4.4 người bị như vậy. Tuy nhiên, chúng ta cũng muốn biết phân phối xấp xỉ của số ca xảy ra. Nếu ta ghi lại số ca AMI hằng ngày trong suốt một năm, biểu đồ tần số (histogram) sẽ trông như thế nào?

Biểu đồ tần số số lượng ca AMI trong 365 ngày tại NYC được hiển thị trong Hình 4.12.³⁹ Trung bình mẫu (4.38) rất gần với trung bình từ dữ liệu trước trong lịch sử (4.4). Độ lệch chuẩn mẫu khoảng 2, và biểu đồ cho thấy khoảng 70% dữ liệu nằm trong khoảng từ 2.4 đến 6.4. Hình dạng của phân phối là đơn đỉnh và lệch phải.



Hình 4.12: Biểu đồ tần số số ca AMI xảy ra trong 365 ngày khác nhau tại thành phố New York.

Phân phối Poisson thường hữu ích để ước lượng số lần xảy ra của một sự kiện trong một quần thể lớn trên một đơn vị thời gian cố định. Ví dụ, hãy xem xét các sự kiện sau:

- Bị đau tim,
- Kết hôn, và
- Bị sét đánh.

Phân phối Poisson giúp ta mô tả số lần những sự kiện như vậy xảy ra trong một ngày cho một quần thể cố định, với điều kiện các cá nhân trong quần thể là độc lập. Phân phối Poisson cũng có thể áp dụng cho các đơn vị thời gian khác, như mỗi giờ hoặc mỗi tuần.

Biểu đồ tần số trong Hình 4.12 xấp xỉ một phân phối Poisson với **tỷ lệ** (rate) bằng 4.4. Tỷ lệ trong phân phối Poisson là số lần xảy ra trung bình trong một quần thể cố định trong mỗi đơn vị thời gian. Trong Ví dụ 4.50, đơn vị thời gian là ngày, quần thể là toàn bộ cư dân Thành phố New York, tỷ lệ lịch sử là 4.4 sự kiện mỗi ngày. Tham số trong phân phối Poisson chính là tỷ lệ – tức là số lượng sự kiện kỳ vọng sẽ xảy ra, và thường được ký hiệu là λ (lambda) hoặc μ . Dựa vào tỷ lệ này, ta có thể mô tả xác suất để quan sát chính xác k sự kiện trong một đơn vị thời gian.

³⁹ Dữ liệu trên là dữ liệu mô phỏng. Trong thực tế, ta nên kiểm tra xem có mối liên hệ giữa các ngày liên tiếp hay không.

4.5 PHÂN PHỐI POISSON

TPHÂN PHỐI POISSON

Giả sử chúng ta đang theo dõi một số sự kiện và số lượng sự kiện quan sát được tuân theo phân phối Poisson với tỷ lệ λ . Khi đó:

$$P(\text{quan sát được } k \text{ sự kiện}) = \frac{\lambda^k e^{-\lambda}}{k!}$$

Trong đó, k có thể nhận các giá trị $0, 1, 2, \dots$, và $k!$ là giai thừa của k (đã được mô tả ở trang 150), $e \approx 2.718$ là cơ số của logarit tự nhiên. Kỳ vọng trung bình và độ lệch chuẩn của phân phối Poisson lần lượt là λ và $\sqrt{\lambda}$.

Chúng ta sẽ dành việc thiết lập đầy đủ các điều kiện chặt chẽ cho phân phối Poisson cho một khóa học sau. Tuy nhiên, dưới đây là một số nguyên tắc đơn giản giúp đánh giá ban đầu xem mô hình Poisson có phù hợp hay không.

Một biến ngẫu nhiên có thể tuân theo phân phối Poisson nếu ta đang đếm số lượng sự kiện xảy ra, quần thể sinh ra các sự kiện này là lớn, các sự kiện xảy ra một cách độc lập với nhau.

Ngay cả khi các sự kiện không hoàn toàn độc lập – ví dụ, thứ Bảy và Chủ nhật là những ngày đặc biệt phổ biến cho việc tổ chức đám cưới – mô hình Poisson vẫn có thể hợp lý nếu ta cho phép tỷ lệ λ thay đổi theo thời gian. Trong ví dụ về đám cưới, tỷ lệ sẽ được mô hình hóa là cao hơn vào cuối tuần so với các ngày trong tuần. Ý tưởng mô hình hóa tỷ lệ λ của phân phối Poisson theo một biến thứ hai, chẳng hạn như ngày trong tuần, chính là nền tảng cho một số phương pháp nâng cao hơn, thuộc lĩnh vực **mô hình tuyến tính tổng quát** (generalized linear models). Các chương 8 và 9 sẽ trình bày nền tảng của mô hình tuyến tính (linear models).

4.5 PHÂN PHỐI POISSON

Bài tập

4.31 Khách hàng tại quán cà phê. Một quán cà phê phục vụ trung bình 75 khách mỗi giờ trong giờ cao điểm buổi sáng.

- Phân phối nào mà chúng ta đã học là phù hợp nhất để tính xác suất một số lượng khách hàng nhất định đến trong vòng một giờ vào thời điểm này trong ngày?
- Trung bình và độ lệch chuẩn của số lượng khách mà quán cà phê này phục vụ trong một giờ vào thời điểm này trong ngày là bao nhiêu?
- Việc chỉ có 60 khách đến quán cà phê này trong một giờ vào thời điểm này trong ngày có được xem là thấp bất thường không?
- Tính xác suất quán cà phê này phục vụ 70 khách trong một giờ vào thời điểm này trong ngày.

4.32 Lỗi đánh máy của người đánh tốc ký. Một người đánh tốc ký cho tòa án rất thành thạo, trung bình sẽ mắc một lỗi đánh máy mỗi giờ.

- Phân phối xác suất nào là phù hợp nhất để tính xác suất một số lượng lỗi đánh máy nhất định mà người đánh tốc ký này mắc trong một giờ?
- Trung bình và độ lệch chuẩn của số lỗi đánh máy mà người đánh tốc ký này mắc là bao nhiêu?
- Việc người đánh tốc ký này mắc 4 lỗi trong một giờ có được xem là bất thường không?
- Tính xác suất người đánh tốc ký này mắc tối đa 2 lỗi trong một giờ.

4.33 Có bao nhiêu xe đến? Từ thứ Hai đến thứ Năm khi không phải ngày lễ, số lượng phương tiện trung bình đến một cửa hàng bán lẻ nhất định từ 2 giờ đến 3 giờ chiều mỗi ngày là 6,5 xe, và số xe đến vào bất kỳ ngày nào tuân theo phân phối Poisson.

- Xác suất chính xác có 5 xe đến vào thứ Hai tuần tới là bao nhiêu?
- Xác suất có 0, 1 hoặc 2 xe đến vào thứ Hai tuần tới từ 2 giờ đến 3 giờ chiều là bao nhiêu?
- Có trung bình 11,7 người đến trong cùng khung giờ từ các phương tiện. Việc số người đến bằng xe trong khung giờ này cũng tuân theo phân phối Poisson có hợp lý không? Giải thích.

4.34 Mất hành lý. Đôi khi một hãng hàng không sẽ làm mất một hành lý. Giả sử một hãng hàng không nhỏ nhận thấy rằng có thể mô hình hóa hợp lý số lượng hành lý bị mất mỗi ngày trong tuần bằng mô hình Poisson với trung bình là 2,2 hành lý.

- Xác suất hãng hàng không sẽ không làm mất hành lý nào vào thứ Hai tuần tới là bao nhiêu?
- Xác suất hãng hàng không sẽ làm mất 0, 1 hoặc 2 hành lý vào thứ Hai tuần tới là bao nhiêu?
- Giả sử hãng hàng không mở rộng trong vòng 3 năm tới, gấp đôi số chuyến bay, và CEO hỏi bạn liệu có hợp lý để tiếp tục sử dụng mô hình Poisson với trung bình 2,2 hay không. Đây là khuyến nghị phù hợp? Giải thích.

4.5 PHÂN PHỐI POISSON

Bài tập chương

4.35 Tiền thắng trong trò chơi roulette. Trong trò chơi roulette, một bánh xe được quay và bạn đặt cược vào nơi bánh xe sẽ dừng lại. Một kiểu cược phổ biến là cược vào ô màu đỏ; loại cược này có xác suất thắng là 18/38. Nếu bánh xe dừng ở ô đỏ, bạn sẽ nhận lại gấp đôi số tiền đã cược. Nếu không, bạn sẽ mất số tiền cược. Giả sử bạn chơi 3 lần, mỗi lần cược 1 đô la. Gọi YY là tổng số tiền thắng hoặc thua. Hãy xây dựng mô hình xác suất cho biến ngẫu nhiên Y .

4.36 Vượt tỷ lệ trên xa lộ I-5, Phần I. Phân phối tỷ lệ của các xe ô tô cá nhân lưu thông trên xa lộ liên bang I-5 ở California gần như tuân theo phân phối chuẩn, với trung bình là 72,6 dặm/giờ và độ lệch chuẩn là 4,78 dặm/giờ.⁴⁰

- Có bao nhiêu phần trăm xe ô tô cá nhân chạy chậm hơn 80 dặm/giờ?
- Có bao nhiêu phần trăm xe ô tô cá nhân chạy trong khoảng từ 60 đến 80 dặm/giờ?
- Tỷ lệ của 5% xe chạy nhanh nhất là bao nhiêu?
- Giới hạn tỷ lệ trên đoạn đường này là 70 dặm/giờ. Ước lượng khoảng bao nhiêu phần trăm xe chạy vượt quá giới hạn tỷ lệ trên đoạn xa lộ này.

4.37 Tuyển sinh đại học. Giả sử một trường đại học thông báo rằng họ đã gửi thư trúng tuyển cho 2.500 sinh viên cho khóa nhập học năm tới. Tuy nhiên, trường chỉ có chỗ ở ký túc xá cho 1.786 sinh viên năm nhất. Nếu xác suất một sinh viên đã được nhận thực sự quyết định nhập học là 70%, thì xác suất xấp xỉ để trường đại học không đủ chỗ ở ký túc xá cho sinh viên năm nhất là bao nhiêu?

4.38 Vượt tỷ lệ trên xa lộ I-5, Phần II. Bài tập 4.36 cho biết phân phối tỷ lệ xe chạy trên xa lộ I-5 ở California gần như tuân theo phân phối chuẩn với trung bình 72,6 dặm/giờ và độ lệch chuẩn 4,78 dặm/giờ. Giới hạn tỷ lệ trên đoạn đường này là 70 dặm/giờ.

- Một cảnh sát giao thông đang ẩn nấp bên lề đường cao tốc. Xác suất để 5 chiếc xe chạy qua đều không vượt tỷ lệ là bao nhiêu? Giả sử tỷ lệ của các xe là độc lập.
- Trung bình, viên cảnh sát sẽ cần quan sát bao nhiêu xe cho đến khi bắt gặp chiếc xe đầu tiên chạy vượt tỷ lệ? Độ lệch chuẩn của số xe cần quan sát đó là bao nhiêu?

4.39 Phí bảo hiểm xe hơi. Giả sử một bài báo cho biết phân phối phí bảo hiểm xe hơi của cư dân bang California xấp xỉ phân phối chuẩn, với trung bình là 1.650 đô la. Bài báo cũng cho biết rằng 25% cư dân California trả trên 1.800 đô la tiền bảo hiểm.

- Z-score tương ứng với 25% cao nhất (hay bách phân vị thứ 75) trong phân phối chuẩn là bao nhiêu?
- Mức phí bảo hiểm trung bình là bao nhiêu? Ngưỡng để lọt vào nhóm 25% cao nhất là bao nhiêu?
- Xác định độ lệch chuẩn của phân phối phí bảo hiểm xe hơi tại California.

4.40 Điểm SAT. Điểm thi SAT (trên thang điểm 1600) có phân phối chuẩn với trung bình là 1.100 và độ lệch chuẩn là 200. Giả sử hội đồng nhà trường trao giấy chứng nhận xuất sắc cho tất cả học sinh đạt ít nhất 1.350 điểm SAT, và giả sử ta chọn ngẫu nhiên một học sinh trong nhóm được vinh danh đó. Xác suất để học sinh này có điểm SAT ít nhất là 1.500 là bao nhiêu? (Phần kiến thức về xác suất có điều kiện ở Mục 3.2 sẽ hữu ích cho câu hỏi này.)

4.41 Phụ nữ đã kết hôn. Khảo sát Cộng đồng Hoa Kỳ (American Community Survey) ước tính rằng 47,1% phụ nữ từ 15 tuổi trở lên đã kết hôn.⁴¹

- Ta chọn ngẫu nhiên 3 người phụ nữ trong độ tuổi này. Xác suất để người phụ nữ thứ ba là người duy nhất đã kết hôn là bao nhiêu?
- Xác suất để cả ba người phụ nữ được chọn ngẫu nhiên đều đã kết hôn là bao nhiêu?
- Trung bình, bạn sẽ cần khảo sát bao nhiêu người phụ nữ để chọn được một người đã kết hôn? Độ lệch chuẩn là bao nhiêu?
- Nếu tỷ lệ phụ nữ đã kết hôn thực tế chỉ là 30%, thì bạn sẽ cần khảo sát bao nhiêu người để chọn được một người đã kết hôn? Độ lệch chuẩn là bao nhiêu?
- Dựa trên câu trả lời ở (c) và (d), việc giảm xác suất của một sự kiện ảnh hưởng như thế nào đến kỳ vọng và độ lệch chuẩn của thời gian chờ đợi cho đến khi thành công?

⁴⁰ S. Johnson and D. Murray. "Empirical Analysis of Truck and Automobile Speeds on Rural Interstates: Impact of Posted Speed Limits". In: *Transportation Research Board 89th Annual Meeting*. 2010.

⁴¹ U.S. Census Bureau, 2010 American Community Survey, Marital Status.

4.5 PHÂN PHỐI POISSON

4.42 Tỷ lệ phản hồi khảo sát. Trung tâm nghiên cứu Pew báo cáo rằng tỷ lệ phản hồi điển hình cho các khảo sát của họ chỉ là 9%. Nếu trong một khảo sát cụ thể, có 15.000 hộ gia đình được liên hệ, xác suất để ít nhất 1.500 hộ đồng ý tham gia khảo sát là bao nhiêu? ⁴²

4.43 Hành lý quá cân. Giả sử cân nặng hành lý ký gửi của hành khách đi máy bay phân bố xấp xỉ chuẩn với trung bình 45 pound và độ lệch chuẩn 3,2 pound. Hầu hết các hãng hàng không sẽ thu phí với hành lý nặng trên 50 pound. Xác định tỷ lệ phần trăm hành khách sẽ phải chịu loại phí này.

4.44 Chiều cao trẻ 10 tuổi – Phần I. Chiều cao của trẻ 10 tuổi (không phân biệt giới tính) phân bố xấp xỉ chuẩn với trung bình 55 inch và độ lệch chuẩn 6 inch.

- Xác suất một trẻ được chọn ngẫu nhiên có chiều cao thấp hơn 48 inch là bao nhiêu?
- Xác suất một trẻ được chọn ngẫu nhiên có chiều cao trong khoảng từ 60 đến 65 inch là bao nhiêu?
- Nếu 10% trẻ cao nhất được xem là "rất cao", thì mức chiều cao ngưỡng để được xem là "rất cao" là bao nhiêu?

4.45 Mua sách trên Ebay. Giả sử bạn đang cân nhắc mua sách giáo khoa môn hóa học đắt tiền trên Ebay. Dữ liệu từ các phiên đấu giá trước cho thấy giá sách này phân bố gần chuẩn với trung bình \$89 và độ lệch chuẩn \$15.

- Xác suất một phiên đấu giá bất kỳ có giá kết thúc trên \$100 là bao nhiêu?
- Ebay cho phép bạn đặt mức giá tối đa mà bạn sẵn sàng trả, để nếu có người khác trả giá cao hơn bạn, hệ thống sẽ tự động nâng giá của bạn lên (tối đa đến mức bạn đã đặt). Nếu bạn chỉ tham gia một phiên đấu giá, đâu là ưu và nhược điểm của việc đặt giá tối đa quá cao hoặc quá thấp? Còn nếu bạn tham gia nhiều phiên đấu giá thì sao?
- Nếu bạn quan sát 10 phiên đấu giá, thì bạn nên sử dụng phân vị khoảng bao nhiêu để đặt mức giá tối đa nhằm có cơ hội thắng ít nhất một phiên đấu giá?
- Nếu bạn sẵn sàng theo dõi tối đa 10 phiên đấu giá, bạn nên đặt giá tối đa khoảng bao nhiêu để có thể khá chắc rằng mình sẽ mua được một cuốn sách trong số đó?

4.46 Chiều cao trẻ 10 tuổi – Phần II. Chiều cao của trẻ 10 tuổi (không phân biệt giới tính) phân bố xấp xỉ chuẩn với trung bình 55 inch và độ lệch chuẩn 6 inch.

- Yêu cầu chiều cao tối thiểu để chơi trò "Batman the Ride" tại công viên giải trí Six Flags Magic Mountain là 54 inch. Bao nhiêu phần trăm trẻ 10 tuổi không đủ chiều cao để chơi trò này?
- Giả sử có bốn trẻ 10 tuổi. Xác suất có ít nhất hai em đủ chiều cao để chơi "Batman the Ride" là bao nhiêu?
- Giả sử bạn đang làm việc tại công viên và muốn hiểu rõ hơn về nhân khẩu học của khách tham quan. Bạn đang đếm người khi họ vào công viên. Xác suất rằng đứa trẻ 10 tuổi đầu tiên đủ chiều cao để chơi là người thứ 3 đi vào công viên là bao nhiêu?
- Xác suất rằng đứa trẻ 10 tuổi thứ năm đủ chiều cao để chơi là người thứ 12 vào công viên là bao nhiêu?

4.47 Chiều cao trẻ 10 tuổi – Phần III. Chiều cao của trẻ 10 tuổi (không phân biệt giới tính) phân bố xấp xỉ chuẩn với trung bình 55 inch và độ lệch chuẩn 6 inch.

- Tỷ lệ trẻ 10 tuổi có chiều cao trên 76 inch là bao nhiêu?
- Nếu có 2.000 trẻ 10 tuổi vào công viên Six Flags Magic Mountain trong một ngày, hãy tính số lượng trẻ dự kiến sẽ cao từ 76 inch trở lên. (Có thể giả định chiều cao của các trẻ là độc lập.)
- Sử dụng phân phối nhị thức, hãy tính xác suất rằng không có trẻ nào trong số 2.000 em cao từ 76 inch trở lên.
- Giả sử số trẻ 10 tuổi cao từ 76 inch trở lên vào công viên trong một ngày tuân theo phân phối Poisson với trung bình là giá trị ở phần (b). Hãy sử dụng phân phối Poisson để xác định xác suất không có em nào cao từ 76 inch trở lên.

4.48 Bài kiểm tra trắc nghiệm. Trong một bài kiểm tra trắc nghiệm có 5 câu hỏi và mỗi câu có 4 lựa chọn (a, b, c, d). Robin chưa học bài gì cả và quyết định đoán ngẫu nhiên câu trả lời. Xác suất rằng:

- Câu đầu tiên Robin trả lời đúng là câu thứ 3?
- Robin trả lời đúng chính xác 3 hoặc 4 câu hỏi?
- Robin trả lời đúng phần lớn số câu hỏi?

⁴² Pew Research Center, Assessing the Representativeness of Public Opinion Surveys, May 15, 2012.