

Chương 5

Cơ sở cho suy luận thống kê

5.1 Ước lượng điểm và Biến động mẫu

5.2 Khoảng tin cậy cho một tỷ lệ

5.3 Kiểm định giả thuyết cho một tỷ lệ

Suy luận thống kê chủ yếu liên quan đến việc hiểu và định lượng sự không chắc chắn của các ước lượng tham số. Mặc dù các phương trình và chi tiết có thể thay đổi tùy theo bối cảnh, nhưng cơ sở cho suy luận là giống nhau trong toàn bộ ngành thống kê.

Chúng ta bắt đầu với một chủ đề quen thuộc: ý tưởng sử dụng tỷ lệ mẫu để ước lượng tỷ lệ của tổng thể. Tiếp theo, chúng ta tạo ra một cái được gọi là *khoảng tin cậy (confidence interval)*, tức là một khoảng giá trị hợp lý mà chúng ta có thể tìm thấy giá trị thực của tổng thể. Cuối cùng, chúng ta giới thiệu *khung kiểm định giả thuyết (hypothesis testing framework)*, cho phép chúng ta đánh giá chính thức các tuyên bố về tổng thể, chẳng hạn như liệu một cuộc khảo sát có cung cấp bằng chứng mạnh mẽ rằng một ứng cử viên có sự ủng hộ của đa số cử tri hay không.



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

www.openintro.org/os

5.1 ƯỚC LƯỢNG ĐIỂM VÀ BIẾN ĐỘNG MẪU

5.1 Ước lượng điểm và Biến động mẫu

Các công ty như Pew Research thường thực hiện khảo sát như một cách để hiểu quan điểm hay mức độ hiểu biết của công chúng về nhiều chủ đề, bao gồm chính trị, kiến thức khoa học, mức độ nhận diện thương hiệu và nhiều lĩnh vực khác. Mục tiêu cuối cùng của việc khảo sát thường là sử dụng các phản hồi thu được để ước lượng ý kiến hoặc sự hiểu biết của toàn bộ dân số.

5.1.1 Ước lượng điểm và sai số

Giả sử một cuộc khảo sát cho thấy tỷ lệ ủng hộ Tổng thống Mỹ là 45%. Chúng ta sẽ xem 45% này là một **ước lượng điểm** của tỷ lệ ủng hộ mà chúng ta có thể thấy nếu thu thập phản hồi từ toàn bộ dân số. Tỷ lệ phản hồi từ toàn bộ dân số này thường được gọi là **tham số cần quan tâm**. Khi tham số là một tỷ lệ, nó thường được ký hiệu là p , và chúng ta thường gọi tỷ lệ mẫu là $\hat{p}$ (đọc là “p-hat”). Trừ khi chúng ta thu thập phản hồi từ mọi cá nhân trong dân số, p vẫn là một ẩn số, và chúng ta sử dụng $\hat{p}$ như một ước lượng cho p . Sai số giữa kết quả khảo sát và tham số thật được gọi là **sai số** của ước lượng. Thông thường, sai số bao gồm hai yếu tố: sai số ngẫu nhiên do lấy mẫu và thiên lệch (bias).

Sai số do lấy mẫu (hay còn gọi là sampling uncertainty) mô tả mức độ mà một ước lượng có thể thay đổi từ mẫu này sang mẫu khác. Ví dụ, một mẫu có thể cho kết quả thấp hơn 1%, trong khi một mẫu khác có thể cao hơn 3%. Phần lớn nội dung thống kê, bao gồm cả cuốn sách này, tập trung vào việc hiểu và định lượng sai số ngẫu nhiên do lấy mẫu. **Kích thước mẫu**, thường được ký hiệu là n , sẽ rất hữu ích trong việc định lượng sai số này.

Thiên lệch (bias) mô tả một xu hướng có hệ thống khiến kết quả khảo sát luôn cao hơn hoặc thấp hơn giá trị thực trong dân số. Ví dụ, nếu chúng ta thực hiện một cuộc khảo sát sinh viên hỏi về việc ủng hộ xây sân vận động mới và đặt câu hỏi là: “*Bạn có ủng hộ trường học của mình bằng cách ủng hộ quỹ xây sân vận động mới không?*”, thì rất có thể kết quả sẽ thiên lệch, vì cách diễn đạt câu hỏi có thể khiến người được hỏi cảm thấy có trách nhiệm phải trả lời “ủng hộ”. Chúng ta cố gắng giảm thiểu thiên lệch thông qua quy trình thu thập dữ liệu cẩn thận - đã được thảo luận trong Chương 1 và là chủ đề của nhiều cuốn sách khác.

5.1.2 Hiểu về biến thiên ước lượng điểm

Giả sử tỷ lệ người trưởng thành ở Mỹ ủng hộ việc mở rộng năng lượng mặt trời là $p = 0,88$, đây là tham số cần quan tâm của chúng ta.² Nếu chúng ta thực hiện một cuộc khảo sát với 1000 người trưởng thành ở Mỹ về chủ đề này, kết quả ước lượng sẽ không hoàn toàn chính xác, vậy thì kết quả khảo sát có thể gần với 88% đến mức nào? Chúng ta muốn hiểu *$\hat{p}$ sẽ dao động ra sao khi p thật sự là 0,88*. Hãy cùng tìm hiểu! Chúng ta có thể mô phỏng các phản hồi từ một mẫu ngẫu nhiên đơn giản gồm 1000 người trưởng thành ở Mỹ, điều này khả thi chỉ vì ta đã biết trước mức độ ủng hộ thực tế là 0,88. Dưới đây là cách chúng ta có thể mô phỏng quá trình này:

1. Vào năm 2018, có khoảng 250 triệu người trưởng thành ở Mỹ. Hãy tưởng tượng viết chữ “ủng hộ” lên 88% trong số 250 triệu mảnh giấy và viết “không ủng hộ” lên 12% còn lại.
2. Trộn đều các mảnh giấy và rút ra 1000 mảnh để đại diện cho mẫu khảo sát gồm 1000 người.
3. Tính tỷ lệ số mảnh giấy ghi “ủng hộ” trong mẫu đó.

Có ai tình nguyện thực hiện mô phỏng này không? Chắc là không. Thực hiện mô phỏng với 250 triệu mảnh giấy sẽ rất mất thời gian và tốn kém. Tuy nhiên, chúng ta có thể mô phỏng quá trình này bằng máy tính.

¹ Đừng nhầm lẫn $\hat{p}$ với “phat” - một từ lóng để chỉ điều gì đó “ngẫu”, như cuốn sách này chẳng hạn!

² Chúng ta chưa thực hiện điều tra dân số để đo chính xác giá trị này, nhưng một khảo sát có cỡ mẫu rất lớn đã chỉ ra rằng mức độ ủng hộ thực tế khoảng 88%.

³ 88% viết dưới dạng tỷ lệ là 0,88. Người ta thường chuyển đổi linh hoạt giữa tỷ lệ phần trăm và tỷ lệ thập phân. Tuy nhiên, các công thức trong cuốn sách này luôn sử dụng tỷ lệ thập phân.

5.1 ƯỚC LƯỢNG ĐIỂM VÀ BIẾN ĐỘNG MẪU

Đoạn mã: Chúng tôi đã viết một đoạn chương trình ngắn trong Hình 5.1 nếu bạn tò mò muốn biết mã máy tính trông như thế nào. Trong mô phỏng này, mẫu khảo sát cho một ước lượng điểm là $\hat{p}_a = 0,894$. Chúng ta biết tỷ lệ thật trong dân số trong mô phỏng là $p = 0,88$, vì vậy ta biết sai số của ước lượng là $0,894 - 0,88 = +0,014$.

```
# 1. Tạo một tập gồm 250 triệu phần tử, trong đó 88% là "ủng hộ"
#    và 12% là "không ủng hộ".
pop_size <- 250000000
possible_entries <- c(rep("support", 0.88*pop_size), rep("not", 0.12*pop_size))
# # 2. Lấy mẫu gồm 1000 phần tử không hoàn lại.
sampled_entries <- sample(possible_entries, size = 1000)
# 3. Tính p-hat: đếm số phần tử là "support", rồi chia cho kích thước mẫu.
sum(sampled_entries == "support") / 1000
```

Hình 5.1: Đối với những bạn tò mò, đây là đoạn mã dùng để mô phỏng một lần giá trị $\hat{p}$ bằng phần mềm thống kê R. Mỗi dòng bắt đầu bằng dấu # là phần **chú thích**, giải thích bằng ngôn ngữ thông thường về chức năng của đoạn mã. Bạn có thể học thêm về phần mềm R tại openintro.org/book/os.

Một lần mô phỏng thì chưa đủ để giúp chúng ta hình dung rõ về phân bố của các ước lượng có thể xảy ra trong thực tế, nên chúng ta cần thực hiện nhiều lần mô phỏng hơn. Trong lần mô phỏng thứ hai, ta thu được $\hat{p}_2 = 0,885$, sai số là $+0,005$. Trong một lần khác, $\hat{p}_a = 0,878$, sai số là $-0,002$. Và ở một lần khác nữa, $\hat{p}_4 = 0,859$, sai số là $-0,021$. Với sự hỗ trợ của máy tính, chúng tôi đã thực hiện 10000 lần mô phỏng và tạo ra một biểu đồ histogram từ tất cả các kết quả trong Hình 5.2. Phân bố các tỷ lệ mẫu này được gọi là **phân phối lấy mẫu**. Ta có thể mô tả phân phối này như sau:

Trung tâm. Trung tâm của phân phối là $\bar{x}_{\hat{p}} = 0,880$, đúng bằng với giá trị tham số p . Lưu ý rằng mô phỏng này bắt chước lấy mẫu ngẫu nhiên đơn giản từ dân số - một chiến lược lấy mẫu trực tiếp và hiệu quả giúp tránh thiên lệch do lấy mẫu.

Độ phân tán. Độ lệch chuẩn của phân phối là $s_{\hat{p}} = 0,010$. Khi nói về phân phối lấy mẫu hay mức độ biến thiên của ước lượng điểm, chúng ta thường dùng thuật ngữ **sai số chuẩn (standard error)** thay vì độ lệch chuẩn, và ký hiệu là $SE_{\hat{p}}$ để chỉ sai số chuẩn tương ứng với tỷ lệ mẫu.

Hình dạng. Phân phối này có hình dạng đối xứng và hình chuông, *rất giống với phân phối chuẩn*.

Những kết quả này thật đáng khích lệ! Khi tỷ lệ dân số là $p = 0,88$ và kích thước mẫu là $n = 1000$, thì tỷ lệ mẫu $\hat{p}$ thường cho ước lượng khá chính xác của tỷ lệ dân số p . Chúng ta cũng quan sát thấy một điều thú vị là biểu đồ histogram giống với phân phối chuẩn.

PHÂN PHỐI LẤY MẪU KHÔNG THỂ QUAN SÁT ĐƯỢC TRỰC TIẾP, NHƯNG CHÚNG TA LUÔN GHI NHỚ CHÚNG

Trong các ứng dụng thực tế, chúng ta không bao giờ thực sự quan sát được phân phối lấy mẫu, nhưng việc luôn hình dung một ước lượng điểm là xuất phát từ một phân phối giả định như vậy lại rất hữu ích. Hiểu được phân phối lấy mẫu sẽ giúp chúng ta mô tả và lý giải các ước lượng điểm mà chúng ta quan sát được.

VÍ DỤ 5.1

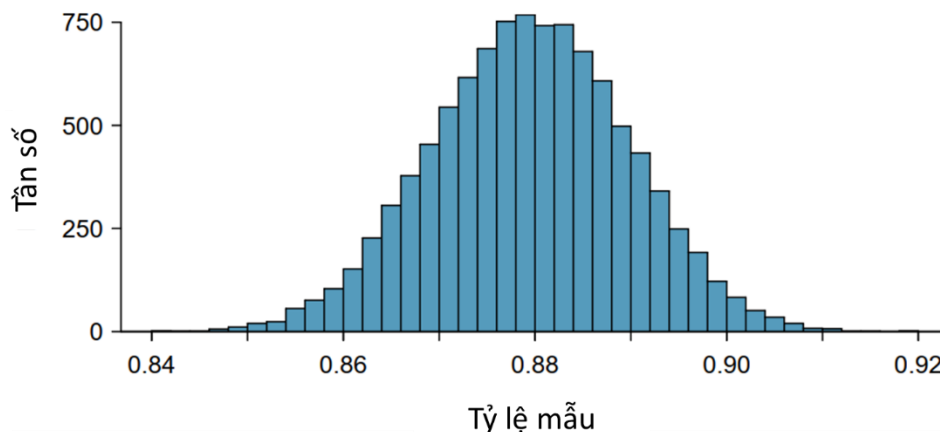
Nếu chúng ta sử dụng một kích thước mẫu nhỏ hơn nhiều, chẳng hạn $n = 50$, bạn có đoán rằng sai số chuẩn của $\hat{p}$ sẽ lớn hơn hay nhỏ hơn khi so với $n = 1000$?

E

Trực giác mách bảo rằng nhiều dữ liệu sẽ tốt hơn ít dữ liệu, và điều đó nói chung là chính xác! Khi $p = 0,88$ và $n = 50$, thì sai số thường gặp sẽ lớn hơn so với khi $n = 1000$.

Ví dụ 5.1 nhấn mạnh một đặc điểm quan trọng mà chúng ta sẽ còn gặp đi gặp lại: Một mẫu càng lớn thì thường cho một ước lượng điểm càng chính xác.

5.1 ƯỚC LƯỢNG ĐIỂM VÀ BIẾN ĐỘNG MẪU



Hình 5.2: Biểu đồ tần suất của 10000 tỷ lệ mẫu, trong đó mỗi mẫu được lấy từ một dân số có tỷ lệ thật là 0,88, với cỡ mẫu $n = 1000$

5.1.3 Định lý Giới Hạn Trung Tâm

Phân phối trong Hình 5.2 trông rất giống với một phân phối chuẩn. Điều này không phải là một hiện tượng ngẫu nhiên; nó là kết quả của một nguyên lý tổng quát được gọi là **Định lý Giới Hạn Trung Tâm**.

ĐỊNH LÝ GIỚI HẠN TRUNG TÂM VÀ ĐIỀU KIỆN THÀNH CÔNG - THẤT BẠI

Khi các quan sát là độc lập và kích thước mẫu đủ lớn, tỷ lệ mẫu $\hat{p}$ sẽ có xu hướng tuân theo phân phối chuẩn với trung bình và sai số chuẩn như sau:

$$\mu_{\hat{p}} = p$$

$$SE_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

Để Định lý Giới Hạn Trung Tâm được áp dụng, kích thước mẫu thường được coi là đủ lớn khi $np \geq 10$ và $n(1-p) \geq 10$, điều này được gọi là **điều kiện thành công – thất bại**.

Định lý Giới Hạn Trung Tâm là một khái niệm vô cùng quan trọng và tạo nền tảng cho phần lớn các phương pháp thống kê. Khi bắt đầu áp dụng định lý này, cần lưu ý hai điều kiện kỹ thuật sau: Các quan sát phải độc lập, và kích thước mẫu đủ lớn, sao cho $np \geq 10$ và $n(1-p) \geq 10$.

VÍ DỤ 5.2

Trước đó, chúng ta đã ước lượng trung bình và sai số chuẩn của $\hat{p}$ bằng dữ liệu mô phỏng khi $p = 0,88$ và $n = 1000$. Hãy kiểm tra xem Định lý Giới Hạn Trung Tâm có áp dụng được không và phân phối lấy mẫu có xấp xỉ phân phối chuẩn hay không?

E

Tính độc lập. Có $n = 1000$ quan sát cho mỗi ước lượng tỷ lệ mẫu $\hat{p}$, và mỗi quan sát đều là rút thăm độc lập. *Cách phổ biến nhất để đảm bảo tính độc lập là lấy mẫu ngẫu nhiên đơn giản.*

Điều kiện thành công - thất bại. Chúng ta có thể xác nhận quy mô mẫu đủ lớn bằng cách kiểm tra điều kiện thành công-thất bại và xác nhận hai giá trị tính toán lớn hơn 10:

$$np = 1000 \times 0,88 = 880 \geq 10$$

$$n(1-p) = 1000 \times (1 - 0,88) = 120 \geq 10$$

Cả hai điều kiện đều được đáp ứng, nên có thể áp dụng Định lý Giới Hạn Trung Tâm, và hợp lý khi mô hình hóa $\hat{p}$ bằng phân phối chuẩn.

5.1 ƯỚC LƯỢNG ĐIỂM VÀ BIẾN ĐỘNG MẪU

CÁCH XÁC MINH CÁC QUAN SÁT TRONG MẪU LÀ ĐỘC LẬP

Trong thí nghiệm, các đối tượng được coi là độc lập nếu họ được phân ngẫu nhiên vào các nhóm điều trị.

Nếu các quan sát được lấy từ một mẫu ngẫu nhiên đơn thì các quan sát này được coi là độc lập.

Nếu mẫu đến từ một quá trình có vẻ ngẫu nhiên, ví dụ như lỗi trình hoạch xảy ra trên dây chuyền sản xuất, thì việc kiểm tra tính độc lập sẽ khó hơn. Trong trường hợp này, bạn nên sử dụng phán đoán tốt nhất của mình.

Một điều kiện bổ sung đôi khi được thêm vào là: kích thước mẫu không vượt quá 10% tổng dân số. Nếu mẫu vượt quá 10% dân số, các phương pháp chúng ta sử dụng có thể ước lượng sai số lấy mẫu hơi cao hơn thực tế, so với các phương pháp nâng cao hơn.⁴ Tuy nhiên, điều này hiếm khi là vấn đề, và khi có thì các phương pháp của ta vẫn thường bảo thủ hơn nên ta xem điều kiện bổ sung này là không bắt buộc.

VÍ DỤ 5.3

Tính trung bình lý thuyết và sai số chuẩn (standard error) của $\hat{p}$ khi $p = 0,88$ và $n = 1000$, theo Định lý Giới Hạn Trung Tâm.

(E)

Trung bình của $\hat{p}$ chính là tỷ lệ dân số: $\mu_{\hat{p}} = 0,88$.

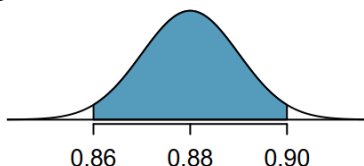
Việc tính toán sai số chuẩn của $\hat{p}$ sử dụng công thức sau:

$$SE_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0,88(1-0,88)}{1000}} = 0,01$$

VÍ DỤ 5.4

Ước tính xác suất tỷ lệ mẫu $\hat{p}$ nằm trong khoảng 0,02 (2%) so với tỷ lệ thật $p = 0,88$. Dựa vào Ví dụ 5.2 và 5.3, ta biết rằng phân phối của $\hat{p}$ xấp xỉ chuẩn $N(\mu_{\hat{p}} = 0,88, SE_{\hat{p}} = 0,010)$.

Sau rất nhiều bài tập ở Mục 4.1, hy vọng ví dụ về phân phối chuẩn này đã trở nên quen thuộc! Chúng ta muốn hiểu tỷ lệ $\hat{p}$ nằm trong khoảng từ 0,86 đến 0,90 là bao nhiêu:



(E)

Với $\mu_{\hat{p}} = 0,88$ và $SE_{\hat{p}} = 0,010$, ta có thể tính điểm Z cho hai điểm giới hạn trái và phải:

$$Z_{0,86} = \frac{0,86 - 0,88}{0,010} = -2$$

$$Z_{0,90} = \frac{0,90 - 0,88}{0,010} = 2$$

Ta có thể sử dụng phần mềm thống kê, máy tính đồ thị, hoặc bảng phân phối chuẩn để tìm diện tích ở hai bên đuôi. Trong mọi trường hợp, ta sẽ thấy mỗi bên đuôi có diện tích là 0,0228. Tổng diện tích hai đuôi là: $2 \times 0,0228 = 0,0456$. Điều đó để lại diện tích ở phần giữa là 0,9544. Nói cách khác, khoảng 95,44% phân phối lấy mẫu trong Hình 5.2 nằm trong khoảng $\pm 0,02$ so với tỷ lệ dân số, $p = 0,8$.

⁴Ví dụ, ta có thể sử dụng cái gọi là **hệ số điều chỉnh dân số hữu hạn** (finite population correction factor): nếu cỡ mẫu là n và kích thước dân số là N , thì ta có thể nhân công thức sai số chuẩn thông thường với $\sqrt{\frac{(N-n)}{(N-1)}}$ để có ước lượng nhỏ hơn và chính xác hơn cho sai số chuẩn thực tế. Khi $n < 0,1 \times N$, yếu tố điều chỉnh này là tương đối nhỏ.

5.1 ƯỚC LƯỢNG ĐIỂM VÀ BIẾN ĐỘNG MẪU

BÀI TẬP THỰC HÀNH 5.5

G

Trong Ví dụ 5.1, chúng ta đã thảo luận rằng cỡ mẫu nhỏ hơn có xu tạo ra ước lượng kém tin cậy hơn.

Hãy giải thích trực giác này thể hiện như thế nào trong công thức: $SE = \sqrt{\frac{p(1-p)}{n}}$.⁵

5.1.4 Áp dụng Định lý Giới Hạn Trung Tâm trong thực tế

Chúng ta thực sự không biết tỷ lệ dân số trừ khi tiến hành một cuộc khảo sát toàn bộ cá thể vì điều này rất tốn kém. Trước đó, chúng ta dùng giá trị $p = 0,88$ dựa trên một cuộc khảo sát do Pew Research thực hiện với 1000 người trưởng thành Mỹ, trong đó tìm thấy: $\hat{p} = 0,887$ ủng hộ việc mở rộng năng lượng mặt trời. Các nhà nghiên cứu có thể tự hỏi: Liệu tỷ lệ mẫu $\hat{p}$ từ cuộc khảo sát này có tuân theo phân phối chuẩn không? Ta sẽ kiểm tra các điều kiện của Định lý Giới Hạn Trung Tâm:

Tính độc lập. Đây là một mẫu ngẫu nhiên đơn giản từ người trưởng thành ở Mỹ, nên các quan sát độc lập.

Điều kiện thành công - thất bại. Để kiểm tra điều kiện này, chúng ta cần biết tỷ lệ dân số p , để xem liệu cả hai giá trị np và $n(1 - p)$ có lớn hơn 10 hay không. Tuy nhiên, trong thực tế chúng ta không biết giá trị p , và đó chính là lý do các nhà khảo sát phải thực hiện lấy mẫu! Trong những trường hợp như vậy, chúng ta thường sử dụng tỷ lệ mẫu $\hat{p}$ như một phương án thay thế hợp lý nhất để kiểm tra điều kiện thành công - thất bại:

$$np = 1000 \times 0,887 = 887$$

$$n(1 - \hat{p}) = 1000 \times (1 - 0,887) = 113$$

Tỷ lệ mẫu $\hat{p}$ đóng vai trò như một thay thế hợp lý cho p trong bước kiểm tra này, và cả hai giá trị tính được đều lớn hơn ngưỡng tối thiểu là 10.

Phép tính gần đúng thay thế này bằng cách sử dụng $\hat{p}$ cho p cũng hữu ích khi tính toán sai số chuẩn của tỷ lệ mẫu:

$$SE_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} \approx \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0,887(1-0,887)}{1000}} = 0,010$$

Kỹ thuật thay thế này đôi khi được gọi là “nguyên lý cắm vào” (plug-in principle). Trong ví dụ này, sai số chuẩn $SE_{\hat{p}}$ không thay đổi nhiều đến mức có thể nhận ra nếu chỉ làm tròn đến 3 chữ số thập phân, so với khi tính với $p = 0,88$ ở phần trước. Nói cách khác, sai số chuẩn được tính ra có xu hướng ổn định hợp lý, ngay cả khi tỷ lệ mẫu thay đổi một chút giữa các lần lấy mẫu khác nhau.

⁵ Vì kích thước mẫu n nằm ở mẫu số (tức phần dưới của phân số), nên khi kích thước mẫu lớn hơn, giá trị toàn bộ biểu thức sẽ có xu hướng nhỏ hơn. Nói cách khác, mẫu càng lớn thì sai số chuẩn càng nhỏ.

5.1 ƯỚC LƯỢNG ĐIỂM VÀ BIẾN ĐỘNG MẪU

5.1.5 Chi tiết thêm về Định lý Giới Hạn Trung Tâm

Chúng ta đã áp dụng Định lý Giới Hạn Trung Tâm trong nhiều ví dụ từ đầu chương đến giờ.

Khi các quan sát là độc lập và kích thước mẫu đủ lớn, phân phối của tỉ lệ mẫu $\hat{p}$ sẽ gần giống với phân phối chuẩn với:

$$\mu_{\hat{p}} = p$$

$$SE_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

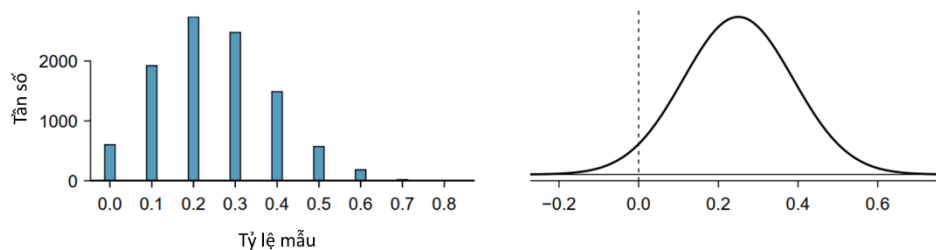
Kích thước mẫu được xem là đủ lớn khi: $np \geq 10$ và $n(1-p) \geq 10$

Trong phần này, chúng ta sẽ khám phá sâu hơn về điều kiện thành công - thất bại và tìm hiểu rõ hơn về Định lý Giới Hạn Trung Tâm.

Một câu hỏi thú vị là: *Chuyện gì xảy ra khi $np \leq 10$ và $n(1-p) \leq 10$?* Giống như ở Mục 5.1.2, chúng ta có thể mô phỏng nhiều mẫu với kích thước khác nhau, chẳng hạn khi tỉ lệ thật là $p = 0,25$. Dưới đây là một mẫu ngẫu nhiên gồm 10 phần tử:

không, không, có, có, không, không, không, không, không, không

Trong mẫu này, chúng ta ghi nhận tỉ lệ “có” là: $\hat{p} = \frac{2}{10} = 0,2$. Chúng ta có thể mô phỏng nhiều tỉ lệ như vậy để hiểu về phân phối lấy mẫu của $\hat{p}$ khi $n = 10$ và $p = 0,25$, và vẽ phân phối đó ở Hình 5.3, so sánh với một phân phối chuẩn có cùng trung bình và độ lệch chuẩn.



Hình 5.3: Trái: mô phỏng tỉ lệ mẫu $\hat{p}$ với $n = 10$, $p = 0,25$;

Phải: phân phối chuẩn với cùng trung bình (0,25) và độ lệch chuẩn (0,137)

	Một đỉnh?	Đường cong trơn?	Đối xứng?
Phân phối chuẩn: $N(0.25, 0.14)$	Có	Có	Có
$n = 10$, $p = 0,25$	Có	Không	Không

Lưu ý rằng điều kiện thành công - thất bại không được thỏa mãn khi $n = 10$ và $p = 0,25$:

$$np = 10 \times 0,25 = 2,5$$

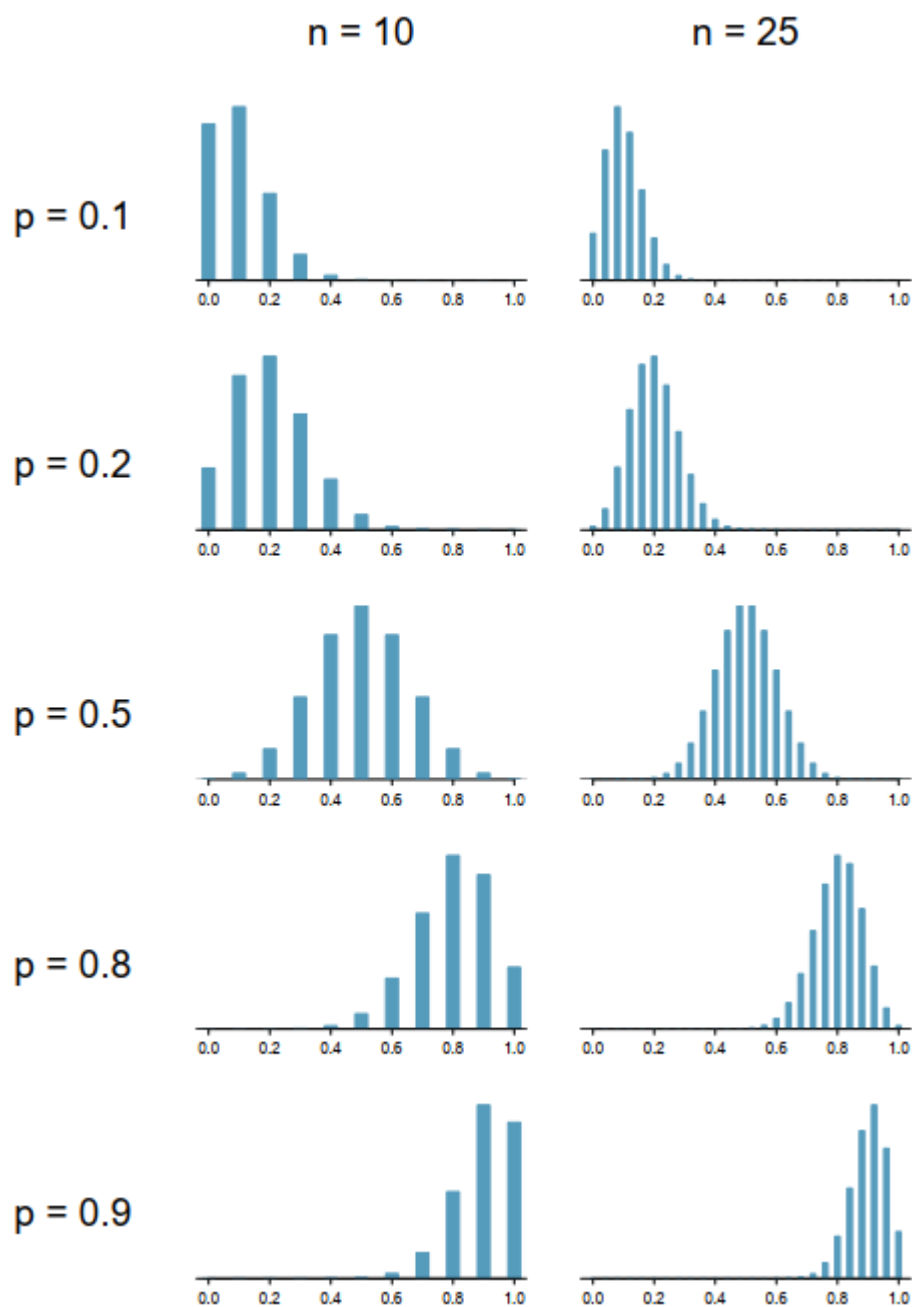
$$n(1-p) = 10 \times 0,75 = 7,5$$

Phân phối lấy mẫu đơn lẻ này không chứng minh rằng điều kiện thành công - thất bại là một nguyên tắc hoàn hảo, nhưng chúng tôi nhận thấy quy tắc này đã đúng khi cảnh báo rằng phân phối chuẩn có thể không phù hợp.

Chúng ta có thể thực hiện thêm một số mô phỏng nữa (được trình bày trong Hình 5.4 và 5.5 và có thể nhận ra một số xu hướng:

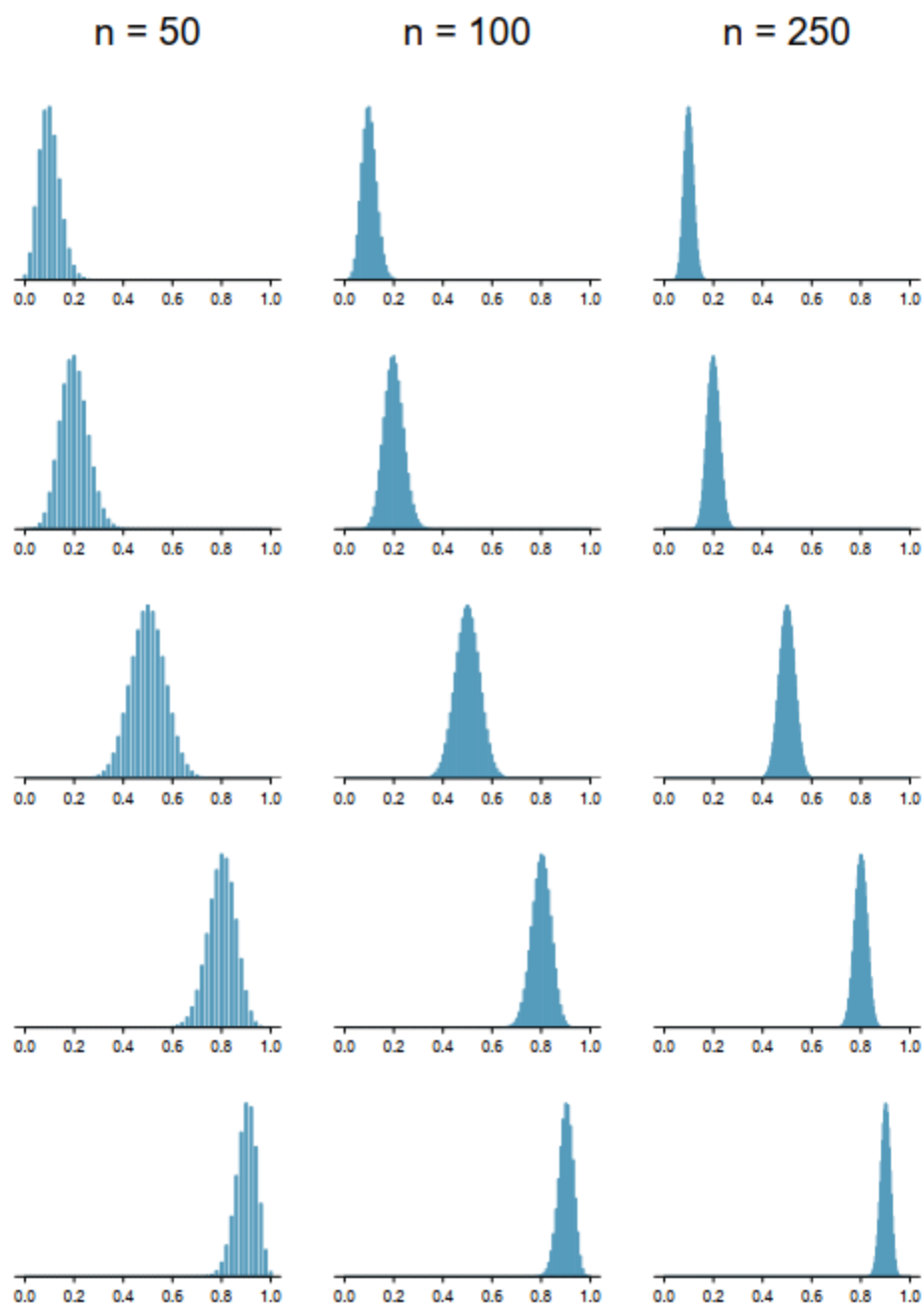
1. Khi np hoặc $n(1-p)$ nhỏ, phân phối sẽ trở nên **rời rạc** hơn, tức là *không liên tục*.
2. Khi np hoặc $n(1-p) < 10$, sự lệch (skewness) trong phân phối trở nên rõ rệt hơn.
3. Càng lớn cả hai giá trị np và $n(1-p)$, phân phối càng giống với phân phối chuẩn. Điều này có thể khó nhận ra hơn ở các mẫu lớn vì phương sai sẽ nhỏ hơn nhiều.
4. Khi np và $n(1-p)$ đều rất lớn, sự rời rạc gần như không còn thấy rõ, và phân phối trông rất giống phân phối chuẩn.

5.1 ƯỚC LƯỢNG ĐIỂM VÀ BIẾN ĐỘNG MẪU



Hình 5.4: Các phân phối mẫu trong nhiều kịch bản khác nhau của p và n .
 Các hàng: $p = 0,10$, $p = 0,20$, $p = 0,50$, $p = 0,80$, and $p = 0,90$.
 Các cột: $n = 10$ and $n = 25$.

5.1 ƯỚC LƯỢNG ĐIỂM VÀ BIẾN ĐỘNG MẪU



Hình 5.5: Các phân phối mẫu trong nhiều kịch bản khác nhau của p và n .
 Các hàng: $p = 0,10$, $p = 0,20$, $p = 0,50$, $p = 0,80$, and $p = 0,90$.
 Các cột: $n = 50$, $n = 100$, and $n = 250$.

5.1 ƯỚC LƯỢNG ĐIỂM VÀ BIẾN ĐỘNG MẪU

Cho đến giờ, chúng ta mới chỉ tập trung vào độ lệch và tính rời rạc của các phân phối. Chúng ta vẫn chưa xem xét cách mà trung bình và sai số chuẩn của các phân phối thay đổi. Hãy dành một chút thời gian để xem lại các biểu đồ, và chú ý đến ba điều sau:

1. Tâm của phân phối luôn nằm tại tỷ lệ quần thể p , được sử dụng để tạo ra mô phỏng. Bởi vì phân phối lấy mẫu của $\hat{p}$ luôn tập trung tại tham số dân số p , điều này có nghĩa rằng tỷ lệ mẫu $\hat{p}$ là **không thiên lệch** khi dữ liệu là độc lập và được rút ra từ một quần thể như vậy.
2. Với một tỷ lệ quần thể p nhất định, độ biến thiên trong phân phối lấy mẫu giảm khi kích thước mẫu n trở nên lớn hơn. Điều này có thể phù hợp với trực giác của bạn: một ước lượng dựa trên cỡ mẫu lớn hơn sẽ có xu hướng chính xác hơn.
3. Với một kích thước mẫu nhất định, độ biến thiên sẽ lớn nhất khi $p = 0,5$. Sự khác biệt có thể hơi tinh tế, vì vậy hãy quan sát kỹ. Điều này phản ánh vai trò của tỷ lệ p trong công thức sai số chuẩn: $SE = \sqrt{\frac{p(1-p)}{n}}$. Sai số chuẩn sẽ lớn nhất khi $p = 0,5$.

Tại bất kỳ thời điểm nào, phân phối của $\hat{p}$ cũng sẽ *không hoàn toàn giống một phân phối chuẩn*, bởi vì $\hat{p}$ luôn nhận giá trị rời rạc (x/n). Đây luôn là vấn đề về mức độ, và chúng ta sẽ sử dụng điều kiện thành công - thất bại tiêu chuẩn với giá trị tối thiểu là 10 cho cả np và $n(1 - p)$ như là một hướng dẫn trong cuốn sách này.

5.1.6 Mở rộng khung phân tích sang các thống kê khác

Chiến lược sử dụng một thống kê mẫu để ước lượng một tham số là khá phổ biến, và đó là một chiến lược mà chúng ta có thể áp dụng cho các thống kê khác ngoài một tỷ lệ. Ví dụ, nếu chúng ta muốn ước lượng mức lương trung bình cho sinh viên tốt nghiệp từ một trường đại học cụ thể, chúng ta có thể khảo sát một mẫu ngẫu nhiên các sinh viên mới tốt nghiệp; trong ví dụ đó, chúng ta sẽ sử dụng trung bình mẫu $\bar{x}$ để ước lượng trung bình dân số μ cho tất cả sinh viên tốt nghiệp. Một ví dụ khác, nếu chúng ta muốn ước lượng sự khác biệt về giá sản phẩm giữa hai trang web, chúng ta có thể chọn một mẫu ngẫu nhiên các sản phẩm có mặt ở cả hai trang, kiểm tra giá của từng sản phẩm, và sau đó tính trung bình của các mức chênh lệch; chiến lược này chắc chắn sẽ cung cấp cho chúng ta một ý tưởng về sự khác biệt thực tế thông qua một ước lượng điểm.

Mặc dù chương này nhấn mạnh vào bối cảnh một tỷ lệ đơn, chúng ta sẽ gặp nhiều bối cảnh khác nhau xuyên suốt cuốn sách này, nơi các phương pháp này sẽ được áp dụng. Các nguyên lý và tư tưởng chung là như nhau, dù các chi tiết có thể thay đổi một chút. Chúng tôi cũng đã rải rác đưa vào một số bối cảnh khác trong phần bài tập để giúp bạn bắt đầu suy nghĩ về cách tổng quát hóa các ý tưởng này.

5.1 ƯỚC LƯỢNG ĐIỂM VÀ BIẾN ĐỘNG MẪU

Bài tập

5.1 Xác định tham số, phần I. Với mỗi tình huống dưới đây, hãy cho biết tham số quan tâm là trung bình hay tỷ lệ. Có thể hữu ích nếu bạn xem xét liệu phản hồi cá nhân là dữ liệu số hay danh mục.

- Trong một khảo sát, một trăm sinh viên đại học được hỏi họ dành bao nhiêu giờ mỗi tuần trên Internet.
- Trong một khảo sát, một trăm sinh viên đại học được hỏi: “Bao nhiêu phần trăm thời gian bạn sử dụng Internet là cho việc học?”
- Trong một khảo sát, một trăm sinh viên đại học được hỏi liệu họ có trích dẫn thông tin từ Wikipedia trong bài viết của họ hay không.
- Trong một khảo sát, một trăm sinh viên đại học được hỏi bao nhiêu phần trăm tổng chi tiêu hàng tuần của họ là cho đồ uống có cồn.
- Trong một mẫu gồm một trăm sinh viên mới tốt nghiệp, phát hiện rằng 85% mong đợi có được một công việc trong vòng một năm kể từ ngày tốt nghiệp.

5.2 Xác định tham số, phần II. Với mỗi tình huống sau, hãy cho biết tham số quan tâm là trung bình hay tỷ lệ.

- Một cuộc thăm dò cho thấy 64% người Mỹ lo lắng rất nhiều về chi tiêu liên bang và thâm hụt ngân sách.
- Một khảo sát báo cáo rằng doanh thu từ tin tức truyền hình địa phương đã tăng 17% trong vòng hai năm, trong khi doanh thu từ báo in giảm 6,4% trong khoảng thời gian này.
- Trong một khảo sát, học sinh trung học và sinh viên đại học được hỏi liệu họ có sử dụng dịch vụ định vị trên điện thoại thông minh hay không.
- Trong một khảo sát, người dùng điện thoại thông minh được hỏi liệu họ có sử dụng dịch vụ taxi dựa trên web hay không.
- Trong một khảo sát, người dùng điện thoại thông minh được hỏi họ đã sử dụng dịch vụ taxi dựa trên web bao nhiêu lần trong năm qua.

5.3 Kiểm soát chất lượng. Là một phần của quy trình kiểm soát chất lượng cho chip máy tính, một kỹ sư tại một nhà máy lấy ngẫu nhiên 212 con chip trong một tuần sản xuất để kiểm tra tỷ lệ hiện tại của các chip có lỗi nghiêm trọng. Cô phát hiện rằng 27 trong số các chip bị lỗi.

- Quần thể đang được xem xét trong bộ dữ liệu này là gì?
- Tham số đang được ước lượng là gì?
- Ước lượng điểm cho tham số là gì?
- Tên của thống kê mà chúng ta sử dụng để đo lường mức độ không chắc chắn của ước lượng điểm là gì?
- Tính giá trị ở phần (d) cho ngữ cảnh này.
- Tỷ lệ lỗi trong lịch sử là 10%. Kỹ sư có nên ngạc nhiên với tỷ lệ lỗi quan sát được trong tuần hiện tại không?
- Giả sử giá trị thật của quần thể được tìm thấy là 10%. Nếu chúng ta sử dụng tỷ lệ này để tính lại giá trị trong phần (e) bằng $p = 0,1$ thay vì $\hat{p}$, thì kết quả có thay đổi nhiều không?

5.4 Chi phí bất ngờ. Trong một mẫu ngẫu nhiên gồm 765 người trưởng thành ở Hoa Kỳ, có 322 người nói rằng họ không thể trang trải một khoản chi phí bất ngờ 400 đô la mà không vay mượn hoặc mắc nợ.

- Quần thể đang được xem xét trong bộ dữ liệu này là gì?
- Tham số đang được ước lượng là gì?
- Ước lượng điểm cho tham số là gì?
- Tên của thống kê mà chúng ta sử dụng để đo lường mức độ không chắc chắn của ước lượng điểm là gì?
- Tính giá trị ở phần (d) cho ngữ cảnh này.
- Một người dẫn chương trình trên kênh tin tức cáo ngửi rằng giá trị thực tế là 50%. Liệu cô ấy có nên ngạc nhiên với dữ liệu này không?
- Giả sử giá trị thực của quần thể được tìm thấy là 40%. Nếu chúng ta sử dụng tỷ lệ này để tính lại giá trị trong phần (e) bằng $p = 0,4$ thay vì $\hat{p}$, thì kết quả có thay đổi nhiều không?

5.1 ƯỚC LƯỢNG ĐIỂM VÀ BIẾN ĐỘNG MẪU

5.5 Lặp lại lấy mẫu nước. Một tổ chức phi lợi nhuận muốn tìm hiểu tỷ lệ hộ gia đình có mức chì cao trong nước uống. Họ dự đoán ít nhất 5% các hộ sẽ có mức chì cao, nhưng không quá khoảng 30%. Họ lấy ngẫu nhiên mẫu từ 800 hộ gia đình và phối hợp với chủ hộ để lấy mẫu nước, sau đó họ tính tỷ lệ các hộ có mức chì cao. Họ lặp lại quy trình này 1000 lần và xây dựng một phân phối của các tỷ lệ mẫu.

- Phân phối này được gọi là gì?
- Bạn có mong đợi hình dạng của phân phối này đối xứng, lệch phải hay lệch trái không? Giải thích lý do của bạn.
- Nếu các tỷ lệ được phân phối xung quanh mức 8%, độ biến thiên của phân phối là bao nhiêu?
- Tên chính thức của giá trị bạn tính được ở phần (c) là gì?
- Giả sử ngân sách của nhóm nghiên cứu bị cắt giảm, và họ chỉ có thể thu thập 250 quan sát cho mỗi mẫu, nhưng vẫn thu thập được 1000 mẫu. Họ xây dựng một phân phối mới của các tỷ lệ mẫu. Độ biến thiên của phân phối mới này sẽ so với phân phối ban đầu (mỗi mẫu gồm 800 quan sát) như thế nào?

5.6 Lặp lại việc lấy mẫu sinh viên. Trong số tất cả sinh viên năm nhất tại một trường đại học lớn, có 16% được ghi danh vào danh sách danh dự (Dean's List) trong năm nay. Như một phần của dự án lớp học, sinh viên lấy ngẫu nhiên mẫu gồm 40 sinh viên và kiểm tra xem họ có nằm trong danh sách danh dự không. Họ lặp lại điều này 1000 lần và xây dựng một phân phối của các tỷ lệ mẫu.

- Phân phối này được gọi là gì?
- Bạn có mong đợi hình dạng của phân phối này đối xứng, lệch phải hay lệch trái không? Giải thích lý do của bạn.
- Tính độ biến thiên của phân phối này.
- Tên chính thức của giá trị bạn tính được ở phần (c) là gì?
- Giả sử các sinh viên quyết định lấy mẫu lại, lần này thu thập 90 sinh viên cho mỗi mẫu, và tiếp tục thu thập 1000 mẫu. Họ xây dựng một phân phối mới của các tỷ lệ mẫu. Độ biến thiên của phân phối mới này sẽ so với phân phối ban đầu (mỗi mẫu gồm 40 sinh viên) như thế nào??

5.2 KHOẢNG TIN CẬY CHO MỘT TỶ LỆ

5.2 Khoảng tin cậy cho một tỷ lệ

Tỷ lệ mẫu $\hat{p}$ cung cấp một giá trị hợp lý duy nhất cho tỷ lệ quần thể p . Tuy nhiên, tỷ lệ mẫu không hoàn hảo và sẽ có *sai số chuẩn* liên quan. Khi đưa ra ước lượng cho tỷ lệ quần thể, cách tốt hơn là cung cấp một *khoảng giá trị* hợp lý thay vì chỉ cung cấp ước lượng điểm.

5.2.1 Bao quát tham số quần thể

Việc chỉ sử dụng một ước lượng điểm giống như câu cá bằng lao trong một hồ nước đục. Ta có thể phóng lao vào nơi ta thấy cá, nhưng có khả năng sẽ trượt. Ngược lại, nếu ta thả lưới ở khu vực đó, ta có cơ hội cao hơn bắt được cá. **Khoảng tin cậy** giống như việc thả lưới, nó đại diện cho một khoảng giá trị hợp lý mà ta có khả năng tìm thấy tham số quần thể.

Nếu ta đưa ra một ước lượng điểm $\hat{p}$, ta có thể sẽ không xác định chính xác tỷ lệ quần thể. Mặt khác, Tuy nhiên, nếu đưa ra một khoảng giá trị hợp lý, tức một khoảng tin cậy, ta có khả năng cao hơn bao trùm được tham số thật.



BÀI TẬP THỰC HÀNH 5.6

Nếu muốn chắc chắn đã bao trùm được tỷ lệ quần thể trong khoảng, ta nên dùng khoảng tin cậy rộng hơn hay hẹp hơn? ⁶

5.2.2 Xây dựng khoảng tin cậy 95%

Tỷ lệ mẫu $\hat{p}$ là giá trị hợp lý nhất cho tỷ lệ quần thể, vì vậy việc xây dựng khoảng tin cậy xung quanh nó là hợp lý. Sai số chuẩn cho ta biết khoảng tin cậy nên rộng bao nhiêu.

Sai số chuẩn đại diện cho độ lệch chuẩn của ước lượng điểm, và khi các điều kiện của Định lý Giới Hạn Trung Tâm được thoả mãn, ước lượng điểm gần như tuân theo phân phối chuẩn. Trong phân phối chuẩn, 95% dữ liệu nằm trong vòng 1,96 độ lệch tiêu chuẩn của trung bình. Dựa trên nguyên tắc này, chúng ta có thể xây dựng một khoảng tin cậy mở rộng 1,96 lỗi chuẩn từ tỷ lệ mẫu để có **độ tin cậy 95%** rằng khoảng này nắm bắt được tỷ lệ dân số:

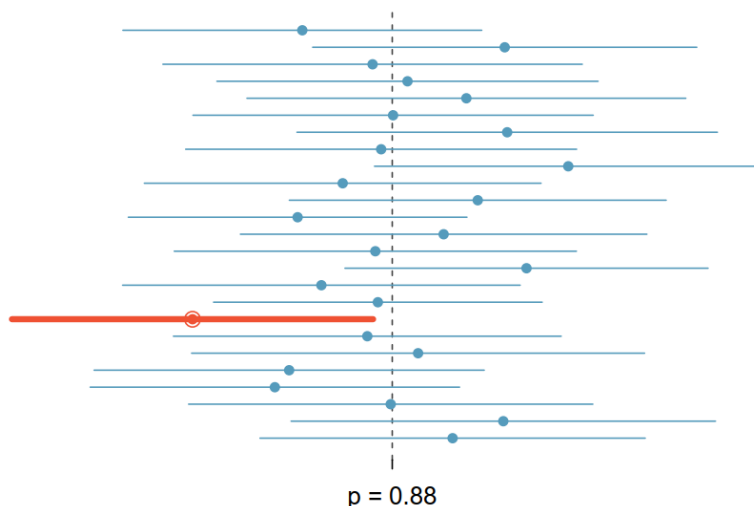
$$\text{ước lượng điểm} \pm 1,96 \times SE$$

$$\hat{p} \pm 1,96 \times \sqrt{\frac{p(1-p)}{n}}$$

Nhưng “95% tin cậy” nghĩa là gì? Giả sử ta lấy nhiều mẫu và xây dựng khoảng tin cậy 95% cho mỗi mẫu, thì khoảng 95% trong số đó sẽ chứa tham số quần thể p . Hình 5.6 minh họa quá trình tạo ra 25 khoảng tin cậy từ 25 mẫu trong mô phỏng được trình bày ở Mục 5.1.2, trong đó 24 khoảng tin cậy kết quả bao gồm tỷ lệ quần thể của mô phỏng là $p = 0,88$, và chỉ có một khoảng không bao gồm giá trị này.

⁶ Nếu chúng ta muốn chắc chắn hơn rằng mình sẽ bắt được cá, chúng ta có thể dùng một chiếc lưới rộng hơn. Tương tự như vậy, chúng ta sử dụng một khoảng tin cậy rộng hơn nếu muốn chắc chắn hơn rằng mình đã bao trùm được tham số.

5.2 KHOẢNG TIN CẬY CHO MỘT TỶ LỆ



Hình 5.6: Hai mươi lăm ước lượng điểm và khoảng tin cậy từ các mô phỏng trong Mục 5.1.2. Các khoảng tin cậy này được biểu diễn tương quan với tỷ lệ quần thể $p = 0,88$. Chỉ có 1 trong số 25 khoảng tin cậy này không bao trùm tỷ lệ quần thể, và khoảng đó đã được tô đậm để đánh dấu.

VÍ DỤ 5.7

Trong hình 5.6 có một khoảng không chứa $p = 0,88$. Điều này có mang ý nghĩa rằng tỷ lệ quần thể được sử dụng trong mô phỏng không thể là $p = 0,88$?

(E)

Cũng giống như một số quan sát có thể xảy ra ngoài 1,96 độ lệch chuẩn so với trung bình, một số ước lượng điểm cũng có thể lệch hơn 1,96 sai số chuẩn so với tham số thực sự. Khoảng tin cậy chỉ cung cấp một khoảng giá trị hợp lý - mặc dù ta có thể cho rằng các giá trị khác là không hợp lý dựa trên dữ liệu, nhưng điều đó không có nghĩa là chúng không thể xảy ra.

KHOẢNG TIN CẬY 95% CHO MỘT THAM SỐ

Khi phân phối của một ước lượng điểm thỏa mãn các điều kiện của Định lý Giới Hạn Trung Tâm và do đó xấp xỉ phân phối chuẩn, ta có thể xây dựng khoảng tin cậy 95% như sau:

$$\text{ước lượng điểm} \pm 1,96 \times SE$$

VÍ DỤ 5.8

Trong Mục 5.1, chúng ta đã tìm hiểu về một cuộc khảo sát của Pew Research, trong đó 88,7% từ một mẫu ngẫu nhiên gồm 1000 người trưởng thành ở Hoa Kỳ ủng hộ việc mở rộng vai trò của năng lượng mặt trời. Hãy tính và diễn giải khoảng tin cậy 95% cho tỷ lệ quần thể.

(E)

Trước đó, chúng ta đã xác nhận rằng $\hat{p}$ xấp xỉ phân phối chuẩn và có sai số chuẩn $SE_{\hat{p}} = 0,010$. Để tính khoảng tin cậy 95%, hãy thay thế ước lượng điểm $\hat{p} = 0,887$ và sai số chuẩn vào công thức khoảng tin cậy 95%:

$$\hat{p} \pm 1,96 \times SE_{\hat{p}} \rightarrow 0,887 \pm 1,96 \times 0,010 \rightarrow (0,8674 - 0,9066)$$

Chúng tôi tin tưởng 95% rằng tỷ lệ thực sự của người trưởng thành ở Hoa Kỳ ủng hộ việc mở rộng năng lượng mặt trời nằm trong khoảng từ 86,7% đến 90,7%. (Thông thường, khi báo cáo khoảng tin cậy, người ta làm tròn đến phần trăm gần nhất hoặc đến một chữ số thập phân gần nhất.)

5.2 KHOẢNG TIN CẬY CHO MỘT TỶ LỆ

5.2.3 Thay đổi mức độ tin cậy

Giả sử chúng ta muốn xem xét các khoảng tin cậy với mức độ tin cậy cao hơn 95%, ví dụ như 99%. Hãy nhớ lại phép ẩn dụ về việc bắt cá: nếu chúng ta muốn chắc chắn hơn rằng mình sẽ bắt được cá, ta nên sử dụng một chiếc lưới rộng hơn. Tương tự như vậy, để xây dựng một khoảng tin cậy 99%, chúng ta cũng cần mở rộng khoảng tin cậy 95%. Ngược lại, nếu chúng ta muốn một khoảng có mức độ tin cậy thấp hơn, chẳng hạn như 90%, thì ta có thể sử dụng một khoảng hẹp hơn một chút so với khoảng 95% ban đầu.

Cấu trúc của khoảng tin cậy 95% cung cấp một hướng dẫn chung để tạo ra các khoảng với các mức độ tin cậy khác nhau. Khoảng tin cậy 95% tổng quát cho một ước lượng điểm tuân theo phân phối chuẩn được viết như sau

$$\text{ước lượng điểm} \pm 1,96 \times SE$$

Khoảng tin cậy này bao gồm ba thành phần: ước lượng điểm, “1,96”, và sai số chuẩn. Việc lựa chọn hệ số $1,96 \times SE$ dựa trên nguyên lý rằng: ước lượng sẽ nằm trong phạm vi $\pm 1,96$ sai số chuẩn so với tham số thật trong khoảng 95% số lần lấy mẫu. Hệ số 1,96 tương ứng với mức độ tin cậy 95%.

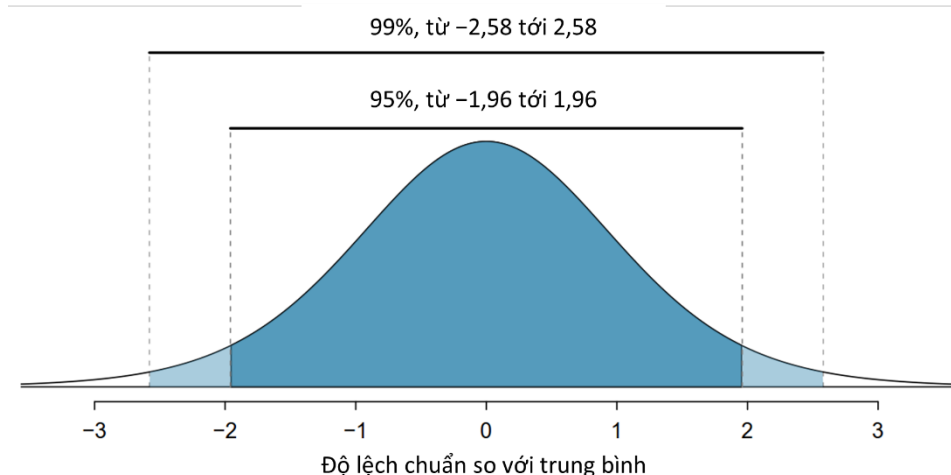
BÀI TẬP THỰC HÀNH 5.9



Nếu X là một biến ngẫu nhiên có phân phối chuẩn, xác suất để giá trị của X nằm trong 2,58 độ lệch chuẩn tính từ trung bình là bao nhiêu? ⁷

Thực hành có hướng dẫn 5.9 nhấn mạnh rằng 99% thời gian, một biến ngẫu nhiên có phân phối chuẩn sẽ nằm trong khoảng $\pm 2,58$ độ lệch chuẩn so với trung bình. Vì vậy, để xây dựng một khoảng tin cậy 99%, ta thay thế hệ số 1,96 trong công thức khoảng tin cậy 95% bằng 2,58. Nói cách khác, công thức cho khoảng tin cậy 99% là

$$\text{ước lượng điểm} \pm 2,58 \times SE$$



Hình 5.7: Diện tích nằm giữa $-z^*$ và z^* tăng lên khi z^* lớn hơn. Nếu mức độ tin cậy là 99%, ta chọn z^* sao cho 99% diện tích của phân phối chuẩn nằm giữa $-z^*$ và z^* , tương ứng với 0,5% ở phần đuôi dưới và 0,5% ở phần đuôi trên $z^* = 2,58$.

⁷Điều này tương đương với việc hỏi tần suất điểm Z sẽ lớn hơn $-2,58$ nhưng nhỏ hơn $2,58$. Để xem hình minh họa, hãy xem Hình 5.7. Để xác định xác suất này, chúng ta có thể sử dụng phần mềm thống kê, máy tính bỏ túi hoặc bảng để tra cứu $-2,58$ và $2,58$ cho phân phối chuẩn: 0,0049 và 0,9951. Do đó, có xác suất $0,9951 - 0,0049 \approx 0,99$ rằng một biến ngẫu nhiên chuẩn không quan sát được X sẽ nằm trong khoảng 2,58 độ lệch chuẩn của μ .

5.2 KHOẢNG TIN CẬY CHO MỘT TỶ LỆ

Phương pháp sử dụng giá trị Z-score trong mô hình phân phối chuẩn để tính mức độ tin cậy là phù hợp khi ước lượng điểm, chẳng hạn như khi $\hat{p}$ tuân theo phân phối chuẩn. Đối với một số ước lượng điểm khác, mô hình phân phối chuẩn không phải là lựa chọn phù hợp; trong những trường hợp như vậy, chúng ta sẽ sử dụng các phân phối thay thế để biểu diễn chính xác hơn phân phối mẫu.

KHOẢNG TIN CẬY VỚI BẤT KỲ MỨC ĐỘ TIN CẬY NÀO

Nếu một ước lượng điểm xấp xỉ phân phối chuẩn và có sai số chuẩn SE , thì khoảng tin cậy cho tham số quần thể được tính theo công thức:

$$\text{ước lượng điểm} \pm z \times SE$$

trong đó z^* tương ứng với mức độ tin cậy đã chọn.

Hình 5.7 cung cấp một hình minh họa về cách xác định z^* dựa trên mức độ tin cậy. Ta chọn z^* sao cho diện tích giữa $-z^*$ và z^* trong phân phối chuẩn hóa, $N(0, 1)$, tương ứng với mức độ tin cậy đã chọn.

BIÊN ĐỘ SAI SỐ

Trong một khoảng tin cậy, biểu thức $z \times SE$ được gọi là **biên độ sai số**.

VÍ DỤ 5.10

Sử dụng dữ liệu trong Ví dụ 5.8 hãy tạo một khoảng tin cậy 90% cho tỷ lệ người trưởng thành ở Hoa Kỳ ủng hộ việc mở rộng sử dụng năng lượng mặt trời. Chúng ta đã xác minh trước đó rằng các điều kiện để sử dụng phân phối chuẩn là thỏa mãn.

E

Đầu tiên, ta xác định z^* sao cho 90% phân phối chuẩn hóa $N(\mu = 0, \sigma = 1)$ nằm giữa $-z^*$ và z^* . Chúng ta có thể thực hiện điều này bằng cách sử dụng máy tính đồ thị, phần mềm thống kê, hoặc bảng xác suất, bằng cách tìm giá trị tương ứng với phần đuôi trên là 5% (và 5% còn lại nằm ở phần đuôi dưới): $z^* = 1,65$. Khi đó, khoảng tin cậy 90% có thể được tính như sau:

$$\hat{p} \pm 1,6449 \times SE_{\hat{p}} \rightarrow 0,887 \pm 1,65 \times 0,0100 \rightarrow (0,8705 - 0,9034)$$

Điều đó có nghĩa là, chúng tôi tin tưởng 90% rằng từ 87,1% đến 90,3% người trưởng thành ở Hoa Kỳ đã ủng hộ việc mở rộng năng lượng mặt trời vào năm 2018.

KHOẢNG TIN CẬY CHO MỘT TỶ LỆ DUY NHẤT

Khi đã xác định rằng khoảng tin cậy cho một tỷ lệ là phù hợp cho một ứng dụng cụ thể, việc xây dựng khoảng tin cậy này sẽ bao gồm 4 bước sau:

Chuẩn bị. Xác định $\hat{p}$ và n , và quyết định mức độ tin cậy bạn muốn sử dụng.

Kiểm tra. Xác minh các điều kiện để đảm bảo rằng $\hat{p}$ gần đạt phân phối chuẩn. Đối với các khoảng tin cậy một tỷ lệ, sử dụng $\hat{p}$ thay cho p để kiểm tra điều kiện thành công – thất bại.

Tính toán. Nếu các điều kiện được thỏa mãn, tính sai số chuẩn SE sử dụng $\hat{p}$, tìm z^* , xây dựng khoảng tin cậy.

Kết luận. Diễn giải khoảng tin cậy trong bối cảnh của bài toán.

5.2 KHOẢNG TIN CÂY CHO MỘT TỶ LỆ

5.2.4 Một số nghiên cứu tình huống khác

Ngày 23 tháng 10 năm 2014, ở New York, một bác sĩ vừa trở về sau khi điều trị cho các bệnh nhân Ebola ở Guinea đã nhập viện vì sốt nhẹ và sau đó được chẩn đoán mắc bệnh Ebola. Ngay sau đó, một cuộc khảo sát do NBC 4 New York/The Wall Street Journal/Marist Poll thực hiện cho thấy 82% người dân New York ủng hộ việc “cách ly bắt buộc 21 ngày đối với bất kỳ ai từng tiếp xúc với bệnh nhân Ebola”. Khảo sát này thu thập phản hồi từ 1,042 người trưởng thành tại New York, diễn ra trong khoảng thời gian từ ngày 26 đến ngày 28 tháng 10 năm 2014.

VÍ DỤ 5.11

Ước lượng điểm trong trường hợp này là gì, và có hợp lý không nếu sử dụng phân phối chuẩn để mô hình hóa ước lượng đó?

- (E) Ước lượng điểm, dựa trên cỡ mẫu $n = 1042$, là $\hat{p} = 0,82$. Để kiểm tra xem liệu $\hat{p}$ có thể được mô hình hóa hợp lý bằng phân phối chuẩn hay không, ta cần kiểm tra tính độc lập (khảo sát dựa trên mẫu ngẫu nhiên đơn giản) và điều kiện thành công – thất bại ($1042 \times \hat{p} \approx 854$ và $1042 \times (1 - \hat{p}) \approx 188$, cả 2 đều lớn hơn 10). Với các điều kiện đã được thỏa mãn, Vì vậy, chúng ta có thể khẳng định rằng phân phối chọn mẫu của $\hat{p}$ hoàn toàn có thể được mô hình hóa bằng phân phối chuẩn một cách hợp lý.

VÍ DỤ 5.12

Ước lượng sai số chuẩn của $\hat{p} = 0,82$ từ khảo sát về Ebola.

- (E) Chúng ta sẽ sử dụng phương pháp xấp xỉ thay thế với $p \approx \hat{p} = 0,82$ để tính sai số chuẩn như sau:

$$SE_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} \approx \sqrt{\frac{0,82(1-0,82)}{1042}} = 0,012$$

VÍ DỤ 5.13

Xây dựng khoảng tin cậy 95% cho p , tỷ lệ người trưởng thành ở New York ủng hộ việc cách ly bất kỳ ai đã tiếp xúc với bệnh nhân Ebola.

- (E) Sử dụng sai số chuẩn $SE = 0,012$ từ Ví dụ 5.12, ước lượng điểm 0,82, và $z^* = 1,96$ tương ứng với mức độ tin cậy 95%, khoảng tin cậy được tính như sau:

$$\text{ước lượng điểm} \pm z^* \times SE \rightarrow 0,82 \pm 1,96 \times 0,012 \rightarrow (0,796 - 0,844)$$

Chúng tôi tin tưởng 95% rằng tỷ lệ người trưởng thành ở New York vào tháng 10 năm 2014 ủng hộ việc cách ly bất kỳ ai đã tiếp xúc với một bệnh nhân Ebola nằm trong khoảng từ 0,796 đến 0,844.

BÀI TẬP THỰC HÀNH 5.14

Trả lời hai câu hỏi sau liên quan đến khoảng tin cậy trong Ví dụ 5.13:⁸

- (G) a) 95% tin cậy có ý nghĩa gì trong ngữ cảnh này?
b) Bạn có nghĩ rằng khoảng tin cậy này vẫn còn hợp lệ đối với quan điểm của người dân New York ngày nay không?

⁸ (a) Nếu chúng ta thực hiện nhiều mẫu khảo sát như vậy và tính toán khoảng tin cậy 95% cho mỗi mẫu, thì khoảng 95% trong số các khoảng tin cậy đó sẽ bao gồm tỷ lệ thực sự người trưởng thành ở New York ủng hộ việc cách ly bất kỳ ai đã tiếp xúc với một bệnh nhân Ebola.

(b) Không nhất thiết. Cuộc khảo sát được thực hiện vào thời điểm có mối lo ngại lớn về an toàn cộng đồng. Giờ đây, khi mọi người đã có thời gian để nhìn nhận lại, họ có thể đã thay đổi quan điểm của mình. Nếu chúng ta muốn ước lượng tỷ lệ hiện tại của người trưởng thành ở New York ủng hộ một khoảng thời gian cách ly như vậy, chúng ta cần thực hiện một cuộc khảo sát mới.

5.2 KHOẢNG TIN CÂY CHO MỘT TỶ LỆ

BÀI TẬP THỰC HÀNH 5.15



Trong cuộc khảo sát của Pew Research về năng lượng mặt trời, họ cũng hỏi về các hình thức năng lượng khác, và có 84,8% trong số 1000 người tham gia khảo sát ủng hộ việc mở rộng sử dụng tuabin gió.⁹

- Có hợp lý không khi mô hình hóa tỷ lệ người Mỹ ủng hộ mở rộng tuabin gió bằng phân phối chuẩn?
- Tạo khoảng tin cậy 99% cho mức độ ủng hộ mở rộng sử dụng tuabin gió để phát điện.

Chúng ta cũng có thể xây dựng khoảng tin cậy cho các tham số khác, chẳng hạn như trung bình của quần thể. Trong những trường hợp này, khoảng tin cậy được tính theo cách tương tự như đối với một tỷ lệ duy nhất: ước lượng điểm cộng/trừ biên độ sai số. Chúng ta sẽ đi sâu vào chi tiết về những nội dung này trong các chương tiếp theo.

5.2.5 Diễn giải khoảng tin cậy

Trong mỗi ví dụ, chúng ta đã mô tả khoảng tin cậy bằng cách đặt nó vào ngữ cảnh dữ liệu cụ thể, đồng thời sử dụng ngôn ngữ mang tính chính xác và học thuật:

Năng lượng mặt trời. Chúng tôi tin tưởng 90% rằng 87,1% đến 90,4% người trưởng thành ở Hoa Kỳ ủng hộ việc mở rộng năng lượng mặt trời vào năm 2018.

Ebola. Chúng tôi tin tưởng 95% rằng tỷ lệ người trưởng thành ở New York vào tháng 10 năm 2014 ủng hộ việc cách ly bất kỳ ai từng tiếp xúc với bệnh nhân Ebola nằm trong khoảng từ 0,796 đến 0,844.

Tuabin gió. Chúng tôi tin tưởng 99% rằng tỷ lệ người trưởng thành ở Hoa Kỳ ủng hộ việc mở rộng sử dụng tuabin gió để phát điện trong năm 2018 nằm trong khoảng từ 81,9% đến 87,7%.

Thứ nhất, hãy chú ý rằng các phát biểu trên luôn nói về tham số quần thể – tức là *tất cả* người trưởng thành ở Hoa Kỳ (trong khảo sát năng lượng) hoặc *tất cả* người trưởng thành ở New York (trong khảo sát cách ly Ebola).

Chúng ta cũng tránh một sai lầm phổ biến khác: đó là *diễn đạt sai* khi nói rằng khoảng tin cậy bao trùm tham số quần thể với một xác suất nào đó. Việc diễn giải theo xác suất là một lỗi thường gặp: Mặc dù có thể hữu ích về mặt trực giác khi nghĩ rằng đây là một xác suất, nhưng thực tế, mức độ tin cậy chỉ đo lường mức độ hợp lý rằng tham số quần thể nằm trong khoảng đã cho – chứ không phải là xác suất mà tham số nằm trong khoảng đó.

Một điểm quan trọng khác cần lưu ý về khoảng tin cậy là: chúng *chỉ áp dụng cho tham số quần thể*. Khoảng tin cậy không đưa ra bất kỳ thông tin nào về các quan sát cá nhân hay về các ước lượng điểm cụ thể. Chúng chỉ cung cấp một khoảng giá trị hợp lý mà trong đó tham số quần thể có thể nằm.

Cuối cùng, hãy nhớ rằng các phương pháp mà chúng ta đã thảo luận chỉ áp dụng cho sai số chọn mẫu, chứ không áp dụng cho thiên lệch. Nếu một bộ dữ liệu được thu thập theo cách mà nó có xu hướng ước lượng thấp hơn hoặc cao hơn một cách có hệ thống so với tham số quần thể, thì các kỹ thuật mà chúng ta đã trình bày sẽ không giải quyết được vấn đề đó. Thay vào đó, chúng ta phải dựa vào quy trình thu thập dữ liệu cẩn thận để giảm thiểu nguy cơ thiên lệch, như đã được thực hiện trong các ví dụ mà chúng ta đã xem xét – đây là một thực hành tiêu chuẩn thường được các nhà khoa học dữ liệu sử dụng để đối phó với thiên lệch.

BÀI TẬP THỰC HÀNH 5.16



Xét khoảng tin cậy 90% từ khảo sát về năng lượng mặt trời: 87,1% đến 90,4%. Nếu chúng ta thực hiện lại cuộc khảo sát, liệu chúng ta có thể nói mình tin tưởng 90% rằng tỷ lệ mới sẽ nằm trong khoảng từ 87,1% đến 90,4% không?¹⁰

⁹ (a) Cuộc khảo sát là một mẫu ngẫu nhiên và số lượng đều ≥ 10 ($1000 \times 0,848 = 848$ và $1000 \times 0,152 = 152$), nên điều kiện độc lập và điều kiện thành công – thất bại đều được thỏa mãn, và $\hat{p} = 0,848$ có thể được mô hình hóa bằng cách sử dụng phân phối chuẩn.

(b) Bài tập thực hành 5.15 đã xác nhận rằng $\hat{p}$ xấp xỉ phân phối chuẩn, nên ta có thể sử dụng công thức khoảng tin cậy:

$$\text{ước lượng điểm} \pm z^* \times SE$$

Trong trường hợp này, ước lượng điểm là $\hat{p} = 0,848$. Đối với khoảng tin cậy 99% ta có $z^* = 2,58$. Tính sai số chuẩn: $SE_{\hat{p}} = \sqrt{\frac{0,848(1-0,848)}{1000}} = 0,0114$. Cuối cùng, ta tính khoảng tin cậy như sau: $0,848 \pm 2,58 \times 0,0114 \rightarrow (0,8186 - 0,8774)$. Điều quan trọng là *luôn luôn* diễn giải khoảng tin cậy: Chúng tôi tự tin 99% rằng tỷ lệ người lớn ở Mỹ ủng hộ việc mở rộng sử dụng tua-bin gió vào năm 2018 nằm trong khoảng từ 81,9% đến 87,7%.

¹⁰ Không, một khoảng tin cậy chỉ cung cấp một khoảng giá trị hợp lý cho một tham số quần thể, chứ không áp dụng cho các ước lượng điểm trong tương lai.

5.2 KHOẢNG TIN CẬY CHO MỘT TỶ LỆ

Bài tập

5.7 Bệnh mạn tính, Phần I. Vào năm 2013, Pew Research Foundation báo cáo rằng “45% người trưởng thành ở Hoa Kỳ cho biết họ đang sống chung với một hoặc nhiều bệnh mạn tính.”¹¹ Tuy nhiên, giá trị này được rút ra từ một mẫu khảo sát, vì vậy nó có thể không phải là ước lượng hoàn hảo cho tham số quần thể liên quan. Nghiên cứu báo cáo sai số chuẩn khoảng 1,2%, và có thể sử dụng mô hình phân phối chuẩn trong trường hợp này. Hãy xây dựng một khoảng tin cậy 95% cho tỷ lệ người trưởng thành ở Hoa Kỳ sống với một hoặc nhiều bệnh mạn tính. Đồng thời diễn giải khoảng tin cậy trong bối cảnh của nghiên cứu.

5.8 Người dùng Twitter và tin tức, Phần I. Một cuộc khảo sát được tiến hành vào năm 2013 cho thấy rằng 52% người trưởng thành sử dụng Twitter ở Hoa Kỳ nhận được ít nhất một phần tin tức từ Twitter.¹² Sai số chuẩn cho ước lượng này là 2,4%, và có thể sử dụng phân phối chuẩn để mô hình hóa tỷ lệ mẫu. Hãy xây dựng một khoảng tin cậy 99% cho tỷ lệ người trưởng thành sử dụng Twitter ở Hoa Kỳ nhận được tin tức từ Twitter, và diễn giải khoảng tin cậy trong bối cảnh cụ thể.

5.9 Bệnh mạn tính, Phần II. Vào năm 2013, Pew Research Foundation báo cáo rằng “45% người trưởng thành ở Hoa Kỳ cho biết họ đang sống chung với một hoặc nhiều bệnh mạn tính”, và sai số chuẩn cho ước lượng này là 1,2%. Hãy xác định mỗi phát biểu sau là đúng hay sai. Đưa ra giải thích để làm rõ từng câu trả lời.

- Chúng ta có thể nói một cách chắc chắn rằng khoảng tin cậy từ Bài tập 5.7 bao gồm tỷ lệ thực sự người trưởng thành ở Hoa Kỳ mắc bệnh mạn tính.
- Nếu chúng ta lặp lại nghiên cứu này 1.000 lần và xây dựng một khoảng tin cậy 95% cho mỗi nghiên cứu, thì khoảng 950 trong số các khoảng tin cậy đó sẽ bao gồm tỷ lệ thực sự người trưởng thành ở Hoa Kỳ mắc bệnh mạn tính.
- Cuộc khảo sát cung cấp bằng chứng thống kê có ý nghĩa (ở mức ý nghĩa $\alpha = 0.05$) rằng tỷ lệ người trưởng thành ở Hoa Kỳ mắc bệnh mạn tính thấp hơn 50%.
- Vì sai số chuẩn là 1,2%, nên chỉ có 1,2% số người tham gia khảo sát thể hiện sự không chắc chắn về câu trả lời của họ.

5.10 Người dùng Twitter và tin tức, Phần II. Một cuộc khảo sát được tiến hành vào năm 2013 cho thấy rằng 52% người trưởng thành sử dụng Twitter ở Hoa Kỳ nhận được ít nhất một phần tin tức từ Twitter, và sai số chuẩn của ước lượng này là 2,4%. Hãy xác định mỗi phát biểu sau là đúng hay sai. Đưa ra giải thích để làm rõ từng câu trả lời.

- Dữ liệu cung cấp bằng chứng thống kê có ý nghĩa rằng hơn một nửa số người trưởng thành sử dụng Twitter ở Hoa Kỳ nhận được tin tức từ Twitter. Sử dụng mức ý nghĩa $\alpha = 0.01$. (Phần này sử dụng các khái niệm từ Mục 5.3 và sẽ được chỉnh sửa trong phiên bản sách sau.)
- Vì sai số chuẩn là 2,4%, ta có thể kết luận rằng 97,6% người trưởng thành sử dụng Twitter ở Hoa Kỳ đã được khảo sát trong nghiên cứu.
- Nếu chúng ta muốn giảm sai số chuẩn của ước lượng, chúng ta nên thu thập ít dữ liệu hơn.
- Nếu chúng ta xây dựng một khoảng tin cậy 90% cho tỷ lệ người trưởng thành sử dụng Twitter ở Hoa Kỳ nhận được tin tức từ Twitter, thì khoảng tin cậy này sẽ rộng hơn so với khoảng tin cậy 99% tương ứng.

¹¹ Pew Research Center, Washington, D.C. The Diagnosis Difference, November 26, 2013.

¹² Pew Research Center, Washington, D.C. Twitter News Consumers: Young, Mobile and Educated, November 4, 2013

5.2 KHOẢNG TIN CÂY CHO MỘT TỶ LỆ

5.11 Chờ đợi tại phòng cấp cứu, Phần I. Một quản lý bệnh viện đang hy vọng cải thiện thời gian chờ khám đã quyết định ước lượng thời gian chờ trung bình tại khoa cấp cứu của bệnh viện. Bà ấy thu thập một mẫu ngẫu nhiên đơn giản gồm 64 bệnh nhân và xác định thời gian (tính bằng phút) từ lúc họ làm thủ tục tại phòng cấp cứu cho đến khi được bác sĩ khám. Một khoảng tin cậy 95% được xây dựng từ mẫu này là (128 phút, 147 phút), dựa trên mô hình phân phối chuẩn cho trung bình. Hãy xác định các phát biểu sau là đúng hay sai, và giải thích lý do.

- a) Chúng tôi tin tưởng 95% rằng thời gian chờ trung bình của 64 bệnh nhân trong mẫu này nằm trong khoảng từ 128 đến 147 phút.
- b) Chúng tôi tin tưởng 95% rằng thời gian chờ trung bình của toàn bộ bệnh nhân tại khoa cấp cứu của bệnh viện này nằm trong khoảng từ 128 đến 147 phút.
- c) 95% của các mẫu ngẫu nhiên sẽ có trung bình mẫu nằm trong khoảng từ 128 đến 147 phút.
- d) Khoảng tin cậy 99% sẽ hẹp hơn khoảng tin cậy 95% vì chúng ta muốn chắc chắn hơn về ước lượng.
- e) Biên độ sai số là 9,5 và trung bình mẫu là 137,5.
- f) Để giảm biên độ sai số của khoảng tin cậy 95% xuống còn một nửa, chúng ta cần tăng gấp đôi cỡ mẫu. (Gợi ý: biên độ sai số đối với trung bình tỉ lệ nghịch với căn bậc hai của cỡ mẫu, tương tự như với tỷ lệ.)

5.12 Sức khỏe tâm thần. Khảo sát Xã hội Tổng quát (General Social Survey) đã hỏi câu: “Trong 30 ngày vừa qua, có bao nhiêu ngày sức khỏe tâm thần của bạn - bao gồm căng thẳng, trầm cảm và các vấn đề cảm xúc - không tốt? Dựa trên câu trả lời từ 1151 cư dân Hoa Kỳ, khảo sát đã báo cáo một khoảng tin cậy 95% là từ 3,40 đến 4,24 ngày, vào năm 2010.

- a) Diễn giải khoảng tin cậy này trong bối cảnh của dữ liệu.
- b) “Tin tưởng 95%” có nghĩa là gì? Giải thích trong bối cảnh của ứng dụng.
- c) Giả sử các nhà nghiên cứu cho rằng mức độ tin cậy 99% sẽ phù hợp hơn cho khoảng này. Liệu khoảng tin cậy mới sẽ rộng hơn hay hẹp hơn so với khoảng tin cậy 95%?
- d) Nếu một khảo sát mới được thực hiện với 500 người Mỹ, bạn nghĩ rằng sai số chuẩn của ước lượng sẽ lớn hơn, nhỏ hơn hay gần như không đổi?

5.13 Đăng ký tài khoản trên trang web. Một trang web đang cố gắng tăng tỷ lệ đăng ký của những người truy cập lần đầu, bằng cách hiển thị thiết kế giao diện mới cho 1% số người truy cập. Trong số 752 người truy cập được chọn ngẫu nhiên trong một tháng và đã xem thiết kế mới, có 64 người đã đăng ký.

- a) Kiểm tra các điều kiện cần thiết để xây dựng một khoảng tin cậy.
- b) Tính sai số chuẩn.
- c) Xây dựng và diễn giải một khoảng tin cậy 90% cho tỷ lệ người truy cập lần đầu sẽ đăng ký với thiết kế mới, giả định rằng hành vi của người dùng mới ổn định theo thời gian.

5.14 Phiếu giảm giá thúc đẩy lượt đến cửa hàng. Một cửa hàng đã chọn ngẫu nhiên 603 khách hàng trong suốt một năm và phát hiện rằng 142 người trong số họ đến cửa hàng vì đã nhận được phiếu giảm giá qua thư. Hãy xây dựng một khoảng tin cậy 95% cho tỷ lệ toàn bộ khách hàng trong năm đến cửa hàng vì đã nhận được phiếu giảm giá qua thư.

5.3 KIỂM ĐỊNH GIẢ THUYẾT CHO MỘT TỶ LỆ

5.3 Kiểm định giả thuyết cho một tỷ lệ

Câu hỏi sau đây trích từ cuốn sách *Factfulness* của Hans Rosling, Anna Rosling Rönnlund và Ola Rosling: *Hiện nay, có bao nhiêu phần trăm trẻ em 1 tuổi trên thế giới đã được tiêm vắc-xin phòng chống một số bệnh?*

- a) 20%
- b) 50%
- c) 80%

Hãy viết ra câu trả lời (hoặc dự đoán) của bạn, và khi đã sẵn sàng, hãy kiểm tra đáp án ở phần chú thích.¹³

Trong phần này, chúng ta sẽ khám phá cách những người có bằng đại học 4 năm trả lời câu hỏi này và các câu hỏi khác về sức khỏe toàn cầu, khi ta học về kiểm định giả thuyết - một khuôn khổ được sử dụng để đánh giá chặt chẽ các ý tưởng và tuyên bố đối lập.

5.3.1 Khung kiểm định giả thuyết

Chúng ta đang quan tâm đến mức độ hiểu biết của con người về sức khỏe và phát triển toàn cầu. Nếu chúng ta đưa ra một câu hỏi trắc nghiệm liên quan đến sức khỏe toàn cầu, thì ta có thể muốn tìm hiểu xem:

H_0 : Con người không bao giờ học những chủ đề cụ thể này và câu trả lời của họ hoàn toàn tương đương với việc đoán ngẫu nhiên.

H_A : Con người có kiến thức giúp họ trả lời tốt hơn đoán ngẫu nhiên, hoặc thậm chí, họ có hiểu lầm khiến họ trả lời tệ hơn cả đoán ngẫu nhiên.

Những ý tưởng đối lập này được gọi là **giả thuyết**. Chúng ta gọi H_0 là giả thuyết không và H_A là giả thuyết thay thế. Khi có chỉ số dưới như 0 trong H_0 , các nhà khoa học dữ liệu thường đọc là “nought”. (ví dụ H_0 được phát âm là “H-nought”).

GIẢ THUYẾT KHÔNG VÀ GIẢ THUYẾT THAY THẾ

Giả thuyết không (H_0) thường đại diện cho một nghi vấn hoặc là một tuyên bố cần được kiểm định. **Giả thuyết thay thế (H_A)** đại diện cho một tuyên bố thay thế đang được xem xét và thường được thể hiện bằng một khoảng giá trị khả dĩ của tham số.

Vai trò của chúng ta với tư cách là nhà khoa học dữ liệu là nghi ngờ: trước khi chấp nhận giả thuyết thay thế, chúng ta cần phải thấy những bằng chứng mạnh mẽ ủng hộ.

Giả thuyết không thường đại diện cho một quan điểm hoài nghi hoặc góc nhìn “không có sự khác biệt”. Trong ví dụ đầu tiên, chúng ta sẽ xem xét liệu người bình thường có trả lời khác gì so với những người đoán ngẫu nhiên trong câu hỏi của Rosling về việc tiêm vắc-xin cho trẻ sơ sinh hay không.

Giả thuyết thay thế thường đại diện cho một quan điểm mới hoặc mạnh mẽ hơn. Trong trường hợp câu hỏi về tiêm vắc-xin cho trẻ sơ sinh, sẽ rất thú vị nếu biết rằng người ta trả lời tốt hơn so với đoán ngẫu nhiên, vì điều đó cho thấy người bình thường có hiểu biết nhất định về thống kê sức khỏe toàn cầu. Ngược lại, nếu phát hiện rằng họ trả lời tệ hơn cả đoán ngẫu nhiên, điều đó cũng rất đáng quan tâm, vì nó cho thấy họ có thể đang tin vào thông tin sai lệch về sức khỏe toàn cầu.

Khung kiểm định giả thuyết là một công cụ rất phổ biến và chúng ta thường sử dụng nó một cách tự nhiên. Khi ai đó đưa ra một tuyên bố khó tin, phản ứng đầu tiên của chúng ta là hoài nghi. Tuy nhiên, nếu có đủ bằng chứng để hỗ trợ tuyên bố đó, chúng ta sẽ gác lại sự nghi ngờ và bác bỏ giả thuyết không để chấp nhận giả thuyết thay thế. Dấu hiệu đặc trưng của kiểm định giả thuyết cũng có thể được tìm thấy trong hệ thống tòa án Hoa Kỳ.

¹³ Câu trả lời đúng là (c): 80% trẻ em 1 tuổi trên thế giới đã được tiêm vắc-xin phòng ít nhất một loại bệnh.

5.3 KIỂM ĐỊNH GIẢ THUYẾT CHO MỘT TỶ LỆ

BÀI TẬP THỰC HÀNH 5.17

G

Một tòa án Hoa Kỳ xem xét hai khả năng về bị cáo: cô ấy vô tội hoặc có tội. Nếu áp dụng khung kiểm định giả thuyết, đâu là giả thuyết không và đâu là giả thuyết thay thế? ¹⁴

Bồi thẩm đoàn đánh giá các bằng chứng để xem liệu chúng có chứng minh được bị cáo có tội một cách thuyết phục hay không. Ngay cả khi bồi thẩm đoàn không bị thuyết phục rằng bị cáo có tội vượt qua mức nghi ngờ hợp lý, điều đó không có nghĩa là họ tin rằng bị cáo vô tội. Điều này cũng giống với kiểm định giả thuyết thống kê: *ngay cả khi chúng ta không bác bỏ được giả thuyết không (H_0), chúng ta cũng không nên coi giả thuyết không là đúng*. Không tìm thấy bằng chứng đủ mạnh cho giả thuyết thay thế không đồng nghĩa với việc chấp nhận giả thuyết không.

Khi xem xét câu hỏi của Rosling về tỷ lệ tiêm vắc-xin ở trẻ 1 tuổi, giả thuyết không cho rằng những người được khảo sát - cụ thể là người trưởng thành có bằng đại học - chỉ chọn đáp án đúng một cách ngẫu nhiên, tương tự như việc đoán mò. Điều này có nghĩa là tỷ lệ p những người chọn đúng đáp án (rằng 80% trẻ 1 tuổi trên thế giới đã được tiêm ít nhất một loại vắc-xin) là khoảng 33,3% - tức một phần ba. Giả thuyết thay thế cho rằng tỷ lệ này khác 33,3%, nghĩa là người tham gia khảo sát có thể biết nhiều hơn hoặc ít hơn so với việc đoán ngẫu nhiên. Ngoài cách diễn đạt bằng lời, ta cũng có thể biểu diễn hai giả thuyết này bằng ký hiệu toán học như sau:

$$H_0: p = 0,333$$

$$H_A: p \neq 0,333$$

Trong thiết lập kiểm định giả thuyết này, mục tiêu của chúng ta là đưa ra kết luận về tham số của tổng thể, được ký hiệu là p . Giá trị mà chúng ta so sánh với tham số này được gọi là **giá trị giả thuyết không** (null value), và trong trường hợp này là 0,333. Thông thường, giá trị giả thuyết không được ký hiệu giống với tham số nhưng thêm chỉ số dưới là "0", tức là $p_0 = 0,333$ (phát âm là "p-nought bằng 0,333").

VÍ DỤ 5.18

Có thể chúng ta thấy khó tin rằng tỷ lệ người chọn đúng đáp án lại chính xác là 33,3%. Vậy nếu không tin vào giả thuyết không, liệu có nên đơn giản là bác bỏ nó không?

E

Không. Mặc dù chúng ta có thể không tin rằng tỷ lệ chính xác là 33,3%, nhưng khung kiểm định giả thuyết yêu cầu phải có bằng chứng mạnh mẽ thì mới được bác bỏ giả thuyết không và đi đến một kết luận thú vị hơn.

Suy cho cùng, việc không tin rằng tỷ lệ là *chính xác* 33,3% cũng không mang lại thông tin hữu ích gì! Ta vẫn sẽ bị mắc kẹt với câu hỏi ban đầu: Liệu người tham gia có làm tốt hơn hay tệ hơn so với việc chọn ngẫu nhiên trong câu hỏi của Rosling? Nếu không có dữ liệu nào chỉ rõ ràng theo hướng này hay hướng kia, thì việc bác bỏ H_0 là vừa không thú vị vừa vô nghĩa.

BÀI TẬP THỰC HÀNH 5.19

G

Một ví dụ khác trong thực tế về kiểm định giả thuyết là đánh giá xem một loại thuốc mới có hiệu quả hơn hay kém hơn so với thuốc hiện có trong việc điều trị một bệnh cụ thể. Trong trường hợp này, ta nên đặt giả thuyết không và giả thuyết thay thế thế nào? ¹⁵

¹⁴ Bồi thẩm đoàn sẽ xem xét liệu các bằng chứng có đủ thuyết phục và mạnh mẽ đến mức không còn nghi ngờ hợp lý nào về việc bị cáo có tội hay không; trong trường hợp đó, bồi thẩm đoàn sẽ bác bỏ giả định vô tội (giả thuyết không) và kết luận bị cáo có tội (giả thuyết thay thế).

¹⁵ Giả thuyết không (H_0) trong trường hợp này là tuyên bố không có *sự khác biệt*: tức là thuốc mới và thuốc cũ có hiệu quả điều trị như nhau. Ngược lại, giả thuyết thay thế (H_A) là thuốc mới có hiệu quả khác với thuốc cũ, nghĩa là nó có thể tốt hơn hoặc kém hơn thuốc hiện tại.

5.3 KIỂM ĐỊNH GIẢ THUYẾT CHO MỘT TỶ LỆ

5.3.2 Kiểm định giả thuyết bằng cách sử dụng khoảng tin cậy

Chúng ta sẽ sử dụng bộ dữ liệu `rosling_responses` để đánh giá kiểm định giả thuyết đánh giá xem người lớn có trình độ đại học trả lời đúng câu hỏi về tiêm chủng cho trẻ sơ sinh có khác với 33,3%. Bộ dữ liệu này tóm tắt câu trả lời của 50 người trưởng thành có trình độ đại học, trong đó 24% trả lời đúng rằng 80% trẻ 1 tuổi trên thế giới đã được tiêm vắc-xin phòng một bệnh nào đó.

Tính đến thời điểm này, các thảo luận của chúng ta chủ yếu mang tính triết lý. Tuy nhiên, giờ đây khi đã có dữ liệu thực tế, ta có thể tự hỏi: liệu dữ liệu này có cung cấp bằng chứng mạnh mẽ rằng tỷ lệ trong toàn bộ quần thể người lớn có trình độ đại học trả lời đúng thực sự khác với 33,3%?

Như đã học ở Mục 5.1, các mẫu khác nhau sẽ có dao động ngẫu nhiên và hiếm khi tỷ lệ mẫu $\hat{p}$, bằng chính xác với tỷ lệ p , tuy nhiên mục tiêu của chúng ta là đưa ra kết luận về p . Câu hỏi đặt ra là: liệu sự lệch 24% so với 33,3% có đơn thuần là do ngẫu nhiên, hay dữ liệu này thực sự cung cấp bằng chứng mạnh mẽ rằng tỷ lệ tổng thể khác với 33,3%?

Ở Mục 5.2, ta đã học cách định lượng mức độ không chắc chắn trong ước lượng thông qua khoảng tin cậy. Phương pháp định lượng sự biến động này cũng rất hữu ích khi áp dụng vào kiểm định giả thuyết.

VÍ DỤ 5.20

Kiểm tra xem có hợp lý để xây dựng khoảng tin cậy cho p bằng dữ liệu mẫu hay không, và nếu có thì hãy xây dựng khoảng tin cậy 95%.

Các điều kiện để ước lượng $\hat{p}$ gần đúng theo phân phối chuẩn đã được thỏa mãn: dữ liệu thu được từ một mẫu ngẫu nhiên đơn (đáp ứng điều kiện độc lập), và $n\hat{p} = 12$ và $n(1 - \hat{p}) = 38$ đều lớn hơn hoặc bằng 10 (thỏa mãn điều kiện thành công – thất bại).

E

Để xây dựng khoảng tin cậy, ta cần xác định ước lượng điểm ($\hat{p} = 0.24$), giá trị tới hạn cho mức tin cậy 95% ($z^* = 1,96$), và sai số chuẩn của $\hat{p}$ ($SE_{\hat{p}} = \sqrt{\hat{p}(1 - \hat{p})/n} = 0,060$). Với các giá trị này, khoảng tin cậy cho p được tính như sau:

$$\begin{aligned} \hat{p} \pm z^* \times SE_{\hat{p}} \\ 0.24 \pm 1,96 \times 0,060 \\ (0,122 - 0,358) \end{aligned}$$

Chúng tôi tin tưởng 95% rằng tỷ lệ người trưởng thành có bằng đại học trả lời đúng câu hỏi cụ thể về việc tiêm chủng cho trẻ sơ sinh nằm trong khoảng từ 12,2% đến 35,8%.

Vì giá trị giả thuyết không trong kiểm định giả thuyết là $p_0 = 0,333$, nằm trong khoảng các giá trị hợp lý của khoảng tin cậy, nên ta không thể kết luận rằng giá trị giả thuyết không này là không hợp lý.¹⁶ Nói cách khác, dữ liệu không cung cấp đủ bằng chứng để bác bỏ giả thuyết rằng hiệu suất trả lời của những người trưởng thành có bằng đại học khác với việc đoán ngẫu nhiên, và do đó chúng ta không bác bỏ giả thuyết H_0 .

VÍ DỤ 5.21

Giải thích vì sao chúng ta không thể kết luận rằng những người trưởng thành có bằng đại học chỉ đơn thuần đoán mò trong câu hỏi về tiêm vắc-xin cho trẻ một tuổi.

E

Mặc dù chúng ta không bác bỏ giả thuyết H_0 , điều đó không có nghĩa là giả thuyết không là đúng. Có thể thực sự tồn tại một sự khác biệt, nhưng chúng ta không phát hiện được sự khác biệt đó do cỡ mẫu khảo sát tương đối nhỏ – chỉ gồm 50 người.

ĐÔI KHI TRONG THỐNG KÊ CÓ THỂ DÙNG PHỦ ĐỊNH KÉP

Trong nhiều giải thích thống kê, chúng ta sử dụng các câu phủ định kép. Ví dụ, ta có thể nói rằng giả thuyết không là *không phi lý* hoặc chúng ta *chưa bác bỏ* giả thuyết không. Việc dùng phủ định kép nhằm truyền đạt rằng mặc dù chúng ta không bác bỏ một quan điểm, điều đó cũng không có nghĩa là chúng ta khẳng định quan điểm đó là đúng.

¹⁶ Có thể lập luận rằng phương pháp này hơi thiếu chính xác. Như chúng ta sẽ thấy trong vài trang tới, sai số chuẩn thường được tính theo một cách hơi khác trong bối cảnh kiểm định giả thuyết cho một tỷ lệ.

5.3 KIỂM ĐỊNH GIẢ THUYẾT CHO MỘT TỶ LỆ

BÀI TẬP THỰC HÀNH 5.22

Hãy chuyển sang câu hỏi thứ hai được nhóm tác giả Rosling đặt ra:

Hiện nay trên thế giới có 2 tỷ trẻ em trong độ tuổi từ 0–15. Theo Liên Hợp Quốc, vào năm 2100 sẽ có bao nhiêu trẻ em ở độ tuổi này?

G

- a) 4 tỷ.
- b) 3 tỷ.
- c) 2 tỷ.

Thiết lập giả thuyết phù hợp để đánh giá xem liệu những người trưởng thành có trình độ đại học có trả lời tốt hơn so với việc đoán ngẫu nhiên hay không cho câu hỏi này. Đồng thời, bạn hãy thử đoán câu trả lời đúng trước khi kiểm tra đáp án ở chú thích nhé!¹⁷

BÀI TẬP THỰC HÀNH 5.23

G

Lần này, chúng ta lấy một mẫu lớn hơn gồm 228 người trưởng thành có trình độ đại học, trong đó có 34 người (tương đương 14,9%) chọn đúng đáp án cho câu hỏi trong Thực hành có hướng dẫn 5.22: 2 tỷ. Liệu chúng ta có thể mô hình hóa tỷ lệ mẫu này bằng phân phối chuẩn và xây dựng một khoảng tin cậy không?¹⁸

VÍ DỤ 5.24

Tính khoảng tin cậy 95% cho tỷ lệ người trưởng thành có trình độ đại học trả lời đúng câu hỏi “số lượng trẻ em vào năm 2100” và đánh giá giả thuyết trong Thực hành có hướng dẫn 5.22.

Để tính sai số chuẩn, ta tiếp tục sử dụng tỷ lệ mẫu $\hat{p}$ thay cho p :

$$SE_{\hat{p}} = \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} = \sqrt{\frac{0,149(1 - 0,149)}{228}} = 0,024$$

Trong Thực hành có hướng dẫn 5.23, ta đã xác định rằng $\hat{p}$ có thể được mô hình hóa bằng phân phối chuẩn, do đó có thể xây dựng khoảng tin cậy 95% một cách chính xác:

E

$$\hat{p} \pm z^* \times SE \rightarrow 0,149 \pm 1,96 \times 0,024 \rightarrow (0,103 - 0,195)$$

Vì giá trị giả thuyết không $p_0 = 0,333$ không nằm trong khoảng tin cậy này, nên một tỷ lệ trong dân số là 0,333 được xem là không hợp lý và chúng ta bác bỏ giả thuyết không. Nói cách khác, dữ liệu cung cấp bằng chứng thống kê có ý nghĩa rằng tỷ lệ thực sự của những người trưởng thành có trình độ đại học trả lời đúng câu hỏi về trẻ em vào năm 2100 là khác với đoán ngẫu nhiên. Vì toàn bộ khoảng tin cậy 95% nằm dưới mức 0,333, chúng ta có thể kết luận rằng những người có trình độ đại học trả lời *tệ hơn* cả đoán mò trong câu hỏi này.

Một điều cần lưu ý là chúng ta đã sử dụng khoảng tin cậy 95%. Vậy nếu dùng mức tin cậy 99% hoặc thậm chí 99,9% thì sao? Hoàn toàn có thể đưa ra kết luận khác nếu sử dụng mức tin cậy khác. Vì thế, khi đưa ra kết luận dựa trên khoảng tin cậy, chúng ta cũng cần chắc chắn rằng mức tin cậy đã sử dụng được nêu rõ.

Kết quả làm tệ hơn cả đoán mò ở câu hỏi này không phải là sự tình cờ: có nhiều câu hỏi về sức khỏe toàn cầu mà ở đó người ta trả lời tệ hơn cả đoán mò. Nhìn chung, các câu trả lời cho thấy con người có xu hướng bi quan về tiến bộ hơn là thực tế. Chủ đề này được thảo luận chi tiết hơn trong cuốn *Factfulness* của Roslings.

¹⁷ Các giả thuyết phù hợp là:

H_0 : Tỷ lệ người trả lời đúng bằng với đoán mò: 1 trên 3, tức là $p = 0,333$.

H_A : Tỷ lệ người trả lời đúng khác với đoán mò, tức là $p \neq 0,333$.

Câu trả lời đúng cho câu hỏi là 2 tỷ. Mặc dù dân số thế giới dự kiến sẽ tăng, nhưng độ tuổi trung bình cũng được dự báo sẽ tăng theo. Điều này có nghĩa là phần lớn sự gia tăng dân số sẽ xảy ra ở các nhóm tuổi lớn hơn, tức là người dân trên toàn thế giới được dự đoán sẽ sống lâu hơn trong tương lai.

¹⁸ Chúng ta kiểm tra cả hai điều kiện, và cả hai đều được thỏa mãn, vì vậy việc sử dụng phân phối chuẩn cho $\hat{p}$ là hợp lý:

Tính độc lập. Vì dữ liệu đến từ một mẫu ngẫu nhiên đơn, nên các quan sát là độc lập.

Điều kiện thành công - thất bại. Ta sử dụng $\hat{p}$ thay cho p để kiểm tra: $n\hat{p} = 34$ và $n(1 - \hat{p}) = 194$. Cả hai đều lớn hơn 10, vì vậy điều kiện này được thỏa mãn.

5.3 KIỂM ĐỊNH GIẢ THUYẾT CHO MỘT TỶ LỆ

5.3.3 Sai lầm trong quyết định

Các kiểm định giả thuyết không phải lúc nào cũng hoàn hảo: chúng ta có thể đưa ra quyết định sai dựa trên dữ liệu thống kê. Ví dụ, trong hệ thống pháp luật, đôi khi người vô tội bị kết án sai và người có tội lại được trắng án. Một điểm khác biệt quan trọng với kiểm định giả thuyết thống kê là chúng ta có các công cụ để định lượng xác suất xảy ra sai sót trong kết luận của mình.

Hãy nhớ rằng luôn có hai giả thuyết thay thế lập: giả thuyết không và giả thuyết thay thế. Trong một kiểm định giả thuyết, chúng ta đưa ra tuyên bố về giả thuyết nào có thể đúng, nhưng cũng có thể chọn sai. Có bốn kịch bản có thể xảy ra, được tóm tắt trong Hình 5.8.

	Kết luận kiểm định	
	Không bác bỏ H_0	Bác bỏ H_0 , ủng hộ H_A
	Kết quả đúng	Sai lầm loại 1
Kết quả	H_0 đúng	Kết quả đúng
	H_A đúng	Sai lầm loại 2

Hình 5.8: Bốn trường hợp khác nhau cho kiểm định giả thuyết.

Sai lầm loại I là khi bác bỏ giả thuyết không trong khi H_0 thực sự đúng. **Sai lầm loại II** là khi không bác bỏ giả thuyết không trong khi giả thuyết thay thế thực sự đúng.

G

BÀI TẬP THỰC HÀNH 5.25

Trong một tòa án ở Hoa Kỳ, bị cáo hoặc là vô tội (H_0) hoặc có tội (H_A). Vậy sai lầm loại 1 đại diện cho điều gì trong bối cảnh này? Còn sai lầm loại 2 thì sao? Hình 5.8 có thể sẽ giúp ích.¹⁹

E

VÍ DỤ 5.26

Làm thế nào để giảm tỷ lệ sai lầm loại 1 trong các tòa án Hoa Kỳ? Điều này ảnh hưởng thế nào đến tỷ lệ sai lầm loại 2?

Để giảm tỷ lệ sai lầm loại 1, chúng ta có thể nâng tiêu chuẩn kết án từ “vượt qua sự nghi ngờ hợp lý” lên “vượt qua mọi khả năng nghi ngờ”, để ít người vô tội bị kết án hơn. Tuy nhiên, điều này cũng khiến việc kết án người có tội trở nên khó khăn hơn, dẫn đến việc tăng tỷ lệ sai lầm loại 2.

G

BÀI TẬP THỰC HÀNH 5.27

Làm thế nào để giảm tỷ lệ sai lầm loại 2 trong các tòa án Hoa Kỳ? Điều này ảnh hưởng thế nào đến tỷ lệ sai lầm loại 1?²⁰

Các bài tập 5.25-5.27 mang đến một bài học quan trọng: nếu chúng ta giảm tần suất mắc một loại sai lầm, thì thường sẽ làm tăng tần suất mắc sai lầm còn lại.

Kiểm định giả thuyết được xây dựng dựa trên việc bác bỏ hoặc không bác bỏ giả thuyết không. Nghĩa là, chúng ta chỉ bác bỏ H_0 nếu có bằng chứng đủ mạnh. Nhưng *bằng chứng mạnh* cụ thể nghĩa là gì? Theo nguyên tắc chung, trong những trường hợp giả thuyết không thực sự đúng, chúng ta không muốn bác bỏ H_0 sai quá 5% số lần. Điều này tương ứng với **mức ý nghĩa (significance level)** là 0,05. Nói cách khác, nếu H_0 đúng, thì mức ý nghĩa cho biết xác suất dữ liệu thu được dẫn đến việc bác bỏ sai H_0 là bao nhiêu. Chúng ta thường ký hiệu mức ý nghĩa này bằng: $\alpha = 0,05$. Chúng ta sẽ thảo luận kỹ hơn về mức ý nghĩa nào là phù hợp trong Mục 5.3.5.

¹⁹ Nếu tòa án mắc sai lầm loại 1, điều này có nghĩa là bị cáo vô tội (H_0 đúng) nhưng lại bị kết án oan. Lưu ý rằng sai lầm loại 1 chỉ có thể xảy ra khi ta đã bác bỏ giả thuyết không (H_0).

Sai lầm loại 2 xảy ra khi tòa án không bác bỏ H_0 (tức là không kết tội người đó) trong khi trên thực tế, người đó có tội (H_A đúng). Lưu ý rằng sai lầm loại 2 chỉ xảy ra khi ta không bác bỏ giả thuyết không.

²⁰ Để giảm tỉ lệ mắc sai lầm loại 2, chúng ta cần kết án nhiều người có tội hơn. Có thể làm được điều này bằng cách giảm tiêu chuẩn kết tội, từ mức “vượt qua sự nghi ngờ hợp lý” xuống chỉ còn “vượt qua một chút nghi ngờ”. Tuy nhiên, việc hạ thấp tiêu chuẩn buộc tội này cũng sẽ khiến nhiều người vô tội bị kết án oan hơn, tức là làm tăng tỉ lệ sai lầm loại 1.

5.3 KIỂM ĐỊNH GIẢ THUYẾT CHO MỘT TỶ LỆ

Nếu chúng ta sử dụng khoảng tin cậy 95% để đánh giá một kiểm định giả thuyết và giả thuyết không xảy ra đúng, chúng ta sẽ phạm lỗi khi ước lượng điểm nằm ít nhất 1,96 sai số chuẩn cách xa tham số tổng thể. Điều này xảy ra khoảng 5% thời gian (2,5% ở mỗi phần đuôi). Tương tự, sử dụng khoảng tin cậy 99% để đánh giá giả thuyết tương đương với mức ý nghĩa $\alpha = 0,01$.

Khoảng tin cậy rất hữu ích trong việc xác định có nên bác bỏ giả thuyết không hay không. Tuy nhiên, phương pháp này không phải lúc nào cũng khả thi. Trong một số phần, chúng ta sẽ gặp các tình huống mà không thể xây dựng khoảng tin cậy. Ví dụ, nếu muốn đánh giá giả thuyết rằng một số tỷ lệ phần trăm bằng nhau, thì việc xây dựng và so sánh nhiều khoảng tin cậy cùng lúc không rõ ràng.

Tiếp theo, chúng ta sẽ giới thiệu một thống kê gọi là *p-value* để giúp mở rộng bộ công cụ thống kê của chúng ta, điều này sẽ giúp chúng ta hiểu rõ hơn về độ mạnh của chứng cứ và làm việc với dữ liệu phức tạp hơn trong các phần sau.

5.3.4 Phương pháp kiểm định dựa trên giá trị *p* (*p-value*)

Giá trị *p* là một cách để đo lường mức độ rõ ràng của bằng chứng chống lại giả thuyết không và ủng hộ giả thuyết thay thế lập. Thông thường, kiểm định giả thuyết thống kê sử dụng phương pháp giá trị *p* thay vì đưa ra quyết định dựa trên khoảng tin cậy.

GIÁ TRỊ *P*

Giá trị *p* là xác suất để quan sát dữ liệu ít nhiều có lợi cho giả thuyết thay thế lập hơn hoặc bằng dữ liệu hiện tại của chúng ta, nếu giả thuyết không là đúng. Thông thường, chúng ta sử dụng một thống kê tóm tắt của dữ liệu, trong phần này là tỷ lệ mẫu, để giúp tính toán giá trị *p* và đánh giá các giả thuyết.

VÍ DỤ 5.28

Nghiên cứu Pew hỏi một mẫu ngẫu nhiên gồm 1000 người trưởng thành ở Mỹ về việc họ có ủng hộ việc tăng sử dụng than để sản xuất năng lượng hay không. Hãy thiết lập các giả thuyết để đánh giá xem liệu đa số người trưởng thành ở Mỹ có ủng hộ hay phản đối việc tăng sử dụng than.

E

Kết quả không đáng chú ý là không có đa số nào rõ ràng: một nửa người Mỹ ủng hộ và nửa còn lại phản đối việc mở rộng sử dụng than để sản xuất năng lượng. Giả thuyết thay thế lập sẽ là có đa số ủng hộ hoặc phản đối (mặc dù chúng ta không biết chính xác là cái nào!). Nếu *p* đại diện cho tỷ lệ ủng hộ, thì chúng ta có thể viết các giả thuyết như sau:

$$H_0: p = 0,5$$

$$H_A: p \neq 0,5$$

Trong trường hợp này, giá trị không (null value) là $p_0 = 0,5$.

Khi đánh giá các giả thuyết về tỷ lệ phần trăm bằng phương pháp giá trị *p*, chúng ta sẽ điều chỉnh một chút cách kiểm tra điều kiện thành công-thất bại và tính sai số chuẩn cho trường hợp tỷ lệ đơn. Những thay đổi này không quá lớn, nhưng hãy chú ý cẩn thận cách chúng ta sử dụng giá trị không, p_0 .

5.3 KIỂM ĐỊNH GIẢ THUYẾT CHO MỘT TỶ LỆ

VÍ DỤ 5.29

Mẫu nghiên cứu của Pew cho thấy 37% người trưởng thành ở Mỹ ủng hộ việc tăng sử dụng than. Bây giờ, chúng ta tự hỏi, liệu 37% có thực sự khác biệt so với giả thuyết không là 50% không? Phân phối mẫu của $\hat{p}$ sẽ trông như thế nào nếu giả thuyết không là đúng?

Nếu giả thuyết không đúng, tỷ lệ trong tổng thể sẽ là giá trị không (null value), 0.5. Trước đó, chúng ta đã học rằng phân phối mẫu của $\hat{p}$ sẽ là phân phối chuẩn khi hai điều kiện sau được đáp ứng:

Độc lập: Cuộc thăm dò dựa trên một mẫu ngẫu nhiên đơn, nên điều kiện độc lập được thỏa mãn.

Thành công-thất bại: Dựa trên kích thước mẫu $n = 1000$, điều kiện thành công-thất bại được đáp ứng vì:

$$np \stackrel{H_0}{=} 1000 \times 0.5 = 500 \qquad n(1-p) \stackrel{H_0}{=} 1000 \times (1-0.5) = 500$$

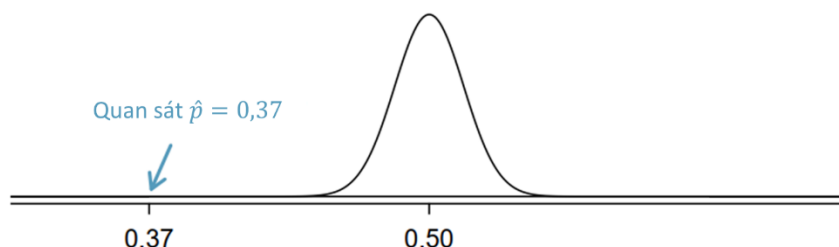
đều ít nhất 10. Lưu ý rằng điều kiện thành công-thất bại đã được kiểm tra dựa trên giá trị không (null value), ($p_0 = 0.5$); đây là khác biệt thủ tục đầu tiên so với khoảng tin cậy.

Nếu giả thuyết không là đúng, phân phối mẫu cho biết rằng tỷ lệ mẫu dựa trên $n = 1000$ quan sát sẽ phân phối theo dạng chuẩn. Tiếp theo, chúng ta có thể tính sai số chuẩn, trong đó lần này sẽ sử dụng lại giá trị không (null value) ($p_0 = 0.5$) trong phép tính:

$$SE_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} \stackrel{H_0}{=} \sqrt{\frac{0.5 \times (1-0.5)}{1000}} = 0.016$$

Điều này là điểm khác biệt thủ tục thứ hai so với khoảng tin cậy: vì phân phối mẫu được xác định dựa trên tỷ lệ không hồi quy, nên giá trị không (null value) p_0 được sử dụng cho tỷ lệ trong phép tính chứ không phải $\hat{p}$.

Cuối cùng, nếu giả thuyết không đúng, thì tỷ lệ mẫu sẽ theo phân phối chuẩn với trung bình 0,5 và sai số chuẩn là 0,016. Phân phối này được thể hiện trong Hình 5.9.



Hình 5.9: Nếu giả thuyết không là đúng, phân phối chuẩn này mô tả phân phối của $\hat{p}$.

KIỂM TRA ĐIỀU KIỆN THÀNH CÔNG-THẤT BẠI VÀ TÍNH SAI SỐ CHUẨN $\hat{p}$ CHO KIỂM ĐỊNH GIẢ THUYẾT

Khi sử dụng phương pháp giá trị p để đánh giá một kiểm định giả thuyết, chúng ta kiểm tra các điều kiện cho $\hat{p}$ và tính sai số chuẩn bằng cách sử dụng giá trị không (null value) p_0 , thay vì dùng tỷ lệ mẫu.

Trong một kiểm định giả thuyết với giá trị p , chúng ta giả định rằng giả thuyết không là đúng, điều này khác với tư duy khi chúng ta tính khoảng tin cậy. Đó là lý do tại sao trong bối cảnh này, chúng ta sử dụng p_0 thay vì $\hat{p}$ khi kiểm tra điều kiện và tính sai số chuẩn.

Khi chúng ta xác định phân phối mẫu dưới giả thuyết không, nó có tên đặc biệt là **phân phối không (null distribution)**. Giá trị p thể hiện xác suất của tỷ lệ mẫu $\hat{p}$ quan sát được, hoặc một tỷ lệ mẫu còn cực đoan hơn, nếu giả thuyết không là đúng. Để tìm giá trị p , chúng ta thường xác định phân phối không, rồi sau đó tính diện tích vùng đuôi trong phân phối đó tương ứng với điểm ước lượng của chúng ta.

5.3 KIỂM ĐỊNH GIẢ THUYẾT CHO MỘT TỶ LỆ

VÍ DỤ 5.30

Nếu giả thuyết không là đúng, hãy xác định xác suất tìm thấy $\hat{p}$ ít nhất cũng xa về phía đuôi như 0,37 dưới phân phối chuẩn, với trung bình $\mu = 0,5$ và sai số chuẩn $SE = 0,016$.

Đây là một bài toán xác suất phân phối chuẩn, trong đó $x = 0,37$. Đầu tiên, chúng ta vẽ một đồ thị đơn giản để biểu diễn tình huống, tương tự như hình 5.9. Vì $\hat{p}$ nằm rất xa về phía đuôi, chúng ta biết diện tích đuôi sẽ rất nhỏ. Để tìm diện tích đó, chúng ta bắt đầu bằng cách tính Z-score using trung bình 0,5 và sai số chuẩn 0,016:

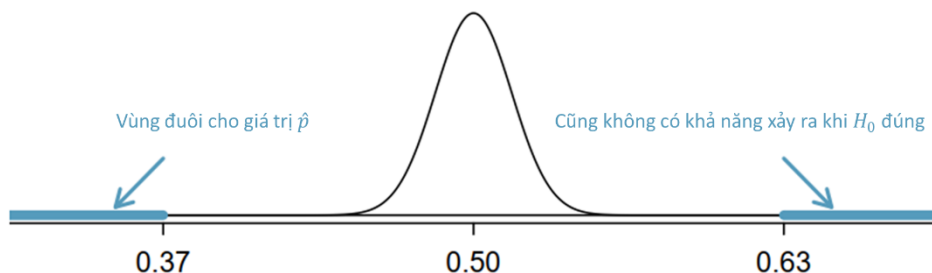
$$Z = \frac{0,37 - 0,5}{0,016} = -8,125$$

E

Chúng ta có thể sử dụng phần mềm để tìm diện tích đuôi: $2,2 \times 10^{-16}$ (0,00000000000000022). Nếu dùng bảng xác suất phân phối chuẩn trong Phụ lục C.1, chúng ta sẽ thấy rằng $Z = -8,125$ vượt khỏi bảng, vì vậy chúng ta sẽ dùng giá trị diện tích nhỏ nhất được liệt kê là 0,0002.

Các giá trị $\hat{p}$ trong đuôi trên vượt quá 0,63, như được thể hiện trong Hình 5.10, cũng biểu thị các quan sát ít nhất cũng cực đoan như giá trị quan sát 0,37. Để tính đến các giá trị này cũng cực đoan hơn theo giả thuyết ban đầu, ta nhân đôi diện tích đuôi dưới để ước lượng giá trị p: $4,4 \times 10^{-16}$ (hoặc nếu dùng phương pháp bảng, là 0,0004).

Giá trị p biểu thị xác suất quan sát được tỷ lệ mẫu cực đoan như vậy một cách ngẫu nhiên, nếu giả thuyết vô hiệu là đúng.



Hình 5.10: Nếu H_0 là đúng, thì các giá trị trên 0,63 cũng không dễ xảy ra hơn các giá trị dưới 0,37.

VÍ DỤ 5.31

Chúng ta nên đánh giá các giả thuyết như thế nào dựa trên giá trị p là $4,4 \times 10^{-16}$? Sử dụng mức ý nghĩa chuẩn $\alpha = 0,05$.

Nếu giả thuyết vô hiệu H_0 đúng, thì khả năng quan sát được sự lệch lệch cực đoan như vậy của $\hat{p}$ so với 0,5 là cực kỳ nhỏ. Điều này có nghĩa là một trong hai điều sau đây phải đúng:

E

1. Giả thuyết vô hiệu là đúng, và chúng ta chỉ tình cờ quan sát được điều gì đó quá cực đoan đến mức xảy ra khoảng một lần trong mỗi 23 quadrillion lần (1 quadrillion = 1 triệu $\times$ 1 tỷ).
2. Giả thuyết thay thế là đúng, điều này phù hợp với việc quan sát một tỷ lệ mẫu khác xa, 0,5.

Kịch bản đầu tiên là vô cùng khó xảy ra, trong khi kịch bản thứ hai có vẻ hợp lý hơn nhiều.

Chính thức, khi chúng ta đánh giá một bài kiểm tra giả thuyết, chúng ta so sánh giá trị p với mức ý nghĩa, trong trường hợp này là $\alpha = 0,05$. Vì giá trị p nhỏ hơn α , chúng ta bác bỏ giả thuyết vô hiệu H_0 . Điều này có nghĩa là dữ liệu cung cấp bằng chứng mạnh mẽ chống lại H_0 . Dữ liệu cũng chỉ ra hướng của sự khác biệt: phần lớn người Mỹ không ủng hộ mở rộng việc sử dụng năng lượng từ than đá.

5.3 KIỂM ĐỊNH GIẢ THUYẾT CHO MỘT TỶ LỆ

SO SÁNH GIÁ TRỊ p VỚI α ĐỂ ĐÁNH GIÁ H_0

Khi giá trị p nhỏ hơn mức ý nghĩa α , bác bỏ H_0 . Chúng ta sẽ kết luận rằng dữ liệu cung cấp bằng chứng mạnh mẽ ủng hộ giả thuyết thay thế.

Khi giá trị p lớn hơn α , không bác bỏ H_0 và báo cáo rằng chúng ta chưa có đủ bằng chứng để bác bỏ giả thuyết vô hiệu.

Trong cả hai trường hợp, quan trọng là mô tả kết luận trong bối cảnh của dữ liệu.



BÀI TẬP THỰC HÀNH 5.32

Một phần lớn người Mỹ ủng hộ hay phản đối giảm vũ khí hạt nhân? Đề ra các giả thuyết để đánh giá câu hỏi này.²¹

VÍ DỤ 5.33

Một mẫu ngẫu nhiên đơn giản gồm 1028 người trưởng thành ở Mỹ vào tháng 3 năm 2013 cho thấy 56% ủng hộ giảm vũ khí hạt nhân. Điều này có cung cấp bằng chứng thuyết phục rằng đa số người Mỹ ủng hộ giảm vũ khí hạt nhân hay không, với mức ý nghĩa 5%?

Trước tiên, kiểm tra các điều kiện:

Independence (Độc lập). Bên khảo sát là mẫu ngẫu nhiên đơn giản của người trưởng thành ở Mỹ, nghĩa là các quan sát độc lập với nhau.

Success-failure (Thành công-thất bại). Trong một bài kiểm tra giả thuyết về tỷ lệ đơn, điều kiện này được kiểm tra bằng tỷ lệ giả thuyết H_0 , là $p_0 = 0,5$ trong ngữ cảnh này: $np_0 = n(1 - p_0) = 1028 \times 0,5 = 514 \geq 10$.

Với các điều kiện này được xác nhận, chúng ta có thể mô hình hóa $\hat{p}$ bằng phân phối chuẩn.

Tiếp theo, có thể tính sai số chuẩn (SE). Giá trị giả thuyết p_0 được sử dụng lại ở đây, vì đây là bài kiểm tra giả thuyết cho một tỷ lệ duy nhất:

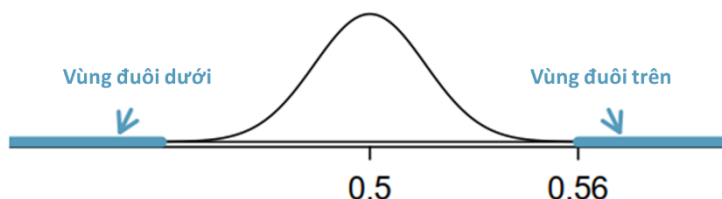
$$SE_{\hat{p}} = \sqrt{\frac{p_0(1 - p_0)}{n}} = \sqrt{\frac{0,5(1 - 0,5)}{1028}} \approx 0,0156$$



Dựa trên mô hình phân phối chuẩn, thống kê kiểm tra có thể được tính như Z-score của ước lượng điểm:

$$Z = \frac{\text{ước lượng điểm} - \text{giá trị không}}{SE} = \frac{0,56 - 0,5}{0,0156} = 3,85$$

Nói chung, việc vẽ phân phối null và các vùng đuôi quan tâm để tính giá trị p sẽ hữu ích:



Diện tích đuôi trên là khoảng 0,0001, và chúng ta nhân đôi diện tích này để lấy giá trị p : 0,0002. Vì giá trị p nhỏ hơn 0,05, chúng ta bác bỏ H_0 . Cuộc khảo sát cung cấp bằng chứng thuyết phục rằng đa số người Mỹ đã ủng hộ các nỗ lực giảm vũ khí hạt nhân vào tháng 3 năm 2013.

²¹ Chúng ta muốn hiểu xem đa số ủng hộ hay phản đối, hoặc cuối cùng, liệu có sự khác biệt hay không. Nếu p là tỷ lệ người Mỹ ủng hộ giảm vũ khí hạt nhân, thì $H_0: p = 0,50$ and $H_A: p \neq 0,50$.

5.3 KIỂM ĐỊNH GIẢ THUYẾT CHO MỘT TỶ LỆ

KIỂM ĐỊNH GIẢ THUYẾT CHO MỘT TỶ LỆ DUY NHẤT

Khi bạn đã xác định rằng kiểm định giả thuyết về một tỷ lệ là thủ tục phù hợp, có bốn bước để hoàn thành bài kiểm tra:

Chuẩn bị. Xác định tham số quan tâm, liệt kê các giả thuyết, xác định mức ý nghĩa, và xác định p và n .

Kiểm tra. Xác minh các điều kiện để đảm bảo p gần như phân phối chuẩn dưới H_0 . Đối với kiểm định giả thuyết về một tỷ lệ, sử dụng giá trị null để kiểm tra điều kiện thành công-thất bại.

Tính toán. Nếu các điều kiện được thỏa mãn, tính sai số chuẩn, một lần nữa sử dụng p_0 , tính Z-score, và xác định giá trị p .

Kết luận. Đánh giá bài kiểm tra giả thuyết bằng cách so sánh giá trị p với α , và đưa ra kết luận trong ngữ cảnh của bài toán.

5.3.5 Lựa chọn mức ý nghĩa thống kê

Lựa chọn mức ý nghĩa cho một bài kiểm tra là rất quan trọng trong nhiều ngữ cảnh, và mức truyền thống là $\alpha = 0,05$. Tuy nhiên, có thể hữu ích để điều chỉnh mức ý nghĩa dựa trên ứng dụng cụ thể. Chúng ta có thể chọn một mức nhỏ hơn hoặc lớn hơn 0,05 tùy thuộc vào hậu quả của các kết luận rút ra từ bài kiểm tra.

Nếu việc mắc Sai số Loại 1 (bác bỏ giả thuyết đúng) nguy hiểm hoặc đặc biệt tốn kém, chúng ta nên chọn mức ý nghĩa nhỏ (ví dụ 0,01). Trong trường hợp này, chúng ta muốn cẩn thận trong việc bác bỏ H_0 , vì vậy chúng ta đòi hỏi bằng chứng rất mạnh hỗ trợ cho H_A trước khi bác bỏ H_0 .

Ngược lại, nếu Sai số Loại 2 (không bác bỏ giả thuyết sai khi giả thuyết đúng) nguy hiểm hoặc tốn kém hơn nhiều so với Sai số Loại 1, thì chúng ta có thể chọn mức ý nghĩa cao hơn (ví dụ 0,10). Ở đây, chúng ta muốn cẩn trọng trong việc không bác bỏ H_0 khi giả thuyết thay thế thực sự đúng.

Ngoài ra, nếu chi phí thu thập dữ liệu nhỏ so với chi phí của Sai số Loại 2, thì cũng có thể là một chiến lược tốt khi thu thập thêm dữ liệu. Trong chiến lược này, Sai số Loại 2 có thể giảm mà không ảnh hưởng đến tỷ lệ Sai số Loại 1. Tất nhiên, việc thu thập dữ liệu bổ sung thường tốn kém, vì vậy thường có một phân tích lợi ích – chi phí cần cân nhắc.

VÍ DỤ 5.34

E

Một nhà sản xuất ô tô đang xem xét chuyển sang một loại thiết bị mới, chất lượng cao hơn, dùng để lắp đặt bản lề cửa xe. Họ nghĩ rằng họ sẽ tiết kiệm được tiền về lâu dài nếu chiếc máy mới này sản xuất ra các bản lề có tỷ lệ lỗi dưới 0,2%. Tuy nhiên, nếu các bản lề bị lỗi nhiều hơn 0,2%, họ sẽ không thu được lợi nhuận đầu tư đủ tốt từ thiết bị mới này và sẽ bị lỗ. Có lý do chính đáng để điều chỉnh mức ý nghĩa trong bài kiểm tra giả thuyết như vậy hay không?

Giả thuyết vô hiệu sẽ là tỷ lệ lỗi của các bản lề là 0,2%, trong khi giả thuyết thay thế là tỷ lệ này khác 0,2%. Quyết định này chỉ là một trong nhiều quyết định có tác động rời rạc đến xe và công ty. Mức ý nghĩa 0,05 dường như hợp lý vì cả Sai số Loại 1 lẫn Sai số Loại 2 đều không quá nguy hiểm hoặc (tương đối) tốn kém hơn nhiều.

5.3 KIỂM ĐỊNH GIẢ THUYẾT CHO MỘT TỶ LỆ

VÍ DỤ 5.35

Cùng nhà sản xuất ô tô đó đang xem xét chọn một nhà cung cấp phần vật tư an toàn, có giá cao hơn một chút, chứ không phải bản lề cửa. Nếu độ bền của các thành phần an toàn này cho thấy tốt hơn so với nhà cung cấp hiện tại, họ sẽ chuyển sang nhà sản xuất mới. Có lý do chính đáng để điều chỉnh mức ý nghĩa trong đánh giá như vậy hay không?

E

Giả thuyết vô hiệu sẽ là các bộ phận của các nhà cung cấp có độ tin cậy ngang nhau. Do liên quan đến an toàn, công ty ô tô nên hăng hái chuyển sang nhà cung cấp có giá cao hơn một chút (bác bỏ H_0), ngay cả khi bằng chứng về việc tăng cường an toàn chỉ mang tính vừa phải. Mức ý nghĩa nhỏ hơn, chẳng hạn như $\alpha = 0,10$, có thể phù hợp.

BÀI TẬP THỰC HÀNH 5.36

G

Một bộ phận bên trong của một máy móc rất đắt để thay thế. Tuy nhiên, máy thường hoạt động bình thường ngay cả khi bộ phận này bị hỏng, vì vậy bộ phận chỉ được thay thế nếu chúng ta cực kỳ chắc chắn rằng nó bị hỏng dựa trên một loạt các phép đo. Xác định các giả thuyết phù hợp cho bài kiểm tra này (bằng ngôn ngữ đơn giản) và đề xuất một mức ý nghĩa phù hợp.²²

TẠI SAO 0,05 LÀ MỨC CHUẨN MẶC ĐỊNH?

Ngưỡng $\alpha = 0,05$ là mức phổ biến nhất. Nhưng tại sao lại thế? Có thể mức tiêu chuẩn nên nhỏ hơn, hoặc lớn hơn. Nếu bạn còn cảm thấy bối rối, bạn đang đọc với một con mắt phản biện – làm tốt lắm! Chúng tôi đã tạo ra một nhiệm vụ 5 phút để giúp làm rõ lý do tại sao là 0,05:

www.openintro.org/why05

5.3.6 Ý nghĩa thống kê so với ý nghĩa thực tiễn

Khi kích thước mẫu trở nên lớn hơn, các ước lượng điểm trở nên chính xác hơn và bất kỳ sự khác biệt thực sự nào về trung bình so với giá trị null cũng dễ dàng được phát hiện và nhận biết hơn. Ngay cả một sự khác biệt rất nhỏ cũng có khả năng được phát hiện nếu chúng ta lấy mẫu đủ lớn. Đôi khi, các nhà nghiên cứu sẽ lấy mẫu lớn đến mức ngay cả những khác biệt nhỏ nhất cũng được phát hiện, thậm chí là những khác biệt không có giá trị thực tiễn. Trong những trường hợp đó, chúng ta vẫn nói rằng **sự khác biệt là có ý nghĩa thống kê**, nhưng không có **ý nghĩa thực tiễn**. Ví dụ, một thử nghiệm trực tuyến có thể xác định rằng việc đặt thêm quảng cáo trên một trang web đánh giá phim làm tăng đáng kể lượng người xem một chương trình truyền hình lên 0,001%, nhưng mức tăng này có thể không mang lại giá trị thực tiễn nào.

Một vai trò của nhà khoa học dữ liệu trong việc thực hiện một nghiên cứu thường bao gồm lập kế hoạch kích thước của nghiên cứu đó. Nhà khoa học dữ liệu có thể bắt đầu bằng cách tham khảo các chuyên gia hoặc tài liệu khoa học để biết sự khác biệt nhỏ nhất có ý nghĩa nào so với giá trị null. Cô ấy cũng sẽ thu thập các thông tin khác, chẳng hạn như một ước lượng sơ bộ về tỷ lệ thực sự p , để có thể ước lượng xấp xỉ sai số chuẩn. Từ đó, cô ấy có thể đề xuất một kích thước mẫu đủ lớn để, nếu có sự khác biệt thực sự mang ý nghĩa, chúng ta có thể phát hiện ra nó. Trong khi các kích thước mẫu lớn hơn vẫn có thể được sử dụng, những phép tính này đặc biệt hữu ích khi xem xét các chi phí hoặc rủi ro tiềm ẩn, chẳng hạn như ảnh hưởng đến sức khỏe của các tình nguyện viên trong một nghiên cứu y học.

²² Ở đây, giả thuyết vô hiệu là bộ phận không bị hỏng, còn giả thuyết thay thế là bộ phận bị hỏng. Nếu chúng ta không có đủ bằng chứng để bác bỏ H_0 , thì chúng ta sẽ không thay thế bộ phận đó. Nghe có vẻ như việc không sửa chữa bộ phận nếu nó bị hỏng (H_0 sai, H_A đúng) không quá gây vấn đề, trong khi việc thay thế bộ phận lại rất tốn kém. Do đó, chúng ta nên yêu cầu bằng chứng rất mạnh chống lại H_0 trước khi thay thế bộ phận. Chọn một mức ý nghĩa nhỏ, chẳng hạn như $\alpha = 0,01$.

5.3 KIỂM ĐỊNH GIẢ THUYẾT CHO MỘT TỶ LỆ

5.3.7 Các kiểm định giả thuyết một phía (chủ đề đặc biệt)

Cho đến nay, chúng ta chỉ xem xét các **kiểm định giả thuyết hai phía**, trong đó chúng ta quan tâm đến việc phát hiện xem p có thể cao hơn hoặc thấp hơn một giá trị không p_0 nào đó. Có một loại kiểm định giả thuyết thứ hai gọi là **kiểm định giả thuyết một phía**. Đối với kiểm định một phía, các giả thuyết có dạng sau:

1. Chỉ có ý nghĩa trong việc phát hiện xem tham số tổng thể *nhỏ hơn* một giá trị p_0 nào đó: Trong trường hợp này, giả thuyết thay thế được viết là $p < p_0$.
2. Chỉ có ý nghĩa trong việc phát hiện xem tham số tổng thể *lớn hơn* một giá trị p_0 nào đó: Trong trường hợp này, giả thuyết thay thế được viết là $p > p_0$.

Trong khi chúng ta điều chỉnh dạng của giả thuyết thay thế, chúng ta vẫn tiếp tục viết giả thuyết không thay đổi bằng dấu bằng trong trường hợp kiểm định một phía.

Trong toàn bộ quá trình kiểm định giả thuyết, chỉ có một điểm khác biệt giữa đánh giá kiểm định một phía so với kiểm định hai phía là cách tính giá trị p . Trong kiểm định một phía, giá trị p được tính bằng diện tích của phần đuôi *liên quan đến hướng của giả thuyết thay thế*, nghĩa là chỉ tính một bên của phân phối. Chính vì lý do này, kiểm định một phía đôi khi rất hấp dẫn: nếu không cần phải nhân đôi diện tích đuôi để có được giá trị p , thì giá trị p sẽ nhỏ hơn và mức độ chứng cứ cần thiết để xác định một phát hiện thú vị theo hướng của giả thuyết thay thế cũng giảm xuống. Tuy nhiên, kiểm định một phía không phải lúc nào cũng tốt đẹp: cái giá phải trả là bất kỳ phát hiện nào thú vị ở hướng ngược lại đều phải bỏ qua.

VÍ DỤ 5.37

Trong Mục 1.1, chúng ta đã gặp một ví dụ trong đó các bác sĩ quan tâm đến việc xác định xem các stent có giúp ích cho những người có nguy cơ đột quỵ cao hay không. Các nhà nghiên cứu tin rằng stent sẽ giúp ích. Đáng tiếc, dữ liệu lại cho thấy điều ngược lại: những bệnh nhân nhận stent thực tế lại tệ hơn. Tại sao việc sử dụng kiểm định hai phía lại quan trọng trong bối cảnh này?

E

Trước khi thực hiện nghiên cứu, các nhà nghiên cứu có lý do để tin rằng stent sẽ giúp ích cho bệnh nhân vì các nghiên cứu hiện có cho thấy stent giúp đỡ những người gặp cơn đau tim. Chắc chắn sẽ rất hấp dẫn khi sử dụng kiểm định một phía trong tình huống này, nhưng nếu làm như vậy, họ sẽ hạn chế khả năng phát hiện những tác hại tiềm tàng đối với bệnh nhân.

Ví dụ 5.37 nhấn mạnh rằng việc sử dụng giả thuyết một phía có thể tạo ra rủi ro bỏ qua dữ liệu ủng hộ kết luận ngược lại. Chúng ta có thể đã mắc sai lầm tương tự khi xem xét dữ liệu câu hỏi của Rosling trong phần này; nếu như chúng ta đã có ý niệm từ trước rằng những người có trình độ đại học sẽ không làm kém hơn so với đoán ngẫu nhiên và do đó sử dụng kiểm định một phía, chúng ta đã bỏ lỡ phát hiện thực sự thú vị rằng nhiều người có kiến thức sai lệch về sức khỏe cộng đồng toàn cầu.

Trong hoàn cảnh nào thì việc sử dụng kiểm định một phía có thể phù hợp? *Rất ít khi*. Nếu bạn bao giờ đó cân nhắc sử dụng kiểm định một phía, hãy trả lời cẩn thận câu hỏi sau:

Điều gì sẽ xảy ra nếu dữ liệu rõ ràng đi ngược lại hướng của giả thuyết thay thế của tôi hoặc của người khác?

Nếu bạn hoặc người khác sẽ thấy có giá trị khi đưa ra kết luận về dữ liệu theo hướng ngược lại với kiểm định một phía, thì nên sử dụng kiểm định hai phía. Những suy nghĩ này có thể rất tinh vi, vì vậy hãy cẩn trọng. Chúng ta sẽ chỉ áp dụng các kiểm định hai phía trong phần còn lại của cuốn sách này.

5.3 KIỂM ĐỊNH GIẢ THUYẾT CHO MỘT TỶ LỆ

VÍ DỤ 5.38

Tại sao chúng ta không thể đơn giản thực hiện một kiểm định một phía theo hướng của dữ liệu?

Chúng ta đã xây dựng một khung phương pháp cẩn thận để kiểm soát lỗi loại 1, đó là mức ý nghĩa α trong kiểm định giả thuyết. Trong ví dụ này, chúng ta sẽ dùng $\alpha = 0.05$ để đơn giản hoá mọi thứ.

Hãy tưởng tượng rằng chúng ta có thể chọn kiểm định một phía sau khi nhìn thấy dữ liệu. Điều gì sẽ xảy ra sai sót?

E

- Nếu p nhỏ hơn giá trị không, thì kiểm định một phía với giả thuyết $p < p_0$ sẽ đồng nghĩa với việc bất kỳ quan sát nào rơi vào đuôi nhỏ hơn 5% của phân phối không đều khiến chúng ta bác bỏ H_0 .
- Nếu p lớn hơn giá trị không, thì kiểm định một phía với giả thuyết $p > p_0$ sẽ đồng nghĩa với việc bất kỳ quan sát nào rơi vào đuôi lớn hơn 5% của phân phối không đều khiến chúng ta bác bỏ H_0 .

Vì vậy, nếu giả thuyết H_0 đúng, thì có khoảng 10% khả năng dữ liệu rơi vào một trong hai đuôi, dẫn đến lỗi kiểm định của chúng ta thực sự là $\alpha = 0,10$, chứ không phải 0,05. Điều này có nghĩa là, nếu không cẩn thận trong việc quyết định khi nào dùng kiểm định một phía, chúng ta đang vô tình làm suy yếu các phương pháp mà chúng ta đã cố gắng xây dựng và áp dụng rất nhiều.

5.3 KIỂM ĐỊNH GIẢ THUYẾT CHO MỘT TỶ LỆ

Bài tập

5.15 Xác định giả thuyết, Phần I. Viết giả thuyết không và giả thuyết thay thế bằng lời nói và sau đó là ký hiệu cho từng tình huống dưới đây.

- Một công ty gia sư muốn hiểu xem phần lớn học sinh có xu hướng cải thiện điểm số của họ (hoặc không) sau khi sử dụng dịch vụ của họ. Họ lấy mẫu 200 học sinh đã sử dụng dịch vụ trong năm qua và hỏi xem điểm số của họ đã tăng hay giảm so với năm trước.
- Các nhà tuyển dụng tại một công ty lo lắng về tác động của March Madness, giải vô địch bóng rổ được tổ chức vào mỗi mùa xuân ở Mỹ, đối với năng suất lao động của nhân viên. Họ ước tính rằng trong một ngày làm việc bình thường, nhân viên trung bình dành khoảng 15 phút của thời gian làm việc để kiểm tra email cá nhân, gọi điện thoại cá nhân, v.v. Họ cũng thu thập dữ liệu về thời gian làm việc của nhân viên dành cho các hoạt động không liên quan đến công việc trong mùa March Madness. Họ muốn xác định xem các dữ liệu này có cung cấp bằng chứng thuyết phục rằng năng suất của nhân viên đã thay đổi trong thời gian March Madness hay không.

5.16 Xác định giả thuyết, Phần II. Viết giả thuyết không và giả thuyết thay thế bằng lời nói và ký hiệu cho từng tình huống dưới đây.

- Kể từ năm 2008, các chuỗi nhà hàng tại California đã được yêu cầu hiển thị số calorie của từng món trong menu. Trước khi các menu bắt đầu hiển thị lượng calorie, trung bình lượng calorie tiêu thụ của khách hàng tại một nhà hàng là 1100 calorie. Sau khi bắt đầu hiển thị số calorie trên menu, một chuyên gia dinh dưỡng đã thu thập dữ liệu về số calorie tiêu thụ tại nhà hàng này từ một mẫu ngẫu nhiên các khách hàng. Những dữ liệu này có cung cấp bằng chứng thuyết phục về sự khác biệt trong trung bình lượng calorie tiêu thụ của khách hàng tại nhà hàng này hay không?
- Tiểu bang Wisconsin muốn hiểu tỷ lệ người trưởng thành của bang đã tiêu thụ rượu trong năm vừa qua, cụ thể là xem tỷ lệ này có khác so với tỷ lệ quốc gia là 70% hay không. Để giúp họ trả lời câu hỏi này, họ tiến hành lấy mẫu ngẫu nhiên 852 cư dân và hỏi về việc tiêu thụ rượu của họ.

5.17 Giao tiếp trực tuyến. Một nghiên cứu cho thấy 60% sinh viên đại học dành 10 hoặc nhiều hơn 10 giờ mỗi tuần để giao tiếp với người khác trực tuyến. Bạn tin rằng điều này không chính xác và quyết định thu thập mẫu của riêng mình để tiến hành kiểm định giả thuyết. Bạn lấy ngẫu nhiên 160 sinh viên trong ký túc xá của mình và phát hiện rằng 70% trong số đó dành 10 giờ hoặc nhiều hơn mỗi tuần để giao tiếp trực tuyến với người khác. Một người bạn của bạn, người đề nghị giúp bạn trong việc kiểm định giả thuyết, đã đề xuất bộ giả thuyết sau đây. Hãy chỉ ra bất kỳ lỗi nào bạn thấy:

$$H_0: \hat{p} < 0,6$$

$$H_A: \hat{p} > 0,7$$

5.18 Kết hôn ở tuổi 25. Một nghiên cứu cho thấy 25% những người 25 tuổi đã kết hôn. Bạn cho rằng điều này không chính xác và quyết định thu thập mẫu của riêng mình để thực hiện kiểm định giả thuyết. Từ một mẫu ngẫu nhiên các người 25 tuổi trong dữ liệu điều tra dân số với kích thước 776, bạn phát hiện rằng 24% trong số họ đã kết hôn. Một người bạn của bạn đề nghị giúp bạn xây dựng kiểm định giả thuyết và đưa ra bộ giả thuyết sau. Hãy chỉ ra bất kỳ lỗi nào bạn thấy:

$$H_0: \hat{p} = 0,24$$

$$H_A: \hat{p} \neq 0,24$$

5.19 Tỷ lệ bắt nạt qua mạng. Thanh thiếu niên đã được khảo sát về việc bị bắt nạt qua mạng, và khoảng tin cậy 95% cho biết từ 54% đến 64% báo cáo đã trải qua việc bị bắt nạt qua mạng.²³ Hãy trả lời các câu hỏi sau dựa trên khoảng tin cậy này.

- Một tờ báo tuyên bố rằng phần lớn thanh thiếu niên đã từng trải qua việc bị bắt nạt qua mạng. Tuy nhiên, tuyên bố này có được hỗ trợ bởi khoảng tin cậy không? Giải thích lý do của bạn.
- Một nhà nghiên cứu giả định rằng 70% thanh thiếu niên đã từng trải qua việc bị bắt nạt qua mạng. Tuyên bố này có được hỗ trợ bởi khoảng tin cậy không? Giải thích lý do của bạn.
- Mà không cần tính toán chính xác khoảng tin cậy, hãy xác định xem tuyên bố của nhà nghiên cứu ở phần (b) có được hỗ trợ dựa trên khoảng tin cậy 90% hay không.

²³ Pew Research Center, A Majority of Teens Have Experienced Some Form of Cyberbullying. September 27, 2018.

5.3 KIỂM ĐỊNH GIẢ THUYẾT CHO MỘT TỶ LỆ

5.20 Chờ đợi tại phòng cấp cứu, phần II. Bài tập 5.11 cung cấp khoảng tin cậy 95% cho trung bình thời gian chờ tại phòng cấp cứu (ER) là (128 phút, 147 phút). Hãy trả lời các câu hỏi sau dựa trên khoảng này.

- Một tờ báo địa phương khẳng định rằng thời gian chờ trung bình tại ER này vượt quá 3 giờ. Liệu tuyên bố này có được hỗ trợ bởi khoảng tin cậy không? Giải thích lý do của bạn.
- Hiệu trưởng của Khoa Y tại bệnh viện này khẳng định rằng thời gian chờ trung bình là 2,2 giờ. Liệu tuyên bố này có được hỗ trợ bởi khoảng tin cậy không? Giải thích lý do của bạn.
- Không cần tính toán chính xác khoảng, hãy xác định xem tuyên bố của hiệu trưởng ở phần (b) có thể được hỗ trợ dựa trên khoảng tin cậy 99% hay không?

5.21 Mức lương tối thiểu, Phần I. Phần lớn người trưởng thành tại Mỹ có tin rằng việc tăng mức lương tối thiểu sẽ giúp nền kinh tế hay có phần lớn người không tin điều này? Một cuộc khảo sát của Rasmussen Reports trên mẫu ngẫu nhiên gồm 1.000 người trưởng thành tại Mỹ cho thấy 42% tin rằng việc này sẽ giúp nền kinh tế. Thực hiện một kiểm định giả thuyết phù hợp để trả lời câu hỏi nghiên cứu này.

5.22 Thiếu ngủ. 400 sinh viên được lấy mẫu ngẫu nhiên từ một trường đại học lớn, và có 289 sinh viên cho biết họ không ngủ đủ giấc. Thực hiện một kiểm định giả thuyết để kiểm tra xem liệu tỷ lệ này có khác đáng kể về mặt thống kê so với 50% hay không, với mức độ ý nghĩa là 0,01.

5.23 Làm ngược, Phần I. Bạn được cho các giả thuyết sau:

$$H_0 : p = 0.3$$

$$H_A : p \neq 0.3$$

Chúng ta biết kích thước mẫu là 90. Với tỷ lệ mẫu nào thì giá trị p sẽ bằng 0,05? Giả sử rằng tất cả các điều kiện cần thiết cho suy luận đều được đáp ứng.

5.24 Làm ngược, Phần II. Bạn được cho các giả thuyết sau:

$$H_0 : p = 0.9$$

$$H_A : p \neq 0.9$$

Chúng ta biết kích thước mẫu là 1429. Với tỷ lệ mẫu nào thì giá trị p sẽ bằng 0,01? Giả sử rằng tất cả các điều kiện cần thiết cho suy luận đều được đáp ứng.

5.25 Kiểm tra Fibromyalgia. Một bệnh nhân tên Diana đã được chẩn đoán mắc hội chứng Fibromyalgia, một hội chứng kéo dài gây đau cơ thể, và đã được kê đơn thuốc chống trầm cảm. Với tư cách là người hay hoài nghi, Diana ban đầu không tin rằng thuốc chống trầm cảm sẽ giúp giảm các triệu chứng của cô ấy. Tuy nhiên, sau vài tháng dùng thuốc, cô quyết định rằng thuốc có tác dụng, vì cô cảm thấy các triệu chứng của mình thực sự đang cải thiện.

- Viết các giả thuyết bằng lời cho vị trí hoài nghi của Diana khi cô bắt đầu dùng thuốc chống trầm cảm.
- Sai lầm loại 1 (Type 1 Error) trong hoàn cảnh này là gì?
- Sai lầm loại 2 (Type 2 Error) trong hoàn cảnh này là gì?

5.26 Cái nào cao hơn? Trong mỗi phần dưới đây, có một giá trị quan tâm và hai kịch bản (I và II). Đối với từng phần, hãy báo cáo xem giá trị quan tâm lớn hơn trong kịch bản nào, hoặc nếu giá trị này bằng nhau trong hai kịch bản.

- Sai số chuẩn của $\hat{p}$ khi (I) $n = 125$ hoặc (II) $n = 500$.
- Khoảng sai số của một khoảng tin cậy khi mức độ tin cậy là (I) 90% hoặc (II) 80%.
- Giá trị p (p-value) cho thống kê Z là 2,5 dựa trên (I) mẫu có $n = 500$ hoặc dựa trên (II) mẫu có $n = 1000$.
- Xác suất mắc sai lầm loại 2 khi giả thuyết thay thế đúng và mức ý nghĩa là (I) 0,05 hoặc (II) 0,10.

²⁴ Rasmussen Reports survey, Most Favor Minimum Wage of \$10.50 Or Higher, April 16, 2019.

5.3 KIỂM ĐỊNH GIẢ THUYẾT CHO MỘT TỶ LỆ

Bài tập chương

5.27 Nghỉ ngơi sau giờ làm việc. Cuộc khảo sát Xã hội Tổng quát (General Social Survey) đã hỏi câu hỏi: “Sau một ngày làm việc trung bình, bạn có khoảng bao nhiêu giờ để thư giãn hoặc tham gia các hoạt động mà bạn thích?” trên một mẫu ngẫu nhiên gồm 1155 người Mỹ. Một khoảng tin cậy 95% cho trung bình số giờ dành cho hoạt động thư giãn hoặc thích là (1,38 – 1,92).

- Diễn giải khoảng này trong bối cảnh dữ liệu.
- Giả sử một nhóm nhà nghiên cứu khác báo cáo một khoảng tin cậy với biên độ sai số lớn hơn dựa trên cùng mẫu gồm 1155 người Mỹ. Mức độ tin cậy của họ so sánh như thế nào với mức độ tin cậy của khoảng đề cập ở trên?
- Giả sử năm sau, một cuộc khảo sát mới hỏi cùng câu hỏi này và lần này kích thước mẫu là 2500. Giả định rằng đặc điểm của tổng thể, liên quan đến thời gian mọi người dành để thư giãn sau giờ làm việc, không thay đổi nhiều trong vòng một năm. Biên độ sai số của khoảng tin cậy 95% được xây dựng dựa trên dữ liệu của cuộc khảo sát mới này sẽ như thế nào so với biên độ sai số của khoảng đã nêu ở trên?

5.28 Mức lương tối thiểu, phần II. Trong bài tập 5.21, chúng ta biết rằng một cuộc khảo sát của Rasmussen Reports gồm 1000 người Mỹ trưởng thành cho thấy 42% tin rằng việc tăng lương tối thiểu sẽ giúp nền kinh tế. Hãy xây dựng một khoảng tin cậy 99% cho tỷ lệ thực sự của các người trưởng thành ở Mỹ tin vào điều này.

5.29 Kiểm tra an toàn thực phẩm. Một kiểm tra viên an toàn thực phẩm được cử đến điều tra một nhà hàng có vài báo cáo của khách hàng về các thực hành vệ sinh kém. Kiểm tra viên sử dụng khung kiểm định giả thuyết để đánh giá xem liệu nhà hàng có không đáp ứng quy định hay không. Nếu quyết định nhà hàng vi phạm nghiêm trọng, giấy phép phục vụ thức ăn của họ sẽ bị thu hồi.

Viết các giả thuyết bằng lời.

- Sai lầm loại 1 (Type 1 Error) trong hoàn cảnh này là gì?
- Sai lầm loại 2 (Type 2 Error) trong hoàn cảnh này là gì?
- Loại lỗi nào nghiêm trọng hơn đối với chủ nhà hàng? Tại sao?
- Loại lỗi nào nghiêm trọng hơn đối với thực khách? Tại sao?
- Là một thực khách, bạn sẽ thích kiểm tra viên yêu cầu bằng chứng rõ ràng hoặc bằng chứng rất rõ ràng trước khi thu hồi giấy phép của nhà hàng? Giải thích lý do.

5.30 Đúng hay sai. Xác định xem các phát biểu dưới đây là đúng hay sai, và giải thích lý do của bạn. Nếu sai, hãy chỉ rõ cách sửa chữa.

- Nếu một giá trị đã cho (ví dụ, giá trị giả thuyết trắng của một tham số) nằm trong khoảng tin cậy 95%, thì nó cũng sẽ nằm trong khoảng tin cậy 99%.
- Giảm mức ý nghĩa (α) sẽ làm tăng xác suất mắc sai lầm loại 1.
- Giả sử giả thuyết vô hiệu là $p = 0.5$ và chúng ta không bác bỏ H_0 . Trong trường hợp này, tỷ lệ dân số thực là 0.5.
- Với kích thước mẫu lớn, ngay cả những khác biệt nhỏ giữa giá trị giả thuyết vô hiệu và ước lượng điểm quan sát, còn gọi là kích thước hiệu ứng, cũng sẽ được xác định là có ý nghĩa thống kê.

5.31 Thất nghiệp và các vấn đề về các mối quan hệ. Một cuộc khảo sát của USA Today/Gallup hỏi một nhóm người Mỹ thất nghiệp hoặc làm việc không đủ giờ nếu họ gặp các vấn đề lớn trong mối quan hệ với vợ/chồng hoặc thành viên gia đình thân thiết khác do không có việc làm (nếu thất nghiệp) hoặc không có công việc toàn thời gian (nếu làm việc không đủ giờ). 27% trong số 1.145 người thất nghiệp và 25% trong số 675 người làm việc không đủ giờ cho biết họ gặp các vấn đề lớn trong các mối quan hệ do tình trạng công việc của họ.

- Các giả thuyết để đánh giá liệu tỷ lệ người thất nghiệp và người làm việc không đủ giờ gặp vấn đề về quan hệ có khác nhau không là gì?
- Giá trị p của phép kiểm định giả thuyết này xấp xỉ bằng 0.35. Hãy giải thích ý nghĩa của điều này trong bối cảnh của phép kiểm định giả thuyết và dữ liệu.

²⁵ National Opinion Research Center, General Social Survey, 2018.

5.3 KIỂM ĐỊNH GIẢ THUYẾT CHO MỘT TỶ LỆ

5.33 Cận thị. Người ta tin rằng khoảng 8% tất cả trẻ em bị cận thị. Trong một mẫu ngẫu nhiên gồm 194 trẻ em, có 21 trẻ bị cận thị. Tiến hành một phép kiểm định giả thuyết để trả lời câu hỏi sau: dữ liệu này có cung cấp bằng chứng cho thấy giá trị 8% là không chính xác không?

5.34 Nhãn dinh dưỡng. Nhãn dinh dưỡng trên một túi khoai tây chiên nói rằng mỗi khẩu phần một ounce (28 gram) có 130 calo và chứa mười gam chất béo, trong đó có ba gam chất béo bão hòa. Một mẫu ngẫu nhiên gồm 35 túi cho ra khoảng tin cậy cho số lượng calo trong mỗi túi từ 128,2 đến 139,8 calo. Có bằng chứng nào cho thấy nhãn dinh dưỡng không cung cấp số liệu chính xác về lượng calo trong các túi khoai tây chiên không?

5.35 Trung bình theo luật lớn (CLT) cho tỷ lệ. Định nghĩa thuật ngữ “phân phối mẫu” của tỷ lệ mẫu, và mô tả cách hình dạng, trung tâm, và độ rộng của phân phối mẫu thay đổi khi kích thước mẫu tăng lên trong trường hợp $p = 0,1$.

5.36 Ý nghĩa thực tiễn so với ý nghĩa thống kê. ác định xem phát biểu sau đúng hay sai, và giải thích lý do của bạn: “Với kích thước mẫu lớn, ngay cả những chênh lệch nhỏ giữa giá trị giả thuyết vô hiệu và ước lượng điểm quan sát cũng có thể có ý nghĩa thống kê.”

5.37 Cùng một quan sát, nhưng khác về kích thước mẫu. Giả sử bạn tiến hành một phép kiểm định giả thuyết dựa trên một mẫu có kích thước $n = 50$, và đạt được p-value là 0,08. Sau đó, bạn xem lại ghi chú và phát hiện ra rằng bạn đã mắc sai lầm cầu thả, kích thước mẫu thực tế phải là $n = 500$. P-value của bạn sẽ tăng, giảm, hay vẫn giữ nguyên? Giải thích.

5.38 Khoảng cách giới tính về lương trong ngành y. Một nghiên cứu đã xem xét mức lương trung bình của nam và nữ bắt đầu làm việc y tế ở 21 vị trí khác nhau.²⁶

- Nếu mỗi giới đều được trả công bằng nhau, thì chúng ta sẽ mong đợi khoảng một nửa số vị trí có nam được trả nhiều hơn nữ và nữ được trả nhiều hơn nam ở nửa còn lại. Viết các giả thuyết phù hợp để kiểm tra tình huống này.
- Trong 19 trong số 21 vị trí, nam trung bình được trả nhiều hơn. Giả sử 21 vị trí này là một mẫu ngẫu nhiên đơn giản, hoàn thành phép kiểm định giả thuyết dựa trên các giả thuyết của phần (a).

²⁶ Lo Sasso AT et al. “The \$16,819 Pay Gap For Newly Trained Physicians: The Unexplained Trend Of Men Earning More Than Women”. In: *Health Affairs* 30.2 (2011).